Detecting motifs with EMOTIF-MAKER and MASIA:

A critical comparison of two tools for finding protein motifs

Daniel Holbert
June 4, 2004
Final Project for Biochem 218, Computational Molecular Biology

Abstract

Motif identification is a crucial method for detecting significant regions of proteins and for classifying newly sequenced proteins. EMOTIF-MAKER and MASIA are two tools that utilize creative and novel methods to computationally find and model motifs. In this paper, I describe these tools and compare selected aspects of their approaches to discovering and modeling protein motifs.

Introduction

The term "protein motif" refers to a highly conserved sequence pattern within a set of related proteins. Motifs often have functional or structural significance, which is presumably the reason why these regions have been preferentially preserved in evolution. Hence, motifs can be very useful for locating important areas of proteins, such as active sites or particular types of folds. Moreover, because motifs are conserved within families, they can be very useful for classifying new proteins. [2]

In the past, motifs have been located and modeled by hand [1], but this process can be prone to human error. Moreover, hand-curation becomes increasingly difficult as protein databases mushroom in size. For these reasons, the field of motif detection is relying more and more on computational techniques. Many researchers have created web-based tools for use this area of research. I've chosen to examine and compare two of these tools, EMOTIF-MAKER and MASIA.

EMOTIF-MAKER

The EMOTIF-MAKER tool, developed by Nevill-Manning and Wu in Stanford's Brutlag lab, uses a modified regular expression scheme to represent motifs. It exhaustively generates all motifs that fit the requirements of its scheme. It returns the

results on a graph that maps specificity vs. coverage, at which point the user can choose an appropriate motif from this graph for further experimentation.

In the EMOTIF-MAKER tool, motifs are modeled using a customized form of regular expression. Every position in the regular expression has a corresponding substitution group of residues that are allowed to appear at that position. This group could contain all of the amino acids, some subset of them, or even just a single residue. The creators of EMOTIF-MAKER wanted their substitution groups to appropriately model the variability of amino acids observed in nature. To achieve this goal, they examined position-specific amino acid diversity within protein families in the BLOCKS and HSSP databases, and they came up with twenty highly conserved groups of amino acids (shown in Figure 1). Many of these groups have obvious chemical significance, such as the [FYW] group of aromatic amino acids. [5] EMOTIF-MAKER restricts its regular expressions to use only these twenty groups, a "wild-card " group that contains all of the residues, and groups for each individual amino acid.

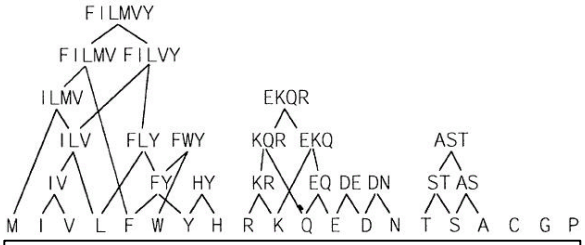

Figure 1: Hierarchy of EMOTIF's allowable substitution groups (From Nevill-Manning et al., [3])

One can string EMOTIF-MAKER's substitution groups together in numerous ways to create motifs that match a given multiple alignment. The most intuitive motif, perhaps, is the one in which each position contains the smallest group that can represent all of the amino acids observed at that position in the multiple alignment. However, if one of the proteins has a sequencing error or has been misclassified into the training set, then the resulting motif may be forced to be much less specific than it should be in order to accommodate for the erroneous sequence. In addition, the training set might contain several subfamilies that could potentially be described by very specific motifs when taken individually, but that require a much more general (and perhaps less useful) motif when they are all grouped together. Consequently, the aforementioned "intuitive" motif frequently might not be the most valuable one, depending on the particular application.

To address this concern, EMOTIF-MAKER doesn't require that its motif models describe all of the sequences in a training set; in fact, it only requires that they cover 30%

of the sequences. (The user can modify this parameter, if desired.) This allows the tool to work around errors in the training set and to generate highly-specific motifs for subfamilies. It generates all possible motifs that satisfy this coverage requirement.

