Apport des outils de la biologie moléculaire pour l'utilisation des diatomées comme bioindicateurs de la qualité des écosystèmes aquatiques lotiques et pour l'étude de leur taxonomie

Lenaïg Kermarrec

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : Biodiversité, écologie, environnement.

Arrêté ministériel : 7 août 2006

Présentée par

Lenaïg KERMARREC

Thèse dirigée par **Agnès BOUCHEZ** et
codirigée par **Jean-François HUMBERT**

préparée au sein du **Laboratoire INRA UMR-CARRTEL**
dans **l'École Doctorale SISEO**

Apport des outils de la biologie moléculaire pour l'utilisation des diatomées comme bioindicateurs de la qualité des écosystèmes aquatiques lotiques et pour l'étude de leur taxonomie

Thèse soutenue publiquement le **04 mai 2012**,
devant le jury composé de :

Mr François ENAULT	MC, CNRS UMR 6023, Clermont-Ferrand, Examinateur
Mr Marc-Henri LEBRUN	DR, INRA UMR BIOGER CPP, Versailles, Rapporteur
Mr Jan PAWLOWSKI	PR, Université de Genève, Genève, Examinateur
Mme Florence PERES	CP, Asconit Consultants, Boulogne-sur-Gesse, Examinateur
Mr Thomas POMMIER	CR, INRA UMR 5557, Lyon, Examinateur
Mr Koen SABBE	PR, Université de Ghent, Ghent, Rapporteur


ASCONIT
CONSULTANTS
INRA

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : Biodiversité, écologie, environnement.

Arrêté ministériel : 7 août 2006

Présentée par

Lenaïg KERMARREC

Thèse dirigée par **Agnès BOUCHEZ** et
codirigée par **Jean-François HUMBERT**

préparée au sein du **Laboratoire INRA UMR-CARRTEL**
dans **l'École Doctorale SISEO**

Apport des outils de la biologie moléculaire pour l'utilisation des diatomées comme bioindicateurs de la qualité des écosystèmes aquatiques lotiques et pour l'étude de leur taxonomie

Thèse soutenue publiquement le **04 mai 2012**,
devant le jury composé de :

Mr François ENAULT	MC, CNRS UMR 6023, Clermont-Ferrand, Examinateur
Mr Marc-Henri LEBRUN	DR, INRA UMR BIOGER CPP, Versailles, Rapporteur
Mr Jan PAWLOWSKI	PR, Université de Genève, Genève, Examinateur
Mme Florence PERES	CP, Asconit Consultants, Boulogne-sur-Gesse, Examinateur
Mr Thomas POMMIER	CR, INRA UMR 5557, Lyon, Examinateur
Mr Koen SABBE	PR, Université de Ghent, Ghent, Rapporteur

La Directive Cadre Européenne sur l'eau impose d'évaluer la qualité des cours d'eau au moyen d'indicateurs chimiques et biologiques dont les diatomées font partie. Les indices basés sur la composition taxonomique et l'abondance relative des taxa de diatomées sont robustes. Cependant, de nombreux échantillons doivent être analysés chaque année alors que l'identification de ces micro-algues en microscopie optique est difficile à cause des incertitudes taxonomiques, et nécessite temps et expertise. Ainsi, des améliorations peuvent encore être apportées pour faciliter le suivi en routine de la qualité de l'eau.

Les techniques de biologie moléculaire sont des outils efficaces pour identifier les microorganismes et pourraient donc être utilisées pour améliorer l'identification des diatomées. Les objectifs de cette thèse étaient donc de compléter les connaissances sur la taxonomie des diatomées d'eau douce par des méthodes moléculaires et de progresser dans le développement d'un outil moléculaire permettant l'identification des diatomées dans des échantillons naturels, en vue de son utilisation en bioindication.

L'étude de la taxonomie de plusieurs groupes de diatomées a été réalisée en combinant des approches morphologiques et des approches moléculaires. Nos travaux ont montré les capacités des séquences ADN pour discriminer les taxa de diatomées et révéler leurs relations phylogénétiques. L'utilisation de séquences ADN a montré que les critères morphologiques utilisés pour identifier les diatomées ne correspondaient pas systématiquement à leurs relations phylogénétiques. L'utilisation de différents marqueurs a permis des discriminations à différents niveaux taxonomiques. Nos résultats ont également révélé l'importance de combiner des approches complémentaires, morphologiques et moléculaires, pour améliorer notre compréhension des relations entre les différents taxa de diatomées et ainsi stabiliser leur taxonomie.

Les séquences ADN permettant une discrimination des taxa de diatomées, nous avons testé un outil moléculaire de séquençage haut-débit, le pyroséquençage 454, dans le but d'identifier les taxa composant les communautés de diatomées. Nous avons ainsi assemblé des bases de séquences de référence bénéficiant d'une identification taxonomique. Nous avons également participé au développement d'outils bioinformatiques nécessaires à l'analyse des données de pyroséquençage. Enfin, nous avons pu tester ces outils pour établir des inventaires taxonomiques de diatomées dans des communautés artificielles (mélanges de souches) et dans des communautés environnementales (biofilms d'eau douce). Ces études ont prouvé le potentiel du pyroséquençage 454 pour étudier les communautés de diatomées à des niveaux taxonomiques précis. La comparaison de différents marqueurs nucléiques a révélé que le marqueur *rbc*L était le marqueur le plus adapté à l'identification des diatomées par pyroséquençage. En effet, en prenant en compte les bases de séquences de référence, la reproductibilité et les biais de la méthode ainsi que le pouvoir résolutif du marqueur, l'utilisation du *rbc*L a permis la meilleure estimation de la composition en diatomées d'échantillons complexes.

Des progrès devront encore être faits avant de pouvoir utiliser les outils moléculaires pour évaluer la qualité de l'eau par les diatomées. Cependant nos différentes études permettront de guider les prochaines analyses de manière à aboutir à un suivi de la qualité de l'eau basé sur des inventaires moléculaires des taxa de diatomées.

The European Water Framework Directive requires assessing the river quality using chemical and biological indicators among which are diatoms. Indices based on taxonomic composition and relative abundance of diatom taxa are robust. However, thousands of diatom samples are analyzed every year while microscopic identification is difficult due to taxonomic uncertainties, and requires time and expertise. Thus, it is still possible to improve the routine monitoring of water quality.

The molecular biology techniques are accurate tools to identify microorganisms and could be used to enhance diatom identification. The objectives of this thesis were therefore to improve our knowledge on the freshwater diatom taxonomy by molecular methods and to progress in the development of a molecular tool in order to identify diatoms in natural samples, thus improving bioindication tools.

The taxonomic study of several groups of diatoms was achieved by combining morphological and molecular approaches. Our results showed the capacity of DNA sequences to discriminate diatom taxa and revealed their phylogenetic relationships. The use of DNA sequences showed that the morphological criteria used to identify diatoms do not correspond systematically to their phylogenetic relationships. The use of different markers allowed discrimination at different taxonomic levels. Our results also revealed the importance of combining complementary approaches, morphological and molecular, to improve our understanding of the relationships between different diatom taxa and thus stabilize their taxonomy.

As DNA sequences allowed discrimination of diatom taxa, we tested a molecular tool of high throughput sequencing, 454 pyrosequencing, in order to identify the diatom composition of communities. We assembled reference libraries of sequences linked to taxonomic identification and we contributed to the development of bioinformatic tools required to analyze data from pyrosequencing. Finally, we tested these tools to establish taxonomic inventories of diatoms in artificial communities (mixtures of strains) and environmental communities (freshwater biofilm samples). The studies showed the potential of 454 pyrosequencing to accurately analyze diatom communities. The comparison of different nucleic markers revealed that the *rbc*L marker was the most suitable to identify diatoms using pyrosequencing. Indeed, taking into account reference libraries, reproducibility and bias of the method, and the resolving power of marker, the use of *rbc*L allowed the best estimation of the diatom composition in complex samples.

Improvements will be necessary to use molecular tools in order to assess water quality using diatoms. However our studies lead the way for future experiments in order to achieve a monitoring of water quality based on molecular inventories of diatom taxa.

En préambule à ce manuscrit, je souhaite adresser tous mes remerciements aux personnes qui, de près ou de loin, m'ont apporté leur aide et qui ont ainsi contribué à la concrétisation de ce travail de thèse. Certaines personnes seront peut-être oubliées, c'est pourquoi, je remercie par avance ceux dont le nom n'apparaît pas dans ces pages.

Je tiens à remercier en premier lieu mes deux directeurs de thèse ***Agnès Bouchez*** *et* ***Jean-François Humbert*** *pour m'avoir fait confiance et pour avoir mis à ma disposition tous les moyens nécessaires afin que cette thèse puisse se réaliser dans de bonnes conditions. Je les remercie également pour les conseils et le soutien qu'ils m'ont apportés.*

Je remercie également ***Florence Pérès***, ***Philippe Blancher***, ***Olivier Monnier*** *et* ***Serge Rochepeau*** *pour m'avoir offert l'opportunité de réaliser cette thèse et maintenant de rejoindre une agence d'****Asconit Consultants****. Etre rattachée à une agence tout en préparant une thèse à plus de 600km n'est pas toujours simple, je remercie donc tous mes collègues d'Asconit qui, d'une manière ou d'une autre, malgré mon isolement géographique, m'ont aidée à participer à la vie de l'entreprise.*

Mes remerciements vont ensuite à ***Jean Marcel Dorioz***, ***Jean Guillard***, *et* ***Bernard Montuelle*** *pour leur accueil au sein de la station d'hydrobiologie lacustre de Thonon-les-Bains.*

Je remercie également ***Koen Sabbe*** *et* ***Marc-Henri Lebrun*** *qui ont très gentiment accepté d'être les rapporteurs de ce travail de thèse ainsi que* ***François Enault***, ***Jan Pawlowski***, ***Thomas Pommier*** *et (à nouveau)* ***Florence Pérès*** *d'avoir accepté de juger mon travail.*

Les conseils reçus par les membres de mon comité de pilotage m'ont beaucoup apporté. Je tiens donc à remercier ***Rosa Trobajo***, ***Michel Coste***, *et* ***David Mann***.

J'adresse également toute ma gratitude et toute ma reconnaissance à ***Alain Franc*** *et* ***Philippe Chaumeil*** *qui ont toujours trouvé un moment dans leur planning chargé pour répondre à mes questions et qui m'ont fait découvrir un nouvel univers, celui de la bioinformatique !*

Merci également à ***Frédéric Rimet*** *pour son expertise morphologique des diatomées.*

Que soit remerciée aussi toute l'équipe de la plateforme Genotoul qui m'a chaleureusement accueillie pour faire le pyroséquençage. C'était une semaine très riche, j'y ai appris beaucoup.

Cette thèse n'aurait pu se faire sans les échantillons environnementaux initiaux. Je remercie donc également toutes les personnes qui m'ont envoyé des échantillons : ***Maurice Bey***, ***Michel Coste***, ***Gilles Gassioles***, ***Didier Guillard*** *et* ***René Le Cohu****. J'en profite également pour remercier* ***Luc Ector*** *pour m'avoir transmis les cultures que j'avais isolées, sous sa direction, au* ***C.R.P. Gabriel Lippmann****.*

Mes remerciements s'adressent également à ceux qui tour à tour ont été mes collègues de bureau : ***Ludwig***, ***Lyria***, ***Thomas***, ***Kevin*** *et* ***Sophie****.*

Je tiens évidemment à remercier toutes les personnes, permanentes ou de passage, que j'ai côtoyées dans le labo de biologie moléculaire.

Les longues heures passées au laboratoire créent des liens, j'en viens donc à remercier les personnes qui sont devenues des amis. Ma plus sincère gratitude va tout particulièrement à ***Isabelle*** *qui a partagé toutes ces longues journées d'extraction de l'algothèque. Merci pour ton aide, pour toutes les petites attentions qui m'ont remonté le moral pendant la rédaction, et surtout pour l'intérêt que tu as porté à mon travail ! Le laboratoire de biologie moléculaire et l'algothèque ne seraient sûrement pas aussi agréables sans la présence de* ***Cécile*** *alors, Cécile, je te remercie pour tous les bons moments passés et pour ton soutien.* ***Noémie***, ***Clément****, j'ai adoré partager nos nombreux problèmes de manips et nos discussions « bio mol » et « bioinfo » ! Ce fut un réel plaisir de partager ces véritables « casse-tête ». Enfin, je tiens à remercier* ***Séverine*** *pour sa bonne humeur, son accueil chaleureux et sa capacité à nous faire sortir du contexte du travail. Heureusement que tu étais là pour t'occuper de nous !*

Comment ne pas remercier ceux qui ont partagé les joies et les colères au quotidien (la semaine et le week-end !). ***Lyria***, ***Thomas, Lidwine, Benjamin*** *et* ***Ludwig****, vous rencontrer a été le bon côté de cette thèse. Vous avez été une véritable bouffée d'oxygène. Il est impensable que j'oublie les bons moments partagés à Thonon ou ailleurs ! Petite pensée particulière pour* ***Benji*** *: je t'abandonne…. à ton tour maintenant !*

Ces remerciements ne seraient pas complets sans les personnes qui sont présentes depuis toujours. Tous mes remerciements vont bien évidemment à « mes copines » : ***Julie, Kristen***, ***Mathilde et Morgane*** *et tous les « Crozonnais » :* ***Adeline***, ***Jean-Marc***, ***Alain*** *et*

J-B *qui sont toujours présents malgré le peu de temps que je leur ai accordé au cours de ces trois années. Je vous adore.*

Je remercie aussi ***ma « Maman » et ma sœur, Nolwenn,*** *pour leur soutien et leurs encouragements mais aussi pour nous avoir régulièrement amené un petit bout de notre Bretagne natale.*

Enfin, ***Julien****, je te remercie tout d'abord pour avoir, à nouveau, accepté de quitter notre « Bout du Monde ». Je n'aurai pas tenu sans ton soutien. Tu mérites une médaille pour avoir accepté de mettre notre vie entre parenthèses pendant ces trois longues années. Maintenant nous allons enfin pouvoir profiter, un petit peu quand même !*

SOMMAIRE

LISTE DES FIGURES ..... **I**

LISTE DES TABLEAUX ..... **V**

GLOSSAIRE ..... **VII**

LISTE DES ABREVIATIONS ..... **IX**

CONTEXTE **1**

CHAPITRE I. SYNTHESE BIBLIOGRAPHIQUE **7**

1 LES DIATOMEES 9

1.1. Les cellules de diatomées *9*

1.2. Classification et diversité *13*

1.3. Cycle de vie *17*

1.3.1. Multiplication végétative 17

1.3.2. Reproduction sexuée 19

1.3.3. Formes de résistances 21

1.4. Mode de vie *21*

1.5. Ecologie *24*

2. LA BIOINDICATION 26

2.1. L'évaluation de la qualité des eaux continentales *26*

2.1.1. Mesures physico-chimiques 26

2.1.2. Indicateurs biologiques 26

2.2. La Directive Cadre Européenne sur l'eau *27*

2.3. La bioindication des rivières par les diatomées *30*

2.3.1. Avantages des diatomées 30

2.3.2. Indices basés sur les diatomées 30

2.4. Limites des indices diatomées *32*

3. APPROCHES MOLECULAIRES 35

3.1. Phylogénies moléculaires des diatomées *35*

3.2. Le concept de « Barcoding DNA » *37*

3.3. Barcoding environnemental *40*

CHAPITRE II. MATERIEL ET METHODES **43**

1. ECHANTILLONS 45

1.1. Sites d'échantillonnage *45*

1.2. Prélèvements in situ *47*

1.3. Cultures de souches de diatomées 47
2. L'APPROCHE MORPHOLOGIQUE 48
2.1. Préparation des frustules 48
2.2. Préparation de lames permanentes 48
2.3. Morphométrie traditionnelle 49
2.4. Morphométrie géométrique 49
3. L'APPROCHE MOLECULAIRE 50
3.1. Séquençage des souches 50
3.1.1. Marqueurs 51
3.1.2. Séquençage et analyse des séquences 55
3.1.3. Bases de référence 57
3.2. Analyse par pyroséquençage 454 59
3.2.1. Principe 59
3.2.2. Echantillons de communautés de diatomées 61
3.2.3. Mise au point du protocole 65
3.2.4. Algorithmes metaMatch 67

CHAPITRE III. TAXONOMIE DES DIATOMEES & APPROCHES MOLECULAIRES 71

1. INTRODUCTION 73
2. ETUDE DE LA TAXONOMIE DE L'ORDRE DES CYMBELLALES. 74
2.1. Article I : Présentation générale de l'étude et synthèse des principaux résultats. 74
2.2. Article I publié dans Diatom Research 26 (3) : 305-315 77
3. ETUDES DE COMPLEXES D'ESPECES 96
3.1. Gomphonema parvulum 96
3.1.1. Article II : Présentation générale de l'étude et synthèse des principaux résultats. 96
3.1.2. Article II : Soumis après révision dans Protist 101
3.2. Nitzschia palea 129
4. CONCLUSION 140

CHAPITRE IV. NOUVELLE APPROCHE DE BIOINDICATION PAR LES DIATOMEES 141

1. INTRODUCTION 143
2. COMPARAISON D'ALGORITHMES POUR L'ASSIGNATION DE READS ISSUS DE NGS 147
2.1. Présentation générale de l'étude et synthèse des principaux résultats. 147
2.2. Article III : en préparation pour la revue BMC Bioinformatics 150
3. COMPARAISON DE MARQUEURS MOLECULAIRES POUR REALISER DES INVENTAIRES D'ESPECES A PARTIR DES NGS : APPROCHE SUR COMMUNAUTES ARTIFICIELLES 167
3.1. Présentation générale de l'étude et synthèse des principaux résultats. 167
3.2. Article IV : soumis dans la revue PLoS ONE 170
4. TEST DE LA NOUVELLE APPROCHE DE BIOINDICATION BASEE SUR LES NGS SUR DES ECHANTILLONS NATURELS DE DIATOMEES 191

4.1. Présentation générale de l'étude et synthèse des principaux résultats. 191

4.2. Article V : en préparation 196

5. CONCLUSION 229

CHAPITRE V. DISCUSSION GENERALE 231

1. TAXONOMIE 233
2. UTILISATIONS DES METHODES MOLECULAIRES 238
3. BIOINDICATION MOLECULAIRE 239
4. VERS UN NOUVEL OUTIL MOLECULAIRE ? 243

BIBLIOGRAPHIE 247

ANNEXE 1 : LISTE DES CULTURES 273

ANNEXE 2 : PROTOCOLES 285

PROTOCOLES D'EXTRACTION 287

PROTOCOLES D'AMPLIFICATION DES MARQUEURS 289

Pour les souches de diatomées 289

Pour les échantillons environnementaux 291

ANNEXE 3 : SEMINAIRES, COLLOQUES NATIONAUX ET INTERNATIONAUX 293

PRÉSENTATION DE POSTERS: 295

COMMUNICATIONS 296

LISTE DES FIGURES

Figure I.1 : *Schéma des caractéristiques générales d'une diatomée pennée*.....8
Figure I.2: *Organisation des parois cellulaires siliceuses des diatomées*. 10
Figure I.3: *Exemple de diatomées centriques (a) et pennées (b).* 12
Figure I.4*: Exemple de valves tératologiques de* Fragilaria *(a) et de* Gomphonema *(b)*. 14
Figure I.5 : *Cycle cellulaire des diatomées.*.....16
Figure I.6: *Division végétative: la règle de MacDonald–Pfitzer*. 16
Figure I.7: *Schéma du cycle de vie des diatomées centriques*.....18
Figure I.8: *Reproduction sexuée chez des diatomées pennées raphidées en culture*.....18
Figure I.9: *Exemples de colonies de diatomées.*.....20
Figure I.10: Gomphonema bourbonense *à l'extrémité de son pédoncule*.....22
Figure I.11: *Carte des hydroécorégions de niveau 1 de France métropolitaine*.....28
Figure I.12: *Arbre phylogénétique des diatomées centriques basé sur les séquences des gènes 18S,* rbc*L et* psb*C.* 34
Figure I.13: *Arbre phylogénétique des diatomées pennées basé sur les séquences des gènes 18S,* rbc*L et* psb*C.* 36
Figure I.14: *Principe du barcoding*. 38
Figure II.1: *Schéma général de la méthodologie*.....44
Figure II.2: *Structure de l'opéron ribosomal des eucaryotes*.....51
Figure II.3: *Principe du séquençage Sanger*. 56
Figure II.4: *Préparation des échantillons pour le pyroséquençage*. 58
Figure II.5: *Appareil de pyroséquençage*.....58
Figure II.6: *Représentation schématique du système enzymatique du pyroséquençage*. 60
Figure II.7: *Rendements d'extraction moyens des 4 méthodes testées*. 66
Figure III.1: *Phylogeny of the Cymbellales inferred from maximum likelihood analysis of 18S rDNA gene sequences using all taxa studied including those from GenBank*.....84
Figure III.2: *Some species of Cymbellales characterizing clades 1 and 2*. 86
Figure III.3: *Some species of Cymbellales characterizing clades 3 and 4*. 88
Figure III.4: *Simplified maximum likelihood phylogenetic tree related to the main morphological data of the branches*:.....90
Figure III.5: *Arbre phylogénétique obtenu par analyse en Maximum Likelihood des séquences de trois marqueurs : ITS,* rbc*L and* cox*1*. 98

Figure III.6 : *Maximum Likelihood tree based on ITS sequences, illustrating the variation of* Gomphonema parvulum *complex sequences.* 106
Figure III.7 : *Maximum Likelihood tree based on* rbc*L sequences, illustrating the variation of* Gomphonema parvulum *complex sequences.* 107
Figure III.8 : *Cleaned valves of* Gomphonema *cf.* lagenula *(1),* G. exilissimum *(2), and* G. parvulum *(3-30), arranged according to the five clades.*..... 108
Figure III.9 : *Valve dimension and density of striae of* Gomphonema parvulum *valves*. 110
Figure III.10: *Mean landmark configurations at different fixed lengths (10, 20, and 30 µm) with an example of the central area focus for each of the five clades.*..... 112
Figure III.11: *Canonical variate analysis using geometric morphometric data for the five clades of* Gomphonema parvulum. 114
Figure III.12 : *Light micrograph showing the position of the 25 landmarks on the pole and the central area used in the geometric morphometric analysis.* 124
Figure III.13 : *Maximum Likelihood tree based on the LSU D1/D2 region, illustrating the phylogeny of* Gomphonema *sequences.* 127
Figure III.14: *Arbre phylogénétique obtenu par analyse des séquences d'ADNr 28S par Maximum Likelihood*. 135
Figure III.15: *Arbre phylogénétique obtenu par analyse des séquences de* rbc*L par Maximum Likelihood*..... 137
Figure III.16: *Arbre phylogénétique obtenu par analyse des séquences de* cox*1 par Maximum Likelihood*..... 139
Figure IV.1: *Procédure d'analyse d'échantillons environnementaux utilisant les nouvelles techniques de séquençage*..... 144
Figure IV.2: *Dendrogam of hierarchical aggregative clustering on pairwise Jaccard distances between 16 inventories*. 160
Figure IV.3: *Shannon index plot showing the amount of variability of each marker.* 174
Figure IV.4: *Proportion of informative reads at different taxonomic levels: from clade to subdivision.* 176
Figure IV.5: *Comparison of the species inventories made for each bulk sample to the real inventory.*..... 178
Figure IV.6: *Comparison of the species inventories made for each marker to the real inventory.*..... 178
Figure IV.7 : *Distribution des distances génétiques calculées en comparant toutes les séquences de chaque base de référence deux à deux (ADNr 18S,* rbc*L et* cox*1) en fonction de*

trois niveaux taxonomiques (intraspécifique, interspécifique / intragénérique, et intergénérique) 193
Figure IV.8: *Box plot of intraspecific, interspecific/intrageneric and intergeneric genetic idendity based on uncorrected p-distances*. 202
Figure IV.9: *Similarities of inventory with actual composition of the mock community and number of informative reads used to compile inventories depending on different identity thresholds for each marker and each taxonomic level (A: species; B: genus)*. 204
Figure IV.10: *Hierarchical clustering analysis (UPGMA) of the samples based on the phylogenetic position of the SSU rDNA and* rbc*L reads*. 208
Figure IV.11: *Non metric Multidimensional Scaling based on Bray-Curtis distances comparing molecular and morphological species inventories of the four samples*. 208
Figure IV.12: *Non metric Multidimensional Scaling based on Bray-Curtis distances comparing molecular and morphological genus inventories of the four samples.* 228

LISTE DES TABLEAUX

Tableau I.1: *Qualités d'eau associées aux notes d'IBD*. 32
Tableau II.1: *Liste des sites échantillonnés.* 46
Tableau II.2: *Liste des amorces utilisées pour amplifier et séquencer les souches de diatomées*..... 54
Tableau II.3: *Composition des échantillons artificiels.* 62
Tableau II.4 : *Description des sites d'échantillonnage des échantillons environnementaux* ... 64
Table III.1: *List of Cymbellales species and strains sequenced in this study, with sampling location, date and their GenBank accession numbers*. 80
Table III.2 : *GenBank accession numbers of Cymbellales and outgroup used in this study*. 82
Table III.3: *p-distances between sequences of the five clades of* Gomphonema parvulum. . 104
Table III.4 : *List of strains sequenced in this study and their GenBank accession numbers.* .. 120
Table III.5 : *Genes and primers used for PCR and sequencing, and the number of positions used after alignment and corrections (including indels)*. 122
Table III.6 : *Accession numbers of the* Gomphonema *sequences downloaded from Genbank.* 126
Tableau III.7 :*Liste des souches de* Nitzschia palea *utilisées pour cette étude*..... 130
Tableau III.8: *Liste des souches de* Nitzschia palea *dont l'ADN a été fourni par R. Trobajo*. . 132
Table IV.1: *Search time, number of perfect matches and informative reads for each library and each method.* 156
Table IV.2: *Number of species in each inventory, number of species detected and present in the mixed samples, and number of false positives at genera level (species belonging to genera not present in the mixed samples).* 158
Table IV.3: *Total number of reads, number of perfect matches and number of species informative reads for each pyrosequencing library (Mix C, PCR1 Mix and PCR2 Mix) and each marker (SSU rDNA,* rbc*L, and* cox*1)*. 176
Table IV.4: *Taxonomic inventories obtained by 454 pyrosequencing of the three bulk samples (mix C, mix PCR1, mix PCR2)*. 180
Table IV.5: *List and proportions (%) of strains used to create the three bulk samples (mix C, mix PCR1, mix PCR2).* 186
Table IV.6: *Environmental DNA samples used for the study and number of reads obtained after removing low-quality sequences*. 200

Table IV.7: *List and proportion of strains used to create mock community.* 200
Table IV.8: *P-values of the P-test comparing phylogenetic information of each pair of sample for both markers (SSU rDNA and* rbc*L).* 206
Table IV.9: *Summary of inventory results*..... 206
Table IV.10: *Comparison of molecular inventories with morphological inventory for each sample*. 210

GLOSSAIRE

Allogamie : Mode de reproduction sexuée dans lequel la fécondation d'un gamète se fait par le gamète d'un autre spécimen, contrairement au mode autogame où les gamètes d'un même spécimen s'autoféconden.

Allométrie : Changements de forme (non linéaire) liés à la taille d'un spécimen.

Anisogamie : Mode de reproduction sexuée qui implique deux gamètes morphologiquement différents.

Apogamie : Mode de reproduction qui n'implique pas de méiose.

Autogamie : Mode de reproduction sexuée où les deux gamètes sont issus du même spécimen.

Cellules Janus : Cellules de diatomées présentant, sur le même frustule, deux valves morphologiquement différentes.

Gamétocyste : Cellule contenant les gamètes.

GenBank : Base de données mettant à disposition les séquences nucléiques et protéiques.

Groupe monophylétique : Groupe de taxa incluant tous les descendants d'un même ancêtre commun.

Groupe paraphylétique : Groupe de taxa qui descendent d'un ancêtre commun mais qui ne comprend pas l'ensemble des descendants de cet ancêtre commun.

Isogamie : Mode de reproduction sexuée qui implique deux gamètes morphologiquement identiques.

Oogamie : Mode de reproduction sexuée qui implique un gamète femelle immobile et un gamète mâle petit et mobile.

Populations allopatriques : Populations appartenant à une même espèce mais séparées l'une de l'autre géographiquement.

Sympatrie : Caractéristique de taxa qui occupent des aires géographiques chevauchantes.

Taille centroïde : Paramètre utilisé en morphométrie géométrique qui correspond à la racine carrée de la somme des distances au carré entre les points de repères et le centre de gravité du spécimen.

LISTE DES ABREVIATIONS

ADN/DNA : Acide DésoxyriboNucléique

ADNr/rDNA : ADN codant pour l'ARN ribosomal

ARN/RNA : Acide RiboNucléique

ARNr/rRNA : ARN ribosomal

ATP : Adénosine TriPhosphate

CBC : Changement de base compensatoire,
Compensatory Base Change

***Cox*1** : Gène codant pour la grande sous unité 1 de la Cytochrome OXydase

DCE : Directive Cadre Européenne

ddNTP : didésoxyriboNucléotides TriPhosphate

DGGE : Electrophorèse sur gel en gradient dénaturant,
Denaturing Gradient Gel Electrophoresis

dNTP : désoxyriboNucléotide TriPhosphate

DOM : Département d'Outre-Mer

DREAL : Directions Régionales de l'Environnement, de l'Aménagement et du Logement

HER : Hydro-EcoRégions

IBD : Indice Biologique Diatomées

IPS : Indice de PolluoSensibilité spécifique

ITS : Internal Transcribed Spacer

NGS : Séquençage nouvelle génération,
Next Generation Sequencing

OTU : Unité taxonomique opérationnelle,
Operational Taxonomic Unit

PCR :	Réaction de polymérisation en chaîne, Polymerase Chain Reaction
***rbc*L** :	Gène codant pour la grande sous-unité de la ribulose-1,5-bisphophate carboxylase.
TCC :	Thonon Culture Collection
UE :	Union Européenne

CONTEXTE

En raison de l'impact croissant des activités humaines, les ressources en eau douce sont aujourd'hui au centre des préoccupations environnementales. L'évaluation de la qualité des masses d'eau est nécessaire à l'utilisation durable de cette ressource, que ce soit pour la consommation (humaine ou animale), l'aquaculture ou les loisirs. Ainsi, l'Union Européenne (UE) a établi un cadre législatif pour la protection des eaux (de surface, souterraines, de transition et côtières) : la Directive Cadre Européenne (DCE) sur l'eau. Des programmes de surveillance ont donc été développés pour évaluer la qualité des masses d'eau. Les objectifs de cette directive sont la prévention et l'amélioration de l'état des écosystèmes aquatiques, la promotion d'une utilisation durable de l'eau, la réduction de la pollution, et l'atténuation des effets des inondations et des sécheresses. Un objectif de «bon état écologique» des eaux de l'UE a été fixé à l'horizon de 2015, sachant qu'il est peu probable que cet objectif soit atteint à cette échéance.

La DCE prévoit d'évaluer l'état chimique des masses d'eau mais également leur état écologique sur la base d'indicateurs biologiques. Ces indicateurs reposent sur l'étude des organismes vivant dans l'écosystème à évaluer. L'observation des communautés biologiques est un élément fondamental d'évaluation de la qualité du milieu. Guelorget & Perthuisot (1984) définissent les bioindicateurs comme « des espèces ou groupes d'espèces, qui par leur présence et/ou leur abondance, sont significatifs d'une ou plusieurs propriétés de l'écosystème dont ils font partie ». Les masses d'eau perturbées peuvent engendrer des conditions défavorables pour la survie et/ou le développement de certains organismes et modifier la structure des communautés. Les organismes ne tolérant pas ces perturbations émigrent, ou disparaissent alors que d'autres organismes plus tolérants se développent. Les principaux intérêts des indicateurs biologiques sont qu'ils informent sur les conséquences des perturbations, par leurs effets directs ou indirects sur les communautés, et qu'ils sont des intégrateurs temporels des perturbations que subit un milieu, ce qui n'est pas le cas pour beaucoup de paramètres chimiques.

Les bioindicateurs utilisés dans le cadre de la DCE, sont les suivants : le phytobenthos, le phytoplancton, les macrophytes, les algues macroscopiques et les angiospermes, la faune benthique invertébrée, et l'ichtyofaune. Le choix de ces bioindicateurs dépend de la catégorie de masses d'eau : rivière, lac, estuaire, ou littoral. Parmi ces indicateurs, les diatomées benthiques (micro-algues du phytobenthos) sont, depuis plusieurs décennies, utilisées pour évaluer la qualité des eaux de rivières. Ces microorganismes ont un cycle de vie rapide et sont donc des bioindicateurs efficaces pour

intégrer des impacts qui ont lieu sur une courte période. De plus, en temps que producteurs primaires, les diatomées sont plus directement affectées par les modifications de disponibilité en nutriments de l'environnement aquatique. L'échantillonnage des diatomées est peu coûteux, facilement préservé et a un très faible impact sur l'environnement. Plusieurs indices de qualité du milieu basés sur les communautés de diatomées ont donc été développés (Kelly & Whitton, 1995; Lenoir & Coste, 1996; Lavoie et al., 2006). Ces indices évaluent l'état de ce compartiment biologique par la composition taxonomique et l'abondance des espèces de diatomées en lien avec leur écologie.

Bien que les indices utilisés actuellement soient robustes, leur application reste difficile car l'identification des diatomées en routine repose sur leur observation en microscopie optique. La détermination des diatomées au niveau de l'espèce n'est réalisable que par des experts en taxonomie ; cette détermination est longue et elle nécessite un degré élevé de compétences et d'expérience, ce qui rend ces analyses coûteuses. D'autre part, la taxonomie des diatomées n'est pas encore stabilisée puisque de nouvelles espèces sont encore décrites régulièrement et que d'autres le seront dans les années futures. Par ailleurs, les taxonomistes spécialistes des diatomées divergent souvent sur la validité et sur les critères d'identification de certaines espèces (Mann et al., 2010). Enfin, l'application de la DCE a engendré une forte augmentation du nombre d'échantillons de diatomées à analyser chaque année, alors que les techniques optiques basées sur l'identification par des experts ont une efficacité limitée. Pour toutes ces raisons, des efforts ont été déployés pour développer des systèmes d'identification nécessitant moins d'interventions humaines. Par exemple, un système d'identification automatique basé sur des critères de forme des diatomées (Jalba et al., 2005) a été testé mais aucun de ces systèmes n'a été jusqu'ici appliqué pour des analyses en routine (Mann et al., 2010).

Dans le même temps, les techniques de biologie moléculaire n'ont cessé d'évoluer depuis le développement, dans les années 80, de la Réaction de Polymérisation en Chaîne, PCR (Saiki et al., 1985, 1988; Mullis et al., 1986). Ainsi ces techniques sont devenues des outils puissants dans le domaine des sciences de la vie et elles ont permis d'obtenir des séquences de plusieurs gènes qui ont par exemple été utilisées pour reconstruire des phylogénies des procaryotes (e.g., Woese, 1987), des protistes (e.g., Baroin et al., 1988),

des plantes supérieures (e.g., Chase et al., 1993) et des animaux (e.g., Simon et al., 1994). En écologie, les techniques de biologie moléculaire ont également permis d'étudier les communautés naturelles procaryotes et eucaryotes par des approches d'empreintes moléculaires (Casamayor et al., 2002; Dorigo et al., 2006) ou de clonage-séquençage (Bérard et al., 2005; Humbert et al., 2009). Le développement constant des techniques de biologie moléculaire est particulièrement prometteur pour la compréhension du fonctionnement des écosystèmes aquatiques (Zinger et al., 2011).

Les outils de biologie moléculaire ont également été appliqués aux diatomées et ont ainsi considérablement amélioré les connaissances sur leur phylogénie (e.g. Medlin & Kaczmarska, 2004; Theriot et al., 2010). De plus, les techniques de biologie moléculaire ont permis de révéler la présence d'espèces cryptiques ou pseudo-cryptiques au sein de complexes d'espèces initialement définis sur la base de critères morphologiques (e.g. Sarno et al., 2005; Evans et al., 2008; Trobajo et al., 2009). Enfin, des techniques comme la PCR quantitative permettent aujourd'hui de détecter et quantifier les diatomées dans des échantillons environnementaux (Créach et al., 2006; Nguyen et al., 2011).

Constatant que les techniques de biologie moléculaire ont permis de réaliser de grandes avancées dans l'étude des microorganismes, y compris les diatomées, la question s'est posée de déterminer si ces approches pouvaient améliorer l'utilisation de ces bioindicateurs. Dans ce contexte général, plusieurs questions sont apparues. Les outils moléculaires peuvent-ils aider à la délimitation des taxa de diatomées ? Quel marqueur nucléique utiliser pour identifier précisément ces taxa ? Quelles technologies utiliser pour la détection des espèces de diatomées dans un échantillon naturel ?

Ainsi, les deux objectifs majeurs de cette thèse ont été :

(i) d'améliorer les connaissances sur la taxonomie de certains groupes de diatomées en combinant des approches moléculaires et morphologiques,

(ii) de développer une approche moléculaire permettant d'identifier les diatomées dans un échantillon environnemental. Dans ce but, il était nécessaire de :

   a. construire les outils moléculaires et informatiques nécessaires à cette approche (bases de séquences de référence, protocoles, outils bioinformatiques, etc.)

b. valider ces outils sur des échantillons synthétiques et environnementaux.

Ce manuscrit de thèse présente donc les différents travaux réalisés pour répondre à ces objectifs :

- Le Chapitre I consiste en une synthèse bibliographique composée de trois parties. Tout d'abord, une description générale de notre modèle d'étude, les diatomées, est présentée. Puis, leur utilisation en bioindication est développée. Enfin, les approches moléculaires sont exposées.

- Le Chapitre II décrit le matériel et les méthodes utilisés au cours de la thèse.

- Le Chapitre III aborde l'utilisation des approches moléculaires en complément des approches morphologiques pour l'étude de la taxonomie des diatomées. Une première partie étudie les relations phylogénétiques au sein de l'Ordre des Cymbellales (Article I) et une seconde étudie plus précisément deux complexes d'espèces : *Gomphonema parvulum* (Article II) et *Nitzschia palea*.

- Le Chapitre IV présente les étapes de développement de l'outil moléculaire pour l'identification des taxa de diatomées présents dans un échantillon complexe. Le premier article compare des algorithmes bioinformatiques (Article III). Le second présente les résultats d'une étude de différents marqueurs moléculaires et des biais liés à l'utilisation des outils moléculaires, sur des communautés synthétiques (Article IV). Le dernier article est dédié au test de l'outil moléculaire sur des échantillons naturels (Article V).

- Enfin, le manuscrit se termine par le Chapitre V qui propose une discussion des principaux résultats et la présentation des perspectives qui en découlent.

CHAPITRE I. SYNTHESE BIBLIOGRAPHIQUE

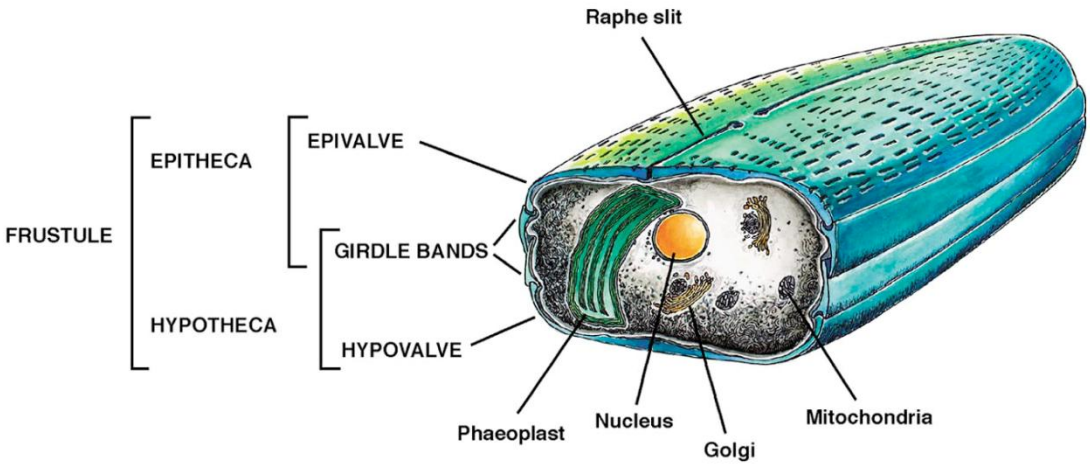


Figure I.1 : *Schéma des caractéristiques générales d'une diatomée pennée*.
Source : Falciatore & Bowler, 2002

1 Les diatomées

Les diatomées ou Bacillariophyta sont des microorganismes eucaryotes unicellulaires. Les connaissances générales acquises sur ces micro-algues sont liées au développement de la microscopie. A la fin du 17e siècle, plusieurs scientifiques européens ont commencé à utiliser la microscopie pour observer des échantillons naturels provenant d'habitats aquatiques. En 1703, une première représentation d'une espèce d'eau douce de *Tabellaria* (identifiée à posteriori) a été publiée par la Royal Society de Londres (Anonymous, 1703). Depuis, les connaissances concernant les diatomées n'ont cessé d'évoluer avec le développement des techniques de microscopie.

1.1. *Les cellules de diatomées*

Les diatomées sont des microorganismes photosynthétiques. La taille de ces algues unicellulaires varie de quelques µm à plus de 500 µm. Comme toutes les cellules eucaryotes, le contenu cellulaire comprend (Figure I.1) :

- un noyau délimité par une membrane et pouvant contenir plusieurs nucléoles;

- des mitochondries et des systèmes membranaires et tubulaires intracytoplasmiques tel que l'appareil de Golgi ;

- des vacuoles et des globules lipidiques (ou gouttelettes lipidiques) en nombre variable qui occupent la majeure partie de l'espace intracellulaire ;

- un ou plusieurs plastes délimités par une enveloppe composée de 4 membranes. La couleur des plastes peut varier du jaune très pâle au brun en raison de la composition en pigments caroténoïdiens (β-carotènes, diatoxanthin, diadinoxanthine et fucoxanthine) qui masquent la couleur des pigments chlorophylliens *a* et *c* (c2, et c1 ou c3) (Stauber & Jeffrey, 1988). Leur nombre et leur forme varient selon les espèces mais restent assez constants au sein de la plupart des espèces. Les chloroplastes peuvent avoir la forme de petits disques, essentiellement chez les espèces centriques et chez quelques pennées ; ou la forme de plus grandes plaques, dans la majorité des diatomées pennées (Round et al., 1990). Ces plastes épousent la forme de la cellule et peuvent contenir des pyrénoïdes (Duke & Reimann, 1977). Les pyrénoïdes sont des globules protéiques dans lesquels se déroulent la voie métabolique de fixation du carbone. La taille et la forme des pyrénoïdes varient en fonction des taxa (Cox, 2011).

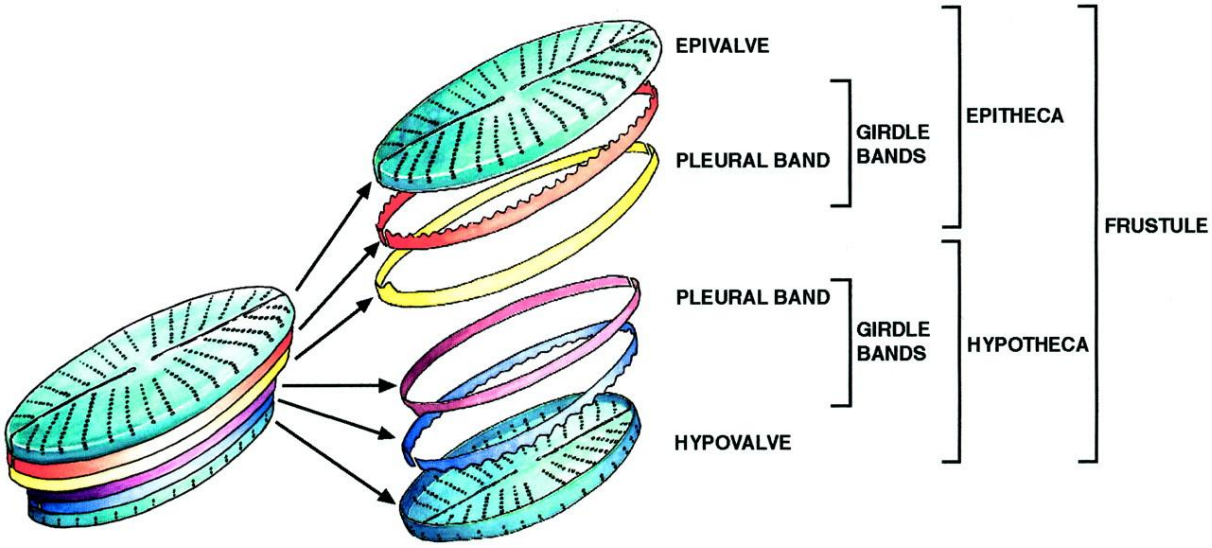


Figure I.2: *Organisation des parois cellulaires siliceuses des diatomées*.
Source : Zurzolo & Bowler, 2001.

La caractéristique principale des diatomées réside dans la présence autour du contenu cellulaire, d'une enveloppe de nature siliceuse dénommée frustule (Figure I.1 et Figure I.2). La paroi cellulaire est une structure commune aux cellules végétales mais la particularité des diatomées est que ce squelette extracellulaire est composé de silice ($SiO_2.H_2O$) et de matériel organique (Duke & Reimann, 1977; Round et al., 1990). Cette coque est formée essentiellement de 2 parties appelées thèques qui s'assemblent comme le fond et le couvercle d'une boîte par l'intermédiaire de ceintures connectives (aussi appelées cingulum) constituées de fines bandes siliceuses dénommées *copulae*. Les thèques sont donc composées de 2 parties, la valve et le cingulum (Figure I.2). Les deux valves sont de tailles différentes. La grande valve est l'épivalve qui, associée à l'épicingulum, forme l'épithèque, tandis que la petite valve est l'hypovalve qui, associée à l'hypocingulum, forme l'hypothèque (Figure I.2).

L'ornementation des deux valves se caractérise par la présence de stries, côtes, cloisons, ponctuations, fentes, soies ou autres protubérances qui associées à la forme générale de l'individu aident à la détermination des taxa. En effet, ces ornementations ne sont pas disposées au hasard mais selon un motif bien défini et elles permettent à la cellule d'échanger avec le milieu (Germain, 1981). Certaines diatomées pennées présentent une fente étroite, le raphé, sur une ou deux valves. Les diatomées sécrètent du mucilage, au travers de ces ouvertures, sous la forme de polysaccharides impliqués dans l'adhésion aux substrats, dans la mobilité des cellules (pour celles qui possèdent un raphé), dans la flottaison et dans la formation des colonies (Round et al., 1990). Cette matrice organique protège également le frustule contre toute forme de dissolution liée aux conditions physico-chimiques du milieu. D'autres substances telles que des toxines peuvent également être produites par quelques espèces de diatomées marines, (par exemple *Pseudo-nitzschia multiseries* et *P. pseudodelicatissima*), et provoquer des troubles digestifs et neurologiques.

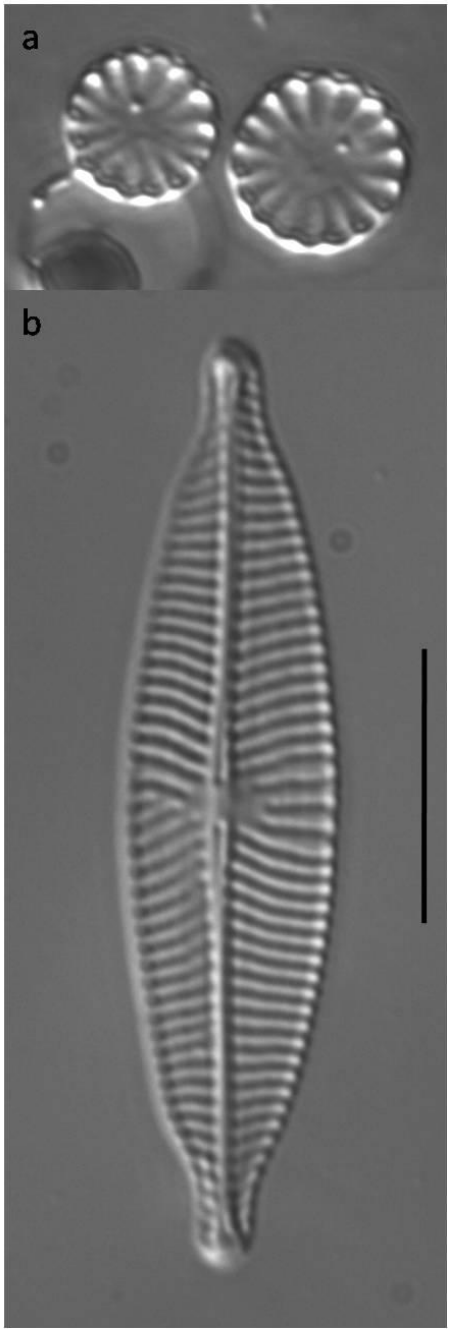


Figure I.3: *Exemple de diatomées centriques (a) et pennées (b).*
Barre d'échelle = 10µm

1.2. *Classification et diversité*

Empire : Eukaryota

Kingdom : Chromista

Infrakingdom : Heterokonta

Phylum : Bacillariophyta

La classification au sein des Bacillariophyta est essentiellement basée sur les caractéristiques morphologiques du frustule. Les diatomées sont traditionnellement classées en deux groupes en fonction de la forme de la cellule : les diatomées centriques ont une symétrie radiale alors que les diatomées pennées ont une symétrie bilatérale (Figure I.3). Dans le groupe des diatomées centriques, deux groupes sont définis :

- les centriques « radiales » ont une forme circulaire ;

- les centriques bi- ou multipolaires présentent des structures qui font apparaître d'autres symétries que la symétrie radiale.

Au sein des pennées, les espèces sont différenciées en fonction de la présence et du nombre de raphés :

- les araphidées ne présentent aucun raphé ;

- les monoraphidées possèdent un raphé sur une valve, l'autre valve en est dépourvu mais peut présenter une ligne médiane appelée pseudo-raphé ;

- les biraphidées possèdent un raphé sur chaque valve.

Bien que la classification des diatomées dépende dans une grande mesure de la morphologie des valves, d'autres critères, caractéristiques de la cellule vivante (par exemple, le nombre et la forme des chloroplastes et des pyrénoïdes), ont également été utilisés (Cox & Williams, 2000, 2006). En ce qui concerne la morphologie des valves, la forme, le type de symétrie, la présence et la forme du raphé, l'agencement et la densité des ornementations sont autant d'éléments qui servent à l'identification taxonomique des diatomées. La détermination des diatomées au niveau de l'espèce peut dépendre de variations très subtiles (Mann et al., 2010), telles que l'écartement des fibules centrales ou de légères variations du contour des pôles. Ces déterminations ne sont donc réalisables qu'en microscopie et par des experts en taxonomie.

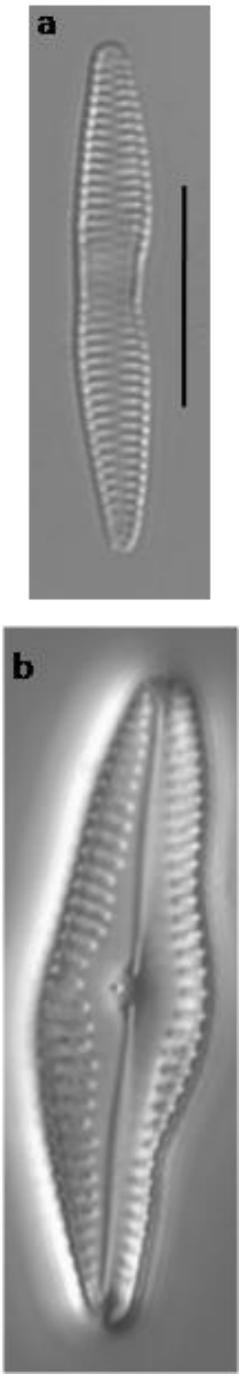


Figure I.4*: Exemple de valves tératologiques de* Fragilaria *(a) et de* Gomphonema *(b). Barre d'échelle = 10µm*

Les caractéristiques du frustule sont très constantes dans une même espèce, mais des changements se produisent au cours du cycle de vie des diatomées (Cox, 2010). Cependant, les diatomistes considèrent généralement que la division cellulaire se déroule avec un haut degré de fidélité dans la morphologie des valves et la variabilité intraspécifique a donc été très peu étudiée. Pourtant cette variabilité a des implications importantes pour l'identification des diatomées au niveau des espèces et à des niveaux intraspécifiques.

D'après Kociolek & Stoermer (2010) les variations au sein des espèces de diatomées ont pour origine trois facteurs : le développement c'est-à-dire le cycle de vie, le polymorphisme génétique et les variables environnementales. Ces variations atteignent parfois des niveaux si importants que des cellules tératologiques (déformées) peuvent être observées (Figure I.4). Les variations morphologiques peuvent être continues, c'est-à-dire qu'un gradient de variations est observé, ou discontinues comme c'est le cas pour les cellules Janus qui possèdent deux valves présentant des morphologies très différentes. Tous ces facteurs doivent être considérés lorsque l'on tente d'identifier les diatomées. En raison de ces variabilités morphologiques, des niveaux infraspécifiques sont souvent définis au sein des espèces de diatomées.

Les diatomées forment ainsi un phylum très diversifié avec plus de 62 000 noms de taxa (à des niveaux spécifiques et infraspécifiques) répertoriés dans le « Catalogue of Diatom Names » (Fourtanier & Kociolek, 2011). Mais cette diversité semble encore sous-estimée. En effet, les complexes d'espèces sont nombreux au sein du phylum des diatomées. Ils sont composés de plusieurs espèces qui sont identiques morphologiquement (espèces cryptiques) ou difficilement reconnues sur la base de la morphologie (espèces pseudo-cryptiques) (Mann & Evans, 2007). Ces complexes d'espèces apparaissent dans les deux classes de diatomées : les centriques (par exemple *Cyclotella meneghiniana*, Beszteri et al., 2007), et les pennées (par exemple *Pseudo-nitzschia delicatissima* et *P. pseudodelicatissima*, Amato et al., 2007). En prenant en compte la présence de ces complexes d'espèces et les espèces restant à décrire, Mann & Droop (1996) estimaient le nombre d'espèces de diatomées à 200 000.

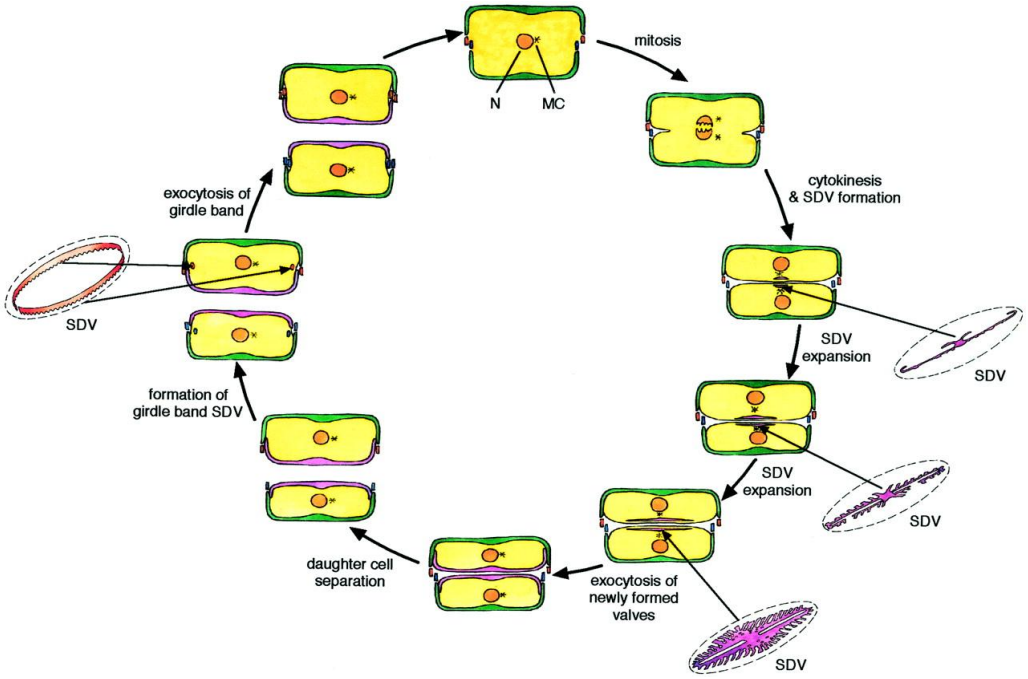


Figure I.5 : *Cycle cellulaire des diatomées.*
N : Noyau ; MC : Centre de microtubules ; SDV : vésicule de dépôt de silice.
Source : Zurzolo & Bowler, 2001

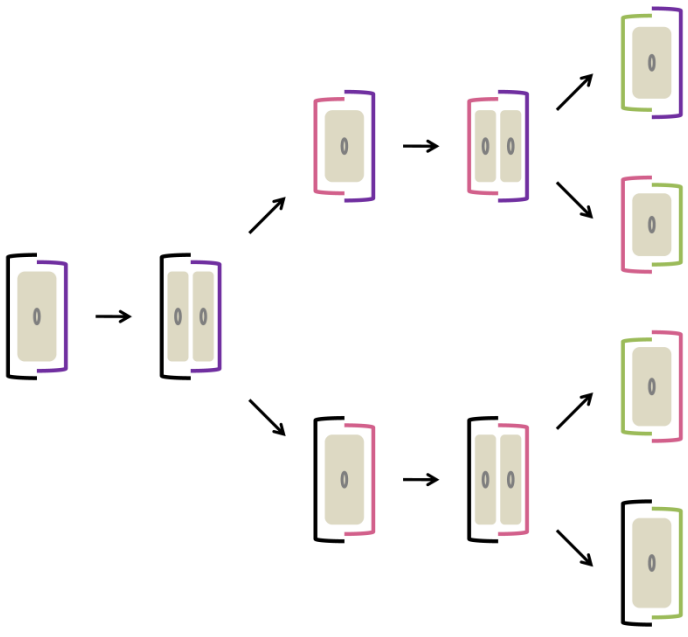


Figure I.6: *Division végétative: la règle de MacDonald–Pfitzer*.

1.3. *Cycle de vie*

Les diatomées sont des organismes diploïdes dont le nombre de chromosomes varient considérablement selon les espèces (de 4 à 52, Kociolek & Stoermer, 1989a). Le cycle de vie des diatomées est composé de deux phases correspondant à deux types de reproduction distincts : une phase de multiplication cellulaire végétative qui peut durer des mois à des années et une phase de reproduction sexuée qui dure quelques heures (Chepurnov et al., 2004). En outre, certaines diatomées peuvent développer des formes de résistances au cours de leur cycle de vie.

1.3.1. Multiplication végétative

Du fait de la structure des frustules, les cellules de diatomées ne peuvent grandir que dans une direction, en agrandissant leur profondeur, c'est-à-dire l'espace entre les deux valves. Ainsi, la croissance des diatomées se fait par ajout de nouvelles ceintures connectives à l'hypocingulum jusqu'à la division végétative (Round et al., 1990). Lorsque les conditions sont favorables au développement des diatomées, celles-ci se multiplient par bipartition après une mitose. La vitesse de multiplication peut être variable selon les espèces de quelques heures à quelques jours (Baars, 1983).

La particularité de la division cellulaire chez les diatomées est la synthèse d'une nouvelle valve au sein de la valve parentale (Figure I.5). Après la mitose, les 2 cellules filles se forment dans la cellule parentale. Chaque cellule fille synthétise une hypovalve avant la séparation des 2 cellules. Les principales étapes ont été schématisées par Zurzolo & Bowler (2001). Brièvement, une vésicule de dépôt de silice (SDV) se forme, puis se propage perpendiculairement jusqu'à former une grande vésicule dans laquelle se forme une nouvelle valve grâce au transport de silice, de protéines et de polysaccharides. Lorsque la formation de la valve est finie, celle-ci est exocytée par fusion des membranes de la vésicule de dépôt de silice et de la membrane plasmique. Enfin, après la séparation, les cellules filles peuvent à nouveau produire de nouvelles ceintures connectives, également grâce aux vésicules de dépôt de silice.

Le processus de formation de la nouvelle valve (à l'intérieur de la valve parentale) engendre une hypovalve plus petite que l'épivalve. La multiplication végétative implique donc une diminution de la taille moyenne de la population au cours des générations (Figure I.6).

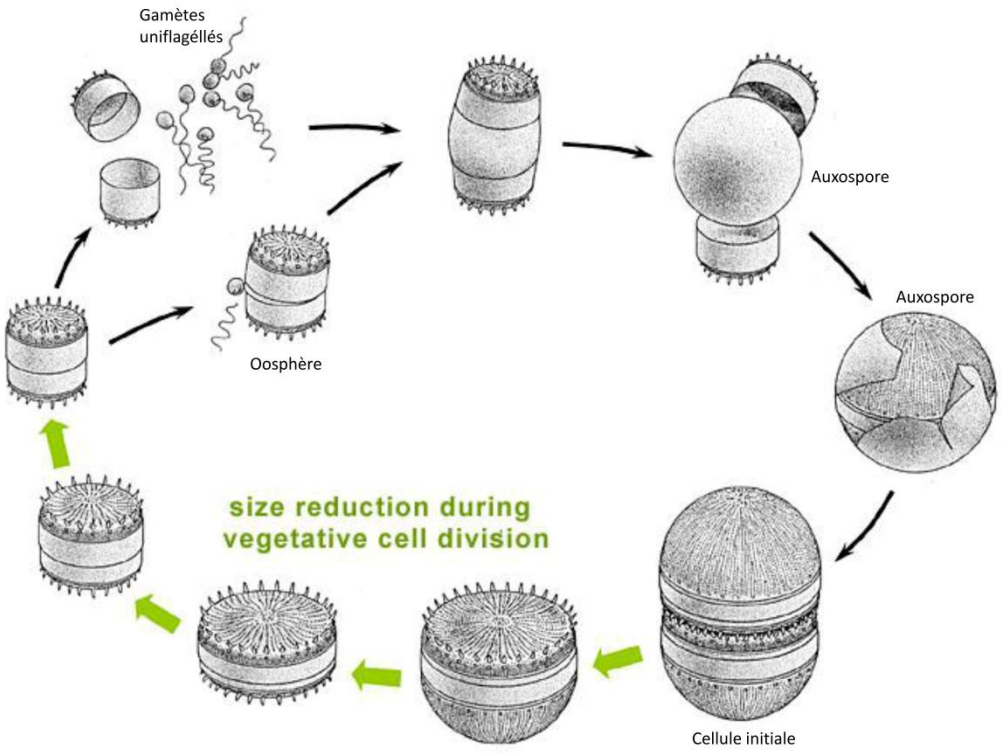


Figure I.7: *Schéma du cycle de vie des diatomées centriques*
Source: Round et al., 1990.

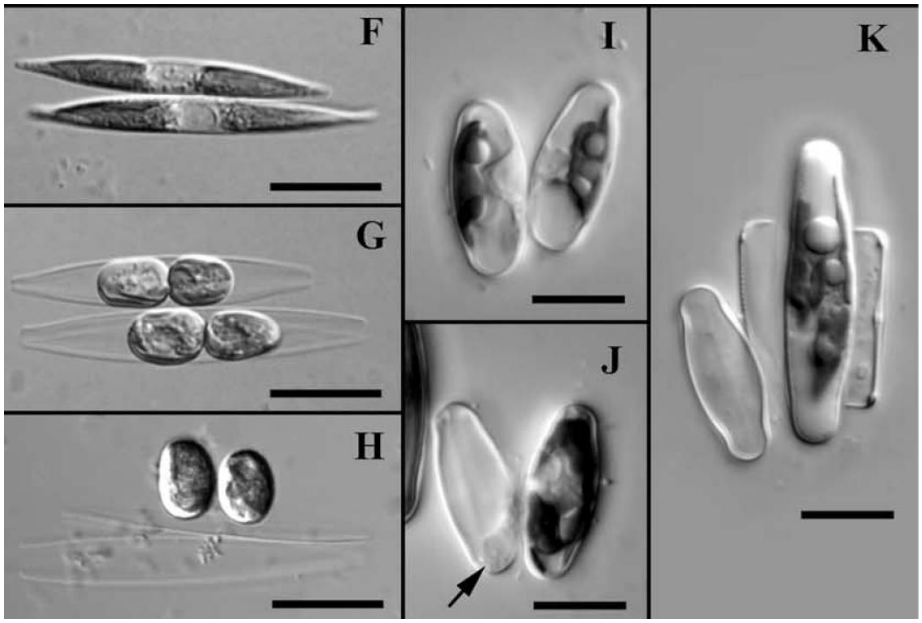


Figure I.8: *Reproduction sexuée chez des diatomées pennées raphidées en culture*.
Barre d'échelle F-H = 20µm, I-K = 10µm
Source : Chepurnov et al., 2004.

Ce principe est appelé la règle (ou l'hypothèse) de MacDonald–Pfitzer (Macdonald, 1869; Pfitzer, 1871).

La forme du frustule peut varier en fonction de la diminution de taille (Kociolek & Stoermer, 2010) alors que les ornementations (densité de stries, de pores) sont peu affectées par ces variations (Cox, 2010) et restent donc un élément important d'identification taxonomique.

1.3.2. Reproduction sexuée

La reproduction sexuée est encore assez mal connue chez les diatomées. La taille « normale » de l'espèce est rétablie par un phénomène sexuel et la production d'un zygote (l'auxospore). L'auxosporulation est donc une reproduction sexuée caractéristique des Bacillariophyta, qui implique une recombinaison génétique et une restauration de la taille de l'espèce. Lorsque les cellules atteignent une taille minimale (environ 30-40% de la taille maximale) et que les conditions environnementales sont appropriées, l'auxosporulation est déclenchée (Edlund & Stoermer, 1997; John, 2000). La taille maximale des cellules initiales ainsi que la taille minimale (en dessous de laquelle les cellules ne peuvent plus se reproduire sexuellement) sont des caractéristiques spécifiques des espèces définies par Geitler (1932) comme les points cardinaux du cycle cellulaire.

La plupart des diatomées sont allogames, mais les modes de reproduction sont différents selon les groupes de diatomées (centriques, pennées araphidées et pennées raphidées) et des exceptions existent dans les trois groupes.

Les diatomées centriques sont généralement anisogames et oogames impliquant des gamètes uniflagellés et des oosphères non mobiles (Figure I.7).

Les diatomées pennées araphidées, comme les diatomées centriques, sont généralement anisogames. De plus, une interaction entre les deux partenaires est nécessaire pour déclencher la méiose. Les deux types de gamètes ne diffèrent pas en taille mais une cellule produit un gamète « passif » alors que l'autre cellule produit un gamète « actif » non flagellé dont les mécanismes de locomotion sont inconnus (Chepurnov et al., 2004).

Les diatomées pennées raphidées (Figure I.8) semblent isogames car les gamétocystes apparaissent identiques morphologiquement (en forme et en taille) mais leur comportement et leur physiologie sont différents (Drebes, 1977; Chepurnov et al., 2004).

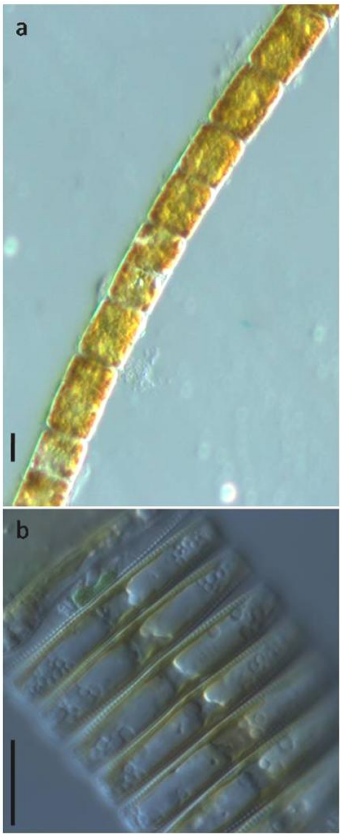


Figure I.9: *Exemples de colonies de diatomées.*
(a) colonie en forme de chaîne : *Melosira varians* ; (b) colonies en forme de ruban : *Fragilaria capucina*. Barre d'échelle = 10µm.

L'appariement de ces diatomées (qui sont mobiles) se fait de manière active puis différents modes de reproduction peuvent se produire. Il peut y avoir une reproduction de « type cis », où un gamétocyste produit deux gamètes actifs tandis que l'autre gamétocyste produit deux gamètes immobiles (Figure I.8 F-H), ou une reproduction de « type trans » où chaque cellule produit un gamète actif et un gamète immobile. Certaines espèces peuvent également ne produire qu'un gamète par gamétocyste (Figure I.8 I-K).

Mais, quel que soit le mode de reproduction, les diatomées forment une auxospore qui va croître et former une ou deux cellules initiales qui sont des cellules de taille maximale. Les cellules n'auront la forme caractéristique de l'espèce qu'après leur première division mitotique (Edlund & Stoermer, 1997; Kociolek, 2011). Certaines diatomées peuvent également produire des auxospores par autogamie, après une méiose par fusion des deux noyaux ou des deux gamètes d'un même gamétocyste, ou par apogamie, sans véritable méiose (Drebes, 1977; Round et al., 1990).

1.3.3. Formes de résistances

En réponse à des conditions environnementales défavorables (azote et phosphore limitant, luminosité et température faible), certaines diatomées peuvent entrer dans une « phase de repos » en formant des spores de résistance (ou hypnospores) ou des cellules de repos. Les cellules de repos sont identiques aux cellules végétatives bien que physiologiquement distinctes (von Dassow & Montresor, 2011), alors que les spores de résistance peuvent avoir une paroi particulière fortement silicifiée, avec peu de pores (Cox, 2011). Ces spores se forment par division mitotique (Edlund & Stoermer, 1997) et peuvent se séparer de la cellule parentale ou rester partiellement ou complètement enfermées dans celle-ci (Round et al., 1990). Les spores germent ensuite lorsque les conditions environnementales sont appropriées. Les spores des diatomées centriques ont cependant une courte période de dormance obligatoire avant de pouvoir germer à nouveau.

1.4. *Mode de vie*

Les diatomées sont unicellulaires, mais elles peuvent aussi bien se présenter sous forme de cellules isolées que sous forme de colonies. Ces colonies peuvent s'organiser sous la forme de chaîne (par exemple les espèces du genre *Melosira,* Figure I.9a), de ruban (*Fragilaria,* Figure I.9b), d'étoile (*Asterionella*), d'éventails (*Meridion*), d'arbre

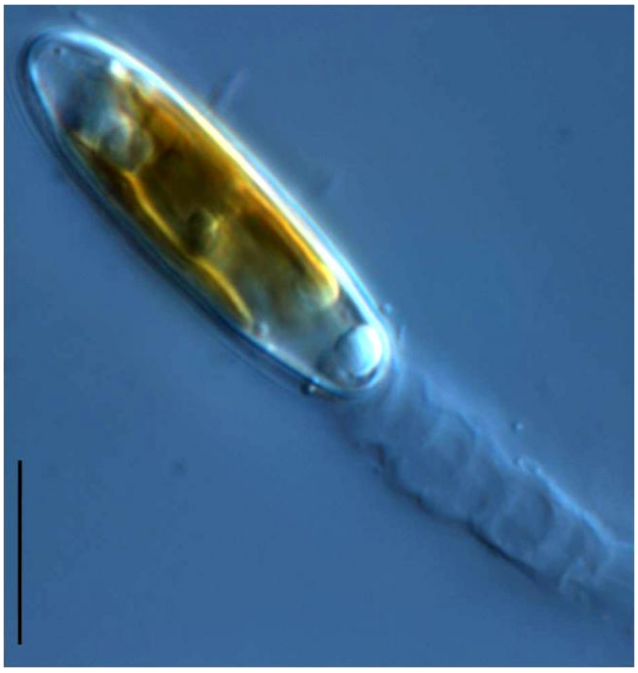


Figure I.10: Gomphonema bourbonense *à l'extrémité de son pédoncule*.
Barre d'échelle = 10µm.

(*Didymosphenia*) ou de zigzags (*Diatoma*).

La liaison entre les cellules au sein de ces colonies repose soit sur des épines siliceuses qui se trouvent sur les bords des valves de cellules adjacentes (par exemple pour des espèces appartenant aux genres *Fragilaria*, *Aulacoseira ou Cymatosira*), soit sur des filaments de polysaccharides ou de chitine (pour des espèces du genre *Coscinosira* ou *Thalassiosira*) (Round et al., 1990).

Par ailleurs, grâce aux substances mucilagineuses élaborées par certaines espèces, d'autres types d'organisation peuvent être observés. Certaines diatomées se développent à l'extrémité de pédoncules mucilagineux (par exemple les espèces des genres *Cymbella* et *Gomphonema*, Figure I.10) ou à l'intérieur de tubes mucilagineux (*Encyonema*) (Round et al., 1990).

Les substances mucilagineuses, en plus de leur rôle dans la constitution et la protection des colonies, sont également impliquées dans la locomotion. En effet, les diatomées sont dépourvues de flagelle, mais les cellules possédant un raphé peuvent se mouvoir en excrétant des substances mucilagineuses qui permettent à la cellule d'adhérer au substrat. Des microfibrilles situées au niveau du raphé permettent alors à la cellule de se déplacer (Drum & Hopkins, 1966). Cette capacité de mouvement détermine les différents habitats colonisables.

En effet, dans les écosystèmes aquatiques, les diatomées peuvent être scindées en deux groupes selon leur mode de vie : les diatomées planctoniques et les diatomées benthiques. Les diatomées planctoniques vivent libres en suspension dans la colonne d'eau et ne possèdent pas de capacités propres pour se déplacer. Elles sont représentées surtout par des centriques isolées (*Stephanodiscus*, *Cyclotella*, *Thalassiosira*) ou associées en chaînes (*Aulacoseira*, *Skeletonema*). Parmi les quelques formes pennées planctoniques se trouvent des colonies rubanées (*Fragilaria*) ou étoilées (*Asterionella*).

Les diatomées benthiques vivent sur divers supports immergés au niveau de la zone photique, où la lumière est suffisante pour assurer la photosynthèse. La majorité des diatomées benthiques correspond à des diatomées pennées dont la plupart sont mobiles. Ces micro-algues se développent au sein d'une matrice complexe (le biofilm), associées à d'autres algues mais aussi à des bactéries et des champignons, formant ensemble le périphyton. Les différentes communautés sont liées par des substances extracellulaires

(notamment des polysaccharides). Différentes catégories de périphyton dépendent du substrat colonisé:

- L'épilithon désigne les espèces vivant sur les substrats durs et inertes.
- L'épiphyton désigne les espèces vivant sur les végétaux aquatiques.
- L'épipélon et l'endopélon sont constitués d'espèces libres vivant à la surface et dans le sédiment.
- L'épipsammon correspond aux espèces vivant à la surface des grains de sable.

1.5. *Ecologie*

Les diatomées colonisent tous les milieux aquatiques à l'exception des eaux les plus chaudes et les plus hypersalines (Round et al., 1990). Mais leur faible besoin en lumière et en humidité peut leur permettre de coloniser d'autres milieux tels que l'air et le sol (Patrick, 1977). Certaines espèces se développent également en tant qu'endosymbiontes de dinoflagellés (Takano et al., 2008), de foraminifères (Lee et al., 1989), ou d'éponges (Bavestrello et al., 2000).

Les diatomées, en tant qu'organismes phototrophes, sont des producteurs importants de matière organique. Elles jouent donc un rôle particulièrement important dans les cycles du carbone (40 à 45% de la production primaire océanique) et de la silice (Mann, 1999). Les diatomées sont à la base de nombreux réseaux trophiques et sont une source de lipides pour un grand nombre d'organismes aquatiques (Julius & Theriot, 2010).

Les diatomées sont dépendantes de la présence de silice soluble dont la cellule a besoin pour construire son frustule, mais aussi de nombreux autres nutriments qui peuvent limiter plus ou moins leur développement (Round et al., 1990). Les différentes espèces ont chacune leurs propres préférences écologiques à certains nutriments. Par exemple, Hillebrand & Sommer (1997) ont détecté des modifications des espèces abondantes en fonction de différentes concentrations d'azote et de phosphore. Plusieurs travaux ont ainsi classé les espèces d'eau douce en fonction de leur préférence en teneur en nutriments (trophie), mais également en fonction de leurs préférences de pH, de salinité et de teneur en matière organique (saprobie) (e.g. Beaver, 1981; Van Dam et al., 1994; Potapova & Charles, 2007).

Différents toxiques rejetés par les activités humaines sont également à l'origine de changements de structure des communautés de diatomées. Les différentes espèces sont

plus ou moins résistantes à certains toxiques tels que les métaux (e.g. Pérès et al., 1997; Duong et al., 2010), les toxiques industriels (e.g. Debenest et al., 2010), pharmaceutiques (e.g. les bactéricides, Morin et al., 2010) ou agricoles (e.g. les herbicides, Pérès et al., 1996, ou les pesticides, Roubeix et al., 2011). Les différentes pressions anthropiques (apports de toxiques et de nutriments) jouent ainsi un rôle important dans la structuration les communautés de diatomées.

D'autres facteurs et processus vont également jouer un rôle majeur dans le développement des communautés de diatomées à l'exemple des caractéristiques hydrauliques, de l'intensité lumineuse et de la température (Patrick, 1977). Ainsi, les diatomées planctoniques sont souvent absentes dans la partie amont des cours d'eau alors qu'elles sont prédominantes dans les parties lentiques des rivières ainsi que dans les milieux lacustres et marins. Le développement des différentes espèces est ainsi lié aux conditions hydrodynamiques (e.g. Wang et al., 2012). Par ailleurs, la quantité de lumière et la température permettant une croissance optimale diffèrent selon les espèces, structurant ainsi fortement les communautés (Stockner, 1967; Patrick, 1971, 1977). Il faut signaler que, bien que la plupart des diatomées soient des organismes phototrophes, certaines espèces sont des organismes hétérotrophes facultatifs (e.g. *Cyclotella cryptica*, *Navicula pavillardi*) ou obligatoires (e.g. *Nitzschia alba*, *N. putrida*) (Hellebust & Lewin, 1977). Des facteurs biotiques tels que le parasitisme ou la prédation ont également une influence sur le développement de ces micro-algues algues et structurent leurs communautés (Patrick, 1977).

Les caractéristiques physico-chimiques des écosystèmes aquatiques ont donc un impact majeur sur la composition des communautés de diatomées ainsi que sur leurs abondances totale ou relative. C'est pourquoi composition et abondance des différentes espèces sont des indicateurs de l'état de l'écosystème dans lequel elles se développent. Ainsi, des modifications au sein de ces communautés peuvent traduire des changements dans leur environnement.

2. La bioindication

2.1. L'évaluation de la qualité des eaux continentales

La qualité de l'eau est déterminante à la fois pour le bon fonctionnement des écosystèmes aquatiques et la pérennité des services écosystémiques qu'ils rendent, et pour les différents usages qui en sont faits par l'Homme (consommation humaine ou animale, aquaculture ou loisirs). Les pollutions, quelles que soient leurs origines (agricole, industrielle et domestique) et leur nature, peuvent se traduire par des atteintes à la biodiversité et au fonctionnement de ces écosystèmes ainsi qu'à leurs usages. L'évaluation de la qualité de l'eau est donc essentielle pour la protection et la gestion des ressources en eau. Cette évaluation repose sur deux types d'approches : les mesures physico-chimiques et les évaluations biologiques.

2.1.1. Mesures physico-chimiques

Si la mesure des variables physico-chimiques de l'eau permet de détecter rapidement et simplement des modifications des conditions environnementales (Vis et al., 1998), ces analyses présentent cependant plusieurs limites :

- Elles sont ponctuelles et ne permettent donc pas de détecter des contaminants qui sont présents de manière intermittente (Blandin, 1986).
- Elles sont limitées analytiquement (seuil de détection, liste des molécules recherchées) et les contaminants, souvent présents en mélange, peuvent avoir des effets antagonistes ou synergiques non détectables par une mesure de leurs seules concentrations.
- Elles ne sont interprétables qu'en tenant compte également des variables biologiques puisque les concentrations mesurées dans l'eau dépendent de leurs consommations par les organismes de l'écosystème (par exemple la consommation du phosphore par le phytoplancton, Anneville & Pelletier, 2000).

2.1.2. Indicateurs biologiques

L'utilisation des bioindicateurs repose sur le principe que la structure des communautés reflète l'état de leur écosystème et que toute altération du milieu provoque un changement de cette structure (Blandin, 1986). Les bioindicateurs ont l'avantage

d'intégrer les variations du milieu sur une période plus longue que les analyses chimiques et sont donc particulièrement appropriés aux rivières où les conditions environnementales peuvent fortement varier spatialement et temporellement (Stevenson & Pan, 1999). De plus, ils permettent d'évaluer les effets de la perturbation qui peuvent ne pas être corrélés simplement aux concentrations (effets synergistes ou antagonistes).

Pour permettre une application large des techniques de bioindication, les communautés d'organismes doivent présenter plusieurs caractéristiques. Leur principale caractéristique est de présenter des tolérances modérées à certaines variables de l'environnement. En effet, les communautés rares avec des tolérances trop étroites sont souvent trop sensibles aux changements environnementaux, ou trop rarement rencontré pour refléter une réponse générale de l'écosystème. Au contraire, les organismes ubiquistes avec des tolérances très larges sont moins sensibles aux perturbations qui peuvent atteindre le reste de l'écosystème (Holt & Miller, 2011). Ces communautés d'organismes doivent donc être cosmopolites pour permettre une application à grande échelle. Ensuite, ces organismes doivent être abondants dans le milieu et facilement échantillonnables (Round, 1991). De plus, ils doivent également avoir un cycle de vie suffisamment court pour montrer des modifications à court terme de l'environnement, afin d'évaluer la qualité du milieu à différentes échelles de temps, des modifications ponctuelles aux changements à long terme (Hellawell, 1978). Enfin, les organismes bioindicateurs doivent être facilement identifiables et quantifiables (Round, 1991).

2.2. *La Directive Cadre Européenne sur l'eau*

En Europe, la Directive 2000/60/CE du parlement européen et du conseil du 23 octobre 2000 (DCE) établit « un cadre pour une politique communautaire dans le domaine de l'eau ». La DCE requiert, pour les états membres, d'évaluer la qualité des écosystèmes aquatiques au moyen d'indicateurs chimiques et biologiques, avec l'objectif d'atteindre un « bon état » chimique et écologique des masses d'eau (cours d'eau, lacs, estuaires, eaux côtières et souterraines) pour l'année 2015.

L'état chimique ne comprend que deux classes : « bon état » et « pas bon état » en fonction des normes de qualité environnementale (NQE) qui sont définies à partir de tests écotoxicologiques, et qui déterminent les seuils de polluants à ne pas dépasser pour obtenir un bon état chimique des eaux.

Figure I.11: *Carte des hydroécorégions de niveau 1 de France métropolitaine*. Source : Wasson et al., 2002.

L'état écologique décrit la qualité de la structure et du fonctionnement des écosystèmes aquatiques. Pour le suivi de l'état écologique, différents bioindicateurs sont requis : les poissons, les macro-invertébrés, les macrophytes, le phytoplancton et le phytobenthos (les diatomées). Ils sont choisis en fonction de leur pertinence pour la masse d'eau considérée. Quel que soit l'indicateur biologique utilisé, le principe d'évaluation est le même : des indices basés sur les bioindicateurs sont calculés et comparés aux valeurs d'indices attendues en conditions naturelles, en référence à des états non perturbés ou faiblement impactés par les activités humaines. L'état écologique est divisé en 5 classes de qualité : du « très bon état » au « mauvais état ».

Des stations de référence ont été définies pour chaque masse d'eau et chaque zone géographique. En effet, les communautés aquatiques présentent naturellement des différences en fonction des caractéristiques du milieu dans lequel elles vivent. En ce qui concerne les cours d'eau, des typologies ont été déterminées (Ector & Rimet, 2005). Ainsi, en France métropolitaine, différentes hydro-écorégions (HER) ont été définies comme des zones géographiques présentant des caractéristiques communes en termes de fonctionnement écologique et donc de communautés aquatiques. Ces HER constituent un cadre pour décrire les conditions abiotiques de différents types de cours d'eau et les communautés de référence associées. En France métropolitaine, 22 HER de niveau 1 ont ainsi été définies (HER-1, Figure I.11) en fonction du relief, de la géologie et du climat (Wasson et al., 2002). Des stations de référence ont été établies pour chacune de ces HER. Concernant les départements d'outre-mer (DOM), ces travaux sont en cours de réalisation. Par exemple, les HER de La Réunion ont été définies en 2004 (Wasson et al., 2004) et les stations de référence pour les cours d'eau ont été proposées en 2008 (Asconit Consultants, 2008).

L'état écologique est défini comme l'écart de l'indice de bioindication par rapport à sa valeur dans la station de référence. Plusieurs bioindicateurs étant utilisés, la classification de l'état écologique d'un site correspond à celle indiquée par le paramètre le plus déclassant.

2.3. *La bioindication des rivières par les diatomées*

2.3.1. Avantages des diatomées

Depuis Kolkwitz & Marsson (1908), qui reliaient les espèces d'algues (incluant les diatomées) à la qualité de l'eau, les diatomées ont prouvé leurs capacités dans les études de pollution, surtout dans les rivières (Round et al., 1990). En effet, les diatomées benthiques présentent un ensemble de caractéristiques attendues pour les bioindicateurs.

