Published online December 1, 2004

6226–6239 *Nucleic Acids Research, 2004, Vol. 32, No. 21*
doi:10.1093/nar/gkh956

EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference

Weidong Tian[1,2], Adrian K. Arakaki[1] and Jeffrey Skolnick[1,*]

[1]Center of Excellence in Bioinformatics, University at Buffalo, The State University of New York, 901 Washington Street, Buffalo, NY 14203-1199, USA and [2]Department of Biology, Washington University in St Louis, One Brookings Drive, St Louis, MO 63130, USA

Received October 8, 2004; Revised and Accepted November 4, 2004

ABSTRACT

EFICAz (Enzyme Function Inference by Combined Approach) is an automatic engine for large-scale enzyme function inference that combines predictions from four different methods developed and optimized to achieve high prediction accuracy: (i) recognition of functionally discriminating residues (FDRs) in enzyme families obtained by a Conservation-controlled HMM Iterative procedure for Enzyme Family classification (CHIEFc), (ii) pairwise sequence comparison using a family specific Sequence Identity Threshold, (iii) recognition of FDRs in Multiple Pfam enzyme families, and (iv) recognition of multiple Prosite patterns of high specificity. For FDR (i.e. conserved positions in an enzyme family that discriminate between true and false members of the family) identification, we have developed an Evolutionary Footprinting method that uses evolutionary information from homofunctional and heterofunctional multiple sequence alignments associated with an enzyme family. The FDRs show a significant correlation with annotated active site residues. In a jackknife test, EFICAz shows high accuracy (92%) and sensitivity (82%) for predicting four EC digits in testing sequences that are <40% identical to any member of the corresponding training set. Applied to *Escherichia coli* genome, EFICAz assigns more detailed enzymatic function than KEGG, and generates numerous novel predictions.

INTRODUCTION

A main goal in the post-genomic era is to identify the function of each newly determined sequence (1). About 40% of the sequences in genomic databases correspond to open reading frames whose annotated functions are missing, incomplete or incorrect (2). Unfortunately, the experimental study of these uncharacterized sequences is costly and time-consuming. Thus, computational methods for the inference of protein function are of great importance to both assist in as well as accelerate the annotation process (3).

Although protein function can be defined on many levels, and predicted based on different properties, here we focus on the inference of enzyme function from sequence. Enzymes represent the most versatile group of all proteins, catalyzing the chemical reactions associated with the metabolism of all living organisms. They constitute a significant fraction of a genome; in higher eukaryotes, the fraction of genes encoding enzymes may be 25–30% (4). The three basic approaches to the inference of enzyme function from sequence are based on (i) homology transfer, (ii) presence of a pattern or motif, and (iii) identification of functional residues.

Homology transfer is the most widely used approach for functional annotation (3). It involves two steps: the detection of homology and the inference of function from homology. With the development of sensitive algorithms such as FASTA (5), BLAST (6), PSI-BLAST (7) and Hidden Markov Models (HMM) (8), the ability to recognize evolutionarily related proteins (i.e. homologs) has been greatly improved. In contrast, the inference of function from homology remains error prone (9) and is only reliable at high levels of sequence identity (3). For example, we have previously shown that, on average, 60% pairwise sequence identity is required to transfer the four EC digits of an enzyme with at least 90% accuracy (10).

A different approach for the inference of enzyme function from sequence is based on the presence of patterns or motifs associated with functional sites. Currently, there are a number of pattern or motif databases developed for functional annotation, such as Prosite (11), PRINTS (12) and BLOCKS (13). However, these databases are not specifically developed to infer enzymatic function. In fact, many Prosite patterns associated with the active site of a specific enzyme are also found in sequences of many unrelated enzymes. For example, we find that the Prosite pattern Aldehyde dehydrogenases glutamic acid active site (PS00687) is present in sequences of 24 different enzymes that correspond to all major reaction types. Recently, a method has been developed aimed at improving the specificity of poorly performing Prosite patterns, but it is limited to those cases in which structural information is available (14).

The biological activity of an enzyme is typically determined by a small number of functional residues, which are usually conserved. Thus, rather than only examining the sequence similarity over the entire sequence, a more sensitive means of inferring function is to also examine whether the residues

*To whom correspondence should be addressed. Tel: +1 716 849 6711; Fax: +1 716 849 6747; Email: skolnick@buffalo.edu

responsible for the given function are conserved. Unfortunately, the functional residues of most enzymes are unknown. Therefore, it is necessary first to determine the residues more likely to characterize an enzyme's function, and then to analyze their degree of conservation. Most methods for the prediction of functional residues involve the selection of conserved residues in a family of proteins, where the families are identified on the basis of sequence and/or structural information (15–19). Thus, these methods implicitly assume that proteins above a certain SIT have the same function. For example, the 'evolutionary trace' method classifies proteins in sub-families mainly based on sequence identity and identifies residues that are conserved in individual sub-families but vary between different sub-families, which are likely to be functionally important (17). However, considering that some enzyme functions diverge even when their sequence identity is high, the classification of proteins by only sequence identity may not be accurate enough to infer functional similarity, and human intervention might be required (16).

The Pfam-based functional subtype analysis developed by Hannenhalli and Russell (20) avoids the problem of inconsistency between sequence and functional similarity by using annotation to define functional sub-families. Their method starts with a Pfam domain (21) associated with enzyme sequences and aims at identifying positions in the domain that distinguish among subtypes of the enzyme with different substrate specificities. However, there are potential problems associated with this method. First, a Pfam family is a collection of evolutionarily related sequences that generally include both enzymes and non-enzymes, the latter being difficult to classify into subtypes. Second, in practice, a Pfam family corresponds to a single domain; in contrast, an enzymatic function may require participation of multiple domains. Thus, using a single Pfam domain may not be sufficient to identify the correct subfamily-specific residues.

Another limitation common to most methods used to identifying functional residues is related to the large number of conserved residues found in protein families having a small number of known members or whose members are mostly close homologs. The greater the number of conserved residues in a family, the more difficult it is to select those responsible for the function of interest.

The goal of the present work is to address the limitations of the current approaches and to develop new methods for highly accurate genome-scale enzyme function inference. Rather than using sequence similarity to infer functional similarity, we directly use the functional annotation of proteins and define each enzyme family as a group of proteins that are evolutionarily related and share four or three digits of their EC numbers. Thus, a given EC number can be related to more than one enzyme family. We recognize of course that the EC classification itself can have problems (22); nevertheless, it is convenient. Employing an automatic Conservation-controlled HMM Iterative procedure for Enzyme Family classification (CHIEFc), we have classified all the sequences in the ENZYME database (23), and have obtained multiple sequence alignments (MSAs) and HMMs associated with different enzyme functions. Next, we apply an 'Evolutionary Footprinting' (EF) approach to identify residues in an enzyme family that can discriminate between sequences having the specified EC number and those sequences with other functions (different EC numbers or non-enzymes). Structural information is not required in any stage of the method.

Our EF method tackles the problem of selecting the functionally discriminating residues (FDRs) even for enzyme families with a high number of conserved residues, by using two sets of MSAs. The first is the MSA of the classified family (homofunctional MSA). The second and larger MSA is obtained by running the associated HMM against the non-redundant combination of the Swiss-Prot (24), TrEMBL (24) and KEGG (25) sequence databases (heterofunctional MSA). The sequences in this (possibly) heterofunctional MSA have a close evolutionary relationship with those in the homofunctional MSA, but they do not necessarily have the same function. Then, for each position in the HMM, we compute a combined conservation score based on both MSAs. Finally, starting from the position with the highest conservation score in the HMM, we gradually select residues that specifically recognize the true members of the enzyme family.

The EF method is based on two hypotheses. First, the divergence of enzyme function is achieved by the modification of an existing active site in the ancestral protein (26). In other words, homologous proteins might share the same architecture of the active site, although they employ different residues to perform different functions. Thus, in a homofunctional or heterofunctional MSA, the active site positions should have lower entropy than the others. Second, it is also possible that the divergence of enzyme function could be achieved by the formation of a new active site at a different position in the sequence that is unrelated to the active site of the ancestral protein (27). If we were to exclusively consider the heterofunctional MSA, we might find positions that are only conserved in a set of sequences functionally unrelated to the original enzyme family. To avoid this pitfall, we focus on those positions that are also conserved in the original family.

By applying the EF method, we show that the number of FDRs in a given enzyme family is significantly reduced, compared with FDRs identified by only using homofunctional MSAs. Moreover, we show that the FDRs identified by the EF method have a strong correlation to known functionally important residues, such as enzyme active site residues. Besides the CHIEFc family based EF approach to identify FDRs, we have developed and benchmarked in a jackknife test, three other enzyme function inference methods based on (i) pairwise sequence comparison using a family specific SIT, (ii) identification of FDRs in Pfam families, and (iv) Multiple Prosite pattern recognition. The results of the jackknife test show that the CHIEFc family based EF approach to identify FDRs can extend the limit of pairwise comparison to remote, yet functionally similar sequences and outperforms the other compared methods. We further improve the accuracy of the Pfam-based and the Prosite-based methods by applying FDR recognition on multiple rather than single Pfam domains and by using only highly specific Prosite patterns, respectively. We have combined these four methods to develop EFICAz, an automatic engine for large-scale enzyme function inference that significantly outperforms any individual method. In the jackknife test, EFICAz shows an accuracy of 92% and a sensitivity of 82% for predicting four EC digits in testing sequences that are <40% identical to any member of the corresponding training set. By way of application, we have carried out the genome-wide enzyme function inference of

the *Escherichia coli* proteome using EFICAz. In total, besides 881 enzyme sequences already included in our database, we assign four (three) digit EC numbers to additional 132 (234) sequences, in contrast to 45 (277) additional sequences annotated in the KEGG database.

