Prediction of Function Divergence in Protein Families Using the Substitution Rate Variation Parameter Alpha

Saraswathi Abhiman, Carsten O. Daub, and Erik L. L. Sonnhammer

Center for Genomics and Bioinformatics, Karolinska Institutet, Stockholm, Sweden

Protein families typically embody a range of related functions and may thus be decomposed into subfamilies with, for example, distinct substrate specificities. Detection of functionally divergent subfamilies is possible by methods for recognizing branches of adaptive evolution in a gene tree. As the number of genome sequences is growing rapidly, it is highly desirable to automatically detect subfamily function divergence.

To this end, we here introduce a method for large-scale prediction of function divergence within protein families. It is called the alpha shift measure (ASM) as it is based on detecting a shift in the shape parameter (alpha [α]) of the substitution rate gamma distribution. Four different methods for estimating α were investigated. We benchmarked the accuracy of ASM using function annotation from Enzyme Commission numbers within Pfam protein families divided into subfamilies by the automatic tree-based method BETE.

In a test using 563 subfamily pairs in 162 families, ASM outperformed functional site-based methods using rate or conservation shifting (rate shift measure [RSM] and conservation shift measure [CSM]). The best results were obtained using the "GZ-Gamma" method for estimating α. By combining ASM with RSM and CSM using linear discriminant analysis, the prediction accuracy was further improved.

Introduction

One of the great challenges in the postgenomic era is to understand the evolution of functional properties among members of the same protein family. New functions can be created either by domain rearrangements or by substituting functionally important residues. Without detailed knowledge of the protein structure and the functional role of each site, it is difficult to distinguish neutral substitutions from ones that modify the function substantially. Yet, predicting a shift in function from sequence data alone would be very useful for large-scale protein annotation. With the wealth of sequence data available today, many protein families are large enough for a statistical analysis of substitution patterns indicative of function shift.

The traditional approach to detect function shift, or adaptive evolution, in sequences is to use the ratio of nonsynonymous to synonymous substitutions (Ka/Ks, also called dN/dS) (Yang 1998) given a protein-coding DNA sequence alignment. However, this approach is limited for very closely related species as silent sites quickly lose signal as they become saturated with substitutions over long evolutionary timescales (Smith JM and Smith NH 1996; Yang and Nielsen 2000).

This problem can be ameliorated by using protein multiple sequence alignments to detect substitution rate variations (Lichtarge et al. 1996; Armon et al. 2001; Blouin et al. 2003; Landau et al. 2005). This has been approached using a probabilistic frameworks on inferred phylogenetic trees (Hannenhalli and Russell 2000; Knudsen and Miyamoto 2001; Truong and Ikura 2002; Gribaldo et al. 2003; Knudsen et al. 2003; Kalinina et al. 2004; Soyer and Goldstein 2004). In a recent analysis of protein enzyme families (Abhiman and Sonnhammer 2005b), 2 such methods were benchmarked in a large-scale test and were shown to be useful not only to infer sites responsible for functional specificity but also for predicting function shift between subfamilies in a protein family.

Functional constraints also affect the rate of amino acid substitutions in protein sequences. It can be hypothesized that a particular function is associated with a certain characteristic distribution of substitution rates and that a function shift induces a change in this distribution. Substitution rates are usually modeled by a gamma distribution (Uzzell and Corbin 1971; Golding 1983; Holmquist et al. 1983; Tamura and Nei 1993; Yang 1993; Gu et al. 1995) that is characterized by a shape parameter alpha (α). This parameter can be estimated from a given sequence alignment with several different methods that can be broadly classified as parsimony-based and maximum likelihood (ML)–based methods. The parsimony-based methods (Uzzell and Corbin 1971; Holmquist et al. 1983; Tamura and Nei 1993; Sullivan et al. 1995; Tourasse and Gouy 1997) using the framework of Fitch (1971) are simple and fast but tend to overestimate α. The minimum number of changes at each site estimated by parsimony will follow a Poisson distribution if the substitution rate is constant across sites or a negative binomial distribution if the rates across sites are gamma distributed. The method of Moments tends to underestimate both the mean and variance of number of changes at each site and in turn overestimate α. The method of Sullivan et al. (1995) uses numerical maximization of a log-likelihood function to estimate α and has also been shown to overestimate α. Both of these methods do not account for unequal branch lengths in the tree.

Approaches that use a combination of likelihood and parsimony methods have also been proposed (Yang and Kumar 1996; Gu and Zhang 1997; Zhang and Gu 1998) to overcome some of the limitations. Yang and Kumar (1996) use the number of differences instead of the number of changes at each site to reduce the overestimation of α. Unequal branch lengths in the tree are not supported, but all branch lengths are set to the average length. The method of Gu and Zhang (1997) estimates the number of substitutions at each site using a likelihood approach that fully takes branch lengths into account and corrects for multiple substitutions. The parameter α is estimated

Present address: Stockholm Bioinformatics Center, AlbaNova University Center, Stockholm University, Stockholm, Sweden.

