

CHAPTER 12

FUNCTIONAL ANNOTATION AND PROTEIN FAMILIES: FROM THEORY TO PRACTICE

Noam Kaplan

The Hebrew University of Jerusalem
kaplann@pob.huji.ac.il

Ori Sasson

The Hebrew University of Jerusalem
ori@cs.huji.ac.il

Michal Linial

The Hebrew University of Jerusalem
michall@cc.huji.ac.il

We discuss two bioinformatics tools, ProtoNet and PANDORA, that deal with different aspects of protein annotations and functional predictions. ProtoNet uses an approach of protein sequence hierarchical clustering to detect remote protein relatives. PANDORA uses a graph-based method to interpret complex protein groups through their annotations.

ORGANIZATION.

Section 1. We introduce the objectives and challenges of computationally inferring function from sequence information. We discuss the shortcomings of some commonly used tools for this purpose. ProtoNet and PANDORA are two tools designed to complement these commonly used tools and alleviate their shortcomings.

Section 2. Then we present ProtoNet. ProtoNet is centered around the concept of homology transitivity. ProtoNet clusters protein sequences into a hierarchy based on sequence similarity and homology transitivity. Thus the hierarchy parallels the evolutionary history of these protein sequences. We then illustrate—using Histone proteins—the application of ProtoNet to detect remote protein relatives.

Section 3. We describe ProTarget, which is built on top of ProtoNet. ProTarget is a useful tool for selection of protein targets that have a high probability of exhibiting to a new fold.

Section 4. Next we present PANDORA. PANDORA starts from a binary protein-annotation matrix and builds a PANDORA graph. Basically, each node of the graph represents a set of proteins that share the same combination of annotations, and the nodes are connected based on their inclusion-intersection relationships. We show how to use PANDORA to assess functional information of protein families.

Section 5. Finally, we discuss the use of PANDORA in large-scale proteomic studies. We mention PAGODA, an advanced option in PANDORA that detects outlier proteins in the sense that they share some annotations but disagree on some other annotations.

1. Introduction

One of the major goals of bioinformatics is to gain biological insights about a protein or gene from sequence information.^{68,93,172,489} The motivation for this is that sequence information is relatively easy to obtain. The reason this goal is realistic follows from the notion that two proteins with highly similar sequences are likely to be evolutionarily related and thus may share some biological properties.^{185,723} Consequently, any knowledge gained with regard to one protein, may allow the inference of biological conclusions regarding any other protein that exhibits high similarity to the former protein. Such biological information is stored in protein databases and is generally referred to as annotations.

Inferring annotations based on sequence similarity opens the door to automatic high-throughput annotation of whole genomes,^{100,489} and significantly reduces the need for tedious and labor-intensive study of every single protein sequenced. As attractive as annotation inference based on similarity is, there are inherent difficulties to be considered.⁵¹⁰ The main challenge is that protein function can often vary significantly with the change of only a few amino acids. For example, changing a few amino acids at an enzyme's active site may drastically alter its function. On the flip side, there are also instances of proteins that share the same function despite having non-significant sequence similarity. But perhaps the most acute difficulty encountered when attempting to transfer functional information among proteins stems from the multi-domains nature of proteins. It is widely accepted that the function of a protein is defined by the composition and organization of its domains rather than from a local significant similarity.^{401,583} Consequently, inference of biological characteristics from one protein to others requires not only high global similarity but also validated biological knowledge about at least one of the proteins. In other words, the protein database that is used must be rich enough to contain a highly similar and well-annotated sequence for every new protein sequence that we wish to learn about.

Naturally, the key drivers that determine the success of annotation inference are the sequence comparison methods used and their sensitivity. The sensitivity of

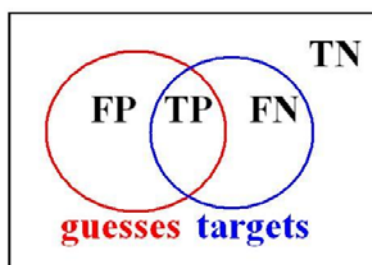


Fig. 1. The 4 categories of guesses vs. targets.

a sequence comparison method determines its ability to detect proteins that belong to the same family and are thus homologous, even when their sequences exhibit low similarity.^{23, 344, 487}

Before dealing with the practice of a bioinformatician in extracting functional information from a set of sequences at hand, let us review the basic notion of success in functional inference. Success can be easily translated to the balance between “precision” and “sensitivity” measures. Say we are looking at a certain set of specimens—*e.g.*, proteins—and some of them have a certain attribute. The set sharing this attribute shall be called the “target” set, and we choose using some prediction method a set of “guesses”—*i.e.*, a set consisting of the best guesses for specimens having the attribute. For example, we may look at a set of proteins and try to predict which proteins are receptors. In this case, the set of receptors that we predict are the guesses and the actual set of receptors are the targets. Given a guess, the specimen space can be divided into 4 categories depicted in Figure 1:

- (1) True Positives (TP)—targets that are correctly predicted.
- (2) True Negatives (TN)—non-targets that are correctly predicted.
- (3) False Positives (FP)—non-targets that are predicted to be targets.
- (4) False Negatives (FN)—targets that are predicted to be non-targets.

Sensitivity is defined as $TP/(TP + FN)$, or the percentage of targets that we guess successfully. Precision is defined as $TP/(TP + FP)$, or the percentage of our guesses that are targets. These two measures are closely linked, and in fact they trade off one for another. Often sensitivity is replaced by the term “coverage” and precision by “purity”. At one end of the spectrum are the extreme cases of having a guess set which includes all specimens. Then the sensitivity is 1.0. However, such a guess set usually also contains a high number of false positives, and thus the precision is usually low. At the other end are cases of a very narrow guess set,

say with only one element and that element is a correct guess. That provides a high precision of 1.0. However, assuming that there is more than a single element having the attribute in question, such a guess set implies many false negatives and thus low sensitivity.

Now that we have defined success, we can move to mention the set of tools that provide the “guess set” for any of the queries. The most widely used method for sequence-based functional annotation is a local alignment using BLAST.²³ While this method is shown to be powerful in terms of precision, BLAST often suffers from a low degree of sensitivity. It is thus unable to detect distant evolutionary relatives. PSI-BLAST²⁴ is a variation on BLAST that has gained popularity. PSI-BLAST increases the sensitivity dramatically, but often at the cost of low precision because of “drifting.” Several other sequence-based functional prediction methods have been developed in an attempt to increase the sensitivity of searches without significant loss of precision.⁶⁴⁷ Biological sequences such as proteins are not uniform in view of their evolutionary history, length, or information content. As such, no single method can be used optimally for all proteins. Instead, to retrieve maximal information on a query, it is advisable to apply alternative tools that provide one or more functional prediction. The ultimate validation for any prediction obtained by computational tools is always in the laboratory.

