CHAPTER 16

MINING NEW MOTIFS FROM CDNA SEQUENCE DATA

Christian Schönbach
RIKEN Genomic Sciences Center
schoen@gsc.riken.jp

Hideo Matsuda
Osaka University
matsuda@ist.osaka-u.ac.jp

General biological databases that store basic information on genome, transcriptome, and proteome are indispensable sequence discovery resources. However, they are not necessarily useful for inferring functions of proteins. To see this, we observe that SWISS-PROT[41]—a protein knowledgebase containing curated protein sequences and functional information on domains and diseases—has grown a mere 26-fold in 15 years, from 3,939 entries in 1986 to 126,147 entries in 2003. Similarly, despite the human draft genome and the mouse draft genome and transcriptome, the number of human and mouse protein sequences with some functional information has remained low—7,471 (7.4%) for man and 4,816 (4.7%) for mouse—compared to an estimated proteome of $0.5\text{–}1.0 \times 10^6$ sequences.[60]

The majority of sequences in the TrEMBL database of SWISS-PROT/TrEMBL, FANTOM,[714] and other similar databases are hypothetical proteins, or are uninformative sequences described as "similar to DKFZ ..." or "weakly similar to KIAA ...." These sequences have no informative homolog that had diverged from a common ancestor, and have matched to a non-informative homolog. Algorithms for identification of motifs are commonly used to classify these sequences, and to provide functional clues on binding sites, catalytic sites, and active sites, or structure/functions relations. For example 5,873 of 21,050 predicted FANTOM1 protein sequences contain InterPro motifs or domains. In fact, the InterPro name is the only functional description of 900 sequences.

Extrapolations from current mouse cDNA data indicate that the proteome is significantly larger than the genome. This underlines the importance of exploring protein sequences, motifs, and modules, to derive potential functions and interactions for these sequences. Strictly defined new protein sequence motifs are either conserved sequences of common ancestry, or are convergence (functional motifs) within several proteins that group together for the first time by similarity search and show statistical significance.

However, some motifs often do not reflect common ancestry. Examples include motifs occurring in paralogs, motifs occurring in mosaic protein sequences, and structural motifs based on secondary structure or active site conservation. Therefore, biological interpretation of motif findings requires additional efforts—for example literature, structural, and phylogenetic analysis.

This chapter presents a case study of mining new motifs from the FANTOM1 mouse cDNA clone collection by a linkage-clustering method, with an all-to-all sequence comparison, followed by visual inspection, sequence, topological, and literature analysis of the motif candidates. Initially the ratio of true positive to false positive new motifs turned out to be about 1:7 due to sequence redundancy and retrotransposons inserted in to the coding sequence. After filtering out those false conserved region, the ratio improves to 1:3.

ORGANIZATION.

Section 1. We briefly introduce the concept of motifs. Then we mention several broad categories of approaches to recognize and discover motifs. We also discuss some of the difficulties involved in discovering new motifs from cDNA sequences.

Section 2. We have designed a pipeline to discover new motifs in the FANTOM[714] data set. This section gives an overview of this pipeline.

Sections 3–9. The next seven sections dive into the details of the key steps of our pipeline. These key steps are: prepare a non-redundant translated data set from the FANTOM data set for motif discovery, cluster the non-redundant sequences into groups of homologous sequences, extract blocks from these groups, form a block graph from these blocks, detect homologous regions using a maximum density subgraph algorithm, eliminate those detected regions that contain known motifs, enrich the remaining blocks with additional sequences that match HMM built from these blocks. These blocks give us our candidate new motifs. Visual inspection are then carried out on these candidate motifs considering issues such as chromosomal localization, secondary structures, cellular localization, phylogenetic relations, and literature to assess which candidate motifs are true and novel. A discussion of the various categories of false positives is also given to illustrate this final manual assessment step.

Sections 10–11. Finally, we close the chapter with an extensive discussion on the true positives and their biological interpretations. We also offer a speculation on the number of new motifs that remain to be discovered.