EMOTIF-MAKER outputs its results visually as points on a coverage-vs-specificity graph (see Figure 2). The user can select a motif from this graph with the right balance of coverage and specificity for his or her particular purpose. For example, if a researcher wanted to search a database for all globin-like proteins, he or she would be well advised to use a globin motif with high coverage. On the other hand, if the researcher wanted to find members of a small protein family to a high degree of certainty, he or she would be wiser to use a highly specific motif.

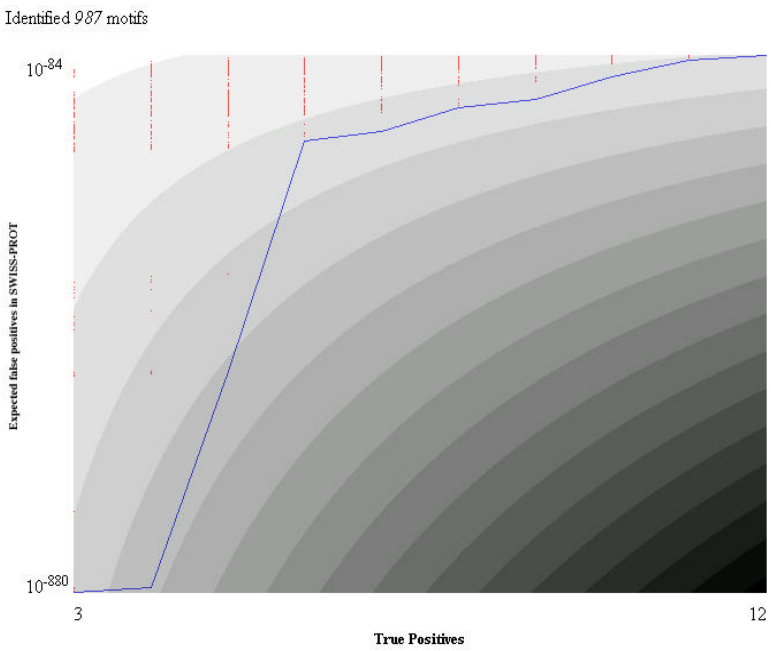

Figure 2: EMOTIF-Maker's graphical output of its results. (From a search at http://fold.stanford.edu/emotif/emotif-maker.html)

MASIA

MASIA is a multi-purpose tool for analyzing multiple sequence alignments. [9] In this paper, I'm focusing only on its macro for detecting Physical-Chemical Property Motifs (PCPM). [8] This macro, maintained by Venkatarajan in the Braun lab at UTMB, detects motifs by measuring the conservation of certain physical and chemical properties at each position in the alignment. The tool defines motifs as regions in which certain property descriptors are well conserved.

Venkatarajan and Braun used a technique known as multidimensional scaling to make it more straightforward to quantify physical-chemical variation between amino

acids. [7] They started by assembling a list of 237 quantitative physical-chemical properties whose values had been measured for all twenty naturally occurring amino acids. After normalizing the amino acids' values for these properties, they constructed a 237-dimensional space with each axis corresponding to one of the physical-chemical properties, and they examined the distribution of amino acids in this space. It turned out that the distribution could be accurately described with only five principal vectors. In fact, when the space was compressed to use just five dimensions based on these vectors, the Euclidian distances between amino acids maintained a 99% correlation coefficient with the original distances in the 237-dimensional property space. Venkatarajan and Braun named their five components E^1 through E^5, and they use these vectors as a basis for defining similarity between amino acids. The resulting similarity index seems to be reliable; it correlates well with established amino acid substitution matrices such as PAM250 and BLOSUM62. Moreover, several of the five component vectors seem to correspond to particular physical properties. For example, amino acid distribution along E^1 is largely determined by hydrophobicity (as shown in Figure 3). However, not all of the components easily reduce to individual properties, nor should they. Rather, they generally represent linear combinations of many different properties. [7]

