

**PUBLIC**

SIXTH FRAMEWORK PROGRAMME  
PRIORITY 1  
LIFE SCIENCES, GENOMICS AND BIOTECHNOLOGY FOR HEALTH



# DIAMONDS PROJECT

Contract no: LSHG-CT-2004-512143

EU Deliverable  
**D3.1**  
Methods for data integration and functional classification and prediction

**Date:** 28<sup>th</sup> February 2006  
Version 2

**Partner responsible: Michal Linial (HUJI)**



## Methods for data integration and functional classification and prediction

This report summarizes the different Methods for data integration that are being exploited by the DIAMONDS partners. We also included an overview of ‘Methods’ used by the community for functional classification and predictions. Several of the methods are elected to become incorporated into the DIAMONDS platform and other will be considered.

Several methods including PANDORA (Kaplan et al., 2003) are using statistical measures for both functional inference and visualization. We will, however, not discuss visualization tools and methods in this report.

The Methods for Data Integration and Predictions are as follows:

### **Functional Classification and Prediction: General View**

*Accurate automatic functional annotation* holds the potential for enormous benefits in speeding up the annotation process of new biological data. At present, there are at least 500 genomes that are either completed or at the final stages of a draft phase. The genomes of an additional 522 genomes (as of October 2005) are currently in the pipeline. This unprecedented number of genomes includes ~200 eukaryotes that are in their final stage of assembly or in progress (<http://www.ncbi.nlm.nih.gov/Genomes>). These new genomes largely outnumber the 18 complete eukaryotic genomes currently available. Therefore, the need for automation in the painstaking task of functional annotation becomes critically important.

In addition to ongoing ‘whole genome’ projects, other types of experimental data are becoming available from numerous high throughput methodologies. In recent years, standardization in the technologies of *SNP arrays*, *DNA micro-array ‘ChIP on Chip’*, *Protein-Protein Interaction TAP data*, and *transcriptomic DNA chips* has increased the quality and reproducibility of the results. Overall, the volume of data that is collectively referred to as “*non-sequence data*” is exponentially growing. However, the quality of the data varies. While the quality of some data sources may be very high other types may be of inherently poor quality. For example, structural genomics projects produce detailed and accurate 3D information from crystallography and NMR spectroscopy. The function of many of these structures, however, is still unknown (Skolnick and Fetrow, 2000). In contrast, data on protein-protein interactions originating from two-hybrid systems suffers from large numbers of false positives and low reproducibility. With the addition of proteomics data from LC MS/MS experiments, protein chips, and subcellular localization data, much of the data which emerges is protein rather than genome centered (Bork et al., 2004).

### **Prediction of Function – list of pitfalls and difficulties:**

The notion of protein function is elusive. In order to apply computational methods, we need to provide an unambiguous definition. We suggest equating *function* to

*annotations*. Annotations are simply categorical biological properties describing the protein's functionality. Annotations can describe various biological aspects of the protein such as its *structure, enzymatic classification, taxonomy, cellular localization*, and more.

Local alignment search tools such as BLAST (Altschul et al., 1997) provide the most straightforward method for performing automatic function prediction on a new sequence (Jones and Swindells, 2002), via function inference. With this method, a protein database is searched for high scoring local alignments with the query protein. The annotations on the sequence which scores the highest alignment are assigned to the query sequence, provided the alignment score passes a predetermined threshold. The underlying logic is simple: proteins with similar sequences are conjectured to have evolved from a single ancestral gene, and thus to have retained similar functionality. However, local alignment searches suffer from some important caveats:

