BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 14 2010, pages 1690–1698
doi:10.1093/bioinformatics/btq266

Genome analysis

Advance Access publication May 30, 2010

DETECT—a Density Estimation Tool for Enzyme ClassificaTion and its application to *Plasmodium falciparum*

Stacy S. Hung[1,2], James Wasmuth[1], Christopher Sanford[1,2] and John Parkinson[1,2,3,*]

[1]Program in Molecular Structure and Function, Hospital for Sick Children, 15-704 MaRS TMDT East, 101 College Street, Toronto, ON, M5G 1L7, [2]Department of Molecular Genetics and [3]Department of Biochemistry, University of Toronto, Toronto, 1 Kings' College Circle, ON, M5S 1A8, Canada

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: A major challenge in genomics is the accurate annotation of component genes. Enzymes are typically predicted using homology-based search methods, where the membership of a protein to an enzyme family is based on single-sequence comparisons. As such, these methods are often error-prone and lack useful measures of reliability for the prediction.

Results: Here, we present DETECT, a probabilistic method for enzyme prediction that accounts for the sequence diversity across enzyme families. By comparing the global alignment scores of an unknown protein to those of all known enzymes, an integrated likelihood score can be readily calculated, ranking the reaction classes relevant for that protein. Comparisons to BLAST reveal significant improvements in enzyme annotation accuracy. Applied to *Plasmodium falciparum*, we identify potential annotation errors and predict novel enzymes of therapeutic interest.

Availability: A standalone application is available from the website: http://www.compsysbio.org/projects/DETECT/

Contact: john.parkinson@utoronto.ca

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

Received on January 7, 2010; revised on May 14, 2010; accepted on May 17, 2010

1 INTRODUCTION

Comprising the entire network of chemical reactions carried out by a living cell, metabolism represents sets of enzymes and pathways that drive the synthesis and degradation of the essential molecules of life. The reconstruction of these networks through genome analyses are beginning to reveal a considerable diversity in metabolic complements, reflecting fundamental biological and ecological adaptations (Chen and Vitkup, 2006; Kharchenko *et al.*, 2006; Pal *et al.*, 2006; Samal *et al.*, 2006). As such, there is increasing interest in the accurate reconstruction of metabolic networks for purposes of metabolic engineering and drug discovery (Lee *et al.*, 2009; Pharkya and Maranas, 2006; Wang *et al.*, 2006).

Prior to the advent of genomic sequencing, metabolic network construction involved painstaking studies focused on individual enzymes derived from an organism of interest. As a result, only a small number of reconstructions were possible (Papoutsakis, 1984), but are very well-developed and serve as models for the reconstruction process in other organisms (Christie *et al.*, 2004). The challenge is now to apply this knowledge to help annotate the increasing numbers of genomes that are currently being sequenced. The process has been aided by the development of the Enzyme Commission (EC) hierarchy in which sequences are mapped to distinct enzymatic functions. Due to the large number of genes encoded by a genome, the annotation process is reliant on the use of automated methods based on sequence similarity or profile discovery. Commonly used tools such as BLAST (Altschul *et al.*, 1990) are useful when looking for homologues based on overall sequence similarity, while more sensitive profile based methods [e.g. PFAM (Bateman *et al.*, 2002), PROSITE (Hulo *et al.*, 2008)] focus on the conservation of domains, sequence patterns and even single residues (Mistry *et al.*, 2007; Zhang *et al.*, 2008).

Although widely applied in databases such as KEGG (Kanehisa *et al.*, 2007) and BioCyc (Caspi *et al.*, 2007), sequence-based methods suffer from limitations when dealing with enzymatic proteins. In particular, homology transfer methods typically ignore the range of sequence variation characterizing individual enzyme classes; while certain enzymes are relatively easy to classify based on sequence conservation, there are numerous examples of enzymes that share high sequence similarity and catalyze different reactions (Rost, 2002). Furthermore, the use of only a single protein match to assign function (usually the highest scoring homolog) greatly impacts the reliability of the assignment. In the absence of some measure of reliability, genes may be indiscriminately annotated with incorrect EC numbers that can lead to non-biologically relevant interpretations (Green and Karp, 2005). Consequently, there is a need to associate enzyme annotations with an appropriate confidence score, which is lacking even in recently developed functional classification tools (Espadaler *et al.*, 2008; Levy *et al.*, 2005). In an attempt to circumvent these issues, enzyme databases such as KEGG and BioCyc devote considerable effort to manual curation to improve on the quality of annotations. However, critical comparisons of these databases reveal numerous discrepancies, highlighting the importance of developing consistent methods and standards for assigning reliable gene annotations (Ginsburg, 2009).

Here, we describe a new probabilistic method for enzyme prediction based on both global and local sequence alignment, which we term Density Estimation Tool for Enzyme ClassificaTion (DETECT). DETECT uses a Bayesian framework that integrates information from density estimation profiles generated for each EC number. Instead of relying on similarity to a single sequence,

[*]To whom correspondence should be addressed.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

© The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

a probability score is calculated from the similarity scores of all relevant proteins for a particular EC number. Each EC number is represented by a family of proteins that have been either biochemically characterized or computationally predicted to have the same catalytic activity. We define an enzyme family as the set of proteins that share a common EC number. A high-scoring match of an unknown protein to a single member of an enzyme family provides evidence for shared catalytic activity, but is not always sufficient to justify the transfer of function. In many cases, the unknown protein will also exhibit varying degrees of similarity to other members of the family and, more importantly, may also show similarity to proteins from different enzyme families. Based on this prior information, we can compute a posterior probability representing the likelihood of membership to a particular enzyme family. By providing a list of probabilities associated with predicted functions for an unknown enzymatic protein, DETECT not only improves enzyme annotation accuracy, but also provides ranked lists of candidate enzymes that may be exploited for the purposes of metabolic network reconstruction. Applying DETECT to the genome of the apicomplexan *Plasmodium falciparum*, the causative agent of malaria, we demonstrate how DETECT improves on existing approaches.

2 MATERIALS AND METHODS

2.1 Protein sequence and enzyme data

Proteins used for generating the density estimation profiles were obtained from SwissProt Version 57.0 which is considered a non-redundant database (Bairoch and Apweiler, 2000). Sequence fragments were excluded by rejecting sequences annotated with a description field containing ‘fragment’. The 397 213 proteins taken from SwissProt were assigned either: (i) a complete EC number, (ii) an incomplete EC number (either partial EC or contained ‘n’ as an EC digit), (iii) multiple EC numbers or (iv) no EC number (representing unannotated enzymes and/or non-enzyme proteins). To avoid ambiguity, multifunctional enzymes (proteins that have been assigned multiple EC numbers) and incompletely annotated enzymes were excluded, leaving 127 478 proteins spanning 2277 complete EC categories. The proteome for *P.falciparum* was obtained from PlasmoDB (version 5.5, http://www.plasmodb.org; Bahl *et al*., 2003). As of January 2009, the *P.falciparum* genome consists of 5459 proteins, 2022 of which have been functionally annotated, leaving 3257 proteins with unknown function.

