Sequence analysis

Determining functional specificity from protein sequences

Jason E. Donald and Eugene I. Shakhnovich*

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA

Received on February 8, 2005; revised on March 15, 2005; accepted on March 17, 2005

Advance Access publication March 29, 2005

ABSTRACT

Motivation: Given a large family of homologous protein sequences, many methods can divide the family into smaller groups that correspond to the different functions carried out by proteins within the family. One important problem, however, has been the absence of a general method for selecting an appropriate level of granularity, or size of the groups.

Results: We propose a consistent way of choosing the granularity that is independent of the sequence similarity and sequence clustering method used. We study three large, well-investigated protein families: basic leucine zippers, nuclear receptors and proteins with three consecutive C2H2 zinc fingers. Our method is tested against known functional information, the experimentally determined binding specificities, using a simple scoring method. The significance of the groups is also measured by randomizing the data. Finally, we compare our algorithm against a popular method of grouping proteins, the TRIBE-MCL method. In the end, we determine that dividing the families at the proposed level of granularity creates very significant and useful groups of proteins that correspond to the different DNA-binding motifs. We expect that such groupings will be useful in studying not only DNA binding but also other protein interactions.

Contact: shakhnovich@chemistry.harvard.edu

Supplementary information: The supplementary material contains: experimental binding specificities, a list of proteins in the proposed clusters, a table listing the percentage of proteins with binding data and from humans, visualizations of nuclear receptor and zinc finger motifs. We expect that such groupings will be useful in studying not only DNA binding but also other protein interactions.

1 INTRODUCTION

One of the major goals of theoretical biology is to organize protein sequences by their functions. Protein function, however, can be defined at several different levels of detail. For example, many computational methods (Abascal and Valencia, 2002; Enright and Ouzounis, 2000; Enright et al., 2002; Krause and Vingron, 1998; Kriventseva et al., 2001; Yona et al., 1999), termed family-based methods, can separate huge numbers of protein sequences into distinct protein families. Proteins within a family typically share a common structural motif, appear to be homologous and share a ‘general’ function. Other methods (Li et al., 2003; Tatusov et al., 1997), termed orthology-based methods, attempt to organize proteins into groups of orthologs (Fitch, 1970). Proteins in each of these much smaller groups share a ‘detailed’ function because orthologs typically play the same role in different organisms.

While these two levels of functional description have been and will continue to be very useful, there is also important information between these two extremes. Some functional details vary within a protein family but are shared not only by orthologs but also by larger groups of paralogs. Examples include specific protein–DNA and protein–protein interactions as well as certain details of enzyme reactions. An intermediate level of functional detail should be broad enough to group these closely related paralogs together but not so broad as to merge every homologous sequence into the same group. We expect that a well-chosen intermediate level of description will be an additional aid in describing protein function.

We were motivated to consider this problem because of our interest in eukaryotic transcription factors. In these families of proteins there is a clear intermediate level of functional description: specificity for different DNA-binding motifs. Paralogs often bind the same DNA sequences but regulate different genes because of interactions with other proteins or expression in different situations. Our goal was to develop a consistent method that could group protein sequences from a particular family of eukaryotic transcription factors by this intermediate level of function.

Every algorithm has a parameter that sets the granularity; that is, how large the groups are allowed to be. Many graph-based algorithms use a particular cutoff in sequence similarity, such as BLAST expectation value (Altschul et al., 1990). Other algorithms use different parameters, such as the inflation parameter in TRIBE-MCL (Enright et al., 2002), to set the granularity. Even if a gene tree (Fitch, 1995) is used instead of a graph to organize the proteins into groups (Wicker et al., 2001), one still needs to choose a level at which to cut the tree.

Both family-based and orthology-based methods have made use of the mathematical concept of a graph. In the standard configuration, each protein is a node in the graph and the nodes are connected by edges weighted by sequence similarity. ‘Unimportant’ connections are then removed. In other situations where graphs have been used, the transition in the size of the largest cluster (known as the giant component) has been used to determine an appropriate level of granularity (Dokholyan et al., 2002). Our idea is to apply this to the sequence similarity graphs of particular protein families.

