Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences

Tian Mi[1], Jerlin Camilus Merlin[1], Sandeep Deverasetty[2], Michael R. Gryk[3], Travis J. Bill[2], Andrew W. Brooks[2], Logan Y. Lee[2], Viraj Rathnayake[2], Christian A. Ross[2], David P. Sargeant[2], Christy L. Strong[2], Paula Watts[2], Sanguthevar Rajasekaran[1,*] and Martin R. Schiller[2,*]

[1]Department of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269-2155, [2]School of Life Sciences, University of Nevada Las Vegas, 4505 Maryland Pkwy., Las Vegas, NV 89154-4004 and [3]Department of Molecular, Microbial, and Structural Biology, University of Connecticut Health Center, 263 Farmington Ave., Farmington, CT 06030-3305, USA

ABSTRACT

Minimotif Miner (MnM available at http:// minimotifminer.org or http://mnm.engr.uconn.edu) is an online database for identifying new minimotifs in protein queries. Minimotifs are short contiguous peptide sequences that have a known function in at least one protein. Here we report the third release of the MnM database which has now grown 60-fold to approximately 300 000 minimotifs. Since short minimotifs are by their nature not very complex we also summarize a new set of false-positive filters and linear regression scoring that vastly enhance minimotif prediction accuracy on a test data set. This online database can be used to predict new functions in proteins and causes of disease.

INTRODUCTION

A common theme in protein activity regulation is the binding of a structural domain of one protein to a short, contiguous peptide segment of another. From a bioinformatics perspective, identifying domain signatures has been incredibly useful in formulating hypotheses regarding the biological function of otherwise uncharacterized proteins. The success of such methods is due in part to the high sequence complexity of these relatively large domains (approximately 100 residues in length), as well as their common evolutionary heritage, which allow for high-confidence domain identification with few false positives. The short, contiguous segments [termed minimotifs or short linear motifs (SLiMs)] are just as useful in identifying the roles of proteins, but are more difficult to identify with high accuracy. Nevertheless, several bioinformatics resources exist for querying protein sequences for the existence of minimotifs, including Minimotif Miner (MnM), the Eukaryotic Linear Motif (ELM) resource and other specialized databases (1–10). It remains an ongoing pursuit to increase both the sensitivity and accuracy of minimotif prediction in proteins.

This article summarizes the latest release and developments of the MnM database and webserver, version 3.0; additional details can be found in the new MnM user guide on the MnM website. Efforts since our last release in 2008 (4) have concentrated on two fronts: improved filters which increase the accuracy of minimotif prediction by removing false positives (11–13), and increasing the size of the MnM database through both manual annotation of minimotifs from the literature and federation with other databases including PhosphoSite, DOMINO, MEROPS, UniProt, PepX, 3DID, PeptiDB and HPRD (14–21). MnM 3 now includes a total of 294 933 minimotif definitions, consisting of 880 consensus minimotifs and 294 053 instances. These minimotifs span three biological activities: trafficking, binding and modifying. Multiple filters have been introduced since our 2008 release of MnM 2, the most important being a combined filtering approach that can result in 90% accuracy of minimotif prediction with few false positives using one scoring

---

threshold or even 38% identification of minimotifs with no false positives with a more stringent threshold (submitted for publication). The score used by this combined filter is now used as the default ranking of the minimotif list, rather than the frequency score used in MnM 2.

Besides their role in recognition by protein domains, minimotifs have a number of important biological roles. In addition to binding, minimotifs are often determinants for post-translational modifications and trafficking proteins to specific parts of cells. Minimotifs are also involved in cell signaling and regulation (22,23). A number of minimotifs are mutated in different disease and pathogens such as viruses tend to exploit host machinery by viral encoded minimotifs (24–26). Due to their role in disease, the actions of several drugs are based on a minimotif-mimetic mechanism (27,28).

RESULTS

Revised minimotif model and new entries in Minimotif Miner 3.0

Prior to adding minimotifs to the MnM 2 database, we first reevaluated our previous model, which presented 22 attributes of a minimotif (12). We have now revised this model to include 28 attributes as shown in Figure 1. This model contains a protein sequence definition and a functional definition where the sequence definition describes the chemistry of the motif. The sequence definition can be an instance or a consensus sequence. Instances are the exact amino acid sequence found in the protein that contains the minimotif; whereas, a consensus sequence is an interpretation of a set of instances that indicates degeneracies at certain positions in the amino acid sequence. The consensus sequence definition format is largely based on that previously proposed by the Seefeld Convention and later modified for MnM (12,29). These modifications include an extensible expanded definition of the covalent chemistry of the minimotif containing the position within the protein, any modified residues and their position in the sequence, and a description of any post-translational modifications of amino acids in the sequence and corresponding accession numbers from the Psi-Mod database (30).

The functional component of the minimotif model is centered around a syntactical triplet where the motif source is the subject, the activity is the verb and the target that engages the minimotif is the object. There are unique properties to this triplet such as an affinity, structure, minimotif reference, database reference for cross referencing external databases and experimental evidence that support the minimotif.