2.2 Generation of probability profiles

To increase efficiency of analyses, proteins from Swiss-Prot were aligned using BLAST and filtered with an E-value <1. Protein pairs from these alignments were globally aligned using the Needleman–Wunsch algorithm (Rice *et al*., 2000), resulting in a total of 55 089 045 alignments covering 4147 species (based on unique Swiss-Prot identifiers). Here, we use global rather than local alignments to avoid individual domain matches that may not be relevant to the catalytic specificity of the enzyme. All self-alignments and alignments with a BLAST bit score <50 were considered unmeaningful and excluded from the DETECT analysis. EC categories were mapped to query and hit proteins of all global alignments. Alignments were grouped by the EC annotation of the query protein. Each EC-specific set of alignments was further divided into one of two categories: (i) positive alignments: aligned protein is annotated with the same EC number as the query protein, or (ii) negative alignments: aligned protein is annotated with a different EC number (or has no EC number at all). A probability profile was generated for each enzyme, consisting of the density estimation values for positive alignment scores and density estimation values for negative alignment scores. Density estimation values were calculated using the R statistical program (http://www.r-project.org/). This was performed for all enzyme categories with 30 members or more (585 EC categories). EC categories with less than 30 members were considered to have inadequate sequence data to produce accurate and meaningful probability profiles.

2.3 Probability score calculation

The probability profiles provide a probabilistic framework for predicting the enzymatic function of an unknown protein based on sequence diversity. For an unknown protein P, pairwise global sequence alignments are generated with every Swiss-Prot protein. The resulting alignments are then categorized by the EC number of the aligned protein (only proteins annotated with EC categories are included in this analysis). The non-redundant set of hit EC numbers $E=\{E_1,E_2,\dots,E_n\}$ represents n potential enzyme activities that are performed by the unknown protein. A probability score corresponding to the likelihood that P has enzyme activity E_i is calculated for each $i=1,2,\dots,n$, based on the respective alignment scores, together with the prior information known about E_i. We integrate these multiple pieces of evidence using Bayes Theorem as follows.

Given unknown protein P and alignments that have been categorized by EC number as described earlier, let F be the EC category of interest and m be the number of alignments between P and proteins with activity F. Now define the alignments as $a_1,a_2,\dots,a_m$ having hit proteins $f_1,f_2,\dots,f_m$ and scores $s_1,s_2,\dots,s_m$. The probability p_i that P belongs to F based on a_i is

$$p_i=P(F|s_i)=\frac{P(F)\cdot P(s_i|F)}{P(F)\cdot P(s_{i|}|F)+P(F')\cdot P(s_i|F')}. \tag{1}$$

$P(F)$ is the probability that a protein belongs to F, calculated as the ratio of proteins with F activity to the total number of proteins. $P(F')$ is the probability that a protein does not belong to F, and equal to $1-P(F)$. $P(s_i|F)$ represents the probability of seeing an alignment with score s_i given that the hit protein has activity F; this is simply the density estimation value taken from the positive hit distribution for the probability profile of F. $P(s_i|F')$ is the probability of seeing an alignment with score s_i given that the hit protein does not have activity F, and is equal to the density estimation value taken from the negative hit distribution for the probability profile of F. Thus, p_i represents the likelihood that P belongs to F given its alignment to protein f_i with score s_i. This gives us m likelihood scores: $p_1,p_2,\dots,p_m$. Combining these probabilities, we obtain a single score that represents the overall likelihood that P is classified with activity F. The integrated likelihood score is calculated as

$$P_{\text{combined}}=1-\prod_{i=1}^{n}(1-p_i) \tag{2}$$

and represents a combined score that expresses increased confidence when additional evidence is available. While formula (2) is relatively simple to implement and has been applied in other contexts such as the prediction of protein-protein interactions (von Mering *et al*., 2005), it is appreciated that an alternative and more elegant solution would be to merge Equations (1) and (2) as previously described (Leontovich *et al*., 2008). However, comparisons between these approaches revealed little significant difference. In many instances, an unknown enzyme will have multiple EC predictions, which can then be ranked according to the integrated likelihood score.

2.4 Five-fold cross-validation and ROC analysis

To compare the performance of DETECT with BLAST and PSI-BLAST, we applied a 5-fold cross-validation approach to predict EC categories for proteins from Swiss-Prot. Proteins annotated with an EC category containing less than 30 protein members were removed, while the remaining 385 142 proteins were randomly assigned into five sets of equal size. For each set, training data were assembled from the other four data sets and used to generate density estimation distributions for all enzyme classes. Using these distributions, DETECT was applied to generate predictions where each EC number is associated with a probability confidence score. Separate BLAST

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

and PSI-BLAST searches were performed for each protein in the test set against the training set and the highest scoring match extracted. PSI-BLAST was run with five iterations and an E-value cutoff of 1. For each method, we defined true positives (TP) as Swiss-Prot annotated enzymes that matched the prediction with a score greater than the assigned cutoff; false positives (FP) as those proteins with incorrect enzyme predictions (including those with no enzyme annotation); true negatives (TN) as proteins neither predicted to have enzymatic activity nor annotated by Swiss-Prot as having enzyme activity; and false negatives (FN) as enzymes annotated by Swiss-Prot but not predicted to have any enzyme activity. Values were determined for a range of score cutoffs (BLAST E-value 0.1 to e^{-300}; DETECT probability 0.1 to 1.0) and were used to generate receiver operating characteristic (ROC) curves. Another method based on enzyme profiles, PRIAM (Claudel-Renard *et al*., 2003), was also considered for the ROC analysis. However, due to the implementation of the algorithm, this was not possible since PRIAM profiles cannot be regenerated for five-fold cross-validation as they have been pre-trained on Swiss-Prot.

2.5 Prediction of malarial enzymes

Each automated method (DETECT, BLAST and PRIAM) was applied to the *P.falciparum* proteome to generate predictions that were compared to the well annotated database of Malaria Parasite Metabolic Pathways (MPMP; Ginsburg, 2006; release 17 December 2008). For the DETECT predictions, density distributions were constructed after the removal of Plasmodium proteins from Swiss-Prot. After removing partial EC numbers, annotations to EC numbers with less than 30 proteins, and proteins that mapped to more than one EC number, 457 protein-EC mappings were obtained from MPMP. The query sequences were classified as either: (i) enzymes based on MPMP annotation or (ii) non-enzymes based on Gene Ontology (GO) evidence. The set of enzymes representing 457 unique proteins were annotated by MPMP with a single complete EC number containing at least 30 protein members. The set of non-enzymes represented 274 unique proteins with annotations that were not in MPMP, did not contain ‘ase’, and mapped to a GO experimental and/or author statement evidence code.