1. What is a Motif?

Motifs are traditionally defined as conserved sequence patterns within a larger set of protein sequences that share common ancestry. Conserved motifs may be used to predict the functions of novel proteins if the relationship among the encoding genes is orthologous.[818] However, the increasing number of paralogs and mosaic proteins evolved from gene duplications and genomic rearrangement mechanisms has led to enlarged interpretations of the term motif and to the concept of modules as conserved building blocks of proteins that have a distinct function.[91,345]

A module can consist of one motif or multiple adjacent motifs. Many mammalian extracellular proteins, proteins involved in signaling cascades, and disease related positionally cloned genes, contain modules with multiple adjacent motifs or domains that are involved in different functions, such as catalytic, adaptor, effector, and/or stimulator functions. For example, C2H2 zinc fingers, leucine zippers, and POU domains are DNA-binding modules.

The variety of domains in multi-domain proteins, structural motifs or active site conservation in a short stretch of sequences seldom reflects common ancestry. Therefore, the biological interpretation of motifs—particularly new motifs—requires additional efforts, for example, literature searching and reading, structural and phylogenetic analysis.

There are so many motif discovery methodologies. References and URLs to some of these are listed in Figure 1. Which one should we use? The methods can be broadly divided into the following five categories:

(1) manual, as in PROSITE;[238] automated, as in PRODOM;[175] and mixed approaches, as in PRINTS[37] and MDS;[424]
(2) regular expressions and profiles, as in PROSITE;
(3) hierarchical clustering-based sequence similarity and derivatives with position weight matrices, as in BLOCKS[343] and PRINTS;
(4) non-linear approaches, as in the hidden Markov models (HMM) of Pfam[65], ProtFam[563], and TIGRFAMs,[313] or the neural networks of ProClass;[370] and
(5) graph-based linkage clustering, as in MDS.

Manual approaches tend to be highly specific but lack broad coverage, whereas results of completely automated methods need careful post-processing and curation to avoid misclassification of sequences. The threshold settings of algorithms in categories 1–3 determine the coverage and specificity of the motifs. Category 4 methods are dependent on the initial seed alignments and number of training set sequences. Method 5 is robust towards cut-off thresholds and data size but may cause biological false positives if conserved regions of paralogs are detected as motif members.

Each method has its strength, weakness, and the potential to miss out novel motifs or motif members. None of the above methods is “the best method” for identifying a known motif or discovering new motifs because the cut-off thresholds are either predefined or not comparable. Thus, the success of any motif analysis depends on applying and comparing multiple existing methodologies. InterPro[30] integrates various motif information and increases the confidence if the results are overlapping.

Several of the mentioned pattern discovery methods—InterProScan[927] and Pfam hmmsearch—are applied during the FANTOM sequence annotation to classify sequences, to assign gene names, and to perform indirect functional assignment by motif-gene ontology mapping. Yet existing motif discovery methods do not yield previously unknown motifs nor refined functional classifications of submotifs that indicate potential new functions of known and new genes. None of the existing methods is designed to detect biologically relevant conserved regions of distantly related proteins without including segments that look similar by chance.

A major problem with existing methods is that cut-off scores are either predetermined by users or are empirically determined by the developers of the motif detection algorithms. In biology, the notion of a conserved region is fuzzy and depends on the hierarchical context of the protein sequences. Therefore, the cut-off thresholds of conserved regions among superfamilies, familes, and subfamilies are different.

2. Motif Discovery

We have designed a pipeline to discover and characterize new motifs in the FANTOM (Functional Annotation of Mouse for RIKEN full-length cDNA clones)[714] dataset. FANTOM is part of a systematic approach to determine the full coding potential of the mouse genome and assign functional annotations to uncharacterized cDNAs. The FANTOM1 data set that is analyzed by us consists of 21,076 full-length clones (see also `http://fantom.gsc.riken.go.jp`). This pipeline is generic and is applicable to other large-scale sequence collections. The pipeline is depicted in Figure 2. It comprises an automated part for discovery of motif sequences using a maximum density subgraph method, and a semi-automated part for exploration of motif sequences. The latter consists of visual inspection, sequence analysis, topological analysis, and literature analysis, of motif candidate members. The thorough case-by-case exploration minimizes the effect of mis-annotations and error propagation.

The maximum density subgraph method (MDS)[550] is a graph-based maximum-linkage clustering method. It avoids the problems of single-linkage and hierarchical clustering, such as similarity by chance and under-clustering—too many small clusters and a few large clusters—caused by a single threshold. The MDS method applies a very low cut-off threshold to detect all related sequence pairs. Irrelevant sequence pairs are filtered out by calculating the density of blocks—the ratio of the sum of similarity scores between ungapped subsequences to the number of the subsequences—in sequence pairs. The blocks are ordered by density and blocks of the highest density cluster are selected first. The process is repeated until no high density cluster blocks are found.