In this chapter we discuss two bioinformatics tools, ProtoNet and PANDORA, that deal with different aspects of protein annotations and functional predictions. ProtoNet uses an approach of protein sequence hierarchical clustering in order to reconstruct an evolutionary tree of all proteins, thereby enabling highly sensitive detection of remote protein relatives.⁷⁴⁶ PANDORA uses a graph-based method in order to provide means of interpreting complex protein groups through their annotations.⁴¹¹

2. ProtoNet — Tracing Protein Families

2.1. The Concept

One of the difficulties in designing a classification method using the information derived from simple pairwise alignment between proteins is incorporating biological knowledge into the method while keeping it fully automatic and unbiased by manual editing. ProtoNet⁷⁴⁶ uses biological reasoning based on the notion of homology transitivity in order to deal with this problem. The definition of homology is simple. Two proteins are considered homologous if and only if they have evolved from a common ancestor protein. An interesting aspect of homology is that it may be considered to be transitive. The reason is that if proteins *A* and *B* are homologous and proteins *B* and *C* are homologous, then proteins *A* and *C* are

homologous by definition. Homology transitivity is a powerful tool for overcoming the difficulty of noticing common evolutionary roots of proteins that do not exhibit high sequence similarity. When comparing two proteins that—by simple pairwise alignment—have a low degree of similarity, there might be a third protein that has a higher degree of similarity to both of them. This third protein can be used to establish a biological connection between these proteins and is often referred to as intermediate sequence.⁶⁴⁷

The ProtoNet system is designed around this concept. ProtoNet takes the protein universe that currently contains over one million sequences as an input. Utilizing an all-against-all comparison all using BLAST, ProtoNet builds a tree-like hierarchical organization of all proteins. This method enables a highly sensitive detection of distant relatives and is assumed to capture the tale of protein family evolutionary history.⁷⁴⁵

2.2. The Method and Principle

ProtoNet uses a clustering technique called hierarchical agglomerative clustering. It comes down to a pretty simple idea, which is to iteratively merge together clusters which exhibit the most similarity. In our case, we start off with each protein being a cluster of its own—*i.e.*, “singleton”—and start a sequence of mergers based on similarities. For clusters that are non singletons, we use the average similarity between clusters as a measure of cluster similarity.

The use of hierarchical agglomerative clustering builds upon homology transitivity, since similar proteins are drawn together into the same cluster, even if they do not exhibit direct similarity. For example, if *A* and *B* show great similarity, and so do *B* and *C*, then all three sequences are most likely end up in the same cluster, thus putting *A* and *C* together. Such an association becomes tricky in the case of multi-domain proteins, which are quite common. In such a case, we might be drawing the wrong conclusion from the similarity information. Any way, with respect to ProtoNet, the process takes place very slowly, one merger at a time. This reduces the risk of less desirable mergers that put together a pair of proteins *A* and *C* with nothing in common other than each sharing a different domain with a protein *B*. Therefore it is more common to see a cluster having proteins with a single domain at the vicinity of other clusters that include such a domain albeit jointly with other non-related domain composition.

Due to the careful averaging rules along the merger process, the final tree is of gigantic proportions, as typically the number of merger steps measures in hundreds of thousands. This creates a problem of presentation of the most informative sub-tree of a homologous family. This difficulty is resolved by “condensing” the

original tree. This condensing of the tree is done based on a criterion called “lifetime.” Intuitively, the lifetime measures the duration on time for which the cluster exists in the larger clustering process. The higher the lifetime, the more “stable” the cluster is in the dynamic process of mergers. Using this criteria for condensation allows us to ignore insignificant clusters, which are very quickly merged into other others. In other words, a set of merges are considered *en-bloc* as a single larger merger. This reduces the size of the tree, and makes it possible to navigate.

2.3. In Practice

In order to demonstrate the way ProtoNet can be used for browsing through the protein space, while keeping track of annotations, we describe a specific example. We look at Histone proteins that have been studied extensively. Histones facilitate the condensation of DNA by wrapping it around them to form a nucleosome. There are four conserved families of histone proteins that participate in this process—H2A, H2B, H3, and H4. Nucleosomes are essential for the compaction of the DNA in all eukaryotes. Histone H1 links the nucleosomes into a high order chromatin structure. Depending on the family, homology between members within each family can range from almost perfect (H4) to 50% (H1) sequence identity. A sixth histone, the poorly characterized H5, plays a role as a linker similar to H1 histone, and as such it is not a genuine part of the nucleosome structure. We start our browsing of the protein space with the protein marked in the SWISS-PROT database⁸⁶ as H1_ONCMY, which is a classical H1 histone protein.

As a biologist, you may find yourself with a sequence of an open reading frame (ORF) or a set of sequences for which you aim to retrieve maximal biological knowledge. The first step in studying a given protein sequence is to perform a sequence search using one of the various search engines. The most commonly used technique is BLAST²³ or its PSI-BLAST²⁴ variant. The result of such a search procedure is a hit list in which the top of the list indicates proteins with high score and associated with a statistical significance of that score in the appropriate database that is searched. The significance is measured by the expectation-value (E-value) and in case the hit list did not provide you any result with an E-value better than a predetermined threshold, let's say $E = 0.01$, you may find yourself going down a dead-end street.

ProtoNet can be used as the search engine for a specific protein in two distinct ways. The trivial case is when your protein is already included in the database of ProtoNet—*i.e.*, it is found in either SWISS-PROT or TrEMBL databases. In this case, the entire information gained by the construction of ProtoNet is available. In a more general setting, your protein is not part of the database, and in that case a

navigation tools

► **Get protein card:**
Returns information about a protein including its sequence, signatures, domains and taxonomy.

