CONSTRUCTION OF A DATABASE OF CO-OCCURRING eMOTIFS BASED ON CONDITIONAL PROBABILITIES

PRASANTH PULAVARTHI
prasanth@stanford.edu
Biochemistry 218

(Submitted 19 March 2000)

Abstract—Classification of a newly discovered protein into a family of proteins enables the determination of its function. The eMOTIF system identifies conserved modular domains that confer functionality or structure to proteins and allows classification of proteins into families based on the conserved domains a protein contains. A program called multeeMOTIF has been developed which analyzes eMOTIFS and determines the conditional probabilities of their occurrences to find pairs of eMOTIFS that occur together a percentage of the time. The proteins that match one eMOTIF compose a large super-family of proteins and proteins that match each additional eMOTIF compose smaller and smaller sub-families. Based on how many and which eMOTIFS an unknown protein matches, it can be assigned to the appropriate sub-family. The more eMOTIFS matched, the more specific the family assignment will be. Those pairs of eMOTIFS that always occur together or occur together a high percentage of the time are observed to be from alignments of the same protein functionality 83% of the time. This may allow assignment of function to alignments of unknown function and consequently proteins belonging to such alignments.

INTRODUCTION

Identification of the function of newly discovered proteins typically involves determining the family to which the protein belongs. Once the family is known, the function of the protein can be postulated to be similar to that of the other proteins in the family. This is because proteins with similar sequences fold into proteins with similar structure and similar structures perform similar functions. Several popular methods are based on global sequence similarity that makes

use of position-specific scoring matrices to characterize the general nature of a protein and the family it belongs to. The eMOTIF system, on the other hand, identifies conserved modular domains that confer functionality or structure to proteins. This approach allows classification of proteins into families and sub-families based on the conserved domains a protein contains.

Using multiple characteristics can identify a protein more precisely than using only one. Databases such as PRINTS use groups of motifs to build characteristic signatures. The multiple motifs provide biological context that can be used to assess the validity of identification. Similarly, multiple eMOTIFS may be used to together to increase the number of subfamilies available for classifying a protein. Here, pairs of eMOTIFS that occur together with high probabilities are found based on their conditional probabilities. Often times both eMOTIFS in a pair are different components of the same overall functionality. Each pair of eMOTIFS composes a subfamily of proteins found by either eMOTIF alone.

METHODS

The eMOTIF-MAKER program was applied to the BLOCKS+ database dated June 10, 2000 and version 28.0 of the PRINTS database to generate the eMOTIFS. The eMOTIFS were then run at an expectation of 10^{-2} against the subset of protein sequences from SWISS-PROT version 39.0 that did not contain B, J, O, U, X, or Z. The results consisted of 44 979 unique proteins that were matched by 136 451 eMOTIFS. The eMOTIFS were derived from 15 043 unique alignments, so multiple matching eMOTIFS were derived from the same alignment. The format of these results was as follows:

Each eMOTIF was listed on one line as:

```
>expectation of the eMOTIF scanning for this eMOTIF|specificity of the
eMOTIF|sensitivity of the eMOTIF|accession number of the parent alignment of
the eMOTIF|descriptive name of the alignment|eMOTIF regular expression
```

Each such line was followed by a list of proteins it matched, one per line as:

```
Swiss-Prot ID|Swiss-Prot accession|matching sequence region |start index|stop index
```

Such a file listing each eMOTIF and the proteins it matched was provided as input to a program called multeeMOTIF. multeeMOTIF was written in Perl to take such an input file and determine the motifs that always occur together. Because the data sets are large, multeeMOTIF operates in stages and generates various intermediate files to allow restarting the process even if it is prematurely terminated. The intermediate files are text files with fields separated by commas.

In the first stage, multeeMOTIF builds an index and a reverse index from the eMOTIF and protein identifiers extracted from the input data file. This allows easy identification of the eMOTIFS that matched each protein. An eMOTIF that was found more than once in a protein is counted as only one match. In this stage, each eMOTIF is assigned a numerical identifier that will be used in the remainder of the intermediate files and output. A file mapping the assigned identifier to the eMOTIF expression, description, and parent alignment is generated as well. In the next stage, multeeMOTIF processes each protein and generates all unique combinations of eMOTIFS matching it. To compute the conditional probabilities, it is necessary to know all the valid combinations of eMOTIFS. Using this procedure rather than taking all possible combinations of all the eMOTIFS saves much work since not all of the ${}_nC_2$ combinations actually exist. In this data set, only 69.2×10^6 of the possible 9.3×10^9 eMOTIF combinations (0.7%) are actually found in the proteins.