To understand the success this simple idea has had for other problems, it is helpful to consider how the size of the giant component correlates with the state of the graph. For strict parameter values, many connections will be removed and the giant component will be small. In this case, all groups will be small and many proteins that share an intermediate level of function will be in different groups. For lax parameter values, the giant component will hold most or all of the proteins in the family, providing no meaningful
information. There will be a phase transition between these two granularities where proteins with the same intermediate function will be grouped together.

We can use this transition to choose the granularity parameter without using any additional information. When the giant component holds ~50% of the proteins, the graph is in the middle of the transition between the unconnected and overly connected extremes. If the phase transition is sharp, the best parameter value will be near the 50% point. We will use this point as our estimate of the granularity parameter.

Other points in the transition may also be important. Without a large amount of experimental data, it is not possible to predict the exact parameter value that will best group the protein sequences. In addition, it is important to note that this transition is unlikely to be meaningful for very small datasets, but a sharp transition in a large dataset provides important information we can use for the organization of protein families.

We will use the method to study three families of eukaryotic transcription factors: proteins containing (1) the basic leucine zipper domain, (2) the zinc finger domain associated with nuclear receptors and (3) exactly three adjacent C2H2 zinc finger motifs. They are, to our knowledge, the largest families with a variety of known distinct DNA-binding specificities. The large number of known, distinct DNA-binding specificities allows us to compare the groups created by our method to groups with distinct DNA-binding specificities.

By analyzing these families and comparing with other methods, we find that the transition in the giant component size can be used to create accurate groups of proteins that reflect an intermediate level of functional detail, the sequence specificity of DNA binding.

2 SYSTEMS AND METHODS

2.1 Protein sequences

Proteins containing a specified eukaryotic transcription factor domain were found in the NCBI non-redundant database. April 24, 2004 edition (Benson et al., 2004): 1255 basic leucine zipper proteins (bzip), 1379 nuclear receptor proteins (nr) and 592 three C2H2 zinc finger proteins (3zf). For both C2H2 zinc fingers and basic leucine zippers, a PROSITE consensus matrix (Hulo et al., 2004) was used by the psfind program (Bucher et al., 1996) to search the database using the default parameters. These consensus matrices are labeled by accession numbers PS50157 and PS50217, respectively. For the nuclear receptor domain, a PROSITE consensus pattern (Hulo et al., 2004), accession number PS00031, was used by a Perl program we developed to search the database. This pattern was extended to 102 amino acids to include the C-terminal extension of the domain.

C2H2 zinc finger patterns were found individually. Proteins were included in this study if they contained three C2H2 zinc fingers in a 106 amino acid range and if no other C2H2 zinc fingers, even weak matches, were found in the protein. C2H2 zinc finger proteins with three consecutive fingers were chosen because this is one of the major groupings of C2H2 zinc finger proteins (Iuchi, 2001) and also allows for a consistent domain size.

Only the subsequence found by the search algorithm was used in the analysis because these are the domains that are involved directly in protein–DNA interactions. Using only subsequences also limits possible confusion that can arise in whole protein alignments from domain transfer and rearrangement events (Liu and Rost, 2003).

2.2 Binding specificities

Binding specificities were found through an extensive literature search. Specificity determinations were accepted only if the experimental methods confirmed at least an affinity for the proposed binding site. Proteins labeled as orthologs in SWISSPROT (Apweiler et al., 2004), such as mouse and human versions of c-JUN, were assumed to have identical binding specificities. Because as orthologs these proteins should share a detailed level of function, we assume that they likewise will share the intermediate level of function.

We would have preferred to also cluster the different consensus DNA-binding sequences and so group the functions in the same way as the sequences. Consensus sequences, however, are reported in a variety of ways. They differ in the length of predicted sequence and the amount of variation in base sequences that is allowed. For this reason, we were unable to cluster the DNA sequences in a meaningful way. For this work we manually grouped the DNA sequences, attempting to match as closely as possible the opinion of the experimental community. Comparing to the available evidence in this way will provide a meaningful test of our results.