Search by Swissprot:
Enter Swissprot ID (e.g. AACT_HUMAN)
or accession number (e.g. P24595):
P06350 search

Search by Protein name:
Enter Protein name: (e.g. Colicin B)
search

Search by keyword:
Step 1: Enter keyword name, accession number or ID
(SwissProt, ENZYME, NCBI Taxonomy, INTERPRO, SCOP, GO)
search

- Search for all possible keywords in DB containing this word, or search by ID/accession numbers if exists

Fig. 2. Protein Search Page.

local BLAST search is activated, followed by a computational procedure that emulates the presence of your protein in the database. In both cases, a connection of your protein to an already preformed cluster is expected. Note that in the case that your protein is remote from any other proteins, it may be reported as a singleton with no connection—this is quite a rare occurrence.

Searching for a specific protein by name or keyword is easy using ProtoNet. Once the ProtoNet main page at `www.protonet.cs.huji.ac.il` loads, we choose “Get Protein Card” from “Navigation Tools” in the main menu. This brings up the window shown in Figure 2. We type in the protein SWISS-PROT accession number (P06350) or the protein name. Alternatively, keyword search is possible. For our specific example, a search of the keywords “histone”, “nucleosome”, or “chromatin” leads you to a full list of proteins that are already annotated by these functionality terms. One of the proteins in this list is H1_ONCMY.

Clicking the search button brings up the data for this specific protein, as shown in Figure 3. The data includes among other things the protein name, sequence, accession number, length—measured in amino-acids—and any PDB⁸⁸⁰ solved structures associated with it. An important control is the button allowing you to go to the cluster corresponding to this protein. This would be the lowest (significant) cluster in the hierarchy containing this protein.

Further down in the page additional information is provided as shown in

Protein P-28997

ProtoNet ID	P-28997
System	SwissProt 40.28
Swissprot ID	H1_ONCMY
Accession number	P06350
Protein name	Histone H1 [Contains: Oncorhyncin II]
Length in amino acids	206
pI	10.97
Molecular weight:	20672Da
PDB	

Go to cluster of protein "P-28997":

Sequence of protein P-28997:

```

1  AEVAPAPAAAAPAKAPKKKAAAKPKKAGPS 30
   VGELIVKAVSASKERSGVSLAALKKSLAAG
61  GYDVEKNNRSRVKIAVKSLVTKGTLVQTKGT 90
   GASGSFKLNNKKAVEAKKPAKKAAPKAKKV
121 AAKKPAAAKKPKKVAAKKAVAAKKSPPKAK 150
   KPATPKKAAKSPKKVKKPAAAAKKAAKSPK
181 KATKAAKPKAAKPKAAKAKKAAPKKK

```

[Get motifs and domains of protein](#)

Fig. 3. Protein page for H1-ONCMY.

Figure 4. Relevant annotations are shown, based on SWISS-PROT⁸⁶ keywords, InterPro²⁹ classes, and GO²⁸⁵ annotations. The taxonomy of the organism in which the protein is found is also shown in a concise form.

It is worth noting that direct links to relevant databases and primary sources is given, as well as a detailed graphical representation of the domain information within a sequence as combined by InterPro. The actual sequence on the protein that is defined as histone by several of the domain-based classification tools is

Keywords	
Swissprot	Acetylation, Chromosomal protein, Nuclear protein, Multigene family, DNA-binding
InterPro accession number	IPR005818, IPR005819
GO	<p>GO cellular component: Cell, Cellular_component, Chromatin, Chromosome, Intracellular, Nucleosome, Nucleus</p> <p>GO molecular function: DNA binding, Ligand binding or carrier, Molecular_function, Nucleic acid binding</p> <p>GO biological process: DNA metabolism, DNA packaging, Biological_process, Cell growth and/or maintenance, Cell organization and biogenesis, Chromatin assembly/disassembly, Chromosome organization and biogenesis (sensu Eukarya), Establishment and/or maintenance of chromatin architecture, Metabolism, Nuclear organization and biogenesis, Nucleobase, nucleoside, nucleotide and nucleic acid metabolism, Nucleosome assembly</p>
NCBI Taxonomy	
<pre> SUPERKINGDOM - eukaryota _ KINGDOM - metazoa _ PHYLUM - chordata _ SUBPHYLUM - craniata _ SUPERCLASS - gnathostomata _ CLASS - actinopterygii _ ORDER - salmoniformes _ SUBORDER - salmonoidei _ FAMILY - salmonidae _ GENUS - oncorhynchus _ SPECIES - oncorhynchus mykiss </pre>	

Fig. 4. Keyword information for protein H1.ONCMY.

illustrated in “domain and motif” presentation window.

From the protein page we can go to a “cluster page” using the button shown in

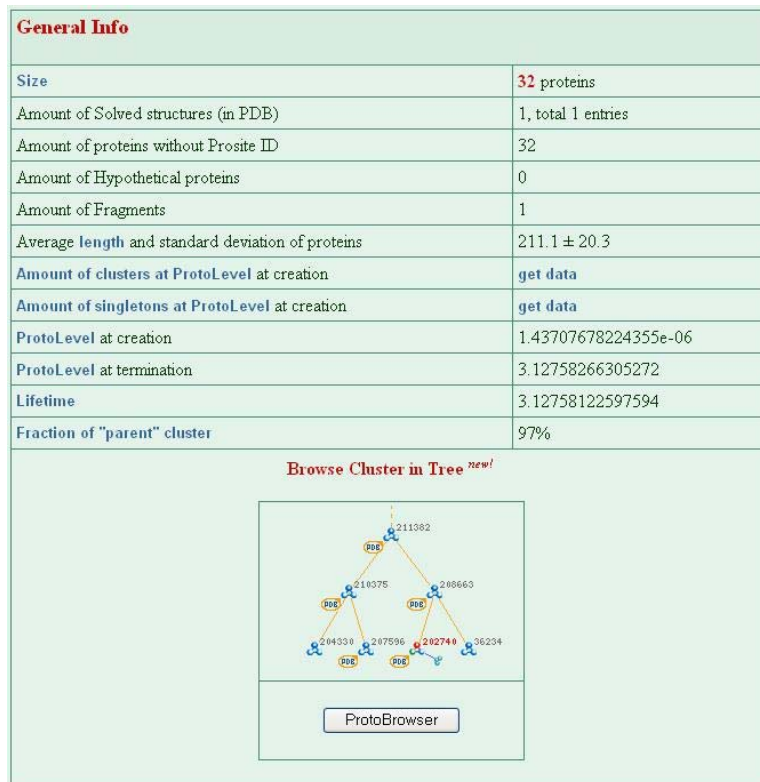


Fig. 5. Cluster page for H1_ONCMY.

