doi:10.1016/j.jmb.2003.08.057 *J. Mol. Biol.* (2003) **333**, 863–882

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

How Well is Enzyme Function Conserved as a Function of Pairwise Sequence Identity?

Weidong Tian[1,2] 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 USA*

[2]*Department of Biology Washington University in St Louis, One Brookings Drive, St Louis, MO 63130 USA*

**Corresponding author*

Enzyme function conservation has been used to derive the threshold of sequence identity necessary to transfer function from a protein of known function to an unknown protein. Using pairwise sequence comparison, several studies suggested that when the sequence identity is above 40%, enzyme function is well conserved. In contrast, Rost argued that because of database bias, the results from such simple pairwise comparisons might be misleading. Thus, by grouping enzyme sequences into families based on sequence similarity and selecting representative sequences for comparison, he showed that enzyme function starts to diverge quickly when the sequence identity is below 70%. Here, we employ a strategy similar to Rost's to reduce the database bias; however, we classify enzyme families based not only on sequence similarity, but also on functional similarity, i.e. sequences in each family must have the same four digits or the same first three digits of the enzyme commission (EC) number. Furthermore, instead of selecting representative sequences for comparison, we calculate the function conservation of each enzyme family and then average the degree of enzyme function conservation across all enzyme families. Our analysis suggests that for functional transferability, 40% sequence identity can still be used as a confident threshold to transfer the first three digits of an EC number; however, to transfer all four digits of an EC number, above 60% sequence identity is needed to have at least 90% accuracy. Moreover, when PSI-BLAST is used, the magnitude of the *E*-value is found to be weakly correlated with the extent of enzyme function conservation in the third iteration of PSI-BLAST. As a result, functional annotation based on the *E*-values from PSI-BLAST should be used with caution. We also show that by employing an enzyme family-specific sequence identity threshold above which 100% functional conservation is required, functional inference of unknown sequences can be accurately accomplished. However, this comes at a cost: those true positive sequences below this threshold cannot be uniquely identified.

Keywords: genome annotation; conservation of protein function; enzyme classification; sequence comparisons; PSI-BLAST

Introduction

In this post-genomic era with many sequenced genomes, functional annotation has become a major aim of Bioinformatics.[1–7] The most widely used functional annotation scheme is based on two steps. The first step is to detect a homologous relationship between pairs of proteins; this can be accomplished by a pairwise sequence similarity search with algorithms such as FASTA,[8] BLAST[9] and PSI-BLAST.[10] The second step of functional annotation is to infer functional similarity from homology. With the continuing development of those methods, the ability of recognizing remote homologies has been greatly improved. However, because there might be only about 1000 major superfamilies in nature,[11,12] most homologous (viz. evolutionarily related) proteins must have different functions, which makes the inference of functional similarity from sequence similarity difficult and perhaps problematic.[13,14] With the rapidly increasing number of completely sequenced genomes

Abbreviation used: EC, enzyme commission.

E-mail address of the corresponding author: skolnick@buffalo.edu and the efforts to annotate gene function, annotation errors could be easily spread if functional annotation is not done carefully; thus, systematic studies that establish the accuracy and reliability of methods that infer functional similarity from homology are urgent and necessary.

Percentage sequence identity and statistical score, such as *E*-value of BLAST or FASTA, are widely used measures for sequence comparison. It has been well established that scores based on the statistical significance relative to random are superior to percentage sequence identity in detecting remote homology.[15] However, there is no clear indication of whether this is also true with respect to assessing functional relationship. In fact, it has been frequently observed that function starts to diverge quickly even at high level of sequence identity at which there is no dispute about homology.[16–20] On the other hand, as a quick and simple measure, sequence identity is also widely used as an indication of functional similarity. For example, it is often implicitly used in dividing a protein family into subfamilies by constructing a phylogenetic tree to derive functionally important residues.[21–24] However, because the lack of a rigorously established sequence identity threshold, the division of a protein family into subfamilies may require human intervention.[24,25] Therefore, it is of great significance to establish the threshold of sequence identity above which functional similarity can be affirmed.

Because the Enzyme Commission (EC) is the best developed and most widely used functional classification scheme,[26] EC numbers have been employed to explore the threshold of sequence identity necessary for accurate function transfer. EC numbers classify the function of an enzyme by four digits. The first digit delineates the main type of enzymatic activity and ranges from 1 to 6: 1, oxidoreductases; 2, transferases; 3, hydrolases; 4, lyases; 5, isomerases; and 6, ligases. The other three digits provide more detail about the reaction that an enzyme catalyzes. The last digit of an EC number usually represents the substrate specificity of a reaction, while the first three digits of the EC number usually describe the overall type of enzymatic reaction. By conducting all-against-all pairwise sequence comparisons and examining EC number match at different sequence identity thresholds, Devos,[20] Wilson,[18] and Todd[17] observed that enzyme function is well conserved. Devos took structure alignments from the FSSP (families of structurally similar proteins) database[27] and discovered that above 50% sequence identity, all four digits of an EC number are well conserved. Wilson performed pairwise sequence, structure and function comparisons on protein domain pairs according to the SCOP (Structural Classification of Proteins) fold classification[28] and found that full conservation of all four EC digits can occur between two proteins with as low as ~40% sequence identity. Todd assessed the functional variation of homologous enzyme superfamilies defined by the CATH (protein class, architecture, topology and homologous superfamily classification) protein structure classification[29] and found that functional variation is rare when the sequence identity is above 40%. Thus, it seemed that 40% sequence identity might be used as a confident threshold for assessing functional conservation.

In contrast, a recent study by Rost[19] argued that the SWISSPROT database,[30] which has been used as a gold standard for the functional annotation of the other databases,[31–34] has many redundant sequences that cover just a small fraction of enzyme functions. Thus, SWISSPROT is a biased database dominated by a few functional families, and the results of enzyme function conservation based only on simple pairwise comparison might be misleading. To reduce the bias in the SWISSPROT database, Rost classified enzyme sequences into families on the basis of their sequence similarity, or detectable evolutionary relationship. Usually, two proteins that are structurally similar to each other are considered as evolutionarily related. It has been established that when the pairwise sequence identity between two proteins is above 30%, they have similar structures and are evolutionarily related.[35–37] However, the relationship between sequence–structure similarities is not clear when the pairwise sequence identity is below 30%, especially in the "twilight zone" (<25% sequence identity).[35–37] To extend sequence comparison into the twilight zone, Rost employed the HSSP (homology-derived structures of proteins) score (a score derived from sequence identity to indicate whether two sequences might have similar structures) to measure the sequence similarity between two proteins.[36,38] He grouped enzyme sequences retrieved from the SWISSPROT database into sequence families based on their HSSP score and then selected representative sequences to construct an unbiased dataset. Finally, enzyme sequences from the unbiased dataset were compared with those from the original dataset to calculate the extent of enzyme function conservation. Rost showed that when the sequence identity is below 70%, both the first digit and all four digits of EC numbers start to quickly diverge;[19] a significantly different conclusion from that of previous studies. This discrepancy in the threshold of enzyme function conservation has raised questions about whether current functional annotation schemes based on sequence similarities can be trusted. Thus, additional evaluation of enzyme function conservation is timely and important.

Obviously, by classifying enzyme sequences into families to reduce the bias, the conclusions of Rost should be closer to the truth. Presumably, a family of sequences should have a clear evolutionary relationship and be functionally similar to each other. However, because the relationship between functional divergence and sequence divergence is in fact not clear (a point that is further addressed here), using only sequence similarity to classify protein families might result in one family being linked to different kinds of function. In fact, it has been frequently observed that function may diverge more quickly than sequence, and homologous proteins may evolve to have different functions and possess different functional sites, especially when sequence identity falls below 40%.[17,18,20,39] Thus, the dataset constructed by Rost that contains only representative sequences might miss some enzyme functions in the calculation of functional conservation. Furthermore, the presence of various functions in one family may also make it difficult to accurately transfer function to new sequences.

Here, we employ both functional similarity and sequence similarity to define a protein family. Instead of trying to define functional similarities, we directly use the functional annotation in the SWISSPROT database and define an enzyme family as a family of sequences that are all above a certain threshold of pairwise sequence similarity and that also have the same function. Enzyme function is defined at two levels: by conservation of the full four digits of the EC number, which include the substrate specificity and more detailed information, such as cofactor or metal of a particular enzyme reaction, and by conservation of the first three digits of the EC number, which generally has a less detailed description of a particular type of enzyme reaction. Employing these criteria, we have classified all enzyme sequences (excluding those sequences with multiple EC numbers, or undetermined EC digits, or identified only by sequence similarity using computational methods) in the SWISSPROT database. We calculate the functional conservation rate of each family by collecting all possible sequence pairs related to the family in the SWISSPROT database at different thresholds of sequence identity and then compare their functional annotation. Finally, we average the enzyme function conservation across all the enzyme families. Our results suggest that for functional annotation, 40% sequence identity can still be used as a confident threshold to transfer the first three digits of an EC number. However, to transfer all four digits of an EC number, above 60% sequence identity is needed to have above 90% accuracy. Moreover, we find that the threshold of the *E*-value for enzyme function conservation changes significantly during the PSI-BLAST iteration process, and in the third iteration of PSI-BLAST, the *E*-value shows only a weak correlation with functional conservation. Furthermore, by conducting a jack-knife analysis, we find that by employing an enzyme family-specific threshold above which 100% functional conservation is required, functional inference of unknown sequence from known sequence can be done accurately. However, because 100% conservation rate is required for establishing the threshold, true positive sequences that have a sequence identity to sequences of known function lower than the threshold cannot be identified. Finally, we apply the family-specific threshold to KEGG annotated enzyme sequences and find that about 58% and 65% of KEGG enzyme sequences can be confirmed with 100% confidence at full four EC digits and the first three EC digits level, respectively. All of our results can be downloaded from our website†.

Results

The conservation of all four digits of the EC numbers is lower than previously anticipated, while the first three digits of the EC numbers are still well conserved

We have calculated the average degree of enzyme function conservation across all the classified enzyme families over different ranges of pairwise sequence identity. To investigate the effects of the definition of sequence identity on the extent of enzyme function conservation, we conduct the calculation based on the global identity, the big identity, the small identity, and the aligned identity of the MM alignment, and the sequence identity reported by the PSI-BLAST alignment, respectively (the definitions of the various types of sequence identity can be found in Methods: Measurements of sequence similarity). The results are shown in Figure 1. It can be clearly seen that: (1) the global identity has the strongest correlation to the degree of enzyme function conservation, while the correlation of the aligned sequence identity of the MM alignment and PSI-BLAST sequence identity to enzyme functional conservation is worse than the other three definitions of sequence identity. For example, for enzyme–enzyme only comparisons, when the pairwise sequence identity falls into the 50–60% range, the conservation rate of the full four digits of an EC number is 87% when the global identity is used, while it is only 68% and 78% when the aligned identity of the MM alignment and the PSI-BLAST identity are used, respectively (Figure 1A). Moreover, when the pairwise sequence identity is in the 40–50% range, the conservation rate of the first three digits of an EC number can still be 93% when the global identity is used, while it is only 74% and 72% when the aligned identity of the MM alignment and the PSI-BLAST identity are used, respectively (Figure 1C). Because the global identity has the strongest correlation to enzyme functional conservation, it is used in what follows to study how well the enzyme function is conserved.