In the third stage, multeeMOTIF calculates the probability P(A) of each motif, the joint probability P(AB) of each combination of two motifs, and the two conditional probabilities, $P(A|B) = P(AB) / P(B)$ and $P(B|A) = P(AB) / P(A)$, for each combination of two motifs. Finally

multeeMOTIF finds all the pairs that satisfy P(A|B) = 1 <= K*P(B|A) where K is a user specified probability factor. If K = 1, then only those motifs that always occur together are returned. If K > 1, then motifs whose probability of co-occurrence is less than 1 are returned. The results are placed in a text file with each line representing a motif combination in the following format:

```
motif A,motif B,P(A|B),P(B|A)
```

These results were imported into Microsoft SQL Server along with an index of the motifs and their parent alignments and descriptions to allow fast running of complex queries. The SQL query used to determine the number of motif pairs derived from different alignments was:

```
SELECT motifa, LEFT(a.description, 40), motifb, LEFT(b.description, 40)
FROM matches_1, motifs a, motifs b
WHERE matches_1.motifa = a.motif AND matches_1.motifb = b.motif
      AND a.accession <> b.accession
```

The SQL query used to determine the number of motif pairs derived from different alignments with different functions was:

```
SELECT motifa, LEFT(a.description, 40), motifb, LEFT(b.description, 40)
FROM matches_1, motifs a, motifs b
WHERE matches_1.motifa = a.motif AND matches_1.motifb = b.motif
      AND a.accession <> b.accession
      AND NOT (a.description LIKE '%'+LEFT(b.description,7)+'%' OR
               b.description LIKE '%'+LEFT(a.description,7)+'%')
```

To determine the number of motif pairs derived from different alignments where one of the motifs was of unknown function, the following WHERE clause was appended to the above query:

```
      AND (LOWER(a.description) LIKE '%unknown%' OR
           LOWER(b.description) LIKE '%unknown%')
```

RESULTS

At a probability factor of 1, 211 693 pairs of motifs occur together. Since a probability factor of 1 means P(A|B) = P(B|A) = 1, these are motifs that always occur together. Of these, 182 724 pairs, or about 86%, are of motifs derived from different alignments. Though they were

from different alignments, in many of the pairs, the motifs had descriptions that were similar or identical, as can be seen in Table 1. Filtering out such pairs yields 30 864 pairs in which the motifs have different alignment descriptions, as listed in Table 2. In the 182 724 pairs that always co-occur derived from different alignments, there are 1190 pairs in which both motifs are from alignments of "unknown function" and 273 pairs in which only one motif alignment has a description of "unknown function", as shown in Table 3.

At a probability factor of 1.1, 275 165 pairs of motifs occur together. Thus 63 472 pairs do not always occur together but do occur with a greater than 1/1.1 = 91% probability. 225 925, or 82%, of the pairs are of motifs derived from different alignments. 37 711 pairs have different alignment descriptions and there are 288 pairs in which only one motif alignment has a description of "unknown function".

DISCUSSION

As seen in Table 2, some of the descriptions, although not deemed similar by a simple comparator, in fact refer to similar functionality. Thus when using a probability factor of 1, at least 1-(30 864/182 724) = 83%, and most likely more than 83%, of pairs derived from different alignments are actually from alignments that have similar function. Since so many of the pairs are of motifs with similar functions, if we know the function of one motif, we can reasonably speculate the function of the other motif to be the same. Thus for the 273 motifs derived from alignments with unknown function, 83% or 226 of them can statistically be assigned a function correctly. Many of the proteins these motifs match are currently classified as "hypothetical", as listed in Table 4, and the functions assigned to the alignment would help identify the function of the proteins.

When the probability factor is increased to 1.1, more pairs are found because they don't have to always co-occur, just most of the time. 82%, as compared to 86% when K = 1, of the eMOTIF pairs are derived from different alignments. Since eMOTIFS are generated by enumerating multiple motifs for the same alignment, some eMOTIFS will be subsets of other eMOTIFS. As the co-occurring criterion is relaxed, more of these subsets will be found. At least 1-(37 711/225 925) = 83% of the pairs derived from different alignments are actually from alignments that have similar function. This rate is the same as when K = 1. Although the percentage of co-occurring pairs derived from the different alignments decreases as K increases, the percentage of pairs that are from alignments that have similar functions stays the same.

At K = 1, using two eMOTIFS to identify proteins does not increase the specificity since the union of the sets of proteins containing each eMOTIF alone is the same as the set of proteins containing both eMOTIFS. By similar logic, the sensitivity will not change either. At K > 1, however, using two eMOTIFS will identify only a subset of the proteins found in the union of the sets of proteins containing either eMOTIF, thus increasing the specificity. When the eMOTIFS do not always occur together, the sensitivity may decrease since there are fewer proteins in the set containing both eMOTIFS than the set of proteins containing at least one of the eMOTIFS.

Using multiple eMOTIFS improves diagnostic ability by determining not just whether a protein has one component that provides some functionality but whether it has several of the components that make up the necessary elements for that functionality. The more components, or eMOTIFS, that are matched, the more accurately the function of the protein can be pinpointed. Thus the proteins that match one eMOTIF can compose a large super-family of proteins and proteins that match each additional eMOTIF compose smaller and smaller sub-families. Based

on how many and which eMOTIFS an unknown protein matches, it can be assigned to the appropriate sub-family. The more eMOTIFS matched, the more specific the family assignment will be. This is an improvement over using the PRINTS database directly since PRINTS, though it identifies the family a protein belongs to and how distant the protein is from the family, does not show how the families are organized into super-families and sub-families.