Comme nous l'avons précédemment évoqué, le phylum des diatomées est retrouvé dans une large gamme d'environnement puisqu'en eau douce, elles sont présentes dans toutes les gammes de qualité d'eau, allant des sources non polluées jusqu'aux eaux les plus dégradées. Par ailleurs, ces microorganismes sont sensibles à la plupart des paramètres physico-chimiques et chaque espèce a ses propres exigences. La grande diversité d'espèces permet donc de couvrir une large gamme de préférences écologiques grâce à ce seul phylum. Ensuite, les diatomées colonisent différents substrats et peuvent donc être échantillonnées sans difficulté dans différents types de rivières (Prygiel, 1991). Enfin, étudiée depuis de nombreuses années, leur taxonomie est devenue suffisamment précise pour utiliser les diatomées dans les programmes de surveillance et de changements environnementaux (Smol & Stoermer, 2010).

2.3.2. Indices basés sur les diatomées

De nombreuses méthodes d'évaluation de la qualité des eaux basées sur les diatomées benthiques (épilithon) sont répertoriées en Europe (Ector & Rimet, 2005). La plupart des indices repose sur la formule de Zelinka & Marvan (1961) qui prend en compte plusieurs paramètres d'un taxon (i) : son abondance relative dans la communauté (A_i), sa sensibilité à la pollution ou à un ou plusieurs paramètres chimiques (I_i) et sa valeur indicatrice (V_i) qui est une fonction de l'amplitude écologique de l'espèce vis-à-vis de la pollution:

$$ID = \frac{\sum_{i=0}^{n} A_i . I_i . V_i}{\sum_{i=0}^{n} A_i . V_i}$$

Les indices diffèrent par l'unité taxonomique retenue et le nombre de taxa utilisés ainsi que par les coefficients de sensibilité et les coefficients indicateurs appliqués aux taxa (Prygiel & Coste, 2000).

En France, les premières applications d'indices diatomiques ont été réalisées en 1974 sur le bassin de la Seine en utilisant 55 espèces (Prygiel & Coste, 2000). Depuis les indices et le nombre de taxa utilisés n'ont cessé d'évoluer et se sont généralisés à tout le territoire métropolitain. Le développement du programme OMNIDIA (Lecointe et al., 1993), permettant le calcul de différents indices diatomées, a favorisé l'utilisation de ces indices.

Les deux principaux indices utilisés en France ont été développés par le Cemagref : l'Indice de Polluosensibilité Spécifique, l'IPS (Cemagref, 1982) et l'Indice Biologique Diatomées, l'IBD (Lenoir & Coste, 1996). Contrairement à d'autres indices plus spécifiques comme le Trophic Diatom Index (Kelly & Whitton, 1995) qui estime l'eutrophisation du milieu, l'IPS et l'IBD informent sur la qualité globale des eaux.

L'IPS est basé sur 3143 taxa de diatomées pour lesquels des valeurs de sensibilité et des valeurs indicatrices ont été définies. La valeur de sensibilité de l'espèce est notée sur cinq, 1.0 correspondant à une sensibilité pour une mauvaise qualité d'eau et 5.0 pour une bonne qualité. La valeur indicatrice est notée sur trois et correspond à l'amplitude écologique d'une espèce: 1 pour espèce ayant une gamme de tolérance de qualité d'eau large et 3 pour une tolérance restreinte.

L'IBD utilise un nombre de taxa plus réduit (1478), excluant les taxa considérés comme rares et regroupant les taxa difficiles à différencier en taxa appariés (Prygiel & Coste, 2000). L'IBD est l'indice requis par la DCE et utilisé en routine pour le suivi de la qualité de l'eau. De mise en œuvre simple, il a été normalisé en 2000 (AFNOR NF T 90-354), puis révisé en 2007. Brièvement, après avoir récolté les diatomées benthiques par brossage de substrats immergés (en priorité des substrats durs tels que des pierres et des galets), l'échantillon est traité afin de détruire les matières organiques et les carbonates de calcium. Cette étape permet ensuite d'observer de manière optimale les frustules en microscopie optique. Le calcul de l'IBD repose sur l'identification, au niveau de l'espèce, et le comptage de 400 individus.

Pour le calcul de l'IBD, 1478 taxa, incluant 476 synonymes et 190 formes tératologiques sont pris en compte et sept classes de qualité sont définies. Le calcul de

l'indice repose sur l'abondance des différentes espèces de l'échantillon (en ‰), sur leur valeur écologique et sur leur probabilité de présence dans chacune des sept classes de qualité. Le logiciel OMNIDIA permet de calculer l'IBD automatiquement en mettant en relation l'inventaire établi par microscopie avec la valeur écologique de chaque taxon et sa probabilité de présence dans les classes de qualité. La valeur de l'indice correspond à une note allant de 0 à 20 à laquelle est associée une qualité d'eau (Tableau I.1).

Tableau I.1: *Qualités d'eau associées aux notes d'IBD.*

IBD (note sur 20)	BD ≥ 17.0	17.0 > IBD ≥ 13.0	13.0 > IBD ≥ 9.0	9.0 >IBD ≥ 5.0	IBD < 5.0
Qualité de l'eau	Très bonne	Bonne	Moyenne	Mauvaise	Très mauvaise

2.4. *Limites des indices diatomées*

Les indices diatomées utilisés actuellement sont robustes mais présentent quelques limites. Tout d'abord, la qualité des données de bioindication dépend principalement de deux facteurs : (i) la façon dont l'échantillon est prélevé afin qu'il représente au mieux la communauté ciblée et (ii) l'analyse de la composition de l'échantillon (Pfrender et al., 2010). Le premier facteur est soumis à une norme (AFNOR NF T 90-354) et n'entraîne donc pas, en théorie, de variations trop importantes de l'indice. Le second facteur est dépendant de la résolution taxonomique utilisée et du taux d'erreur d'identification (Pfrender et al., 2010). Les identifications sont réalisées en microscopie optique sur les caractéristiques subtiles du frustule, ce qui nécessite une forte expertise. Blandin, en 1986, évoquait déjà l'inconvénient de ces déterminations précises qui ne peuvent être réalisées que par des spécialistes. Avec l'application de la DCE, le nombre d'échantillons à analyser annuellement pour la surveillance des milieux a fortement augmenté alors que le nombre de diatomistes reste quant à lui limité, avec même une tendance à la diminution.

Les effets de certains facteurs environnementaux sur la morphologie des frustules (Round et al., 1990; Kociolek & Stoermer, 2010) peuvent gêner l'identification morphologique des diatomées, de même que l'utilisation de différentes classifications pour les diatomées peut engendrer des divergences dans l'identification des espèces (Mann et al., 2010). Or, la précision de l'identification est cruciale car des diatomées

morphologiquement proches peuvent avoir des sensibilités différentes aux conditions environnementales (Ivorra et al., 2002; Vanelslander et al., 2009). Par ailleurs, lors d'exercices d'intercalibration (Prygiel et al., 2002; Kahlert et al., 2009), les complexes de plusieurs espèces (par exemple *Achnanthidium minutissimum* (Kützing) Czarnecki; *Fragilaria capucina* Desmazières; *Nitzschia palea* (Kützing) Smith) ont été des sources d'erreurs importantes et ont généré une grande variabilité dans les résultats des indices. L'opérateur est également une source d'incertitude dans toutes les étapes de l'analyse, de l'échantillonnage jusqu'à l'identification (Besse-Lototskaya et al., 2006).

Pour certains pays comme la France, qui se doivent d'appliquer la DCE dans des régions tropicales, se pose de surcroît le problème de l'existence de flores différentes de celles du continent européen et encore mal connues. Les indices habituellement utilisés sont inadaptés pour décrire la qualité des milieux tropicaux et ils doivent donc être adaptés aux particularités environnementales, biologiques et physico-chimiques de ces régions (Asconit Consultants, 2008).

Evaluer la qualité de l'eau à large échelle, nécessite donc exactitude, précision, rapidité, tout ceci à un coût raisonnable en raison du très grand nombre d'échantillons à analyser (Pfrender et al., 2010). Pour toutes ces raisons, les techniques basées sur l'identification des diatomées par microscopie sont devenues limitantes. Il devient donc nécessaire de développer de nouveaux outils d'identification des diatomées afin de faciliter le suivi de l'état écologique des rivières et c'est dans ce cadre général que s'est placé mon travail de thèse.

Les contraintes actuelles pourraient être minimisées en développant un nouvel outil qui ne serait plus basé sur un caractère variable tel que la morphologie et qui serait plus facilement automatisable. Dans ce cadre, les approches moléculaires basées sur l'ADN ont un fort potentiel pour évaluer rapidement la composition d'un échantillon. De plus basées sur des séquences de quatre nucléotides, l'ADN fournit des informations moins ambigües que les variations morphologiques qui sont graduelles.

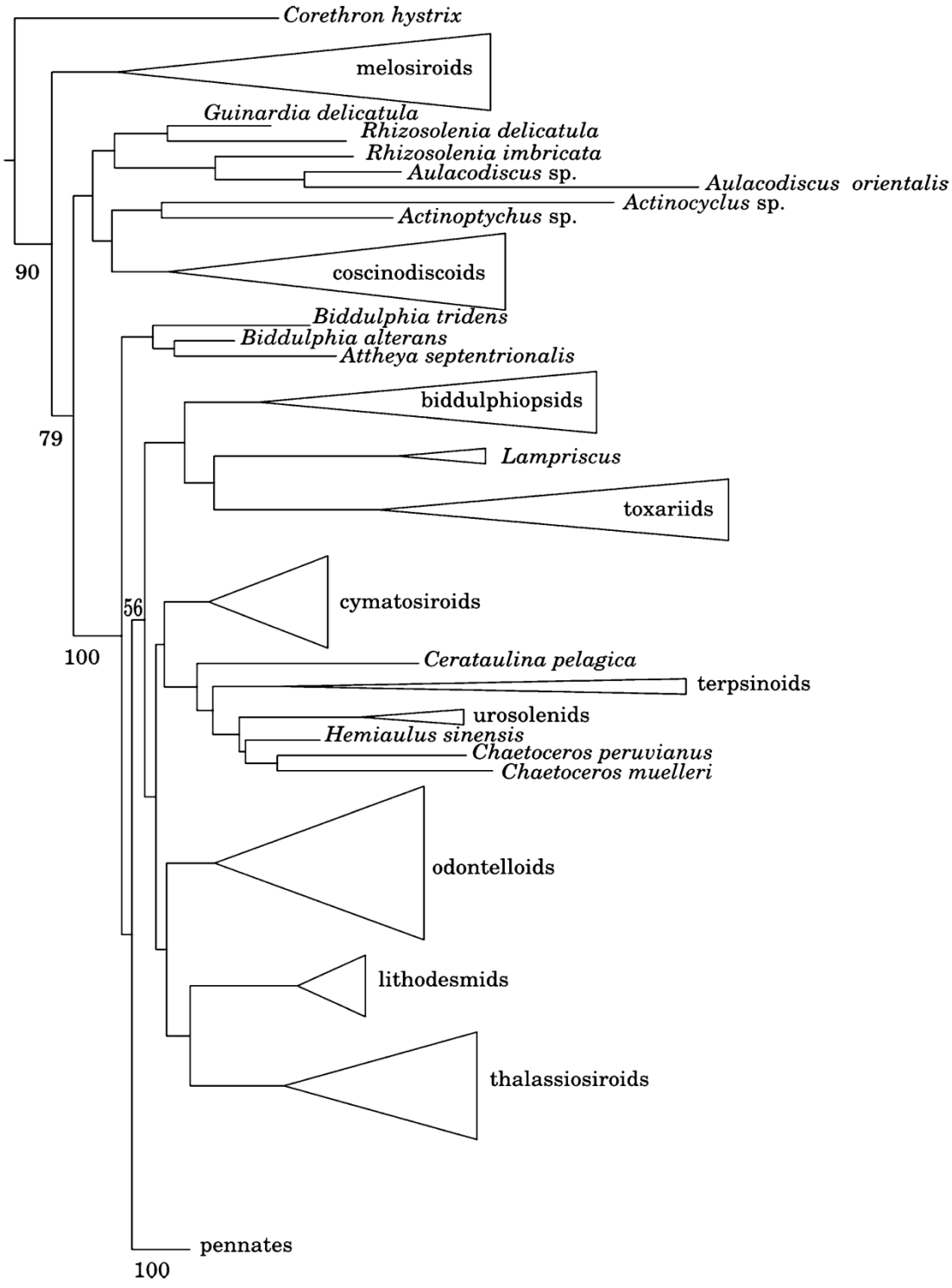


Figure I.12: *Arbre phylogénétique des diatomées centriques basé sur les séquences des gènes 18S,* rbc*L et* psb*C.*
Source : Theriot et al., 2011

3. Approches moléculaires

La généralisation des approches moléculaires a été rendue possible grâce au développement, au milieu des années 80, de la PCR (Saiki et al., 1985, 1988; Mullis et al., 1986) et du séquençage Sanger (Sanger et al., 1977). Plus récemment, les avancées des technologies de séquençage à haut débit ont ouvert des perspectives nouvelles dans l'utilisation de ces technologies pour l'analyse d'échantillons environnementaux (Hudson, 2008).

Bien que les diatomées soient étudiées depuis plus d'un siècle, l'utilisation des approches moléculaires sur ces micro-algues est récente et reste limitée essentiellement à l'étude de leurs relations phylogénétiques. Quelques génomes de diatomées (nucléaires ainsi que les génomes des organites) ont été séquencés (e.g. Armbrust et al., 2004; Oudot-Le Secq et al., 2007; Bowler et al., 2008; Ravin et al., 2010). Les génomes nucléaires ont une longueur de plus de 30 Mbp tandis que les génomes des organites ont une taille d'environ 120 kbp et 45 kbp pour les chloroplastes et les mitochondries respectivement. Ainsi la taille des génomes limite les études sur génome entier et la plupart des analyses moléculaires des diatomées sont réalisées après amplification d'une région d'intérêt.

3.1. *Phylogénies moléculaires des diatomées*

Les séquences ADN apportent des informations sur l'histoire évolutive en se basant sur le concept d'horloge moléculaire développé par Zuckerkandl & Pauling (1965). La comparaison des séquences ADN permet de produire des arbres phylogénétiques qui reflètent le degré de proximité entre les séquences. Un exemple d'arbre phylogénétique des diatomées centriques basé sur les séquences de trois gènes (18S, *rbc*L et *psb*C) est présenté en Figure I.12.

En ce qui concerne les diatomées, les premières analyses par des approches moléculaires remontent à la fin des années 80 (Medlin et al., 1988). Depuis ces travaux basés sur l'utilisation des séquences codant pour la petite sous-unité de l'ARNr, les séquences de l'opéron ribosomal sont couramment analysées pour étudier la phylogénie des diatomées. Plusieurs auteurs ont donc utilisé les ADNr 18S et 28S pour étudier la phylogénie et la taxonomie des diatomées (Sorhannus et al., 1995; Beszteri et al., 2001; Kooistra et al., 2003; Alverson & Kolnick, 2005; Sarno et al., 2005).

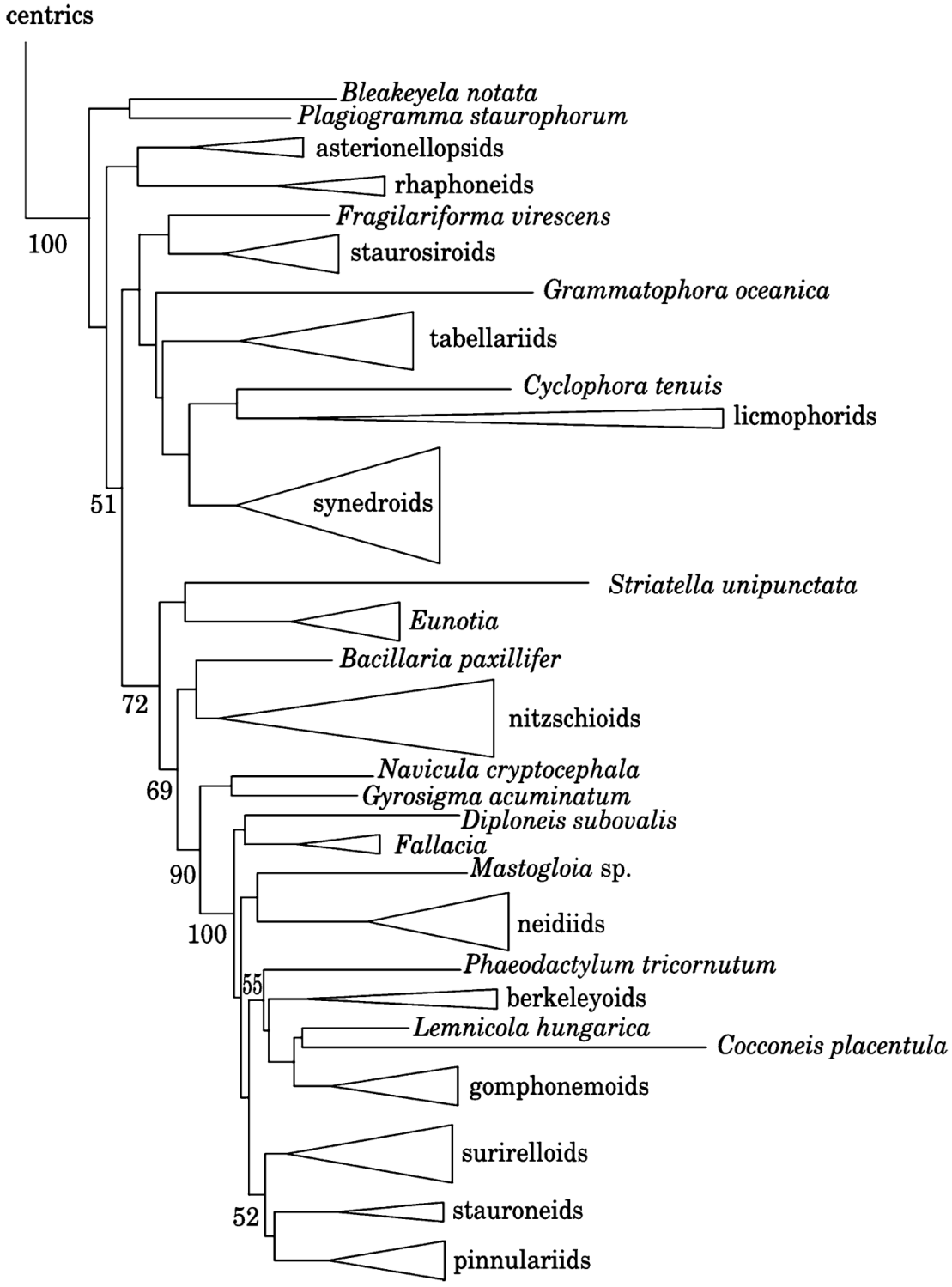


Figure I.13: *Arbre phylogénétique des diatomées pennées basé sur les séquences des gènes 18S,* rbc*L et* psb*C.*
Source : Theriot et al., 2011

D'autres marqueurs ont également été séquencés pour étudier les relations entre diatomées, de leur position au sein des heterokonts jusqu'aux relations entre les espèces d'un même genre: le *rbc*L (Daugbjerg & Andersen, 1997a; Edgar & Theriot, 2004; Bruder & Medlin, 2007, 2008), le *rpo*A (Fox & Sorhannus, 2003) et le *psb*C (Theriot et al., 2010) sur le génome chloroplastique ainsi que le *cox*1 (Ehara et al., 2000; Evans et al., 2007; Kaczmarska et al., 2008) sur le génome mitochondrial.

Les arbres phylogénétiques sont basés soit sur un marqueur nucléique, soit sur les données associées de plusieurs marqueurs. Par exemple, l'arbre phylogénétique présenté en Figure I.13 est le résultat d'une analyse basée sur trois gènes : l'ADNr 18S, le *rbc*L et le *psb*C (Theriot et al., 2011). La classification traditionnelle distinguait trois grands groupes chez les diatomées (les centriques, les araphidées et les raphidées). Il en est de même pour les données moléculaires (Medlin & Kaczmarska, 2004) mais la composition de ces groupes diffère. Le premier regroupe les Coscinodiscophyta comprenant les diatomées centriques radiales, le second les Mediophyceae contenant les diatomées centriques bi- ou multipolaires et les Thalassiosirales, et enfin le troisième les Bacillariophyceae composé des diatomées pennées. Ces trois groupes sont connus sous le nom de l'hypothèse CMB (Coscinodiscophyta, Mediophyceae, Bacillariophyceae), hypothèse qui a suscité beaucoup de débats dans la communauté des taxonomistes et a été fortement critiquée (Theriot et al., 2009).

Il reste encore beaucoup d'incertitudes concernant les relations phylogénétiques chez les diatomées. En effet, Theriot et al. (2011) indiquent que des données supplémentaires (morphologie, séquences) sont nécessaires pour obtenir des résultats robustes. Par ailleurs, les analyses phylogénétiques restent limitées par l'étape de mise en culture des espèces à étudier, préalable au séquençage, sachant que cette étape est difficile et donc que seules quelques espèces sont actuellement disponibles en culture.

3.2. *Le concept de « Barcoding DNA »*

Les espèces vivantes sont traditionnellement décrites et caractérisées sur la base de critères morphologiques. Mais comme nous l'avons vu, cette approche a des limites dans les groupes de microorganismes où les caractères morphologiques sont peu accessibles ainsi que chez certains groupes d'organismes où les caractères morphologiques sont très peu variables (Pompanon et al., 2011).

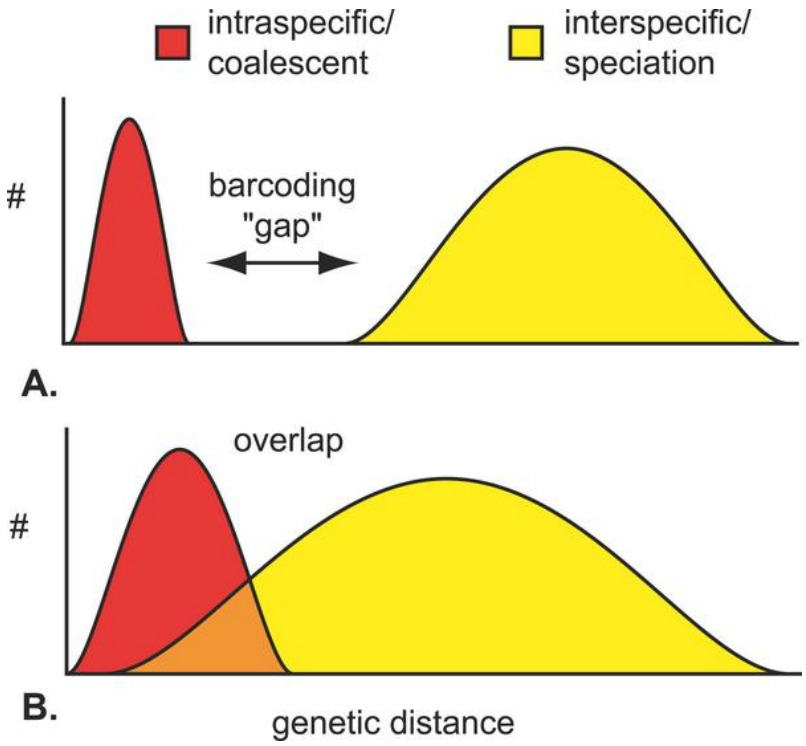


Figure I.14: *Principe du barcoding*.
Source : Meyer & Paulay, 2005

Pour cette raison, chez certains groupes comme les bactéries, des critères moléculaires (séquences d'ADNr 16S) ont rapidement été privilégiés pour distinguer des Unités Taxonomiques Opérationnelles (OTU) et ainsi étudier leur biodiversité (McCaig et al., 1999; Kroes et al., 1999; Ward, 2002). Depuis quelques années, l'utilisation des approches moléculaires à des fins d'identification taxonomique s'est étendue aux eucaryotes sous le concept de « barcoding DNA » (Hebert et al., 2003). Cette approche basée sur l'utilisation de l'ADN permet d'identifier plus facilement des organismes présentant par exemple différentes formes lors de leur cycle de vie (larves, œufs) ou des spécimens endommagés, rendant difficile leur identification, sur critères morphologiques.

Même si le barcoding utilise des séquences ADN comme la phylogénie moléculaire, son but est différent. La phylogénie moléculaire étudie l'histoire évolutive des espèces alors que le barcoding est un outil d'identification. Le but du barcoding est donc d'identifier un spécimen inconnu, en comparant sa séquence (qui est un barcode ou code-barre ADN) à une base de séquences de référence composée de spécimens connus (Hajibabaei et al., 2007).

Le concept du barcoding suppose cependant que les séquences provenant d'individus d'une même espèce soient plus semblables entre elles que les séquences provenant d'individus appartenant à des espèces différentes. Le fragment d'ADN utilisé comme barcode doit donc présenter une variabilité interspécifique plus importante que sa variabilité intraspécifique. Cet écart entre les variations intra- et interspécifiques est appelé le « barcoding gap » (Figure I.14 A). Ainsi, plus les variations intra- et interspécifiques se chevauchent (Figure I.14 B), moins le barcode est efficace pour identifier les espèces et peut induire des erreurs d'identification. Le choix du barcode est donc crucial dans l'utilisation du concept de barcoding.

Le barcoding a d'abord été développé sur des animaux comme les oiseaux (Hebert et al., 2004) et les poissons (Ward et al., 2005) en utilisant comme barcode un fragment du gène mitochondrial *cox*1. Ce gène peut être amplifié aisément par PCR en utilisant des amorces universelles pour les espèces animales. De plus, il est présent en de nombreuses copies ce qui facilite son séquençage à partir de matériel ancien ou dégradé (Hajibabaei et al., 2005). Le barcoding a rapidement été testé pour remplacer des analyses morphologiques dans le cadre de la bioindication. Sweeney et al. (2011) ont testé son efficacité pour évaluer la biodiversité des macro-invertébrés aquatiques et la qualité de l'eau. Leurs résultats ont montré que le barcoding est un outil qui peut fournir une

description plus précise des espèces présentes dans un échantillon, que l'identification morphologique.

Cependant, il est très vite apparu que le gène *cox*1 ne pouvait être utilisé pour l'ensemble des eucaryotes car chez certains groupes d'organismes ce gène n'évolue pas assez rapidement, ce qui ne permet pas d'obtenir une résolution au niveau de l'espèce (Kress et al., 2005). D'autres propositions de barcodes ont donc été faites pour les plantes (Kress & Erickson, 2007; Chase et al., 2007) ainsi que pour certains champignons (Seena et al., 2010).

En ce qui concerne les diatomées, plusieurs barcodes ont été proposés: une région du gène *cox*1 (Evans et al., 2007), les 5,8S-ITS2 (Moniz & Kaczmarska, 2010), une région du gène *rbc*L associée à l'ADNr 28S en barcode secondaire (Hamsher et al., 2011) et une région polymorphe (V4) de l'ADNr 18S (Zimmermann et al., 2011). Le fait que plusieurs séquences aient été testées, reflète l'intérêt du barcoding pour l'identification des diatomées. Par rapport aux limites déjà évoquées sur l'utilisation des seuls critères morphologiques pour identifier les diatomées, le barcoding, même s'il ne peut pas régler tous les conflits taxonomiques, peut offrir une solution aux difficultés d'identification de ces micro-algues. Le barcoding ciblant une région du génome, les diatomées pourraient être identifiées à n'importe quel stade de leur cycle de vie. Ainsi le polymorphisme, les formes tératologiques et les formes de résistance ne poseraient plus de problème d'identification (Mann et al., 2010). De plus, la standardisation de la taxonomie et de l'identification qui résulterait de l'utilisation des barcodes faciliterait les études sur les diatomées. Plusieurs exemples illustrent la façon dont les barcodes ADN peuvent être utilisés pour stabiliser la taxonomie (Jahn et al., 2008; Evans & Mann, 2009).

Le barcoding des diatomées reste cependant soumis aux mêmes limites que les analyses phylogénétiques, notamment en raison de la nécessité de mise en culture de souches pour créer des bases de séquences de référence.

3.3. *Barcoding environnemental*

Les taxonomistes utilisent le barcoding de manière « stricte » c'est-à-dire pour identifier un spécimen au niveau taxonomique de l'espèce. Valentini et al., (2009) ont proposé un concept élargi du barcoding pour une utilisation en écologie. Cette approche moins restreinte englobe toutes les identifications (à n'importe quel niveau taxonomique)

d'organismes présents dans un échantillon environnemental en utilisant un court fragment d'ADN.

Jusqu'à récemment, le séquençage d'échantillons environnementaux, contenant différents organismes, nécessitait une étape de clonage avant le séquençage. En effet, la technique classique de séquençage Sanger exigeait la séparation préalable des différents fragments d'ADN à séquencer. Au cours des dernières années, de nouvelles technologies de séquençage (les NGS) ont été développées. Selon les techniques, le principe peut être différent, mais toutes permettent le séquençage de différents ADN en mélange. Ainsi, l'étape de clonage longue et coûteuse n'est plus nécessaire. De plus, les NGS sont des techniques à haut débit, fournissant des millions de séquences, ce qui permet d'appréhender directement la diversité d'échantillons environnementaux à des échelles jusqu'alors inaccessibles (Sogin et al., 2006).

L'émergence des NGS offre le potentiel d'étendre l'utilisation des barcodes à des applications sur des échantillons environnementaux. A ce jour, quelques études ont utilisé les NGS à partir d'un fragment d'ADN préalablement amplifié par PCR, pour étudier la biodiversité dans différents échantillons environnementaux : champignons ou bactéries dans le sol (Buée et al., 2009; Uroz et al., 2010), bactéries ou protistes des écosystèmes marins (Galand et al., 2009; Stoeck et al., 2009) ou fèces d'animaux (Soininen et al., 2009; Valentini et al., 2009).

Hajibabaei et al. (2011) ont utilisé le pyroséquençage 454, une technologie NGS, pour des analyses de macro-invertébrés benthiques d'eau douce. Leurs résultats ont démontré l'utilité d'une approche de barcoding environnemental en permettant de comparer la composition en macro-invertébrés benthiques d'une région urbaine à celle d'une zone protégée. Ces nouvelles technologies montrent donc un fort potentiel pour remplacer le travail laborieux de comptage en microscopie pour les études de diatomées (Mann et al., 2010). Un tel système, basé sur les barcodes, ne nécessiterait plus l'expertise en diatomées requise par la microscopie, mais plutôt des compétences en biologie moléculaire, ainsi que l'accès à des plateformes de séquençage haut-débit pour la production des données, et à des outils bioinformatiques pour le traitement de ces données. Or, ces types de compétences et d'outils deviennent actuellement plus répandus que les diatomistes.

Ainsi, au cours de ce travail de thèse, nous avons utilisé des approches moléculaires pour déterminer dans quelle mesure celles-ci pouvaient simplifier et améliorer l'utilisation des diatomées en tant que bioindicateurs. Tout d'abord, nous avons évalué les avantages que les outils moléculaires pouvaient apporter à l'identification des diatomées. Puis, nous avons développé une méthode basée sur le barcoding et les NGS visant à analyser la composition en diatomées d'un échantillon environnemental. Et enfin, nous avons testé cette méthode en comparant les résultats moléculaires aux résultats des analyses classiques de morphologie.

CHAPITRE II. MATERIEL ET METHODES

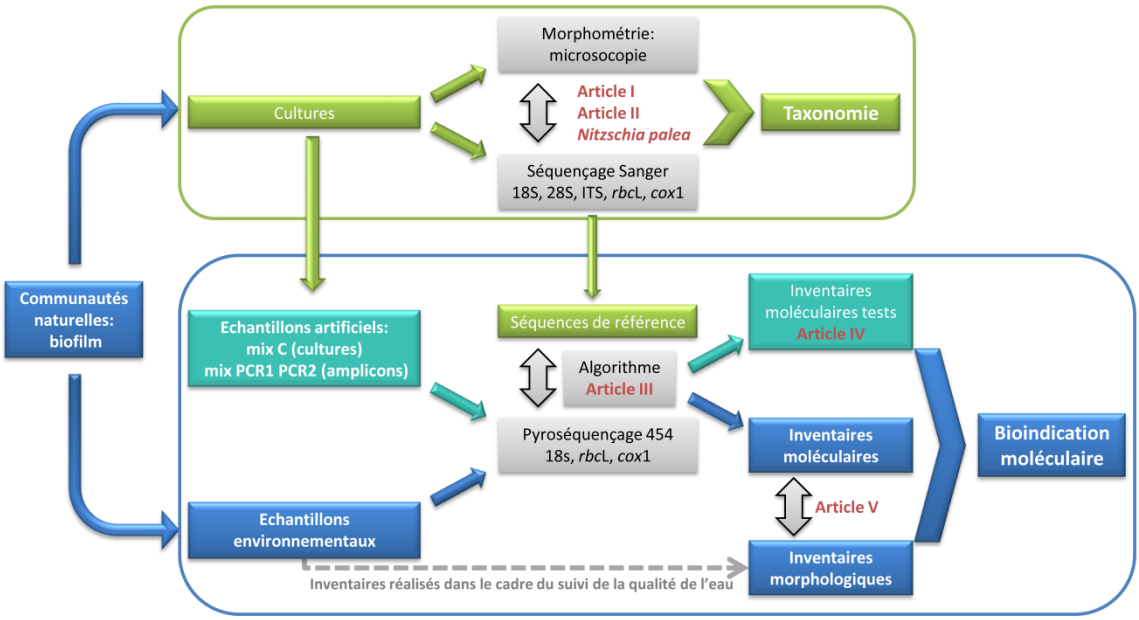


Figure II.1: *Schéma général de la méthodologie*.

Les objectifs principaux de la thèse étaient d'améliorer les connaissances taxonomiques des diatomées et de développer un nouvel outil pour l'identification des diatomées dans le cadre de la bioindication. Ainsi nous avons utilisé des approches moléculaires que nous avons comparées aux approches morphologiques utilisées couramment pour l'étude des diatomées. La méthodologie générale mise en place au cours de la thèse est présentée en Figure II.1. A partir de différentes communautés naturelles, nous avons mis en culture des souches de diatomées que nous avons identifiées par microscopie optique et que nous avons caractérisées par séquençage de différents marqueurs moléculaires. La comparaison des données morphologiques et moléculaires nous ont permis d'étudier la taxonomie de certains groupes de diatomées (CHAPITRE III). En parallèle, les séquences des différentes cultures ont complété les bases de séquences de référence existantes. De plus, nous avons créé des échantillons artificiels de diatomées à partir de ces souches caractérisées par leurs critères morphologiques et par leurs séquences. Ces échantillons artificiels ont servi aux étapes de développement d'un outil moléculaire en permettant la mise au point d'un algorithme de comparaison de séquences et l'étude des biais liés aux marqueurs et à la techniques de pyroséquençage 454 (CHAPITRE IV). Les échantillons environnementaux ont ensuite servi à tester la méthode développée au cours de la thèse, sur des échantillons naturels (CHAPITRE IV).

1. Echantillons

1.1. *Sites d'échantillonnage*

Pour répondre aux différents objectifs de la thèse, plusieurs sites d'échantillonnage ont été utilisés au cours de l'année 2009 dans les DOM, Mayotte et La Réunion, ainsi qu'en France métropolitaine, par Asconit Consultants et la DREAL. Les sites d'échantillonnage ont été choisis car ils présentaient des espèces abondamment représentées dans les échantillons benthiques d'eaux douces prélevés dans la cadre de la bioindication. Ce choix s'est basé sur les inventaires de diatomées réalisés en 2008, présentant des espèces d'intérêt ayant une abondance moyenne ou élevée. Les échantillons utilisés au cours de cette thèse sont listés dans le Tableau II.1.

Tableau II.1: *Liste des sites échantillonnés.*

	Code échantillon	Rivière	Site	Date de prélèvement
Mayotte	DJAV	Djalimou	aval	18.04.2009
	BOIN	Bouyouni	intermédiaire	19.04.2009
	DEAV	Dembeni	aval	18.04.2009
	COIN	Coconi	intermédiaire	18.04.2009
	COAV	Coconi	aval	18.04.2009
	KWAM	Kwalé	amont	18.04.2009
	LOAV	Longoni	aval	20.04.2009
	BOAM	Bouyouni	amont	19.04.2009
La Réunion	BGLA	rivière des Galets	Marla	30.04.2009
	BDNA	Saint Denis	amont prise AEP	28.04.2009
	BBEA	Bras des étangs		24.04.2009
	BCVA	Bras Caverne	Amont confluence riv. du Mât	29.04.2009
	BLGA	Langevin	amont cascade grand Galet	25.04.2009
	BLGC	Langevin	amont prise EDF	25.04.2009
	BSZ1	Sainte Suzanne	Amont Bassin Bœuf	19.05.2009
	BSZ2	Sainte Suzanne	Cascade Niagara	19.05.2009
France métropolitaine	SOLV	Le Sânon	Solvay	18.06.2009
	MEAR	Meurthe	Art-sur-Meurthe	17.06.2009
	SECH	Seille	Chambrey	17.06.2009
	MOAR	Moselle	Ars-sur-Moselle	17.06.2009
	ORJO	Orne	Joeuf	17.06.2009
	CHLO	Chiers	Longlaville	17.06.2009
	MOBA	Moselle	Bainville aux Miroirs	18.06.2009
	NANO	Canal de Nantes à Brest	Nort-sur-Erdre	18.08.2009
	DOGU	Don	Guémené-Penfao	18.08.2009
	VIFE	La Vie	Fenouillet	17.08.2009
	ISGE	Isac	Genrouet	18.08.2009
	LACL	Le Lay	La Clay	17.08.2009
	LOLE	Le Lourdon	Lentigny	19.08.2009
	GIVA	Le Gier	La Valla	19.08.2009
	GIGI	Le Gier	Givors	19.08.2009
	SALY	La Saône	Lyon	13.10.2009
	SABE	La Saône	Saint Bernard	13.10.2009
	SAFL	La Saône	Fleurville	13.10.2009

1.2. *Prélèvements* in situ

Les prélèvements ont été réalisés selon la norme NF T90-354 de décembre 2007 qui définit la procédure d'échantillonnage des diatomées *in situ* dans le but d'évaluer la qualité biologique d'un cours d'eau.

Les échantillons prélevés ont été divisés en trois sous-échantillons:

- un sous-échantillon vivant (non fixé) pour l'isolement de cellules.

- un sous-échantillon fixé au formaldéhyde pour évaluer la diversité de l'échantillon par microscopie.

- un sous-échantillon fixé à l'éthanol 70% pour évaluer la diversité génétique de l'échantillon par des approches moléculaires.

En outre, de l'eau de rivière a été prélevée sur chaque site d'échantillonnage pour compléter le milieu de culture destiné aux souches isolées.

1.3. *Cultures de souches de diatomées*

Afin d'accroître la diversité des souches en culture, trois milieux de culture différents, adaptés aux diatomées, ont été utilisés: le milieu DV (Hughes & Lund, 1962), le milieu WC (Guillard & Lorenzen, 1972) et le milieu EBMII (Rimet, testé dans la collection de cultures de Thonon-les-Bains).

Après dilution des échantillons, différentes cellules ont été isolées à l'aide d'une micropipette effilée et d'un microscope inversé. Ces cellules ont été placées séparément dans un milieu composé à 50% d'un milieu de culture et 50% d'eau filtrée (0.2 µm) du site correspondant à l'échantillon. Après quatre à sept jours, les cultures ont été observées par balayage au microscope inversé afin de s'assurer du développement de la culture et de son caractère mono-spécifique. Les cultures ont ensuite été transférées dans 100% de milieu de culture, maintenue à 20°C dans une chambre de culture ayant un cycle jour : nuit de 15 : 9 heures, et repiquées mensuellement. Dans le cadre de cette thèse, 210 cultures ont été obtenues. Par ailleurs, 112 cultures, isolées lors d'un précédent projet, ont été transmises par le Centre de Recherche Public Gabriel Lippmann (Luxembourg) et 34 cultures de diatomées supplémentaires provenaient de la TCC (Thonon Culture Collection, http://www.inra.fr/carrtel-collection). La liste des cultures utilisées pour cette thèse est donnée en Annexe 1. Toutes ces cultures ont ensuite été transférées à la TCC.

Après avoir récolté suffisamment de cellules pour réaliser les analyses morphologiques et moléculaires, le développement des cultures a été ralenti en les stockant à 7°C avec un cycle jour : nuit de 12 : 12 heures afin de réduire la fréquence des repiquages (repiquage semestriel).

2. L'approche morphologique

Les diatomées en culture ont été identifiées morphologiquement en microscopie optique au laboratoire. Les études en microscopie électronique présentées dans le paragraphe CHAPITRE III.2 ont été réalisées par le Centre de Recherche Public Gabriel Lippmann (Luxembourg).

2.1. *Préparation des frustules*

Afin d'observer de manière optimale les frustules, le contenu cellulaire ainsi que les matières organiques entourant le frustule, doivent être éliminés. Le traitement traditionnel au peroxyde d'hydrogène et à l'acide chlorhydrique utilisé pour les échantillons environnementaux n'étant pas efficace sur des souches en culture, un autre protocole a été utilisé : les cellules ont été traitées par de l'acide nitrique (30%) et chauffées pendant environ dix heures. Après décantage, le surnageant a été enlevé et les frustules ont été rincés par trois cycles de dilution/décantation avec de l'eau distillée. Les frustules nettoyés ont été stockés dans l'eau distillée.

2.2. *Préparation de lames permanentes*

Pour observer les frustules, ceux-ci doivent être montés dans une résine de montage à indice de réfraction élevé. Pour cela, quelques gouttes de matériel nettoyé ont été placées sur des lamelles. Après évaporation du liquide, ces lamelles ont été retournées sur une goutte de mélange Naphrax©/Toluène déposé sur une lame. Le toluène a été évaporé sur une plaque chauffante à ~ 200 °C. Les lamelles ont ensuite été ajustées pour éliminer l'excès d'air et de résine et pour obtenir la répartition des diatomées dans un plan horizontal.

2.3. *Morphométrie traditionnelle*

Après solidification et refroidissement de la résine, les lames ont été observées au microscope Zeiss Axio Imager A1 couplé à une caméra Zeiss AxioCam 321 HRc pour photographier les frustules. Les principales caractéristiques morphologiques permettant la détermination des diatomées au niveau de l'espèce sont : la forme générale des frustules, la forme des pôles, la longueur, la largeur, la densité de strie et la densité de fibules. Les souches de diatomées ont donc été photographiées et ces caractéristiques ont été comparées à la bibliographie pour identifier les espèces de diatomées en culture.

2.4. *Morphométrie géométrique*

La morphométrie géométrique permet de décrire quantitativement la morphologie d'un spécimen en utilisant les coordonnées de points de repère ou « landmarks » relevés dans un espace bi ou tridimensionnel (Rohlf & Marcus, 1993). Cette technique a été appliquée dans un espace bidimensionnel à l'étude d'un complexe d'espèces : *Gomphonema parvulum* (CHAPITRE III.3.1). Le « package » de logiciels tps (Rohlf, 2007) permet de réaliser ces analyses morphométriques en utilisant la méthode « Thin Plate Spline » ou méthode des « plaques minces flexibles ». Les coordonnées des points de repères d'un spécimen sont placées sur une « plaque » théorique. L'énergie nécessaire pour déformer cette plaque jusqu'aux points d'un spécimen de référence, correspond à la variation morphologique.

Sur les images photographiées en microscopie optique, des points de repère ont été placés en utilisant le logiciel tpsDig (version 2.16). Pour l'étude des valves de *Gomphonema parvulum*, nous avons placé 25 points de repère : 10 autour des pôles et 15 dans l'aire centrale du frustule (Figure III.12, Article II).

Les coordonnées de ces points dépendent de la forme du spécimen mais également de sa position, de sa taille et de son orientation (Rohlf, 2002). Pour comparer différents spécimens, il est nécessaire de s'affranchir de ces paramètres en ajustant la conformation des différents spécimens par rapport à une conformation de référence. Dans ce but, une méthode de superposition procruste est utilisée. Nous avons appliqué à nos coordonnées la méthode itérative des moindres carrés généralisés, GLS, aussi appelée Generalized Procrustes Analysis, GPA (Rohlf, 2002). Cette méthode réalise une translation (pour éliminer les différences de positions), une standardisation par la « taille centroïde » (pour

éviter l'effet de la taille) et une rotation (pour réorienter les spécimens). Une conformation moyenne (consensus) est ensuite calculée. Ces deux étapes sont répétées jusqu'à ce que l'alignement des spécimens soit optimal et que les coordonnées du consensus ne varient plus.

TpsRegr (version 1.46) a ensuite été utilisé pour déterminer l'effet de l'allométrie sur la variation totale de la forme de nos spécimens. En effet, au cours des divisions végétatives, la taille des frustules de diatomées diminue et cette diminution de taille peut s'accompagner de modifications de forme du frustule (Cox, 2010). Ce type de variation n'est pas corrigé par la méthode GLS. Pour déterminer les variations entre groupes de diatomées, il était donc nécessaire de déterminer au préalable la part de variation due à la diminution de taille.

Ensuite, nous avons effectué une Analyse en Composantes Principales (ACP) sur les données morphométriques (les déformations relatives ou relative warps) en utilisant tpsRelw (version 1.46). L'ACP permet ainsi de visualiser les groupes de spécimens en fonction des variations morphologiques. Le logiciel tpsRelw permet également de visualiser les variations morphologiques (le positionnement des points de repères) le long d'un axe de l'ACP.

Après avoir éliminé la première composante principale qui était fortement corrélée avec les différences de taille, la séparation des clades a été testée par une analyse de variance multiple (MANOVA) et une analyse canonique des variables (CVA) avec Past version 2.06 (Hammer et al., 2001).

3. L'approche moléculaire

3.1. *Séquençage des souches*

Après la caractérisation morphologique, une caractérisation moléculaire a été réalisée sur les souches de diatomées. Dans ce but, plusieurs marqueurs nucléiques ont été séquencés par séquençage Sanger : les ADNr 18S et 28S ainsi que les ITS et le 5.8S de l'opéron ribosomal, le gène *rbc*L du génome chloroplastique et le gène *cox*1 du génome mitochondrial. Les séquences de ces fragments nucléiques ont ensuite été analysées.

3.1.1. Marqueurs

(a) Opéron ribosomal

Les ARNr ont un rôle central dans la synthèse protéique et sont présents universellement dans les cellules procaryotes et eucaryotes. L'opéron ribosomal, qui contient les gènes codant ces ARNr, présente un nombre variable de copies. Chaque répétition est composée du gène codant la petite sous-unité de l'ARNr (ADNr SSU, ou ADNr 18S), du gène codant pour l'ARNr 5.8S, du gène de la grande sous-unité de l'ARNr (ADNr LSU, ou ADNr 28S), et de trois régions intergéniques (Figure II.2). Ces régions intergéniques sont les ITS (Internal Transcribed Spacer) 1 et 2 entourant le 5.8S et l'ETS (External Transcribed Spacer) an amont du 18S (Figure II.2). Les différentes copies adjacentes de l'opéron ribosomal sont séparées par d'autres régions intergéniques, les NTS (Non Transcribed Spacer) (Hillis & Dixon, 1991).

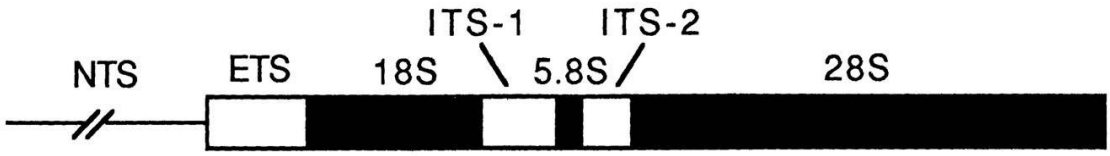


Figure II.2: *Structure de l'opéron ribosomal des eucaryotes*.
Source : Hillis & Dixon, 1991.

L'ADNr 18S est la région nucléique la plus largement utilisée pour étudier les relations phylogénétiques. Des milliers de séquences (partielles ou complètes) de procaryotes, eucaryotes unicellulaires et pluricellulaires peuvent être trouvées dans des bases de données telles que GenBank. Chez les diatomées, la taille de l'ADNr 18S est d'environ 1800 pb et son taux de substitution est d'environ 1% par 18 à 26 millions d'années (Kooistra & Medlin, 1996). Cette région présente une évolution lente (Mann & Evans, 2007) avec des domaines aux vitesses d'évolution variables (Elwood et al., 1985) et a été utilisée de ce fait pour étudier les relations phylogénétiques des diatomées à des niveaux taxonomiques relativement élevés : des principales classes de diatomées (Medlin et al., 1993; Medlin & Kaczmarska, 2004; Theriot et al., 2009, 2010) aux différentes familles et genres (Beszteri et al., 2001; Bruder & Medlin, 2007, 2008). Cette région offre

l'avantage d'être aisément séquencée grâce aux amorces universelles disponibles pour les eucaryotes et au nombre de copies important de ce gène dans le génome nucléaire. De plus, les bases de séquences, telles que GenBank, possèdent de nombreuses séquences de ce gène ce qui permet d'élargir le nombre de taxa de référence utilisés. Ainsi une région particulièrement variable de l'ADNr 18S, la région V4, a été proposée récemment comme barcode pour les diatomées (Zimmermann et al., 2011).

La région nucléique codant pour l'ARNr 28S est plus longue et comprend plus de domaines variables que celle du 18S. Mais en raison de sa grande taille (plus de 3300 pb), peu de séquences complètes sont disponibles pour les eucaryotes. En ce qui concerne les diatomées, différentes longueurs de fragment ont été séquencées : les domaines variables D1/D2 (e.g. Bruder & Medlin, 2007), D1/D3 (e.g. Lundholm et al., 2002) et D1/D4 (Kooistra et al., 2010). L'ADNr 28S semble ainsi fournir plus d'informations phylogénétiques que l'ADNr 18S (Van Der Auwera & De Wachter, 1998). Les séquences d'ADNr 28S ont été utilisées pour étudier les relations phylogénétiques entre les espèces de diatomées et au sein de complexes d'espèces (Sarno et al., 2007; Ellegaard et al., 2008; Trobajo et al., 2009; Poulíčková et al., 2010).

Les séquences des ITS sont transcrites mais sont ensuite excisées. La pression de sélection est ainsi moins forte sur les ITS que sur les régions codant les ARNr 18S et 28S. La vitesse d'évolution de ces séquences est donc supérieure aux autres régions de l'opéron ribosomal. Les séquences ITS des diatomées sont tellement variables qu'il est difficile d'aligner sans ambigüité certaines séquences (Behnke et al., 2004). Les ITS comprenant ici, l'ITS1, le 5.8S et l'ITS2, ne sont donc pas adaptés à l'étude des relations phylogénétiques profondes. Cependant ces séquences ont prouvé leur utilité, chez les diatomées, pour des analyses moléculaires intraspécifiques : des études génétiques de populations d'une même espèce et des études de biogéographie (Zechman et al., 1994; Orsini et al., 2004; Evans et al., 2007; Vanormelingen et al., 2007; Kaczmarska et al., 2008; Vanelslander et al., 2009). De plus, certaines variations de séquence et de structure secondaire de l'ITS2 des eucaryotes semblent corrélées à la définition biologique de l'espèce (Coleman, 2009). En effet, les organismes qui diffèrent par au moins 1 CBC (Compensatory Base Change) dans leurs séquences d'ITS2 ne sont pas interféconds. Cette corrélation a également été démontrée chez les diatomées (Amato et al., 2007). Un fragment 5.8S-ITS2 a d'ailleurs été proposé comme barcode pour les diatomées (Moniz & Kaczmarska, 2009). Cependant, les ITS posent un problème de variabilité

intragénomique entre les différentes copies de l'opéron ribosomal (Alvarez & Wendel, 2003), ce qui rend leur séquençage direct difficile et leur utilisation dans le cadre des approches de barcoding délicate.

(b) Gènes des organites

Chez les eucaryotes, le génome nucléaire n'est pas le seul génome de la cellule puisque les chloroplastes et les mitochondries possèdent également leur propre génome.

Du fait de la dégénérescence du code génétique (divers codons peuvent coder pour le même acide aminé), les taux d'évolution sont différents en fonction de la position de la base dans le codon. Le taux de mutation à la troisième position est plus élevé que le taux à la première ou seconde position, parce que les changements de nucléotides à la troisième position sont, dans la plupart des cas, des mutations neutres qui n'ont pas d'influence sur l'acide aminé codé. Les séquences codantes, des trois génomes, sont donc généralement plus facilement alignables que les séquences des gènes ribosomaux.

Le gène codant pour l'enzyme Rubisco (la ribulose-1,5-bisphosphate carboxylase) est responsable de la fixation de dioxyde de carbone dans le cycle de Calvin. Le gène codant la grande sous-unité de cette enzyme, le gène *rbc*L, se trouve en copie unique dans le génome du chloroplaste. Sa longueur est d'environ 1450 pb pour les diatomées, et les insertions ou délétions sont extrêmement rares (Soltis et al., 2000), ce qui facilite l'alignement des séquences. Ce gène a été utilisé pour des études phylogénétiques des diatomées en complément des gènes ribosomaux (Daugbjerg & Andersen, 1997a; Edgar & Theriot, 2004; Bruder & Medlin, 2007, 2008), mais également pour des études de complexes d'espèces (Trobajo et al., 2010). En comparaison de l'ADNr 18S qui est plus conservé, le *rbc*L semble plus adapté, chez les diatomées, à des études taxonomiques plus précises. Un fragment de ce gène a d'ailleurs été proposé comme barcode pour les diatomées (Hamsher et al., 2011).

Le gène codant pour la sous-unité 1 de la cytochrome oxydase (*cox*1) appartient au génome mitochondrial. Il a été utilisé pour des approches de barcoding chez les animaux (Hebert et al., 2003; Witt et al., 2006) et chez certaines algues (Saunders, 2005). En revanche, les séquences de *cox*1 ne sont pas appropriées pour la plupart des espèces de plantes supérieures en raison d'un taux d'évolution beaucoup plus lent chez ces dernières que chez les animaux.

Tableau II.2: *Liste des amorces utilisées pour amplifier et séquencer les souches de diatomées*.

	Cible	Amorces		Références
		Code	Séquence (5'-3')	
PCR	18S	1F	AACCTGGTTGATCCTGCCAGTA	Medlin et al., 1988
		1528R	CTTCTGCAGGTTCACCTAC	
	ITS	ITS 5F	GGAAGTAAAAGTCGTAACAAGG	White et al., 1990
		ITS 4R	TCCTCCGCTTATTGATATGC	
	28S	D1R	ACCCGCTGAATTTAAGCATA	Scholin et al., 1994
		D2C	CCTTGGTCCGTGTTTCAAGA	
	*rbc*L	DPrbcL1	AAGGAGGAADHHATGTCT	Daugbjerg & Andersen, 1997b
		DPrbcL7	AAASHDCCTTGTGTWAGTYTC	
	*cox*1	GazF2	CAACCAYAAAGATATWGGTAC	Saunders, 2005
		KEdtmR	AAACTTCWGGRTGACCAAAAA	Evans et al., 2007
Séquençage	18S	528F	GCGGTAATTCCAGCTCCAA	Elwood et al., 1985
		536R	AATTACCGCGGCKGCTGGCA	
		1055F	GGTGGTGCATGGCCGTTCTT	
		1055R	ACGGCCATGCACCACCACCCAT	
	*rbc*L	NDrbcL6	GTAAATGGATGCGTA	Daugbjerg & Andersen, 1997b
		15R	ACACCAGACATACGCATCCA	Jones et al., 2005
		16F	TTAGAAGATATGCGTATT	

Le gène *cox*1 est également utilisé pour des études phylogénétiques chez les diatomées (Ehara et al., 2000) et il a permis de distinguer les espèces de *Sellaphora* (Evans et al., 2007). Contrairement aux plantes, le *cox*1 des diatomées est suffisamment variable pour permettre de distinguer les taxa au niveau de l'espèce. Une limite à l'utilisation du *cox*1 est le faible nombre de séquences disponibles dans les banques de séquences. De plus, diverses études ont révélé la difficulté de séquençage des diatomées avec les amorces disponibles actuellement (Moniz & Kaczmarska, 2009; Trobajo et al., 2010; Hamsher et al., 2011).

3.1.2. Séquençage et analyse des séquences

Afin de séquencer les différents marqueurs choisis (ADNr 18S, ITS, ADNr 28S, *rbc*L et *cox*1), l'ADN de plusieurs souches de diatomées a été extrait. Pour cela un protocole d'extraction simplifié a été utilisé. Ce protocole reposait tout d'abord sur une lyse cellulaire, puis sur l'élimination des débris cellulaires par centrifugation, et enfin, sur la précipitation de l'ADN en utilisant le GenEluteTM-LPA (Sigma-Aldrich). Le protocole détaillé de l'extraction est disponible en Annexe 2. Les cinq marqueurs ont été amplifiés à l'aide des amorces PCR listées en Tableau II.2. Les programmes et les mélanges réactionnels des PCR sont également disponibles en Annexe 2. Après vérification de la qualité des amplicons sur gel d'agarose, les produits PCR ont été purifiés puis séquencés par Beckman Coulter Genomics (Takeley, United Kingdom), GATC (Konstanz, Germany) ou le Génoscope (Evry, France) selon le principe du séquençage Sanger (Sanger et al., 1977).

Cette technique est basée sur la synthèse d'ADN avec incorporation de dNTP, et de ddNTP. Les ddNTP ne possèdent pas de groupe hydroxyle en 3', et provoquent ainsi un arrêt de l'élongation du brin d'ADN (Figure II.3). La séparation des fragments néosynthétisés est ensuite effectuée en fonction de leur taille par électrophorèse capillaire. Marqués par un fluorochrome différent, chaque ddNTP est identifiable et la séquence d'ADN est lue à partir des résultats de migration. Actuellement, cette technique permet la lecture de séquences de 500 à 1000 pb (Ahmadian et al., 2006). Compte tenu de leurs tailles, les séquences des marqueurs 28S, ITS et *cox*1 étaient obtenues en une seule lecture (et donc un seul couple d'amorces), alors que les marqueurs 18S et *rbc*L ont nécessité, du fait de leur grande taille, l'utilisation d'amorces internes pour obtenir la totalité de la séquence (Tableau II.2).

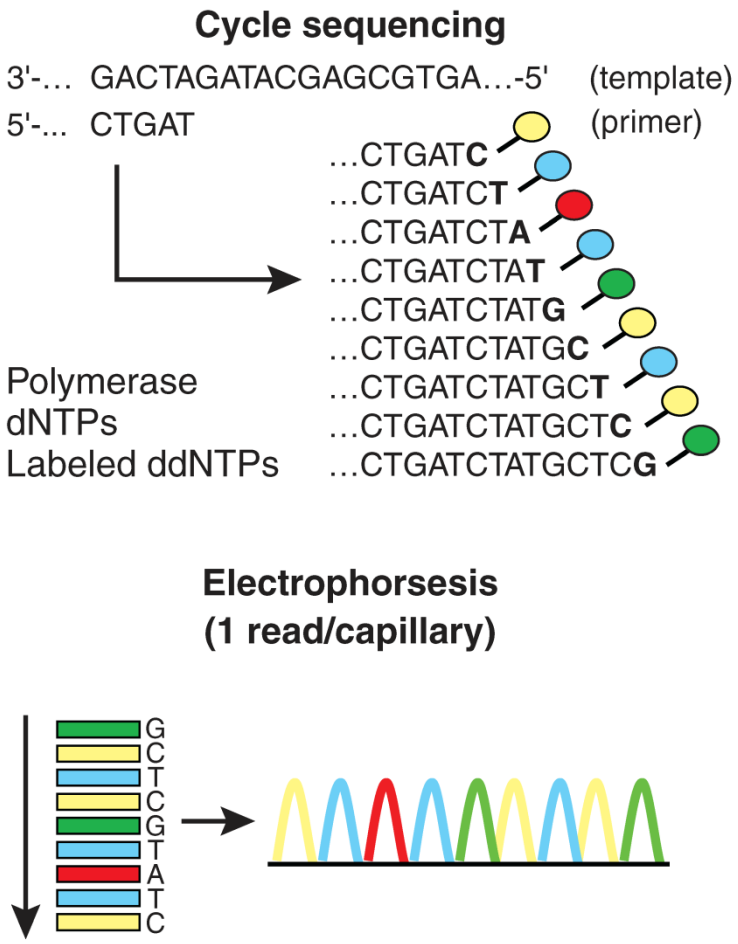


Figure II.3: *Principe du séquençage Sanger*.
Source : Shendure & Ji, 2008.

Les séquences (lues dans les deux directions) ont été assemblées, alignées en utilisant la procédure CLUSTAL W (Thompson et al., 1994) disponible dans le logiciel BioEdit v.7.0.5 (Hall, 1999), puis corrigées manuellement. Pour améliorer la robustesse de certaines analyses d'ADNr 18S, la structure secondaire de l'ARNr 18S a été utilisée pour aligner les séquences de 18S en utilisant le service d'alignement SILVA (Pruesse et al., 2007) disponible en ligne. Les alignements 18S ont ensuite été transférés dans le programme ARB (Ludwig et al., 2004). Différentes analyses ont enfin été réalisées sur les alignements en utilisant le programme MEGA5 (Tamura et al., 2011) : calcul de distance génétique, recherche de modèle d'évolution de séquence optimal et construction d'arbres phylogénétiques.

3.1.3. Bases de référence

Les bases de référence permettent d'associer des séquences nucléiques à des souches types bien identifiées. Nos bases de référence contiennent des séquences de diatomées obtenues à partir des cultures caractérisées au cours de la thèse et des séquences disponibles dans GenBank. Parmi ces dernières, une partie d'entre elles est de mauvaise qualité ou est associée à des souches mal identifiées ce qui peut introduire de nombreux biais dans les analyses ultérieures. Nous avons donc réalisé un tri important parmi les séquences porvenant de GenBank, sur la base de la qualité de ces séquences (présence ou absence de N), sur la qualité de l'identification (les séquences environnementales ont été éliminées), ainsi que sur leur mauvaise position dans un arbre phylogénétique, ce qui pouvait suggérer une mauvaise assignation taxonomique.

Nous avons donc créé trois bases de référence, soit une par marqueur (18S, *rbc*L et *cox*1) et donc par génome de diatomées (nucléaire, chloroplastique et mitochondrial). Ces bases ont été mises à jour régulièrement au cours de la thèse. Les versions finales étaient composées de 1412 séquences de 18S (dont 303 séquencées au cours de la thèse), 1071 séquences de *rbc*L (357) et 266 séquences de *cox*1 (96). Ces bases de séquences correspondaient respectivement à 508, 407 et 63 espèces.

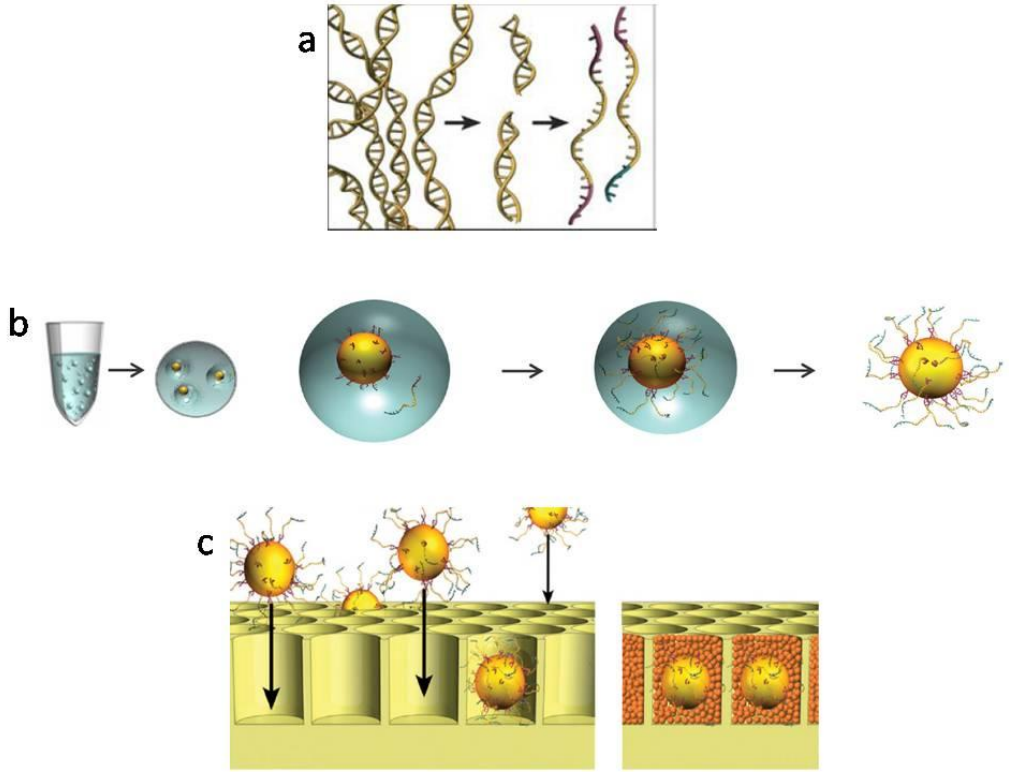


Figure II.4: *Préparation des échantillons pour le pyroséquençage*.
(a) Fragmentation de l'ADN, ajout d'adaptateurs et séparation de l'ADNdb en ADNsb ;
(b) PCR en émulsion; (c) Transfert des microbilles dans la PTP et ajout des enzymes (immobilisées sur des microbilles) nécessaires au pyroséquençage.
Source : (a) Margulies et al., 2005 ; (b) et (c) © Roche.

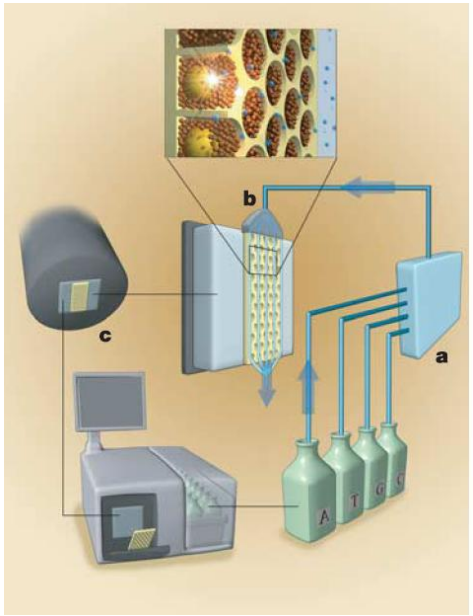


Figure II.5: *Appareil de pyroséquençage*.
(a) : assembleur de fluides; (b) : chambre de flux; (c) : caméra CCD couplée à un ordinateur.
Source : Margulies et al., 2005.

3.2. *Analyse par pyroséquençage 454*

3.2.1. Principe

Le principe du pyroséquençage 454 (Life Sciences) est différent du principe du séquençage Sanger. La technique de pyroséquençage est un séquençage caractérisé par la révélation en temps réel de la synthèse d'un brin d'ADN. La technologie 454 (Margulies et al., 2005) repose sur plusieurs avancées techniques: la PCR en émulsion, les plaques en fibre optique (PicoTiterplate, PTP), le pyroséquençage, et les technologies informatiques pour l'analyse des images.

Tout d'abord, les fragments d'ADN doivent être préparés. L'ADN génomique ou, dans notre cas, les amplicons de grande taille sont fragmentés par nébulisation puis des adaptateurs sont fixés, par ligation, aux extrémités des fragments (Figure II.4 : a). Des microbilles présentant à leur surface des amorces complémentaires à un des adaptateurs fixent une molécule d'ADNsb à la fois. Les microbilles liées à l'ADNsb sont mises en émulsion en présence des réactifs de PCR (Figure II.4 : b). Chaque gouttelette correspond à un « micropuits » qui contient une microbille liée à une molécule d'ADN. Dans chaque gouttelette est réalisée une amplification du fragment lié à la bille. Après amplification et tri, les microbilles porteuses de l'ADN amplifié sont transférées dans une plaque en fibre optique (PTP) contenant environ 1,6 millions de puits (Figure II.4 : c). Les puits possèdent un diamètre qui assure le dépôt d'une seule microbille par puits. La PTP est ensuite placée dans une chambre conçue pour créer un flux de réactifs de séquençage au-dessus des ouvertures des puits, et pour lire la réaction de pyroséquençage (Figure II.5).

Le réaction de séquençage se fait selon le principe du pyroséquençage en suivant l'activité de l'ADN polymérase, en temps réel, par bioluminescence (Nyrén, 1987; Hyman, 1988; Ronaghi, 1996; Ronaghi et al., 1998). L'ADNsb à séquencer est hybridé avec une amorce de séquençage. Ce fragment hybridé est incubé avec un mélange d'enzymes et de réactifs immobilisés sur des microbilles (Figure II.4 : c). Une cascade enzymatique, impliquant une ADN polymérase, une ATP sulfurylase, une luciférase et une apyrase (Figure II.6), permet de suivre la synthèse du brin d'ADN complémentaire au brin d'intérêt (Ahmadian et al., 2006).

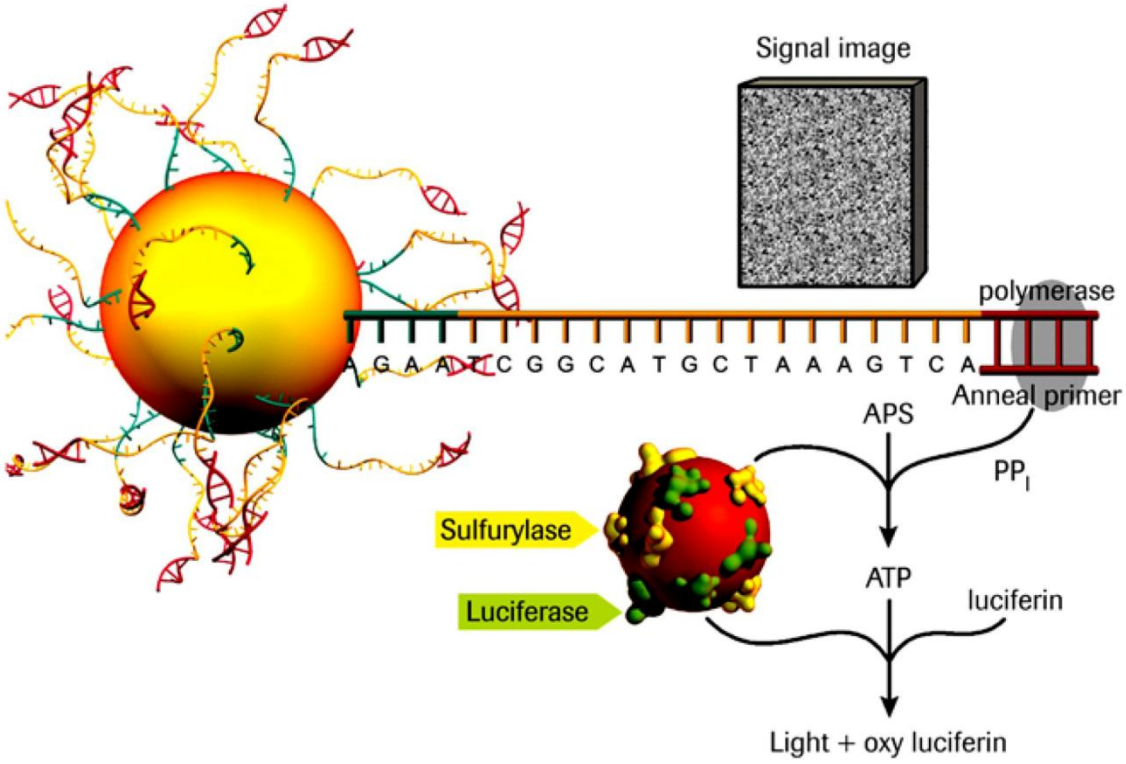


Figure II.6: *Représentation schématique du système enzymatique du pyroséquençage*.
Source : © Roche.

Les dNTP ne sont pas ajoutés tous ensemble comme dans une réaction de séquençage Sanger, mais l'un après l'autre par flux successifs. Si le dNTP ajouté dans le milieu réactionnel correspond à celui attendu par l'ADN polymérase (complémentaire de la base du fragment d'intérêt), il est incorporé dans le brin en cours de synthèse tout en libérant un pyrophosphate inorganique (PPi) (Figure II.6). Les PPi libérés sont proportionnels en molarité au nombre de dNTP incorporés (Ronaghi, 1996). Les PPi sont transformés en ATP par l'ATP sulfurylase (Figure II.6). En présence de luciférine, les ATP sont utilisés par la luciférase pour produire de l'Oxyluciférine et émettre un signal lumineux qui peut être capté par une caméra CCD (Ronaghi et al., 1998). L'intensité du signal lumineux est fonction du nombre de nucléotides incorporés sur le brin d'ADN en synthèse (Margulies et al., 2005). Enfin les dNTP non incorporés ainsi que l'ATP sont dégradés par un flux contenant une apyrase. Un nouveau cycle d'addition d'un autre dNTP peut alors recommencer.

La caméra CCD (Figure II.5) permet de capturer les images après addition de chaque nucléotide. Connaissant l'ordre dans lequel les 4 nucléotides sont ajoutés automatiquement, l'analyse des différentes images capturées permet de déduire la séquence des différents fragments d'ADN d'intérêt (illustré par un pyrogramme ou flowgram).

Une séquence obtenue par pyroséquençage sera appelée « read » dans la suite du manuscrit pour la différencier plus aisément des séquences obtenues par séquençage Sanger.

3.2.2. Echantillons de communautés de diatomées

(a) Echantillons artificiels

Dans le but de développer des outils moléculaires, nous avons simulé des échantillons environnementaux en créant trois échantillons artificiels à partir de souches de diatomées. La complète maîtrise de la composition des échantillons était nécessaire pour évaluer la qualité de notre approche basée sur l'utilisation du pyroséquençage. Ces échantillons artificiels ont été créés à partir de 30 souches de diatomées représentant 12 genres et 21 espèces. Ainsi, certains genres étaient représentés par plusieurs espèces et certaines espèces par plusieurs clades différents, afin de simuler la diversité présente dans un échantillon naturel, et d'estimer la précision d'identification de l'outil moléculaire.

Tableau II.3: *Composition des échantillons artificiels.*

Code	Taxon name	Mix C	Mix PCR1	Mix PCR2
TCC 477	*Amphora montana*	1,8	1,8	4,8
TCC 501	*Cocconeis placentula*	0,6	0,6	4,8
TCC 582	*Cyclotella meneghiniana*	2,3	2,3	4,8
TCC 508	*Fistulifera saprophila*	1,9	1,9	4,8
TCC 547	*Fragilaria capucina*	1,7	1,7	4,8
TCC 452	*Gomphonema bourbonense*	2,9	0,4	4,8
TCC 458	*Gomphonema bourbonense*		2,9	2,4
TCC 527	*Gomphonema clavatum*	1,7	1,7	4,8
TCC 507	*Gomphonema clevei*	0,5	0,5	4,8
TCC 429	*Gomphonema parvulum* clade A	26,4	5,3	1,0
TCC 426	*Gomphonema parvulum* clade B		5,3	1,0
TCC 492	*Gomphonema parvulum* clade B		5,3	1,0
TCC 595	*Gomphonema parvulum* clade C		5,3	1,0
TCC 612	*Gomphonema parvulum* clade C		5,3	1,0
TCC 536	*Gomphonema pumilum*	0,8	0,4	2,4
TCC 540	*Mayamaea permitis*	3,9	3,9	4,8
TCC 515	*Navicula cryptocephala*	6,5	6,4	4,8
TCC 537	*Nitzschia acidoclinata*	10,9	5,5	2,4
TCC 543	*Nitzschia acidoclinata*		5,5	2,4
TCC 521	*Nitzschia cf. frustulum*	1,9	1,9	4,8
TCC 700	*Nitzschia draveillensis*	0*	0,1	4,8
TCC 481	*Nitzschia inconspicua*	2,6	1,3	2,4
TCC 487	*Nitzschia inconspicua*		1,3	2,4
TCC 516	*Nitzschia lorenziana*	0*	0,1	4,8
TCC 435	*Nitzschia palea clade B*	26,5	8,8	1,6
TCC 570	*Nitzschia palea clade G*		8,8	1,6
TCC 583	*Nitzschia palea clade J*		8,8	1,6
TCC 472	*Pinnularia* acrosphaeria	1,8	1,8	4,8
TCC 461	*Sellaphora seminulum*	4,6	4,6	4,8
TCC 520	*Ulnaria ulna*	1,0	1,0	4,8

Dans le but d'évaluer les biais potentiels dus à l'extraction d'ADN et aux étapes de PCR, un premier échantillon artificiel a été créé avant ces deux étapes, alors que deux autres ont été réalisés après l'étape de PCR. Le mélange « avant extraction », mix C, a été réalisé en mélangeant 30 souches de diatomées. Ce mélange a ensuite été récolté par centrifugation et séparé en deux : un sous-échantillon pour l'analyse moléculaire et un sous-échantillon pour l'analyse morphologique. L'ADN total a été extrait et les trois marqueurs ont été séquencés à partir de ce mélange. Pour évaluer la composition de ce mélange par microscopie, nous avons traité le mélange de cultures à l'acide nitrique, de la même manière que pour des souches pures (paragraphe II.2.1), puis nous avons réalisé deux comptages de plus de 400 valves en microscopie optique (Tableau II.3).

Les deux autres échantillons artificiels ont été constitués après extraction et PCR. Chacune des 30 cultures a été centrifugée indépendamment et leurs ADN ont été extraits et amplifiés séparément. Nous avons ensuite effectué des mélanges des 30 amplicons, les deux échantillons se distinguant par les proportions relatives de chaque espèce. Nous avons dosé les amplicons au NanoDrop 1000 avant de faire les mélanges pour créer des échantillons aux proportions connues (Tableau II.3). Pour le mix PCR1, les 30 produits PCR ont été mélangés en différentes proportions avec l'objectif de détecter si l'ADN présent en faible proportion pouvait être détecté par pyroséquençage, alors que le mix PCR2 a été créé en mélangeant les 30 produits PCR en proportions égales.

Nous n'avons pas réalisé de mélange entre les étapes d'extraction et de PCR car, nos cultures n'étant pas axéniques, il n'était pas possible d'évaluer exactement les proportions d'ADN de chacune des souches de diatomées. Nous n'avons donc pas pu distinguer les biais dus à l'extraction d'ADN de ceux dus à l'amplification par PCR.

(b) Echantillons naturels

Pour tester la méthode moléculaire développée au cours de la thèse, quatre échantillons naturels ont été utilisés. Ces échantillons benthiques ont été prélevés en brossant la surface de plusieurs pierres dans quatre rivières différentes, selon la norme DCE.

Tableau II.4 : *Description des sites d'échantillonnage des échantillons environnementaux*

	LACL	CHLO	COIN	BDNA
Rivière	Le Lay	Chiers	Coconi	Saint Denis
Site	La Clay	Longlaville	intermédiaire	Amont prise AEP
Date de prélèvement	17.08.2009	17.06.2009	18.04.2009	28.04.2009
Nombre de taxa	56	29	61	21
Effectif	400	413	410	500
Diversité (indice de Shannon)	4,62	3,55	5,18	2,90
IBD	12,2	6,3	13,8	16,7
IPS	11,5	6,7	12,8	12,9
Température de l'Eau (°C)	26,5	14,5	25,6	21,7
pH (unité pH)	8,5	7,6	7,3	8,1
Conductivité (µS/cm)	493,0	512,0	129,0	98,3
Oxygène dissous (mg(O2)/L)	7,3	7,2	9,2	8,9
Taux de saturation en O2 (%)	92	71	122	102

Deux échantillons ont été récoltés dans des rivières de France métropolitaine (Le Lay et la Chiers) où la flore des diatomées est étudiée depuis de nombreuses années et donc relativement bien connue. Les deux autres échantillons ont été prélevés dans des rivières des îles tropicales françaises de l'Océan Indien (la rivière Coconi à Mayotte, et la rivière Saint Denis à La Réunion) où les diatomées sont moins connues.

Ces échantillons on été choisis en fonction des diatomées préalablement préparées et déterminées par microscopie optique selon les recommandations de la norme NF EN 13946 (AFNOR, 2003). Les diatomées de ces échantillons ont donc été nettoyées avec 35-40% de peroxyde d'hydrogène puis avec 37% d'acide chlorhydrique. Après rinçage, les frustules nettoyés ont été montés avec de la résine synthétique (Naphrax ©) et des lames ont été préparées pour la microscopie optique. Pour déterminer la composition spécifique de la communauté de diatomées, 400 valves ont été identifiées et comptées par Asconit Consultants et la DREAL.

Pour chaque zone géographique (tempérée et tropicale), nous avons choisi un échantillon présentant une faible richesse spécifique (CHLO et BDNA) et un échantillon présentant une forte richesse (LACL et COIN) (Tableau II.4). De plus, un duplicat de l'échantillon provenant de France métropolitaine et présentant la plus forte diversité (LACL) a été réalisé pour tester la reproductibilité de la méthode. Les cinq échantillons ont ensuite été analysés par pyroséquençage à la plateforme génomique du Génopole Toulouse/Midi-Pyrénées sur un appareil GS FLX Titanium PicoTiterPlate 454 (Roche), selon les recommandations du fabricant et avec l'aide d'Eugénie Robe.

3.2.3. Mise au point du protocole

L'analyse d'échantillons complexes (communautés) est soumise à différents biais techniques, c'est pourquoi il est nécessaire de réaliser des tests préalables pour tenter de les identifier.

Différentes méthodes d'extraction ont donc été testées, afin de choisir une méthode présentant un bon rendement d'extraction, tout en permettant de conserver la diversité présente dans l'échantillon. Quatre méthodes d'extraction ont été évaluées :

- la méthode d'extraction utilisée pour extraire l'ADN des cultures (méthode utilisant le GenEluteTM-LPA, Sigma-Aldrich),

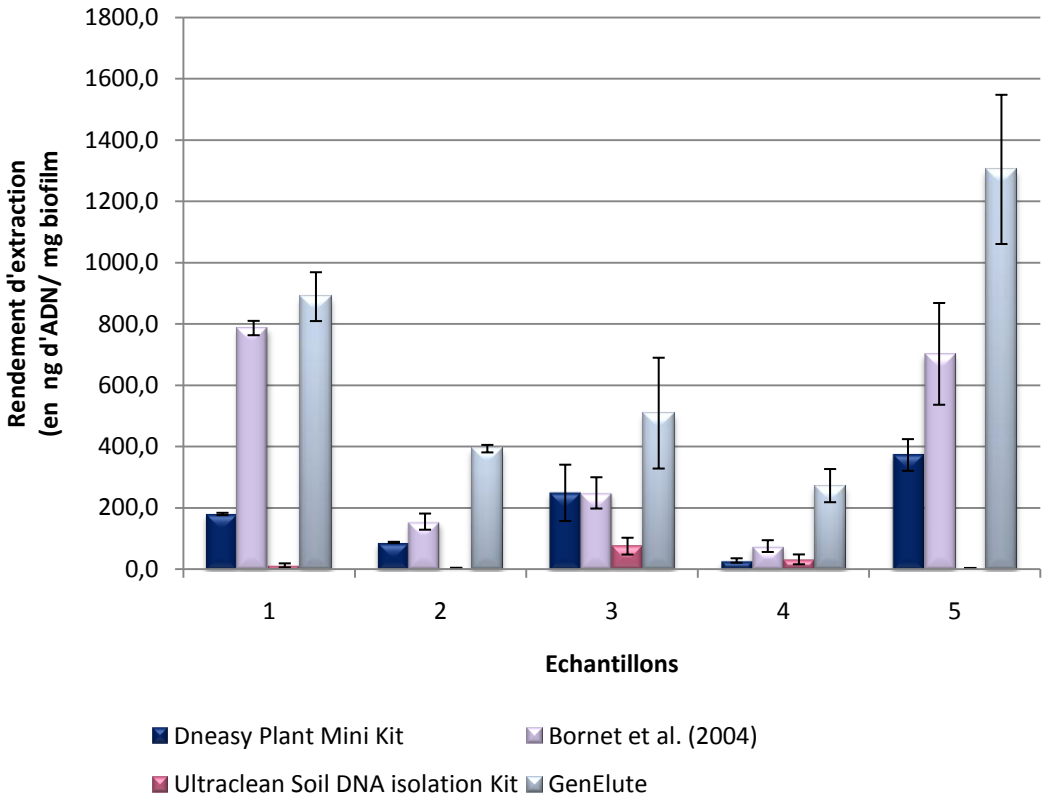


Figure II.7: *Rendements d'extraction moyens des 4 méthodes testées*.

- une méthode d'extraction légèrement modifiée à partir de Bornet et al. (2004),
- un kit d'extraction d'ADN de plantes: DNeasy Plant Mini Kit (Qiagen),
- et un kit d'extraction d'ADN issu du sol : UltraClean Soil DNA Isolation Kit (MoBio).

Pour comparer ces méthodes d'extraction, trois réplicats de cinq échantillons (quatre échantillons naturels et un mélange de cultures) ont été utilisés. A partir des échantillons, un culot de cellules a été obtenu par centrifugation. Les échantillons ont ensuite été séparés pour les trois réplicats des quatre méthodes d'extraction. Les rendements obtenus à partir des différents protocoles sont présentés en Figure II.7. La méthode présentant le rendement le plus important était la méthode utilisant le GenEluteTM-LPA, suivie de la méthode modifiée de Bornet et al. (2004). Sachant que les méthodes d'extraction peuvent donner une image déformée de la composition de la communauté étudiée (Carrigg et al., 2007; Feinstein et al., 2009), nous avons réalisé une DGGE (Denaturing Gradient Gel Electrophoresis) en utilisant un fragment de l'ADNr 18S, sur les produits de PCR obtenus avec chaque méthode d'extraction et pour chaque réplicat. Les rendements d'amplification révélés sur gel d'agarose ont montré une meilleure amplification pour la méthode dérivée de celle de Bornet et al. (2004). La DGGE n'ayant révélé aucune différence de diversité pour les différentes méthodes d'extraction, nous avons choisi la méthode modifiée à partir de Bornet et al. 2004. En effet, le rendement d'extraction était moins important que pour la méthode de GenElute mais le rendement d'amplification était légèrement meilleur.