METHODS

Collection of enzyme sequences

All databases containing biological information derived from computational analysis are error-prone (28), and even annotations based on experimental evidence have been refuted in a number of cases (29). However, it is reasonable to use as a reference for functional annotation a database manually curated by numerous experts, such as Swiss-Prot (24). Thus, we adopt the EC number assignments for Swiss-Prot entries provided by the Enzyme database (23), as the standard of truth for our study. Release 33 of the Enzyme database of October 2003 includes 4208 enzymatic reactions; 1861 of them are associated with sequences, corresponding to 41 225 sequences in total. Among those sequences, 2375 are annotated as 'fragment' by Swiss-Prot (Release 42); to avoid ambiguity, they are not included in our analysis. Thus, we obtain 38 850 sequences to construct the function-oriented enzyme family database; they correspond to 1861 distinct full four EC digits and 188 first three EC digits, respectively.

CHIEFc and MSA construction

We divide the 38 850 enzyme sequences into 1861 and 188 EC groups according to their full four EC digits and the first three EC digits, respectively. Enzyme sequences associated with multiple EC numbers are included in every corresponding EC group. For each EC group, we apply the procedure depicted in Figure 1. First, after classification by EC number, we further classify the sequences by evolutionary relationship, employing complete linkage clustering (30) to divide them into subgroups using a cut-off of 30% sequence identity. With complete linkage clustering, all objects in a cluster must be similar to one another above a certain threshold, and no object can be in more than one cluster. Pairwise sequence comparisons are carried out by the Myers/Miller (MM) global alignment algorithm (31), under which the amount of memory required to align two sequences becomes a linear rather than a quadratic function of the lengths of the sequences. Then, we construct an MSA using Clustal W (32) based on the subgroup that maximizes the 'number of sequences/average sequence identity' ratio (seed subgroup). The rationale behind this criterion to select the seed subgroup is that a larger and more diverse set of sequences will allow the construction of more robust HMMs in the next step of the procedure. To reduce redundancy in the MSA, only sequences below 85% pairwise sequence identity are used. From this MSA, we build an HMM using the 'hmmbuild' program obtained from the HMMER software package (21) using the global option, followed by application of the 'hmmcalibrate' program to tune the statistical *E*-value for improvement of search sensitivity. We run the 'hmmsearch' program with this HMM to search against all the sequences in the original EC group (1315 and 2813 sequences for the most populated four EC digit and three EC digit groups, respectively) and select sequences with an *E*-value <0.01. Using the HMM as a template, we obtain an HMM-based MSA by grouping the alignments of the selected sequences, which is used to build a new HMM. We iterate the HMM construction and search process until either sequence convergence (i.e. no new sequences with an *E*-value <0.01 are found) or loss of residue conservation (i.e. no 'potential active site' residues are completely conserved in an MSA) is reached. Here, a 'potential active site' residue is defined as any charged or polar amino acid plus phenylalanine, i.e. any amino acid type except Gly, Pro, Ala, Val, Leu, Met and Ile, because these residues are not observed in any active site, as annotated by Swiss-Prot (data not shown). The obtained HMM and MSA define a family for the EC group under analysis. After removing subgroups whose sequences have all been added to the seed subgroup, we select the next subgroup with the highest 'number of sequences/average sequence identity' ratio and repeat the above-described procedures to define families until all the sequences in the EC group are classified. In the end, for each EC group, we obtain a number of families whose sequences all have the same EC number and are clearly evolutionarily related. Moreover, for families with enough sequences, we obtain an MSA that has at least one completely conserved 'potential active site' residue and also a corresponding

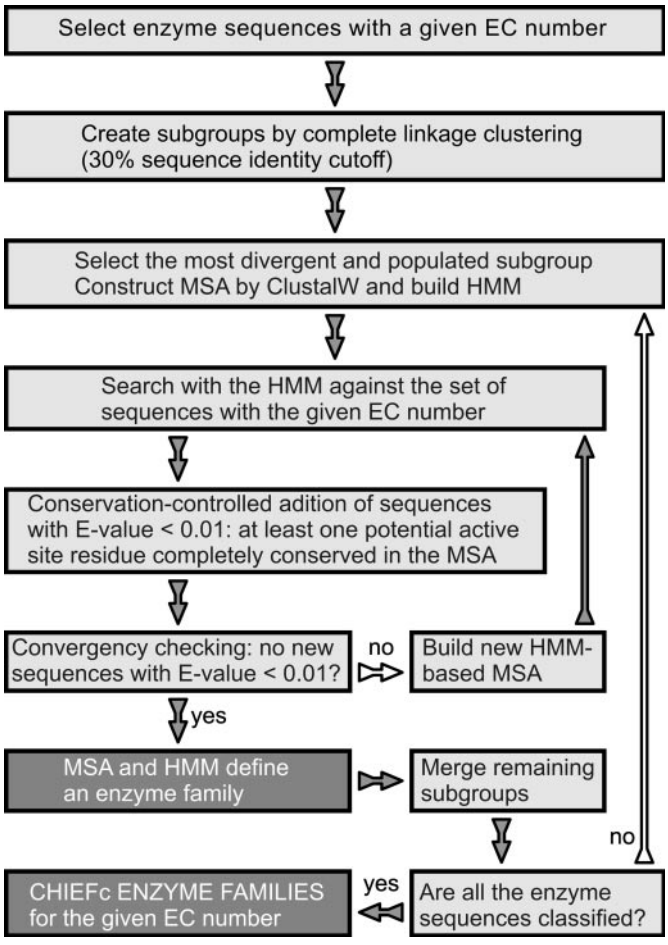

Figure 1. Overview of the procedure to build enzyme families by CHIEFc.

HMM. We name the protocol CHIEFc, for Conservation-controlled HMM Iterative procedure for Enzyme Family classification.

EF method to select FDRs in enzyme families

We have developed an EF method to select FDRs in enzyme families derived by CHIEFc. The EF method requires two MSAs for each enzyme family: a homofunctional MSA and a heterofunctional MSA. The homofunctional MSA is the original MSA associated with the CHIEFc family. The heterofunctional MSA is obtained by searching with the HMM of the family (E-value <0.01) against a non-redundant combination of the Swiss-Prot (Release 42), TrEMBL (Release 25) and KEGG (Release 28) databases (1 383 915 sequences), and grouping the alignments of the selected homologous sequences. Homologous sequences in the heterofunctional MSA might have functions different from the one associated to the homofunctional MSA. Since both types of MSAs are based on the same HMM, we calculate a combined conservation score for each position x in the corresponding HMM: $C(x)_{\text{combined}} = C(x)_{\text{homofunctional}} + C(x)_{\text{heterofunctional}}$. Here, $C(x) = e^{-H(x)}$, and $H(x)$ is the Shannon entropy: $H(x) = \sum_i -P_i(x)\log P_i(x)$. $P_i(x)$ is the observed frequency of amino acid type i in position x of the corresponding MSA. Based on our hypothesis, mutations of functional residues might be frequent in the heterofunctional MSA because of functional divergence. In contrast, they must be rare in the homofunctional MSA because of the maintenance of the same function. Thus, we use 20 amino acid types to calculate the sequence entropies in the homofunctional MSA. However, for the heterofunctional MSA, we use 10 amino acid groups derived from the BLOSUM62 mutation matrix (33); they are A, G, P, C, (I, L, M, V), (D, E), (S, T), (N, H), (F, W, Y), and (R, Q, K), respectively. Then, we rank the conservation degree of each position by their Z-score: $Z_C = C - \text{sd}(C)/\overline{C}$. Here, $\overline{C}$ and sd(C) are the average and the SD of the conservation degree, respectively.

Sequences in the heterofunctional MSA whose biochemical functions differ from the one of the original enzyme family play a critical role in the EF method, although not all the sequences in this MSA have available functional annotation. We exploit the functional information provided by Swiss-Prot to prepare three sets of sequences: 'enzymes', 'incomplete enzymes' and 'non-enzymes'. The 'enzymes' set includes the 38 850 enzyme sequences selected as described above. The 'incomplete enzymes' set comprises 8495 Swiss-Prot sequences that have missing EC digits and do not contain 'hypothetical', 'putative', 'probable', 'by similarity' or 'by homology' in the DE (definition) line, or that have complete EC numbers but are annotated as 'fragment'. The 'non-enzymes' set contains 25 096 sequences that fulfill the following conditions: (i) they lack EC number information, (ii) they do not contain 'hypothetical', 'putative', 'probable', 'by similarity' or 'by homology' in the DE line, (iii) they do not contain any enzyme name in the KW (keyword) line, and (iv) they can be detected with E-value <10 by running a three-iteration PSI-BLAST search (7) of any sequence in the 'enzymes' set against the complete Swiss-Prot database (135 694 sequences). By imposing the fourth condition, we mimic the real annotation situation, in which, typically, a database search is first performed, and then a criterion is applied to identify the true hits. The list of Swiss-Prot sequences included in the 'enzymes', 'incomplete enzymes' and 'non-enzymes' sets can be found at http://www.bioinformatics.buffalo.edu/eficaz/index.html.

