Enzyme-specific profiles for genome annotation: PRIAM.

C. Claudel-Renard, Claude C. Chevalet, Thomas Faraut, D. Kahn

Nucleic Acids Research, 2003, Vol. 31, No. 22 ***6633–6639***
DOI: 10.1093/nar/gkg847

Enzyme-specific profiles for genome annotation: PRIAM

Clotilde Claudel-Renard, Claude Chevalet, Thomas Faraut and Daniel Kahn[1],*

Laboratoire de Génétique Cellulaire, INRA and [1]Laboratoire des Interactions Plantes Micro-organismes, INRA/CNRS, BP27, 31326 Castanet-Tolosan Cedex, France

Received June 24, 2003; Revised August 26 2003; Accepted September 23, 2003

ABSTRACT

The advent of fully sequenced genomes opens the ground for the reconstruction of metabolic pathways on the basis of the identification of enzyme-coding genes. Here we describe PRIAM, a method for automated enzyme detection in a fully sequenced genome, based on the classification of enzymes in the ENZYME database. PRIAM relies on sets of position-specific scoring matrices ('profiles') automatically tailored for each ENZYME entry. Automatically generated logical rules define which of these profiles is required in order to infer the presence of the corresponding enzyme in an organism. As an example, PRIAM was applied to identify potential metabolic pathways from the complete genome of the nitrogen-fixing bacterium *Sinorhizobium meliloti*. The results of this automated method were compared with the original genome annotation and visualised on KEGG graphs in order to facilitate the interpretation of metabolic pathways and to highlight potentially missing enzymes.

INTRODUCTION

With the availability of complete genome sequences, the possibility has arisen to aim at a comprehensive inventory of the biochemical functions potentially performed by an organism, thus shedding light on its metabolism (1). The method generally used involves similarity searches in primary sequence databases, which in favourable cases allows the inference of biochemical function on the basis of homology (2). This is, however, fraught with difficulties because: (i) homologous enzymes may have evolved different activities, particularly if paralogous, i.e. if they have diverged as the result of gene duplication (3); (ii) it can be difficult to determine whether two genes are orthologous, i.e. if they have diverged as the result of speciation (3); and (iii) on occasion even orthologous genes may code for enzymes with different specificity. There is therefore a need for better enzyme descriptors that would account for existing enzymes and discriminate against different related enzymes. Such improved descriptors could in turn be used for the systematic identification of enzyme-coding genes from genome sequences.

Specialised enzyme databases, such as ENZYME (4) or BRENDA (5), report enzyme-specific information as well as lists of polypeptides involved in every reported enzyme activity. Enzymes are organised according to the EC (Enzyme Commission) classification which has been developed and maintained by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (6). Proteins listed under the same enzymatic function, i.e. with the same EC number, may sometimes display a variety of sequences that must be taken into account. Once enzyme-coding genes are detected, they can be reported in pathway-oriented databases such as WIT (7), ECOCYC (8), METACYC (9), KEGG (10) or UMBBD (11). In the present work, we have automatically developed descriptors for all entries in the ENZYME database and show how their combined use can help genome annotators to infer metabolic pathways potentially active in an organism.

METHODS

Let us first describe the general outline of the PRIAM methodology. First, enzyme-specific sequence collections are extracted from the ENZYME database (4). Secondly, the modules characteristic of each collection are automatically detected using the MKDOM program (12). Thirdly, we generate enzyme-specific rules defining which module(s) are required in order to infer the presence of a given enzyme. Fourthly, each of these modules is described using a specific position-specific scoring matrix, or 'profile'. Finally, we show how the PRIAM program allows enzyme activity to be inferred using these profiles together with enzyme-specific rules on a complete genome. The resulting predictions are mapped onto KEGG metabolic charts (10) for easy inspection and end-user interpretation. In order to demonstrate the utility of this tool, PRIAM was tested on the recently sequenced *Sinorhizobium meliloti* genome (http://sequence.toulouse.inra.fr/S.meliloti).

Definition of enzyme-specific sequence collections

We define an enzyme-specific sequence collection as the set of all protein sequences participating in a given enzyme activity and thus sharing the same EC number. We extracted these enzyme collections from the ENZYME database (release 27.0). For most EC numbers, ENZYME provides a list of

*To whom correspondence should be addressed Tel: +33 561285329; Fax: +33 561285061; Email: dkahn@toulouse.inra.fr relevant protein sequences, generating 1606 sequence collections. The number of sequences in each collection ranges from one to hundreds of sequences. For example, the alcohol dehydrogenase (EC 1.1.1.1) collection contains 165 sequences. We wish to emphasise here that all sequences belonging to a collection do not necessarily share similarity because of the occurrence of (i) oligomeric enzymes; (ii) non-homologous enzymes catalysing the same reaction; and (iii) multifunctional enzymes. Therefore, sequence relationships may be complex when comparing the various sequences involved in a particular enzyme activity.