Key words: protein evolution, adaptive evolution, enzyme, protein function, protein subfamily, substitution rates, gamma distribution, alpha parameter, function shift.

E-mail: Abhiman.Saraswathi@cgb.ki.se.

Mol. Biol. Evol. 23(7):1406–1413. 2006
doi:10.1093/molbev/msl002
Advance Access publication May 3, 2006

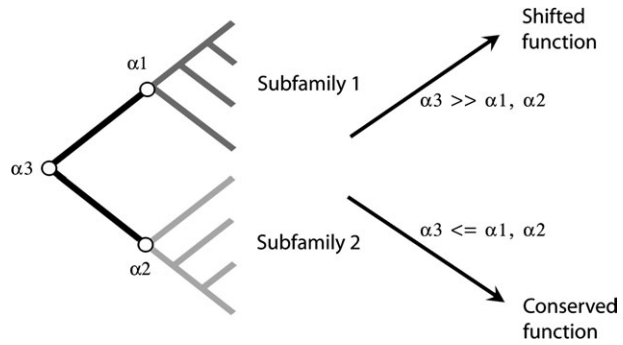

FIG. 1.—A schematic gene tree showing the division of a protein family into subfamilies. For the ASM calculation, the parameter α was estimated at the ancestral nodes for subfamily 1 (α_1) and subfamily 2 (α_2) and at the node that joins them (α_3). Function shift between the 2 subfamilies can be inferred if α_3 is much larger than α_1 and α_2 (See also Siltberg and Liberles 2002).

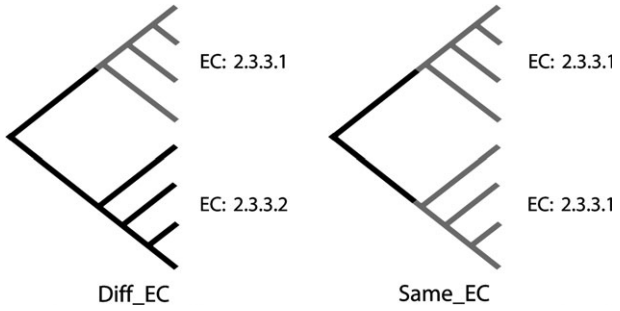

FIG. 2.—Schematic showing the assignment of subfamily pairs into the classes "different function" (Diff_EC) or "same function" (Same_EC) based on EC numbers.

using an ML approach, which uses the expected number of substitutions inferred at each site. This method has been shown to give lower estimates of α compared with the above-mentioned methods but still gives higher estimates than the ML method of Yang. The ML methods (Yang 1994; Gu et al. 1995; Kelly and Rice 1996) using the framework of Felsenstein (Felsenstein 1981) are unbiased but take huge amount of computational time.

The α parameter is inversely related to the rate variation among sites. A small α indicates a high degree of rate variation where most sites are highly conserved, whereas a large α indicates a low degree of rate variation where most sites evolve around the neutral rate (Yang 1996). In general, the mean substitution rate of a protein is inversely correlated to the rate variation (Zhang and Gu 1998), that is, slowly evolving proteins are more likely to have a high rate variation among sites. This property is captured well by the α parameter.

Estimating the α parameter for all branches in a gene tree has been used to estimate the likelihood at each branch that adaptive evolution has taken place (Siltberg and Liberles 2002). It can also be used to estimate function shift between 2 subfamilies. If the combined tree have a much larger α value than the individual subfamilies, this indicates a shift in function between them (Gu 1999; Gaucher et al. 2001; Gu 2001; Siltberg and Liberles 2002) (see fig. 1). The rationale for this is that the lower α parameters of the subfamilies indicate that they have diversified to become more specific than the ancestral family (e.g., substrate specificity of enzymes). This can be considered a nonstationary covarion process (i.e., variable positions under one clade are not the same as those of another clade) (Galtier 2001; Pupko and Galtier 2002) and is similar to the Type I and Type II functional divergence (Gu 1999) at the subfamily level instead of at individual alignment positions.

In the study presented here, we introduce and evaluate a novel measure for the function divergence between protein subfamilies, the alpha shift measure (ASM), which is based on the substitution rate distribution shape parameter α. We demonstrate the ability of the ASM to correctly predict cases of function divergence in a large-scale test of discriminating between protein subfamilies with same and different functions. Finally, we combine the ASM with previously proposed functional site-based measures to further increase the prediction accuracy.