(2) The conservation of the full four digit EC number is lower than previously anticipated, while the first three digits of the EC numbers are still well conserved. Following the strategy used by Devos,[20] Wilson,[18] and Todd,[17] we also conduct a simple all-against-all pairwise sequence comparison without family classification and employ the global identity to calculate the enzyme function

† http://bioinformatics.buffalo.edu/resources/enzyme_conservation

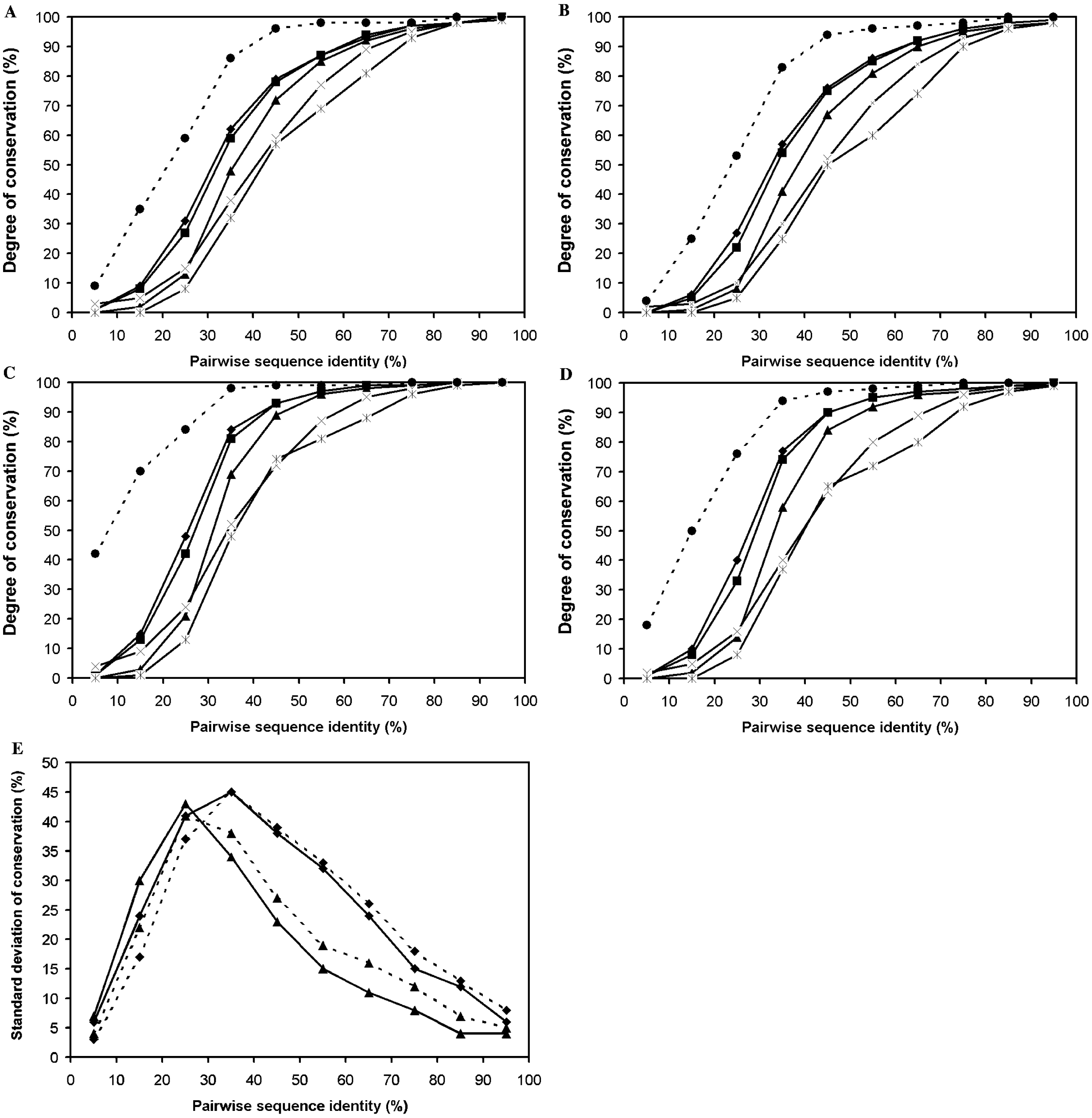

Figure 1. The enzyme function conservation in terms of sequence identity. Two pools of sequence pairs are used to derive the extent of enzyme function conservation: one that includes only enzyme–enzyme sequence pairs (A and C), and another that includes both enzyme–enzyme and enzyme–non-enzyme sequence pairs (B and D). The extent of functional conservation is calculated at two levels of enzyme functions: the all four EC digits (A and B), and the first three EC digits (C and D). In calculating the extent of enzyme function conservation, the functional conservation of each enzyme family is calculated first and then is averaged across all enzyme families. To investigate the effects of the definition of sequence identity on enzyme function conservation, five definitions of sequence identity are used: the global sequence identity (diamond symbol in continuous line, ◆), the big sequence identity (square symbol in continuous line, ■), the small sequence identity (triangle symbol in continuous line, ▲), and the aligned sequence identity of the MM alignment (star symbol in continuous line, ∗), and the reported sequence identity of the PSI-BLAST alignment (cross symbol in continuous line, ×). (The definitions can be found in Methods: Measurements of sequence similarity.) To compare with the previous estimation of enzyme function conservation, a simple all-against-all pairwise sequence comparison without family classification is conduced using the global sequence identity (circle symbol in broken line, ●). The standard deviation of enzyme function conservation as a function of global sequence identity is shown in E. The standard deviation of the conservation of all four EC digits and the first three EC digits are shown with diamond symbol (◆) and triangle symbol in continuous line (▲), respectively, with the continuous line and broken line representing the results in the condition of only enzyme–enzyme comparisons and both enzyme–enzyme and enzyme–non-enzyme comparisons, respectively.

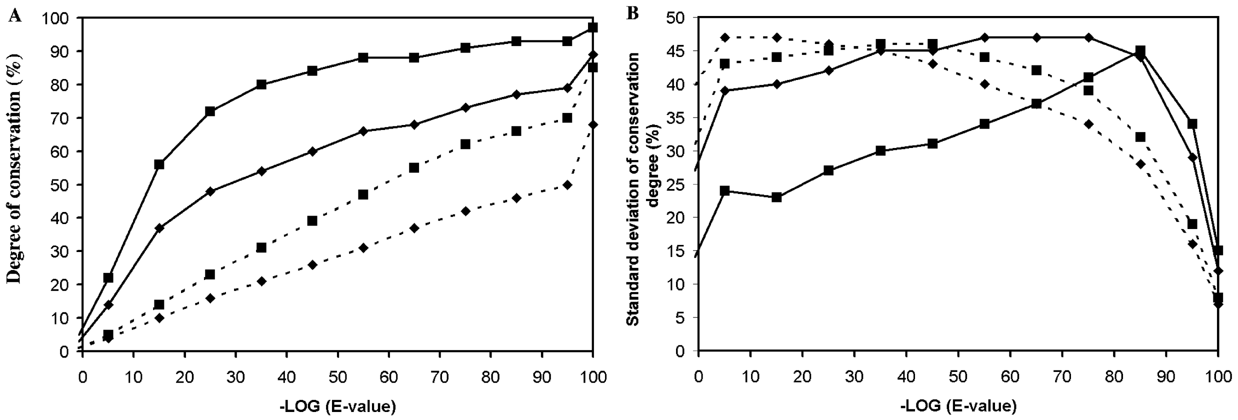

Figure 2. The enzyme function conservation in terms of the PSI-BLAST *E*-value. Following a similar strategy to calculate the enzyme function conservation in terms of sequence identity, the PSI-BLAST *E*-value is used to evaluate the extent of enzyme function conservation. A, The result of function conservation in terms of PSI-BLAST *E*-value when only enzyme–enzyme comparisons are used, with the standard deviation shown in B. Two sets of PSI-BLAST *E*-value are employed. The square symbol (■) and the diamond symbol (◆) in continuous line represent the results of the full four EC digits conservation and the first three EC digits conservation, respectively, when the PSI-BLAST *E*-value obtained in the first round of PSI-BLAST search is used. The triangle symbol (▲) and the cross symbol (×) in the broken line represent the results of the full four EC digits conservation and the first three EC digits conservation, respectively, when the PSI-BLAST *E*-value obtained in the third round of PSI-BLAST search is used.

conservation. Similar to the previous observations, for all-against-all enzyme–enzyme comparison, the conservation rate of the full four digits of an EC number can still be 96% when the global identity is in the 40–50% range, while the conservation rate of the first three digits of an EC number is 98% when the global identity falls in the 30–40% range (Figure 1A and C). In contrast, following our strategy to reduce the bias in the SWISSPROT database by both sequence and functional similarity, we find that for enzyme–enzyme comparisons, the conservation rate of the full four digits of an EC number starts to be below 90% when the global identity is below 60% (Figure 1A); this makes the extent of conservation of all four digits of an EC number lower than previously anticipated. However, the first three digits of an EC number are still well conserved; for enzyme–enzyme only comparisons, the average conservation rate of the first three digits of an EC number can still be 93% in the range of 40–50% of global identity (Figure 1C). This result is in contrast to Rost's observation that both the full four digits and the first digit of an EC number start to diverge when sequence identity is below 70%.[19] When enzyme–non-enzyme comparisons are also included in the calculation, we find that the degree of function conservation is a little lower than that when only enzyme–enzyme comparison is counted (Figure 1B and D). However, the degree of function conservation for the full four digits and the first three digits can still be above 90% when the global identity is above 60% and 40%, respectively (Figure 1B and D).

Thus, our analysis suggests that for function transferability, 40% global identity can still be used as a confident threshold to transfer the first three digits of an EC number; however, to transfer the full four digits of an EC numbers with above 90% accuracy, then above 60% global identity is required.

The *E*-value at the third iteration of PSI-BLAST has a weak correlation to enzyme function conservation

Following the similar strategy used to calculate the enzyme function conservation as a function of the global identity, we also calculate the enzyme function conservation as a function of the PSI-BLAST *E*-value, e.g. e^{-60}–e^{-50} (Figure 2). Our results show that the *E*-value in the third iteration of the PSI-BLAST search has a weak correlation to the degree of enzyme function conservation. For example, for enzyme–enzyme comparison, even when the *E*-value is less than e^{-100}, the average conservation rate of the full four digits and the first three digits of an EC number is only 68% and 85%, respectively (Figure 2A).

In contrast, the *E*-value in the first round of the PSI-BLAST search has a much better correlation to enzyme function conservation. For example, the average function conservation rates of all four digits and the first three digits of an EC number are 89% and 97%, respectively, when the *E*-value is less than e^{-100} in the first round of the PSI-BLAST search. Moreover, the average conservation rate of the first three digits of an EC number can be 88% when the *E*-value is between e^{-60} and e^{-50} in the first round of the PSI-BLAST search, in contrast to only around 47% in the third iteration of the PSI-BLAST search in the same *E*-value level (Figure 2A). The significant changes of the threshold of *E*-value for enzyme function conservation

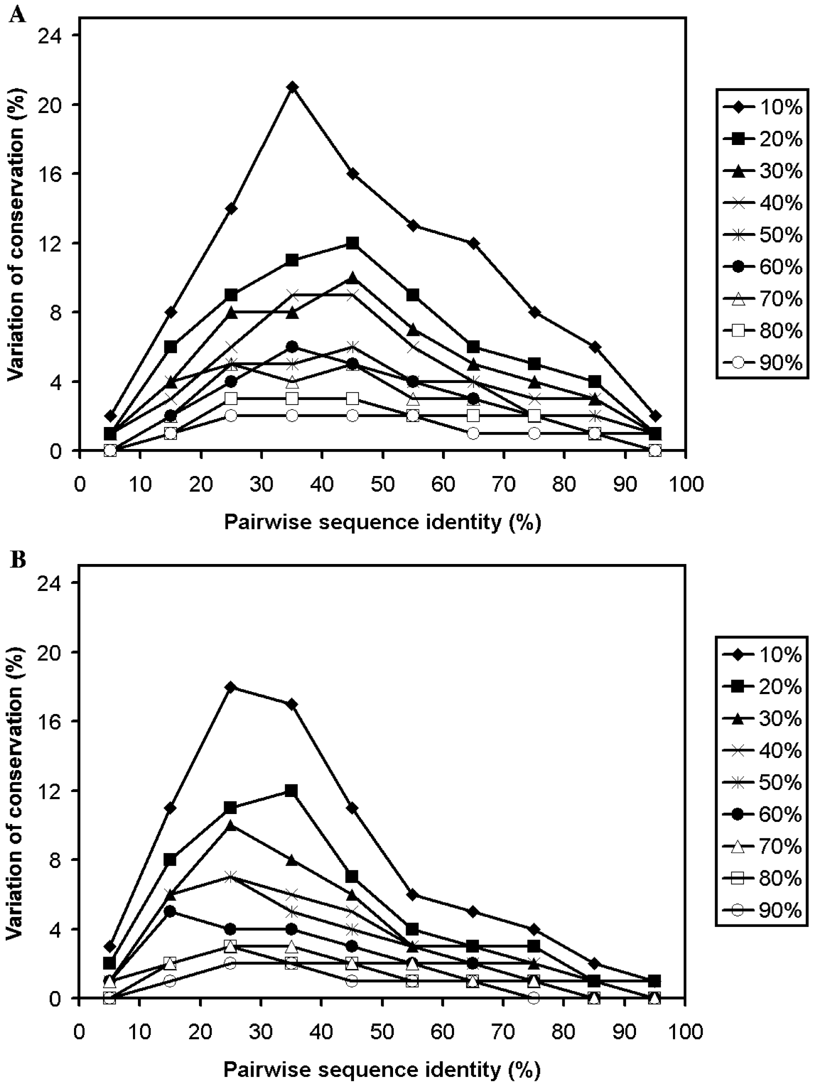

Figure 3. The variation of function conservation by the selection of enzyme families. A, The conservation variation of full four EC digits by using different subsets of full four EC digits enzyme families at different levels of sequence identity. B, The conservation variation of the first three EC digits by using different subsets of the first three EC digits enzyme families at different levels of sequence identity.

at different steps of PSI-BLAST search suggests that *E*-value is not a good measure for transferring functions by PSI-BLAST.