Enfin un test des conditions d'amplification a été réalisé pour ne pas atteindre le plateau PCR, car il est préférable d'arrêter l'amplification durant la phase exponentielle d'amplification, pour ne pas modifier les proportions des différents taxa. Différents nombre de cycles d'amplification ont donc été testés pour les différents marqueurs. Pour l'ADNr 18S, l'amplification choisie était de 28 cycles, 32 cycles pour le *rbc*L et 30 cycles pour le *cox*1. Les conditions d'amplification des trois marqueurs sont disponibles en Annexe 2.

3.2.4. Algorithmes metaMatch

Plusieurs algorithmes de comparaison de séquences sont actuellement disponibles. Beaucoup utilisent des heuristiques qui permettent une comparaison rapide qui peut se

faire au détriment de l'exactitude de la comparaison. Nous avons fait le choix de privilégier l'exactitude de la comparaison et donc l'utilisation de nouveaux algorithmes. Les algorithmes de comparaison de séquences « *metaMatch* », utilisés au cours de cette thèse, ont été développés par Alain Franc et Philippe Chaumeil (INRA UMR BIOGECO, Bordeaux). Ces programmes ont été créés dans le but d'obtenir des inventaires précis des taxa présents dans un échantillon environnemental et dans une base de référence, à partir de reads obtenus par pyroséquençage d'échantillons environnementaux.

Plusieurs versions de *metaMatch* sont disponibles et chacune est composée de deux programmes. Un premier programme compare chaque read aux séquences de référence et le second lie ce résultat aux informations taxonomiques des séquences de référence pour obtenir l'identification taxonomique du read.

Le premier programme *metaMatch* est codé en langage C. Quatre versions existent :

- *metaMatch*_PS : P, « Perfect » pour correspondance parfaite ; et S, « Single core » car il fonctionne sur un processeur simple ;
- *metaMatch*_PM : P, « Perfect » ; et M, « Multi-cores » car il est parallélisé et fonctionne sur des clusters de calcul;
- *metaMatch*_IS : I « Imperfect » pour correspondance imparfaite et S, « Single core » ;
- *metaMatch*_IM : I « Imperfect » et M, « Multi-cores »;

Chacun de ces programmes est accompagné de son programme *metaMatch*_Taxo (*metaMatch*_PS_Taxo, *metaMatch*_PM_Taxo, *metaMatch*_IS_Taxo et *metaMatch*_IM_Taxo) qui est codé en R.

Les algorithmes *metaMatch*_PS et *metaMatch*_IM ont été utilisés pour étudier les échantillons artificiels et environnementaux (CHAPITRE IV) et sont donc présentés en détails.

*metaMatch*_PS recherche uniquement les correspondances parfaites (100% d'identité) entre les reads issus du pyroséquençage et les séquences de la base de référence. *metaMatch*_PS fonctionne en utilisant l'invite de commandes disponible sous Windows. Pour l'utilisation de *metaMatch*_PS, deux fichiers sous format Fasta sont nécessaires : un fichier Fasta correspondant aux reads (issus du pyroséquençage) et un fichier Fasta correspondant aux séquences de référence. *metaMatch*_PS compare chaque read à chaque

séquence de la base. Toute insertion, délétion ou mutation entre un read et une séquence de référence, provoque une correspondance imparfaite. *metaMatch*_PS produit une matrice contenant en ligne tous les reads et en colonne toutes les séquences de la base de référence. Dans chacune des cellules sont indiqués les résultats de la comparaison avec « -1 » si le read n'a pas de correspondance parfaite avec la séquence de référence, et la position de la première base commune (position sur la séquence de référence) si le read correspond parfaitement à une séquence de référence. *metaMatch*_PS écrit ensuite le fichier correspondant à la matrice en ne conservant que les reads ayant au moins une correspondance avec une séquence de la base de référence.

Ce fichier de sortie, ainsi qu'un fichier contenant les informations taxonomiques des séquences de référence sont lus par *metaMatch*_PS_Taxo. Pour chaque read, *metaMatch*_PS_Taxo fournit la liste des noms de taxa correspondant. Un read peut correspondre parfaitement à plusieurs séquences de référence. Si toutes ces séquences de référence appartiennent au même taxon, ce taxon est considéré comme « informatif » et le nom de ce taxon est attribué au read. Sinon, le read est considéré comme « non informatif » et la liste des taxa correspondants au read est attribuée. Enfin, une liste des reads et des assignations est créée.

La seconde version de *metaMatch* a été enrichie d'un module calculant les distances entre les reads et les séquences de la base de référence et correspond aux *metaMatch*_I*. Nous avons utilisé *metaMatch*_IM sur 10 ou 300 « cœurs de calcul ». Pour l'utilisation de *metaMatch*_IM, les deux mêmes fichiers sous format Fasta sont nécessaires (un fichier Fasta de reads et un fichier Fasta de séquences de référence). *metaMatch*_IM compare chaque read à chaque séquence de la base et recherche d'abord la présence d'un « mot » commun de plus de 30 pb. Si ce mot commun n'est pas trouvé, la recherche de correspondance est arrêtée pour ce read. Si un mot commun est trouvé, *metaMatch*_IM calcule la distance génétique entre ce read et la séquence de référence. Toute base différente (mutation, insertion, délétions) entre un read et une séquence de référence entraîne une distance de « 1 ». Ainsi la distance génétique due à une insertion ou une délétion est proportionnelle à sa longueur (c'est-à-dire qu'il n'y a pas de différence de valeur entre l'ouverture d'un « gap » et son extension). *metaMatch*_IM génère une matrice par cœur de calcul utilisé, dans notre cas, 10 ou 300 matrices. Ces matrices (fichiers ASCII) contiennent en ligne les reads et en colonne les séquences de la base de référence. Dans chacune des cellules sont indiqués les résultats de la comparaison avec

« -1 » si le read n'a pas de correspondance avec la séquence de référence, et la distance génétique si le read correspond à une séquence de référence.

*metaMatch*_IM_Taxo relie le fichier contenant la taxonomie des séquences de référence et les matrices issues de *metaMatch*_IM. L'utilisation d'un seuil d'identité de séquence pour sélectionner les reads est nécessaire pour faire fonctionner *metaMatch*_IM_Taxo. Pour chaque read, *metaMatch*_IM_Taxo donne la liste des séquences ayant une identité de séquence (calculée en fonction de la distance génétique et de la longueur du read) supérieure au seuil défini. Enfin, deux inventaires des taxa détectés (nom du taxon et le nombre de reads correspondant à ce taxon) sont créés. Ces deux inventaires diffèrent par leur critère de sélection des reads. Pour l'inventaire « all », un nom de taxon est assigné à un read, si toutes les séquences de référence sélectionnées (c'est-à-dire présentant une homologie supérieure au seuil) appartiennent au même taxon. Si différents noms de taxa apparaissent dans la liste créée par *metaMatch*_IM_Taxo, le read n'est pas utilisé dans l'inventaire. Cet inventaire permet de minimiser les erreurs d'identification dues aux erreurs de séquençage ou aux erreurs d'identification des séquences de référence. En ce qui concerne l'inventaire « best », le nom du taxon assigné au read est le nom correspondant à la séquence qui présente la plus forte homologie avec le read (au dessus du seuil défini). Si l'identité de séquence la plus forte correspond à deux taxa différents, ce read est rejeté car il est « non informatif ». La sélection des reads est moins stricte pour l'inventaire « best » que pour l'inventaire « all » mais l'inventaire « best » permet de retenir plus de reads correspondant à des taxa morphologiquement différents mais présentant des séquences proches. En effet, la définition d'un seuil d'identité de séquence indiquant la séparation des taxa est difficile. Tous les taxa de diatomées n'ont pas évolué à la même vitesse, et les taxa proches peuvent être plus ou moins différenciés génétiquement. Il a par exemple été démontré que la vitesse d'évolution du 18S n'est pas la même pour toutes les diatomées (Kooistra & Medlin, 1996). Le nombre de reads retenu pour chaque taxon par l'inventaire « best » est donc moins biaisé par ces différences que le nombre retenu par l'inventaire « all ». Au cours de la thèse, nous n'avons pas optimisé la quantification des reads, et nous avons donc uniquement utilisé l'inventaire « all » (Article V).

CHAPITRE III. TAXONOMIE DES DIATOMEES & APPROCHES MOLECULAIRES

1. Introduction

La classification traditionnelle des diatomées était à l'origine une aide à l'identification plus qu'une description des relations entre les diatomées. Les caractères utilisés pour l'identification continuent ainsi d'influencer les descriptions taxonomiques (Cox, 2009). Dans les flores, les différentes classifications disponibles pour certaines identifications de diatomées, sont une des sources de désaccords entre les taxonomistes (Mann et al., 2010). Les divisions ou regroupements d'espèces et de genres (très courants dans la taxonomie des diatomées) créent une grande confusion lors de l'identification. Ainsi, dans le but d'utiliser les diatomées en tant qu'indicateur écologique de la qualité des eaux et de standardiser l'utilisation des indices diatomées, il est essentiel d'utiliser une systématique stable. Pour tendre vers une telle classification, il est nécessaire de développer nos connaissances sur les différentes espèces des écosystèmes à surveiller. Jahn et al. (2008) et Evans & Mann (2009) ont indiqué que les séquences ADN peuvent être utilisées pour stabiliser la taxonomie des diatomées. L'étude des relations phylogénétiques au sein des diatomées permet de mieux définir les relations entre les taxa, et par conséquent, de stabiliser les classifications utilisées pour leur identification dans le cadre de la bioindication.

De nombreuses espèces de l'Ordre des Cymbellales et de la famille des Bacillariaceae sont abondamment représentées dans les échantillons benthiques d'eaux douces. Cependant plusieurs complexes ou groupes de taxa, par exemple *Encyonema minutum/silesiacum, G. tergestinum/rosentockianum*, *Nitzschia palea s.l.* etc., posent de grandes difficultés d'identification en microscopie optique et parfois même en microscopie électronique. Les relations phylogénétiques de certaines espèces ne sont pas toujours claires car leur taxonomie a régulièrement été soumise à de profonds changements. Par exemple, la validité du genre *Gomphoneis*, défini par Cleve (1894), fait toujours débat au sein de la communauté des diatomistes.

Trois exemples d'utilisation des outils moléculaires, pour améliorer la taxonomie à des niveaux taxonomiques différents, sont présentés : une étude de la phylogénie des deux principales familles des Cymbellales (Article I), une étude d'un complexe d'espèces appartenant à cet Ordre : *Gomphonema parvulum* (Article II), et une étude d'un second complexe d'espèces appartenant à la famille des Bacillariaceae : *Nitzschia palea*.

2. Etude de la taxonomie de l'Ordre des Cymbellales.

Les données de cet article (souches, séquences, et photographies) ont été acquises lors d'un précédent projet au Centre de Recherche Public Gabriel Lippmann, sous la direction de Luc Ector. Les analyses ainsi que la rédaction de l'article ayant été réalisées durant ma thèse et les résultats étant bien intégrés à mon sujet de thèse, ces derniers s'inscrivent naturellement dans ce manuscrit.

2.1. *Article I : Présentation générale de l'étude et synthèse des principaux résultats.*

L'ordre des Cymbellales défini par D.G. Mann est composé de quatre familles : Anomoeoneidaceae D.G. Mann, Cymbellaceae Greville, Gomphonemataceae Kützing, et Rhoicospheniaceae Chen & Zhu (Round et al., 1990). Ces familles et les genres qui les composent sont définis sur la base de la morphologie des frustules. Les membres des Cymbellales montrent de grandes similarités au niveau de leurs caractéristiques cytoplasmiques et cellulaires mais présentent également une grande diversité de formes et de symétries (Cox, 2002). Des méthodes cladistiques ont précédemment été appliquées sur les Cymbellaceae et les Gomphonemataceae pour préciser leurs relations phylogénétiques (Kociolek & Stoermer, 1988, 1993). Ces résultats ont amené Cox (2002) à proposer un scénario d'évolution des genres de l'Ordre des Cymbellales qu'il serait intéressant de tester par une approche moléculaire. La question à laquelle nous avons tenté de répondre était tout simplement la suivante: quelles sont les relations phylogénétiques au sein des deux principales familles de Cymbellales : les Cymbellaceae et les Gomphonemataceae ?

Cette étude s'appuie sur l'étude des séquences de l'ADNr 18S et sur l'étude de la morphologie de plusieurs taxa par microscopie optique et électronique. A partir de populations de diatomées benthiques récoltées dans des cours d'eau en Espagne, en Italie, au Luxembourg et au Portugal, 20 cultures monoclonales appartenant à différents genres de Cymbellaceae (*Cymbella*, *Encyonema*) et de Gomphonemataceae, (*Gomphoneis*, *Gomphonema* et *Reimeria*) ont été obtenues. Les ADN de ces souches ont été extraits, amplifiés et séquencés afin d'obtenir leurs séquences d'ADNr 18S. Après plusieurs échecs de mise en culture de *Didymosphenia geminata*, nous avons choisi de réaliser une amplification de l'ADNr 18S directement sur une cellule pour obtenir la séquence de

cette espèce. Les séquences obtenues à partir des cultures et de la cellule de *D. geminata* ont été comparées aux séquences disponibles dans la base « SSU database Silva 102 ». Nos séquences et les séquences proches disponibles dans cette base ont été alignées en fonction de leurs structures primaires et secondaires grâce à l'outil d'alignement disponible dans le logiciel Arb (Ludwig et al., 2004). Les souches, ainsi que la population d'origine de la cellule de *D. geminata* ont par ailleurs été étudiées et photographiées en microscopie optique et électronique à balayage, afin de pouvoir discuter des critères morphologiques les plus importants permettant de définir les genres et les familles.

L'analyse de l'ADNr 18S a mis en évidence plusieurs groupes, certains correspondants aux groupes créés sur la base de la morphologie des frustules et d'autres ne reflétant pas ces groupes.

Par exemple, les souches appartenant aux genres *Encyonema,* qui partagent une même symétrie, l'absence de champs apicaux de pores et la présence d'un intermissio, forment un groupe monophylétique. Il en est de même pour les souches du genre *Placoneis*.

Au contraire, l'espèce *Gomphonema rosenstockianum*, qui présente le même type de symétrie que les autres espèces de *Gomphonema,* est génétiquement séparée de ces dernières. La séquence de l'ADNr 18S d'une autre espèce de *Gomphonema* (*G. micropus*) présente un plus fort pourcentage d'identité avec la séquence d'ADNr 18S de *Reimeria sinuata* qu'avec celles des autres espèces de *Gomphonema*. Pourtant, ces deux espèces montrent peu de caractéristiques identiques : leurs symétries, leurs stries et le nombre de champs apicaux de pores sont différents. Ces deux espèces partagent cependant la présence d'un stigma, d'un intermissio ainsi que des champs apicaux de pores non traversés par le raphé. En outre, elles montrent une absence de strie ou une strie très courte au centre de la valve du côté opposé au stigma. Les résultats du séquençage de l'ADNr 18S mettent également en doute la validité du genre *Gomphoneis* car la séquence de *G. minuta* est insérée au sein d'un groupe comprenant les séquences d'espèces de *Gomphonema*.

Concernant les espèces du genre *Cymbella*, celles-ci sont toutes regroupées dans un même clade. Cependant, celles-ci sont divisées en deux sous-clades dont un comprend également les séquences de *Cymbopleura naviculiformis* et de la cellule de *Didymosphenia geminata*. Ces genres présentent beaucoup de caractéristiques différentes au niveau de leurs symétries, du nombre de stigma, de la présence de champs apicaux de

pores, mais sont toutes dépourvues d'un intermissio. Nos résultats montrent donc que le genre *Cymbella* est paraphylétique.

En conclusion, les résultats de cette étude montrent que l'Ordre des Cymbellales devrait être organisé selon de nouveaux critères morphologiques puisque les principales familles, Cymbellaceae et Gomphonemataceae, telles qu'elles sont définies actuellement, apparaissent paraphylétiques. Or, en accord avec Williams & Kociolek (2007), les groupes paraphylétiques sont artificiels et ne devraient pas être admis dans une classification. Par conséquent, soit les espèces *D. geminata* et *Cymbopleura naviculiformis* doivent être incluses dans le genre *Cymbella*, soit les espèces de *Cymbella* qui forment un groupe avec *D. geminata* et *Cymbopleura naviculiformis* doivent être combinées dans un nouveau genre, séparé des autres espèces de *Cymbella*. De plus, les espèces *Gomphonema micropus* et *G. rosenstockianum* devraient faire l'objet d'une étude plus précise, incluant d'autres marqueurs moléculaires, pour éclaircir leurs positions phylogénétiques et ainsi définir le genre auquel elles appartiennent.

2.2. *Article I publié dans Diatom Research 26 (3) : 305-315*

A preliminary phylogenetic analysis of the Cymbellales based on 18S rDNA gene sequencing

Lenaïg Kermarrec[1,2,3*], Luc Ector[3], Agnès Bouchez[1], Frédéric Rimet[1] & Lucien Hoffmann[3]

[1]*Institut National de la Recherche Agronomique, Unité Mixte de recherche, Centre Alpin de Recherche sur les Réseaux Trophiques et les Ecosystèmes Limniques, 75 avenue de Corzent, F-74203 Thonon cedex, France*

[2]*Asconit Consultants, 3 boulevard Clairfont, F-66350 Toulouges, France*

[3]*Public Research Centre - Gabriel Lippmann, Department of Environment and Agro-Biotechnologies (EVA), 41 rue du Brill, L-4422 Belvaux, Luxembourg*

Running title: Phylogenetic evaluation of Cymbellales

*Corresponding author. Email: lenaig.kermarrec@asconit.com

Abstract

Although diatom taxa have been observed and described for many years using light and electron microscopy, several taxa have called for some clarifications and taxonomic reassessments. This is the case for the order Cymbellales D.G. Mann, which is widely represented in freshwater. The phylogenetic relationships among taxa belonging to this order are not always clear because their taxonomic status has been repeatedly revised. In this study, diatom cells were isolated from rivers in Italy, Luxembourg, Portugal and Spain. In total, 21 18S rDNA gene sequences, representing six genera of Cymbellales (*Cymbella* C. Agardh, *Didymosphenia* M. Schmidt, *Encyonema* Kützing, *Gomphoneis* Cleve, *Gomphonema* Ehrenberg and *Reimeria* Kociolek & Stoermer) were determined. These sequences were analyzed along with other known GenBank diatom 18S rDNA gene sequences. The results indicate that the Cymbellaceae Greville and Gomphonemataceae Kützing, especially the genus *Gomphonema*, are paraphyletic, and that the significance of some of the morphological characteristics traditionally used for classification purposes requires a reassessment. These results also demonstrate the importance of a polyphasic approach combining both morphological and molecular data in attempting to improve the taxonomy and classification system of diatoms.

Keywords: *18S rDNA, Bacillariophyceae, Cymbellales, freshwater diatoms, phylogeny*

Introduction

Diatoms are probably the most species-rich group of algae. Thus, Mann & Droop (1996) suggested that the total number of diatom species worldwide is probably not less than 200 000. A recent compilation of names of diatom genera, species and taxa at infraspecific ranks, indicates that over 60 000 diatom taxa have been described to date (Kociolek, 2011; Fourtanier & Kociolek, 2011). This diversity of species and their relationships have been traditionally explored using light microscopy (LM) and scanning electron microscopy (SEM). According to Cox (2010), the development of morphogenetic studies revealed differences among close features and similarities among divergent characters. The morphological characteristics that define the taxonomic groups can therefore be subject to discussion. This problem also accounts for the repeated shifts in classification and phylogenetic relationships that have occurred.

The Cymbellales D.G. Mann is commonly found in freshwater environments. According to Round et al. (1990), four families belong to this order: Anomoeoneidaceae D.G. Mann, Cymbellaceae Greville, Gomphonemataceae Kützing and Rhoicospheniaceae Chen & Zhu. Members of this order show a high diversity of shape with different symmetry in the apical and transapical axes. Cladistic methods have been used to clarify the phylogeny of freshwater Cymbellaceae and Gomphonemataceae (Kociolek & Stoermer, 1988, 1989b, 1993). Based on morphological features, Cox (2002) presented a scenario of possible evolution for the different genera in the order Cymbellales which assumed a common isopolar ancestor, dorsiventrality and the position of stigmata in relation to primary and secondary sides of the valve, and a separation of two groups depending on the presence or absence of intermissio. However, Cox indicated that a molecular approach would be interesting to test this scenario. Indeed, molecular biology techniques offer interesting alternatives to microscopic observations for determining phylogenetic relationships. Several authors have used 18S rDNA gene sequence comparisons to clarify diatom phylogeny at the species or genus level (Beszteri et al., 2001; Kooistra et al., 2003; Edgar & Theriot, 2004; Alverson & Kolnick, 2005; Bruder & Medlin, 2007; Bruder et al., 2008), or to reassess the relationships between major classes (Medlin & Kaczmarska, 2004; Theriot et al., 2009). Molecular studies of species belonging to the order Cymbellales have already been carried out at higher taxonomic (Medlin & Kaczmarska, 2004) and generic levels (Bruder & Medlin, 2008), but with only a few sequences from the Cymbellales.

Table III.1: *List of Cymbellales species and strains sequenced in this study, with sampling location, date and their GenBank accession numbers.*

Taxon	code	Location (Site, Country)	Date	Accession number
Cymbella excisa	TCC772	Carrión (Triollo, Spain)	1 Nov 06	JN790273
Cymbella tumida	TCC713	Pisuerga (Medina de Fernamental, Spain)	2 Nov 06	JN790274
Encyonema silesiacum	TCC674	Brusago (Trento, Italy)	27 Oct 06	JN790275
Encyonema silesiacum	TCC717	Yuso (Boca de Huérgano, Spain)	1 Nov 06	JN790276
Encyonema silesiacum	TCC718	Yuso (Boca de Huérgano, Spain)	1 Nov 06	JN790277
Encyonema sp.	TCC655	Attert (Colmar-Berg, Luxembourg)	11 Nov 06	JN790278
Gomphoneis minuta	TCC715	Yuso (Boca de Huérgano, Spain)	1 Nov 06	JN790279
Gomphonema acuminatum	TCC737	Ribeira de Arão (Pereira, Portugal)	5 Dec 06	JN790280
Gomphonema bourbonense	TCC773	Ribeira De Seixe (Foz Do Arroio, Portugal)	4 Dec 06	JN790281
Gomphonema micropus	TCC704	Bernesga (Alija de la Ribera, Spain)	2 Nov 06	JN790282
Gomphonema parvulum	TCC725	Ribeira De Seixe (Foz Do Arroio, Portugal)	4 Dec 06	JN790283
Gomphonema parvulum	TCC664	Schlirbech (Aval Esch-sur-Sûre, Luxembourg)	10 Oct 06	JN790284
Gomphonema parvulum	TCC736	Ribeira de Arão (Pereira, Portugal)	5 Dec 06	JN790285
Gomphonema parvulum	TCC734	Ribeira da Cerca (Moinho do Bispo, Portugal)	4 Dec 06	JN790286
Gomphonema rosenstockianum	TCC775	Ribeira da Perna Seca (Parcanhão, Portugal)	5 Dec 06	JN790287
Gomphonema rosenstockianum	TCC776	Ribeira de Arão (Pereira, Portugal)	5 Dec 06	JN790288
Gomphonema rosenstockianum	TCC740	Bar. da Água Velha (Água Velha, Portugal)	5 Dec 06	JN790289
Reimeria sinuata	TCC719	Esla (Mansilla de las Mulas, Spain)	2 Nov 06	JN790290
Reimeria sinuata	TCC735	Ribeira de Arão (Pereira, Portugal)	5 Dec 06	JN790291
Reimeria sinuata	TCC721	Ribeira De Seixe (Zambujeira De Baixo, Portugal)	15 Dec 06	JN790292
Didymosphenia geminata	TCC777 [a]	Brusago (Trento, Italy)	8 May 07	JN790293

Note: Codes correspond to the strain codes of Thonon Culture Collection (http://w3.pierroton.inra.fr/r-syst/). [a] One cell was analyzed, not cultured

This study attempts to address the phylogenetic relationships among Cymbellales by means of molecular phylogeny, in combination with classical morphological analyses. The gene sequences of the 18S rDNA from different species of diatoms were determined and their genetic interrelationships are presented. The associations between molecular data and morphological characters of the diatom frustules were also studied.

Materials and methods

Isolation and culture of different strains.

Benthic samples were collected by scraping material from the surface of stones from rivers in various regions of Italy, Luxembourg, Portugal and Spain. A subsample of each collection was added to a medium composed of 25% milli-Q water, 25% "WC" liquid medium (Guillard & Lorenzen, 1972) and 50% of water from the sampling site (pre-filtered through 0.2 µm pore size filter). Monoclonal strains were established by micropipetting single cells under an inverted microscope. Non axenic unialgal cultures were maintained in WC liquid medium at 13°C in a growth chamber with a 16:8 h light:dark photoperiod at an irradiance of ~30–60 µE m^{-2} s^{-1}.

It was not possible to grow *Didymosphenia geminata* (Lyngbye) M. Schmidt under these culture conditions. Consequently, a single cell of this large species was isolated from an epilithic sample obtained from Brusago stream in the Province of Trento, northern Italy, where massive mat formation of *D. geminata* was observed along this oligotrophic, siliceous watercourse in 2006 and 2007 (Beltrami et al., 2008), and directly subjected to polymerase chain reaction (PCR) analysis. In total, 20 monoclonal diatom cultures and one *D. geminata* cell were subjected to phylogenetic analyses (Table III.1).

Sequencing reactions were performed in both directions using a Big Dye Terminator v3.1 cycle sequencing kit (Applied Biosystems). Primers 1F, 1528R (Medlin et al., 1988), and 528F, 1055F, 536R and 1055R (Elwood et al., 1985) were used to retrieve the complete sequence of both strands. After purification, sequencing products were subjected to electrophoresis on a 3130 Genetic Analyser (Applied Biosystems). The six sequences of each of the strains and of *D. geminata* were manually assembled using BioEdit Sequence Alignment Editor v. 7.0.5 (Hall, 1999).

Table III.2 : *GenBank accession numbers of Cymbellales and outgroup used in this study*.

Taxon	Accession number
Achnanthidium minutissimum	AM502032
Cymbella affinis	AM502018
Cymbella affinis	AM502009
Cymbella aspera	AM502016
Cymbella cymbiformis	AJ535156
Cymbella lanceolata	AM502026
Cymbella proxima	AM502017
Cymbopleura naviculiformis	AM501997
Cymbopleura naviculiformis	AM502004
Encyonema caespitosum	AM502035
Encyonema minutum	AM501961
Encyonema triangulum	AJ535157
Gomphonema acuminatum	AM502019
Gomphonema affine	AM502002
Gomphonema affine	AM502033
Gomphonema cf. *angustatum*	AM502005
Gomphonema micropus	AM501964
Gomphonema micropus	AM501965
Gomphonema parvulum	AJ243062
Gomphonema cf. *parvulum*	AM501995
Gomphonema productum	AM501993
Gomphonema truncatum	AM501956
Uncultured Gomphonema sp.	AB473923
Uncultured Gomphonema sp.	AB473924
Placoneis elginensis	AM501953
Placoneis hambergii	AM502030
Placoneis sp.	AM502014

Phylogenetic analysis.

The sequences were aligned using the SILVA web aligner (Pruesse et al., 2007) and transferred to the ARB program package (Ludwig et al., 2004). Sequences of Cymbellales (Table III.2) available in the SSU database Silva 102 were included in the analysis. The choice of outgroup to root the trees was guided by the criteria of Verbruggen & Theriot (2008). One of the immediate sister sequence of Cymbellales with 18S rDNA gene was *Achnanthidium minutissimum* (Kützing) Czarnecki (Table III.2). The final dataset included 48 sequences that correspond to 28 taxa identified at the species level and 4 taxa identified at the genus level. The alignment was manually corrected considering primary and secondary structural similarity using the alignment tool of the ARB software. The alignment is available from the authors.

Maximum likelihood (ML) and neighbor joining (NJ) trees were calculated using software available on the Phylogeny.fr website (Dereeper et al., 2008). PhyML software (Guindon & Gascuel, 2003) was used with the optimal model of sequence evolution (General Time Reversible) selected according to the Akaike information criterion in jmodeltest 0.1 (Posada, 2008). For NJ tree, the BioNJ software (Gascuel, 1997) was used with the Jukes & Cantor (1969) correction. Maximum parsimony (MP) phylogenetic tree was constructed using Phylip v. 3.6 (Felsenstein, 2005). For all analyses, 1000 bootstrap (BS) replicates were performed in order to assess the robustness of the nodes.

Light and scanning electron microscopy.

Diatom material was prepared according to the recommendations of the (European Committee for Standardization, 2003). Diatoms from cultures and from the epilithic sample obtained from Brusago stream were cleaned with 35–40% H_2O_2 and 37% HCl. The frustules were then rinsed with demineralized water. The cleaned diatoms were dried and mounted with synthetic resin (Naphrax©) and slides were prepared for LM. Light micrographs were taken using a Leica DMR microscope (× 100 oil immersion) and a Leica DC 500 camera. For SEM, clean diatom frustules and valves were filtered over a 3 µm pore size filter. The filters were then mounted on 12 mm aluminium stubs, coated with ~350 Å of gold (MED 020 Modular High Vacuum Coating System, Bal-Tec, Balzers) and examined using a Leica Stereoscan 430i microscope operating at a working distance of 10 mm and an accelerating voltage (iProb) of 20 kV.

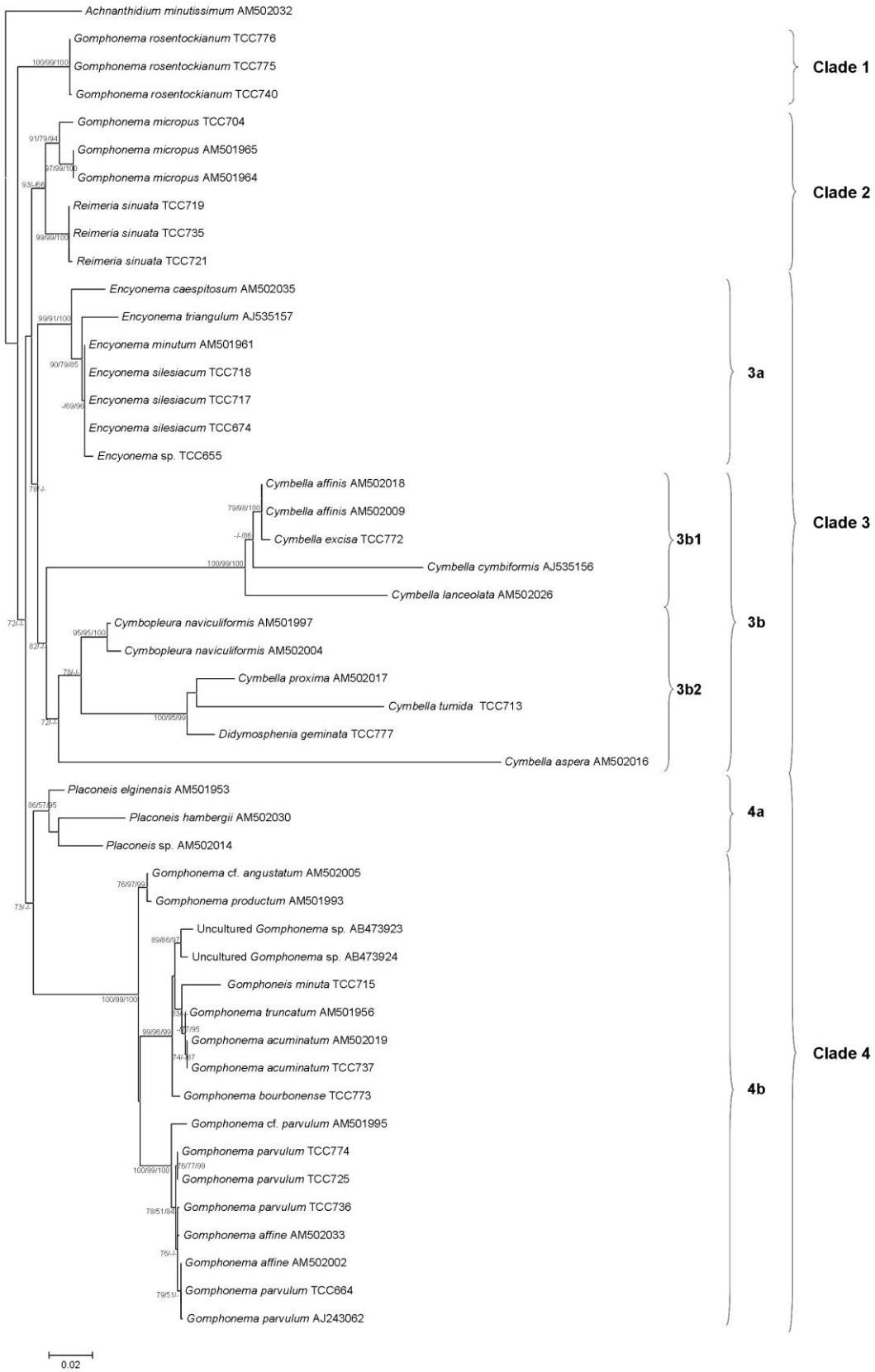


Figure III.1: *Phylogeny of the Cymbellales inferred from maximum likelihood analysis of 18S rDNA gene sequences using all taxa studied including those from GenBank.*
Numbers on branches show bootstrap values obtained by maximum likelihood, maximum parsimony and neighbor joining analyses, respectively. Scale bar shows genetic distances.

Results

Phylogenetic data. The 18S rDNA gene sequences from diatom cultures and from a single cell of *D. geminata* were obtained in both directions. After alignment and corrections, 1668 positions (including indels) were used to construct the phylogenetic trees. Sequence similarities among all the sequences were high (84–100%) and four main clades emerged in the ML tree (Figure III.1).

Clade 1 was composed by strains identified as *Gomphonema rosenstockianum* Lange-Bertalot & Reichardt. This clade diverged at the base of the Cymbellales in the three phylogenetic trees and formed a strongly supported group with BS values of 100/99/100 obtained by ML/MP/NJ analyses, respectively.

Clade 2 included species of both *Reimeria sinuata* (Gregory) Kociolek & Stoermer and *Gomphonema micropus* Kützing. Only in the MP tree, was this clade divided and the two genera diverged one after another. In this clade, each genus, composed of only one species, was well-supported by the BS values of each of the three phylogenetic analyses (ML/MP/NJ) with BS 91/79/94 for *G. micropus* and 99/99/100 for *R. sinuata*. *G. micropus* appeared more closely related to *R. sinuata* than to other *Gomphonema* species.

In clade 3, two subclades appeared. Subclade 3a consisted of strains belonging to the genus *Encyonema* Kützing, which formed a strongly supported monophyletic group with BS of 99/91/100 for each of the three phylogenetic analyses (ML/MP/NJ): *E. caespitosum* Kützing, *E. minutum* (Hilse) Mann, *E. silesiacum* (Bleisch) Mann, *E. triangulum* (Ehrenberg) Kützing and *E.* sp. showing a high identity (>97%, data not shown). Subclade 3b comprised, in addition to several *Cymbella* species, *Cymbopleura naviculiformis* (Auerswald ex Heiberg) Krammer and *D. geminata*. This subclade was modified depending on the phylogenetic analysis, but in ML tree subclade 3b (BS = 82) could be divided in at least two groups. Group 3b1 was composed by species of the genus *Cymbella* C. Agardh, namely *C. affinis* Kützing, *C. excisa* Kützing, *C. cymbiformis* C. Agardh and *C. lanceolata* (C. Agardh) Kirchner. This group was homogeneous and presented high BS support (100/99/100) in the three phylogenetic trees. Group 3b2 was composed by different genera such as *Cymbopleura* (Krammer) Krammer (*C. naviculiformis*), *Didymosphenia* M. Schmidt (*D. geminata*) and *Cymbella* [*C. tumida* (Brébisson ex Kützing) Van Heurck, *C. aspera* (Ehrenberg) Cleve and *C. proxima* Reimer]. *Didymosphenia geminata* grouped with *C. tumida* and *C. proxima* with the ML,

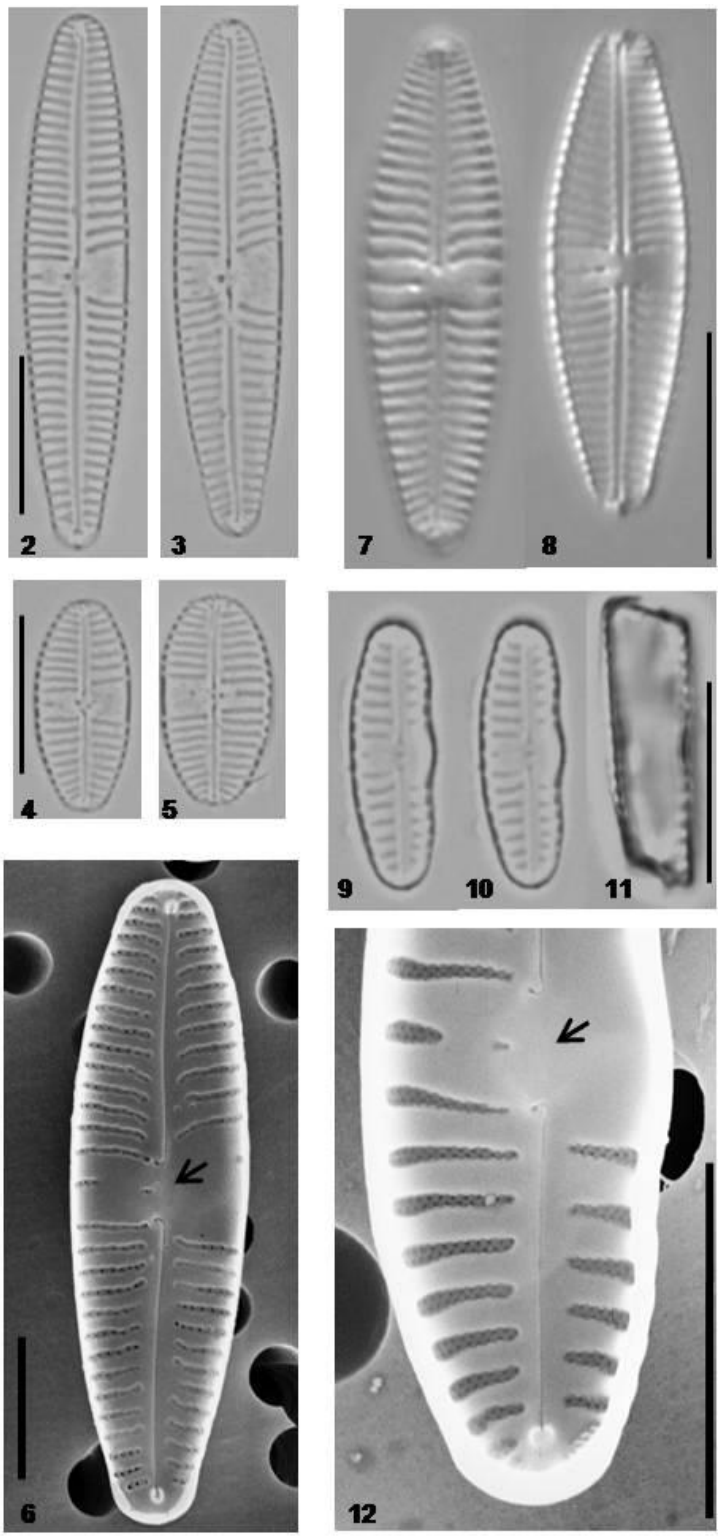


Figure III.2: *Some species of Cymbellales characterizing clades 1 (Figs 2–6) and 2 (Figs 7–12).* Figs 2–6. *Gomphonema rosenstockianum*, strain TCC740. Figs 2–5. Valve views, LM. Fig. 6. Internal view of valve showing the intermissio (arrow),SEM. Figs 7–8. Valve view of *Gomphonema micropus*, strain TCC704, LM. Figs 9–12. *Reimeria sinuata*. Figs 9–10. Valve views of strain TCC735, LM. Fig. 11. Girdle view of strain TCC735, LM. Fig. 12. Internal view of valve showing the intermission (arrow), strain TCC719, SEM. Scale bars = 10 µm (Figs 2–5, 7–11); 5 µm (Figs 6, 12).

MP and NJ analyses. The position of *Cymbopleura naviculiformis* was the same for both ML and NJ trees (grouped with *D. geminata*, *C. proxima* and *C. tumida*), but was at the base of group 3b2 in the MP tree. The 18S rDNA gene sequences confirmed that *D. geminata* and *C. naviculiformis* were more closely related to species of the genus *Cymbella* than to *Gomphonema* Ehrenberg, *Gomphoneis* Cleve or *Encyonema*.

Clade 4 could be divided in two subclades. Subclade 4a (supported by BS values of 88, 57 and 95 in the ML, MP and NJ trees, respectively) was composed only by three sequences of *Placoneis* Mereschkowsky. The position of these sequences changed depending on the analysis performed. *Placoneis* grouped with *Cymbella* in the MP tree, diverged alone in the NJ tree, and grouped with *Gomphonema* in the ML tree. Thus, the position of the genus *Placoneis* was not resolved. Subclade 4b comprised sequences of several species of *Gomphonema*, namely *G. acuminatum* Ehrenberg, *G. affine* Kützing, *G.* cf. *angustatum* (Kützing) Rabenhorst, *G. bourbonense* Reichardt, *G. parvulum* (Kützing) Kützing, *G. productum* (Grunow) Lange-Bertalot & Reichardt, *G. truncatum* Ehrenberg), and one sequence of *Gomphoneis minuta* (Stone) Kociolek & Stoermer. Subclade 4b appears to be composed of different species groups that formed a well-supported subclade with high BS support values (100/99/100). *Gomphoneis minuta* formed a group with *Gomphonema acuminatum*, *G. truncatum* and *G. bourbonense* in the three trees with high BS values (99/96/99), but its position changed depending on the analysis performed.

Morphological data. Within these clades, the morphology of strains isolated for this study was investigated using LM and SEM (Figure III.2 and Figure III.3) to find associations between molecular and morphological data. The main morphological data of the genera were then related to the phylogenetic tree in Figure III.4.

Clade 1 clearly identifies *G. rosenstockianum* (Figure III.2), which shows symmetry to the apical axis and slightly heteropolar valves, a narrow basal pole and a wider head pole. All strains have a straight central raphe presenting slightly expanded (Figure III.2: 2–5) to hooked central raphe endings in internal view (Figure III.2: 6) and hooked external polar endings. Internal views show intermissio (i.e. interruption in the raphe slits, Figure III.2: 6, arrow). Frustules have an apical pore field at the basal pole and uniseriate striae. All strains possess one stigma. This *Gomphonema* species is mainly characterized by the absence of striae in the center of one side of the valve (Figure III.2: 2–6).

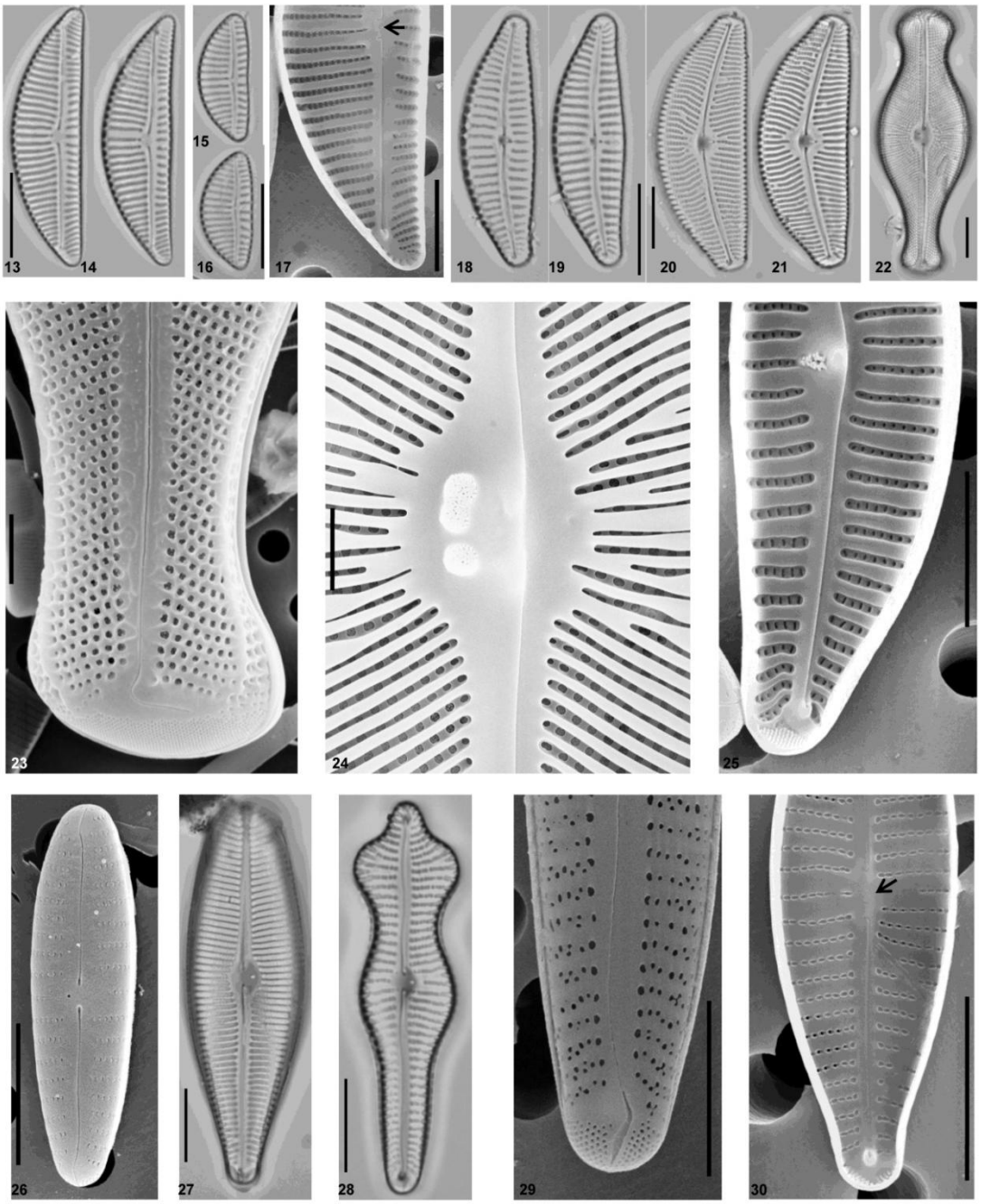


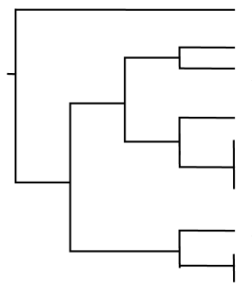
Figure III.3: *Some species of Cymbellales characterizing clades 3 (Figs 13–25) and 4 (Figs 26–30).*

Figs 13–17. *Encyonema silesiacum* complex. Figs 13–14. Valve views of strain TCC717, LM. Figs 15–16. Valve views of strain TCC655, LM. Fig. 17. Internal valve view of strain TCC717 showing the intermission (arrow), SEM. Figs 18–19, 25. *Cymbella excisa*, strain TCC772. Figs 18–19. Valve views, LM. Fig. 25. Internal view of the central area and the pole, SEM. Figs 20–21. Valve views of *Cymbella tumida*, strain TCC713, LM. Figs 22–24. *Didymosphenia geminata*. Fig. 22. Valve view, LM. Fig 23. External detail of the basal pole with apical pore field, SEM. Fig. 24. Internal view of the central area, SEM Fig. 26. Valve view of *Gomphonema bourbonense*, strain TCC 773, SEM. Fig. 27. Valve view of *Gomphoneis minuta*, strain TCC715, LM. Figs 28–29. *Gomphonema acuminatum*, strain TCC 737. Fig. 28. Valve view, LM. Fig. 29. External detail of the basal pole with apical pore field, SEM. Fig. 30. Internal view of the central area with intermission (arrow) and apex of *Gomphonema parvulum*, strain TCC664, SEM. Scale bars = 10 µm (Figs 13–16, 18–22, 27–28); 5 µm (Figs 17, 23–26, 29–30).

Clade 2 is represented by *G. micropus* (Figure III.2: 7–8) and *Reimeria sinuata* (Figure III.2: 9–12). *Gomphonema micropus* shows symmetry to the apical axis and slightly heteropolar valves, a narrow basal pole and a wider head pole. As with *G. rosenstockianum*, *G. micropus* is characterized by either a very abbreviated stria (Figure III.2: 7) or the absence of striae (Figure III.2: 8) in the center of the valve side opposite the stigma. *Reimeria sinuata* isolated during this study (Figure III.2: 9–12) show symmetry to the transapical axis, asymmetry to the apical axis (moderately dorsiventral valve), with the dorsal valve margin slightly arcuate. Cells of *R. sinuata* are trapezoid in girdle view (Figure III.2: 11). The raphe is straight with bulbous proximal ends. An intermissio is observed in SEM internal view (Figure III.2: 12). Striae are biseriate and one stigma is present near the center. An apical pore field is present at each pole on the ventral side (slightly visible on Figure III.2: 12). In addition, *R. sinuata* shows an absence of striae at the center on one side of the valve.

In clade 3, the subclade 3a is represented by several species of the genus *Encyonema* (Figure III.3: 13–17). They all share strongly dorsiventral cells, which present valves asymmetrical to the apical plane and symmetrical to the transapical plane. The dorsal valve margin is highly arched, while the ventral margin and the raphe are nearly straight. Both distal ends of the raphe curved towards the ventral side, and the proximal ends of the raphe curved dorsally. Species of this genus were characterized by the absence of an apical pore field and the presence of intermissio (Figure III.3: 17, arrow). All three strains of the *E. silesiacum* complex and *Encyonema* sp. shared these characteristics. By contrast, subclade 3b is made of *Cymbella* and *D. geminata* (Figure III.3: 18–25). All sharing differentiated apical pore fields not bisected by the external raphe ends (Figure III.3: 23), a stigma along the ventral margin, the lack of intermissio (Figure III.3: 24–25), and distal raphe ends deflected dorsally.

In clade 4, the subclade 4b grouped several species of *Gomphonema* and *Gomphoneis* (Figure III.3: 26–30). All strains of *G. acuminatum*, *G. bourbonense*, *G. parvulum* and *Gomphoneis minuta* show the main features of the Gomphonemataceae, including heteropolar valves with a narrow basal pole and a wider head pole (only slightly wider in *G. bourbonense*, Figure III.3: 26). All the strains show a central raphe (straight or sinuous), uni- or biseriate striae, one stigma, an apical pore field at the basal pole (Figure III.3: 29–30) and an intermissio (Figure III.3: 30, arrow).



	Symmetry		Striae	Stigma number	Apical pore field		Intermissio
	apical axis	transapical axis			number	not bisected	
G. rosentockianum	symmetrical	asymmetrical	uniseriate	1	1	0	1
G. micropus	symmetrical	asymmetrical	uniseriate	1	1	0	1
R. sinuata	asymmetrical	symmetrical	biseriate	1	2	0	1
Encyonema	asymmetrical	symmetrical	uniseriate	0 or 1	0	N/A	1
Cymbopleura	asymmetrical	symmetrical	uniseriate	0	0	N/A	0
Cymbella	asymmetrical	symmetrical	uniseriate	1 or more	2	1	0
Didymosphenia	symmetrical	asymmetrical	uniseriate	2 or more	1	1	0
Placoneis	symmetrical	symmetrical	uni/biseriate	0	0	N/A	1
Gomphonema	symmetrical	asymmetrical	uni/biseriate	1	1	0	1
Gomphoneis	symmetrical	asymmetrical	biseriate	1	1	0	1

Figure III.4: *Simplified maximum likelihood phylogenetic tree related to the main morphological data of the branches*:
symmetry, type of striae, number of stigma, number of apical pore fields bisected (1), not bisected(0), or not applicable (N/A) and the presence (1) or absence (0) of intermissio.

Discussion

Only two main families of the order Cymbellales, Cymbellaceae and Gomphonemataceae, were studied. The Cymbellales have complex relationships and the order requires a taxonomic revision because the main families are paraphyletic.

Cymbellaceae.

The Cymbellaceae family is distributed in 14 genera (Fourtanier & Kociolek, 2011): *Afrocymbella* Krammer, *Brebissonia* Grunow, *Cymbella*, *Cymbopleura* (Krammer) Krammer, *Delicata* Krammer, *Encyonema*, *Encyonopsis* Krammer, *Gomphocymbella* O.F. Müller, *Gomphocymbellopsis* Krammer, *Navicella* Krammer, *Navicymbula* Krammer, *Oricymba* Jüttner, Krammer, Cox, Van de Vijver & Tuji., *Placoneis* and *Pseudencyonema* Krammer. Four of these genera were included in this study. The genus *Encyonema* in the subclade 3a forms a monophyletic group separated from the genus *Cymbella* which was well-supported by our phylogenetic analyses. *Encyonema* species have no apical pore field and can form mucilage tubes (Round et al., 1990), features not shared by other genera of the Cymbellales.

Krammer (2002) has defined *Cymbella* sensu stricto to include only more or less dorsiventral Naviculaceae Kützing with a cymbelloid raphe (polar fissures turned to the dorsal side, proximal raphe endings ventrally bent or displaced), apical pore fields and the ability to produce mucilaginous stalks, ventral stigma, and a dorsal plastid (Cantonati et al., 2010). *Cymbella* species form the subclade 3b with other genera represented by two species, *D. geminata* and *Cymbopleura naviculiformis*. Kociolek & Stoermer (1988) considered the genus *Didymosphenia* to be more closely related to *Cymbella* than to other gomphonemoid diatoms. By contrast, Round et al. (1990) assigned *Didymosphenia* to the Gomphonemataceae with other genera characterized by an asymmetrical frustule to the transapical axis as *Gomphonema* and *Gomphoneis*. The 18S rDNA gene sequence analyses confirmed the conclusions of Kociolek & Stoermer (1988) and the scenario of Cox (2002) in which the genera *Cymbella* and *Didymosphenia* are grouped together.

For Krammer (2003), *Cymbopleura* contained only the free-living *Cymbella* taxa (no apical pore field) that are symmetrical or slightly asymmetrical with a raphe similar to the genus *Cymbella* and without stigma. The close relationship between *Cymbella* and *Cymbopleura* was confirmed by our phylogenetic analysis of 18S rDNA gene sequences.

The genus *Cymbella* is paraphyletic based on the subclade 3b. Therefore, either *D. geminata* and *Cymbopleura naviculiformis* should probably be included in the genus *Cymbella* or the *D. geminata*, *Cymbopleura naviculiformis* and the other *Cymbella* taxa, *C. proxima* and *C. tumida* (subclade 3b2), should be combined together in a new genus, separated from *Cymbella*.

The genus *Placoneis*, including *P. elginensis* (Gregory) Cox, *P. hambergii* (Hustedt) Bruder and *Placoneis* sp., form a monophyletic group. Bruder & Medlin (2007) investigated this genus with the 18S, 28S rDNA and ribulose 1,5-bisphosphate carboxylase large subunit (*rbc*L) gene sequences. The monophyly of the genus *Placoneis* was well-supported, but its relationship to other genera in the order varied with the markers used. Our results show that its relationship to other genera varied also with the phylogenetic methods used. Our phylogenetic analyses confirm more sequences are needed to determine the exact taxonomic position of *Placoneis* within the Cymbellales.

Our results show that the Cymbellaceae are surely polyphyletic. When considering clade 3, which can be considered as "typical" Cymbellaceae, the dorsiventral curved shape seems to be a basal characteristic which was lost for *Didymosphenia.* Apical pore fields are also a basal characteristic of clade 3, but one which was lost several times (*Encyonema*, *Cymbopleura*).

Gomphonemataceae.

The Gomphonemataceae family (Fourtanier & Kociolek, 2011) is composed of five genera: *Didymosphenia*, *Gomphoneis*, *Gomphonema*, *Gomphopleura* Reichelt ex Tempère and *Reimeria*. Three of these genera were included in our study. As we demonstrated before, *Didymosphenia* is closer related to the Cymbellaceae than the Gomphonemataceae. Heteropolarity of valves, characteristics of the Gomphonemataceae, has arisen several times during the evolutionary process of the Cymbellales, as asymmetry has been acquired on several occasions over the evolutionary diatom history (Kociolek & Spaulding, 2003).

Species belonging to the genera *Gomphoneis* and *Gomphonema* are characterized by the same raphe structure and by the presence of apical pore fields.

Cleve (1894) erected the genus *Gomphoneis* for species formerly considered as belonging to *Gomphonema*, but differed from it both in the structure and in the presence of longitudinal lines. In the original description of *Gomphoneis*, Cleve did not mention

the presence of double-punctate striae; however, several authors (Dawson, 1974; Merino et al., 1994; Tuji, 2005) considered biseriate striae as an important diagnostic feature of this genus and, for this reason, several *Gomphonema* species have been transferred to *Gomphoneis*. According to Compère (1995), Passy et al. (1997), Reichardt (2007, 2008) and Beltrami et al. (2010) the presence of biseriate striae alone does not justify the assignment of several *Gomphonema* species to the genus *Gomphoneis*. Our preliminary phylogenetic analysis, which is based on a single species of *Gomphoneis,* does not seem to support its separation from *Gomphonema*. More taxa should be investigated because some *Gomphonema* species (e.g., *G. acuminatum*) are more closely related to *Gomphoneis minuta* than to other *Gomphonema* species (e.g., *G. productum*). This could form the basis for a revised interpretation of *Gomphoneis* because with *Gomphonema*, they both share several features, such as heteropolar valves, stigma, intermissio, apical pore field and mucilage stalks.

Bruder & Medlin (2008) considered that *Gomphonema* formed a monophyletic clade, but also observed that *G. micropus* diverged at the base of the group. They pointed out that the long branch leading to *G. micropus* and the low bootstrap support for this species with the remaining *Gomphonema* spp. warranted a separation from the genus *Gomphonema*. With the addition of more taxa, the separation of *Gomphonema* was warranted as suggested by Bruder & Medlin (2008). *Gomphonema micropus* was the only *Gomphonema* species in the study of Bruder & Medlin (2008) where the areolae open externally in small round poroids. The striae are also formed by round, dot-like areolae in external view in *G. rosenstockianum* and small, round foramina that are alternating biseriate in *Reimeria sinuata*. Nevertheless, areolae having C-shaped foramina are also present in many other species of *Reimeria* (not investigated in this preliminary study), for instance *R. fontinalis* Levkov or *R. lacus-idahonensis* Kociolek & Stoermer (Levkov & Ector, 2010). Recently, Fránková et al. (2009) observed that *G. tergestinum* (Grunow) Fricke and *R. sinuata* have very similar morphological features, especially when the former has small valves. Krammer & Lange-Bertalot (1985) stated that *R. sinuata* is an intermediate species between *Cymbella* and *Gomphonema*. *Gomphonema rosenstockianum*, *G. tergestinum* and all *Reimeria* species share a common characteristic, which is the absence of striae or very short striae at center on one side of the valve. Our results indicate that *G. micropus* is related to *R. sinuata* and that *G. rosenstockianum* diverged at the base of the Cymbellaceae and Gomphonemataceae. Further studies are

necessary to determine the taxonomic position of *G. rosenstockianum* (and the *G. tergestinum* species complex) in the order Cymbellales and to define the exact generic position of *G. micropus*.

These results indicate that the Gomphonemataceae and especially the *Gomphonema* genus are paraphyletic. In future, it will be necessary to sequence additional strains and to probably redefine genera on the basis of their phylogenetic relationships, once these are more firmly established.

Conclusion

The results of this preliminary study concerning the phylogenetic relationships of the Cymbellales based on 18S rDNA gene data clearly show that both the Gomphonemataceae and Cymbellaceae are polyphyletic, and that several of their genera should be re-evaluated. Recently, Nakov & Theriot (2009) assessed the relationships within the Cymbellales using nuclear and chloroplastic genes, and reported similar results. They stressed the need to increase both the number of taxa sampled and the number of molecular markers in order to reach a multigene phylogeny of the diatoms (Theriot et al., 2010).

It would be very interesting to extend this study to other genera that share the morphological features of *Gomphonema* and *Cymbella*. For instance, *Gomphocymbella* is a genus that displays a heteropolar and curved frustule shape, a foot pole with apical pore fields, and a stigma on the dorsal side composed of one or several puncta. Another such genus is *Gomphocymbellopsis*, which has heteropolar and dorsiventral valves, pore fields in both the foot and head poles, and distinct, isolated stigma visible at the end of the central striae on the ventral side.

Finally, the genus *Krsticiella* Levkov has also been recently described at the boundary between *Cymbella* and *Gomphonema* (Levkov & Lange-Bertalot, 2007). Some species of *Gomphonema*, such as *G. cymbelliclinum* Reichardt & Lange-Bertalot, which display a dorsiventral curved shape, are also good examples of this close relationship between *Cymbella* and *Gomphonema* (Reichardt, 1999). As underlined by Cox (2009), the traditional classification of diatoms originated as an aid for identification rather than an arrangement of species relationships. Sequencing these taxa could help to validate a scenario explaining the evolution of morphological features in the order Cymbellales as

the one proposed in Cox (2002), and therefore would enable a better definition of diatom genera.

Acknowledgements

We would like to thank Julie Mathu for technical help. Thanks are due to Christophe Bouillon for technical support with both light and scanning electron microscopy. This study (project DIAMED) was funded by the Centre de Recherche Public – Gabriel Lippmann (Luxembourg), the CEMAGREF (France), the Conseil Général des Alpes Maritimes (France), and the Universities of Aveiro and Evora (Portugal), of Barcelona, León, Vigo (Spain), and of Camerino and Trento (Italy). We are grateful to Eduardo A. Morales for constructive comments on the paper.

3. *Etudes de complexes d'espèces*

3.1. *Gomphonema parvulum*

3.1.1. Article II : Présentation générale de l'étude et synthèse des principaux résultats.

Comme nous l'avons précédemment introduit dans le CHAPITRE I, la définition précise des espèces de diatomées, sur des critères morphologiques, est d'autant plus problématique que les espèces de diatomées ont une grande variabilité morphologique au cours de leur cycle de vie. Chaque diatomée répond aux conditions environnementales au cours de la division végétative. Par exemple, des changements de la taille des pores de *Cocconeis placentula* Ehrenberg en fonction de la salinité ont été observés (Leterme et al., 2010). Les formes tératologiques sont d'autres exemples de ces variations morphologiques survenant au cours du cycle végétatif, en réponse aux modifications des conditions environnementales (e.g. la revue bibliographique de Falasco et al., 2009). Cette variabilité morphologique doit pourtant être prise en compte pour une identification correcte des diatomées. Beaucoup d'espèces de diatomées présentent plusieurs morphotypes difficiles à distinguer. Ainsi, la présence de complexes d'espèces dans les échantillons est une source majeure d'erreurs lors de l'évaluation des indices diatomées (Prygiel et al., 2002; Kahlert et al., 2009). En effet, différents complexes d'espèces sont composés de groupes morphologiquement proches qui peuvent présenter des préférences écologiques différentes (Potapova & Hamilton, 2007; Vanelslander et al., 2009).

Le genre *Gomphonema* est un genre de Cymbellales abondant dans les échantillons benthiques d'eau douce. Toutefois, plusieurs taxa sont difficiles à identifier en microscopie optique. C'est le cas du complexe d'espèces *Gomphonema parvulum*. Ce groupe est connu pour sa forte variabilité morphologique et sa large gamme de tolérance écologique. *Gomphonema parvulum* est largement distribué et, est parmi les taxa les plus fréquemment cités dans la littérature (Finlay et al., 2002). Les taxonomistes ont souvent des difficultés à distinguer morphologiquement les taxa appartenant à ce complexe d'espèces. Leur distinction taxonomique est pourtant importante, car ces espèces ont souvent des valeurs écologiques différentes dans les indices diatomées. Par exemple, une valeur de sensibilité de 2.0 et une valeur indicatrice de 1.0 sont attribuées à *G. parvulum* var. *parvulum* pour l'IPS alors que *G. exilissimum* (anciennement *G. parvulum* var. *exilissimum*) présente des valeurs différentes : une valeur de sensibilité de 5.0 et une

valeur indicatrice de 1.0. Au sein de ce complexe, *G. lagenula* (anciennement *G. parvulum* var. *lagenula*) présente également des valeurs différentes (sensibilité 2.0, valeur indicatrice 3.0) et est considérée comme une espèce tropicale.

Comme nous l'avons confirmé dans l'article précédent, les techniques de biologie moléculaire sont des alternatives intéressantes aux observations microscopiques pour l'étude des diatomées. En effet, nous avons démontré que l'utilisation de l'ADNr 18S apporte des informations supplémentaires permettant de mieux étudier les relations entre les familles et les genres de diatomées. L'objectif de cette seconde étude était d'étudier la diversité du complexe *G. parvulum* par une approche polyphasique basée à la fois sur la morphologie et sur des marqueurs moléculaires, et ainsi révéler le niveau de précision qui peut être atteint par l'utilisation de marqueurs moléculaires. En incluant dans cette étude des souches provenant de milieux tempérés et insulaires tropicaux, nous avons pu également utiliser les outils d'identification pour mieux comprendre la biogéographie au sein de ce complexe d'espèces.

Pour répondre à ces objectifs, nous avons choisi de combiner les approches morphométriques et moléculaires. Les séquences d'ADNr 18S, 28S, des ITS, et des gènes *rbc*L et *cox*1, de plusieurs souches de *G. parvulum* et d'autres espèces de *Gomphonema* des rivières d'îles tropicales françaises (Mayotte, La Réunion) et du continent européen (France métropolitaine, Espagne, Italie, Luxembourg et Portugal), ont été déterminées et comparées avec des séquences disponibles dans GenBank. Les séquences d'ADNr 18S n'étant pas suffisamment polymorphes pour fournir des informations robustes sur la diversité au sein des *G. parvulum*, celles-ci n'ont finalement pas été utilisées. En parallèle, nous avons étudié la morphologie de nos souches de *G. parvulum* par des approches de morphométrie conventionnelle, et de morphométrie géométrique (basée sur les pôles et l'aire centrale).

Le séquençage des différents marqueurs et le calcul des distances génétiques intra- et intergroupes nous a permis de diviser nos souches de *G. parvulum* en cinq clades. La faible différenciation génétique de deux de ces clades ainsi que leur séparation géographique ont suggéré qu'ils représentaient deux populations distinctes, isolées géographiquement, d'une même entité taxonomique.

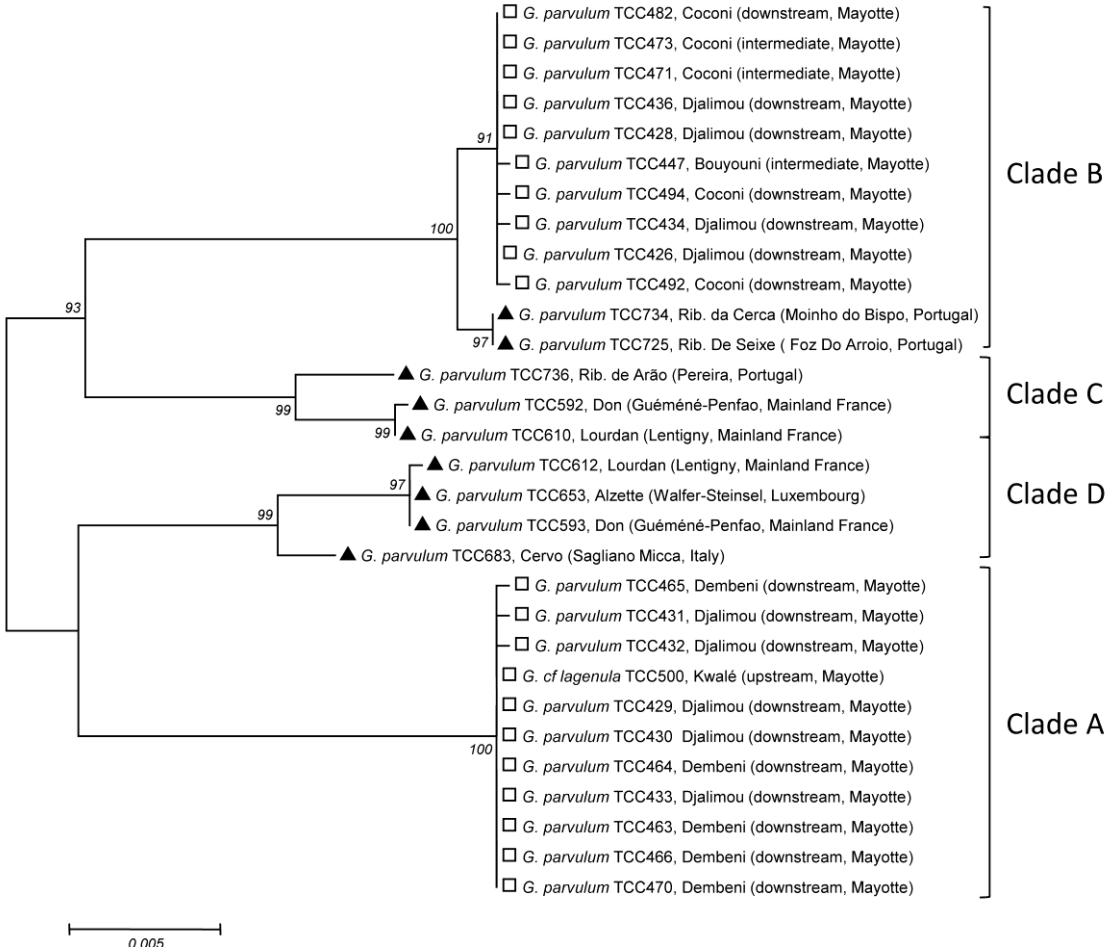


Figure III.5: *Arbre phylogénétique obtenu par analyse en Maximum Likelihood des séquences de trois marqueurs : ITS,* rbc*L and* cox*1*.
Les valeurs de bootstrap sont indiquées sur les branches ; les carrés blancs correspondent aux souches tropicales et les triangles noirs correspondent aux souches tempérées.

Par ailleurs, les espèces *G.* cf. *lagenula* et *G. exilissimum* (anciennement considérées comme des variétés de *G. parvulum*) étaient nettement incluses dans un des clades de *G. parvulum*, indiquant que la séparation de ces deux espèces pourrait ne pas être justifiée phylogénétiquement. Cependant, l'étude d'un plus grand nombre de souches serait nécessaire pour conclure définitivement sur le statut taxonomique de ces deux espèces.

Avec l'utilisation des données morphométriques géométriques et conventionnelles, nous avons confirmé que les clades principaux présentent également des différences morphologiques, mais qu'il y a un continuum morphologique entre les frustules des clades. En effet, cette étude a révélé des différences significatives pour la densité de stries, la longueur et la largeur des frustules de certains clades, mais les gammes de variation de ces caractères sont très chevauchantes. De plus, il est apparu que les variations de l'aire centrale étaient un critère important pour caractériser la diversité des souches de *G. parvulum*, alors que c'est la forme des pôles qui est un des critères classiquement utilisé pour différencier les variétés de *G. parvulum*.

Le groupe *G. parvulum* possède une large répartition géographique. Finlay et al. (2002) ont référencé la présence de *G. parvulum* en Asie, en Europe, en Amérique centrale et en Amérique du Nord. Cependant, la répartition géographique de nos clades a démontré qu'il existe une différenciation en fonction de l'origine tropicale ou tempérée des souches (Figure III.5). Les clades de *G. parvulum* pourraient donc présenter des profils biogéographiques différents. Toutefois, au niveau d'une zone géographique restreinte, aucune différenciation nette n'est observée puisque différents clades peuvent vivre en sympatrie en zone tempérée ou en zone tropicale.

Les différents marqueurs moléculaires nous ont apporté des résultats concordants. De plus, nous avons observé de faibles variations génétiques intraclades et de plus fortes variations interclades. Nos résultats suggèrent donc que les échanges de matériel génétique entre les clades sont limités, d'autant plus que la séparation géographique n'explique pas cette séparation dans le cas de clades vivants en sympatrie. Pour définir correctement le statut taxonomique de ces clades, des informations complémentaires concernant la reproduction et l'écologie devraient être acquises. De plus, il serait nécessaire de confirmer nos résultats sur un échantillonnage de souches plus important.

3.1.2. Article II : Soumis après révision dans Protist

Using a polyphasic approach to explore the diversity and geographical distribution of the *Gomphonema parvulum* (Kützing) Kützing complex (Bacillariophyta)

Lenaïg Kermarrec[1,2], Agnès Bouchez[2], Frédéric Rimet[2] and Jean-François Humbert[3]

[1]*Asconit Consultants, 3 bd Clairfont, F-66350 Toulouges, France*

[2]*INRA, UMR CARRTEL, 75 av de Corzent, F-74203 Thonon-les-Bains cedex, France*

[3]*INRA, UMR BIOEMCO, site de l'ENS, 46 rue d'Ulm, F-75005 Paris, France*

Running title: *Gomphonema parvulum* characterization

Author for correspondence: Agnès Bouchez: agnes.bouchez@thonon.inra.fr

Soumis après révision le 20 janvier 2012.

Abstract

Several diatom taxa have recently been shown to require some clarifications and taxonomic reassessments despite having been observed and described for many years using light and electron microscopy. One such is the *Gomphonema parvulum* complex, which included species displaying considerable morphological variability, making it difficult to identify them by microscopic examination. As this species complex is widely distributed, the possibility of biogeographical differentiation has also been raised. In this context, we isolated 39 *G. parvulum* s.l. strains from rivers located in tropical islands and in Europe. By sequencing approach four DNA fragments (LSU rDNA, ITS, *rbc*L and *cox*1), we showed that the *G. parvulum* complex contains five clades. Geometric and conventional morphometric analyses showed that these clades display morphological diversity, but also that there was a continuum among the frustules of the five clades. The geographical distribution of these clades demonstrated clear differentiation at a large spatial scale, consisting of a tropical *versus* temperate differentiation. Additional findings concerning the inter-clade and intra-clade variability allowed us to propose several hypotheses about the colonization of tropical islands by *Gomphonema parvulum*.

Key words: Diatoms, *Gomphonema parvulum*, genetic diversity, morphological and morphometric variability, biogeography.

Introduction

Diatoms (Bacillariophyta) are major constituents of the communities of aquatic ecosystems. Many diatom species contain several different morphotypes that are difficult to differentiate. For example, in the *Achnanthidium minutissimum* (Kützing) Czarnecki species complex, six morphologically closely related groups display different ecological profiles (Potapova & Hamilton, 2007). Beside this known morphological variability, molecular techniques have recently revealed that many diatom species also contain cryptic or pseudo-cryptic species (Sarno et al., 2005; Amato et al., 2007; Evans et al., 2008; Trobajo et al., 2009; Poulíčková et al., 2010).

In addition to the problem of assessing the taxonomic diversity in diatom species complexes, there is also that of possible differences in the geographical distribution of species belonging to the same species complex. For instance, it has been shown that some marine *Skeletonema* species (Kooistra et al., 2008) are widely distributed, whereas some cryptic species are only found in a few locations. Moreover, a high degree of genetic isolation between distant populations has also been identified in a cosmopolitan marine species, *Pseudo-nitzschia pungens*, as a result of limited gene flow (Casteleyn et al., 2010). In freshwater ecosystems, Heino et al. (2010) found that diatoms were not ubiquitously distributed over a wide geographical range in boreal streams. None of the papers dealing with these questions concerned diatoms inhabiting tropical island freshwater ecosystems, despite their potential interest for biogeographical studies, as has been shown for the Australasian region (Sabbe et al., 2001; Kilroy et al., 2007) or the Antarctic and Sub-Antarctic freshwaters (Sabbe et al., 2003; Van de Vijver et al., 2005). Telford et al. (2006) indicated that cosmopolitan distribution is attained slowly, allowing endemic taxa specializing in rare or isolated habitats to evolve, because diatoms can evolve more quickly than previously thought (Theriot et al., 2006). We therefore carried out our study in geographically-isolated ecosystems with very different climatic characteristics – tropical islands versus the European continent – in order to assess the diversity of a diatom complex, and more particularly to investigate its geographical distribution.

Table III.3: *p-distances between sequences of the five clades of* Gomphonema parvulum.

LSU / ITS	*rbc*L / *cox*1	Clade A	Clade B	Clade B'	Clade C	Clade D
Clade A		0 – 0-0.2	0.9-1.1	0.7-0.8	0.7-1.1	0.6-1.1
		0 – 0-0.2	5.2-5.5	5.2-5.3	3.1-3.4	4.4-4.7
Clade B		0.4	0 – 0-0.2	0.5-0.6	1.0-1.5	0.6-0.9
		2.5-2.7	0 – 0-0.3	0-0.2	3.6-3.9	4.2-4.7
Clade B'		0.4	0	0 – 0	0.8-1.2	0.7-1.0
		2.5-2.7	0	0 – /	4.2-4.5	3.6-3.8
Clade C		0.6-0.8	0.2-0.4	0.2-0.4	0-0.2 – 0-0.4	0.7-1.1
		2.7-3.0	1.7-1.8	1.7-1.8	0-0.5 – 0-1.3	4.4-5.0
Clade D		0.2	0.2	0.2	0.6	0 – 0.1-0.6
		2.3-2.7	2.2-2.3	2.2-2.3	1.5-2.0	0-0.8 – 0-1.4

p-distances of sequences of nuclear markers (LSU rDNA, top and ITS, bottom) are shown in white square, whereas p-distances of sequences of organelle genomes (*rbc*L, top and *cox*1, bottom) are shown in gray squares. "/" indicates that only one sequence of this clade was obtained and so the intra-clade p-distances cannot be calculated.

We focused our work on the diatom species complex, *Gomphonema parvulum* (Kützing) Kützing, which is a very common species that inhabits both temperate and tropical freshwater ecosystems. Due to its high morphological variability, this species is very difficult to identify. In a paper published on the diversity of this species, Wallace & Patrick (1950) suggested that this morphological variability was the result of phenotype plasticity, because they did not find any clear morphological distinction between the different varieties, and no obvious geographical or ecological differentiation. Since this pioneering work, many varieties have been identified in this species complex on the basis of morphological variations (Dawson, 1972; Krammer & Lange-Bertalot, 1991; Reichardt, 1999). Currently 36 records are available for *G. parvulum* in the compilation of names of diatoms (Fourtanier & Kociolek, 2011). Among these varieties and forms, several have no valid status, and others have undergone taxonomic rearrangements (some have now been classified as distinct species), but this diversity of names is indicative of the complexity of the *G. parvulum* complex.

To study the diversity of the *G. parvulum* species complex and to assess the putative existence of geographical differentiation in its distribution, 39 strains were isolated from rivers in Europe and on small tropical islands. A molecular approach was used to define clades, and these clades were then morphologically characterised. The molecular approach consisted of sequencing four DNA fragments: two belonged to the nuclear genome (D1-D2 of LSU rDNA and the 18S-28S ITS regions), one to the mitochondrial genome (*cox*1), and one to the chloroplast genome (*rbc*L). These last three fragments had already been used as diatom barcodes (Evans et al., 2007; Moniz & Kaczmarska, 2009, 2010; Hamsher et al., 2011). LSU rDNA was used to define the *G. parvulum* group in the genus, and the other fragments to study the genetic variability within this group. The clades were then characterized from the conventional morphology of the strains and using geometric morphometry. Numerous recent papers have used this methodology, which enhances diatom identification (Beszteri et al., 2005; Falasco et al., 2009; Fránková et al., 2009; Novais et al., 2009; Poulíčková et al., 2010). Finally, we discuss the geographical differentiation of the clades on the basis of these results.

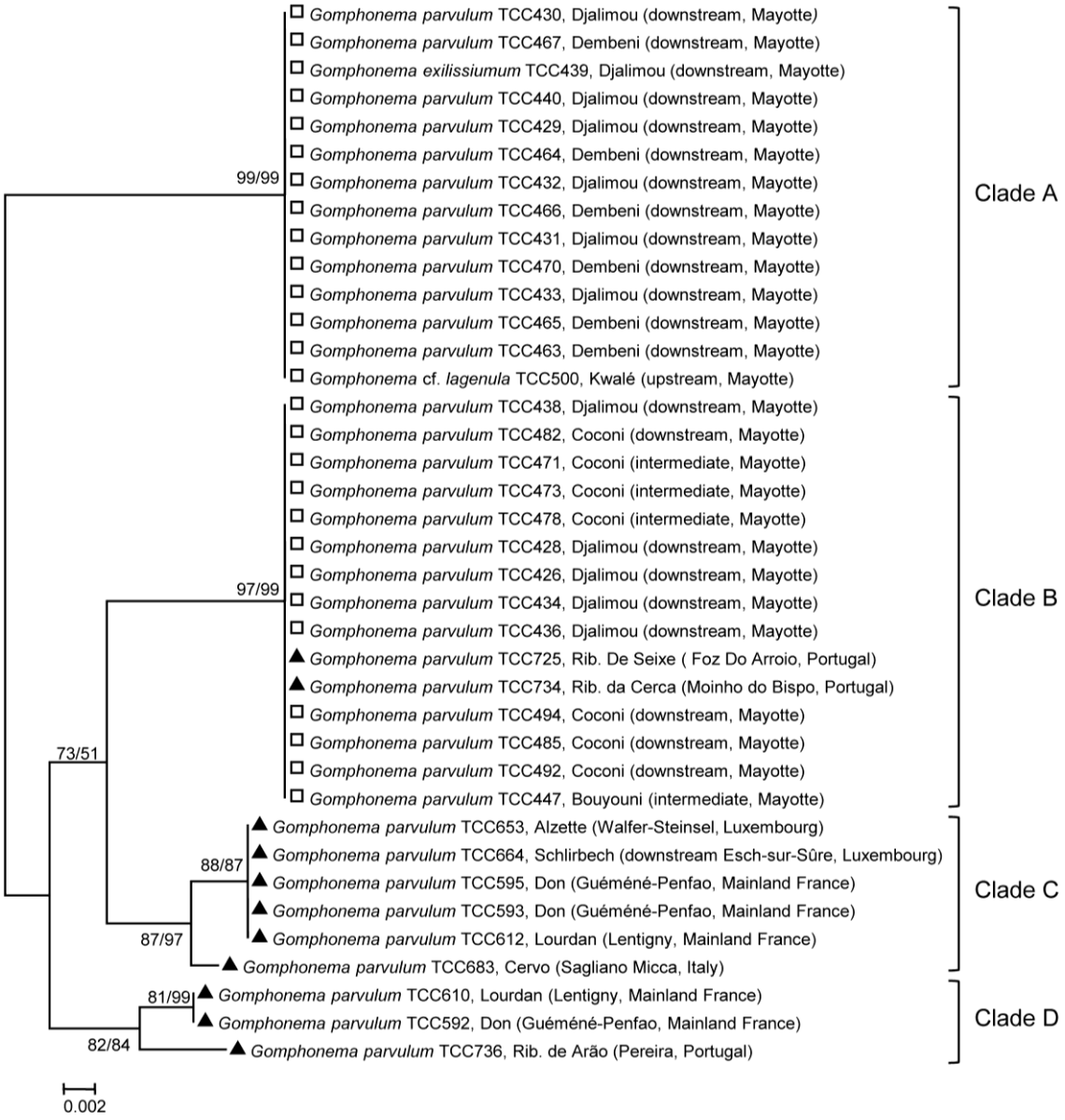


Figure III.6 : *Maximum Likelihood tree based on ITS sequences, illustrating the variation of* Gomphonema parvulum *complex sequences.*
Maximum Likelihood and Neighbor-Joining (ML/NJ) bootstrap values are shown on the branches; empty squares correspond to the tropical strains, and black filled triangles to the temperate strains.