Figure 3. The cluster page corresponds to a grouping—using the ProtoNet algorithms explained earlier—of proteins together into a set of proteins based on their similarity. Clicking this button brings up a cluster description card as shown in Figure 5. The most important information on a cluster is its size and composition. In our case the cluster is numbered 202740 and consists of 32 proteins.

Other information available in the cluster page relates to the number of solved structures and how many PDB entries relate to them, the number of hypothetical proteins, and the number of fragments. Additional statistics relating to the ProtoNet clustering process is given.

Below this information we have a localized view of the clustering tree. This allows us to go up and down the tree, browsing for the correct level of granularity that we are interested in. The reason we might need to do that relates directly to the



Keyword description	NumIN	Keyword Deviation from Expectation	Keyword frequency among proteins (%)		NumOUT	General keyword frequency (%)
			in this cluster	in children and appearances in merged proteins		
■ 2 keywords of InterPro All						
Histone H1/H5 (IPR005818) /all clusters /	32	200.19	100%		59	0.07%
Histone H5 (IPR005819) /all clusters /	29	197.23	90.62%		48	0.06%

Fig. 6. InterPro keyword breakdown for H1LONCMY protein.

issue of sensitivity and precision discussed previously. Due to the different rates of evolutionary diversion and due to the characteristics of specific applications, there is no single “resolution”—or in terms of BLAST search, there is no single E-value threshold—that can be used universally. In the ProtoNet realm, relaxing your search threshold is equivalent to going up the tree and restricting it means going down the tree. The nice thing with this tree is that going one step up or one step down does not involve coming up with arbitrary thresholds—*e.g.*, going from E-value of 0.01 to 0.001 is quite arbitrary, and might not change the results in some cases, while changing them dramatically in other cases.

In our specific example, we have 32 proteins in the cluster, and we would like to know if this is the appropriate resolution for looking at Histone proteins. ProtoNet provides us a way to quickly break down the protein to groups based on keywords—as before, SWISS-PROT, InterPro, and GO keywords among others are supported. Looking at the InterPro keywords for this cluster, we get the result shown in the Figure 6.

What we see here is that the protein matches two families in InterPro—where one is actually a subfamily of the other—and indeed all cluster members are classified as H1/H5. However, we can notice that 59 proteins in this family are outside the cluster. That is, the sensitivity is rather low with a very high precision. This brings us to the conclusion we are better off go one step up the tree. We reach cluster 211832 with 87 proteins. This cluster captures the vast majority of protein annotated as H1/H5 by InterPro (87 out of 91), and all proteins in the cluster share this annotation. A statistical estimation for the deviation from expectation is provided at each of the steps. This measure is based on a distribution of the keywords and accounts for keyword abundance.







Keyword description	NumIN	Keyword Deviation from Expectation	Keyword frequency among proteins (%)		NumOUT	General keyword frequency (%)
			in this cluster	in children and appearances in merged proteins		
13 keywords of InterPro All						
Histone H1/H5 (IPR005818) / all clusters /	90	243.54	52.63%		1	0.07%
Histone H5 (IPR005819) / all clusters /	73	214.72	42.69%		4	0.06%
Histone H3 (IPR000164) / all clusters /	51	182.53	29.82%		1	0.04%
Linker histone, N-terminal (IPR003216) / all clusters /	39	161.17	22.8%		0	0.03%
High mobility group proteins HMG-I and HMG-Y (IPR000116) / all clusters /	8	72.98	4.67%		0	0%
Histone-fold/TFIID-TAF/NF-Y domain (IPR004822) / all clusters /	48	71.04	28.07%		252	0.26%

Fig. 7. Keyword breakdown for cluster 223741.

Going further up in the tree does not improve the sensitivity level because we do not stumble upon a cluster with 91 proteins covering this exact family. The next significant merge brings us to cluster 223741, with 171 proteins. Part of the keyword breakdown for this cluster is shown in Figure 7. Interestingly, we can see that 90 out of the 91 H1/H5 proteins are detected here, and we also have 51 (out of 52) H3 proteins, and 39 (out of 39) Linker histone N-terminal proteins. Other proteins that are included share the properties of being DNA binding proteins that are instrumental in gene expression regulation. In addition, about 20% of the proteins marked “histone-fold TFIID-TAF-NF-Y domain” are included in the cluster. This cluster can be thus considered having a high level of functional abstraction.

To summarize, depending on the level of detail we are interested in, we can look at cluster 211832 if we are interested in H1/H5 proteins only, or we can go to cluster 223741 for a higher level view of a wider family. Similarly we can go

classify your protein

Paste your protein sequence:

```
AEVAPAPAAAAPAKAPKKKAAAAPKKAGPSV GELIVK
AVSASKERSGVSLAALKKSLAAGGYDVEKMNSRVKIA
VKSLVTRGTLVQTRGTGASGSFKLNKKAVEAKKPAKK
AAAPKAKKVAAKKPAAAAPKKVAAKKAVAAKSPKK
AKKPATPKKAAKSPKKVKKPAAAAPKAAKSPKKATKA
AKPKAAKPKAAKAKKAAPKK
```

Name of your protein (Maximum 20 chars):
my_histone

Search

[switch to the advanced mode](#)

Fig. 8. Classify-your-protein window.

higher in the tree to find larger clusters containing this cluster as well as other histone proteins.

The significance of clustering “histone-fold TFIID-TAF-NF-Y domain” with the major group of H1/H5 and H3 proteins cannot be fully appreciated at that stage. A use of PANDORA—to be described in the next section—can illuminate on the functional connectivity between the classical histones and of this group of proteins. A direct link is given from each cluster page to PANDORA.

As mentioned above, another entry point into this process is by showing a specific protein sequence. This is achieved by “Classify your protein” in the main menu of ProtoNet. In the example shown in Figure 8, we enter the sequence of the same protein studied above with the last amino-acid omitted—admittedly this is an artificial example, but it simplifies our presentation.

The output from this search is the most appropriate cluster in the hierarchy for

the given sequence, determined by a combination of BLAST search and a slight variation of the ProtoNet algorithm. Our specific search is trivial, and brings us to cluster 202740 as one might expect.

