

- To join AI Seminar Mailing List, please send an email to [xli1@cs.uic.edu](mailto:xli1@cs.uic.edu), with the subject “joining AI Seminars”. Thanks!

# Protein Function Prediction Using Data Mining

Yi Zhang

Department of Computer Science

University of Illinois at Chicago

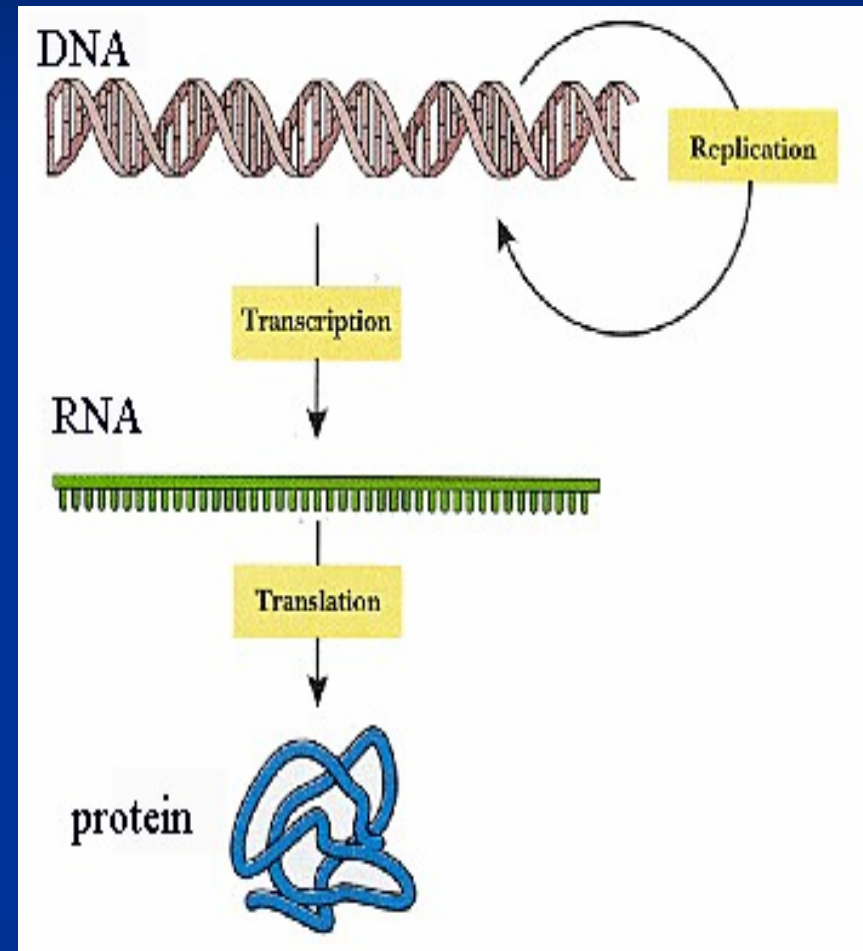
Sep. 2005

# Outline

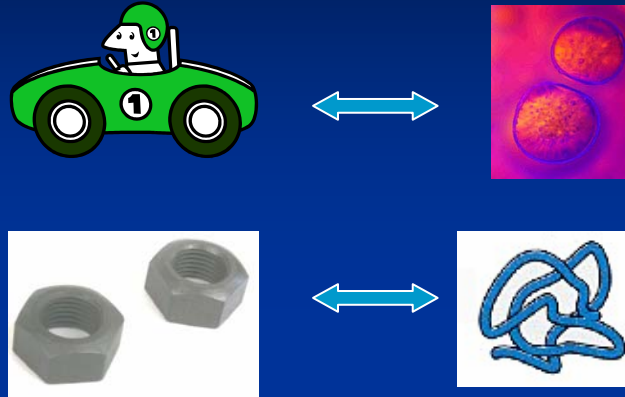
- Background
- Motivation
- Data mining
- Problem description
- Related work
- Proposed work

# Background

- Protein
  - Basic molecules in the cell.
  - A protein consists of a chain of amino acids. There are 20 different amino acids.
  - Protein structure.



# Protein Function



- In the cell, each protein is performing its specific function to make the cell work.
- Understanding the protein's functions is the key to understanding the cell.
- Functional annotation scheme, Enzyme Commission(EC) system. Each function is classified into 4 level hierarchy. (E.g. 2.1.3.37).

# Traditional Function Annotation

- Experiments
- Computational methods
  - By sequence similarity.
  - Identification by motifs, domains, protein fingerprints, and other sequence based features.
  - By protein families.
  - By protein structures.

# Prediction By Similarity

- Sequence similarity

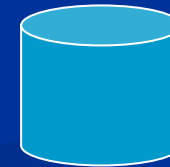
```
VTISCTGSSSNIGAH--VKWYQQLPG  
VTISCTGSSSNIGSWTVNWYQQLPG
```

- Similarity=>Homology=>Function

```
VTISCTGSSSNIG  
AHVKWYQQLP  
GAVTCGSQPLG  
NIAHKW.....
```

New Sequence

Database  
Scanning  
Tool (BLAST)