The motif source, activity and target have a number of attributes previously modeled, but here we have renamed the 'required modification' to 'motif modification' to better distinguish this attribute from 'activity modification'. Motif modification is when a motif needs to be covalently modified to engage the target such as when a minimotif must be phosphorylated to bind 14–3-3. Whereas, an activity modification describes a situation where the target is an enzyme that covalently modifies the minimotif as when a minimotif becomes myristoylated. The description of these modifications requires more detail than in our original model. To accurately describe these modifications the new model includes
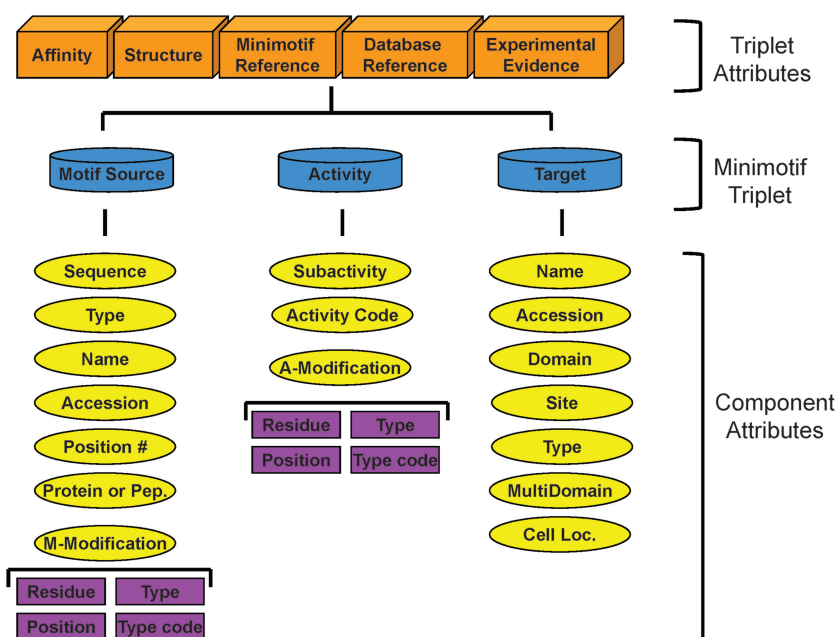
Figure 1. Revised minimotif model. The key elements of the minimotif syntax are colored blue. Orange boxes indicate attributes that are unique to specific minimotif triplets. Yellow ovals are for different attributes of minimotif triplet elements. All attributes except those in the purple boxes were previously described in our minimotif model and the purple boxes are new attributes to define motif modifications and activity modifications (12).

Table 1. Growth of minimotif entries in MnM

Category	MnM	MnM 2	MnM 3
Total			
Motif sequences	462	5089	294 933
Consensus sequences	312	858	880
Instance sequences	44	4229	294 053
Post-translational modifications	116	663	210 949
Binding	162	4689	4922
Trafficking	34	195	228
Required for cell process	–	–	47
Unique			
Motif sequences	312	2224	185 833
Motif proteins	<312	1211	49 671
Motif targets	<312	687	2620

the 'residue' that is modified, the sequence 'position number' of the sequence, the 'type' of modification and the 'type code' number of the modification, which for the most part makes use of accession ids in the Psi-Mod database (30).

Since the release of MnM 2, the total number of minimotif sequences has increased from about 5300 to almost 300 000 (Table 1) (4). The majority of these new entries have been gleaned from federation with other open databases and some were manually annotated from the primary literature using the MimoSA annotation helper application designed for this task (31). All minimotif entries are annotated using our revised minimotif syntax model (Figure 1) (12). The majority of growth comes from addition of instances. We have focused on instances because the Minimotif Miner query engine can be used to generate consensus sequence from any set of instances (12). Most of the minimotifs added from external sources are for post-translation modifications. Minimotifs are found in approximately 50 000 different proteins in many different species; most minimotifs are for mammalian organisms, although MnM does contain some bacterial, yeast and invertebrate minimotifs as well. The number of domains that interact with or associate with minimotifs in MnM is approximately 2600 suggesting that there are still many minimotifs yet to be discovered.

There is a minimal set of attributes necessary to define a minimotif for entry into MnM. The minimotif sequence types are classified as either an instance or a consensus sequence, which each have a minimal set of attributes. Consensus sequence definitions must have an amino acid sequence less than 15 residues long, activity and subactivity, literature reference, one or more experimental techniques in support of the minimotif, and annotation of any post-translational modifications to the minimotif sequence (residue modified, position in sequence, type of modification and Psi-Mod id for the modification, if available). Instances contain this attribute set, but also must have a name of the sequence harboring the minimotif, whether the source protein is a peptide fragment or a protein, and if a protein, must have an accession number to one of the available protein databases. While we prefer to have information about the target molecule

that is associated with the minimotif, this is not required in the minimal set because there is value for such database entries in that this information can be used to identify unknown targets by mining-based approaches. For example, an instance of a phosphorylation site on a protein substrate can be used with kinase consensus sequences in the database to predict the target kinase. The 28 attributes of minimotifs are stored in a MySQL database. For the approximately 6000 manually annotated minimotifs, all 28 attributes were entered, except in the cases where information was not available from the literature. For example, some minimotifs do not have structures or affinities. We note that many of the minimotifs imported from external databases have the minimal set of information required to define a minimotif, but are often missing many of the other attributes defined in our model; we only imported minimotifs that have the minimal set of attributes.

Minimotif filtering to reduce false-positive predictions

The major difficulty in identifying functional minimotifs within a protein sequence of interest is the high false-positive rate—that is, a large number of predicted minimotifs do not perform the predicted biological function, but coincidentally share the minimotif sequence signature present in other biologically active proteins. These false-positive predictions are notoriously difficult to filter out based on sequence definitions alone, due to the inherently low-sequence complexity of minimotifs (7,32). However, additional context information (beyond amino acid sequence) can be used to narrow the search and effectively filter these false positives, increasing the accuracy of minimotif prediction (11,13). Such context information is routinely employed by individual researchers when evaluating minimotif prediction results. For instance, a researcher studying nuclear import in mouse neurons would quickly discard motif predictions regulating bacterial cell division. In this case, the researcher would be imparting context-specific information about molecular function, cellular function and taxonomy to rule out an obvious false positive. While effective, such a filtering technique is highly inefficient both in the time it takes an individual to prune the results list, as well as in the breadth of understanding required to effectively filter all false positives. Over the past 2 years, several contextual filters have been added to the MnM web service, which have been demonstrated to be highly effective in improving the accuracy of minimotif prediction thereby increasing the ease of interpretation of the Minimotif Miner results (11,13).