Valuable global information on ProtoNet tree can also be retrieved from the “horizontal view” option. The properties of all clusters that are created at a certain level of the hierarchy are summarized in addition to the statistical information regarding the compactness of clusters and the properties of neighboring clusters in that specific level. More advanced queries such as to find the cluster in which two proteins—for example, histone H1 and histone H2A—are first merged may be very useful in looking for remote connectivity of proteins.

Navigating ProtoNet in its full capacity is aided by the detailed “Site Map” tour while additional options are only available by activating the “Advance mode”.

3. ProtoNet-Based Tools for Structural Genomics

An exciting potential application of ProtoNet is ProTarget at www.protarget.cs.huji.ac.il, a valuable tool for structural genomics projects. Structure prediction is a process of guessing the three-dimensional structure of a protein given its sequence information. We conjecture that proteins in the same cluster are more likely to possess a similar structure. Validation of this conjecture leads to the development of a tool for Structural Genomics target selection.⁶⁷⁵ Rational navigation in the ProtoNet tree allows the user to select protein targets that have a high probability to belong to a new superfamily or a new fold. The selection of targets can be done iteratively and changed dynamically by the user. Consequently, recently solved structures can be marked and ProTarget algorithm can then provide a new updated list of the best candidates for structural determination.⁵¹¹

Automatic detection of protein families is a challenging and unsolved problem. ProtoNet offers some automation but does not go all the way in determining what a family is—thus the need to go up and down the tree. One way to enhance ProtoNet toward automatic family detection is to study quantitative aspects of clusters and their annotations. A first step in this direction is done with PANDORA described in the following section.

4. PANDORA — Integration of Annotations

4.1. *The Concept*

A multitude of tools have been developed to assign annotations to proteins based on sequence properties, and these are constantly being improved. Such automatic methods are important for large-scale proteomic and genomic research, since they

reduce bottlenecks associated with obtaining specific biological knowledge that is needed for each protein and each gene.

Large-scale proteomic and genomic research is often characterized by dealing with large sets of proteins or genes. Experimental methods such as DNA microarrays, 2D electrophoresis, and mass spectrometry provide means of dealing with complex biological processes and diseases by simultaneous inspection of hundreds of genes or proteins simultaneously. Computational proteomic family research can often deal with families of hundreds of proteins spanning several different proteomes—ProtoNet clusters, for example. Although high-throughput functional annotation eliminates the need to study each individual protein in these multi-protein sets, it shifts the bottleneck to the biological analysis of the results, a phase that requires manual inspection of the protein set, and often does not provide high-level biological insights about the proteins of the set.

PANDORA is a web-based tool to aid biologists in the interpretation of protein sets without the need of examining each individual protein. Furthermore, a major goal is to provide a global view of the protein set and of relevant biological subsets within it that are often hard to detect through normal manual inspection.

The general approach that PANDORA uses is based on annotations. In PANDORA, annotations are treated as binary properties that can be assigned to proteins. For example, for a given annotation “kinase”, a protein may either have or not have the property “kinase”, but cannot be half kinase. Each protein may have any amount of annotations assigned to it.

Annotations are often derived from different sources, each source possessing different characteristics. PANDORA uses annotation sources that cover a wide range of biological aspects: function, structure, taxonomy, cellular localization, biological pathways and more. Understanding the unique characteristics of each annotation source can greatly enhance the use of PANDORA.

4.2. The Method and Principle

The input to PANDORA is a set of P proteins. Based on the annotation sources chosen by the user, each protein may have any amount of annotations assigned to it by these sources. As mentioned, each annotation is treated as a binary property. Let the total amount of annotations used by this set of proteins be designated as K . Now, this data can be represented in a binary matrix, with P columns and K rows. Each cell in the matrix is binary, and can be occupied by a 1 or a 0, designating whether the protein represented by that column has the annotation represented by that row, as depicted in Part I of Figure 9.

Each column in the matrix represents all the annotations given to a single

- (I) A binary protein-annotation matrix—indicating which protein possesses which property—underlies every PANDORA graph.

	Protein A	Protein B	Protein C	Protein D	Protein E
Kinase	1	1	1	0	0
Membrane	1	0	0	0	0
Transcription	0	0	1	1	0
Nuclear	0	0	1	1	0
Tyrosine Kinase	1	1	0	0	0
Viral Envelope	0	0	0	0	1

- (II) A PANDORA graph derived from the protein-annotation matrix above. Each node of such a graph contains proteins that share a unique combination of annotations. An edge is then drawn between two nodes if one node is a superset of the other. The edge is directed and is implicitly indicated by placing the first node above the second node.

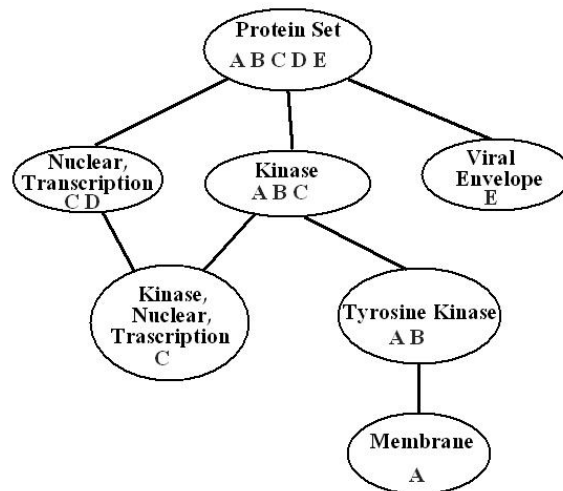


Fig. 9. A PANDORA graph and its underlying protein-annotation matrix.

protein. Looking at the columns provide us with the “conventional” view of the annotations: For each protein, we look at the list of its annotations. Looking at the

rows provide us with a more global view of the data. Each row represents a set of proteins that share an annotation that is represented by that row. These protein sets are the building blocks of PANDORA's graph. Each of these sets is checked for intersection with the other sets, and a graph is constructed. The graph constructed is an intersection-inclusion directed acyclic graph (DAG). This is a hierarchical graph representing intersection and inclusion relations between sets. In our case, each node in the graph represents a set of proteins that share a unique combination of annotations. The edges of the graph represent the hierarchy: An edge between two nodes shows that the upper node is a superset of the lower node—also referred to as the “parent” node. How is this graph constructed? The basic protein sets—represented by rows in the matrix—are checked for intersection amongst them. If two sets are equal, they are merged and become one set of proteins that has both annotations. If there is some partial degree of intersection between the nodes, a new “intersection node” is created. This node represents the set of proteins that belong to both nodes, and has both annotations. Also, this node is added to the graph hierarchy as a “daughter” of both nodes.