Future Directions

This work has only been a preliminary version of multeeMOTIF. There are several improvements that should be undertaken in the future. Among these are improvements to the implementation of multeeMOTIF. It is currently implemented in Perl and works with flat text files. Though the size of these files is extremely large (the various intermediate files take up over 2 gigabytes of disk space), the performance is decent on a dual processor PIII server with a gigabyte of RAM. However, generation of the final list based on the user specified probability factor currently cannot be done fast enough for use with a web-based interactive query system. Thus the databases used by an interactive query system must be populated with lists pre-generated using pre-specified probability factors. Future implementation should look at either using binary files or moving more stages of the process into an SQL server.

Another issue that should be addressed in a future version is the current limitation of multeeMOTIF to two eMOTIFS. multeeMOTIF should support pairing 3, 4 or even more eMOTIFS, such that, for example, $P(A|B) = P(B|A) = P(A|C) = P(B|C) = P(C|A) = P(C|B) = 1$. While two eMOTIFS are better than one eMOTIF, even more eMOTIFS will further increase the ability to divide families into sub-families and pinpoint protein functionality more accurately.

Additionally, the validity of the functions assigned to alignments and proteins needs to be verified to see if the statistically suggested 83% correctness holds. This can be done either by formulating a test set or by obtaining biological data about the true functionality of the proteins that were assigned functions.

APPENDIX

MOTIF A	Alignment Description	MOTIF B	Alignment Description
56896	Trehalase	93071	Trehalase
15948	Adhesin family signature III	43528	Adhesin B signature V
135049	Anion exchanger family signature VI	48197	Anion exchanger family

Table 1. A few of the 182 724 eMOTIF pairs derived from different alignments that had descriptions that were similar or identical.

MOTIF A	Alignment Description	MOTIF B	Alignment Description
88769	P2X3 purinoceptor signature VIII	2952	ATP P2X receptor
135431	Homoserine dehydrogenase	94344	Aspartokinase
131518	Plasmodium circumsporozoite protein	122391	CAP protein
131734	Plant globin signature IV	131733	Leghaemoglobin

Table 2. Several of the 30 864 pairs of eMOTIFS derived from different alignments that had different alignment descriptions. But as can be seen, even though the simple description comparator could not identify them, several of the descriptions indicate similar functionality.

MOTIF A	Alignment Description	MOTIF B	Alignment Description
15501	Domain of unknown function	125231	Guanylate kinase
75837	Calcium channel signature I	133997	Protein of unknown function
111501	RNA methyltransferase trmA	5379	Domain of unknown function

Table 3. Several of the 273 pairs of eMOTIFS derived from different alignments that have one eMOTIFS derived from an alignment of “unknown function”. 83% of the time both eMOTIFS are derived from alignments that have the same or similar function. Thus it is likely that the function of the unknown alignment in each pair is the same as the known alignment.

Proteins Matched by eMOTIF 133997 (“protein of unknown function”):		
Q58124	Y714_METJA	HYPOTHETICAL PROTEIN
Q58736	YD40_METJA	HYPOTHETICAL PROTEIN
Q57758	Y310_METJA	HYPOTHETICAL PROTEIN

Table 4. Many of the proteins matched by eMOTIFS derived from alignments of “unknown function” are hypothetical proteins whose function is not known. Since the function of the

alignment can be postulated by seeing which other eMOTIFS it co-occurs with, the function of these proteins may also be speculated.

ACKNOWLEDGEMENTS

The author would like to thank Doug Brutlag for his helpful insights and guidance and Jimmy Huang for providing the data and assistance with working with the data.

REFERENCES

Attwood, T. K., Croning, M. D., Flower, D. R., Lewis, A. P., Mabey, J. E., Scordis, P., Selley, J. N. and Wright, W. (2000). PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.*, 28(1), 225-227.

Attwood, T. K., Beck, M. E., Flower, D. R., Scordis, P. and Selley, J. N. (1998). The PRINTS protein fingerprint database in its fifth year. *Nucleic Acids Res.*, 26(1), 304-308.

Attwood, T. K., Flower, D. R., Lewis, A. P., Mabey, J. E., Morgan, S. R., Scordis, P., Selley, J. N. and Wright, W. (1999). PRINTS prepares for the new millennium. *Nucleic Acids Res.*, 27(1), 220-225.

Henikoff, S. and Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucleic Acids Res.*, 19(23), 6565-6572

Henikoff, S., Henikoff, J. G., Alford, W. J. and Pietrokovski, S. (1995). Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene*, 163(2), GC17-26.

Huang, J. Y. and Brutlag, D. L. (2001). The eMOTIF Database. *Nucleic Acids Res.*, 29, 202-204.

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