Sections 3 to 9 are devoted to a more detailed exposition of the main steps of our pipeline for motif discovery.

3. Preparations for Computational Motif Detection

Implementation of our motif discovery pipeline requires that we have a UNIX or LINUX operating system and several locally installed programs and databases as listed in Figure 3. The starting point of our motif discovery is the preparation of a non-redundant set of translated sequences. We have chosen DECODER[268] to predict the open reading frame (ORF) because of its ability to correct frame shifts. In retrospect, we recommend the application of multiple programs—*e.g.*, ESTSCAN[384] and OrfFinder[882]—because the positions of the ORF can differ significantly depending on the algorithm used.

DECODER prediction yields 21,050 potential coding sequences. Since the clone set shows redundancies that can lead to false positive motifs, we cluster the putative translations using DDS[371] and ClustalW.[826] From each cluster we select the longest sequence as the representative of the cluster. As a result, we obtain 15,631 non-redundant sequences.

4. Extraction of Homologous Sequences and Clustering

The non-redundant sequences are compared against each other by BLASTP[24] using a E-value of 0.1, the BLOSUM62 matrix, and the SEG filter option[900] to remove low-complexity regions. The E-value is set low to detect all possible sequence pairs.

Each sequence pair is then analyzed using a clustering algorithm[551] based on graphy theory to extract homologous groups of sequences. Each sequence is considered as a vertex. If the similarity between any pair of sequences exceeds the user-defined E-value threshold of 0.1, an arc is drawn between the two vertices corresponding to the sequence pair. The algorithm repeatedly extracts subgraphs whose vertices are connected with at least a fraction P—a user-defined ratio, here $P = 40\%$—of the other vertices until the subgraphs cover the whole graph or no further subgraphs can satisfy the conditions. The groups of subgraphs may overlap with each other if some sequences, such as the multi-domain containing sequences, share two or more homologous regions with a different set of sequences.

The method is equivalent to complete-linkage clustering if P is set to 100%. In contrast, single-linkage clustering requires only one arc to any member in a group and P becomes virtually 0% when the number of members is large. The linkage-clustering method with all-to-all sequence comparison[551] results in 2,196 homologous groups of non-redundant sequences.

5. Detection of Homologous Regions with Maximum-Density Subgraphs

Next, we extract all subsequences of at least 20 amino acid residues length in a sequence. Then we perform ungapped pairwise alignments among all subsequence pairs to obtain blocks. A block must contain at least four subsequences. Subsequence pairs may overlap with each other if some sequences share two or more homologous regions with a different set of sequences. In this case, the overlapping pairs are merged to the same blocks step by step in descending order of their similarity scores. However, the merge is not performed if it cause the accidental join of two independent pairs (non-overlapping or partialy-overlapping pairs in the previous merge step).[550]

The alignments are scored using the BLOSUM50 score matrix. A block graph is constructed by regarding blocks as vertices. Two vertices are connected by weighted arcs if the corresponding blocks show at least the user-defined level of similarity according to their BLOSUM50 score.

Highly connected components in the block graph are detected using a maximum-density subgraph algorithm (MDS).[550] Here, "density" is a graph-theoretic term that is defined as the ratio of the sum of the similarity scores between blocks to the number of blocks. Homologous regions longer than 20 amino acids are obtained by combining overlapping blocks. The MDS algorithm yields 465 blocks that contain at least 4 sequences, and a total of 1,531 motif candidates (i.e., sequences which share similar regions over more than 20 amino acid residues). The 465 blocks occur in the 3,202 conceptually translated mouse cDNA sequences and the blocks overlap 12,251 conserved regions. Conserverd regions are defined as those regions detected by HMMER in Pfam, BLASTP in ProDom and InterPro Scan in InterPro databases.

6. Visualization of Graph-Based Clustering

The original publication[424] of the MDS method does not have room to visualize the graph-based clustering. So we take the opportunity of this chapter to illustrate the visualization. For this purpose, we conduct a small experiment with 35 known members of the Inhibitor of Growth (ING) subfamilies and two control sequences from yeast, YNJ7_YEAST and YHP0_YEAST, that share a PHD domain with ING members but are otherwise unrelated.