The average enzyme function conservation does not differ by the selection of enzyme families

Figures 1E and 2B show the standard deviation of conservation rate at different levels of sequence identity as a function of the global sequence identity and the PSI-BLAST *E*-value, respectively. Apparently, functional conservation varies among different enzyme families. However, the average enzyme function conservation does not depend on which enzyme families are selected, provided there are a sufficient number of enzyme families considered. When using 10% of the total number of enzyme families, the average conservation rate varies significantly among different subsets (Figure 3A and B). However, when additional enzyme families are included for the calculation, the results become stable: when using more than 50% of the total number of enzyme families, the variation among different subsets is less than 5% at all level of sequence identity (Figure 3A and B).

When the conservation rate of an individual enzyme family is not 100% at a level of sequence identity, it varies with the selection of enzyme sequences

Although the average enzyme function conservation does not differ with the selection of enzyme families (Figure 3A and B), for an individual enzyme family, generally, the inclusion of more enzyme sequences makes the sequence identity threshold required to accurately transfer function more stable (Figure 4). However, when the sequence identity level is high, there is not much variation of the conservation rate: the average conservation variation for individual enzyme families is less than 5% for all subsets when the global sequence identity is above 60% (Figure 4A).

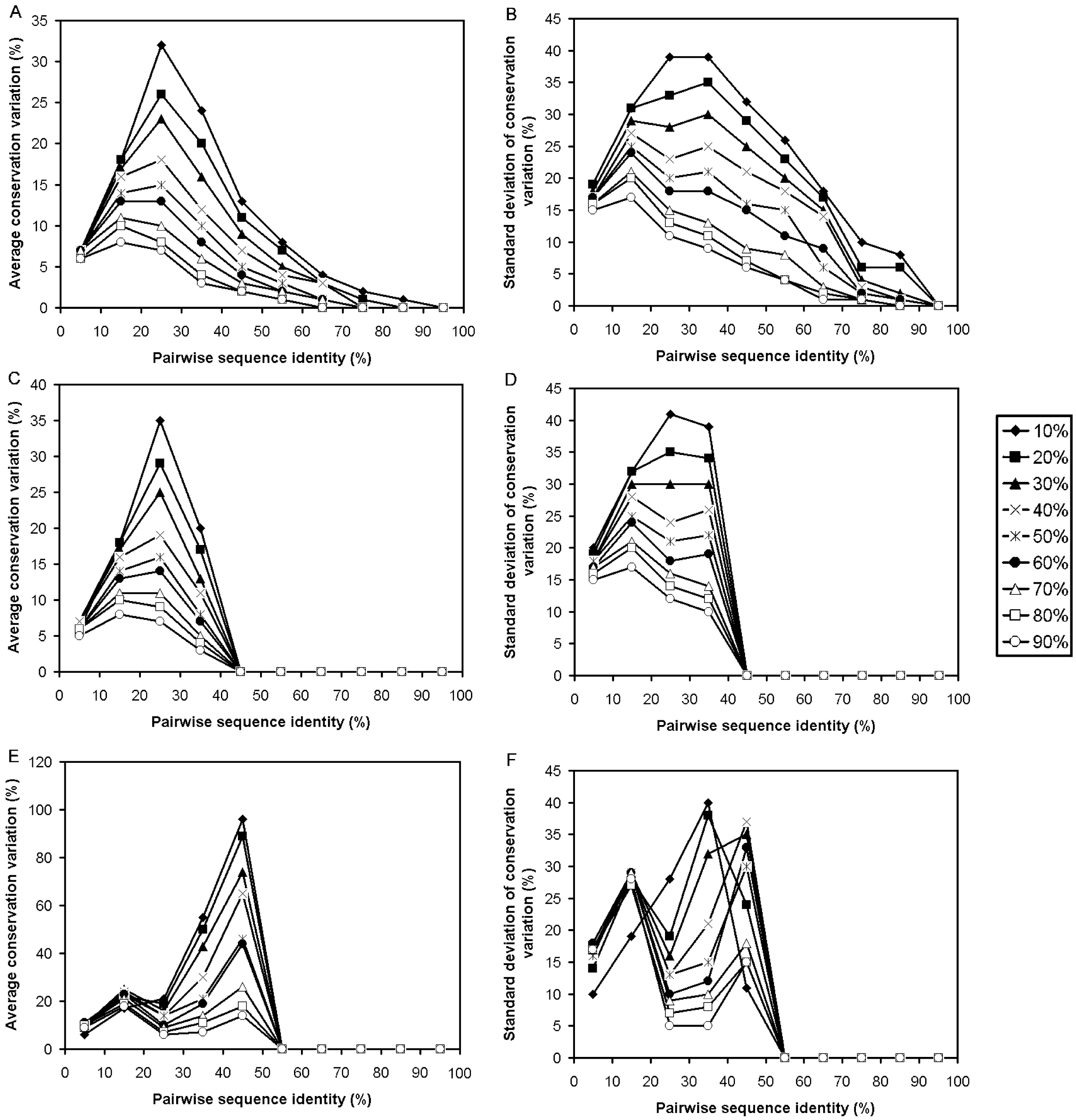

Figure 4. The variation of functional conservation of individual enzyme families by the selection of enzyme sequences. A, C and E show the average conservation variation by using different subsets of enzyme sequences (the symbol for different subsets is shown in the text box) at different levels of sequence identity, for all full four EC digits enzyme families that have more than 30 sequences, those full four EC digits enzyme families that are absolutely conserved above 40% global sequence identity, and those full four EC digits enzyme families that are conserved above 50% but not above 40% global sequence identity, respectively. The corresponding standard deviation of conservation is shown in B, D, and F, respectively.

This is because when the sequence identity is above 60%, the enzyme function tends to be conserved (Figure 1). Actually, when we focus on those enzyme families that are absolutely conserved above 40% global sequence identity, it can be clearly seen that there is no conservation variation for all subsets when the sequence identity is above 40% (Figure 4C). In contrast, when we focus on those enzyme families that are absolutely conserved above 50% but not above 40% global sequence identity, it can be clearly seen that the function conservation rate varies significantly with the selection of enzyme sequences when the level of global sequence identity is 40–50%: the average conservation variation is around 70% and 20% when 30% and 70% of the total number of known enzyme sequences in the family are used, respectively (Figure 4E). Because there are only a small number of enzyme families that are absolutely conserved above 50% sequence identity but not above 40% global sequence identity (12 families), it is possible that the conservation variation at 40–50% sequence identity level may change when more enzyme families are used. Nevertheless, the conservation rate of individual enzyme family can vary significantly with the selection of enzyme sequences when it is not 100% conserved at a particular level of sequence identity.

When both enzyme–enzyme and enzyme–non-enzyme are included in the comparison, above 70% and 50% global sequence identity is required to have more than 90% of the enzyme families absolutely conserved at the level of full four EC digits and the first three EC digits, respectively

As we have shown in Figure 4, the conservation rate of an individual enzyme family at a level of sequence identity is sensitive to the selection of enzyme sequences if it is not 100% conserved. Therefore, in order to accurately transfer function from sequence comparison, it is desirable that we use the enzyme family-specific sequence identity threshold above which a 100% conservation rate can be obtained. We have obtained the family-specific sequence identity threshold for all the classified full four EC digits and the first three EC digits enzyme families. As an alternative indication of enzyme function conservation, we plot the percentage of absolutely conserved enzyme families above different sequence identity levels (Figure 5). For enzyme–enzyme only comparisons, when the global identity is above 60% and 40%, there are more than 90% of the enzyme families absolutely conserved at the level of full four EC digits and the first three EC digits EC, respectively. When both enzyme–enzyme and enzyme–non-enzyme comparisons are included, above 70% and 50% global identity is now required to have more than 90% of the enzyme families absolutely conserved at the level of full four digits and the first three digits of the EC number, respectively. However, there are still more than 70% of the families absolutely conserved at the level of the full four EC digits and the first three EC digits, when the global sequence identity is above 40% and 30%, respectively (Figure 5). This suggests that by using a family-specific sequence identity threshold, the majority of enzyme functions can still be transferred at relatively low levels of sequence identity with confidence. However, we have to point out that the majority of the conserved enzyme families are families with a small number of sequences. Thus, it is possible that the current status of enzyme function conservation might change when more enzyme sequences are deposited into the database.

The enzyme family-specific sequence identity threshold can be used to transfer function accurately from sequence comparison; however, because a 100% conservation rate is required, true positive sequences can be missed

After all, the goal of studying enzyme function conservation is to transfer enzyme functions accurately. To test whether the calculated enzyme family-specific sequence identity threshold can be applied to accurately transfer function, we conduct a jack-knife analysis. The results do not differ by the selection of "training" and "test" sequence for both the full four EC digits and the first three EC digits transfer (Figure 6A and B). Generally, the percentage of "transferable" sequence pairs (the "test"-"training" sequence pair with a sequence identity above the threshold of the family to which the training sequence belongs) decreases with the decrease of the level of sequence identity,

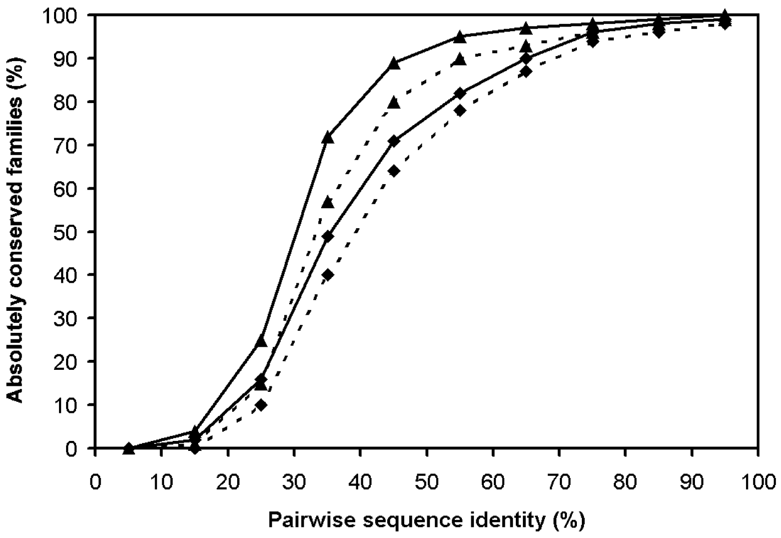

Figure 5. The percentage of 100% conserved enzyme families above different levels of sequence identity. The diamond symbol (◆) and the triangle symbol (▲) represent the results of the full four EC digits and the first three EC digits enzyme families, respectively. The results from only enzyme–enzyme comparisons and from both enzyme–enzyme and enzyme–non-enzyme comparisons are shown as continuous line and broken line, respectively.

