The use of evolutionary biology concepts for genome annotation

Etienne Danchin, Anthony Levasseur, Virginie Lopez Rascol, Philippe Gouret, Pierre Pontarotti

The Use of Evolutionary Biology Concepts for Genome Annotation

ETIENNE G.J. DANCHIN[1], ANTHONY LEVASSEUR[2,3]
VIRGINIE LOPEZ RASCOL[2], PHILIPPE GOURET[2],
AND PIERRE PONTAROTTI[2]*
[1]*Glycogenomics and Biomedical Structural Biology, AFMB Laboratory, UMR 6098, CNRS, Universités d'Aix-Marseille I et II, 13288 Marseille, France*
[2]*Phylogenomics Laboratory, EA 3781 Evolution Biologique Université de Provence, 13331 Marseille, France*
[3]*UMR 1163 INRA de Biotechnologie des Champignons Filamenteux, IFR86-BAIM, Universités de Provence et de la Méditerranée, ESIL, 13288 Marseille, France*

ABSTRACT The past decade has seen the completion of numerous whole-genome sequencing projects, began with bacterial genomes and continued with eukaryotic species from different phyla: fungi, plants and animals. Besides, more biological information are produced and are shared thanks to information exchange systems, and more biological concepts, as well as more bioinformatics tools, are available. In this article, we will describe how the evolutionary biology concepts, as well as computer science, are useful for a better understanding of biology in general and genome annotation in particular. The genome annotation process consists of taking the raw DNA produced, for example, by the genome sequencing projects, adding the layers of analysis and interpretation necessary to extract its biological significance and placing it in the context of our understanding of biological processes. Genome annotation is a multistep process falling into two broad categories: structural and functional annotation. *J. Exp. Zool. (Mol. Dev. Evol.) 308B:26–36, 2007.* © 2006 Wiley-Liss, Inc

How to cite this article: Danchin EGJ, Levasseur A, Rascol VL, Gouret P, Pontarotti P. 2007. The use of evolutionary biology concepts for genome annotation. J. Exp. Zool. (Mol. Dev. Evol.) 308B:26–36.

The first step in genome annotation is to identify the structural characteristics, boundaries and organization of protein-coding gene, RNA-coding gene, sequences from retroviral origin and other features. When, for example, a protein-coding gene is found, the next step is to predict the protein or the proteins (in the case of a gene that give rise to several spliced variants), and when the protein is deciphered the functional annotation is possible through different methods of analysis. The main goal of this paper consists in emphasizing important evolutionary biology concepts for structural and functional annotation. We will show how such concepts could improve predictions and should be integrated in future annotation platforms.

STRUCTURAL AND FUNCTIONAL ANNOTATION USING EVOLUTIONARY BIOLOGY CONCEPTS

The annotation of protein-coding sequences can be split into two complementary tasks, structural annotation and functional annotation.

Grant sponsor: STREP FungWall European Consortium; Grant number: LSH 2003 1.2.5.5; Grant sponsor: ANR05 BIO-006 ARD.

*Correspondence to: P. Pontarotti, Phylogenomics Laboratory, EA 3781 Evolution Biologique Université de Provence, Case 36, Pl. V. Hugo, 13331 Marseille, Cedex 03, France.
E-mail: pontarot@up.univ-mrs.fr

Received 10 April 2006; Accepted 27 June 2006
Published online 2 October 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jez.b.21131.

Structural annotation

The structural annotation consists in localizing genome features such as coding sequences. When a coding sequence is localized, the next step is to predict the intron/exon organization and infer the sequence of the corresponding protein. This step is very important for the functional annotation as for example a missed domain could be dramatic for the functional inference. The most efficient programs for protein sequence prediction use ab initio along with similarity-based programs (Mathe et al., 2002). However, such programs require that homologous proteins are found in biological databases. In fact, these approaches are partially based on evolutionary biology concepts but this has never been clearly stated—even if this is intuitive. The approach we developed for the structural annotation is strongly based on such concepts as the annotation is based, as for the mixed ab initio/similarity program, on the finding of similar proteins.

When proteins sharing significant similarities are found, this indicates that the proteins could be homologous, which means that they originate from a common ancestral gene. This common ancestor evolved toward the genes coding for these proteins, as well as the other members of the family, by substitution in the coding or the non-coding region, 5′and 3′ exon extension, shift in the acceptor and donor sites, exon(s) losses and gains. All these events have to be modeled by the algorithm used for the structural annotation. Practically, the sequence containing the gene to be annotated is used to search against protein databases with the BLASTX algorithm (Altschul et al., '97); the sequence is divided into non-overlapping segments that match different parts of one or several proteins. Each segment that corresponds to independent genic units is then treated independently.

The regions of higher similarity with the protein (the most significant BLAST hits) are located in the fragment of sequence to annotate: such regions should correspond to coding exons. Then all the donor and acceptor splicing sites around the islands of similarity are searched. This includes also the search for the 5′ and 3′ exons and the non-conserved ones.