3 RESULTS AND DISCUSSION

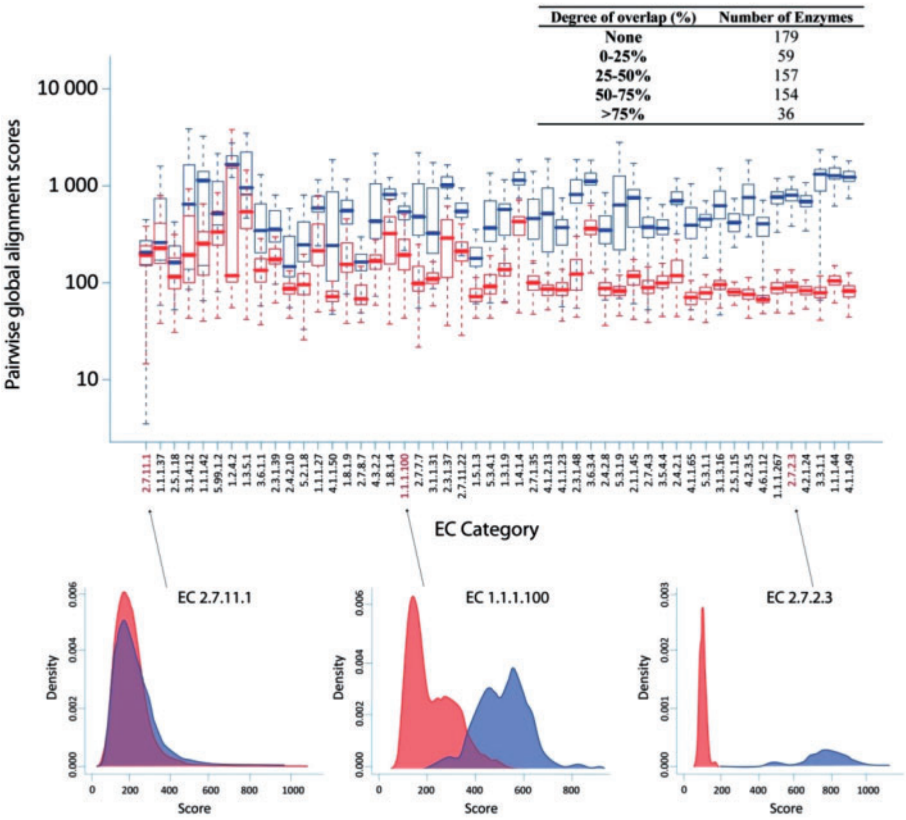
3.1 Assessing enzyme diversity

Previous sequence based methods for genome annotations have relied on single score cutoffs to define enzyme classifications. As such these methods typically ignore the range of sequence variation characterizing individual enzyme classes, which can make certain enzymes easier to discriminate on the basis of sequence similarity than others. To explore sequence diversity within different enzyme families, we performed a systematic set of global alignments for each enzyme against every other protein within the Swiss-Prot database (Boeckmann *et al*., 2005). The Swiss-Prot database was chosen as it provides manually curated high-quality protein annotations (Schnoes *et al*., 2009). From Swiss-Prot, we identified 127 478 proteins annotated with one of 2277 complete EC categories. Of these, 1285 (56%) were associated with 10 or fewer distinct protein sequences, while only 80 (4%) were represented by >400 (Supplementary Fig. S1). For each of these 127 478 enzymes, we initially performed a BLAST search against Swiss-Prot to identify potential sequence matches. Each match was globally aligned using the Needleman–Wunsch algorithm. Results were split into two groups: (i) aligned protein is annotated with the same EC number as query protein (‘positive alignment’); or (ii) aligned protein is annotated with a different EC number than query protein, or has not been annotated with an EC number (‘negative alignment’). For each EC category, similarity scores of the positive and negative alignments for each individual protein belonging to that category are combined to yield two density distributions. To reduce potential bias caused by sampling relatively small datasets, density profiles were created only for those enzyme families which contain 30 distinct proteins or more (585 EC numbers represented by 115 407 proteins; Supplementary Table S1). The results of this analysis for fifty representative enzymes are shown in Figure 1. Across all enzymes we note a wide spectrum of sequence diversity suggesting that some enzymes are easier to discriminate than others. For example, for proteins belonging to the enzyme family, EC:2.7.2.3 (phosphoglycerate kinase), scores of alignments to other family members may be readily discriminated from the alignment scores of non-family members. Conversely, for proteins belonging to EC:2.7.11.1 (serine/threonine kinases), it is difficult to distinguish family members on the basis of their alignment scores. Interestingly, we found that for enzymes on both ends of the sequence diversity spectrum, one or two EC categories predominated the negative alignments for that enzyme. This could be problematic for enzymes considered difficult to distinguish based on alignment scores, and may be explained by the level of specificity present within the EC hierarchy where a new reaction must be assigned on the basis of a very small difference in sequence specificity. For instance, EC:1.1.1.37 has most of its negative alignments to EC:1.1.1.27; while both enzymes perform similar activities, they differ in sequence by only a few key catalytic residues. We found this was the case for many enzymes with low discrimination profiles. For enzymes with highly discriminatory profiles, however, the predominant negative EC category was typically associated with a completely different enzyme activity showing no similarity in function or substrates. The enzyme EC:3.3.1.1 (Adenosylhomocysteinase), for instance, hits mostly EC:1.1.1.86 (Ketol-acid reductoisomerase) in the negative distribution. These observed differences in alignment score distributions may at least in part be attributable to the degree of substrate specificity associated with the enzyme. As an example, we note that EC:2.7.2.3 has a very specific function in the glycolysis pathway where binding of 3-phospho-D-glycerate leads to its phosphorylation. On the other hand the reaction involving EC:2.7.11.1 is thought to involve ∼30% of all cellular proteins (Cohen, 2000). Within our set of 585 enzymes, 401 (69%) exhibit some degree of overlap between their positive and negative alignments (Fig. 1). This suggests that homology-based methods used to predict enzyme function could be improved through accounting for such sequence similarity overlaps. Motivated by these considerations, in the following section we describe to the development of DETECT, a novel homology-based approach for automated enzyme annotation.