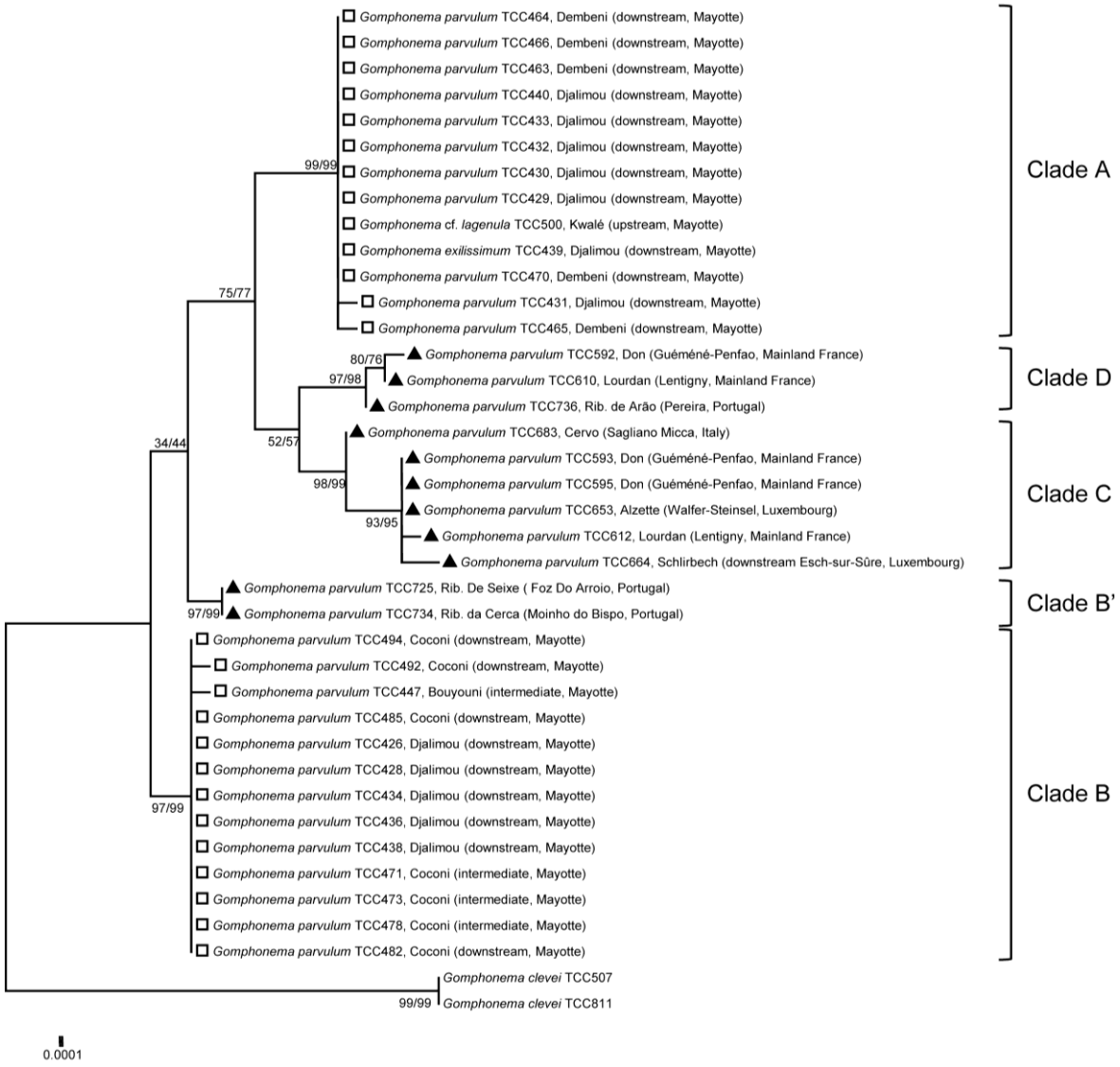


Figure III.7 : *Maximum Likelihood tree based on* rbc*L sequences, illustrating the variation of* Gomphonema parvulum *complex sequences.*
Maximum Likelihood and Neighbor-Joining (ML/NJ) bootstrap values are shown on the branches; empty squares correspond to the tropical strains, and black filled triangles to the temperate strains.

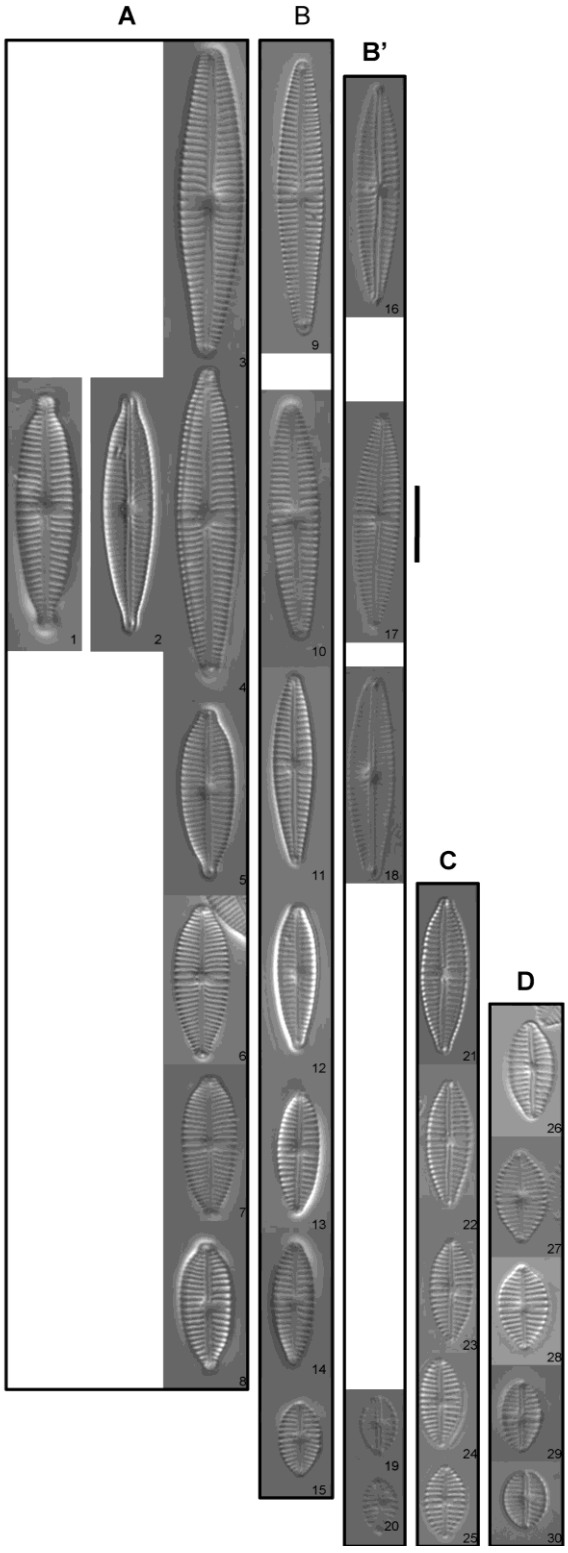


Figure III.8 : *Cleaned valves of* Gomphonema *cf.* lagenula *(1),* G. exilissimum *(2), and* G. parvulum *(3-30), arranged according to the five clades.*
Scale bar = 10 µm. Clade A: 1: TCC500; 2: TCC439; 3: TCC465; 4: TCC432; 5: TCC429; 6:TCC464; 7:TCC465; 8:TCC470. Clade B: 9: TCC473; 10: TCC478; 11: TCC494; 12: TCC492; 13:TCC471; 14: TCC438; 15: TCC482. Clade B': 16-17, 19-20: TCC725; 18: TCC734. Clade C: 21: TCC595; 22: TCC593; 23: TCC612; 24: TCC653; 25: TCC683. Clade D: 26, 28: TCC610; 27: TCC592; 29-30: TCC736.

Results

Molecular differentiation

Whichever method was used (Maximum Likelihood, ML or Neighbor Joining, NJ) the same tree topologies were obtained using the LSU rDNA sequences (Supplementary material). With the exception of one *Gomphonema* cf. *parvulum* sequence found in Genbank (AM710551, strain AT_161.15), which is now identified as *Gomphonema* sp. in the Algaterra project website, all the *G. parvulum* strains were clustered in the same group, which also contained the *G.* cf. *lagenula* Kützing and *G. exilissimum* (Grunow) Lange-Bertalot and Reichardt sequences. Two other *G. affine* Kützing sequences (AM710558, strain AT_219Gel06 and AM710589, strain AT_196Gel03) grouped with *G. parvulum*, whereas our *G. affine* strain TCC526 was clearly separated. The two Genbank *G. affine* strains have also been re-identified, since sequence publication, as *G.* cf. *affine* (Algaterra project). We therefore did not use these sequences in further analyzes. The LSU rRNA sequences positioned the *G. parvulum* group as the sister group of *G. affine* and *G. clevei* Fricke. The LSU rDNA sequences displayed low polymorphism in the *G. parvulum* group including *G.* cf. *lagenula* and *G. exilissimum* (uncorrected pairwise distances, p-distances between 0 and 0.8% in Table III.3).

We studied the *Gomphonema parvulum* group, defined on the basis of the LSU results, with 3 other markers: ITS, *cox*1 and *rbc*L. Four clades containing the same sequences were identified in the ITS and *cox*1 trees, whichever method (ML or NJ) was used (e.g. ITS tree in Figure III.6). In clade A, sequences of *G. parvulum*, *G.* cf. *lagenula* and *G. exilissimum* strains were found, whereas clades B, C and D contained only sequences of *G. parvulum*. On the other hand, 5 clades were found in the *rbc*L tree (Figure III.7). Clades A, C and D were the same as previously, but here, clade B was divided into two subgroups, designated clades B and B', respectively. In all the trees, the clades were supported by high bootstrap values, but the phylogenetic relationships between them were weakly supported, and differed in the different trees (Figure III.6 and Figure III.7).

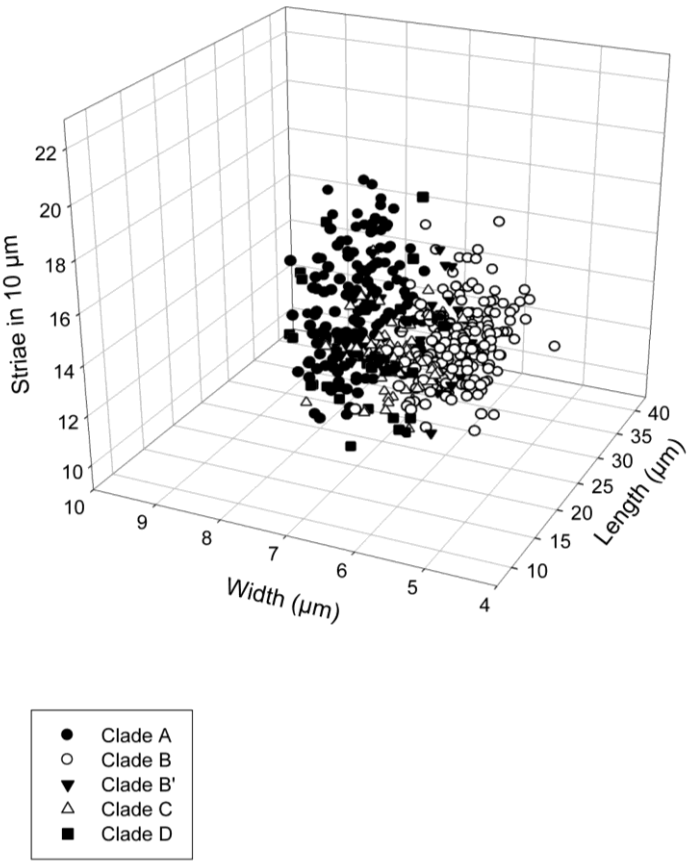


Figure III.9 : *Valve dimension and density of striae of* Gomphonema parvulum *valves.*

For ITS, *cox*1 and *rbc*L, the genetic distances were greater than those found with the LSU rDNA sequences (Table III.3), but each clade was very homogeneous (Table III.3: intraclade distances < 0.8% for ITS, < 1.4% for *cox*1, and <0.6% for *rbc*L). For example, for the ITS sequences, a total of only six different sequences were obtained from all 38 strains, two of which were found only once (TCC683 and TCC736). For the three regions studied (ITS, *cox*1 and *rbc*L), the sequence variability was lower in clades A and B (Table III.3: intraclade distance 0% for ITS, < 0.3% for *cox*1 and <0.2 % for *rbc*L) than in clades C and D (Table III.3: intraclade distance < 0.8% for ITS, < 1.4% for *cox*1 and <0.6 % for *rbc*L).

Morphometric differentiation

Morphological and morphometric studies were performed on all *Gomphonema parvulum* strains previously clustered by the molecular approach. These strains all displayed high morphological variability, as shown in Figure III.8. The cells appeared to range between heteropolar and slightly heteropolar, and had elliptical, lanceolate to linear-lanceolate valves, and rounded, rostrate or capitate poles. The striae were always almost perpendicular to the raphe. The central area was asymmetrical, with a single long stria on one side terminating in a distinct stigma, and a short stria on the other side. Variations were observed between strains, and even between the valves of cells belonging to the same clone, suggesting that in some cases, different life stages of the same clone could have been erroneously assigned to different varieties. Strains were therefore only identified to species level.

For our morphometric examination of the valves (Figure III.9), clade B' was not used for the statistical analysis due to the limited data available for this clade (only two strains). It appeared that there were significant differences in stria density (Mann-Whitney comparisons with the Bonferroni correction, $p<0.05$) between strains belonging to clades A/B and those belonging to clades C/D. Clades C and D displayed a higher density of striae (12 - 19 and 14 – 22 respectively) than clades A and B (10 – 17 and 8 – 17 respectively). Due to automixis of strains belonging to clades A and B (which was not observed for clades C and D), we were able to compare a range of sizes (width and length) in these two clades. A significant difference was found between clades A and B for both length (Mann-Whitney with Bonferroni correction $p<0.003$) and valve width ($p<0.001$). The size range for clade B (length: 9 - 36 µm; width: 4.9 – 7.2) was smaller than that of clade A (length: 16 - 42 µm; width: 6.3 – 9.6).

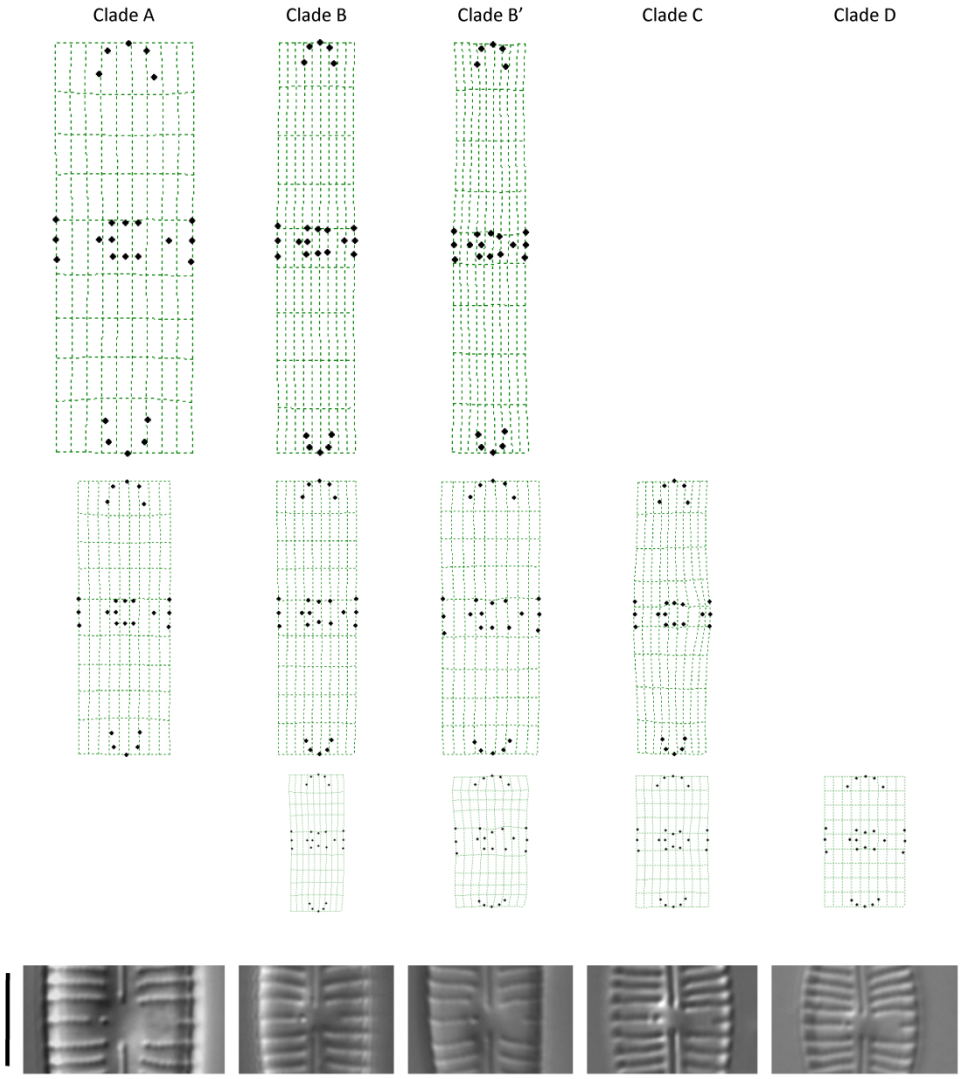


Figure III.10: *Mean landmark configurations at different fixed lengths (10, 20, and 30 µm) with an example of the central area focus for each of the five clades.*
Scale bar = 5 µm.

However, there was a considerable overlap between these clades for the stria density, length and width.

The multivariate regression of procrustes coordinates on centroid size confirmed the occurrence of a strong allometric effect accounting for almost 68% of the total shape variation. Determinations of the relative contribution of each landmark showed that they made differing contributions to the frustule variation. The landmarks placed on the stigma and on the two raphe central endings contributed 10.7%, 11.9% and 11.2%, respectively, to the total variance. The contributions of the other landmarks ranged between 0.4 and 6.4%. Mean landmark configurations for the five clades at different fixed lengths are shown in Figure III.10. Due to the different size ranges, all configurations could not be obtained. Clade A presented poles that were more rostrate than the other clades, which often had rounded poles although some rostrate poles were also observed (Figure III.8). Examples of picture of the central area are shown for each clade in Figure III.10. Clade A showed a straighter central area, whereas the other clades had central raphe endings slightly inside the central area. Different configurations of the 2 striae near the stigma were also observed. These striae were aligned with the stigma in clades A, C and D, whereas they form a symmetric convex line with the stigma in clade B, and an asymmetric convex line in clade B'.

Biogeography

Using LSU rDNA, no geographical differentiation was found between the strains reflecting whether they were of tropical or temperate origin. Using ITS, *rbc*L and *cox*1 sequences, only sequences from tropical strains were found in clade A, whereas only sequences from the European continent were found in clades C and D. Clade B contained sequences from tropical strains, plus two Portuguese strains (in the ITS and *cox*1 trees, e.g. Figure III.6), which were divided into two subgroups B and B' depending on their geographical origin in the *rbc*L tree (Figure III.7). At a large geographical scale, in the phylogenetic tree obtained from the concatenation of the ITS, *cox*1 and *rbc*L sequences (data not shown), each of the fives clades corresponded to only one geographic origin (either temperate or tropical). Even when tropical and temperate clades were grouped together (clade A with clade D and clades B and B' with clade C), there was a clear biogeographical differentiation between the tropical and temperate strains. In contrast, at smaller geographical scales, no such differentiation could be detected.

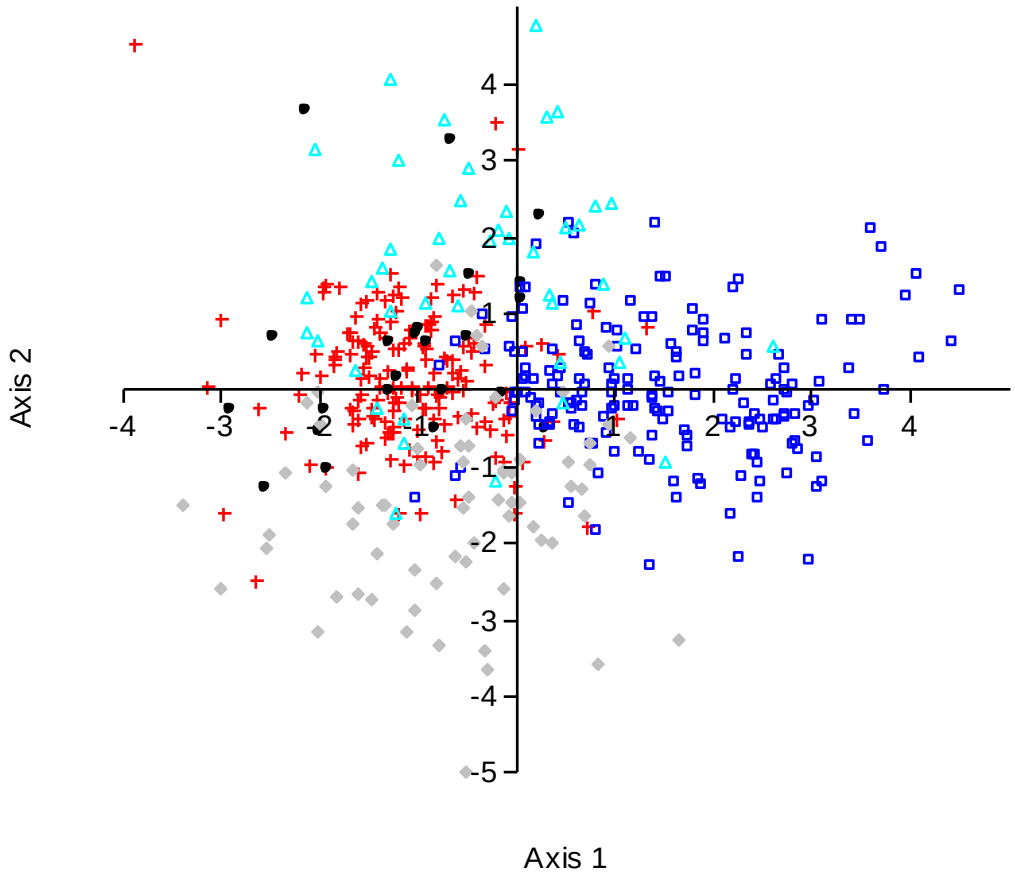


Figure III.11: *Canonical variate analysis using geometric morphometric data for the five clades of* Gomphonema parvulum.
Squares represent clade A; crosses clade B; dots clade B'; diamonds clade C, and triangles clade D.

For example, the sequences of twelve strains isolated in Djalimou were distributed within the two tropical clades (A and B), and in the same way, some of the sequences distributed within clades C and D came from the same mainland French sites (Figure III.6 and Figure III.7). Large phylogenetic distances were sometimes found between strains isolated from the same geographic area (e.g. the *cox*1 sequences of TCC494 and TCC432, which were both from Mayotte Island, showed a 5.5% distance), whereas the same ITS and *cox*1 sequences were also found in strains isolated from very distant geographic areas (e.g. TCC734 from Portugal and TCC473 from Mayotte).

Canonical variate analysis (Figure III.11) of our morphometric data, successfully separated the 2 tropical clades A and B and the 2 temperate clades C and D (Hotelling's pairwise comparisons with Bonferroni correction, $p<0.001$). Moreover, tropical clades A and B were also significantly different from temperate clades C and D ($p<0.001$). Interestingly, temperate clade B', which was genetically closely related to tropical clade B, was significantly separated from tropical clades A and B ($p<0.001$), but not from temperate clades C and D ($p>0.05$), thus displaying a clear differentiation between tropical and temperate clades. Similarly, the 2 tropical clades were clearly differentiated. In contrast, it was more difficult to differentiate between the temperate clades on the basis of the morphological data, but this could have been linked to the smaller number of strains in these clades.

Discussion

The goals of this study were to enhance the characterization of the *Gomphonema parvulum* complex by a polyphasic approach combining genetic, morphological and morphometric methods, and to perform a biogeographical study of this species complex by comparing strains isolated from rivers located on tropical islands in the Southern hemisphere to those from rivers in Europe.

Molecular clades of the G. parvulum *species complex*

The genetic data clearly demonstrated that five clades could be distinguished within the set of strains. The LSU rDNA sequences showed that the morpho-group of *G. parvulum* strains forms a consistent phylogenetic group including *G. exilissimum* and *G* cf. *lagenula,* and seems to be closely related to *G. affine* and *G. clevei*. In contrast, *G. micropus* Kützing, which was formerly considered to be a variety of *G. parvulum,* has

previously been demonstrated by other phylogenetic analyses (Bruder & Medlin, 2008; Kermarrec et al., 2011) to be very distant from this group. But LSU rDNA results alone were not sufficiently variable to be used to assess the diversity within this group. Four clades were identified among the *G. parvulum* strains using ITS and *cox*1 sequences, and five using *rbc*L sequences. Moreover, using the ITS and *cox*1 sequences, a clear difference was found between intra- and interclade distances. These findings raise the question of the taxonomic status of these clades, which could be classified either as different populations belonging to the same species or as distinct species. No genetic differentiation was revealed between clades B and B' by two markers (LSU and ITS), whereas slight divergences were observed with the *rbc*L and *cox*1 sequences. Clades B and B' therefore seemed to be populations of the same species. Evans et al. (2009) had previously demonstrated that geographical patterns of genetic diversity can exist between different populations of the same diatom species. By looking at the genetic distances between the four clades (A, B, C and D) estimated from their ITS, *cox*1 and *rbc*L sequences, it appeared that they were comparable to those found between clearly-established species. For example, Moniz & Kaczmarska (2010) studied intra- and interspecies distances in several families of diatoms, and defined the intraspecies distance as ranging from 0 to 17%, and the interspecies distance as ranging from 2 to 45% in pennate diatoms. Similarly, the interclade distances (3.1-5.5%) for the *cox*1 sequences were equal to or even greater than the interspecies distances estimated for *Sellaphora* (Evans et al., 2007). Finally, for the last sequenced region (*rbc*L), Trobajo et al. (2010) found little divergence between reproductively-isolated groups of *Nitzschia palea* (1%). Although we would need more strains and mating experiments to definitively conclude about the taxonomic level of these clades, the genetic distances between the four clades (A, B, C and D) are therefore congruent with interspecies distances.

Morphological characterization of G. parvulum *clades*

Our morphometric analysis supported the clustering of our strains obtained by molecular methods, even though vegetative division generated morphological variations. Different ranges of size and stria density, which are generally used to define diatom taxa, were observed. Moreover our data showed that the main variations occurred in the central area of the cells, although differentiation between *G. parvulum* varieties has conventionally been based mainly on the shape of the poles. The central area has previously shown its usefulness for discriminating between *Navicula cryptocephala* and

N. trivialis (Veselá et al., 2009) and *Gomphonema rosentockianum* and *G. tergestinum* (Novais et al., 2009) by means of geometric morphometric analyses. Even though more strains would be necessary to confirm our results (in order to avoid the possible effect of strain-specific signals in the poorly-represented clades), the central area seems to be an important characteristic for distinguishing between the different clades of *G. parvulum*. In agreement with Cox (2010), who referred to unpublished findings that heteropolar diatoms display diverse shapes at the head pole during size reduction, our study confirmed that the shape of the poles cannot be used alone to clarify taxonomy in the *G. parvulum* complex. In addition, although the four main clades (A, B, C and D) appeared to be differentiated by our morphometric approaches, there was a degree of overlap between them, making it difficult to distinguish clearly on the basis of frustule morphology alone. Moreover clades B and B', which genetic data suggests are populations of the same species, were in fact separated by our morphometric analyses, whereas no separation was observed between clades B', C and D, demonstrating the limits of light microscopy for distinguishing between groups in the *G. parvulum* complex.

Biogeography of G. parvulum *clades*

A clear biogeographical differentiation between the strains belonging to clade A and those belonging to clades C and D was observed. For clades B and B', the two temperate strains appeared to be mixed with the tropical strains on the basis of their ITS and *cox*1 sequences, but *rbc*L sequences made it possible to distinguish between the two clades that reflects the geographic origin of the strains. More strains would be necessary to confirm these initial findings, but our data already suggest that some *G. parvulum* clades are not ubiquitous, and that a tropical versus temperate biogeographical differentiation may exist for some diatom species. The same kind of tropical versus temperate differentiation had previously been reported for freshwater bacterial species by Humbert et al. (2009), which might suggest that the processes that lead to this biogeographical differentiation also affect other freshwater microbial communities.

The large genetic interclade distances found between clades A and B, and the fact that in the tree resulting from the concatenation of the ITS, *cox*1 and *rbc*L sequences (data not shown), clades D and B were clustered together with a significant bootstrap value, suggest that the presence of the clades A and B in the tropical island of Mayotte resulted from two distinct colonization events. Moreover, the very slight variability found within clades A and B supports the hypothesis that a small number of propagules were involved

in these colonization events, even if we cannot exclude the possibility that other processes (e.g. strong selective pressures, genetic drift…) could have also contributed to the very low level of genetic variability found in clades A and B. Finally, the slight differences in the *rbc*L, ITS and *cox*1 fragments found between the tropical and the temperate strains belonging to clades B and B' suggest that these two populations have only recently been isolated (more recently than the separation of clades A and B), however this too needs to be validated using a larger number of strains. All these findings suggest that the *G. parvulum* complex, which was hitherto thought to be ubiquitously distributed, in fact displays a much more complicated pattern of distribution. Ten years ago, Kociolek & Spaulding (2000) stated that freshwater diatom research was predominantly based on the paradigm of cosmopolitan geographic distribution of species and environmental selection but, drawing on the literature they disagreed with this point of view. Our results confirm the findings of other recent studies of diatoms revealing biogeographic differentiation patterns (Kilroy et al., 2007; Vyverman et al., 2007; Casteleyn et al., 2008; Vanormelingen et al., 2008; Evans et al., 2009) and the possible existence of endemic species (Sabbe et al., 2001, 2003).

Conclusion

The *Gomphonema parvulum* complex consists of at least four clades. Interestingly, on the basis of their genetic characteristics the morphospecies *G. lagenula* and *G. exilissimum* (formerly both thought to be *G. parvulum* varieties) were both clearly included in clade A. However, only one *G.* cf. *lagenula* strain and one *G. exilissimum* strain were used in this study, and so it will be necessary to look at more strains of these two species, before we can definitively conclude whether they constitute one species or two different species. Diatom biogeography depends on the taxa delineation used as the basis of classification. Van de Vijver & Beyens (1999) observed that the sub-Antarctic diatom communities were characterized by having a high proportion of cosmopolitan species. A few years later, the application of a more precise morphological taxonomy revealed a clear biogeographic pattern (Sabbe et al., 2003; Van de Vijver et al., 2005). Although the *G. parvulum* complex was ubiquitously distributed, the use of a finer taxonomy also identified a more restricted distribution of some of the clades belonging to this complex. Four main clades displaying different genetic, morphological, and geographical characteristics have been defined in this study. The congruence between the

findings from molecular markers, the low intraclade genetic variation, and the higher interclade variation all suggest that gene flow is limited between sympatric clades A/B and sympatric clades C/D. Moreover, two geographically separated populations (B and B') of the same clade showed less genetic differentiation than was found between sympatric clades. If we consider a diatom species to be an evolutionary lineage, the genetic differentiation of these clades is indicative of lineage separation (De Queiroz, 2007). The morphological approach provided complementary criteria to support speciation. According to Alverson (2008) and De Queiroz (2007), multiple congruent lines of evidence lead to more robust species limits. The division into distinct sibling species will therefore require further evidence, with confirmation from broader sampling, and from mating experiments and ecological studies.

Methods

Isolation and culture of the different strains

Fifty-six *Gomphonema* strains were isolated from samples collected from rivers on two tropical islands in the Indian Ocean (Mayotte and La Réunion) and in Europe (Luxembourg, Italy, Mainland France, Portugal and Spain). Monoclonal strains were obtained by isolation using sterile Pasteur micropipettes, and cultured in a medium containing 50% 0.2-µm filtered water from the river site and 50% culture medium. Non-axenic, unialgal cultures (Table III.4) were maintained in 100% culture medium in a growth chamber at 20°C and with a 15/9 light/dark cycle. Because of the size reduction cycle, it was not possible to maintain all the strains after this, which made it impossible to conduct any further mating experiments. Some of the strains are still available from the Thonon Culture Collection (http://www6.inra.fr/carrtel-collection).

Among these 56 strains collected from 21 localities, the following species were identified by light microscopy examination: *G. acuminatum* Ehrenberg (2 strains), *G. affine* (1), *G. angustum* C. Agardh (1), *G. bourbonense* E. Reichardt (8), *G. clavatum* Ehrenberg (1), *G. clevei* (2), *G. micropus* (1), *G. parvulum* s.l. (39: 37 *G. parvulum* s.s., 1 *G. exilissimum* and 1 *G.* cf. *lagenula*), *G. pumilum* (Grunow) Reichardt and Lange-Bertalot (1).

Table III.4 : *List of strains sequenced in this study and their GenBank accession numbers.*

Taxon name	code	River (Site, Country) where sampled	date sampled (dd.mm.yyyy)	Accession number			
				28S	ITS	*rbc*L	*cox*1
Gomphonema acuminatum	TCC 680	Regnana (upstream de Bedollo, Italy)	27.10.2006	JQ354609	/	/	/
Gomphonema acuminatum	TCC 737	Rib. de Arão (Pereira, Portugal)	05.12.2006	JQ354616	/	/	/
Gomphonema affine	TCC 526	Saint Denis (upstream, La Réunion)	28.04.2009	JQ354599	/	/	/
Gomphonema angustum	TCC 460	Dembeni (downstream, Mayotte)	18.04.2009	JQ354580	/	/	/
Gomphonema bourbonense	TCC 441	Bouyouni (intermediate, Mayotte)	19.04.2009	JQ354573	/	/	/
Gomphonema bourbonense	TCC 450	Bouyouni (intermediate, Mayotte)	19.04.2009	JQ354575	/	/	/
Gomphonema bourbonense	TCC 451	Bouyouni (intermediate, Mayotte)	19.04.2009	JQ354576	/	/	/
Gomphonema bourbonense	TCC 452	Bouyouni (intermediate, Mayotte)	19.04.2009	JQ354577	/	/	/
Gomphonema bourbonense	TCC 453	Bouyouni (intermediate, Mayotte)	19.04.2009	JQ354578	/	/	/
Gomphonema bourbonense	TCC 458	Dembeni (downstream, Mayotte)	18.04.2009	JQ354579	/	/	/
Gomphonema bourbonense	TCC 513	Bouyouni (upstream, Mayotte)	19.04.2009	JQ354597	/	/	/
Gomphonema bourbonense	TCC 514	Bouyouni (upstream, Mayotte)	19.04.2009	JQ354598	/	/	/
Gomphonema clavatum	TCC 527	Saint Denis (upstream, La Réunion)	28.04.2009	JQ354600	/	/	/
Gomphonema clevei	TCC 469	Dembeni (downstream, Mayotte)	18.04.2009	JQ354586	/	/	JQ354710
Gomphonema clevei	TCC 507	Longoni (downstream, Mayotte)	20.04.2009	JQ354596	/	JQ354682	JQ354718
Gomphonema clevei	TCC 811	Saint Denis (upstream, La Réunion)	28.04.2009	/	/	JQ354694	/
Gomphonema exilissimum	TCC 439	Djalimou (downstream, Mayotte)	18.04.2009	JQ354571	JQ354627	JQ354665	/
Gomphonema cf. lagenula	TCC 500	Kwalé (upstream, Mayotte)	18.04.2009	JQ354595	JQ354643	JQ354681	JQ354717
Gomphonema micropus	TCC 704	Bernesga (Alija de la Ribera, Spain)	02.11.2006	JQ354611	/	/	/
Gomphonema parvulum	TCC 426	Djalimou (downstream, Mayotte)	18.04.2009	JQ354561	JQ354617	JQ354655	JQ354695
Gomphonema parvulum	TCC 428	Djalimou (downstream, Mayotte)	18.04.2009	JQ354562	JQ354618	JQ354656	JQ354696
Gomphonema parvulum	TCC 429	Djalimou (downstream, Mayotte)	18.04.2009	JQ354563	JQ354619	JQ354657	JQ354697
Gomphonema parvulum	TCC 430	Djalimou (downstream, Mayotte)	18.04.2009	JQ354564	JQ354620	JQ354658	JQ354698
Gomphonema parvulum	TCC 431	Djalimou (downstream, Mayotte)	18.04.2009	JQ354565	JQ354621	JQ354659	JQ354699
Gomphonema parvulum	TCC 432	Djalimou (downstream, Mayotte)	18.04.2009	JQ354566	JQ354622	JQ354660	JQ354700
Gomphonema parvulum	TCC 433	Djalimou (downstream, Mayotte)	18.04.2009	JQ354567	JQ354623	JQ354661	JQ354701
Gomphonema parvulum	TCC 434	Djalimou (downstream, Mayotte)	18.04.2009	JQ354568	JQ354624	JQ354662	JQ354702
Gomphonema parvulum	TCC 436	Djalimou (downstream, Mayotte)	18.04.2009	JQ354569	JQ354625	JQ354663	JQ354703
Gomphonema parvulum	TCC 438	Djalimou (downstream, Mayotte)	18.04.2009	JQ354570	JQ354626	JQ354664	/

Gomphonema parvulum	TCC 440	Djalimou (downstream, Mayotte)	18.04.2009	JQ354572	JQ354628	JQ354666	/
Gomphonema parvulum	TCC 447	Bouyouni (intermediate, Mayotte)	19.04.2009	JQ354574	JQ354629	JQ354667	JQ354704
Gomphonema parvulum	TCC 463	Dembeni (downstream, Mayotte)	18.04.2009	JQ354581	JQ354630	JQ354668	JQ354705
Gomphonema parvulum	TCC 464	Dembeni (downstream, Mayotte)	18.04.2009	JQ354582	JQ354631	JQ354669	JQ354706
Gomphonema parvulum	TCC 465	Dembeni (downstream, Mayotte)	18.04.2009	JQ354583	JQ354632	JQ354670	JQ354707
Gomphonema parvulum	TCC 466	Dembeni (downstream, Mayotte)	18.04.2009	JQ354584	JQ354633	JQ354671	JQ354708
Gomphonema parvulum	TCC 467	Dembeni (downstream, Mayotte)	18.04.2009	JQ354585	JQ354634	JQ354672	JQ354709
Gomphonema parvulum	TCC 470	Dembeni (downstream, Mayotte)	18.04.2009	JQ354587	JQ354635	JQ354673	JQ354711
Gomphonema parvulum	TCC 471	Coconi (intermediate, Mayotte)	18.04.2009	JQ354588	JQ354636	JQ354674	JQ354712
Gomphonema parvulum	TCC 473	Coconi (intermediate, Mayotte)	18.04.2009	JQ354589	JQ354637	JQ354675	JQ354713
Gomphonema parvulum	TCC 478	Coconi (intermediate, Mayotte)	18.04.2009	JQ354590	JQ354638	JQ354676	/
Gomphonema parvulum	TCC 482	Coconi (downstream, Mayotte)	18.04.2009	JQ354591	JQ354639	JQ354677	JQ354714
Gomphonema parvulum	TCC 485	Coconi (downstream, Mayotte)	18.04.2009	JQ354592	JQ354640	JQ354678	/
Gomphonema parvulum	TCC 492	Coconi (downstream, Mayotte)	18.04.2009	JQ354593	JQ354641	JQ354679	JQ354715
Gomphonema parvulum	TCC 494	Coconi (downstream, Mayotte)	18.04.2009	JQ354594	JQ354642	JQ354680	JQ354716
Gomphonema parvulum	TCC 592	Don (Guéméné-Penfao, Mainland France)	18.08.2009	JQ354602	JQ354644	JQ354683	JQ354719
Gomphonema parvulum	TCC 593	Don (Guéméné-Penfao, Mainland France)	18.08.2009	JQ354603	JQ354645	JQ354684	JQ354720
Gomphonema parvulum	TCC 595	Don (Guéméné-Penfao, Mainland France)	18.08.2009	JQ354604	JQ354646	JQ354685	/
Gomphonema parvulum	TCC 610	Lourdan (Lentigny, Mainland France)	19.08.2009	JQ354605	JQ354647	JQ354686	JQ354721
Gomphonema parvulum	TCC 612	Lourdan (Lentigny, Mainland France)	19.08.2009	JQ354606	JQ354648	JQ354687	JQ354722
Gomphonema parvulum	TCC 653	Alzette (Walfer-Steinsel, Luxembourg)	13.10.2006	JQ354607	JQ354649	JQ354688	JQ354723
Gomphonema parvulum	TCC 664	Schlirbech (downstream Esch-sur-Sûre, Luxembourg)	10.10.2006	JQ354608	JQ354650	JQ354689	/
Gomphonema parvulum	TCC 683	Cervo (Sagliano Micca, Italy)	15.12.2006	JQ354610	JQ354651	JQ354690	JQ354724
Gomphonema parvulum	TCC 725	Rib. De Seixe (Foz Do Arroio, Portugal)	04.12.2006	JQ354612	JQ354652	JQ354691	/
Gomphonema parvulum	TCC 727	Rib. De Seixe (Foz Do Arroio, Portugal)	04.12.2006	JQ354613	/	/	/
Gomphonema parvulum	TCC 734	Rib. da Cerca (Moinho do Bispo, Portugal)	04.12.2006	JQ354614	JQ354653	JQ354692	JQ354725
Gomphonema parvulum	TCC 736	Rib. de Arão (Pereira, Portugal)	05.12.2006	JQ354615	JQ354654	JQ354693	JQ354726
Gomphonema pumilum	TCC 536	Bras des Etangs (La Réunion)	24.04.2009	JQ354601	/	/	/

Table III.5 : *Genes and primers used for PCR and sequencing, and the number of positions used after alignment and corrections (including indels).*

	Gene	Primer Name	Primer Sequence (5'-3')	References	Alignment length
PCR	LSU	D1R	ACC CGC TGA ATT TAA GCA TA	Scholin et al., 1994	534 positions
		D2C	CCT TGG TCC GTG TTT CAA GA		
	ITS	ITS 5F	GGA AGT AAA AGT CGT AAC AAG G	White et al., 1990	681 positions
		ITS 4R	TCC TCC GCT TAT TGA TAT GC		
	*rbc*L	DPrbcL1-F	AAG GAG GAA DHH ATG TCT	Daugbjerg & Andersen, 1997b	1223 positions
		DPrbcL7-R	AAA SHD CCT TGT GTW AGT YTC		
	*cox*1	GazF2	CAA CCA YAA AGA TAT WGG TAC	Saunders, 2005	641 positions
		KEdtmR	AAA CTT CWG GRT GAC CAA AAA	Evans et al., 2007	
Sequencing	*rbc*L	NDrbcL6	GTA AAT GGA TGC GTA	Daugbjerg & Andersen, 1997b	
		15R	ACA CCA GAC ATA CGC ATC CA	Jones et al., 2005	
		16F	TTA GAA GAT ATG CGT ATT		

Molecular methods

Cells were harvested from the cultures by centrifuging. After lysing the cells, nucleic acids were coprecipitated using GenElute^{TM}-LPA (Sigma-Aldrich) according to the Manufacturer's instructions, and dissolved in TE buffer. Several gene fragments were amplified: D1-D2 region of the LSU rDNA gene, ITS1-5.8S-ITS2, and partial *rbc*L and *cox*1 genes. All PCRs were performed using the primers listed in Table III.5, and according to the authors' specifications in a PCR thermal cycler (T personal, Biometra, Göttingen, Germany). Additional internal sequencing primers were used to retrieve the complete sequences of *rbc*L (Table III.5). Sanger sequencing was performed by Beckman Coulter Genomics (Takeley, United Kingdom), GATC (Konstanz, Germany), and Genoscope (Evry, France).

Sequence alignment and phylogenetic analysis

Sequences from both directions, together with internal fragments of the *rbc*L, were manually assembled using BioEdit Sequence Alignment Editor v. 7.0.5 (Hall, 1999). LSU rDNA sequences from Genbank were added to this analysis for species belonging to the genus *Gomphonema* Ehrenberg. The accession numbers of these sequences are provided in the Supplementary Material. No *Gomphonema parvulum* sequences were available on Genbank for ITS, *rbc*L or *cox*1. Sequences were aligned using the CLUSTAL W procedure (Thompson et al., 1994), which is available in BioEdit Sequence Alignment Editor v. 7.0.5 (Hall, 1999). The use of the secondary structure was not necessary to align the LSU D1/D2 region and ITS rDNA sequences, because there was no ambiguity in the alignment. Uncorrected pairwise distances were calculated from alignments (lengths in Table III.5) using MEGA5 (Tamura et al., 2011).

In order to root the phylogenetic trees, outgroups were chosen according to the criteria of Verbruggen & Theriot (2008). These outgroups included sequences of *G. micropus* for the LSU rDNA sequences, and sequences of *G. clevei* for the *rbc*L and *cox*1 genes. Only the tree obtained from ITS sequences was unrooted, because no good quality outgroup was available. ML and NJ phylogenetic trees were calculated using MEGA5 (Tamura et al., 2011). The ML analyses were carried out using the optimum model of sequence evolution selected according to the Akaike information criterion with 1000 bootstrap replicates. The NJ analyses were also done using the correction of Jukes & Cantor (1969) with 1000 bootstrap replicates.

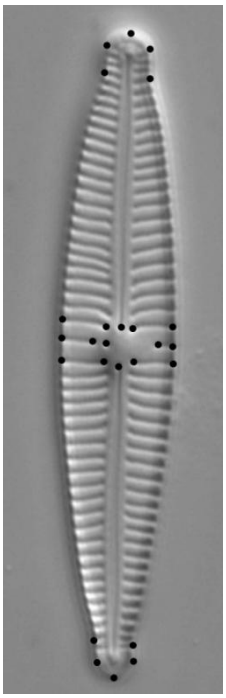


Figure III.12 : *Light micrograph showing the position of the 25 landmarks on the pole and the central area used in the geometric morphometric analysis.*

Morphological examination and geometric morphometric analysis

Diatoms from each of the 39 *G. parvulum* s.l. cultures were cleaned with 30% nitric acid. The frustules were rinsed with demineralized water. The diatoms were then mounted in synthetic resin (Naphrax©), and slides were prepared for light microscopy. Ten to 15 light micrographs of each strain were taken using a Zeiss Axio Imager A1 microscope and a Zeiss AxioCam HRc camera. The length, width, and density of striae were measured for each frustule.

Geometric morphometric analyses were performed on the same micrographs (10 to 15 per clone, i.e. a total of 488 frustules). The tps-series software (Rohlf, 2007) was used to analyze the morphology of *G. parvulum* s.l.. Twenty-five landmarks were digitized on the valve poles (10) and on the central area (15) using tpsDig software (version 2.16) (as shown in Figure III.12). For the poles, landmarks were located at the extremities of the valve, and for each side and each pole on the maximum convex and concave curvature. Concerning the central area, landmarks were positioned on the stigma, on the raphe central endings, and at each extremity of the 6 striae surrounding the central area.

The coordinates were adjusted by General Procrustes Analysis (GPA), because the cells were not all at comparable life cycle stages. Using GPA eliminated variations introduced by differences in the position, orientation or size of the specimens, TpsRegr (version 1.46) was used to determine the allometric effect on the total variation. We performed a Principal Component Analysis (PCA) on these data using tpsRelw (version 1.46). PCA residuals were visualized with Past (version 2.06, Hammer et al., 2001). The separation between clades was tested by Multivariate Analysis of Variance (MANOVA) and Canonical Variate Analysis (CVA) in Past after omitting the first principal component, which was strongly correlated with the size differences. These analyses provided an ordination that maximized the separation between the group means relative to the variation within groups.

Acknowledgements

This work is part of the @SPEED-ID project (“accurate SPEciEs Delimitation and Identification of eukaryotic biodiversity using DNA markers”) proposed by F-Bol, the French Barcode of Life initiative. Corinne Cruaud and Arnaud Couloux from Genoscope are thanked for sequencing. We also thank Luc Ector and the CRP Gabriel Lippmann

Centre for providing strains, and Florence Pérès, Maurice Bey, Michel Coste, Gilles Gassioles, Didier Guillard and René Le Cohu for sending fresh samples. This work was funded by ONEMA (French National Agency for Water and Aquatic Environments) and ANRT (French National Agency for Research and Technology). Monika Ghosh is acknowledged for improving the English version of the manuscript. Dr. Bank Beszteri and the anonymous reviewer are thanked for helping in improving the manuscript.

Supplementary material

Table III.6 : *Accession numbers of the* Gomphonema *sequences downloaded from Genbank.*

Taxon name	**GenBank accession number**
	LSU
Gomphonema acuminatum Ehrenberg	AM710575
Gomphonema affine Kützing become *G.* cf. *affine*	AM710558
Gomphonema affine become *G.* cf. *affine*	AM710589
Gomphonema cf. *angustatum* (Kützing) Rabenhorst	AM710561
Gomphonema micropus Kützing	AM710520
Gomphonema micropus	AM710519
Gomphonema cf. *parvulum* (Kützing) Kützing become *G.* sp	AM710551
Gomphonema productum (Grunow) Lange-Bertalot and Reichardt	AM710549
Gomphonema truncatum Ehrenberg	AM710598

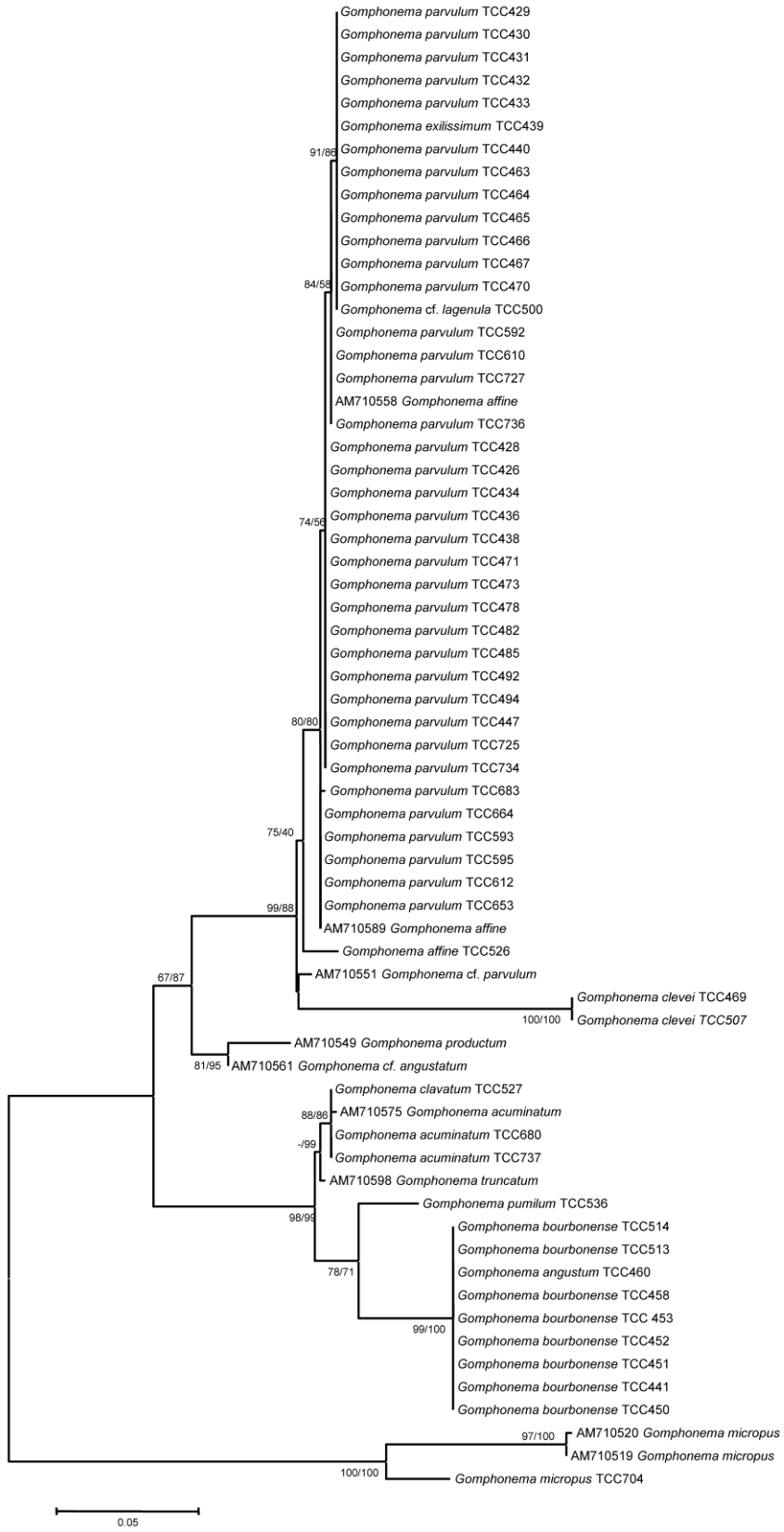


Figure III.13 : *Maximum Likelihood tree based on the LSU D1/D2 region, illustrating the phylogeny of* Gomphonema *sequences.*
Maximum Likelihood and Neighbor-Joining (ML/NJ) bootstrap values are shown on the branches.

3.2. *Nitzschia palea*

La famille des Bacillariaceae Ehrenberg comprend 594 espèces réparties en 18 genres dont les principaux sont : *Denticula* Kützing, *Fragilariopsis* Hustedt, *Hantzschia* Grunow, *Nitzschia* Hassall, *Psammodictyon* D.G. Mann, *Pseudo-nitzschia* H. Peragallo et *Tryblionella* W. Smith (Round et al., 1990). Au sein de cette famille des Bacillariaceae, les relations phylogénétiques sont complexes (Rimet et al., 2011). La plupart des espèces de la famille des Bacillariaceae appartiennent au genre Nitzschia (345 espèces selon la base de données « Algaebase », http://www.algaebase.org/). La délimitation de certaines espèces de *Nitzschia* est problématique. Parmi celles-ci, le complexe d'espèce *N. palea* (Kützing) W. Smith présente de nombreuses formes et variétés : 44 noms de taxa lui sont associés dans le « Catalogue of Diatom Names » (Fourtanier & Kociolek, 2011). Ce complexe est fréquemment observé dans les échantillons prélevés dans le cadre de la bioindication, et est largement distribué géographiquement (Finlay et al., 2002).

Comme pour le complexe *G. parvulum*, l'identification exacte des variétés de *N. palea* est importante car ces variétés présentent des préférences écologiques différentes. Par exemple, *N. palea* a été considérée comme une espèce résistante à la pollution organique (Khan, 1990) ou résistante à différents toxiques (Duong et al., 2010; Bere & Tundisi, 2011). Au contraire, *N. palea* var. *debilis* semble être indicatrice de faibles concentrations en nutriments (Potapova & Charles, 2007). Une valeur de sensibilité de 1.0 et une valeur indicatrice de 3.0 pour l'IPS sont attribuées à *N. palea* tandis que sa variété *debilis* possède une valeur de sensibilité de 3.0 et une valeur indicatrice de 1.0. De plus, il a été montré que la variété *debilis* n'était ni une réponse phénotypique aux conditions environnementales ni un clade particulier du complexe *N. palea* (Trobajo et al., 2009). En outre, dans cette même étude, des expériences de croisement entre différentes souches ont révélé que le complexe *N. palea* était composé de plusieurs espèces qui n'étaient pas interfécondes alors que la morphologie des souches présentait un continuum. L'objectif de nos travaux était donc d'étudier à nouveau la diversité au sein de ce complexe d'espèces en complétant les données moléculaires disponibles sur un plus large échantillonnage de souches.

Tableau III.7 :*Liste des souches de* Nitzschia palea *utilisées pour cette étude*.
Les * indiquent les souches pour lesquelles R. Trobajo a tenté des expériences de croisement.
Toutes ces souches ont été testées deux à deux à deux reprises.

Code TCC	Taxon	Origine du prélèvement : Pays (Rivière ou Lac, site)	Date de prélèvement
TCC139-1	*Nitzschia palea*	France métropolitaine (Lac Léman)	-
TCC139-2	*Nitzschia palea*	France métropolitaine (Lac Léman)	-
TCC139-3	*Nitzschia palea*	France métropolitaine (Lac Léman)	-
TCC425	*Nitzschia palea*	Mayotte (Djalimou, aval)	18.04.2009
TCC427	*Nitzschia palea*	Mayotte (Djalimou, aval)	18.04.2009
TCC435	*Nitzschia palea*	Mayotte (Djalimou, aval)	18.04.2009
TCC437	*Nitzschia palea*	Mayotte (Djalimou, aval)	18.04.2009
TCC456	*Nitzschia palea*	Mayotte (Dembeni, aval)	18.04.2009
TCC457	*Nitzschia palea*	Mayotte (Dembeni, aval)	18.04.2009
TCC468*	*Nitzschia palea*	Mayotte (Dembeni, aval)	18.04.2009
TCC475	*Nitzschia palea*	Mayotte (Coconi, intermédiaire)	18.04.2009
TCC476	*Nitzschia palea*	Mayotte (Coconi, intermédiaire)	18.04.2009
TCC480*	*Nitzschia palea*	Mayotte (Coconi, aval)	18.04.2009
TCC486	*Nitzschia palea*	Mayotte (Coconi, aval)	18.04.2009
TCC491	*Nitzschia palea*	Mayotte (Coconi, aval)	18.04.2009
TCC493	*Nitzschia palea*	Mayotte (Coconi, aval)	18.04.2009
TCC523*	*Nitzschia palea*	La Réunion (Saint Denis, amont prise AEP)	28.04.2009
TCC528	*Nitzschia palea*	La Réunion (Saint Denis, amont prise AEP)	28.04.2009
TCC531	*Nitzschia palea*	La Réunion (Bras des étangs)	24.04.2009
TCC563	*Nitzschia palea*	La Réunion (Sainte Suzanne, Amont Bassin Bœuf)	19.05.2009
TCC568*	*Nitzschia palea*	La Réunion (Sainte Suzanne, Cascade Niagara)	19.05.2009
TCC570*	*Nitzschia palea*	La Réunion (Sainte Suzanne, Cascade Niagara)	19.05.2009
TCC573*	*Nitzschia palea*	La Réunion (Sainte Suzanne, Cascade Niagara)	19.05.2009
TCC577	*Nitzschia palea*	France métropolitaine (Le Sânon, Solvay)	18.06.2009
TCC583	*Nitzschia palea*	France métropolitaine (Chiers, Longlaville)	17.06.2009
TCC585	*Nitzschia palea*	France métropolitaine (Moselle, Bainville aux Miroirs)	18.06.2009
TCC588	*Nitzschia palea*	France métropolitaine (Canal de Nantes à Brest, Nort-sur-Erdre)	18.08.2009
TCC594*	*Nitzschia palea*	France métropolitaine (Don, Guéméné-Penfao)	18.08.2009
TCC598	*Nitzschia palea*	France métropolitaine (La Vie, Fenouillet)	17.08.2009
TCC599*	*Nitzschia palea*	France métropolitaine (Isac, Genrouet)	18.08.2009
TCC600*	*Nitzschia palea*	France métropolitaine (Isac, Genrouet)	18.08.2009
TCC601*	*Nitzschia palea*	France métropolitaine (Isac, Genrouet)	18.08.2009
TCC602*	*Nitzschia palea*	France métropolitaine (Isac, Genrouet)	18.08.2009
TCC603	*Nitzschia palea*	France métropolitaine (Isac, Genrouet)	18.08.2009
TCC609*	*Nitzschia palea*	France métropolitaine (Le Lourdon, Lentigny)	19.08.2009
TCC613*	*Nitzschia palea*	France métropolitaine (Le Lourdon, Lentigny)	19.08.2009
TCC614	*Nitzschia palea*	France métropolitaine (Le Lourdon, Lentigny)	19.08.2009
TCC620	*Nitzschia palea*	France métropolitaine (Le Gier, La Valla)	19.08.2009
TCC623	*Nitzschia palea*	France métropolitaine (Le Gier, La Valla)	19.08.2009

TCC627	*Nitzschia palea*	France métropolitaine (Le Gier, Givors)	19.08.2009
TCC636	*Nitzschia palea*	Luxembourg (Eischbaach, Boevange-sur-Attert)	06.09.2006
TCC638	*Nitzschia palea*	Luxembourg (Eischbaach, Boevange-sur-Attert)	06.09.2006
TCC639	*Nitzschia palea*	Luxembourg (Eischbaach, Boevange-sur-Attert)	06.09.2006
TCC641	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC642	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC643	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC646	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC647*	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC648*	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC649	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC650	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC651	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC652	*Nitzschia palea*	Luxembourg (Aalbach, aval Dreiborn)	12.10.2006
TCC703	*Nitzschia palea*	Espagne (Bernesga, Alija de la Ribera)	02.11.2006
TCC708	*Nitzschia palea*	Espagne (Valdavia, Osorno)	01.11.2006
TCC812	*Nitzschia palea*	Mayotte (Dembeni, aval)	18.04.2009

Tableau III.8: *Liste des souches de* Nitzschia palea *dont l'ADN a été fourni par Rosa Trobajo*.

Code	Taxon	Origine du prélèvement : Pays (Rivière ou Lac, site)	Date d'isolement
Japan A1	*Nitzschia palea*	near Skura River, Tokyo, Japan	02.2002
Japan A2	*Nitzschia palea*	near Skura River, Tokyo, Japan	02.2002
Japan B1	*Nitzschia palea*	near Skura River, Tokyo, Japan	02.2002
Japan B2	*Nitzschia palea*	near Skura River, Tokyo, Japan	02.2002
Japan C	*Nitzschia palea*	Sakura River, Tokyo, Japan	06.2003
Japan D	*Nitzschia palea*	Sakura River, Tokyo, Japan	08.2002
Japan E	*Nitzschia palea*	Hiso River, Tokyo, Japan	06.2003
Japan F	*Nitzschia palea*	Okinawa island, Japan	-
Spain A1	*Nitzschia palea*	Cala Castell, Girona, Spain	08.2004
Spain A2	*Nitzschia palea*	Cala Castell, Girona, Spain	08.2004
Spain A3	*Nitzschia palea*	Cala Castell, Girona, Spain	08.2004
Spain A4	*Nitzschia palea*	Cala Castell, Girona, Spain	08.2004
Spain A5	*Nitzschia palea*	Cala Castell, Girona, Spain	08.2004
Spain B	*Nitzschia palea*	Terri stream, Cornella de Terri, Girona, Spain	01.2004
Belgium1	*Nitzschia palea*	Destelbergen, Ghent, Belgium	07.2005
Belgium2	*Nitzschia palea*	Destelbergen, Ghent, Belgium	07.2005
Belgium3	*Nitzschia palea*	Destelbergen, Ghent, Belgium	07.2005
Belgium4	*Nitzschia palea*	Destelbergen, Ghent, Belgium	07.2005
Paraguay	*Nitzschia palea*	Estancia Santa, Asuncion, Presidente Hayes, Paraguay	10.2004
Brazil	*Nitzschia palea*	Amazonas River, Manaus, Brazil	09.2004
Egypt	*Nitzschia palea*	Roda Island, R. Nile, Giza, Egypt	09.2004
India	*Nitzschia palea*	Kochi, Kerala, India	09.2004
New Scot1	*Nitzschia palea*	Dunsapie Loch, Edinburgh, UK	2008
New Scot2	*Nitzschia palea*	Threipmuir reservoir, near Edinburgh, UK	2008
New Spain1	*Nitzschia palea*	River Ter, near Girona, Spain	2008
New Spain2	*Nitzschia palea*	Sant Martí Vell stream, near Girona, Spain	2008
Spain C	*Nitzschia palea*	Letutxe dam, Bilbao, Spain	11.2004
SriLanka1	*Nitzschia palea*	Dambulla Rock Temple, Sri Lanka	10.2004
SriLanka2	*Nitzschia palea*	Dambulla, Sri Lanka	10.2004
UK	*Nitzschia palea*	Burnham Beeches, Buckinghamshire, UK	11.2004

Matériel et méthodes

Nous avons utilisé 22 souches de *N. palea* isolées de rivières tropicales (La Réunion et Mayotte) et 34 souches isolées de rivières du continent européen (Tableau III.7). En complément, nous avons utilisé l'ADN de souches de *N. palea* isolées par Rosa Trobajo (IRTA, Espagne) à partir d'échantillons provenant de rivières de différentes origines (Belgique, Brésil, Inde, Japon, Paraguay, Espagne, Sri Lanka et Royaume-Uni (Tableau III.8). Contrairement au complexe *G. parvulum*, le séquençage direct des ITS des *N. palea* n'a pas abouti en raison de la variabilité intragénomique de ce marqueur. Nous avons donc séquencé l'ADN de quatre autres régions nucléiques: les ADNr 18S et 28S, le gène *rbc*L et le gène *cox*1. Les souches isolées (Tableau III.7) ont été séquencées pour ces 4 marqueurs. Nous avons également séquencé l'ensemble des souches du Tableau III.8 pour l'ADNr 18S, et complété les données précédentes de Rosa Trobajo (Trobajo et al., 2009, 2010) pour les 3 autres marqueurs. Comme pour le complexe *G. parvulum*, les séquences de l'ADNr 18S n'étaient pas suffisamment polymorphes, pour permettre une distinction des différents groupes constituant ce complexe. Nous n'utilisons donc pas ces séquences dans l'analyse qui suit. Nous avons utilisé 131 séquences d'ADNr 28S, 114 séquences de *rbc*L et 60 séquences de *cox*1 du genre *Nitzschia* comprenant respectivement 76, 79, et 47 séquences de *N. palea*. Pour chaque marqueur, les séquences ont été alignées et analysées par « Maximum Likelihood » (ML) et « Neighbour joining » (NJ).

En raison du manque de caractéristiques morphologiques des *N. palea*, nous n'avons pas réalisé d'analyses de morphométrie géométrique. Les frustules de *N. palea* ont une forme générale linéaire à lancéolée avec des pôles légèrement capités à arrondis. Les stries sont difficilement discernables en microscopie optique (Krammer & Lange-Bertalot, 1988) et les fibules sont irrégulièrement espacées. De précédentes analyses morphométriques (longueur, largeur, densités de stries et de fibules) en microscopie électronique à balayage, ont révélé des différences morphologiques au sein du complexe *N. palea*, mais pas de séparation nette entre les différentes morphologies (Trobajo et al., 2009).

Des tests de croisements au sein du complexe *N. palea* ont été effectués par Rosa Trobajo sur quelques souches (souches indiquées par des étoiles dans le Tableau III.7). Ces souches ont fait, à deux reprises, l'objet de tests de croisement deux à deux.

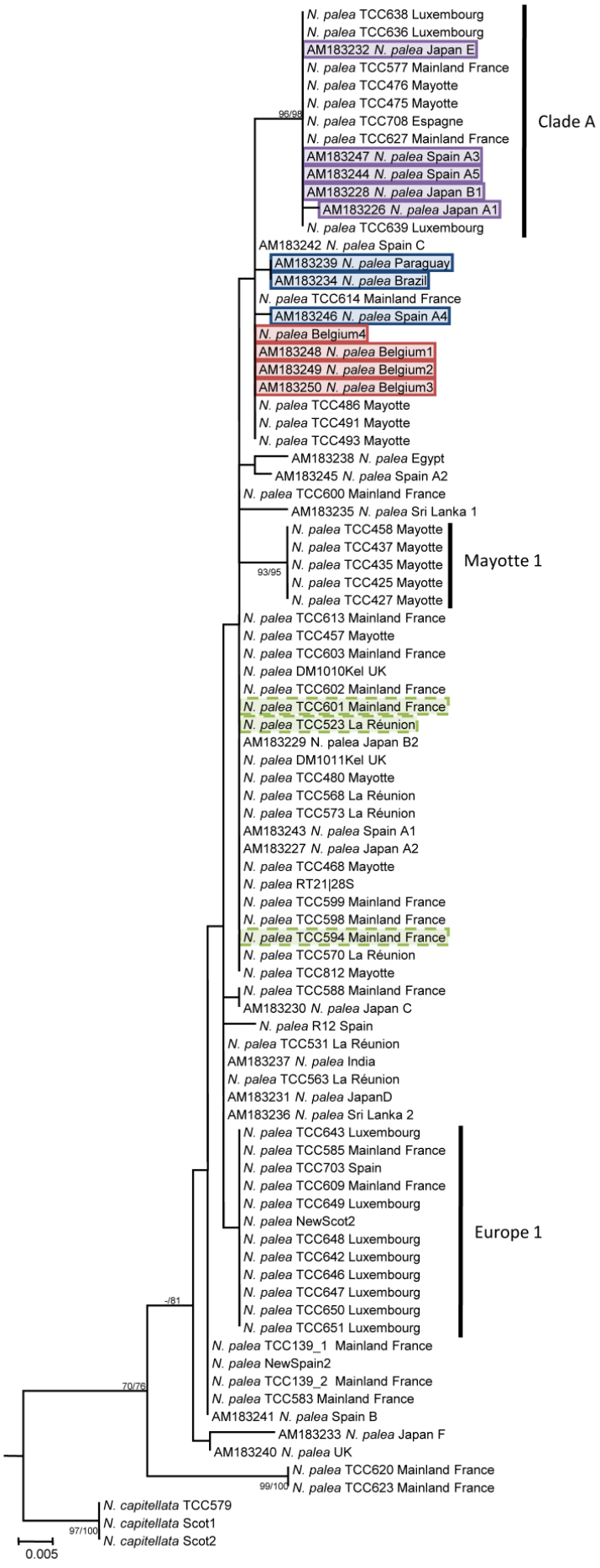

N. palea TCC638 Luxembourg
N. palea TCC636 Luxembourg
AM183232 N. palea Japan E
N. palea TCC577 Mainland France
N. palea TCC476 Mayotte
N. palea TCC475 Mayotte
96/98
N. palea TCC708 Espagne
N. palea TCC627 Mainland France
AM183247 N. palea Spain A3
AM183244 N. palea Spain A5
AM183228 N. palea Japan B1
AM183226 N. palea Japan A1
N. palea TCC639 Luxembourg
Clade A
AM183242 N. palea Spain C
AM183239 N. palea Paraguay
AM183234 N. palea Brazil
N. palea TCC614 Mainland France
AM183246 N. palea Spain A4
N. palea Belgium4
AM183248 N. palea Belgium1
AM183249 N. palea Belgium2
AM183250 N. palea Belgium3
N. palea TCC486 Mayotte
N. palea TCC491 Mayotte
N. palea TCC493 Mayotte
AM183238 N. palea Egypt
AM183245 N. palea Spain A2
N. palea TCC600 Mainland France
AM183235 N. palea Sri Lanka 1
N. palea TCC458 Mayotte
N. palea TCC437 Mayotte
93/95
N. palea TCC435 Mayotte
N. palea TCC425 Mayotte
N. palea TCC427 Mayotte
Mayotte 1
N. palea TCC613 Mainland France
N. palea TCC457 Mayotte
N. palea TCC603 Mainland France
N. palea DM1010Kel UK
N. palea TCC602 Mainland France
N. palea TCC601 Mainland France
N. palea TCC523 La Réunion
AM183229 N. palea Japan B2
N. palea DM1011Kel UK
N. palea TCC480 Mayotte
N. palea TCC568 La Réunion
N. palea TCC573 La Réunion
AM183243 N. palea Spain A1
AM183227 N. palea Japan A2
N. palea TCC468 Mayotte
N. palea RT21|28S
N. palea TCC599 Mainland France
N. palea TCC598 Mainland France
N. palea TCC594 Mainland France
N. palea TCC570 La Réunion
N. palea TCC812 Mayotte
N. palea TCC588 Mainland France
AM183230 N. palea Japan C
N. palea R12 Spain
N. palea TCC531 La Réunion
AM183237 N. palea India
N. palea TCC563 La Réunion
AM183231 N. palea JapanD
AM183236 N. palea Sri Lanka 2
N. palea TCC643 Luxembourg
N. palea TCC585 Mainland France
N. palea TCC703 Spain
N. palea TCC609 Mainland France
N. palea TCC649 Luxembourg
N. palea NewScot2
N. palea TCC648 Luxembourg
N. palea TCC642 Luxembourg
N. palea TCC646 Luxembourg
N. palea TCC647 Luxembourg
N. palea TCC650 Luxembourg
N. palea TCC651 Luxembourg
Europe 1
-/81
N. palea TCC139_1 Mainland France
N. palea NewSpain2
70/76
N. palea TCC139_2 Mainland France
N. palea TCC583 Mainland France
AM183241 N. palea Spain B
AM183233 N. palea Japan F
AM183240 N. palea UK
N. palea TCC620 Mainland France
99/100
N. palea TCC623 Mainland France
N. capitellata TCC579
97/100
N. capitellata Scot1
N. capitellata Scot2
0.005

Figure III.14: *Arbre phylogénétique obtenu par analyse des séquences d'ADNr 28S par Maximum Likelihood*.
Les valeurs de bootstrap supérieures à 70, obtenues par les différentes méthodes d'analyses (ML/NJ) sont indiquées sur les branches. Pour rendre l'arbre plus lisible, seules les branches correspondant aux séquences de *N. capitellata* et *N. palea* sont montrées. Les séquences de R. Trobajo correspondant aux 3 groupes, clade A, mating group 1 et mating group 2, sont colorées respectivement en violet, rouge et bleu. Les nouvelles souches ayant montré un appariement sexuel sont colorées en vert.

Résultats et discussion

Les analyses phylogénétiques par ML et NJ réalisées pour chacun des 3 marqueurs nucléiques sur des séquences de différentes espèces du genre *Nitzschia* ont révélé que toutes les souches appartenant à l'espèce morphologique *N. palea* formaient un groupe monophylétique robuste. De plus, nous avons démontré que parmi les séquences du genre *Nitzschia*, l'espèce *N. capitellata* était l'espèce la plus proche génétiquement des *N. palea* (Figure III.14 et Figure III.15). Au sein du groupe des *N. palea*, nous avons observé une grande diversité génétique et de nombreux clades sont apparus quel que soit le marqueur utilisé (Figure III.14, Figure III.15 et Figure III.16). Quelques clades sont supportés par de fortes valeurs de « bootstrap ». Par exemple, le clade A, défini par les séquences d'ADNr 28S de Trobajo et al. (2009) et complété par de nouvelles séquences, présentait de fortes valeurs de bootstrap dans les arbres obtenus par les séquences d'ADNr 28S et de *cox*1 (Figure III.14 et Figure III.16). Cependant, ce clade n'a pu être identifié par l'analyse des séquences de *rbc*L (Figure III.15).

Par ailleurs, les deux groupes définis sur la base d'expériences de croisement (« mating group 1 » et « mating group 2 », Trobajo et al. 2009) étaient supportés par de fortes valeurs de bootstrap dans l'analyse des séquences de *rbc*L (BS 94/96 et 92/98). Nos résultats étaient ainsi concordants avec ceux obtenus par Trobajo et al. (2009, 2010).

En ajoutant des séquences aux données de Trobajo et al. (2009, 2010), nous pensions compléter les groupes définis auparavant et préciser la définition de ceux composés d'une ou deux souches. Certaines de nos souches ont, en effet, complété certaines branches des arbres phylogénétiques obtenus dans les deux études précédentes, par exemple le clade Europe 2. Cependant, l'ajout de nos souches a également entraîné l'émergence de nouveaux clades (Mayotte 1 et Europe 1) et de nouvelles branches composées d'une ou deux souches. L'apport de nouvelles données moléculaires a ainsi révélé une diversité supplémentaire au sein du complexe *N. palea*.

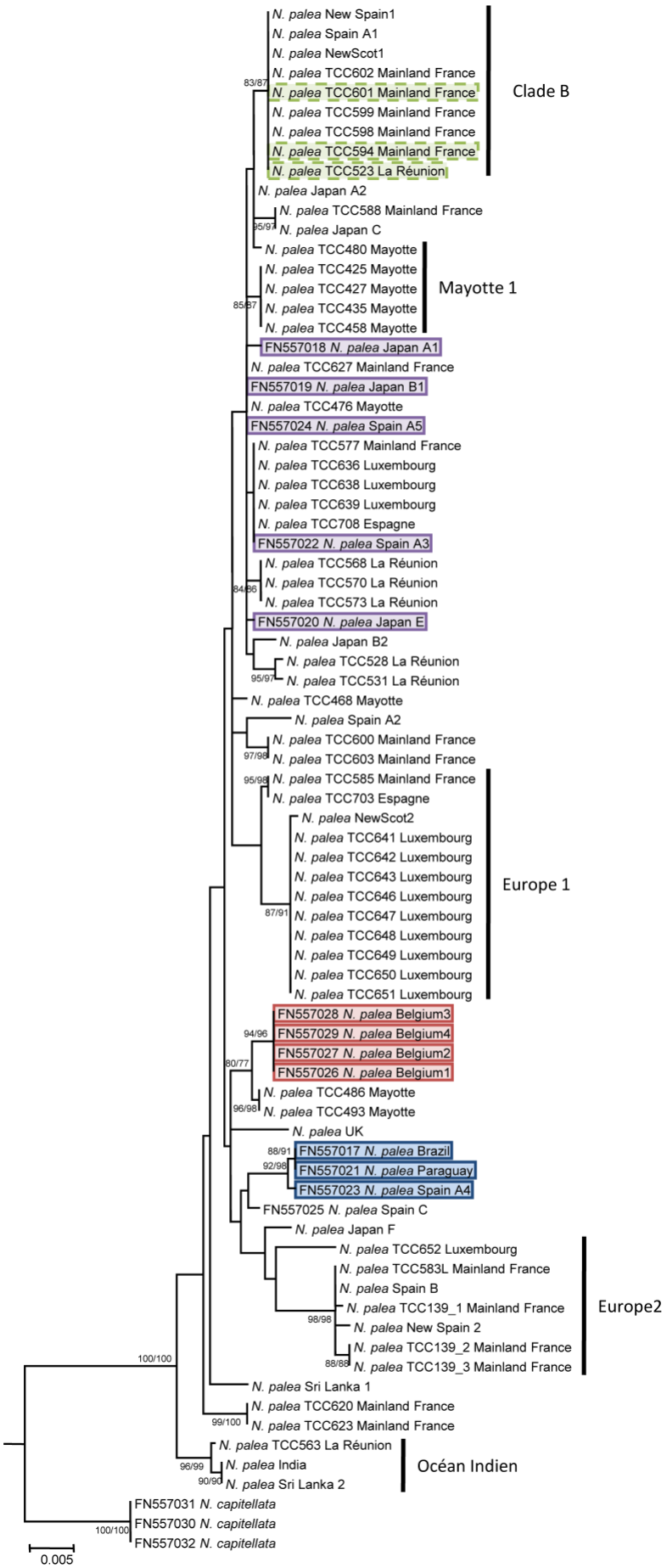

N. palea New Spain1
N. palea Spain A1
N. palea NewScot1
N. palea TCC602 Mainland France
N. palea TCC601 Mainland France
N. palea TCC599 Mainland France
N. palea TCC598 Mainland France
N. palea TCC594 Mainland France
N. palea TCC523 La Réunion
Clade B
N. palea Japan A2
N. palea TCC588 Mainland France
N. palea Japan C
N. palea TCC480 Mayotte
N. palea TCC425 Mayotte
N. palea TCC427 Mayotte
N. palea TCC435 Mayotte
N. palea TCC458 Mayotte
Mayotte 1
FN557018 N. palea Japan A1
N. palea TCC627 Mainland France
FN557019 N. palea Japan B1
N. palea TCC476 Mayotte
FN557024 N. palea Spain A5
N. palea TCC577 Mainland France
N. palea TCC636 Luxembourg
N. palea TCC638 Luxembourg
N. palea TCC639 Luxembourg
N. palea TCC708 Espagne
FN557022 N. palea Spain A3
N. palea TCC568 La Réunion
N. palea TCC570 La Réunion
N. palea TCC573 La Réunion
FN557020 N. palea Japan E
N. palea Japan B2
N. palea TCC528 La Réunion
N. palea TCC531 La Réunion
N. palea TCC468 Mayotte
N. palea Spain A2
N. palea TCC600 Mainland France
N. palea TCC603 Mainland France
N. palea TCC585 Mainland France
N. palea TCC703 Espagne
N. palea NewScot2
N. palea TCC641 Luxembourg
N. palea TCC642 Luxembourg
N. palea TCC643 Luxembourg
N. palea TCC646 Luxembourg
N. palea TCC647 Luxembourg
N. palea TCC648 Luxembourg
N. palea TCC649 Luxembourg
N. palea TCC650 Luxembourg
N. palea TCC651 Luxembourg
Europe 1
FN557028 N. palea Belgium3
FN557029 N. palea Belgium4
FN557027 N. palea Belgium2
FN557026 N. palea Belgium1
N. palea TCC486 Mayotte
N. palea TCC493 Mayotte
N. palea UK
FN557017 N. palea Brazil
FN557021 N. palea Paraguay
FN557023 N. palea Spain A4
FN557025 N. palea Spain C
N. palea Japan F
N. palea TCC652 Luxembourg
N. palea TCC583L Mainland France
N. palea Spain B
N. palea TCC139_1 Mainland France
N. palea New Spain 2
N. palea TCC139_2 Mainland France
N. palea TCC139_3 Mainland France
Europe2
N. palea Sri Lanka 1
N. palea TCC620 Mainland France
N. palea TCC623 Mainland France
N. palea TCC563 La Réunion
N. palea India
N. palea Sri Lanka 2
Océan Indien
FN557031 N. capitellata
FN557030 N. capitellata
FN557032 N. capitellata
0.005

Figure III.15: *Arbre phylogénétique obtenu par analyse des séquences de* rbc*L par Maximum Likelihood*.
Les valeurs de bootstrap supérieures à 70, obtenues par les différentes méthodes d'analyses (ML/NJ) sont indiquées sur les branches. Pour rendre l'arbre plus lisible, seules les branches correspondant aux séquences de *N. capitellata* et *N. palea* sont montrées. Les séquences de R. Trobajo correspondant aux 3 groupes (clade A, mating group 1 et mating group 2) sont colorées respectivement en violet, rouge et bleu. Les nouvelles souches ayant montré un appariement sexuel sont colorées en vert.

Des tests de croisements effectués par Rosa Trobajo, seuls les couples de souches TCC601 et TCC523, et TCC601 et TCC594 ont montré des appariements et des échanges de gamètes. Cependant, aucun développement d'auxospore n'a été observé. Ces trois souches présentaient 100% d'identité pour leurs séquences d'ADNr 28S et de *rbc*L et ne sont différenciables que par leurs séquences de *cox*1 qui présentent des p-distances de 2.6% entre TCC601 et TCC523 et de 2.4% entre TCC601 et TCC594. Ces résultats indiquent que les marqueurs *rbc*L et 28S ne sont pas suffisamment polymorphes pour discriminer tous les groupes composant le complexe d'espèces *N. palea*. Au contraire, le marqueur *cox*1 semble présenter une vitesse d'évolution suffisante pour analyser la diversité de ce complexe. Les différences génétiques entre les séquences de *cox*1 des trois souches TCC523, TCC594 et TCC601 sont plus importantes que les différences génétiques entre les clones Paraguay et Brazil qui sont interféconds (1%).

Hélas, le faible succès du séquençage du gène *cox*1 n'a pas permis d'obtenir autant de séquences de *cox*1 que de séquences d'ADNr 28S ou de *rbc*L, ce qui limite les comparaisons entre les différents groupes.

Trobajo et al. (2009) concluaient, à partir du manque d'interaction sexuelle entre les groupes reproductifs 1 et 2, que des phénomènes de spéciation s'étaient produits de longue date dans ce complexe d'espèces. L'appariement entre ces trois souches sans formation d'auxospore pourrait révéler un processus de spéciation plus récent au sein du clade B que celui qui s'est produit entre les groupes 1 et 2. La comparaison des variabilités génétiques des séquences de *cox*1 au sein du clade A (0 à 4,7%) avec celle du clade B, laissent supposer que le clade A pourrait lui aussi contenir plusieurs groupes, d'autant plus qu'aucune reproduction n'avait été observée par Trobajo et al. (2009), et que ce groupe n'est pas supporté dans l'analyse des séquences de *rbc*L (Figure III.15).

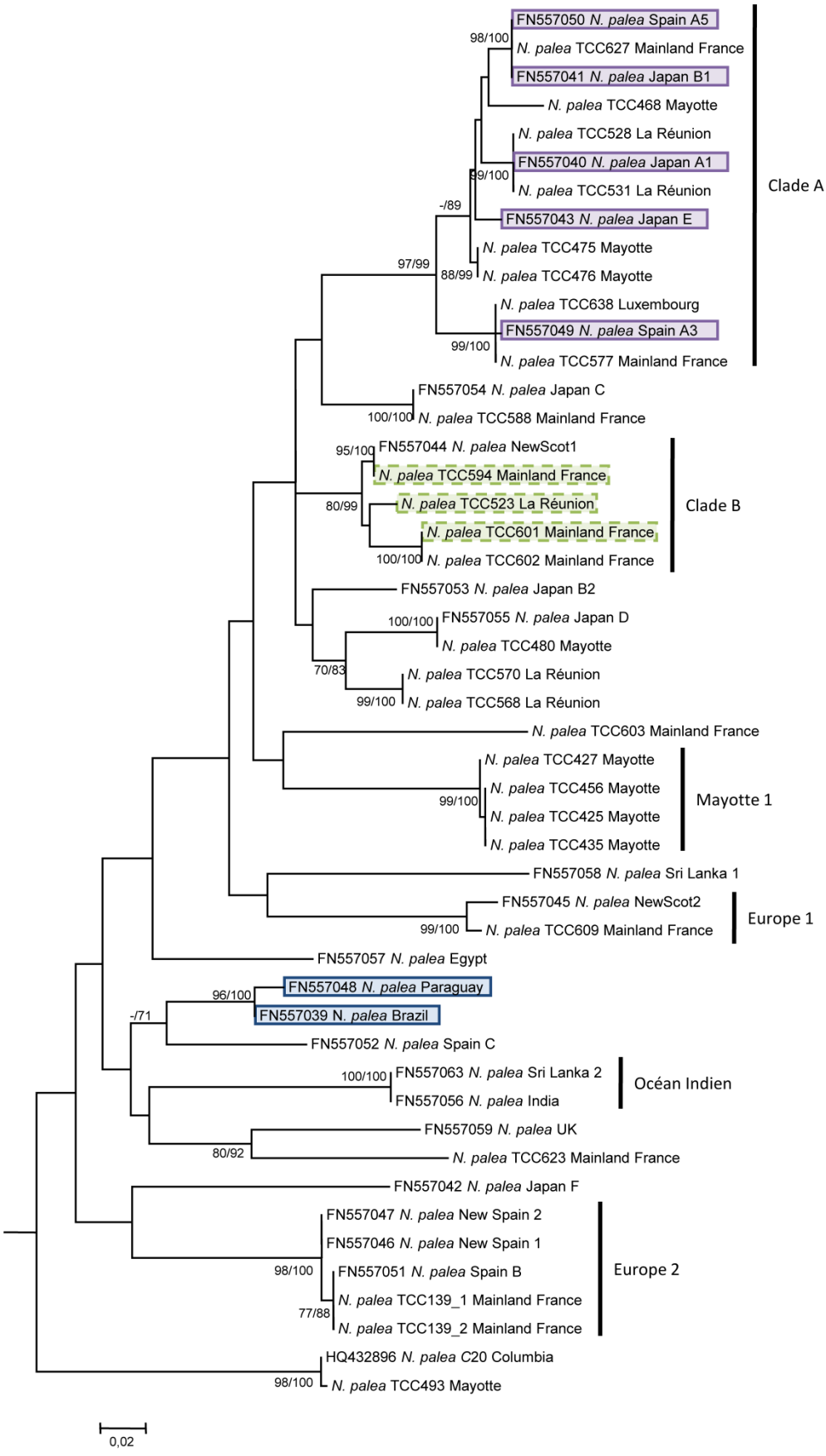

FN557050 N. palea Spain A5
98/100
N. palea TCC627 Mainland France
FN557041 N. palea Japan B1
N. palea TCC468 Mayotte
N. palea TCC528 La Réunion
FN557040 N. palea Japan A1
99/100
N. palea TCC531 La Réunion
-/89
FN557043 N. palea Japan E
N. palea TCC475 Mayotte
97/99
88/99
N. palea TCC476 Mayotte
N. palea TCC638 Luxembourg
FN557049 N. palea Spain A3
99/100
N. palea TCC577 Mainland France
Clade A
FN557054 N. palea Japan C
100/100
N. palea TCC588 Mainland France
95/100
FN557044 N. palea NewScot1
N. palea TCC594 Mainland France
80/99
N. palea TCC523 La Réunion
N. palea TCC601 Mainland France
100/100
N. palea TCC602 Mainland France
Clade B
FN557053 N. palea Japan B2
100/100
FN557055 N. palea Japan D
N. palea TCC480 Mayotte
70/83
N. palea TCC570 La Réunion
99/100
N. palea TCC568 La Réunion
N. palea TCC603 Mainland France
N. palea TCC427 Mayotte
99/100
N. palea TCC456 Mayotte
N. palea TCC425 Mayotte
N. palea TCC435 Mayotte
Mayotte 1
FN557058 N. palea Sri Lanka 1
FN557045 N. palea NewScot2
99/100
N. palea TCC609 Mainland France
Europe 1
FN557057 N. palea Egypt
96/100
FN557048 N. palea Paraguay
-/71
FN557039 N. palea Brazil
FN557052 N. palea Spain C
100/100
FN557063 N. palea Sri Lanka 2
FN557056 N. palea India
Océan Indien
FN557059 N. palea UK
80/92
N. palea TCC623 Mainland France
FN557042 N. palea Japan F
FN557047 N. palea New Spain 2
FN557046 N. palea New Spain 1
98/100
FN557051 N. palea Spain B
77/88
N. palea TCC139_1 Mainland France
N. palea TCC139_2 Mainland France
Europe 2
HQ432896 N. palea C20 Columbia
98/100
N. palea TCC493 Mayotte
0,02

Figure III.16: *Arbre phylogénétique obtenu par analyse des séquences de* cox*1 par Maximum Likelihood*.
Les valeurs de bootstrap supérieures à 70, obtenues par les différentes méthodes d'analyses (ML/NJ) sont indiquées sur les branches. Pour rendre l'arbre plus lisible, seules les branches correspondant aux séquences de *N. palea* sont montrées. Les séquences de R. Trobajo correspondant aux 2 groupes (clade A et mating group 2) sont colorées respectivement en violet et bleu. Les nouvelles souches ayant montré un appariement sexuel sont colorées en vert.

