A Structure Based Algorithm for Improving Motifs Prediction

Sudipta Pathak[1], Vamsi Krishna Kundeti[2], Martin R. Schiller[3], and Sanguthevar Rajasekaran[1,*]

[1] Department of Computer Science, University of Connecticut
{sup11002,rajasek}@engr.uconn.edu
[2] Intel Corporation
vamsi.k.kundeti@intel.com
[3] School of Life Sciences, University of Nevada Las Vegas
martin.schiller@unlv.edu

Abstract. Minimotifs are short contiguous peptide sequences in proteins that are known to have functions. There are many repositories for experimentally validated minimotifs. MnM is one of them. Predicting minimotifs (in unknown sequences) is a challenging and interesting problem in biology. Minimotifs stored in the MnM database range in length from 5 to 15. Any algorithm for predicting minimotifs in an unknown query sequence is likely to have many false positives owing to the short lengths of the motifs looked for. Our team has developed a series of algorithms (called *filters*) in the past to reduce the false positives and improve the prediction accuracy. All of these algorithms are based on sequence information. In a recent paper we have demonstrated the power of structural information in characterizing motifs. In this paper we present an algorithm that exploits structural information for reducing false positives in motifs prediction. We test the validity of our algorithm using the minimotifs stored in the MnM database. MnM is a web system for minimotif search that our team has built. It houses more than 300,000 minimotifs. Our new algorithm is a learning algorithm that will be trained in the first phase and in the second phase its accuracy will be measured. For any input query protein sequence, MnM identifies a list of putative minimotifs in the query sequence. We currently employ a series of sequence based algorithms to reduce the false positives in the predictions of MnM. For every minimotif stored in MnM, we also store a number of attributes pertinent to the motif. One such attribute is the *source* of the minimotif. The source is nothing but the protein in which the minimotif is present. For the analysis of our new algorithm we only employ those minimtofis that have multiple sources for positive control. Random data is used as negative data. The basic idea of our algorithm is the hypothesis that a putative minimotif is likely to be valid if its structure in the query sequence is very similar to its structure in its source protein. Another important feature of our algorithm is that it is specific to individual minimotifs. In other words, a unique set of parameters is learnt for every minimotif. We feel that this is a better approach than learning a common set of parameters for all the minimotifs together. Our findings reveal that in most of the cases the occurrences of the minimotifs in their source proteins are structurally similar. Also, typically,

[*] Corresponding author.

A. Ngom et al. (Eds.): PRIB 2013, LNBI 7986, pp. 242–252, 2013.

the occurrences of a minimotif in its source protein and a random protein are dissimilar. Our experimental results show that the parameters learnt by our algorithm can significantly reduce false positives.

1 Introduction

Genetic linkage analysis and other approaches have identified many mutations that are associated with inherited human disease. Many of these mutations are in protein coding regions. An effective strategy for treating many diseases is to identify a drug that interferes with the protein that contains the mutation. Thus, it is important to understand the function of the protein such that drugs can be designed to interfere with its function. Analysis of protein and DNA sequence is an important approach for predicting protein function, thus an important part of the pipeline in drug discovery.

Analysis of DNA and protein sequences often involves the identification of patterns. As a new tool for predicting new causes of disease, our group has built and operates the Minimotif Miner (MnM) website/database (Balla, et al. 2006, Rajasekaran, et al. 2009). MnM can be used to predict potential minimotifs and thus new functions in proteins. These are not domain motifs, but the short functional motif determinants for binding other molecules, the signatures for regulatory posttranslational modifications on proteins, and the short sequence elements that code for protein trafficking. These motifs are readily cross-mapped with disease-associated single nucleotide polymorphisms (SNPs) on the MnM website, thus any scientist can determine a motif that is introduced or eliminated by a disease-associated mutation. One of the principle problems with this approach is that the short motifs are not very complex and false-positives overwhelm the true motifs. In fact all the motif search systems currently available (such as ELM [12], Scansite [7], Prosite [13], Dilimot [14], etc.) suffer from this problem. If this approach were refined, then the approach may be very useful for identifying new drug targets.