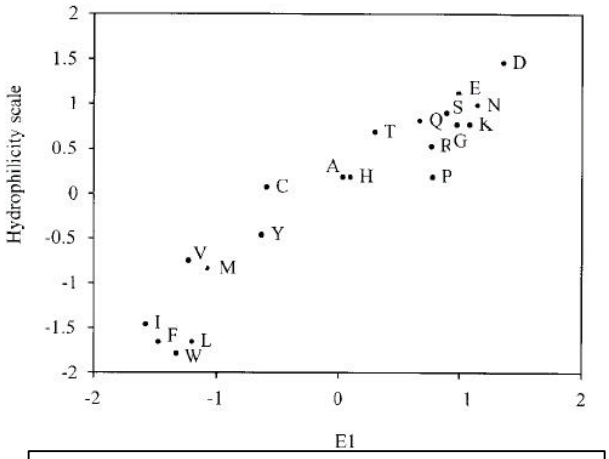

Figure 3: Graph of the amino acids' values along MASIA's E^1 vector vs. their hydrophobicity (From Venkatarajan et al., [5])

In MASIA, a motif is defined as a series of positions that show a minimum level of conservation in at least one of the five descriptors, E^1 to E^5. MASIA motifs must be longer than a minimum length L, and they are allowed to have internal regions with insignificant levels of conservation up to a length of G. L and G are user-definable parameters. MASIA uses the default values of 4 and 2, respectively, based on empirical results. [8]

MASIA's motifs are stored as sequence profiles. To construct a profile, MASIA iterates across every position in a multiple alignment, storing for each position the average values, standard deviations, and relative entropies for all five E-components. The resulting motif profile can be used to probabilistically search databases for matching sequences. [8]

Discussion

EMOTIF-MAKER and MASIA's PCPM macro are designed to solve the same overall problem: given a multiple sequence alignment, they attempt to discover motifs. However, the tools' methods are quite different. I'll examine the differences in two main areas. First, I consider their techniques for modeling conservation on the scale of individual amino acid positions. Second, I compare their strategies for generating complete motifs and sets of motifs.

Modeling amino acid conservation

Before one can model a motif, one must first decide on a way to model degrees of amino acid conservation for a particular position. EMOTIF-MAKER and MASIA take very different approaches to this problem: EMOTIF-MAKER uses predefined substitution groups, whereas MASIA measures conservation more quantitatively by using statistical methods in multidimensional property-space. Both approaches have comparative advantages and disadvantages.

While EMOTIF-MAKER's fixed substitution groups have evidence demonstrating their empirical relevance, [5] there is something to be said for a more versatile approach whose results are more specific to the observed data, as in MASIA. On the one hand, this sort of technique often runs the risk of overfitting the training data; that is, if one tries too hard to exactly model a training set, it's possible to create a motif that matches the observed data to such a high degree that it loses its biological significance. However, MASIA avoids this overfitting problem by rooting its models in statistically conserved physical and chemical properties, and this helps to keep it from losing touch with the biological significance. By measuring values for its five principal vectors, MASIA very precisely models the degree of conservation at an amino acid

position. When compared to this scheme, EMOTIF-MAKER's fixed substitution groups seem to be a somewhat more limited method for modeling position-specific diversity.

Of course, MASIA isn't perfect at modeling amino acid conservation. In particular, it is susceptible to biases in the training sequences. It assigns the same weight to all sequences when it computes averages to construct a profile, and this could lead to undesirable results for training data that contain subsets of especially highly related sequences. For example, a hypothetical training set that contains four closely related human sequences and one distant bacterial sequence would probably have many positions at which the human sequences all agreed with each other but differed from the bacterial sequence. As a result, the average values incorporated into the motif profile would be skewed in favor of the human sequences because these sequences would "outvote" the bacterial sequence. One simple way in which MASIA could correct for this problem would be to first measure sequence-wide similarity in the training set and then to proportionally down-weight similar sequences when computing average values. [4]

In contrast, EMOTIF-MAKER deals with sequence bias in a very capable way that follows directly from its use of substitution groups. EMOTIF-MAKER always generates all possible motifs that cover at least 30% of the training set based on its substitution groups, regardless of the relative prevalence of particular variations. As a result, EMOTIF-MAKER is resistant to problems resulting from overrepresentation of particular variations. Indeed, rather than having problems with sequence bias, EMOTIF-MAKER capitalizes on it: if a highly-related subfamily represents at least 30% of the data set, then EMOTIF-MAKER will generate an additional very specific motif for this subfamily, in addition to the other motifs that describe the dataset as a whole. Overall, EMOTIF-MAKER deals with sequence bias much better than MASIA.