Detection of homologous modules in enzyme collections

These potentially complex sequence relationships are best captured by detecting which modules are required in order to adequately describe the whole sequence set. In the framework of this paper, modules are defined as the longest homologous segments shared within an enzyme collection. The MKDOM2 program (12) allows for such modules to be identified by an exhaustive PSI-BLAST search (13) within the sequence set. We used default MKDOM2 parameters, except for the E-value that was changed to 10^{-4} in order to group more distantly related sequences. The MKDOM2 program is based upon the hypothesis that the smallest sequence in the database corresponds to a single module. In the first step, the shortest sequence in the collection is used as the first PSI-BLAST query, thus generating the first module family; in a second step, the corresponding modules are deleted from the sequence set. The shortest remaining sequence is used as a PSI-BLAST query and the process is iterated until all sequences are decomposed into modules that are grouped into families. Because enzyme collections are usually relatively small sequence sets, the resulting modules are not necessarily decomposed into individual domains: they tend to be longer than domains as defined in ProDom (14), Pfam (15) or SCOP (16).

Module selection and logical rules for enzyme inference

All modules are not necessarily relevant to characterise the whole collection, some of them being found only in a limited subset. We therefore selected the longest modules involving the largest number of sequences, in such a way that all sequences of the collection are represented. We thus obtained 2435 module families characterising 1606 enzyme collections. Most enzyme collections are efficiently represented by a single module type (Fig. 1). This corresponds to the frequent situation where all sequences from an enzyme collection share one typical homologous region. However, in some instances, more than one module is required. This happens particularly with oligomeric enzymes for which different descriptors are required for different subunits. In this case, the relevant subunits [possibly fused, Marcotte *et al.* (17)] should be detected simultaneously in order to infer the presence of a given enzyme: an 'AND' rule is applied. Another situation arises in the case of non-homologous enzymes catalysing the same reaction, so that different descriptors are required for each different enzyme type. In the simple case where each enzyme type is characterised by only one descriptor, any one of them is sufficient for enzyme inference ('OR' rule). More generally, both AND and OR operators will be required when non-homologous oligomeric enzymes are found to catalyse the same reaction (a rather rare occurrence in the ENZYME database). In order to automatically derive the relevant logical rule *P*, we test which modules are required in order to account for the presence of a given enzyme in a set of organisms (see algorithm in Appendix A). The 'AND' rule is applied whenever two modules occur systematically together in the same organisms. An example of an enzyme collection requiring an 'AND' rule is shown in Figure 2 for homocitrate synthase (EC 4.1.3.21). The black and red striped modules were selected as representative of all homocitrate synthase sequences: both modules are required, either in one or in two polypeptide chain(s).

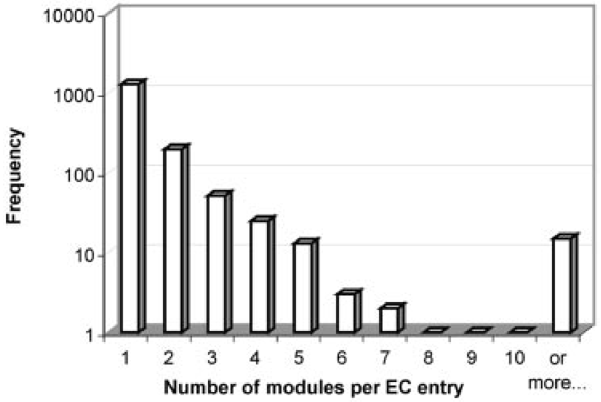

Figure 1. Distribution of the number of modules selected for each enzyme collection in PRIAM (logarithmic scale).

