ProtoNet 6.0: Organizing 10 million protein sequences in a compact hierarchical family tree

Journal:	*Nucleic Acids Research*
Manuscript ID:	NAR-02320-DATA-E-2011.R1
Manuscript Type:	6 Database Issue
Key Words:	Protein family, Annotation, Clustering , Protein structure, Functional inference

ProtoNet 6.0: Organizing 10 million protein sequences in a compact hierarchical family tree

Nadav Rappoport[1], Solange Karsenty[1], Amos Stein[1], Nathan Linial[1] and Michal Linial[2]

[1]School of Computer Science and Engineering, [2]Deptartment of Biological Chemistry, Institute of Life Sciences, The Sudarsky Center for Computational Biology, The Hebrew University of Jerusalem, Israel

*Corresponding author

Corresponding author details:
Michal Linial
E-Mail: michall@cc.huji.ac.il

Department of Biological Chemistry, Institute of Life Sciences
The Hebrew University
Givat Ram Campus
Jerusalem, 91904
Israel

Phone: 972-2-6585425
FAX: 972-2-6586448

Table 1
Figures 3

Running title: ProtoNet 6.0: Clustering the protein sequence space

ABSTRACT

ProtoNet 6.0 (http://www.protonet.cs.huji.ac.il) is a data structure of protein families that cover the protein sequence space. These families are generated through an unsupervised bottom-up clustering algorithm. This algorithm organizes large sets of proteins in a hierarchical tree that yields high quality protein families. The 2012 ProtoNet (Version 6.0) tree includes over 9 million proteins of which 5.5% come from UniProtKB/SwissProt and the rest from UniProtKB/TrEMBL. The hierarchical tree structure is based on an all-against-all comparison of 2.5 million representatives of UniRef50. Rigorous annotation-based quality tests prune the tree to most informative 162,088 clusters. Every high quality cluster is assigned a ProtoName that reflects the most significant annotation of its proteins. These annotations are dominated by GO terms, UniProt/Swiss-Prot keywords and InterPro. ProtoNet 6.0 operates in a default mode. When used in the advanced mode, this data structure offers the user a view of the family tree at any desired level of resolution. Systematic comparisons with previous versions of ProtoNet are carried out. They show how our view of protein families evolves, as larger parts of the sequence space become known. ProtoNet 6.0 provides numerous tools to navigate the hierarchy of clusters.

INTRODUCTION

ProtoNet (1) was launched in 2002. The goal of this system was to achieve an automatic hierarchical clustering of the protein sequences space. It covered 94,000 protein sequences from Swiss-Prot. Now, almost 10 years later, our census of proteins has grown tremendously. Thus, the UniProtKB database of protein sequences (2) includes over 17 millions proteins (UniProt, August 2011) of which 0.53 million proteins form the UniProtKB/Swiss-Prot section. While the size of UniProtKB/Swiss-Prot grew from 2002 by a factor of 5 (SwissProt release 40.0, October 2001), the TrEMBL section (TrEMBL Release 18.0) went from 550,000 to 16.5 million sequences, a 30-fold increase during the same period.

Notably, even in the curated high quality UniProtKB/Swiss-Prot section, only 25% of the proteins carry evidence at the protein or transcript levels, while 70% of the sequences are inferred from homology and about 3% remain questionable and marked as predicted or even uncertain proteins. The situation with the millions of sequences from UniProtKB/TrEMBL is far less satisfying. Only 3% carry some experimental supporting evidence and the majority of sequences (74%) are only based on prediction. With this immense growth in the number of protein sequences, it is clear that only unsupervised methods can cope with this data set. We need algorithms that can automatically trace the functional and evolutionary relatedness among protein sequences (3).

Assigning biological functions to proteins is a major obstacle and a challenging task (4,5). Despite important progress in structural genomics, enzyme classifications and

phylogenomics, the goal of automatic functional inference is far from being reached (3,6-8). Numerous motif recognition algorithms, statistical model-based and clustering methods were developed during the last two decades for the purpose of handling the growing number of sequences. These methods differ in their coverage, the level of manual curation involved and even in the basic definition of a domain family. For example, Pfam (9), SMART (10), EVEREST (11), PANTHER (12) and Gene3D (13) are based on thousands of profile hidden Markov models (profile HMMs). New sequences that pass a pre-determined threshold of similarity are assigned to the corresponding model domain family. Additional resources are based on algorithms that search for signature, regular expressions or Position Specific Scoring Matrix (PSSM) fingerprints. Representative databases that follow this paradigm include PROSITE (14), PRINTS (15), ProDom (16), BLOCKS (17). The above resources are based on sequence data.

In addition, integrative resources such as PIRSF (18), CDD (19) and InterPro (20) take a different approach to the end of attaining higher coverage of the protein space. They accomplish this by merging a variety of external sources with a focus on protein families, domains and functional sites. The classifications of SCOP (21), CATH (22) and SUPERFAMILY (23) rest on 3D-structural information. A functional perspective is offered by the ontology-based resource of Gene Ontology (GO) (24).

The available data is highly redundant, which creates a major difficulty in this area. Thus the main archive of UniProt database contains 25 million sequences (25) which represent about 17 million unique proteins. The UniRef50 with only 4 million sequence is created by grouping together proteins with >50% identical amino acids.