In our previous work we have proposed a series of algorithms (called *filters*) (see e.g., [8,9]) to reduce false positives. Examples include protein-protein interaction filter [8], molecular function filter [9], cell function filter [9], etc. These algorithms are all based on sequence information. As is well known, in addition to sequences, structures also contain a rich amount of useful information. In this paper we propose an algorithm for reducing false positives in the prediction of minimotifs. We have tested the accuracy of this algorithm using the minimotifs in MnM. Our empirical tests indicate that the new algorithm is very effective. An interesting feature of our algorithm is that its predictions are specific to individual motifs.

The rest of this paper is organized as follows. In the next section we provide some preliminaries on protein structures. Followed by this we describe our algorithm. Subsequently we provide the results and discussions.

1.1 Some Preliminaries

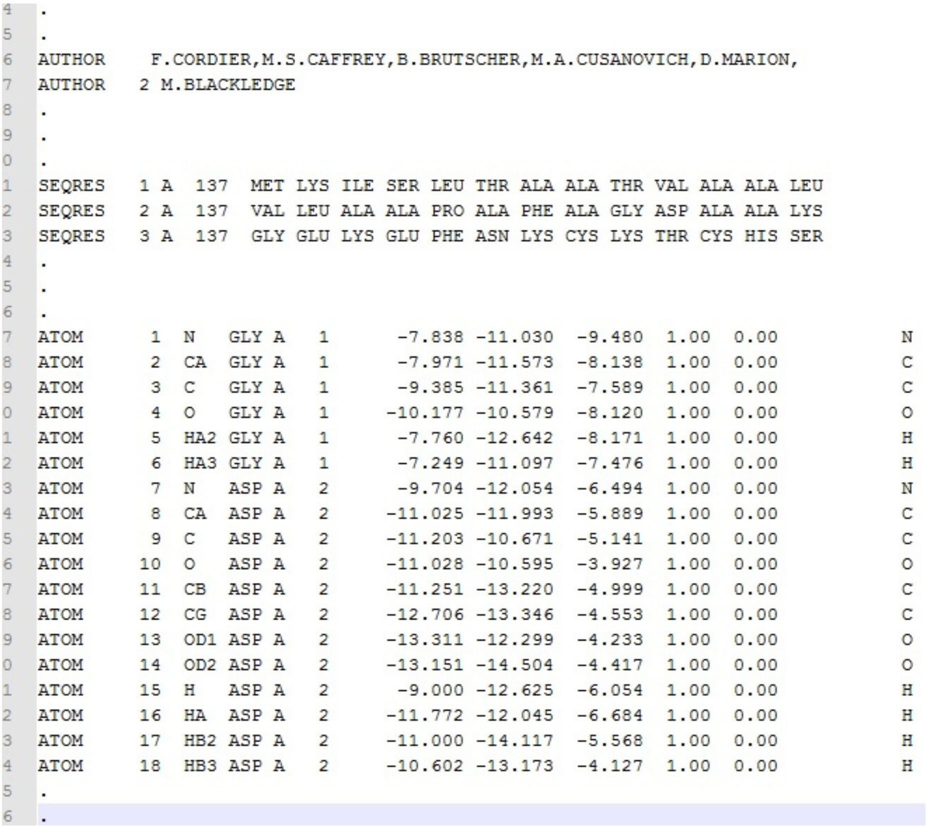
Every Protein has its primary and secondary structures. Primary structure of a protein is its sequence. The secondary structure consists of helices, sheets, etc. Some of the proteins might have quaternary structures. Protein architecture is one of the most fundamental research topics because the 3D protein structure is responsible for the cell functional properties in all living systems. Amino acid residues are the building blocks of protein primary structure.