Different protein solutions are constructed using the different predicted exons and splicing sites, the protein solution giving the best alignment with the presumed homologous proteins is considered the best candidate (this is done using BLAST algorithm). In the future, instead of using one protein for the comparison, it would be of great interest to use the protein family alignment (or HMM). This could allow us to avoid apomorphies (derived characters) that could be present in the protein most similar to the gene to annotate. This solution will also permit the method to be less sensible to wrongly annotated proteins present in biological databases (i.e. incorrect gene models with missed or additional exons or frameshifts). It should be noted that some proteins are orphan (no similar sequence is found in the databases) in this case ab initio programs have to be used.

Functional annotation

As for the structural annotation, the approaches based on evolutionary biology concepts should be developed for improving functional annotation. Ancestrally a gene product has a given function: this function can change in the daughter genes (gene originating via descent transmission or duplication) due to mutational events on the gene. The functional shifts can be revealed at biochemical level as well as higher levels of the organism organization (e.g. cellular processes, physiology or social organization). These functional shifts are also called co-option (Ganfornina and Sanchez, '99).

It should be noted that a peculiar case of co-option can occur without shift of the original function; in other words, no mutational event is required for the apparition of a novel function. A given gene product like an enzyme (e.g. the gene products constituting the proteasome) can indeed be used in a new pathway (antigen presentation to MHC acquired during the evolution of vertebrate) without changing its basic biochemical activity (Danchin et al., 2004). However, co-option events occur mainly following mutations that: (1) change the coding sequence properties and therefore possibly the activity and/or cellular localization of the concerned protein (2) change the transcriptional pattern (tissue/level of expression) through modification in regulatory regions (see for review True and Carroll, 2002).

The emergence of a new biochemical activity is illustrated by the *epsilon crystallin/LDHB4*. This gene encodes a product that acts both as an enzyme (lactate dehydrogenase) and as lens crystalline (without known enzymatic activity) in the case of birds and some reptiles. In contrast, in the other species it only acts as lactate dehydrogenase indicating that this gene has been co-opted as crystalline specifically in the sauropsidae lineage. The subcellular localization (targeting) innovation has been described in the case of the cox gene family (Schmidt et al., 2003). After duplication event, one of the gene copies still encodes for a protein located in the mitochondria whereas the paralogous gene (the other copy) codes for a product located in the Golgi apparatus. Similar evolutionary scenario has been described for several genes involved in the MHC peptide presentation (Danchin et al., 2004) and in the case of alanine glyoxylate aminotransferase (Birdsey et al., 2004). The shift in tissue/territory expression has been described in the case of the Zeta Crystallin/quinone-oxidoreductase gene coding for a protein which acts as crystallin in the guinea pig lens and also acts as a quinone oxidoreductase in the lens and other tissues, thanks to two different promoters, one used for non-lens expression, the other one used for lens expression. In some species, the crystallin and quinone-oxdoreductase are found to be encoded by two related but distinct genes.

These shifts either in molecular function, subcellular localization or transcriptional tissue-specific activity can have an impact at different functional level of the organism. For example a neo-expression of a ‘‘master’’ regulator gene can permit the expression of several genes from the same cascade in a new cellular environment. One of the most famous examples is that of the Dll gene and the butterfly eyespots (True and Carroll, 2002).

These important events in the evolutionary history can be deciphered by a phylogenetic analysis. Using the sequence shift information the gene genealogy can be reconstructed; then the function genealogy can be superimposed to the gene genealogy. As more information and more refined methods are available for biological sequence data, reconstructing a tree that deciphers the evolutionary history of genes, and proteins will be more straightforward and accurate than reconstructing a tree that traces back the evolutionary history of function. Moreover, sequence-based phylogenetic tree can help in inferring function by superimposing functional information on the phylogenetic tree (Engelhardt et al., 2005).

However, beside the phylogeny itself other information can be useful for functional annotation in a phylogenetic tree based on sequence information.

ORTHOLOG/PARALOG INFORMATION

In most of the cases, the species from which the sequences originated are known as well as the species tree. Therefore, homologous genes issued via speciation or duplication can be differentiated (orthologs and paralogs). Bibliographic analysis (for example, Collette et al., 2003) indicated that orthologs have more chance to keep the similar function compared to paralogs. This can be also argued theoretically since after duplication either one of the copy is lost, or both duplicates undergo subfunctionalization, or one of the duplicate evolves toward a new function (Force et al., ’99). By function, Force et al. meant biochemical function or expression pattern meaning that a functional shift corresponds, for the authors, either to a functional biochemical shift or to a transcriptional shift. In the later case, we witness a semantic problem as the authors confound pattern of expression and its potential functional output. Therefore, at the molecular level paralogs can be either biochemically subfunctionalized or neofunctionalized. They will have therefore a different biochemical function, but in the case of neofunctionalization, one of the copies will retain the ancestral function. Note that the paralog that undergoes neofunctionalization can be identified by the evolutionary shift analysis (see after).