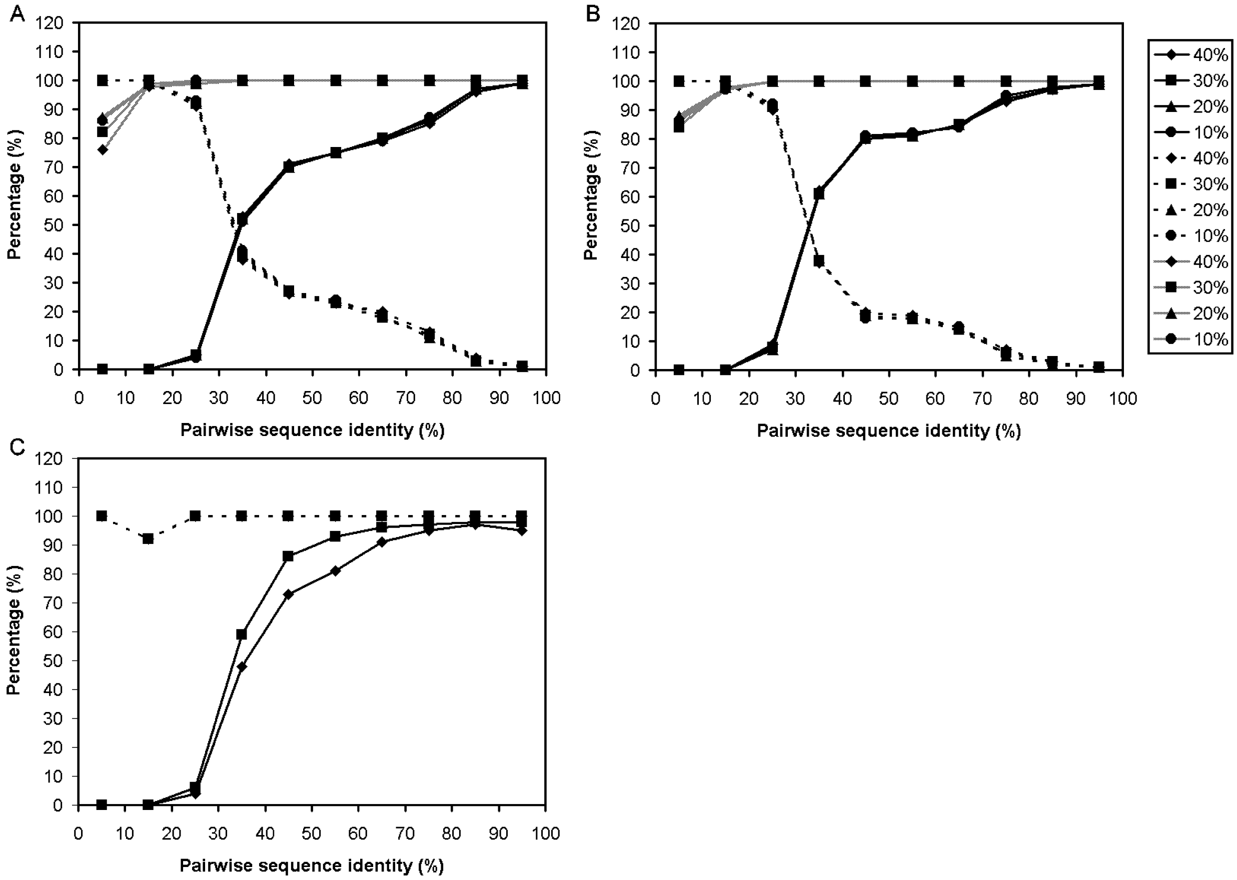

Figure 6. The application of enzyme family-specific sequence identity threshold for functional inference of unknown sequences. The results of jack-knife analysis for the full four EC digits and the first three EC digits transfer are shown in A and B, respectively. The average percentage of transferable sequence pairs, missed sequence pairs, and the prediction accuracy at different levels of sequence identity are shown in continuous line, broken line, and gray line, respectively. The symbols of different test sets are shown in the text box. The standard deviation of the above numbers is small and is not shown here. C, The prediction results of KEGG annotated enzyme sequences by employing the enzyme family-specific sequence identity threshold obtained from the 22,645 SWISSPROT sequences. The diamond symbol (◆) and the square symbol (■) represent the results of the full four EC digits and the first three EC digits prediction, respectively. The percentage of transferable sequence pairs and the prediction accuracy at different sequence identity levels are shown in continuous line and broken line, respectively.

and it drops close to zero when the sequence identity is below 30%. However, it can be clearly seen that despite the decrease of transferable sequence pairs when the global sequence identity drops, the prediction accuracy is not affected by the level of sequence identity: when the sequence identity is above 20%, it is always 100% for both full EC number and the first three EC number transfers (Figure 6A and B). However, the price for accurate functional transfer is that all those true positive test sequences that have a sequence identity with the training sequence below its family-specific identity threshold cannot be identified, and the number increases when the sequence identity drops: at the level of 40–50%, the percentage of the "missed" sequence pairs (the test-training sequence pair with a sequence identity below the threshold of the family of the training sequence but with the same function) is about 30% and 20% for full EC number and the first three EC number transfer, respectively, in contrast to around 40% for both full EC number and the first three EC number transfer at the level of 30–40% sequence identity (Figure 6A and B). Nevertheless, considering the concerns about the substantial amount of misannotations, it is better to have a small number of "safe" annotations than a large number of "unsure" annotations.

The percentage of transferable sequence pairs at different levels of sequence identity reflects the possibility of transferring functions accurately by employing a family-specific identity threshold given a pair of sequences at a particular sequence identity level. However, we note that the distribution of enzyme sequences in the SWISSPROT database is biased and different from that in a real genome;[19] the test-training sequence pairs here might be dominated by those sequences from a small number of enzyme families. As a result, this probability might change when more enzyme sequences are deposited into the database.

Employing a family-specific sequence identity threshold to evaluate the functional annotation of enzyme sequences in the KEGG genome database, about 58% and 65% of the KEGG annotated enzyme sequences have an accurate annotation of the full four EC digits and the first three EC digits, respectively

In Figure 6A and B, we have shown that the family-specific sequence identity threshold is applicable for accurate functional transfer by sequence comparison. Based on these results, we have employed the enzyme family-specific sequence identity threshold to validate the annotation of KEGG enzyme sequences with the result shown in Figure 6C. Among 25,326 KEGG enzyme sequences, 24,697 enzyme sequences can find at least one sequence out of the 22,645 selected SWISSPROT enzyme sequences by running a PSI-BLAST search. Then, by employing the enzyme family-specific sequence identity threshold, 14,813 (58%) and 16,389 (65%) KEGG sequences can be predicted with the full four EC digits and the first three EC digits, respectively. By comparing the predicted function with the KEGG annotation, we can see that the "prediction accuracy" is 100% when the sequence identity between the KEGG sequence and the SWISSPROT sequence is above 20% (Figure 6C). Therefore, those KEGG sequences can then be merged with SWISSPROT enzyme sequences to construct a bigger "clean" enzyme family database that can be found on our website†. However, there are also a large number of KEGG sequences with functions unconfirmed at the level of the full four EC digits (42%), or the first three EC digits (35%), by this method. This does not mean that the annotations of those "unconfirmed" KEGG sequences are wrong, but suggests that those annotations need to be confirmed either by experiment or by more sophisticated measures other than only simple sequence comparison.

In Figure 6C, it can be seen that the percentage of transferable sequence pairs is different from that obtained by using the SWISSPROT enzyme sequences as the test sequences in the jack-knife analysis (Figure 6A and B). For example, for full EC number transfer at 60–70% sequence identity level, the percentage of transferable sequence pairs is 91% and 80% by using the KEGG annotated enzyme sequences and the SWISSPROT enzyme sequences as the test sequences, respectively; for first three EC digits transfer at the 50–60% sequence identity level, this percentage is 96% and 84% by using the KEGG annotated enzyme sequences and the SWISSPROT enzyme sequences as the test sequences, respectively. As mentioned above, this percentage could be regarded as the probability of transferring enzyme functions accurately by employing a family-specific identity threshold. Compared with the SWISSPROT database, the distribution of enzyme sequences in the KEGG database should be closer to that in real genomes; therefore, it is possible that the results obtained by evaluating KEGG annotated enzyme sequences might reflect the limit of accurate enzyme function transfer to unknown sequences from simple sequence comparison based on our current knowledge.

Discussion and Conclusion

How well is enzyme function conserved?

It has been established that use of a statistical score, such as the BLAST *E*-value is superior to percentage sequence identity in detecting remote homology,[15] or structural similarities, by sequence comparison. However, because functional divergence can happen at high levels of sequence identity,[16–20] where there is no dispute about homology, the statistical score might not be advantageous over percentage sequence identity for functional inference. In fact, as shown in Figure 2, the *E*-value is not a good measure for transferring function, especially when multiple iterations of PSI-BLAST are conducted. On the other hand, because the percentage of sequence identity is simple, quick and widely accepted by the biologist, several groups have examined the extent to which pairwise sequence identity functional similarity can be inferred. Devos,[20] Wilson,[18] and Todd[17] conducted an all-against-all pairwise enzyme sequence comparison, and concluded that when the sequence identity is above 40%, above 90% accuracy can be achieved when transferring the full four digits EC number. However, by reducing the database bias with sequence classification, Rost[19] discovered that the enzyme function conservation is much lower than previously anticipated. He showed that less than 30% of all the pairs found at 50% sequence identity had identical EC numbers, and when the sequence identity is below 70%, the conservation of both the first EC number and the full EC number starts to be below 90%. This is significantly different from other groups' results.

By employing the global identity and running a simple all-against-all pairwise sequence comparison, here we have demonstrated that the conservation of the full EC number and the first three digits of the EC number can be above 90% when the global identity is above 40% and 30%, respectively, which is in agreement with Todd's observation.[17] However, after classifying enzyme sequences by both functional and sequence similarity and averaging the extent of function conservation across all enzyme families, we have confirmed Rost's conclusion that enzyme function conservation is lower than previously anticipated. However, our results also differ from Rost's conclusion in that we find that the degree of enzyme function conservation is not as poor as what he observed. Even in the presence of non-enzymes in the comparison, although transferring all four EC digits with above 90% accuracy requires above 60% global sequence identity, 40% global sequence identity can still be used to transfer the first three digits of an EC number with above 90% accuracy (Figure 1). Although function conservation varies among different enzyme families, these obtained thresholds of enzyme function conservation are stable and do not differ by the selection of enzyme families (Figure 3).

The effects of the definition of sequence identity on the results of enzyme function conservation

We noticed that employing different definitions of sequence identity leads to different results for enzyme function conservation. Among the five definitions of sequence identity, we found that the global identity has the strongest correlation to function conservation, while the correlation of the aligned sequence identity of the MM alignment and PSI-BLAST sequence identity to enzyme function conservation is worse than the other three definitions of sequence identity we examined (Figure 2). The reasons why the aligned sequence identity and the PSI-BLAST sequence identity has worse correlation to functional conservation can be that the length of the aligned residues in the MM alignment, or the length of the PSI-BLAST alignment might be very short; however, the corresponding sequence identity is high. For example, using PSI-BLAST to search with sequence DEOB_BUCAI (E.C.5.4.2.7, phosphopentomutase) finds the sequence PPB4_BACSU (E.C.3.1.3.1, alkaline phosphatase) with an *E*-value of 3.4, which produces a PSI-BLAST alignment with 19 residues and a reported sequence identity of 63%. In contrast, the global identity of the MM alignment between these two sequences is 14%. Thus, the presence of short alignments with high sequence identity contaminates the pool of sequence pairs having high sequence identity, and makes the extent of function conservation worse than using the global identity. On the other hand, although the aligned region in the MM alignment or the PSI-BLAST alignment might have reasonable length, it might not represent the functional region of the enzyme, or just part of the functional region. For example, the MM alignment between the sequence FENR_SYNY3 (E.C.1.18.1.2, ferredoxin–NADP(+) reductase, 413 residues) and the sequence PYS1_SYNEL (non-enzyme, phycobilisome 8.9 kDa linker polypeptide, 77 residues) has an aligned identity of 61%, with a reasonably long aligned region of 74 residues. However, because PYS1_SYNEL does not have any enzymatic activity, but is just a structural protein, the aligned region between these two proteins cannot be the functional domain of FENR_SYNY3, but might be a common structural domain for both proteins. Another example, the MM alignment between the sequence THER_BACTH (E.C.3.4.24.27, thermolysin, 316 residues) and the sequence NPRE_PAEPO (E.C.3.4.24.28, bacillolysin, 517 residues) has an aligned identity of 67%, with 310 aligned residues. However, although their aligned identity is high, because these two proteins have different substrate specificities, the substrate specificity of NPRE_PAEPO might not only be determined by this aligned region, but also by the other parts of the sequence. Overall, for complete transfer of functions between two sequences, the global identity is more accurate, especially when the function of the aligned domain between two proteins is unknown.