We classify the sequences in the heterofunctional MSA as 'true' members, 'false' members or 'unknown' members of the enzyme family. A sequence is considered a 'true' member if it is included in the 'enzymes' set and its EC number matches with one of the enzyme family, and it is considered a 'false' member if (i) it is included in the 'incomplete enzymes' or 'enzymes' sets and its (partial) EC number does not match with one of the enzyme family of interest, or (ii) it is included in the 'non-enzymes' set. The rest of the sequences for which there is no functional annotation available are considered 'unknown'.

In the EF method, we evaluate the different positions of the homo and hetero functional MSAs, following a descending order of conservation Z-score (Z_C), to seek the minimal set of residues that can discriminate 'false' from 'true' members of the enzyme family in the heterofunctional MSA. We restrict the analysis to positions in which a residue of the 'potential active site' type is conserved in at least 50% of the sequences of the homofunctional MSA. This residue (and other amino acid types in its BLOSUM62 group, if they appear in the analyzed position of the homofunctional MSA) is considered a tentative FDR. Thus, conservative substitutions (as defined by the BLOSUM62 group) observed in the analyzed position of the homofunctional MSA are considered valid alternatives for this tentative FDR. Sequences containing or lacking this/these amino acid type/s in the corresponding position of the heterofunctional MSA are classified as 'positives' or 'negatives', respectively. Then, we evaluate the functional discrimination ability of the residue/s in the given position by calculating (i) prediction accuracy = (true positives)/(true positives + false positives), (ii) prediction sensitivity = (true positives)/(true positives + false negatives), and (iii) Matthews correlation coefficient (34), $\text{MCC} = (\text{TP}\times\text{TN} - \text{FP}\times\text{FN})/\sqrt{[(\text{TP}+\text{FN})(\text{TP}+\text{FP})(\text{TN}+\text{FP})(\text{TN}+\text{FN})]}$. Here, TP is true positives, TN is true negatives, FP is false positives and FN is false negatives.

If the accuracy is 100%, i.e. the residues can discriminate all 'false' members, we promote them from tentative to validated FDRs for the enzyme family, and stop evaluating other positions. If not, we keep these positions and continue to the next one according to a descending order of Z_C. The procedure finishes when we find the minimal set of FDRs corresponding to a prediction accuracy of 100%. We also account for a slight fraction of functional residue misalignments that might happen in true sequences. Thus, sequences in the heterofunctional MSA having a mismatch in a single discriminating position that can be fixed by allowing one residue shift are still considered 'positives', provided that the FDRs are exactly matched in all the other positions.

In 7% of the enzyme families, it is not possible to achieve 100% accuracy even when all the positions of the MSAs are analyzed. In those cases, we stop when the Z_C of the position under analysis drops below zero, and then we trace back to find the minimal set of FDRs that maximizes the product of prediction accuracy and the MCC (this product ranges from −1 to 1, as the MCC does, but emphasizes the accuracy of the predictions). For each of these families, we record all the

mismatched EC numbers, i.e. the EC numbers of the false positives identified by the FDRs. Thus, a list of mismatched EC numbers is associated with each enzyme family that does not achieve 100% accuracy.

Enzyme function inference by CHIEFc family based FDR recognition

We denote the application of the CHIEFc family based FDRs obtained using the EF method to infer the EC number of a query sequence as 'CHIEFc family based FDR recognition'. The CHIEFc family based FDR recognition is a two-step procedure. The first step is the scanning of the whole enzyme family library to detect the families that predict the query sequence as their putative member. The query sequence is provisionally predicted as member of a given family if it satisfies two conditions: (i) it is recognized by the corresponding HMM with E-value <0.01 and (ii) it matches the set of associated FDRs for the specified family. Let us call f an enzyme family that predicts the query sequence as its member in the first step of the procedure. After this first step, a given query sequence can be provisionally assigned to have more than one EC number. The second step is the simultaneous analysis of the provisional predictions to decrease the number of erroneous multiple assignments. Each predicted EC number associated to an enzyme family f whose FDRs show 100% accuracy in discriminating known functionally similar and dissimilar sequences is included in the final prediction for the given query sequence. For each enzyme family f that does not achieve 100% accuracy, we retrieve its corresponding list of mismatched EC numbers and check whether any of them coincides with any of the EC numbers provisionally assigned to the query sequence in the first step. If that is the case, the EC number associated to f is not included in the final prediction for the given query sequence. By this procedure, EC numbers that we find difficult to be accurately distinguished from each other are not assigned to the same query sequence. Therefore, the accuracy is improved and the ability to detect true multienzymes, i.e. sequences having multiple enzymatic functions, is not affected.

Relationship between FDRs and annotated active site residues

We use the functional annotation from Swiss-Prot to understand the biological significance of the FDRs detected using the EF method. We first select Swiss-Prot sequences whose FT (feature) lines contain the ACT_SITE (active site residue) key name, but lack the keywords 'by similarity', 'potential' or 'probable' in the description field. By checking the 38 850 sequences in the 'enzymes' set, we obtain 1859 sequences with annotated active site residues, which correspond to 367 and 299 four EC digits and three EC digits CHIEFc enzyme families, respectively. For each of these two sets of enzyme families, we determine F_{obs}, the fraction of families whose FDRs include at least one annotated active site residue. Then, we perform a 1 000 000-repetition simulation to determine $\overline{F}_r$, the expected value of F_{obs} when the FDRs are selected at random instead of using the EF method. Each repetition consists of two steps: (i) we randomly select n_i out of m_i residues for each enzyme family i, where n_i is the number of FDRs for the family i and m_i is the total number of residues of the 'potential active site' type in the MSA of family i, and (ii) we calculate F_r, the fraction of families whose n_i randomly selected residues match at least one annotated active site residue of the family i. Finally, we average the 1 000 000 F_r values to obtain $\overline{F}_r$. The increase in F_{obs} over $\overline{F}_r$ is assessed as a Z-score: $Z_F = F_{obs} - \mathrm{sd}(F_r)/\overline{F}_r$, where $\mathrm{sd}(F_r)$ is the SD of F_r.

Enzyme function inference by pairwise sequence comparison

For enzyme function inference by pairwise sequence comparison, we derive an enzyme family-specific SIT for each CHIEFc family in our library, according to our previous work (10). The SIT for a given enzyme family is a discrete sequence identity value t that can vary from 20 to 90%, with a 10% increment. It is defined as the minimum value that satisfies the following two conditions: (i) on comparing with every member of the family, there are no sequences in Swiss-Prot having a different function (enzyme or non-enzyme) with sequence identity above t, and (ii) there is at least one pair of family members with sequence identity in the $[t, t + 10]$ interval. When there is only one sequence in a CHIEFc family, we define the corresponding SIT as 60 and 40% for four EC digits and three EC digits families, respectively. These values allow an average of 90% accuracy for enzyme functional inference purely based on pairwise sequence comparison (10). We predict a query sequence as a member of a given family when it has a sequence identity to any member of the family greater than the corresponding SIT. We denote this approach as 'CHIEFc family specific SIT evaluation'.

Enzyme function inference by Prosite pattern recognition

For 'Multiple Prosite pattern recognition', we first prepare a list of Prosite (Release 18) (11) patterns/profiles associated with enzyme functions. We run the 'ps_scan.pl' program (obtained from the Prosite distributed package and used with default settings) to search the enzyme-function related patterns/profiles against the 'enzymes' set of sequences. Then, we analyze all the matches and define a combination of patterns/profiles to detect each enzyme function. We exclude 67 patterns that do not have functional discrimination ability because they are present in sequences belonging to more than 20 different enzymes. The complete list of excluded Prosite patterns can be found as Supplementary Material.

We have modified the Multiple Prosite pattern recognition approach to increase its accuracy. For each combination of Prosite patterns associated with enzyme functions, we carry out an extensive search against Swiss-Prot sequences. Only those combinations of Prosite patterns that are specific for a unique EC number are selected for functional inference. For example, sequences related to EC.1.1.1.1 are associated with four combinations of Prosite patterns: (i) PS00044 and PS00059, (ii) PS00059, (iii) PS00060 and PS00913, and (iv) PS00913. However, after extensive search, only a combination of PS00044 and PS00059 is found to be specific to EC.1.1.1.1. We denote this modification of the Multiple Prosite pattern recognition approach as 'High specificity multiple Prosite pattern recognition'. A list of Prosite pattern

combinations specific for four and three digit EC numbers is available as Supplementary Material.

Enzyme function inference by Pfam family based FDR recognition

The EF method is not restricted to only CHIEFc families to identify FDRs. In fact, we also apply the EF method to Pfam domains (21) that are associated with enzyme functions to identify the FDRs corresponding to specific enzyme functions. We denote this approach as 'Single Pfam family based FDR recognition'.

To construct the Pfam domain based enzyme family library, we first carry out an extensive search with the sequences in the 'enzymes' set (38 850 sequences) against 'Pfam_ls', the Pfam HMM library of global alignment models (Release 9). For the search, we employ the 'hmmpfam' program from the HMMER software package, using an *E*-value <0.01 to detect a hit. Of 5724 HMMs in the 'Pfam_ls' library, 1623 are associated with enzyme sequences. Thus, for each enzyme function, we have a set of single Pfam domain based MSAs (the homofunctional MSAs) and the corresponding HMMs. Then, we obtain the heterofunctional MSAs by searching the HMM of the corresponding Pfam domain against the non-redundant combination of the Swiss-Prot, TrEMBL and KEGG databases. Next, we apply the EF method to obtain FDRs in every Pfam domain for each of its corresponding enzyme functions. The application of the 'Single Pfam family based FDR recognition' approach for enzyme function inference of a query sequence is performed the same as that in the 'CHIEFc family based FDR recognition' approach.