At the transcriptional level: in the case of neotranscription events, one of the copies will retain the ancestral transcription pattern; in the case of subtranscription the two copies will have a complementary pattern that will recapitulate the one of the preduplicate copy and the non-duplicate ortholog.

EVOLUTIONARY SHIFT AND FUNCTIONAL SHIFT (SEE FOR REVIEW GAUCHER ET AL., 2002, YANG AND BIELAWSKY, 2000)

Phylogenetic methods, as well as computational methods, in general that generate hypothesis about function from sequence evolution can be valuable. Patterns of replacement including change in the rate of replacement are likely to be important to these methods. For example, the functional importance of sites is intuitively inversely related to the evolutionary rate of amino acid replacements. This intuition arises from one interpretation of the neutral theory of evolution in which the sites of the greatest functional significance are under the strongest selective constraint. An organism that experiences a replacement at one of these site is less likely to survive and therefore reproduce. In some cases, the extent to which function constrains the evolution of a protein sequence can be estimated by measuring the ratio of non-synonymous to synonymous substitution during its evolution. This ratio is also used to detect positive selection in coding DNA which in turn could be linked to functional shift. To assess more broadly the possible functional significance of sequence evolution, new approaches that consider amino acid replacements (non-synonymous substitution) alone and that are based on the measuring of the ratio of non-synonymous to synonymous substitution have emerged.

Method based on amino acid replacement

These methods begin by analyzing how the evolutionary rates of amino acid replacements differ among sites in a protein sequence (site to site rate heterogeneity), with a statistical formalism in which the rate varies among sites according to a gamma distribution. In a conventional analysis of sequence evolution using the gamma model, rapidly and slowly evolving sites remain rapid or slow across the entire evolutionary tree. Because of this the model is termed homogeneous. A homogeneous evolutionary rate is expected when the functional constraints at sites are constant for the entire evolutionary history. However, if the function of the protein is changing, some residues might be subjected to altered functional constraints in various places of the phylogenetic tree. This in turn implies that the evolutionary rates at these sites will be different in different branches of the tree (heterotachy). To capture this phenomenon, the constraint of fixed rates per site along the phylogeny must be relaxed to allow the identities of fast and slow sites to change over time that are to allow site-specific rate shifts. This process is a non-homogeneous gamma model. Rate shifted sites are the residues that have either enhanced or reduced selective constraint as a possible consequence of the change of function during protein evolution.

Comparison of silent and replacement sites

Beside the approach described above, another possible effective approach is to contrast the rates at which synonymous (silent) d_S and non-synonymous (replacement) d_N mutations are fixed in the history of a given gene. The silent rate d_S provides a benchmark against which we can decide whether the replacement rate d_N is accelerated or diminished possibly by natural selection on the protein (Miyata and Yasunaga, '80). Thus $d_N < d_S$, $d_N = d_S$, $d_N > d_S$, represent negative (purifying) selection, neutral evolution and positive selection, respectively. A problem with this criterion is its lack of discriminative power. Most proteins have highly conserved region where replacements are not tolerated, and d_N is almost 0. Furthermore, adaptive evolution may occur in an episodic fashion and only in a narrow window of time. Comparison of a pair of genes averaging the d_N and d_S rates over all sites in the protein and over the whole time period separating the two sequences, fails to infer positive selection, because the signal of positive selection is overwhelmed by the ubiquitous purifying selection. To boost the power of the detection method, work has focused on detecting selection that affects individual sites (Miyata and Yasunaga, '80) as opposed to the whole protein or particular lineage as opposed to the whole phylogeny (see for example Yang, '98) taking into account models of variable selection pressure among sites and lineages (see, for example, Yang and Nielsen, 2002).

Synonymous substitutions are most of the time neutral (and therefore occur at relatively rapid rate). Hence d_N over d_S ratio can be used to detect only recent functional divergence, as synonymous sites rapidly become saturated with mutation. For a typical vertebrate nuclear encoded gene, this type of analysis has been useful to detect events only as far back as around 150 million years ago. It should be noted that these methods have been used in few cases for older events (Rodriguez-Trelles et al., 2003; Bos, 2005).

The sites deciphered by methods based on amino acid replacement or by methods based on comparison of silent and replacement sites can be further evaluated for their roles in functional divergence by mapping them onto the available tertiary (or three-dimensional) structures of their protein.

It should be noted, however, that few examples of relaxed or positive selection have been linked to actual functional shifts (manuscript in preparation). Beside amino acid substitution, larger-scale gene rearrangements can give rise, for example, to domain loss or gain occurred during protein evolution. Theses events could have a big impact on the overall protein function.

