ProtoNet 4.0: A hierarchical classification of one million protein sequences

Noam Kaplan[1*], Ori Sasson[2], Uri Inbar[2], Moriah Friedlich[2], Menachem Fromer[2], Hillel Fleischer[2], Elon Portugaly[2], Nathan Linial[2] and Michal Linial[1]

[1] Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Israel

[2] School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel

[*] Corresponding author

Corresponding author details:
Noam Kaplan
E-Mail: kaplann@cc.huji.ac.il
Address: The Hebrew University, Department of Biological Chemistry, Givat Ram Campus, Jerusalem, Israel, 91904
Telephone: 972-2-6585433

ABSTRACT

ProtoNet is an automatic hierarchical classification of the protein sequence space. In 2004 the ProtoNet (version 4.0) presents the analysis of over one million proteins merged from SwissProt and TrEMBL databases. In addition to rich visualization and analysis tools to navigate the clustering hierarchy, we incorporated several improvements that allow a simplified view of the scaffold of the proteins. An unsupervised biologically valid method that was developed resulted in a condensation of the ProtoNet hierarchy to only 12% of the clusters. A large portion of these clusters were automatically assigned high confidence biological names according to their correspondence with functional annotations. ProtoNet is available at: http://www.protonet.cs.huji.ac.il

INTRODUCTION

ProtoNet (1) (launched at 2002) is an automatic hierarchical clustering of the SwissProt and TrEMBL (2) protein databases. The clustering process is based on an all-against-all BLAST (3) similarity search. The similarities' E-score is used to perform a continuous bottom-up clustering process by joining at each step the two most similar protein clusters, resulting in a hierarchy of protein clusters at various degrees of biological granularity. This hierarchy is structured as a collection of trees, in which the root clusters contain all the proteins of the tree and the rest of the clusters

represent subdivisions of the proteins into smaller groups. Browsing this global hierarchical organization of the protein world provides several interesting biological insights about protein families and the evolution of structural and functional relations between proteins (4). Furthermore, ProtoNet can be used to assess the function of novel protein sequences, by finding the best matching cluster for the new sequence. We describe here several new developments in ProtoNet 4.0 including the increase of the sequences analyzed from 114,033 proteins in SwissProt (version 2.1) to 1,072,911 sequences (SwissProt and TrEMBL, version 4.0) and the improved methodology for simplification of the scaffold of the protein hierarchy.
ProtoNet is available at: http://www.protonet.cs.huji.ac.il

HIERARCHY CONDENSATION

Due to the immense size of the ProtoNet hierarchy and the number of protein clusters, it would be very difficult to navigate in such a large hierarchy. Furthermore, it is obvious that many of the clusters are biologically irrelevant and uninteresting (for example huge root clusters containing hundreds of thousands non-related proteins). In order to get a condensed yet biologically relevant view of the hierarchy, we searched for some process-intrinsic parameter that would indicate which clusters are biologically relevant. The parameter found measures the stability of the cluster in the process, assuming that a stable cluster would be also more relevant biologically. We found this assumption to be correct, and that if we select a small subset of the clusters that show high stability, they would retain the biological validity of ProtoNet (Kaplan et al., in preparation).
The default condensation of the ProtoNet hierarchy leaves 12% of the clusters. However, the ProtoNet website now offers an "advanced mode", in which advanced users can control the level of condensation and the method by which it is done, resulting in a larger or smaller set of clusters as required. Note that the condensation causes the trees to change from binary trees to non-binary trees, and some browsing options have been developed accordingly (see "Web Enhancements").

DATABASE UPDATES

ProtoNet has gone through a major update of all database sources. Primarily, the protein database from which the ProtoNet tree is constructed has been updated and extended to include the TrEMBL protein database as well as SwissProt. This results in a leap from 114,033 to 1,072,911 protein sequences. Although TrEMBL is not manually validated by experts, it provides a much more extensive view of the protein world including whole genomes and thorough representation of several key organisms (see Table 1).

Table 1

CLUSTER NAMES

Assigning a biological function to proteins is a major objective in bioinformatics. We have developed an automatic high-confidence method that assigns a functional annotation to ProtoNet protein clusters. The method finds a functional annotation from either InterPro (5), GO (6), SwissProt or ENZYME (7) databases that best fits the proteins of the cluster and assigns it a score relative to how well it fits the cluster. If this score is above a certain threshold, the annotation is assigned as the cluster's name. Understandably, not all protein clusters would have an existing annotation that fits them well. Clusters whose best fitting annotation does not pass the threshold remain nameless, possibly suggesting a novel functional group or clusters that are associated with mixed functions. By applying this method we were able to assign biological names to 78% of the clusters that contain 20 proteins or more. Due to the fact that a high threshold is used, the annotation can be assigned with high confidence to the cluster.

WEBSITE ENHANCEMENTS

Several enhancements have been made to the ProtoNet website in order to allow easier and more in-depth analysis of the ProtoNet trees.

BROWSING CLUSTER NAMES

Cluster names are extremely useful for quickly browsing the ProtoNet trees, eliminating the need to check the proteins of each cluster in order to get an impression of the hierarchy. Furthermore, the assignment of a biological function to clusters suggests an easy scheme of assigning function to proteins with unknown function: a protein can be assigned the function of all clusters to which it belongs. This scheme can be used not only on each of the 1,072,911 proteins from which the ProtoNet hierarchy is constructed, but also for new protein sequences given by the user (using the "Classify your protein" option in the website, which finds the most suitable cluster for a new protein sequence given by the user).