Methods and Data

Protein Subfamilies with Same and Different Functions

Protein domain family multiple sequence alignments were downloaded from the Pfam database version 12 (Bateman et al. 2004). The full sequence alignments of the Pfam families were divided into subfamilies using the software BETE version 1.1 (Sjolander 1998) that uses relative entropy as a distance metric for phylogenetic tree estimation and cuts the tree into subfamilies based on an encoding cost function. Each subfamily pair was then assigned to 1 of 2 possible function categories based on the Enzyme Commission (EC) number annotation (Bairoch 2000) of the constituent sequences: one category containing subfamily pairs where all sequences in both the subfamilies were annotated with the same EC number (Same_EC) and a second category where the annotations were the same within the subfamilies but different between them (Diff_EC) (see fig. 2). The data set consisted of 563 subfamily pairs, with 129 pairs assigned to the Diff_EC category (corresponding to 35 Pfam families) and 434 pairs assigned to the Same_EC category (corresponding to 127 Pfam families). In total, 514 subfamilies from 162 protein families were used. To exclude proteins with incomplete annotation, we applied 2 additional conditions: all the proteins in a subfamily should have EC number annotation at all 4 levels and also have the same domain architecture. The process of data generation has been described in more detail elsewhere (Abhiman and Sonnhammer 2005b). The data set here is smaller than in the previous study because subfamily pairs that failed to be processed by at least one of the α estimation methods were discarded. The subfamily pairs were sorted based on the size of smaller subfamily in each category and divided into 3 equal parts. The top quantile (large subfamily pairs) was used to analyze the effect of subfamily size on the predictive performance.

Modification of Alignments and Generation of Tree Topologies

As the methods we used for estimating the substitution rate parameter α do not allow gaps in the protein sequence alignments, we modified the alignments obtained from Pfam. In a first step, we removed whole sequences if a

sequence contained more than 50% gaps. For the remaining sequences, we removed all the columns of an alignment that contained any gaps. The tree topologies, that were required by all 4 methods, were estimated with the neighbor-joining method available in the ClustalW software (Thompson et al. 1994). Nested and shuffled subtree topologies were generated by custom written scripts to analyze the effect of wrong tree topologies.

Estimation of α Parameter and ASM

For each protein subfamily pair, we estimated the substitution rate parameter α for both subfamilies as well as for the combined alignment of the 2 subfamilies, resulting in 3 individual substitution rate parameter values α1, α2, and α3 (fig. 1). We employed 4 different methods to estimate the substitution rate parameter α: the ML-based method "GZ-Gamma" (G) (Gu and Zhang 1997) and 3 parsimony-based methods, namely the method of Moments (M) (Tamura and Nei 1993), the method of Sullivan (S) (Sullivan et al. 1995), and the method by Yang (Y) (Yang and Kumar 1996), all 3 implemented in the "pamp" program of the PAML 3.13 package (Yang 1997). The ASM was then computed for each protein subfamily pair as:

$$\mathrm{ASM} = \alpha 3 - \left(\frac{\alpha 1 + \alpha 2}{2}\right)$$

A number of other equations for expressing the relation between α3 and {α1; α2} were also explored, for example, taking the ratio instead of the difference. We chose the equation above because it performed slightly better than other variants and because a difference is expected to behave more stably than a ratio. In some cases when the estimation of α failed for a certain method and the ASM could not be calculated, we excluded the subfamily pair from any further analysis. Typical reasons for such failures were small or short subfamilies containing less than 4 sequences or too few alignment positions to make meaningful α estimates.

Rate Shift Measure and Conservation Shift Measure

Amino acid sites that are evolving with different evolutionary rates in 2 subfamilies of the same protein family can be described as rate shifting sites (RSS). We identified the RSS for all protein families and their corresponding subfamily pairs with the LRT software (Knudsen and Miyamoto 2001). Sites showing different conservation patterns between 2 subfamilies of the same protein family—conserved within both subfamilies but with 2 different amino acids—can be described as conservation shifting sites (CSS). We identified these CSS in a way described elsewhere (Abhiman and Sonnhammer 2005b). The rate shift measure (RSM) and conservation shift measure (CSM) were then determined by normalizing the RSS and CSS to the alignment lengths.

Evaluation of Predictions

In the presented study, we evaluated the ability of ASM to predict the function divergence of protein subfamilies derived from the same protein family and compared it with previously proposed methods, the RSM and the CSM. For this, we calculated the ASM for each subfamily pair and its corresponding combined alignment. By taking the knowledge about the subfamily functions into account, we subdivided the ASM values into 2 classes, one with conserved and one with diverged function (as described in detail above). For a given ASM threshold value, we then counted the number of protein subfamily pairs that were predicted to fall into the category of the different function (ASM $>$ threshold) and at the same time were annotated through their EC numbers to fall into the different function category (true positives). In the same way, we counted the number of true negative protein subfamily pairs (TN, same function according to ASM and EC numbers), false positives (FP, same function according to ASM but different function according to EC number), and false negatives (FN, different function according to ASM but same function according to EC number). We applied the same procedure to RSM and CSM values and calculated sensitivity and specificity values according to

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

Combination of Methods

We also tested whether the predictions obtained from the ASM, RSM, and CSM can be enhanced by integrating all these measures into one and compared the combined measure with the individual ones. This combined measure was obtained by a linear combination of the 3 measures, where the influence of each measure was determined by a linear discriminant analysis (LDA) approach. For this, we calculated the specificities and sensitivities from the individual and combined measures and evaluated their accuracies with a cross-validation procedure: 1) the data set containing all protein subfamily pairs was randomly split up into 2 equally large parts (test and training sets); 2) the classification functions were determined for the training set by LDA; and 3) sensitivities and specificities were calculated for the test set. This whole procedure was repeated a thousand times.