However, in order to study sequence homologies and the evolution of protein families, they must be viewed at a much finer level of granularity.

In order to deal with the enormous number of known protein sequences, ProtoNet 6.0 generates automatically, with no supervision a consistent classification tree. This system covers over 9 million proteins from UniProtKB. To address the expected future growth in the number of protein sequences, the system is equipped with a protocol for maintenance and updating. A system-provided confidence parameter quantifies the quality of every cluster in ProtoNet 6.0. Additional tools for analysis and visualization enhance the user's navigation options through the ProtoNet tree. These tools provide a rich biological context for the observed parts of the tree.

We describe here the newly introduced capabilities and improvements compared with the previous version (26) where one million proteins were classified (1,072,911 sequences, UniProt Release, Feb. 2005, ProtoNet Version 4.0).

PROTEIN SEQUENCES DATABASE

All database sources used in ProtoNet 6.0 has been thoroughly updated. The most critical aspect is the use of UniRef50 clusters as our basic objects. On average a UniRef50 cluster contains 4 proteins. Thus, the 2,478,328 UniRef50 proteins that are included in ProtoNet 6.0 represent over 9 million sequences. In comparison the number of protein sequences in ProtoNet 4.0 is 1,072,911.

PROTONET TREE CONSTRUCTION

The basic algorithm of ProtoNet was previously described (1,27). It starts by pre-calculating an all-against-all BLAST similarity score (28) for all protein

representatives from the UniRef50 resource (called cluster seed proteins). The similarities' E-scores were used to produce a continuous hierarchical bottom-up clustering process. At each step, the two most similar protein clusters are joined (the exact algorithm is described (29)). Importantly, BLAST E-score with an extremely relaxed threshold is considered throughout the ProtoNet construction (E-score=100). The bottom-up agglomerative clustering of the ProtoNet algorithm benefits from such relaxed E-score distances in constructing a robust family tree. A key ingredient of ProtoNet 6.0 that is essential for handling such a large number of proteins is the Constrained Memory-ProtoNet algorithm (29).

The result is a hierarchy of protein clusters at various degrees of biological granularity. This hierarchy is structured as a collection of trees that form what we call ProtoNet Tree (actually it is a ProtoNet forest). The root clusters contain all the proteins of the tree while other clusters represent subdivisions of proteins into smaller groups. The hierarchical definitions allow the user to navigate from a protein to the sub-family and the super-family levels in order to discover specific functions and evolutionary signals.

THE HIERARCHY'S QUALITY

The entire protocol to construct ProtoNet is unsupervised and therefore no annotations are included. However, measuring the correspondence between a given cluster and specific annotations that are provided by external expert systems is essential for the *supervised validation* of the automatically generated ProtoNet clusters.

We thus define the notion of a correspondence score (CS). The CS for a specific cluster and a given keyword is a measure of correlation between two. Formally, let us

fix a cluster C in the ProtoNet tree and a keyword K (from a specific source such as InterPro). Let c be the set of proteins in cluster C and let k be the set of proteins in the system annotated by keyword K. We define the CS as:

Correspondence Score (cluster C for keyword K) = $$CS(C, K) = \frac{|c \cap k|}{|c \cup k|}$$

The cluster receiving the maximal score for keyword K (called K's *best cluster*) is considered the cluster that best represents K within the ProtoNet tree. The score for a given cluster on keyword K ranges from 0 (no correspondence) to 1 (the cluster C is comprised of all the proteins with keyword K). The CS values are used as a quality measure for the ProtoNet tree. For example, we may consider the distribution of CS value over all ProtoNet clusters or over clusters of size that exceeds some cutoff threshold. In order to obtain a biologically relevant view of the hierarchy, we applied several tests that allow us to focus only on the clusters that are enriched with some coherent biological information.

The main algorithmic difference between ProtoNet 6.0 and the earlier version ProtoNet 4.0 (26) is the use of CM-ProtoNet (29). We refined the clusters' quality test by evaluating the CM-ProtoNet method over a single-linkage performance (that is implemented in ClusTr (30)). The tests were carried out on 3.2 million proteins from ProtoNet 5.1 (Table 1). In addition, we tested the impact of selecting UniRef50 as cluster seed proteins for ProtoNet. It can be seen that CM-ProtoNet outperforms the other methods that were applied to the same set of proteins. Notice that the main improvement of MC-ProtoNet comes from enhanced sensitivity. The performance of the *Single linkage* algorithm drops drastically due to a low sensitivity. We tested three choices of cluster seeds: UniRef50 representatives (the choice that we finally adopted), UniRef90 (proteins sharing >90% sequence identity) and the complete

redundant protein sequences. It is remarkable that the quality of clustering with respect to all three choices remains essentially unchanged.

The same tests with respect to a set of keywords from Pfam Clan (9) validated the high performance of the MC-ProtoNet algorithm over other clustering methods (not shown). We confirmed that the protocol that was applied to construct the ProtoNet 6.0 produces a stable tree with a collection of biologically coherent families and super-families.

SELECTING STABLE CLUSTERS

The ProtoNet tree is huge, and the immense number of its protein clusters makes it quite impractical to navigate the tree. In order to deal with this difficulty, we pruned the tree. The basic idea is that many clusters that are created along the process of generating the tree are biologically irrelevant and uninteresting. For example, a root cluster in the ProtoNet forest contains thousands of unrelated proteins.