Why is enzyme function conservation obtained from our analysis different from Rost's results?

First, we employ the global identity to evaluate the extent of enzyme function conservation, while Rost used the identity derived from the BLAST alignment, a gapped-local alignment. As discussed above, the PSI-BLAST identity is worse for determining the extent of function conservation (Figure 1).

Second, we conduct both functional and sequence similarity classification, while Rost only did sequence similarity classification, which can result in distant homologs with different EC numbers being grouped into the same family. Thus, the selection of the representative sequences after family classification in his procedure may result in missing some enzyme functions in the constructed "unbiased" dataset. In our studies, all types of enzyme functions are compared to derive the level of enzyme function conservation.

Third, Rost employed the pairwise BLAST algorithm to compare each of 1973 sequences in the constructed unbiased dataset with all the original 26,342 sequences that are used to construct the unbiased dataset. Obviously, for each of 1973 sequences, the majority of the 26,342 sequences are unrelated and have no sequence similarity. However, for some of those unrelated sequences, the pairwise BLAST algorithm may generate very short alignments with high sequence identity, which can contaminate the pool of related sequence pairs having high sequence identity and with reasonable alignment length; thus, it can worsen the results of function conservation. Moreover, in Rost's studies, because the original dataset has a bias in the distribution of enzymes, this might still have some effects on the results of enzyme function conservation. For example, the large number of sequences from some functional diverse families might dilute the contribution of those enzymes with a small number of sequences but well conserved functions. In contrast, in our analysis, we calculate the function conservation of each enzyme family and then average the enzyme function conservation across all the enzyme families. This can guarantee that all types of enzyme functions contribute evenly to the final results.

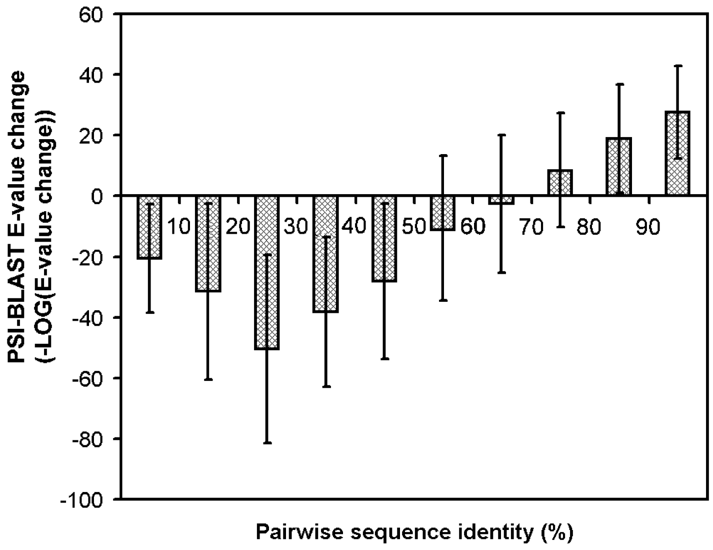

Figure 7. The effects of PSI-BLAST iteration on the PSI-BLAST *E*-value. The change of PSI-BLAST *E*-value is shown according to different levels of the global sequence identity.

The effects of PSI-BLAST iteration on transferring of enzyme functions

Our analysis suggests that the *E*-value tends to have a weaker correlation to enzyme function conservation during the iteration process of PSI-BLAST (Figure 2). To understand the reason, we track the change of a hit sequence's *E*-value from the first round to the third round of the PSI-BLAST search. Because the MM alignment between the query sequence and the hit sequence is stable and independent of the iteration process, in Figure 5, we examined the *E*-value changes in terms of the global sequence identity. It is surprising that the *E*-value changes significantly during the iteration process (Figure 7). For a hit sequence with high global identity to a query sequence (above 70%), the *E*-value increases during the iteration of PSI-BLAST search; the average increase of the *E*-value is e^{28} for sequences above 90% sequence identity to the query sequence. However, for hit sequences with low global identity to the query sequence, the *E*-value decreases significantly during the iteration; the average decrease of the *E*-value is even e^{-50} for sequence with 20–30% global identity to the query sequence, i.e. when the *E*-value of a hit sequence is 1 at the first iteration, it could be e^{-50} at the third iteration. A hit sequence with e^{-50} would usually be regarded as significant. However, because of their low global identity, their function might be very different. For example, the sequences of GOX_ASPNG (E.C.1.1.3.4, glucose oxidase) and MDL2_PRUSE (E.C.4.1.2.10, mandelonitrile lyase) have a global identity of around 23%. The PSI-BLAST search with GOX_ASPNG against the SWISSPROT + TREMBL + PDB database hits MDL2_PRUSE with an *E*-value of e^{-10} in the first round, and in the third round of the PSI-BLAST search, the *E*-value of MDL2_PRUSE becomes e^{-115}, which would be thought to be very significant. However, apparently, they do not have similar functions. Probably, the structure similarity, or evolutionary relationship between distant homologs is detected during the iterations of the PSI-BLAST search, while the functional similarity might not be directly inferred from the *E*-value when multiple iterations of PSI-BLAST search are conducted. This suggests that using the *E*-value from PSI-BLAST to infer functional similarity should be used with caution.

The application of enzyme function conservation for functional inference of unknown sequences by sequence comparison

The main goal of studying enzyme function conservation is to explore the relationship between sequence divergence and functional divergence and derive the threshold of functional inference of unknown sequences. However, because enzyme sequences in fact only account for around 20% of the total number of protein sequences in a genome, the relationship between sequence and enzyme function might not be applicable to other protein functions, such as structural proteins. Thus, the results of enzyme function conservation may only be used for enzyme-related functional inference, i.e. the functional inference when the known sequence is an enzyme. Generally, we would like to know to what extent of pairwise sequence identity could the function of an unknown sequence be transferred from an enzyme. As we have shown in Figure 1, to transfer all four digits and the first three digits of an EC number to an unknown sequence with 90% confidence, above 60% and 40% global sequence identity is required.

However, because functional conservation varies among different enzyme families, rather than using a general threshold of enzyme function conservation for functional inference, we can also take advantage of the information of individual enzyme family to which the known sequence belongs, and develop a family-specific threshold. With a jackknife analysis, we have shown that the prediction accuracy is almost always 100% and not affected by the level of sequence identity as long as it is above the family-specific threshold (see Figure 6A and B). However, because 100% conservation rate is required for establishing the threshold, this comes at a cost: those true positive sequences below this threshold cannot be uniquely identified. It is possible that more functional inference can be made if we lower the conservation rate to establish the family-specific sequence identity threshold, e.g. using 90% instead of a 100% conservation rate. However, as we have shown in Figure 4E, when the conservation rate of an individual enzyme family is not 100%, it is quite sensitive to the particular set of sequences chosen, suggesting that a substantial number of misannotations could occur if we were to loosen the criteria below 100% accuracy. Considering the big concerns about the pollution of current database by easily spread functional annotation errors,[13,16–20] it is better to have a smaller number of safe annotations than a larger number of unsure annotations.

For functional inference of those missed true positives that cannot be identified by simple sequence comparison, more sophisticated computational techniques need to be developed, such as constructing multiple sequence alignment to identify signatures that might be associated with functional information or integrating protein structural features.[40–45] However, the quality of the functional signature relies on the quality (accurate annotation) and quantity (number) of the sequences in the family. Therefore, we have applied the family-specific sequence identity threshold on KEGG annotated enzyme sequences to expand the original SWISSPROT enzyme database. We have confirmed the function of about 58% and 65% out of 25,326 KEGG enzyme sequences at the level of full EC number and the first three EC numbers, respectively, which together with the original SWISSPROT enzyme sequences can greatly facilitate the accurate functional inference of those missed true positive sequences by more sophisticated measures. A list of those sequences can be found on our website†.

Enzyme functions that are indistinguishable by sequence comparison at a high level of sequence identity

As shown in Figure 5, for enzyme–enzyme only comparisons, 90% of the classified full EC number enzyme families are absolutely conserved when the global sequence identity is above 60%. To understand more about those 10% errant enzyme families, we have listed in the Appendix all enzyme functions that are indistinguishable from each other by sequence comparison with a 60% global pairwise sequence identity cutoff. For a complete list of combinations of enzyme functions that are indistinguishable from each other at lower global sequence identity, see our website.

Above 60% global sequence identity, there are only a few enzyme functions indistinguishable from each other at the first digit of the EC number: EC 2.4.1.10-EC 3.2.1.26, EC 2.4.1.19-EC 3.2.1.1, EC 2.4.1.25-EC 3.2.1.1, and EC 2.4.1.119-EC 5.3.4.1, respectively (see Appendix). From the point of view of reaction type, these enzyme functions are very different: EC 2.4.1.10 (levansucrase), EC 2.4.1.19 (cyclomaltodextrin glucanotransferase), EC 2.4.1.25 (4-α-glucanotransferase) and EC 2.4.1.119 (dolichyl-diphosphooligosaccharide–protein glycosyltransferase) are all responsible for “hexosyl group transfer”; EC 3.2.1.26 (β-fructofuranosidase) and EC 3.2.1.1 (α-amylase) are both responsible for “hydrolysis of *O*-glycosyl bond”; EC 5.3.4.1 (protein disulfide isomerase) is responsible for “isomerization”. However, if we look at the substrates and products of the reaction that they catalyze, they do have something in common. EC 2.4.1.19 and EC 3.2.1.1 both can function to degrade starch or glycogen,[46,47] and EC 2.4.1.19 even has an alternative name: *Bacillus macerans* amylase. Similar to EC 2.4.1.19, EC 2.4.1.25 also was reported to play an important role in starch metabolism.[48] EC 2.4.1.10 and EC 3.2.1.26 both can function on sucrose to produce β-D-fructose and glucose,[49,50] though β-D-fructose is then transferred to (2,6-β-D-fructosyl)*n* by EC 2.4.1.10. For EC 2.4.1.119 and EC 5.3.4.1, it seems that these two enzyme functions are very different: EC 2.4.1.119 is a glycotransferase, while EC 5.3.4.1 catalyzes the rearrangement of disulfide bonds in proteins. However, EC 5.3.4.1 can also act as a subunit of a triacylglycerol transfer protein, which facilitates the transfer of lipids to newly synthesized core lipoproteins.[51] Thus, because the definition of EC number is arbitrary, the disagreement of EC number might not necessarily be related to functional disagreement. Nevertheless, this indeed adds to the difficulties of functional inference by computational techniques.

The majority of enzyme functions indistinguishable from each other above 60% only differ at the last digit of the EC number; i.e. mainly by the substrate of the reaction they catalyze, such as EC 1.13.11.31 (arachidonate 12-lipoxygenase) and EC 1.13.11.33 (arachidonate 15-lipoxygenase), EC 1.1.1.149 (20-α-hydroxysteroid dehydrogenase) and EC 1.1.1.50 (3-α-hydroxysteroid dehydrogenase (B-specific)), and EC 1.1.1.27 (L-lactate dehydrogenase) and EC 1.1.1.37 (malate dehydrogenase). There are also enzyme functions that differ in the cofactor of the reaction: such as EC 1.1.1.1 (alcohol dehydrogenase) and EC 1.1.1.2 (alcohol dehydrogenase (NADP +)), EC 1.2.1.3 (aldehyde dehydrogenase (NAD +)) and EC 1.2.1.5 (aldehyde dehydrogenase (NAD(P) +)), and EC 1.6.6.1 (nitrate reductase (NADH)) and EC 1.6.6.2 (nitrate reductase (NAD(P)H)). Some enzyme functions differ in the metal that play

† http://bioinformatics.buffalo.edu/resources/enzyme_conservation

an important role in the reaction, such as EC 1.11.1.13 (Mn-dependent peroxidase) and EC 1.11.1.7 (peroxidase).