Results

We tested the ASM's ability to predict a function shift between 2 protein subfamilies by applying it to the large-scale benchmark based on 162 protein families in Pfam. Before describing this test, however, we first evaluate and characterize the 4 methods for estimating α.

Global Comparison of Substitution Rate Distribution Parameter Values

For all 514 subfamilies, we calculated the substitution rate distribution parameter α using the 4 methods

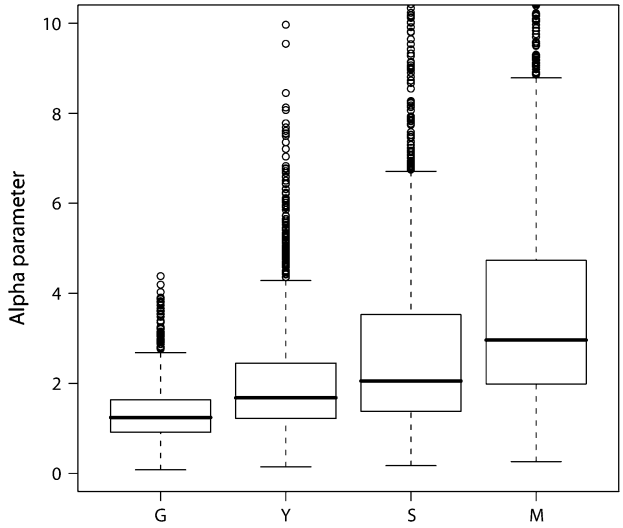

FIG. 3.—The distribution of α parameter values estimated for all protein families by 4 methods (G, GZ-Gamma; Y, Yang; S, Sullivan; and M, Moments). The upper and lower borders of a box correspond to the first and third quartile, respectively. The middle bar shows the median. The upper bar is placed at third quartile + 1.5 × (box height) or the maximum value, whereas the lower bar is placed at first quartile − 1.5 × (box height) or the minimum value. Data points outside the upper or lower bar are shown explicitly as circles.

GZ-Gamma (G), Yang (Y), Sullivan (S), and Moments (M). By comparing the median α values obtained by the different methods, we observed that Moments produced the highest α values, followed by Sullivan, Yang, and GZ-Gamma (fig. 3). We noted that Moments and Sullivan sometimes produce very high α values.

Another way to compare the methods is to calculate pairwise correlation coefficients of the computed α values for all subfamilies (table 1). This revealed that Moments and Sullivan produce very similar results ($r = 0.96$), whereas GZ-Gamma is very different from these 2 ($r \sim 0.3$). The Yang method is in the middle between these extremes, with $r \sim 0.6$ to the other methods.

Evaluation of the ASM

With the data set of subfamily pairs with known function shifts and nonshifts, we could evaluate the performance of the proposed ASM method. The first question was to compare the different methods for calculating α. For each subfamily pair in the test set, we calculated 4 ASM values using the G, Y, S, and M methods.

The accuracy of each ASM variant was determined as described in Methods and Data for a set of thresholds to cover the entire sensitivity/specificity range and were plotted as receiver operating characteristic (ROC) curves (fig. 4*A*). This shows that all ASM variants are able to classify the protein subfamily pairs much better than random. The ASM based on the GZ-Gamma method outperformed the other methods over the whole sensitivity/specificity range and thus appears generally the most accurate. In other words, it seems that the low α estimates of the GZ-Gamma produce the most accurate ASM classifier. This observation, however, does not necessarily mean that the underlying α values are more accurately measured by the GZ-Gamma method.

Table 1
Pairwise Correlation of α Parameter Estimation by 4 Different Methods

	Y	S	M
G	0.57	0.28	0.31
Y		0.64	0.63
S			0.96

NOTE.—For all 514 protein subfamilies in the test set, the substitution rate distribution shape parameter α was estimated using the methods G, Y, S, and M. Pearson's correlation coefficient *r* is shown for each pairwise comparison.