We have modified the 'Single Pfam family based FDR recognition' approach to increase its accuracy by using a combination of domains rather than a single one. We denote this modification as 'Multiple Pfam family based FDR recognition'. Since a significant fraction of the enzymes are multidomain proteins, very often an enzyme sequence is classified as a member of different Pfam families, even when only one domain might be involved in catalysis. Even more, enzyme sequences sharing the same EC number may be associated with different combinations of Pfam domains. In this approach, we first collect all different combinations of Pfam domains observed in enzyme sequences with a given EC number. Then, we remove those combinations that include at least one domain whose corresponding Pfam family associated FDRs do not achieve 100% accuracy. Thus, we obtain a list of the observed combinations of accurate Pfam families that concurrently detect the enzyme sequences of a given EC number. To assign an EC number to a query sequence, we require the sequence to be detected as 'positive' by all the Pfam families that are included in at least one of the combinations of domains in the list.

EFICAz

In EFICAz, we combine the predictions of four independent enzyme function inference methods: (i) CHIEFc family based FDR recognition, (ii) CHIEFc family specific SIT evaluation, (iii) High specificity multiple Prosite pattern recognition, and (iv) Multiple Pfam family based FDR recognition. Since the four methods are highly accurate and their predictions do not overlap completely between each other, EFICAz infers a particular enzyme function when one or more of the four component methods predicts that enzyme function. Furthermore, by requiring the consensus of two or more components of EFICAz to predict a particular enzyme function, we can increase the confidence of the predictions. We term this set of predictions, the 'higher confidence' subset of EFICAz. We have implemented EFICAz as a Web server at http://www.bioinformatics.buffalo.edu/eficaz/.

Benchmark of enzyme function inference by jackknife test

We carry out a jackknife test to benchmark different approaches for enzyme function inference. We prepared training and testing supersets of sequences based on the 'enzymes', 'incomplete enzymes' and 'non-enzymes' sequence sets. Because some enzyme functions have only a few sequences, for benchmark purposes, we select those four digits EC numbers that have more than 10 sequences each (684 EC numbers), which correspond to 34 823 sequences in total (90% of the sequences in the 'enzymes' set). About 3% (1017 sequences) of these 34 823 enzyme sequences are multienzymes, mostly linked to two EC numbers (836 sequences). We randomly select 80% of the 34 823 enzyme sequences to be included in the training superset; the remaining 20% is included in the testing superset. Similarly, we randomly select 80 and 20% of the 25 096 sequences in the 'non-enzymes' to be included in the training and testing supersets, respectively. All the 8495 sequences in the 'incomplete enzymes' set are included in the training superset.

The training set for a given EC number is defined as the collection of enzyme sequences from the training superset that belongs to that EC number. The testing set for a given EC number is composed of those sequences that can be detected with *E*-value <10 by running a three-iteration PSI-BLAST (7) search of each member of the corresponding training set against the sequences in the testing superset. We train and test the following enzyme function inference approaches: (i) CHIEFc family based FDR recognition, (ii) CHIEFc family specific SIT evaluation, (iii) Multiple Prosite pattern recognition, (iv) High specificity multiple Prosite pattern recognition, (v) Single Pfam family based FDR recognition, (vi) Multiple Pfam family based FDR recognition, and (vii) EFICAz.

We reduce the bias resulting from enzyme families with a large number of sequences by calculating the average performance per EC number (accuracy, sensitivity and MCC). Furthermore, to reduce the bias resulting from the abundance of closely related homologous sequences, for each EC number, we first select the testing sequences whose sequence identities to any member of their corresponding training sets are not higher than a given limit. Then, based on the selected testing sequences, we calculate the performance of each method for each of those EC numbers. Finally, we report the accuracy, sensitivity and MCC values for each method, at different levels of maximal testing to training sequence identity, averaged per EC number. The reported values are the averages of three repetitions of the jackknife analysis.

Genome-wide enzyme function inference on *E.coli* genome

The sequences of the protein coding genes for the genome-wide enzyme function inference on the *E.coli* K-12 proteome

by EFICAz are retrieved from the KEGG database (Release 28). We compare our prediction results with the annotation of KEGG (Release 28) and Swiss-Prot (Release 42) databases. We select KEGG database for comparison instead of other *E.coli* annotation databases, because KEGG collects and combines information from several public sources (including organism-specific databases) that is subjected to internal reannotation for linking to metabolic pathways and EC number assignment (25). Thus, KEGG provides EC number information for many genes that are annotated with enzyme descriptions in other databases, but lack explicit EC numbers.

RESULTS

Enzyme family classification

By applying CHIEFc to the sequences in the Enzyme database, we have obtained 2944 four EC digits and 2054 three EC digits enzyme families. Most four EC digits enzyme types correspond to unique four EC digits families (1479 out of 1861 different four EC digits), while only 36 out 188 three EC digits enzyme types correspond to unique three EC digits families. The detailed enzyme classification results can be found at http://www.bioinformatics.buffalo.edu/eficaz/.

In 20% of the cases, a four EC digit enzyme is related to multiple families. This is due to three different reasons, the first being convergent evolution. For example, we classify EC 1.1.1.1 (Alcohol dehydrogenase) into three families, which is consistent with the Pfam classification of this enzyme: 'ADH_zinc_N', 'adh_short' and 'Fe-ADH'. These three families do not share any sequence or structural similarity and do not interact with each other. The second reason is related to our family construction procedure. We require at least one 'potential active site' residue to be completely conserved in the MSA associated with an enzyme family. Accordingly, if merging two families would result in the loss of conservation, then we separate them into different families. For example, EC 2.7.4.3 (Adenylate kinase) is classified into two families, which are evolutionarily related and have similar structure, but combining them would result in loss of conservation. Third, enzyme complexes contribute to the family multiplicity because different subunits annotated with the same EC number can have unrelated sequences. For example, EC 3.6.3.14 (H^+-transporting ATPase) is a large multi-subunit complex, classified into 70 families in our database, probably because a combination of the above-mentioned factors.

Among the classified families, 1890 four EC digits families and 1392 three EC digits families have an associated MSA and HMM. These families correspond to 1371 four EC digits and 179 three EC digits, respectively. The remaining families lack MSAs because there is only one sequence in the family, or the sequences in the family are too closely related. The following analyses are based on those families with associated MSAs.

Divergence of enzyme functions

The EF method for the selection of FDRs in enzyme families requires two MSAs: a homofunctional MSA (the original MSA associated to the family) and a heterofunctional MSA, obtained by searching with the HMM of the family against a non-redundant sequence database (see Methods). As its name indicates, a heterofunctional MSA usually includes sequences with different functions. To investigate the extent of functional divergence in heterofunctional MSAs, we check the annotation of Swiss-Prot sequences in the heterofunctional MSA of each enzyme family and plot the cumulative relative distribution of the number of different functions in heterofunctional MSAs (Figure 2A and B). Here, we mainly focus on divergence of enzyme functions; therefore, the functions of all non-enzymes are described with a single functional category 'non-enzyme'. The heterofunctional MSAs corresponding to 64% of the four EC digits families include two or more types of functions, with some alignments associated with more than 50 different functions (Figure 2A). This indicates that functional divergence occurs frequently among homologous enzyme sequences. The heterofunctional MSAs of three EC digits families show less divergence, with 43% having two or more types of function (Figure 2B).

Figure 2C and D show the number of different functions in heterofunctional MSAs versus the average pairwise sequence identity (sequence diversity) of the homofunctional MSAs, for four EC digits and three EC digits families, respectively. The more diverse an HMM profile is, the more distantly related sequences it will find. Since functional divergence occurs more frequently among distantly related sequences, a positive correlation might be expected between the diversity of the HMM profile (i.e. diversity of the homofunctional MSA) and the number of different functions in the corresponding heterofunctional MSA. However, we find that the median number of different functions in heterofunctional MSAs is independent of the sequence diversity of homofunctional MSAs (Figure 2C and D). For example, the heterofunctional MSA of the *trans*-cinnamate 4-monooxygenase family (EC 1.14.13.11) includes sequences from more than 20 types of EC numbers, although the corresponding homofunctional MSAs has an average pairwise sequence identity of above 80%. These findings further stress the difficulty of inferring functions from sequence similarity and stimulated us to develop a functional inference approach based on the identification of FDRs.

FDRs identified by the EF method

We have applied the EF method to select FDRs for all CHIEFc families with MSAs. The average accuracy and sensitivity of the method applied to four EC digits enzyme families is 99 and 95%, respectively, and for three EC digits enzyme families, it is 99 and 96%, respectively (data not shown). This suggests that the selected residues not only can discriminate 'false' function from 'true' function but also can identify almost all 'true' members of the family, allowing the application of our approach to enzyme function inference. The median number of FDRs for four and three EC digits enzyme families is 9 and 5, respectively.