3.2 Density estimation tool for enzyme classification (DETECT)

In the previous section, we noted that the distribution of alignment scores may be used to discriminate members of enzyme families. While score increases with the length alignment, it is not true that the length of the enzyme is correlated to its sequence variability (Supplementary Fig. S2). Consequently, it should be possible to exploit these distributions to classify potentially unknown proteins. For example, from Figure 1, we note that if a protein aligns to a member of EC:1.1.1.100 with a score between 200 and 600, there is a

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016



Degree of overlap (%)	Number of Enzymes
None	179
0-25%	59
25-50%	157
50-75%	154
>75%	36

Fig. 1. Boxplots show ranges of global sequence alignment scores for a representative set of 50 gold standard enzymes. Alignments scores were divided into enzymes from the same EC category (blue) and enzymes from different EC categories (red). Enzymes are ordered by the percent overlap between the interquartile ranges of the positive and negative hit scores. Boxplots close to the right side of the graph represent enzymes that are readily distinguishable on the basis of sequence alignment scores. In contrast, enzymes closer to the left-hand side represent enzymes that are more difficult to discriminate. The lower graphs represent density distributions of alignment scores for three enzymes representing the cross-spectrum of enzyme diversity. Again, colours indicate alignment scores to enzymes from the same EC category (blue) and different EC categories (red). The Table inset shows a summary of enzyme overlap in global alignment scores that match the same EC (positive dataset) or a different EC (negative dataset). Overlap is averaged between positive and negative datasets (i.e. 10% overlap means if score range was 500, then the overlap was 50). Numbers have been presented for enzymes that have been annotated for at least 30 proteins. For further information see 'Materials and Methods' section.

level of uncertainty as to whether the protein belongs to that enzyme class. Furthermore, as score increases, there is a corresponding increase in confidence associated with the classification. Hence, through comparing the distribution of alignment scores of an unknown protein to the set of proteins associated with an enzyme family in comparison to those not belonging to that family, it is possible to obtain a probability of the unknown protein belonging to that enzyme class. For example, given an alignment with a score of 450, an unknown protein is more likely to belong to 1.1.1.100 than to another EC number since the score distribution of positive alignments is denser than that of negative alignments. Conversely, for an alignment with a score of 300, an unknown protein is unlikely to belong to 1.1.1.100. Hence by incorporating information on the sequence diversity across the various enzyme families, it is possible to compute a probability score for the likely association of an unknown protein to each enzyme family.

Based on these ideas, we have developed a novel algorithm, which we term DETECT which applies a Bayesian statistical framework to the enzyme density profiles generated above, to assign a probability score to an enzyme annotation. For a query protein, q, a BLAST search first retrieves all matches of q to Swiss-Prot. For each EC category within the set of matches, the Needleman–Wunsch algorithm was used to generate global alignment scores between $q$and each protein assigned that EC category. These scores are then used to derive a pair of density estimation values from the density distribution plot for that EC category generated earlier (Fig. 2A). The two values correspond to the density of positive and negative alignment scores previously generated for that EC category. Multiple pairs of values generated from all matches to proteins from the same EC category, are combined within a Bayesian framework, to generate an integrated likelihood score (ILS) that q belongs to that EC category. Briefly, for each pair of scores, a likelihood score is generated based on prior probabilities associated with that EC category. The final ILS is then obtained from the product of all likelihood scores. See Materials and Methods for further details.

To evaluate performance, we compared DETECT with BLAST and PSI-BLAST using 5-fold cross-validation to predict enzyme activities for proteins in Swiss-Prot. Due to difficulties in differentiating between enzymes with multiple catalytic sites and those with a single catalytic site performing multiple types of

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

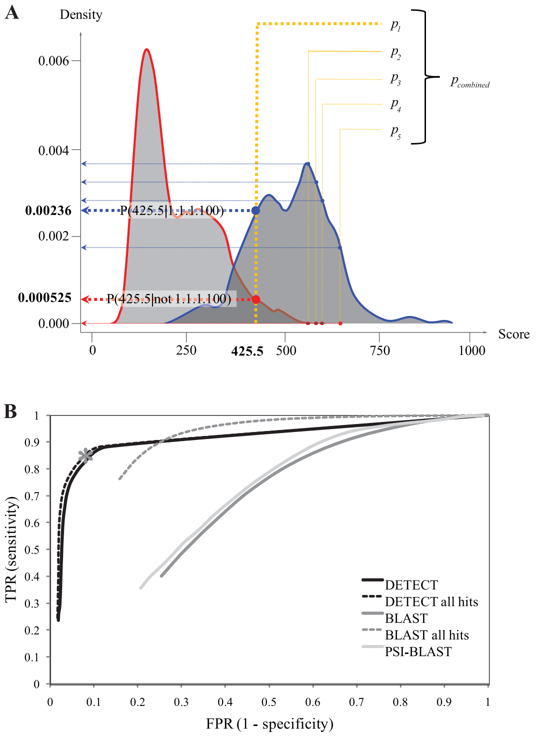


Fig. 2. (**A**) Plot showing the density estimates of P(*alignment score* | positive alignment to EC:1.1.1.100) and P(*alignment score* | negative alignment to EC:1.1.1.100). Using the density estimation values of a given *alignment score*, we can obtain the probability the protein that has aligned to an EC:1.1.1.100 protein belongs to the same enzyme category. For brevity, 'positive alignment to EC:1.1.1.00' is abbreviated '1.1.1.100', 'negative alignment to EC:1.1.1.100' is abbreviated '$\overline{1.1.1.100}$', and 'alignment score' is abbreviated as 'score'. In this formula, $P(1.1.1.100|score)=\frac{P(score|1.1.1.100)\cdot P(1.1.1.100)}{P(score|1.1.1.100)\cdot P(1.1.1.100)+P(score|\overline{1.1.1.100})\cdot P(\overline{1.1.1.100})}$ (**B**) ROC curves for DETECT, BLAST and PSI-BLAST based on data generated from 5-fold cross-validation analyses.

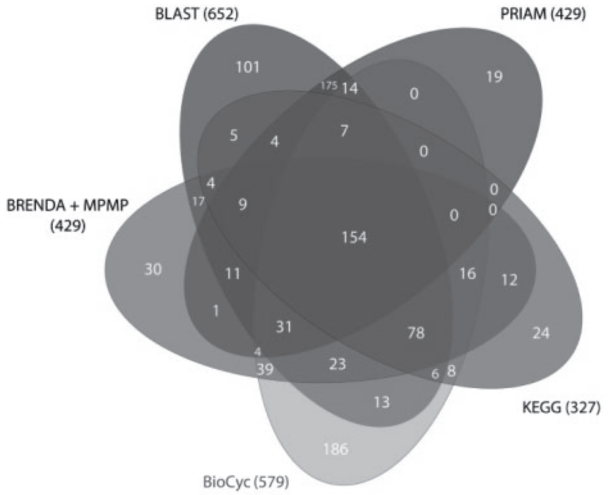


Fig. 3. Overlap in enzyme curations and predictions for *P.falciparum* as annotated by four database resources: KEGG (Kanehisa *et al.*, 2007), BioCyc (Caspi *et al.*, 2007), BRENDA (Schomburg *et al.*, 2004) and the Malaria Parasite Metabolic Pathways (MPMP; Ginsburg, 2006), and two typically applied automated annotation methods: BLAST (top scoring match against all enzymes in Swiss-Prot using a cutoff of e^{-10}) and PRIAM (again using a cutoff of e^{-10}; Claudel-Renard *et al.*, 2003).

reactions, proteins assigned multiple EC categories were not included. Applying a range of score cutoffs, DETECT predictions were defined as the EC category with the highest ILS; BLAST and PSI-BLAST predictions were the highest scoring hit from the training set. Proteins predicted to have two or more EC numbers were not included in the performance analysis. DETECT and BLAST were also evaluated based on the entire list of predictions (not just top-scoring prediction). From the ROC curves in Figure 2B, it is clear that the DETECT approach significantly outperforms the BLAST-based methods with greater sensitivity and specificity across all score thresholds. The gain in performance when the entire list of predictions is considered is almost negligible, illustrating the discriminatory power of DETECT and highlighting the usefulness of a ranked list of ILSs. In contrast, BLAST (similar to PSI-BLAST) predict with much lower accuracy when the top-scoring hit is considered.