The secondary structure of a protein mainly contains the following information: Helix, Sheet, Connectivity Details (disulfide bonds, prolines and other peptides found in cis conformations, etc.), Crystallographic and Coordinate Transformation information (transformation from orthogonal coordinates, transformations expressing non-crystallographic symmetry, etc.), Coordinate Information (collection of atomic coordinates), etc. There exist databases that contain the above information for a subset of the known proteins. An example is the World Wide Protein Data Bank [2]. PIR [15], developed by National Biomedical Research Foundation (NBRF), is one of the earliest primary protein databases. Later in 1988 Martinsried Institute for Protein sequences collected the protein sequences from PIR and developed a web server. Swiss-prot [3] is one of the well known primary protein databases maintained collaboratively by Swiss Institute of Bioinformatics(SIB) and European Bioinformatics Institute(EBI)/European Molecular Biology Laboratory(EMBL). Swiss-prot provides a lot of information including functions of proteins, structures of their domains, post-translational modifications information, etc. This database is a valuable resource produced by PIR from sequences extracted from the Brookhaven Protein Data Bank (PDB). The significance of this database is that it makes available the protein sequence information in the PDB for keyword interrogation and for similarity searches. It includes bibliographic references, MEDLINE cross-references active site, secondary structure and binding site annotations. Also there are composite databases like Non-Redundant DataBase (NRDB)by NCBI (National Center for Biotechnology Information)[5], BLAST (Basic Local Alignment Search Tool) service[16], OWL from the UK EMBnet National Node and the UCL Specialist Node[6] etc. Secondary databases are a consequence of analysis of the sequences of the primary databases, mainly based from Swiss-prot. Prosite [13] is the first among all the secondary databases. This consists of entries about protein families, domains, functional sites, amino acid patterns, etc. This was introduced by Swiss Institute of Bioinformatics and this is mainly based on Swiss-prot.

Along with the above databases a number of web based tools have been developed to allow investigators to search for motifs in a protein query sequence. Scansite [7] is one such tool which includes ten different programs. The Motif Scan ensemble of programs computationally identifies all motifs within a given user-specified protein, while the Database Search ensemble of programs finds all proteins in a protein database, such as Swiss-prot, that match a given motif. One of the most successful tools in this area of research is Minimotif Miner (MnM) that our team has built [8,9,10,11]. All of the known motif search tools suffer from a high false positive rate especially when the motif length is small. We offer a novel solution to this problem in this paper that utilizes structural information.

1.2 Implementation of the Algorithm

To implement the algorithm we make use of Worldwide Protein Data Bank (wwPDB). PDB contains more than eighty thousand proteins and their structural information. We downloaded the entire PDB from the following link: `ftp://ftp.wwpdb.org/pub/pdb/data/structures/divided/`. A typical PDB file contains thousands of lines like the ones shown in Figure 1.



```
.
.
AUTHOR    F.CORDIER,M.S.CAFFREY,B.BRUTSCHER,M.A.CUSANOVICH,D.MARION,
AUTHOR   2 M.BLACKLEDGE
.
.
.
SEQRES   1 A  137  MET LYS ILE SER LEU THR ALA ALA THR VAL ALA ALA LEU
SEQRES   2 A  137  VAL LEU ALA ALA PRO ALA PHE ALA GLY ASP ALA ALA LYS
SEQRES   3 A  137  GLY GLU LYS GLU PHE ASN LYS CYS LYS THR CYS HIS SER
.
.
.
ATOM      1  N   GLY A   1      -7.838 -11.030  -9.480  1.00  0.00           N
ATOM      2  CA  GLY A   1      -7.971 -11.573  -8.138  1.00  0.00           C
ATOM      3  C   GLY A   1      -9.385 -11.361  -7.589  1.00  0.00           C
ATOM      4  O   GLY A   1     -10.177 -10.579  -8.120  1.00  0.00           O
ATOM      5  HA2 GLY A   1      -7.760 -12.642  -8.171  1.00  0.00           H
ATOM      6  HA3 GLY A   1      -7.249 -11.097  -7.476  1.00  0.00           H
ATOM      7  N   ASP A   2      -9.704 -12.054  -6.494  1.00  0.00           N
ATOM      8  CA  ASP A   2     -11.025 -11.993  -5.889  1.00  0.00           C
ATOM      9  C   ASP A   2     -11.203 -10.671  -5.141  1.00  0.00           C
ATOM     10  O   ASP A   2     -11.028 -10.595  -3.927  1.00  0.00           O
ATOM     11  CB  ASP A   2     -11.251 -13.220  -4.999  1.00  0.00           C
ATOM     12  CG  ASP A   2     -12.706 -13.346  -4.553  1.00  0.00           C
ATOM     13  OD1 ASP A   2     -13.311 -12.299  -4.233  1.00  0.00           O
ATOM     14  OD2 ASP A   2     -13.151 -14.504  -4.417  1.00  0.00           O
ATOM     15  H   ASP A   2      -9.000 -12.625  -6.054  1.00  0.00           H
ATOM     16  HA  ASP A   2     -11.772 -12.045  -6.684  1.00  0.00           H
ATOM     17  HB2 ASP A   2     -11.000 -14.117  -5.568  1.00  0.00           H
ATOM     18  HB3 ASP A   2     -10.602 -13.173  -4.127  1.00  0.00           H
.
.
```