The BLAST scores of the sequences are computed from SWISS-PROT/TrEMBL NRDB (SWISS-PROT 40.14, 03-Apr-2002). The thresholds for drawing an arc between the sequence pairs and extracting subgraphs whose vertices are connected are set to $E < 10^{-20}$ and $P = 80\%$. The results are shown and explained in Figure 4.

7. Filtering Out Known Motifs

The detected blocks are then searched for already reported conserved regions with HMMER in Pfam (Release 5.5), BLASTP in ProDom (Release 2000.1), and InterProScan in InterPro (Release 2.0) databases. Blocks that overlap with one or more residues of known conserved regions (motifs or domains reported in Pfam, InterPro, or ProDom) are discarded. The remaining 49 blocks, containing 139 sequences and 216 conserved regions, are labeled as new motif candidates with the original discovery date.

8. Extension of New Motif Candidates

In order to expand the number of motif members, we construct new candidates from the conserved blocks using the HMMER hmmbuild program, with the –f option for a local alignment of multiple domains HMM. The HMM profile is searched with hmmsearch, with E-value < 0.1, against the SPTR-NRDB database, the 10,603 DECODER predicted FANTOM1 translations, and the 1,908 DECODER predicted translation of the EST assemblies that are not included in GenBank nor SPTR.

The hmmsearch for the 49 motif candidates increases the number of sequence from 139 to 277 sequences. The HMM expanded candidate motif sequences are aligned and displayed together with their HMM score, E-value, start position, end position, and chromosomal localization information if available to facilitate visual inspections.

9. Motif Exploration and Extended Sequence Analysis

The interpretation of 49 motif candidates is a manual process that requires biological expertise. Visual inspections of all conserved regions are carried out under consideration of species distribution, chromosomal localization, secondary structures, cellular localization, phylogenetic relations, and publications.

On the basis of the inspections, 7 of the 49 motif candidates are assessed as true and new motifs (MDS00105, MDS00113, MDS00132, and MDS00145–MDS00148). These 7 motifs are present in 28 FANTOM and 108 SPTR derived sequences. The remaining 42 of the 49 motif candidates are assessed as false positive motifs in the sense that they are either not true or not new.

The 42 false positives fall into the following 7 categories:

(1) Two motifs overlap with a published domain or motif that has not yet been incorporated into InterPro, Pfam, and ProDom releases at the time of the analysis.
(2) One motif turns out to be a low complexity region that is missed by the SEG filter.
(3) Twenty-four motifs are generated by sequence redundancy. The detection of redundant sequences after applying a clustering program shows that one should not rely on a single program. In retrospect, we should have applied two different clustering algorithms.
(4) Alternative splicing or the presence of unspliced introns cause another three false positives.
(5) Eight motifs are detected only in mouse sequences of the FANTOM clone set. Since mouse-specific motifs are unlikely to occur, the motif members may be derived from paralogs.
(6) The last category of false positives comprises repeat elements because the cDNA sequences have not been masked before predicting open reading frames. Depending on the data sources, it is recommended to check at the beginning of the pipeline for computational translated repeat elements using the RepeatMasker, which can be obtained from `http://ftp.genome.washington.edu/RM/RepeatMasker.html`.
    When we to scanned the candidate motif sequences retrospectively for repeat elements and compared the positions of the repeats within coding regions (CDS) to the motif candidate positions four (8%) of the motif candidates contained B1, B2, intracisternal A-particle LTR, and mammalian apparent LTR-retrotransposon repeat elements, respectively. Details and sequence alignments are shown at `http://motif.ics.es.osaka-u.ac.jp/MDS/falsepositives.html`.

For the 7 true positive motifs, we search PubMed with the informative gene or protein names of their member sequences to collect articles that contain biochemical, structural, or disease information. In addition, we carry out for all sequences additional sequence analyses. Secondary structure analyses of sequences are performed with locally installed ANTHEPPROT V5.0 rel.1.0.5 software,[195] DSC package,[432] and on the external PredictProtein server.[722] Functional sites—for example, phosphorylation and N-glycosylation sites—are predicted using ANTHEPROT from the PROSITE database. The locally installed PSORTII program is also used to predict the cellular localization of proteins. Chromosomal localization information is retrieved from the FANTOM map of RIKEN clones to human and mouse chromosomes as well as LocusLink[684]; it is used to judge orthology and paralogy. Multiple sequence alignments of motif member sequences are performed using locally installed ClustalW 1.8. The alignments motif sequences and motif member sequences are post-processed with the coloring software MView 1.41[112]. In some cases, the alignments are also edited by hand to improve the alignment quality. The colored alignments by amino acid properties are helpful in inferring possible functions. Phylogenetic trees are constructed from the motif member sequences by the maximum-likelihood method using MOLPHY ProtML[6] to re-assess common ancestry. The tree is obtained by the "Quick Add Search," using the Jones-Taylor-Thornton model[400] of amino acid substitution, and 300 top ranking trees are retained (options -jf -q -n 300). Bootstrap values of the tree are calculated by analyzing 1,000 replicates using the resampling of the estimated log-likelihood (RELL) method.[433]