A process of repeated pair-wise merging yields a tree of size roughly twice the number of leaves (see illustration in Figure 1A). Therefore, starting with the 2.5 million UniRef50 seed proteins we obtain 5.0 million clusters. We applied several computational procedures that are aimed at reducing this number. Our aim is to simplify the navigation in the system while maintaining the hierarchical structure and with essentially no loss in clusters' quality.

To this end we sought intrinsic parameters of ProtoNet that measure the *stability* of a cluster. One such parameter is Life Time (LT), which is the difference between the time (i.e. merging steps) in which a cluster is created and the time it is merged to a larger cluster. This number reflects the relative height of a cluster in the merging tree.

The level of the tree (called ProtoLevel, PL) is used as an internal monotonic *timer for merging*, along the clustering process (which is reflected by the index of the cluster, Figure 1A). Individual protein sequences have PL=0 and for the root of the ProtoNet tree PL=100. The idea is that *stable clusters* tend to be more relevant biologically. We thus used a tradeoff between the number of clusters that are retained and the reduction in the performance of the clusters, measured by the average of the CS for all clusters. A minimal reduction in the average CS score for the InterPro keyword annotations was attained for LT <1.0. We thus set the LT=1.0 as a default parameter (see *"Advanced Navigation"*).

Figure 1 illustrates the pruning process at different LT cutoffs (marked x, y). Evidently, fewer valid clusters (colored red) remain as LT is increased. Figure 1B shows a cluster summary at different LT cutoffs. Note that the statistical parameters of the analyzed clusters depend on the choice of LT values.

The pruned version of the ProtoNet 6.0 tree at a LT=1.0 and PL=90 has 162,088 high quality stable clusters. With these parameters the original number of 5 million clusters (including leaves) is reduced by a factor of about 30.

ANNOTATION INFERENCE

Functional inference in ProtoNet 6.0 is done by an automatic high-confidence method that infers the functional annotation of a cluster by integrating the annotations of its individual proteins. The method builds on functional annotations from multiple resources including InterPro, Gene Ontology (GO) (24), UniProt keywords (2), ENZYME (31) and more. We consider all the annotations that cover >1% of the proteins and focus on those that best fit the proteins of the cluster.

Evidently, automatic inference cannot be error-free. Thus, a predetermined specificity threshold is calculated for the keywords associated with the cluster's proteins. Such annotation is assigned as the ProtoName (Figure 2). To avoid faulty inference, we calculated ProtoName for clusters in which >20% of the proteins share the specific annotation where this annotation shows an enrichment of p-value <0.005. Recall that presenting additional names for a cluster often hints at a novel overlooked function or the presence of multi-domain proteins that exhibit multi-functionality.

Each of the ~162,000 stable clusters was assigned a ProtoName. On average, a cluster is associated with 9.7 possible names. Most names are derived from Taxonomy (33%), UniProt (19%), GO (18%), InterPro (17%) and the rest includes information from structural classifications (e.g., SCOP (21) and CATH (22)) or ENZYME-based annotations (31). A partition of the unique clusters according to their annotation types is shown (Figure 2). Notably, most annotation types contribute to some ProtoName. This suggests that the integration of knowledge from diverse annotation sources substantially improves the performance of the ProtoNet tree.

GENOMIC VIEW ON PROTEIN CLUSTERS

A huge number of organisms are represented in UniProtKB (Figure 2B). Still, a third of the protein sequences originate from a relatively small number of organisms that were completely sequenced. A substantial number of all these sequences (mostly from multi-cellular organisms) also serve as genetic model organisms. Therefore, we included a selected list of over 30 organisms on which the user can choose to focus. These organisms represent all superkingdoms.

WEBSITE PROPERTIES

Several added features in the ProtoNet 6.0 website make it easier to reach an in-depth analysis of the ProtoNet tree. We describe these new features in 'simplified mode' and in 'advanced mode' (Figure 3).

(i) **BROWSING CLUSTER NAMES**

Cluster names are now available for browsing. One can choose a keyword of interest and view clusters that are named by it. Note that a keyword of low statistical significance will be absent in ProtoName. Figure 2 shows the contribution of the major annotation resources that are included in determining the ProtoName.

(ii) HYPOTHETICAL & PUTATIVE PROTEINS

The assignment of a biological function to clusters suggests a safe scheme of assigning function to proteins with unknown function. Naively, the protein can be assigned the function of all clusters to which it belongs. This can be applied for 'hypothetical' and 'putative' proteins within the clusters. It can also be used for a new user-provided protein sequence (with the *"Classify your protein"* option). We provide a list of all the proteins that are marked as hypothetical and putative proteins in the summary table (Figure 3).

(iii) PROTONET TREE RESOLUTION

Following the pruning process described above, ProtoNet is no longer a binary tree. To cope with this non-binary condensed version, we introduced the *ProtoBrowser* page that zooms in on the tree only in the vicinity of the cluster that is being analyzed. A selected branch is shown in the context of related neighboring branches. The user hovers the mouse over a cluster to display essential information such as the cluster

size, the number of proteins according to selected species (if a 'genomic view' was activated). An example of such ProtoBrowser tree views is shown (Figures 3).