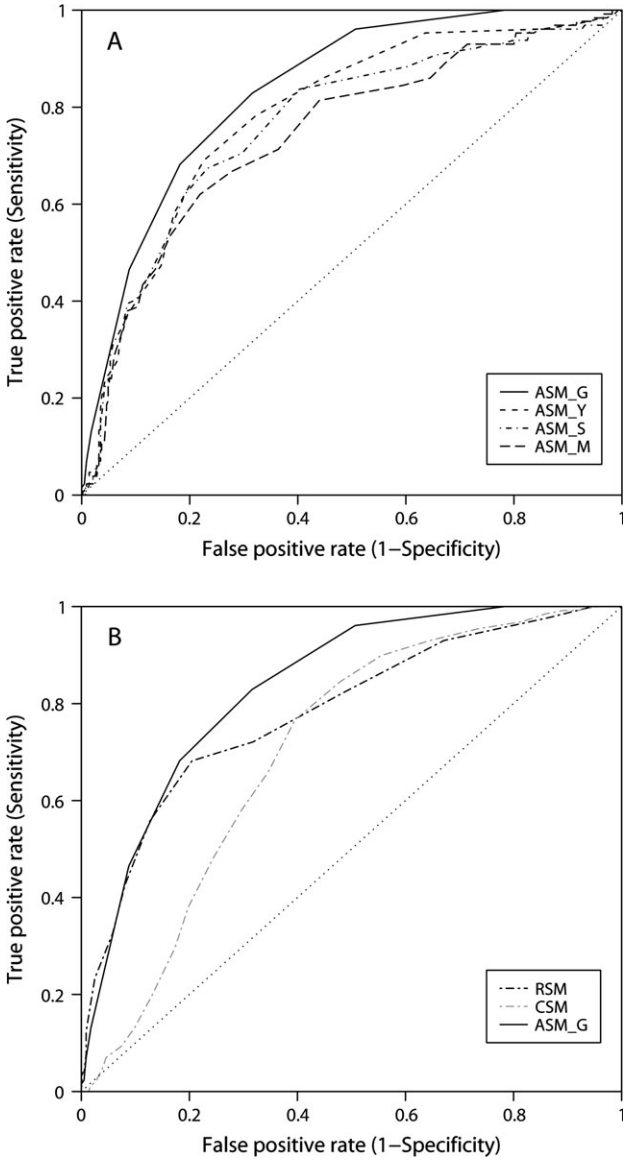

FIG. 4.—Performance comparison of function shift predictors using the large test set based on EC number annotation in Pfam families. ROC curves show the ability of the ASM to discriminate between positive (Diff_EC) and negative (Same_EC) cases of function divergence. (*A*) Performance of the 4 ASM variants. ASM_G, ASM using α estimated by the GZ-Gamma method; ASM_Y, by Yang; S, by Sullivan; ASM_M, by Moments. (*B*) Performance comparison of the best ASM method (ASM_G) to the functional site-based methods RSM and CSM. The dashed diagonal line indicates the performance of a random classifier.