10. Biological Interpretation of Motifs

In general the impact of bioinformatics-aided functional predictions depends on a close collaboration with biologists who put the new findings into the functional context of existing data. Since the scope of the book is on Bioinformatics, we give only an abridged version of the biological findings that have been published by Kawaji *et al.*.[424] The sequence alignments of all motifs described in this section are available at `http://motif.ics.es.osaka-u.ac.jp/MDS`.

Three of the 7 new motifs given in Section 9 have been found in hypothetical proteins. Since we lack experimental information on these proteins, we briefly summarize the predicted functions. MDS00132 members are encoded by mouse 2210414H16Rik and 330001H21Rik, and human DKFZP586A0522 loci (SPTR accessions Q9H8H3 Q9H7R3, Q9Y422, AAH08180). The human proteins belong to the generic methyltransferase family (InterPro IPR001601) and contain, adjacent to the N-terminal located MDS00132, a SAM (S-adenosyl-L-methionine) binding motif (IPR000051). Considering the 80% sequence identity and 90% similarity to DKFZP586A0522 over 146 residues (data not shown), it is likely that hypothetical proteins 2210414H16Rik and 3300001H21Rik belong to the methyltransferase family.

Motif MDS00146 comprises 21 members of hypothetical proteins or fragment derived from human, mouse, rat, fruitfly, and worm. Three members, mouse 1200017A24Rik (SPTR accession Q9DB92) and human DOCK8 (Q8NF50) previously represented by BA165F24.1 and FLJ00026, carry at their C-terminus an aminoacyl-transfer RNA synthetases class-II signature (IPR002106), indicating possible involvement in the protein synthesis.

Motif MDS00147 is located at the N-terminus of four mouse and two human hypothetical proteins. No other motifs have been detected in the sequences. This motif is an example where even motif analysis fails to add functional information. However, from the perspective of experimental biologists, the non-informative motifs are most interesting as they provide new discovery targets for protein interactions and biochemical reactions.

Motif MDS00105 is specific for the ING family, comprising three subfamilies: the ING1/ING1L subfamily,[308, 733, 929] the ING3 subfamily, and the ING1-homolog subfamily including distant homologues in *D. melanogaster*, *A. thaliana*, and *S. pombe*. The tree submotifs allow classification of sequences into the subfamilies which represent binding sites for distinct subfamily-specific protein-protein interaction candidates with HAT, HDAC, MYC,[340] and other cell cycle related proteins, while the unique regions of each subfamily member may modulate interactions. Motif MDS00105 is a candidate for protein-protein interaction experiments to define the physiological roles of the three ING subfamily members.

Motif MDS00145 is specific for mammalian 1-acyl-SN-glycerol-3-phosphate acyltransferases AGPAT3 and AGPAT4. RIKEN clones 4930526L14 and 2210417G15 represent Agpat3 (Chr 10 41.8 cM), which is the ortholog of human AGPAT3 on Chr 21q22.3. The FANTOM1 mapping of RIKEN clones 4930526L14 and 2210417G15 to Chr 16 69.90–71.20 appears to be caused by an Agpat3 related sequence on Chr 16. Agpat4 (clone 1500003P24) has been mapped to mouse Chr 17, 7.3–8.2 cM and a syntenic region on human Chr 6 that contains AGPAT4 and is close to MAP3K3. AGPAT4 is a paralog of AGPAT1 (Chr 6p21) located in major histocompatibility complex class III region. According to PSORT predictions,[597] AGPAT3 and AGPAT4 are endoplasmic reticulum (ER) membrane proteins. The latter is in concordance with previous findings for human AGPAT1.[13] The Pfam-defined acyltransferase domain is located between second and third transmembrane helix and shared by all AGPAT members. The divergence of AGPATs in the ER-sided region around MDS00145 motif of APGAT3 and AGPAT4 suggests a regulatory function of MDS00145 for transacylation specificity or activity.

