

# ProtoNet: Navigating the Hierarchical Clustering of the Protein Space

**Ori Sasson<sup>1\*</sup>, Hillel Fleischer<sup>1</sup>, Elon Portugaly<sup>1</sup>, Yonatan Bilu<sup>1</sup>,  
Nathan Linial<sup>1</sup>, Michal Linial<sup>2</sup>**

<sup>1</sup> School of Computer Science and Engineering, Hebrew University, Jerusalem 91904 Israel

<sup>2</sup> Department of Biological Chemistry, Institute of Life Sciences, Hebrew University, Jerusalem 91904 Israel

\*Corresponding Author. Email: ori@cs.huji.ac.il

## Abstract

The ProtoNet site provides an automatic hierarchical clustering of the protein space. The clustering is based on an all-against-all BLAST similarity test. With this similarity measure we proceed to perform a continuous bottom-up clustering process by applying alternative rules for merging clusters. The outcome of this clustering process is a classification of the input proteins into a hierarchy of clusters of varying degrees of granularity. This clustering can be used for function prediction, for define superfamilies and subfamilies and for large-scale protein annotation purposes.

ProtoNet is accessible in the form of an interactive Web site at <http://www.protonet.cs.huji.ac.il>.

## Introduction

Recent years have seen an explosive growth in the amount of biological data gathered by the scientific community. Specifically, the amount of publicly available protein sequences increased rapidly, much as a result of large-scale sequencing projects including that of the human genome.

The large volumes of data collected give rise to the need to classify and sort data in a global automated manner. However, currently used methods, mostly those based on automated procedures are limited in their success to infer function (Bork and Koonin, 1998). While a multitude of techniques exist for comparing sequences (Needleman and Wunch 1970; Smith and Waterman 1981; Lipman and Perason 1985; Altschul et al. 1990), the issue of determining whether two biological sequences share the same function is yet to be resolved in general. It is well known that sequence similarity or

structural similarity imply a high likelihood of having the same biological function.

The shortcomings of plain sequence analysis algorithms give rise to the concept of classification and clustering. Clustering sequences provides a way to identify homologies using the simple observation that homology is by definition a transitive relation. By clustering sequences into groups based on similarity, it is expected to discover relations that are not visible with simple sequence comparison. The rationale is that if a sequence A is similar to B, and B is similar to C, sequences A and C might share a similar function, even without exhibiting high sequence similarity. In other words, clustering may reveal facets of the relationship between protein sequences that are not visible in sequence comparison since sequence similarity is not transitive, whereas homology (and biological function) is. As is well known, and indicated below, transitivity has its perils and must be conducted with great care.

Protein classification algorithms are roughly divided to those based on motif and domain analyses and those that rely on whole protein analysis (reviewed by Kriventseva et al., 2001). The latter have to deal with the problems that many proteins are multiple domains. Consequently, transitivity may result in classifying together non-related proteins that share some highly conserved domains (but not others).

The advantages of clustering proteins and the observation that it may provide insight based on transitivity have been studied and implemented in various systems, such as ProtoMap (Yona et al. 1999), Systers and ClusTr (Krause et al 2000) and Picasso (Heger and Holm 2001).

Our clustering methods depend on standard similarity measure, namely gapped BLAST. We rely on the notion of restricted transitivity in order to perform a continuous process of clustering. As this process progresses, we discover larger clusters, making use of weaker similarity. In our previous work (e.g. Yona et al. 1999), we used predetermined thresholds for constructing the hierarchy and that resulted with discrete, somewhat arbitrary stages. Herein, we allowed the procedure of clustering to progress continuously, so the resulting clusters have different levels of granularity.

## **Methods**

This section details the computational aspects of our work, namely the required pre-computation, and the clustering algorithms.

### ***Pre-Computation***

The basis for our clustering process is a sequence comparison measure for each pair of proteins in Swissprot 39 (~94,000 proteins). We use standard gapped BLAST

based on BLOSUM62 and on filtration of low complexity sequences.

From the BLAST output we obtain a set of pairs, each associated with an E-Score value. For pairs with very low (or no) similarity, the E-Score value is very large (or infinite), and we ignore it. We arbitrarily select a high E-Score value, and cut off higher values. In this work, this threshold value is set at 10, clearly above any expected significance. Comparison of two proteins with E-Score 10 or above, will rarely show any significant similarity.