Construction of position-specific scoring matrices (profiles)

For each selected family, we generated position-specific scoring matrices (profiles) using PSI-BLAST 2.0.11 (13) run against all homologous module sequences. This is possible even in the particular case when only one module sequence is available, which occurred for 470 modules out of 2435 families. Indeed one of the important issues in deriving a profile is to combine the prior knowledge of residue relationships, as embodied in the usual substitution matrices, with the information provided by the multiple alignment. When few observations are available, a greater emphasis should be given to the prior knowledge of residue relationships. The solution adopted and implemented by the PSI-BLAST program is based on a pseudo-count method. As noted by the authors, in the extreme case where only one sequence is available, the scoring scheme reduces to the usual substitution matrix [see Altschul *et al.* (13) for a detailed description]. It is therefore technically possible and theoretically founded to derive a profile even in this special case, which allows all modules to be treated with the same profile methodology. These profiles can be used either as new PSI-BLAST queries or as targets for RPS-BLAST homology searches (18).

Complete genome analysis with PRIAM

The profiles described above can be used to systematically search for the presence of enzymes in a genome. We have designed the PRIAM program (profils pour l'identification automatisée du métabolisme) to automatically identify

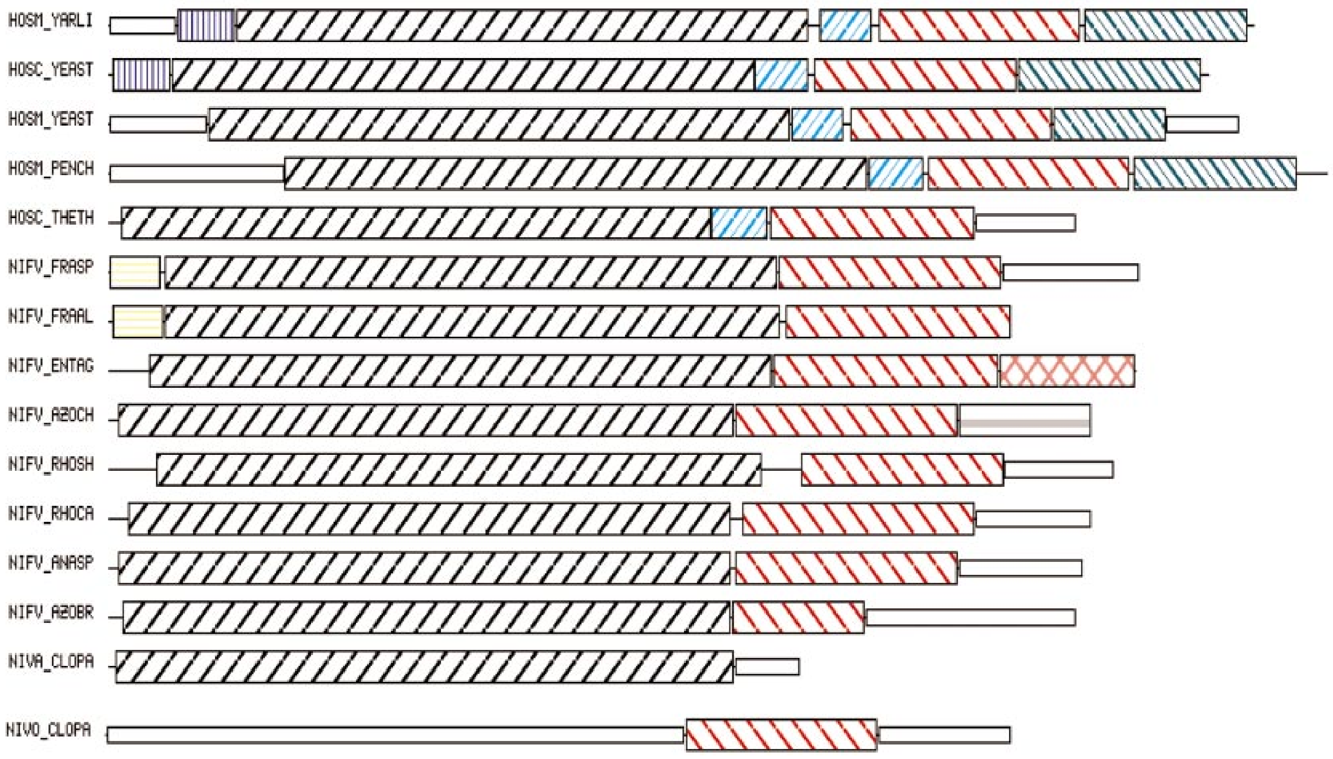

Figure 2. Example of modular enzymes implying an 'AND' rule. Modules detected for homocitrate synthase (EC 4.1.3.21) are displayed using the XDOM program (35). Both black and red striped modules are required in order to infer the presence of homocitrate synthase.

enzymes encoded in a genome. The first step is a homology search between each protein and PRIAM profiles using the RPS-BLAST program (18), filtered for E-values below 10^{-10}. In a second step, the best non-overlapping matches are reported for each protein in order to detect potential multi-enzymes (see Appendix B). The processing of all proteins encoded in a genome thus generates a list of PRIAM matches. In a third step, we test whether the enzyme-specific rule *P* defined above is fulfilled for each entry in the ENZYME database. Finally, this generates a list of predicted enzymes which can be mapped onto metabolic charts, such as found in the KEGG database (10).

