Minimotif Miner: a tool for investigating protein function

Sudha Balla[1], Vishal Thapar[1], Snigdha Verma[1], ThaiBinh Luong[1], Tanaz Faghri[1], Chun-Hsi Huang[1], Sanguthevar Rajasekaran[1], Jacob J del Campo[2], Jessica H Shinn[2], William A Mohler[2], Mark W Maciejewski[3], Michael R Gryk[3], Bryan Piccirillo[4], Stanley R Schiller & Martin R Schiller[3,4]

In addition to large domains, many short motifs mediate functional post-translational modification of proteins as well as protein-protein interactions and protein trafficking functions. We have constructed a motif database comprising 312 unique motifs and a web-based tool for identifying motifs in proteins. Functional motifs predicted by MnM can be ranked by several approaches, and we validated these scores by analyzing thousands of confirmed examples and by confirming prediction of previously unidentified 14-3-3 motifs in EFF-1.

Protein domains are highly conserved throughout evolution. It is logical to expect that their binding partners or substrates would also be conserved. These shorter motifs (typically <15 residues) provide complementary information about protein function. For example, the Pro-X-X-Pro sequence forms a left-handed helix that binds to SH3 domains. Identifying a Pro-X-X-Pro motif can be as insightful as identifying an SH3 domain.

There are many databases that focus on identifying small subsets of short motifs. For example, several web-based tools can be used to identify putative or experimentally determined protein phosphorylation sites[1–3]. Eukaryotic Linear Motif (ELM) can predict a broader range of motifs, but it is difficult for users to choose which of the many selected motifs to experimentally pursue[4]. Scansite uses experimental data to generate positional scoring matrices for ranking predicted motif sites[5].

To extend the present motif-search tools, we have generated a motif database and web-based tool (Minimotif Miner; MnM), which complements other databases, allowing researchers to screen a protein for known motifs. A new aspect of MnM is that identified motifs can be ranked by several approaches. Collectively, these scores can be used to reduce the amount of false positives. MnM also provides the frequencies of motif occurrences in proteomes.

There are many databases that catalog small collections of motifs by specific functions such as phosphorylation. These databases are useful if one suspects that a protein may have such functionality. They are not, however, designed to identify unknown functions in proteins. Because the broad functional range of motifs exceeds the knowledge base of most scientists, a comprehensive motif database would be of great value. Analogous to the Pfam collection of protein domains, we have generated a database of motifs with a broad functional spectrum. All motifs in MnM have been experimentally validated and published. In cases for which there are reports of different consensus motifs for a particular function (for example, a general and a more specific consensus motif), we have included all such consensus sequences in the database. In most cases, motif definitions were not derived from a comprehensive analysis of all amino-acid permutations for the motif, a limitation of motif analysis.

At this point MnM contains 462 total motifs of which 312 are unique (**Table 1** and **Supplementary Methods** online), but this version of the database is far from comprehensive. The motifs in MnM fall into three general functional categories: motifs that allow post-translational modifications, binding motifs for other proteins and molecules, and protein trafficking motifs. There is a broad range of motifs thus far identified in these categories and subcategories.

Table 1 | General motif categories of MnM database

Motif type	Number of subcategories
Post-translational modification	116 (total)
Phosphorylation	73
Lipid attachment	8
Phosphatase	1
Glycosylation	3
Proteolysis	26
Other	5
Binding	162 (total)
Protein-protein	146
DNA-RNA	4
Sugar	1
Small molecule	11
Trafficking	34 (total)
Nuclear	3
ER	19
Peroxisomes	4
Endocytosis	8

The database contains a total of 312 unique motifs.

[1]Department of Computer Science and Engineering, University of Connecticut, Storrs, Connecticut 06269, USA. [2]Departments of Genetics and Developmental Biology, [3]Molecular, Microbial and Structural Biology, and [4]Neuroscience, University of Connecticut Health Center, 263 Farmington Ave., Farmington, Connecticut 06030, USA. Correspondence should be addressed to M.R.S. (schiller@nso.uchc.edu).