It was earlier established (see e.g. Portugaly et al. 2001) that similarity among clusters at such low levels of confidence does include a good deal of significant biological information. In previous systems, this information was present, but very noisy, and its interpretation required an expert's view. A major advance of the present study is that the levels of noise are reduced and the important biological information becomes much more apparent.

### ***Clustering Methodology***

Our clustering method is an adaptation of the widely accepted hierarchical clustering paradigm. The clustering algorithm starts out with each protein as a singleton cluster. It then iteratively merges pairs of clusters, always selecting the pair of clusters whose merging has the lowest merging score. We use three different merging scores, producing three different clustering hierarchies.

For the purpose of performing the clustering, we identify proteins with running numbers starting from 1. The singleton associated with a certain protein shares the same number (so cluster 1204 is a singleton containing protein 1204).

## Merging Rules

When merging two clusters, we are interested in performing the most beneficial merge. The corresponding notion in the case of clustering proteins is based on the E-Score of pairs of proteins in a cluster. However, in order to avoid bias towards larger or smaller clusters, we focus on notions of ‘average’ E-score. We consider the following merging scores for two clusters, based on the E-scores of pairs of proteins from the union of them:

- **Arithmetic** - The merging score is the arithmetic mean of E-Scores for all pairs.
- **Geometric** - The merging score equals the geometric mean of E-Scores for all pairs.
- **Harmonic** - The merging score equals the harmonic mean of E-Scores for all pairs.

A simple inequality ties all these averages. The harmonic mean is less than or equal to the geometric mean which in turn is less than or equal to the arithmetic mean.

This comes into play in the clustering process by the way weak similarities are considered. The arithmetic gives the most weight to weak similarity scores (large E-values). This weight decreases as we progress through the geometric mean, and finally harmonic mean. Similarly, the harmonic mean gives much more weight to strong similarities (small E-values) than the rest of the tested means.

## ProtoNet Web-site

The ProtoNET website is available at <http://www.protonet.cs.huji.ac.il>. The website allows easy navigation through the hierarchy of clusters created and a set of queries that assist biological validation of the clustering quality.

## Searching for Proteins

ProtoNet allows searching for proteins based on their ProtoNet identifier, their Swissprot name, or the full protein name.

**Figure 1** shows an example of a protein page in ProtoNet.

The information displayed for a single protein includes the internal ProtoNet identification, the Swissprot name, identification number and update date (where applicable), the Prosite accession number (where applicable), the full protein name and length in amino-acids. Furthermore, the actual sequence is displayed and annotated with color codes, and the protein's taxonomy is presented.

In addition, the protein motifs and domains as described by Prosite, ProDom, Pfam, SMART, Prints and Interpro are shown graphically along the protein sequence. The exact amino-acids covered by the domains are also indicated.

The protein page allows you to jump to the corresponding singleton cluster in each, for each of the merging rules (marked as Harmonic, Geometric, and Arithmetic).

## Navigating the Cluster Hierarchy

The cluster hierarchy is navigated using the cluster page. The cluster page shows the list of proteins in the cluster as well as a summary of the taxonomy for this cluster. Separate pages provide additional information about the cluster, a listing of neighbors, a breakdown of keywords (including SwissProt and InterPro keywords), and a tool for aligning any protein pairs from this cluster.

The top of the cluster page provides a graphical navigation tool for browsing the cluster hierarchy at high resolution. This navigation pane shows the current cluster, to its right the clusters whose merging created

it, and to its left the sequence of larger clusters generated from subsequent mergings. The navigation tool marks differently merges of clusters of size 2 and above. Furthermore, a view of the number of singletons at any new merge is provided by pop-up tool box.

A commutative list of keywords for each cluster is provided. The breakdown page for all keywords provides an easy way to track the ‘purity’ of generated clusters (of course, this applies to clusters whose proteins are annotated with keywords in the respective databases). Partition of the proteins in a cluster according to their major phylogenetic groups is provided.