Distribution

PRIAM is available from http://genopole.toulouse.inra.fr/bioinfo/priam/.

RESULTS AND DISCUSSION

Combination of profiles characterising each ENZYME entry

Following the methodology described above, 2435 profiles were selected for 1606 entries from the ENZYME database. Most ENZYME entries (1296 enzymes) are simply characterised by a single profile (Fig. 1). Among the 310 enzymes requiring more than one profile, 38 enzymes follow an 'AND' rule imposing the simultaneous presence of two or more modules for enzyme inference. The only enzyme with a rule implying both 'AND' and 'OR' rules corresponds to L-serine dehydratase (EC 4.3.1.17), with two profiles for both subunits of the prokaryotic enzyme and one profile for the eukaryotic enzyme type. PRIAM thus provides a comprehensive 'profile' view of the ENZYME database that, together with enzyme-specific rules, provides a systematic basis for enzyme inference.

Application of PRIAM to complete genomes

The PRIAM methodology was tested on five complete microbial genomes which were exhaustively annotated in the SWISS-PROT database (19): *Buchnera aphidicola* (20), *Escherichia coli* (21), *Haemophilus influenzae* (22), *Mycoplasma genitalium* (23) and *M.pneumoniae* (24) (as listed in http://www.expasy.org/sprot/hamap/bacteria.html). These genomes were selected as the best 'standards of truth' currently available because they have been expertly annotated using both state-of-the-art methodology and current biological knowledge in the framework of the HAMAP project (19). For each genome, we counted how many enzyme activities predicted by PRIAM were reported in SWISS-PROT (true positive, TP) or not reported (probable false positive, FP), and how many activities reported in SWISS-PROT were missed by PRIAM (false negative, FN). For each genome, the specificity TP/(TP + FP) and sensitivity TP/(TP + FN) were calculated for PRIAM at various RPS-BLAST E-value thresholds. An excellent average specificity and sensitivity of 93% could be reached for an E-value of 10^{-30} (Fig. 3). At this threshold, PRIAM compares favourably with KEGG orthology assignments which rely on both human inspection and automated screening with the GFIT program (10,25) (Table 1). Jacknife tests were performed by recalculating all

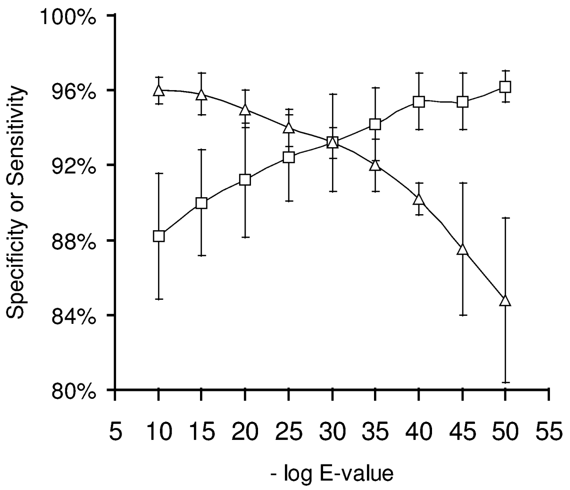

Figure 3. Calibration of PRIAM methodology on complete genomes. The specificity and sensitivity of enzyme predictions were calculated using as standards the enzyme sets from the complete genome annotation found in SWISS-PROT for *B.aphidicola, E.coli, H.influenzae, M.genitalium* and *M.pneumoniae* (19). The mean and standard deviation of PRIAM specificity (squares) and sensitivity (triangles) were calculated for different RPS-BLAST E-values.

PRIAM profiles, excluding sequences from one genome at a time, and testing both specificity and sensitivity on the same genome (Table 1). The resulting average specificity and sensitivity were 86 and 89%, respectively, to be compared with values of 89 and 91% for KEGG semi-automated orthology assignments (calculated similarly against SWISS-PROT annotation). Note that these tests were performed using PRIAM in a fully automated mode, whereas KEGG assignments also rely on some human expertise. We expect that PRIAM profiles will become even more reliable as more enzyme sequences are introduced into the ENZYME database.