Fig. 1. PDB Format

Figure 1 displays the information for the structure of 1C2N. The HEADER, TITLE and AUTHORS records provide information about the investigators involved in defining the structure and other information on the file. The SEQRES records provide the sequences of the peptide chains. We are interested in the ATOM records. The first amino acid GLY (Glycine, symbol G) spans 7 atoms (lines 1-7) and the rest of the atoms correspond to amino acid ASP (Aspartic Acid, symbol D). The 3rd column in each line indicates the type of the atom and the C-alpha atom is indicated by CA (highlighted). The columns 7, 8, and 9 indicate the (X,Y,Z) coordinates of the atom. In the example of Figure 1, the CA atoms have the coordinates (-7.971, -11.573, -8.138) and (-11.025, -11.993, -5.889), respectively.

The Minimotif Miner (MnM) database contains more than three hundred thousand motifs. We only employ those motifs with multiple sources. Let M_i be such a minimotif that occurs in the following set of source proteins: $S_i = \{s_1, s_2, s_3, \ldots, s_n\}$. Note that, if some motif M_i occurs as a substring in some protein s_j it does not mean that s_j is a source of M_i. Whether this is the case or not can only be experimentally validated. On the contrary, M_i may occur multiple times in its source protein s_j. It is not mandatory that all of these occurrences of M_i in s_j are motifs. At least one of these occurrences of M_i is a motif. So it is not enough for us to know only the source protein ID for a motif. We have to know the location l_k of motif M_i in source s_j. The MnM database provides all such information.

PDB is a much smaller and a slowly growing database than Swissprot/Uniprot. This means that there are many motifs in MnM for which we do not have a valid PDB ID. MnM uses a variety of IDs for proteins including Uniprot/Swissprot and Refseq. The mapping between MnM and PDB is done using the mapping files obtained from the following link : `http://www.bioinf.org.uk/pdbsprotec/mapping.txt`.

We have implemented our algorithm using the Center of Gravity algorithm for computing the distance between two structures [1]. The Center of Gravity algorithm is described in the next subsection.

1.3 Center of Gravity Algorithm

This algorithm can be applied to compute the distance between two point sets in any n-dimensional Euclidian space. We explain the algorithm for 3-dimensional case because of simplicity and the scope of our work.

Input : This algorithm takes as input two sets of (x,y,z) coordinates. These are given by $S^{(x,y,z)_i} = \{(x_1^i,y_1^i,z_1^i),(x_2^i,y_2^i,z_2^i),\ldots,(x_n^i,y_n^i,z_n^i)\}$ and $S^{(x,y,z)_j} = \{(x_1^j,y_1^j,z_1^j),(x_2^j,y_2^j,z_2^j),\ldots,(x_n^j,y_n^j,z_n^j)\}$.