As shown in Part II of Figure 9, the PANDORA graph shows all the protein annotations in the matrix in a graphical manner, making it easy to detect relevant biological subsets of proteins that share a unique combination of annotations.

Consider what is visible in a PANDORA graph. All possible annotation combinations that have underlying proteins are shown. Each unique combination of annotations has its own node. Theoretically a graph with K annotations may have up to 2^K nodes—the amount of all the possible combinations. Thus even if there are only 20 annotations the graph can potentially have more than a million nodes! But this worst-case scenario never happens, mainly due to the fact that annotations are seldom randomly dispersed because they hold biological “meanings”—*e.g.*, the same annotations tend to appear on the same kinds of proteins. While cases of such extreme complexity never appear, the actual graphs associated with some annotations are very complex. At times so complex that you are probably be better off to abandon it and check the proteins one by one.

PANDORA offers a method to mitigate this complexity, by using variable resolution. Resolution is a parameter that can be used to simplify the graph. Recall that the graph shows the entire data that is in the binary matrix. Changing the graph would mean losing data. However, this is not necessarily a bad thing, because it is often not useful to see all the tiny details at once. Think about how you interpret complex data: first look at the data in low-detail to see the bigger picture showing main groups and relations, and then focus on specific parts of the data that are relevant to you and view them in high detail. This is exactly the concept behind varying resolution. Resolution is defined as the number of proteins that will be

considered as an insignificant error when building the graph. Basically what this means is that if we set the resolution to 2 proteins, the graph will be simplified under the constraint that there are no errors in accuracy of more than 2 proteins. For example, one possible simplification would be if two nodes differ by only one protein. These nodes could be considered to be equal and merged. Although this is not entirely accurate, the error is relatively small—an error of 1 protein—and can be considered insignificant for most purposes. So, the higher the value of the resolution parameter is set, the graph becomes simpler but less accurate, thus providing a global view of the data. It is possible that there may be multiple ways to construct a graph with a specific resolution. In such a situation, we arbitrarily pick one of the possible ways—we believe that if a view in a certain resolution is meaningful, it is unlikely to depend much on which way is picked.

Once we have a simplification of the graph, PANDORA provides “zooming” in order to allow focusing on areas of interest. Zooming is a simple concept: you choose a node from a “low-detail” graph, and display it in a new window in higher detail. This allows you to focus on subsets that are relevant to your biological question one at a time, and study them separately without overloading the graph.

4.3. In Practice

A typical example of large protein sets whose study can gain from PANDORA is protein clusters such as the ones created by ProtoNet. In the previous section we considered the example of histone proteins by initiating a search in the ProtoNet tree starting by a H1 histone representative. For simplicity and consistency let us carry over the discussion of this example into this section. We use PANDORA to gain further functional understanding on the cluster with 171 proteins that was already discussed (cluster 223741). Inspecting Figure 7 and the summary of keyword appearances on the proteins indicate the presence of H1 and H1/H5 as well as H3 proteins but none of the keywords appear on a majority of the cluster’s proteins. Now let’s see the PANDORA graph of this cluster depicted in Figure 10:

To understand how to read PANDORA graph we should remember that the groups of proteins represented may have inclusion and intersection relations with other nodes. These relations are represented by the graph’s hierarchy: If node *A* is connected to node *B* which is beneath it, *A* is a superset of *B*. This provides a simple yet important rule to follow: Each of the proteins of a node share not only the keywords of that node, but also those of all its ancestors in the graph. The Basic Set (BS, all 171 proteins in the cluster) appears at the top of the graph. Clicking this node opens a window that lists the protein of this set as a list.

PANDORA captures also the biological “quality” or significance of the pro-

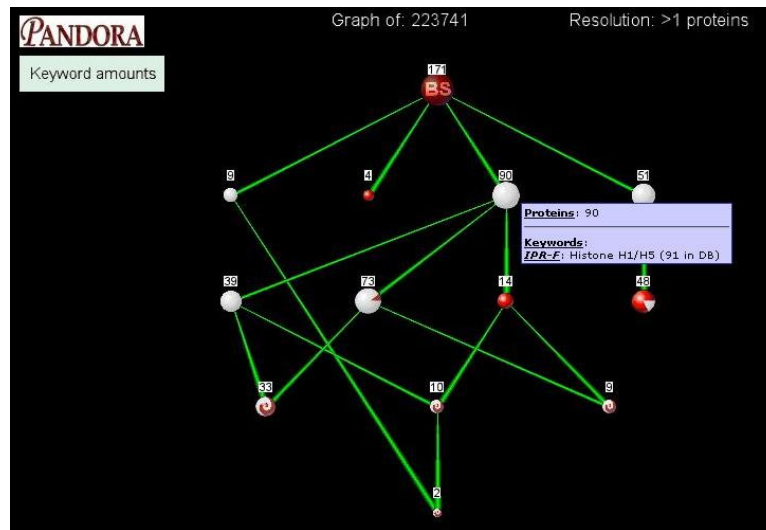


Fig. 10. PANDORA view of cluster 223741 by InterPro keywords.

tein set. To understand this concept, consider a case where your protein set of 50 proteins share a common keyword. How significant is this biologically? Well, this depends on what is the keyword that they share. Obviously, if they share a keyword that appears only 50 times in the database, this should be considered significant. Conversely, if the annotation is highly abundant in the database—*e.g.*, “enzyme” or “membranous”—it may be less interesting biologically. To implement this concept we use sensitivity measure as explained in the Section 1. Sensitivity is defined here as the fraction of proteins in the node that have a common keyword out of the total amount of proteins in the database that have that keyword. The coloring of a node represents the sensitivity for that node’s keyword: White represents the proteins in the node that have the keyword (TP); red represents proteins not in the node that have the keyword (FN). Therefore, a node that is completely white has a sensitivity of 1, because there are no other proteins in the database that have the keyword. Conversely, a node that contains a small fraction of the proteins having that keyword is coloured white only in a small portion and the rest of it is coloured red. The exact number of instances in the database of every keyword is visible via a “tool-tip” opened from the node, as shown in Figure 10. Sensitivity is more complex in the case of intersection nodes. To avoid this complexity, such nodes appear as a red-white swirl.