The original implementation of MnM 1.0 did not attempt to filter any false positives, but ranked minimotif predictions in descending order of sequence complexity. A scoring metric for location of a minimotif on a protein surface, and evolutionary conservation among divergent species was also provided. MnM 2 allowed the user to filter the results list based on particular minimotif activities of interest and also separated minimotif instances from consensus definitions. Neither of these functionalities formally removes false positives, they

simply assist the user in filtering based on his/her own knowledge.

Our first step toward knowledge-based filtering was to model minimotif definitions in a richer way—formally modeling the source and target proteins (12). Tying the source/target protein information to the minimotif definition provides a relationship to the taxonomy of the observed activity, a relationship to other related species through homology databases/searches, a relationship to molecular and cellular function through the source/target annotation and the use of the Gene Ontology definitions, and a relationship to other proteins in the same biological pathway through protein–protein interaction databases (21,33–37). In this manner, the use of context-specific definitions can be applied to filter false-positives computationally, removing that burden from the user.

Molecular/cellular function. Knowing that the source protein and target interact directly is highly helpful in identifying true positives in a minimotif search, as minimotif activities require an interaction between the source and target. In the absence of such direct interaction, filtering for source/target pairs, which are active in the same molecular/cellular pathway can also be useful. Functions of source/target pairs are accessed from the Gene Ontology database, which allows for filtering based on the molecular/cellular function of the source/target pairs (38). This filtering technique can be used to restrict results to only source/target pairs, which share a common function, or can be extended to source/target pairs that share a related function. Three thresholds are provided in MnM 2.1 for varying the relatedness of the functions to be filtered (13). The best performing cellular function filter is estimated to result in ∼26% sensitivity with 6% selectivity for a combined discrimination ratio of 4.6 whereas the best performing molecular function filter has a discrimination ratio of 2.9 (Table 2). Sensitivity is the percentage of true positives that are not filtered out, whereas selectivity is the percentage of true negatives that are not removed by the filter (11,13). The discrimination ratio is sensitivity/selectivity.

Protein–protein interactions. MnM 2.2 allowed the user to filter results based on known protein–protein interaction (PPI) networks (11). The logic behind this filter is that minimotif predictions are filtered on the basis of

experimental verification of the interaction between source and target proteins. MnM makes use of six external databases containing more than 300 000 non-redundant PPIs: DiP, Entrez Gene, HPRD, MINT, VirusMINT and IntAct (21,33–35,37,39). In the most stringent use of the PPI filter, only exact matches between source and target are reported, and the predicted minimotif represents a hypothetical mode of interaction for the known PPI. While effective, this stringent filter is limited due to the relatively small number of established PPIs. For this reason, the user can extend the filter to include homologous proteins for both the known source and known target of the PPI. This can be done in one of two ways: one, by accessing homologous protein clusters via the HomoloGene database; two, by using BLAST similarity searches to predict homologous proteins not included in HomoloGene (39). Ten default BLAST thresholds are provided in the motif filtering dialog box (Figure 2) accounting for a total of 12 possible PPI filtering choices. The base-level PPI filter is estimated to result in ∼62% sensitivity with 2% selectivity for a combined discrimination ratio of 29; this is the best performing filter and significantly reduces false positives (Table 2).

Genetic interactions. A genetic interaction (GI) helps to identify that there is a functional relationship between two proteins. In some cases, this can be due to direct interactions or modifications of one protein by another. If a minimotif source and predicted target protein have a GI, this prediction can provide a mechanistic explanation for the observed relationship. Since the two proteins of a GI have this relationship, these proteins are more likely to have a minimotif than two unrelated proteins. This concept was implemented in three different GI filters on MnM 2.3 (submitted for publication). The basic GI filter identifies those motif/target pairs where there is a known GI and was the GI filter with the highest accuracy; the GI-node based filter extends the GIs for the sources and target an additional interaction away to a path length of 2; the GI-HomoloGene filter takes advantage of orthologous GIs. The basic GI filter had a discrimination ratio of 7.3, which was better than the GI-node and GI-HomoloGene filters with ratios of 4.5 and 2, respectively. The primary difference comes from a poorer selectivity in removing true negatives (∼3% versus ∼12%); similar sensitivities of 21% and 24% were observed for these filters. The basic GI filter also had a better discrimination ratio than

Table 2. Comparison of different minimotif filters

Minimotif filter	Area under ROC curve	*P*-value	Discrimination ratio	Reference
Frequency score	0.7	0.08	ND	(11,13)
Cellular function	0.7	0.12	4.6	(13)
Cellular function + frequency score	0.9	0.0002	–	(13)
Molecular function	0.8	0.03	2.9	(13)
Molecular function + frequency score	0.9	0.002	–	(13)
Protein–protein interaction	0.9	0.001	12.5	(11)
Genetic interaction	–	–	7.3	Submitted for publication
Surface prediction filter	0.3	1	–	(7)
Multifilter	0.94	$9.7\ e^{-278}$	–	Submitted for publication

Filter	Subtypes	Not Applicable	Include	Exclude
Motif Activity	Binds	◉	○	○
	Post-Translational	◉	○	○
	Requires	◉	○	○
Motif Sub-Activity	Trafficks	◉	○	○
	Phosphorylated	◉	○	○
	Proteolyzed	◉	○	○
Motif Class	Concensus	◉	○	○
	Instance	◉	○	○
Protein-Protein Interaction	PP Interaction	◉	○	○
	PPI-HomoloGene	◉	○	○
	PPI-Homology-10*	◉	○	○
	PPI-Homology-20*	◉	○	○
	PPI-Homology-30*	◉	○	○
	PPI-Homology-40*	◉	○	○
	PPI-Homology-50*	◉	○	○
	PPI-Homology-100*	◉	○	○
	PPI-Homology-150*	◉	○	○
	PPI-Homology-200*	◉	○	○
	PPI-Homology-500*	◉	○	○
	PPI-Homology-1000*	◉	○	○
Molecular/Cellular	Cellular1	◉	○	○
	Cellular2	◉	○	○
	Cellular3	◉	○	○
	Molecular1	◉	○	○
	Molecular2	◉	○	○
	Molecular3	◉	○	○
Genetic Interaction	GI	◉	○	○
	GI-Homologene*	◉	○	○
	GI-Node*	◉	○	○
Combined Filter	Maximum Accuracy*	◉	○	○
	Minimum False Positives*	◉	○	○
	* May take several minutes to process			
	Apply Filter(s)			

Figure 2. Screenshot of minimotif filter selection page. Screenshot of MnM 3 filter section for choosing approaches for filtering out false-positive minimotifs.

the cellular and molecular function filters, but was not as high as the PPI filter (Table 2).