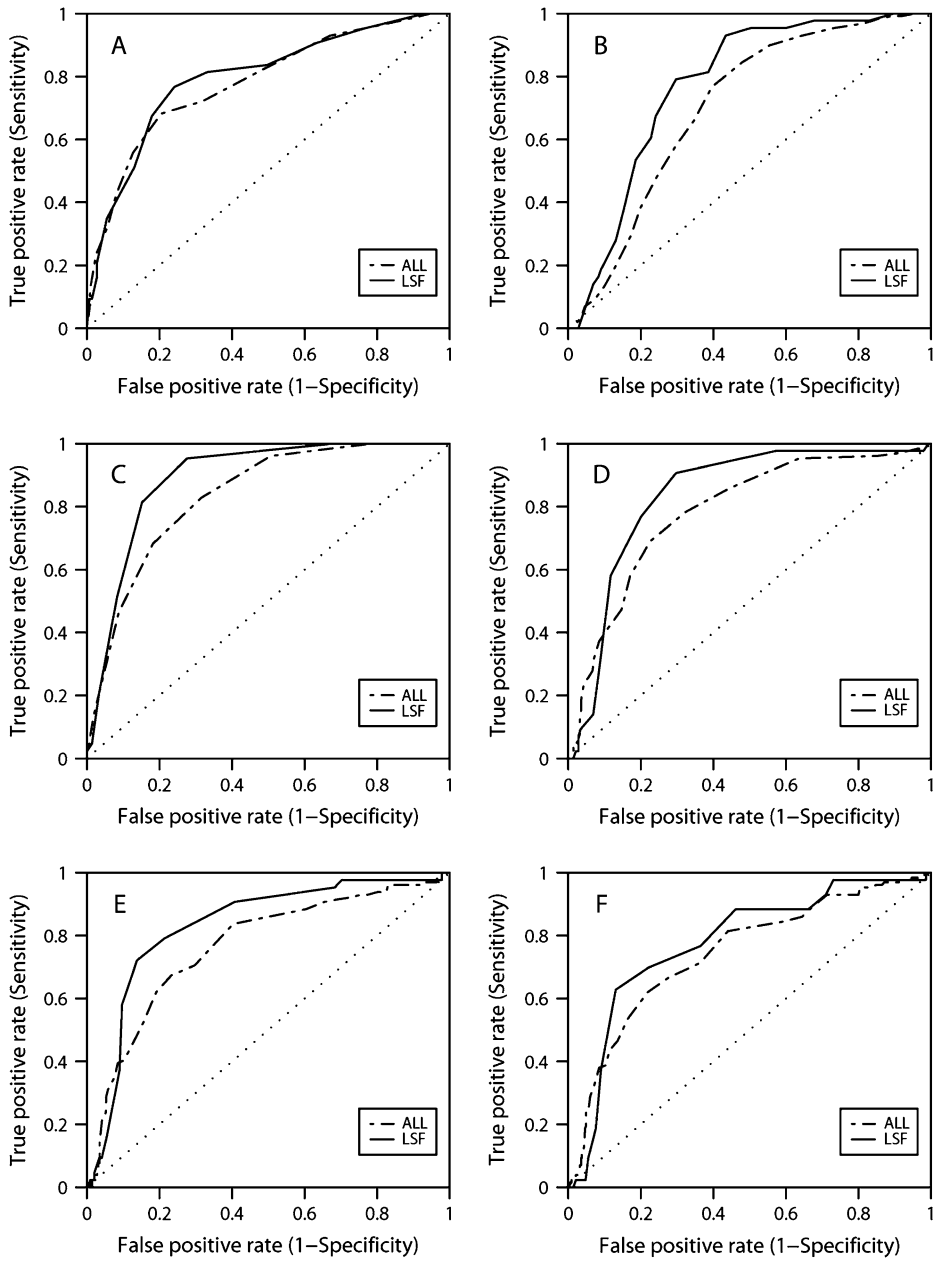

FIG. 5.—Analysis of the effect of subfamily size on performance. ROC curves show that performance is improved for the larger subfamily pairs (LSF) compared with all subfamily pairs (ALL). LSF correspond to the top third based on the number of members in the smaller subfamily. (*A*) RSM, (*B*) CSM, (*C*) ASM_G, (*D*) ASM_Y, (*E*) ASM_S, and (*F*) ASM_M. The dashed diagonal line indicates the performance of a random classifier.

Comparison of ASM to RSM and CSM

Next, we compared the best ASM variant (ASM_G) with previously presented function shift predictors—the RSM and the CSM (Abhiman and Sonnhammer 2005b). These methods are based on the fraction of sites that appear to have shifted function. The comparison was done using the same ROC curves as above. As seen in figure 4*B*, at low false positive rate (high specificity), the RSM is considerably more sensitive than the CSM and parallels the ASM. However, above a false positive rate of 0.2, ASM clearly outperforms RSM, which above 0.4 performs roughly equally with CSM.

Influence of Subfamily Size and Subtree Topology

In order to determine the importance of protein subfamily size for prediction accuracy, we grouped the subfamily pairs into 3 equally sized quantiles according to the number of sequences the smaller subfamily contained (large, medium, and small). We then evaluated the ASM together with the RSM and CSM on only the group with the largest protein subfamily pairs (fig. 5). We observed increased prediction accuracy by all classifiers compared with the results obtained for all protein subfamily pairs, underscoring the importance of using many sequences for these types of analyses. We noted, however, that RSM

performance was almost as good for all subfamilies as for large ones only.

We have also analyzed the influence that the topology of the subtrees have on the prediction accuracy. On a randomly chosen subset of the data (1/3 of the data set), we estimated the α parameters and in turn the ASM measure by using shuffled tree topology instead of the true tree topology. We observed that the predictive performance of ASM was unchanged over wide range of specificity and sensitivity values (Supplementary Figure 1, Supplementary Material online).

Predictions by Combining ASM, RSM, and CSM

The ASM, RSM, and CSM are using different signals for function shift prediction and may therefore complement each other. A simple way to exploit this to improve accuracy is to combine them using LDA. All methods individually and combined were evaluated in terms of sensitivity and specificity in cross-validation tests (fig. 6) on the complete data set.

Of the 3 individual measures, the ASM gave the highest sensitivity values (fig. 6*A*), whereas the CSM gave the lowest. The LDA combination of all 3 methods increased the median sensitivity by almost 4 percentage points compared with ASM. Looking at specificity (fig. 6*B*) gives a somewhat different picture. Here the RSM is best, even slightly more specific than the LDA combination of the 3 methods. LDA is thus mainly of use for increasing sensitivity, that is, the accuracy of correctly detecting shifted function.