BROWSING THE TREE

Subtree View: In order to cope with the change to a non-binary tree, we have introduced the ProtoBrowser (Figure 1), which shows the tree in the vicinity of the cluster that is being displayed. Instead of presenting only the branch of the tree to which the cluster belonged to, the new display allows easy navigation to neighboring clusters and an enhanced global overview of biological protein families.

Figure 1

Functionality View: PANDORA (8) is a web-based tool that allows in-depth biological analysis of large protein sets. When trying to biologically interpret large protein clusters that contain hundreds of proteins, PANDORA is a natural choice. ProtoNet now offers a direct link from its cluster page to PANDORA, providing the ability to easily understand what biological groups the cluster is built from and analyze them by several different biological aspects.

Similarity View: When viewing a protein cluster, it is sometimes helpful to obtain an in-depth look into the sequence similarity between the proteins of the cluster. This could allow the user to tell if there is a natural partitioning into subgroups or the cluster inner similarity is uniform, for example (see example in Shachar and Linial, 2004). In order to address this, the website offers a cluster similarity matrix (Figure 2), showing an all-against-all color coded matrix of all protein pairs in the cluster, colored according to the BLAST E-score between the two proteins. This also facilitates access to the BLAST result, which can be obtained simply by clicking on the appropriate cell in the matrix.

Figure 2

MAINTENANCE AND UPDATING

The ProtoNet source databases are generally updated twice a year. The next ProtoNet release is planned to include the UniProt Ref 100 protein database. Other future plans include allowing the users to select a subset of proteins from ProtoNet according to their needs (e.g. selecting for study only the proteins from the SwissProt database or only the proteins of a specific species) and expanding links to further biological databases such as OMIM (9) and DIP (10).

ACKNOWLEDGEMENTS

We thank the entire current and previous ProtoNet team for their endless support. Special thanks to Alex Savenok for the Web design as well as for the development of the visualization tools. We thank the fellowship support by the Sudarsky Center for Computational Biology (SCCB) to NK, UI, MF and MF. This study is partially supported by the EU NoE BioSapiens consortium and the CESG consortium supported by the NIH.

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. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O`Donovan, C., Phan, I. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365-370.
3. 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**, 3389-2402.
4. Shachar, O. and Linial, M. (2004) A robust method to detect structural and functional remote homologues. *Proteins*, in press.
5. Camon, E., Magrane, M., Barrell, D., Lee V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262-D266.

6. Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P. et al. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315-318.
7. Bairoch, A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304-305.
8. Kaplan, N., Vaaknin, A. and Linial, M. (2003) PANDORA: keyword-based analysis of proteins sets by integration of annotation sources. *Nucleic Acids Res.*, **31**, 5617-5626.
9. McKusick, V.A. (1998) Mendelian Inheritance in Man. Johns Hopkins University Press, Baltimore.
10. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. and Eisenberg D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449-451.

FIGURE LEGENDS

Figure 1.
The ProtoBrowser allows viewing the near vicinity of a cluster in the ProtoNet hierarchy. Blue triangle-shaped icons represent protein clusters. The cluster currently being viewed is the cluster A260586, which appears in the center in red. Clusters that include proteins with 3D solved structures as marked PDB.

Figure 2.
Example of a cluster similarity matrix. Colored cells represent different degrees of similarity, ranging from white (no similarity: BLAST E-score higher than 100) to dark blue (high similarity, BLAST E-score close to 0). It is evident that the cluster A222801 is roughly divided into 3 subsets: in the upper left of the diagonal there are proteins which show no similarity to each other or to any protein in the cluster; in the center of the diagonal there is a subset of proteins that are similar to each other but to no other proteins; and at the bottom right of the diagonal there is another subset of proteins that are similar to each other but not to other proteins.

TABLES

Table 1.

Species	Proteins in ProtoNet 2.1	Proteins in ProtoNet 4.0
Homo sapiens	8,507	47,641
Mus musculus	5,678	41,813
Drosophila melanogaster	2,049	22,603
Arabidopsis thaliana	1,680	39,367
Plasmodium falciparum	153	8,434

FIGURES

Figure 1.

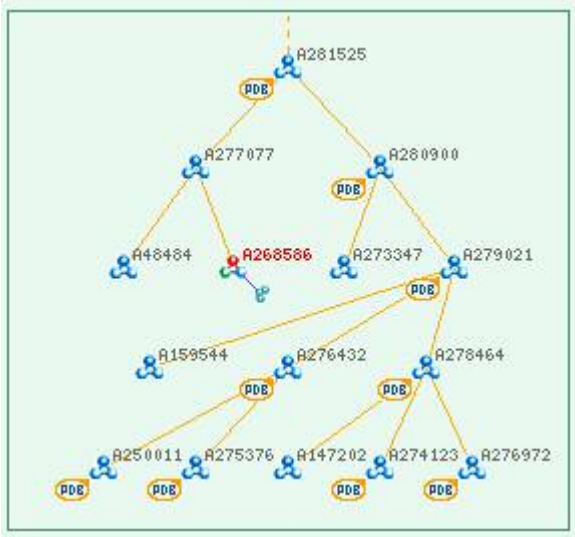

Figure 2.