A table summarizing the statistical features of the cluster under inspection is provided. It includes the number of merging steps, the number of clusters in this level of the hierarchy, the average length of protein within this cluster, the fraction of the current cluster considering the next merging steps and other statistics. These features can be used in comparing the global properties of the clusters at each of the merging procedures described.

**Figure 2** shows an example of a cluster page.

### ***Classifying New Proteins***

ProtoNet allows users to insert their own protein sequences, which are stored in the system database. Such new sequences are inserted into the clustering, and are available to the owning user when navigating the clustering hierarchy.

### ***Inserting New Sequences***

The issue of updating the clustering is an important one. Large numbers of new proteins are discovered, and it is of high interest to ‘map’ them into an existing

clustering, since this can shed light on their function and structure.

ProtoNet implements an algorithm for inserting a new protein into an existing clustering system, which is based on performing a BLAST computation of the new protein sequence against all sequences stored in the database. The clustering for the new protein is approximated using the protein closest this new sequence in the BLAST similarity test.

### ***Following the navigating tour***

In order to ease the task of navigating the clustering system, ProtoNet provides a powerful history mechanism that allows the user to return to any of the most recently visited proteins and clusters (20 steps).

## **Conclusion**

ProtoNet provides a tool for comprehensive analysis of all protein sequences in Swissprot database. ProtoNet allows a dynamic view of the protein clusters at different level of granularity and according to varying merger rules. The ability to insert new sequences allows users to study their own sequences against the various clustering systems.

## **Acknowledgements**

This study could not be done without the outstanding effort of the ProtoNet team: Shmuel Brody, Lilach Traivish, Hagit Mor and the excellent work of Alexander Savenok in designing the website. Special thanks to Edna Wigderson for and suggestions and fruitful discussions.

This work was supported by the Israeli Ministry of Defense, the Israeli Ministry of Science and the Horowitz Foundation.



## PROTEIN 40222

About protein 40222	
ProtoNet ID	40222
Swissprot ID	MAUF_METFL
Swissprot accession number	<a href="#">Q50418</a> , <a href="#">Q50419</a>
Prosite accession number	
InterPro accession number	
Update date in Swissprot	01 October 2000
Protein name	METHYLAMINE UTILIZATION PROTEIN MAUF (FRAGMENTS).
Length in amino acids	153
PDB	

### Go to cluster of protein "40222"

Select type of classification:

Geometric ▾

Go to cluster 40222

### Sequence of protein 40222:

```
1 MSIDAKLSRQASGSKAGVACVPDAYSFSEA 30
  RSPGTRFALMLSAVAAGLAGGAMLHSAMSA
61 TSALTGLFIVLALAGGFLLSTWSPCGYSSLS 90
  LLRPAGRYSFASVLRWSPTFFTHATVQVWL
121 ARWQERAVEVDGFLLLSVGSAALIAAGAVR 150
  AGV
```