Combined filter approach. When examining the effectiveness of the cellular / molecular function filters, we discovered that combining two filters provided greater accuracy in minimotif prediction than either filter alone (13). We have recently extended this idea by training a linear combination of all the filters to maximize both the accuracy and specificity in minimotif search. Using this approach with one threshold, the resulting combined filter allows us to increase accuracy to 90%, while a

more stringent threshold retained 39% of the true positives while rejecting all false positives in a large test data set. The elimination of all false positives represents an important milestone in minimotif prediction. MnM 3 provides access to all minimotif filters and now ranks all minimotif predictions using this new combined filter score. The results can also be filtered according to either the threshold for maximizing accuracy or for maximum stringency in removing all false positives (Figure 2). In the minimotif results table, minimotifs with scores above the threshold of 0.91 are highlighted green (produce no false positives on a test data set), between 0.24 and 0.90 are highlighted yellow (produce high recovery of true minimotifs with only 2% false positives on a test data set). Experimentally validated minimotifs are distinguished from predictions by highlighting the minimotifs blue in the results table.

Uses of Minimotif Miner

The major workflow in minimotif miner is to search a single protein query for the presence of minimotifs. This is geared toward identifying new functions in proteins, minimotif determinants of protein–protein interactions, or for matching post-translational modifications with potential enzymes that catalyze such modifications. Many different subsets of minimotifs can be selected by using the filtering section of the MnM results page (Figure 2). Once a custom filter combination is selected with radio buttons, selection of the 'Apply Filter(s)' button will re-populate the motif results table with the search results. As the filtering approaches have grown, this has now been moved from the bottom of the protein sequence menu to a separate expandable section on the results page.

To provide an example of the uses of different filters we explore a sample analysis of HIV-1 Nef, (NP_057857). We chose Nef because it is a well-studied protein and we could evaluate minimotif predictions using HIVToolbox (40). Although we expected better filtering from the new algorithm, only a small portion of known minimotifs have been identified and added to MnM, thus we would only expect that a subset of predicted minimotifs would have been previously experimentally validated. In the old MnM 2 output, the minimotifs were rank ordered by frequency score. This ordering often strongly selects for a high ranking of instances, which generally are far more information-rich than consensus sequences. The new ranking in MnM 3 depends on many types of different data and filter testing indicates that the new filters are superior to the minimotif ranking used in MnM 2.

The new default filter ranks 21 minimotif predictions for Nef with a score between 0.24 and 0.91 where few false positives were observed when a test data set was analyzed (Figure 3). This figure shows 19 minimotifs that are colored blue indicating support by experiments in the literature; we note that two of these minimotifs have scores below 0.24 and three do not have scores because of missing information. Of the other seven high scoring predicted minimotifs, two of these minimotifs were for previously known interactions of Nef with the SH3 domain of Fyn and with AIP-1; both were annotated and added to

the MnM database (41,42). One was a for a c-Raf1 binding motif consensus sequence where there is a verified instance. A minimotif was identified for binding to the β subunit of AP1, AP2 and AP3, which was previously known and has now been added to MnM; the motif predicted to interact with the AP2 and AP3 μ subunit was not previously identified (43–46). MnM predicted phosphorylation of Nef by PKCα at three sites: 15, 80 and 103, none of which were present in MnM. In support of these predictions, Nef is known to be phosphorylated at Thr 15, which is inhibited by a PKC inhibitor (47,48). We note that only 29% of >7000 HIV isolate sequences have a Thr in this position of Nef; whereas 98% of these viruses have a Ser at position 103 [analysis with HIVToolbox (40)]. Ser 103 was suggested to be phosphorylated by PKCα *in vitro* (49). Thr at position 80 was predicted as a PKCα site, but no evidence supporting phosphorylation of this site by PKCα could be identified in the literature. MnM also predicted a novel interactions of Nef with the C-terminal SH3 domain of Grb2 and a site that binds peptidylprolyl isomerase. In summary, of the 24 minimotifs identified by the MnM analysis (including one with three distinct sites) with scores above a major false-positive threshold, 21 had previously been demonstrated and we cannot rule out the possibility that the other three have not yet been discovered. Although Nef is well studied, most proteins have many minimotifs predicted with scores above 0.24 that are yet to be investigated.

Some scientists may want to analyze many protein sequences at once. We have now enabled this type of workflow as an email service for batch query input mode on the MnM input page. The input file for the request must contain a list of protein accession numbers from one or more various data sources (UniProtKB, MIM, RefSeq, Ensemble, UniGene, MIM, PIR, Entrez Gene) and/or protein sequences; this format is indicated in a hyperlink in this section of the input page.

Another workflow is identifying minimotifs that play a role in human disease, or organism diversity as originally reported (27,50). In MnM 2, this was accomplished by mapping missense SNPs located in protein coding regions from the dbSNP database (4,39). In the View menu on the results page, the 'View SNPs' selection reveals known SNPs in the Protein Sequence window highlighted blue and capitalized. When any SNP is clicked, the SNP has a green highlight and the amino acid change is shown. Any combination of SNPs can be selected. The 'View motifs from New SNPs' found under the 'View' menu item will create a new table that identifies any minimotifs that are introduced or eliminated by the selected SNPs. Since many SNPs are for disease-associated mutations, this tool can be used to formulate new hypotheses about disease mechanisms.

