

Title: Integrated Protein Family Classification for Functional Genomics

A. Tutor: Cathy H. Wu, Ph.D.

Education

- Ph.D. in Molecular Plant Pathology, Purdue University, 1984
- M.S. in Computer Science, University of Texas at Tyler, 1989

Primary Expertise

- Principal Investigator, Protein Information Resource
- Conduct bioinformatics research since 1990 and develop several protein family classification systems including MOTIFIND neural networks, GeneFIND identification system, and ProClass/iProClass databases

Professional Appointments

- Director of Bioinformatics and Vice President, National Biomedical Research Foundation; Affiliate Professor of Georgetown University Medical Center, George Mason University, and University of Maryland
- Assistant Professor of Computer Science, University of Texas at Tyler (1989-94)
- Assistant Professor (1990-94), Associate Professor (1994-98), Professor (1998-99) of Biomathematics, University of Texas Health Center at Tyler

Teaching Experience

- Bioinformatics: Co-director and co-instructor of a bioinformatics course at Georgetown University Medical Center (Spring 2000 & 2001); Taught a two-day bioinformatics workshop (>150 attendees), National Yang Ming University, Taiwan (October 2000)
- Computer Science: Taught many graduate and undergraduate courses, including Relational Databases, Operating Systems, Object-Oriented Programming, and Programming Languages at University of Texas at Tyler (1989-94)
- Mentored numerous bioinformatics projects at Georgetown U and U of Texas at Tyler

Publications

- Over 60 referred journal articles and conference papers, and a book - "Neural Networks and Genome Informatics," Elsevier Science, 2000

Professional Activities

- Member, Board of Directors, International Society for Computational Biology
- Co-Chair, CBGI-2001 (Atlantic Symposium on Computational Biology and Genome Information Systems)
- Bioinformatics review panels, National Science Foundation (NSF) and Department of Energy (DOE)
- More than 10 invited presentations last year

B. Tutorial Presentation

Motivation. Protein family classification is now well recognized as an effective means for large-scale genomic sequence annotation and functional characterization with several advantages. The classification approach (1) improves the sensitivity of protein identification, (2) provides complete clustering for database organization, (3) detects and corrects genome annotation errors systematically, (4) drives other functional annotations, and (5) stimulates research in molecular evolution, genomics, and proteomics. There have been developments of

many family-based sequence analysis algorithms (such as profile methods and hidden Markov models) and a proliferation of protein family databases (such as ProSite, Pfam, and InterPro). We have developed a framework that supports a high level of integration of the protein classification data and methods, including the integration of (1) family classification at the global, whole protein (superfamily and family) level with the local, structural/functional unit (domain and motif) level, (2) protein sequence families with function and structural classes, and (3) sequence-based and annotation-based database searches. The objective of such a framework is to facilitate discovery of comprehensive protein family relationships, which we believe is crucial to our understanding of protein evolution, structure, and function.

Goals. The goals of the tutorial are for the attendees:

- To familiarize with the concepts and terminology of protein family classification
- To appreciate the significance of integrated classification approach for protein sequence analysis and functional characterization beyond BLAST
- To review the major protein classification databases and family search algorithms
- To learn how databases and identification system could be integrated and what are some of the issues involved in database and infrastructure development
- To exchange ideas on issues and future directions of research and development

Contents

- Overview (30 min): Protein family classification concept and significance of the classification approach
- Family Classification Databases (1 hr): Overview of protein classification databases and iProClass case study for integrated classification database
- Family Identification Tools (1 hr): Overview of family search algorithms and GeneFIND case study for integrated family identification system
- Discussion (30 min): Database interoperability and infrastructure (PIR ontology, XML, and database schema) and open discussion

C. Intended Audience: The tutorial is aimed at both users who apply bioinformatics databases and analytical tools to formulate hypothesis and answer scientific questions, and developers who design and implement bioinformatics resource for functional genomics and proteomics studies. The audiences may be either biologists or computer scientists from industry or academia. The level of the tutorial is medium to advanced. Although introductory materials will be presented in the overview, audience who has some bioinformatics background would benefit the most. The attendees are expected to have basic biology (concepts of proteins, protein evolution, function, and structure) and basic bioinformatics knowledge (molecular databases, database information retrieval, and sequence similarity search). Background in algorithms (such as hidden Markov models, neural networks) or experience in computer programming and relational databases would help but are not required.

D. Length: The tutorial is for a half-day.

E. Detailed Outline of the Presentation: The tutorial will be given using a PowerPoint slide show with many real-world examples. The effect of a live demo will be provided with screen shots (and accompanying Web site URLs).

Overview (30 min):

- Family classification concept
 - Molecular evolution: homolog, ortholog, paralog
 - Family relationships based on sequence similarity: whole protein (superfamily/family hierarchical clustering), structural/functional units (domain), and functional sites (motif)
- Significance of classification approach
 - Detect genome annotation errors: e.g., IMP dehydrogenases and hisI bifunctional enzymes
 - Provide family-function relationships: e.g., ASK/SAT/CYSN (EC 2.7.7.4 & 2.7.1.25) domains

Family Classification Databases (1 hr):

- Overview of classification databases
 - Whole protein clustering: e.g., PIR superfamily, Protomap
 - Domains: e.g., Pfam, ProDom
 - Motifs: e.g., ProSite, Prints
 - Integrated: e.g., InterPro, Metafam, iProClass
 - Structural classification: SCOP and CATH
- Integrated family classification database - iProClass case study
 - Design principals: as an integrated resource to provide comprehensive family relationships at both global and local levels, as well as structural/functional classifications and features
 - Database functionality illustrated with example reports and analyses (superfamily and sequence reports, and text and sequence searches)

Family Identification Tools (1 hr):

- Overview of family search and alignment algorithms
 - Motif pattern matching: ProSite regular expressions
 - Profile method and hidden Markov models: position-specific scoring
 - ClustalW multiple sequence alignment method
 - MOTIFIND Neural network: sequence encoding, network architecture and learning
- Integrated family identification system – GeneFIND case study
 - System design: multi-level filter system for rapid and accurate protein classification and motif identification by combining several sequence search and alignment programs
 - System functionality illustrated with example reports and analyses
 - New development: ProAnnotator system to integrate sequence-based and annotation-based database searches

Discussion (30 min):

- Database interoperability and infrastructure – PIR case study
 - Controlled vocabulary and ontology development: alignment of thesaurus of terms
 - XML exchange format and associated DTD
 - Database development: overview of database and application design and implementation
- Open discussion