3.3 Comparison to current prediction methods

One of the central aims of developing DETECT is to improve the automated prediction of enzyme function from genomic information. Here, we are interested in DETECT's ability to accurately predict metabolic enzymes from the annotated genome of *P.falciparum*. To date, a number of resources have been constructed that describe genome-scale reconstructions of enzyme complements. For *P.falciparum*, these include KEGG (Kanehisa *et al.*, 2007), BioCyc (Caspi *et al.*, 2007), BRENDA (Schomburg *et al.*, 2004) and the MPMP (Ginsburg, 2006). Each dataset exhibits a wide range in coverage and accuracy, reflecting potential errors and biases in their respective curations (Fig. 3). Due its high standard of curation and coverage, we chose MPMP as a suitable baseline to examine the performance of DETECT relative to the other resources, simple BLAST analyses and predictions derived from a profile-based tool termed PRIAM (Claudel-Renard *et al.*, 2003; Table 1 and Supplementary Tables S1–S3). For this analysis, predictions of DETECT and BLAST are comparable due to the lack of a good negative training set for *P.falciparum*. Consequently, the improved specificity observed with Swiss-Prot for DETECT is difficult to assess. Of 765 proteins annotated with EC numbers by MPMP, DETECT predicts activities for 622. Of these, 325 (52.3%) matched the MPMP/BRENDA annotations. However, within this set of 622 proteins are 198 which are annotated by MPMP/BRENDA with EC numbers not present within the set of 582 EC categories used by the DETECT algorithm (Supplementary Table S2), ignoring these increased correct predictions to 76.7%. Furthermore, considering only predictions with an ILS in excess of 0.2 (a defined cutoff, providing a good compromise between specificity and sensitivity as derived with reference to Fig. 2B), DETECT correctly predicts 246 of 276 proteins (89.1%). BLAST unsurprisingly provided a higher number of matches—352 (58.4%) compared to DETECT. However, PRIAM was found to correctly predict only 117 of the 424 proteins annotated with one of the 582 EC categories, while KEGG and BioCyc match 233 (55.0%) and 150 (35.4%), respectively. We also performed a comparison to a generic automatic annotation tool based on the transfer of Gene Ontology terms, Blast2GO (Conesa, 2005).

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

Table 1. Matches of various resources and methods to MPMP annotations

Dataset	Predictions for 765 proteins annotated by MPMP	Correct predictions (%)	Predictions for 582 categories[a]	Correct predictions for 582 categories[a] (%)	Additional predictions (not in MPMP)
DETECT	622	325 (52.3)	424	325 (76.7)	2451
DETECT (>0.2)	318	246 (77.4)	276	246 (89.1)	465
BLAST	603	352 (58.4)	365	309 (84.7)	2342
PRIAM	258	169 (65.5)	147	117 (79.6)	143
KEGG	470	389 (82.8)	264	233 (88.3)	201
BioCyc	285	263 (92.3)	160	150 (93.8)	21

[a] The 582 categories refer to those EC categories with at least 30 representatives in the Swiss-Prot dataset.

Table 2. Lists of proteins correctly annotated by either DETECT or BLAST

Protein ID	DETECT prediction	ILS	Positives	Negatives	MPMP annotation	BLAST prediction	PlasmoDB annotation
PFF1145c	2.7.10.2	0.97	134	744	2.7.10.2	2.7.11.25	Phosphatidylinositol 4-kinase, putative
PF10_0320	3.1.3.16	0.94	7	28	3.1.3.16	4.6.1.1	Lipoate-protein ligase A type 2
MAL13P1.301	4.6.1.2	0.27	49	109	4.6.1.2	3.6.3.1	m1-family aminopeptidase
PF11_0395	4.6.1.2	0.91	49	92	4.6.1.2	3.6.3.1	M18 aspartyl aminopeptidase
PF13_0141	1.1.1.37	0.97	292	268	1.1.1.27	1.1.1.27	L-lactate dehydrogenase
MAL13P1.122	5.2.1.8	0.85	4	176	2.1.1.43	2.1.1.43	SET domain protein, putative
PFB0505c	2.3.1.180	0.30	163	201	2.3.1.41; 2.3.1.85	2.3.1.41	β−Ketoacyl-acyl carrier protein synthase III precursor, putative
PF11_0242	6.5.1.1	0.42	1	1004	2.7.11.17	2.7.11.17	calcium-dependent protein kinase, putative
PF11_0060	3.1.3.48	1.00	2	515	2.7.11.17	2.7.11.17	exported serine/threonine protein kinase
PFD0740w	2.7.10.2	0.68	150	907	2.7.11.22	2.7.11.22	Ser/Thr protein kinase, putative
PFF0275c	6.1.1.7	1.00	1	520	2.7.4.6	2.7.4.6	Protein kinase, putative
PFA0340w	6.1.1.7	1.00	1	137	2.7.7.60	2.7.7.60	Casein kinase II, α-subunit
PFF0745c	5.2.1.8	0.93	1	62	3.1.-.-; 3.1.13.1	3.1.13.1	Adenylate kinase
PFL0475w	3.1.26.5	1.00	1	50	3.1.4.17	3.1.4.35	Glycosyltransferase family 28 protein, putative

From the list of 276 proteins with DETECT ILS scores >0.2 and members of the 582 EC categories with ≥30 members, four were correctly identified by DETECT but not BLAST (*E*-value <1), and 10 were correctly identified by BLAST but not DETECT. Grey backgrounds indicate predictions matching MPMP annotations. Positives and negatives indicate the number of proteins used to formulate the DETECT prediction. Seven of the incorrect DETECT predictions were based on matches to less than five proteins.

Using MPMP as the gold standard, DETECT was found to have significantly greater accuracy and coverage than Blast2GO. Of the 301 Blast2GO annotations only 154 matched MPMP annotations (51.1%) compared to an accuracy of 76.7% for DETECT (data not shown).