- (I) *Excessive transfer of annotations*. In some cases, similarity is restricted to a local region in the sequence. While only annotations that are functionally linked to the region of similarity should be transferred, annotations which are not related to the local region of similarity will be transferred as well, even though they are not shared by both proteins. This difficulty arises even when using manual inference of the annotations, as it is not possible to conclusively determine what annotation is linked to the region of similarity. The reason is that the connection between specific segments of the protein to its function is often unknown. Excessive transfer of annotations occurs more frequently for annotations that describe a high-level functionality than for annotations that are motif-based and can be localized in sequence.
- (II) *Annotation errors in the source database*. Due to the fact that many databases employ computational methods in the assignment of annotations, isolated cases of false annotation assignment occur. Studies have shown that once an erroneous annotation is introduced into a database, it tends to propagate via automatic annotation inference methods that are based on sequence similarity (Linial, 2003). If the best matching sequence has been assigned a false annotation, the annotation will be transferred to the new protein sequence.
- (III) *Threshold relativity*. Various scoring methods exist for assessing the quality of an alignment. The score threshold used for annotation is usually arbitrary and fails to reflect the relativity that scoring methods tend to exhibit (different thresholds are suitable for different groups of proteins).
- (IV) *Low sensitivity/specificity*. Depending on the annotation threshold that is used, simple local alignment methods are usually outperformed by advanced supervised methods in terms of sensitivity/specificity. This is due to the fact that advanced methods take into account features which are shared by the family of proteins to which the protein belongs, while a simple local alignment search does not consider this data.
- (V) *Paralogs versus orthologs*. Two different proteins in one species that resulted from a gene duplication event might possess significant sequence similarity but will often have different functions. In contrast, two proteins from different species that may have almost undetected similarity can still share the same function or a similar one. Sequence comparison methods frequently fail to distinguish between these two instances (Sonnhammer and Koonin, 2002).

## Methods for Prediction of Function:

The method for inference of functional annotations of protein sequences consists of two parts:

- (i) An automatic organization of protein sequence databases for families representing functional and evolutionary relations amongst the proteins. For illustrating the method we will refer to ProtoNet. Identical procedures are valid for other classification methods that are hierarchical such as Systems (Krause et al., 2005) and additional high quality classification such as PIRSF (Apweiler et al., 2004).
- (ii) (ii) An automatic method for predicting the function of a new protein based on its localization in the protein tree (Kaplan et al., 2005; Sasson et al., 2003).

### Protein classification method for functional prediction: Example ProtoNet

Given a set of proteins (typically a protein from a database such as UniProt (Bairoch et al., 2005)), ProtoNet aims at organizing the proteins into a hierarchy of trees, each tree representing a biologically-related group of proteins and its division into functional subgroups. Much work was done in the field of protein classification, and in particular hierarchical clustering e.g. Systems (Krause et al., 2005), CLusTr (Kriventseva et al., 2001).

In contrast to a non-hierarchical functional grouping, hierarchical representation of proteins provides a much more accurate view on protein functional relations, because functionality encompasses several degrees of granularity, from very general effects at the organism level to very specific descriptions of biochemical function. To achieve this organization, we use the following three phases:

- (I) All-against-all BLAST. A matrix is constructed so that it holds the e-values resulting from NCBI-BLAST comparisons (McGinnis and Madden, 2004) on all possible pairs of sequences. E-values greater than 100 are set to be equal 100.
- (II) Clustering. We use the well known paradigm of hierarchical agglomerative clustering (Kaufman and Rousseeuw, 1990) using group average linkage. We use arithmetic averaging, and define the score between two clusters. At each step of the clustering method, the pair of clusters that has the lowest score is merged.
- (III) Pruning. An automatic unsupervised method is applied to distinguish biologically valid clusters from clusters that are artifacts of the method. Following the method presented in (Kaplan et al., 2004), the resulting hierarchy is automatically pruned according to an intrinsic measure. The pruning method has been shown to eliminate 88% of the clusters while maintaining the validity of the remaining cluster, as measured by its correspondence to external classification benchmark data sources.

ProtoNet is available at: <http://www.protonet.cs.huji.ac.il>.