Motif MDS00148 spans a 35–37 amino acids long extracellular oriented loop region between two transmembrane domains that is conserved among members of the mammalian solute carrier family 21 (organic anion transporters), organic anion transporter polypeptide-related (OATPRP) and related *Drosophila* and *C. elegans* organic anion transporters.[838] MDS00148 represents a novel module with some structural similarities to the kazal-type domain. MDS00148 might have evolved from a kazal-like protease inhibitor domain but acquired different functionality related to substrate binding in the Na^+-independent transport of organic anions, conjugated and unconjugated bile acids when transferred into an ancestral transmembrane SLC21 family protein.

Motif MDS00113 includes 20 members with conserved sequences of 43 amino acids length that carry either a leucine zipper signature characteristic for Fos related antigen 1 (FRA1) or a leucine zipper-like motif (16 members). We analyze 13 representative members in detail. Ten out of the 13 contain a leucine heptad repeat in their sequences. Since the leucine repeat could have occurred by chance, it would be risky to infer from it the leucine zipper function of DNA binding. We therefore re-analyze the sequences for features of known and characterized leucine zippers occurring in transcription factors:

(1) alpha helical coiled-coil region[639] with mostly 3,4-hydrophobic repeat of apolar amino acids at positions a and d of the helix,
(2) overlap of the coiled-coil region with the leucine heptad repeat,
(3) a coiled-coil trigger sequences[257] or the 13-residue trigger motif,[407,905]
(4) a basic DNA binding region preceding the heptad repeat and
(5) a nuclear localization signals.[351]

When these these criteria are applied, only the FRA1 sequences qualify as basic leucine zipper with DNA binding function.[468,469] Eleven sequences bear a coiled-coil trigger sequence or trigger motif. Given the sequence conservation with FRA1, it is conceivable that an ancestral functional basic leucine zipper region was subjected to recombination and mutation events degenerated the basic leucine zipper. We therefore suggest that the tandem coiled-coil containing proteins bind to proteins, rather than DNA, in a similar fashion as the group D basic helix-loop-helix (HLH) proteins.[36]

11. How Many New Motifs are Waiting to be Discovered?

Let us conclude this chapter with a speculation on the number of new motifs that are waiting to be discovered. To answer this question, we estimate the motif coverage and do some extrapolations.

The coverage of the 7 MDS motifs in Section 9 is 0.224% (28 sequences out of 12,511 sequences comprising 10,603 DECODER predicted FANTOM1 translations and 1,908 DECODER predicted translation of EST assemblies thatare not included in GenBank nor SPTR). If we extrapolate from the 136 hits of the 7 motifs to 707,571 sequences of the non-redundant SPTR database, excluding the 10,465 FANTOM1 sequences, the estimated number of new MDS motif containing sequences would be 927 (0.133% of 697,106 sequences). Since the number of sequences per MDS motif varies from 4—the minimum number of motif containing sequences that our method detects—to 57 (MDS00148), the estimated number of not-yet-discovered MDS motifs in the current release of SPTR should range from 16 to 231.

The low number of new motifs may reflect a constraint on the number of possible functions and interactions for a given protein in the proteome. In addition, some of the new motifs may be lineage-specific due to species-specific expansion of regulatory genes.[578]

Any way, each motif discovery strategy provides different results and views. We have presented an alternative strategy that has potential to extract many new motifs from existing data. Motifs can provide a rich data source of functional clues that support initial steps of protein-protein interaction, and regulatory and active site target selections in a drug discovery process.

However, their discovery on transcriptome or genome-scale requires careful data preparation. Otherwise, too much time is spent on filtering out false positive motifs. Before embarking on a motif discovery and exploration journey, one should keep mind that proteins function in a cellular context. One or multiple functions are the results of protein structure, which is dependent on the protein sequence, transcription, translation, post-translational modifications, and cellular localization at a given time within a complex network. Motif discovery can give at best some answers for one layer of complexity—at the level of protein sequence—that can enable us to ask new questions.