The PRIAM methodology has been used for the functional annotation of several complete genomes including those of *S.meliloti* and *Ralstonia solanacearum* (26), and for the interpretation of *S.meliloti* proteome analysis (27). *Sinorhizobium meliloti* is a nitrogen-fixing bacterium able to establish a symbiotic relationship with alfalfa. It is of interest to analyse *S.meliloti* metabolism as it is intimately coupled to plant host metabolism during symbiosis. We applied PRIAM to the complete set of *S.meliloti* proteins (28). PRIAM processing took 1 h 40 min with a 699 MHz Pentium III personal computer under Linux. At an E-value threshold of 10^{-10}, PRIAM predicted enzyme activity for 1460 out of 6204 proteins, among which 13 proteins were predicted as bifunctional multienzymes. These 1460 predicted proteins correspond to 660 different enzyme activities, emphasizing the extent of paralogy in the *S.meliloti* genome as observed earlier (28) (see http://genopole.toulouse.inra.fr/bioinfo/priam/).

Comparison with manual annotation of the *S.meliloti* genome

PRIAM results were compared with manual annotation which initially identified 532 enzyme activities encoded by 808 genes in the *S.meliloti* genome (28–31). The vast majority of these were also predicted by PRIAM, since only 39 enzyme activities were missed. Among these, seven activities could not be predicted because no corresponding sequence was available in the ENZYME database. Other activities were missed because of a different substrate specificity. As an example, let us focus on glutamate metabolism. The SMa0680 and SMa0682 proteins were manually annotated as arginine decarboxylase (EC 4.1.1.19), while PRIAM rather suggests ornithine decarboxylase activity (EC 4.1.1.17). Another discrepancy was observed for 1-pyrroline-5-carboxylate dehydrogenase (EC 1.5.1.12) that was not found by PRIAM, while manual annotation identified a candidate dehydrogenase (SMc02181) with two possible substrates: 1-pyrroline-5-carboxylate (1.5.1.12) and proline (1.5.99.8). In this case, PRIAM delivers only the best matching activity (EC 1.5.99.8).

PRIAM also predicted 167 additional enzyme activities that were not proposed during manual annotation. In many cases, PRIAM suggested a more precise EC for proteins which were annotated with truncated EC numbers such as 1.1.–.–.. These annotations may therefore require re-evaluation. As an example of a new prediction made by PRIAM, a candidate was found for glutamate decarboxylase (EC 4.1.1.15), the second step of the γ-aminobutyrate shunt. This similarity originates from a profile normally matching mammalian sequences (E-value = 2×10^{-52}). The corresponding protein RhbB was previously annotated as L-2,4-diaminobutyrate decarboxylase and proposed to be involved in rhizobactin siderophore synthesis (32). Another candidate, SMb21414, could also be found using a PRIAM profile for glutamate decarboxylase, although this protein matches better with EC

Table 1. Specificity and sensitivity of PRIAM-based enzyme detection in five complete genomes, using SWISS-PROT annotation as a standard

Genome	PRIAM		PRIAM jacknife		KEGG orthology	
	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
B.aphidicola	97%	93%	86%	91%	87%	80%
E.coli	93%	93%	92%	88%	89%	91%
H.influenzae	94%	94%	84%	91%	88%	93%
M.genitalium	92%	92%	86%	87%	93%	95%
M.pneumoniae	90%	94%	85%	87%	91%	95%

The RPS-BLAST E-value was set at 10^{-30}. Jacknife analysis was performed with PRIAM profiles in which sequences from the corresponding genome were omitted. Specificity and sensitivity of KEGG orthology assignments (retrieved from http://www.genome.ad.jp/kegg/kegg2.html; 10,25) were calculated similarly against SWISS-PROT for comparison.