Output: Distance between $S^{(x,y,z)_i}$ and $S^{(x,y,z)_j}$. We call it CoG distance.

Algorithm:

```
  BEGIN
  Compute $(x,y,z)$ coordinates of the centroid of $S^{(x,y,z)_i}$.
  This is given by $(x_c^i,y_c^i,z_c^i)$;
  Compute $(x,y,z)$ coordinates of the centroid of $S^{(x,y,z)_j}$.
  This is given by $(x_c^j,y_c^j,z_c^j)$;
  for each of the coordinates $(x_q^i,y_q^i,z_q^i) \in S^{(x,y,z)_i}$ do
      compute the Euclidian distance between $(x_c^i,y_c^i,z_c^i)$ and $(x_q^i,y_q^i,z_q^i)$.
  Let the set of distances be given by $D_i^{Euclidian} = \{d_1^i,d_2^i,\ldots,d_n^i\}$;
  for each of the coordinates $(x_q^j,y_q^j,z_q^j) \in S^{(x,y,z)_j}$ do
      compute the Euclidian distance between $(x_c^j,y_c^j,z_c^j)$ and $(x_q^j,y_q^j,z_q^j)$.
  Let the set of distances be given by $D_j^{Euclidian} = \{d_1^j,d_2^j,......,d_n^j\}$;
  CoG distance is given by $D_{ij}^{CoG} = \sqrt{(d_1^i-d_1^j)^2+(d_2^i-d_2^j)^2+\ldots+(d_n^i-d_n^j)^2}$.
END
```

2 Methods

Our algorithm is based on the following hypothesis: Positive occurrences of the same motif in different sources are structurally similar. Also, the structure of a positive occurrence of a motif and any of its negative occurrences will be dissimilar. To compute the distance between two structures we employ the center of gravity algorithm proposed in [1].

Our algorithm is a learning algorithm that has to be trained with a set of positive and negative examples in the first phase. We evaluate its accuracy in the second phase. A special feature of our algorithm is that it learns the relevant parameters for each individual motif separately. It turns out there is only one parameter that is learnt. This parameter is nothing but a distance threshold between two structures. Let M be any motif. If O_1 and O_2 are the structures corresponding to two positive occurrences of M, then we expect the distance between O_1 and O_2 to be 'small'. On the other hand, if O_1 corresponds to a positive occurrence and O_2 corresponds to a negative occurrence, then we expect the distance between them to be 'large'. Since any learning algorithm requires multiple positive and negative examples to learn from, and our algorithm is motif-specific, we only employ those validated minimotifs in MnM that have multiple sources. Each such source serves as a positive example. Finding negative examples for any biological experiment is in general a challenge since we may not be able to be sure that any data is negative. Like in our previous works on filters, in this paper also we employ random data as negative data. As has been argued before, a random data has a very high probability of being negative.

If M is a motif under concern and if its known sources are $S_1, S_2, \ldots, S_n$, we first get all the occurrences of M in each of the sources. Let these occurrences be $O_1, O_2, \ldots, O_m$. Our hypothesis states that the O_is are structurally similar. Since a motif can occur more than once in the same source, it is the case that $m \geq n$. By structure information we mean a point set in 3D. Specifically, by structure we mean the set of coordinates of the alpha carbon atoms in the motif sequence. This information is available in the PDB files. In this paper we consider only the alpha carbon atoms. Note that including other atoms would only improve the prediction accuracy further. In the final version of the paper we will include other atoms as well.