Les données moléculaires n'ont pas indiqué de relation nette entre la variabilité génétique et l'origine géographique des souches. Certains clades correspondaient à une origine géographique proche, comme par exemple le clade Océan Indien observé dans les analyses du *rbc*L et du *cox*1 (Inde, Sri Lanka et La Réunion), ou les clades européens ou encore mahorais observés dans les trois arbres. Mais plusieurs exemples inverses sont également observés : par exemple, les clades A et B paraissaient ubiquistes étant composés de souches provenant d'origines géographiques différentes. D'autre part, comme pour le complexe *G. parvulum*, des souches isolées à partir du même échantillon benthique (les souches provenant de la rivière Isac en France métropolitaine) étaient réparties dans différentes branches. L'isolement géographique ne pouvant expliquer en totalité la séparation des clades, d'autres facteurs tels que l'écologie ou la géologie des sites devraient être étudiés pour expliquer leur répartition.

Conclusion

Au sein du complexe *N. palea*, ni l'étude de la morphologie ou de la reproduction, ni l'utilisation des séquences de trois marqueurs nucléiques n'ont permis de définir des groupes robustes et cohérents. Ce travail a ainsi montré les limites de l'utilisation des séquences ADN traitées dans cette étude. Les relations entre les différents clades étant faiblement supportées et certaines branches n'incluant qu'une ou deux souches, nous n'avons pas été en mesure de mieux définir les groupes composant le complexe *N. palea*. Cependant, notre étude a révélé une importante diversité génétique au sein du complexe *N. palea.* Puisque cette diversité génétique est accompagnée dans certains cas d'une perte des capacités de reproduction sexuée croisée, des analyses moléculaires et des croisements de nouvelles souches seraient nécessaires pour définir les différents taxa qui composent ce complexe d'espèces. Bien que le séquençage des ITS nécessite une étape de clonage dans le cas des *N. palea* (compte-tenu de leur variabilité intragénomique), l'analyse des CBC des ITS2 pourrait également apporter des informations complémentaires sur ce complexe d'espèces.

4. Conclusion

Les travaux détaillés dans ces trois études illustrent la manière dont les séquences ADN peuvent améliorer notre compréhension des relations phylogénétiques entre les diatomées. Ces résultats montrent également qu'il est parfois difficile d'identifier les correspondances entre les critères morphologiques choisis pour la taxonomie des diatomées et la phylogénie moléculaire. En effet, les caractères morphologiques identiques les plus évidents tels que l'asymétrie des valves de *Didymosphenia* et *Gomphonema* ou la forme des pôles de *Gomphonema parvulum*, ne reflètent pas leurs relations phylogénétiques. Des caractères morphologiques communs peuvent être acquis par différentes lignées (convergence morphologique) et des caractères différents peuvent correspondre à des variabilités morphologiques intraspécifiques. En outre, les résultats concernant le complexe *Gomphonema parvulum* démontrent que la délimitation des taxa est différente en fonction de l'utilisation d'outils moléculaires ou d'outils morphologiques. Par exemple, les outils moléculaires distinguent les clades vivant en sympatrie alors que la morphologie ne permet pas cette séparation. Au contraire, les observations en microscopie distinguent les formes *lagenula* et *exilissimum* des formes *parvulum* alors que les séquences ADN ne le permettent pas. L'approche moléculaire permet donc une discrimination des taxa de diatomées, mais celle-ci est différente de la discrimination morphologique. De plus, nous avons observé, sur le complexe d'espèces *Nitzschia palea*, que les données moléculaires, comme les données morphologiques, ne sont pas toujours suffisantes pour étudier les relations entre des lignées en cours de spéciation. Une approche polyphasique, combinant à la fois des données morphologiques et des données moléculaires, mais également des données de reproduction et des données cytologiques en lien avec les données environnementales, paraît donc utile pour améliorer et stabiliser la taxonomie des diatomées.

D'autres part, ces résultats confirment que les séquences ADN sont un outil utilisable pour différencier la plupart des taxa de diatomées. Différentes régions nucléiques permettent une discrimination des diatomées à différents niveaux taxonomiques : l'ADNr 18S pour étudier les familles et les genres de diatomées, l'ADNr 28S pour les genres et les espèces, et les marqueurs ITS, *cox*1 et *rbc*L pour des niveaux spécifiques voire certains niveaux intraspécifiques. Le choix du marqueur est donc une étape déterminante lorsqu'il s'agit d'étudier les taxa de diatomées avec des outils moléculaires.

CHAPITRE IV. NOUVELLE APPROCHE DE BIOINDICATION PAR LES DIATOMEES

1. Introduction

Comme nous l'avons illustré dans le chapitre précédent, les caractéristiques morphologiques de certains taxa sont difficilement discernables en microscopie optique. Dans le contexte de la bioindication, des analyses en microscopie optique sont appliquées pour évaluer la composition des communautés de diatomées. Pour améliorer les identifications des taxa, d'autres outils tels que la morphométrie géométrique, la microscopie électronique ou les croisements, sont efficaces mais ne sont pas adaptés à une utilisation en routine. Au contraire, les outils de biologie moléculaire sont adaptés pour les analyses de routine, et apportent des informations complémentaires à la détermination morphologique des espèces de diatomées.

Les données moléculaires sont, depuis de nombreuses années, utilisées pour étudier la diversité ainsi que la structure et la composition des communautés microbiennes. Différentes méthodes sont ainsi utilisées en fonction des informations recherchées (Dorigo et al., 2005). Certaines méthodes produisent des empreintes moléculaires des communautés (van Hannen et al., 1998; Casamayor et al., 2000; Díez et al., 2001a) permettant de comparer différentes communautés mais sans identifier les taxa qui les composent. La méthode du « clonage-séquençage » (Díez et al., 2001b; Romari & Vaulot, 2004) permet également d'étudier les communautés de microorganismes en identifiant les séquences ou en les regroupant en OTUs. Néanmoins, pour bien représenter la diversité d'une communauté il est nécessaire d'obtenir suffisamment de séquences en passant par le clonage, ce qui rend cette méthode coûteuse et laborieuse. Les puces à ADN, basées sur le principe d'hybridation de séquences ADN complémentaires, ont aussi démontré leur potentiel pour identifier les communautés phytoplanctoniques d'échantillons naturels (e.g. .Medlin et al., 2006), mais ces méthodes requièrent une sélection des taxa à cibler. Des méthodes de PCR quantitative ont également été développées pour quantifier les diatomées dans des échantillons environnementaux. Cette technique permet soit de quantifier quelques taxa ciblés (Créach et al., 2006; Andree et al., 2011), soit d'évaluer la proportion des diatomées (Godhe et al., 2008). Cependant, il est impossible d'identifier et quantifier chacun des taxa de diatomées présents dans un échantillon naturel en utilisant la PCR quantitative.

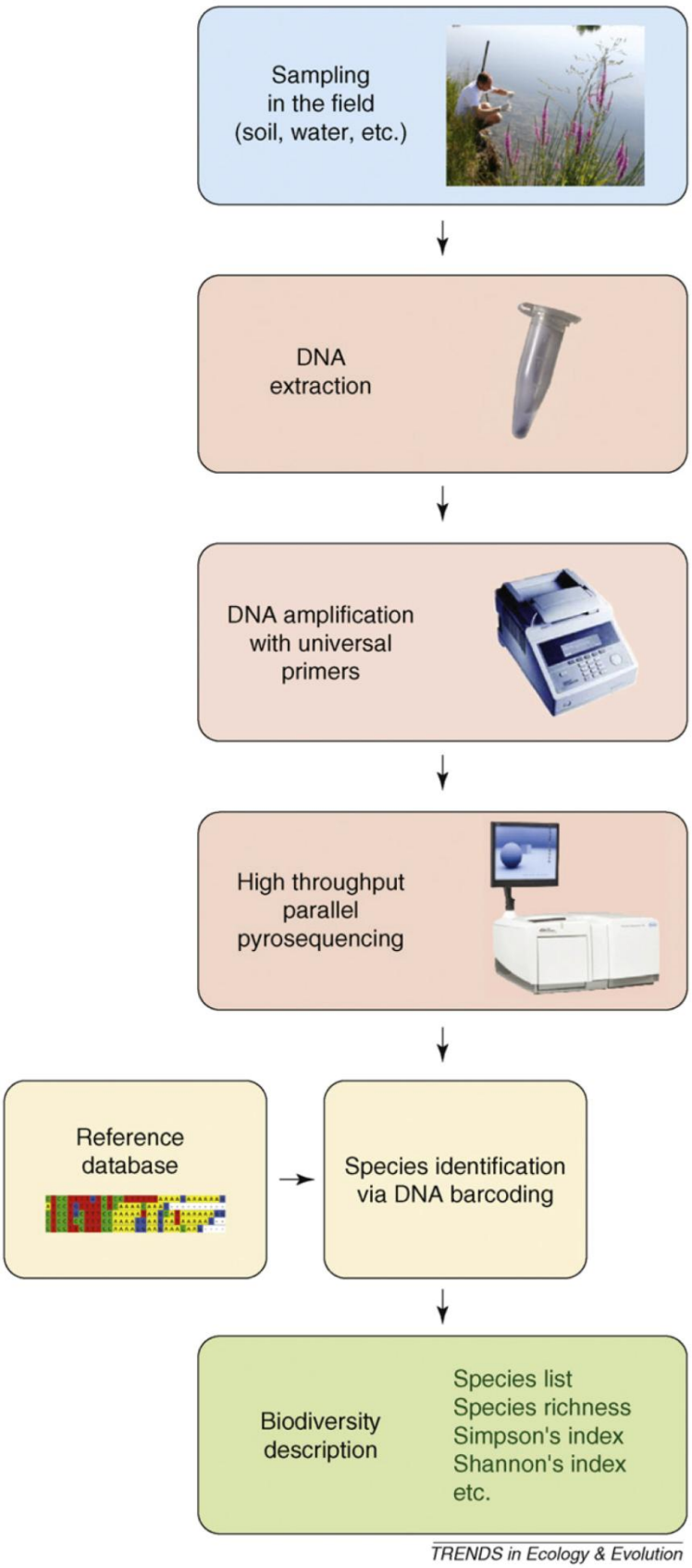


Figure IV.1: *Procédure d'analyse d'échantillons environnementaux utilisant les nouvelles techniques de séquençage*.
Source : Valentini et al., 2009.

Comme nous l'avons précédemment introduit dans le Chapitre I, l'émergence de nouvelles technologies de séquençage à haut débit telles que les NGS, révolutionne actuellement l'évaluation de la biodiversité. Ces techniques couplées aux barcodes ADN offrent des perspectives nouvelles pour la bioindication. Valentini et al. (2009) ont proposé une méthodologie pour analyser la biodiversité présente dans un échantillon environnemental (Figure IV.1) en analysant par séquençage haut-débit un marqueur nucléique, et en comparant les séquences obtenues à des bases de séquences de référence. Cette approche a déjà été utilisée sur les macro-invertébrés et a indiqué des résultats prometteurs pour les applications de bioindication (Hajibabaei et al., 2011). L'utilisation du séquençage massif sur des échantillons de l'environnement pourrait donc également permettre d'améliorer en termes d'efficacité et de précision les identifications de diatomées dans les programmes de bioindication. Nous avons donc choisi de développer et tester un outil moléculaire, basé sur les NGS. Pour cela, nous devons optimiser les différentes étapes du processus qui conduisent d'un site naturel à l'évaluation de la diversité en diatomées sur ce site (Figure IV.1) : échantillonnage, méthodes moléculaires (extraction, amplification, séquençage) et identification taxonomique des séquences.

Dans le cas de l'utilisation des diatomées en tant que bioindicateur de la qualité des eaux, l'échantillonnage est normalisé. Nous n'avons donc pas testé cette étape et nous nous sommes focalisés sur l'optimisation des méthodes moléculaires et des méthodes bioinformatiques nécessaires au développement d'un outil moléculaire.

Tout d'abord, notre choix de séquençage massif s'est porté sur le pyroséquençage 454. Nous avons choisi cette technologie, car elle est la technique NGS qui permet d'obtenir les plus longues séquences (reads), donc les plus informatives pour différencier des taxa. Pour évaluer la méthode et les biais associés, nous avons utilisé des échantillons artificiels dont la composition était totalement maîtrisée (CHAPITRE II.3.2.2(a)), puis nous avons analysé des échantillons naturels (CHAPITRE IV.4).

En raison du grand nombre de reads obtenus grâce au pyroséquençage 454, un outil bioinformatique efficace, adapté à la comparaison de ces nombreux reads à des bases de référence, est indispensable. Nous avons donc commencé par comparer les algorithmes développés par A. Franc et P. Chaumeil à d'autres algorithmes permettant de comparer et d'assigner les séquences (Article III). Ensuite, nous nous sommes focalisés sur le choix du marqueur à utiliser. En effet, nous avions démontré dans le chapitre précédent que les différents marqueurs moléculaires pouvaient décrire les taxa de diatomées à des degrés de

précision différents. Nous avons donc comparé trois marqueurs nucléiques : l'ADNr 18S, le *rbc*L et le *cox*1 en prenant en compte leurs bases de séquences de référence, leurs pouvoirs de résolution ainsi que les biais méthodologiques liés à leur utilisation (Article IV). Enfin, nous avons testé notre méthode d'analyse et nos marqueurs moléculaires sur des échantillons naturels (Article V).

2. *Comparaison d'algorithmes pour l'assignation de reads issus de NGS*

2.1. *Présentation générale de l'étude et synthèse des principaux résultats.*

Le séquençage par des NGS implique l'utilisation d'outils bioinformatiques pour traiter la quantité considérable de données produites. Disposer d'outils bioinformatiques adaptés est donc crucial pour développer une approche de bioindication basée sur les NGS. Il était nécessaire de disposer d'un outil, permettant de comparer de très nombreux reads issus du pyroséquençage d'un échantillon, à des séquences de référence de taxonomie connue, pour assigner un nom aux reads, afin d'établir l'inventaire taxonomique de l'échantillon. De nombreuses méthodes bioinformatiques existent, que ce soit pour comparer des séquences à des référence (BLAST, Altschul et al., 1990 ; UBLAST, Edgar, 2010 ; BLAT, Kent, 2002), pour grouper les séquences en OTUs (DOTUR, Schloss & Handelsman, 2005 ; UCLUST, Edgar, 2010) ou pour positionner les séquences dans un arbre phylogénétique (STAP, Wu et al., 2008 ; PPLACER, Matsen et al., 2010). De plus, de nombreux « pipelines » ont été développés à partir de ces outils pour analyser les données issues des échantillons environnementaux, par exemple MOTHUR (Schloss et al., 2009) ou PANGEA (Giongo et al., 2010). Cependant, beaucoup de ces outils sont spécifiquement adaptés à un type de séquences ou à un type de technologie. Par exemple, le pipeline PANAM (Taib et al.) est dédié à l'analyse de séquences d'ADNr SSU et le programme d'alignement SOAP (Li et al., 2008) est spécifiquement adapté aux courts reads produits par la technologie de séquençage Illumina-Solexa.

L'objectif de ce travail était donc de tester un nouvel ensemble d'algorithmes, adapté à cette étude, *metaMatch,* développé par A. Franc et P. Chaumeil. Ces outils permettent de comparer, sans heuristique, des reads issus de n'importe quelle technologie NGS, à une base de référence quel que soit le marqueur. Nous avons comparé cet outil à deux autres outils utilisant des heuristiques pour un alignement local : BLAST (Altschul et al., 1990) qui est couramment utilisé pour comparer et assigner les séquences (Buée et al., 2009; Stoeck et al., 2009; Hajibabaei et al., 2011), et UBLAST (Edgar, 2010) récemment développé pour accélérer la comparaison des séquences.

Pour cela nous avons analysé nos trois échantillons artificiels (le mélange de cultures et les deux mélanges PCR, (CHAPITRE II.3.2.2(a)) en séquençant le marqueur 18S, qui est le marqueur le plus représenté dans les bases de séquences, et qui permet ainsi d'avoir une base de référence bien renseignée. Pour éviter les biais liés à la taxonomie des diatomées, nous avons testé la version de *metaMatch*_PS qui utilise uniquement les « matchs parfaits » c'est-à-dire les séquences qui sont 100% identiques à au moins une séquence de la base de référence. Nous avons donc également paramétré BLAST et UBLAST pour une recherche de reads présentant 100% d'identité avec des séquences de la base de référence. De plus, ces deux algorithmes recherchent des fragments de séquences (des « mots ») qui sont communs entre les séquences inconnues et les séquences de référence. Ces « mots » servent de point d'ancrage pour aligner les deux séquences. Nous avons donc testé deux versions de ce paramètre : une recherche d'identité avec une taille de « mot » de 11 pb (paramètre par défaut pour BLAST) et une recherche avec une taille de 130 bp car notre plus petit read avait une taille de 131 pb. Enfin, comme *metaMatch* utilise la totalité du read pour la comparaison, nous avons rajouté cette indication grâce à une option disponible dans UBLAST (mais non disponible dans BLAST). Les trois algorithmes ont été utilisés avec la même base de référence de séquence 18S (1412 séquences) élaborée à partir des séquences de nos souches et de séquences de GenBank curées.

Après la comparaison des reads aux séquences de référence, les reads sans assignation exacte ont été éliminés. En effet, il est possible qu'un read corresponde à plusieurs séquences de la base de référence. Si les correspondances se font toujours sur le même taxon, ce read est considéré comme « informatif ». Par contre, si ce read correspond à plusieurs taxa, nous considérons qu'il n'est pas informatif. Les inventaires sont réalisés à partir de l'identification des seuls reads informatifs, afin d'obtenir les inventaires de diatomées en présence / absence, les plus précis possibles.

Les résultats révèlent que UBLAST est l'outil le plus rapide. La comparaison des librairies de reads à notre base de référence de 1412 séquences prend moins de dix minutes pour 100 000 reads. BLAST et *metaMatch*_PS s'exécutent plus lentement mais avec des temps similaires pour les deux algorithmes, selon les paramètres choisis pour BLAST (longueur du « mot » pour les points d'ancrage).

Nous avons également comparé la proportion de reads informatifs obtenus par les trois outils. UBLAST obtient la plus forte proportion (100%) de reads informatifs alors

que *metaMatch* et BLAST présentent environ 50% de reads informatifs. Cette différence provient de la conception des algorithmes. BLAST et *metaMatch* sont élaborés pour trouver toutes les correspondances possibles entre les reads et les séquences de référence. Au contraire, pour améliorer la rapidité de l'outil, UBLAST est conçu pour trouver une correspondance forte entre un read et une base de référence. Ainsi, UBLAST termine la recherche de correspondance lorsqu'une forte correspondance est trouvée, sans continuer sa recherche sur la suite de la base. Les informations nécessaires au tri des reads, en reads informatifs et reads non informatifs ne sont donc pas disponibles. Il est à noter que ce paramètre peut être modifié mais UBLAST emploie alors un algorithme très proche de BLAST.

Enfin, nous avons comparé les inventaires des trois échantillons obtenus par les trois algorithmes. Une méthode de classification ascendante hiérarchique (HAC) utilisant les distances de Jaccard, nous a permis de visualiser les similarités entre les inventaires. Nous avons ainsi constaté que les inventaires sont très dépendants de l'algorithme utilisé pour comparer les reads aux références. En effet, les trois échantillons sont toujours regroupés en fonction des options de recherche et des algorithmes utilisés. Les inventaires les plus proches de l'inventaire réel sont les inventaires des trois échantillons obtenus à partir de *metaMatch*, puis plus éloignés, les inventaires fournis par BLAST, et enfin ceux obtenus par UBLAST.

Nous avons donc validé l'exactitude de *metaMatch* pour filtrer les reads informatifs et assigner un nom à un read. Nous avons observé que d'autres outils permettent des analyses plus rapides mais ce gain de temps est au détriment de la précision de l'inventaire obtenu. Dans le cadre de la bioindication, la précision de l'assignement est cruciale d'autant plus que des taxa proches peuvent présenter des préférences écologiques différentes. Pour les étapes suivantes du développement de l'outil moléculaire, afin de privilégier la précision à la rapidité, nous avons utilisé les différentes versions de *metaMatch* (matchs parfaits ou imparfaits, calculs sur processeur simple ou sur cluster).

2.2. *Article III : en préparation pour la revue BMC Bioinformatics*

metaMatch: a tool for metabarcoding

Lenaïg Kermarrec[1,2], Philippe Chaumeil [3,4], Frédéric Rimet [1], Jean-Marc Frigerio [3,4], Agnès Bouchez [1], and Alain Franc * [3,4]

[1] *INRA, UMR CARRTEL, 74200 Thonon-les-Bains,*

[2] *Asconit Consultants, 66350 Toulouges, France,*

[3] *INRA, UMR BioGeCo, 33610 Cestas, France*

[4] *University of Bordeaux 1, UMR BioGeCo, 33400 Talence, France.*

Email: Lenaïg Kermarrec - lenaig.kermarrec@asconit.com; Philippe Chaumeil - philipe.chaumeil@pierroton.inra.fr; Frédéric Rimet - frederic.rimet@inra.thonon.fr; Jean-Marc Frigerio - frigerio@pierroton.inra.fr; Agnès Bouchez - agnes.bouchez@thonon.inra.fr; Alain Franc - alain.franc@pierroton.inra.fr

*Corresponding author: alain.franc@pierroton.inra.fr

Abstract

Background: Microbial ecology has been revolutionized since the advent of metagenomic and Next-Generation Sequencing (NGS). Additionally, barcoding is becoming an essential tool for microbes identification. Nevertheless there is an urging need for metabarcoding algorithms. Algorithms classifying sequences by comparison to a reference library, are the most widely used tools for assessing community composition of environmental samples. However, almost all these algorithms use heuristics designed to speed up the database exploration phase, at the cost of being less strict with the quality of the match between a query and a reference. Nevertheless accurate identification of the species in a community is a crucial step for metabarcoding and biodiversity studies. There is, therefore, a need for a high quality automatic taxonomic inventory of environmental samples from NGS approaches targeting a specific DNA region.

Results: We present here a tool dedicated to metabarcoding, *metaMatch*, that can assess the taxonomic inventory of a community sample from an NGS read with (i) the best possible quality and (ii) at a speed similar to BLAST. We used this tool to analyse three mock communities of diatoms, the exact composition of which was actually known, as the strains artificially assembled in a sample had been taken from a strain collection, with known morphological and genetic characters. We observed that inventories based on NGS read libraries are dependent on the algorithm used to assign species names to reads.

Conclusions: Presented software permits algorithmic community taxonomic inventories with sets of reads from a mock community and a reference database as inputs, with speed comparable to BLAST, and significantly better accuracy (fewer false positives) than BLAST and UBLAST. The main reason for this improvement is that there is no heuristics for speeding the database searching process, and software relies on efficiency of existing algorithms for finding a perfect match of a word within a text. In order to derive metabarcoding inventories, heuristic algorithms can initially be run on a large reference database to determine the high taxonomic levels and *metaMatch* can then be used to obtain an expert taxonomic inventory for a particular group.

Background

One of the main limitations in ecological studies has been the cost of obtaining observational data, for plants, animals, fungi, and protists, when they were taxonomically identified using floras and faunas. Species recognition using molecular markers has been developing exponentially as a result of the introduction of DNA barcoding (Hebert et al., 2003), although molecular markers have in fact been used for decades (Hillis et al., 1996). It relies on the use of markers, the variability of which is consistent with taxa circumscription (see Doolittle 1999, and reference therein). Current needs and development are focusing on metabarcoding, targeting microbial communities with a short DNA region (Stoeck et al., 2009; Uroz et al., 2010; Koskinen et al., 2011) by NGS (Next-Generation Sequencing). Therefore, several programs have recently been developed to align sequences from NGS to reference sequences with accurate taxonomic identification. Concerning long reads (such as from 454 pyrosequencing), BLAST (Altschul et al., 1997) is a standard tool for matching sequences against very large databases. As an example, MEGAN (Huson et al., 2007), a widely used analysis program of large metagenomic data, starts from a read comparison performed by BLAST (Altschul et al., 1997) or similar tool. A suite called USearch (Edgar, 2010) focuses on speed and has been designed for NGS. It contains software called UBLAST which implements such an alignment with higher speed (it will be discussed in this paper, as it focuses on a small number of hits rather than all high quality hits)

In the context of metabarcoding of community samples, a barcode (or a nucleic marker) is amplified and sequenced using NGS. Reads are then matched against existing sequence databases designed with high quality taxonomic identification (e.g. BOLD, see http://www.boldsystems.org/views/login.php, or CBoL data portal, see http://bol.uvm.edu/). Ways to put sequences into taxonomic entities are either through quantifying their similarity to reference sequences or operational taxonomic units (OTU)-based methods (see Schloss & Westcott 2011 and references therein). Some workflows combines the two approaches by adding a clustering step before the assignation in order to assign only one sequence of each OTU and therefore reduce the analysis time (e.g. Jaguc software package Nebel et al., 2011). In the context of metabarcoding and ecological studies, accurate sequence comparisons are therefore crucial to correctly describe community biodiversity.

The organisms used here to test the reliability of an automated molecular method for species identification within microbial communities are freshwater diatoms, a phylum of eukaryotic unicellular algae. As for many protists, diatom taxonomy is not perfectly stable: there are species complex and often classification shifts, a great diversity has not yet been described and new diatoms are still being discovered (Mann et al., 2010). Overall, diatoms are selected here as a model because species identification is possible on a morphological base, from the features of the silica cell wall, which gives a reference for comparisons to molecular identification. Such a dual approach is not available for many microorganisms, as unicellular. Moreover diatoms are present in most environmental samples (freshwater or marine) and are widely used as indicator species making them a good candidate for applying pyrosequencing approaches.

Here, we describe a new tool, complementary to algorithms using heuristics in the framework of metabarcoding, called *metaMatch*, which yields exact matches of reads versus reference databases of a reasonable size (i.e. a few thousand specimens) with high taxonomic quality. This makes it possible to compile a quick community inventory indicating the presence/absence of taxa that have been adequately sampled in the reference database. We compare this tool with BLAST and UBLAST for accuracy and speed. BLAST and UBLAST includes some heuristics that are appropriate for exploring large databases, whereas *metaMatch* favours accuracy without heuristics. As a benchmark for accuracy, *metaMatch*, BLAST and UBLAST have been run on data sets from sequencing mock communities, where specific composition was exactly known, built from organisms the sequence has been known. This allows as well an estimate of discrepencies between Sanger sequences on one hand, and reads from NGS sequences on the other with optimal local alignment, likely to be due to sequencing errors or technological artefacts, as reads must be parts of Sanger reference sequences, and should be recognized as such.

Implementation

Model: diatom samples

The marker selected for identification is crucial in biodiversity assessment. The SSU rDNA has limitations concerning the separation of diatom species complexes (Moniz & Kaczmarska, 2009; Hamsher et al., 2011). However, we used the SSU rDNA because the

associated dataset in GenBank is probably the largest available for diatoms with the greatest taxonomic breadth, as this ribosomal region has been commonly used for phylogenetic studies (e.g. Medlin et al., 1988, 1993; Kooistra & Medlin, 1996; Beszteri et al., 2001). Moreover universal primers were available for this marker and success rate in amplification and sequencing were high. A reference SSU rDNA database (REF) was therefore compiled from 1412 diatom sequences either collected in our laboratory (303 sequences) or taken from GenBank (1109) with expert taxonomic identification. These sequences corresponded to 508 distinct species. Diatom strains (303) had been isolated from freshwater environments, and their cultures are maintained in the Thonon Culture Collection (http://www.inra.fr/carrtel-collection). These strains have been identified to species level on the basis of morphological criteria, and characterized genetically by Sanger sequencing of SSU rDNA. Sequences from GenBank were curated by eliminating poor quality sequences, environmental sequences and sequences thought to have been misidentified.

We built three artificial environmental samples (mock communities) designated mix C, mix PCR1, and mix PCR2. Thirty strains corresponding to 21 species were used. Strains were mixed and harvested by centrifugation in order to create a mix of cultures before extraction and amplification steps (Mix C). The DNA of mix C were extracted together and SSU rDNA (~ 1800 bp) of the whole community was amplified using 1F and 1528R primers (Medlin et al., 1988) according to author specifications. At the same time, each culture was also independently centrifuged. DNA of each culture was extracted and SSU rDNA was amplified separately using 1F and 1528R primers (Medlin et al., 1988) according to author specifications. Mixes PCR1 and PCR2 were made after extraction and amplification steps from pooled PCR products to create two mixes with different proportions for strains. The three mixed samples were then sequenced after DNA fractioning and tagging at the GenoToul Genomic facility (Toulouse, France) on a GS FLX Titanium PicoTiterPlate 454 (Roche) according to the Manufacturer's instructions. Reads from the three mixed samples were sorted according to sample tag sequences into three libraries. Reads were eliminated if they were too short, contained ambiguous bases, or displayed low complexity using PyroCleaner software version 1.0 (Mariette et al., 2011).

Algorithm for perfect match

The *metaMatch* tool is composed of two programs: *metaMatch*_PS (P for perfect, and S for single core) and *metaMatch*_PS_Taxo (Taxo for taxonomic inventory). *metaMatch*_PS has as entry a Fasta file of reads from a NGS library obtained for one mixed sample (designated QRY, for query) and a Fasta file of a reference database (REF) including taxonomic information for a given marker. It has as an output a text file written on the disk, which is described below. This output file plus a file giving a precise taxonomic description of each specimen in REF (each row is a reference sequence, and columns are Class, Order, Family, Genus, Species and Strain) are inputs of *metaMatch*_PS_Taxo. *metaMatch*_PS_Taxo yields as output file a list of the taxa of the reference database that are present in the mixed sample. This list is built from perfect matches between the library (QRY) and reference database (REF) obtained by pairwise comparisons (brute force). Perfect match between two strings is one of the oldest and most studied problems in bioinformatics (Gusfield, 1997), and several types of algorithms have been produced. Here, we selected C language for its speed. The procedure for *metaMatch*_PS involves "brute force", as it is based on a loop on all pairs (str1, str2), where string str1 is a reference in the database, and string str2 a read. We then use the function `strstr`, which returns a pointer to the first occurrence of str2 in str1 when it exists, or a null pointer if str2 is not included in str1. The pseudo code reads (without pointers)

```
for str2 in QRY
    write(query name)
    for str1 in REF
          y <- strstr(str1,str2)
          write (y)
          write(\t)
    end
    write(\n)
end
```

Let us suppose we have n reads and p references in the database (typically, $n \sim 10^5$ or $\sim 10^6$ and $p \sim 10^3$). We then produced an $n \times p$ matrix M, with -1 in cell $M(i, j)$, if the read in row i does not match the reference in column j, and the position of the first occurrence if the read in row i perfectly matches a substring of the reference in column j. Any edit, such as an SNP or indel, between the read and reference will yield a mismatch.

Table IV.1: *Search time, number of perfect matches and informative reads for each library and each method.*

Library	Reads*	Method	search time	Perfect matches	Informative reads (see text)
Mix PCR1	39 090	*metaMatch*	1 h 41	2378 (6,0%)	1175 (49,4%)
		BLAST_11	2 h 30 min	7786 (19,9%)	4120 (52,9%)
		BLAST_130	40 min	6687 (17,1%)	3280 (49,1%)
		USEARCH_11	9 min	2393 (6,1%)	2393 (100%)
		USEARCH_130	5 min	2340 (6,0%)	2340 (100%)
Mix PCR2	83 747	*metaMatch*	3 h 44	5394 (6,4%)	2945 (54,6%)
		BLAST_11	6 h 00 min	17408 (20,8%)	9879 (56,7%)
		BLAST_130	2 h 02 min	14081 (16,8%)	7319 (52,0%)
		USEARCH_11	22 min	5540 (6,6%)	5540 (100%)
		USEARCH_130	12 min	5514 (6,6%)	5514 (100%)
Mix C	38 423	*metaMatch*	1 h 35	3253 (8,5%)	1419 (43,6%)
		BLAST_11	4 h 00	7786 (20,3)	4120 (52,9%)
		BLAST_130	1 h 13 min	9615 (25,0%)	4133 (43,0%)
		USEARCH_11	11 min	3371 (8,8%)	3371 (100%)
		USEARCH_130	6 min	3231 (8,4%)	3231 (100%)

*Number of reads calculated on the total number of reads and the proportion of three markers pooled into equimolar concentrations.
Mixes PCR1, PCR2, and C correspond to the three reads libraries, and BLAST_11, UBLAST_11, BLAST_130 and UBLAST_130 to searches with word size settings of 11 bp and 130 bp, respectively. Times should be read as comparisons between methods. All implementations were on the same computer.

Then, in the second step, only reads with at least one perfect match are conserved. For each of them, the list of taxa for which there is at least one perfect match is given by *metaMatch*_PS_Taxo, from taxonomic knowledge in the reference database, REF. A given read may match several reference sequences perfectly. If all reference sequences with perfect matches belong to the same taxon, this taxon is given as an output, and this read is considered to be "informative" and the taxon is included in the species inventory of the mixed sample. Otherwise, the output is the list of taxa with perfect matches. Finally, a list of all these outputs is provided (see Supplementary material for an example) and the species inventory, obtained from "informative" reads, may be sorted from it.

Taxonomy Quality assessment

Libraries were matched to REF without any reassembling of SSU rDNA fragment using *metaMatch*_PS, BLAST (Altschul et al., 1997) (BLASTN version 2.2.24+) and UBLAST (Edgar, 2010) for benchmark. BLAST and UBLAST were calibrated at 100% identity of matches. *metaMatch* uses a 100% query coverage, we used therefore the same setting for UBLAST as this option is available. This option is not available in BLAST. BLAST and UBLAST used heuristics that find word-matches (anchors) between the reads from QRY and reference sequences from REF. If a perfect match of the anchor is met, extension is implemented. Thus, the word-size parameter controls the sensitivity and speed of the search. We used, therefore, both the default word size setting of BLAST (11 bp) and a word size of 130 bp, because our read sizes ranged from 131 to 616 bp. As BLAST and UBLAST works on both the sequence and its reverse complement, REF was also reversed to test *metaMatch* under the same conditions.

As the species composition had been controlled when building the 3 mixed samples, their taxonomic inventory is fully known (labeled as KNOWN). Each run of the selected algorithm (*metaMatch*, BLAST and UBLAST) produced an inventory. We therefore obtained as many inventories as the number of times that BLAST, UBLAST or *metaMatch* had been run. We computed inventories of informative reads after removing singletons because it was demonstrated that most of them were artifactual (Tedersoo et al., 2010). Pairwise distances between KNOWN and communities yielded by *metaMatch*, BLAST and UBLAST have been computed by Jaccard index. Then, a hierarchical aggregative clustering has been implemented on this matrix.

Table IV.2: *Number of species in each inventory, number of species detected and present in the mixed samples, and number of false positives at genera level (species belonging to genera not present in the mixed samples).*

Library	Method	Species number in inventory	Species number in KNOWN (/21)	False positives (at genera level)
Mix PCR1	*metaMatch*	23	19	1
	BLAST_11	35	19	8
	BLAST_130	34	19	5
	USEARCH_11	48	18	15
	USEARCH_130	43	18	10
Mix PCR2	*metaMatch*	27	20	1
	BLAST_11	49	20	16
	BLAST_130	42	20	9
	USEARCH_11	59	20	17
	USEARCH_130	60	20	15
Mix C	*metaMatch*	16	14	1
	BLAST_11	35	19	8
	BLAST_130	28	15	5
	USEARCH_11	46	15	15
	USEARCH_130	39	16	7

Mixes PCR1, PCR2, and C correspond to the three reads libraries, and BLAST_11, UBLAST_11, BLAST_130 and UBLAST_130 to searches with word size settings of 11 bp and 130 bp, respectively.

Results

Search Time

The analysis times in the three read libraries differed depending on the number of reads and the "search" method used. UBLAST analyses on the reference database containing 1412 sequences were significantly faster (Table IV.1) than those carried out using BLAST and *metaMatch*. Then *metaMatch* analyses were faster than those carried out using BLAST with the word size setting of 11 bp (Table IV.1). However, using a word size of 130 bp for BLAST reduced the time to significantly less than the search time using *metaMatch*. We therefore considered the search time for procedures *metaMatch* and BLAST to be similar.

Proportions of reads

For the three libraries, BLAST identified a higher number of perfect matches than *metaMatch* and UBLAST. Between 6 and 8.3% of reads were selected as perfect matches by *metaMatch* and between 6 to 8.8% by UBLAST whereas BLAST retrieved between 16.8 and 25% of reads depending on the sample.

One or several species names were then assigned to each of these reads by comparison to the taxonomic information of REF. For *metaMatch* and BLAST which are designed to find all possible matches in the reference libraries, the fraction of informative reads (i.e. reads corresponding only to one species) among those having perfect matches was around 50% and similar with both algorithms. On the other hand, 100% of reads were considered as informative by UBLAST because this search algorithm is designed to terminate the search as soon as the first match is found. All reads are informative by construction with UBLAST, irrespective of the possibility of several perfect matches on sequences from different taxa.

Comparison of inventories

The number of species detected depended on the search method used and on the read library. *metaMatch* detected 23, 27 and 16 species for Mixes PCR1, PCR2 and C, respectively; BLAST detected 35, 49 and 35 species with a word size of 11bp and, 34, 42 and 28 with a word size of 130 bp; and UBLAST detected 48, 59 and 46 species with a word size of 11bp and, 43, 60 and 39 with a word size of 130 bp (see Table IV.2). The correct number of species (KNOWN inventory) was 21.

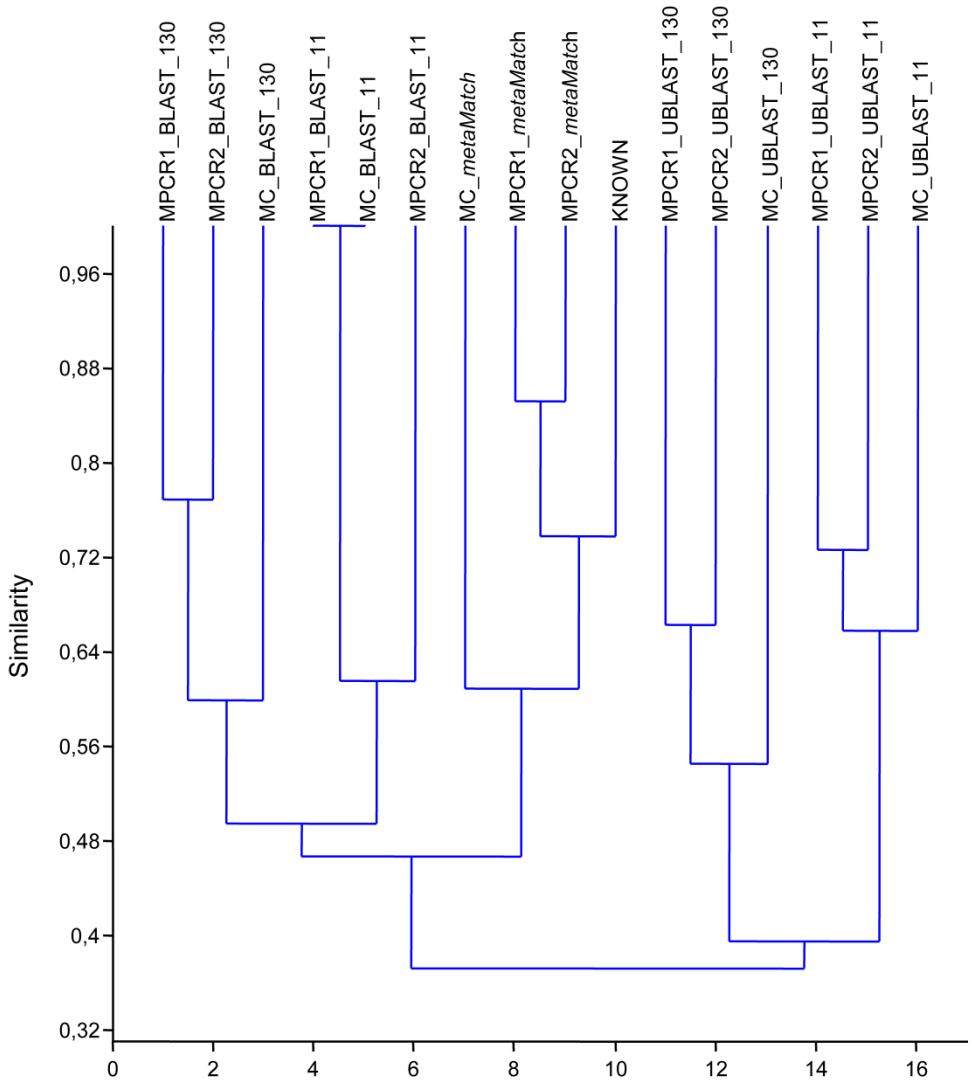


Figure IV.2: *Dendrogam of hierarchical aggregative clustering on pairwise Jaccard distances between 16 inventories*.

KNOWN is the expected inventory as assembled from known strains; Mx_*metaMatch* are inventories obtained as automatic outputs of algorithm *metaMatch*; Mx_BLAST_.. are outputs obtained by running BLAST; Mx_UBLAST_.. are outputs obtained by running UBLAST on same data (see text for details).

Some false positives (species present in the inventory which were not present in the mixed samples) and false negatives (species present in the mixed samples but not detected after analysis of reads) appeared with the three search methods.

The hierarchical agglomerative clustering using the Jaccard distances (see Figure IV.2) showed that inventories are dependent on the algorithm used to assign species names to reads. The three different samples were grouped depending on the algorithm and the parameter applied during analysis process. All three inventories from *metaMatch* were in the same clade than KNOWN and the only ones in this clade. They are closer to KNOWN than any of the inventories produced by BLAST or UBLAST. Thus, *metaMatch* yields metabarcoding inventories that are the closest to the actual composition as assessed by morphological criteria, followed by inventories provided by BLAST (Figure IV.2).

As expected, the impossibility to filter informative and non informative reads for the UBLAST search lead to a weaker assessment of the species composition.

False negatives and false positives

Almost all the species in the mixed samples were detected by *metaMatch*. Only one species of KNOWN was not detected in Mix PCR2, two in Mix PCR1 and five in Mix C (Table IV.2). Indeed, some reads matched both with these undetected species (false negatives), and also with other allied species belonging to the same genus included in REF. These reads were therefore not classified as informative reads, and were discarded. Some false positives were included in the inventories. Most of the false positives from *metaMatch* were closely related to species used in the mixed samples, except one which did not belong to genera present in the mixed samples (Table IV.2). BLAST and UBLAST detected roughly the same species of KNOWN than *metaMatch*. But these two heuristic algorithms added also many false positives (Table IV.2). Unlike *metaMatch*, BLAST and UBLAST overestimated the species richness, and included distant species (i.e. not belonging to genera present in the mock community) in the inventories (Table IV.2: between 5 and 16 with BLAST and between 7 and 17 with UBLAST).

Discussion

The results presented here quantify a comparison between the proposed tool (*metaMatch*), and a standard one (BLAST, and UBLAST for NGS). Unsurprinsigly, UBLAST outperforms BLAST and *metaMatch* for speed, whereas BLAST and

metaMatch run with similar speed. *metaMatch* outperforms both BLAST and UBLAST for accuracy. BLAST is commonly used to assign sequences to taxa (Buée et al., 2009; Stoeck et al., 2009) and UBLAST was recently developed to speed up sequence comparison in metabarcoding (Edgar, 2010). *metaMatch* has the additional advantage that it does not require specific settings in implementation, as it contains no heuristics.

As the mock community has been built from strains, the sequences of which are present in the reference library, it was expected that most of the reads would lead to a perfect match (informative or not). Surprisingly, a low number of perfect matches has been found. The imperfect matches were due to the errors which occurred during amplification and pyrosequencing steps and to the sequence lengths of the reference database. Indeed, SSU rDNA sequenced using 454 pyrosequencing were longer than SSU rDNA of reference sequenced using Sanger sequencing. For example, ~42 % of the reads of the artificial samples contained one PCR primer in their sequence. On the other hand, only 16.2 % of the reference sequences of the species included in the artificial samples contained the forward or the reverse PCR primer. BLAST led to incorrect perfect matches, as it quantifies the quality of match of the coverage. If coverage is lower than 100%, a 100% homology is not a perfect match. This made it possible to find more reads with perfect matches (around 20%) than *metaMatch* and UBLAST. Using *metaMatch* and UBLAST, we used perfect matches on the whole read size (100% coverage of the query), leading to the removal of reads which only overlapped reference sequences. BLAST used therefore the highest number of reads but the same proportion of reads than *metaMatch* was discarded after filtering reads into informative and non informative reads. On the contrary, due to the design of UBLAST, it was not possible to apply this filter on its output. However, one option in UBLAST is available to search matches on the entire database but this strategy is very similar to BLAST and may increase the analysis time.

Thus, *metaMatch* selected the lowest number of reads to compute inventories. The exact algorithm is the strictest. Moreover, many of the reads analysed by BLAST and UBLAST as informative led to false positives. There were fewer false positives when the "wordsize" setting was increased, but they were not completely eliminated. Several distant species (i.e. not belonging to the genera used to create the mock communities) were added in all inventories showing the lower accuracy of UBLAST and, to a lesser extent, of BLAST. Although *metaMatch* selected fewer reads, this strict algorithm added fewer false positives and provided an inventory that was closer to the KNOWN result

than BLAST or UBLAST did. Several parameters can be adjusted in BLAST and UBLAST to balance within the trade-off between speed and accuracy. Using *metaMatch*, we combined 100% homology on 100% coverage of a query with a filter of informative reads, which led to a better estimation of community composition without prior settings. Moreover, the number of false positives could be decreased using another marker presenting higher polymorphism than SSU rDNA more adapted to identification at the species level (*rbc*L, results not shown).

NGS methods provide so many sequences that a drastic selection can be applied to enhance the result accuracy. Selected reads were therefore sufficient to compile a confident inventory of the species in a particular diatom community. *metaMatch* outperforms both BLAST and UBLAST on accuracy. Moreover speed is not a problem as it is easy to distribute *metaMatch* on several cores or even a grid like EGI (European Grid Instrument), just by splitting the QRY file and sending each subfile on one core (SGE program available). Accurate identification of the species in a community is a crucial step for biodiversity studies. Behnke et al. (2011) used SSU rDNA to analyse microbial eukaryote communities, and reported a high pyrosequencing error rate, which may explain the severity of the read selection by *metaMatch*: using *metaMatch*, focusing on long reads and exact matches with full coverage made it possible to disregard pyrosequencing errors. Almost all sequencing errors, indels and SNPs were rejected. Only sequencing errors which corresponded to another sequence in REF were conserved and led to false positives, which were often close species. Moreover the filter of informative reads allows selection of non ambiguous reads only. The divergences observed between the inventories obtained using the 2 different PCR mixes (mix PCR1 and mix PCR2), and the pooled-culture sample (mix C) demonstrate the presence of extraction and PCR biases. These two steps increased the number of false negatives, but did not increase the number of false positives. Interestingly, the selection of informative reads that displayed 100% identity with only one species automatically discarded chimera.

The quality of the reference database is a crucial requirement in order to obtain reliable inventories in metabarcoding. The use of *metaMatch* requires cleaned sequence reference databases, including the maximum number of species that could potentially be detected in the community of interest. To study the communities present in a natural environmental sample, several sequences of each species will be necessary in order to cope with within species variability. With a reference database containing sequences

reflecting intraspecies variability, *metaMatch* can be used to detect different populations of a particular species. However, quantification of match imperfection is compulsory in such a case. Therefore, *metaMatch* is currently being enriched by a module computing edit distances between reads and references through local alignment, focusing on algorithms efficient when the distance between the reference and the read in the region common to both of them is due to a few SNPs and short indels only.

Conclusions

The development of barcoding was triggered by the need to identify species unequivocally. Microbial ecology is developing exponentially from avalanche of data from NGS. Therefore, species level, whatever the incessant discussions about species definition, remains a building block for community ecology. Then, in the context of metabarcoding, it is a key objective to provide tools for accurate species inventory in microbial communities. We demonstrated here that the results (inventories) were mainly dependent on the algorithm and on the parameters used to assign the sequences. The undefined coverage of the BLAST search and the single match of UBLAST, tested in this study, limited the accuracy of the inventories they produced. The analysis of large data sets, such as those supplied by NGS, requires a compromise between speed and accuracy. Using artificially built communities from strains from a collection, we propose here a tool dedicated to the purpose of high taxonomic quality inventories, *metaMatch*. Its main advantages are to provide accurate inventories without adjusting parameters and with a suitable search time.

To assess the taxonomic composition of microbial communities in metabarcoding, BLAST or UBLAST can initially be run on a large reference database to determine the composition at a high taxonomic level and to sort for large taxonomic groups (such as filtering the diatoms contained in an environmental freshwater sample), and *metaMatch* can be then applied within these groups to obtain an expert taxonomic inventory for a particular group from a dedicated, high quality taxonomic reference database. On an environmental sample, composed by different lineages, *metaMatch*, used as complementary and nested approach to heuristics methods such as BLAST or UBLAST, can provide an accurate inventory of the species in a community that are described in a reference database. *metaMatch* does not evaluate the diversity of taxa which are not in

REF but this algorithm can be used for different purposes, e.g. to obtain inventories of known species from environmental samples required in biomonitoring program.

Availability and requirements

Software *metaMatch* and documentation are available at http://w3.pierroton.inra.fr/biogeco/Bio2/

Author's contributions

Experiment has been designed and discussed through all steps by all participants. L.K, F.R. and A.B. designed the artificial community assembly and inventories comparisons. A.F., P.C. and J.M.F. designed data analysis and wrote the code. L.K. produced the reads. L.K. and A.F. implemented data analysis, and wrote the paper.

Acknowledgements and funding

All authors acknowledge network R-Syst for developing collaborations and discussions. Pyrosequencing and production of reads has been made at Genotoul facilities, by L.K. with help of Eugénie Robe and Olivier Bouchez.

This work has been supported by "projet innovant CODAL" (given to A.B. & A.F.) from Division Ecology of Forests, Grassland and Freshwaters of INRA, and projet innovant (given to A.F.) from Division Plant Health and Environment of INRA. Network R-Syst is supported by both Divisions.

Supplementary material: Example of output file provided by metaMatch.

"GP7CDQH01B0MFW|521 -> Nitzschia_acidoclinata"
"GP7CDQH01CLPA7|447 -> Navicula_cryptocephala"
"GP7CDQH01D1RJG|514 -> Gomphonema_parvulum"
"GP7CDQH01D7RD8|524 -> Gomphonema_bourbonense"
"GP7CDQH01DQEVR|419 -> Cyclotella_meneghiniana"
"GP7CDQH01DX5VR|463 -> Gomphonema_clevei"
"GP7CDQH01D5MEQ|516 -> Ulnaria_ulna"
"GP7CDQH01DVYQQ|486 -> Cyclotella_meneghiniana"
"GP7CDQH01CMCMV|457 -> Ulnaria_ulna"
"GP7CDQH01C2TX1|519 -> Gomphonema_parvulum"
"GP7CDQH01E09H2|274 -> Denticula_kuetzingii or Nitzschia_amphibia or Nitzschia_inconspicua"
"GP7CDQH01DLAQB|522 -> Gomphonema_affine or Gomphonema_parvulum"
"GP7CDQH01CYH03|420 -> Gomphonema_affine or Gomphonema_parvulum"
"GP7CDQH01D6VIN|443 -> Cyclotella_meneghiniana"
"GP7CDQH01AO1EE|455 -> Gomphonema_bourbonense"
"GP7CDQH01D7S60|504 -> Gomphonema_bourbonense"
"GP7CDQH01B6III|524 -> Gomphonema_affine or Gomphonema_parvulum"
"GP7CDQH01D7QMT|541 -> Fistulifera_pelliculosa or Fistulifera_saprophila"
"GP7CDQH01DOC9O|463 -> Ulnaria_ulna"
"GP7CDQH01BGQIX|485 -> Nitzschia_inconspicua"
"GP7CDQH01ET61M|475 -> Gomphonema_affine or Gomphonema_parvulum"
"GP7CDQH01DV0D4|486 -> Mayamaea_atomus or Mayamaea_permitis"
"GP7CDQH01CNBNR|470 -> Fragilaria_capucina"
"GP7CDQH01CHVQT|488 -> Cocconeis_pediculus or Cocconeis_placentula"
"GP7CDQH01CQ4F2|514 -> Mayamaea_permitis"
"GP7CDQH01AX4SK|494 -> Gomphonema_affine"
"GP7CDQH01DKVJ2|178 -> Centronella_reicheltii or Fragilaria_austriaca or Fragilaria_bidens or Fragilaria_capucina or Fragilaria_crotonensis or Fragilaria_nanana or Fragilaria_unassigned or Fragilariforma_virescens or Mayamaea_atomus or Mayamaea_fossalis or Mayamaea_permitis or Phaeodactylum_tricornutum or Phaeodactylum_unassigned or Ulnaria_ulna"
"GP7CDQH01A25WO|499 -> Mayamaea_atomus or Mayamaea_permitis"
"GP7CDQH01BVR9B|517 -> Gomphonema_affine or Gomphonema_parvulum"
"GP7CDQH01CEWEX|451 -> Fragilaria_capucina or Fragilaria_unassigned"
"GP7CDQH01DUKPZ|451 -> Gomphonema_affine or Gomphonema_parvulum"
"GP7CDQH01B7IM1|483 -> Gomphonema_affine or Gomphonema_parvulum"
"GP7CDQH01CJCXQ|518 -> Mayamaea_permitis"
"GP7CDQH01DNTH1|521 -> Gomphonema_affine or Gomphonema_parvulum or Gomphonema_unassigned"
"GP7CDQH01DPL1H|260 -> Nitzschia_acidoclinata"
"GP7CDQH01CWQPZ|426 -> Gomphonema_bourbonense"
"GP7CDQH01A8BHZ|497 -> Nitzschia_palea"
"GP7CDQH01D3DLZ|466 -> Gomphonema_affine or Gomphonema_parvulum"
"GP7CDQH01DJUV7|485 -> Navicula_cryptocephala"
"GP7CDQH01AIRD8|513 -> Gomphonema_bourbonense"
"GP7CDQH01EALQA|471 -> Sellaphora_pupula"

3. Comparaison de marqueurs moléculaires pour réaliser des inventaires d'espèces à partir des NGS : approche sur communautés artificielles

3.1. *Présentation générale de l'étude et synthèse des principaux résultats.*

Comme nous l'avons indiqué précédemment, les différents marqueurs moléculaires transmettent des informations à différentes profondeurs taxonomiques. Ainsi, pour développer un nouvel outil de bioindication par des méthodes moléculaires, le choix du marqueur nucléique est crucial. En outre, les différentes étapes des méthodes moléculaires peuvent engendrer des erreurs et des biais qui peuvent modifier la description de la biodiversité d'un échantillon : l'extraction (Martin-Laurent et al., 2001; Carrigg et al., 2007), l'amplification PCR (von Wintzingerode et al., 1997; Suzuki et al., 1998; Qiu et al., 2001; Kanagawa, 2003; Wu et al., 2010) et le pyroséquençage 454 (Margulies et al., 2005; Kunin et al., 2010; Tedersoo et al., 2010). Comme les erreurs de pyroséquençage, peuvent être dépendantes de la région ciblée (Behnke et al., 2011), ces biais doivent être évalués pour les différents marqueurs afin d'optimiser la méthodologie.

Ainsi, l'objectif de cette étude était de tester et de comparer l'efficacité de trois marqueurs moléculaires, pour déterminer les inventaires de diatomées dans des échantillons de communautés, et pour évaluer les biais liés à l'utilisation des méthodes moléculaires après amplification de ces marqueurs.

Dans ce but, nous avons analysé, par pyroséquençage 454, nos trois communautés artificielles de diatomées (un mélange de cultures et deux mélanges de produits PCR, CHAPITRE II.3.2.2(a)) en utilisant trois marqueurs nucléiques différents: l'ADNr 18S, le gène *rbc*L et le gène *cox*1. Nous avons également élaboré des bases de séquences de référence pour chacun des trois marqueurs testés en compilant les séquences de notre collection de culture et des séquences provenant de GenBank (CHAPITRE II.3.1.3). Pour chaque échantillon et chaque marqueur, les reads ont été comparés aux bases de référence (en utilisant *metaMatch*_PS puis *metaMatch*_PS_Taxo) afin d'établir un inventaire (en présence / absence) des taxa de diatomées présents dans l'échantillon.

Nous avons comparé les marqueurs de différentes manières:

- Les bases de séquences de référence ont été étudiées en observant le polymorphisme des séquences.
- Après avoir assigné un ou plusieurs noms de taxa à chaque read, les pouvoirs résolutifs des marqueurs ont été comparés à différents niveaux taxonomiques (du clade à la sous-division).
- Pour chaque marqueur, les inventaires obtenus pour les trois échantillons ont été comparés, entre eux et à la composition réelle des échantillons artificiels, afin d'évaluer les biais liés aux techniques de biologie moléculaire.
- Pour chaque échantillon, les inventaires obtenus pour les trois marqueurs ont été comparés afin de déterminer le marqueur le plus efficace pour décrire l'assemblage de diatomées.

Bien que la base de données de référence des séquences d'ADNr 18S soit la plus importante et couvre une large gamme d'espèces de diatomées (508 espèces), les inventaires obtenus par pyroséquençage en utilisant ce gène sont éloignés de l'inventaire réel. En raison de la distribution du polymorphisme (alternance de régions conservées et de régions polymorphes), ce gène présente un faible pouvoir de résolution avec seulement 50% de reads informatifs au niveau de l'espèce. De plus, l'ADNr 18S est inefficace pour discriminer certains complexes d'espèces. Enfin, les erreurs de séquençage aboutissent à la présence de nombreux faux positifs (espèces détectées par l'analyse mais non présentes dans la communauté artificielle) dans les inventaires d'espèces de diatomées.

Ensuite, nos analyses ont démontré que le fragment du gène *cox*1 semble avoir le pouvoir résolutif le plus important avec 100% de reads informatifs à tous les niveaux taxonomiques étudiés. De plus, ce segment *cox*1 se révèle capable de discriminer les clades des complexes d'espèces. Toutefois, les inventaires d'espèces obtenus à partir des séquences de *cox*1 sont les plus éloignés du résultat attendu car la base de référence est très peu représentative de la diversité des diatomées. En effet, seulement 63 espèces sont représentées dans la base de référence dont onze des 21 espèces du mélange artificiel. Les difficultés de séquençage, dues à la faible efficacité des amorces PCR disponibles actuellement, limitent le développement de la base de référence de *cox*1, et donc l'utilisation de ce gène pour étudier les communautés de diatomées.

Enfin, les inventaires d'espèces et de clades obtenus à partir des séquences de *rbc*L sont étroitement liés aux inventaires attendus. Ce gène *rbc*L a l'avantage d'allier un fort pouvoir résolutif (> 94% de reads informatifs au niveau de l'espèce) à une importante base de référence (1071 séquences correspondant à 407 espèces). De plus, la distribution du polymorphisme le long de ce fragment nucléique laisse penser qu'il serait possible de cibler un plus petit fragment du gène sans perdre de pouvoir résolutif. Ainsi, le protocole de pyroséquençage pourrait être simplifié en éliminant l'étape de fractionnement et l'analyse pourrait être améliorée en ciblant un fragment bien représenté dans la base de référence. Cependant, l'analyse du mélange de cultures (l'échantillon le plus proche d'un échantillon naturel) a mis en évidence la difficulté de détection des espèces faiblement représentées (< 1%). Ce biais, qui n'est pas retrouvé avec les échantillons en mélange de produits PCR, provient donc des étapes d'extraction et de PCR. Ainsi, même si des optimisations des protocoles sont encore nécessaires pour l'utilisation du *rbc*L, celui-ci paraît, actuellement, être le meilleur gène disponible pour identifier les espèces de diatomées présentes au sein d'une communauté.

3.2. *Article IV : soumis dans la revue PLoS ONE*

New Biomonitoring Approaches based on Next Generation Sequencing : a Test for Freshwater Diatom Communities

Lenaïg Kermarrec[1,2], Alain Franc[3,4], Frédéric Rimet[1], Philippe Chaumeil[3,4], Jean-François Humbert[5] and Agnès Bouchez[1]

[1] INRA, UMR CARRTEL, 75 av de Corzent, BP 511, F-74203 Thonon-les-Bains cedex, France

[2] Asconit Consultants, 3 bd Clairfont, F-66350 Toulouges, France

[3] INRA, UMR BIOGECO, 69 route d'Arcachon, F-33612 Cestas cedex, France

[4] University of Bordeaux 1, UMR BIOGECO, 33400 Talence, France.

[5] INRA, UMR BIOEMCO, site de l'ENS, 46 rue d'Ulm, F-75005 Paris, France

Corresponding author: Agnès Bouchez: agnes.bouchez@thonon.inra.fr

Soumis dans PLoS ONE le 9 janvier 2012

Abstract

Several diatom indices can be used to assess the ecological quality of rivers, but they are time consuming and require a high degree of taxonomic expertise. In the context of devising new automated tools for assessing the species composition of diatom communities, we tested the capacities of 454 pyrosequencing methods on three nucleic markers (SSU rDNA, *rbc*L, and *cox*1). To do this, we used three artificial samples of a mock community composed of 30 freshwater diatom strains belonging to 21 species. To detect any methodological bias, one sample was made directly from pooled cultures, whereas the other two samples consisted of pooled PCR products. The reads for each sample and each marker were compared to a DNA reference library to obtain taxonomic inventories, which were then compared to the real inventory. Although the DNA reference library was largest for SSU rDNA, this gene has a low resolving power at all taxonomic levels because of its conserved regions. Due to sequencing errors, some false positives and false negatives were observed. For c*ox*1, polymorphism was distributed throughout the sequence. It also had the highest resolving power, and was suitable for distinguishing diatoms to an infraspecies level. However, the species inventories obtained from *cox*1 sequences were the furthest from the expected results, because of its incomplete DNA reference library. In contrast, inventories obtained from *rbc*L data were closely related to the expected findings. R*bc*L has the advantage of offering high resolving power with a large DNA reference library, making it possible to achieve accurate identification despite methodological bias. We therefore conclude that 454 pyrosequencing is suitable for identifying diatoms in assemblages, and that *rbc*L seems to be the best marker available at present.

Introduction

Freshwater ecosystems worldwide are subjected to growing anthropic pressures, which are damaging their biodiversity and their functioning, and also compromising the ecosystem services they provide (Allan, 2004; Holland et al., 2011). These disruptions have stimulated the implementation of monitoring programs, intended to assess the chemical and the biological qualities of these ecosystems in numerous countries. In Europe, for example, the EC implemented the Water Framework Directive at the beginning of the 2000s, with the goal of achieving "good status" for all surface water by 2015 (Directive 2000/60/EC). Many of these monitoring programs are based on the use of biological indicators, which require an estimation of the taxa composition (taxonomic identification and counting) of the targeted communities. Compiling such inventories is time and money consuming, and requires a high level of taxonomic expertise (Gardner et al., 2008; Mandelik et al., 2010).

Due to difficulties in the accurate taxa identification of many micro- and macro-organisms, Hebert et al. (2003) have proposed that the DNA barcode approach should be used to identify all organisms. The potential of this barcode approach has been demonstrated for freshwater communities used as bioindicators, as recently shown by Sweeney et al. (2011) for macroinvertebrate communities, and by Mann et al. (2010) for diatom species. However, one of the main limitations of using DNA barcoding to characterize environmental samples is linked to its use of the Sanger sequencing, which requires the construction of clone libraries for microorganisms before they can be sequenced. This is time consuming, and can introduce further bias. For this reason, the introduction of Next Generation Sequencing (NGS), and that of 454 pyrosequencing in particular, has provided new perspectives for the use of barcodes for biomonitoring ecosystems, as it provides a quick community inventory of taxa by comparison to a DNA reference library. In this way, Hajibabaei et al. (2011) have shown that a 454-sequencing approach is a very promising tool for the biomonitoring of freshwater ecosystems in an urban region using benthic macroinvertebrate communities. However, this pioneer study highlighted the urgent need for data mining to make it possible to use NGS for biomonitoring. In order to establish taxonomic inventories, these novel molecular methods require the following successive steps: DNA extraction and marker amplification, massive sequencing and automated bioinformatic assignment of sequences using taxonomic DNA reference libraries. All these steps have their limits and specific

requirements that need to be explored to avoid bias that could mar taxonomic inventories and consequently environmental evaluations. To date, few studies have explored the necessary features and biases associated with automated molecular approaches (Tedersoo et al., 2010; Behnke et al., 2011; Gilles et al., 2011), and work is still required with regard to their application to environmental communities.

We present here a "proof of concept" test to evaluate the suitability of NGS for the taxonomic characterization of the diatom communities used in biomonitoring indices. We took into account the fact that the suitability of such a molecular approach for taxa identification depends on (i) the quality (technical quality and taxonomic reliability) and representativeness of the DNA reference libraries, (ii) the resolving power of the marker used, (iii) the sources of methodological bias linked to molecular techniques and (iv) the quality of assignment generated by the bioinformatic method. To do this, we worked in a controlled framework using mock communities (known taxonomic compositions, known sequences), and an algorithm for taxonomic assignation that is appropriate for such controlled samples. In this context, we compared the efficiency of the three markers most commonly used for molecular diatom identification (SSU rDNA, small subunit ribosomal DNA) from the nuclear genome, *rbc*L from the chloroplast genome, and *cox*1 from the mitochondrial genome) to perform taxonomic inventories on three artificial bulk samples made up of strain cultures. The 30-strain assemblage chosen for these samples was representative of the range of variation observed in environmental samples: from the inter-genera to intra-species level. The mix of strains was prepared either prior to DNA extraction and PCR amplification (C mix), or post PCR amplification (PCR1 and PCR2 mixes), in order to detect any bias due to these biomolecular steps. Three DNA reference libraries linking a taxonomical identification to marker sequences obtained by Sanger sequencing (SSU rDNA, *rbc*L, and *cox*1) were compiled and used for reads assignment. Based on selected reads, taxa inventories were made for each marker and for each sample, and their accuracies evaluated by comparison to the known inventory of our mock communities. By taking all our findings together, we were able to identify the most powerful genetic marker for assessing the composition of diatom assemblages, and to identify some of the limitations and bias generated by the methodologies tested. Finally, our results shed some light on the boundaries and prospects for using taxonomic inventories based on NGS of environmental communities for biomonitoring purposes.

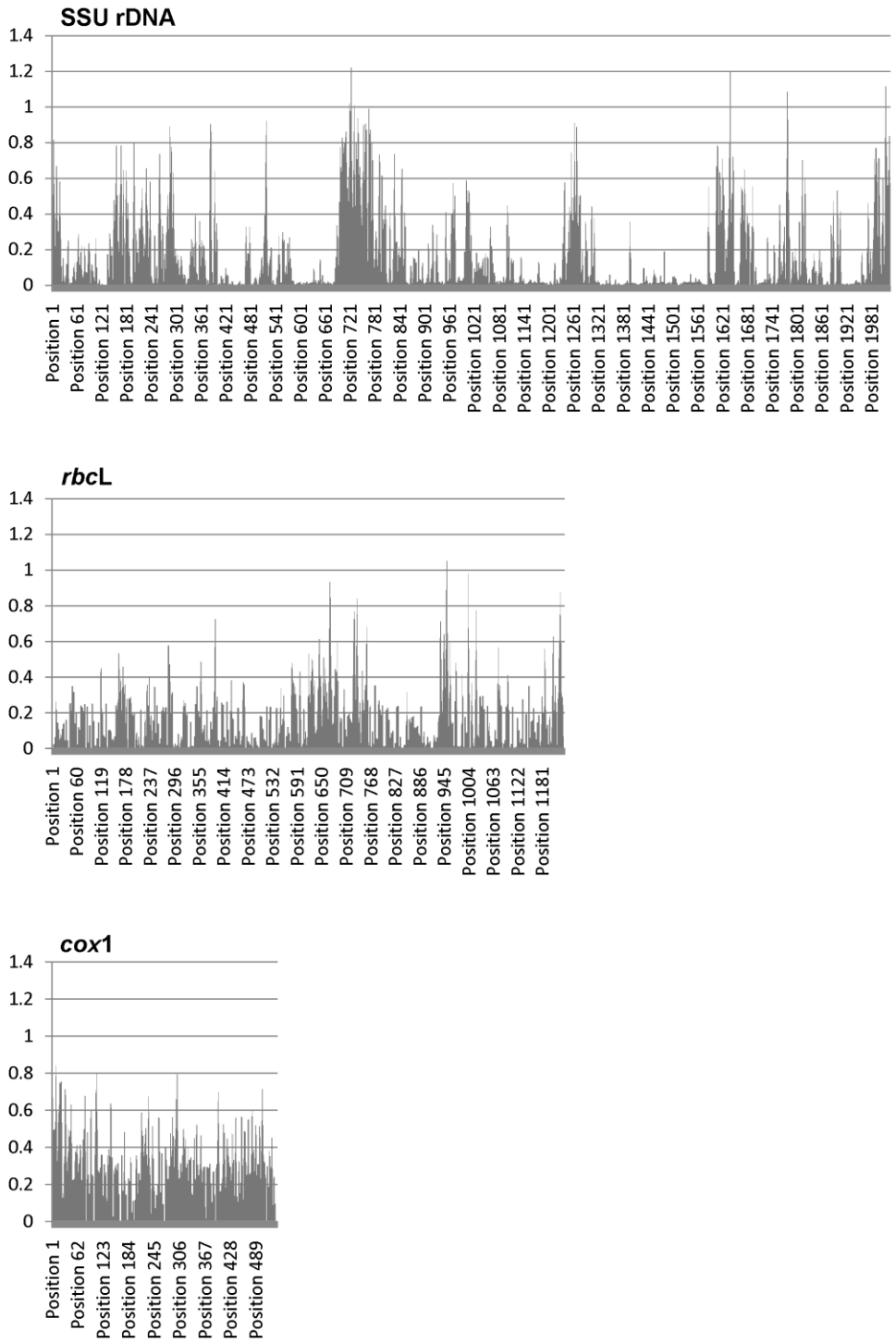


Figure IV.3: *Shannon index plot showing the amount of variability of each marker.* The Shannon index was calculated for each column of the alignments of the SSU rDNA, *rbc*L, and *cox*1 sequences from the reference DNA libraries. To make the figure easier to read, averages over the three columns are shown.

Results

Genetic polymorphism in the three markers (SSU rDNA, rbcL, and cox1) and the effectiveness of these markers for taxonomic identification

We included sequences from our culture collection together with sequences selected from GenBank in our three DNA reference libraries (SSU rDNA, *rbc*L, and *cox*1). The final alignment of the three DNA libraries (after removing shorter sequences and extremities) contained 1172 sequences for the SSU rDNA marker (3019 columns including gaps), 703 sequences for the *rbc*L marker (1242 columns), and 190 sequences for the *cox*1 marker (555 columns). These sequences corresponded to 407, 302, and 43 species, respectively. The sequence polymorphism of each genetic marker was estimated by calculating a Shannon index value for each position (column) of the three alignments (Figure IV.3). The SSU rDNA gene and, to a lesser extent, the *rbc*L sequences displayed alternating highly-conserved and polymorphic regions. In contrast, in the *cox*1 fragment, the polymorphism was distributed throughout the sequences, and the mean Shannon index value was higher than for the two other markers.

For each mix, a variable number of reads was obtained by pyrosequencing after removing the low-quality sequences, with an average of 161 256 reads per mix (Table IV.3). The average length value of these reads was 421 bp, with minimum and maximum values of 131 and 616 bp, respectively. By using *metaMatch*_PS algorithm (Kermarrec et al., technical report), each read was compared to the sequences of our three DNA libraries (SSU rDNA, *rbc*L, and *cox*1). Around 3.8% of the reads displayed a perfect match (100% sequence identity) with at least one sequence in our three DNA libraries (Table IV.3). Using perfect matches makes it possible to avoid diatom taxonomy problems. However, the low percentage of perfect matches is probably attributable to the presence of amplification and sequencing errors (including chimeras), together with incomplete overlapping of reads on the reference libraries. Most of the selected reads were obtained for the SSU rDNA marker, whereas very few perfect matches were found for *cox*1.

Table IV.3: *Total number of reads, number of perfect matches and number of species informative reads for each pyrosequencing library (Mix C, PCR1 Mix and PCR2 Mix) and each marker (SSU rDNA,* rbc*L, and* cox*1).*

		Number of reads	Perfect matches	Species informative reads
Mix C	SSU rDNA	115269	3206	1426 (44%)*
	*rbc*L		1004	944 (94%)
	*cox*1		175	175 (100%)
Mix PCR1	SSU rDNA	117269	2336	1185 (50%)
	*rbc*L		2049	1954 (95%)
	*cox*1		141	141 (100%)
Mix PCR2	SSU rDNA	251229	5248	2957 (56%)
	*rbc*L		3820	3705 (97%)
	*cox*1		233	233 (100%)

*: % of informative species reads among sequences displaying a perfect match.

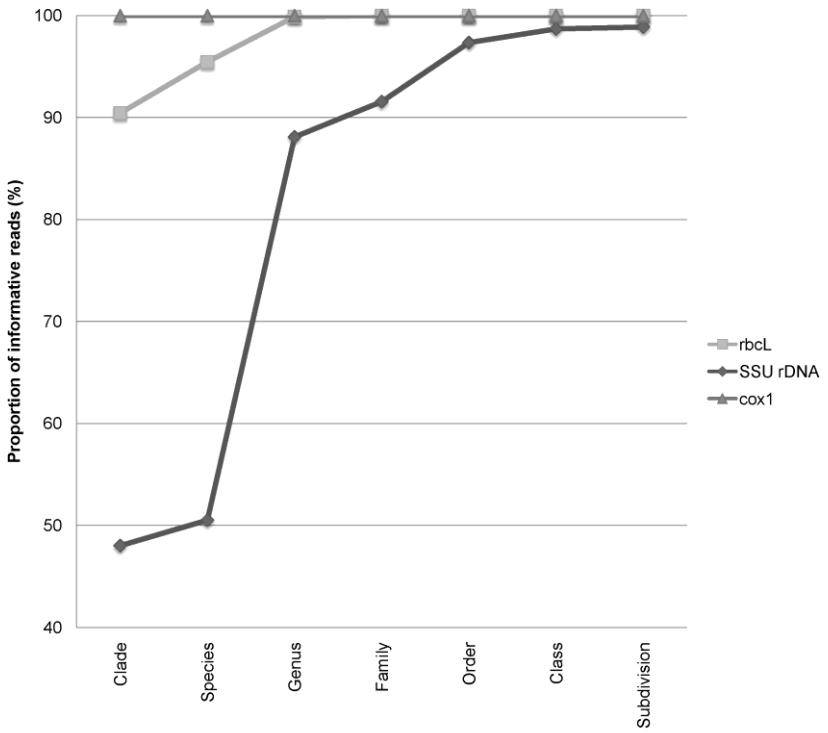


Figure IV.4: *Proportion of informative reads at different taxonomic levels: from clade to subdivision.*

In order to establish taxonomic inventories, we selected among these perfect matches only the "informative" reads, which were associated with a single species. For the *rbc*L, and the *cox*1, almost all reads (>94%) displaying a perfect match with a sequence from our reference libraries were associated with single species (Table IV.3). On the other hand, a high proportion (around 50%) of SSU rDNA reads displaying 100% sequence identity with sequences from our reference libraries, were associated with several species (Table IV.3). When considering the proportion of "informative" reads at different taxonomic levels (Figure IV.4), we confirmed the very high taxonomic resolution of *cox*1 compared to the lower resolution of *rbc*L, and the much poorer resolution of the SSU rDNA, which only provided a full taxonomic resolution at the class level with lower resolution at the clade and species levels.

Bias related to molecular biology techniques

With the aim of testing the potential impact of DNA extraction, PCR amplification, and pyrosequencing on the taxonomic inventories, we estimated the similarities between the expected species inventory of our mock community (21 known species) and those obtained after pyrosequencing the three mixes. Hierarchical agglomerative clustering (HAC) using Jaccard distances on species inventories (Figure IV.5) demonstrated that these inventories were different from the real inventory for all three markers. Only the inventory of the PCR1 mix obtained using *rbc*L corresponded exactly to the expected result. Methodological bias therefore occurred for the three markers and for the different mixes. As expected, the estimation of the species composition of our assemblages was better for the PCR mixes (PCR1 and PCR2) than for the culture mix C. The inventories of the PCR1 and PCR2 mixes were not the same, and only the *rbc*L inventories of the PCR1 and PCR2 mixes were closely related. When the three mixes were compared, their Jaccard similarities for *cox*1, and to a lesser degree those for *rbc*L, were higher than those obtained for SSU rDNA (Figure IV.5).

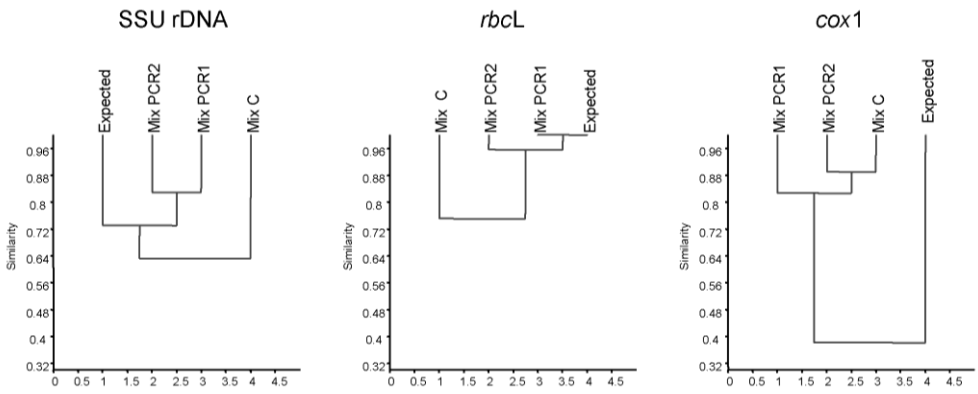


Figure IV.5: *Comparison of the species inventories made for each bulk sample to the real inventory.*

Dendrograms of hierarchical agglomerative clustering on pairwise Jaccard distances between the three mix inventories and the known inventory at species level, presented for each marker (SSU rDNA, *rbc*L, and *cox*1). "Expected" is the inventory of the mock community assembled using known strains.

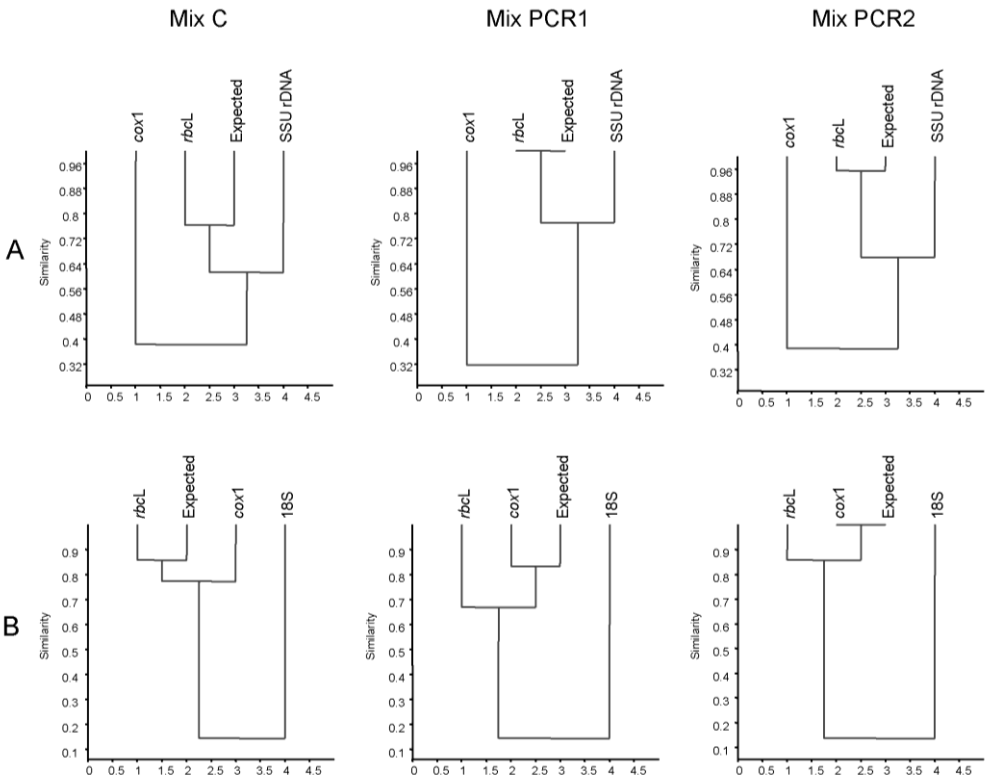


Figure IV.6: *Comparison of the species inventories made for each marker to the real inventory.*

Dendrograms of hierarchical agglomerative clustering on pairwise Jaccard distances between the three marker inventories and the known inventory. Dendrograms are presented for each sample (C, PCR1 and PCR2 mixes) at two taxonomical levels: species (A) and clades (B). "Expected" is the inventory of the mock community assembled using known strains.

Effectiveness of our DNA markers for correctly describing the species composition of our assemblages

With the aim of testing the effectiveness of our DNA markers for establishing the taxonomic inventories, we compared the species inventories provided by each marker to the real inventory. This was done for each of our three artificial communities by HAC using Jaccard distances (Figure IV.6A). The best estimation of the species composition was provided by the *rbc*L marker, followed by that provided by SSU rDNA. As expected, the small number of perfect matches for the *cox*1 marker made it impossible to obtain a good estimation of the species composition.

Almost all the species in the samples were detected by SSU rDNA (Table IV.4). Five species were false negatives (i.e. the species was not detected) in mix C, but only one in the PCR1 and PCR2 mixes. Other errors in the inventories correspond to false positives (i.e. to species not actually included in the mixes). In almost all cases, these species belonged to the same genera, and were closely related to species that had been included, for example *Gomphonema affine* Kützing, which is a sister species of *Gomphonema parvulum* (Kützing) Kützing.

Less than 43% of the 21 species contained in the artificial communities were retrieved using the *cox*1 marker (Table IV.4), and no false positive was added. Almost all the false negatives were due to the absence of the exactly corresponding sequences in the *cox*1 reference library, as we were not able to obtain the Sanger sequences of these strains, due to the poor amplification efficiency of the *cox*1 primers available. For two of these species we were able to add sequences available for other strains of the same species to the reference library. However, due to intraspecies variations, no read displayed perfect matches, and so these species could not be included in the inventories.

In the case of the *rbc*L analyses, the inventory of the C mix included five false negatives (Table IV.4). With the exception of *Nitzschia frustulum*, they were all species present at a low abundance in the mix of cultures (< 1% of the individuals counted under the microscope). The inventory of the PCR1 mix corresponded exactly to the expected inventory, and there was only one false positive (*Gomphonema capitatum* Ehrenberg) in the inventory of the PCR2 mix.