Database	URL	Motifs/ Domains	Comment
GRAPH-BASED LINKAGE CLUSTERING			
MDS	motif.ics.es.osaka-u.ac.jp/MDS	7	curated
REGULAR EXPRESSIONS			
PROSITE	www.expasy.ch/prosite	1,517	curated
PROFILES & HIDDEN MARKOV MODELS			
PFAM	pfam.wustl.edu	3,621	
ProtFam	mips.gsf.de/proj/protfam		
TIGRFAMs	www.tigr.org/TIGRFAMs	1,415	
BLOCKS	blocks.fhcrc.org/blocks	2,101	curated
PRINTS	www.bioinf.man.ac.uk/dbbrowser/PRINTS	1,650	curated
SBASE	www3.icgeb.trieste.it/~sbasesrv/main.html		
	SBASE-A (consolidated domains)	2,425	
	SBASE-B (unconsolidated domains)	739	
CONSENSUS SEQUENCES			
ProDom	prodes.toulouse.inra.fr/prodom/doc/prodom.html	108,076	fully automated
DOMO	www.infobiogen.fr/services/domo	8,877	fully automated
NEURAL NET			
PROCLASS	pir.georgetown.edu/gfserver/proclass.html	6,171	curated
INTEGRATED DATABASE			
Interpro	www.ebi.ac.uk/interpro	4,691	curated
HITS	hits.isb-sib.ch	4,547	fully automated
SMART	smart.embl-heidelberg.de	631	curated
eMOTIF	dna.stanford.edu/identify	170,294	
MetaFAM	metafam.ahc.umn.edu	2,793	fully automated

Number of motifs/domains as of 7 April 2002.

Fig. 1. Motif Databases

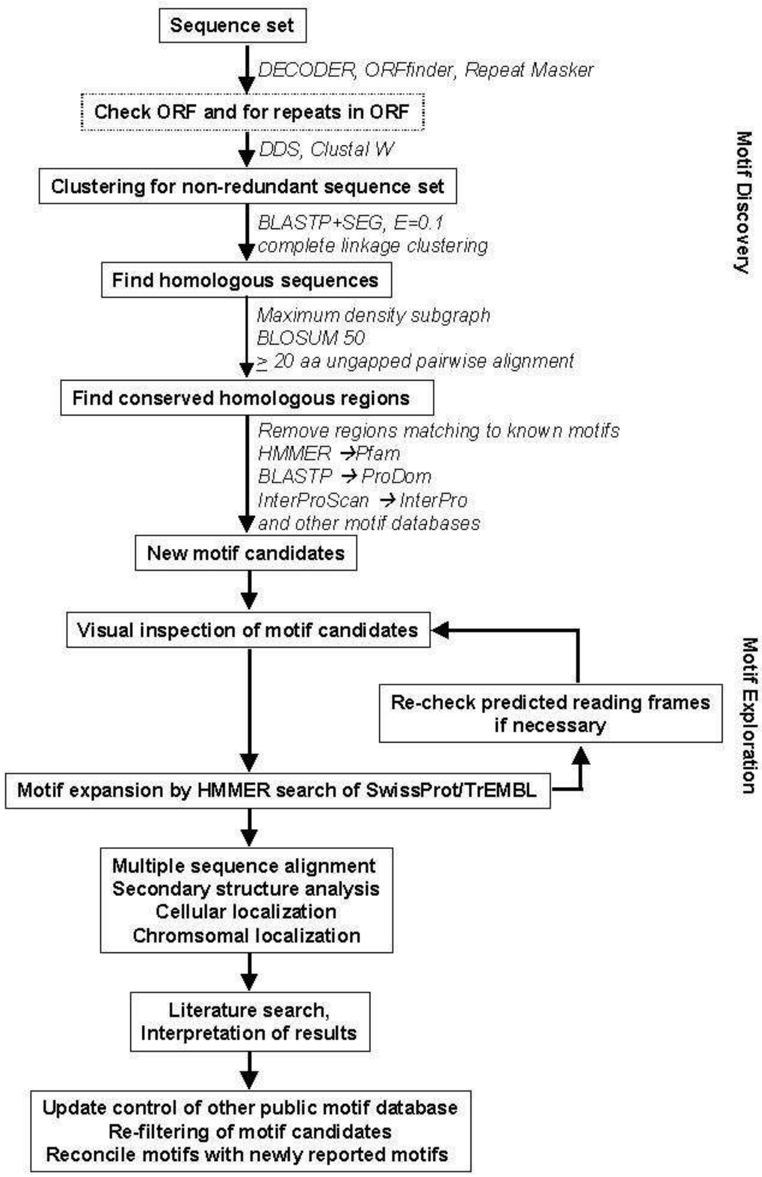

Fig. 2. A flowchart of the maximum-density subgraph motif discovery and exploration process.