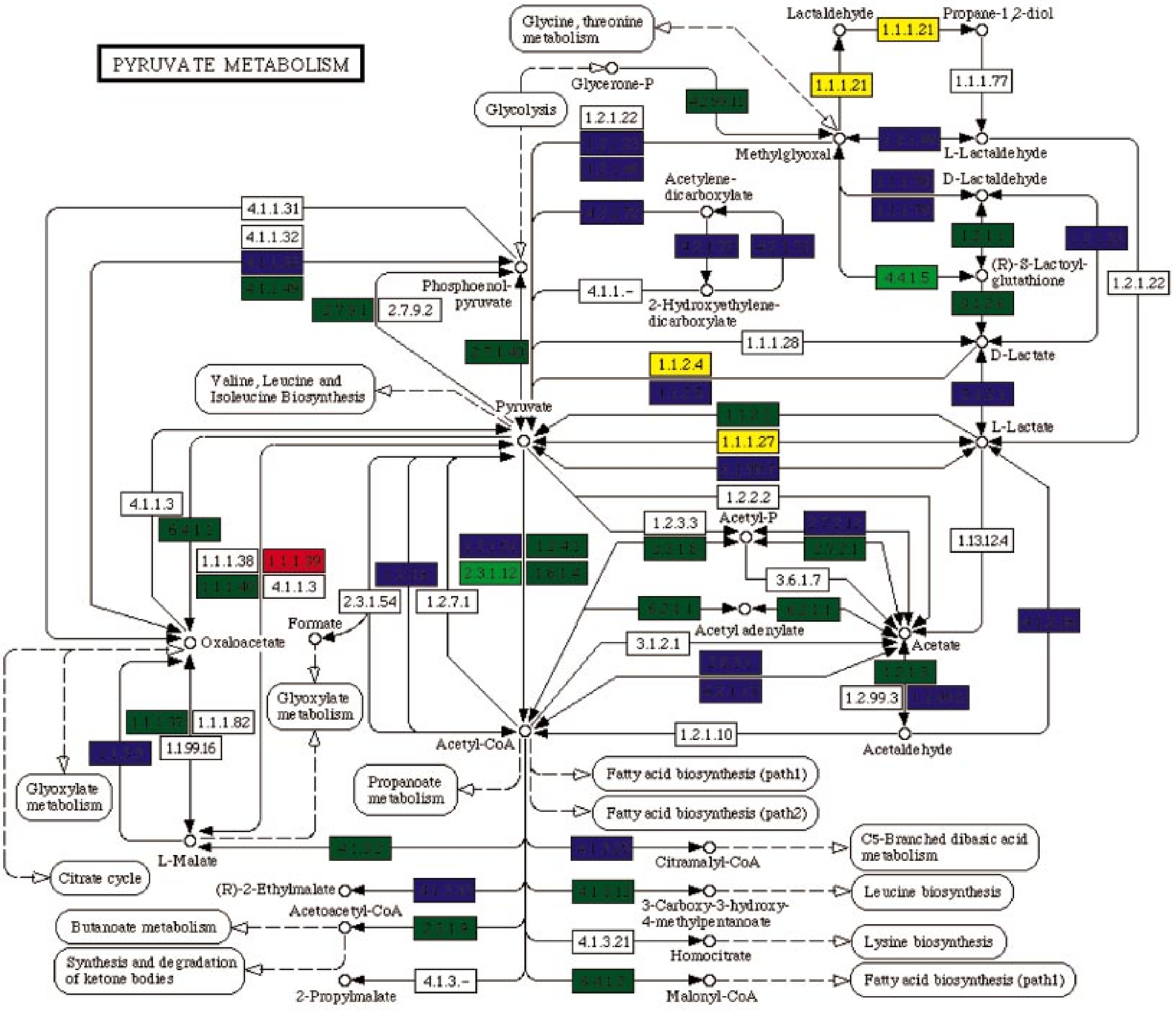

Figure 4. Example of a colour-coded KEGG metabolic chart used for interpreting *S.meliloti* metabolism (http://genopole.toulouse.inra.fr/bioinfo/priam/). Green boxes correspond to EC numbers found by both annotation and PRIAM (dark or pale green: E-value below or above 10^{-30}, respectively). Orange and yellow boxes correspond to enzymes not annotated, yet detected by PRIAM with an E-value below or above 10^{-30}, respectively. Red boxes indicate annotated enzymes not detected by PRIAM. Blue boxes correspond to EC numbers for which no sequence information was available in the ENZYME database, hence for which no PRIAM profile could be built. Yellow, orange and red boxes indicate a discrepancy between PRIAM analysis and current genome annotation, suggesting metabolic steps that should be reinvestigated.

4.1.1.28 (aromatic amino acid decarboxylase). PRIAM has thus pointed to candidate genes for an enzymatic activity which has indeed been found experimentally in *Rhizobium* bacteroids (33). Furthermore, PRIAM also suggests a precise annotation for a few proteins which were previously overlooked. For example, in the pentose phosphate pathway, PRIAM suggests a candidate phosphoketolase (SMc04146) that was missed previously (EC 4.1.2.9 or 4.1.2.22).