Table IV.4: *Taxonomic inventories obtained by 454 pyrosequencing of the three bulk samples (mix C, mix PCR1, mix PCR2).*
Species used to constitute the artificial communities are shown in gray.

	Mix C			PCR1 Mix			PCR2 Mix		
Species	SSU rDNA	*rbc*L	*cox*1	SSU rDNA	*rbc*L	*cox*1	SSU rDNA	*rbc*L	*cox*1
Amphiprora paludosa							X		
Amphora montana	X	X		X	X		X	X	
Cocconeis pediculus							X		
Cocconeis placentula				X	X		X	X	
Cyclotella meneghiniana	X	X		X	X		X	X	
Cylindrotheca closterium	X			X			X		
Fistulifera saprophila	X	X		X	X		X	X	
Fragilaria capucina	X	X		X	X		X	X	
Gomphonema acuminatum							X		
Gomphonema affine	X			X			X		
Gomphonema bourbonense	X	X	X	X	X	X	X	X	X
Gomphonema capitatum								X	
Gomphonema clavatum	X	X		X	X		X	X	
Gomphonema clevei	X			X	X		X	X	X
Gomphonema parvulum	X	X	X	X	X	X	X	X	X
Gomphonema pumilum	X			X	X		X	X	
Mayamaea permitis	X	X	X	X	X	X	X	X	X
Navicula cryptocephala	X	X		X	X		X	X	
Nitzschia acidoclinata	X	X	X	X	X	X	X	X	X
Nitzschia amphibia				X			X		
Nitzschia draveillensis	X	X		X	X		X	X	
Nitzschia frustulum				X	X		X	X	
Nitzschia inconspicua		X	X	X	X	X	X	X	X
Nitzschia lorenziana					X			X	
Nitzschia palea	X	X	X	X	X	X	X	X	X
Nitzschia unassigned	X						X		
Pinnularia acrosphaeria	X	X	X	X	X	X	X	X	X
Sellaphora auldreekie				X					
Sellaphora pupula				X			X		
Sellaphora seminulum		X	X	X	X		X	X	X
Surirella minuta	X								
Tryblionella apiculata	X								
Ulnaria ulna	X	X		X	X		X	X	
Number of species	21	16	8	25	21	7	28	22	9
False negative species	5	5	13	1	0	14	1	0	12
False positive species	5	0	0	5	0	0	8	1	0

With regard to the detection of clades, we worked on two species complexes; *Gomphonema parvulum* and *Nitzschia palea* (Kützing) Smith. Four clades of *G. parvulum* and 12 clades for *N. palea* were defined in the DNA reference library, but only three clades of each complex were used to create the samples. The HAC using Jaccard distances (Figure IV.6B) were based only on these two complexes, because clades had not been defined for the other species. When comparing the efficiency of our three DNA markers for estimating the taxa composition in our assemblages (Figure IV.6B), *cox*1, and *rbcL* were the two most efficient markers, and the SSU rDNA was always the least efficient. The SSU rDNA marker did not distinguish the complexes at the intraspecies level. For the C mix, the inventories from *rbc*L, and *cox*1 contained only one error each (one false positive for *rbc*L, and one false negative for *cox*1).

Discussion

The objective of this study was to test the efficiency of 454 pyrosequencing using three markers (SSU rDNA, *rbc*L, and *cox*1) for making taxonomic inventories of diatom communities. This was tested in three different ways: (i) the three DNA reference libraries were compared in terms of their usefulness for making taxonomic inventories, (ii) the inventories obtained for the three bulk samples of the mock community were compared to assess the bias linked with the molecular biology techniques, and (iii) to determine the most suitable marker for describing the diatom assemblages, the three markers were compared both in terms of their discriminatory efficiency at different taxonomic levels and in terms of the quality of the inventories produced.

DNA reference libraries for taxonomic inventories

The reference library of SSU rDNA sequences is the largest reference library, because this ribosomal region has been the main marker used for phylogenetic studies for a long time (e.g. Medlin et al., 1988, 1993; Kooistra & Medlin, 1996; Beszteri et al., 2001). It has been used so often, because the presence of conserved regions facilitates the design of universal primers and the presence of variable regions allows phylogenetic study. In the same way, the *rbc*L reference library comprised many sequences, because this gene has been used for cryptic diversity assessment (Evans et al., 2008; Trobajo et al., 2010) and for phylogenetic studies (Bruder & Medlin, 2007; Theriot et al., 2010). Recently, *rbc*L has been proposed as the reference barcode for diatoms, with LSU as secondary barcode

(Hamsher et al., 2011). As previously pointed out by Moniz & Kaczmarska (2009), Trobajo et al. (2010), and Hamsher et al. (2011), there has still been little success with *cox*1 fragment sequencing for diatoms. The Shannon index values demonstrate its high variability distributed throughout the entire length of the fragment, and explain why it is difficult to design universal primers for the diatoms as a whole. Primer design is still critical for the use of DNA barcoding, and the *cox*1 primers available do not reach the 95% amplification success proposed by Hajibabaei et al. (2005). This problem has resulted in a small DNA reference library (only 266 sequences).

According to Mann et al. (2010), new attempts to isolate and grow diatom strains are required to complete the DNA reference libraries required for a barcoding approach. A consensus should be also reached to define the primers used to create the DNA reference libraries. Currently, our reference libraries are composed by different fragment lengths and we used long amplicons for pyrosequencing. Using the *metaMatch* algorithm, we performed perfect matches on the whole read size. Reads which only overlapped the reference sequences were therefore discarded. These observations will have to be taken into consideration in future studies in order to optimize the design of the barcode used for 454 pyrosequencing (i.e. not including the extremities of the markers, which are often of varying lengths depending on the primers used for sequencing the reference libraries).

Bias related to molecular biology techniques

Our results revealed differences between the real inventory and the inventories obtained after pyrosequencing even when the analysis process had been designed to minimize artifacts. For an accurate analysis of molecular data from environmental samples, the methodological sources of bias have to be clearly defined in order to optimize protocols and decrease the effects of artifacts. Some of the divergences (false positives) were due to pyrosequencing errors leading to a false identification. Moreover, the differences between inventories from the PCR1 and PCR2 mixes showed that species abundance influenced the detection of species. Bias occurring during pyrosequencing had previously been documented and lead to an overestimation of diversity, which could be minimized using appropriate bioinformatic tools (e.g. Quince et al., 2009; Kunin et al., 2010). The stringency of the filters applied to our reads (using reads displaying 100% identity with a reference sequence) allowed us to reduce the effect of pyrosequencing errors, but depending on the marker used, this strategy did not allow us to avoid them totally. Pyrosequencing errors can generate false positive matches between species

displaying a high sequence similarity. From our data, it appeared that the SSU rDNA inventories were more subject to these artifacts than the *rbc*L inventories, which displayed very few differences from the real inventory for the PCR1 and PCR2 mixes. The low variation displayed by SSU rDNA combined with sequencing errors gave rise to more assignment errors. The use of a highly polymorphic marker such as *rbc*L combined with a suitable bioinformatic tool should make it possible to reduce the influence of pyrosequencing biases.

Artifacts due to DNA extraction and to PCR amplification (observed by comparing mix C with the PCR1 and PCR2 mixes) reduce the number of species detected. Major contributors to PCR errors are chimera and mutations (Acinas et al., 2005), which were discarded by our bioinformatic tool *metaMatch*, because we used only reads displaying 100% identity with a reference sequence. In addition to the sampling effect (related to the small number of reads used due to the 100% sequence identity threshold), some false negatives were not detected because of their low abundance. For mix C analyzed using *rbc*L four out of five species with abundance < 1% were not detected. This failure to detect species with a low abundance (<1%) had previously been observed by Hajibabaei et al. (2011). However, abundance does not explain all the results. For example, *Nitzschia draveillensis* was not detected by microscopy, but was detected by pyrosequencing, and *Nitzschia frustulum* was not detected by pyrosequencing despite an abundance of 1.9%, whereas other species with similar abundances (e.g. *Fistulifera saprophila*) were detected. Other factors, such as extraction and PCR bias, are also involved in detection divergences. The extraction method can influence the community profiles observed (DeSantis et al., 2005; Carrigg et al., 2007; Feinstein et al., 2009). We chose our extraction method from among others, because it was the most efficient method, but without any loss of diversity (observed by DGGE, data not shown). However, we assumed that the factors responsible for differential extraction rates (such as differences in cell wall resistance) may occur, whatever extraction protocol is used. Moreover, PCR divergences (due to primer selection, polymerase errors, G+C content, and cycle number) also arose, and may influence species richness and evenness (Reysenbach et al., 1992; Suzuki & Giovannoni, 1996; Polz & Cavanaugh, 1998; Acinas et al., 2005; Jumpponen, 2007; Huber et al., 2009; Engelbrektson et al., 2010). Thus, further optimizations of the extraction and PCR steps, in addition to lowering the identity threshold, will be required to enhance the detection of all species.

Resolving power and efficiency of our DNA markers for describing the composition of a diatom community

We demonstrated that the *cox*1 fragment, in which the polymorphism was distributed throughout the sequenced fragment, seemed to have the highest resolving power for diatoms, and was able to distinguish diatoms at an intraspecies level. This marker therefore appears to be suitable for identifying diatom taxa, as previously demonstrated by Evans et al. (2007). However, the species inventories obtained from *cox*1 sequences were distant from the true inventory due to the lower representativeness of the *cox*1 reference library. The main limitation on the use of *cox*1 is its DNA reference library. We therefore conclude that the *cox*1 gene cannot be used to study diatom communities until more efficient primers are available to enlarge the reference library. The design of good primers could constitute an avenue of research intended to improve the *cox*1 reference library, and thus to enhance the detection of diatom species, especially since the official barcode proposed by Hebert et al. (2003) was the *cox*1 gene.

We also showed that SSU rDNA reads provided poor resolution at the clade and species levels for diatoms, probably due to the presence of many conserved regions. Moreover, the low polymorphism of SSU rDNA makes it more sensitive to sequencing errors. Indeed, even a few sequencing errors in some short SSU reads led to the misidentification of sequences and, therefore, impaired the quality of inventories. Zimmermann et al. (2011) used the V4 region of SSU rDNA, a highly polymorphic region, and proposed this region for barcoding environmental samples. Nevertheless, they pointed out the limits of the V4 region for identifying all cryptic species as we did for our clade analysis. As diatom taxa belonging to the same species can have different sensitivity to environmental conditions (Ivorra et al., 2002; Vanelslander et al., 2009), a marker showing a higher resolving power than SSU rDNA, i.e. intraspecies resolution, could therefore be more useful and, above all, more accurate in the context of biomonitoring programs.

Finally, the inventories obtained from the *rbc*L sequences provided the best estimation of the true composition of the mock community. This marker had the advantage of combining high resolving power with a large DNA reference library. As a result of the polymorphism distribution, the use of a shorter fragment would probably be sufficiently efficient, because almost all reads were informative at both the clade and species levels. Our findings confirmed the proposition of Hamsher et al. (2011), who

recommended the use of a 748-bp fragment *rbc*L-3P as a diatom barcode. However, for routine diatom analyses using pyrosequencing, a shorter fragment (~400 bp in order to remove the fractioning step) would be more appropriate, making it possible to use more reads to construct the taxonomic inventories.

Conclusion

Here, we have presented a preliminary study in which we tested the feasibility and the accuracy of 454 pyrosequencing for the detection of diatom taxa in bulk samples of mock communities. The process tested in this study makes it possible to identify diatom taxa at any taxonomic level. With regard to the DNA reference libraries, resolving power and methodological bias, we demonstrated that 454 pyrosequencing could potentially be used for diatom communities. Currently, *rbc*L appears to offer the best compromise for studying diatom communities using 454 pyrosequencing. Further efforts will be required to design a shorter *rbc*L marker, to optimize laboratory protocols, and to continue to develop the *rbc*L DNA reference library. However, this approach is promising for determining taxonomic inventories of environmental communities of diatoms in the context of biomonitoring programs. *Cox*1 sequences also gave promising results, but these results are limited by its small DNA reference library. Another way to improve diatom identification could also be to design new primers targeting the *cox*1 gene.

The next test step will now be to apply the same sequencing method to “real” environmental communities. The possibility of microscopic identification means that it is possible to compare inventories of diatoms obtained with pyrosequencing data to the usual microscopic determinations of environmental samples. To study such complex samples, a bioinformatic tool able to take into account intraspecies variability will be essential. For this purpose, the *metaMatch* algorithm is currently being enriched by a module that computes the edit genetic distances, where this consists of only a few SNPs and short indels. This tool will make it possible to detect intraspecies variability whilst maintaining accuracy. All these steps should lead to the generation of automatic taxonomic inventories of diatom environmental communities, which will be faster and more reliable, and will therefore participate to the development of a new generation of tools for freshwater biomonitoring, which may be able to meet the increasing demand for ecosystem monitoring.

Table IV.5: *List and proportions (%) of strains used to create the three bulk samples (mix C, mix PCR1, mix PCR2).*

Taxon name	River (Site, Country) of sampling	Mix C	Mix PCR1	Mix PCR2
Amphora montana	Coconi (intermediate, Mayotte)	1,8	1,8	4,8
Cocconeis placentula	Kwalé (upstream, Mayotte)	0,6	0,6	4,8
Cyclotella meneghiniana	Seille (Chambrey, Mainland France)	2,3	2,3	4,8
Fistulifera saprophila	Longoni (downstream, Mayotte)	1,9	1,9	4,8
Fragilaria capucina	Langevin (upstream Grand Galet, La Réunion)	1,7	1,7	4,8
Gomphonema bourbonense	Bouyouni (intermediate, Mayotte)	2,9	0,4	4,8
Gomphonema bourbonense	Dembeni (downstream, Mayotte)		2,9	2,4
Gomphonema clavatum	Saint Denis (upstream, La Réunion)	1,7	1,7	4,8
Gomphonema clevei	Longoni (downstream, Mayotte)	0,5	0,5	4,8
Gomphonema parvulum clade A	Djalimou (downstream, Mayotte)	26,4	5,3	1,0
Gomphonema parvulum clade B	Djalimou (downstream, Mayotte)		5,3	1,0
Gomphonema parvulum clade B	Coconi (downstream, Mayotte)		5,3	1,0
Gomphonema parvulum clade C	Don (Guéméné-Penfao, Mainland France)		5,3	1,0
Gomphonema parvulum clade C	Le Lourdan (Lentigny, Mainland France)		5,3	1,0
Gomphonema pumilum	Bras des étangs (La Réunion)	0,8	0,4	2,4
Mayamaea permitis	Bras Caverne (upstream riv. du Mât, La Réunion)	3,9	3,9	4,8
Navicula cryptocephala	Bouyouni (upstream, Mayotte)	6,5	6,4	4,8
Nitzschia acidoclinata	Bras Caverne (upstream riv. du Mât, La Réunion)	10,9	5,5	2,4
Nitzschia acidoclinata	Bras Caverne (upstream riv. du Mât, La Réunion)		5,5	2,4
Nitzschia cf. frustulum	riv des Galets (Marla, La Réunion)	1,9	1,9	4,8
Nitzschia draveillensis	Pisuerga (Melgar de Fernamental, Spain)	0*	0,1	4,8
Nitzschia inconspicua	Coconi (downstream, Mayotte)	2,6	1,3	2,4
Nitzschia inconspicua	Coconi (downstream, Mayotte)		1,3	2,4
Nitzschia lorenziana	Bouyouni (upstream, Mayotte)	0*	0,1	4,8
Nitzschia palea clade B	Djalimou (downstream, Mayotte)	26,5	8,8	1,6
Nitzschia palea clade G	Sainte Suzanne (Cascade Niagara, La Réunion)		8,8	1,6
Nitzschia palea clade J	Chiers (Longlaville, Mainland France)		8,8	1,6
Pinnularia acrosphaeria	Coconi (intermediate, Mayotte)	1,8	1,8	4,8
Sellaphora seminulum	Dembeni (downstream, Mayotte)	4,6	4,6	4,8
Ulnaria ulna	riv des Galets (Marla, La Réunion)	1,0	1,0	4,8

Materials and methods

Preparation of the bulk samples

This work was based on cultures. Non-axenic strains from the Thonon Culture Collection (TCC, http://www.inra.fr/carrtel-collection) were maintained at 20°C in a growth chamber with a 15/9 light/dark cycle. Thirty strains corresponding to 21 species (Table IV.5) were used to simulate an environmental community. Two of these species were represented by strains from various identified clades (Table IV.5). Three bulk samples were made from these strains. In order to evaluate potential bias due to the DNA extraction and PCR amplification steps, one sample was prepared prior to these steps, while the other two were prepared afterwards. The bulk sample mix C was made directly from a balanced mix of the 30 cultures that were harvested by centrifuging, prior to the extraction and amplification steps. The other two bulk samples were constituted after the extraction and amplification steps. Each of the 30 cultures was independently centrifuged and submitted separately to DNA extraction and PCR amplification. We then prepared a mix of the 30 PCR products. In the PCR1 mix, PCR products were mixed in varying proportions in order to find out whether DNA present in small proportions could be detected by pyrosequencing, whereas in the PCR2 mix, PCR products were mixed in equal proportions. No mix was created between the extraction and amplification steps, because our strains were non-axenic cultures, which made it impossible to achieve an accurate evaluation of diatom DNA quantities. Consequently, in this study we could not distinguish between bias due to DNA extraction and that due to PCR amplification.

Microscopy

The real composition of mix C was checked microscopically. A sub-sample was cleaned with 30% nitric acid. The frustules were then rinsed with demineralized water, mounted in synthetic resin (Naphrax©), and slides were prepared for light microscopy. A minimum of 450 valves were identified on two different slides using a Zeiss Axio Imager A1 microscope (x100 immersion) to determine the real proportions of all the species in this bulk sample (Table IV.5).

DNA methods

The nucleic acid extractions were slightly modified from the method of Bornet et al. (2004). This method was chosen among four tested methods as that most efficient for DNA extraction from samples of diatom strains without any loss of diversity (observed

with DGGE). The three markers amplified were those most commonly used for diatom identification: SSU rDNA, partial *rbc*L, and *cox*1. PCR were performed using 1F and 1528R (Medlin et al., 1988) for SSU rDNA, DP*rbc*L1 and DP*rbc*L7 (Daugbjerg & Andersen, 1997b) for *rbc*L and PC1, and pB1 (Ehara et al., 2000) for *cox*1 according to the authors' specifications. Primers were chosen in order to obtain fragment sizes compatible with the following protocol. Preliminary tests were performed to define the cycle number of each reaction in order to avoid reaching the amplification asymptote. Amplification was processed for mix C and for each culture separately, and each of the PCR products was purified using the QIAquick PCR Purification Kit (Qiagen) according to the Manufacturer's specifications, and quantified using a Nanodrop 1000 (Thermo Scientific). The PCR1 and PCR2 mixes of the bulk samples were then constituted from the PCR products of each culture pooled according to the proportions shown in Table IV.5 for each marker. The PCR products of the three markers were then pooled in equimolar concentrations for each of the three bulk samples. After DNA fractioning, the three samples were tagged to make it possible to assign sequences to the respective samples, and were pyrosequenced on $3/8^{th}$ of a run at the Genomic Platform of Génopole Toulouse/Midi-Pyrénées on a GS FLX Titanium PicoTiterPlate 454 (Roche) according to the Manufacturer's instructions.

DNA reference libraries

DNA reference libraries were built so that diatom sequences could be linked to their taxonomic identification. For the three markers used in this study, three libraries were made of collected diatom sequences obtained both from our laboratory and from GenBank. The diatom strains from the TCC had previously been identified to species level on the basis of morphological criteria, and characterized genetically by Sanger sequencing of the three markers. Sequences from GenBank were curated by eliminating poor quality sequences, environmental sequences, and supposed misidentified sequences. Our final DNA reference libraries comprised 1412 SSU rDNA sequences (303 from TCC), 1071 *rbc*L sequences (357 from TCC), and 266 *cox*1 sequences (96 from TCC), corresponding to 508, 407 and 63 species, respectively. Unfortunately, due to the poor amplification efficiency of the *cox*1 primers, we were not able to obtain the *cox*1 sequences corresponding to ten strains from the bulk samples, which were therefore not included in the corresponding reference library. The three DNA reference libraries were also aligned, and we removed shorter sequences and extremities to obtain same sequence

lengths in order to calculate the Shannon diversity index for each column of the three alignments.

Read assignment

Reads were sorted into libraries according to their sample tag sequences. Reads were removed if they were too short, contained an ambiguous base, or showed low complexity using PyroCleaner software version 1.0 (Mariette et al., 2011). For this "proof of concept" study intended to test the ability of an NGS approach to construct taxa inventories, we did not want to maximize the number of reads that could be used. On the contrary, we wanted to take advantage of the sequencing depth to select only un-ambiguous reads so as to avoid the presence of chimeras, and the taxonomic difficulties inherent in diatom taxonomy. Indeed, on our samples, an assignment tested with a threshold less than 100% had been shown to generate additional bias (data not shown). As our bulk samples were perfectly controlled, and the exact sequences characterizing them were included in the libraries, we used a fast and simple assignation method involving no heuristic and based on 100% homology. Each read library was therefore compared to the three DNA reference libraries using the *metaMatch_PS* and *metaMatch_inv* algorithms (Kermarrec et al., technical report). Only reads displaying a perfect match (100% sequence identity) with at least one sequence from one DNA library were selected.

Presence/absence inventories

In order to compile species inventories, we considered that a read was "informative" for a given taxonomic level if it matched one or several sequences of a single taxon. Among the set of reads with perfect matches, we selected only these "informative" reads to determine a presence/absence inventory for each sample and for each marker. For this study we focused on presence/absence inventories, and abundances were not used; this was to avoid the bias related to relative quantification identified by Amend et al. (2010). To compare the inventories with the real diatom composition, pairwise Jaccard distance matrices were computed on inventories, and hierarchical aggregative clustering was applied to these matrices using PAST version 2.06 (Hammer et al., 2001).

Acknowledgements

All authors acknowledge network R-Syst for developing collaborations and discussions. Pyrosequencing and the production of reads were carried out at the Genotoul

facility by L. K. with help from Eugénie Robe and Olivier Bouchez. We would also like to thank Corinne Cruaud and Arnaud Couloux from Genoscope for Sanger sequencing. Monika Ghosh is acknowledged for improving the English version of the manuscript.

Funding

This work was funded by ”projet innovant CODAL” (awarded to A.B. & A.F.) by the Ecology of Forests, Grassland and Freshwaters Division of INRA, and “projet innovant” (awarded to A.F.) by the Plant Health and Environment Division of INRA. Network R-Syst is supported by both Divisions. Some sequences were obtained by the project @SPEED-ID “accurate SPEciEs Delimitation and Identification of eukaryotic biodiversity using DNA markers” proposed by F-Bol, the French Barcode of life initiative.

4. Test de la nouvelle approche de bioindication basée sur les NGS sur des échantillons naturels de diatomées

4.1. *Présentation générale de l'étude et synthèse des principaux résultats.*

Nous avons démontré dans le paragraphe précédent, que l'utilisation du pyroséquençage 454 était possible sur des échantillons artificiels. Cette étape était particulièrement importante pour tester les capacités de notre méthode et déterminer ses biais. Cependant, l'objectif de l'outil moléculaire est de permettre une étude des communautés de diatomées des échantillons naturels collectés *in situ*. Après cette étape de test sur des communautés « maîtrisées », il était donc nécessaire de valider la méthode en « conditions réelles », c'est-à-dire en analysant des échantillons naturels, collectés sur le terrain, comme pour la bioindication. Ainsi, l'objectif de cette étude était de tester la méthode présentée précédemment pour déterminer les inventaires de diatomées dans des échantillons naturels. Grâce à la possibilité de réaliser en parallèle un inventaire basé sur la morphologie des frustules, nous disposions d'un moyen de valider les inventaires de la nouvelle approche proposée ici.

Nous avons choisi quatre échantillons très différents : deux échantillons provenant de France métropolitaine et deux échantillons provenant d'îles tropicales (Mayotte et La Réunion). Pour chaque zone géographique, un échantillon présentait une faible richesse spécifique et l'autre une forte richesse, d'après l'inventaire effectué en microscopie.

Jusqu'à présent nous avons utilisé une méthode de comparaison des reads aux séquences de référence utilisant un seuil de 100% d'identité pour l'assignation des reads, dans le contexte de nos communautés artificielles. Ce seuil nécessite une optimisation pour l'analyse d'échantillons naturels dont la composition est inconnue. La première étape de cette étude était donc de déterminer les seuils d'identité à appliquer pour analyser des échantillons environnementaux. Travailler à un plus faible niveau de similarité permet d'utiliser plus de reads pour réaliser les inventaires, car un nom peut ainsi être assigné aux reads présentant quelques erreurs de séquençage ou une variabilité intraspécifique.

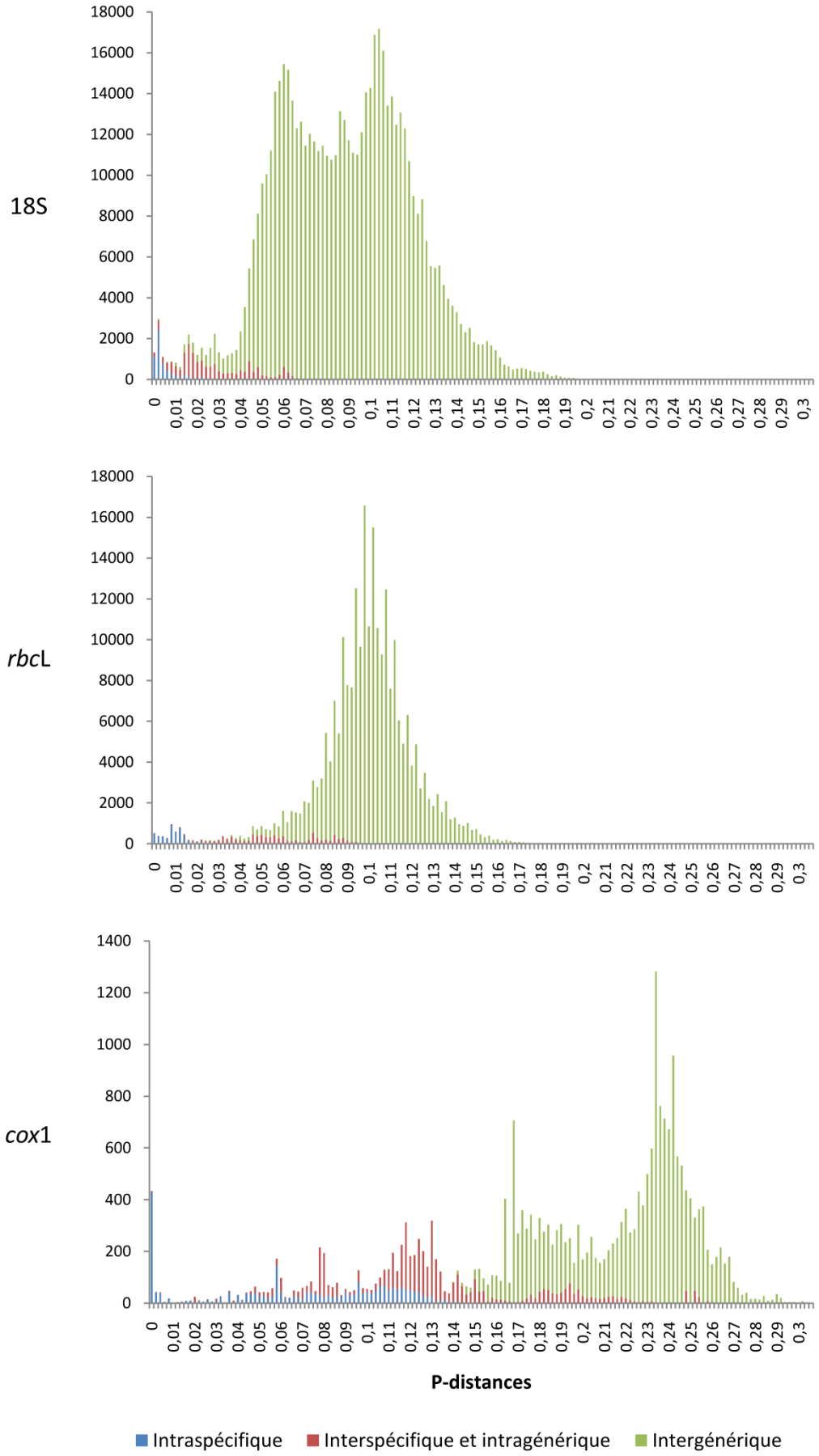

18S
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
0 0,01 0,02 0,03 0,04 0,05 0,06 0,07 0,08 0,09 0,1 0,11 0,12 0,13 0,14 0,15 0,16 0,17 0,18 0,19 0,2 0,21 0,22 0,23 0,24 0,25 0,26 0,27 0,28 0,29 0,3
rbcL
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
0 0,01 0,02 0,03 0,04 0,05 0,06 0,07 0,08 0,09 0,1 0,11 0,12 0,13 0,14 0,15 0,16 0,17 0,18 0,19 0,2 0,21 0,22 0,23 0,24 0,25 0,26 0,27 0,28 0,29 0,3
cox1
1400
1200
1000
800
600
400
200
0
0 0,01 0,02 0,03 0,04 0,05 0,06 0,07 0,08 0,09 0,1 0,11 0,12 0,13 0,14 0,15 0,16 0,17 0,18 0,19 0,2 0,21 0,22 0,23 0,24 0,25 0,26 0,27 0,28 0,29 0,3
P-distances
Intraspécifique
Interspécifique et intragénérique
Intergénérique

Figure IV.7 : *Distribution des distances génétiques calculées en comparant toutes les séquences de chaque base de référence deux à deux (ADNr 18S,* rbc*L et* cox*1) en fonction de trois niveaux taxonomiques (intraspécifique, interspécifique / intragénérique, et intergénérique)*

En effet, les différents spécimens d'une même espèce de diatomées présentent une variabilité intraspécifique (Moniz & Kaczmarska, 2010; Zimmermann et al., 2011). Or, la totalité de cette variabilité intraspécifique ne peut être représentée dans la base de référence. Les trois marqueurs testés précédemment ont à nouveau été utilisés: l'ADNr SSU, le *rbc*L et le *cox*1. Les bases de référence des trois marqueurs ont été alignées et des distances génétiques (p-distances) ont été calculées en comparant toutes leurs séquences deux à deux. La distribution des distances génétiques en fonction de différents niveaux taxonomiques, a été observée pour nous aider à définir des seuils de distances interspécifiques et intergénériques (Figure IV.7). Aucun « barcoding gap » net n'étant apparu, plusieurs seuils ont été testés sur notre mélange de cultures et nous avons comparé les inventaires obtenus par *metaMatch*_IM (inventaires « all », CHAPITRE II.3.2.4) avec la composition réelle de cette communauté artificielle. En parallèle, nous avons suivi le nombre de reads informatifs utilisés pour les inventaires. Le choix du seuil d'identité a été basé sur un compromis entre la précision de l'identification et la quantité d'information (nombre de reads informatifs) utilisée pour créer l'inventaire. Pour *cox*1, nous avons observé que les différents seuils testés ne pouvaient être efficacement comparés à cause de la faible représentativité de sa base de référence. Nous avons donc cessé d'utiliser ce marqueur pour ce test de méthode. Concernant l'ADNr 18S, un seuil de 99% d'identité a été choisi pour les deux niveaux taxonomiques (espèce et genre) car ce seuil permettait la meilleure estimation de la composition de la communauté artificielle. Enfin, des seuils de 99% d'identité pour les espèces et de 98% pour les genres permettaient une bonne utilisation du gène *rbc*L pour déterminer la composition de notre échantillon artificiel tout en conservant un nombre suffisant de reads.

La seconde étape de l'étude était de tester la reproductibilité de la méthode bien que celle-ci ait préalablement été démontrée par Porazinska et al. (2010). Nous avons testé notre méthode et les seuils définis précédemment sur deux échantillons d'une même communauté naturelle pour comparer la reproductibilité de chacun de nos marqueurs. Aucune différence significative n'a été détectée ($p < 0.05$, P-test, Martin, 2002) entre les deux sous-échantillons pour les deux marqueurs. Cependant, nous avons constaté que le marqueur *rbc*L était plus reproductible que l'ADNr 18S au niveau de l'espèce et au

niveau du genre. Les distances entre les inventaires des deux sous-échantillons étaient dues à des détections différentielles d'espèces faiblement représentées. La plus faible reproductibilité du marqueur 18S pourrait être liée au plus faible nombre de reads informatifs utilisés pour réaliser les inventaires.

Enfin, nous avons utilisé notre méthode pour analyser les échantillons naturels utilisés en bioindication (CHAPITRE II.3.2.2(b)). Les deux marqueurs ont montré des différences significatives entre les compositions phylogénétiques des différents échantillons (hormis entre les deux échantillons tropicaux analysés avec l'ADNr 18S). L'ADNr 18S permettait une bonne différenciation entre les échantillons tropicaux et les échantillons tempérés alors que le *rbc*L apportait une différenciation plus précise ne révélant pas l'origine géographique des communautés.

Nous avons ensuite comparé les inventaires réalisés aux listes floristiques obtenues par microscopie. Quel que soit le marqueur et l'échantillon, des différences ont été observées entre les inventaires sur données morphologiques et sur données moléculaires. Pour chaque échantillon, tous les inventaires présentaient des taxa communs, des taxa uniquement détectés par la morphologie (que nous appellerons des positifs morphologiques), et des taxa uniquement détectés par le pyroséquençage (que nous appellerons des positifs moléculaires). Cependant, pour chaque échantillon, les trois inventaires (morphologique, 18S et *rbc*L) étaient corrélés au niveau de l'espèce (corrélation de Spearman, $p<0,05$), exceptés les inventaires d'un échantillon (COIN).

La présence de positifs morphologiques au niveau de l'espèce était en partie expliquée par l'absence de certaines séquences dans la base de référence. En effet, les espèces non représentées dans notre base de référence ne pouvaient pas être détectées par notre outil. Cette limite était amplifiée pour les échantillons d'origine tropicale, dont les diatomées ont été moins étudiées que les diatomées du continent européen. Néanmoins, même en se focalisant uniquement sur les espèces présentes dans nos bases de référence, des divergences d'inventaires subsistaient. Ces positifs morphologiques peuvent être dus à la persistance des frustules après la mort des spécimens. En effet, les frustules de diatomées sont si bien préservés au cours du temps qu'ils sont couramment utilisés dans les études de paléolimnologie (Millet et al., 2010; Konfirst et al., 2011). Ainsi, il est possible que des frustules de spécimens morts soient présents dans les échantillons benthiques alors même que leur ADN est totalement dégradé.

Concernant les positifs moléculaires observés, ceux-ci peuvent être expliqués par une richesse taxonomique (estimée par l'indicateur Chao1) non détectée par le comptage en microscopie. En effet, les divergences entre les inventaires moléculaires et morphologiques étaient plus importantes pour les deux échantillons présentant une forte richesse taxonomique que pour les deux échantillons faiblement diversifiés. De plus, comme pour les positifs morphologiques, les positifs moléculaires peuvent aussi être expliqués par la persistance d'ADN dans les échantillons après la disparition des spécimens (Dejean et al., 2011). Enfin, bien que notre sélection de reads soit stringente, nous ne pouvons pas exclure, que des reads contenant des erreurs de séquençage puissent correspondre à des séquences de référence, et puissent ainsi apparaître dans les inventaires alors que ces taxa ne sont pas présents dans l'échantillon.

Nous avons obtenu des résultats similaires au niveau des genres et les mêmes explications peuvent s'appliquer aux divergences d'inventaires à ce niveau taxonomique. Cependant, en raison de la meilleure représentativité des genres dans la base de référence, nous nous attendions à obtenir de meilleures correspondances entre les inventaires morphologiques et moléculaires au niveau du genre. Il est possible que ces divergences soient également expliquées par la difficulté de définir un seuil d'identité entre genres. En effet, contrairement aux espèces, les définitions de genres sont plus diverses, ce qui gêne la définition d'un bon seuil d'identité intergénérique. Pour améliorer les analyses au niveau du genre, il serait sûrement plus utile de modifier notre sélection de reads informatifs.

4.2. *Article V : en préparation*

New biomonitoring approaches for diatoms: Applications on benthic freshwater samples.

Lenaïg Kermarrec[1,2,3], Alain Franc[4,5], Philippe Chaumeil[4,5], Frédéric Rimet[2,3], Jean-François Humbert[6] and Agnès Bouchez[2,3]

[1] *Asconit Consultants, 3 bd Clairfont, F-66350 Toulouges, France*

[2] *INRA, UMR CARRTEL, 75 av. de Corzent, BP 511, F-74203 Thonon-les-Bains cedex, France*

[3] *Université de Grenoble, UMR CARRTEL, 75 av. de Corzent, BP 511, F-74203 Thonon-les-Bains cedex, France*

[4] *INRA, UMR BIOGECO, 69 route d'Arcachon, F-33612 Cestas cedex, France*

[5] *University of Bordeaux 1, UMR BioGeCo, 33400 Talence, France*

[6] *INRA, UMR BIOEMCO, site de l'ENS, 46 rue d'Ulm, F-75005 Paris, France*

Corresponding author : Agnès Bouchez : agnes.bouchez@thonon.inra.fr

Abstract

In order to evaluate human influences and impacts of global changes on natural ecosystems, several biomonitoring programs have been implemented. Diatoms are among the main bioindicators used to assess the ecological quality of rivers. However, as for many bioindicators, the taxonomic estimation of their communities is difficult and time-consuming. As high-throughput sequencing has proved its suitability to study microorganism community, we tested the reliability of the 454 pyrosequencing to determine diatom inventories in environmental samples in the framework of bioindication. Three markers were used: the SSU rDNA, the *rbc*L and the *cox*1. First, we studied reference libraries of the three markers to help defining thresholds between intraspecific and interspecific, and between intrageneric and intergeneric genetic distances. Thresholds were tested on a mock community and applied to five natural samples (among which one in duplicate) to assign taxa names to environmental sequences. For each sample, inventories obtained from pyrosequencing were compared to the floristic list obtained by microscopy approach. Divergences were observed between morphological and molecular inventories. Some limits linked to the molecular method could be overcome extending our DNA reference libraries and others are not more important than the bias of the current method based on the morphology. In conclusion, 454 pyrosequencing could be used in biomonitoring programs after some optimizations.

Introduction

The assessment and monitoring of water resources is a major concern for environmental agencies in charge to evaluate the health of these ecosystems, which are submitted to increasing local and global anthropogenic pressures (Meybeck, 2003; Foley et al., 2005). Among all the biological indicators used for that, diatoms are of first interest because they are directly impacted by changes occurring in physico-chemical parameters and consequently provide an integrated reflection of water quality (Van Dam et al., 1994; Kovács et al., 2006). The main problems linked to the use of diatoms for bioindication are that their identification and enumeration by microscopic examination are time-consuming and require a high level of taxonomic expertise. Moreover, the discrimination of taxa, which are closely related, is very difficult and can lead to misidentifications that degrade the accuracy of diatom index (Besse-Lototskaya et al., 2006).

In order to avoid these problems, alternative approaches based on the use of molecular approaches using DNA sequences have been developed in the recent years and among them, environmental barcoding is considered as one of the most promising tools for biomonitoring applications. For diatom communities, several sequences belonging to the nuclear, mitochondrial and chloroplast genomes have already shown their potential as barcodes (Evans et al., 2007; Moniz & Kaczmarska, 2010; Hamsher et al., 2011; Zimmermann et al., 2011). But until yet, the cost of the Sanger sequencing, which restricts the number of sequences available per sample, was one of the main limits in the routine use of diatom barcodes for the assessment of water quality.

The development of Next Generation Sequencing (NGS) methods recently opens a new area in the use of barcoding approaches. Indeed, these methods allow obtaining a very large quantity of data per sequencing run and consequently a deep description of the microbial communities, including the rare biosphere. The potential of these methods, in particular of 454 pyrosequencing, for barcoding approaches in the framework of bioindication, has been recently highlighted for macroinvertebrates communities (Hajibabaei et al., 2011) and for diatoms (Kermarrec et al. submitted). In this last paper, the feasibility of the 454 pyrosequencing to determine the composition of diatom communities was tested on mock communities by using three markers: the SSU rDNA for which a large reference database is available, the *rbc*L gene which has proved its capacities to discriminate diatom taxa (Hamsher et al., 2011; Stoof-Leichsenring et al., 2012) and the *cox*1 which is the first barcode proposed by Hebert et al. (2003).

Only reads displaying 100% identity with reference libraries were used to study the methodological bias without taking into account the taxonomic uncertainties. In these artificial conditions, pyrosequencing appeared sufficiently accurate to determine diatom inventories and *rbc*L seemed to be the best available nucleic marker.

In the present work, reference databases of the same three markers were studied to help defining thresholds between intraspecific and interspecific, and between intrageneric and intergeneric genetic distances. Thresholds were tested on a mock community and the selected thresholds were applied to natural samples. Reads from 454 pyrosequencing were compared to reference databases and a species name was assigned to each read using *metaMatch*_IM algorithms designed for this study. The inventories obtained for each marker were compared to the floristic list obtained by microscopy approach (as required for biomonitoring programs) to test the reliability of this molecular tool on natural samples.

Material and methods

Samples and microscopy

Four sampling locations were selected for this study (Table IV.6). Benthic samples were collected by scraping material from the surface of stones in four rivers: two (LACL and CHLO) in Mainland France, and two (COIN and BDNA) in tropical islands of Indian Ocean (Mayotte and La Réunion respectively). These samples were chosen because diatom taxa from European continent have been studied since many years while tropical communities are less known. For each geographic origin (temperate and tropical), we expected to have one sample with low diversity (CHLO and BDNA) and one sample with high diversity (LACL and COIN) as diatom communities (identification and counting) of these samples was previously used in biomonitoring programs. An additional sample was produced by mixing 30 diatom strains corresponding to 21 species to create a mock community (Table IV.7). A sub-sample was cleaned with 30% nitric acid. The frustules were then rinsed with demineralized water, dried and mounted with synthetic resin (Naphrax©) and slides were prepared for light microscopy. A minimum of 400 valves was counted to determine floristic list of the mock community as it has been done for environmental samples.

Table IV.6: *Environmental DNA samples used for the study and number of reads obtained after removing low-quality sequences*.

Sample ID	River (Site, Country) of sampling	Number of reads
LACL replicat 1	Lay (La Clay, Mainland France)	91,284
LACL replicat 2	Lay (La Clay, Mainland France)	165,89
CHLO	Chiers (Longlaville, Mainland France)	179,34
COIN	Coconi (intermediate, Mayotte)	207,632
BDNA	Saint Denis (upstream, La Réunion)	127,648
Artificial sample	/	115,269

Table IV.7: *List and proportion of strains used to create mock community. Proportions were determined by light microscopy*.

Taxon name	Mix C
Amphora montana	1,8
Cocconeis placentula	0,6
Cyclotella meneghiniana	2,3
Fistulifera saprophila	1,9
Fragilaria capucina	1,7
Gomphonema bourbonense	2,9
Gomphonema clavatum	1,7
Gomphonema clevei	0,5
Gomphonema parvulum	26,4
Gomphonema pumilum	0,8
Mayamaea permitis	3,9
Navicula cryptocephala	6,5
Nitzschia acidoclinata	10,9
Nitzschia cf. frustulum	1,9
Nitzschia draveillensis	0*
Nitzschia inconspicua	2,6
Nitzschia lorenziana	0*
Nitzschia palea	26,5
Pinnularia acrosphaeria	1,8
Sellaphora seminulum	4,6
Ulnaria ulna	1,0

*Strains used to create the pooled-culture mix, but not detected by microscopy.

DNA reference libraries

Reference libraries were made of collected diatom sequences from our laboratory and from GenBank. Diatom strains from the Thonon Culture Collection (TCC, http://www.inra.fr/carrtel-collection) had previously been identified at the species level on the basis of morphological criteria, and characterized genetically by Sanger sequencing of several markers. Sequences from GenBank were curated by removing poor quality sequences, environmental sequences and supposed misidentified sequences. Our final DNA reference libraries comprised 1412 SSU rDNA sequences, 1071 *rbc*L sequences and 266 *cox*1 sequences, corresponding to 508, 407, and 63 species respectively.

For each library, reference sequences were also aligned and we removed shorter sequences and extremities to obtain same sequence lengths in order to calculate uncorrected pairwise distances (p-distances). P-distances were computed using MEGA5 (Tamura et al., 2011) on final alignments datasets containing 1086 sequences of SSU rDNA, 704 sequences of *rbc*L and 229 sequences of *cox*1. Intraspecific, interspecific/intrageneric and intergeneric genetic distances and their relative distributions were plotted using PAST (version 2.06, Hammer et al., 2001) in order to help us define thresholds of interspecific and intergeneric distances.

DNA methods

The nucleic acid extractions were slightly modified from the method of Bornet et al. (2004). PCR were performed on several markers commonly used for diatom identification: SSU rDNA using 1F and 1528R (Medlin et al., 1988), partial *rbc*L using DP*rbc*L1 and DP*rbc*L7 (Daugbjerg & Andersen, 1997b) and partial *cox*1 using PC1 and pB1 (Ehara et al., 2000) according to the author's specifications. PCR products of each marker were purified separately using the QIAquick PCR Purification Kit (Qiagen) according to the manufacturer specifications, and quantified using a Nanodrop 1000 (Thermo Scientific). The three markers were then pooled into equimolar concentrations. After fractioning, PCR products from each sample were tagged and pyrosequenced at the Genomic Platform of Génopole Toulouse/Midi-Pyrénées on a GS FLX Titanium PicoTiterPlate 454 (Roche) according to the Manufacturer's instructions.

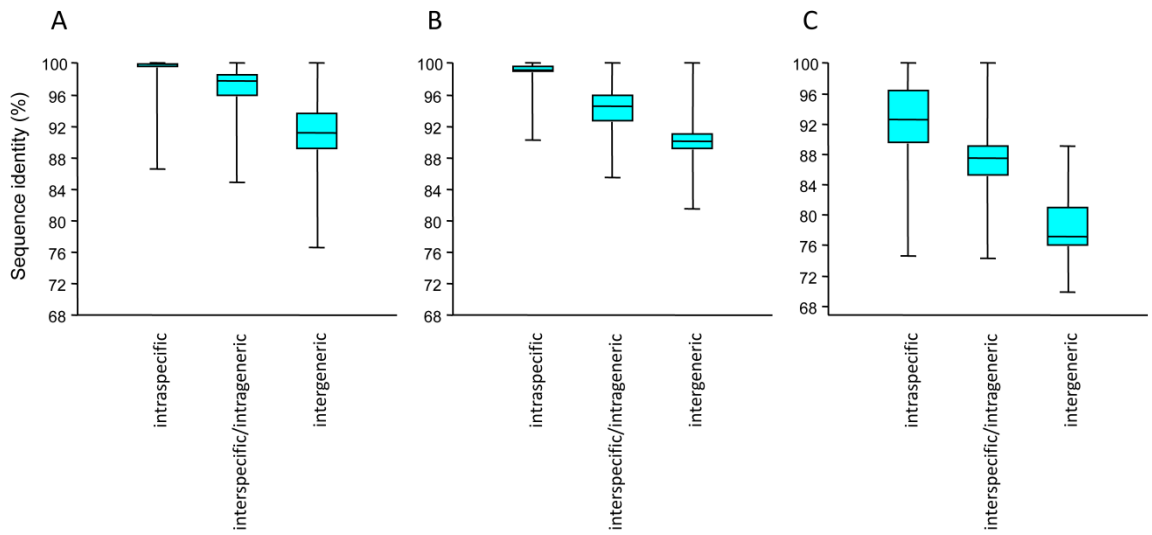


Figure IV.8: *Box plots of intraspecific, interspecific/intrageneric and intergeneric genetic idendity based on uncorrected p-distances*.
A: SSU rDNA; B: *rbc*L; C: *cox*1.

Read assignment and inventories

Sequence data were sorted into libraries according to sample tag sequences. Reads were removed when they were too short, contained ambiguous bases, or showed low complexity using PyroCleaner software version 1.0 (Mariette et al., 2011). Each library was compared to the three reference databases using *metaMatch*_IM algorithms with an identity threshold. To create inventories, a taxon name is assigned to a read only if all the selected reference sequences (that is to say having a homology greater than the threshold) belong to the same taxon. If different taxon names appear in the list created by *metaMatch*_IM, the read was not used in the inventory. Then, we removed singletons from the inventories as there was demonstrated that most of them were artifactual (Tedersoo et al., 2010; Behnke et al., 2011). As quantitative assessment of species composition provides supplementary bias due to the variable copy number of marker, we have focused on inventories with presence/absence for the definition of threshold in order to define the more accurate threshold. Although imperfect, we used the relative quantification for environmental samples, since this parameter is used in bioindication. The presence/absence species inventories from pyrosequencing were compared among them or to morphological data, computing pairwise Jaccard distance matrices using PAST (version 2.06, Hammer et al., 2001) whereas species inventories comprising relative abundances were compared using Bray-Curtis distance matrices. A non-metric multidimensional scaling (MDS) was used to compare inventories of environmental samples as this method place data points in two dimensional system such that the ranked differences (based on Bray-Curtis distances) between inventories were preserved. For each sample, molecular end morphological inventories have been also compared using Spearman correlation to notice significant correlation between them with PAST (version 2.06, Hammer et al., 2001). As absence of singletons in molecular inventories prevents richness analysis, we removed one read per species as previously done by Unterseher et al. (2011) to obtain inventories usable for richness estimation. Then, the estimation of species richness (Chao1) for morphological and molecular data was performed using "vegan" R package (Oksanen et al., 2011). Moreover we compared common taxa (i.e. taxa detected by both approaches), molecular-detected taxa (i.e. species not detected by morphological counts but detected by pyrosequencing analysis) and morphology-detected taxa (i.e. species detected by morphological counts but not by pyrosequencing analysis) for each sample.

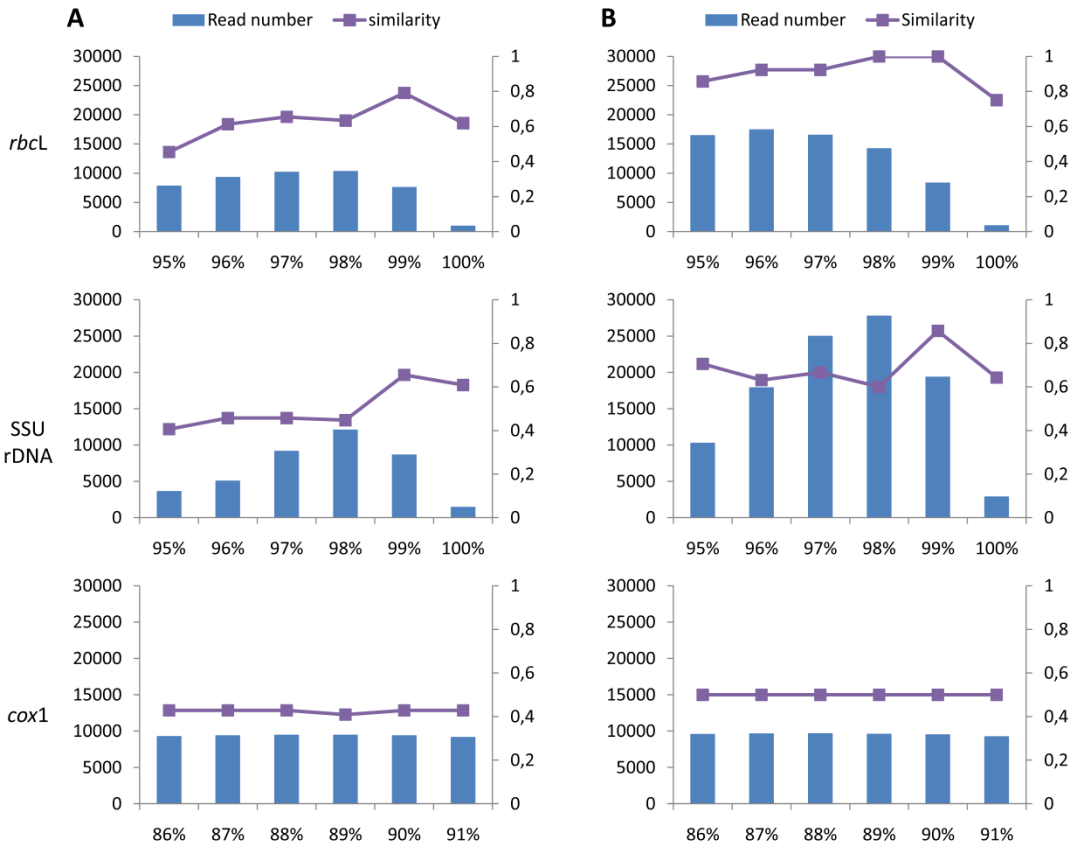


Figure IV.9: *Similarities of inventory with actual composition of the mock community and number of informative reads used to compile inventories depending on different identity thresholds for each marker and each taxonomic level (A: species; B: genus).*
Similarities were based on Jaccard distances between inventory and the real composition of the artificial sample.

At the same time, for each marker, we analyzed the phylogenetic information of samples using Fast Unifrac Web interface (Hamady et al., 2010) using a reference tree based on sequences from reference libraries. We compared samples using P-test (Martin, 2002) and using a hierarchical clustering analysis (UPGMA) based on a distance matrix generated by the weighted Unifrac algorithm.

Results

Threshold definition

Uncorrected pairwise distances (p-distances) of the DNA reference libraries were used to plot sequence identity depending on taxonomy (intraspecific, interspecific / intrageneric, and intergeneric genetic distances, Figure IV.8). Frequencies of distances were also plotted to help to define distance threshold (data not shown). Overlaps between taxonomic levels were observed for each marker and no clear barcoding gap both between intra- and interspecific distances and between intra- and intergeneric distances was observed.

A total of 115 269 reads was obtained for the artificial sample after low-quality sequences were removed. Read length were comprised between 131 and 596 bp with an average length of 414 bp. For each marker, we compared the inventories provided by several identity thresholds to the actual composition of the mock community. For each marker and each taxonomic level (genus and species), the similarities based on Jaccard distances between inventory from each threshold and the real composition of the artificial sample were presented in Figure IV.9. We compare the evolution of the similarity and the number of informative reads used to compile inventories depending on the identity threshold. So we chose the identity thresholds providing the best similarity with the actual composition (minimizing false positives and false negatives) and using the largest number of reads. We chose therefore threshold of 99% for SSU rDNA at the two levels (genus and species), 99% and 98% for *rbc*L gene at species and genus level respectively. Concerning *cox*1, similar results (low estimation of communities) were obtained whatever the threshold. No false positive was added, but the same false negatives appeared because these taxa were absent from the DNA reference library. We do not have defined suitable threshold and therefore, we did not analyze this marker for environmental samples.

Table IV.8: *P-values of the P-test comparing phylogenetic information of each pair of sample for both markers (SSU rDNA and* rbc*L).*

SSU rDNA*rbc*L	BDNA	CHLO	COIN	LACL1	LACL2
BDNA		≤ 2.0e-03	≤ 2.0e-03	≤ 2.0e-03	≤ 2.0e-03
CHLO	≤ 2.0e-03		≤ 2.0e-03	≤ 2.0e-03	≤ 2.0e-03
COIN	1.0	≤ 2.0e-03		≤ 2.0e-03	≤ 2.0e-03
LACL1	≤ 2.0e-03	≤ 2.0e-03	≤ 2.0e-03		0,24
LACL2	≤ 2.0e-03	≤ 2.0e-03	≤ 2.0e-03	1.0	

Table IV.9: *Summary of inventory results*.
Number of frustules or informative reads used to compile inventories, number of genus and species identified and richness estimation (Chao1 estimator). A: temperate samples; B: tropical samples.

		LACL				CHLO		
A	Morphology	SSU	*rbc*L	SSU	*rbc*L	Morphology	SSU	*rbc*L
Frustule number	400	/	/	/	/	413	/	/
Number of species informative read	/	125	2850	222	5016	/	1376	25579
Number of genus informative read		842	7230	741	13378	/	2414	47748
Genus number	21	14	14	18	13	16	15	17
Species number	54	16	25	27	25	27	19	24
Chao1 mean	77,7	17	25	28,4	25	34	20,5	25,5

	COIN			BDNA		
B	Morphology	SSU	*rbc*L	Morphology	SSU	*rbc*L
Frustule number	412	/	/	500	/	/
Number of species informative read	/	107	194	/	3818	6862
Number of genus informative read	/	327	513	/	5123	16358
Genus number	22	7	8	12	15	14
Species number	59	8	13	21	25	23
Chao1 mean	64,5	8	13,5	22,7	25,75	23,6

Reproducibility

To test the reproducibility of the methods we used two sub-samples from the same sample (LACL). Different numbers of reads were retrieved for the two samples after removing low quality reads: 91 284 for LACL1 and 165 890 for LACL2 (Table IV.6) including the three markers. Read lengths were comprised between 135 and 611 bp (with an average length of 426 bp), and between 130 and 612 bp (with an average length of 428 bp) for LACL1 and LACL2 respectively.

The P-test comparing phylogenetic information of samples, available on the Fast UniFrac Web interface, showed no significant difference (Table IV.8) between the two sub-samples using both markers (SSU rDNA and *rbc*L).

Then, we compared species inventories, comprising relative proportions of each species, from LACL1 and LACL2. *Rbc*L appeared more reproducible with 94.9% of similarity between the two sub-samples whereas results from SSU rDNA looked more variables with only 71.7% of similarity between the two sub-samples. Species causing differences in inventories were represented by a maximum of 8 or 10 reads for *rbc*L and SSU rDNA respectively. We performed the same analysis on genus inventories and observed similar results with a similarity of 97.6% for *rbc*L and a similarity of 88.2% for SSU rDNA. As for species inventories, some species weakly represented (less than 9 and 6 reads for *rbc*L and SSU rDNA respectively) produced divergences of inventories.

Environmental samples

A total of 91 284, 165 890, 179 340, 207 632 and 127 648 reads for LACL1, LACL2, CHLO, COIN and BDNA respectively for the three markers were obtained by pyrosequencing after low-quality sequences were removed. Read length were comprised between 130 and 625 bp with an average length of 425 bp.

The inventories (species and genus) obtained by the two approaches (morphological investigation and 454 pyrosequencing) were presented in supplementary material. Contrary to the artificial samples, the inventories from environmental samples obtained using SSU rDNA were based on very few informative reads (Table IV.9) due to the non specific primers which targeted eukaryotes whereas *rbc*L primers targeted only heterokont algae.

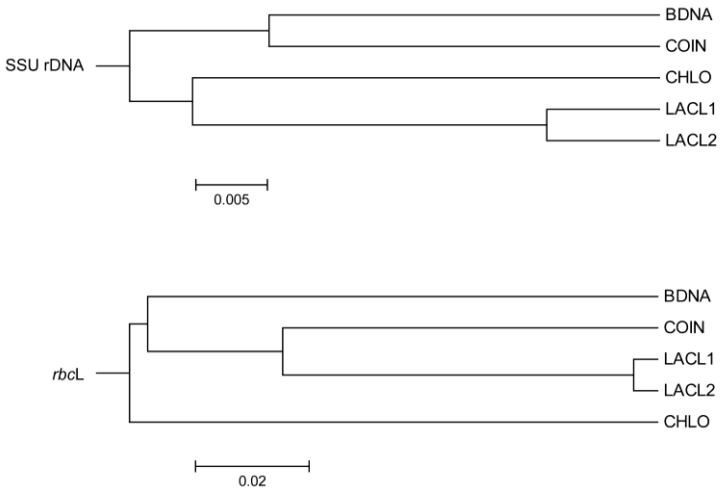


Figure IV.10: *Hierarchical clustering analysis (UPGMA) of the samples based on the phylogenetic position of the SSU rDNA and* rbc*L reads.*
UPGMA was based on a distance matrix generated by the weighted UniFrac algorithm.

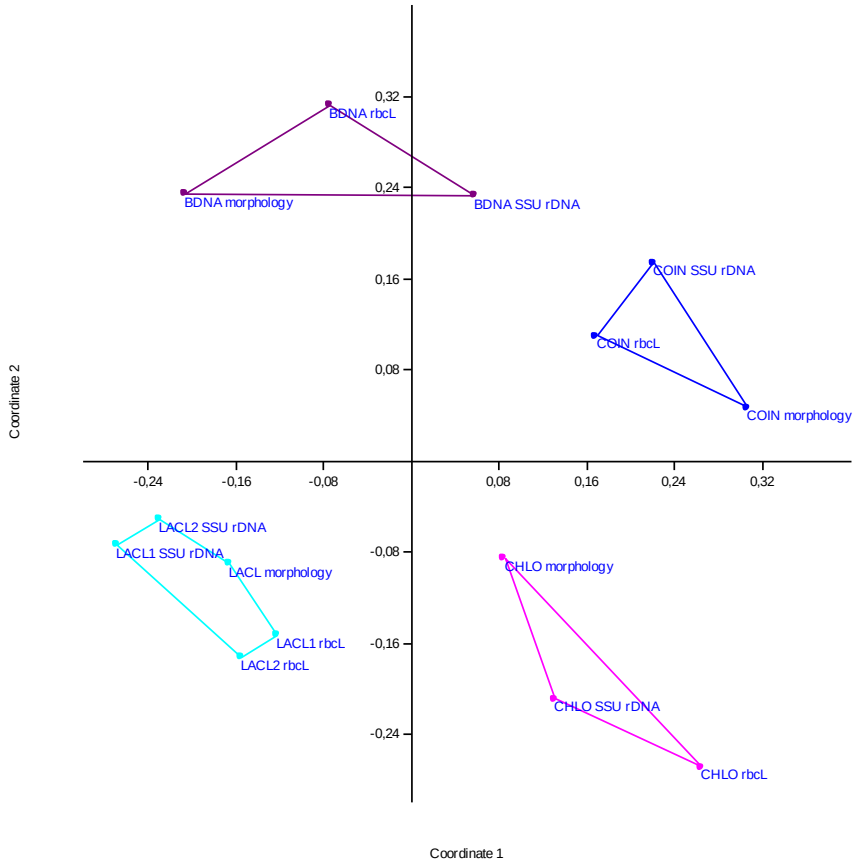


Figure IV.11: *Non metric Multidimensional Scaling based on Bray-Curtis distances comparing molecular and morphological species inventories of the four samples*.

Moreover the low polymorphism of the SSU rDNA marker combined with our stringent selection of reads (informative reads) decreased the number of informative reads that were used. However, phylogenetic composition of the samples showed significant differences with both markers. Only the two tropical samples analyzed using SSU rDNA had no significant difference (Table IV.8). When we compared sample clusters provided by each marker, differences were observed (Figure IV.10). SSU rDNA grouped samples depending on their origin (tropical/temperate) whereas *rbc*L marker did not. Moreover, the differentiation level of samples provided by *rbc*L was higher than differentiation level from SSU rDNA as illustrated by the distance scale.

We compared all inventories using a non metric MDS based on Bray-Curtis distances (Figure IV.11). As expected, whatever the sample, there were differences between morphological inventories and the different molecular inventories. The non metric MDS plots showed that species inventories of each sample could be grouped. These results were confirmed by the Spearman correlation which showed significant correlation between morphological and the two molecular inventories for LACL, CHLO and BDNA samples (p values <0.05) but not for COIN inventories (p values >0.01). The comparison of genus inventories (supplementary material) provided similar results with significant correlation between morphological and the two molecular inventories for the four samples (p values <0.05).

To assess differences between molecular and morphological results, we compared more precisely the inventories at the two taxonomic levels (Table IV.9 and Table IV.10). All molecular inventories presented common taxa with morphological inventories, molecular-detected taxa, and missed some morphology-detected species. Few common species or genera were detected by the two methods (Table IV.9 and Table IV.10). We observed that a low percent of species from morphological inventories were represented in the DNA reference libraries whereas almost all genera are included in the reference libraries (Table IV.10). As expected the percent of species from tropical samples represented in the DNA reference libraries was lower (23.7 to 42.9%) than the species from temperate samples (53.7 to 77.8%). When we only considered species included in the reference libraries, we still observed many differences. Between 18.8% and 42.9% of species detected in morphological counts and represented in the reference libraries, were detected by pyrosequencing approach for high diversity samples (Table IV.10).

Table IV.10: *Comparison of molecular inventories with morphological inventory for each sample*.

Number of common taxa between the molecular and morphological inventories and number of specific species to each inventory are presented. The proportion of species detected by morphological inventories and represented in the reference libraries are shown for each sample. And, proportion of taxa detected by molecular approach (whatever the relative proportion) taking into account only taxa represented in the reference library and, only abundant taxa (>1%) represented in the reference library.

A: taxonomic level: species; B: taxonomic level: genus.

A	LACL				CHLO		COIN		BDNA	
	SSU	*rbc*L	SSU	*rbc*L	SSU	*rbc*L	SSU	*rbc*L	SSU	*rbc*L
Number of common species	7	11	12	11	11	14	3	6	6	6
Molecular-detected species	9	14	15	14	8	10	5	7	19	17
Morphology detected species	47	43	42	43	16	13	56	53	15	15
Proportion of species represented in the reference library (%)	59	54	59	54	78	78	27	24	43	43
Proportion of species detected by molecular method (%)	22	38	38	38	52	67	19	43	67	67
Proportion of abundant species detected by molecular method (%)	60	67	80	67	91	100	25	43	80	80

B	LACL				CHLO		COIN		BDNA	
	SSU	*rbc*L	SSU	*rbc*L	SSU	*rbc*L	SSU	*rbc*L	SSU	*rbc*L
Number of common genera	8	8	11	11	12	13	6	7	8	8
Molecular-detected genera	6	6	7	2	3	4	1	1	7	6
Morphology detected genera	13	13	10	10	4	3	16	15	4	4
Proportion of genera represented in the reference library (%)	95	90	95	90	94	94	82	77	100	100
Proportion of genera detected by molecular method (%)	40	42	55	58	80	87	33	41	67	67
Proportion of abundant genera detected by molecular method (%)	50	50	70	70	100	100	36	56	88	88

These proportions were higher (between 52.4% and 66.7%) for low diversity samples. Interestingly, when we focused on species with abundance greater than 1%, we observed a better agreement between morphological and molecular inventories except for the sample COIN (high diversity tropical sample). Indeed, between 66.7 and 100% of abundant species detected by morphological inventories were detected by our molecular inventories for the LACL, CHLO and BDNA samples.

Species richness estimation (Chao1) indicated that we did not detect the whole diatoms richness using microscopy (Table IV.9). Richness estimators did not allow an accurate evaluation of richnesss from molecular inventories but showed that species richness that can be detected by the molecular method was almost reached.

Concerning genus inventories (Table IV.10), despite a good representativeness of genera in the DNA reference libraries, a low proportion of genera detected by morphology were also detected by pyrosequencing analyses. We observed similar results to species results with few common genera between the morphological and molecular inventories (from 33.3% to 57.9%) for high diversity samples (LACL and COIN) and a best conformity for low diversity samples (> 80% for CHLO and 66.7% for BDNA). Moreover, as for species analyses, this agreement was improved when we focused on abundant genera except for COIN sample.

Discussion

The objectives of this study was to define distance thresholds for several nucleic markers in order to assign species name to diatom sequences and to test the efficiency of this threshold on 454 pyrosequencing data for describing diatom communities in the context of bioindication.

Threshold definition

We determined thresholds to assign a name (genus and species) to SSU rDNA sequences and *rbc*L sequences of diatoms. However, we could not establish such thresholds for sequences of *cox*1 because of the low number of available reference sequences. Speciation is a gradual process for which it is difficult to establish a common limit (identity threshold) for all diatom taxa because diatom lineages did not evolve at the same rate, and close taxa may be more or less differentiated genetically.

For example, it has been shown that the rate of evolution of SSU rDNA is different among diatoms (Kooistra & Medlin, 1996). Thus, whatever the marker, we could not observe a clear barcoding gap for all diatoms. Moreover for SSU rDNA for which polymorphism is variable depending on the region of the marker (Elwood et al., 1985), the definition of threshold cannot be exact for all the regions. This led to the definition of the same average threshold for the two taxonomic levels, which cannot be exact.The use of a unique polymorph region with diatom-specific primers, as used by Zimmermann et al. (2011), would facilitate the threshold definition and the use of more informative reads. Indeed the best threshold is a tradeoff between number of read used to compile inventory and similarity to the actual composition. However, our SSU rDNA and *rbc*L identity thresholds allowed us to obtain an estimation of the composition of our mock and environmental communities and allowed a differentiation of sample communities. Our thresholds correspond to previous studies of diatoms. For example, Moniz & Kaczmarska (2009) found a maximum distance of 1% within species using a ~ 1600 bp of SSU rDNA. Concerning *rbc*L marker, Hamsher et al. (2011) noted between 0.14 and 7.3% sequence divergences between *Sellaphora* species. In the bioindication framework, as there is no perfect threshold, we recommend using an identity threshold of 99% for *rbc*L reads to study freshwater diatom communities at the species level using 454 pyrosequencing.

The determination of a genus identity threshold is all the more difficult that genera are wider taxonomic groups and includes different lineages. Thus, despite a good representativeness in DNA reference libraries, it was difficult to assign a genus name to reads with our stringent read selection. Broad range of intrageneric distances have previously been described (Moniz & Kaczmarska, 2009; Zimmermann et al., 2011). We applied the same selection for species and genera, but our method (filters applied to read selection) was probably not suitable for genus analyzes. Another selection filter would probably be more appropriate to this taxonomic level.

Reproducibility

No significant difference was observed between the two sub-samples with both markers although we demonstrated that the *rbc*L marker seemed to be more reproducible than the SSU rDNA for the species and genus levels. For the two markers, the distances between inventories from the two sub-samples were due to differential detection of rare species. We demonstrated in a previous study that inventories from SSU rDNA sequences were more influenced by sequencing errors than *rbc*L (Kermarrec et al. submitted). This

limit of the use of the whole SSU rDNA could explain the difference of reproducibility between the two markers. As Porazinska et al. (2010) demonstrated that reads numbers of a SSU rDNA fragment were very reproducible, we assume that our results were linked to the fragment used in our study. The use of the whole SSU rDNA with the low specificity of primers provided a low number of informative reads which decreased the reproducibility. We therefore believe that the use of a more appropriate fragment would enhance the reproducibility.

Divergence sources between morphological and molecular data

We demonstrated that molecular and morphological approaches provided divergent inventories even if the different inventories from a sample were correlated. The morphology-detected species may be partly explained by the representativeness of the DNA reference libraries. Indeed a high proportion of species detected by morphological counts was not in the DNA reference libraries and cannot be added in the molecular inventories using our method. This limit was amplified for tropical samples for which the diversity of diatoms is less known and so for which few reference sequences were available. The incompleteness of reference libraries impeded therefore the validity of taxonomic assignment by molecular methods. As already underlined by Mann et al. (2010), to go to molecular studies of diatom communities, we need to increase the recovery of our taxonomic DNA libraries especially if we want to study samples from different geographic origins or from different ecological habitat with very different flora.

Because inventories still differed when we analyzed only species or genera present in our DNA reference libraries, these ones were not the only source of divergences. Some morphology-detected taxa may be due to the persistence of frustules after the presence of the living cells. Indeed, the frustules of diatoms are preserved over time and can be used as paleaoindicators (e.g. Sylvestre, 2002; Laslandes et al., 2006). Thus, in the benthic samples, frustules of old cells could also been preserved and counted during diatom analysis by microscopy. This hypothesis is confirmed by the morphological detection of some planktonic species in our benthic samples, for example three frustules of *Cyclostephanos dubius* (Fricke) Round in LACL sample; one of *Cyclotella meneghiniana* Kützing in CHLO sample and one of *Stephanodiscus hantzschii* Grunow in COIN sample. These planktonic species were represented in our two DNA reference libraries but no DNA was detected. Therefore, these frustules could be artifacts from microscopic observations and not representative of the living benthic community. We demonstrated

that better correlation was observed between the two approaches when we focused on abundant taxa (>1%) of the morphological composition. As planktonic species, some other taxa poorly represented may be not representative of the living community and could explain why DNA was not perceived.

Similarly, the presence of extra-cellular DNA can affect the composition of retrieved diatom community and can explain the divergences between inventories due to molecular-detected taxa. DNA can be recovered in freshwater ecosystems during few days after its release (Dejean et al., 2011) and can therefore compromise the determination of the benthic community. In addition, the molecular-detected species can be explained by the depth of pyrosequencing which allows detecting species at lower abundances. Chao1 indicator revealed that frustule counts did not cover the whole estimated species richness. This hypothesis was confirmed by the differences between the molecular and morphological inventories which were higher for two samples with high taxonomic richness than for the two samples with lower richness.

Finally, the taxonomic uncertainty of some diatom taxa provides divergences in our inventories. For example, in the morphological inventories from CHLO sample, *Achnanthidium saprophilum* (Kobayasi & Mayama) Round & Bukhtiyarova, belonging to *Achnanthidium minutissimum* (Kützing) Czarnecki complex (Potapova & Hamilton, 2007) was identified whereas pyrosequencing detected *A. minutissimum* using the two markers. In the same way, *Amphora libyca* Ehrenberg and *A. ovalis* (Kützing) Kützing were identified by morphological inventories of LACL samples whereas SSU rDNA and *rbc*L pyrosequencing revealed the DNA presence of *A. copulata* (Kützing) Schoeman & Archibald which can be a synonymous of *A. libyca* or of some varieties of *A. ovalis* (Fourtanier & Kociolek, 2011). We assumed that misidentification of some reference sequences or misidentification of sample may be the source of these differences. But it is also possible that our thresholds were not suitable to differentiate these sister species or that these species were different morphs of the same species.

Consequences on bioindication purpose

We demonstrated that the molecular method was more congruent with morphological inventories when we focused on abundant species. PCR-based approaches presented bias which lead to an underestimation of the rare taxa (Gonzalez et al., 2012). However, the bioindication is different of traditional ecosystem functioning studies. Indeed, the rare

taxa that can have important ecological function are not used in biomonitoring programs because they are not considered as representative of the environment. For example, in the calculation of the Biological Diatom Index, only taxa higher than 7.5 ‰ are considered in order to avoid the ecologically insignificant specimen (Prygiel et al., 2002). Therefore, the limits due to PCR and sequencing bias should not prevent their use in bioindication. On the other hand, the persistence of DNA should not cause difficulties as its persistence is lesser than those of frustules. Moreover, targeting RNA instead of DNA would allow targeting living cells, as RNA is more subjected to degradation than DNA, and excluding dormant cells which are not active cells. Finally, the taxonomic uncertainties damaged molecular inventories but also the bioindication based on morphology (Besse-Lototskaya et al., 2006).

We demonstrated that the SSU rDNA marker differentiated samples at a larger scale than *rbc*L. The SSU rDNA would be more appropriate to compare samples at a high level such as comparing tropical / temperate communities whereas phylogenetic information from *rbc*L differentiates samples from the same area. We confirmed therefore our previous results (Kermarrec et al, submitted) that, in the framework of bioindication, the whole SSU rDNA is less suitable for diatom community pyrosequencing than *rbc*L marker. Indeed, the indices based on bioindicator communities are adapted to one country and results are compared with a reference sample at a local geographic scale.

Overall, 454 pyrosequencing of *rbc*L marker could be used in bioindication, even if it needs several optimizations of laboratory protocols and reference libraries.

Conclusion

We defined thresholds to names environmental sequences of diatoms. In addition to our molecular method, these thresholds will be used to group OTUs to determine the unreferenced richness and will provides many data on diatoms diversity. Many improvements still need to be made before substituting the morphological method by the pyrosequencing method. However, we demonstrated that 454 pyrosequencing is a tool suitable for the study of diatom communities in the context of bioindication using *rbc*L marker, or for global studies using SSU rDNA. As the variable copy number of SSU rDNA and *rbc*L among diatom taxa currently bother the quantification of relative proportion of diatom taxa, new improvements would be necessary. The relative

quantification could be enhancing applying correction factor to overcome variable copies of markers. The chloroplast number could be used to improve quantification using *rbc*L. And, concerning SSU rDNA, correction factors depending on cell sizes would be also tested as rDNA copy number seemed to be correlated with size in picoeukaryotes (Zhu et al., 2005) or with biovolumes of cells for diatoms and dinoflagellates (Godhe et al., 2008),.

Acknowledgements

We thank Florence Peres, Maurice Bey, Michel Coste, Gilles Gassioles, Didier Guillard and René Le Cohu for sending samples and for identifying and counting diatoms morphologically.

This work was funded by "projet innovant CODAL" (awarded to A.B. & A.F.) by the Ecology of Forests, Grassland and Freshwaters Division of INRA, and "projet innovant" (awarded to A.F.) by the Plant Health and Environment Division of INRA. Pyrosequencing and the production of reads were carried out at the Genotoul facility by L. K. with help from Eugénie Robe and Olivier Bouchez. This work was also funded by ONEMA (French National Agency for Water and Aquatic Environments) and ANRT (French National Agency for Research and Technology). Some sequences were obtained by the project @SPEED-ID "accurate SPEciEs Delimitation and Identification of eukaryotic biodiversity using DNA markers" proposed by F-Bol, the French Barcode of life initiative. Corinne Cruaud and Arnaud Couloux from Genoscope are thanked for sequencing.