In addition to the graphical coloring method for quality assessment for a set

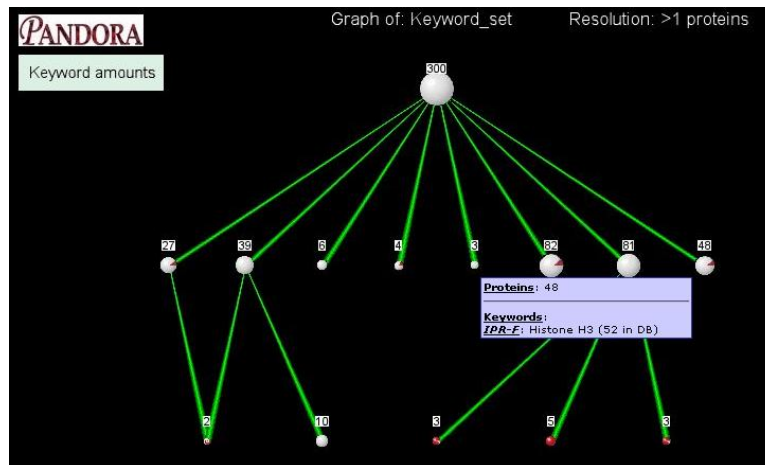


Fig. 11. PANDORA for InterPro annotation “Histone-fold/TFIID-TAF/NF-Y domain” (300 proteins).

of proteins, one can use the “show statistics” option. A table of the keywords that participate in the graph are sorted by the $\log(\text{observed}/\text{expected})$ score, where the “expected” score is defined as the frequency of the keyword in the total population of protein sequences. This provides further means of assessing significance of keywords in your set.

Now that we know how to read a PANDORA graph we can come back to the tested example of cluster 224741 from ProtoNet. The relationships by the intersection and inclusion representation now becomes evident; see Figure 10. For example, the node in the graph that marks Histone H1/H5 (90 proteins) is the parent of 3 other nodes of Histone H5 (73 proteins); Linker histone, N-terminal (39 proteins); and Proline-rich extensin (14 proteins). The first two share 33 proteins that carry both terms of Histone 5 and Linker histone, N-terminal.

We can now turn to learn more on the basic set of the proteins in cluster 223741 through additional annotation sources. For example, annotation sources covering biochemical function, cellular localization and participation in biological processes are provided by Gene Ontology database. Other annotation sources provide information about the 3D structure, taxonomy, and more.

PANDORA is also useful to visually appreciate the connectivity of a specific node, as it reflects a keyword or a unification of keywords. For example, it can be seen that the Histone 3 set of proteins (51 proteins) are rather separated from the rest and the group; furthermore, 48 proteins that are marked “Histone-fold/TFIID-

Keyword type	Keyword	Amount	Expected	Log ₂ (obs/exp)
InterPro: Domain	Histone-fold/TFIID-TAF/NF-Y domain	300	0.7892	8.5704
InterPro: Domain	Transcription factor CBF/NF-Y/archaeal histone	39	0.1026	8.5703
InterPro: Family	Histone-like transcription factor CBF/NF-Y/archaeal histone, subunit A	9	0.0237	8.5689
InterPro: Domain	TATA box binding protein associated factor (TAF)	6	0.0158	8.5689
InterPro: Domain	Transcription initiation factor TFIID	3	0.0079	8.5689
InterPro: Family	Histone H2A	81	0.2210	8.5177
InterPro: Domain	Histone H2B	82	0.2289	8.4848
InterPro: Family	Histone H3	48	0.1368	8.4548
InterPro: Domain	Histone H4	27	0.0816	8.3702
InterPro: Domain	Transcription factor TAFII-31	4	0.0132	8.2433
InterPro: Family	Histone-like transcription factor/archaeal histone/topoisomerase	10	0.0816	6.9372

Fig. 12. PANDORA statistical list for annotations associated with Histone-fold/TFIID-TAF/NF-Y domain.

TAF/NF-Y domain” are included within Figure 10. The red coloring of the node for the 48 proteins reflects the low sensitivity of this group—only 48 out of total of 300 proteins in the database. A rationale for the connection of the “Histone-fold/TFIID-TAF/NF-Y domain” to the cluster that is mostly composed of H1/H5 and H3 proteins can be sought by applying PANDORA to all proteins that are listed as “Histone-fold/TFIID-TAF/NF-Y domain” (300 proteins).

Figure 11 shows the PANDORA graph of annotation “Histone-fold/TFIID-TAF/NF-Y domain” (300 proteins). Most of the nodes are white, suggesting the high sensitivity of these nodes. The right most is the already discussed 48 proteins set that are combined with the term “H3 histone”. The rest of the nodes global information is presented by clicking on the “statistical view” option; see Figure 12.

Now we can understand better the link of “Histone-fold/TFIID-TAF/NF-Y domain” to the cluster of histones. The 300 proteins that are included under this InterPro term combine the core of the nucleosomes—H2A, 2B, H3 and H4—and other histone-like proteins that are active in transcription by binding to promoter region, such as TAF and CBP. Our initial observation that this group of proteins is not tightly linked to the H1/H5 histones has corroborated itself.

Using PANDORA you can easily test potential functional connectivity. Continuing with our histone-fold example, one can ask whether bacterial DNA-binding proteins—recall that bacteria have no nucleosomes—share functional properties with the group of “Histone-fold/TFIID-TAF/NF-Y domain” that we have just discussed. We can collect a set of proteins that we would like in a combined PANDORA graph—for example, the 124 proteins that are annotated Histone-like bacterial DNA-binding protein, as well as the 31 proteins marked Histone-like transcription factor/archaeal histone/topoisomerase.