Of the 246 DETECT predictions with an ILS>0.2, which match MPMP/BRENDA annotations, four were incorrectly predicted by BLAST (Table 2). These include MAL13P1.301 and PF11_0395, annotated by MPMP as guanylyl cyclases (GCs - EC:4.6.1.2). GCs are responsible for the synthesis of cyclic GMP, an important secondary messenger molecule. In eukaryotes, GCs are highly conserved and the presence of multiple isoforms is common. The two Plasmodium GCs, however, are significantly divergent from mammalian homologues in both structure and function (Baker and Kelly, 2004). In addition to containing two GC catalytic domains, the sequences of MAL13P1.301 and PF11_0395 appear to have regions similar to P-type ATPases (EC:3.6.3.-; Fig. 4). On the basis of only local similarity, applying BLAST would incorrectly assign these proteins to EC:3.6.3.1. On the other hand, from Figure 4, we note the distributions of MAL13P1.301 and PF11_0395 alignment scores to GCs produce significantly greater overlap to the positive distribution for EC:4.6.1.2 than for EC:3.6.3.1 resulting in higher ILS's (MAL13P1.301: ILS = 0.27 and 0.0024 for EC:4.6.1.2 and EC:3.6.3.1, respectively; PF11_039: ILS = 0.91 and 0.55 for EC:4.6.1.2 and EC:3.6.3.1, respectively). While a role for the P-type ATPase domain has yet to be found, a previous study has suggested the two isoforms of GC encode bifunctional enzymes in *P.falciparum* (Carucci *et al*., 2000) The diverse nature of these sequences relative to their hosts, suggests that these two GCs may represent suitable targets for therapeutic intervention. The serine/threonine-specific protein phosphatase (EC:3.1.3.16) activity of PF10_0320, was also correctly predicted by DETECT but not by BLAST. Such phosphatases have been shown to play an important role both in parasite growth and development (Ward *et al*., 1994) and also in the invasion process (Rangachari *et al*., 1986). The density estimation profiles resulted in the assignment of a high probability for EC:3.1.3.16 (ILS = 0.94) relative to the top BLAST prediction of EC:6.1.1.17 (ILS = 10^{-5}). Conversely, there were 10 examples where BLAST but not DETECT correctly predicts enzyme function (Table 2). Note that the majority of these proteins were correctly predicted by DETECT based on the top 2nd hit, further supporting the usefulness of having a ranked list of enzymes. Additionally, while only two of 246 correct DETECT predictions had less than five positive matches, we note that seven of the ten incorrect predictions were based on fewer than five positive matches. This suggests that the number of positive matches should be used as additional criteria to filter DETECT predictions. Intriguingly, the remaining three incorrect DETECT predictions were based on a large number

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

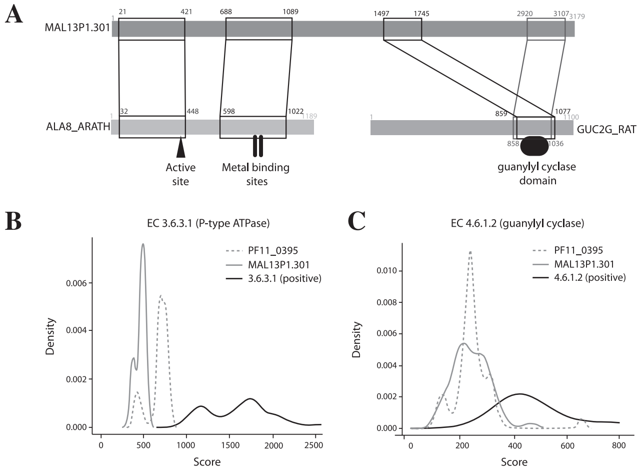


Fig. 4. (**A**) Plasmodium protein MAL13P1.301 appears to be a bifunctional enzyme based on an alignment to a P-Type ATPase (ALA8_ARATH) and guanylyl cyclase (GUC2G_RAT). (**B** and **C**) Overlap of the density distribution of Plasmodium alignment scores to P-type ATPase (B) or guanylyl cyclase (C) enzymes compared to the density distribution of positive alignment scores for the respective enzyme class.

of positive sequence matches to their respective EC categories. This might suggest that the MPMP annotations are incorrect. For instance, PFB0505c is reported in PlasmoDB as a β-ketoacyl-ACP synthase III (fabH; EC:2.3.1.180) but annotated by MPMP as a β-ketoacyl-ACP synthase I (fabB; EC:2.3.1.41) and fatty-acid synthase (FASN; EC:2.3.1.85). Based on a large number of alignments to both EC:2.3.1.180 and EC:2.3.1.41, DETECT predicts the gene to have FabH activity, whereas BLAST predicts it to have FabB activity. Although FabH activity has been demonstrated for *P.falciparum* (Waller *et al*., 1998), recent experimental studies focusing on a putative FabB/F enzyme indicates that FabF and not FabB activity is also present (Sharma *et al*., 2009). Another potential misannotation involves PF13_0141, assigned L-lactate dehydrogenase activity (LDH; EC:1.1.1.27) by MPMP but predicted by DETECT to be malate dehydrogenase (MDH; EC:1.1.1.37). In *Plasmodium*, LDH is an essential enzyme required for the production of ATP (Gomez *et al*., 1997), while MDH plays a crucial role in pathogenicity (Chan and Sim, 2004). Members of both families have been characterized and sequenced from a wide variety of organisms representative of all domains of life (Madern, 2002). While structurally distinct, MDH and LDH share significant sequence similarity. Furthermore, analysis of the MDH family identified two distinct groups of closely related enzymes (Goward and Nicholls, 1994), responsible for a bimodal distribution of within-family alignment scores (Fig. 5). This is further supported by a group of LDH members that share greater sequence similarity to the MDH family than the rest of the LDH's (Pazos *et al*., 2006). Based on global alignments, PF13_0141 produces many significant matches to both LDH and MDH family members. Inspection of the resulting density distribution profiles indicates a better fit to the distribution of positive alignment scores for EC:1.1.1.37 than that of EC:1.1.1.27 (Fig. 5), resulting in a higher ILS (0.97 versus 0.31). Interestingly, a recent study uncovered additional MDH activity for the Plasmodium LDH based on molecular function (Wiwanitkit, 2007), suggesting that perhaps PF13_0141 has broad specificity.

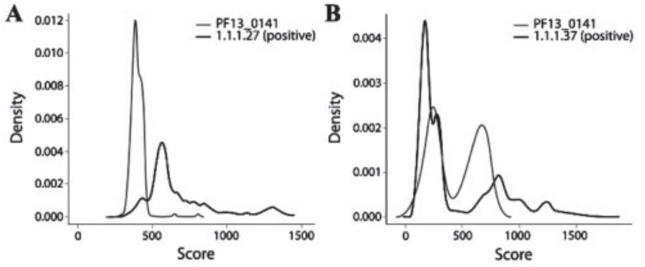


Fig. 5. The distribution of scores for alignments of PF13_0141 against (**A**) EC:1.1.1.27 and (**B**) EC:1.1.1.37 proteins compared to the respective within-family alignments indicates that PF13_0141 is more likely to belong to EC:1.1.1.37 based on density estimation profiles.