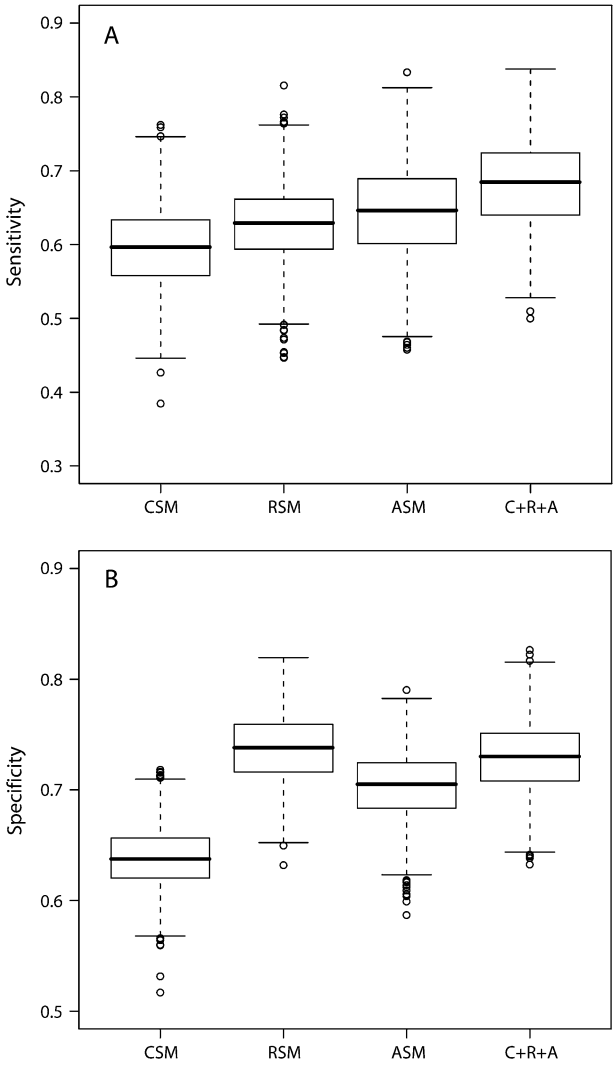

FIG. 6.—Cross-validation analysis of the prediction accuracies for CSM, RSM, and ASM (GZ-Gamma) and the LDA combination of them (C + R + A). (*A*) Sensitivity. (*B*) Specificity. The upper and lower borders of a box correspond to the first and thirrd quartile, respectively. The middle bar shows the median. The upper bar is placed at the third quartile + 1.5 × (box height) or the maximum value, whereas the lower bar is placed at first quartile − 1.5 × (box height) or the minimum value. Data points outside the upper or lower bar are shown explicitly as circles.

Discussion

This paper presents a novel approach, the ASM, to predict the shift of protein function between protein subfamilies. Previously, we had explored other approaches based on sequence conservation signals derived from protein multiple sequence alignments. We here proposed and evaluated a complimentary approach based on detecting changes in the distribution of amino acid substitution rates that may accompany a function shift. We showed that the ASM outperforms the previously proposed predictors, RSM and CSM.

For the calculation of the ASM presented in this study, we employed 4 different methods to estimate the substitution rate parameter α. Our results on the α values estimated from these methods confirmed previous comparisons of α estimation methods (Gu and Zhang 1997), that is, that the parsimony-based methods (Moments, Sullivan, and Yang) tend to give higher α estimates than the ML-based methods (GZ-Gamma and ML-Yang [Yang 1996]). Simulation studies have shown (Gu and Zhang 1997) that ML-Yang gives slightly better estimates than the GZ-Gamma method but at the same time performed hundred times worse in terms of runtime. Due to the large number of alignments used in our study and the practical limitation of the ML-Yang method, we were not able to include this method in our analysis.

The ASM is related to the RSM in the sense that both measures are based on the substitution rates obtained from protein alignments. However, both measures cover different aspects of the 2 alignments under consideration. The RSM identifies sites between 2 subfamilies that evolve with different rates and calculates the rate shift based only on these sites. The ASM, in contrast, incorporates the whole subfamily by calculating substitution rates for all the sites of the alignment. In addition, the ASM also includes the ancestral family that both subfamilies arose from.

When comparing the ASM to the previously proposed CSM and RSM, the ASM outperforms the other 2 measures in its ability to predict function shift. We showed that combining the ASM with the RSM and CSM leads to improved prediction accuracy. It is important to keep in mind, however, that in contrast to the CSM and RSM approaches, it is not possible with the ASM to actually pinpoint the residues responsible for the shift of function. Another

limitation is that for certain cases, it is not possible to use the ASM because there is not enough data to estimate α, yet it may still be possible to use the CSM and RSM. Hence, the combination of all 3 measures gives the most complete and accurate picture of the protein family under consideration.

We employed EC numbers to assign known function annotations to the test set protein families used in our analysis. Even though EC numbers are recognized as high-quality function classifiers, one should be aware that there are some disadvantages (Babbitt 2003). Enzymes with the same EC number occasionally have different functions, and enzymes with different EC numbers are sometimes highly similar in sequence and in the reactions they catalyze (Nahum and Riley 2001; Tian and Skolnick 2003). This dependency on EC numbers probably affected our results negatively, although we estimate the false negative rate (Same_EC cases that really are Diff_EC) to be at most a few percent. This would mean that an improved accuracy could be achieved by our measures if trained on more pure data sets. Another potential caveat is that we only train our method on enzymes, yet only around 20% of the proteins in a genome are enzymes. However, we believe that our method will be general enough for most globular proteins as the structural constraints are similar for all such proteins. One could argue that proteins with protein–protein binding moieties (Sjolander 1998) have characteristics quite similar to enzymes.

Even though many methods are currently available for predicting functional divergence between nucleotide or protein sequences, they have been applied only to small number of families with closely related sequences. Large-scale comparative analysis of these methods is necessary to investigate their applicability for genome annotation efforts. To enable the application of the new ASM method and its combination with the RSM and the CSM for the biological community, we are aiming towards integrating it into the publicly available FunShift database (Abhiman and Sonnhammer 2005a). This way a more refined function annotation for many protein families can be achieved, which will be welcomed in the light of novel sequences coming from new genome projects.