In Figure 3, we plot the average number of FDRs per family against intervals of average sequence identity of homofunctional MSAs. If we rely only on homofunctional MSAs to select FDRs, we can observe a strong positive correlation between the number of selected residues and the sequence identity of the homofunctional MSA. In contrast, using heterofunctional MSAs together with homofunctional MSAs significantly reduces the number of needed residues, and such correlation is now greatly diminished. For example, even for families whose homofunctional MSAs have an average sequence

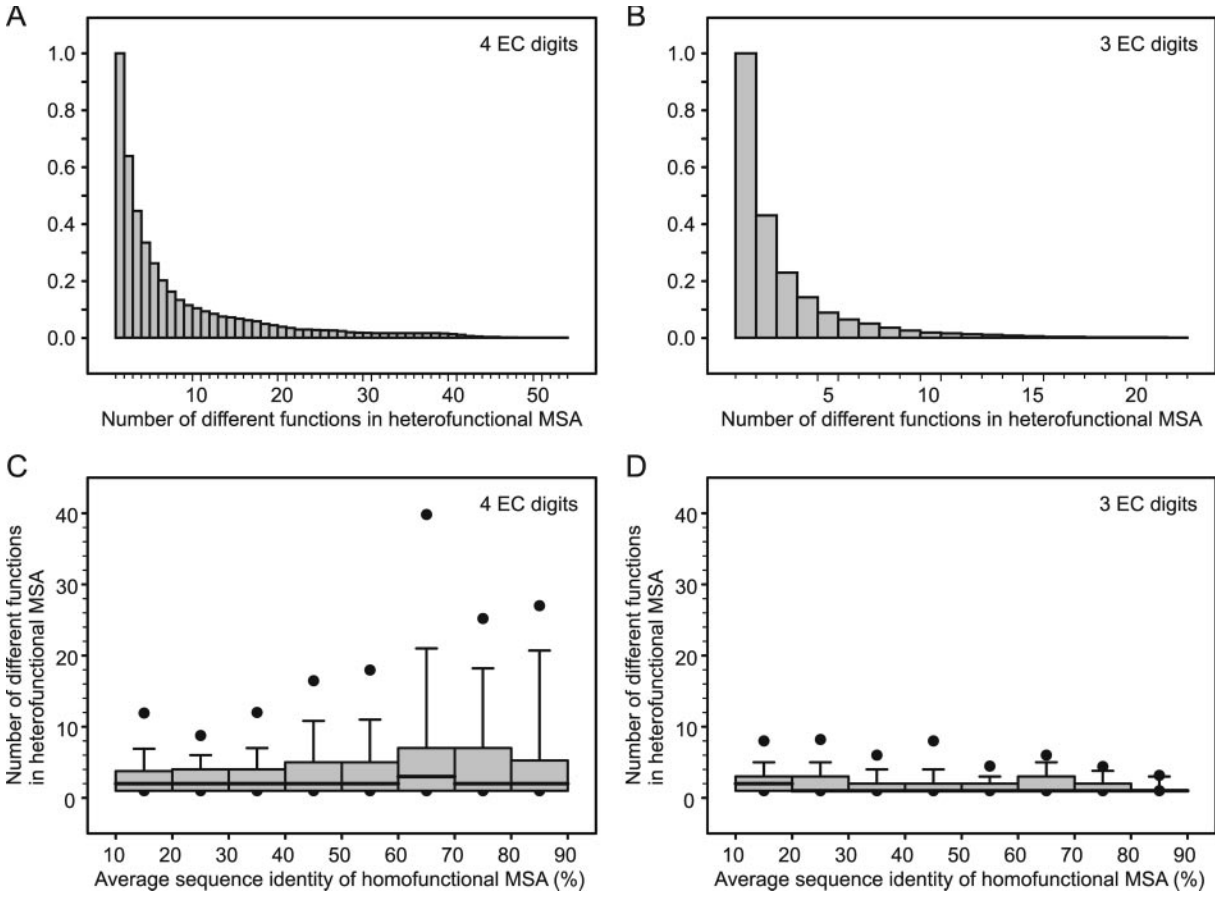

Figure 2. Divergence of enzyme functions. Cumulative relative distributions of the number of different enzyme functions in the heterofunctional MSAs associated with four EC digits (**A**) and three EC digits CHIEFc enzyme families (**B**). Box-and-whisker plots showing the distributions of the number of different enzyme functions in heterofunctional MSAs versus the average pairwise sequence identity (sequence diversity) of the corresponding homofunctional MSAs associated with four EC digits (**C**) and three EC digits CHIEFc enzyme families (**D**). From top to bottom, the statistics represented in the box-and-whisker plots are 95th percentile (black circle), 90th percentile (whisker, top), 75th percentile (box, top), median (thick line), 25th percentile (box, bottom), 10th percentile (whisker, bottom) and 5th percentile (closed circle).

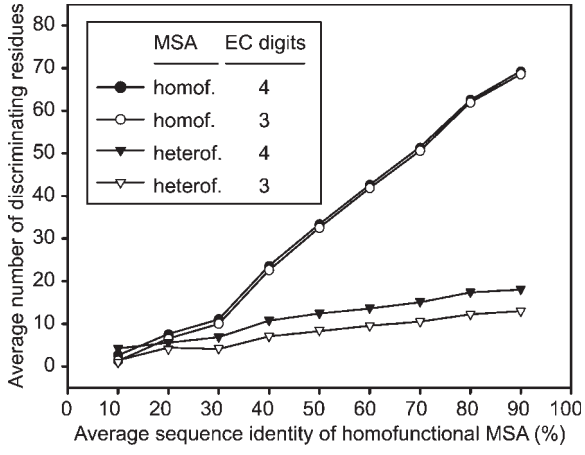

Figure 3. Average number of FDRs in CHIEFc enzyme families versus the average pairwise sequence identity of their corresponding homofunctional MSAs. The FDRs are selected based on the conservation score of either homofunctional MSAs alone (closed circles and open circles, to discriminate four EC digits and the first three EC digits, respectively), or both homofunctional MSAs and heterofunctional MSAs (closed triangles and open triangles, to discriminate four EC digits and the first three EC digits, respectively).

identity of above 80%, the average number of selected residues by the EF method is below 20, in contrast to more than 70 when homofunctional information alone is used. Using both types of alignments, the number of residues needed to discriminate three EC digits is less than that needed to discriminate four EC digits. In contrast, using only homofunctional MSA to select FDRs, we do not observe such a difference.

Correlation between functionally important and functionally discriminating residues

We stress that the EF method does not attempt to predict all functional residues needed for a specific function, but aims at selecting the minimal set of residues that can discriminate sequences with a 'false' function from those with a 'true' function. In certain cases, the selected residues may be fold determinant rather than functionally important. However, it is reasonable to expect an enrichment of functionally relevant residues in the set of FDRs. We investigate the correlation between functional importance and discrimination ability of the residues selected by the EF method, by analyzing F_{obs}, the observed fraction of CHIEFc families whose FDRs include at least one residue annotated as an active site in Swiss-Prot.

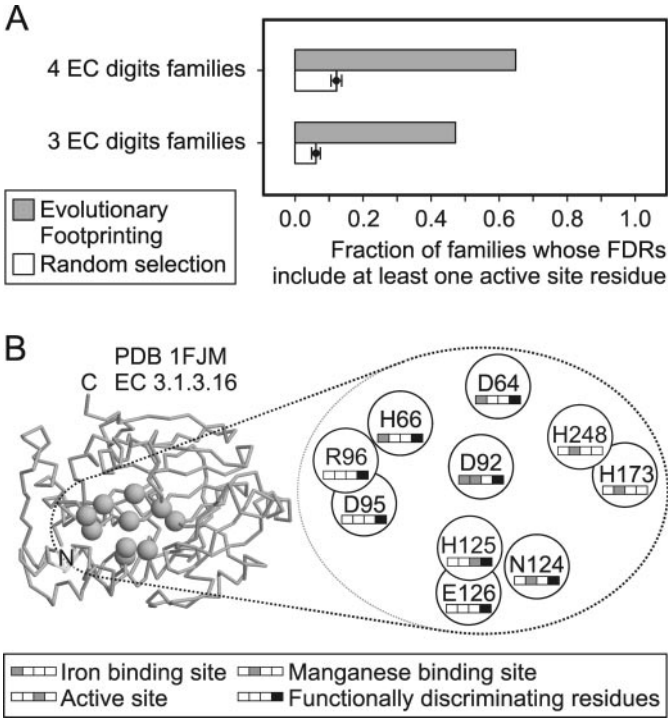

Figure 4. Correlation between functionally important residues and FDRs. (**A**) Fraction of CHIEFc families whose FDRs include at least one residue annotated as active site in Swiss-Prot. Two strategies for obtaining the FDRs are compared. The EF method (gray bars) and random selection (open bars, with error bars representing SD of the mean). (**B**) Functional annotation and spatial location of the FDRs for the phosphoprotein phosphatase CHIEFc family, mapped on the 3D structure of PDB entry 1FJM.

For 65% (47%) of the four (three) EC digits families, the EF method selects FDRs that include at least one active site residue, while the random procedure does so for only 12% (6%) of the families (Figure 4A). Thus, the FDRs for four (three) EC digits families are more than five (seven) times richer in active site residues than randomly selected residues, with a significance Z_f of 34 (32). Still, the FDRs for 35% (53%) of the four (three) EC digits families do not include active residues. This can be partially due to incomplete annotation. In fact, the majority of the families in this study have only one annotated active site residue in Swiss-Prot, which means some FDRs may be active site residues that are not yet annotated. The FDRs tend to be clustered in space and often include other functionally important residues. This tendency is exemplified in Figure 4B, where we show the functional annotation and the spatial location of the FDRs for the phosphoprotein phosphatase family (EC 3.1.3.16), mapped on the 3D structure of one of the representatives of this enzyme family, obtained from the entry 1FJM in the PDB (35). Five out of eight FDRs are annotated as functionally important, while two functionally relevant residues are not included in the set of FDRs (His248 and His173).