(iv) INTEGRATION OF ANNOTATION SOURCES

The functional analysis of a cluster is performed using PANDORA (32) visualization, which allows in-depth analysis of large protein sets. The system allows direct export from the cluster page to PANDORA. Using PANDORA it is possible to assess the functional relevance of the proteins in the clusters from numerous biological aspects. The annotation sources used by PANDORA were updated, and now offer ~200,000 different annotations, spanning several different biological domains.

Specifically, PANDORA extracts most of the annotations from UniProtKB. For structural annotations CATH (22), SCOP and Gene3D (13) are considered. The functional domain is covered by the 4 layers of the ENZYME classification (31) and the Gene Ontology (GO) structure with the 3 main functional branches: Cellular component, Biochemical function and Biological process (24).

The protein families are forwarded to PANDORA analysis tool that statistically analyzes a given cluster by means of the annotations that are assigned to its proteins (32). On average, each protein sequence in ProtoNet is associated with 6.6 different annotation types (11 and 10 annotations for human and mouse, respectively). PANDORA supports also each of the dozen domain and family resources of the InterPro collection.

In a typical application of PANDORA the user concentrates on any of the 200,000 annotations with the query *"Get clusters containing proteins with a given keyword"* (e.g., InterPro domain: GTPase-binding/formin homology 3). In response one receives

an integrated view of all proteins that are associated with this annotation, not only those that belong to the UniRef50 seed proteins (see below).

(v) EXPANDED PROTEINS

The ProtoNet tree is started with the representative proteins of UniRef50. The cluster view offers a list of the proteins of the cluster. Two levels of expansion are provided: the list of proteins according to the UniRef representatives and the complete UniProtKB list. On average, the passage from UniRef50 to UniRef90 and from UniRef90 to the UniProtKB full list results in a 2.5 folds and 1.8 fold expansion, respectively. Cluster A4686503 contains 487 proteins that have a mamimal CS for the keyword *Cadherins* of CATH homologous superfamily (CS=0.767). This cluster is expanded to 2349 proteins. Similarly, the expanded list of proteins can be conveniently viewed via PANDORA (see *"Integration of Annotation Sources"*). For example, 557 proteins in the ProtoNet 6.0 database are annotated *Cadherins* according to the CATH homologous superfamily, but using PANDORA, this list is expanded to a total of 2298 proteins.

(vi) PHYLOGENETIC TREE VIEWER

The user can select one or several organisms and have the branches in the ProtoNet tree that include the selected organisms highlighted. Navigation through the selection of complete proteomes is illustrated in Figure 3. It is shown for a few selected mammals (human, mouse, rat). Only branches that include proteins from the *selected* organisms become visible, though all 'faded' clusters can still be analyzed. In Figure 3, the indicated cluster (Cluster ID 4201544) contains 310 proteins. The number of proteins that is covered by the selected proteins is listed (Figure 3). At any stage the user can reset or remove or change the taxonomical based selection.

(vii) COMPARING VERSIONS

The user may select to navigate each of the main releases of ProtoNet. Maintenance of the different versions allows assessing the changes in the clusters along the constant growth in protein sequences. For example, with the same threshold of PL=90 and LT=1 there are 5,245 and 74,446 stable clusters by ProtoNet versions of SwissProt 40.28 and UniProt 8.1, respectively.

(viii) ADVANCED NAVIGATION

The advance mode provides additional control for the user on the parameters of the visualization that concern: (i) the ProtoBrowser. (ii) ProtoNet tree condensation. The user can choose to activate the ProtoBrowser at a different resolution. While the simplified mode (Figure 3, upper panel) shows two levels above and below the observed cluster (marked in red font in the tree, Figure 3), in 'advanced mode', the number of presented surrounding tree layers is a user-selected parameter. By moving up the tree one observes how the cluster grows in size and becomes more diverse.

The user can change the tree resolution by modifying the parameters of the tree condensation protocol (see *"Selecting Stable Clusters"*). Such change of parameters turns a binary tree to a non-binary tree, and some browsing options help the user in following this modification.

Other capacities of the 'advanced mode' reflect certain intrinsic properties of the ProtoNet Tree. The user can retrieve the ProtoNet clusters at a specific ProtoLevel (PL) (Figure 3, lower panel). This determines the number of clusters to be presented but it also (indirectly) allows the user to focus on a PL that is maximally enriched by proteins with unknown function. While a careful biological interpretation of the

ProtoNet 6.0 clusters is beyond the scope of this paper, we should note a significant explosion of proteins of unknown function that appears at PL above 90.

Additional queries address the connectivity of selected proteins in the tree. In particular, one can get the lowest common cluster of any two proteins. Search for the appearance of a specific protein within a cluster, search for all the clusters that are associated with a selected keyword and more.

A TEST CASE - METAGENOME TO FUNCTION

Global Ocean Sampling (GOS) sequences is a huge collection of (mostly) unidentified marine metagenome sequences that covers nearly all known prokaryotic protein families (33). We now illustrate a test case of one of hypothetical protein GOS_6351915.

