BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 20 2012, pages 2687–2688
doi:10.1093/bioinformatics/bts510

Sequence analysis
Advance Access publication August 24, 2012

EFICAz$^{2.5}$: application of a high-precision enzyme function predictor to 396 proteomes

Narendra Kumar and Jeffrey Skolnick*
Center for Study of Systems Biology, School of Biology, Georgia Institute of Technology, Atlanta, GA 30318, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: High-quality enzyme function annotation is essential for understanding the biochemistry, metabolism and disease processes of organisms. Previously, we developed a multi-component high-precision enzyme function predictor, EFICAz^2 (enzyme function inference by a combined approach). Here, we present an updated improved version, EFICAz$^{2.5}$, that is trained on a significantly larger data set of enzyme sequences and PROSITE patterns. We also present the results of the application of EFICAz$^{2.5}$ to the enzyme reannotation of 396 genomes cataloged in the ENSEMBL database.

Availability: The EFICAz$^{2.5}$ server and database is freely available with a use-friendly interface at http://cssb.biology.gatech.edu/EFICAz2.5.

Contact: skolnick@gatech.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 5, 2012; revised on July 22, 2012; accepted on August 10, 2012

1 INTRODUCTION

Understanding cellular functions such as metabolic pathways requires the precise assignment of enzymatic functions to proteins in their respective genomes. The ever-increasing accumulation of sequence data that is the result of the genome-sequencing projects requires rapid computational methods to assign enzyme functions (Friedberg, 2006). Enzymes are classified by the Enzyme Commission (EC) using a hierarchical system of four-digit codes, such as EC 1.2.3.4, for each type of enzymatic activity (Webb, 1992). High-precision EC number assignment is of utmost importance for studies such as metabolic pathway reconstruction, understanding evolutionary relationships in pathways (Ginsburg, 2009) and metabolite prediction, etc. (Arakaki *et al.*, 2008).

Previously, we developed an automatic engine for the proteomic scale inference of enzyme function, EFICAz (enzyme function inference by a combined approach) (Tian *et al.*, 2004). EFICAz combines the predictions from four independent methods, namely: (i) CHIEFc family-based Functionally Discriminating Residue (FDR) identification, (ii) multiple PFAM-based FDR recognition, (iii) CHIEFc SIT evaluation and (iv) high-specificity multiple PROSITE patterns. The main strength of EFICAz is its ability to identify FDRs as residues that discriminate the members of a homo-functional family (consisting of members with a specific EC number) from a hetero-functional family (sequences with another EC number). Subsequently, we enhanced our approach to EFICAz^2 (Arakaki *et al.*, 2009) with the inclusion of support vector machine (SVM) models for discrimination between homo-functional and hetero-functional EC family members. This led to the inclusion of two additional components: (v) CHIEFc family-based SVM evaluation and (vi) multiple PFAM family-based SVM evaluation.

Recent years have seen a huge influx of sequence data as a result of automated genome-sequencing projects and subsequent assignment of new sequences by the EC. CatFam (Yu *et al.*, 2009), PRIAM (Claudel-Renard *et al.*, 2003) and Enzymedetector (Quester and Schomburg, 2011) are some of the programs capable of genome-level enzyme prediction. However, all of them are sequence similarity-based methods, whereas EFICAz also uses PROSITE, PFAM, followed by detection of FDRs and is optimized for high precision. Here, we present the latest implementation of our enzyme function prediction engine, EFICAz$^{2.5}$, using the latest, significantly larger, data set of sequences from SWISSPROT and sequence patterns from PROSITE (Sigrist *et al.*, 2010). We use enzyme annotations of the SWISSPROT (O'Donovan *et al.*, 2002) component of UNIPROT sequences (Magrane and Consortium, 2011) for training EFICAz$^{2.5}$. Subsequently, we applied EFICAz$^{2.5}$ to annotate 396 fully sequenced genomes in the ENSEMBL (Flicek *et al.*, 2011) database belonging to bacteria, fungi, protists, metazoans, plants and vertebrates. The results are cataloged and are available at http://cssb.biology.gatech.edu/enzymes.

2 METHODS

Annotations in the public databases can be error prone, especially those that include automated annotations. Accordingly, we used sequences in the ENZYME data set of SWISSPROT (O'Donovan *et al.*, 2002) release 2011_01 (January 2011) that are manually curated, and thus, it is reasonable to use their EC-recommended EC numbers for training. This release consisted of 220 485 sequences belonging to 2757 four-digit EC (4EC) families and 211 three-digit EC (3EC) families. Sequences were divided into enzyme and non-enzyme sets. The enzyme set was classified into EC families at the 3EC and 4EC digit levels. To train EFICAz$^{2.5}$, we followed the protocol described in detail by Tian *et al.*, 2004 and Arakaki *et al.*, 2009. CHIEFc families were constructed using an iterative procedure, which aims to divide the sequences in each EC group into evolutionarily related families based on the conservation of functionally discriminating residues. The PFAM database (Punta *et al.*, 2012) release 23 was used for training the components 'Multiple PFAM based FDR recognition' and 'multiple PFAM based SVM evaluation'. We used release 20.68 (January 2011) of the PROSITE database (Sigrist *et al.*, 2010) for training the 'High specificity multiple PROSITE pattern', which detects sequence

*To whom correspondence should be addressed.

patterns in the enzyme subset of sequences. All components are trained separately for sequences annotated up to 3EC and 4EC numbers. The classification tree as reported in Arakaki *et al.*, 2009 is implemented to integrate predictions generated by each of six components into a final EC number prediction.