ORTHOLOGS/PARALOGS EVOLUTIONARY SHIFT AND TRANSCRIPTION SHIFT

Several studies indicate that ubiquitous genes evolve more slowly (rate of protein sequence evolution) than tissue-specific genes (Hastings, '96; Duret and Mouchiroud, 2000; Zhang and Li, 2004; Balandraud et al., 2005; Lemos et al., 2005). This has been shown at the general level by comparing the evolutionary rate of ubiquitous genes vs. tissue-specific genes (Duret and Mouchiroud, 2000; Zhang and Li, 2004; Lemos et al., 2005), or at the level of specific families (Hastings, '96; Balandraud et al., 2005). Indeed the paralog with a broader distribution evolves slower that the paralog with narrower distribution. One can hypothesize that the ubiquitous expression could correspond to the ancestral distribution. One copy has a distribution closer to the ancestral one while the other copies undergo shifts in their expression pattern which are different and more limited. Therefore, an evolutionary shift at sequence level could indicate a shift in the transcription pattern and indicate therefore a biochemical and transcriptional shift with a possible shift at higher organization level.

HOW CAN WE RETRIEVE THE FUNCTIONAL INFORMATION?

Here we have two questions:

1. Where to find the information (how it is defined and organized)?
2. How to use it the most efficiently possible?

Where and how to find the information?

Information can be found in published articles and in several curated databases. The use of an ontology dedicated to biological function is valuable for such a task. In computer science, ontology is the product of an attempt to formulate an exhaustive and rigorous conceptual schema about a domain. Ontology is typically a hierarchical data structure containing all the relevant entities and their relationships and rules within that domain (e.g. a domain ontology). We use in biology the computer science usage of the term ontology which is derived from the much older usage of the term ontology in philosophy the study of the nature of being, reality, and substance. Several ontologies exist in biology the most famous one being gene ontology (GO) (Ashburner et al., 2000). It provides an important resource for describing the functional characteristics of sequences. GO contains three ontologies, describing the biological process, cellular compartment and molecular function properties of the sequences; it should be noted that other ontologies such as ontology of cell types are developed. Each ontology is a directed acyclic graph of functional term nodes where edges between nodes describe relationship between them. The terms are organized in a hierarchical way according to parent–child relationships. This allows a progressive functional description, matching the current level of experimental characterization of the corresponding gene product.

One of the problem of using evolutionary-based approach along with GO is that the filiations relationships in GO are analogy filiations and are not completely based on evolutionary relationship. We give here as an example, a simplified graph found in GO in the case of TNGR receptor I: BINDING → PROTEIN BINDING → CYTOKINE BINDING → TNF BINDING → TNF RECPTOR ACTIVITY and in parallel, a second hierarchy: SIGNAL TRANSDUCER ACTIVITY → RECEPTOR ACTIVITY → TRANSMEMBRANE RECEPTOR ACTIVITY → DEATH RECEPTOR ACTIVITY → TNF RECEPTOR ACTIVITY. Most of the nodes are not evolutionarily linked; for example, different receptors have different origins. Indeed, not all receptors have protein-binding function, and not all proteins having receptor activity are signal transducers. As opposed, all vertebrates are bilaterian animals, all bilaterian animals are metazoan eukaryotes, all insects are arthropods, etc.

Why is it interesting to have an ontology based on evolutionary biology?

An ontology based on evolutionary biology is informative since as said above protein classification via phylogenetic analysis allows deciphering the gene history which in turn is related to the history of function. A gene has a given function in an ancestral species. The gene will evolve via speciation and duplication events and give rise to a series of homologous genes (paralogs and orthologs). Structural shift in the children gene can give rise to new function (at different level of organization). If the function is known for homologs and that the evolutionary history is known through a phylogenetic tree, then the function can be inferred for certain nodes and, therefore, for some leafs representing genes present in modern species. If we take the case of molecular function, closer to the root broader is the definition of the molecular function. If the phylogenetic history of the function is described in the GO graph, this will permit to annotate all the nodes at more or less precise levels. In the case where the phylogenetic history of the function cannot be deciphered from GO which is actually the case, only the terminal nodes can be annotated.

The tyrosine kinase family is taken here as an example (Fig. 1). This family is a huge family all the experimentally identified members are involved in protein tyrosine kinase activity. Therefore, this information can be reported for all the nodes of the tree, and then the family diverges into two monophylogenetic groups: receptor tyrosine kinases (RTKs) and non-receptor (cytosolic) tyrosine kinases. The RTK ancestor emerged via a shuffling of a tyrosine kinase and a member of the immunoglobulin family. Therefore, in the case of RTK, the entire node can be annotated as RTK as it is likely that the ancestor of this family was already an RTK. The RTK family is split in different groups: FGFR, VGR as well as other. In the case of FGFR, all the experimentally known members of this monophylogenetic group interact with fibroblast growth factor (FGF) ligand. Thus it is likely that the ancestor of this phylum already interacted with FGF and therefore all the members of this family should be involved in such functions. The FGFR phylum includes four subphyla: FGFR1, FGFR2, FGFR3 and FGFR4. The duplication that gave rise to these four paralogous groups arose in the vertebrate lineage after the separation of the cephalochordates and chordates ancestors, but before the jawed vertebrate radiation. The FGFR during this evolution became specialized in recognizing different subsets of FGF ligands (Coulier et al., '97).

For example, if a new member of the kinase family is found and if it does not belong to the FGFR or the VGR family and is included in the RTK subgroups, using a phylogenetic functional analysis approach we can propose that the gene should be annotated as an RTK. This will be impossible if the functional information is not hierarchically based on evolutionary biology.