3.4 Expanding the metabolome of P. falciparum

As noted above, BLAST proved to be very effective in confirming MPMP predictions. However, a major challenge in enzyme annotation is to be able to discriminate false positive predictions. By calculating an ILS that accounts for observed diversity in sequence similarity distributions, DETECT provides a useful method for the generation of ranked lists of enzyme predictions that can be combined with metabolic network reconstruction (Caspi *et al*., 2007; Kanehisa *et al*., 2007) methods to help prioritize candidate enzymes for more detailed studies. Applied to *Plasmodium*, DETECT predicts activities for an additional 2451 genes (Supplementary Table S3). However, using an ILS > 0.2, relying on predictions involving at least six positive sequence matches and ignoring the highly promiscuous EC categories—2.7.11.1 (serine/threonine kinase: 47 annotations), 2.7.13.3 (histidine kinase: 735 annotions) and 2.7.7.6 (RNA polymerase: 70 annotations), results in a high confidence list of 88 predictions worthy of follow-up investigations.

The conserved Plasmodium protein of unknown function, PF11_0207, is predicted by DETECT to be EC:5.99.1.2 (DNA topoisomerase) with an ILS ~0.48. Biochemical evidence has been found for topoisomerase (Topo) I and II in *P.falciparum* (Cheesman *et al*., 1994; Tosh and Kilbey, 1995), and a number of putative annotations for Topo are reported in PlasmoDB. BLAST, on the other hand predicts PF11_0207 to be EC:6.5.1.1 (DNA ligase) with an E-value $<10^{-14}$. While many eukaryotes express three DNA ligase isoforms, *P.falciparum* expresses only one (Cheesman *et al*., 1994). The gene for the sole ligase, DNA ligase I, has been identified, and no evidence suggests the presence of additional isoforms in the parasite. The density profile for PF11_0207 alignment scores against the positive score distribution for EC:5.99.1.2 shows a large degree of overlap, compared to a single alignment for EC:6.5.1.1, suggesting a more likely membership to the Topo family (Supplementary Fig. S3). The accurate identification of the genes encoding Topo may be important as studies have suggested that certain Topo II poisons may act selectively against the human malaria parasite (Chavalitshewinkoon *et al*., 1993; Gamage *et al*., 1994).

DETECT also serves to reinforce the predictions of existing tools when searching for genes with unidentified enzyme activity. For instance, PFI1475w, a merozoite surface protein, is predicted by BLAST to be EC:3.1.11.6 (Exonuclease VII) with only moderate statistical support (E-value <0.001). With an ILS~0.56, DETECT also makes the same prediction, where analysis of the density profile for matching PFI1475w alignment scores against the positive score

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

distribution of EC:3.1.11.6 indicates a good fit (Supplementary Fig. S4). The putative Exonuclease VII has not been biochemically characterized in *P.falciparum*, but genes for Exonuclease I and III have been identified, suggesting other exonucleases are yet to be identified.

Even in instances of less than six positive matches, the ILS can produce informative predictions. For example two genes annotated with 'unknown function' are PFD0485w and PFL1535w, which DETECT predicts to have glucose-6-phosphate isomerase (GPI) activity (EC:5.3.1.9) with high confidence scores of 0.90 and 0.91, respectively. One may dismiss these as spurious predictions given that the experimentally characterized GPI (Srivastava *et al.*, 1992) has been annotated in *P.falciparum* as PF14_0341. This enzyme, which catalyzes the reversible conversion of glucose-6-phosphate to fructose-6-phosphate, is necessary for glycolysis in *P.falciparum*. As evidenced by the absence of mitochondrial activity in the parasite (Mather *et al.*, 2007), the energy needs of *P.falciparum* are met entirely through the anaerobic consumption of exogenous glucose. While PF14_0341 is the only gene that has been annotated to have GPI activity, purification studies detected the presence of several isozymes in *P.falciparum* (Srivastava *et al.*, 1992). The appearance of multiple isozymes is a common feature for Plasmodial glycolytic enzymes (Maeda *et al.*, 2009) suggesting that PFD0485w and PFL1535w may represent isozymes for GPI. Global alignments of PFD0485w and PFL1535w to non-Plasmodium members of the GPI family have optimal scores based on their fit to the positive distribution for EC:5.3.1.9 (Supplementary Fig. S5).

Finally, in cases where enzyme activity may be expected (e.g. through biochemical assays or hole filling algorithms), DETECT provides a powerful approach to identify and prioritize candidate proteins that may be at the extreme of sequence homology detection. The MPMP database identifies 428 enzyme categories, of which 404 have been annotated to one or more proteins. Of the remaining categories, DETECT predicts a protein for four although all with low ILSs: EC:1.1.1.25 (PF14_0424, ILS $\sim 10^{-7}$), EC:1.4.4.2 (PFL2095w, ILS ~0.001), EC:2.7.1.23 (PFI0650c, ILS ~0.001) and EC:3.6.1.1 (PF14_0541; PFL1700c; PF11_0190; and PF11_0202, ILS:1; 1; ~0.003; ~0.0006, respectively). PF14_0541 and PFL1700c are both annotated in PlasmoDB as putative V-type H+-translocating pyrophosphatases, confirming the DETECT predictions, but currently lack the appropriate EC designation. As a component of the shikimate pathway, not present in mammals EC:1.1.1.25 (shikimate dehydrogenase) is of particular interest from a drug target point of view. Evidence for the requirement of the shikimate pathway was originally provided by isolation of *P.falciparum* mutants requiring pABA for growth (McConkey *et al.*, 1994). While the presence of shikimate dehydrogenase has been reported at very low levels in *P.falciparum*, no protein has previously been identified. Although possessing only a low ILS, PF14_0424, produces reasonable alignments to the active site of three shikimate dehydrogenase proteins found in bacteria (Supplementary Fig. S6). As such, PF14_0424 represents a suitable target for confirming shikimate dehydrogenase activity.

3.5 Applications and future work

By focusing on *P.falciparum* as a model organism, we have shown that sequence diversity is an important factor in the accurate identification of enzymes. Through providing a ranked list of ILSs, DETECT facilitates the prioritization of potential novel enzymes that can help guide species-specific metabolic reconstructions. A future goal is to apply DETECT to the prediction of multifunctional enzymes. Current Swiss-Prot annotations indicate that there are over 4000 proteins that catalyze more than one reaction. A potential approach involves assessing all high-scoring predictions made by DETECT and examining the potential of multiple catalytic domains within the protein of interest.

ACKNOWLEDGEMENTS

The authors would like to thank Ping-Zhao Hu and three anonymous reviewers for helpful suggestions and advice.

Funding: Canadian Institutes of Health Research (CIHR #MOP84556). CIHR New Investigators award; Early Researcher award from the Ontario Ministry of Research and Innovation (to J.P.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bahl,A. *et al.* (2003) PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. *Nucleic Acids Res.*, **31**, 212–215.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Baker,D.A. and Kelly,J.M. (2004) Purine nucleotide cyclases in the malaria parasite. *Trends Parasitol.*, **20**, 227–232.