PRIAM-based pathway analysis

The results of PRIAM analysis can be mapped onto metabolic charts, for example from the KEGG database (10). As an example, let us consider pyruvate metabolism in *S.meliloti*: Figure 4 displays the enzymatic steps which were predicted with PRIAM, proposed during genome annotation (28) or both. In this pathway, one enzyme is missed by PRIAM analysis: the NAD-dependent malic enzyme (EC 1.1.1.39), indicated in red on Figure 4. This activity is known to be carried out by the *dme* gene product in *S.meliloti* (34), which PRIAM predicts as an NADP-dependent malic enzyme (EC 1.1.1.40). As a matter of fact, the profiles generated for both enzymes are very similar, making it difficult to discriminate between the NAD- and NADP-dependent malic enzymes. Moreover, the *dme* gene product also possesses NADP-dependent activity (34). This example points out two limitations inherent to sequence-based enzyme prediction. The first limitation results from the existence of closely related enzymes with different substrate specificity. The second limitation is linked to the relaxed substrate specificity exhibited by some enzymes. The PRIAM methodology is unable to predict such relaxed specificity, as it selects only the best matching profile for functional inference. It is therefore useful to confront the results of PRIAM analysis with available annotation, as can be readily visualised on metabolic charts.

ACKNOWLEDGEMENTS

We are grateful to Professor Minoru Kanehisa for providing the KEGG database and metabolic charts, and to our colleague Florence Servant for helpful comments. This work was supported in part by the French Ministry of Education and by the Toulouse-Midi Pyrénées Genopole.