For annotation and reannotation of genomes for enzyme functions, genomic sequences were downloaded from the ftp site of the ENSEMBL genomes database (release 13) for 249 bacterial, 26 fungi, 15 protists, 16 plants and 35 metazoans. An additional 55 vertebrate species genomes were downloaded from ENSEMBL release 67. All alternate spliced versions for a gene were considered in the detection of enzyme function.

3 RESULTS AND DISCUSSION

EFICAz$^{2.5}$ is the latest implementation of the high-precision enzyme function prediction method EFICAz that was trained with the latest data sets for the SWISSPROT, PFAM and PROSITE. We implemented changes in the software code that make it easier for a user to upgrade by requiring only the replacement of the directory containing the library files. EFICAz$^{2.5}$ takes about 4 min for a typical 200-residue protein compared with 6 min by EFICAz^{2}. The results on the web interface are more informative and include the EC name, EC number and relevant links to the EC, BRENDA and KEGG resources. The results of user queries performed using SWISSPROT ID are stored at the backend, and subsequent queries for the same ID immediately retrieve the results. This version uses a total number of 220 485 sequences whose EC numbers are provided by the ENZYME data set of curated SWISSPROT sequences as compared with only 136 167 sequences available in 2008 when the previous version, EFICAz^{2}, was built. In terms of the diversity of enzymatic functions, the current release contains a total of 2757 4EC numbers compared with 2354 in the previous version. The diversity of functions is seen only at the 4EC level, as the number of 3EC digit functions only increases from 209 to 211. Benchmarking results when 80% of the sequences from every EC function were used for training while testing on the remaining 20% of sequences (only families with >5 sequences were used) shows a similar trend of high precision and recall values as a function of the maximum test to training sequence identity (MTTSI). At a sequence similarity of >40% for 3 EC level, recall and precision values were 0.88 and 0.85, respectively, and reached almost 1.0 at a MTTSI of >60%. Even at a MTTSI of 20%, the precision and recall were 0.8 and 0.6, respectively, at the 4EC level. However, both drop rapidly as the MTTSI became <20% (see Supplementary Fig. S1). This is not surprising, as this lies well within the twilight zone for function prediction. Supplementary Figure S2 shows a comparison of EFICAz$^{2.5}$ and EFICAz^{2} for all sequences in the latest release in SWISSPROT.

In another comparison analysis, we compared the EC annotations made by EFICAz^{2} and EFICAz$^{2.5}$ for the human proteome. The latest release of the human proteome consists of 2856 sequences annotated with EC numbers in ENZYME database. EFICAz$^{2.5}$ could predict all four digits of 2642 sequences correctly compared with 2288 sequences by EFICAz^{2}. In addition, EFICAz$^{2.5}$ predicted 30 sequences at a low confidence level. At the 3EC level, EFICAz$^{2.5}$ predicted 2758 sequences correctly as compared with 2,413 by EFICAz^{2}.This shows that EFICAz$^{2.5}$ has better prediction ability compared with its predecessor. In addition, EFICAz$^{2.5}$ predicts EC numbers for a total of 8886 of 50 475 sequences in the translated TrEMBL component of the human proteome.

Advancements in sequencing technologies have resulted in the full genome sequencing of a variety of organisms. After sequencing, the next obvious step is the annotation of the functions of genes. Among various roles played by proteins, enzymatic function is the most important from the biochemical point of view. We took this opportunity to reannotate 4 044 586 sequences from a total of 396 genomes from the ENSEMBL database using EFICAz$^{2.5}$. The results will be updated with each release of ENSEMBL and ENSEMBL genomes.

Prediction results are obtained at two levels: viz. ‘high confidence’ and ‘all prediction’ (Tian *et al.*, 2004). Supplementary Table S1 shows a summary of predictions at the 3EC and 4EC digit levels for various groups of organisms. In any case, extant accurate enzyme annotations represent the lower bound of enzymatic functions for an organism. This data can be used by the biochemist and computational biologist alike. We believe EFICAz$^{2.5}$ will be a useful tool for the scientific community.

Funding: This work was supported in part by grant no. GM-48835 of the Division of General Medical Sciences of the National Institutes of Health.

Conflict of Interest: none declared.

REFERENCES

Arakaki,A.K. *et al.* (2009) EFICAz2: enzyme function inference by a combined approach enhanced by machine learning. *BMC Bioinformatics*, **10**, 107.

Arakaki,A.K. *et al.* (2008) Identification of metabolites with anticancer properties by computational metabolomics. *Mol. Cancer*, **7**, 57.

Claudel-Renard,C. *et al.* (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.*, **31**, 6633–6639.

Flicek,P. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.

Friedberg,I. (2006) Automated protein function prediction—the genomic challenge. *Brief. Bioinform.*, **7**, 225–242.

Ginsburg,H. (2009) Caveat emptor: limitations of the automated reconstruction of metabolic pathways in Plasmodium. *Trends Parasitol.*, **25**, 37–43.

Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford), 2011, bar009.

O’Donovan,C. *et al.* (2002) High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinform.*, **3**, 275–284.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Quester,S. and Schomburg,D. (2011) EnzymeDetector: an integrated enzyme function prediction tool and database. *BMC Bioinformatics*, **12**, 376.

Sigrist,C.J. *et al.* (2010) PROSITE, a protein domain database for functional characterization and annotatio. *Nucleic Acids Res.*, **38**, D161–D166.

Tian,W. *et al.* (2004) EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Res.*, **32**, 6226–6239.

Webb,E. (1992) *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification Of Enzymes*. Published for the international union of Biochemistry and Molecular Biology by Academic Press, San Diego.

Yu,C. *et al.* (2009) Genome-wide enzyme annotation with precision control: catalytic families (CatFam) databases. *Proteins*, **74**, 449–460.