Applying the ProtoNet option *'Paste your new sequence'* in basic mode with default parameters finds this sequence in cluster 4033656 (26 proteins, 5 named 'predicted protein' and additional 2 proteins named 'putative') all of which belong to InterPro entry of "Longin" and to additional keywords that specifies the relevance to SNARE-like (based on SCOP). However, upon moving up the tree to a larger cluster with 107 proteins (Cluster ID 4312270), the dominating keyword (ProtoName) is changed to InterPro IPR016444: Synaptobrevin that metazoa/fungi. The taxonomy of the merged cluster includes only metazoa and fungi (excluding green plans).

Activating the 'advanced mode' for a condensed tree (LT threshold=10) indicates that GOS_6351915 sequence belongs to a larger cluster (213 proteins, Root) where the most significant annotation (Cluster ID 4446624 and CS=0.965) is from CATH topology of Beta-Lactamase and homologous group of CATH 3.30.450.50. Analyzing

this very stable cluster via PANDORA shows that the dominating features are *membrane* and *coiled coil*. The significant p-value for other functional annotations such as v-SNARE, trafficking, synaptic vesicles, ER and golgi confirm that GOS_6351915 sequence is a genuine member of the SNARE family. We postulate with high confidence that this sequence is a Synaptobrevin-like protein that is probably derived from the unicellular species of marine-centric diatom.

MAINTENANCE AND UPDATING

ProtoNet will be updated once a year. A partition in UniProt to the sections of UniProt/Swiss-Prot and UniProt/TrEMBL will be implemented. This will allow users to focus, as needed, on each UniProt section, separately. Future ProtoNet releases will incorporate additional annotation resources from KEGG, STRING, OMIM and GO evidence codes. To provide the user with control over the confidence level, annotations evidence (e.g., experimental, inferred) will be added for each protein in our database.

ProtoNet 6.0 had also incorporated few fundamental technical improvements in the automation, database design and technologies. These improvements concern the automation for the future updates and releases.

ACKNOWLEDGEMENTS

We thank the ProtoNet team and especially Yaniv Loewenstein for his support in establishing the system's performance. We thank Eden Dror for his help in maintaining and improving the databae of ProtoNet. We thank the fellowship support by the Sudarsky Center for Computational Biology (SCCB) to NR. This study is

partially funded by the EU FRVII Prospects consortium and the Israel Science Foundation ISF 592/07.

REFERENCES

1. Sasson, O., Vaaknin, A., Fleischer, H., Portugaly, E., Bilu, Y., Linial, N. and Linial, M. (2003) ProtoNet: hierarchical classification of the protein space. Nucleic Acids Res, 31, 348-352.

2. The UniProt Consortium (2011) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res, 38, D142-148.

3. Loewenstein, Y., Raimondo, D., Redfern, O.C., Watson, J., Frishman, D., Linial, M., Orengo, C., Thornton, J. and Tramontano, A. (2009) Protein function annotation by homology-based inference. Genome Biol, 10, 207.

4. Fleischmann, W., Moller, S., Gateau, A. and Apweiler, R. (1999) A novel method for automatic functional annotation of proteins. Bioinformatics, 15, 228-233.

5. Friedberg, I. (2006) Automated protein function prediction--the genomic challenge. Brief Bioinform, 7, 225-242.

6. Brown, D.P., Krishnamurthy, N. and Sjolander, K. (2007) Automated protein subfamily identification and classification. PLoS Comput Biol, 3, e160.

7. Watson, J.D., Sanderson, S., Ezersky, A., Savchenko, A., Edwards, A., Orengo, C., Joachimiak, A., Laskowski, R.A. and Thornton, J.M. (2007) Towards fully automated structure-based function prediction in structural genomics: a case study. J Mol Biol, 367, 1511-1522.

8. Pazos, F. and Sternberg, M.J. (2004) Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci U S A, 101, 14754-14759.

9. Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K. et al. (2010) The Pfam protein families database. Nucleic Acids Res, 38, D211-222.

10. Letunic, I., Doerks, T. and Bork, P. (2009) SMART 6: recent updates and new developments. Nucleic Acids Res, 37, D229-232. http://www.ncbi.nlm.nih.gov/pubmed/18978020

11. Portugaly, E., Linial, N. and Linial, M. (2007) EVEREST: a collection of evolutionary conserved protein domains. Nucleic Acids Res, 35, D241-246.

12. Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S. and Thomas, P.D. (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res, 38, D204-210.

13. Yeats, C., Lees, J., Carter, P., Sillitoe, I. and Orengo, C. (2011) The Gene3D Web Services: a platform for identifying, annotating and comparing structural domains in protein sequences. Nucleic Acids Res, 39, W546-550.

14. Sigrist, C.J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A. and Hulo, N. (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res, 38, D161-166.

15. Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P. et al. (2003) PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res, 31, 400-402.

16. Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S. and Kahn, D. (2005) The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res, 33, D212-215.

17. Henikoff, J.G., Greene, E.A., Pietrokovski, S. and Henikoff, S. (2000) Increased coverage of protein families with the blocks database servers. Nucleic Acids Res, 28, 228-230.

18. Wu, C.H., Nikolskaya, A., Huang, H., Yeh, L.S., Natale, D.A., Vinayaka, C.R., Hu, Z.Z., Mazumder, R., Kumar, S., Kourtesis, P. et al. (2004) PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Res, 32, D112-114.

19. Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R. et al. (2010) CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res, 39, D225-229.

20. Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L. et al. (2009) InterPro: the integrative protein signature database. Nucleic Acids Res, 37, D211-215.

21. Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res, 36, D419-425.

22. Cuff, A.L., Sillitoe, I., Lewis, T., Redfern, O.C., Garratt, R., Thornton, J. and Orengo, C.A. (2009) The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res, 37, D310-314.

23. Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D. et al. (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res, 33, D247-251.

24. Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C. and Apweiler, R. (2009) The GOA database in 2009--an integrated Gene Ontology Annotation resource. Nucleic Acids Res, 37, D396-403.

25. Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R. and Wu, C.H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics, 23, 1282-1288.

26. Kaplan, N., Sasson, O., Inbar, U., Friedlich, M., Fromer, M., Fleischer, H., Portugaly, E., Linial, N. and Linial, M. (2005) ProtoNet 4.0: a hierarchical classification of one million protein sequences. Nucleic Acids Res, 33, D216-218.

27. Sasson, O., Kaplan, N. and Linial, M. (2006) Functional annotation prediction: all for one and one for all. Protein Sci, 15, 1557-1562.

28. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25, 3389-3402.

29. Loewenstein, Y., Portugaly, E., Fromer, M. and Linial, M. (2008) Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. Bioinformatics, 24, i41-49.

30. Petryszak, R., Kretschmann, E., Wieser, D. and Apweiler, R. (2005) The predictive power of the CluSTr database. Bioinformatics, 21, 3604-3609.

31. Bairoch, A. (2000) The ENZYME database in 2000. Nucleic Acids Res, 28, 304-305.

32. Rappoport, N., Fromer, M., Schweiger, R. and Linial, M. (2010) PANDORA: analysis of protein and peptide sets through the hierarchical integration of annotations. Nucleic Acids Res, 38, W84-89.

33. Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J.A., Heidelberg, K.B., Manning, G., Li, W. et al. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol, 5, e16.

TABLES

Table 1.

Database	Clustering	CS	Specificity	Sensitivity
UniRef90	MC-ProtoNet	**0.89**	**0.96**	**0.92**
	Single Linkage	0.78	0.93	0.24
	ProtoNet 4.0	0.75	0.94	0.79
UniRef50	MC-ProtoNet	**0.88**	**0.96**	**0.91**
	Single Linkage	0.72	0.91	0.79
SwissProt	MC-ProtoNet	**0.90**	**0.96**	**0.94**
	Single Linkage	0.81	0.90	0.91

Clustering performance evaluation based on Pfam keywords. Tests were performed on UniRef90 (1.8M), UniRef50 (960K) and SwissProt (220K). CS, correspondence score.

FIGURES LEGENDS

Figure 1

ProtoNet clusters following pruning at selected thresholds. (A) A scheme of the binary tree following low and high condensations (LT≥x and LT≥y). The high level of compression (LT=5) results in a smaller number of stable clusters. (B) Each panel represents a cluster summary according to a selected threshold (LT). Low (LT=0.2) and high condensation level (LT=5.0) differ in their cluster size and other statistical properties. Details on the cluster size, depth (by ProtoLevel, PL), the number of hypothetical proteins, solved structures in the PDB database and more are shown.

Figure 2

The contribution of annotation types to ProtoNet clusters. (A) About 40 annotation types that cover different aspects of function are included. Some of the minor annotation sources were combined and depicted as 'others'. (B) The major annotation types and their coverage as measured by the fraction of proteins that are assigned with the indicated annotation type are listed. In ProtoNet 6.0, a total of 143,849,828 annotations (74,416,565 without taxonomy) is associated with the ~9 million protein sequences.

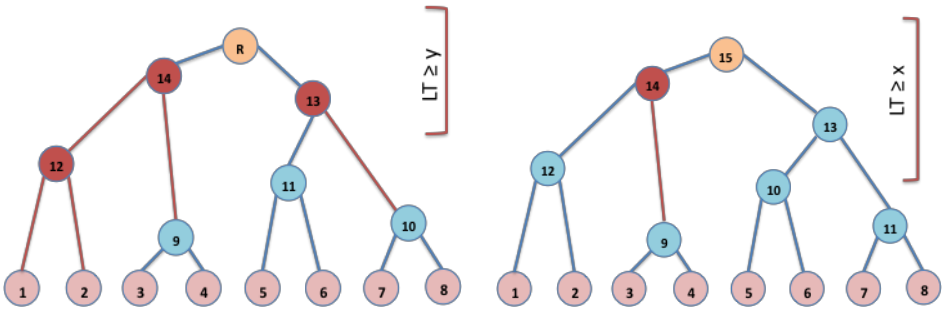
Figure 3

ProtoNet cluster page and a tree viewer in simplified and advanced modes. (Top) From the cluster page (Cluster ID 4201544) the user can focus on the ProtoName and

the collection of additional high quality annotations that are associated with this cluster. The number of proteins from the selected organisms is indicated with a framed T-symbol (for Taxonomy). Similarity, clusters that include proteins with 3D solved structures as marked by a symbol for PDB. Each cluster provides a short summary as a popup box with the number of proteins and the appearance of pre-selected organisms. The red edges in the tree indicate the branches that include the selected organisms. All other branched are faded. (Bottom) Using the advanced mode, the number of clusters in the ProtoNet tree is listed according to the predetermined LT and PL values. There are several sorting options according to the cluster size and the properties of the tree. An interactive use of the condensation levels allows inspecting the near vicinity of a subjected cluster in the ProtoNet hierarchy.