REFERENCES

1. Bork,P., Ouzounis,C., Casari,G., Schneider,R., Sander,C., Dolan,M., Gilbert,W. and Gillevet,P.M. (1995) Exploring the *Mycoplasma capricolum* genome: a minimal cell reveals its physiology. *Mol. Microbiol.*, **16**, 955–967.
2. Tatusov,R.L., Mushegian,A.R., Bork,P., Brown,N.P., Hayes,W.S., Borodovsky,M., Rudd,K.E. and Koonin,E.V. (1996) Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.*, **6**, 279–291.
3. Fitch,W.M. (2000) Homology a personal view on some of the problems. *Trends Genet.*, **16**, 227–231.
4. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305
5. Schomburg,I., Chang,A. and Schomburg,D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.*, **30**, 47–49.
6. Nomenclature Committee. (1992) *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. Academic Press, San Diego, CA.
7. Overbeek,R., Larsen,N., Pusch,G.D., D'Souza,M., Selkov,E., Jr., Kyrpides,N., Fonstein,M., Maltsev,N. and Selkov,E. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, **28**, 123–125.
8. Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The EcoCyc database. *Nucleic Acids Res.*, **30**, 56–58.
9. Karp,P.D., Riley,M., Paley,S.M. and Pellegrini-Toole,A. (2002) The MetaCyc database. *Nucleic Acids Res.*, **30**, 59–61.
10. Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.
11. Ellis,L.B., Hershberger,C.D., Bryan,E.M. and Wackett,L.P. (2001) The University of Minnesota Biocatalysis/Biodegradation database: emphasizing enzymes. *Nucleic Acids Res.*, **29**, 340–343.
12. Gouzy,J., Corpet,F. and Kahn,D. (1999) Whole genome protein domain analysis using a new method for domain clustering. *Comput. Chem.*, **23**, 333–340.
13. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
14. Servant,F., Bru,C., Carrere,S., Courcelle,E., Gouzy,J., Peyruc,D. and Kahn,D. (2002) ProDom: automated clustering of homologous domains. *Brief Bioinform.*, **3**, 246–251.
15. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
16. LoConte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
17. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
18. Marchler-Bauer,A., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al.* (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.*, **31**, 383–387.
19. Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J., Lachaize,C. *et al.* (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.*, **27**, 49–58.
20. Shigenobu,S., Watanabe,H., Hattori,M., Sakaki,Y. and Ishikawa,H. (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*, **407**, 81–86.
21. Blattner,F.R., Plunkett,G.,3rd, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
22. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
23. Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.
24. Himmelreich,R., Hilbert,H., Plagens,H., Pirkl,E., Li,B.C. and Herrmann,R. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.*, **24**, 4420–4449.
25. Bono,H., Ogata,H., Goto,S. and Kanehisa,M. (1998) Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. *Genome Res.*, **8**, 203–210.
26. Salanoubat,M., Genin,S., Artiguenave,F., Gouzy,J., Mangenot,S., Arlat,M., Billault,A., Brottier,P., Camus,J.C., Cattolico,L. *et al.* (2002) Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature*, **415**, 497–502.
27. Djordjevic,M.A., Chen,H.C., Natera,S., Van Noorden,G., Menzel,C., Taylor,S., Renard,C., Geiger,O. and Weiller,G.F. (2003) A global analysis of protein expression profiles in *Sinorhizobium meliloti*: discovery of new genes for nodule occupancy and stress adaptation. *Mol. Plant–Microbe Interact.*, **16**, 508–524.
28. Galibert,F., Finan,T.M., Long,S.R., Puhler,A., Abola,P., Ampe,F., Barloy-Hubler,F., Barnett,M.J., Becker,A., Boistard,P. *et al.* (2001) The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science*, **293**, 668–672.
29. Barnett,M.J., Fisher,R.F., Jones,T., Komp,C., Abola,A.P., Barloy-Hubler,F., Bowser,L., Capela,D., Galibert,F., Gouzy,J. *et al.* (2001) Nucleotide sequence and predicted functions of the entire *Sinorhizobium meliloti* pSymA megaplasmid. *Proc. Natl Acad. Sci. USA*, **98**, 9883–9888.
30. Capela,D., Barloy-Hubler,F., Gouzy,J., Bothe,G., Ampe,F., Batut,J., Boistard,P., Becker,A., Boutry,M., Cadieu,E. *et al.* (2001) Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021. *Proc. Natl Acad. Sci. USA*, **98**, 9877–9882.
31. Finan,T.M., Weidner,S., Wong,K., Buhrmester,J., Chain,P., Vorholter,F.J., Hernandez-Lucas,I., Becker,A., Cowie,A., Gouzy,J. *et al.* (2001) The complete sequence of the 1,683-kb pSymB megaplasmid from the N_2-fixing endosymbiont *Sinorhizobium meliloti*. *Proc. Natl Acad. Sci. USA*, **98**, 9889–9894.
32. Lynch,D., O'Brien,J., Welch,T., Clarke,P., Cuiv,P.O., Crosa,J.H. and O'Connell,M. (2001) Genetic organization of the region encoding regulation, biosynthesis and transport of rhizobactin 1021, a siderophore produced by *Sinorhizobium meliloti*. *J. Bacteriol.*, **183**, 2576–2585.
33. Miller,R.W., McRae,D.G. and Joy,K. (1991) Glutamate and gamma-aminobutyrate metabolism in isolated *Rhizobium* bacteroids. *Mol. Plant–Microbe Interact.*, **4**, 37–45.
34. Voegele,R.T., Mitsch,M.J. and Finan,T.M. (1999) Characterization of two members of a novel malic enzyme class. *Biochim. Biophys. Acta*, **1432**, 275–285.
35. Gouzy,J., Eugene,P., Greene,E.A., Kahn,D. and Corpet,F. (1997) XDOM, a graphical tool to analyse domain arrangements in any set of protein sequences. *Comput. Appl. Biosci.*, **13**, 601–608.

APPENDIX A

Pseudo-code generating the logical rule for an enzyme collection

Let n be the total number of modules for an enzyme collection. We apply the following algorithm to select which module is used in the logical rule P:

```
// Initialisation
for i = 1…n {
   let $O_i$ be the list of organisms represented by module $i$;
   let $P_i$ define the logical rule matching module $i$
   }
sort $O_i$ by decreasing $|O_i|$
set $P$ = FALSE, $O = O_1$;
// Rule generation
for i = 2…n {
   if $O_i == O_{i-1}$ set $P_i = P_i$ AND $P_{i-1}$ ;
   // revise $P_i$ when verified in all species $\in O_{i-1}$
   else if $O_i \notin O$ set $P = P$ OR $P_{i-1}$ ;
   // revise $P$ only when $P_i$ is verified in new species $\notin O$
   else set $P_i = P_{i-1}$ ;
   // ignore $P_i$ if verified only in subset of $O$
   $O = O \cup O_i$
   }
set $P = P$ OR $P_n$
```

APPENDIX B

Pseudo-code for multienzyme detection, applied to each protein *p*

```
let $n$ be the total number of PRIAM matches for protein $p$;
let $M_i$ be the matches sorted by increasing E-value ( $i = 1…n$);
set $M = M_1$ ;
for i = 2…n {
   if $M$ and $M_i$ overlap on $p$ by less than 20 amino acids
   set $M = M \cup M_i$;
   // $M_i$ corresponds to a new region on query protein $p$
   }
```