

Predicting Specificity-determining Residues in Eukaryotic Transcription Factors

Jason E. Donald,¹ Eugene I. Shakhnovich²

Keywords: functional specificity, DNA binding, transcription factors, mutual information, graph theory

1 Motivation.

Certain residues are more important than others in carrying out the function of a protein. For example, a single mutation in a transcription factor can cause a change in DNA binding specificity. While crystal structures can be helpful, even when a structure is known other experiments are still needed to understand which residues determine the functional specificity of the protein.

We would like to be able to predict which residues are specificity-determining without carrying out expensive and time-consuming mutagenesis studies. Predicted residues could then be experimentally verified and used to design proteins of novel function.

2 Method.

The mutations that have occurred in nature provide an avenue for predicting these specificity-determining residues. Certain residue positions have been used by natural proteins to modify function. A detailed analysis of these naturally occurring mutations should allow us to predict which residues determine the specificity of the protein.

Previously, Mirny and Gelfand[5] presented a method for predicting specificity determining residues. They studied sets of orthologous bacterial transcription factors that are believed to carry out the same function. The sequences of each set are then compared to paralogous sequences to see which mutations have occurred. Kalinina, et. al.[3], extended the analysis to use a Bernoulli estimator to calculate which predictions are significant.

In many large protein families, especially in eukaryotes, however, there are several problems with defining orthologous relationships and how they relate to protein function. Paralogous sequences can carry out the same general function, such as DNA binding specificity, but differ in time of expression or interaction partners. Also, for many protein families the number of protein sequences is far greater than the extent of our knowledge of their functions. For these reasons, we wanted to test others ways of grouping protein sequences.

In previous work[1], we developed a method that uses graph theory to group protein sequences in a particular protein family into distinct functional groups. The method was able to group several families of eukaryotic transcription factors better than an advanced method of sequence clustering, TRIBE-MCL[2, 6]. We now use these protein groupings within the statistical method of Mirny and Gelfand. We use the combined method on two large, well-studied families of eukaryotic transcription factors, basic leucine zippers and nuclear receptors, so that we can compare our results to the known specificity-determining

¹Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138 USA. E-mail: jdonald@fas.harvard.edu

²Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138 USA. E-mail: eugene@belok.harvard.edu

positions and the results of the original orthology-based method[5]. We also make new predictions.

3 Results.

When we compare our results to known specificity determining residues, we find that we predict the very residues that are known to modulate DNA binding specificity for these two families of proteins. The most important positions tend to be ranked most highly by our method. On the other hand, the original orthology-based method[5] does not predict the known specificity determining residues for these protein families.

We are also able to make some novel predictions. One position in basic leucine zippers contacts the phosphate groups and may play an important role in positioning the protein on the DNA. We are aware of only one experimental study that has mutated this position, and it has only substituted hydrophobic groups while the majority of basic leucine zipper proteins use a basic residue at this position.

Likewise, for the nuclear receptor family, we find a residue position that is involved in dimerization in all known crystal structures. How nuclear receptors dimerize, head-to-head or head-to-tail, determines which promoters they can bind to. This position therefore appears to be very important for nuclear receptor function. To our knowledge, this position has not been studied in mutagenesis experiments. The evolutionary trace method also did not detect its importance for determining specificity[4].

4 Conclusions.

We present a method of finding specificity determining residues automatically given a family of related proteins. Because it does not depend on orthology relationships, it is able to effectively find the specificity determinants of two eukaryotic transcription factor families. We expect the method to be useful for many other families of proteins as well, and plan to present a web database of the results for a wide range of protein families in the future.

5 References and bibliography.

References

- [1] Donald, J.E. and Shakhnovich, E.I. 2005. Determining Functional Specificity from Protein Sequences. *Bioinformatics. In submission.*
- [2] Enright, A.J., Van Dongen, S., and Ouzounis, C.A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* 30:1575–84.
- [3] Kalinina, O.V., Mironov, A.A., Gelfand, M.S., and Rakhmaninova, A.B. 2004. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Science* 13:443–456.
- [4] Lichtarge, O., Yamamoto, K.R., and Cohen, F.E. 1997. Identification of Functional Surfaces of the Zinc Binding Domains of Intracellular Receptors. *Journal of Molecular Biology* 274:325–337.
- [5] Mirny, L.A. and Gelfand, M.S. 2002. Using Orthologous and Paralogous Proteins to Identify Specificity-determining Residues in Bacterial Transcription Factors. *Journal of Molecular Biology* 321:7–20.
- [6] Van Dongen, S. 2000. Graph Clustering by Flow Simulation. Ph.D. thesis, University of Utrecht, The Netherlands.