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

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 non-synonymous 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 (\(\alpha\)). 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 underestimate \(\alpha\). 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 \(\alpha\). The method of Sullivan et al. (1995) uses numerical maximization of a log-likelihood function to estimate \(\alpha\) and has also been shown to overestimate \(\alpha\). 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 \(\alpha\). 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 \(\alpha\) is estimated...
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 $\alpha$ 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 $\alpha$ parameter is inversely related to the rate variation among sites. A small $\alpha$ indicates a high degree of rate variation where most sites are highly conserved, whereas a large $\alpha$ 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 $\alpha$ parameter.

Estimating the $\alpha$ 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 $\alpha$ 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 $\alpha$ 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 covariance 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 $\alpha$. 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 $\alpha$ 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 $\alpha$ 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 $\alpha$ Parameter and ASM

For each protein subfamily pair, we estimated the substitution rate parameter $\alpha$ for both subfamilies as well as for the combined alignment of the 2 subfamilies, resulting in 3 individual substitution rate parameter values $\alpha_1$, $\alpha_2$, and $\alpha_3$ (fig. 1). We employed 4 different methods to estimate the substitution rate parameter $\alpha$: 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:

$$\text{ASM} = \alpha_3 - \left(\frac{\alpha_1 + \alpha_2}{2}\right)$$

A number of other equations for expressing the relation between $\alpha_3$ and $\{\alpha_1; \alpha_2\}$ 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 $\alpha$ 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 $\alpha$ 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

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$
$$\text{Specificity} = \frac{TN}{TN + FP}$$

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 $\alpha$.

Global Comparison of Substitution Rate Distribution Parameter Values

For all 514 subfamilies, we calculated the substitution rate distribution parameter $\alpha$ using the 4 methods
GZ-Gamma (G), Yang (Y), Sullivan (S), and Moments (M). By comparing the median \(a\) values obtained by the different methods, we observed that Moments produced the highest \(a\) values, followed by Sullivan, Yang, and GZ-Gamma (fig. 3). We noted that Moments and Sullivan sometimes produce very high \(a\) values.

Another way to compare the methods is to calculate pairwise correlation coefficients of the computed \(a\) values for all subfamilies (table 1). This revealed that Moments and Sullivan produce very similar results (\(r \approx 0.96\)), whereas GZ-Gamma is very different from these 2 (\(r \approx 0.3\)). The Yang method is in the middle between these extremes, with \(r \approx 0.6\) 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 \(a\). 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. 4A). 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 \(a\) estimates of the GZ-Gamma produce the most accurate ASM classifier. This observation, however, does not necessarily mean that the underlying \(a\) values are more accurately measured by the GZ-Gamma method.

### Table 1

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>S</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0.57</td>
<td>0.28</td>
<td>0.31</td>
</tr>
<tr>
<td>Y</td>
<td>0.64</td>
<td></td>
<td>0.63</td>
</tr>
<tr>
<td>S</td>
<td></td>
<td>0.96</td>
<td></td>
</tr>
</tbody>
</table>

*Note:* For all 514 protein subfamilies in the test set, the substitution rate distribution shape parameter \(a\) was estimated using the methods G, Y, S, and M. Pearson’s correlation coefficient \(r\) is shown for each pairwise comparison.
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 4B, 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 $\alpha$ 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. 6A), 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. 6B) 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.

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 $\alpha$. Our results on the $\alpha$ values estimated from these methods confirmed previous comparisons of $\alpha$ estimation methods (Gu and Zhang 1997), that is, that the parsimony-based methods (Moments, Sullivan, and Yang) tend to give higher $\alpha$ 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 \( \alpha \), 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. 2005b. Large-scale prediction of function shift in protein families with a focus on enzymatic function. Proteins 60:758–68.


Michele Vendruscolo, Associate Editor

Accepted April 27, 2006