2.1 Steps in the Algorithm

1. Get a list of all the validated motifs in the MnM database that have multiple sources.
2. Let M be any motif whose sources are $S_1, S_2, \ldots, S_q$. For these source proteins Refseq IDs are available in MnM.
3. We keep only those sources for which structure information is available in PDB. This is done using a Refseq ID$\rightarrow$ PDB ID mapping table.
4. For a given motif M, let its sources for which we are able to get PDB IDs be $S_1, S_2, \ldots, S_n$. We pick one of these sources as the reference for our experiment and call it S_{ref}. The others are used as positive controls. In other words, they serve as positive examples in learning.
5. For each of the positive controls and the reference we apply the Center of Gravity algorithm to perform the following tasks:

a. Compute the Center of Gravity of the alpha carbon atoms in the motif sequence.
b. Compute the Euclidean distances between each of the alpha carbon atoms and the center of gravity. Let these distances in sorted order be $d_1, d_2, \ldots, d_l$, where l is the length of the motif. Note that for every amino acid in the motif there is a single alpha carbon atom. Also note that we will get one such sorted set $\{d_1, d_2, \ldots, d_l\}$ for each of the positive controls.
c. Let the set of distances for the reference S_{ref} be given by $\{d_1^{ref}, d_2^{ref}, \ldots, d_l^{ref}\}$.
d. Calculate the Euclidean distance between $\{d_1^{ref}, d_2^{ref}, \ldots, d_l^{ref}\}$ and $\{d_1, d_2, \ldots, d_l\}$ for each positive control. Let the Euclidean distance for the jth positive control be d_j.
e. Take an average over all the d_js. This is called the *positive mean.*

6. For a given motif M scan through the PDB to look for proteins which are not known to be source proteins for M and in which M occurs as a substring. In other words, exclude the set of positive controls and the reference from the set of all proteins in PDB where M occurs as a substring. This new set is used as the set of negative controls for the motif M. Let this set be $\{N_1, N_2, \ldots, N_t\}$.
7. For each of these negative controls and the reference protein we again apply the Center of Gravity algorithm and compute a distance as in step 5. This will give us the Euclidean distance between $\{d_1^{ref}, d_2^{ref}, \ldots, d_l^{ref}\}$ and $\{d_1, d_2, \ldots, d_l\}$ for each negative control. Let the Euclidean distance for the kth negative control be d_k. We get an average over all of these d_ks and obtain the *negative mean.*
8. We have to come up with a threshold using which we can separate the true positives and false positives. One possibility is to use the negative mean as the threshold. In this case we compute how many of the positive distances d_js are above the negative mean and how many of the negative distances d_ks are above the negative mean.

We expect that a large fraction of positive control distances will be below the negative mean based on our hypothesis.

3 Results

We have tested our algorithm on a collection of almost 650 motifs (that have multiple sources). We have performed two types of analyses. The first analysis is to test the statistical significance of the results obtained using ROC plots. The second analysis measures the accuracy of predictions.

3.1 ROC Plots

For each motif M_i we compute its negative mean $D^-_{M_i}$ and use it as a threshold for predictions. We calculate the number $Count^+_{M_i}$ of distance values below the threshold value, from among the true positive occurrences. This count gives us the true positive rate (TPR). We also calculate the number $Count^-_{M_i}$ of distance values below the same threshold value from among the false positives (i.e., negative control). This number will give us the false positive rate (FPR). According to our hypothesis there should be a good structural similarity between occurrences of a motif in its source proteins. This means that the CoG distance between any two occurrences of the motif in its sources is supposed to be smaller compared to the CoG distance of a true positive occurrence of the motif and a false occurrence of the same motif. We plot FPR (as horizontal axis) vs TPR (as vertical axis) curve and calculate the area under the curve (AUC) for various threshold values. We do this for all the 650 motifs. Table 1 summarizes the outcomes of our experiment.

Table 1. Areas Under the Curves

$AreaUnderCurve(AUC)$	Number of Motifs (as a %)
$> 90\% and \leq 100\%$	29.629
$> 80\% and \leq 90\%$	7.407
$> 70\% and \leq 80\%$	9.259
$> 60\% and \leq 70\%$	7.407
$> 50\% and \leq 60\%$	18.518
$< 50\%$	27.777

Out of the 650 motifs (each having 67.85 positive controls on an average) used for analysis, 216 have got an area under the curve (AUC) between 0.9 and 1. For almost 58 motifs the AUC is exactly 1. This demonstrates the power of our algorithm. The idea is to use our new algorithm only for those motifs for which the AUC is at a level comfortable to a biologist.

