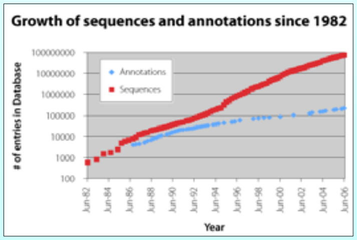
Computational Approaches for Protein Function Prediction

Gaurav Pandey, Michael Steinbach, Rohit Gupta, Tushar Garg, Lakshmi N. Ramakrishnan, Vipin Kumar
Department of Computer Science and Engineering, University of Minnesota (200 Union Street SE, Minneapolis MN 55455 USA)

e-mail: gaurav@cs.umn.edu website: http://www.cs.umn.edu/~kumar/dmbio

OPPORTUNITY

- The knowledge of protein function is a crucial link in the development of new drugs, better crops, and synthetic biochemicals such as biofuels
- With the rapid advances in sequencing technology, over-whelming number of proteins being characterized for several hundred organisms
- Hard to perform high-throughput annotation by experimental methods
- Consequently
  - Most of these proteins are unannotated or 'hypothetical'
  - The annotation gap seems to be widening at an exponential rate



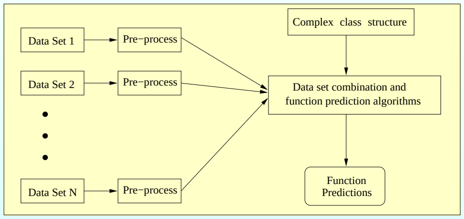
http://www.ctwatch.org/quarterly/articles/2006/08/genome-sequencing-vs-moores-law/

- Recently, several high throughput experimental techniques have been developed in molecular biology
- Immense amounts of genomic and proteomic data has been accumulated in standardized databases available for free access

- Hence, there is both an unprecedented opportunity and urgent need to develop computational approaches to study the functions of proteins
  - Unprecedented amount of readily accessible biological data
    - Protein sequences and structures
    - Genome sequences
    - Phylogenetic profiles
    - Protein-protein interactions
    - Gene expression data
    - Bio-medical text and literature
  - Urgent need to complement the experimental determination of protein function by providing potential hypotheses generated by computational methods

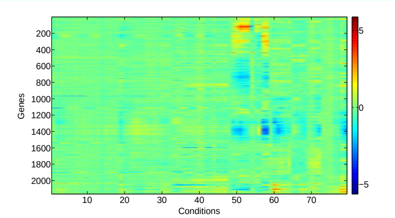
CHALLENGES AND APPROACHES

- We address protein function prediction as a data mining problem using techniques such as classification and clustering
- Currently working on various forms of genomic data for budding yeast (S. cerevisiae):
  - Protein-protein interaction networks
  - Gene expression data
  - Phylogenetic profiles
  - Soon to start: mass spectrometry data
- Open to research on other organisms and other types of biomedical data



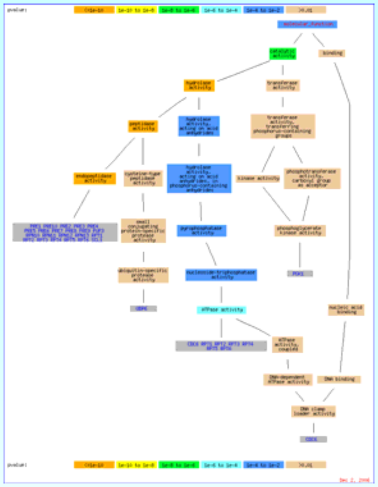
Issues being addressed by the group

- We are currently addressing the following issues that are important for accurate prediction of protein function:
  - Effective pre-processing of biological data
  - Factoring of the complex structure and inter-relationships of functional classes into the prediction algorithms
  - Integrating multiple sources and types of biological data
- Several other interesting issues, such as effective benchmarking amd evaluation of results, will be addressed in the near future
- Open to research on other issues also



Pre-processing of biological data

- Several commonly used types of biological data have well-known quality issues. Examples:
  - Microarray data: noise, different scales of experiments
  - Protein interaction maps: false positive interactions, incompleteness of maps
- Innovative techniques from other domains of data mining used to address these issues:
  - Normalization of microarray data
  - Graph transformation applied to interaction networks



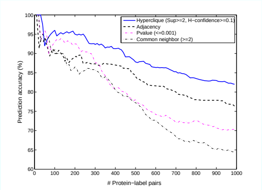
Handling the complex structure of functional classes

- Use of reliable functional hierarchies such as Gene Ontology raises several important challenges:
  - Multiple functional annotations of a protein
  - Hierarchical arrangement of functional classes
  - Rare functional classes
- Current approach: Modeling the inter-relationships between different functional classes by measuring their semantic similarity in the hierarchy and the data

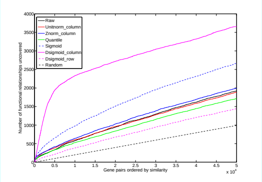
Integrating multiple sources and types of biomedical data

- Different types of biological data often provide supplementary and sometimes even complementary information about biological processes. Thus, combining different sources and types of biological data provides several advantages:
  - *Global picture of biological processes *Noise reduction and quality improvement *More reliable predictions
- Several approaches for effective protein function prediction using this integration
  - Combination of results from several data sets and factoring the inter-relationships between functional classes
  - Exploiting interdependencies between different data sets and results

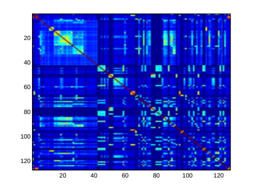
RESULTS



Graph transformation using assoication-analysis based approaches produces better function prediction results from interaction networks than other transformation techniques, due to better handling of both false positive and false negative interactions



Normalization of gene expression data using known and novel techniques enhances the functional content of microarray data sets



Different types of semantic similarity measures for pairs of functional classes arranged in a hierarchy are expected to improve multi-label function prediction (work in procgress).. Shown here: Lin et al's similarity measure applied to a set of 127 GO biological processes identified by Myers et al

REFERENCES

- *Computational Approaches for Protein Function Prediction: A Survey*, Gaurav Pandey, Vipin Kumar, Michael Steinbach, Technical Report 06-028, October 2006, Department of Computer Science, University of Minnesota
- *Association Analysis-based Transformations for Protein Interaction Networks: A Function Prediction Case Study*, Gaurav Pandey, Michael Steinbach, Rohit Gupta, Tushar Garg, Vipin Kumar, Technical Report 07-007, March 2007, Department of Computer Science, University of Minnesota
- *Comparative Study of Various Genomic Data Sets for Protein Function Prediction and Enhancements Using Association Analysis*, Rohit Gupta, Tushar Garg, Gaurav Pandey, Michael Steinbach and Vipin Kumar, To appear in the proceedings of the Workshop on Data Mining for Biomedical Informatics, held in conjunction with SIAM International Conference on Data Mining, 2007
- M. Ashburner et al., *Gene Ontology: tool for the unification of biology*, Nature Genetics 25, 1, 25–29.
- E. Nabieva et al., *Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps*, Bioinformatics 21, Suppl. 1, i1–i9
- M. Kuramochi and G. Karypis, *Gene classification using expression profiles: A feasibility study,* International Journal on Artificial Intelligence Tools. Vol. 14, No. 4, pp. 641 - 660, 2005

ACKNOWLEDGEMENTS

This work has been supported by NSF grants #CRI-0551551, #IIS-0308264 and #ITR-0325949. Computational resources for this work were provided by MSI.