Figure 1

A

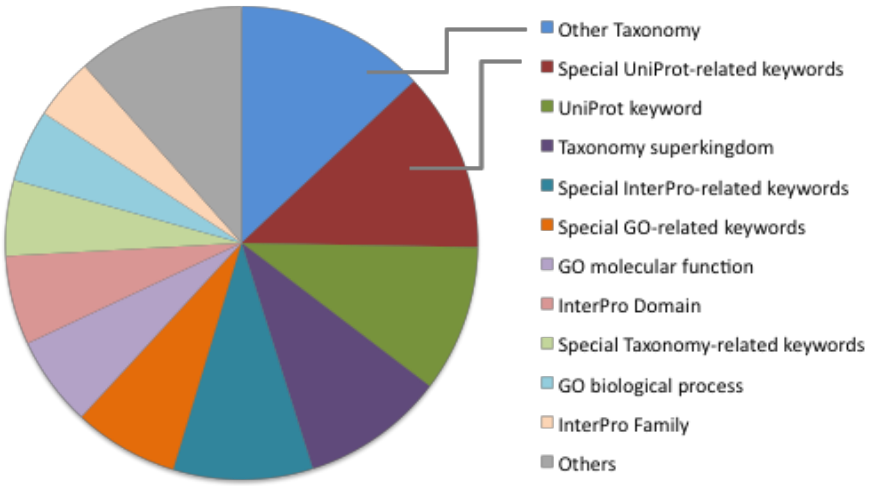


B

	LT=0.2	LT=5
Size	387 proteins	636 proteins
Fraction of UniProtKB/Swiss-Prot proteins	46.2%	32.2%
Fraction of UniProtKB/TrEMBL proteins	53.7%	67.7%
Number of Solved structures (in PDB)	16 proteins, 70 PDB's	17 proteins, 74 PDB's
Number of proteins without Prosite ID	68	176
Number of Hypothetical proteins	6	23
Number of Fragments	0	0
Average length and standard deviation of proteins	398 ± 63.2	376.4 ± 132.1
Number of clusters at ProtoLevel at creation	351879	get data
Number of leaves at ProtoLevel at creation	195678	get data
ProtoLevel at creation	0.798639	48.9337
ProtoLevel at termination	1.54009	60.7214
Lifetime	0.313586	6.3886