Supplementary Material

Supplementary Figure 1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by Pfizer Corporation and the Swedish Knowledge Foundation.

Literature Cited

Abhiman S, Sonnhammer EL. 2005a. FunShift: a database of function shift analysis on protein subfamilies. Nucleic Acids Res 33:D197–200.

Abhiman S, Sonnhammer EL. 2005b. Large-scale prediction of function shift in protein families with a focus on enzymatic function. Proteins 60:758–68.

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 307:447–63.

Babbitt PC. 2003. Definitions of enzyme function for the structural genomics era. Curr Opin Chem Biol 7:230–7.

Bairoch A. 2000. The ENZYME database in 2000. Nucleic Acids Res 28:304–5.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. 2004. The Pfam protein families database. Nucleic Acids Res 32:D138–41.

Blouin C, Boucher Y, Roger AJ. 2003. Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. Nucleic Acids Res 31: 790–7.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–76.

Fitch WM. 1971. Rate of change of concomitantly variable codons. J Mol Evol 1:84–96.

Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18:866–73.

Gaucher EA, Miyamoto MM, Benner SA. 2001. Function-structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. Proc Natl Acad Sci USA 98: 548–52.

Golding GB. 1983. Estimates of DNA and protein sequence divergence: an examination of some assumptions. Mol Biol Evol 1:125–42.

Gribaldo S, Casane D, Lopez P, Philippe H. 2003. Functional divergence prediction from evolutionary analysis: a case study of vertebrate hemoglobin. Mol Biol Evol 20:1754–9.

Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16:1664–74.

Gu X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. Mol Biol Evol 18: 453–64.

Gu X, Fu YX, Li WH. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol Biol Evol 12:546–57.

Gu X, Zhang J. 1997. A simple method for estimating the parameter of substitution rate variation among sites. Mol Biol Evol 14:1106–13.

Hannenhalli SS, Russell RB. 2000. Analysis and prediction of functional sub-types from protein sequence alignments. J Mol Biol 303:61–76.

Holmquist R, Goodman M, Conroy T, Czelusniak J. 1983. The spatial distribution of fixed mutations within genes coding for proteins. J Mol Evol 19:437–48.

Kalinina OV, Mironov AA, Gelfand MS, Rakhmaninova AB. 2004. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. Protein Sci 13:443–56.

Kelly C, Rice J. 1996. Modeling nucleotide evolution: a heterogeneous rate analysis. Math Biosci 133:85–109.

Knudsen B, Miyamoto MM. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc Natl Acad Sci USA 98:14512–7.

Knudsen B, Miyamoto MM, Laipis PJ, Silverman DN. 2003. Using evolutionary rates to investigate protein functional divergence and conservation. A case study of the carbonic anhydrases. Genetics 164:1261–9.

Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N. 2005. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 33:W299–302.

Lichtarge O, Bourne HR, Cohen FE. 1996. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257:342–58.

Nahum LA, Riley M. 2001. Divergence of function in sequence-related groups of Escherichia coli proteins. Genome Res 11:1375–81.
Pupko T, Galtier N. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc Biol Sci 269:1313–6.
Siltberg J, Liberles DA. 2002. A simple covarion-based approach to analyze nucleotide substitution rates. J Evol Biol 15:588–94.
Sjolander K. 1998. Phylogenetic inference in protein superfamilies: analysis of SH2 domains. Proc Int Conf Intell Syst Mol Biol 6:165–74.
Smith JM, Smith NH. 1996. Synonymous nucleotide divergence: what is "saturation"? Genetics 142:1033–6.
Soyer OS, Goldstein RA. 2004. Predicting functional sites in proteins: site-specific evolutionary models and their application to neurotransmitter transporters. J Mol Biol 339:227–42.
Sullivan J, Holsinger KE, Simon C. 1995. Among-site rate variation and phylogenetic analysis of 12S rRNA in sigmodontine rodents. Mol Biol Evol 12:988–1001.
Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–26.
Thompson JD, Higgins DG, Gibson TJ. 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 22:4673–80.
Tian W, Skolnick J. 2003. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333:863–82.
Tourasse NJ, Gouy M. 1997. Evolutionary distances between nucleotide sequences based on the distribution of substitution rates among sites as estimated by parsimony. Mol Biol Evol 14:287–98.
Truong K, Ikura M. 2002. Identification and characterization of subfamily-specific signatures in a large protein superfamily by a hidden Markov model approach. BMC Bioinformatics 3:1.
Uzzell T, Corbin KW. 1971. Fitting discrete probability distributions to evolutionary events. Science 172:1089–96.
Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396–401.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–14.
Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367–72.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–6.
Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–73.
Yang Z, Kumar S. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol Biol Evol 13:650–9.
Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32–43.
Zhang J, Gu X. 1998. Correlation between the substitution rate and rate variation among sites in protein evolution. Genetics 149:1615–25.

Michele Vendruscolo, Associate Editor

Accepted April 27, 2006