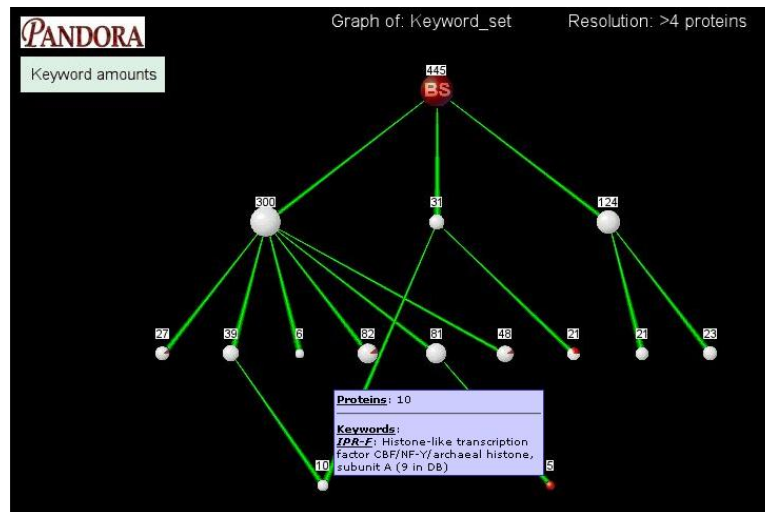


Fig. 13. PANDORA set of histone-like proteins from bacterial, archeal, and histone-fold/TFIID-TAF/NF-Y domain.

Figure 13 illustrates the result of such a query. In the Basic Set (BS is of 445 proteins), one can see that the bacterial—the rightmost note—shares no common keyword with “Histone-fold/TFIID-TAF/NF-Y domain”. However, out of the 31 archaeal proteins 10 are shared with a new intersecting node of Histone-like transcription factor CBF/NF-Y/archaeal histone, subunit A. Note that in this case as in the previous examples, we gain biological understanding without the need to get the information through inspecting individual proteins.

5. PANDORA-Based Tools for Functional Genomics

We have illustrated the power of PANDORA for assessing valuable functional information in view of protein families. However, PANDORA is a generic tool that is suitable for any large-scale proteomic study that results in a large list of genes and proteins. Indeed, the input for PANDORA is either a ProtoNet cluster or any “User Set”. The sets may be proteins that appear in our database—currently SWISS-PROT and TrEMBL—but also any of your sequence that is locally BLASTed against our database. The best matching protein above a certain threshold for each sequence is returned as input for PANDORA search.

Recall that PANDORA deals with binary properties that a protein may either have or not have. However, there are many biological properties that are naturally not binary, but quantitative. To be able to provide PANDORA capacity to quantita-

tive measures a new advanced addition has been implemented. A classic example is taken from proteomic experiments that compare expression levels of proteins at different states or following a pharmacological treatment. In this case, it would be important to ask not only whether a protein's expression has changed or not, but also by how much. For this purpose we allow users to input values for quantitative properties on proteins.

A representative graph is shown in Figure 14. Looking at the graph, you notice the colorful bars beneath each node. These are color histograms indicating the distribution of the property on the proteins of the node. This provides a visual cue that simplifies the task of identifying nodes that are "interesting" in terms of the quantitative property. For example, if our quantitative property is "change in expression level", a node that shows increased expression means that there is a group of proteins that share some biological traits—the annotations of the node—and whose expression level is increased. Naturally this can be very helpful in obtaining biological understanding of such complex results and sets. Point the mouse over the color bars to see normal histograms of the distribution. On the upper left corner you see a color legend, with tool-tips showing the value range represented by each color.

It is important to mention that you are not limited to any specific kind of quantitative property. The values that are entered with the proteins can signify anything—disease linkage, toxicity, or even protein length. Any quantitative property that is interesting to look at in the context of biological sets can be used.

An additional tool that is essential for large-scale functional genomics is called PAGODA (Probing a Group of Disagreed Annotation). The basic principle is to apply the consistency of all annotations that are associated with proteins to automatically detect nodes that are outliers and are in disagreement with the rest of the annotations. This tool is still under development and will be available as part of PANDORA's advanced options.

Acknowledgements

This study is partially supported by the Sudarsky Center for Computational Biology. We thank all former and present members of the ProtoNet team for their enthusiasm and endless effort in developing, maintaining, and advancing our web-based systems.

Supporting Servers

ProtoNet (version 3.0)	www.protonet.cs.huji.ac.il
PANDORA (version 1.1)	www.pandora.cs.huji.ac.il

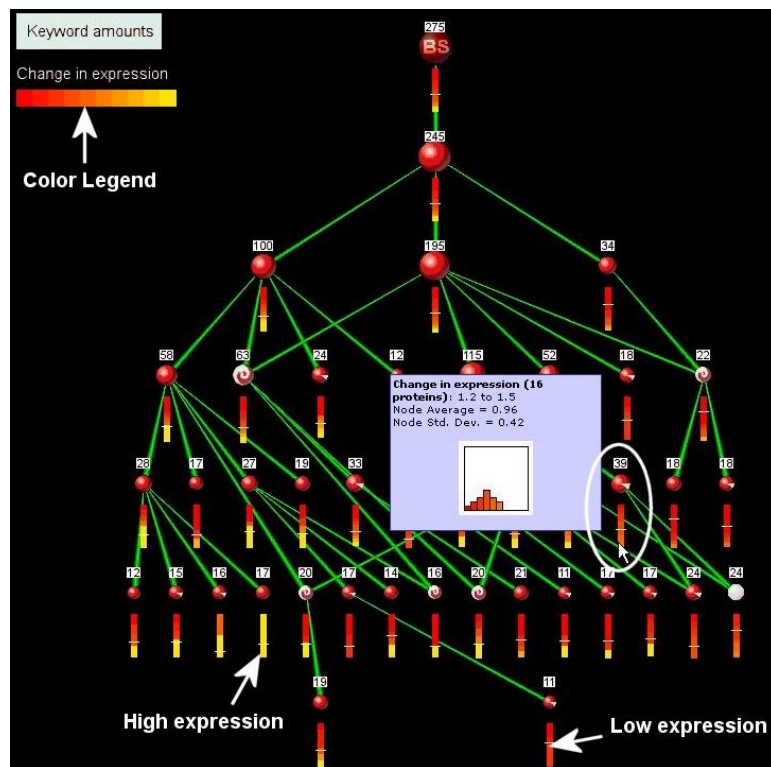


Fig. 14. PANDORA graph for the expression proteomic profile of 275 proteins analyzed by the quantitative option.

Related Web-sites and Resources

EBI GO Annotation	www.ebi.ac.uk/GOA
ENZYME	www.expasy.org/enzyme
Gene Ontology (GO)	www.ebi.ac.uk/go
InterPro	www.ebi.ac.uk/interpro
NCBI Taxonomy	www.ncbi.nlm.nih.gov
SCOP	scop.mrc-lmb.cam.ac.uk/scop
SWISS-PROT	www.expasy.org/swissprot
ProTarget	www.protarget.cs.huji.ac.il