3.2 Accuracy Calculation

Accuracy is defined in the following equation:

$$Accuracy = \frac{Number\ of\ +ve\ distances\ below\ threshold\ +\ Number\ of\ -ve\ distances\ above\ threshold}{Total\ number\ of\ distances}$$

Table 2 shows number of motifs in different intervals of accuracy.

We have almost 147 motifs with a prediction accuracy between 90% and 100%. Here again the filter corresponding to the new algorithm is to be used for only those motifs for which the accuracy is at an acceptable level. Figure 2 displays the ROC plots for a randomly chosen subset of the motifs. We show two ROC plots for each category of Table 2.

Table 2. Accuracy

Accuracy	Number of Motifs (as a %)
$> 90\% and \leq 100\%$	22.727
$> 80\% and \leq 90\%$	9.09
$> 70\% and \leq 80\%$	19.696
$> 60\% and \leq 70\%$	15.151
$> 50\% and \leq 60\%$	33.333
$< 50\%$	0

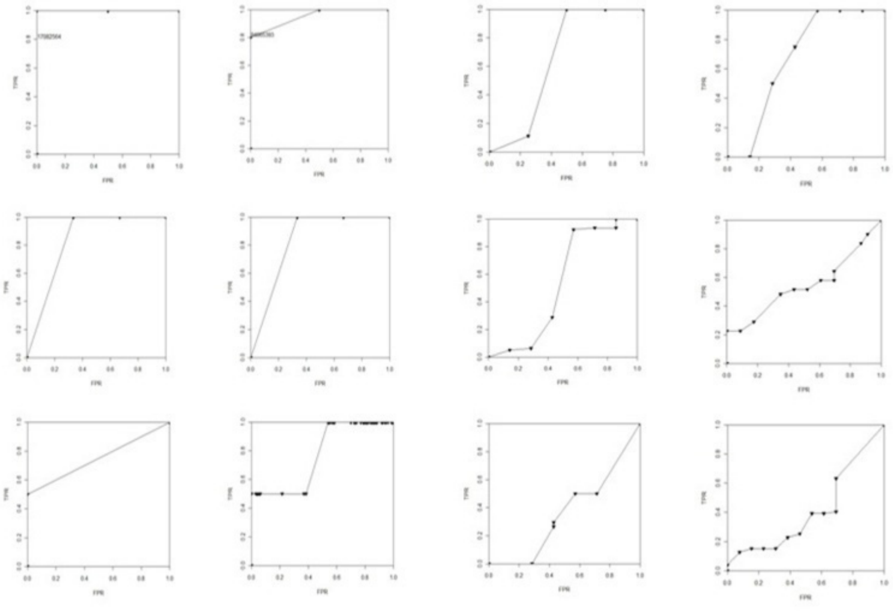


Fig. 2. ROC plots

We plan to integrate the entire data and code as a part of the MnM web system. We will associate a threshold and accuracy/AUC with each of the motifs in the MnM database. Once a user enters a protein query *Q*, MnM reports the putative motifs in *Q*. For any motif *M* if the query is one of the known sources then *M* is reported as a true prediction with an accuracy of 100%. One the contrary, if *Q* is not one of the known sources of M, the filter checks to see if Q is present in PDB. If Q is found in PDB we apply the center of gravity algorithm to compute the CoG distance D_M^Q for M in Q. If the difference between D_M^Q and the CoG distance of M in the reference protein is below the threshold set for M in the MnM database, then M is reported to be a true motif. Accuracy of prediction and AUC value is also reported by MnM. If D_M^Q is above the threshold we will not report M as a putative motif.