Therefore, the evolutionary-based approach ontology should be developed, not only at the molecular level but also at more complex levels such as at cellular level, tissues level or organ.

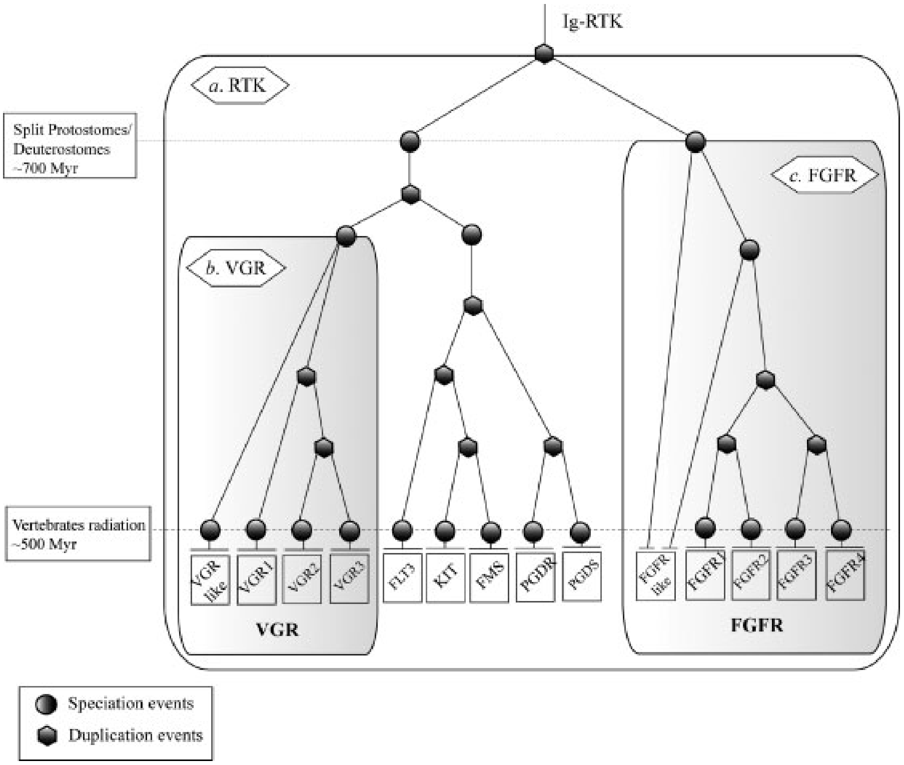

Fig. 1. The RTK family: an example for an ontology based on evolutionary biology. Based on the phylogenetic history, nodes and their corresponding leaves can be functionally annotated as (**a**) RTK, (**b**) VGR or (**c**) FGFR. The function of a new member of the kinase family could be predicted according to its evolutionary history and its belonging to a particular node of the RTK family.

PHYLOGENETIC-BASED FUNCTIONAL ANNOTATION (PBFA)

PBFA which is only a part of the use of the evolutionary concept for functional annotation can be summarized in the following steps:

- creation of a dataset of sequences similar to the query sequence,
- multiple alignment of these sequences with elimination of data distorting the evolutionary signal,
- phylogenetic reconstruction,
- inference of function,

Different platforms taking into account the PBFA have already been developed, i.e. Orthostrapper (Storm and Sonnhammer, 2002), RIO (Zmasek and Eddy, 2002), SIFTER (Engelhardt et al., 2005) and FIGENIX (Gouret et al., 2005). In the following section, we present as a general example, the detailed protocol applied in the FIGENIX platform.

Clustering homologous proteins

The first step in a PBFA analysis in particular (and for any phylogenetic analysis) generally requires the identification of proteins related to the protein of interest (the query) (Fig. 2). At this step of the analysis, one has to be sure that all the proteins share a common ancestor since similarity between two proteins can be due to convergence, if the similarity score is high this is likely. The similarity search is done usually by BLAST and PSI-BLAST to increase the search sensitivity (Altschul et al., '97). Given a seed sequence, PSI-BLAST iteratively searches a sequence database to identify and align putative homologs from which a profile (HMM) is constructed for database search in the next iteration. This step is important since sufficient sequence identity is needed to enable the generation of an accurate multiple sequence alignment (MSA) and therefore a phylogenetic tree.

Multiple sequence alignment (MSA)

At this step, an MSA of the sequences retrieved from the database search is constructed. The accuracy of the alignment is critical since it is the source of the phylogenetic signal for the tree construction. We have concrete and detailed data in the performance of MSA methods through the use of benchmark datasets (Thompson et al., 2005,) which allow ranking the different alignment tools. Among the various tools, MUSCLE seems to be one of the best performing (Edgar, 2004a,b).