The search resulted in 11 different binding specificities for 144 bzip proteins, 4 different specificities for 209 nr proteins and 7 different specificities for 53 3zf proteins. This corresponds to 11.5, 15.2, and 9.0% of protein sequences in the respective transcription factor families. Binding specificities used in this study are available in the supplementary material. The percentages of proteins with binding specificities in the clusters are reported in Supplementary Tables 6–9.

2.3 Sequence comparison

A specific type of sequence similarity, sequence identity, is defined as the number of amino acid residue positions in a pairwise alignment divided by the length of possible matching positions. Like Tian and Skolnick (2003), we find that ‘global’ sequence identity gives the best definition. ‘Global’ sequence identity is defined as the number of identical residues divided by the total length of the alignment. Sequence identity was determined using the align0 program in the FASTA package (Pearson and Lipman, 1988).

The sequence alignments will be reliable because the sequences in the graph share a high degree of sequence identity. It has been found that sequence alignments between sequences with the level of identity below 25% are often inaccurate (Thompson et al., 1999). The high level of sequence identity occurs because we are only comparing homologous DNA-binding domains in each graph.

Other measures of similarity such as pairwise BLAST expectation values (E-values) (Altschul et al., 1990) give similar results, but sequence identity was chosen because of its simplicity of definition. When using BLAST (blastall version 2.2.9), the average pairwise log E-value was used, following the example of Enright et al. (2002).

2.4 Clustering methods

We use single linkage clustering because of its simplicity. In this method, a sequence similarity cutoff value is chosen and any connections with similarity values below this cutoff value are removed. Proteins that are still connected, even indirectly, are considered to be part of the same group. We consider all sequence identity cutoff values from 0 to 1.

One of the most sophisticated other methods of grouping proteins on a graph is TRIBE-MCL (Enright et al., 2002). This method uses the Markov cluster (MCL) algorithm of Van Dongen (Enright et al., 2002; Van Dongen, 2000). The algorithm uses a graph where the undirected edges contain the log of the average pairwise BLAST E-values. It converts these E-values into probabilities of transitions between nodes during a random walk, and then calculates the probabilities of choosing different paths. Iterative ‘inflation’ and ‘expansion’ operations on the matrix of probabilities separate the graph into different clusters by increasing the probability of steps within a cluster and decreasing the probability of moving between clusters. A parameter used during the ‘inflation’ operation, the inflation parameter, is primarily used to affect the granularity of the graph.

For the nuclear receptor family of proteins, the TRIBE-MCL algorithm gives only a fully connected graph in the entire range of the inflation parameter (from 1.01 to 5). In order to test the algorithm at different levels of granularity, we needed to be able to separate this large group. The MCL website FAQ provides an additional way to make the graphs more fine-grained.
in Section 5.3 (http://micans.org/mcl/man/mclfaq.html#faq5.3). We followed the example and used a parameter of 3.0. TRIBE-MCL was then able to separate the proteins at different granularities. These are the TRIBE-MCL results presented for the nuclear receptor family.

2.5 Visualization

A subset of the proteins in the graph is shown visually with each node colored by its experimentally determined DNA specificity. Proteins were selected for visualization of the clusters if they came from humans and had known binding specificities. Human proteins were chosen because of the large amount of experimental work that has been devoted to these proteins or to the proteins of closely related organisms.

In such visualizations, dotted lines indicate clusters that were combined in the graph by sequences other than those displayed. These dotted lines only indicate that the clusters are combined and are not meant to show the actual points of attachment. The number of human proteins with binding data in each cluster can be found in Supplemental Tables 6–9.

Separate visualizations were also made for the TRIBE-MCL results. The output of the algorithm only lists the group to which each protein belongs. Edges were added between proteins in the same group in order to display the TRIBE-MCL results in a similar manner to the results from our method. The program Graphviz (Gansner and North, 2000) was used to display all of the visualizations.

2.6 Scoring

We used the following matrix method to quantify the score. For a graph of total proteins with proteins having a known binding specificity, an lower triangular matrix, termed the connectivity matrix, was created. We use a lower triangular matrix because the full matrix is symmetric. If nodes and in the same group, the value at position of the matrix is 1. If the nodes are not in the same group, the value is 0.

We likewise made a lower triangular matrix from the DNA-binding specificities, termed the DNA-binding matrix. If two proteins and have been determined experimentally to bind the same or very similar DNA sequences, the value again is 1, and it is 0, otherwise.