Jackknife test to benchmark enzyme function inference by different approaches

To benchmark the performance of the different enzyme function inference approaches listed in Table 1, we carry out a jackknife test (see Methods for a description of each approach). In Figure 5A–C, we compare the average accuracy, sensitivity and average MCC of the CHIEFc family based FDR recognition method (the application of the CHIEFc family based FDRs obtained using the EF method to predict the EC number of a query sequence) to the following approaches: (i) CHIEFc family specific SIT evaluation, (ii) Multiple Prosite pattern recognition, and (iii) Single Pfam family based FDR recognition. In Figure 5D–F, we similarly compare EFICAz to the following approaches: (i) CHIEFc family specific SIT evaluation, (ii) High Specificity multiple Prosite pattern recognition, (iii) Multiple Pfam family based FDR recognition, and (iv) CHIEFc family based FDR recognition. In Figure 5G–I, we compare the entire set of EFICAz predictions with a subset of higher confidence EFICAz predictions. All the shown results correspond to four EC digits predictions; the three EC digits predictions follow the same trends (data not shown).

Table 1. Enzyme function inference methods

(i)	CHIEFc family specific SIT evaluation
(ii)	Multiple Prosite pattern recognition
(iii)	High specificity multiple Prosite pattern recognition
(iv)	Single Pfam family based FDR recognition
(v)	Multiple Pfam family based FDR recognition
(vi)	CHIEFc family based FDR recognition
(vii)	EFICAz

[a]CHIEFc: Conservantion-controlled HMM Iterative procedure for Enzyme Family classification.

[b]SIT: Sequence Identity Threshold.

[c]FDR: Functionally Discriminating Residue.

[d]EFICAz: Enzyme Function Inference by Combined Approach.

In this benchmark, we evaluate the prediction performance according to the maximal testing to training sequence identity. The rationale is based on the observation that Swiss-Prot is not an evenly distributed database; most enzyme sequences in Swiss-Prot have at least one closely related enzyme sequence also included in the database, and their functions could be easily predicted by a simple pairwise comparison. On the other hand, as we have pointed out in our previous work (10), Swiss-Prot enzyme sequences are dominated by entries belonging to only a few enzyme functions. Therefore, we also reduce this bias by calculating the average performance per EC number. The global statistics obtained without adopting these normalization procedures to avoid both sources of bias would be optimistically misleading.

In Figure 5A, it can be observed that by using a family specific SIT, the accuracy of a pairwise sequence comparison can be close to 100%. However, the sensitivity of this method is only 21% for sequences whose maximal sequence identity to their training sequences is <30% (Figure 5B). In fact, this low sensitivity motivated us to develop the EF method in an effort to extend the limit of functional inference to distantly related sequences. The Multiple Prosite pattern recognition is not a good approach for enzyme functional inference; it has both the worst prediction accuracy (Figure 5A) and the worst prediction sensitivity (Figure 5B). In contrast, those methods based on FDR recognition by EF (Single Pfam family based FDR recognition and CHIEFc family based FDR recognition) show similar and improved accuracy-sensitivity tradeoffs, which is reflected in their MCCs (Figure 5C). Both methods can extend

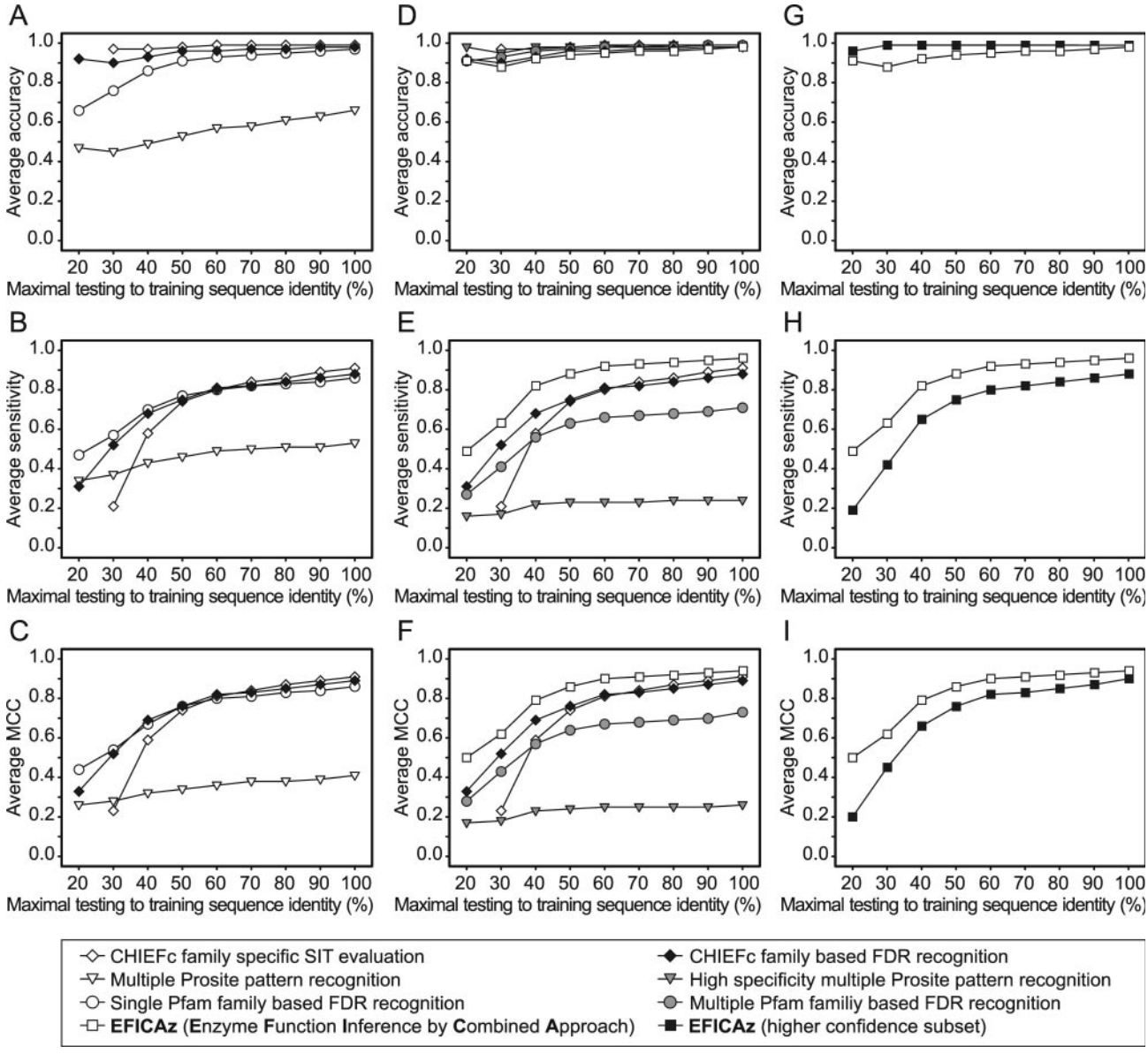

Figure 5. Benchmark of different enzyme function inference approaches by jackknife test. Accuracy (**A**, **D**, **G**), sensitivity (**B**, **E**, **H**) and Matthews Correlation Coefficient (**C**, **F**, **I**) values for different enzyme function inference methods, at different levels of maximal testing to training sequence identity, averaged per EC number. See Methods for a full description of the jackknife procedure. The plotted values are the averages of three repetitions of the jackknife analysis; the corresponding SDs are omitted for clarity, they range from 0.01 to 0.09, with a median value of 0.01.

the limit of enzyme function inference to distantly related sequences, with sensitivities (for 30% of maximal testing to training sequence identity) of 52% for the CHIEFc family based approach and 57% for the Single Pfam family based approach (Figure 5B). However, the CHIEFc family based approach achieves an accuracy of 90% compared with 76% of the Single Pfam family based approach (Figure 5A). Considering that our main goal is to maintain a high level of accuracy, the CHIEFc family based FDR recognition is the best performer of the four compared methods.

We have modified the Single Pfam family based FDR recognition and the Multiple Prosite pattern recognition approaches with the purpose of increasing their levels of accuracy (see Methods). Figure 5D shows significantly higher accuracies at 30% of maximal testing to training sequence identity, for the Multiple Pfam family based FDR recognition (93%) and the High Specificity multiple Prosite pattern recognition approaches (95%) when compared with their respective parent methods in Figure 5A (76 and 45%, respectively). We can observe in Figure 5D that the following four methods (i) CHIEFc family specific SIT evaluation, (ii) High Specificity multiple Prosite pattern recognition, (iii) Multiple Pfam family based FDR recognition, and (iv) CHIEFc family based FDR recognition have accuracies of at least 90%, independent of the level of maximal testing to training sequence identity. Although these methods show high accuracy, their sensitivities are more diverse, with the CHIEFc family based FDR recognition approach displaying the highest sensitivity of the

four, up to 60% maximal testing to training sequence identity. We unite the predictions of these four highly accurate approaches to create EFICAz. EFICAz has a significantly higher sensitivity than its constituents (Figure 5E), and yet retains an accuracy not worse than the least accurate method (Figure 5D). Consequently, EFICAz achieves the highest MCC of all the analyzed methods (Figure 5C and F).