The next step involves removing potential non-homologs. To accomplish this, the MSA should be examined to identify critical motifs or conserved residues followed by removal of sequences not matching the consensus structure of the family as a whole. The next step involves alignment masking to prevent the intrusion of too variable regions. Two main steps are used in the lab:

i. When known domains are found through a PFAM database search with HMMER (Bateman et al., 2000, 2004), we concatenate the different domains after having tested that the evolutionary history of the various different domains is congruent. If a particular domain has a ‘‘divergent’’ evolutionary history compared to the others, the corresponding portion of the multiple alignment is deleted.
ii. When no known domains are found, the ‘‘alignable’’ portions of full lengths sequences are kept.

In both cases, we delete columns that appear unreliable or include many gaps (for more details about the whole process see Gouret et al., 2005) who describe how the MSA step has been implemented in the FIGENIX platform). We have to add here a note of caution: even if it is our choice to mask highly variable regions of a multiple alignment, this could have a potential impact for the search of the actual phylogeny. Regions outside the conserved core can play important functional roles such as determining binding specificity. The binding pocket positions are not always structurally conserved across all the superfamily members, and may shift (along with changes in substrate specificity) to form different pockets and clefts in different subgroups. In these cases the information outside the conserved core may be necessary for the tree topology accuracy even more in the case of the closer members. The next step consists in eliminating sequences distorting the phylogenetic signal. One of the most known examples of distortion of phylogenetic signal is the composition in amino acids bias. Indeed the phylogenetic signal is based on the protein evolution model which is calculated from matrix obtained from the alignment of different protein

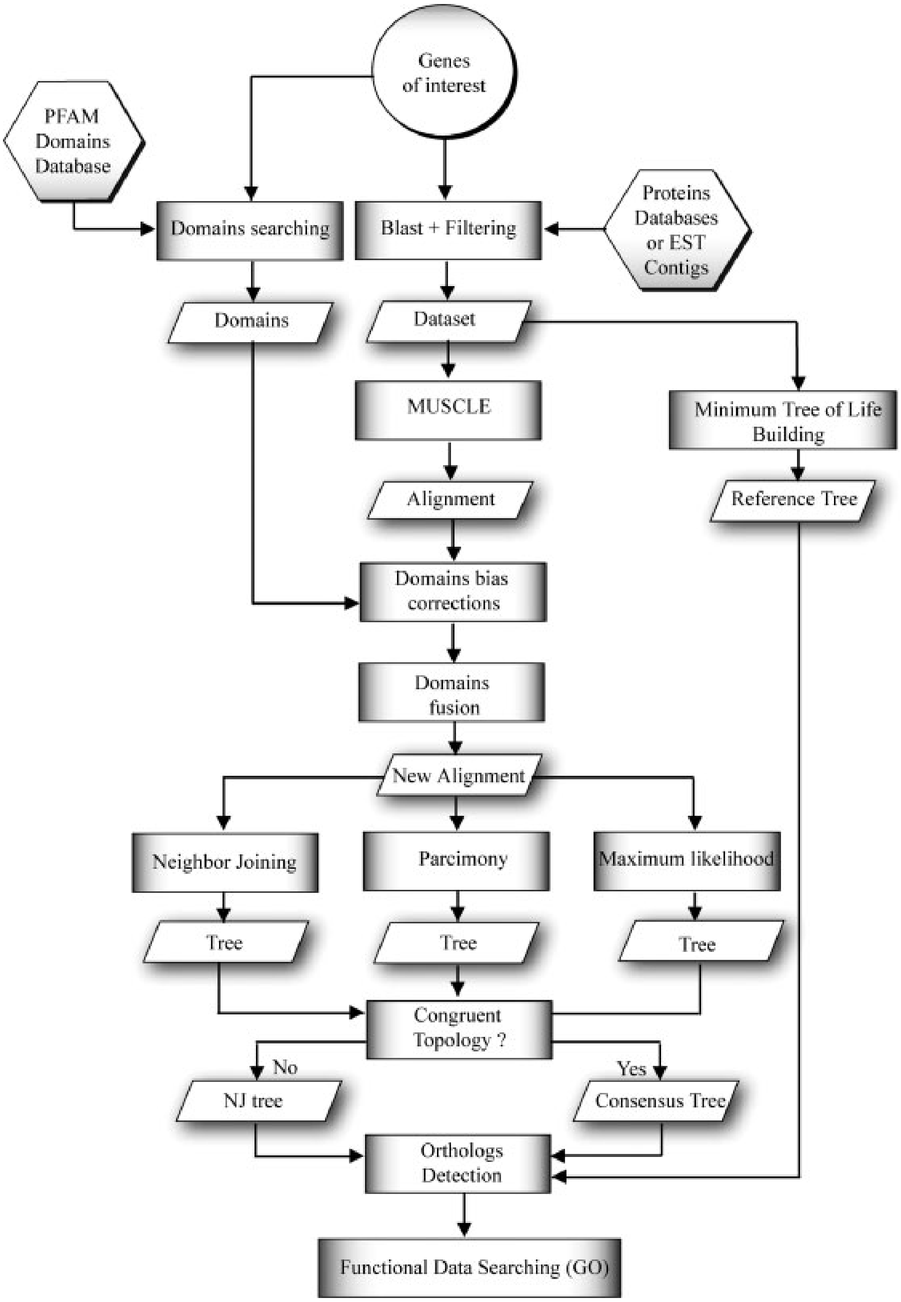

Fig. 2. Schematic representation of the PBFA pipeline. This flowchart summarizes the main steps of the phylogenetic-based functional annotation pipeline (see section *Multiple sequence alignment* for detailed description).

families; if the amino acid composition is not homogeneous, the probability law of the substitutions from one particular amino acid to another will be different and, therefore, the phylogenetic output will also be different. The puzzle test (Schmidt et al., 2002) can reveal sequences with different amino acid composition and such sequences can be eliminated from further analysis.