4 Conclusion and Future Work

In this paper we have presented a novel structure based algorithm for reducing false positives in the prediction of minimotifs. Our algorithm is a motif-specific learner. We live in an era of personalized medicine and hence this approach is very relevant. The statistical significance of the results obtained as well as the accuracy of the new algorithm demonstrate that the new algorithm is indeed very effective. The outcomes of this work points to the following directions for future work. We want to consider the coordinate information of all the atoms in the amino acids. We want to see the best possible set of features to come up with a better classification accuracy. As mentioned earlier this could only improve the result. Also, we choose the positive reference arbitrarily. We want to extend our the work by choosing each of the positive instances as a possible reference. We will calculate the area under curve and accuracy for each one of them. Finally we choose the best of these scores and the reference associated with it.

Acknowledgements. This work has been supported in part by the following grants: NSF 0829916 and NIH R01-LM010101. We also thank David Sargent for many fruitful discussions.

References

1. Kundeti, V.K., Rajasekaran, S.: A Statistical Technique to Predict Structural Characteristics of Short Motifs, BECAT Tech. Report
2. Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description Version 3.30 Document Published by the wwPDB
3. UniProt Documentation, `http://www.ebi.ac.uk/uniprot/Documentation/`
4. Database of protein domains, families and functional sites, `http://prosite.expasy.org/prosite.html/`
5. Non-redundant databases (NRDB)
6. OWL database, `http://www.bioinf.man.ac.uk/dbbrowser/OWL/index.php`
7. Obenauer, J.C., Cantley, L.C., Yaffe, M.B.: Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Research 31(13), 3635–3641 (2003)
8. Rajasekaran, S., Merlin, J.C., Kundeti, V., Oommen, A., Mi, T., Oommen, A., Vyas, J., Alaniz, I., Chung, K., Chowdhury, F., Deverasatty, S., Irvey, T.M., Lacambacal, D., Lara, D., Panchangam, S., Rathnayake, V., Watts, P., Schiller, M.R.: A computational tool for identifying minimotifs in protein-protein interactions and improving the accuracy of minimotif predictions. Proteins: Structure, Function, and Bioinformatics 79(1), 153–164 (2010)

9. Rajasekaran, S., Mi, T., Merlin, J.C., Oommen, A., Gradie, P., Schiller, M.R.: Partitioning of minimotifs based on function with improved prediction accuracy. PLoS ONE 5(8), e12276 (2010)
10. Rajasekaran, S., Balla, S., Gradie, P., Gryk, M.R., Kadaveru, K., Kundeti, V., Maciejewski, M.W., Mi, T., Rubino, N., Vyas, J., Schiller, M.R.: Minimotif miner 2nd release: a database and web system for motif search. Nucleic Acids Research 37, D185–D190 (2009)
11. Balla, S., Thapar, V., Verma, S., Luong, T., Faghri, T., Huang, C.-H., Rajasekaran, S., del Campo, J.J., Shinn, J.H., Mohler, W.A., Maciejewski, M.W., Gryk, M.R., Piccirillo, B., Schiller, S.R., Schiller, M.R.: Minimotif Miner, a tool for investigating protein function. Nat. Methods 3, 175–177 (2006) (PMID: 16489333)
12. Via, A., Gould, C.M., Gemünd, C., Gibson, T.J., Helmer-Citterich, M.: A structure filter for the Eukaryotic Linear Motif Resource. BMC Bioinformatics 10, 351 (2009), doi:10.1186/1471-2105-10-351
13. Sigrist, C.J.A., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., Bucher, P.: PROSITE: A documented database using patterns and profiles as motif descriptors. Oxford Journals (2002), doi: 10.1093/bib/3.3.265
14. Neduva, V., Russell, R.B.: DILIMOT: discovery of linear motifs in proteins. Nucleic Acids Res. (2006), doi: 10.1093/nar/gkl159
15. Sidman, K.E., George, D.G., Barker, W.C., Hunt, L.T.: The protein identification resource (PIR). Nucleic Acids Research 16(5) (1988)
16. Altschul, S.F., Gish, W., Myers, W.M.E.W., Lipmanl, D.J.: Basic Local Alignment Search Tool. J. Mol. Biol. 215, 403–410 (1990)