DISCUSSION AND CONCLUSIONS

We have expanded the model for Minimotifs to contain 28 attributes that offers a number of advantages. Some advantages are the segregation of specific attributes are that

[X]	MOTIF EXPRESSION	ANNOTATION	OCCURRING POSITIONS	COMB. FILTER SCORE	# IN PROTEOME
	MGGKW	This instance motif in Nef is myristoylated by NMT1; The Glycine at position 2 in Nef must	1	0.49	1
	DDPERE	This instance motif in Nef binds C-Raf1	174	0.49	1
	EEEE	This instance motif in Nef binds AP1 mu1	62	0.49	1
	ENTSLL	This instance motif in Nef binds AP2	160	0.49	1
	FPD	This instance motif in Nef binds Peroxisomal acy-CoA thioesterase 1	121	0.49	1
	NKGENTSLLHPVSLH	This instance motif in Nef binds V1H	157	0.49	1
	PLRP	This instance motif in Nef binds the SH3 domain of fyn	75	0.49	1
	PQVPLR	This instance motif in NEF binds the SH3_1 domain of FYN	72	0.49	1
	PQVPLRP	This instance motif in NEF binds the SH3_1 domain of FYN	72	0.49	1
	PQVPLRPMTYKA	This instance motif in Nef binds the SH3 domain of Lck	72	0.49	1
	PVEP	This instance motif in Nef binds the SH3 domain of Hck	147	0.49	1
	SKSSV	This instance motif in Nef is phosphorylated by PKA; The Serine at position 6 in Nef must I	6	0.49	1
	TVRE	This instance motif in Nef is phosphorylated by PKC alpha; The Threonine at position 15 in	15	0.49	1
	YPLT	This instance motif in Nef binds AIP-1	135	0.49	3
	EE	This instance motif in Nef binds Arf1	62, 64, 154	0.41	47
	RMRRAEPA	This instance motif in Nef binds GATA-1	19	0.21	1
	TVR	This instance motif in NEF is phosphorylated by PKCalpha; The Threonine at position 15 in	15	0.21	6
	MGGKWS	This instance motif in NEF is lipidated by N-myristoyl; The Glycine at position 2 in NEF mu	1	null	1
	PQVP	This instance motif in NEF is modified by an unknown target;	72	null	1
	QVPLRPMTYK	This instance motif in NEF binds HLA class I histocompatibility antigen. A-11 alpha chain	73	null	1
	DDxxxE	This consensus motif binds C-Raf1	174	0.49	5
	[DE]xxxL[LI]	This consensus motif binds the Trunk domain of AP1,AP2,AP3 beta subunits	160	0.49	1
	PxxxR	This consensus motif binds the #2 SH3 domain of Grb2	25	0.49	9
	GP	This instance motif in Capsid binds the peptidylprolyl isomerase domain of peptidylprolyl i	130	0.49	39
	[DE]D	This consensus motif binds AP2	174	0.49	34
	Yxx[FILMVWY]	This consensus motif binds AP1,AP2,AP3, AP4 mu subunits and is trafficked by endocytos	143	0.49	83
	[ST]x[KR]	This consensus motif is phosphorylated by PKCalpha; The Threonine at position 1 must be	15, 80, 103	0.41	86
	Px[AP]x[PV]R	This consensus motif binds the #3 SH3 domain of CIN85	129	0.21	1
	RRA	This consensus motif binds Sar1 and is trafficked by Endoplasmic Reticulum Export	21	0.21	3
	PxTP	This consensus motif is phosphorylated by CDK1/CylcinB1 complex; The Threonine at pos	69	0.21	1
	PxVxL	This consensus motif binds the Chromo Shadow domain of HP-1	72	0.21	12
	PxxPxK	This consensus motif binds the #1 SH3 domain of Crk	147	0.21	1
	Px[ST]P	This consensus motif is phosphorylated by Erk; The Serine at position 3 must be modified	69	0.21	6
	Rx[KR]R	This consensus motif is proteolyzed by Furin;	19	0.21	8
	Rx[RK]R	This consensus motif is proteolyzed by Furin;	19	0.21	8
	[FWY]x[FWY]xxxx[FWY]	This consensus motif binds Caveolin 1	113	0.21	10
	[ILV]ExD	This consensus motif is proteolyzed by Granzyme B;	148	0.21	3
	[ED]Y[IFLV]	This consensus motif is phosphorylated by EGFR; The Tyrosine at position 2 must be mod	201	0.21	4
▽	HxxK	This consensus motif binds the #2 SH3 domain of GADS	89	0.21	18
	YxxQ	This consensus motif is phosphorylated by Stat3; The Tyrosine at position 1 must be modif	115	0.21	12
	IY	This instance motif in peptide is phosphorylated by cFGR; The Tyrosine at position 2 in pep	114	0.21	24
	SxD	This consensus motif is phosphorylated by CamKII; The Serine at position 1 must be modi	34	0.21	11
	[PSAT]x[QE]E	This consensus motif binds the TRAF domain of TRAF2	60	0.21	19
	PxxxD	This consensus motif binds platelet fibrinogen receptor	147	0.21	17
	RxxR	This consensus motif is proteolyzed by Prohormone Convertase 7;	19	0.21	33
	PxP	This consensus motif binds Vif	129	0.21	26
	DxxG	This consensus motif binds the G-protien domain of Phosphate	28	0.21	8
	RxGx	This consensus motif binds the phosphatase domain of PP1alpha	29	0.21	41
	GPx	This consensus motif is proteolyzed by TTP;	130	0.21	39
	TQ	This instance motif in Chk2 (isoform a) is phosphorylated by ATM; The Threonine at positi	117	0.21	18
	[EDY]Y	This consensus motif is dephosphorylated by TC-PTP; ; The Tyrosine at position 2 must b	201	0.21	29
	ExT	This consensus motif binds Skp2; The Threonine at position 3 must be modified with a O-p	160	0.21	14
	Yxx[LIV]	This consensus motif is phosphorylated by JAK2; The Tyrosine at position 1 must be modi	143	0.21	53
	[ST]xxE	This consensus motif is phosphorylated by Casein Kinase II; The Serine at position 1 must	15	0.21	87
	Yxx[VILFWYM]	This consensus motif binds the SH2 domain of ZAP70; The Tyrosine at position 1 must be	143	0.21	83
	Txx[IL]	This consensus motif binds the #2 FHA domain of Rad53; The Threonine at position 1 mus	162	0.21	76
	[ST]xxx[ST]	This consensus motif is phosphorylated by GSK3 beta; The Threonine at position 1 must b	44	0.21	91
	[ST]xx[DE]	This consensus motif is phosphorylated by Casein Kinase II; The Serine at position 1 must	15	0.21	126
	PxxPx[KR]	This consensus motif binds the SH3 domain of Src	72, 147	0.17	7
	YxL	This consensus motif binds the #2 FHA domain of Rad53; The Tyrosine at position 1 must	135, 143	0.17	29
	TP	This instance motif in ELK-1 is phosphorylated by Erk; The Threonine at position 1 in ELK-	71, 128	0.17	41
▽	RxL	This consensus motif binds Cdk2	35, 196	0.17	56
	Sxx[ST]	This consensus motif is phosphorylated by Casein Kinase I; The Threonine at position 4 m	6, 45	0.17	53
	Rxx[ST]	This consensus motif is phosphorylated by CamKII; The Threonine at position 4 must be m	77, 184	0.17	28
	[ST]P	This consensus motif is phosphorylated by Erk; The Threonine at position 1 must be modif	71, 128	0.17	76
	[ST]Q	This consensus motif is phosphorylated by ATM; The Serine at position 1 must be modified	103, 117	0.17	50
	Txx[SA]	This consensus motif binds the FHA domain of KAPP; The Threonine at position 1 must be	51, 80	0.17	55
	[KR]xx[ST]	This consensus motif is phosphorylated by RSK; The Serine at position 4 must be modifiec	77, 184	0.17	89
	[RK]x[RK]	This consensus motif binds Sar1 and is trafficked by Endoplasmic Reticulum Export	17, 92	0.17	138
	[TS]xxxx[VI]	This consensus motif binds the SH2 domain of SH2D1A	6, 80, 128, 163	0.09	112
▽	MG	This instance motif in Src is myristoylated by an unknown target; The Glycine at position 2	1	null	4
	MGxxx[ST]	This consensus motif is myristoylated by an unknown target; The Glycine at position 2 mus	1	null	4
▽	GMD	This instance motif in SPS1 binds to ATP;	172	null	6
▽	PVTP	This instance motif in CARML is modified by an unknown target;	69	null	1
	Px[ILMV]PxRP	This consensus motif binds the SH3 domain of Las17p	72	null	1
	NYT	This instance motif in CSG is glycolysated by Glycosyl; The Asparagine at position 643 in (	126	null	3
	LDL	This instance motif in unknown protein (Q51451) binds the 14-3-3 domain of 1433B	110	null	1
	RExxE	This consensus motif binds Iron	178	null	1
	EEV	This instance motif binds the SH2 domain of an unknown target	64	null	3
▽	HH	This instance motif in AIFM1 binds to FAD;	192	null	4
▽	SLL	This instance motif in QORX binds to NADP;	163	null	3