RECEIVED 9 AUGUST 2005; ACCEPTED 18 JANUARY 2006; PUBLISHED ONLINE 17 FEBRUARY 2006; DOI:10.1038/NMETH856

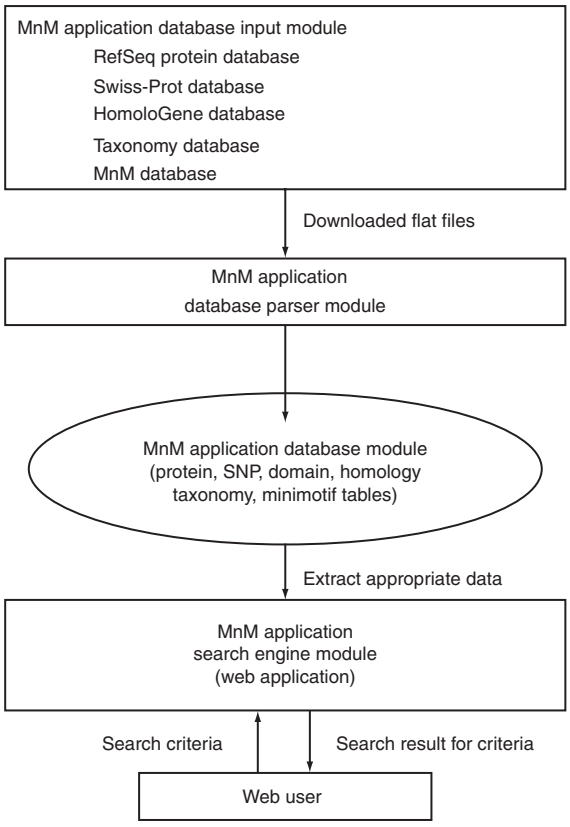

Figure 1 | Schematic representation of the modules of the MnM application. The input module shows the databases that are read into MnM application database parser module. The MnM application database parser module imports the MnM database, several US National Center for Biotechnology Information (NCBI) databases (RefSeq, LocusLink, HomoloGene, Taxonomy, OMIM and dbSNP), the Swiss-Prot database and Pfam database into the MnM application database module[6,13–15]. The database parser module organizes these data into cross-referenced tables of the MnM application database.

For example, there are over 100 protein-protein interaction motifs, but only three consensus glycosylation motifs in MnM.

MnM can be used to search known or new proteins for the presence of motifs in the MnM database (**Fig. 1**). The MnM input window accepts protein queries as a RefSeq accession number or as a protein sequence string. The species to be queried against should be selected to allow proper calculation of motif frequencies. The motif analysis can also be restricted to analysis of motifs that function in specific subcellular compartments. A three-panel output window contains search results (**Supplementary Fig. 1** online).

One of the major limitations in predicting short functional motifs is evaluating which of the many identified motifs in each protein are real and which are false positive. Short motifs often have some degree of degeneracy. Thus, the presence of motifs in a protein may reflect a conserved functional role of the motif, a yet-to-be-discovered structural-functional role or a nonfunctional 'false positive'. To begin to address this limitation, we have implemented evolutionary conservation, surface prediction and frequency scores to rank motifs identified in a protein query (**Supplementary Methods**). Motifs can be ranked (highest scoring on top) using any of these methods by MnM.

Given their functionalities, motifs might be evolutionarily conserved, similarly to protein domains. We wanted to rank motifs in protein queries by the degree of evolutionary conservation. Available BLAST algorithms are not well suited for analyzing the conservation of short motifs, which often have degenerate definitions and can shift position within homologous genes. We therefore developed an algorithm to create evolutionary conservation scores (ECS). MnM calculates ECS scores for queries entered with RefSeq accession numbers and homologs in a Homologene cluster[6]. This algorithm generates positive scores for conserved motifs and negative scores if a motif is absent in one or more homologs. The algorithm also takes into account the overall conservation of homologs.

The three-dimensional structures of several motifs bound to proteins suggest that for motifs to be functional, they need to be on the surface of the protein (for example, Pro-X-X-Pro motifs). Therefore, we implemented surface prediction scoring (SPS) to rank motifs. We adapted the surface prediction algorithm of Naderi-Manesh with an improved information theory algorithm[7,8]. The algorithm is derived from the analysis of 215 different structures in the Protein Data Bank (PDB) and generates scores between 0 and 1; higher scores suggest the motif is on the protein surface. Validation of the scoring algorithm yielded a prediction accuracy of 75% in the training set and 64% in 395 protein structures from the Nh3D test dataset[9].

Frequency scores measure the occurrence of motifs in queries relative to the proteome, but also consider the amino-acid composition and complexity of the motif. Frequency scores greater than 1 indicate the degree to which a motif is over-represented; a frequency score of 1 indicates that a motif is at its proteomic frequency; and frequency scores less than 1 indicate under-representation. The motif table output shows the expected count, observed count and enrichment factor (actual/expected) of the motif in the entire proteome (**Supplementary Table 1** online). To our knowledge, there is no other resource that reports numbers of motifs in proteomes, and this information is useful for gauging the potential motif specificity.