The percentage of sequences predicted by only one of the four constituents of EFICAz at 40% maximal testing to training sequence identity is 20% for CHIEFc family based FDR recognition, 8% for CHIEFc family specific SIT evaluation, 5% for Multiple Pfam family based FDR recognition and 2% for High Specificity multiple Prosite pattern recognition. Thus, although CHIEFc family based FDR recognition is the main component of EFICAz, the four methods contribute to the high sensitivity of the combined approach. Furthermore, by requiring the consensus of two or more components of EFICAz to predict a particular enzyme function, we can achieve an accuracy of almost 100% independent of the level of sequence diversity (Figure 5G), with sensitivity (Figure 5H and E) and MCC (Figure 5I and F) comparable to those of the best component of EFICAz.

Consistent with the 3% of multienzymes included in the jackknife enzyme sequences (see Methods), about 2% of the sequences predicted by EFICAz are multienzymes, among which 70% have all the EC numbers correctly assigned, and 98% have at least one EC number correctly assigned. This result indicates that EFICAz is able to annotate multi-EC enzymes, which are often excluded from enzyme inference analysis.

Genome-wide enzyme function inference by EFICAz

We have employed EFICAz for the genome-wide enzyme functional inference on the *E.coli* K12 proteome. The total number of protein-coding genes in *E.coli* is 4289. Among them, 881 are included in the ENZYME database (Release 33) and annotated with four EC digits in the Swiss-Prot database (Release 42). Besides these genes, we predict 132 (234) additional genes with four (three) EC digits, with 49 (94) of those predictions resulting from the consensus of at least two components of EFICAz. In contrast, the KEGG database provides annotation for an additional 45 and 277 genes with four EC and first three EC digits, respectively. A comparison of our four EC digit predictions with the annotations in KEGG shows that (i) 18 genes are annotated by both approaches, with two genes showing disagreement between EFICAz and KEGG, (ii) 27 genes are annotated only by KEGG, and (iii) 114 genes are annotated only by EFICAz. Similarly, the analysis of the three EC digit predictions shows that (i) 104 genes are annotated by both approaches, with 21 genes showing mismatched annotations, (ii) 191 genes are annotated only by KEGG, and (iii) 130 genes are annotated only by EFICAz. A spreadsheet including all the EFICAz predictions for the *E.coli* genes that lack complete EC number annotation in the Swiss-Prot database is provided as Supplementary Material and is available at http://www.bioinformatics.buffalo.edu/eficaz/ecoli/index.html.

To analyze the differences between our results and KEGG annotations, we focus on the four EC digit assignments. Among the 18 genes predicted by both approaches, the two genes whose annotations disagree are b2979 (*glcD*) and b3583 (*sgbE*). KEGG annotation for b2979 is 'glycolate oxidase subunit glcD [EC 1.1.3.15]', although only GlcF, the iron–sulfur subunit of the *E.coli* glycolate oxidase complex is catalytic (36). In contrast, the EFICAz prediction for b2979 is EC 1.1.2.4 (D-lactate dehydrogenase). Indeed, the glycolate oxidase complex of *E.coli* can act as a D-lactate dehydrogenase (37), and it has been suggested that GlcD could be responsible for this activity (36). The KEGG annotation for b3583 is 'probable sugar isomerase sgbE [EC 5.1.3.4]', while EFICAz predicts it as both EC 5.1.3.4 (L-ribulose-phosphate 4-epimerase) and EC 4.1.2.17 (L-fuculose-phosphate aldolase). Although the Swiss-Prot annotation for b3583 is 'probable sugar isomerase [EC 5.1.-.-]', it has been shown that SgbE catalyzes the L-ribulose-phosphate 4-epimerase reaction (38). Thus, in this case, the EC 4.1.2.17 assignment by EFICAz appears to be a false positive.

We have investigated the reasons as to why EFICAz cannot predict the 27 gene products that are only annotated with four EC digits by KEGG. We distinguish three general cases. First, the gene product is not recognized by any of the CHIEFc or Pfam family HMMs associated with the EC number assigned by KEGG. In most of these cases, KEGG incorrectly assigns EC numbers to non-catalytic subunits of enzyme complexes (e.g. b4348, a non-catalytic subunit of the type I site-specific deoxyribonuclease complex, is annotated as EC 3.1.21.3 by KEGG). Second, the gene product is recognized by an HMM associated with the EC number assigned by KEGG, but not all the corresponding FDRs are matched. In some of these cases, KEGG assigns a four digit EC number that we find is difficult to distinguish from other/s (at least the conservation of a large number of FDRs is satisfied), while EFICAz assigns the corresponding partial three digit EC number (e.g. KEGG annotates b0600 as EC 2.6.1.1, which we find difficult to distinguish from EC 2.6.1.57). Third, the EC number assigned by KEGG is only related to the CHIEFc or Pfam families with only one sequence (therefore, without an associated HMM), and the default SIT of 60% happens to be too conservative to identify the gene by the 'CHIEFc family specific SIT evaluation' component of EFICAz. For example, b0969 is annotated in KEGG as a 'putative, sulfite reductase [EC 1.8.99.3]'. Its closet homolog in our enzyme database is DSVC_DESVH, which belongs to a CHIEFc family for EC 1.8.99.3 with only one sequence. Since the sequence identity between b0969 and DSVC_DESVH is only 36% (below the 60% SIT), EFICAz cannot infer the function of b0969.

Among the 114 genes with four EC digits predicted only by EFICAz, we find the following KEGG annotations: (i) partial EC numbers (or descriptions of enzymatic activities that can be converted to partial EC numbers) for 69 genes, with 54 of them matching the corresponding partial EC numbers of our annotations, (ii) descriptions not related to enzymes for four genes, and (iii) 'hypothetical protein' without a functional description for 38 genes. Both Swiss-Prot and KEGG suffer from the problem of annotation lag; therefore, it is possible that some gene products annotated in these databases as hypothetical proteins are true enzymes in light of recent evidence. In fact, by performing a literature search, we could validate the EFICAz predictions for two gene products annotated as hypothetical proteins by both KEGG and

Swiss-Prot: b0581 and b1119. The EFICAz predictions for b0581 (*ybdK*) and b1119 (*ycfX*) are EC 6.3.2.2 (Glutamate-cisteine ligase) and EC 2.7.1.2 (Glucokinase), respectively. These assignments are in agreement with very recently published articles that find YbdK to have γ-glutamyl:cysteine ligase activity (39), and YcfX to be a rudimentary glucokinase of ambiguous substrate specificity (40).

DISCUSSION

Enzyme family classification using CHIEFc is one of the reasons for the high accuracy of EFICAz

In our method of enzyme functional inference, EFICAz, we combine four methods among which two are based on FDR recognition. These two FDR recognition approaches differ from that one is based on enzyme families classified by CHIEFc, while another is based on multiple Pfam families. As shown in Figure 5D–F, when we consider the whole range of sequence diversity, the CHIEFc family based FDR recognition method is the most important component of EFICAz. Thus, one of the major reasons for the success of EFICAz in enzyme functional inference is attributed to CHIEFc, our enzyme family classification protocol. CHIEFc is deliberately designed for the purpose of enzyme functional inference and has the following advantages. First, by defining enzyme families based on both functional annotation and evolutionary relationship, we avoid the inclusion of 'false' members. This increases the specificity of the FDRs for their corresponding enzyme function. Second, by introducing a conservation-controlled selection of new sequences to be included in a CHIEFc family, we ensure the final alignment to have high quality. This enhances the reliability of the FDRs. Third, unlike Pfam families that are based on single domains, we start from whole enzyme sequences to classify them in CHIEFc families and generate sequence alignments. Considering that in general it is not known in advance which regions of a multi-domain enzyme are actually functionally important, this procedure avoids the chance of missing discriminating residues that might be located in linker regions between individual domains. Fourth, the MSA and HMM of CHIEFc families are constructed only from sequences with the same EC number. In contrast, the sequences used to construct the MSA and HMM of Pfam domains may be from a variety of different functions. Thus, in cases where only a small number of sequences are related to the enzyme function under study, the Pfam domain alignments might be dominated by sequences with different functions.

The EF method is the major reason for the high sensitivity of EFICAz

Since functionally important residues are usually highly conserved in an MSA of proteins sharing a given function, a standard way of inferring their function is to identify conserved residues in the MSA and then verify their conservation in a query sequence, as a measure of functional similarity. However, a common problem of this approach is that there may not be a sufficient number of sequences in the MSA, or the sequences in the MSA may not be divergent enough, making the identification of the functionally important residues difficult. To overcome this problem, we have developed an EF method that is mainly based on the assumption that homologous proteins with divergent enzyme functions frequently share the same architecture of the active site, although different functional residues are used to perform different functions (26). Thus, the EF method exploits information about conservation in heterofunctional MSAs to facilitate the identification of FDRs in homofunctional MSAs. On the other hand, the inclusion of non-enzymes in the heterofunctional MSAs, which is not a common practice in the derivation of methods for enzyme function inference, imposes more restrictions on the selection of FDRs, but reflects better the real genome annotation scenario. As shown in Figure 3, the combined information of heterofunctional and homofunctional MSA conservation significantly reduces the number of FDRs corresponding to a given enzyme family compared with using homofunctional MSAs alone. This reduction in the number of discriminating residues is reflected in an increased sensitivity of the methods for enzyme function inference based on FDR recognition compared with other approaches (see Figure 5B and E). Consequently, the two components based on FDR recognition contribute the most to the high sensitivity of EFICAz.