## Methods for quality control of classification – Validity and Performance

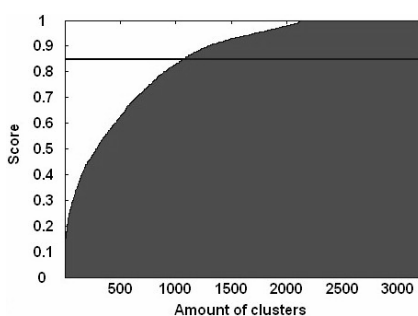
The validity of clusters can be determined in comparison to other classifications, e.g. InterPro (Mulder et al., 2002). At present, the InterPro classifier uses a combination of 12 supervised detection methods based on state-of-the-art methods such as Hidden Markov Models (HMMs), Position Specific Scoring Matrices (PSSMs) and profiles (Mulder et al., 2005).

In order to determine if ProtoNet is able to detect the weak functional relationships that are detected by InterPro, we perform the following test: For each InterPro annotation (each InterPro entry can be thought of as an annotation), we consider the set of all proteins that were assigned that annotation ( $S$ ). Next, we define a score between a cluster  $C$  and the set  $S$  (this score is also known as the Jaccard coefficient) and we refer to it as *Correspondence Score* (CS). The CS for a certain cluster and a given keyword (a biological term that annotates proteins) measures the correlation between the cluster and the keyword, using the well-known intersect-union ratio:

$$score(C, S) = \frac{|C \cap S|}{|C \cup S|}$$

Let  $C$  be a cluster in the ProtoNet tree, and  $K$  be a keyword (from a specific source such as InterPro) that annotates some proteins in the system. The score for a given cluster on annotation  $S$  ranges from 0 (no correspondence) to 1 (the cluster contains exactly all of the proteins with annotation  $S$ , i.e. maximally corresponds to the keyword). Finally, we find the highest scoring cluster for each InterPro annotation.

The figure shows an area plot describing the distribution of the scores for the highest scoring cluster of each InterPro annotation. ProtoNet is able to produce clusters that are extremely consistent with the InterPro classification (mean score 0.85). For more results see (Kaplan et al., 2004) and (Shachar and Linial, 2004).



Note that in term of ‘prediction’ the method described above is an unsupervised one. Supervised methods are given a training set upon which they learn a pattern and then use it to perform prediction. Therefore supervised methods are only able to detect predefined. In contrast, unsupervised approach expected to detect previously unknown families and previously unknown relationships between families. Example for a new finding by this method (Furman et al. 2006).

## Integration for Functional Annotation: Prediction

Given that the protein clusters and the hierarchies are highly coherent with other classifications that can be used in order to annotate a new sequence. We discuss the integration of:

- InterPro (Mulder et al., 2002)
- SCOP (Hubbard et al., 1999)
- GOA (Camon et al., 2004)
- ENZYME (Bairoch, 2000)),

When provided with a new sequence, it is localized to an existing cluster. Once it is localized, we can learn about its functionality from its relative position in the hierarchy. To do this, we first assign to each cluster the annotations of its member proteins, which adhere to the following two conditions:

- (a) The annotation is shared by at least 75% of the proteins in the cluster and
- (b) The annotation achieves a p-value lower than 0.001 under the assumption that the annotations are distributed hypergeometrically (see Appendix for a formal definitions)

These two requirements ensure that only annotations that are *statistically significant* and *represent a majority of the proteins of the cluster* will be assigned to the cluster. Furthermore, these requirements provide a secondary measure of caution to prevent clusters that are not biologically coherent due to methodical flaws (i.e. mixed groups of functionally unrelated proteins) from being used to infer annotations.

**Inference:** Once the clusters are assigned annotations, the new sequence is assigned the annotations of the cluster to which it belongs and the annotations of all of the cluster's parents in the hierarchy. By doing this, robustness is used in order to avoid most of the pitfalls noted previously. One pitfall which is difficult to overcome is the issue of correctly inferring the function of paralogs which evolved into having a new function. Such sequences might be misclassified in our method, but this is inevitable regardless of the method used.