All these facts taken together make function annotation complicated when employing computational approaches. As a result, for some particular type of enzyme sequences, experimental approaches are essential to further clarify their functions. On the other hand, other than simple sequence comparison, more sophisticated computation approaches, such as constructing a phylogenetic tree to highlight those residues that are associated with specific protein functions or careful examination of protein structural features, such as active-site residue clusters or characteristic surface properties, might be necessary.[40–45]

Conclusions

Here, we have classified enzyme families based on both function and sequence similarities and studied the conservation of enzyme function by averaging the function conservation across all the enzyme families. Our results suggest that for function annotation on genome sequences, a 40% sequence identity can still be used as a confident threshold to transfer the first three digits of the EC number; however, to transfer all four digits of an EC number, above 60% sequence identity is needed to have above 90% accuracy. Compared with sequence identity, the weak correlation of *E*-value at the third iteration of PSI-BLAST suggests that functional annotation based on *E*-value should be done with particular caution. By using enzyme family-specific sequence identity thresholds above which 100% conservation rate is required, it is possible to transfer functions accurately. We have applied the enzyme family-specific sequence identity threshold to the KEGG annotated enzyme sequences and about 58% and 65% of the KEGG enzyme sequences have been confirmed at the full EC number and the first three EC number level, respectively.

Methods

Collection of enzyme sequences

Following the strategy employed by Rost to collect enzyme sequences, we retrieved 33,024 sequences that have annotated EC numbers in the "DE" line of the newest version of the SWISSPROT database-sprot40.dat.[30] Then, we removed those sequences that: (1) contain EC numbers with undetermined digits (−); (2) have more than one EC number; (3) have keywords with "probable", "hypothetical", "putative", "by homology", or "by similarity"; (4) have the keyword "fragment".

These criteria remove 10,379 sequences and result in a set of 22,645 sequences that will be subject to further classification and which correspond to 1549 different EC numbers. The 22,645 enzyme sequences are strongly dominated by a few enzymes. For example, the three biggest enzyme groups, EC 1.6.5.3 (NADH dehydrogenase (ubiquinone)), EC 1.9.3.1 (cytochrome *c* oxidase), and EC 3.6.1.34 (H^+-transporting ATP synthase), have 956, 520, and 1122 sequences, respectively. Together, these three enzyme groups account for 11% of the entire number of enzyme sequences in the SWISSPROT database. However, 70% of the total EC numbers have less than ten sequences each, and together they only account for 17% of the total number of enzyme sequences in the SWISSPROT database. Consequently, the results of enzyme function conservation from a simple all-against-all pairwise sequence comparison based on the current database might be misleading.

Measurements of sequence similarity

The most widely used way to measure sequence similarity between two proteins is their sequence identity. We use two different alignment schemes to obtain sequence identity: (1) the global alignment; and (2) the PSI-BLAST alignment. The global pairwise sequence alignment is conducted by align0, a program that employs the Myers/Miller (MM) global alignment algorithm,[52] without penalizing for end-gaps. The PSI-BLAST alignment is a gapped-local alignment,[10] which is generated automatically between the query sequence and the hit sequence after a PSI-BLAST search. Because the definition of sequence identity is arbitrary, for global alignment, we explore four definitions of sequence identity: the percentage of identical residues between two proteins in terms of (1) the alignment length (including gaps), the global identity, (2) the length of the protein with the greater number of residues, the big identity, (3) the length of the protein with the smaller number of residues, the small identity, and (4) the length of aligned residues (not counting gaps), the aligned identity.[19] The PSI-BLAST sequence identity is simply adapted from its output, which is calculated by the percentage of identical residues between two proteins in terms of the length of the gapped-local alignment. In all cases, the sequence identity is the ratio of the number of identical aligned residues divided by the total number of residues given by one of the above definitions.

As a second means of purely sequence-based family classification, we also employ the statistical significant score (*E*-value) given by PSI-BLAST as a measure of sequence similarity.

Enzyme family classification

We define an enzyme family as a family of evolutionarily related sequences whose function is conserved. Based on this definition, we first conduct functional classification, by collecting enzyme sequences into groups according to different levels of enzyme function, i.e. the full four digits or the first three digits of their EC number. Then, we conduct a sequence similarity classification within each functional group. For sequence similarity classification, we calculate the HSSP-distance of all sequence pairs within each functional group with the HSSP-threshold *J* set to be -2^{19}. Then, we conduct a complete linkage analysis to cluster enzyme sequences into families by requiring that the HSSP-distance of all sequence pairs in the family must be larger than 0. Following this strategy, we have classified the selected 22,645 SWISSPROT enzyme sequences into 2431 and 1794 families, with full four digits EC number and the first three digits EC number, respectively.

Typically, enzymes that catalyze the same reaction often show significant sequence and structural similarity; However, although they have the identical EC number, some enzymes may have evolved independently of one another and have different catalytic mechanisms and little structure similarity.[53–55] As a result, for a particular enzyme function, there might be more than one family. For example, there are three families associated with EC 1.1.1.1 (alcohol dehydrogenase), which correspond to three Pfam families: Adh_short, Fe_ADH and Adh_zinc,[27] respectively. By running protein structure prediction with threading, these three families hit three different protein structure templates that have no sequence similarity to each other (our unpublished data).

Evaluation of enzyme function conservation

After enzyme family classification, we need to calculate the extent of sequence and function conservation of each enzyme family. To do this, we first collect a list of sequence pairs between sequences from the enzyme family and all the other SWISSPROT enzyme sequences (excluding those removed enzyme sequences) at a certain level of sequence identity, e.g. 40–50% global identity. Then, we compare the functional match of each sequence pair to determine how many of the collected sequence pairs at a particular level of sequence identity are from the same enzyme family:

$$\text{“degree of enzyme family function conservation} \\ = (\text{number of pairs with same function})/ \\ (\text{number of all pairs})\text{”}$$

Then, we average the enzyme function conservation across all the enzyme families to obtain the average degree of enzyme function conservation by:

$$\text{“average degree of enzyme function conservation} \\ = (\sum \text{degree of enzyme family function conservation})/ \\ (\text{number of enzyme families at that level} \\ \text{of sequence identity})\text{”}$$

In practice, it is computationally expensive to compare each enzyme sequence with all the other 22,645 enzyme sequences by the MM algorithm. On the other hand, this is also unnecessary. For standard procedure of functional annotation of unknown sequences, usually the first step is a database search, such as PSI-BLAST, to collect possible homologous sequences. Then functional inference from homology is made if possible. Apparently, for each enzyme sequence, only a few out of the 22,645 enzyme sequences can be picked up by PSI-BLAST search and only these sequences are really what we need to pay attention to. Therefore, for each enzyme sequence, we run a PSI-BLAST search against the SWISSPROT + TREMBL + PDB database (over 820,000 sequences in total) with three iterations and pick up only those SWISSPROT sequences (excluding those removed enzyme sequences) with an *E*-value less than 10 for comparison with the query enzyme sequence by the MM algorithm. During the iteration process of PSI-BLAST search, to avoid results drifting from the inclusion of false positive sequences, we set the iteration parameter (*H*-value) to be e^{-10}, the same parameter used by Rost.[19] The hit SWISSPROT sequences can have any function: either an enzyme that belongs to the selected 22,645 SWISSPROT sequences, or a non-enzyme that has no "EC" identifier in the SWISSPROT functional annotation. Thus, after the PSI-BLAST search, for each enzyme family, we obtain two pools of sequence pairs: one that includes only enzyme–enzyme pairs, and the other that includes not only enzyme–enzyme but also enzyme–non-enzyme pairs. These two pools of sequence pairs correspond to two degrees of enzyme function conservation, with the results from comparisons of both enzyme–enzyme and enzyme–non-enzyme pairs representing a more realistic situation when functional transfer is to be assessed.

Bootstrap analysis to assess the stability of the threshold of enzyme function conservation

To assess whether the average enzyme function conservation differs by the selection of enzyme families, we conduct a bootstrap analysis. We randomly choose a subset of enzyme families, ranging from 10%, 20%,…to 90% of the total number of enzyme families with full four EC digits and the first three EC digits, respectively. Then we calculate the degree of enzyme function conservation by averaging the conservation rate across all enzyme families of each subset. This procedure is repeated 100 times. Then for each subset, the variation of function conservation at different level of sequence identity is calculated by:

$$\text{“conservation variation} \\ = \text{the maximum conservation rate} \\ -\text{the minimum conservatorium rate”}$$

Bootstrap analysis to assess the stability of the conservation rate of individual enzyme families

To evaluate whether the conservation rate of individual enzyme families differs by the selection of enzyme sequences, we select those full EC number enzyme families that have more than 30 sequences each (167 enzyme families in total). For each enzyme family, we randomly choose a subset of enzyme sequences, ranging from 10%, 20%,…to 90% of the total, and calculate its function conservation rate. This procedure is repeated 100 times. Then, for each enzyme family's subset, we obtain a conservation variation at different levels of sequence identity, by:

$$\text{“conservation variation} \\ = \text{the maximum conservation rate} \\ -\text{ the minimum conservatorium rate”}$$

which is averaged across all enzyme families.

To study whether the conservation variation of individual enzyme families is affected by its conservation rate, we pick up two sets of enzyme families: one is composed of enzyme families that are absolutely conserved above 40% (137 enzyme families in total); another is composed of enzyme families that are absolutely conserved above 50% but not above 40% sequence identity (12 enzyme families in total). Then, following the above procedures, we calculate the average conservation variation with respect to individual enzyme family by using different subsets of enzyme sequences at different sequence identity levels.

Jack-knife analysis to test the usefulness of enzyme family-specific sequence identity threshold in functional inference

The enzyme family-specific sequence identity threshold is defined as the sequence identity above which the conservation rate is 100% and also at which a conservation rate is available. For example, an enzyme family might have a conservation rate of 100% and 50% at 60–70% and 30–40% sequence identity level, respectively, and no conservation rate available at any other sequence identity levels. Then the family-specific sequence identity threshold is 60%. Here, the conservation rates are obtained when both enzyme–enzyme and enzyme–non-enzyme are included in the calculation. To apply this threshold for functional inference, we calculate the global sequence identity between the "unknown" and "known" enzyme sequence. If it is bigger than the threshold of the enzyme family to which the known sequence belongs, then the function of the known sequence is transferred to the unknown sequence. Otherwise, no function is transferred.

To validate this procedure, we conduct a jack-knife analysis. From the original 22,645 SWISSPROT enzyme sequences, we randomly choose a subset of 60%, 70%, 80% and 90% of the total number of enzyme sequences as the training set with the remaining 40%, 30%, 20% and 10%, respectively of the total number of enzyme sequences as the test set, respectively. From each training set, we conduct enzyme family classification and calculate the conservation rate of each enzyme family to derive the family-specific sequence identity threshold. Then, for each test sequence, we run a PSI-BLAST search against the SWISSPROT + TREMBLE + PDB database with three iterations to pick up the functionally known training sequences. After PSI-BLAST search of all test sequences, we pool the test-training sequence pairs together according to different level of sequence identity. For each test-training sequence pair, we attempt the functional inference of the test sequence by applying the family-specific sequence identity threshold of the training sequence. If function can be transferred, this sequence pair is called transferable. If the sequence pair is non-transferable, however, the test and training sequence do have the same function, it is called missed. Then, at different levels of global sequence identity, we calculate:

"percentage of transferable sequence pairs

= (number of transferable sequence pairs)/

(number of all sequence pairs)"

"prediction accuracy = (number of transferable

sequence pairs with right prediction)/

(number of transferable sequence pairs)"

and:

"percentage of missed sequence pairs

= (number of missed sequence pairs)/

(number of all sequence pairs with same function)"

For example, for one subset with 70% of the total number of enzyme sequences as the training set and the remaining 30% sequences as the test set, at 30–40% sequence identity level, we collect 45,362 test-training sequences pairs in total. Among those sequence pairs, functional inference can be done at the level of the full four EC numbers for 23,165 sequence pairs (transferable) and they all have right predictions. However, there are actually 39,001 sequence pairs with same function. In other words, 15,836 sequence pairs are missed. Thus, the percentage of transferable sequence pairs, the prediction accuracy, and the percentage of missed sequence pairs at 30–40% sequence identity level are 51%, 100%, and 41%, respectively. This procedure is repeated 100 times to obtain the average of the three percentages at different levels of sequence identity.