Figure 2

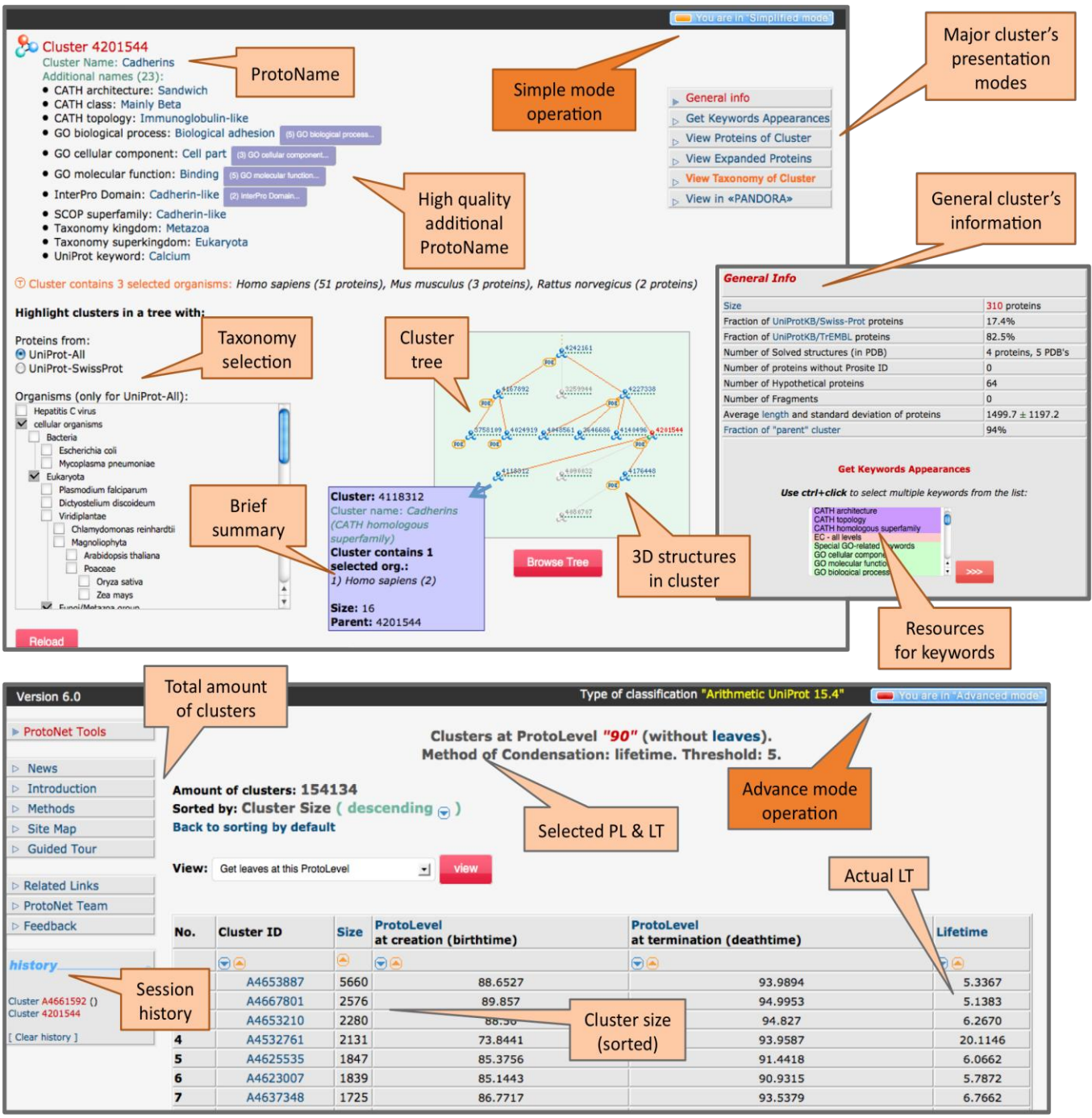
A



B

Major annotation sources	Coverage (% of proteins)	Amount of different annotations
ENZYME (6/09)	11	5,190
SMART (6.0)	14	720
GENE3D (6.1.0)	27	1,024
CATH (3.2.0)	27	3,338
SCOP (1.75)	35	7,821
GO (Gene Ontology) (1.7)	52	27,050
UniProtKB/SWP (15.4)	63	949
PFAM (23.0)	73	10,640
InterPro (21.0)	77	18,638
NCBI Taxonomy (6/09)	100	442,867

Figure 3


You are in "Simplified mode"
Cluster 4201544
Cluster Name: Cadherins
Additional names (23):
CATH architecture: Sandwich
CATH class: Mainly Beta
CATH topology: Immunoglobulin-like
GO biological process: Biological adhesion
GO cellular component: Cell part
GO molecular function: Binding
InterPro Domain: Cadherin-like
SCOP superfamily: Cadherin-like
Taxonomy kingdom: Metazoa
Taxonomy superkingdom: Eukaryota
UniProt keyword: Calcium
ProtoName
Simple mode operation
General info
Get Keywords Appearances
View Proteins of Cluster
View Expanded Proteins
View Taxonomy of Cluster
View in «PANDORA»
Major cluster's presentation modes
High quality additional ProtoName
General cluster's information
Cluster contains 3 selected organisms: Homo sapiens (51 proteins), Mus musculus (3 proteins), Rattus norvegicus (2 proteins)
Highlight clusters in a tree with:
Proteins from:
UniProt-All
UniProt-SwissProt
Taxonomy selection
Organisms (only for UniProt-All):
Hepatitis C virus
cellular organisms
Bacteria
Escherichia coli
Mycoplasma pneumoniae
Eukaryota
Plasmodium falciparum
Dictyostelium discoideum
Viridiplantae
Chlamydomonas reinhardtii
Magnoliophyta
Arabidopsis thaliana
Poaceae
Oryza sativa
Zea mays
Cluster tree
Brief summary
Cluster: 4118312
Cluster name: Cadherins (CATH homologous superfamily)
Cluster contains 1 selected org.:
1) Homo sapiens (2)
Size: 16
Parent: 4201544
Browse Tree
3D structures in cluster
General Info
Size 310 proteins
Fraction of UniProtKB/Swiss-Prot proteins 17.4%
Fraction of UniProtKB/TrEMBL proteins 82.5%
Number of Solved structures (in PDB) 4 proteins, 5 PDB's
Number of proteins without Prosite ID 0
Number of Hypothetical proteins 64
Number of Fragments 0
Average length and standard deviation of proteins 1499.7 ± 1197.2
Fraction of "parent" cluster 94%
Get Keywords Appearances
Use ctrl+click to select multiple keywords from the list:
CATH architecture
CATH topology
CATH homologous superfamily
EC - all levels
Special GO-related keywords
GO cellular component
GO molecular function
GO biological process
>>>
Resources for keywords
Reload
Version 6.0
Total amount of clusters
Type of classification "Arithmetic UniProt 15.4"
You are in "Advanced mode"
ProtoNet Tools
News
Introduction
Methods
Site Map
Guided Tour
Related Links
ProtoNet Team
Feedback
history
Cluster A4661592 ()
Cluster 4201544
[Clear history]
Session history
Clusters at ProtoLevel "90" (without leaves).
Method of Condensation: lifetime. Threshold: 5.
Advance mode operation
Amount of clusters: 154134
Sorted by: Cluster Size (descending)
Back to sorting by default
Selected PL & LT
View: Get leaves at this ProtoLevel
view
Actual LT
No. | Cluster ID | Size | ProtoLevel at creation (birthtime) | ProtoLevel at termination (deathtime) | Lifetime
A4653887 5660 88.6527 93.9894 5.3367
A4667801 2576 89.857 94.9953 5.1383
A4653210 2280 88.50 94.827 6.2670
4 A4532761 2131 73.8441 93.9587 20.1146
5 A4625535 1847 85.3756 91.4418 6.0662
6 A4623007 1839 85.1443 90.9315 5.7872
7 A4637348 1725 86.7717 93.5379 6.7662
Cluster size (sorted)