Get motifs and domains of protein

Figure 1: ProtoNet Protein Card

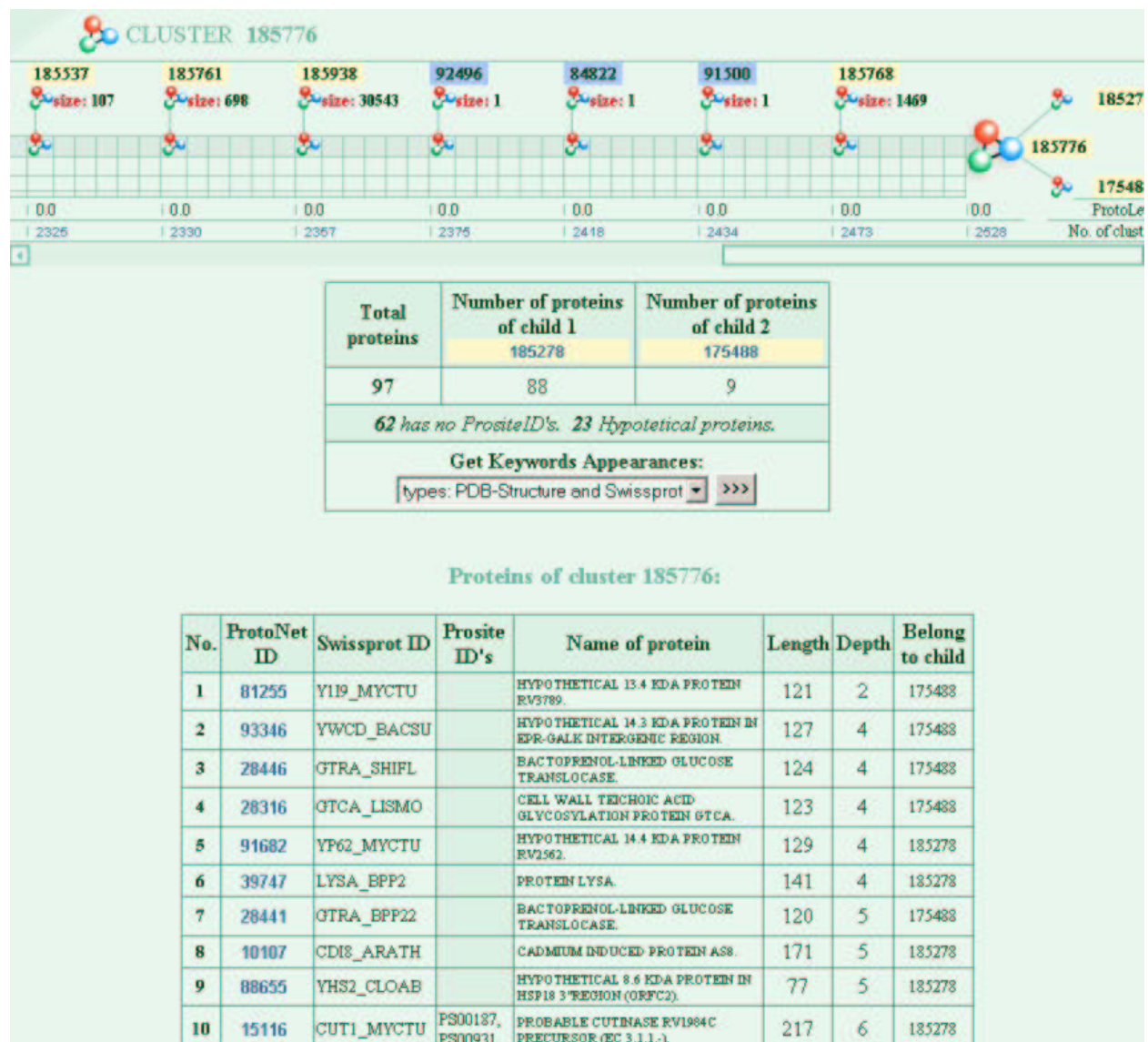


Figure 2: ProtoNet Cluster Page

## References

- Altschul S.F., Carrol R.J. and Lipman D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215: 410-410.
- Bork P. and Koonin E. V. (1998) Predicting functions from protein sequences-where are the bottlenecks? *Nature Genet.* 18:313-318.
- Needleman S.B., and Wunsch C.D. (1970). A general method applicable to the search of similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48:443-453.
- Heger A., and Holm L. (2001) Picasso: generating a covering set of protein family profiles. *Bioinformatics* 17:272-279.
- Holm L. and Sander C. (1997). Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* 25:231-234.
- Krause A., Stoye J., and Vingron, M. (2000) The SYSTERS Protein Sequence Cluster Set. *Nucleic Acids Research* 28: 270-272.
- Kriventseva E.V., Biswas M.; and Apweiler R. (2001). Clustering and analysis of protein families. *Current Opinion in Structural Biology*, 11:3:334-339.
- Lipman D.J. and Pearson, W.R. (1985) Rapid and sensitive protein similarity. *Science* 227: 1435-1441.
- Portugaly E., Linial M. (2000) Estimating the probability for a protein to have a new fold: A statistical computational model. *Proc. Natl. Acad. Sci. USA* 97:5161-5166
- Smith, T.F., and Waterman, M.S. (1981). Comparison of Biosequences. *Advances in Applied Mathematics* 2:428-489.
- Yona, G., Linial, N. and Linial M. (1999) ProtoMap – Automated classification of all proteins sequences: a hierarchy of protein families, and local maps of the protein space. *Proteins: Structure, Function, and Genetics* 37:360-378.