The two matrices were then subtracted: the connectivity matrix minus the DNA-binding matrix. A value of means that there is a false positive. A value of means that there is a false negative. A value of 0 where the position is 1 in both the matrices signifies a true positive; otherwise, a value of 0 corresponds to a true negative. The score is given by the number of true positives over the sum of the number of true positives, false positives and false negatives.

2.7 Randomized specificities

To test the significance of the scores, scores were calculated when the binding specificities were randomized among the proteins with known experimental data. The same numbers of proteins match a given DNA-binding specificity before and after randomization. For each score, 1000 random DNA-binding matrices were created and used to calculate random scores. These values were then used to calculate a mean and standard deviation. Distances of the true score from the random mean score were then quantified as a Z-score, the difference in units of the standard deviation.

2.8 Phylogenetic relationships

As part of the analysis of the results, CLUSTALW (Thompson et al., 1994) was used to determine approximate phylogenetic relationships between the proteins with a given domain. Phylip (Felsenstein, 1989), version 3.6, was used to plot an unrooted gene tree using these relationships.

3 RESULTS

We present a method that classifies protein amino acid sequences into groups that match an intermediate level of functional detail. For the primary analysis, we use single linkage clustering and sequence identity to make the graphs (Section 2). Each of these graphs is then compared to the known experimental DNA specificities, the intermediate level of functional detail for the transcription factors studied here.

One feature of each graph, the size of the giant component (the largest group), is used as a guideline for the graph granularity. We provide three ways of evaluating the agreement between a graph and the experimental binding data: (1) figures showing the distribution of binding specificities on a subset of the proteins in the graph, (2) scores based on the known binding specificities and (3) the significance of the scores by a comparison to randomly shuffled binding specificities.

3.1 Giant component size

Figure 1 shows the giant component sizes for the three different protein domains under consideration: bzip, zinc fingers of nr and 3zf domains. The giant component size for each family is normalized to the number of nodes (proteins) in the entire graph so that they can more easily be compared.

The transitions from fully connected to sparsely connected graphs occur at different cutoff values. Since an intermediate level of functional description requires the groups to be of an intermediate size, it appears that no single cutoff value can be used to achieve the same granularity for every family. Instead, each family will need its own cutoff value.

3.2 Visualizations

In order to get a qualitative view of the graphs, we look specifically at human proteins with known binding specificities. The different DNA specificities are labeled by the different colors of the nodes. Because some clusters shown are connected by sequences not included in the figures, dotted lines indicate where such connections occur.

The groups in Figure 2 and Supplementary Figures 1 and 2 were formed where the giant components are in the middle of their transitions, specifically, where the giant component comes closest to holding 50% of the proteins in the graph. The giant components at these points contain 46, 49, and 40% of the proteins for bzip, nr and 3zf, respectively. The cutoff value at these points served as
Our initial guess for where proteins will be grouped together by an intermediate level of functional detail.

Visual examination shows that the clusters appear to be both non-random and accurate in clustering the proteins by their binding specificities. The zinc finger protein figure (Supplementary Figure 2) clearly has one overly connected cluster and a one well-formed cluster. The nuclear receptors (Supplementary Figure 1) have some correctly separated clusters but are limited by the small number of binding specificities. The basic leucine zippers (Fig. 2) in particular have many correctly separated clusters.

### 3.3 Scoring

In order to also have a quantitative measure of a graph’s accuracy in clustering proteins, we introduced a scoring method based on the DNA-binding specificities of the proteins. We used a matrix based scoring method (Section 2) that gives the score as the number of true positives (correct grouping of a protein) over true positives plus false positives (missed grouping).

The scores along with the normalized sizes of the giant component for each domain are shown in Figures 3a and 4a. The basic leucine zipper’s maximum score clearly falls near the midpoint of the giant component transition, which appears to be the middle giant component transition. The nuclear receptor’s peak score is found in the transition of the giant component, our initial guess. Likewise, the nuclear receptors (Supplementary Figure 1) have some correctly separated clusters but are limited by the small number of binding specificities.