Phylogenetic tree construction

An important point has to be underlined: a protein is usually composed of different domains and the domain could have different evolutionary histories due to exon shuffling. The exon shuffling can occur via homologous genes; therefore, such events cannot be identified on the alignment alone but the phylogenetic analysis will decipher such an event as the topology between the two domains' phylogenetic trees will be different. If the shuffling occurred via non-homologous genes, the alignment with the rest of the family will be partial and the sequence will be eliminated from the analysis. As each domain can carry a specific function, such information can be added in the phylogenetic tree.

There are two main classes of phylogenetic tree construction methods: distancebased (neighbor joining) and characterbased (maximum parsimony, maximum likelihood and Bayesian method). Distance-based methods compute matrix of pairwise distances between sequences in an alignment and thereafter ignore the sequences themselves constructing a tree based entirely on the original distance computation. The distance of character-based computation can be calculated using different matrices. These matrices use maximum likelihood estimates based on family alignments (examples: Dayhoff PAM matrix model, the JTT matrix model), or a model based on the genetic code plus a constraint on changing to a different category of amino acids. The distances can also be corrected for gamma-distributed and gamma-plus-invariant-sites-distributed rates of change in different sites. Rates of evolution can vary among sites in a prespecified way, and also according to a Hidden Markov model.

Unfortunately, no biological datasets exist to assess phylogenetic tree method directly. The community has, therefore, no way of knowing the true evolutionary tree underlying a protein superfamily. For this reason, all experimental validations of phylogenetic inference methods have been performed on simulated data and results relevant to protein superfamily are inconclusive. In the lab we have chosen to calculate the trees with the three different methods (NJ, maximum parsimony and maximum likelihood).

Given the same multiple sequences alignment, two reconstruction methods will produce at least two trees and sometimes many more, for example, the maximum parsimony tree will produce many hundreds of equally parsimonious trees. Closely related subgroups are found reliably by most tree methods, with most of the differences between trees being found to the deeper node in the tree. To avoid any systematic biases of particular method, we combine bootstrap analysis with different tree methods. The next step is to compare the topology by an adequate algorithm such as the Hasegawa test (Kishino and Hasegawa, '89) and look for the congruence of the trees. When the three trees are congruent, a fusion is realized; in the case if one of the tree is not congruent with the other, we fused only two trees. In the case where the three trees are not congruent, no fusion is possible; the default choice is the maximum likelihood. In any case, the nodes supported by the three methods can be underlined. The next step is to differentiate between orthologs and paralogs among sequences in the tree. Several approaches not based on phylogenetic analysis claim to find orthology and one of the most popular is based on clustering method, such as Inparanoid (Remm et al., 2001). The method of clustering requires complete genome and gives erroneous information when in the case of lineage-specific differential paralog loss (see for examples Danchin et al., 2006). This is not the case for orthologs and paralogs identification based on phylogeny. Furthermore, when phylogenetic trees are constructed, specific algorithms have to be applied to distinguish between orthologs and paralogs, two such tools have been developed for this task (Zmasek and Eddy, 2002; Dufayard et al., 2005).

Inference of function

This step consists in deducing the function from the phylogenetic tree, taking into account the orthologous/paralogous phylogenetic shift criteria. PBFA approaches have already been proposed to address the systematic errors of protein function prediction and improve the accuracy of functional classification (Sjölander, 2004).

When an ad hoc phylogenetic tree is constructed, functional information can be used to label the proteins in the tree and different methods can be used for the inference. This is done by retrieving experimentally verified functional data for orthologs and paralogs to the query sequence on WEB databases (GO, MGI and NCBI's dbEST) and by deducing the function for the non-annotated protein in the tree. One way to determine the ancestral function and therefore predict today's function is to use statistical model of molecular function evolution to propagate all observed molecular function annotations throughout phylogeny. This has been developed by Engelhard et al. (2005). This method has to be distinguished from other methods that exploit phylogenic information that has been described such as orthostrapper (Storm and Sonnhammer, 2002) and RIO (Zmasek and Eddy, 2002) as these methods simplify the problem by extracting pairwise comparison from the phylogeny and by using heuristics to convert these comparisons into annotations. Note also that the laboratory has developed a software platform called FIGENIX that simplifies the problem by propagating the annotation from one ortholog to another ortholog (for a comparative analysis with other softwares, see Gouret et al., 2005). It should be noted that SIFTER (Engelhardt et al., 2005) takes into account the paralogy/orthology criteria but not the evolutionary shift criteria. Several software able to detect evolutionary shift as described above are available e.g. DIVERGE (Gu and Vander Velden, 2002) or PAML (Yang, '97). They should be combined with SIFTER-like approaches in the analysis in the future.