Generating entire motifs

Beyond their differences in modeling amino acid conservation, EMOTIF-MAKER and MASIA have very diverse functionality at the whole-motif level, as well. Principal among their differences at this level is the fact that EMOTIF-MAKER generates many motifs, while MASIA generates only one for each sufficiently similar region.

EMOTIF-MAKER is more flexible for most datasets because of its ability to generate many motifs, most of which selectively ignore some of the sequences from the data set. As mentioned before, this improves the tool's robustness against errors in the data, and it also gives EMOTIF-MAKER the power to automatically detect subfamilies within data sets. In contrast, MASIA only returns a single "average" motif for each adequately conserved region. As a result, its motifs are more susceptible to errors in the data because they necessarily incorporate information from all of the sequences in the multiple alignment.

Nonetheless, there are cases in which the motifs returned by MASIA could be more useful than those generated by EMOTIF-MAKER. In particular, if one wanted to search for members of a diverse family with relatively low sequence conservation, it might be easier to use a single overarching property-based set of MASIA motifs. MASIA's motifs preserve the physical and chemical information about important sites while remaining detached from the actual sequences, and this can be especially useful if the sequences have relatively little in common. [8] In comparison, EMOTIF-MAKER might not be as easily able pick up on a motif in such a diverse sets of sequences, especially if the diversity doesn't directly map to the tool's fixed substitution groups.

Conclusion

EMOTIF-MAKER and MASIA stand out in different ways. The main comparative benefits of EMOTIF-MAKER are its resistance to errors in the data and the sheer number of motifs that it produces, at various levels of specificity. MASIA, on the other hand, is intriguing because it is rooted in a comprehensive and quantitative description of amino acids' physical and chemical properties, and this helps it detect a broad spectrum of biochemically significant patterns. Future work in this field might look into the feasibility of combining aspects of these two programs.

Availability

EMOTIF-MAKER is available on the web at http://motif.stanford.edu/emotif/, and MASIA is available at http://www.scsb.utmb.edu/masia/masia.html

References

1. Brutlag, Doug. (2004) Functional Genomics & Proteomics: Using Protein Motifs to Represent Function. Bioc 218 / BMI 231, Lecture notes 5/13/2004.

2. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J.A., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235-238.

3. Huang,J.Y. and Brutlag,D.L. (2001) The EMOTIF database. *Nucleic Acids Res.*, **29**, 202–204.

4. Henikoff S. and Henikoff J.G. (1994) Position-based sequence weights. *J Mol Biol.*, **243**, 574-578

5. Nevill-Manning,C.G., Wu,T.D. and Brutlag,D.L. (1998) Highly specific protein sequence motifs for genome analysis. *Proc. Natl Acad. Sci. USA*, **95**, 5865–5871.

6. Schein,C.H., Izumi,T., Oezguen,N., Feng,Y.L. and Braun,W. (2002) Total sequence decomposition and genomic cross-networking distinguishes functional modules in apurinic/apyrimidinic endonucleases. *BMC Bioinformatics*, **3**, 37.

7. Venkatarajan,M.S. and Braun,W. (2001) New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. *J. Mol. Model.*, **7**, 445–453.

8. Venkatarajan,M.S., Schein,C.H. and Braun,W. (2003) Identifying property based sequence motifs in protein families and superfamilies: application to DNase-1 related endonucleases. *Bioinformatics*, **19**, 1381–1390.

9. Zhu,H., Schein,C.H. and Braun,W. (2000) MASIA: recognition of common patterns and properties in multiple aligned protein sequences. *Bioinformatics*, **16**, 950–951.