The application of the enzyme family-specific sequence identity threshold to evaluate the annotation of KEGG enzyme sequences

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database for systematic analysis of gene functions in terms of the networks of genes and molecules, i.e. pathways including both metabolic pathways and regulatory pathways.[56] For each metabolic pathway, KEGG has a reference pathway including all possible enzymatic reactions and also a number of organism-specific reconstructed pathways composed of annotated enzymes in each corresponding genome. However, sequence comparison is the main annotation method used by KEGG. To evaluate the annotation of KEGG sequences, we have selected 25,326 KEGG sequences that have been annotated with a unique EC number by KEGG and are not identical with any enzyme sequences in the SWISSPROT database. Then the function of those 25,326 KEGG sequences is predicted by employing the enzyme family-specific sequence identity threshold calculated from the 22,645 SWISSPROT enzyme sequences. After prediction, we calculate the percentage of transferable sequence pairs and the prediction accuracy at different level of global sequence identity. The prediction accuracy is calculated by comparing the KEGG annotation with the predicted function.

Acknowledgements

Thanks to Dr Arakaki Adrian and Dr Kihara Daisuke for stimulating discussions and comments. This research was supported, in part, by NIH grant no. GM-48835 of the Division of General Medical Sciences of the National Institutes of Health and the Oishei Foundation.

References

1. Andrade, M. A. & Sander, C. (1997). Bioinformatics: from genome data to biological knowledge. *Curr. Opin. Biotechnol.* **8**, 675–683.
2. Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. & Yuan, Y. (1998). Predicting function: from genes to genomes and back. *J. Mol. Biol.* **283**, 707–725.
3. Galperin, M. Y. & Koonin, E. V. (1999). Functional genomics and enzyme evolution. Homologous and

analogous enzymes encoded in microbial genomes. *Genetica*, **106**, 159–170.

4. Huynen, M., Snel, B., Lathe, W. & Bork, P. (2000). Exploitation of gene context. *Curr. Opin. Struct. Biol.* **10**, 366–370.
5. Koonin, E. V., Tatusov, R. L. & Rudd, K. E. (1996). Protein sequence comparison at genome scale. *Methods Enzymol.* **266**, 295–322.
6. Skolnick, J., Fetrow, J. S. & Kolinski, A. (2000). Structural genomics and its importance for gene function analysis. *Nature Biotechnol.* **18**, 283–287.
7. Skolnick, J. & Fetrow, J. S. (2000). From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends Biotechnol.* **18**, 34–39.
8. Pearson, W. R. (2000). Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* **132**, 185–219.
9. Altschul, S. F. & Gish, W. (1996). Local alignment statistics. *Methods Enzymol.* **266**, 460–480.
10. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
11. Brenner, S. E., Chothia, C. & Hubbard, T. J. (1997). Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.* **7**, 369–376.
12. Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature*, **357**, 543–544.
13. Brenner, S. E. (1999). Errors in genome annotation. *Trends Genet.* **15**, 132–133.
14. Devos, D. & Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends Genet.* **17**, 429–431.
15. Brenner, S. E., Chothia, C. & Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
16. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2002). Plasticity of enzyme active sites. *Trends Biochem. Sci.* **27**, 419–426.
17. Todd, A. E., Orengo, C. A. & Thornton, J. M. (1999). Evolution of function in superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143.
18. Wilson, C. A., Kreychman, J. & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297**, 233–249.
19. Rost, B. (2002). Enzyme function less conserved than anticipated. *J. Mol. Biol.* **318**, 595–608.
20. Devos, D. & Valencia, A. (2000). Practical limits of function prediction. *Proteins: Struct. Funct. Genet.* **41**, 98–107.
21. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
22. Innis, C. A., Shi, J. & Blundell, T. L. (2000). Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis. *Protein Eng.* **13**, 839–847.
23. Madabushi, S., Yao, H., Marsh, M., Kristensen, D. M., Philippi, A., Sowa, M. E. & Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139–154.
24. del Sol Mesa, A., Pazos, F. & Valencia, A. (2003). Automatic methods for predicting functionally important residues. *J. Mol. Biol.* **326**, 1289–1302.
25. Hannenhalli, S. S. & Russell, R. B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.* **303**, 61–76.
26. Bairoch, A. (2000). The ENZYME database in 2000. *Nucl. Acids Res.* **28**, 304–305.
27. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R. *et al.* (2002). The Pfam protein families database. *Nucl. Acids Res.* **30**, 276–280.
28. Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acids Res.* **30**, 264–267.
29. Orengo, C. A., Bray, J. E., Buchan, D. W., Harrison, A., Lee, D., Pearl, F. M. *et al.* (2002). The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics*, **2**, 11–21.
30. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
31. Attwood, T. K., Croning, M. D., Flower, D. R., Lewis, A. P., Mabey, J. E. & Scordis, P. (2000). PRINTS-S: the database formerly known as PRINTS. *Nucl. Acids Res.* **28**, 225–227.
32. Mewes, H. W., Frishman, D., Gruber, C., Geier, B., Haase, D. & Kaps, A. (2000). MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.* **28**, 37–40.
33. Ursing, B. M., van Enckevort, F. H., Leunissen, J. A. & Siezen, R. J. (2002). EXProt: a database for proteins with an experimentally verified function. *Nucl. Acids Res.* **30**, 50–51.
34. Wu, C. H., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y. & Hu, Z. Z. (2002). The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucl. Acids Res.* **30**, 35–37.
35. Yang, A. S. & Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *J. Mol. Biol.* **301**, 679–689.
36. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94.
37. Lindahl, E. & Elofsson, A. (2000). Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* **295**, 613–625.
38. Schneider, R. & Sander, C. (1996). The HSSP database of protein structure–sequence alignments. *Nucl. Acids Res.* **24**, 201–205.
39. Olmea, O., Rost, B. & Valencia, A. (1999). Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.* **293**, 1221–1239.
40. Landgraf, R., Xenarios, I. & Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307**, 1487–1502.
41. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). Evolutionarily conserved Galphabetagamma binding surfaces support a model of the G protein–receptor complex. *Proc. Natl Acad. Sci. USA*, **93**, 7507–7511.
42. Sjolander, K. (1998). Phylogenetic inference in protein superfamilies: analysis of SH2 domains. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 165–174.
43. Armon, A., Graur, D. & 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.

44. Fetrow, J. S. & Skolnick, J. (1998). Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* **281**, 949–968.
45. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* **5**, 1001–1013.
46. Bovetto, L. J., Backer, D. P., Villette, J. R., Sicard, P. J. & Bouquelet, S. J. (1992). Cyclomaltodextrin glucanotransferase from *Bacillus circulans* E 192. I. Purification and characterization of the enzyme. *Biotechnol. Appl. Biochem.* **15**, 48–58.
47. Abdel-Naby, M. A. (1993). Immobilization of *Aspergillus niger* NRC 107 xylanase and beta-xylosidase, and properties of the immobilized enzymes. *Appl. Biochem. Biotechnol.* **38**, 69–81.
48. Pazur, J. H. & Okada, S. (1968). The isolation and mode of action of a bacterial glucanosyltransferase. *J. Biol. Chem.* **243**, 4732–4738.
49. Chambert, R. & Petit-Glatron, M. F. (1991). Polymerase and hydrolase activities of *Bacillus subtilis* levansucrase can be separately modulated by site-directed mutagenesis. *Biochem. J.* **279**, 35–41.
50. Zarate, V. & Belda, F. (1996). Characterization of the heterologous invertase produced by *Schizosaccharomyces pombe* from the *SUC2* gene of *Saccharomyces cerevisiae*. *J. Appl. Bacteriol.* **80**, 45–52.
51. Freedman, R. B., Hirst, T. R. & Tuite, M. F. (1994). Protein disulphide isomerase: building bridges in protein folding. *Trends Biochem. Sci.* **19**, 331–336.
52. Myers, E. W. & Miller, W. (1988). Optimal alignments in linear space. *Comput. Appl. Biosci.* **4**, 11–17.
53. Doolittle, R. F. (1994). Convergent evolution: the need to be explicit. *Trends Biochem. Sci.* **19**, 15–18.
54. Galperin, M. Y., Walker, D. R. & Koonin, E. V. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome Res.* **8**, 779–790.
55. Smith, M. W., Feng, D. F. & Doolittle, R. F. (1992). Evolution by acquisition: the case for horizontal gene transfers. *Trends Biochem. Sci.* **17**, 489–493.
56. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* **27**, 29–34.

Appendix A

Table A1. Enzyme functions indistinguishable from each other by sequence identity (above 60% global identity)

Enzyme 1	Enzyme 2	Example gene of enzyme 1	Example gene of enzyme 2	ID (%)
Global sequence identity 90–100%				
1.13.11.31 (Arachidonate 12-lipoxygenase)	1.13.11.33 (Arachidonate 15-lipoxy genase)	LOX2_RABIT	LOX1_RABIT	99
2.4.1.19 (Cyclomaltodextrin glucanotransferase)	3.2.1.1 (α-Amylase)	CDGU_BACCI	AMYR_BACS8	97
3.1.3.6 (3′-nucleotidase)	3.1.4.21 (Aspergillus nuclease S_1)	NUP3_PENSQ	NUP1_PENCI	99
3.4.21.35 (Tissue kallikrein)	3.4.21.77 (Semenogelase)	KLK3_MACMU	KLK3_HUMAN	90
Global sequence identity 80–90%				
1.11.1.13 (Mn-dependent peroxidase)	1.11.1.7 (Peroxidase)	PEM4_PHACH	PEM1_PHACH	83
1.13.11.31 (Arachidonate 12-lipoxygenase)	1.13.11.33 (Arachidonate 15-lipoxygenase)	LOX2_BOVIN	LOX1_HUMAN	86
1.14.13.30 (Leukotriene-B4 20-monooxygenase)	1.14.99.8 (Unspecific monooxygenase)	CPF2_HUMAN	CPF8_HUMAN	81
1.14.14.1 (Unspecific monooxygenase)	1.14.15.3 (Alkane-1 monooxygenase)	CP44_RABIT	CP45_RABIT	86
2.4.1.119 (Dolichyl-diphosphooligosaccharide-protein glycosyltransferase)	5.3.4.1 (Protein disulfide isomerase)	GSBP_CHICK	PDI_BOVIN	83
2.4.1.19 (Cyclomaltodextrin glucanotransferase)	3.2.1.1 (α-Amylase)	CDGT_BACS0	AMYR_BACS8	86
2.4.1.203 (Zeatin O-β-D-glucosyltransferase)	2.4.1.204 (Zeatin *O*-β-D-xylosyltransferase)	ZOG_PHALU	ZOX_PHAVU	85
3.4.21.35 (Tissue kallikrein)	3.4.21.54 (Gamma-renin)	KLK8_MOUSE	KLKG_MOUSE	80
3.4.22.15 (Cathepsin L)	3.4.22.43 (Cathepsin V)	CATL_PIG	CSL2_HUMAN	80
3.4.22.25 (Glycyl endopeptidase)	3.4.22.30 (Caricain)	PAP4_CARPA	PAP3_CARPA	80
3.4.24.27 (Thermolysin)	3.4.24.28 (Bacillolysin)	THER_BACST	NPRE_BACCL	85
3.4.24.42 (Atrolysin C)	3.4.24.46 (Adamalysin)	HRT2_CRORU	ADAM_CROAD	80
Global sequence identity 70–80%				
1.1.1.188 (Prostaglandin-F synthase)	1.3.1.20 (*trans*-1,2-Dihydrobenzene-1,2-diol dehydrogenase)	PGF2_BOVIN	DBDD_HUMAN	77
1.13.11.31 (Arachidonate 12-lipoxygenase)	1.13.11.33 (Arachidonate 15-lipoxygenase)	LOX2_RAT	LOX1_HUMAN	74
1.14.13.30 (Leukotriene-B4 20-monooxygenase)	1.14.99.8 (Unspecific monooxygenase)	CPF2_HUMAN	CPF1_RAT	78
1.14.14.1 (Unspecific monooxygenase)	1.14.15.3 (Alkane-1 monooxygenase)	CP44_RABIT	CP41_RAT	73
1.2.1.12 (Glyceraldehyde 3-phosphate dehydrogenase (phosphorylating))	1.2.1.59 (Glyceraldehyde-3-phosphate dehydrogenase (NAD(P)) (phosphorylating))	G3P2_ANASQ	G3P2_SYNY3	74

