

Automatic Construction of 3D Structural Motifs for Protein Function Prediction

Mike P. Liang
*Department of Genetics and
Stanford Medical Informatics*
mliang@smi.stanford.edu

Doug L. Brutlag
*Department of Biochemistry and
Stanford Medical Informatics*
brutlag@smi.stanford.edu

Russ B. Altman
*Department of Genetics and
Stanford Medical Informatics*
altman@smi.stanford.edu

Abstract

Structural genomics initiatives are on the verge of generating a vast number of protein structures. The biological roles for many of these proteins are still unknown, and high-throughput methods for determining their function are necessary. Understanding the function of these proteins will have profound impact in drug development and protein engineering.

Current methods for protein function prediction on structures require manual creation of structural motifs. Thus only few structural motifs are available. The lack of structural motifs limits the use of these methods for function prediction at a structural-genomics scale.

To overcome this limitation, we describe a method for automatically creating a library of three dimensional structural motifs. Automatically generating a library of structural motifs can be used for structural-genomic scale function prediction on protein structures.

1. Introduction

Sequencing of the human genome and other model organism has brought a multitude of gene sequences, many of whose function are still unknown [1]. To help elucidate the function of these gene products, structural genomics initiatives aim to determine the structures of all proteins [2] [3]. With the impending explosion of available structural information, automatic methods for predicting functional sites on the protein structure are necessary.

Current methods for modeling functional sites using 3D structural information, such as PROCAT [4] and FFF [5] require manual identification of conserved functional residues and manual assembly of the training set. This manual process is a limiting step to creating a large number of functional site models necessary for genomic-scale functional annotation.

The FEATURE system [6] overcomes the manual identification of conserved residues by automatically

identifying conserved physicochemical properties. Properties around sites with common function and statistically different from those in a set of background non-sites are discovered and used to predict functional sites. However, FEATURE still requires manual selection of the training set of sites and non-sites.

In this poster, we present SeqFEATURE, a method for automatically creating a library of 3D models of functional sites from 1D sequence motifs. We describe its application to calcium binding sites and serine protease catalytic sites. Results show that using 3D structural information around sequence motifs can improve sensitivity while maintaining good positive predictive value in calcium binding. In addition, preliminary results show the method can describe functional sites composed of multiple sequence motifs such as in the case of serine protease.

2. Method

SeqFEATURE automatically creates 3D structural models of functional sites from 1D sequence motifs. There are many different sequence motif databases, such as PROSITE [7]. In our current implementation, we use eMotif [8], although other sequence pattern databases can easily be substituted. eMotif provides a set of sequence patterns with varying specificity and sensitivity for a given set of multiply aligned sequences.

SeqFEATURE maps the 1D sequence motifs of eMotif onto the 3D structure of proteins containing that sequence motifs. The mapped portion of the structure containing the sequence motif is called the motif fragment.

Site selection: The alpha carbons of the motif fragment are used to calculate the geometric centroid of the fragment. This centroid is used as a site center in the training set.

Non-site selection: The atom density around the centroid is measured and a random location in the same protein is selected. If the new location has similar atom density, but not near the motif fragment, it is used as a non-site center in the training set.

The automatically selected training set is fed into the FEATURE system to generate a 3D structural model of the functional site.

Multiple motif sites: For functional sites with multiple sequence motifs describing different parts of the site, a 3D structural model is constructed for each sequence motif as described earlier. Structures with high scoring hits from all the 3D structural models are combined in predicting the functional site.

Evaluation metrics: To evaluate the performance, the sensitivity and positive predictive value of the models are calculated. Sensitivity represents the percentage of gold standard sites that are correctly identified by the model. Positive predictive value represents the percentage of the predictions that correctly identify a site.

3. Results

SeqFEATURE is applied to the calcium binding site and the serine protease catalytic site.

Calcium binding: The eMotifs corresponding to the EF-Hand calcium binding motif were used as the sequence motif to characterize. The EF-Hand represents only a small subset of proteins that bind calcium. A 3D structural model of the EF-hand sequence motif was automatically constructed using SeqFEATURE. The resulting model was able to detect more calcium binding structures than just the EF-hands alone while maintaining similar positive predictive value. In addition, the structural model allows a trade off between sensitivity and PPV to be chosen by selecting an appropriate score cutoff.

Serine protease: The eMotifs corresponding to the subtilisin subfamily of serine proteases was used as the sequence motif to characterize. Three structural models corresponding to the three sequence motifs of the catalytic residues HIS, SER, and ASP were created. Preliminary results show that each of these models has higher scores for data sets enriched with serine proteases than for background structures. In addition, combining the predictions from all three models detected five subtilisin structures, two which were not detected by the original sequence motifs. Two false positives were detected, but high scoring hits of the three models were not co-located. These false positives may be easily filtered out with a co-locality metric.

4. Discussion and Conclusion

SeqFEATURE provides an automatic way of constructing 3D structural motifs from 1D sequence motifs. Initial results show it increases the sensitivity and PPV of detecting functional sites than just the 1D sequence motif alone.

In the case of calcium binding, the sequence motifs identifying only the EF-hand subfamily of calcium binding proteins were used in training to automatically build a more general calcium binding detector. By using structural information, the automatically created 3D structural motif detected calcium binding structures that were not just the EF-hand structures from which it was trained.

SeqFEATURE was also applied to the serine proteases. Automatically constructing three 3D structural models for each of the three sequence motifs containing the catalytic residues of subtilisin, SeqFEATURE was able to detect more subtilisin structures than just the sequence motif. The models did not detect many structures because the site selection was not near the catalytic ASP for one of the models. Future work will include changing the site selection to use a more information theoretic approach.

Preliminary results show SeqFEATURE's applicability to creating a library of 3D structural models. Future work includes improving SeqFEATURE to select better sites and adding a co-locality filter to remove false positives due to spurious non-local high scoring hits. In addition, SeqFEATURE will be validated on other functional sites and on a library of models. Automatic construction of a library of structural models of functional sites will be apposite for structural-genomic scale functional annotation of protein structures.

References

- [1] E. S. Lander, et al., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860-921., 2001.
- [2] S. E. Brenner, "A tour of structural genomics," *Nat Rev Genet*, vol. 2, pp. 801-9., 2001.
- [3] S. K. Burley, et al., "Structural genomics: beyond the human genome project," *Nat Genet*, vol. 23, pp. 151-7., 1999.
- [4] A. C. Wallace, R. A. Laskowski, and J. M. Thornton, "Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases," *Protein Sci*, vol. 5, pp. 1001-13., 1996.
- [5] J. S. Fetrow and J. Skolnick, "Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases," *J Mol Biol*, vol. 281, pp. 949-68., 1998.
- [6] L. Wei and R. B. Altman, "Recognizing protein binding sites using statistical descriptions of their 3D environments," *Pac Symp Biocomput*, vol., pp. 497-508., 1998.
- [7] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. J. Sigrist, K. Hofmann, and A. Bairoch, "The PROSITE database, its status in 2002," *Nucleic Acids Res*, vol. 30, pp. 235-8., 2002.
- [8] C. G. Nevill-Manning, T. D. Wu, and D. L. Brutlag, "Highly specific protein sequence motifs for genome analysis," *Proc Natl Acad Sci U S A*, vol. 95, pp. 5865-71., 1998.