Bateman,A. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.

Boeckmann,B. *et al.* (2005) Protein variety and functional diversity: Swiss-Prot annotation in its biological context. *C R Biol.*, **328**, 882–899.

Carucci,D.J. *et al.* (2000) Guanylyl cyclase activity associated with putative bifunctional integral membrane proteins in Plasmodium falciparum. *J. Biol. Chem.*, **275**, 22147–22156.

Caspi,R. *et al.* (2007) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res*., **36,** D623–D631.

Chan,M. and Sim,T.S. (2004) Functional characterization of an alternative [lactate dehydrogenase-like] malate dehydrogenase in Plasmodium falciparum. *Parasitol Res.*, **92**, 43–47.

Chavalitshewinkoon,P. *et al.* (1993) Structure-activity relationships and modes of action of 9-anilinoacridines against chloroquine-resistant Plasmodium falciparum in vitro. *Antimicrob Agents Chemother.*, **37**, 403–406.

Cheesman,S. *et al.* (1994) The gene encoding topoisomerase II from Plasmodium falciparum. *Nucleic Acids Res.*, **22**, 2547–2551.

Chen,L. and Vitkup,D. (2006) Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biol.*, **7,** R17.

Christie,K.R. *et al.* (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res.*, **32**, D311–D314.

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

Cohen,P. (2000) The regulation of protein function by multisite phosphorylation–a 25 year update. *Trends Biochem. Sci.*, **25**, 596–601.

Espadaler,J. *et al.* (2008) Prediction of enzyme function by combining sequence similarity and protein interactions. *BMC Bioinformatics*, **9,** 249.

Gamage,S.A. *et al.* (1994) Synthesis and in vitro evaluation of 9-anilino-3,6-diaminoacridines active against a multidrug-resistant strain of the malaria parasite Plasmodium falciparum. *J. Med. Chem.*, **37**, 1486–1494.

Ginsburg,H. (2006) Progress in in silico functional genomics: the malaria Metabolic Pathways database. *Trends Parasitol.*, **22**, 238–240.

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

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

Gomez,M.S. *et al*. (1997) Substrate and cofactor specificity and selective inhibition of lactate dehydrogenase from the malarial parasite P. falciparum. *Mol. Biochem. Parasitol.*, **90**, 235–246.

Goward,C.R. and Nicholls,D.J. (1994) Malate dehydrogenase: a model for structure, evolution, and catalysis. *Protein Sci.*, **3**, 1883–1888.

Green,M.L. and Karp,P.D. (2005) Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucleic Acids Res.*, **33**, 4035–4039.

Hulo,N. *et al*. (2008) The 20 years of PROSITE. *Nucleic Acids Res.*, **36**, D245–D249.

Kanehisa,M. *et al*. (2007) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **35** (Web server issue), W18–W185.

Kharchenko,P. *et al*. (2006) Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics*, **7**, 177.

Lee,D.S. *et al*. (2009) Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple Staphylococcus aureus genomes identify novel antimicrobial drug targets. *J. Bacteriol.*, **191**, 4015–4024.

Leontovich,A.M. *et al*. (2008) The comparative analysis of statistics, based on the likelihood ratio criterion, in the automated annotation problem. *BMC Bioinformatics*, **9**, 31.

Levy,E.D. *et al*. (2005) Probabilistic annotation of protein sequences based on functional classifications. *BMC Bioinformatics*, **6**, 302.

Madern,D. (2002) Molecular evolution within the L-malate and L-lactate dehydrogenase super-family. *J. Mol. Evol.*, **54**, 825–840.

Maeda,T. *et al*. (2009) Pyruvate kinase type-II isozyme in Plasmodium falciparum localizes to the apicoplast. *Parasitol. Int.*, **58**, 101–105.

Mather,M.W. *et al*. (2007) Mitochondrial drug targets in apicomplexan parasites. *Curr. Drug Targets*, **8**, 49–60.

McConkey,G.A. *et al*. (1994) Auxotrophs of Plasmodium falciparum dependent on p-aminobenzoic acid for growth. *Proc. Natl Acad. Sci. USA*, **91**, 4244–4248.

Mistry,J. *et al*. (2007) Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics*, **8**, 298.

Pal,C. *et al*. (2006) Chance and necessity in the evolution of minimal metabolic networks, *Nature*, **440**, 667–670.

Papoutsakis,E.T. (1984) Equations and calculations for fermentations of butyric acid bacteria. *Biotechnol. Bioeng.*, **26**, 174–187.

Pazos,F. *et al*. (2006) Phylogeny-independent detection of functional residues. *Bioinformatics*, **22**, 1440–1448.

Pharkya,P. and Maranas,C.D. (2006) An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metab. Eng.*, **8**, 1–13.

Rangachari,K. *et al*. (1986) Control of malarial invasion by phosphorylation of the host cell membrane cytoskeleton. *Nature*, **324**, 364–365.

Rice,P. *et al*. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.

Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.

Samal,A. *et al*. (2006) Low degree metabolites explain essential reactions and enhance modularity in biological networks. *BMC Bioinformatics*, **7**, 118.

Schnoes,A.M. *et al*. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, **5**, e1000605.

Schomburg,I. *et al*. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32,** D431–D433.

Sharma,S. *et al*. (2009) Beta-ketoacyl-ACP synthase I/II from Plasmodium falciparum (PfFabB/F)–is it B or F? *IUBMB Life*, **61**, 658–662.

Srivastava,I.K. *et al*. (1992) Identification and purification of glucose phosphate isomerase of Plasmodium falciparum. *Mol. Biochem. Parasitol.*, **54**, 153–164.

Tosh,K. and Kilbey,B. (1995) The gene encoding topoisomerase I from the human malaria parasite Plasmodium falciparum. *Gene*, **163**, 151–154.

von Mering,C. *et al*. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, **33**, D433–D437.

Waller,R.F. *et al*. (1998) Nuclear-encoded proteins target to the plastid in Toxoplasma gondii and Plasmodium falciparum. *Proc. Natl Acad. Sci. USA*, **95**, 12352–12357.

Wang,Q. *et al*. (2006) Genome-scale in silico aided metabolic analysis and flux comparisons of Escherichia coli to improve succinate production. *Appl. Microbiol. Biotechnol.*, **73**, 887–894.

Ward,G.E. *et al*. (1994) Staurosporine inhibits invasion of erythrocytes by malarial merozoites. *Exp. Parasitol.*, **79**, 480–487.

Wiwanitkit,V. (2007) Plasmodium and host lactate dehydrogenase molecular function and biological pathways: implication for antimalarial drug discovery. *Chem. Biol. Drug Des.*, **69**, 280–283.

Zhang,T. *et al*. (2008) Accurate sequence-based prediction of catalytic residues. *Bioinformatics*, **24**, 2329–2338.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016