Figure 3. Results table in MnM 3 from analysis of HIV-1 Nef protein. Nef (NP_057857) was analyzed to produce the minimotif predictions shown. Column 1 shows the minimotif sequence, column 2 shows the function of the minimotif, column 3 shows the amino acid position(s) for the start residue in the minimotif, column 4 shows the combined filter score, and column 5 shows the number of occurrences of each motif in the entire HIV-1 proteome. Rows colored blue are for minimotifs that are experimentally validated, yellow are above a threshold for high accuracy prediction, and red are below this threshold or do not have data to calculate a combined filter score (null).

it reduces ambiguities and faulty annotations, can be readily used to identify missing data, and allows the use of many different types of controlled vocabularies. The rich model enables easy mining of data through SQL queries in a number of ways. For example, MnM allows a widget-based custom query builder to mine Query Engine MnM database. Using this tool, consensus sequences or position-specific scoring matrices can be generated for minimotifs where many different instances were studied, often in separate laboratories. In this manner, our model that maintains the instance information in its raw form becomes a rich source for automated generation of consensus motifs.

MnM offers a unique resource that is synergistic with other minimotif search tools. MnM is a broad-based minimotif resource that covers all types of minimotifs from any species with now approximately 300 000 minimotifs. A brief comparison with some other motifs tools is highlighted, but there are far too many tools to present a comprehensive review. The Eukaryotic Linear Motif Server is the closest broad-based minimotif resource with 170 consensus motifs and 1817 instances (6). Phospho-ELM, an associated database that focuses on phosphorylation sites has approximately 42 000 instances (5). These tools use a different, but overlapping set of approaches to help reduce false positives. Other sites such as Scansite and DomPep use position specific scoring matrices for predicting new instances, but focus on a set of protein binding domains (3,10). MOTIPs can be used to search proteomes for minimotifs (9). SLIMSearch 2.0 and MyHits allow proteome search of user-defined motifs (8,51).

Minimotif Miner 3 is an important improvement over MnM 2. The number of minimotif sequences has increased two orders of magnitude, vastly improving the sensitivity of minimotif search. This large increase in the number of potential minimotifs could potentially hinder researchers rather than help if not for the aid of filtering mechanisms to reduce the number of false positives. The new filtering mechanisms recently introduced, based on protein–protein interactions, molecular function, cellular function, genetic interactions and the combined filter, greatly improve the accuracy and specificity of MnM 3 search results.