The three scoring methods have been statistically analyzed and each scoring method has global statistical significance (**Supplementary Methods**). Given limitations of each scoring method, we suggest that users select motifs with high scores for the three methods and consider the biological function of the protein and motif.

We analyzed EFF-1 to determine whether an MnM prediction could be experimentally validated. The *Caenorhabditis elegans* EFF-1 protein is necessary and sufficient for most somatic cell fusions in development, and *eff-1* mutant worms can be rescued morphologically by a cloned copy of the *eff-1* gene[10,11]. The mechanism by which EFF-1 functions is unknown and EFF-1 is not homologous to other known proteins. We predict that the carboxy-terminal domain of EFF-1 resides in the cytoplasm. Truncated mutant (*eff-1*(*zz1*)) or alternatively spliced isoforms of *eff-1* lacking the C terminus do not mediate normal cell fusion (**Supplementary Fig. 2** online). MnM analysis identified 23 potential motifs in the C-terminal domain of EFF-1. Among the highest scoring sites are two 14-3-3–binding motifs, at positions 631–636 and 650–654 with frequency scores of 71.6 and 11.4 (calculated with the Drosophila proteome) and SPS scores of 1.0 and 0.35, respectively (**Fig. 2a**).

EFF-1 bound to glutathione-*S*-transferase (GST)–14-3-3, and a phospho–14-3-3 motif antibody recognized EFF-1 (**Fig. 2b** and **Supplementary Fig. 2**). When we mutated either one or both of the

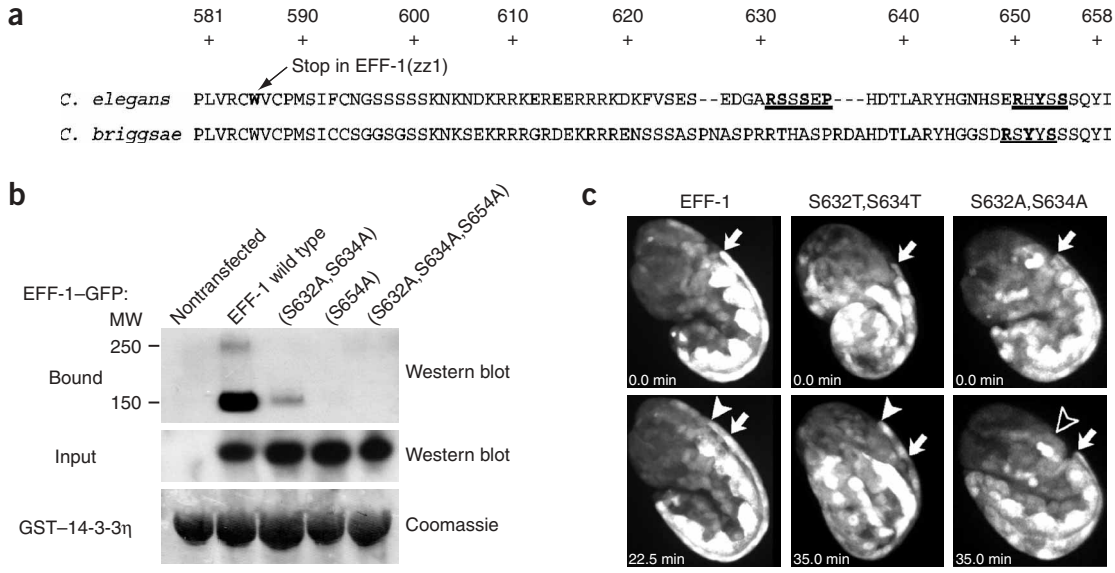

Figure 2 | The 14-3-3 motifs are required for EFF-1 function. (**a**) Sequence alignment of *C. elegans* (RefSeq: NP_741029) and *Caenorhabditis briggsae* (NCBI: CAE57697) EFF-1 cytosolic C termini. The 14-3-3 motifs are underlined; the key motif residues are bolded. (**b**) GST-pulldown analysis using EFF-1 mutant proteins as indicated. Input sample represent 1/50th of the amount used in bound samples and the blot was probed with anti-GFP. (**c**) Time-lapse images of embryonic cell fusions, showing normal fusions with wild-type EFF-1 and S632T,S634T EFF-1, but defective fusion with S632A,S634A EFF-1 (**Supplementary Videos 1–3** online). White arrows and filled arrowheads indicate edges of unfused and fused fields of cells labeled with cytoplasmic GFP. Open arrowhead in failed fusion of S632A,S634A EFF-1 indicates the expected extent of GFP diffusion during normal cell fusion.