To better understand the underlying reasons for the breadth of the transition and the shift in the score in the 3zf family, we plotted unrooted gene trees of the basic leucine zipper and 3zf families (Section 2). As seen in Supplementary Figure 3b, the basic leucine zipper proteins are well distributed and at similar evolutionary distances. The 3zf protein tree, however, shows a distinct cluster at a different distance from the rest of the tree (circled in Supplementary Figure 3a). The known binding specificity data is almost exclusively contained within the circled distinct families (Supplementary Figure 3a).

We therefore hypothesized that the zinc fingers we are considering may contain distinct families of homologous sequences that would need separate cutoffs. To test this hypothesis, we repeated the analysis on the 53 proteins with DNA-binding specificity data to see if a true family of 3zf proteins would be amenable to our method.

Figure 4c shows the resulting scores and giant component sizes for this subset. The transition is now sharp, as it is in the basic leucine zipper and nuclear receptor cases. The best score (0.74447) occurs at a sequence identity cutoff of 0.64, which is also the estimate provided by our method. At this point the giant component contains 40% of the proteins. Note that the cutoff is much more stringent than the full set of 3zf proteins and the score is much higher than the previous estimate. Supplementary Figure 8 shows that the large, over-connected group in Supplementary Figure 2 is now split into smaller groups that reflect the different DNA-binding specificities.

### 3.4 Randomized specificities

In order to determine the significance of the clusters at each sequence identity cutoff value, we compared the scores with those found when the binding specificities were randomly distributed to the proteins. If a graph has a similar score whether or not the binding specificities are randomly distributed, the graph is not giving meaningful information about the way the proteins bind. The calculations show that the scores are very significant; the maximum z-scores were 70.6, 33.3 and 43.9 for bzip, nr and 3zf, respectively. In particular, the results show that the scores in the transition in the giant component are highly significant (Figs 3b and d and 4b).

### 3.5 Other methods

We also used BLAST E-values (Altschul et al., 1990) for sequence comparisons to see if a more advanced sequence comparison algorithm would affect the results. A comparison of Supplementary Figure 4 with Figure 3a shows that similar results were obtained for the basic leucine zippers. The other BLAST E-value results are likewise similar to the results found using sequence identity. The midpoint of the giant component transition occurred at log E-values of −5.0, −26.1 and −19.1 for bzip, nr and 3zf, respectively. Similar to the sequence identity results, the BLAST results show that no single cutoff will give the same granularity for all three families.

Likewise, the TRIBE-MCL algorithm (Enright et al., 2002) was used to see if a different clustering algorithm could provide better results. The algorithm uses BLAST E-values for sequence similarity comparisons. The TRIBE-MCL manual (http://www.ebi.ac.uk/research/egg/tribe/manual.txt) lists five suggested inflation parameters. This parameter is used to affect the granularity of the graphs, the size of the clusters. For small (‘tight’) groups, they suggest using 4.0 or 5.0, but for large (‘broad’) groups, the suggestion is 1.1, 2.0 or 3.0. Without further information it is not possible to predict which parameter should be chosen. We tested the algorithm at each of the five suggested values of the inflation parameter for each of the protein families (see Section 2 for more...
Functional specificity from protein sequences

4 DISCUSSION

We have presented a method of grouping large numbers of protein sequences by an intermediate level of function, here the DNA-binding specificity of transcription factors. First, we discuss two families, the basic leucine zippers and the nuclear receptors, where the method accurately predicts the granularity at which the groups will best match the experimental binding data. Second, we consider the case of the three consecutive C2H2 zinc finger family, where the method appeared initially to be unsuccessful. Third, we compare our method with other algorithms. Finally, we consider applications and extensions of the method.

4.1 Basic leucine zippers and nuclear receptors

The best scores clearly fall within the transition in the giant component for both bzip and nr. The best bzip score uses a sequence identity cutoff of 0.442. The closest giant component size to the midpoint (46%) corresponds to this highest scoring cutoff value. The score at this value is also highly significant (Fig. 3b).