AVAILABILITY

The MnM database can be accessed through single protein or batch queries using the MnM user interface. The entire database is not currently available for download, but the MnM investigators are open to collaborations that involve using the database.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Puntervoll,P., Linding,R., Gemund,C., Chabanis-Davidson,S., Mattingsdal,M., Cameron,S., Martin,D.M.A., Ausiello,G., Brannetti,B., Costantini,A. *et al.* (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, **31**, 3625–3630.
2. Davey,N.E., Haslam,N.J., Shields,D.C. and Edwards,R.J. (2010) SLiMFinder: a web server to find novel, significantly over-represented, short protein motifs. *Nucleic Acids Res.*, **38**, W534–W549.
3. Obenauer,J.C., Cantley,L.C. and Yaffe,M.B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.
4. Rajasekaran,S., Balla,S., Gradie,P., Gryk,M.R., Kadaveru,K., Kundeti,V., Maciejewski,M.W., Mi,T., Rubino,N., Vyas,J. *et al.* (2009) Minimotif miner 2nd release: a database and web system for motif search. *Nucleic Acids Res.*, **37**, D185–D190.
5. Dinkel,H., Chica,C., Via,A., Gould,C.M., Jensen,L.J., Gibson,T.J. and Diella,F. (2011) Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Res.*, **39**, D261–D267.
6. Gould,C.M., Diella,F., Via,A., Puntervoll,P., Gemünd,C., Chabanis-Davidson,S., Michael,S., Sayadi,A., Bryne,J.C., Chica,C. *et al.* (2010) ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.*, **38**, D167–D180.
7. 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. *et al.* (2006) Minimotif Miner: a tool for investigating protein function. *Nat. Methods*, **3**, 175–177.
8. Davey,N.E., Haslam,N.J., Shields,D.C. and Edwards,R.J. (2011) SLiMSearch 2.0: biological context for short linear motifs in proteins. *Nucleic Acids Res.*, **39**, W56–W60.
9. Lam,H.Y.K., Kim,P.M., Mok,J., Tonikian,R., Sidhu,S.S., Turk,B.E., Snyder,M. and Gerstein,M.B. (2010) MOTIPS: automated motif analysis for predicting targets of modular protein domains. *BMC Bioinformatics*, **11**, 243.
10. Li,L., Zhao,B., Du,J., Zhang,K., Ling,C.X. and Li,S.S.-C. (2011) DomPep-A general method for predicting modular domain-mediated protein-protein interactions. *PLoS One*, **6**, e25528.
11. Rajasekaran,S., Merlin,J.C., Kundeti,V., Mi,T., Oommen,A., Vyas,J., Alaniz,I., Chung,K., Chowdhury,F., Deverasatty,S. *et al.* (2010) A computational tool for identifying minimotifs in protein-protein interactions and improving the accuracy of minimotif predictions. *Proteins*, **79**, 153–164.
12. Vyas,J., Nowling,R.J., Maciejewski,M.W., Rajasekaran,S., Gryk,M.R. and Schiller,M.R. (2009) A proposed syntax for Minimotif Semantics, version 1. *BMC Genomics*, **10**, 360.
13. Rajasekaran,S., Mi,T., Merlin,J.C., Oommen,A., Gradie,P. and Schiller,M.R. (2010) Partitioning of minimotifs based on function with improved prediction accuracy. *PLoS One*, **5**, e12276.
14. Stein,A., Céol,A. and Aloy,P. (2011) 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, **39**, D718–D723.
15. Vanhee,P., Reumers,J., Stricher,F., Baeten,L., Serrano,L., Schymkowitz,J. and Rousseau,F. (2010) PepX: a structural database of non-redundant protein-peptide complexes. *Nucleic Acids Res.*, **38**, D545–D551.
16. London,N., Movshovitz-Attias,D. and Schueler-Furman,O. (2010) The structural basis of peptide-protein binding strategies. *Structure*, **18**, 188–199.

17. Hornbeck,P.V., Chabra,I., Kornhauser,J.M., Skrzypek,E. and Zhang,B. (2004) PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, **4**, 1551–1561.

18. Ceol,A., Chatr-Aryamontri,A., Santonico,E., Sacco,R., Castagnoli,L. and Cesareni,G. (2007) DOMINO: a database of domain-peptide interactions. *Nucleic Acids Res.*, **35**, D557–D560.

19. Rawlings,N.D., Barrett,A.J. and Bateman,A. (2010) MEROPS: the peptidase database. *Nucleic Acids Res.*, **38**, D227–D233.

20. The Universal Protein Resource (UniProt) in 2010. (2010) *Nucleic Acids Res.*, **38**, D142–D148.

21. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

22. Diella,F., Haslam,N., Chica,C., Budd,A., Michael,S., Brown,N.P., Trave,G. and Gibson,T.J. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front. Biosci.*, **13**, 6580–6603.

23. Lieber,D.S., Elemento,O. and Tavazoie,S. (2010) Large-scale discovery and characterization of protein regulatory motifs in eukaryotes. *PLoS One*, **5**, e14444.

24. Kadaveru,K., Vyas,J. and Schiller,M.R. (2008) Viral infection and human disease- insights from minimotifs. *Front Biosci.*, **13**, 6455–6471.

25. Davey,N.E., Travé,G. and Gibson,T.J. (2011) How viruses hijack cell regulation. *Trends Biochem. Sci.*, **36**, 159–169.

26. Evans,P., Dampier,W., Ungar,L. and Tozeren,A. (2009) Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Med. Genomics*, **2**, 27.

27. Kadaveru,K., Vyas,J. and Schiller,M.R. (2008) Viral infection and human disease - insights from minimotifs. *Front. Biosci.*, **13**, 6455–6471.

28. Parthasarathi,L., Casey,F., Stein,A., Aloy,P. and Shields,D.C. (2008) Approved drug mimics of short peptide ligands from protein interaction motifs. *J. Chem. Inf. Model.*, **48**, 1943–1948.