14-3-3 motifs in EFF-1, the resulting mutant proteins had markedly reduced affinity to GST–14-3-3η (**Fig. 2b**). These experiments indicate that the 14-3-3 motifs in EFF-1 bind to 14-3-3η *in vitro*. To determine whether the 14-3-3 motifs in EFF-1 were necessary for *in vivo* cell fusion, we mutated the serine residues in each 14-3-3 motif to alanines, which should abrogate binding of the motif to 14-3-3, and to threonine, which can substitute for serine residues in the 14-3-3–binding motif[12]. *eff-1* mutant worms were not rescued by the mutant *eff-1* encoding an EFF-1 with either the S632A,S634A or S654A mutations, but were rescued by mutant *eff-1* encoding a protein containing the S632T,S634T or S654T mutations (**Supplementary Fig. 2**). Worms with threonine-substituted mutants of EFF-1 showed no delay in embryonic cell fusions examined by time-lapse imaging (**Fig. 2c** and **Supplementary Videos 1–3** online). These data strongly suggest that the two 14-3-3–binding motifs in EFF-1 are essential for function and that the nematode genes *par-5* and *ftt-2*, which encode 14-3-3–like proteins, may have a role in cell fusion. In addition, these experiments validate the ability of MnM and frequency scoring to provide insight into the function of proteins through motif identification.

We have yet to be successful in eliminating all false positive motifs because of the short length and degeneracy of most motifs. We note that the purpose of this analysis is to predict new putative functions in proteins, similar to a yeast two-hybrid screen, and that motifs should be confirmed experimentally. Although the ECS, SPS and frequency scores can be used in a complementary manner to rank motifs, these scores each have intrinsic bias. Identification of interdomain regions and enhancement of motif definitions can be used in the future to reduce false positive motif predictions. Presently, interdomain regions can be identified using the ELM server. An analysis with Scansite can also be used as an additional filter to reduce false positive predictions. Users can also select relevant motifs based on relationships to the known biology of a protein. We suspect that a combination of these factors is presently the best approach for selecting biologically relevant motifs. A **Supplementary Discussion** is available online.

MnM is available via our website (http://mnm.engr.uconn.edu).

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This research was funded in part by US National Institutes of Health grants MH65567 to M.R.S., CCR-9912395 and ITR-0326155 to S.R., EB001496 to M.R.G. and HD43156 to W.A.M., and a Muscular Dystrophy Association grant to W.A.M. We thank R. Holz for providing the GST–14-3-3η construct.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Blom, N., Gammeltoft, S. & Brunak, S. *J. Mol. Biol.* **294**, 1351–1362 (1999).
2. Kreegipuu, A., Blom, N. & Brunak, S. *Nucleic Acids Res.* **27**, 237–239 (1999).
3. Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S. & Brunak, S. *Proteomics* **4**, 1633–1649 (2004).
4. Puntervoll, P. *et al. Nucleic Acids Res.* **31**, 3625–3630 (2003).
5. Obenauer, J.C., Cantley, L.C. & Yaffe, M.B. *Nucleic Acids Res.* **31**, 3635–3641 (2003).
6. Wheeler, D.L. *et al. Nucleic Acids Res.* **28**, 10–14 (2000).
7. Naderi-Manesh, H., Sadeghi, M., Arab, S. & Moosavi Movahedi, A.A. *Proteins* **42**, 452–459 (2001).
8. Kloczkowski, A., Ting, K.L., Jernigan, R.L. & Garnier, J. *Proteins* **49**, 154–166 (2002).
9. Thiruv, B., Quon, G., Saldanha, S.A. & Steipe, B. *BMC Struct. Biol.* **5**, 12 (2005).
10. Mohler, W.A. *et al. Dev. Cell* **2**, 355–362 (2002).
11. delCampo, J.J. *et al. Curr. Biol.* **15**, 413–423 (2005).
12. Tzivion, G. & Avruch, J. *J. Biol. Chem.* **277**, 3061–3064 (2002).
13. Pruitt, K.D. & Maglott, D.R. *Nucleic Acids Res.* **29**, 137–140 (2001).
14. Bateman, A. *et al. Nucleic Acids Res.* **32**, D138–D141 (2004).
15. Bairoch, A. & Boeckmann, B. *Nucleic Acids Res.* **19** (Suppl.), 2247–2249 (1991).