(*continued*)

Table A1 continued

Enzyme 1	Enzyme 2	Example gene of enzyme 1	Example gene of enzyme 2	ID (%)
1.4.1.2 (Glutamate dehydrogenase)	1.4.1.4 (Glutamate dehydrogenase (NADP$^+$))	DHE2_PORGI	DHE4_BACFR	72
1.6.6.1 (Nitrate reductase (NADH))	1.6.6.2 (Nitrate reductase (NAD(P)H))	NIA1_ARATH	NIA_BETVE	72
2.3.1.74 (Naringenin-chalcone synthase)	2.3.1.95 (Trihydroxystilbene synthase)	CHS1_CAMSI	THS1_ARAHY	74
2.4.1.119 (Dolichyl-diphosphooligosaccharide-protein glycosyltransferase)	5.3.4.1 (Protein disulfide isomerase)	GSBP_CHICK	PDI_CHICK	73
2.4.1.19 (Cyclomaltodextrin glucanotransferase)	3.2.1.1 (α-Amylase)	CDGT_BACCI	AMYR_BACS8	71
2.4.1.25 (4-α-Glucanotransferase)	3.2.1.1 (α-Amylase)	MALQ_PYRKO	AMYA_PYRAB	71
3.2.1.39 (Glucan *endo*-1,3-β-D-glucosidase)	3.2.1.73 (Licheninase)	E132_SOLTU	GUB_NICPL	70
3.2.1.8 (*endo*-1,4-β-Xylanase)	3.2.1.91 (Cellulose 1,4-beta-cellobiosidase)	XYNA_THESA	XYNX_CLOTM	75
3.4.21.35 (Tissue kallikrein)	3.4.21.54 (Gamma-renin)	KLK1_MOUSE	KLKG_MOUSE	69
3.4.21.35 (Tissue kallikrein)	3.4.21.77 (Semenogelase)	KLK2_HUMAN	KLK3_HUMAN	77
3.4.22.15 (Cathepsin L)	3.4.22.43 (Cathepsin V)	CATL_BOVIN	CSL2_HUMAN	78
3.4.22.2 (Papain)	3.4.22.30 (Caricain)	PAPA_CARPA	PAP3_CARPA	72
3.4.24.17 (Stromelysin 1)	3.4.24.22 (Stromelysin 2)	MM03_HORSE	MM10_HUMAN	74
3.4.24.27 (Thermolysin)	3.4.24.28 (Bacillolysin)	THER_BACST	NPRS_BACST	70
3.5.1.1 (Asparaginase)	3.5.1.38 (Glutaminase-(asparagin-)ase)	ASPG_PSEFL	ASPQ_PSES7	79
3.6.1.34 (H(+)-transporting two-sector ATPase)	3.6.3.15 (Sodium-transporting two-sector ATPase)	ATPB_ANASP	ATPB_PROMO	70
Global sequence identity 60–70%				
1.1.1.1 (Alcohol dehydrogenase)	1.1.1.2 (Alcohol dehydrogenase (NADP$^+$))	ADH1_ALLMI	ADH4_RANPE	62
1.1.1.149 (20-α-Hydroxysteroid dehydrogenase)	1.1.1.188 (Prostaglandin-F synthase)	PE2R_RAT	PGF2_BOVIN	67
1.1.1.149 (20-α-Hydroxysteroid dehydrogenase)	1.1.1.50 (3-α-Hydroxysteroid dehydrogenase (B-specific))	PE2R_RAT	DIDH_RAT	67
1.1.1.149 (20-α-Hydroxysteroid dehydrogenase)	1.3.1.20 (*trans*-1,2-Dihydrobenzene-1,2-diol dehydrogenase)	PE2R_RAT	DBDD_HUMAN	68
1.1.1.188 (Prostaglandin-F synthase)	1.1.1.50 (3-α-Hydroxysteroid dehydrogenase (B-specific))	PGF2_BOVIN	DIDH_RAT	69
1.1.1.206 (Tropine dehydrogenase)	1.1.1.236 (Tropinone reductase)	TRN1_DATST	TRN2_DATST	61
1.1.1.27 (L-lactate dehydrogenase)	1.1.1.37 (Malate dehydrogenase)	LDH_BOTBR	MDH_RHILV	68
1.1.1.50 (3-α-Hydroxysteroid dehydrogenase (B-specific))	1.3.1.20 (*trans*-1,2-Dihydrobenzene-1,2-diol dehydrogenase)	DIDH_RAT	DBDD_HUMAN	69
1.13.11.31 (Arachidonate 12-lipoxygenase)	1.13.11.33 (Arachidonate 15-lipoxygenase)	LOXE_MOUSE	LOX1_HUMAN	66
1.14.14.1 (Unspecific monooxygenase)	1.14.15.3 (Alkane-1 monooxygenase)	CP48_RAT	CP42_RAT	69
1.2.1.12 (Glyceraldehyde 3-phosphate dehydrogenase (phosphorylating))	1.2.1.59 (Glyceraldehyde-3-phosphate dehydrogenase (NAD(P)) (phosphorylating))	G3PA_ARATH	G3P2_SYNY3	60
1.2.1.22 (Lactaldehyde dehydrogenase)	1.1.1.70 (Aldehyde dehydrogenase (NAD$^+$))	ALDB_ECOLI	DHA2_ALCEU	68
1.2.1.3 (Aldehyde dehydrogenase (NAD$^+$))	1.2.1.5 (Aldehyde dehydrogenase (NAD(P)$^+$))	DHA1_BOVIN	DHA6_HUMAN	68
1.3.1.19 (*cis*-1,2-Dihydrobenzene-1,2-diol dehydrogenase)	1.3.1.56 (*cis*-2,3-Dihydrobiphenyl-2,3-diol dehydrogenase)	BNZE_PSEPU	BPHB_COMTE	61
1.3.5.1 (Succinate dehydrogenase (ubiquinone))	1.3.99.1 (Succinate dehydrogenase)	DHSB_RECAM	DHSB_RICPR	63
1.3.5.1 (Succinate dehydrogenase (ubiquinone))	1.3.99.1 (Succinate dehydrogenase)	DHSA_DROME	DHSA_RICPR	60
1.5.99.10 (Dimethylamine dehydrogenase)	1.5.99.7 (Trimethylamine dehydrogenase)	DHDM_HYPSX	DHTM_METME	63
1.6.6.1 (Nitrate reductase (NADH))	1.6.6.2 (Nitrate reductase (NAD(P)H))	NIA1_ARATH	NIA7_HORVU	63
2.1.1.53 (Putrescine N-methyltransferase)	2.5.1.16 (Spermidine synthase)	PMT2_TOBAC	SPD1_PEA	61
2.3.1.74 (Naringenin-chalcone synthase)	2.3.1.95 (Trihydroxystilbene synthase)	CHS1_GERHY	THS1_ARAHY	69
2.4.1.10 (Levansucrase)	3.2.1.26 (β-Fructofuranosidase)	SACB_ZYMMO	INVB_ZYMMO	61
2.4.1.19 (Cyclomaltodextrin glucanotransferase)	3.2.1.1 (α-Amylase)	CDG1_PAEMA	AMYR_BACS8	66
2.4.1.25 (4-α-Glucanotransferase)	3.2.1.1 (α-Amylase)	MALQ_PYRKO	AMYA_PYRFU	69
2.7.1.11 (6-phosphofructokinase)	2.7.1.90 (Diphosphate-fructose-6-phosphate 1-phosphotransferase)	K6P1_STRCO	PFP_AMYME	68
2.7.7.7 (DNA-directed DNA polymerase)	2.7.7.7 (DNA-directed DNA polymerase)	DPOD_PLAFK	DPOD_PLAFK	64
2.8.1.1 (Thiosulfate sulfurtransferase)	2.8.1.2 (3-Mercaptopyruvate sulfurtransferase)	THTR_MOUSE	THTM_RAT	60
3.2.1.1 (α-Amylase)	3.2.1.98 (Glucan 1,4-alpha-maltohexaosidase)	AMY_BACAM	AMT6_BACS7	63

(*continued*)

Table A1 continued

Enzyme 1	Enzyme 2	Example gene of enzyme 1	Example gene of enzyme 2	ID (%)
3.2.1.21 (β-Glucosidase)	3.2.1.86 (6-Phospho-beta-glucosidase)	BGL1_BACSU	ABGA_CLOLO	63
3.2.1.39 (Glucan *endo*-1,3-beta-D-glucosidase)	3.2.1.73 (Licheninase)	E13B_HEVBR	GUB_NICPL	64
3.4.17.1 (Carboxypeptidase A)	3.4.17.15 (Carboxypeptidase A2)	CBP1_HUMAN	CBP2_RAT	62
3.4.21.1 (Chymotrypsin)	3.4.21.32 (Brachyurin)	CTR1_PENVA	COGS_UCAPU	64
3.4.21.2 (Chymotrypsin C)	3.4.21.71 (Pancreatic elastase II)	CLCR_HUMAN	EL2A_HUMAN	62
3.4.21.35 (Tissue kallikrein)	3.4.21.54 (Gamma-renin)	KLK1_RAT	KLKG_MOUSE	64
3.4.21.35 (Tissue kallikrein)	3.4.21.77 (Semenogelase)	KLK1_HUMAN	KLK3_HUMAN	60
3.4.21.74 (Venombin A)	3.4.21.95 (Snake venom factor V activator)	VSP1_AGKCO	VSPA_DABRU	61
3.4.22.2 (Papain)	3.4.22.25 (Glycyl endopeptidase)	PAPA_CARPA	PAP4_CARPA	68
3.4.22.2 (Papain)	3.4.22.6 (Chymopapain)	PAPA_CARPA	PAP2_CARPA	61
3.4.22.25 (Glycyl endopeptidase)	3.4.22.6 (Chymopapain)	PAP4_CARPA	PAP2_CARPA	67
3.4.22.30 (Caricain)	3.4.22.6 (Chymopapain)	PAP3_CARPA	PAP2_CARPA	65
3.4.23.38 (Plasmepsin I)	3.4.23.39 (Plasmepsin II)	PLM1_PLAFA	PLM2_PLAFA	65
3.4.24.15 (Thimet oligopeptidase)	3.4.24.16 (Neurolysin)	MEPD_HUMAN	NEUL_RABIT	60
3.4.24.17 (Stromelysin 1)	3.4.24.22 (Stromelysin 2)	MM03_HORSE	MM10_MOUSE	68
3.4.24.42 (Atrolysin C)	3.4.24.44 (Atrolysin E)	HRTD_CROAT	HRTE_CROAT	63
3.4.24.53 (Trimerelysin II)	3.4.24.72 (Fibrolase)	HR2_TRIFL	FIBR_AGKCO	61
3.6.1.34 (H(+)-transporting two-sector ATPase)	3.6.3.15 (Sodium-transporting two-sector ATPase)	ATPA_ANASP	ATPA_PROMO	60
3.6.3.10 (Hydrogen/potassium-exchanging ATPase)	3.6.1.37 (Sodium/potassium-exchanging ATPase)	ATHA_CANFA	A1A1_ANGAN	60
4.1.1.25 (Tyrosine decarboxylase)	4.1.1.27 (Aromatic-L-amino-acid decarboxylase)	TYD2_PETCR	TYD1_PAPSO	65
4.2.99.9 (*O*-succinylhomoserine (thiol)-lyase)	4.4.1.8 (Cystathionine beta-lyase)	METB_HELPJ	METC_LACLA	63
6.2.1.4 (Succinate-CoA ligase (GDP-forming))	6.2.1.5 (Succinate-CoA ligase (ADP-forming))	SUCA_DICDI	SUCD_COXBU	60

Edited by B. Honig

(Received 16 December 2002; received in revised form 8 May 2003; accepted 26 August 2003)