The aforementioned procedure was applied to over 10,000 unannotated predicted proteins from the honey bee genome. A ProtoNet like approach including about 200,000 sequences was applied ([www.protobee.cs.huji.ac.il](http://www.protobee.cs.huji.ac.il)) and for ~75% of the honey bee proteins some biological annotation was successfully assigned (Kaplan and Linial, unpublished). In addition to the high sensitivity/specificity results compared to other methods and the threshold relativity which the clustering method is able to take into account, it seems that this method succeeds in avoiding many of the common pitfalls of local alignment searches.

#### **Additional methods for integration and functional prediction:**

Several new approaches for automatic function prediction were introduced recently in order to advance beyond the shortcomings of simple local alignment searches (Edgar and Sjolander, 2004; Godzik, 2003; Han et al., 2005; Yang, 2004). While the relative performance of these methods is difficult to benchmark, it is clear that they are all superior to the naïve approach.

An approach which is related to the one presented in this work is prediction by phylogenomic methods, using the evolutionary context of a sequence for function prediction (Engelhardt et al., 2005). The use of the evolutionary context is analogous to the use of the classification hierarchy in this work.

An interesting advantage of the inference mode discussed above over the naïve local similarity search approach is that *any kind of functional annotation* can be assigned to the new sequence. This means that any data that is available through the underlying database of proteins is available for use in annotation.

By using UniProt as its underlying protein database many high quality sources are available including InterPro, UniProt keywords, GO, ENZYME and SCOP and more. This not only offers a wider and constantly-growing range of available annotations, but also overcomes inconsistencies between different sources.

Integration by STRING (Von Mering et al. 2003) is based on a somewhat different model. There a confidence score assigned to each predicted association. These scores are derived by benchmarking a reference set of ‘true’ annotations. Functional grouping of proteins is based on KEGG (Kanehisa, 2002) that is based on manual annotations.

### **Integration of functional sources as tools**

The concept of integration in addition to the clear need for quality inference (as discussed in this report) is to provide a clear view on sets of sequences. Biological interpretation of such sets is time-consuming and requires intimate knowledge of each protein. Furthermore, it is difficult to gain a global view of the protein set and to detect biologically significant subsets.

*PANDORA* was developed in order to allow in-depth biological analysis of such large protein sets. This is obtained through annotation analysis, with the implementation of two main ideas:

1. representation of all protein-keyword relations with a Concept DAG (Directed Acyclic Graph), and
2. integration of several annotation sources covering different biological aspects, such as: function, 3D structure, cellular localization, taxonomy and participation in biological processes.

*PANDORA* is based on the proteins that appear in the SwissProt database (version 41.21) and TrEMBL (version 24.8) and is available in [www.pandora.cs.huji.ac.il](http://www.pandora.cs.huji.ac.il)

A statistical model for the quality of the integration was developed. The user can see an integrated view including the quality of the set in term of consistency of any annotation and combination of annotation.

The idea of statistical and graphical integration was explored by *PANDORA*, by *STRING* (von Mering et al., 2003) and several other efforts.

**Functional inference for multi-domain proteins:** One problem that remains partially unaddressed is the problem of multiple domains. A protein often consists of several

domains and in eukaryotes this phenomena is extensive. It can be viewed as belonging to several protein families. In ProtoNet (and other mentioned classification systems) proteins are the basic entities. As a result of this, every protein appears once and can therefore belong to several families only if they contain each other. This issue is irresolvable in the current scheme.

However, this issue is addressed in a related work called EVEREST ([www.everest.cs.huji.ac.il](http://www.everest.cs.huji.ac.il)), in which protein domains are the basic entities that are clustered. There, integration of methods and annotations is presented only as a visualization aid. For most experimentalists, it is suggested to use multiple sources and to view one in view of the others.