CONCLUSION

Evolutionary-based annotation could have a major impact for the future, and we have described in this article only a part of the concepts of evolutionary biology. Several avenues are possible. The first one is to develop an ontology based on evolutionary biology concept. The second consists in completing the molecular function prediction by Bayesian Phylogenomics model by using evolutionary shift analysis. The third is to include other levels of the ''functional organization ''in the tree with, for e.g., cell, tissues, organ lineage. In future, the next steps will consist in integrating all evolutionary biology concepts in automatic platform dedicated to genome annotation.

ACKNOWLEDGMENTS

Etienne G.J. Danchin is funded by the STREP FungWall European consortium (LSH 2003 1.2.5.5) and Anthony Levasseur, by the ANR05 BIO-006 ARD.

LITERATURE CITED

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29.

Balandraud N, Gouret P, Danchin EG, Blanc M, Zinn D, Roudier J, Pontarotti P. 2005. A rigorous method for multigenic families' functional annotation: the peptidyl arginine deiminase (PADs) proteins family example. BMC Genomics 6:153.

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. 2000. The Pfam protein families database. Nucleic Acids Res 28:263–266.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. 2004. The Pfam protein families database. Nucleic Acids Res 32:D138–141.

Birdsey GM, Lewin J, Cunningham AA, Bruford MW, Danpure CJ. 2004. Differential enzyme targeting as an evolutionary adaptation to herbivory in carnivora. Mol Biol Evol 21:632–646.

Bos DH. 2005. Natural selection during functional divergence to LMP7 and proteasome subunit X (PSMB5) following gene duplication. J Mol Evol 60:221–228.

Collette Y, Gilles A, Pontarotti P, Olive D. 2003. A co-evolution perspective of the TNFSF and TNFRSF families in the immune system. Trends Immunol 24:387–394.

Coulier F, Pontarotti P, Roubin R, Hartung H, Goldfarb M, Birnbaum D. 1997. Of worms and men: an evolutionary perspective on the fibroblast growth factor (FGF) and FGF receptor families. J Mol Evol 44:43–56.

Danchin E, Vitiello V, Vienne A, Richard O, Gouret P, McDermott MF, Pontarotti P. 2004. The major histocompatibility complex origin. Immunol Rev 198:216–232.

Danchin EG, Gouret P, Pontarotti P. 2006. Eleven ancestral gene families lost in mammals and vertebrates while otherwise universally conserved in animals. BMC Evol Biol 6:5.

Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perriere G. 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics 21: 2596–2603.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol 17: 68–74.

Edgar RC. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Edgar RC. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.

Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. 2005. Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol 1:e45.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531–1545.

Ganfornina MD, Sanchez D. 1999. Generation of evolutionary novelty by functional shift. Bioessays 21:432–439.

Gaucher EA, Gu X, Miyamoto MM, Benner SA. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci 27:315–321.

Gouret P, Vitiello V, Balandraud N, Gilles A, Pontarotti P, Danchin EGJ. 2005. FIGENIX: Intelligent automation of genomic annotation: expertise integration in a new software platform. BMC Bioinformatics 6:198.

Gu X, Vander Velden K. 2002. DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics 18:500–501.

Hastings KE. 1996. Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. J Mol Evol 42:631–640.

Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol 29:170–179.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol 22:1345–1354.

Mathe C, Sagot MF, Schiex T, Rouze P. 2002. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res 30:4103–4117.

Miyata T, Yasunaga T. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol 16:23–36.

Remm M, Storm CE, Sonnhammer EL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314:1041–1052.

Rodriguez-Trelles F, Tarrio R, Ayala FJ. 2003. Convergent neofunctionalization by positive Darwinian selection after ancient recurrent duplications of the xanthine dehydrogenase gene. Proc Natl Acad Sci USA 100:13413–13417.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504.

Schmidt TR, Doan JW, Goodman M, Grossman LI. 2003. Retention of a duplicate gene through changes in subcellular targeting: an electron transport protein homologue localizes to the golgi. J Mol Evol 57:222–228.

Sjölander K. 2004. Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics 20:170–179.

Storm CE, Sonnhammer EL. 2002. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics 18:92–99.

Thompson JD, Koehl P, Ripp R, Poch O. 2005. BALIBASE 3.0-latest developments of the multiple alignment benchmark. Proteins 61:127–136.

True JR, Carroll SB. 2002. Gene co-option in physiological and morphological evolution. Annu Rev Cell Dev Biol 18: 53–80.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–556.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496–503.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19:908–917.

Zhang L, Li WH. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol 21:236–239.

Zmasek CM, Eddy SR. 2002. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics 3:14.