By combining four accurate methods, EFICAz achieves high sensitivity while keeping high accuracy

EFICAz is a combination of four accurate methods: CHIEFc family specific SIT evaluation, CHIEFc family based FDR recognition, Multiple Pfam family based FDR recognition and High specificity multiple Prosite pattern recognition. Each of the four methods has been specifically developed for high accuracy. For example, compared with Single Pfam family based FDR recognition, the Multiple Pfam family based FDR recognition significantly improves the prediction accuracy in the benchmark test (Figure 5A and D). Therefore, EFICAz is accurate because each of its four constituent methods is highly accurate. On the other hand, each component of EFICAz alone generates a number of successful predictions that cannot be obtained by the rest of the components. Thus, the high sensitivity of EFICAz (Figure 5E) is due to the fact that predictions generated by any pair of its component methods do not overlap completely. By requiring the consensus of two or more components instead of inferring a particular enzyme function when anyone or more of the four component methods predicts it, EFICAz can achieve an accuracy of almost 100% (Figure 5G) and reasonable levels of sensitivity (Figure 5H). We have explored more elaborate ways of combining the predictions of the four components of EFICAz (data not shown), but those evaluated in Figure 5G–I are the combinations that offer the most useful tradeoff between accuracy and sensitivity.

EFICAz provides a platform for automatic and accurate enzyme function annotation

To illustrate the application of our method to a real case, we have employed EFICAz for the genome-wide enzyme functional inference of the *E.coli* K12 proteome. It should be stressed that *E.coli* is one of the most intensively studied organisms (41); therefore, it is more difficult to generate novel predictions in *E.coli* than in other less well-characterized organisms. The comparison of our predictions with KEGG annotations suggests that while putting the emphasis in the

accuracy of functional inference, EFICAz can assign more detailed enzymatic function than KEGG and is capable of generating novel predictions.

In the development of EFICAz, we have tried to overcome the problems and weaknesses observed in other methods for systematic genome-scale inference of enzyme function. The annotation scheme of KEGG is mainly based on orthologous relationships and the examination of functional linkage between genes (25). However, the assessment of orthologous relationships still depends on sequence comparisons, which we have previously shown are problematic when used to infer functional similarity even at high levels of sequence identity (10). In contrast, instead of relying only on sequence similarity, EFICAz is based on four accurate methods that have all been developed and optimized to achieve high prediction accuracy. Moreover, three out of the four components of EFICAz are based on the recognition of patterns or residues that at least in the case of the CHIEFc family based FDR recognition approach are correlated with important functional features of enzymes (Figure 4).

Another recently developed approach, PRIAM, is an enzyme function-specific profile-based method based on homology transfer using a single *E*-value cut-off common for all families (42). However, we previously found that *E*-value is not a reliable measure for functional inference; in addition, using a common cut-off is not accurate because of the inconsistency between sequence and function divergence for different families (10). In contrast, the sequence comparison component of EFICAz uses enzyme family specific SITs, which we have shown to be very accurate (Figure 5A and D). The functional subtype analysis, another method that could be extended for automatic genome annotation, identifies positions in a Pfam domain that discriminate between enzyme subtypes with different specificities (20). However, although certain enzyme subtypes might require discriminating residues located in different domains to be distinguished from other subtypes, this method is limited to the analysis of single Pfam domains. In contrast, the component of EFICAz based on the analysis of Pfam families uses a specific combination of these families rather than a single family to generate a prediction, with a significant increase in accuracy (Figure 5A and D).

Genome annotation requires the utilization of highly reliable approaches to minimize the problematic propagation of annotation errors (43). By stressing the importance of accuracy, EFICAz can contribute to the complex task of high quality annotation of enzyme function on a genome-scale. In the near future, we plan to apply EFICAz to all available complete genomes and compile the predictions in a public database. Although complete genomes always have a certain level of annotation when they are released into the public domain, it is advantageous to periodically re-annotate them using automatic, reproducible protocols (44) such as EFICAz. These predictions might serve as a first step for metabolic pathway reconstruction protocols, or to identify sequences that can be investigated by researchers interested in particular enzyme functions.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

A.K.A. acknowledges partial support from the Pew Latin American Fellows Program in Biomedical Sciences. We gratefully acknowledge Mike McManaman for helping us to set the EFICAz web server. This research was supported in part by National Institutes of Health Grants GM-48835 and U54AI-057158.

REFERENCES

1. Orengo,C.A., Todd,A.E. and Thornton,J.M. (1999) From protein structure to function. *Curr. Opin. Struct. Biol.*, **9**, 374–382.
2. Kenyon,G.L., DeMarini,D.M., Fuchs,E., Galas,D.J., Kirsch,J.F., Leyh,T.S., Moos,W.H., Petsko,G.A., Ringe,D., Rubin,G.M. *et al.* (2002) Defining the mandate of proteomics in the post-genomics era: workshop report. *Mol. Cell. Proteomics*, **1**, 763–780.
3. Rost,B., Liu,J., Nair,R., Wrzeszczynski,K.O. and Ofran,Y. (2003) Automatic prediction of protein function. *Cell. Mol. Life Sci.*, **60**, 2637–2650.
4. Jimenez-Sanchez,G., Childs,B. and Valle,D. (2001) Human disease genes. *Nature*, **409**, 853–855.
5. Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.
6. Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Meth. Enzymol.*, **266**, 460–480.
7. 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.
8. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
9. Devos,D. and Valencia,A. (2001) Intrinsic errors in genome annotation. *Trends Genet.*, **17**, 429–431.
10. Tian,W. and Skolnick,J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.*, **333**, 863–882.
11. Sigrist,C.J., Cerutti,L., Hulo,N., Gattiker,A., Falquet,L., Pagni,M., Bairoch,A. and Bucher,P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinformatics*, **3**, 265–274.
12. Attwood,T.K., Croning,M.D., Flower,D.R., Lewis,A.P., Mabey,J.E., Scordis,P., Selley,J.N. and Wright,W. (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.*, **28**, 225–227.
13. Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.
14. Via,A. and Helmer-Citterich,M. (2004) A structural study for the optimisation of functional motifs encoded in protein sequences. *BMC Bioinformatics*, **5**, 50.
15. Casari,G., Sander,C. and Valencia,A. (1995) A method to predict functional residues in proteins. *Nature Struct. Biol.*, **2**, 171–178.
16. del Sol Mesa,A., Pazos,F. and Valencia,A. (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, **326**, 1289–1302.
17. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.
18. Armon,A., Graur,D. and Ben-Tal,N. (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.*, **307**, 447–463.
19. Landgraf,R., Xenarios,I. and Eisenberg,D. (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.*, **307**, 1487–1502.
20. Hannenhalli,S.S. and Russell,R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.*, **303**, 61–76.
21. 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.
22. Babbitt,P.C. (2003) Definitions of enzyme function for the structural genomics era. *Curr. Opin. Chem. Biol.*, **7**, 230–237.

23. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res*., **28**, 304–305.
24. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*., **28**, 45–48.
25. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27–30.
26. Gerlt,J.A. and Babbitt,P.C. (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu. Rev. Biochem*., **70**, 209–246.
27. Todd,A.E., Orengo,C.A. and Thornton,J.M. (2002) Plasticity of enzyme active sites. *Trends Biochem. Sci*., **27**, 419–426.
28. Bork,P. and Koonin,E.V. (1998) Predicting functions from protein sequences—where are the bottlenecks? *Nature Genet*., **18**, 313–318.
29. Iyer,L.M., Aravind,L., Bork,P., Hofmann,K., Mushegian,A.R., Zhulin,I.B. and Koonin,E.V. (2001) Quoderat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. *Genome Biol*., **2**, RESEARCH0051.
30. Sneath,P.H.A. and Sokal,R.R. (1973) *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. W. H. Freeman, San Francisco.
31. Myers,E.W. and Miller,W. (1988) Optimal alignments in linear space. *Comput. Appl. Biosci*., **4**, 11–17.
32. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
33. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
34. Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
35. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
36. Pellicer,M.T., Badia,J., Aguilar,J. and Baldoma,L. (1996) glc locus of *Escherichia coli*: characterization of genes encoding the subunits of glycolate oxidase and the glc regulator protein. *J. Bacteriol*., **178**, 2051–2059.
37. Lord,J.M. (1972) Glycolate oxidoreductase in *Escherichia coli*. *Biochim. Biophys. Acta*, **267**, 227–237.
38. Yew,W.S. and Gerlt,J.A. (2002) Utilization of L-ascorbate by *Escherichia coli* K-12: assignments of functions to products of the yjf-sga and yia-sgb operons. *J. Bacteriol*., **184**, 302–306.
39. Lehmann,C., Doseeva,V., Pullalarevu,S., Krajewski,W., Howard,A. and Herzberg,O. (2004) YbdK is a carboxylate-amine ligase with a gamma-glutamyl:cysteine ligase activity: crystal structure and enzymatic assays. *Proteins*, **56**, 376–383.
40. Miller,B.G. and Raines,R.T. (2004) Identifying latent enzyme activities: substrate ambiguity within modern bacterial sugar kinases. *Biochemistry*, **43**, 6387–6392.
41. Donnenberg,M.S. and Whittam,T.S. (2001) Pathogenesis and evolution of virulence in enteropathogenic and enterohemorrhagic *Escherichia coli*. *J. Clin. Invest*., **107**, 539–548.
42. Claudel-Renard,C., Chevalet,C., Faraut,T. and Kahn,D. (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res*., **31**, 6633–6639.
43. Brenner,S.E. (1999) Errors in genome annotation. *Trends Genet*., **15**, 132–133.
44. Ouzounis,C.A. and Karp,P.D. (2002) The past, present and future of genome-wide re-annotation. *Genome Biol*., **3**, COMMENT2001.