For nrs, the highest score occurs in the sequence identity cutoff range of 0.570–0.575. The giant component size closest to 50% for the nuclear receptors (49%) occurs at a sequence identity of 0.586, very close to the best scoring values. Like the bzip family, scores in this range are highly significant (Fig. 3c and d). Using only sequence information, the giant component provides an accurate way of choosing a granularity parameter for these families of proteins.

4.2 Three consecutive C2H2 zinc fingers

For the 3zf case, however, the best score occurs at a much more stringent cutoff (sequence identity cutoff of 0.7) than the 50% giant component level (sequence identity cutoff of 0.5). It is important to note that the transition in the giant component size is broader for the 3zf family than for the other two families that were considered (Fig. 1). The original 3zf dataset appears to contain distinct families of proteins that transition at different parameter values. When only
a related subset of the 3zf proteins is considered, we again observe a sharp transition that gives the highest score, correctly separating the proteins (Fig. 4c and Supplementary Figure 8).

The C2H2 zinc finger is an ancient, modular domain. It is likely that there are different families of three consecutive C2H2 zinc fingers with distinct evolutionary lineages in the dataset. Since different protein families transition from fully connected to sparsely connected graphs at different parameter values (Fig. 1), the transition in the giant component will be broader if different protein families are analyzed together. Clearly one must be careful when using this method to only include one related family of proteins.

4.3 TRIBE-MCL

We chose to compare our method with TRIBE-MCL (Enright et al., 2002), an advanced algorithm that has been used successfully at both the family-based and orthology-based levels of granularity (Enright et al., 2002; Li et al., 2003). We used the inflation parameters suggested by the original authors to determine if any of these would be able to group each of the families by the intermediate level of functional detail. As discussed earlier, all of the inflation parameters created graphs that did not score as well as our method (Supplemental Table 1). We also found that the best inflation parameter is distinct for different families. Without a large amount of experimental information, we would not be able to choose an appropriate inflation parameter, and therefore we would not be able to group the families well at the intermediate level of functional detail. Our method, however, explicitly determines an appropriate level of granularity. The transition in the giant component provides a consistent, family-dependent way of choosing this level of granularity that matches the experimental data.

4.4 Applications

The groups of protein sequences should be useful wherever an intermediate level of functional information is needed. One particular application is the use of statistical methods that determine which residues determine the functional specificity of the protein. In the case of the DNA-binding proteins considered here, statistical analyses of these groups could be used to find which amino acid positions uniformly differ between the groups. These positions are likely to play a role in imparting DNA specificity. Mirny and Gelfand (2002) used such a statistical analysis on a list of known prokaryotic DNA-binding proteins. Our clusters provide a larger body of data to use for an analogous study of eukaryotic DNA-binding proteins.
Functional specificity from protein sequences

We also expect these clusters to be helpful in designing experiments that test DNA-binding specificity. Just as structural genomics methods have found groups of proteins with a fold that needs to be determined experimentally, our method provides different groups of proteins without any known DNA-binding specificity. One can test the function of a protein in such an unstudied group to discover binding specificities that are likely to be applicable to the entire cluster. Also, the method can be used to predict DNA-binding specificity from known experimental data. A high level of identity in one family may imply a shared function while the same level of identity in another family does not. Our method provides a dependable, consistent way of predicting function on this intermediate level.

Although this method has only been applied to transcription factors, many different families of protein have an intermediate level of functional detail that might be organized by our method. Certain enzymes, for example, catalyze certain subclasses of substrates; certain protein–protein interaction domains bind to the same classes of protein surfaces. Our method is likely to be applicable to a wide range of families of protein domains. In the future, we plan to create an online database that implements our method on a large number of protein domain families.

5 CONCLUSIONS

In summary, our method presents a consistent, non-arbitrary way of grouping protein sequences by an intermediate level of functional detail that agrees with experimental results. Although our algorithm is simple, it performs better than more advanced methods because it considers the granularity of the graph that is needed to describe the level of function. In the future, we hope to use this systematic method to help select proteins to test for specificity, to aid in the determination of the function of new proteins and to study the reasons for the differences in function between the different groups.

ACKNOWLEDGEMENTS

This work was supported by the NIH and NSF. J.E.D. is the recipient of an NSF Graduate Research Fellowship. We would like to thank reviewers for helpful comments during the submission process.

REFERENCES