Example for such rich visualization scheme is shown in the website. EVEREST shows visualization integrating several functional annotation of **Pfam**, **GO**, **SCOP**, **ENZYME**, **Swissprot keywords**, **Taxonomy** and more.

#### References:

- 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 Research* 25, 3389-3402.
- Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., *et al.* (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 32, D115-119.
- Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Res* 28, 304-305.
- Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., *et al.* (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Res* 33, D154-159.
- Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I., and Marcotte, E. M. (2004). Protein interaction networks from yeast to human. *Curr Opin Struct Biol* 14, 292-299.
- Camon, E., Barrell, D., Lee, V., Dimmer, E., and Apweiler, R. (2004). The Gene Ontology Annotation (GOA) Database--an integrated resource of GO annotations to the UniProt Knowledgebase. *In Silico Biol* 4, 5-6.
- Edgar, R. C., and Sjolander, K. (2004). COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics* 20, 1309-1318.
- Godzik, A. (2003). Fold recognition methods. *Methods Biochem Anal* 44, 525-546.
- Han, S., Lee, B. C., Yu, S. T., Jeong, C. S., Lee, S., and Kim, D. (2005). Fold recognition by combining profile-profile alignment and support vector machine. *Bioinformatics* 21, 2667-2673.
- Hubbard, T. J., Ailey, B., Brenner, S. E., Murzin, A. G., and Chothia, C. (1999). SCOP: a Structural Classification of Proteins database. *Nucleic Acids Res* 27, 254-256.
- Jones, D. T., and Swindells, M. B. (2002). Getting the most from PSI-BLAST. *Trends Biochem Sci* 27, 161-164.
- Kanehisa, M. (2002). The KEGG database. *Novartis Found Symp* 247, 91-101; discussion 101-103, 119-128, 244-152.
- Kaplan, N., Friedlich, M., Fromer, M., and Linial, M. (2004). A functional hierarchical organization of the protein sequence space. *BMC Bioinformatics* 5, 196.

- 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.
- Kaplan, N., Vaaknin, A., and Linial, M. (2003). PANDORA: keyword-based analysis of protein sets by integration of annotation sources. *Nucleic Acids Res* 31, 5617-5626.
- Linial, M. (2003). How incorrect annotations evolve--the case of short ORFs. *Trends Biotechnol* 21, 298-300.
- McGinnis, S., and Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* 32, W20-25.
- Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., *et al.* (2002). InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform* 3, 225-235.
- Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., *et al.* (2005). InterPro, progress and status in 2005. *Nucleic Acids Res* 33, D201-205.
- 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.
- Shachar, O., and Linial, M. (2004). A robust method to detect structural and functional remote homologues. *Proteins* 57, 531-538.
- Skolnick, J., and Fetrow, J. S. (2000). From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends Biotechnol* 18, 34-39.
- Sonnhammer, E. L., and Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* 18, 619-620.
- von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 31, 258-261.
- Yang, Z. R. (2004). Biological applications of support vector machines. *Brief Bioinform* 5, 328-338.

## Appendix

### Inference by a statistical measure

The annotation of new unannotated sequences is performed as described in the text. First, Annotations that are assigned to the clusters can be taken from the following sources: UniProt keywords, InterPro, GO and enzyme E.C. numbers. As a result, in the hierarchical platform each protein is assigned the annotations that were given to the cluster to which it belongs and the annotations that were assigned to all the cluster's parents in the hierarchy.

The p-value for a cluster  $C$  and an annotation  $a$  given the database  $D$  is calculated according to the hypergeometric distribution:

$$pvalue(a, C, D) = \sum_{i=|C \cap A|}^{\min(|A|, |C|)} \frac{\binom{|A|}{i} \binom{|D|-|A|}{|C|-i}}{\binom{|D|}{|C|}}$$

where  $A$  is the set of all proteins in the database that have annotation  $a$ .

Similar procedures can be applied to any set where a coherence is expected (pathway, network, subcellular organelle etc).