Components	URL/Contact
GENERAL TOOLS	
BioPERL	`www.bioperl.org`
Apache HTTP server	`www.apache.org`
PostgresSQL	`www.postgresql.org` Results are stored in a relational database. Note that the added value is low and can increase maintenance costs compared to semi-structured flat file format.
SPECIFIC TOOLS	
DECODER	Available from `rgscerg@gsc.riken.go.jp`
DDS	Available as part of the AAT package at `ftp://ftp.tigr.org/pub/software/AAT`
BLASTP	`ftp://ncbi.nlm.nih.gov/toolbox/ncbi_tools`
InterProScan	`ftp://ftp.ebi.ac.uk/pub/databases/interpro/iprscan`
HMMER	`hmmer.wustl.edu`
MDS	Available from `matsuda@ist.osaka-u.ac.jp`
SEView	`www.isrec.isb-sib.ch/ftp-server/SEView`
ClustalW	`ftp://ftp.ebi.ac.uk/pub/software/unix/clustalw`
MView	`mathbio.nimr.mrc.ac.uk/˜nbrown/mview`
PSORTII	`psort.ims.u-tokyo.ac.jp`
DCS	`ftp://ftp.dcs.aber.ac.uk/pub/users/rdk/dsc`
PredictProtein	`maple.bioc.columbia.edu/pp`
MINIMUM DATABASE SETS	
InterPro	`ftp://ftp.ebi.ac.uk/pub/databases/interpro`
ProDoM	`ftp://ftp.toulouse.inra.fr/pub/prodom/current_release`
SPTR	`ftp://ftp.ebi.ac.uk/pub/databases/sp_tr_nrdb`

Fig. 3. Tools for motif discovery

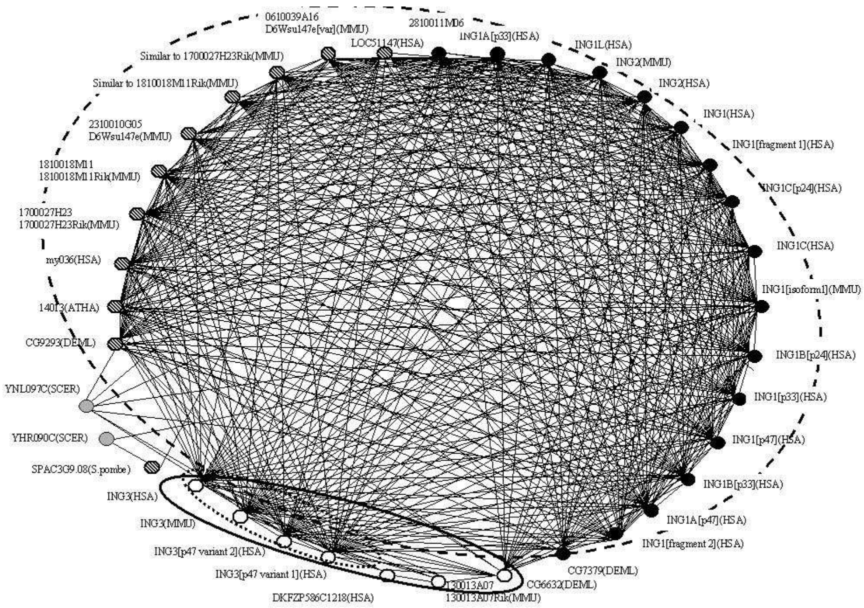

Fig. 4. The maximum density subgraph of ING subfamily members. ING members split into three clusters which are denoted by ovals. The first cluster (broken line) contains 27 members ING1/ING1L and ING1-homolog subfamilies. The second cluster (solid line) is composed of 7 ING3 subfamily members. The third cluster (dotted line) contains 4 ING3 members that belong to both cluster 1 and 2. White circles indicate ING3 members. ING1-homolog members are symbolized by hatched circles. The black circles denote ING1/ING1 members. The two yeast sequences, YNL097C (YNJ7_YEAST) and YHR090C (YHP0_YEAST), that contain a PHD domain but are otherwise unrelated to ING1 family are shown in grey circles. SPAC3G9.08 (O42871) belongs to the ING1-homolog subfamily but was not included into cluster 1 under the threshold setting of this experiment ($P = 80\%$ and E-value $< 10^{-20}$).