29. Aasland,R., Abrams,C., Ampe,C., Ball,L.J., Bedford,M.T., Cesareni,G., Gimona,M., Hurley,J.H., Jarchau,T., Lehto,V.P. *et al.* (2002) Normalization of nomenclature for peptide motifs as ligands of modular protein domains. *FEBS Lett.*, **513**, 141–144.

30. Montecchi-Palazzi,L., Beavis,R., Binz,P.-A., Chalkley,R.J., Cottrell,J., Creasy,D., Shofstahl,J., Seymour,S.L. and Garavelli,J.S. (2008) The PSI-MOD community standard for representation of protein modification data. *Nat. Biotechnol.*, **26**, 864–866.

31. Vyas,J., Nowling,R.J., Meusburger,T., Sargeant,D., Kadaveru,K., Gryk,M.R., Kundeti,V., Rajasekaran,S. and Schiller,M.R. (2010) MimoSA: a system for minimotif annotation. *BMC Bioinformatics*, **11**, 328, doi: 10.1186/1471-2105-11-328.

32. Gould,C.M., Diella,F., Via,A., Puntervoll,P., Gemünd,C., Chabanis-Davidson,S., Michael,S., Sayadi,A., Bryne,J.C., Chica,C. *et al.* (2010) ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.*, **38**, D167–D180.

33. Chatr-Aryamontri,A., Ceol,A., Peluso,D., Nardozza,A., Panni,S., Sacco,F., Tinti,M., Smolyar,A., Castagnoli,L., Vidal,M. *et al.* (2009) VirusMINT: a viral protein interaction database. *Nucleic Acids Res.*, **37**, D669–D673.

34. Ceol,A., Chatr Aryamontri,A., Licata,L., Peluso,D., Briganti,L., Perfetto,L., Castagnoli,L. and Cesareni,G. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.

35. Xenarios,I., Salwínski,L., Duan,X.J., Higney,P., Kim,S.-M. and Eisenberg,D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.

36. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **38**, D5–D16.

37. Aranda,B., Achuthan,P., Alam-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.

38. The Gene Ontology Consortium. (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**, D331–D335.

39. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **38**, D5–D16.

40. Sargeant,D., Deverasatty,S., Luo,Y., Baleta,A.V., Zobrist,S., Rathnayake,V., Russo,J.C., Vyas,J., Muesing,M.A. and Schiller,M.R. (2011) HIVToolbox, an integrated web application for investigating HIV. *PLoS One*, **6**, e20122.

41. Costa,L.J., Chen,N., Lopes,A., Aguiar,R.S., Tanuri,A., Plemenitas,A. and Peterlin,B.M. (2006) Interactions between Nef and AIP1 proliferate multivesicular bodies and facilitate egress of HIV-1. *Retrovirology*, **3**, 33.

42. Arold,S., Franken,P., Strub,M.P., Hoh,F., Benichou,S., Benarous,R. and Dumas,C. (1997) The crystal structure of HIV-1 Nef protein bound to the Fyn kinase SH3 domain suggests a role for this complex in altered T cell receptor signaling. *Structure*, **5**, 1361–1372.

43. Lindwasser,O.W., Smith,W.J., Chaudhuri,R., Yang,P., Hurley,J.H. and Bonifacino,J.S. (2008) A diacidic motif in human immunodeficiency virus type 1 Nef is a novel determinant of binding to AP-2. *J. Virol.*, **82**, 1166–1174.

44. Coleman,S.H., Madrid,R., Van Damme,N., Mitchell,R.S., Bouchet,J., Servant,C., Pillai,S., Benichou,S. and Guatelli,J.C. (2006) Modulation of cellular protein trafficking by human immunodeficiency virus type 1 Nef: role of the acidic residue in the ExxxLL motif. *J. Virol.*, **80**, 1837–1849.

45. Greenberg,M., DeTulleo,L., Rapoport,I., Skowronski,J. and Kirchhausen,T. (1998) A dileucine motif in HIV-1 Nef is essential for sorting into clathrin-coated pits and for downregulation of CD4. *Curr. Biol.*, **8**, 1239–1242.

46. Coleman,S.H., Van Damme,N., Day,J.R., Noviello,C.M., Hitchin,D., Madrid,R., Benichou,S. and Guatelli,J.C. (2005) Leucine-specific, functional interactions between human immunodeficiency virus type 1 Nef and adaptor protein complexes. *J. Virol.*, **79**, 2066–2078.

47. Bandres,J.C., Luria,S. and Ratner,L. (1994) Regulation of human immunodeficiency virus Nef protein by phosphorylation. *Virology*, **201**, 157–161.

48. Bodéus,M., Marie-Cardine,A., Bougeret,C., Ramos-Morales,F. and Benarous,R. (1995) In vitro binding and phosphorylation of human immunodeficiency virus type 1 Nef protein by serine/threonine protein kinase. *J. Gen. Virol.*, **76(Pt 6)**, 1337–1344.

49. Coates,K. and Harris,M. (1995) The human immunodeficiency virus type 1 Nef protein functions as a protein kinase C substrate in vitro. *J. Gen. Virol.*, **76(Pt 4)**, 837–844.

50. Schiller,M.R. (2007) Minimotif Miner: a computation tool to investigate protein function, disease, and genetic diversity. In: Coligan,J.E., Dunn,B.M., Speicher,D.W. and Winkler,H. (eds), *Current Protocols in Protein Science*. John Wiley & Sons, Inc., New York, pp. 2.12.1–2.12.14.

51. Pagni,M., Ioannidis,V., Cerutti,L., Zahn-Zabal,M., Jongeneel,C.V., Hau,J., Martin,O., Kuznetsov,D. and Falquet,L. (2007) MyHits: improvements to an interactive resource for analyzing protein sequences. *Nucleic Acids Res.*, **35**, W433–W437.