Supplementary materials: *Species inventories*

Species	Morphology	CHLO SSUrDNA	*rbc*L
Achnanthidium minutissimum	0	5	9
Achnanthidium saprophilum	6	0	0
Amphora pediculus	2	0	0
Amphora veneta	1	0	0
Amphora unassigned	0	2	0
Aulacoseira granulata	1	0	0
Cocconeis pediculus	20	61	72
Cocconeis placentula	0	0	2
Craticula accomoda	0	0	2
Craticula cuspidata	0	71	0
Cyclotella meneghiniana	1	0	0
Cylindrotheca closterium	0	4	0
Discostella pseudostelligera	1	0	0
Discostella unassigned	0	0	2
Eolimna subminuscula	109	623	9806
Fistulifera pelliculosa	0	10	0
Fistulifera saprophila	17	299	10882
Gomphonema bourbonense	0	0	5
Gomphonema parvulum	33	58	744
Gomphonema pumilum	4	0	0
Mayamaea atomus	0	6	0
Mayamaea permitis	11	2	852
Navicula cryptotenella	0	0	41
Navicula gregaria	17	58	74
Navicula lanceolata	3	0	114
Navicula tripunctata	1	4	9
Navicula veneta	9	52	380
Nitzschia agnewii	1	0	0
Nitzschia amphibia	1	0	0
Nitzschia capitellata	5	22	1897
Nitzschia dissipata	3	0	0
Nitzschia palea	15	2	448
Nitzschia pusilla	0	0	16
Nitzschia unassigned	0	0	35
Planothidium frequentissimum	2	0	0
Planothidium lanceolatum	0	5	30
Pseudogomphonema unassigned	0	3	0
Reimeria sinuata	1	0	5
Rhoicosphenia abbreviata	2	0	0
Sellaphora sp.	22	0	0
Sellaphora minima	114	89	123
Sellaphora seminulum	11	0	28
Surirella angusta	0	0	3

Species	Morphology	LACL SSUrDNA_1	SSUrDNA_2	*rbc*L_1	*rbc*L_2
Achnanthes ploenensis	1	0	0	0	0
Achnanthidium eutrophilum	3	0	0	0	0
Achnanthidium minutissimum	3	0	2	0	0
Actinocyclus normanii	1	0	0	0	0
Amphora capitellata	0	0	5	0	0
Amphora copulata	30	0	0	0	0
Amphora inariensis	22	0	0	0	0
Amphora libyca	0	3	11	27	34
Amphora montana	0	0	0	0	4
Amphora ovalis	1	0	0	0	0
Amphora pediculus	40	7	7	115	150
Amphora unassigned	0	3	4	27	55
Aulacoseira granulata	2	0	2	0	0
Aulacoseira pusilla	2	0	0	0	0
Aulacoseira subarctica	1	0	0	0	0
Bacillaria paxillifer	2	7	9	96	173
Craticula importuna	0	0	0	0	2
Caloneis lewisii	0	0	3	0	0
Craticula molestiformis	2	0	0	0	0
Cyclostephanos dubius	3	0	0	0	0
Cyclostephanos invisitatus	1	0	0	0	0
Cyclotella meduanae	1	0	0	0	0
Cyclotella meneghiniana	2	0	0	0	0
Cymatosira belgica	1	0	0	0	0
Cymbella aspera	0	0	3	0	0
Discostella pseudostelligera	5	0	0	0	0
Encyonema caespitosum	1	0	0	0	0
Encyonema lange-bertalotii	0	0	2	0	0
Encyonema silesiacum	1	0	0	0	0
Epithemia sorex	0	0	0	5	0
Fistulifera saprophila	0	0	0	0	3
Fragilaria crotonensis	0	0	0	0	3
Gomphonema parvulum	3	0	0	19	9
Gyrosigma acuminatum	0	36	42	48	86
Gyrosigma limosum	0	0	6	0	0
Gyrosigma sciotense	9	0	0	0	0
Mayamaea permitis	1	0	0	0	0
Melosira varians	0	0	5	0	0
Navicula amphiceropsis	1	0	0	0	0
Navicula antonii	1	0	0	0	0
Navicula capitatoradiata	1	0	0	8	22
Navicula caterva	3	0	0	0	0
Navicula cincta	0	6	10	22	29
Navicula cryptotenella	43	7	10	680	1166

Navicula cryptotenelloides	0	0	10	552	968
Navicula erifuga	1	0	0	0	0
Navicula gregaria	1	0	0	0	0
Navicula phyllepta	0	0	3	0	0
Navicula rostellata	1	0	0	0	0
Navicula symmetrica	1	0	2	106	182
Navicula tripunctata	5	26	49	268	533
Navicula unassigned	0	0	0	25	49
Navicula veneta	0	0	0	5	22
Navicula viridula	0	0	0	78	122
Nitzschia acicularis	3	0	0	0	0
Nitzschia amphibia	7	4	5	0	0
Nitzschia communis	0	0	2	0	0
Nitzschia dissipata	10	0	4	10	24
Nitzschia draveillensis	0	0	0	3	0
Nitzschia filiformis	12	7	6	462	864
Nitzschia fonticola	3	0	0	0	0
Nitzschia gracilis	1	0	0	0	0
Nitzschia inconspicua	3	0	0	0	0
Nitzschia liebetruthii	5	0	0	0	0
Nitzschia cf. *frustulum*	39	0	0	0	0
Nitzschia palea	2	0	0	172	371
Nitzschia paleacea	18	0	0	0	0
Nitzschia pusilla	0	0	0	2	17
Nitzschia recta	1	0	0	0	0
Nitzschia sigma	0	2	0	0	0
Nitzschia sociabilis	18	0	0	0	0
Nitzschia subacicularis	6	0	0	0	0
Nitzschia supralitorea	0	2	3	0	0
Nitzschia unassigned	0	4	0	39	54
Paralia sulcata	1	0	0	0	0
Planothidium lanceolatum	0	0	0	3	0
Rhoicosphenia abbreviata	4	0	3	0	0
Sellaphora minima	32	5	11	70	74
Sellaphora seminulum	34	0	0	0	0
Skeletonema subsalsum	0	4	3	0	0
Stephanodiscus hantzschii	3	0	0	0	0
Stephanodiscus estibulis	2	0	0	0	0
Tetraselmis unassigned	0	2	0	0	0
Ulnaria ulna	0	0	0	8	0

Species	COIN		
	Morphology	SSUrDNA	*rbc*L
Achnanthidium eutrophilum	1	0	0
Achnanthidium exiguum	1	0	0
Achnanthidium minutissimum	6	0	0
Amphora montana	4	0	6
Amphora pediculus	2	0	0
Amphora subturgida	2	0	0
Cocconeis euglypta	1	0	0
Diadesmis contenta	10	0	0
Diadesmis paracontenta	4	0	0
Diadesmis sp.	1	0	0
Eolimna sp.1	5	0	0
Eolimna sp.2	8	0	0
Eunotia incisa	5	0	0
Eunotia sp.	4	0	0
Eunotia viola	8	0	0
Frustulia crassinervia	4	0	0
Frustulia saxonica	1	0	0
Frustulia sp.	3	0	0
Frustulia vulgaris	2	0	0
Gomphonema affine	0	0	2
Gomphonema angustum	30	0	0
Gomphonema bourbonense	22	8	2
Gomphonema brasiliense	21	0	0
Gomphonema clevei	6	0	0
Gomphonema designatum	6	0	0
Gomphonema parvulum	7	20	43
Gomphosphenia sp.	2	0	0
Hantzschia amphioxys	1	0	0
Luticola mutica	2	0	0
Mayamaea permitis	4	0	0
Mayamaea unassigned	1	0	0
Navicula cruxmeridionalis	2	0	0
Navicula cryptocephala	2	9	17
Navicula eichorniaephila	18	0	0
Navicula erifuga	7	0	0
Navicula escambia	29	0	0
Navicula heimansioides	2	0	0
Navicula incertata	1	0	0
Navicula lundii	2	0	4
Navicula phyllepta	0	0	0
Navicula quasidisjuncta	22	0	0
Navicula rostellata	1	0	0
Navicula symmetrica	0	2	50
Navicula unassigned	0	0	18

Navicula viridula	0	6	0
Naviculadicta nanogomphonema	26	0	0
Nitzschia amphibia	0	0	5
Nitzschia clausii	1	0	0
Nitzschia dissipata	5	0	0
Nitzschia frustulum var frustulum	4	0	0
Nitzschia hantzschiana	9	0	0
Nitzschia lange-bertalotii	5	0	0
Nitzschia lorenziana	0	0	6
Nitzschia palea	3	0	13
Nitzschia palea var debilis	7	0	0
Nitzschia scalpelliformis	5	0	0
Nitzschia sp.1	12	0	0
Nitzschia sp.2	35	0	0
Nitzschia supralitorea	10	0	0
Nupela sp.	4	0	0
Pinnularia acrosphaeria	0	8	0
Pinnularia graciloides	1	0	0
Pinnularia sp.	2	0	0
Pseudogomphonema sp.	0	3	0
Rhopalodia gibberula	4	0	0
Sellaphora auldreekie	0	0	3
Stauroneis unassigned	3	0	0
Stephanodiscus hantzschii	1	0	0
Tryblionella debilis	13	0	0
Ulnaria biceps	2	0	0
Ulnaria ulna	0	51	25

Species	Morphology	BDNA SSUrDNA	BDNA *rbc*L
Achnanthidium coarctatum	0	2	0
Achnanthidium minutissimum	0	3	8
Achnanthidium navaroii (sp. nov.)	4	0	0
Adlafia muscora	2	0	0
Cocconeis placentula	79	21	113
Eolimna ruttneri	5	0	0
Fistulifera saprophila	0	0	3
Fragilaria capucina	0	25	163
Fragilaria nanana	0	8	0
Fragilaria vaucheriae	15	0	0
Gomphonema affine	0	14	2
Gomphonema bourbonense	0	6	3
Gomphonema clevei	122	1266	3066
Gomphonema aff. *designatum*	1	0	0
Gomphonema parvulum	0	655	0
Gomphonema pumilum	0	0	5
Mayamaea permitis	1	0	0
Melosira varians	0	122	0
Navicula capitatoradiata	0	0	62
Navicula cryptotenella	1	0	0
Navicula cryptotenelloides	0	17	15
Navicula escambia	2	0	0
Navicula gregaria	1	20	0
Navicula radiosa	0	2	0
Navicula sp.	5	0	0
Navicula symmetrica	0	3	53
Navicula tripunctata	0	7	4
Navicula unassigned	0	2	2
Navicula viridula	0	4	0
Navicula ranomafanensis	1	0	0
Nitzschia amphibia	4	566	1415
Nitzschia bourbonensis (sp. nov.)	105	0	0
Nitzschia frustulum	0	502	1140
Nitzschia palea	0	0	3
Nitzschia solgensis	2	0	0
Nitzschia tropica	53	0	0
Nitzschia tubicola	0	0	3
Pseudo-nitzschia cuspidata	0	3	0
Pseudogomphonema sp.	0	9	0
Rhopalodia contorta	0	0	13
Rhopalodia gibba	0	13	323
Rhopalodia hirundiniformis	7	0	0
Sellaphora minima	86	9	85
Sellaphora seminulum	2	0	6

Staurosirella aff. *pinnata*	2	0	0
Surirella angusta	0	0	2
Ulnaria ulna	0	532	373
Urosolenia eriensis	0	7	0

Genus inventories

Genus	Morphology	LACL SSUrDNA_1	SSUrDNA_2	*rbc*L_1	*rbc*L_2
Achnanthes	1	0	0	0	0
Achnanthidium	6	0	9	0	2
Actinocyclus	1	0	0	0	0
Amphora	93	747	1388	50	84
Aulacoseira	5	0	3	0	6
Bacillaria	2	151	284	7	9
Caloneis	0	4	0	0	3
Craticula	2	0	6	0	0
Cyclostephanos	4	0	0	0	0
Cyclotella	3	0	0	0	2
Cylindrotheca	0	0	0	2	3
Cymatosira	1	0	0	0	0
Cymbella	0	0	2	0	3
Denticula	5	0	0	0	0
Encyonema	2	6	0	3	7
Epithemia	0	8	0	4	0
Fistulifera	0	9	26	0	0
Gomphonema	3	34	15	4	4
Gyrosigma	9	789	1507	122	170
Mayamaea	1	0	3	0	0
Melosira	0	0	0	0	5
Navicula	58	3840	6941	182	376
Nitzschia	128	1531	3063	27	36
Paralia	1	0	0	0	0
Planothidium	0	6	0	0	0
Pseudogomphonema	0	0	0	0	3
Rhoicosphenia	4	0	0	0	0
Sellaphora	66	94	131	8	20
Skeletonema	0	2	0	5	5
Stephanodiscus	5	0	0	0	0
Tetraselmis	0	0	0	2	0
Thalassiosira	0	0	0	5	3
Ulnaria	0	9	0	0	0

	CHLO		
Genus	Morphology	SSUrDNA	*rbc*L
Achnanthidium	6	22	5
Amphora	3	8	3
Aulacoseira	1	0	0
Cocconeis	20	185	144
Craticula	0	7	71
Cyclotella	1	0	0
Cylindrotheca	0	0	4
Discostella	1	2	2
Eolimna	109	17669	623
Fistulifera	17	20162	843
Fragilaria	0	2	0
Gomphonema	37	1460	242
Mayamaea	11	1552	83
Navicula	30	1165	207
Nitzschia	25	3809	42
Planothidium	2	53	5
Pseudogomphonema	0	0	3
Reimeria	1	8	0
Rhoicosphenia	2	0	0
Sellaphora	147	1625	137
Surirella	0	17	0
Ulnaria	0	2	0

Genus	Morphology	COIN SSUrDNA	COIN *rbc*L
Achnanthidium	8	0	0
Amphora	8	8	0
Cocconeis	1	0	0
Diadesmis	15	0	0
Eolimna	13	0	0
Eunotia	17	23	4
Frustulia	10	0	0
Gomphonema	92	115	133
Gomphosphenia	2	0	0
Hantzschia	1	0	0
Luticola	2	0	0
Mayamaea	5	0	0
Navicula	86	165	125
Naviculadicta	26	0	0
Nitzschia	96	64	3
Nupela	4	0	0
Pinnularia	3	12	8
Pseudogomphonema	0	0	3
Rhopalodia	4	0	0
Sellaphora	0	6	0
Stauroneis	3	0	0
Stephanodiscus	1	0	0
Tryblionella	13	0	0
Ulnaria	2	120	51

Genus	Morphology	BDNA SSUrDNA	BDNA *rbc*L
Achnanthidium	4	18	7
Adlafia	2	0	0
Amphora	0	0	2
Cocconeis	79	335	79
Cymbella	0	5	11
Eolimna	5	0	0
Fistulifera	0	4	0
Fragilaria	15	265	36
Gomphonema	123	5620	2831
Mayamaea	1	0	0
Melosira	0	2	122
Navicula	10	306	319
Nitzschia	164	4967	1139
Pseudo-nitzschia	0	0	4
Pseudogomphonema	0	0	9
Reimeria	0	29	0
Rhopalodia	7	3865	13
Sellaphora	88	193	12
Staurosirella	2	0	0
Surirella	0	6	0
Ulnaria	0	743	532
Urosolenia	0	0	7

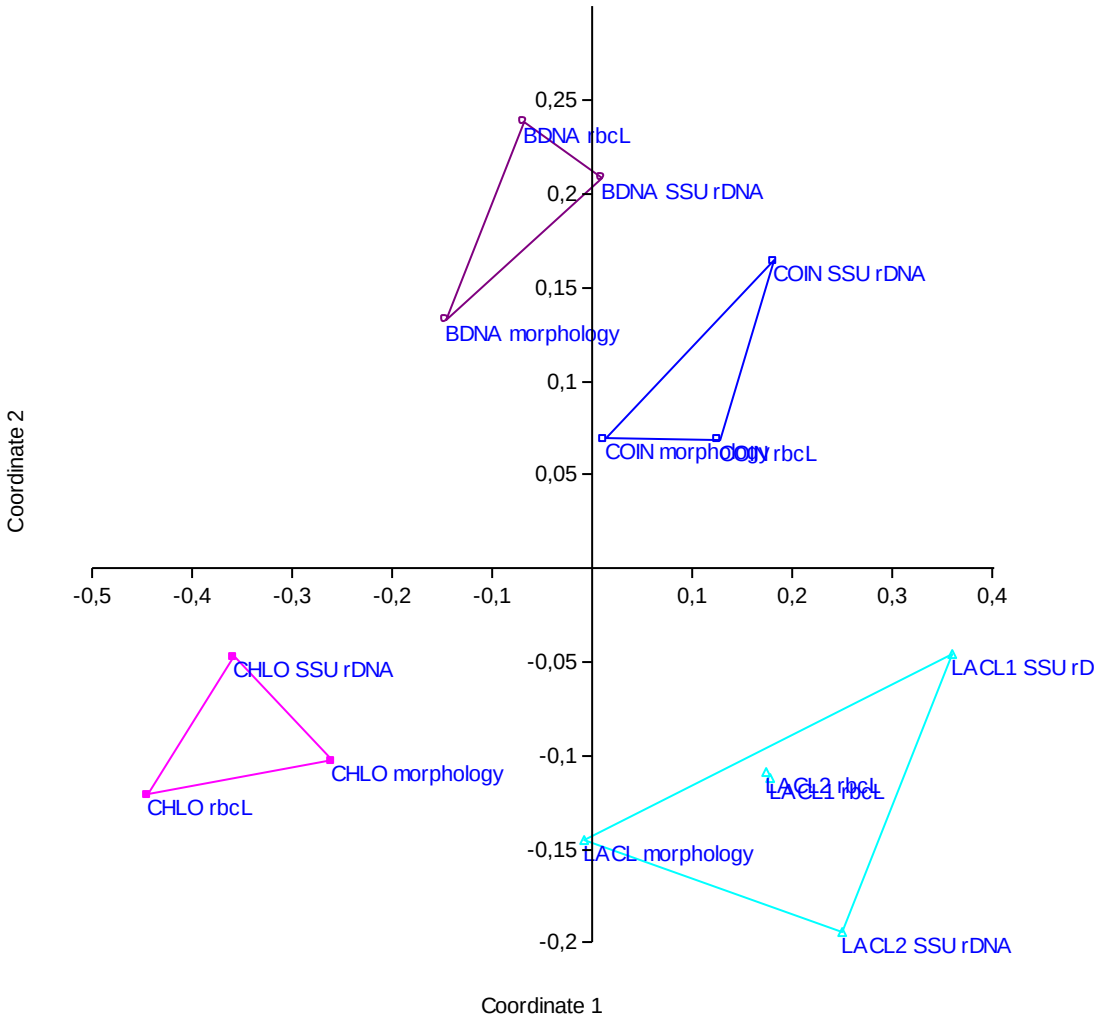


Figure IV.12: *Non metric Multidimensional Scaling based on Bray-Curtis distances comparing molecular and morphological genus inventories of the four samples.*

5. Conclusion

Nos différentes études ont révélé que les méthodes moléculaires de séquençage à haut débit associées à des méthodes bioinformatiques adaptées sont de puissants outils pour l'étude des communautés de diatomées. Les algorithmes *metaMatch* sont particulièrement appropriés pour établir, à partir de reads obtenus en NGS, des inventaires précis des taxa composant les communautés de diatomées, lorsque ceux-ci sont référencés dans des bases de séquences. Ces outils offrent l'avantage d'être suffisamment rapides et simples d'utilisation. De plus, nous avons démontré que l'utilisation du pyroséquençage 454 est adaptée à l'étude des communautés de diatomées dans des échantillons naturels tels que les échantillons benthiques utilisés pour la bioindication.

Concernant les différents marqueurs nucléiques, les études présentées ont montré que même si le *cox*1 permet une bonne discrimination des taxa de diatomées, ce marqueur ne peut être utilisé en raison de la faible quantité de séquences de référence disponibles actuellement. Ce manque de séquences de référence, bénéficiant d'une identification taxonomique sûre, est, pour tous les marqueurs, la limite actuelle la plus importante à l'utilisation de notre outil moléculaire. Mann & Droop (1996) évaluaient le nombre d'espèce de diatomées à 200 000. Actuellement, 62 000 nom de taxa sont référencés dans le « Catalogue of Diatom Names » (Fourtanier & Kociolek, 2011). Les bases de référence préparées au cours de la thèse ne couvrent donc qu'une infime partie de cette diversité. Nous avons donc besoin d'étendre de manière importante ces bases de référence pour améliorer l'utilisation de notre outil moléculaire, ce qui passe par un effort accru d'isolement et de séquençage de souches de diatomées. Le séquençage de nombreux marqueurs nucléiques sur chaque souche de diatomées peut s'avérer long et coûteux. Le choix d'un marqueur principal serait donc plus raisonnable. Pour des études générales de la diversité des diatomées, le marqueur 18S semble apporter des informations intéressantes. Le pouvoir résolutif de ce marqueur est faible, mais il permettrait d'obtenir des profils globaux de la diversité des diatomées. Par contre, en raison des biais liés aux méthodes moléculaires, le gène *rbc*L paraît être le marqueur le plus adapté dans le cadre de la bioindication. En effet, son pouvoir résolutif, la distribution de son polymorphisme, sa facilité de séquençage, les séquences déjà disponibles et sa reproductibilité en pyroséquençage, aboutissent à la meilleure estimation de la composition en diatomées d'un échantillon benthique naturel. Il serait donc intéressant de cibler en priorité ce

marqueur pour le séquençage de nouvelles souches de diatomées afin d'enrichir la base de référence.

De nombreuses améliorations devront encore être faites avant de pouvoir utiliser le barcoding environnemental comme nouvelle approche de bioindication. Cependant nos différentes études ont démontré le potentiel des méthodes moléculaires et permettent ainsi d'orienter les futures études à mener pour aboutir à un suivi de la qualité de l'eau, basé sur des inventaires moléculaires des taxa de diatomées.

CHAPITRE V. DISCUSSION GENERALE

Le sujet central de ce travail de thèse était l'utilisation des outils moléculaires pour améliorer l'identification des diatomées dans la cadre de la bioindication, dans l'objectif de progresser dans le développement d'un nouvel outil de bioindication, basé sur ces techniques. Utiliser une telle méthode de bioindication, nécessitait d'obtenir des connaissances supplémentaires sur l'apport des données moléculaires dans l'identification des diatomées et de développer des outils adaptés à l'étude d'échantillons environnementaux. Dans ce contexte général, deux axes principaux ont été développés. Le premier était basé sur l'étude de souches de diatomées pour améliorer nos connaissances sur leur détermination, et pour vérifier l'intérêt des marqueurs nucléiques pour cette identification. Parallèlement, ces souches ont enrichi les bases de séquences de diatomées existantes. Le second axe, plus appliqué, était basé sur l'analyse d'échantillons complexes de diatomées, afin de définir la méthodologie à mettre en place pour une nouvelle méthode de bioindication sur échantillons environnementaux. Ces différentes approches ont apporté des connaissances générales sur la taxonomie des diatomées ainsi que sur les méthodologies utilisables pour l'étude des communautés environnementales de diatomées.

1. Taxonomie

La systématique a pour but de classer les organismes. Le niveau taxonomique « espèce » est un concept dont les définitions sont nombreuses (Mayden, 1997). La notion d'espèce la plus courante chez les eucaryotes est l'espèce biologique définie par Mayr en 1942, qui se base principalement sur les capacités de reproduction sexuée entre les différents individus. Comme nous l'avons présenté dans le CHAPITRE I, la reproduction des diatomées est peu connue et difficilement observable en conditions naturelles, puisque la phase de reproduction sexuée est très courte par rapport à la phase de multiplication végétative. De plus, la capacité d'autogamie ou d'apogamie de certaines espèces de diatomées (e.g. Sabbe et al., 2004; Trobajo et al., 2006) ne permet pas d'établir une taxonomie uniquement basée sur ce concept.

Taxonomie morphologique

Les espèces de diatomées sont donc essentiellement définies sur un concept morphologique facilement applicable grâce au frustule qui constitue un ensemble de marqueurs phénotypiques. De plus, les frustules sont bien conservés au cours du temps

dans différents environnements et peuvent également être fossilisés (Smol & Stoermer, 2010) ce qui permet de retracer les événements passés.

Cependant, nous avons démontré dans le CHAPITRE III que ce concept morphologique a des limites chez les diatomées à tous les niveaux taxonomiques, car les variations morphologiques intraspécifiques ont été insuffisamment étudiées. Les définitions des taxa sont réalisées autour de « types » qui sont des références au nom de taxon (illustration ou spécimen) selon le code international de nomenclature botanique (McNeill et al., 2006). Mais les références aux types à des fins de nomenclature ont entraîné des descriptions de taxa trop étroites, car les types ont été considérés comme représentatifs de taxa, alors qu'ils peuvent ne pas être au centre de la gamme de variation d'un taxon (Cox, 2009). En effet, nous avons démontré que la différence entre des espèces telles que *Gomphonema parvulum, G. exilissimum* ou *G. lagenula* (Article II) pourrait être remise en cause par des analyses moléculaires. La description de nouvelles espèces a parfois été faite sur des divergences uniquement phénotypiques qui ne reflètent pas de divergences phylogénétiques. Mais, nous avons également démontré que le phénomène inverse est possible, en révélant de fortes variations génétiques au sein d'une espèce morphologique (CHAPITRE III.3 : clades de *G. parvulum* et *Nitzschia palea*) tout comme au sein d'un même genre (Article I : Genres paraphylétiques). Les données moléculaires fournissent donc des preuves phylogénétiques qui ne sont pas toujours en cohérence avec la systématique traditionnelle des diatomées (Cox, 2009).

Taxonomie moléculaire

Le concept de taxonomie moléculaire s'est développé avec les techniques de biologie moléculaire et en particulier avec le séquençage. Pour les bactéries, qui ne présentent pas de reproduction sexuée et qui ont des caractères morphologiques difficilement exploitables, ce concept est largement appliqué. Dès 1987, les propriétés d'hybridation de l'ADN ont été utilisées pour la taxonomie des espèces procaryotiques (Wayne et al., 1987) et les séquences d'ADNr 16S sont devenues une alternative aux méthodes morphologiques (Cristea-Fernström et al., 2007). La taxonomie ADN (Blaxter, 2004) se base sur ce concept d'espèce moléculaire pour lequel les individus d'une même espèce ont des séquences plus semblables que les individus appartenant à des espèces différentes. Ainsi, des seuils de 97% d'identité (Stackebrandt & Goebel, 1994) entre séquences d'ADNr 16S sont généralement appliqués pour regrouper les séquences bactériennes en

taxa « moléculaires ». Cette approche est particulièrement utile pour les groupes de taxa pour lesquels il manque un système taxonomique détaillé.

Pour les diatomées, qui ont un système taxonomique détaillé, qui cependant ne reflète pas systématiquement les divergences génétiques, le développement du séquençage de marqueurs nucléiques spécifiques a fourni les séquences de référence nécessaires à la définition d'espèces moléculaires. Ces séquences ont essentiellement été utilisées pour identifier, au niveau moléculaire, des taxa définis à partir du concept d'espèce morphologique. Ainsi, plusieurs auteurs ont comparé les distances génétiques des diatomées à différents niveaux taxonomiques (e.g. Evans et al., 2007; Moniz & Kaczmarska, 2009; Trobajo et al., 2010; Zimmermann et al., 2011).

Au cours de ce travail de thèse, nous avons défini des seuils d'identité pour assigner à des séquences ADN des noms de taxa définis sur la base de leur morphologie (Article V). Un seuil de 99% peut être appliqué aux marqueurs 18S et *rbc*L pour déterminer les taxa au niveau de l'espèce et des seuils de 99% pour le 18S et de 98% pour le *rbc*L peuvent être utilisés au niveau du genre. La définition, pour l'ADNr 18S, d'un même seuil pour les deux niveaux taxonomiques reflète la difficulté de définir un seuil moyen du fait de différentes vitesses d'évolution au sein des différentes lignées de diatomées. Ces seuils ne peuvent donc pas être exacts, mais ils ont été choisis comme les meilleurs compromis. Nous n'avons pas pu définir de tels seuils pour les séquences de *cox*1 en raison du faible nombre de séquences de référence dont nous disposions. Ces seuils pourront être utilisés soit en faisant correspondre les séquences à des noms de taxa pour étudier la diversité connue (cas du barcoding ADN), soit en regroupant les séquences en OTUs pour étudier la diversité totale des diatomées, sans correspondance directe avec les espèces morphologiques (cas de la taxonomie ADN). Ces OTUs pourraient ensuite être comparées aux bases de référence pour identifier les espèces les plus proches.

Du fait de la variabilité intraclonale des ITS (Alvarez & Wendel, 2003) et des difficultés de séquençage qui l'accompagnent, nous n'avons pas choisi d'utiliser ce marqueur pour identifier les espèces d'une communauté de diatomées. Néanmoins, ce marqueur peut s'avérer utile pour l'étude de souches dans le contexte de la taxonomie moléculaire. Il a précédemment été utilisé pour étudier les plantes (e.g. Baldwin, 1992; Sun et al., 1994), les champignons (e.g. Cooke & Duncan, 1997; Kelly et al., 2011), mais également les diatomées (Behnke et al., 2004; Casteleyn et al., 2008). Nous avons confirmé sa capacité de discrimination des taxa proches sur le complexe *Gomphonema*

parvulum (Article II). L'intérêt principal des ITS, pour la taxonomie moléculaire, réside dans leur structure secondaire qui peut être utilisée pour définir les espèces. En effet, les CBC sembleraient être de bons marqueurs pour distinguer les espèces isolées par la reproduction sexuée (Amato et al., 2007; Coleman, 2009; Ruhl et al., 2010). Ce type de données moléculaires serait donc lié au concept d'espèce biologique. Sorhannus et al. (2010) ont d'ailleurs observé, pour *Thalassiosira weissflogii,* des résultats concordants entre la présence de CBC et les séquences d'un gène supposé être impliqué dans la reconnaissance entre les gamètes. Chez les diatomées, malgré la difficulté de séquençage des ITS, l'étude des CBC serait donc intéressante pour stabiliser la taxonomie en apportant une information liée au concept d'espèce biologique, tout en évitant l'étude souvent lourde et délicate de la reproduction.

Morphologie vs. séquences ADN

Les études menées au cours de la thèse étaient principalement basées sur une comparaison des données obtenues par des observations en microscopie et des données obtenues par séquençage ADN. Ces deux types de méthodes ont précédemment révélé des diversités phytoplanctoniques différentes (Savin et al., 2004). Nous avons, en effet, confirmé que les deux méthodes fournissent des informations différentes, que ce soit sur des souches (CHAPITRE III) ou sur des échantillons environnementaux (Article V). Cependant, nous avons également observé des concordances entre les indications obtenues par la morphologie et par les informations génétiques. Par exemple, nous avons démontré que certains genres de Cymbellales, comme *Encyonema* ou *Placoneis*, définis sur la base de leur morphologie, correspondaient effectivement à des groupes phylogénétiques monophylétiques (Article I). De plus, par des analyses de morphométrie géométrique nous avons pu identifier des différences morphologiques entre des clades définis par analyses moléculaires (Article II). Enfin, lors des études sur les échantillons complexes (Article V), des correspondances ont également été trouvées entre les inventaires morphologiques et moléculaires. Des différences morphologiques existent parfois entre les groupes moléculaires mais ce ne sont pas automatiquement celles qui ont été retenues par la taxonomie morphologique. Par exemple, au sein des complexes d'espèces, certains groupes présentent entre eux des variations morphologiques (Lundholm et al., 2003; Sarno et al., 2005; Evans et al., 2008; Kooistra et al., 2010) qui n'avaient pas été prises en compte lors de la description morphologique initiale. En

accord avec Cox (2009), le choix des caractères morphologiques utilisés pour identifier les différents taxa peut donc altérer la systématique des diatomées.

Ces écarts entre la classification basée sur la morphologie du frustule et celle basée sur les séquences ADN sont liés aux critères morphologiques choisis par les taxonomistes, et aux différents facteurs qui les influencent. Les séquences des marqueurs nucléiques quant à elles, nous renseignent sur l'évolution de ces gènes ou régions intergéniques en fonction de la pression sélective qu'ils subissent qui est différente pour chaque marqueur nucléique. Les pressions subies par ces marqueurs sont différentes des pressions environnementales qui génèrent les variations morphologiques.

Ces différences entre taxonomie morphologique et taxonomie moléculaire, tout comme les différences entre les observations en microscopie optique et en microscopie électronique ou les différences entre les marqueurs nucléiques, aboutissent à des représentations différentes de l'écologie des taxa ou de leur biogéographie. En effet, notre compréhension de l'écologie des taxa est influencée par la taxonomie utilisée pour les distinguer. Mann & Droop (1996) indiquaient que l'utilisation d'un concept d'espèce plus étroit apporterait probablement une plus grande précision des données en écologie et paléolimnologie. Des différences écologiques ont, par exemple, été observées entre des clades appartenant à la même espèce morphologique (Vanelslander et al., 2009).

Par ailleurs, notre compréhension des distributions géographiques est également dépendante de la taxonomie appliquée. Par exemple, les taxa *Thalassiosira gravida* (eaux polaires) et *T. rotula* (eaux tempérées), considérés précédemment comme deux espèces distinctes, ont pu être récemment regroupés en observant que les variations morphologiques étaient dépendantes de l'environnement, notamment de la température de l'eau (Sar et al., 2011). Au contraire, les complexes *Gomphonema parvulum* et *Nitzschia palea* étudiés sur la base de la morphologie paraissaient être distribués largement et ne pas présenter de différenciation biogéographique (Finlay et al., 2002). En les étudiant sur la base moléculaire, nous avons déterminé que les clades composant ces complexes pouvaient présenter des profils biogéographiques particuliers (CHAPITRE III.3). Le même type de résultats a été obtenu pour d'autres micro-algues, par exemple *Synura petersenii* (Boo et al., 2010), ou d'autres diatomées telle que *Pseudo-nitzschia pungens* (Casteleyn et al., 2008). Les fortes capacités de dispersion des diatomées, liées à leur petite taille (inférieure à 1mm, Finlay & Fenchel, 2004), ne semblent donc pas empêcher les processus de spéciation allopatriques.

Les deux méthodes de classification sont complémentaires puisqu'elles nous renseignent sur des facteurs différents qui aboutissent à des entités taxonomiques différentes difficilement comparables. Cependant, à la classification morphologique des diatomées sont reliées de nombreuses connaissances sur les taxa telles que leur biogéographie ou leur écologie. Il est donc utile de conserver une correspondance à la morphologie pour relier les nouvelles données moléculaires à ces connaissances.

En conclusion, différents types de taxonomie peuvent être utilisés pour classer les diatomées. Cependant, la spéciation est un processus progressif vis-à-vis duquel aucun marqueur, qu'il soit phénotypique ou génotypique n'est parfait. L'ensemble des méthodes apportent des informations qui, combinées dans une approche polyphasique, peuvent aider à atteindre une taxonomie qui soit la plus représentative possible de l'évolution des taxa de diatomées. Une taxonomie précise et stable est nécessaire pour mieux comprendre la diversité des diatomées et les facteurs qui influencent leur distribution et leur dynamique.

2. Utilisations des méthodes moléculaires

Les analyses de biodiversité (rapides et exactes) sont un enjeu majeur pour les programmes de bioindication mais également pour comprendre le fonctionnement général des écosystèmes. Nous avons testé l'utilisation des méthodes moléculaires sur des échantillons complexes. Pour cela, nous avons testé quelques étapes du processus sur des échantillons artificiels, puis nous avons testé notre outil sur des échantillons naturels. Nous avons choisi de baser nos études sur les outils moléculaires existants (marqueurs étudiés précédemment, amorces d'amplification disponibles, etc.) et sur les dernières technologies de séquençage haut-débit (NGS). Les différentes études réalisées au cours de la thèse ont ainsi révélé que les NGS couplées à des outils bioinformatiques adaptés sont des voies intéressantes d'analyse des communautés de diatomées.

Notre cadre d'étude était l'évaluation de la qualité de l'eau par les diatomées, mais il est important de noter que nos travaux, axés sur les méthodes de barcoding environnemental, peuvent être généralisés aux approches de diversité des communautés de diatomées et des autres communautés environnementales de protistes. Par exemple, la profondeur de séquençage, fournie par les NGS, permettrait d'évaluer la répartition géographique de certains protistes et ainsi, de valider ou réfuter la proposition de Finlay et al. (2002), qui suppose que les profils biogéographiques observés sont liés à la non

détection des taxa rares. Le pyroséquençage de communautés de diatomées peut être réalisé pour détecter les espèces de diatomées référencées dans les bases de séquences en utilisant les algorithmes *metaMatch* et les marqueurs nucléiques 18S et *rbc*L (CHAPITRE IV). Mais, de telles analyses peuvent également être réalisées sur d'autres protistes pour lesquelles des bases de séquences de référence sont disponibles.

D'autre part, la comparaison de différents échantillons naturels (Article V) a révélé que le marqueur 18S pouvait être utile pour une étude générale de la diversité des diatomées. Du fait de son pouvoir résolutif plus faible, il pourrait être utilisé pour étudier les processus globaux qui structurent les communautés de diatomées à une échelle plus large. Cette approche permettrait d'étudier la structure des communautés de diatomées en fonction des distributions géographiques et des différents types d'écosystèmes comme cela a été précédemment réalisé pour les communautés bactériennes (e.g. Pommier et al., 2007; Humbert et al., 2009).

Enfin, la méthode utilisée au cours de ce travail de thèse, pourrait également être appliquée aux études de paléolimnologie. En effet, du fait de leur capacité bioindicatrice et de la persistance de leurs frustules, les diatomées sont couramment utilisées pour reconstruire les évènements écologiques passés, à partir de carottes de sédiments lacustres. Récemment, Stoof-Leichsenring et al. (2012) ont démontré, par une approche de clonage-séquençage, les capacités des études génétiques basées sur les diatomées pour la paléolimnologie. Ainsi, le pyroséquençage du marqueur *rbc*L, couplé à l'utilisation des programmes *metaMatch* et de notre base de séquences de référence pourrait être utilisés pour suivre les modifications temporelles des communautés de diatomées dans un lac. En complément, le marqueur 18S pourrait être utilisé, sans correspondance à des bases de référence, pour comparer les changements de communautés dans différents lacs.

3. Bioindication moléculaire

L'avantage principal des outils moléculaires est leur capacité d'automatisation dans le cas d'analyses de routine. Nous avons utilisé les méthodes moléculaires pour développer les éléments nécessaires à un nouvel outil de bioindication et déterminer les futurs axes de recherche à développer.

La recherche de correspondance entre séquences inconnues et séquences de référence est une étape centrale de notre méthodologie, tout comme l'identification morphologique

par un taxonomiste est une étape essentielle de la bioindication actuelle. Un algorithme sélectionne les reads provenant du pyroséquençage, qui correspondent aux références de manière à leur assigner un nom de taxa exact. Nous avons démontré que l'outil bioinformatique utilisé pour analyser les reads est important, puisque nous avons observé une forte dépendance des inventaires aux algorithmes et options utilisés (Article III). Nous avons également montré que les algorithmes *metaMatch* sont les plus précis, de ce fait, leur utilisation est appropriée pour réaliser des inventaires de taxa de diatomées de qualité. Le développement des algorithmes devra se poursuivre pour tendre vers des outils encore plus précis, plus rapides et plus accessibles aux gestionnaires de la qualité de l'eau.

La méthodologie que nous avons utilisée est basée sur le barcoding environnemental et les NGS. Le principal but du barcoding ADN est d'aider à l'identification de spécimens par comparaison de leurs séquences à une base de séquences de référence (DeWalt, 2011). Nos études ont démontré les capacités des outils moléculaires mais également perçu leurs limites.

Tout d'abord, notre méthode est limitée à la détection des taxa précédemment séquencés et référencés, tout comme la bioindication morphologique est limitée à l'analyse des taxa précédemment décrits. Cependant, les bases de référence assemblées au cours de la thèse ne couvrent qu'une petite partie des 62 000 taxa référencés dans le « Catalogue of Diatom Names » (Fourtanier & Kociolek, 2011). Ainsi, actuellement les taxa décrits morphologiquement sont beaucoup plus nombreux que les taxa présents dans les bases de séquences, et permettent donc une meilleure identification des taxa qui composent une communauté naturelle. La première limite de notre méthodologie est donc l'étendue des bases de référence. Dans le but de les enrichir, les efforts doivent être centrés sur la mise en culture et le séquençage de nouvelles souches représentatives de la diversité des diatomées, en particulier des taxa bioindicateurs, ce qui demande la coopération des taxonomistes (Mann et al., 2010).

Les limites des bases de référence sont évidemment dépendantes du marqueur utilisé, car tous les marqueurs n'ont pas le même nombre de séquences disponibles actuellement. De plus, certains marqueurs sont difficiles à séquencer en raison de la variabilité intraclonale (cas des ITS) ou de la faible efficacité des amorces PCR disponibles (cas du *cox*1). Par exemple, pour le marqueur *cox*1, certaines amorces (amorces pC1 et pB1, Ehara et al., 2000) sont efficaces pour amplifier le marqueur mais ne sont pas efficaces en

séquençage Sanger (observations sur la TCC, Evans et al., 2007). Une alternative pour les marqueurs *cox*1 et ITS serait d'ajouter une étape de clonage avant le séquençage Sanger. Cependant cette étape longue et onéreuse limiterait le développement rapide des bases de référence.

Outre les bases de séquences de référence, le choix du marqueur nucléique étudié est crucial pour la précision des données qu'il apporte. Les études réalisées au cours de la thèse ont révélé que les différents marqueurs moléculaires avaient des résolutions taxonomiques différentes, que ce soit pour étudier les relations entre groupes de diatomées (CHAPITRE III), ou pour évaluer la composition d'un échantillon complexe (CHAPITRE IV). Par exemple, les ADNr 18S et 28S ont apporté des informations sur des niveaux taxonomiques tels que le genre ou l'espèce (Articles I et II), permettant des analyses globales de communautés (Article V). Mais ceux-ci ne sont pas suffisamment variables pour des études taxonomiques plus fines. Au contraire, les marqueurs ITS, *cox*1 et *rbc*L nous apportent des données sur les relations spécifiques et intraspécifiques des taxa de diatomées (Articles II et IV), et donc sur les modifications précises des structures des communautés.

De plus, nous avons montré que les différences de distribution du polymorphisme le long des marqueurs nucléiques influencent les résultats de notre méthode. En effet, les différentes régions de l'ADNr 18S sont des sources d'erreur d'assignation (Article IV) et provoquent une importante perte d'information lors de la sélection des reads informatifs (Article V). Behnke et al. (2011) ont révélé des différences de taux d'erreurs de pyroséquençage entre deux régions hypervariables de l'ADNr 18S (V4 et V9). Ainsi, pour un même marqueur, la région ciblée est donc également importante. Dans le cadre de la bioindication, en prenant en compte les bases de référence actuelles, les biais liés au marqueur, la distribution du polymorphisme et la résolution taxonomique, nous avons conclu qu'un fragment du marqueur *rbc*L serait le plus approprié à l'étude des communautés de diatomées. Ce marqueur a d'ailleurs été proposé comme barcode principal pour l'identification des diatomées (Hamsher et al., 2011), et a prouvé ses capacités pour étudier la diversité des diatomées en paléolimnologie (Stoof-Leichsenring et al., 2012).

Des optimisations méthodologiques seront nécessaires pour une utilisation en routine. Tout d'abord, les biais d'extraction et d'amplification (Article IV) devront être diminués. Ensuite, l'utilisation de fragments ADN plus courts (~ 400 pb) sera nécessaire pour être

en adéquation avec les technologies NGS. Cette taille de fragment permettrait d'éliminer l'étape de fragmentation d'ADN et ainsi d'optimiser le nombre de reads utilisables pour compiler les inventaires. En effet, nous avons observé que les extrémités des fragments étaient faiblement représentées dans les bases de référence ce qui empêchait une bonne comparaison des reads contenant ces extrémités. Le séquençage Sanger en utilisant les amorces PCR ne permet pas d'obtenir les extrémités des marqueurs. Pour obtenir ces séquences, un séquençage dans les deux sens de lecture sur la totalité de la longueur est nécessaire. Ceci est très coûteux pour les longs fragments ADN et n'est pas systématiquement réalisé. De plus, toutes les séquences d'un même marqueur ne sont pas toujours obtenues à partir des mêmes amorces. Ainsi, nos bases de référence sont composées par des séquences de longueurs différentes. En utilisant les algorithmes *metaMatch*, nous avons effectué des recherches de correspondances sur la totalité de la longueur des reads afin d'obtenir une meilleure identification. Ce paramètre, particulièrement important pour conserver la précision de l'outil, nous permettait d'utiliser uniquement les reads entièrement inclus dans les séquences de référence, mais pas les reads chevauchants, partiellement inclus. Par exemple, ~ 43% des reads des échantillons artificiels contenaient une amorce PCR dans leur séquence, alors que seulement 10% des séquences de référence des espèces incluses dans ces échantillons artificiels contenaient une amorce. Ces différences de longueur ont donc fortement limité les données utilisables. L'utilisation d'un fragment plus court, mieux représenté dans les bases de référence nous permettrait donc d'optimiser l'utilisation du pyroséquençage.

Zimmermann et al. (2011) ont développé de nouvelles amorces ciblant une région polymorphe (V4) de l'ADNr 18S mieux représentée dans la base de référence. Il serait donc intéressant de tester ces amorces avec notre méthode pour évaluer l'apport d'un fragment plus court et éliminer les biais dus aux régions conservées de ce marqueur (Article IV). De nouvelles amorces pour le marqueur *rbc*L pourraient également être définies grâce à nos bases de référence et à l'aide de nouveaux programmes tels que la série de logiciel EcoPCR (Ficetola et al., 2010). Ainsi, les amorces existantes ou de nouveaux couples d'amorces devront être testés pour optimiser l'utilisation du pyroséquençage 454.

Le coût des analyses moléculaires est un facteur important à prendre en compte dans le cadre de la bioindication. Le coût de la méthode mise en place au cours de la thèse est nettement supérieur au coût actuel d'une analyse morphologique. Cependant, après une

optimisation méthodologique, celle-ci pourrait être aisément automatisée (via l'utilisation de la robotique de laboratoire), afin de traiter de nombreux échantillons en parallèle, ce qui diminuerait considérablement le coût de ces analyses. De plus, la diminution des coûts de séquençage permet d'envisager d'atteindre des coûts comparables pour les méthodes morphologiques et moléculaires.

Les diatomées ne sont pas les seuls bioindicateurs pour lesquelles de nouvelles approches de bioindications sont recherchées. En effet, des travaux sont également réalisés pour développer de telles approches pour l'analyse de la biodiversité des macro-invertébrés. En effet, tout comme pour les diatomées, l'identification, basée sur la morphologie des taxa bioindicateurs, prend du temps, et n'atteint que rarement une résolution au niveau de l'espèce. Récemment, Sweeney et al. (2011) ont démontré l'efficacité du barcoding ADN pour évaluer la biodiversité des macro-invertébrés aquatiques et la qualité de l'eau. Ensuite, Hajibabaei et al. (2011) ont confirmé les capacités du pyroséquençage 454 pour identifier avec précision les espèces de macro-invertébrés d'eau douce présentant plus de 1% d'abondance dans un mélange. Comme pour les diatomées, des efforts sont nécessaires pour optimiser les outils moléculaires. Néanmoins, Pfrender et al. (2010) pensent qu'un effort coordonné entre taxonomistes, biologistes moléculaires, bioinformaticiens, écologues (d'eau douce), et gestionnaires permettrait de produire un outil moléculaire pour l'évaluation des systèmes d'eau douce par les macro-invertébrés, dans les 5 à 10 prochaines années.

Les résultats obtenus sur les macro-invertébrés, ainsi que nos résultats sur les diatomées démontrent le potentiel des techniques NGS pour les programmes de bioindication et l'intérêt de continuer les recherches dans cette voie.

4. Vers un nouvel outil moléculaire ?

Le barcoding environnemental n'est pas encore prêt pour réaliser les suivis environnementaux requis dans le cadre de la bioindication. Les principales sources de divergences entre l'approche classique et le barcoding environnemental sont la représentativité des bases de référence, les écarts entre taxonomie moléculaire et taxonomie morphologique, et les différences de persistance des frustules et de l'ADN

dans le milieu naturel. Ces différences de persistance induisent des erreurs essentiellement au niveau des espèces de faible abondance, qui ne sont pas utilisées pour le calcul de l'IBD (Prygiel & Coste, 2000) ou qui n'ont qu'un faible impact dans le calcul final des indices puisque c'est l'abondance relative qui est utilisée (e.g. Kelly & Whitton, 1995). Ainsi ce problème ne gêne pas l'utilisation du pyroséquençage 454 dans le cadre de la bioindication.

En revanche, nous avons évoqué régulièrement les limites des bases de référence. L'efficacité de l'identification est très dépendante de la représentativité de la base de référence : s'il n'y a pas de séquence de référence pour une espèce, les identifications de cette espèce ne peuvent pas aboutir. De plus, la méthodologie que nous avions mise en place était basée à la fois sur la taxonomie moléculaire, puisque nous utilisions des séquences, mais également sur la taxonomie morphologique puisque les séquences des bases de référence étaient identifiées en fonction de la morphologie des souches correspondantes. Comme nous l'avons précédemment indiqué, ces deux types de taxonomie ne peuvent pas être totalement concordants. Pour enrichir les bases de référence, il est nécessaire de séquencer de nombreuses cultures. Or, réaliser des cultures pour toute la diversité des diatomées est un travail long et laborieux. De plus, un important travail taxonomique serait nécessaire pour créer des bases de référence en accord avec les deux types de taxonomie. Pourtant, le développement de nouveaux outils de bioindication est urgent. Aussi, une méthode pouvant répondre plus rapidement aux attentes des gestionnaires de la qualité de l'eau serait plus appropriée que la méthode basée sur les séquences de référence reliées à la morphologie.

Une alternative serait d'élaborer un outil que ne serait plus basé sur la taxonomie morphologique. Il s'agirait de relier directement une séquence à une qualité de milieu. La même méthode de pyroséquençage pourrait être utilisée sur un marqueur particulier, tel qu'un fragment du *rbc*L, mais sans assigner un nom de taxa aux séquences. En effet, à partir des HERs déjà définies, il serait possible d'analyser en pyroséquençage des échantillons de référence (très peu impactés par les activités humaines) et d'observer les changements de séquences le long des gradients de perturbation comme cela a été réalisé pour le développement des indices basés sur la morphologie. Des coefficients de sensibilité et des coefficients indicateurs pourraient être appliqués aux séquences comme ils ont été appliqués aux taxa morphologiques. La base de référence pour cet outil contiendrait les séquences des reads (obtenus dans les échantillons de référence et le long

des gradients de perturbation) reliées aux coefficients utiles à la bioindication. Pour établir la qualité de l'eau d'un nouvel échantillon, il faudrait alors comparer les reads provenant de cet échantillon à la nouvelle base de référence pour assigner des informations écologiques aux reads. Un indice pourrait être calculé, tout comme l'IBD, en se basant uniquement sur les séquences ADN. Les seuils d'identité définis (Article V) pourraient être appliqués pour établir des correspondances entre les reads inconnus et la nouvelle base de référence. Ce sont les modifications de séquences qui renseigneraient sur les conditions environnementales à la place de la morphologie des frustules.

Bien qu'il soit dommage de ne pas utiliser la quantité considérable de données qui existent déjà sur l'écologie des diatomées, ce type d'indice éviterait de se heurter au problème des correspondances entre les données morphologiques et moléculaires. Les variations morphologiques peuvent être des réponses aux conditions environnementales et ainsi révéler une modification du milieu qui ne peut être détectée par un outil moléculaire. Cependant, cette perte d'information serait moindre comparée à la quantité d'informations supplémentaires apportées par l'approche moléculaire (populations présentant des préférences écologiques différentes) et qui ne pourrait être détectées par la morphologie.

Un tel outil, basé uniquement sur les séquences ADN, nécessitera encore de nombreuses mises au point, mais l'ensemble des travaux de la thèse constituent une base pour développer ce nouvel outil moléculaire.

BIBLIOGRAPHIE

A

ACINAS, S.G., SARMA-RUPAVTARM, R., KLEPAC-CERAJ, V. & POLZ, M.F. (2005) PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. *Applied and environmental Microbiology*, 71, 8966–8969.

AFNOR (2003) Qualité de l'eau - Guide pour l'échantillonnage en routine et le prétraitement des diatomées benthiques de rivières. Dans *NF EN 13946* p. 19.

AHMADIAN, A., EHN, M. & HOBER, S. (2006) Pyrosequencing: History, biochemistry and future. *Clinica Chimica Acta*, 363, 83–94.

ALGATERRA PROJECT. Algaterra project website. www.algaterra.org/ [consulté 11 novembre 2011].

ALLAN, J. (2004) Landscapes and riverscapes: The influence of land use on stream ecosystems. *Annual Review of Ecology Evolution and Systematics*, 35, 257–284.

ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W. & LIPMAN, D.J. (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215, 403–410.

ALTSCHUL, S.F., MADDEN, T.L., SCHÄFFER, A.A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25, 3389 –3402.

ALVAREZ, I. & WENDEL, J.F. (2003) Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution*, 29, 417–434.

ALVERSON, A.J. (2008) Molecular Systematics and the Diatom Species. *Protist*, 159, 339–353.

ALVERSON, A.J. & KOLNICK, L. (2005) Intragenomic nucleotide polymorphism among small subunit (18S) rDNA paralogs in the diatom genus *Skeletonema* (Bacillariophyta). *Journal of Phycology*, 41, 1248–1257.

AMATO, A., KOOISTRA, W.H.C.F., LEVIALDI GHIRON, J.H., MANN, D.G., PRÖSCHOLD, T. & MONTRESOR, M. (2007) Reproductive Isolation among Sympatric Cryptic Species in Marine Diatoms. *Protist*, 158, 193–207.

AMEND, A.S., SEIFERT, K.A. & BRUNS, T.D. (2010) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Molecular Ecology*, 19, 5555–5565.

ANDREE, K.B., FERNANDEZ-TEJEDOR, M., ELANDALOUSSI, L.M., QUIJANO-SCHEGGIA, S., SAMPEDRO, N., GARCES, E., ET AL. (2011) Quantitative PCR Coupled with Melt Curve Analysis for Detection of Selected *Pseudo-nitzschia* spp. (Bacillariophyceae) from the Northwestern Mediterranean Sea. *Applied and Environmental Microbiology*, 77, 1651–1659.

ANNEVILLE, O. & PELLETIER, J.P. (2000) Recovery of Lake Geneva from eutrophication: quantitative response of phytoplankton. *Archiv für Hydrobiologie*, 148, 607–624.

ANONYMOUS (1703) Two letters from a Gentleman in the Country, relating to Mr Leuwenhoeck's Letter in Transaction, no 283. *Philosophical Transactions of the Royal Society of London*, 23, 1494–1501.

ARMBRUST, E.V., BERGES, J.A., BOWLER, C., GREEN, B.R., MARTINEZ, D., PUTNAM, N.H., ET AL. (2004) The Genome of the Diatom *Thalassiosira pseudonana*: Ecology, Evolution, and Metabolism. *Science*, 306, 79–86.

ASCONIT CONSULTANTS (2008) Réalisation du suivi et validation des sites de référence pour les eaux douces (cours d'eau) de la Réunion. Dans DIREN Réunion, Trois Ilets.

B

BAARS, J.W.M. (1983) Autecological investigations on freshwater diatoms. 1. Generation times of some species. *Archiv für Hydrobiologie Supplement*, 67, 11–18.

BALDWIN, B.G. (1992) Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: An example from the compositae. *Molecular Phylogenetics and Evolution*, 1, 3–16.

BAROIN, A., PERASSO, R., QU, L.H., BRUGEROLLE, G., BACHELLERIE, J.P. & ADOUTTE, A. (1988) Partial phylogeny of the unicellular eukaryotes based on rapid sequencing of a portion of 28S ribosomal RNA. *Proceedings of the National Academy of Sciences*, 85, 3474 –3478.

BAVESTRELLO, G., ARILLO, A., CALCINAI, B., CATTANEO-VIETTI, R., CERRANO, C., GAINO, E., ET AL. (2000) Parasitic diatoms inside antarctic sponges. *The Biological Bulletin*, 198, 29 –33.

BEAVER, J. (1981) Apparent ecological characteristics of some common freshwater diatoms. Dans p. 550. Ontario Ministry of the Environment Report, Ontario Ministry of the Environment, Ontario.

BEHNKE, A., ENGEL, M., CHRISTEN, R., NEBEL, M., KLEIN, R.R. & STOECK, T. (2011) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environmental Microbiology*, 13, 340–349.

BEHNKE, A., FRIEDL, T., CHEPURNOV, V.A. & MANN, D.G. (2004) Reproductive compatibility and rDNA sequence analyses in the *Sellaphora pupula* species complex (Bacillariophyta). *Journal of Phycology*, 40, 193–208.

BELTRAMI, M., CAPPELLETTI, C., CIUTTI, F., HOFFMANN, L. & ECTOR, L. (2008) The diatom *Didymosphenia geminata*: distribution and mass occurrence in the province of Trento (Northern Italy). *Verhandlungen Internationale Vereinigung für theoretische und angewandte Limnologie*, 30, 593–597.

BELTRAMI, M.E., BLANCO, S., SCHEUDECKER, N., CIUTTI, F., CAPPELLETTI, C., MANCINI, L., ET AL. (2010) *Gomphonema vidalii* sp. nov. a new diatom from mediterranean streams. *Diatom Research*, 25, 29–42.

BÉRARD, A., DORIGO, U., HUMBERT, J.F. & MARTIN-LAURENT, F. (2005) Microalgae community structure analysis based on 18S rDNA amplification from DNA extracted directly from soil as a potential soil bioindicator. *Agronomy for Sustainable Development*, 25, 285–291.

BERE, T. & TUNDISI, J.G. (2011) Toxicity and sorption kinetics of dissolved cadmium and chromium III on tropical freshwater phytoperiphyton in laboratory mesocosm experiments. *Science of The Total Environment*, 409, 4772–4780.

BESSE-LOTOTSKAYA, A., VERDONSCHOT, P.F.M., COSTE, M. & VAN DE VIJVER, B. (2011) Evaluation of European diatom trophic indices. *Ecological Indicators*, 11, 456–467.

BESSE-LOTOTSKAYA, A., VERDONSCHOT, P.F.M. & SINKELDAM, J.A. (2006) Uncertainty in diatom assessment: sampling, identification and counting variation. *Hydrobiologia*, 566, 247–260.

BESZTERI, B., ACS, E., MAKK, J., KOVACS,, G., MARIALIGETI, K. & KISS, K.T. (2001) Phylogeny of six naviculoid diatoms based on 18S rDNA sequences. *International Journal of Systematic and Evolutionary Microbiology*, 51, 1581–1586.

BESZTERI, B., JOHN, U. & MEDLIN, L.K. (2007) An assessment of cryptic genetic diversity within the *Cyclotella meneghiniana* species complex (Bacillariophyta) based on nuclear and plastid genes, and amplified fragment length polymorphisms.

European Journal of Phycology, 42, 47–60.

BESZTERI, B., ACS, E. & MEDLIN, L.K. (2005) Conventional and geometric morphometric studies of valve ultrastructural variation in two closely related *Cyclotella* species (Bacillariophyta). *European Journal of Phycology*, 40, 89–103.

BLANDIN, P. (1986) Bioindicateurs et diagnostic des systèmes écologiques. *Bulletin d'écologie*, 17, 215–307.

BLAXTER, M. (2004) The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 359, 669–679.

BOO, S.M., KIM, H.S., SHIN, W., BOO, G.H., CHO, S.M., JO, B.Y., ET AL. (2010) Complex phylogeographic patterns in the freshwater alga *Synura* provide new insights into ubiquity vs. endemism in microbial eukaryotes. *Molecular Ecology*, 19, 4328–4338.

BORNET, B., ANTOINE, E., BARDOUIL, M. & MARCAILLOU-LE BAUT, C. (2004) ISSR as new markers for genetic characterization and evaluation of relationships among phytoplankton. *Journal of Applied Phycology*, 16, 285–290.

BOWLER, C., ALLEN, A.E., BADGER, J.H., GRIMWOOD, J., JABBARI, K., KUO, A., ET AL. (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature*, 456, 239–244.

BRUDER, K. & MEDLIN, L.K. (2007) Molecular assessment of phylogenetic relationships in selected species/genera in the naviculoid diatoms (Bacillariophyta). I. The genus *Placoneis*. *Nova Hedwigia*, 85, 331–352.

BRUDER, K. & MEDLIN, L.K. (2008) Morphological and molecular investigations of naviculoid diatoms. II. Selected genera and families. *Diatom Research*, 23, 283–329.

BRUDER, K., SATO, S. & MEDLIN, L.K. (2008) Morphological and molecular investigations of naviculoid diatoms IV. *Pinnularia* vs. *Caloneis*. *Diatom*, 24, 8–24.

BUÉE, M., REICH, M., MURAT, C., MORIN, E., NILSSON, R.H., UROZ, S. & MARTIN, F. (2009) 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist*, 184, 449–456.

ℂ

CANTONATI, M., LANGE-BERTALOT, H., SCALFI, A. & ANGELI, N. (2010) *Cymbella tridentina* sp. nov. (Bacillariophyta), a crenophilous diatom from carbonate springs of the Alps. *Journal of the North American Benthological Society*, 29, 775–788.

CARRIGG, C., RICE, O., KAVANAGH, S., COLLINS, G. & O'FLAHERTY, V. (2007) DNA extraction method affects microbial community profiles from soils and sediment. *Applied Microbiology and Biotechnology*, 77, 955–964.

CASAMAYOR, E.O., MASSANA, R., BENLLOCH, S., ØVREÅS, L., DÍEZ, B., GODDARD, V.J., ET AL. (2002) Changes in archaeal, bacterial and eukaryal assemblages along a salinity gradient by comparison of genetic fingerprinting methods in a multipond solar saltern. *Environmental Microbiology*, 4, 338–348.

CASAMAYOR, E.O., SCHAFER, H., BANERAS, L., PEDROS-ALIO, C. & MUYZER, G. (2000) Identification of and Spatio-Temporal Differences between Microbial Assemblages from Two Neighboring Sulfurous Lakes: Comparison by Microscopy and Denaturing Gradient Gel Electrophoresis. *Applied and Environmental Microbiology*, 66, 499–508.

CASTELEYN, G., CHEPURNOV, V.A., LELIAERT, F., MANN, D.G., BATES, S.S., LUNDHOLM,

N., ET AL. (2008) *Pseudo-nitzschia pungens (*Bacillariophyceae): A cosmopolitan diatom species? *Harmful Algae*, 7, 241–257.

CASTELEYN, G., LELIAERT, F., BACKELJAU, T., DEBEER, A.-E., KOTAKI, Y., RHODES, L., ET AL. (2010) Limits to gene flow in a cosmopolitan marine planktonic diatom. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 12952–12957.

CEMAGREF (1982) Etude des méthodes biologiques quantitatives d'appreciation de la qualite des eaux. Dans Cemagref, Pierre-Bénite p. 28.

CHASE, M.W., COWAN, R.S., HOLLINGSWORTH, P.M., VAN DEN BERG, C., MADRINAN, S., PETERSEN, G., ET AL. (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon*, 56, 295–299.

CHASE, M.W., SOLTIS, D.E., OLMSTEAD, R.G., MORGAN, D., LES, D.H., MISHLER, B.D., ET AL. (1993) Phylogenetics of Seed Plants: An Analysis of Nucleotide Sequences from the Plastid Gene *rbc*L. *Annals of the Missouri Botanical Garden*, 80.

CHEPURNOV, V.A., MANN, D.G., SABBE, K. & VYVERMAN, W. (2004) Experimental Studies on Sexual Reproduction in Diatoms, 237, 91 – 154.

CLEVE, P.T. (1894) Synopsis of the naviculoid diatoms. Part I. Stockholm.

COLEMAN, A.W. (2009) Is there a molecular key to the level of ''biological species" in eukaryotes? A DNA guide. *Molecular Phylogenetics and Evolution*, 50, 197–203.

COMPÈRE, P. (1995) *Gomphonema zairense* sp. nov. from the Tshopo waterfalls (Kisangani, Zaire). *Diatom Research*, 10, 31–37.

COOKE, D.E.L. & DUNCAN, J.M. (1997) Phylogenetic analysis of *Phytophthora* species based on ITS1 and ITS2 sequences of the ribosomal RNA gene repeat. *Mycological Research*, 101, 667–677.

COX, E.J. (2002) Diatoms: the evolution of morphogenetic complexity in single-celled plants. Dans *Developmental Genetics and Plant Evolution* p. 459–492Taylor & Francis. Q. C. B. Cronk, R. M. Bateman and J. A. Hawkins, London.

COX, E.J. (2009) What's in a name? – Diatom classification should reflect systematic relationships. *Acta Botanica Croatica*, 68, 443–454.

COX, E.J. (2010) Morphogenetic information and the selection of taxonomic characters for raphid diatom systematics. *Plant Ecology and Evolution*, 143, 271–277.

COX, E.J. (2011) Morphology, Cell Wall, Cytology, Ultrastructure and Morphogenetic Studies Overview and Specific Observations. Dans *The Diatom World, Cellular Origin, Life in Extreme Habitats and Astrobiology* p. 21–45. J. Seckbach and J.P. Kociolek.

COX, E.J. & WILLIAMS, D. (2000) Systematics of naviculoid diatoms: the interrelationships of some taxa with a stauros. *European Journal of Phycology*, 35, 273–282.

COX, E.J. & WILLIAMS, D.M. (2006) Systematics of naviculoid diatoms (Bacillariophyta): a preliminary analysis of protoplast and frustule characters for family and order level classification. *Systematics and Biodiversity*, 4, 385–399.

CRÉACH, V., ERNST, A., SABBE, K., VANELSLANDER, B., VYVERMAN, W. & STAL, L.J. (2006) Using quantitative PCR to determine the distribution of a semicryptic benthic diatom, *Navicula phyllepta* (Bacillariophyceae). *Journal of Phycology*, 42, 1142–1154.

CRISTEA-FERNSTRÖM, M., OLOFSSON, M., CHRYSSANTHOU, E., JONASSON, J. & PETRINI, B. (2007) Pyrosequencing of a short hypervariable 16S rDNA fragment for the identification of nontuberculous mycobacteria—a comparison with conventional 16S rDNA sequencing and phenotyping. *Acta Pathologica, Microbiologica et Immunologica Scandinavica*, 115, 1252–1259.

D

VAN DAM, H., MERTENS, A. & SINKELDAM, J. (1994) A coded checklist and ecological indicator values of freshwater diatoms from The Netherlands. *Aquatic Ecology*, 28, 117–133.

VON DASSOW, P. & MONTRESOR, M. (2011) Unveiling the mysteries of phytoplankton life cycles: patterns and opportunities behind complexity. *Journal of Plankton Research*, 33, 3 –12.

DAUGBJERG, N. & ANDERSEN, R.A. (1997a) A molecular phylogeny of the Heterokont algae based on analyses of chloroplast-encoded *rbc*L sequence data. *Journal of Phycology*, 33, 1031–1041.

DAUGBJERG, N. & ANDERSEN, R.A. (1997b) Phylogenetic Analyses of the *rbc*L Sequences from Haptophytes and Heterokont Algae Suggest Their Chloroplasts are Unrelated. *Molecular Biology and Evolution*, 14, 1242–1251.

DAWSON, P.A. (1972) Observations on the structure of some forms of *Gomphonema parvulum* Kütz. *British Phycological Journal*, 7, 255–271.

DAWSON, P.A. (1974) Observations on diatom species transferred from *Gomphonema* C. A. agardh to *Gomphoneis* Cleve. *British Phycological Journal*, 9, 75–82.

DEBENEST, T., GAGNÉ, F., PETIT, A.-N., KOHLI, M., EULLAFROY, P. & BLAISE, C. (2010) Monitoring of a flame retardant (tetrabromobisphenol A) toxicity on different microalgae assessed by flow cytometry. *Journal of Environmental Monitoring.*, 12, 1918–1923.

DEJEAN, T., VALENTINI, A., DUPARC, A., PELLIER-CUIT, S., POMPANON, F., TABERLET, P. & MIAUD, C. (2011) Persistence of environmental DNA in freshwater ecosystems. *PLoS One*, 6, e23398.

VAN DER AUWERA, G. & DE WACHTER, R. (1998) Structure of the Large Subunit rDNA from a Diatom, and Comparison Between Small and Large Subunit Ribosomal RNA for Studying Stramenopile Evolution. *Journal of Eukaryotic Microbiology*, 45, 521–527.

DEREEPER, A., GUIGNON, V., BLANC, G., AUDIC, S., BUFFET, S., CHEVENET, F., ET AL. (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research*, 36, W465–469.

DESANTIS, T.Z., STONE, C.E., MURRAY, S.R., MOBERG, J.P. & ANDERSEN, G.L. (2005) Rapid quantification and taxonomic classification of environmental DNA from both prokaryotic and eukaryotic origins using a microarray. *FEMS Microbiology Letters*, 245, 271–278.

DEWALT, R.E. (2011) DNA barcoding: a taxonomic point of view. *Journal of the North American Benthological Society*, 30, 174–181.

DÍEZ, B., PEDRÓS-ALIÓ, C., MARSH, T.L. & MASSANA, R. (2001) Application of Denaturing Gradient Gel Electrophoresis (DGGE) To Study the Diversity of Marine Picoeukaryotic Assemblages and Comparison of DGGE with Other Molecular Techniques. *Applied and Environmental Microbiology*, 67, 2942–2951.

DÍEZ, B., PEDRÓS-ALIÓ, C. & MASSANA, R. (2001) Study of Genetic Diversity of Eukaryotic Picoplankton in Different Oceanic Regions by Small-Subunit rRNA Gene Cloning and Sequencing. *Applied and Environmental Microbiology*, 67, 2932–2941.

DOOLITTLE, W.F. (1999) Phylogenetic Classification and the Universal Tree. *Science*, 284, 2124 –2128.

DORIGO, U., FONTVIEILLE, D. & HUMBERT, J.-F. (2006) Spatial variability in the abundance and composition of the free-living bacterioplankton community in the

pelagic zone of Lake Bourget (France). *FEMS Microbiology Ecology*, 58, 109–119.

DORIGO, U., VOLATIER, L. & HUMBERT, J.-F. (2005) Molecular approaches to the assessment of biodiversity in aquatic microbial communities. *Water Research*, 39, 2207–2218.

DREBES, G. (1977) Sexuality. Dans *The Biology of Diatoms* p. 250–283Werner D. Blackwell Scientific Publications, Oxford.

DRUM, R. & HOPKINS, J. (1966) Diatom locomotion: An explanation. *Protoplasma*, 62, 1–33.

DUKE, E.L. & REIMANN, B.E.F. (1977) The ultrastructure of the diatom cell. Dans *The Biology of Diatoms* p. 65–109Werner D. Blackwell Scientific Publications, Oxford.

DUONG, T.T., MORIN, S., COSTE, M., HERLORY, O., FEURTET-MAZEL, A. & BOUDOU, A. (2010) Experimental toxicity and bioaccumulation of cadmium in freshwater periphytic diatoms in relation with biofilm maturity. *Science of The Total Environment*, 408, 552–562.

E

ECTOR, L. & RIMET, F. (2005) Using bioindicators to assess rivers in Europe: An overview. Dans *Modelling community structure in aquatic ecosystems.* p. 7–19Lek S., Scardi M., Verdonschot P.F.M., Descy J.-P., Park Y.S. Springer, Verlag.

EDGAR, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26, 2460-2461.

EDGAR, S.M. & THERIOT, E.C. (2004) Phylogeny of *Aulacoseira* (Bacillariophyta) based on molecules and morphology. *Journal of Phycology*, 40, 772–788.

EDLUND, M.B. & STOERMER, E.F. (1997) Ecological, evolutionary, and systematic significance of diatom life histories. *Journal of Phycology*, 33, 897–918.

EHARA, M., INAGAKI, Y., WATANABE, K.I. & OHAMA, T. (2000) Phylogenetic analysis of diatom *cox*I genes and implications of a fluctuating GC content on mitochondrial genetic code evolution. *Current Genetics*, 37, 29–33.

ELLEGAARD, M., GODHE, A., HÄRNSTRÖM, K. & MCQOID, M. (2008) The species concept in a marine diatom: LSU rDNA–based phylogenetic differentiation in *Skeletonema marinoi /dohrnii* (Bacillariophyceae) is not reflected in morphology. *Phycologia*, 42, 156–167.

ELWOOD, H.J., OLSEN, G.J. & SOGIN, M.L. (1985) The small-subunit ribosomal RNA gene sequences from the hypotrichous ciliates *Oxytricha nova* and *Stylonychia pustulata*. *Molecular Biology and Evolution*, 2, 399 –410.

ENGELBREKTSON, A., KUNIN, V., WRIGHTON, K.C., ZVENIGORODSKY, N., CHEN, F., OCHMAN, H. & HUGENHOLTZ, P. (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J*, 4, 642–647.

EUROPEAN COMMITTEE FOR STANDARDIZATION (2003) Water quality – Guidance standard for the routine sampling and pre-treatment of benthic diatoms from rivers. European Standard EN 13946. Dans p. 14. European Committee for Standardization,, Brussels.

EVANS, K.M., CHEPURNOV, V.A., SLUIMAN, H.J., THOMAS, S.J., SPEARS, B.M. & MANN, D.G. (2009) Highly Differentiated Populations of the Freshwater Diatom *Sellaphora capitata* Suggest Limited Dispersal and Opportunities for Allopatric Speciation. *Protist*, 160, 386–396.

EVANS, K.M. & MANN, D.G. (2009) A proposed protocol for nomenclaturally effective

DNA barcoding of microalgae. *Phycologia*, 48, 70–74.

EVANS, K.M., WORTLEY, A.H. & MANN, D.G. (2007) An Assessment of Potential Diatom ‘‘Barcode’’ Genes (*cox*1, *rbc*L, 18S and ITS rDNA) and their Effectiveness in Determining Relationships in *Sellaphora* (Bacillariophyta). *Protist*, 158, 349–364.

EVANS, K.M., WORTLEY, A.H., SIMPSON, G.E., CHEPURNOV, V.A. & MANN, D.G. (2008) A molecular systematic approach to explore diversity within the *Sellaphora pupula* species complex (Bacillariophyta). *Journal of Phycology*, 44, 215–231.

F

FALASCO, E., BLANCO, S., BONA, F., GOMÀ, J., HLÚBIKOVÁ, D., NOVAIS, M.H., ET AL. (2009) Taxonomy, morphology and distribution of the *Sellaphora stroemii* complex (Bacillariophyceae). *Fottea*, 9, 243–256.

FALASCO, E., BONA, F., BADINO, G., HOFFMANN, L. & ECTOR, L. (2009) Diatom teratological forms and environmental alterations: a review. *Hydrobiologia*, 623, 1–35.

FALCIATORE, A. & BOWLER, C. (2002) Revealing the molecular secrets of marine diatoms. *Annual Review of Plant Biology*, 53, 109–130.

FEINSTEIN, L.M., SUL, W.J. & BLACKWOOD, C.B. (2009) Assessment of Bias Associated with Incomplete Extraction of Microbial DNA from Soil. *Applied and Environmental Microbiology*, 75, 5428–5433.

FELSENSTEIN, J. (2005) PHYLIP (Phylogeny Inference Package). Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

FICETOLA, G.F., COISSAC, E., ZUNDEL, STÉPHANIE, RIAZ, T., SHEHZAD, W., BESSIÈRE, J., ET AL. (2010) An *In silico* approach for the evaluation of DNA barcodes. *BMC Genomics*, 11, 434.

FINLAY, B.J. & FENCHEL, T. (2004) Cosmopolitan Metapopulations of Free-Living Microbial Eukaryotes. *Protist*, 155, 237–244.

FINLAY, B.J., MONAGHAN, E.B. & MABERLY, S.C. (2002) Hypothesis: The Rate and Scale of Dispersal of Freshwater Diatom Species is a Function of their Global Abundance. *Protist*, 153, 261–273.

FOLEY, J.A., DEFRIES, R., ASNER, G.P., BARFORD, C., BONAN, G., CARPENTER, S.R., ET AL. (2005) Global Consequences of Land Use. *Science*, 309, 570 –574.

FOURTANIER, E. & KOCIOLEK, J.P. (2011) Catalogue of diatom names [online]. Http://research.calacademy.org/ research/diatoms/names/index.asp [consulté 10 février 2011].

FOX, M.G. & SORHANNUS, U.M. (2003) *Rpo*A: A Useful Gene for Phylogenetic Analysis in Diatoms. *Journal of Eukaryotic Microbiology*, 50, 471–475.

FRÁNKOVÁ, M., POULÍČKOVÁ, A., NEUSTUPA, J., PICHRTOVÁ, M. & MARVAN, P. (2009) Geometric morphometrics - a sensitive method to distinguish diatom morphospecies: a case study on the sympatric populations of *Reimeria sinuata* and *Gomphonema tergestinum* (Bacillariophyceae) from the River Bečva, Czech Republic. *Nova Hedwigia*, 88, 81–95.

G

GALAND, P.E., CASAMAYOR, E.O., KIRCHMAN, D.L. & LOVEJOY, C. (2009) Ecology of the rare microbial biosphere of the Arctic Ocean. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 22427–22432.

GARDNER, T.A., BARLOW, J., ARAUJO, I.S., ÁVILA-PIRES, T.C., BONALDO, A.B., COSTA, J.E., ET AL. (2008) The cost-effectiveness of biodiversity surveys in tropical forests. *Ecology Letters*, 11, 139–150.

GASCUEL, O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14, 685 –695.

GEITLER, L. (1932) Der Formwechsel der pennaten Diatomeen (Kieselalgen). Fischer, Jena.

GERMAIN, H. (1981) Flore des diatomées. Diatomophycées. Eaux douces et saumâtres du Massif armoricain et des contrées voisines d'Europe occidentale. Société nouvelle des éditions Boubée, Paris.

GILLES, A., MEGLECZ, E., PECH, N., FERREIRA, S., MALAUSA, T. & MARTIN, J.-F. (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, 12, 245.

GIONGO, A., CRABB, D.B., DAVIS-RICHARDSON, A.G., CHAULIAC, D., MOBBERLEY, J.M., GANO, K.A., ET AL. (2010) PANGEA: pipeline for analysis of next generation amplicons. *ISME J*, 4, 852–861.

GODHE, A., ASPLUND, M.E., HÄRNSTRÖM, K., SARAVANAN, V., TYAGI, A. & KARUNASAGAR, I. (2008) Quantification of Diatom and Dinoflagellate Biomasses in Coastal Marine Seawater Samples by Real-Time PCR. *Applied and Environmental Microbiology*, 74, 7174–7182.

GONZALEZ, J.M., PORTILLO, M.C., BELDA-FERRE, P. & MIRA, A. (2012) Amplification by PCR Artificially Reduces the Proportion of the Rare Biosphere in Microbial Communities. *PLoS ONE*, 7, e29973.

GUELORGET, O. & PERTHUISOT, J.P. (1984) Indicateurs biologiques et diagnose écologique dans le domaine paralique. *Bulletin d'écologie*, 15, 67–76.

GUILLARD, R.R.L. & LORENZEN, C.J. (1972) Yellow-green algae with chlorophyllide c. *Journal of Phycology*, 8, 10–14.

GUINDON, S. & GASCUEL, O. (2003) A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, 52, 696 –704.

GUSFIELD, D. (1997) Algorithms on Strings, Trees and Sequences. Computer Science and Computational Biology. Cambridge University Press., New York.

H

HAJIBABAEI, M., DEWAARD, J.R., IVANOVA, N.V., RATNASINGHAM, S., DOOH, R.T., KIRK, S.L., ET AL. (2005) Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360, 1959 –1967.

HAJIBABAEI, M., SHOKRALLA, S., ZHOU, X., SINGER, G.A.C. & BAIRD, D.J. (2011) Environmental Barcoding: A Next-Generation Sequencing Approach for Biomonitoring Applications Using River Benthos. *PLoS One*, 6, e17497.

HAJIBABAEI, M., SINGER, G.A.C., HEBERT, P.D.N. & HICKEY, D.A. (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in genetics*, 23, 167–172.

HALL, T. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41, 95–98.

HAMADY, M., LOZUPONE, C. & KNIGHT, R. (2010) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME Journal*, 4, 17–27.

HAMMER, Ø., HARPER, D.A.T. & RYAN, P.D. (2001) PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica*, 4, 9 pp.

HAMSHER, S.E., EVANS, K.M., MANN, D.G., POULÍČKOVÁ, A. & SAUNDERS, G.W. (2011) Barcoding Diatoms: Exploring Alternatives to COI-5P. *Protist*, 162, 405–422.

VAN HANNEN, E.J., VAN AGTERVELD, M.P., GONS, H.J. & LAANBROEK, H.J. (1998) Revealing genetic diversity of eukaryotic microorganisms in aquatic environments by denaturing gradient gel electrophoresis. *Journal of Phycology*, 34, 206–213.

HEBERT, P.D.N., CYWINSKA, A., BALL, S.L. & DEWAARD, J.R. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B*, 270, 313–321.

HEBERT, P.D.N., STOECKLE, M.Y., ZEMLAK, T.S. & FRANCIS, C.M. (2004) Identification of Birds through DNA Barcodes. *Plos Biology*, 2, e312.

HEINO, J., BINI, L.M., KARJALAINEN, S.M., MYKRÄ, H., SOININEN, J., VIEIRA, L.C.G. & DINIZ-FILHO, J.A.F. (2010) Geographical patterns of micro-organismal community structure: are diatoms ubiquitously distributed across boreal streams? *Oikos*, 119, 129–137.

HELLAWELL, J.M. (1978) Biological surveillance of rivers: a biological monitoring handbook. Water Research Centre.

HELLEBUST, J.A. & LEWIN, J. (1977) Heterotrophic nutrition. Dans *The Biology of Diatoms* p. 169–197Werner D. Blackwell Scientific Publications, Oxford.

HILLEBRAND, H. & SOMMER, U. (1997) Response of epilithic microphytobenthos of the Western Baltic Sea to in situ experiments with nutrient enrichment. *Marine Ecology Progress Series*, 160, 35–46.

HILLIS, D.M. & DIXON, M. (1991) Ribosomal DNA: Molecular Evolution and phylogenetic inference. *Quarterly Review of Biology*, 66, 411–453.

HILLIS, D.M., MORITZ, C. & MABLE, B. (1996) Molecular Systematics. 2nd Edition. Sinauer Associates, Sunderland, Massachusetts.

HOLLAND, R.A., EIGENBROD, F., ARMSWORTH, P.R., ANDERSON, B.J., THOMAS, C.D., HEINEMEYER, A., ET AL. (2011) Spatial covariation between freshwater and terrestrial ecosystem services. *Ecological Applications*, 21, 2034–2048.

HOLT, E.A. & MILLER, S.W. (2011) Bioindicators: Using organisms to measure environmental impacts. *Nature Education Knowledge*, 2, 8.

HUBER, J.A., MORRISON, H.G., HUSE, S.M., NEAL, P.R., SOGIN, M.L. & MARK WELCH, D.B. (2009) Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. *Environmental Microbiology*, 11, 1292–1302.

HUDSON, M.E. (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, 8, 3–17.

HUGHES, J.C. & LUND, J.W.G. (1962) The rate of growth of *Asterionella formosa* Hass. in relation to its ecology. *Archiv für Mikrobiologiet*, 42, 117–129.

HUMBERT, J.-F., DORIGO, U., CECCHI, P., LE BERRE, B., DEBROAS, D. & BOUVY, M. (2009) Comparison of the structure and composition of bacterial communities from temperate and tropical freshwater ecosystemsemi_1960 2339. *Environmental Microbiology*, 11, 2339–2350.

HUSON, D.H., AUCH, A.F., QI, J. & SCHUSTER, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Research*, 17, 377 –386.

HYMAN, E.D. (1988) A new method of sequencing DNA. *Analytical Biochemistry*, 174, 423–436.

I

IVORRA, N., BARRANGUET, C., JONKER, M., KRAAK, M.H.S. & ADMIRAAL, W. (2002) Metal-induced tolerance in the freshwater microbenthic diatom *Gomphonema parvulum*. *Environmental Pollution*, 116, 147–157.

J

JAHN, R., MANN, D.G., EVANS, K.M. & POULÍČKOVÁ, A. (2008) The identity of *Sellaphora bacillum* (Ehrenberg) D.G. Mann. *Fottea*, 8, 121–124.

JALBA, A.C., WILKINSON, M.H.F., ROERDINK, J.B.T.M., BAYER, M.M. & JUGGINS, S. (2005) Automatic diatom identification using contour analysis by morphological curvature scale spaces. *Machine Vision and Applications*, 16, 217–228.

JOHN, J. (2000) A guide to diatoms as indicators of urban stream health. *LWRRDC Occasional Paper*, 14/99 (Urban Sub Program, Report No.7).

JONES, H.M., SIMPSON, G.E., STICKLE, A.J. & MANN, D.G. (2005) Life history and systematics of *Petroneis* (Bacillariophyta), with special reference to British waters. *European Journal of Phycology*, 40, 61–87.

JUKES, T. & CANTOR, C. (1969) Evolution of Protein Molecules. Academy Press.

JULIUS, M.L. & THERIOT, E.C. (2010) The diatoms: a primer. Dans *The Diatoms: Applications for the Environmental and Earth Sciences* p. 8–22Cambridge University Press. John P. Smol, Eugene F. Stoermer.

JUMPPONEN, A. (2007) Soil Fungal Communities Underneath Willow Canopies on a Primary Successional Glacier Forefront: rDNA Sequence Results Can Be Affected by Primer Selection and Chimeric Data. *Microbial Ecology*, 53, 233–246.

K

KACZMARSKA, I., REID, C., MARTIN, J.L. & MONIZ, M.B.J. (2008) Morphological, biological, and molecular characteristics of the diatom *Pseudo-nitzschia delicatissima* from the Canadian maritimes. *Botany*, 86, 763–772.

KAHLERT, M., ALBERT, R.-L., ANTTILA, E.-L., BENGTSSON, R., BIGLER, C., ESKOLA, T., ET AL. (2009) Harmonization is more important than experience – results of the first Nordic-Baltic diatom intercalibration exercise 2007 (stream monitoring). *Journal of Applied Phycology*, 21, 471–482.

KANAGAWA, T. (2003) Bias and Artifacts in Multitemplate Polymerase Chain Reactions (PCR). *Journal of Bioscience and Bioengineering*, 96, 317–323.

KELLY, L.J., HOLLINGSWORTH, P.M., COPPINS, B.J., ELLIS, C.J., HARROLD, P., TOSH, J. & YAHR, R. (2011) DNA barcoding of lichenized fungi demonstrates high identification success in a floristic context. *New Phytologist*, 191, 288–300.

KELLY, M.G. & WHITTON, B.A. (1995) The Trophic Diatom Index: a new index for monitoring eutrophication in rivers. *Journal of Applied Phycology*, 7, 433–444.

KENT, W.J. (2002) BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12, 656 –664.

KERMARREC, L., ECTOR, L., BOUCHEZ, A., RIMET, F. & HOFFMANN, L. (2011) A preliminary phylogenetic analysis of the Cymbellales based on 18S rDNA gene sequencing. *Diatom Research*, 26, 305–315.

KHAN, I.S.A.N. (1990) Assessment of Water Pollution using Diatom Community

Structure and Species Distribution — A Case Study in a Tropical River Basin. *Internationale Revue der gesamten Hydrobiologie und Hydrographie*, 75, 317–338.

KILROY, C., BIGGS, B.J.F. & VYVERMAN, W. (2007) Rules for macroorganisms applied to microorganisms: patterns of endemism in benthic freshwater diatoms. *Oikos*, 116, 550–564.

KOCIOLEK, J.P. (2011) Microscopic in size: macroscopic in impact. Diatom-human interactions. Dans p. 257–283. Springer Netherlands.

KOCIOLEK, J.P. & SPAULDING, S. (2000) Freshwater diatom biogeography. *Nova Hedwigia*, 71, 223–241.

KOCIOLEK, J.P. & SPAULDING, S.A. (2003) Eunotioid and asymmetrical naviculoid diatoms. Dans *Freshwater Algae of North America*. p. 655–668. Academic Press. Wehr J.D.; Sheath R.G., London.

KOCIOLEK, J.P. & STOERMER, E.F. (1988) A preliminary investigation of the phylogenetic relationships among the freshwater, apical pore field-bearing Cymbelloid and Gomphonemoid diatoms (Bacillariophyceae). *Journal of Phycology*, 24, 377–385.

KOCIOLEK, J.P. & STOERMER, E.F. (1989a) Chromosome numbers in diatoms: a review. *Diatom Research*, 4, 47–54.

KOCIOLEK, J.P. & STOERMER, E.F. (1989b) Phylogenetic relationships and evolutionary history of the diatom genus Gomphoneis. *Phycologia*, 28, 438–454.

KOCIOLEK, J.P. & STOERMER, E.F. (1993) Freshwater gomphonemoid diatom phylogeny: preliminary results. *Hydrobiologia*, 269/270, 31–38.

KOCIOLEK, J.P. & STOERMER, E.F. (2010) Variation and polymorphism in diatoms: the triple helix of development, genetics and environment. A review of the literature. *Vie Et Milieu-Life and Environment*, 60, 75–87.

KOLKWITZ, R. & MARSSON, M. (1908) Ökologie der pflanzlichen Saprobien. Borntraeger, Berlin.

KONFIRST, M., SJUNNESKOG, C., SCHERER, R. & DORAN, P. (2011) A diatom record of environmental change in Fryxell Basin, Taylor Valley, Antarctica, late Pleistocene to present. *Journal of Paleolimnology*, 46, 257–272.

KOOISTRA, W.H.C.F. & MEDLIN, L.K. (1996) Evolution of the Diatoms (Bacillariophyta) IV. A Reconstruction of Their Age from Small Subunit rRNA Coding Regions and the Fossil Record. *Molecular Phylogenetics and Evolution*, 6, 391–407.

KOOISTRA, W.H.C.F., SARNO, D., BALZANO, S., GU, H., ANDERSEN, R.A. & ZINGONE, A. (2008) Global Diversity and Biogeography of *Skeletonema* Species (Bacillariophyta). *Protist*, 159, 177–193.

KOOISTRA, W.H.C.F., SARNO, D., HERNANDEZ-BECERRIL, D.U., ASSMY, P., DI PRISCO, C. & MONTRESOR, M. (2010) Comparative molecular and morphological phylogenetic analyses of taxa in the Chaetocerotaceae (Bacillariophyta). *Phycologia*, 49, 471–500.

KOOISTRA, W.H.C.F., DE STEFANO, M., MANN, D.G., SALMA, N. & MEDLIN, L.K. (2003) Phylogenetic position of *Toxarium*, a pennate-like lineage within centric diatoms (Bacillariophyceae). *Journal of Phycology*, 39, 185–197.

KOSKINEN, K., HULTMAN, J., PAULIN, L., AUVINEN, P. & KANKAANPÄÄ, H. (2011) Spatially differing bacterial communities in water columns of the northern Baltic Sea. *FEMS Microbiology Ecology*, 75, 99–110.

KOVÁCS, C., KAHLERT, M. & PADISÁK, J. (2006) Benthic diatom communities along pH and TP gradients in Hungarian and Swedish streams. *Journal of Applied Phycology*, 18, 105–117.

KRAMMER, K. (2002) *Cymbella*.A.R.G. Gantner Verlag K.G. Horst Lange-Bertalot.

KRAMMER, K. (2003) *Cymbopleura*, *Delicata*, *Navicymbula*, *Gomphocymbellopsis*, *Afrocymbella* A.R.G. Gantner Verlag K.G. Horst Lange-Bertalot.

KRAMMER, K. & LANGE-BERTALOT, H. (1985) Naviculaceae: neue und wenig bekannte Taxa. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart.

KRAMMER, K. & LANGE-BERTALOT, H. (1988) Süsswasserflora von Mitteleuropa. Bacillariophyceae, Part 2. Teil: Bacillariaceae, Epithemiaceae, Surirellaceae. Fischer, Stuttgart.

KRAMMER, K. & LANGE-BERTALOT, H. (1991) Bacillariophyceae. 4. Teil: Achnanthaceae. Kritische Ergänzungen zu *Navicula* (Lineolatae) und *Gomphonema*. Gustav Fischer Verlag.

KRESS, W.J. & ERICKSON, D.L. (2007) A Two-Locus Global DNA Barcode for Land Plants: The Coding *rbc*L Gene Complements the Non-Coding *trn*H-*psb*A Spacer Region. *PLoS ONE*, 2, e508.

KRESS, W.J., WURDACK, K.J., ZIMMER, E.A., WEIGT, L.A. & JANZEN, D.H. (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 8369 –8374.

KROES, I., LEPP, P.W. & RELMAN, D.A. (1999) Bacterial diversity within the human subgingival crevice. *Proceedings of the National Academy of Sciences*, 96, 14547 –14552.

KUNIN, V., ENGELBREKTSON, A., OCHMAN, H. & HUGENHOLTZ, P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, 12, 118–123.

L

LASLANDES, B., SYLVESTRE, F., SIFEDDINE, A., TURCQ, B., ALBUQUERQUE, A.L.S. & ABRÃO, J. (2006) Enregistrement de la variabilité hydroclimatique au cours des 6500 dernières années sur le littoral de Cabo Frio (Rio de Janeiro, Brésil). *Comptes Rendus Geoscience*, 338, 667–675.

LAVOIE, I., CAMPEAU, S., GRENIER, M. & DILLON, P.J. (2006) A diatom-based index for the biological assessment of eastern Canadian rivers: an application of correspondence analysis (CA). *Canadian journal of fisheries and aquatic sciences*, 63, 1793–1811.

LECOINTE, C., COSTE, M. & PRYGIEL, J. (1993) « Omnidia »: software for taxonomy, calculation of diatom indices and inventories management. *Hydrobiologia*, 269-270, 509–513.

LEE, J.J., MCENERGY, M.E., TER KUILE, B., EREZ, J., ROETGER, R., ROCKWELL, R.F., ET AL. (1989) Identification and distribution of endosymbiotic diatoms in larger foraminifera. *Micropaleontology*, 35, 353–366.

LENOIR, A. & COSTE, M. (1996) Develpopment of a practical diatom index of overall water quality applicable to the French National Water Board Network. Dans *Use of Algae for monitoring rivers II* p. 29–45. Whitton, B. a., Rott, E., Institut für Botanik, Universität Innsbruck.

LETERME, S.C., ELLIS, A.V., MITCHELL, J.G., BUSCOT, M.-J., POLLET, T., SCHAPIRA, M. & SEURONT, L. (2010) Morphological flexibility of *Cocconeis placentula* (Bacillariophyceae) nanostructure to changing salinity level. *Journal of Phycology*, 46, 715–719.

LEVKOV, Z. & ECTOR, L. (2010) A comparative study of *Reimeria* species (Bacillariophyceae). *nova_hedwigia*, 90, 469–489.

LEVKOV, Z. & LANGE-BERTALOT, H. (2007) Diatoms of lakes Prespa and Ohrid: about

500 taxa from ancient lake system. A.R.G. Gantner.

LI, R., LI, Y., KRISTIANSEN, K. & WANG, J. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24, 713 –714.

LUDWIG, W., STRUNK, O., WESTRAM, R., RICHTER, L., MEIER, H., YADHUKUMAR, L., ET AL. (2004) ARB: a software environment for sequence data. *Nucleic Acids Research*, 32, 1363 –1371.

LUNDHOLM, N., DAUGBJERG, N. & MOESTRUP, Ø. (2002) Phylogeny of the Bacillariaceae with emphasis on the genus *Pseudo-nitzschia* (Bacillariophyceae) based on partial LSU rDNA. *European Journal of Phycology*, 37, 115–134.

LUNDHOLM, N., MOESTRUP, Ø., HASLE, G.R. & HOEF-EMDEN, K. (2003) A study of the *Pseudo-nitzschia pseudodelicatissima*/*cuspidata* complex (Baccillariophyceae): what is P. pseudodelicatissima? *Journal of Phycology*, 39, 797–813.

M

MACDONALD, J.D. (1869) I.—On the structure of the Diatomaceous frustule, and its genetic cycle. *Annals And Magazine of Natural History*, 3, 1–8.

MANDELIK, Y., ROLL, U. & FLEISCHER, A. (2010) Cost-efficiency of biodiversity indicators for Mediterranean ecosystems and the effects of socio-economic factors. *Journal of Applied Ecology*, 47, 1179–1188.

MANN, D.G. (1999) The species concept in Diatoms. *Phycologia*, 38, 437–495.

MANN, D.G. & DROOP, S.J.M. (1996) 3. Biodiversity, biogeography and conservation of diatoms. *Hydrobiologia*, 336, 19–32.

MANN, D.G. & EVANS (2007) Molecular genetics and the neglected art of diatomics. Dans *Unravelling the Algae – the past, present and future of algal molecular systematics* p. J. Brodie and J.M. Lewi.

MANN, D.G., SATO, S., TROBAJO, R., VANORMELINGEN, P. & SOUFFREAU, C. (2010) DNA barcoding for species identification and discovery in diatoms. *Cryptogamie Algologie*, 31, 557–577.

MARGULIES, M., EGHOLM, M., ALTMAN, W.E., ATTIYA, S., BADER, J.S., BEMBEN, L.A., ET AL. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437, 376–380.

MARIETTE, J., NOIROT, C. & KLOPP, C. (2011) Assessment of replicate bias in 454 pyrosequencing and a multi-purpose read-filtering tool. *BMC Research Notes*, 4, 149.

MARTIN, A.P. (2002) Phylogenetic Approaches for Describing and Comparing the Diversity of Microbial Communities. *Applied and Environmental Microbiology*, 68, 3673 –3682.

MARTIN-LAURENT, F., PHILIPPOT, L., HALLET, S., CHAUSSOD, R., GERMON, J.C., SOULAS, G. & CATROUX, G. (2001) DNA Extraction from Soils: Old Bias for New Microbial Diversity Analysis Methods. *Applied and Environmental Microbiology*, 67, 2354–2359.

MATSEN, F., KODNER, R. & ARMBRUST, E.V. (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11, 538.

MAYDEN, R.L. (1997) A hierarchy of species concepts: The denouement in the saga of the species problem. Dans *Species: The Units of Biodiversity.* (eds M.F. Claridge, H.A. Dawah & M.R. Wilson), p. 381–424Chapman & Hall. London.

MCCAIG, A.E., GLOVER, L.A. & PROSSER, J.I. (1999) Molecular Analysis of Bacterial Community Structure and Diversity in Unimproved and Improved Upland Grass

Pastures. *Applied and Environmental Microbiology*, 65, 1721 –1730.

MCNEILL, J., BARRIE, F.R., BURDET, H.M., DEMOULIN, V., HAWKSWORTH, D.L., MARHOLD, K., ET AL. (2006) International Code of Botanical Nomenclature (Vienna Code). A.R.G. Gantner, Verlag KG.

MEDLIN, L.K., ELWOOD, H.J., STICKEL, S. & SOGIN, M.L. (1988) The characterization of enzimatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene*, 71, 491–499.

MEDLIN, L.K. & KACZMARSKA, I. (2004) Evolution of the diatoms. V. Morphological and cytological support for the major clades and a taxonimic revision. *Phycologia*, 43, 245–270.

MEDLIN, L.K., METFIES, K., MEHL, H., WILTSHIRE, K. & VALENTIN, K. (2006) Picoeukaryotic Plankton Diversity at the Helgoland Time Series Site as Assessed by Three Molecular Methods. *Microbial Ecology*, 52, 53–71.

MEDLIN, L.K., WILLIAMS, D.M. & SIMS, P.A. (1993) The evolution of the diatoms (Bacillariophyta). I. Origin of the group and assessment of the monophyly of its major divisions. *European Journal of Phycology*, 28, 261–275.

MERINO, V., GARCÍA, J., HERNÁNDEZ-MARINÉ, M. & FERNÁNDEZ, M. (1994) Morphology and ultrastructure of *Gomphoneis rhombica* (Fricke) comb. nov.. *Diatom Research*, 9, 335–347.

MEYBECK, M. (2003) Global analysis of river systems: from Earth system controls to Anthropocene syndromes. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358, 1935 –1955.

MEYER, C.P. & PAULAY, G. (2005) DNA Barcoding: Error Rates Based on Comprehensive Sampling. *PLoS Biol*, 3, e422.

MILLET, L., GIGUET-COVEX, C., VERNEAUX, V., DRUART, J.-C., ADATTE, T. & ARNAUD, F. (2010) Reconstruction of the recent history of a large deep prealpine lake (Lake Bourget, France) using subfossil chironomids, diatoms, and organic matter analysis: towards the definition of a lake-specific reference state. *Journal of Paleolimnology*, 44, 963–978.

MONIZ, M.B.J. & KACZMARSKA, I. (2009) Barcoding diatoms: Is there a good marker? *Molecular Ecology Resources*, 9, 65–74.

MONIZ, M.B.J. & KACZMARSKA, I. (2010) Barcoding of Diatoms: Nuclear Encoded ITS Revisited. *Protist*, 161, 7–34.

MORIN, S., PROIA, L., RICART, M., BONNINEAU, C., GEISZINGER, A., RICCIARDI, F., ET AL. (2010) Effects of a bactericide on the structure and survival of benthic diatom communities. *Vie Et Milieu-Life and Environment*, 60, 109–116.

MULLIS, K., FALOONA, F., SCHARF, S., SAIKI, R., HORN, G. & ERLICH, H. (1986) Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, 51, 263 –273.

N

NAKOV, T. & THERIOT, E. (2009) Preliminary molecular phylogeny of the Cymbellales (Bacillariophyceae). Dans *Abstract book of the 20th North American diatom symposium* p. 28–29. Milford, IA.

NEBEL, M.E., WILD, S., HOLZHAUSER, M., HUTTENBERGER, L., REITZIG, R., SPERBER, M. & STOECK, T. (2011) Jaguc - a software package for environmental diversity analyses. *Journal of bioinformatics and computational biology*, 9, 749–773.

NGUYEN, T.N.M., BERZANO, M., GUALERZI, C.O. & SPURIO, R. (2011) Development of molecular tools for the detection of freshwater diatoms. *Journal of*

Microbiological Methods, 84, 33–40.

NOVAIS, M.H., BLANCO, S., HLÚBIKOVÁ, D., FALASCO, E., GOMÀ, J., DELGADO, C., ET AL. (2009) Morphological examination and biogeography of the *Gomphonema rosenstockianum* and *G. tergestinum* species complex (Bacillariophyceae). *Fottea*, 9, 257–274.

NYRÉN, P. (1987) Enzymatic method for continuous monitoring of DNA polymerase activity. *Analytical Biochemistry*, 167, 235–238.

O

OKSANEN, J., BLANCHET, F.G., KINDT, R., LEGENDRE, P., MINCHIN, P.R., O'HARA, R.B., ET AL. (2011) vegan: Community Ecology Package. R package version 2.0-1.

ORSINI, L., PROCACCINI, G., SARNO, D. & MONTRESOR, M. (2004) Multiple rDNA ITS-types within the diatom *Pseudo-nitzschia delicatissima* (Bacillariophyceae) and their relative abundances across a spring bloom in the Gulf of Naples. *Marine Ecology-Progress Series*, 271, 87–98.

OUDOT-LE SECQ, M.-P., GRIMWOOD, J., SHAPIRO, H., ARMBRUST, E., BOWLER, C. & GREEN, B. (2007) Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*: comparison with other plastid genomes of the red lineage. *Molecular Genetics and Genomics*, 277, 427–439.

P

PASSY, S.I., KOCIOLEK, J.P. & LOWE, R.L. (1997) Five new *Gomphonema* species (Bacillariophyceae) from rivers in South Africa and Swaziland. *Journal of Phycology*, 33, 455–474.

PATRICK, R. (1971) The effects of increasing light and temperature on the structure. *Limnology and Oceanography*, 16, 405–421.

PATRICK, R. (1977) Ecology of freshwater diatoms - Diatom communities. Dans *The Biology of Diatoms* p. 284–332Werner D. Blackwell Scientific Publications, Oxford.

PÉRÈS, F., COSTE, M., RICARD, F. & BOUDOU, A. (1997) Effects of methylmercury and inorganic mercury on periphytic diatom communities in freshwater indoor microcosms. *Journal of Applied Phycology*, 9, 215–227.

PÉRÈS, F., FLORIN, D., GROLLIER, T., FEURTET-MAZEL, A., COSTE, M., RIBEYRE, F., ET AL. (1996) Effects of the phenylurea herbicide isoproturon on periphytic diatom communities in freshwater indoor microcosms. *Environmental Pollution*, 94, 141–152.

PFITZER, E.H.H. (1871) Untersuchungen ŭber Bau und Entwicklung der Bacillariaceen (Diatomaceen). Bonn, Marcus.

PFRENDER, M.E., HAWKINS, C.P., BAGLEY, M., COURTNEY, G.W., CREUTZBURG, B.R., EPLER, J.H., ET AL. (2010) Assessing macroinvertebrate biodiversity in freshwater ecosystems: advances and challenges in DNA-based approaches. *Quarterly Review of Biology*, 85, 319–340.

POLZ, M.F. & CAVANAUGH, C.M. (1998) Bias in Template-to-Product Ratios in Multitemplate PCR. *Applied and Environmental Microbiology*, 64, 3724–3730.

POMMIER, T., CANBÄCK, B., RIEMANN, L., BOSTRÖM, K.H., SIMU, K., LUNDBERG, P., ET AL. (2007) Global patterns of diversity and community structure in marine bacterioplankton. *Molecular Ecology*, 16, 867–880.

POMPANON, F., COISSAC, E. & TABERLET, P. (2011) Metabarcoding, une nouvelle façon d'analyser la biodiversité. *Biofutur*, 319, 30–32.

PORAZINSKA, D.L., SUNG, W., ROBIN M. GIBLIN-DAVIS & THOMAS, W.K. (2010) Reproducibility of read numbers in high-throughput sequencing analysis of nematode community composition and structure. *Molecular Ecology Resources*, 10, 666–676.

POSADA, D. (2008) jModelTest: Phylogenetic Model Averaging. *Molecular Biology and Evolution*, 25, 1253 –1256.

POTAPOVA, M. & CHARLES, D.F. (2007) Diatom metrics for monitoring eutrophication in rivers of the United States. *Ecological Indicators*, 7, 48–70.

POTAPOVA, M. & HAMILTON, P.B. (2007) Morphological and ecological variation within the *Achnanthidium minutissimum* (Bacillariophyceae) species complex. *Journal of Phycology*, 43, 561–575.

POULÍČKOVÁ, A., VESELÁ, J., NEUSTUPA, J. & ŠKALOUD, P. (2010) Pseudocryptic Diversity versus Cosmopolitanism in Diatoms:a Case Study on *Navicula cryptocephala* Kütz. (Bacillariophyceae) and Morphologically Similar Taxa. *Protist*, 161, 353–369.

PRUESSE, E., QUAST, C., KNITTEL, K., FUCHS, B.M., LUDWIG, W., PEPLIES, J. & GLÖCKNER, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35, 7188 –7196.

PRYGIEL, J. (1991) Use of benthic diatoms in surveillance of the Artois-Picardie basin hydrobiological quality. Dans *Use of algae for monitoring rivers* p. . Whitton, B. A., Rott, E., Friedrich G.

PRYGIEL, J., CARPENTIER, P., ALMEIDA, S.F.P., COSTE, M., DRUART, J.-C., ECTOR, L., ET AL. (2002) Determination of the biological diatom index (IBD NF T 90–354): results of an intercomparison exercise. *Journal of Applied Phycology*, 14, 27–39.

PRYGIEL, J. & COSTE, M. (2000) Guide méthodologique pour la mise en oeuvre de l'Indice Biologique Diatomées NF T 90-354. Agence de l'eau Artois Picardie, Douai.

Q

QIU, X., WU, L., HUANG, H., MCDONEL, P.E., PALUMBO, A.V., TIEDJE, J.M. & ZHOU, J. (2001) Evaluation of PCR-Generated Chimeras, Mutations, and Heteroduplexes with 16S rRNA Gene-Based Cloning. *Applied and Environmental Microbiology*, 67, 880–887.

DE QUEIROZ, K. (2007) Species Concepts and Species Delimitation. *Systematic Biology*, 56, 879–856.

QUINCE, C., LANZÉN, A., CURTIS, T.P., DAVENPORT, R.J., HALL, N., IAN M HEAD, ET AL. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*, 6, 639–641.

R

RAVIN, N.V., GALACHYANTS, Y.P., MARDANOV, A.V., BELETSKY, A.V., PETROVA, D.P., SHERBAKOVA, T.A., ET AL. (2010) Complete sequence of the mitochondrial genome of a diatom alga *Synedra acus* and comparative analysis of diatom mitochondrial genomes. *Current Genetics*, 56, 215–223.

REICHARDT, E. (1999) Zur Revision der Gattung *Gomphonema*: Die Arten um *G. affine/insigne*, *G. angustatum/micropus*, *G. acuminatum* sowie gomphonemoide Diatomeen aus dem Oberoligozän in Böhmen. A.R.G. Gantner.

REICHARDT, E. (2007) Neue und wenig bekannte *Gomphonema*-Arten (Bacillariophyceae) mit Areolen in Doppelreihen. *Nova Hedwigia*, 85, 103–137.

REICHARDT, E. (2008) Gomphonema intermedium Hustedt sowie drei neue, ähnliche arten. *Diatom Research*, 23, 105–115.

REYSENBACH, A.L., GIVER, L.J., WICKHAM, G.S. & PACE, N.R. (1992) Differential amplification of rRNA genes by polymerase chain reaction. *Applied and Environmental Microbiology*, 58, 3417–3418.

RIMET, F., KERMARREC, L., BOUCHEZ, A., HOFFMANN, L., ECTOR, L. & MEDLIN, L.K. (2011) Molecular phylogeny of the family Bacillariaceae based on 18S rDNA sequences: focus on freshwater *Nitzschia* of the section Lanceolatae. *Diatom Research*, 26, 273–291.

ROHLF, F.J. (2002) Geometric morphometrics and phylogeny. Dans *Morphology, shape and phylogeny*. p. 175–193Taylor and Francis. MacLeod, N., and P. L. Forey, eds., London.

ROHLF, F.J. (2007) TPS Series. Department of Ecology and Evolution, State University of New York at Stony Brook.

ROHLF, F.J. & MARCUS, L.F. (1993) A revolution morphometrics. *Trends in Ecology & Evolution*, 8, 129–132.

ROMARI, K. & VAULOT, D. (2004) Composition and temporal variability of picoeukaryote communities at a coastal site of the English Channel from 18S rDNA sequences. *Limnology and Oceanography*, 49, 784–789.

RONAGHI, M. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.*, 242, 84–89.

RONAGHI, M., UHLEN, M. & NYREN, P. (1998) A sequencing method based on real-time pyrophosphate. *Science*, 281, 363–365.

ROUBEIX, V., MAZZELLA, N., SCHOULER, L., FAUVELLE, V., MORIN, S., COSTE, M., ET AL. (2011) Variations of periphytic diatom sensitivity to the herbicide diuron and relation to species distribution in a contamination gradient: implications for biomonitoring. *Journal of Environmental Monitoring.*, 13, 1768–1774.

ROUND, F.E. (1991) Use of diatoms for monitoring rivers. Dans *Use of algae for monitoring rivers* p. . Whitton, B. A., Rott, E., Friedrich G.

ROUND, F.E., CRAWFORD, R.M. & MANN, D.G. (1990) The diatoms: biology & morphology of the genera. Cambridge University Press.

RUHL, M.W., WOLF, M. & JENKINS, T.M. (2010) Compensatory base changes illuminate morphologically difficult taxonomy. *Molecular Phylogenetics and Evolution*, 54, 664–669.

S

SABBE, K., CHEPURNOV, V.A., VYVERMAN, W. & MANN, D.G. (2004) Apomixis in *Achnanthes* (Bacillariophyceae); development of a model system for diatom reproductive biology. *European Journal of Phycology*, 39, 327–341.

SABBE, K., VANHOUTTE, K., LOWE, R.L., BERGEY, E.A., BIGGS, B.J.F., FRANCOEUR, S., ET AL. (2001) *Ellerbeckia arenaria* (Bacillariophyceae): formation of auxospores and initial cells. *European Journal of Phycology*, 36, 307–320.

SABBE, K., VERLEYEN, E., HODGSON, D., VANHOUTTE, K. & VYVERMAN, W. (2003) Benthic diatom flora of freshwater and saline lakes in the Larsemann Hills and

Rauer Islands, East Antarctica. *Antarctic Science*, 15, 227–248.

SAIKI, R., GELFAND, D., STOFFEL, S., SCHARF, S., HIGUCHI, R., HORN, G., ET AL. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239, 487–491.

SAIKI, R., SCHARF, S., FALOONA, F., MULLIS, K., HORN, G., ERLICH, H. & ARNHEIM, N. (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230, 1350–1354.

SANGER, F., NICKLEN, S. & COULSON, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74, 5463–5467.

SAR, E.A., SUNESEN, I., LAVIGNE, A.S. & LOFEUDO, S. (2011) *Thalassiosira rotula*, a heterotypic synonym of Thalassiosira gravida: morphological evidence. *Diatom Research*, 26, 109–119.

SARNO, D., KOOISTRA, W.H.C.F., BALZANO, S., HARGRAVES, P.E. & ZINGONE, A. (2007) Diversity in the genus *Skeletonema* (Bacillariophyceae): III. phylognenetic position and morphological variability of *Skeletonema costatum* and *Skeletonema grevillei*, with the description of *Skeletonema ardens* sp. nov. *Journal of Phycology*, 43, 156–170.

SARNO, D., KOOISTRA, W.H.C.F., MEDLIN, L.K., PERCOPO, I. & ZINGONE, A. (2005) Diversity in the genus *Skeletonema* (Bacillariophyceae). II. An assessment of the taxonomy of *S. costatum*-like species with the description of four new species. *Journal of Phycology*, 41, 151–176.

SAUNDERS, G.W. (2005) Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Philosophical Transactions of the Royal Society B*, 360, 1879–1888.

SAVIN, M.C., MARTIN, J.L., LEGRESLEY, M., GIEWAT, M. & ROONEY-VARGA, J. (2004) Plankton Diversity in the Bay of Fundy as Measured by Morphological and Molecular Methods. *Microbial Ecology*, 48, 51–65.

SCHLOSS, P.D. & HANDELSMAN, J. (2005) Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Applied and Environmental Microbiology*, 71, 1501 –1506.

SCHLOSS, P.D. & WESTCOTT, S.L. (2011) Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. *Applied and Environmental Microbiology*, 77, 3219 –3226.

SCHLOSS, P.D., WESTCOTT, S.L., RYABIN, T., HALL, J.R., HARTMANN, M., HOLLISTER, E.B., ET AL. (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, 75, 7537 –7541.

SCHOLIN, C.A., HERZOG, M., SOGIN, M. & ANDERSON, D.M. (1994) Identification of group- and strain-specific genetic markers for globally distributed *Alexandrium* (dinophyceae). II. Sequence analysis of a fragment of the LSU rRNA gene. *Journal of Phycology*, 30, 999–1011.

SEENA, S., PASCOAL, C., MARVANOVÁ, L. & CÁSSIO, F. (2010) DNA barcoding of fungi: a case study using ITS sequences for identifying aquatic hyphomycete species. *Fungal Diversity*, 44, 77–87.

SHENDURE, J. & JI, H. (2008) Next-generation DNA sequencing. *Nature biotechnology*, 26, 1135–1145.

SIMON, C., FRATI, F., BECKENBACH, A., CRESPI, B., LIU, H. & FLOORS, P. (1994) Evolution, Weighting, and Phylogenetic Utility of Mitochondrial Gene Sequences and a Compilation of Conserved Polymerase Chain Reaction Primers. *Annals of*

the Entomological Society of America, 87, 651–701.

SMOL, J.P. & STOERMER, E.F. (2010) The Diatoms: Applications for the Environmental and Earth Sciences. Cambridge University Press.

SOGIN, M.L., MORRISON, H.G., HUBER, J.A., WELCH, D.M., HUSE, S.M., NEAL, P.R., ET AL. (2006) Microbial diversity in the deep sea and the underexplored ‘‘rare biosphere’’. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 12115–12120.

SOININEN, E.M., VALENTINI, A., COISSAC, E., MIQUEL, C., GIELLY, L., BROCHMANN, C., ET AL. (2009) Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. *Frontiers in Zoology*, 6, 16.

SOLTIS, D.E., SOLTIS, P.S., CHASE, M.W., MORT, M.E., ALBACH, D.C., ZANIS, M., ET AL. (2000) Angiosperm phylogeny inferred from 18S rDNA, *rbc*L, and *atp*B sequences. *Botanical Journal of the Linnean Society*, 133, 381–461.

SORHANNUS, U., GASSE, F., PERASSO, R. & TOURANCHEAU, A.B. (1995) A preliminary phylogeny of diatoms based on 28S ribosomal RNA sequence data. *Phycologia*, 34, 65–73.

SORHANNUS, U., ORTIZ, J.D., WOLF, M. & FOX, M.G. (2010) Microevolution and speciation in *Thalassiosira weissflogii* (Bacillariophyta). *Protist*, 161, 237–249.

STACKEBRANDT, E. & GOEBEL, B.M. (1994) Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic Bacteriology*, 44, 846 –849.

STAUBER, J.L. & JEFFREY, S.W. (1988) Photosynthetic pigments in fifty-one species of marine diatoms. *Journal of Phycology*, 24, 158–172.

STEVENSON, R.J. & PAN, Y. (1999) Assessing environmental conditions in rivers and streams with diatoms. Dans *The Diatoms: Applications for the Environmental and Earth Sciences* (eds E.F. Stoermer & J.P. Smol), p. 11–40. Cambridge University Press, Cambridge, UK.

STOCKNER, J.G. (1967) Observations of thermophilic algal communities in Mount Rainier and Yellowstone National Parks. *Limnology and Oceanography*, 12, 13–17.

STOECK, T., BEHNKE, A., CHRISTEN, R., AMARAL-ZETTLER, L., RODRIGUEZ-MORA, M.J., CHISTOSERDOV, A., ET AL. (2009) Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biology*, 7, 72.

STOOF-LEICHSENRING, K.R., EPP, L.S., TRAUTH, M.H. & TIEDEMANN, R. (2012) Hidden diversity in diatoms of Kenyan Lake Naivasha: a genetic approach detects temporal variation. *Molecular Ecology*.

SUN, Y., SKINNER, D.Z., LIANG, G.H. & HULBERT, S.H. (1994) Phylogenetic analysis of Sorghum and related taxa using internal transcribed spacers of nuclear ribosomal DNA. *Theoretical and Applied Genetics*, 89, 5463–5467.

SUZUKI, M. & GIOVANNONI, S. (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology*, 62, 625–630.

SUZUKI, M., RAPPÉ, M.S. & GIOVANNONI, S.J. (1998) Kinetic Bias in Estimates of Coastal Picoplankton Community Structure Obtained by Measurements of Small-Subunit rRNA Gene PCR Amplicon Length Heterogeneity. *Applied and Environmental Microbiology*, 64, 4522–4529.

SWEENEY, B.W., BATTLE, J.M., JACKSON, J.K. & DAPKEY, T. (2011) Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? *Journal of the North American Benthological Society*, 30, 195–216.

SYLVESTRE, F. (2002) A high-resolution diatom reconstruction between 21,000 and 17,4000 ^{14}C yr BP from the southern Bolivian Altiplano (18–23°S). *Journal of Paleolimnology*, 27, 45–57.

T

TAIB, N., BRONNER, G. & DEBROAS, D.PANAM: Phylogenetic Analysis of Next generation AMplicons.

TAKANO, Y., HANSEN, G., FUJITA, D. & HORIGUCHI, T. (2008) Serial Replacement of Diatom Endosymbionts in Two Freshwater Dinoflagellates, Peridiniopsis spp. (Peridiniales, Dinophyceae). *Phycologia*, 47, 41–53.

TAMURA, K., PETERSON, D., PETERSON, N., STECHER, G., NEI, M. & KUMAR, S. (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution*, 28, 2731–2739.

TEDERSOO, L., NILSSON, R.H., ABARENKOV, K., JAIRUS, T., SADAM, A., SAAR, I., ET AL. (2010) 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist*, 188, 291–301.

TELFORD, R.J., VANDVIK, V. & BIRKS, H.J.B. (2006) Dispersal Limitations Matter for Microbial Morphospecies. *Science*, 312, 1015.

THERIOT, E.C., ASHWORTH, M., RUCK, E., NAKOV, T. & JANSEN, R.K. (2010) A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. *Plant Ecology and Evolution*, 143, 278–296.

THERIOT, E.C., CANNONE, J.J., GUTELL, R.R. & ALVERSON, A.J. (2009) The limits of nuclear-encoded SSU rDNA for resolving the diatom phylogeny. *European Journal of Phycology*, 44, 277–290.

THERIOT, E.C., FRITZ, S.C., WHITLOCK, C. & CONLEY, D.J. (2006) Late Quaternary rapid morphological evolution of an endemic diatom in Yellowstone Lake, Wyoming. *Paleobiology*, 32, 38 –54.

THERIOT, E.C., RUCK, E., ASHWORTH, M., NAKOV, T. & JANSEN, R.K. (2011) Status of the Pursuit of the Diatom Phylogeny: Are Traditional Views and New Molecular Paradigms Really That Different? Dans *The Diatom World, Cellular Origin, Life in Extreme Habitats and Astrobiology* p. 119–142. J. Seckbach and J.P. Kociolek.

THOMPSON, J.D., HIGGINS, D.G. & GIBSON, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22, 4673 –4680.

TROBAJO, R., CLAVERO, E., CHEPURNOV, V.A., SABBE, K., MANN, D.G., ISHIHARA, S. & COX, E.J. (2009) Morphological, genetic and mating diversity within the widespread bioindicator *Nitzschia palea* (Bacillariophyceae). *Phycologia*, 48, 443–459.

TROBAJO, R., MANN, D.G., CHEPURNOV, V.A., CLAVERO, E. & COX, E.J. (2006) Taxonomy, life cycle and auxosporulation of *Nitzschia fonticola* (Bacillariophyta). *Journal of Phycology*, 42, 1353–1372.

TROBAJO, R., MANN, D.G., CLAVERO, E., EVANS, K.M., VANORMELINGEN, P. & MCGREGOR, R.C. (2010) The use of partial cox1, rbcL and LSU rDNA sequences for phylogenetics and species identification within the *Nitzschia palea* species complex (Bacillariophyceae). *European Journal of Phycology*, 45, 413–425.

TUJI, A. (2005) Taxonomy of the *Gomphoneis tetrastigmata* species complex. *Bulletin of the National Science Museum, Tokyo, Series B (Botany)*, 31, 89–108.

U

UNTERSEHER, M., JUMPPONEN, A., ÖPIK, M., TEDERSOO, L., MOORA, M., DORMANN, C.F. & SCHNITTLER, M. (2011) Species abundance distributions and richness estimations in fungal metagenomics – lessons learned from community ecology. *Molecular Ecology*, 20, 275–285.

UROZ, S., BUÉE, M., MURAT, C., FREY-KLETT, P. & MARTIN, F. (2010) Pyrosequencing reveals a contrasted bacterial diversity between oak rhizosphere and surrounding soil. *Environmental Microbiology Reports*, 2, 281–288.

V

VALENTINI, A., MIQUEL, C., NAWAZ, M.A., BELLEMAIN, E., COISSAC, E., POMPANON, F., ET AL. (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trn L approach. *Molecular Ecology Resources*, 9, 51–60.

VALENTINI, A., POMPANON, F. & TABERLET, P. (2009) DNA barcoding for ecologists. *TRENDS in Ecology and Evolution*, 24, 110–117.

VANELSLANDER, B., CRÉACH, V., VANORMELINGEN, P., ERNST, A., CHEPURNOV, V.A., SAHAN, E., ET AL. (2009) Ecological differentiation between sympatric pseudocryptic species in the estuarine benthic diatom Navicula phyllepta (Bacillariophyceae). *Journal of Phycology*, 45, 1278–1289.

VANORMELINGEN, P., CHEPURNOV, V.A., MANN, D.G., COUSIN, S. & VYVERMAN, W. (2007) Congruence of morphological, reproductive and ITS rDNA sequence data in some Australasian *Eunotia bilunaris* (Bacillariophyta). *European Journal of Phycology*, 42, 61–79.

VANORMELINGEN, P., VERLEYEN, E. & VYVERMAN, W. (2008) The diversity and distribution of diatoms: from cosmopolitanism to narrow endemism. *Biodiversity and Conservation*, 17, 393–405.

VERBRUGGEN, H. & THERIOT, E.C. (2008) Building trees of algae: some advances in phylogenetic and evolutionary analysis. *European Journal of Phycology*, 43, 229–252.

VESELÁ, J., NEUSTUPA, J., PICHRTOVÁ, M. & POULÍČKOVÁ, A. (2009) Morphometric study of *Navicula* morphospecies (Bacillariophyta) with respect to diatom life cycle. *Fottea*, 9, 307–316.

VAN DE VIJVER, B. & BEYENS, L. (1999) Biogeography and ecology of freshwater diatoms in Subantarctica: a review. *Journal of Biogeography*, 26, 993–1000.

VAN DE VIJVER, B., GREMMEN, N.J.M. & BEYENS, L. (2005) The genus *Stauroneis* (Bacillariophyceae) in the Antarctic region. *Journal of Biogeography*, 32, 1791–1798.

VIS, C., HUDON, C., CATTANEO, A. & PINEL-ALLOUL, B. (1998) Periphyton as an indicator of water quality in the St Lawrence River (Québec, Canada). *Environmental Pollution*, 101, 13–24.

VYVERMAN, W., VERLEYEN, E., SABBE, K., VANHOUTTE, K., STERKEN, M., HODGSON, D.A., ET AL. (2007) Historical processes constrain patterns in global diatom diversity. *Ecology*, 88, 1924–1931.

W

WALLACE, J.H. & PATRICK, R. (1950) A consideration of *Gomphonema parvulum* Kutz. *Butler University Botanical Studies*, 9, 23.

WANG, P., SHEN, H. & XIE, P. (2012) Can Hydrodynamics Change Phosphorus Strategies of Diatoms?—Nutrient Levels and Diatom Blooms in Lotic and Lentic Ecosystems. *Microbial Ecology*, 63, 369–382.

WARD, B.B. (2002) How many species of prokaryotes are there? *Proceedings of the National Academy of Sciences*, 99, 10234 –10236.

WARD, R.D., ZEMLAK, T.S., INNES, B.H., LAST, P.R. & HEBERT, P.D.. (2005) DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360, 1847 –1857.

WASSON, J.-G., CHANDESRIS, A. & PELLA, H. (2004) Hydro-écorégions de l'île de la Réunion. Propositions de régionalisation des écosystèmes aquatiques en vue de l'application de la Directive Cadre Européenne sur l'Eau. Dans p. 18. Cemagref, Lyon.

WASSON, J.-G., CHANDESRIS, A., PELLA, H. & BLANC, L. (2002) Les hydro-écorégions de France métropolitaine. Approche régionale de la typologie des eaux courantes et éléments pour la définition des peuplements de référence d'invertébrés. Dans Cemagref, Lyon.

WAYNE, L.G., BRENNER, D.J., COLWELL, R.R., GRIMONT, P.A.D., KANDLER, O., KRICHEVSKY, M.I., ET AL. (1987) Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *International journal of systematic bacteriology*, 37, 463–464.

WHITE, T., BRUNS, T., LEE, S. & TAYLOR, J. (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. Dans *PCR Protocols: A Guide to Methods and Applications* p. 315–322. Academic Press.

WILLIAMS, D.M. & KOCIOLEK, J.P. (2007) Pursuit of a natural classification of diatoms: History, monophyly and the rejection of paraphyletic taxa. *European Journal of Phycology*, 42, 313–319.

VON WINTZINGERODE, F., GÖBEL, U.B. & STACKEBRANDT, E. (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiology Reviews*, 21, 213–229.

WITT, J.D.S., THRELOFF, D.L. & HEBERT, P.D.N. (2006) DNA barcoding reveals extraordinary cryptic diversity in an amphipod genus: implications for desert spring conservation. *Molecular Ecology*, 15, 3073–3082.

WOESE, C.R. (1987) Bacterial evolution. *Microbiology and Molecular Biology Reviews*, 51, 221–271.

WU, D., HARTMAN, A., WARD, N. & EISEN, J.A. (2008) An Automated Phylogenetic Tree-Based Small Subunit rRNA Taxonomy and Alignment Pipeline (STAP). *PLoS ONE*, 3, e2566.

WU, J.-Y., JIANG, X.-T., JIANG, Y.-X., LU, S.-Y., ZOU,, F. & ZHOU, H.-W. (2010) Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method. *BMC Microbiology*, 10, 255.

Z

ZECHMAN, F.W., ZIMMER, E.A. & THERIOT, E.C. (1994) Use of ribosomal DNA Internal Transcribed Spacers for phylogenetic studies in diatoms. *Journal of Phycology*,

30, 507–512.

ZELINKA, M. & MARVAN, P. (1961) Zur Prazisierung der biologischen Klassifikation der Reinheit flisender Gewasser. *Archiv für Hydrobiologie*, 57, 389–407.

ZHU, F., MASSANA, R., NOT, F., MARIE, D. & VAULOT, D. (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiology Ecology*, 52, 79–92.

ZIMMERMANN, J., JAHN, R. & GEMEINHOLZER, B. (2011) Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. *Organisms Diversity and Evolution*, 11, 173–192.

ZINGER, L., GOBET, A. & POMMIER, T. (2011) Two decades of describing the unseen majority of aquatic microbial diversity. *Molecular Ecology*, no.

ZUCKERKANDL, E. & PAULING, L. (1965) Evolutionary divergence and convergence in proteins. Dans *Evolving Genes and Proteins* p. 97–166. Academic Press, New York.

ZURZOLO, C. & BOWLER, C. (2001) Exploring Bioinorganic Pattern Formation in Diatoms. A Story of Polarized Trafficking. *Plant Physiology*, 127, 1339 –1345.

ANNEXE 1 : LISTE DES CULTURES

Code TCC	Taxon	Origine du prélèvement : Pays (Rivière ou Lac, site)	Date de prélèvement
TCC7	*Fragilaria capucina* var. *vaucheriae*	France métropolitaine (La Loue)	16.09.2003
TCC107	*Craticula accomoda*	France métropolitaine (Lac Léman)	-
TCC131	*Fragilaria* sp.	Angleterre (Esthwaite Water)	-
TCC134	*Ulnaria ulna var. acus*	France métropolitaine (Lac d'Annecy)	-
TCC139-1	*Nitzschia palea*	France métropolitaine (Lac Léman)	-
TCC139-2	*Nitzschia palea*	France métropolitaine (Lac Léman)	-
TCC139-3	*Nitzschia palea*	France métropolitaine (Lac Léman)	-
TCC301	*Fragilaria crotonensis*	France métropolitaine (Lac du Bourget)	01.04.2008
TCC302	*Fragilaria crotonensis*	France métropolitaine (Lac du Bourget)	01.04.2008
TCC303	*Fragilaria crotonensis*	France métropolitaine (Lac du Bourget)	01.04.2008
TCC304	*Fragilaria crotonensis*	France métropolitaine (Lac du Bourget)	01.04.2008
TCC306	*Fragilaria ulna* var. *angustissima*	France métropolitaine (Lac du Bourget)	08.02.2008
TCC353	*Cyclotella costei*	France métropolitaine (Lac Léman, SHL2)	06.2009
TCC354	*Stephanodiscus parvus*	France métropolitaine (Moselle, Hauconcourt)	06.2009
TCC355	*Stephanodiscus parvus*	France métropolitaine (Lac Léman, SHL2)	06.2009
TCC356	*Cyclotella meneghiniana*	France métropolitaine (Moselle, Hauconcourt)	17.06.2009
TCC357	*Cyclotella meneghiniana*	France métropolitaine (Moselle, Hauconcourt)	17.06.2009
TCC358	*Cyclotella meneghiniana*	France métropolitaine (Moselle, Hauconcourt)	17.06.2009
TCC359	*Cyclotella meneghiniana*	France métropolitaine (Meurthe, Art-sur-Meurthe)	17.06.2009
TCC360	*Cyclotella meneghiniana*	France métropolitaine (Meurthe, Art-sur-Meurthe)	17.06.2009
TCC361	*Encyonema minutum*	France métropolitaine (Lac Léman, Port de l'INRA)	01.03.2009
TCC362	*Asterionella formosa*	France métropolitaine (Lac d'Annecy)	30.06.2009
TCC363	*Cyclotella meneghiniana*	France métropolitaine (Seille, Chambrey)	17.06.2009
TCC364	*Thalassiosira weissflogii*	France métropolitaine (Seille, Chambrey)	17.06.2009
TCC365	*Fragilaria crotonensis*	France métropolitaine (Lac Léman, SHL2)	20.07.2009
TCC366	*Mayamaea fossalis* var. *fossalis*	France métropolitaine (canal de la Haute-Deûle, Noyelles-Godault)	01.12.2004
TCC367	*Fragilaria capucina* var. *vaucheriae*	France métropolitaine (Lac Léman, Port de l'INRA)	01.03.2009
TCC425	*Nitzschia palea*	Mayotte (Djalimou, aval)	18.04.2009
TCC426	*Gomphonema parvulum*	Mayotte (Djalimou, aval)	18.04.2009
TCC427	*Nitzschia palea*	Mayotte (Djalimou, aval)	18.04.2009
TCC428	*Gomphonema parvulum*	Mayotte (Djalimou, aval)	18.04.2009
TCC429	*Gomphonema parvulum*	Mayotte (Djalimou, aval)	18.04.2009
TCC430	*Gomphonema parvulum*	Mayotte (Djalimou, aval)	18.04.2009
TCC431	*Gomphonema parvulum*	Mayotte (Djalimou, aval)	18.04.2009
TCC432	*Gomphonema parvulum*	Mayotte (Djalimou, aval)	18.04.2009

TCC433	*Gomphonema parvulum*	Mayotte (Djalimou, aval)	18.04.2009
TCC434	*Gomphonema parvulum*	Mayotte (Djalimou, aval)	18.04.2009
TCC435	*Nitzschia palea*	Mayotte (Djalimou, aval)	18.04.2009
TCC436	*Gomphonema parvulum*	Mayotte (Djalimou, aval)	18.04.2009
TCC437	*Nitzschia palea*	Mayotte (Djalimou, aval)	18.04.2009
TCC438	*Gomphonema parvulum*	Mayotte (Djalimou, aval)	18.04.2009
TCC439	*Gomphonema exilissimum*	Mayotte (Djalimou, aval)	18.04.2009
TCC440	*Gomphonema parvulum*	Mayotte (Djalimou, aval)	18.04.2009
TCC441	*Gomphonema bourbonense*	Mayotte (Bouyouni, intermédiaire)	19.04.2009
TCC442	*Nitzschia amphibia*	Mayotte (Bouyouni, intermédiaire)	19.04.2009
TCC443	*Nitzschia amphibia*	Mayotte (Bouyouni, intermédiaire)	19.04.2009
TCC444	*Nitzschia amphibia*	Mayotte (Bouyouni, intermédiaire)	19.04.2009
TCC445	*Nitzschia amphibia*	Mayotte (Bouyouni, intermédiaire)	19.04.2009
TCC446	*Navicula cryptocephala*	Mayotte (Bouyouni, intermédiaire)	19.04.2009
TCC447	*Gomphonema parvulum*	Mayotte (Bouyouni, intermédiaire)	19.04.2009
TCC448	*Navicula lundii*	Mayotte (Bouyouni, intermédiaire)	19.04.2009
TCC449	*Cocconeis euglypta*	Mayotte (Bouyouni, intermédiaire)	19.04.2009
TCC450	*Gomphonema bourbonense*	Mayotte (Bouyouni, intermédiaire)	19.04.2009
TCC451	*Gomphonema bourbonense*	Mayotte (Bouyouni, intermédiaire)	19.04.2009
TCC452	*Gomphonema bourbonense*	Mayotte (Bouyouni, intermédiaire)	19.04.2009
TCC453	*Gomphonema bourbonense*	Mayotte (Bouyouni, intermédiaire)	19.04.2009
TCC454	*Nitzschia palea*	Mayotte (Dembeni, aval)	18.04.2009
TCC455	*Nitzschia palea*	Mayotte (Dembeni, aval)	18.04.2009
TCC456	*Nitzschia palea*	Mayotte (Dembeni, aval)	18.04.2009
TCC457	*Nitzschia palea*	Mayotte (Dembeni, aval)	18.04.2009
TCC458	*Gomphonema bourbonense*	Mayotte (Dembeni, aval)	18.04.2009
TCC459	*Craticula molestiformis*	Mayotte (Dembeni, aval)	18.04.2009
TCC460	*Gomphonema angustum*	Mayotte (Dembeni, aval)	18.04.2009
TCC461	*Sellaphora seminulum*	Mayotte (Dembeni, aval)	18.04.2009
TCC462	*Sellaphora seminulum*	Mayotte (Dembeni, aval)	18.04.2009
TCC463	*Gomphonema parvulum*	Mayotte (Dembeni, aval)	18.04.2009
TCC464	*Gomphonema parvulum*	Mayotte (Dembeni, aval)	18.04.2009
TCC465	*Gomphonema parvulum*	Mayotte (Dembeni, aval)	18.04.2009
TCC466	*Gomphonema parvulum*	Mayotte (Dembeni, aval)	18.04.2009
TCC467	*Gomphonema parvulum*	Mayotte (Dembeni, aval)	18.04.2009
TCC468	*Nitzschia palea*	Mayotte (Dembeni, aval)	18.04.2009
TCC469	*Gomphonema clevei*	Mayotte (Dembeni, aval)	18.04.2009
TCC470	*Gomphonema parvulum*	Mayotte (Dembeni, aval)	18.04.2009
TCC471	*Gomphonema parvulum*	Mayotte (Coconi, intermédiaire)	18.04.2009
TCC472	*Pinnularia acrosphaeria*	Mayotte (Coconi, intermédiaire)	18.04.2009
TCC473	*Gomphonema parvulum*	Mayotte (Coconi, intermédiaire)	18.04.2009
TCC474	*Nitzschia inconspicua*	Mayotte (Coconi, intermédiaire)	18.04.2009
TCC475	*Nitzschia palea*	Mayotte (Coconi, intermédiaire)	18.04.2009
TCC476	*Nitzschia palea*	Mayotte (Coconi, intermédiaire)	18.04.2009
TCC477	*Amphora montana*	Mayotte (Coconi, intermédiaire)	18.04.2009
TCC478	*Gomphonema parvulum*	Mayotte (Coconi, intermédiaire)	18.04.2009

TCC479	*Amphora montana*	Mayotte (Coconi, intermédiaire)	18.04.2009
TCC480	*Nitzschia palea*	Mayotte (Coconi, aval)	18.04.2009
TCC481	*Nitzschia inconspicua*	Mayotte (Coconi, aval)	18.04.2009
TCC482	*Gomphonema parvulum*	Mayotte (Coconi, aval)	18.04.2009
TCC483	*Mayamaea permitis*	Mayotte (Coconi, aval)	18.04.2009
TCC484	*Mayamaea permitis*	Mayotte (Coconi, aval)	18.04.2009
TCC485	*Gomphonema parvulum*	Mayotte (Coconi, aval)	18.04.2009
TCC486	*Nitzschia palea*	Mayotte (Coconi, aval)	18.04.2009
TCC487	*Nitzschia inconspicua*	Mayotte (Coconi, aval)	18.04.2009
TCC488	*Nitzschia inconspicua*	Mayotte (Coconi, aval)	18.04.2009
TCC489	*Nitzschia inconspicua*	Mayotte (Coconi, aval)	18.04.2009
TCC490	*Navicula* sp.	Mayotte (Coconi, aval)	18.04.2009
TCC491	*Nitzschia palea*	Mayotte (Coconi, aval)	18.04.2009
TCC492	*Gomphonema parvulum*	Mayotte (Coconi, aval)	18.04.2009
TCC493	*Nitzschia palea*	Mayotte (Coconi, aval)	18.04.2009
TCC494	*Gomphonema parvulum*	Mayotte (Coconi, aval)	18.04.2009
TCC495	*Navicula* sp.	Mayotte (Coconi, aval)	18.04.2009
TCC496	*Lemnicola hungarica*	Mayotte (Kwalé, amont)	18.04.2009
TCC497	*Pinnularia microstauron*	Mayotte (Kwalé, amont)	18.04.2009
TCC498	*Nitzschia amphibia*	Mayotte (Kwalé, amont)	18.04.2009
TCC499	*Placoneis elginensis*	Mayotte (Kwalé, amont)	18.04.2009
TCC500	*Gomphonema* cf. l*agenula*	Mayotte (Kwalé, amont)	18.04.2009
TCC501	*Cocconeis placentula*	Mayotte (Kwalé, amont)	18.04.2009
TCC502	*Navicula viridula var. rostellata*	Mayotte (Longoni, aval)	20.04.2009
TCC503	*Fistulifera saprophila*	Mayotte (Longoni, aval)	20.04.2009
TCC504	*Fistulifera saprophila*	Mayotte (Longoni, aval)	20.04.2009
TCC505	*Fistulifera saprophila*	Mayotte (Longoni, aval)	20.04.2009
TCC506	*Fistulifera saprophila*	Mayotte (Longoni, aval)	20.04.2009
TCC507	*Gomphonema clevei*	Mayotte (Longoni, aval)	20.04.2009
TCC508	*Fistulifera saprophila*	Mayotte (Longoni, aval)	20.04.2009
TCC509	*Fistulifera saprophila*	Mayotte (Longoni, aval)	20.04.2009
TCC510	*Nitzschia hantzschiana*	Mayotte (Longoni, aval)	20.04.2009
TCC511	*Cocconeis placentula*	Mayotte (Longoni, aval)	20.04.2009
TCC512	*Nitzschia linearis*	Mayotte (Bouyouni, amont)	19.04.2009
TCC513	*Gomphonema bourbonense*	Mayotte (Bouyouni, amont)	19.04.2009
TCC514	*Gomphonema bourbonense*	Mayotte (Bouyouni, amont)	19.04.2009
TCC515	*Navicula cryptocephala*	Mayotte (Bouyouni, amont)	19.04.2009
TCC516	*Nitzschia lorenziana*	Mayotte (Bouyouni, amont)	19.04.2009
TCC517	*Sellaphora minima*	Mayotte (Bouyouni, amont)	19.04.2009
TCC518	*Cocconeis placentula*	Mayotte (Bouyouni, amont)	19.04.2009
TCC519	*Melosira varians*	La Réunion (riv des Galets, Marla)	30.04.2009
TCC520	*Ulnaria ulna*	La Réunion (riv des Galets, Marla)	30.04.2009
TCC521	*Nitzschia* cf. *frustulum*	La Réunion (riv des Galets, Marla)	30.04.2009
TCC522	*Ulnaria ulna*	La Réunion (riv des Galets, Marla)	30.04.2009
TCC523	*Nitzschia palea*	La Réunion (Saint Denis, amont prise AEP)	28.04.2009
TCC524	*Sellaphora minima*	La Réunion (Saint Denis, amont prise AEP)	28.04.2009

TCC525	*Fistulifera saprophila*	La Réunion (Saint Denis, amont prise AEP)	28.04.2009
TCC526	*Gomphonema affine* var. *affine*	La Réunion (Saint Denis, amont prise AEP)	28.04.2009
TCC527	*Gomphonema clavatum*	La Réunion (Saint Denis, amont prise AEP)	28.04.2009
TCC528	*Nitzschia palea*	La Réunion (Saint Denis, amont prise AEP)	28.04.2009
TCC529	*Sellaphora minima*	La Réunion (Saint Denis, amont prise AEP)	28.04.2009
TCC530	*Navicula capitoradiata*	La Réunion (Bras des étangs)	24.04.2009
TCC531	*Nitzschia palea*	La Réunion (Bras des étangs)	24.04.2009
TCC532	*Melosira varians*	La Réunion (Bras des étangs)	24.04.2009
TCC533	*Nitzschia fonticola*	La Réunion (Bras des étangs)	24.04.2009
TCC534	*Fistulifera saprophila*	La Réunion (Bras des étangs)	24.04.2009
TCC535	*Fistulifera saprophila*	La Réunion (Bras des étangs)	24.04.2009
TCC536	*Gomphonema pumilum* var. *pumilum*	La Réunion (Bras des étangs)	24.04.2009
TCC537	*Nitzschia acidoclinata*	La Réunion (Bras Caverne, Amont confluence riv. du Mât)	29.04.2009
TCC538	*Nitzschia acidoclinata*	La Réunion (Bras Caverne, Amont confluence riv. du Mât)	29.04.2009
TCC539	*Fragilaria capucina* var. *vaucheriae*	La Réunion (Bras Caverne, Amont confluence riv. du Mât)	29.04.2009
TCC540	*Mayamaea permitis*	La Réunion (Bras Caverne, Amont confluence riv. du Mât)	29.04.2009
TCC541	*Fragilaria capucina* var. *vaucheriae*	La Réunion (Bras Caverne, Amont confluence riv. du Mât)	29.04.2009
TCC542	*Fragilaria capucina* var. *vaucheriae*	La Réunion (Bras Caverne, Amont confluence riv. du Mât)	29.04.2009
TCC543	*Nitzschia acidoclinata*	La Réunion (Bras Caverne, Amont confluence riv. du Mât)	29.04.2009
TCC544	*Sellaphora minima*	La Réunion (Bras Caverne, Amont confluence riv. du Mât)	29.04.2009
TCC545	*Nitzschia frustulum*	La Réunion (Langevin, amont cascade grand Galet)	25.04.2009
TCC546	*Fistulifera saprophila*	La Réunion (Langevin, amont cascade grand Galet)	25.04.2009
TCC547	*Fragilaria capucina* var. *vaucheriae*	La Réunion (Langevin, amont cascade grand Galet)	25.04.2009
TCC548	*Fistulifera saprophila*	La Réunion (Langevin, amont cascade grand Galet)	25.04.2009
TCC549	*Nitzschia* cf. *frustulum*	La Réunion (Langevin, amont cascade grand Galet)	25.04.2009
TCC550	*Nitzschia frustulum*	La Réunion (Langevin, amont cascade grand Galet)	25.04.2009
TCC551	*Nitzschia frustulum*	La Réunion (Langevin, amont cascade grand Galet)	25.04.2009
TCC552	*Fistulifera saprophila*	La Réunion (Langevin, amont cascade grand Galet)	25.04.2009
TCC553	*Fragilaria capucina* var. *vaucheriae*	La Réunion (Langevin, amont cascade grand Galet)	25.04.2009
TCC554	*Fistulifera saprophila*	La Réunion (Langevin, amont cascade grand Galet)	25.04.2009
TCC555	*Fragilaria capucina* var. *vaucheriae*	La Réunion (Langevin, amont prise EDF)	25.04.2009
TCC556	*Fistulifera saprophila*	La Réunion (Langevin, amont prise EDF)	25.04.2009
TCC557	*Fistulifera saprophila*	La Réunion (Langevin, amont prise EDF)	25.04.2009

TCC558	*Fragilaria capucina* var. *vaucheriae*	La Réunion (Langevin, amont prise EDF)	25.04.2009
TCC559	*Fragilaria capucina*	La Réunion (Sainte Suzanne, Amont Bassin Bœuf)	19.05.2009
TCC560	*Nitzschia* cf. *acidoclinata*	La Réunion (Sainte Suzanne, Amont Bassin Bœuf)	19.05.2009
TCC561	*Sellaphora seminulum*	La Réunion (Sainte Suzanne, Amont Bassin Bœuf)	19.05.2009
TCC562	*Fragilaria capucina*	La Réunion (Sainte Suzanne, Amont Bassin Bœuf)	19.05.2009
TCC563	*Nitzschia palea*	La Réunion (Sainte Suzanne, Amont Bassin Bœuf)	19.05.2009
TCC564	*Achnanthidium minutissimum*	La Réunion (Sainte Suzanne, Amont Bassin Bœuf)	19.05.2009
TCC565	*Nitzschia amphibia*	La Réunion (Sainte Suzanne, Amont Bassin Bœuf)	19.05.2009
TCC566	*Nitzschia acidoclinata*	La Réunion (Sainte Suzanne, Amont Bassin Bœuf)	19.05.2009
TCC568	*Nitzschia palea*	La Réunion (Sainte Suzanne, Cascade Niagara)	19.05.2009
TCC569	*Sellaphora seminulum*	La Réunion (Sainte Suzanne, Cascade Niagara)	19.05.2009
TCC570	*Nitzschia palea*	La Réunion (Sainte Suzanne, Cascade Niagara)	19.05.2009
TCC571	*Nitzschia amphibia*	La Réunion (Sainte Suzanne, Cascade Niagara)	19.05.2009
TCC572	*Fistulifera pelliculosa*	La Réunion (Sainte Suzanne, Cascade Niagara)	19.05.2009
TCC573	*Nitzschia palea*	La Réunion (Sainte Suzanne, Cascade Niagara)	19.05.2009
TCC574	*Nitzschia amphibia*	La Réunion (Sainte Suzanne, Cascade Niagara)	19.05.2009
TCC575	*Nitzschia tubicola*	La Réunion (Sainte Suzanne, Cascade Niagara)	19.05.2009
TCC576	*Nitzschia* sp.	France métropolitaine (Le Sânon, Solvay)	18.06.2009
TCC577	*Nitzschia palea*	France métropolitaine (Le Sânon, Solvay)	18.06.2009
TCC578	*Navicula capitoradiata*	France métropolitaine (Le Sânon, Solvay)	18.06.2009
TCC579	*Nitzschia capitellata*	France métropolitaine (Meurthe, Art-sur-Meurthe)	17.06.2009
TCC580	*Navicula tripunctata*	France métropolitaine (Meurthe, Art-sur-Meurthe)	17.06.2009
TCC581	*Mayamaea permitis*	France métropolitaine (Meurthe, Art-sur-Meurthe)	17.06.2009
TCC582	*Cyclotella meneghiniana*	France métropolitaine (Seille, Chambrey)	17.06.2009
TCC583	*Nitzschia palea*	France métropolitaine (Chiers, Longlaville)	17.06.2009
TCC584	*Ulnaria ulna*	France métropolitaine (Moselle, Bainville aux Miroirs)	18.06.2009
TCC585	*Nitzschia palea*	France métropolitaine (Moselle, Bainville aux Miroirs)	18.06.2009
TCC586	*Nitzschia cf. pusilla*	France métropolitaine (Canal de Nantes à Brest, Nort-sur-Erdre)	18.08.2009
TCC587	*Navicula cryptotenelloides*	France métropolitaine (Canal de Nantes à Brest, Nort-sur-Erdre)	18.08.2009
TCC588	*Nitzschia palea*	France métropolitaine (Canal de Nantes à Brest, Nort-sur-Erdre)	18.08.2009

TCC589	*Fragilaria capucina*	France métropolitaine (Canal de Nantes à Brest, Nort-sur-Erdre)	18.08.2009
TCC590	*Navicula capitoradiata*	France métropolitaine (Canal de Nantes à Brest, Nort-sur-Erdre)	18.08.2009
TCC591	*Navicula veneta*	France métropolitaine (Don, Guéméné-Penfao)	18.08.2009
TCC592	*Gomphonema parvulum*	France métropolitaine (Don, Guéméné-Penfao)	18.08.2009
TCC593	*Gomphonema parvulum*	France métropolitaine (Don, Guéméné-Penfao)	18.08.2009
TCC594	*Nitzschia palea*	France métropolitaine (Don, Guéméné-Penfao)	18.08.2009
TCC595	*Gomphonema parvulum*	France métropolitaine (Don, Guéméné-Penfao)	18.08.2009
TCC596	*Cyclotella atomus*	France métropolitaine (La Vie, Fenouillet)	17.08.2009
TCC597	*Sellaphora minima*	France métropolitaine (La Vie, Fenouillet)	17.08.2009
TCC598	*Nitzschia palea*	France métropolitaine (La Vie, Fenouillet)	17.08.2009
TCC599	*Nitzschia palea*	France métropolitaine (Isac, Genrouet)	18.08.2009
TCC600	*Nitzschia palea*	France métropolitaine (Isac, Genrouet)	18.08.2009
TCC601	*Nitzschia palea*	France métropolitaine (Isac, Genrouet)	18.08.2009
TCC602	*Nitzschia palea*	France métropolitaine (Isac, Genrouet)	18.08.2009
TCC603	*Nitzschia palea*	France métropolitaine (Isac, Genrouet)	18.08.2009
TCC604	*Navicula symmetrica*	France métropolitaine (Le Lay, La Clay)	17.08.2009
TCC605	*Navicula symmetrica*	France métropolitaine (Le Lay, La Clay)	17.08.2009
TCC606	*Nitzschia* sp.	France métropolitaine (Le Lay, La Clay)	17.08.2009
TCC607	*Navicula cryptotenella*	France métropolitaine (Le Lay, La Clay)	17.08.2009
TCC608	*Pinnularia subgibba*	France métropolitaine (Le Lourdon, Lentigny)	19.08.2009
TCC609	*Nitzschia palea*	France métropolitaine (Le Lourdon, Lentigny)	19.08.2009
TCC610	*Gomphonema parvulum*	France métropolitaine (Le Lourdon, Lentigny)	19.08.2009
TCC611	*Sellaphora* sp.	France métropolitaine (Le Lourdon, Lentigny)	19.08.2009
TCC612	*Gomphonema parvulum*	France métropolitaine (Le Lourdon, Lentigny)	19.08.2009
TCC613	*Nitzschia palea*	France métropolitaine (Le Lourdon, Lentigny)	19.08.2009
TCC614	*Nitzschia palea*	France métropolitaine (Le Lourdon, Lentigny)	19.08.2009
TCC615	*Planothidium lanceolata*	France métropolitaine (Le Lourdon, Lentigny)	19.08.2009
TCC616	*Pinnularia subgibba*	France métropolitaine (Le Lourdon, Lentigny)	19.08.2009
TCC617	*Planothidium lanceolata*	France métropolitaine (Le Lourdon, Lentigny)	19.08.2009
TCC618	*Sellaphora minima*	France métropolitaine (Le Gier, La Valla)	19.08.2009
TCC619	*Nitzschia acidoclinata*	France métropolitaine (Le Gier, La Valla)	19.08.2009
TCC620	*Nitzschia palea*	France métropolitaine (Le Gier, La Valla)	19.08.2009
TCC621	*Sellaphora minima*	France métropolitaine (Le Gier, La Valla)	19.08.2009
TCC622	*Navicula cryptocephala*	France métropolitaine (Le Gier, La Valla)	19.08.2009
TCC623	*Nitzschia palea*	France métropolitaine (Le Gier, La Valla)	19.08.2009

TCC624	*Sellaphora minima*	France métropolitaine (Le Gier, La Valla)	19.08.2009
TCC625	*Melosira varians*	France métropolitaine (Le Gier, Givors)	19.08.2009
TCC626	*Ulnaria ulna*	France métropolitaine (Le Gier, Givors)	19.08.2009
TCC627	*Nitzschia palea*	France métropolitaine (Le Gier, Givors)	19.08.2009
TCC628	*Fistulifera saprophila*	France métropolitaine (Le Gier, Givors)	19.08.2009
TCC629	*Staurosira construens*	France métropolitaine (Le Gier, Givors)	19.08.2009
TCC630	*Navicula cincta*	France métropolitaine (La Saone, Lyon)	13.10.209
TCC631	*Melosira varians*	France métropolitaine (La Saone, Saint Bernard)	13.10.209
TCC632	*Nitzschia* sp.	France métropolitaine (La Saone, Fleurville)	13.10.209
TCC633	*Ulnaria ulna*	Luxembourg (Eischbaach, Boevange-sur-Attert)	06.09.2006
TCC634	*Ulnaria ulna*	Luxembourg (Eischbaach, Boevange-sur-Attert)	06.09.2006
TCC635	*Ulnaria acus*	Luxembourg (Eischbaach, Boevange-sur-Attert)	06.09.2006
TCC636	*Nitzschia palea*	Luxembourg (Eischbaach, Boevange-sur-Attert)	06.09.2006
TCC637	*Eolimna subminuscula*	Luxembourg (Eischbaach, Boevange-sur-Attert)	06.09.2006
TCC638	*Nitzschia palea*	Luxembourg Eischbaach, Boevange-sur-Attert)	06.09.2006
TCC639	*Nitzschia palea*	Luxembourg (Eischbaach, Boevange-sur-Attert)	06.09.2006
TCC640	*Cyclotella meneghiniana*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC641	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC642	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC643	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC644	*Cyclotella meneghiniana*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC645	*Mayamaea permitis*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC646	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC647	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC648	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC649	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC650	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC651	*Nitzschia palea*	Luxembourg (Viichtbacch, Boevange-sur-Attert)	06.09.2006
TCC652	*Nitzschia palea*	Luxembourg (Aalbach, aval Dreiborn)	12.10.2006
TCC653	*Gomphonema parvulum*	Luxembourg (Alzette, Walfer-Steinsel)	13.10.2006
TCC654	*Ulnaria ulna*	Luxembourg (Alzette, Walfer-Steinsel)	13.10.2006
TCC655	*Encyonema lange-bertalotii*	Luxembourg (Attert, Colmar-Berg)	11.10.2006
TCC656	*Ulnaria ulna*	Luxembourg (Attert, Colmar-Berg)	11.10.2006

TCC657	*Fistulifera saprophila*	Luxembourg (Clerve, aval step Clervaux)	10.10.2006
TCC658	*Cyclotella meneghiniana*	Luxembourg (Wiltz / Clerve, Kautenbach)	10.10.2006
TCC659	*Navicula cincta*	Luxembourg (Lenningerbach, amont Ehnen)	12.10.2006
TCC660	*Cyclotella meneghiniana*	Luxembourg (Mamer, aval Thillsmillen)	13.10.2006
TCC661	*Eolimna subminuscula*	Luxembourg (Mamer, aval Thillsmillen)	13.10.2006
TCC662	*Fragilaria capucina*	Luxembourg (Our, Vianden)	11.10.2006
TCC663	*Aulacoseira subartica*	Luxembourg (Our, Vianden)	11.10.2006
TCC664	*Gomphonema parvulum*	Luxembourg (Schlirbech, aval Esch-sur-Sûre)	10.10.2006
TCC665	*Nitzschia* sp.	Luxembourg (Schlirbech, aval Esch-sur-Sûre)	10.10.2006
TCC666	*Fragilaria capucina*	Luxembourg (Sûre, Camping Heiderscheidergrund)	10.10.2005
TCC667	*Achnanthidium minutissimum*	Luxembourg (Sûre, Camping Heiderscheidergrund)	10.10.2006
TCC668	*Discotella* sp.	Italie (Canal, Vérone)	28.10.2006
TCC669	*Fragilaria capucina*	Italie (Canal, Vérone)	28.10.2006
TCC670	*Ulnaria ulna*	Italie (Avisio, Lavis)	27.10.2006
TCC671	*Fragilaria rumpens*	Italie (Brusago)	27.10.2006
TCC672	*Eucocconeis laevis*	Italie (Brusago)	27.10.2006
TCC673	*Fragilaria capucina*	Italie (Brusago)	27.10.2006
TCC674	*Encyonema silesiacum*	Italie (Brusago)	27.10.2006
TCC675	*Encyonema silesiacum*	Italie (Brusago)	27.10.2006
TCC676	*Achnanthidium minutissimum*	Italie (Brusago)	27.10.2006
TCC677	*Fragilaria capucina*	Italie (Brusago)	27.10.2006
TCC678	*Encyonema silesiacum*	Italie (Brusago)	27.10.2006
TCC679	*Achnanthidium* cf. *minutissimum*	Italie (Brusago)	27.10.2006
TCC680	*Gomphonema coronatum*	Italie (Regnana, Amont de Bedollo)	27.10.2006
TCC681	*Fragilaria capucina*	Italie (Regnana, Amont de Bedollo)	27.10.2006
TCC682	*Fragilaria capucina*	Italie (Regnana, Amont de Bedollo)	27.10.2006
TCC683	*Gomphonema parvulum*	Italie (Cervo, Sagliano Micca)	15.12.2006
TCC684	*Achnanthidium minutissimum*	Italie (Cenischia, Plain St Nicolas)	19.12.2006
TCC685	*Achnanthidium helvetica*	Italie (Cenischia, Plain St Nicolas)	19.12.2006
TCC686	*Fragilaria capucina*	Italie (Rocciamelone, Foresto)	19.12.2006
TCC687	*Achnanthidium minutissimum*	Italie (Rocciamelone, Foresto)	19.12.2006
TCC688	*Achnanthidium minutissimum*	Italie (Rocciamelone, Foresto)	19.12.2006
TCC689	*Achnanthidium minutissimum*	Italie (Rocciamelone, Foresto)	19.12.2006
TCC690	*Cyclotella meneghiniana*	Italie (Bormida, Cortemilia)	28.12.2006
TCC691	*Staurosira venter*	Italie (Bormida, Cortemilia)	28.12.2006
TCC692	*Navicula cryptocephala*	Italie (Malone, Lombardore)	23.12.2006
TCC693	*Achnanthidium minutissimum*	Italie (Torrente, Fersina)	27.01.2007
TCC695	*Ulnaria ulna*	Italie (Torrente, Fersina)	27.01.2007
TCC696	*Achnanthidium minutissimum*	Italie (Adige, Verona)	28.01.2007
TCC697	*Fistulifera saprophila*	Italie (Adige, Verona)	28.01.2007
TCC698	*Diatoma tenuis*	Italie (Leno, Rovereto)	27.01.2007
TCC699	*Ulnaria ulna*	Espagne (Carrión, Embalse de Compuerto)	01.11.2006
TCC700	*Nitzschia draveillensis*	Espagne (Pisuerga, Melgar de Fernamental)	01.11.2006
TCC701	*Pinnularia anglica*	Espagne (Tuerto, Astorga)	31.10.2006
TCC702	*Amphora pediculus*	Espagne (Bernesga, Hostal de san Marcos)	02.11.2006

TCC703	*Nitzschia palea*	Espagne (Bernesga, Alija de la Ribera)	02.11.2006
TCC704	*Gomphonema micropus*	Espagne (Bernesga, Alija de la Ribera)	02.11.2006
TCC705	*Fragilaria capucina*	Espagne (Valdavia, Osorno)	01.11.2006
TCC706	*Achnanthidium minutissimum*	Espagne (Valdavia, Osorno)	01.11.2006
TCC707	*Nitzschia dissipata*	Espagne (Valdavia, Osorno)	01.11.2006
TCC708	*Nitzschia palea*	Espagne (Valdavia, Osorno)	01.11.2006
TCC709	*Fistulifera saprophila*	Espagne (Valdavia, Osorno)	01.11.2006
TCC710	*Cyclotella meneghiniana*	Espagne (Valdavia, Osorno)	01.11.2006
TCC711	*Surirella angusta*	Espagne (Valdavia, Osorno)	01.11.2006
TCC712	*Navicula veneta*	Espagne (Pisuerga, Medina de Fernamental)	02.11.2006
TCC713	*Cymbella tumida*	Espagne (Pisuerga, Medina de Fernamental)	02.11.2006
TCC714	*Fistulifera saprophila*	Espagne (Pisuerga, Medina de Fernamental)	02.11.2006
TCC715	*Gomphoneis minutum*	Espagne (Yuso, Boca de Huérgano)	01.11.2006
TCC716	*Ulnaria ulna*	Espagne (Yuso, Boca de Huérgano)	01.11.2006
TCC717	*Encyonema silesiacum*	Espagne (Yuso, Boca de Huérgano)	01.11.2006
TCC718	*Encyonema silesiacum*	Espagne (Yuso, Boca de Huérgano)	01.11.2006
TCC719	*Reimeria sinuata*	Espagne (Esla, Mansilla de las Mulas)	02.11.2006
TCC720	*Cocconeis euglypta*	Espagne (Esla, Mansilla de las Mulas)	02.11.2006
TCC721	*Reimeria sinuata*	Portugal (Rib. De Seixe, Zambujeira De Baixo)	05.12.2006
TCC722	*Fragilaria capucina*	Portugal (Rib. De Seixe, Zambujeira De Baixo)	05.12.2006
TCC723	*Reimeria sinuata*	Portugal (Rib. De Seixe, Zambujeira De Baixo)	05.12.2006
TCC724	*Achnanthidium minutissimum*	Portugal (Rib. De Seixe, Vale de Agoia)	04.12.2006
TCC725	*Gomphonema parvulum*	Portugal (Rib. De Seixe, Foz Do Arroio)	04.12.2006
TCC726	*Achnanthidium minutissimum*	Portugal (Rib. De Seixe, Foz Do Arroio)	04.12.2006
TCC727	*Gomphonema parvulum*	Portugal (Rib. De Seixe, Foz Do Arroio)	04.12.2006
TCC728	*Fragilaria capucina*	Portugal (Rib. De Seixe, Foz Do Arroio)	04.12.2006
TCC729	*Fragilaria capucina*	Portugal (Rib. De Seixe, Foz Do Arroio)	04.12.2006
TCC730	*Gomphonema rosenstockianum*	Portugal (Rib. De Seixe, Arroio De Baixo)	05.12.2006
TCC731	*Gomphonema* cf. *micropus*	Portugal (Rib. De Seixe, Arroio De Baixo)	05.12.2006
TCC732	*Mayamaea permitis*	Portugal (Rib. de Aljezur, Cerca Dos Pomares)	04.12.2006
TCC733	*Achnanthidium minutissimum*	Portugal (Rib. da Cerca, Moinho do Bispo)	04.12.2006
TCC734	*Gomphonema parvulum*	Portugal (Rib. da Cerca, Moinho do Bispo)	04.12.2006
TCC735	*Reimeria sinuata*	Portugal (Rib. de Arão, Pereira)	05.12.2006
TCC736	*Gomphonema parvulum*	Portugal (Rib. de Arão, Pereira)	05.12.2006
TCC737	*Gomphonema acuminatum*	Portugal (Rib. de Arão, Pereira)	05.12.2006
TCC738	*Sellaphora seminulum*	Portugal (Rib. de Arão, Pereira)	05.12.2006
TCC739	*Nitzschia acidoclinata*	Portugal (Rib. de Arão, Pereira)	05.12.2006
TCC740	*Gomphonema rosenstockianum*	Portugal (Bar. da Água Velha, Água Velha)	05.12.2006
TCC741	*Gomphonema rosenstockianum*	Portugal (Bar. da Água Velha, Água Velha)	05.12.2006
TCC742	*Mayamaea fossalis*	France métropolitaine (Lac Léman, estuaire de la Chamberonne)	04.08.2010
TCC743	*Fragilaria capucina* var. *vaucheriae*	France métropolitaine (Lac Léman, estuaire du Foron)	03.08.2010
TCC745	*Mayamaea atomus*	France métropolitaine (Lac Léman, estuaire du Foron)	03.08.2010
TCC746	*Achnanthidium minutissimum*	France métropolitaine (Lac Léman, estuaire du Foron)	03.08.2010

TCC747	*Fragilaria capucina* var. *vaucheriae*	France métropolitaine (Lac Léman, estuaire du Foron)	03.08.2010
TCC751	*Pseudostaurosira elliptica*	France métropolitaine (Lac Léman, proche port de Riex)	04.08.2010
TCC752	*Fragilaria capucina* var. *vaucheriae*	France métropolitaine (Lac Léman, au niveau du parc de Rovorée)	03.08.2010
TCC754	*Nitzschia intermedia*	France métropolitaine (Lac Léman, au niveau du parc de Rovorée)	03.08.2010
TCC755	*Cyclotella meneghiniana*	France métropolitaine (Lac Léman, estuaire de la Venoge)	03.08.2010
TCC756	*Nitzschia palea*	France métropolitaine (Lac Léman, estuaire de la Venoge)	03.08.2010
TCC772	*Cymbella excisa*	Portugal (Rib. De Seixe, Foz Do Arroio)	01.11.2006
TCC773	*Gomphonema bourbonense*	Portugal (Rib. da Perna Seca, Parcanhão)	04.12.2006
TCC775	*Gomphonema rosenstockianum*	Portugal (Rib. de Arão, Pereira)	05.12.2006
TCC776	*Gomphonema rosenstockianum*	Italie (Brusago)	05.12.2006
TCC809	*Fistulifera saprophila*	Portugal (Rib. de Arão, Pereira)	05.12.2006
TCC810	*Melosira* sp.	France métropolitaine (Orne, Joeuf)	17.06.2009
TCC811	*Gomphonema clevei*	La Réunion (Saint Denis, amont prise AEP)	28.04.2009
TCC812	*Nitzschia palea*	Mayotte (Dembeni, aval)	18.04.2009

ANNEXE 2 : PROTOCOLES

Protocoles d'extraction

GenEluteTM-LPA (Sigma-Aldrich)

- Ajouter au culot cellulaire : 300 µl de TE et 200 µL de tampon de lyse.

- Suspendre le culot cellulaire.

- Faire un choc thermique en plaçant les tubes 15 min à – 80°C puis 2 min à 55°C.

- Plonger 20s dans un bain ultrasons.

- Ajouter 50 µL de SDS 10% et 10 µL de protéinase K (20mg.mL-1)

- Incuber à 37°C pendant 1h puis à 55°C pendant 20 min.

- Centrifuger 3 min à 13000rpm et 4°C .

- Transférer le surnageant dans 1 tube de 2 mL.

- Ajouter 0,1 volume d'acétate de sodium 3M, pH 5.2.

- Ajouter 1 µL de GenEluteTM-LPA (Sigma-Aldrich, 25µg.µL-1).

- Laisser à température ambiante pendant 5 à 10 min.

- Précipiter avec 2,5 volumes d'éthanol absolu froid.

- Centrifuger 10 min à 12000g et 4°C.

- Retirer le surnageant.

- Laver le culot avec 200µL d'éthanol 80%.

- Centrifuger 10 min à 12000g et 4°C.

- Retirer le surnageant. Laver le culot au minimum 2 fois avec de l'éthanol 80%.

- Evaporer le reste d'alcool 20 min au Speedvac.

- Resuspendre dans 20-30µL de TE en fonction de la taille du culot d'ADN.

- Incuber 1h, à 37°C.

- Mesurer la concentration d'ADN au Nanodrop.

- Stocker à –20° C.

CTAB - Phénol – Chloroforme, dérivé de Bornet et al. (2004)

- Préchauffer le CTAB 10% à 65°C.

- Ajouter au culot cellulaire : 800µL de tampon CTAB, 200µL de CTAB 10% et 15 µL de protéinase K (20mg/mL)

- Suspendre le culot cellulaire.

- Incuber 1h 30 à 65° C.

- Ajouter un volume de phénol/chloroforme/alcool isoamylique (25:24:1).

- Mélanger et centrifuger 5 min à 10000g et 4°C.

- Transférer la phase aqueuse dans un nouveau tube.

- Ajouter un volume de chloroforme/alcool isoamylique (24:1).

- Mélanger et centrifuger 5 min à 10000g et 4° C.

- Transférer la phase aqueuse dans un nouveau tube.

- Ajouter 0,1 volume d'acétate de sodium 3 M et 0,6 volume d'isopropanol froid (-20°C).

- Agiter par inversion.

- Laisser précipiter toute la nuit à –20° C.

- Centrifuger 15 min à 12000 g et 4° C.

- Eliminer le surnageant.

- Rincer avec 300 µl d'éthanol 70% et centrifuger 10 min à 12000 g et 4° C.

- Eliminer le surnageant. Laver le culot 2 fois avec de l'éthanol 70%.

- Sécher le culot d'ADN au Speedvac.

- Reprendre le culot d'ADN dans du TE (volume choisi selon la taille du culot d'ADN).

- Incuber 1 heure à 37°C.

- Mesurer la concentration d'ADN au Nanodrop.

- Stocker à –20° C.

Protocoles d'amplification des marqueurs

Pour les souches de diatomées

ADNr 18S

Mélange réactionnel

Tampon PCR (avec MgCl2):	1X
Chaque dNTP:	200 µM
Chaque amorce:	1 µM
Taq Polymerase:	1 U
ADN:	50 ng
H_2O:	pour 50 µl

Programme d'amplification

Dénaturation initiale: 94° C, 6 min

30 cycles: Dénaturation: 94° C, 45 s

Hybridation: 57° C, 45 s

Extension: 72° C, 2 min

Extension finale: 72°C, 5 min

ITS1-5.8S-ITS2

Mélange réactionnel

Tampon PCR (avec MgCl2):	1X
Chaque dNTP:	200 µM
Chaque amorce:	0.65µM
Taq Polymerase:	1 U
ADN:	50 ng
H_2O:	pour 50 µl

Programme d'amplification

Dénaturation initiale: 94° C, 6 min

30 cycles: Dénaturation: 94° C, 45s

Hybridation: 53° C, 30s

Extension: 72° C, 30s

Extension finale: 72°C, 5 min

ADNr 28S

Mélange réactionnel

Tampon PCR (avec MgCl2):	1X
Chaque dNTP:	200 µM
Chaque amorce:	0,1 µM
Taq Polymerase:	1,25 U
ADN:	50 ng
H_2O:	pour 50 µl

Programme d'amplification

Dénaturation initiale: 94° C, 5 min

35 cycles: Dénaturation: 94° C, 1min
Hybridation: 45° C, 45 s
Extension: 72° C, 1 min

Extension finale: 72°C, 10 min

rbcL

Mélange réactionnel

Tampon PCR (avec MgCl2):	1X
Chaque dNTP:	200 µM
Chaque amorce:	1 µM
Taq Polymerase:	1.5 U
ADN:	50 ng
H_2O:	pour 50 µl

Programme d'amplification

Dénaturation initiale: 94° C, 6 min

30 cycles: Dénaturation: 94° C, 45 s
Hybridation: 52° C, 45 s
Extension: 72° C, 2 min

Extension finale: 72°C, 5 min

cox1

Mélange réactionnel

Tampon PCR (avec MgCl2)	1X
Chaque dNTP:	200 µM
Chaque amorce:	0.3 µM
Taq Polymerase:	1,25 U
ADN:	50 ng
H_2O:	pour 50 µl

Programme d'amplification

Dénaturation initiale: 95° C, 3 min

35 cycles: Dénaturation: 95° C, 30s
Hybridation: 50° C, 1 min
Extension: 72° C, 1 min 30s

Extension finale: 72°C, 5 min

Pour les échantillons environnementaux

ADNr 18S:

Mélange réactionnel

Tampon PCR (avec MgCl2):	1 X
Chaque dNTP:	200 µM
Chaque amorce:	0,5 µM
Pfu:	1,25 U
ADN:	25 ng
H_2O:	pour 50 µl

Programme d'amplification

Dénaturation initiale: 95 °C, 2 min

30 cycles: Dénaturation: 95 °C, 45 s
Hybridation: 55 °C, 45 s
Extension: 72 °C, 4 min

Extension finale: 72 °C, 5 min

rbcL

Mélange réactionnel

Tampon PCR (avec MgCl2):	1 X
Chaque dNTP:	200 µM
Chaque amorce:	1 µM
Taq Polymerase:	1,5 U
ADN:	25 ng
H_2O:	pour 50 µl

Programme d'amplification

Dénaturation initiale: 94 °C, 6 min

30 cycles: Dénaturation: 94 °C, 45 s
Hybridation: 51 °C, 45 s
Extension: 72 °C, 2 min

Extension finale: 72 °C, 5 min

cox1

Mélange réactionnel

Tampon PCR (avec MgCl2):	1 X
Chaque dNTP:	200 µM
Chaque amorce:	1 µM
Taq Polymerase:	1,5 U
ADN:	25 ng
H_2O:	pour 50 µl

Programme d'amplification

Dénaturation initiale: 95° C, 2 min

35 cycles: Dénaturation: 95° C, 45s
Hybridation: 47° C, 45 s
Extension: 72° C, 1 min 30 s

Extension finale: 72°C, 5 min

ANNEXE 3 : SEMINAIRES, COLLOQUES NATIONAUX ET INTERNATIONAUX

Présentation de posters:

2009

- Development of molecular tools to use diatoms as bioindicators of watercourses quality and to study their taxonomy and ecology. **Kermarrec L.**, Bouchez A., Rimet F., Monnier O., Humbert J.-F.. Geneva Barcoding Day, Geneva (Suisse), 22 Juin 2009.

- Preliminary phylogenetic analysis of Cymbellales from sequences of the 18S rRNA gene. **Kermarrec L.**, Ector L., Bouchez A., Rimet F. & Hoffmann L.. Van Heurck Symposium, Meise (Belgique), 23 - 26 Août 2009.

2010

- Genotypic and morphometric characterization of temperate and tropical *Gomphonema parvulum* isolates (Bacillariophyta). **Kermarrec L.**, Rimet F., Humbert J.-F., Bouchez A.. Journées Internationales de Limnologie, Thonon-les-Bains (France), 5 - 8 Octobre 2010.

2011

- Séquençage nouvelle génération pour caractériser les communautés de diatomées bioindicatrices : comparaison de 3 marqueurs moléculaires. **Kermarrec L.**, Rimet F., Franc A., Chaumeil P., Humbert J.-F. and Bouchez A.. Colloque Génomique Environnementale, Lyon (France), 28 - 30 novembre 2011.

Communications

2009

- Développement d'outils moléculaires pour l'utilisation des diatomées comme bioindicateurs de la qualité des écosystèmes aquatiques lotiques. **Kermarrec L.**. Séminaire de laboratoire, Thonon-les-Bains (France), 2 Octobre 2009

2010

- Genetic diversity within the *Gomphonema parvulum* complex. **Kermarrec L.**. Journée des Doctorants, Thonon-les-Bains (France), 25 Mai 2010.

- Genotypic and morphometric characterization of temperate and tropical *Gomphonema parvulum* isolates. **Kermarrec L.**, Rimet F., Humbert J.-F., Bouchez A.. International Diatom Symposium, Saint-Paul (USA), 29 Août - 3 Septembre 2010.

- Caractérisation génotypique et morphométrique de souches tempérées et tropicales de *Gomphonema parvulum*. **Kermarrec L.**, Rimet F., Humbert J.-F., Bouchez A.. 29ème colloque de l'Association des Diatomistes de Langue Française, Québec (Canada), 6 - 10 Septembre 2010.

- Bioindication par les diatomées en rivière : perspectives pour les milieux tropicaux. **Kermarrec L.**, Rimet F., Bouchez A.. Bioindication des écosystèmes aquatiques d'eau douce en milieux tropicaux, Fort de France (Martinique, France), 1 - 3 Décembre 2010.

2011

- Séquençage nouvelle génération pour déterminer les communautés de diatomées : comparaison de 3 marqueurs moléculaires. **Kermarrec L.**, Rimet F., Franc A., Chaumeil P., Humbert J.-F. and Bouchez A.. 30ème colloque de l'Association des Diatomistes de Langue Française, Boulogne-sur-Mer (France), 6 - 8 Septembre 2011.

- Diversité in vivo, multi-coeurs in silico: relier métagénomique et taxonomie moléculaire. Franc A., Chaumeil P., Frigerio J.-M., **Kermarrec L.**, Rimet F., & Bouchez A.. Journée masses de données et calculs intensifs, Paris (France), 7 Octobre 2011.

- Efficiency of 3 genetic markers to determine the composition of a diatom assemblage using Next Generation Sequencing. **Kermarrec L.**, Rimet F., Franc A.,

Chaumeil P., Humbert J.-F. and Bouchez A.. 4th International Barcode of Life Conference, Adelaïde (Australie), 30 Novembre - 3 Décembre 2011.

- A pipeline for automatic taxonomic community inventories on data from NGS: An example with freshwater diatoms. **Kermarrec L.**, Chaumeil P., Rimet F., Frigerio J.-M., Laval V., Lebrun M.H., Bouchez A. and Franc A.. 4th International Barcode of Life Conference, Adelaïde (Australie), 30 Novembre - 3 Décembre 2011.