Sequence Database

EC 2.3.1.37

Function Prediction

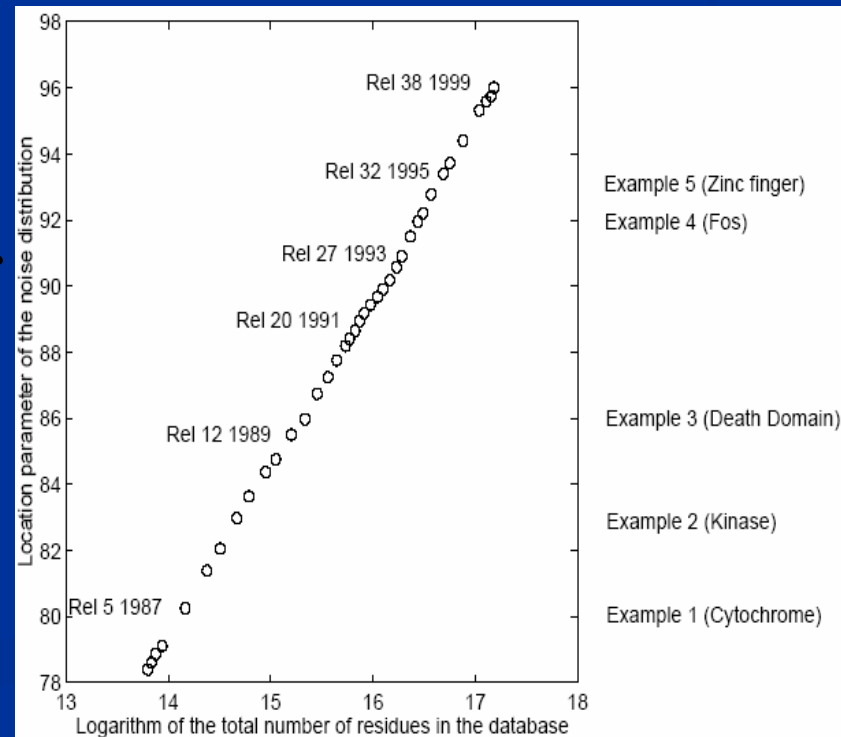
```
Protein1 EC 2.3.1.37  
Protein2 EC 2.3.1.37  
Protein3 EC 2.3.1.37  
Protein4 EC 2.3.1.37
```

Similar Sequences

# Prediction By Similarity -- Problems

- Low similarity.
- No consensus function.
- Noise in database search, i.e. random sequence similarity. (R. Spang and M. Vingron 2001)

Protein1 EC 2.3.1.37  
Protein2 EC 2.3.1.47  
Protein3 EC 2.3.1.47  
Protein4 EC 2.3.1.37



# Prediction By Sequence Based Features

- Multiple Sequence Alignment

```
VTISCTGSSSNIGAH--VKWYQQLPG  
NYISCTGSSSNIGSWTVNWOPQLPG  
LRISCSSTGFIFSS-YAMYWVRQAPG
```

- Motif/domain/fingerprints/functional sites

```
ISC  
ISC  
ISC
```

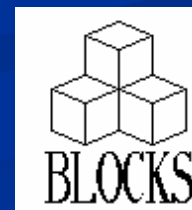
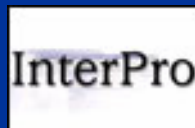
```
W  
W  
W
```

```
QLPG  
QLPG  
QAPG
```

- Position Specific Scoring Matrix

- Hidden Markov Model

- Public Domain/Motif Databases



# Protein Families and Protein Structures Databases

- Protein families databases

COGs

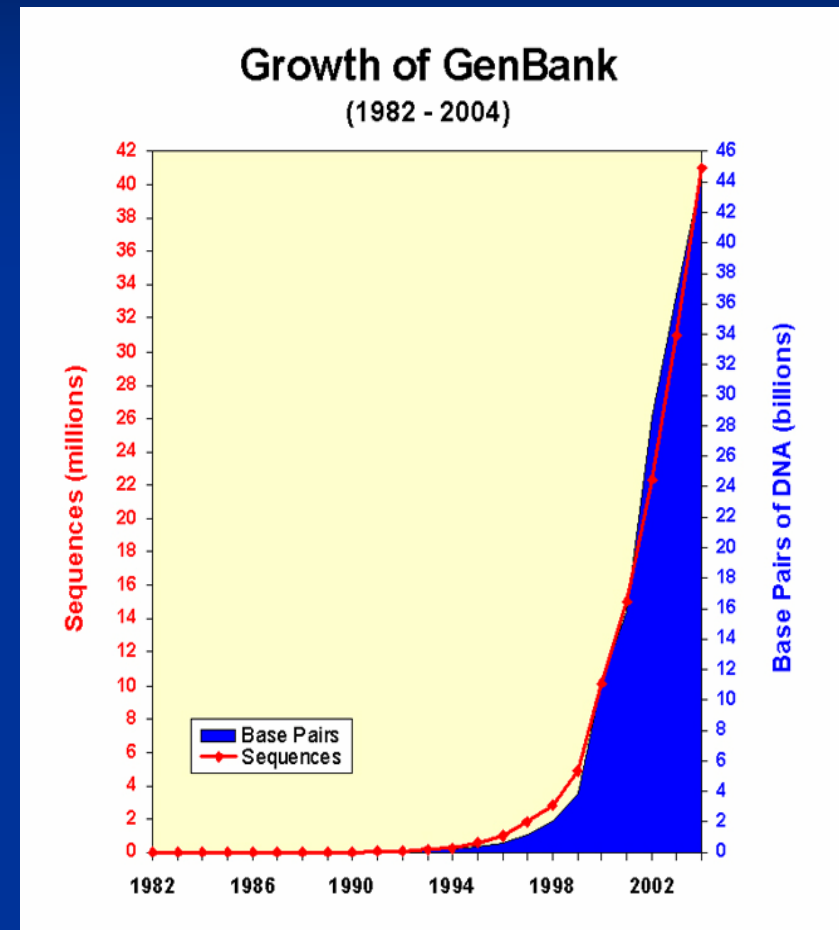


- Protein structures databases



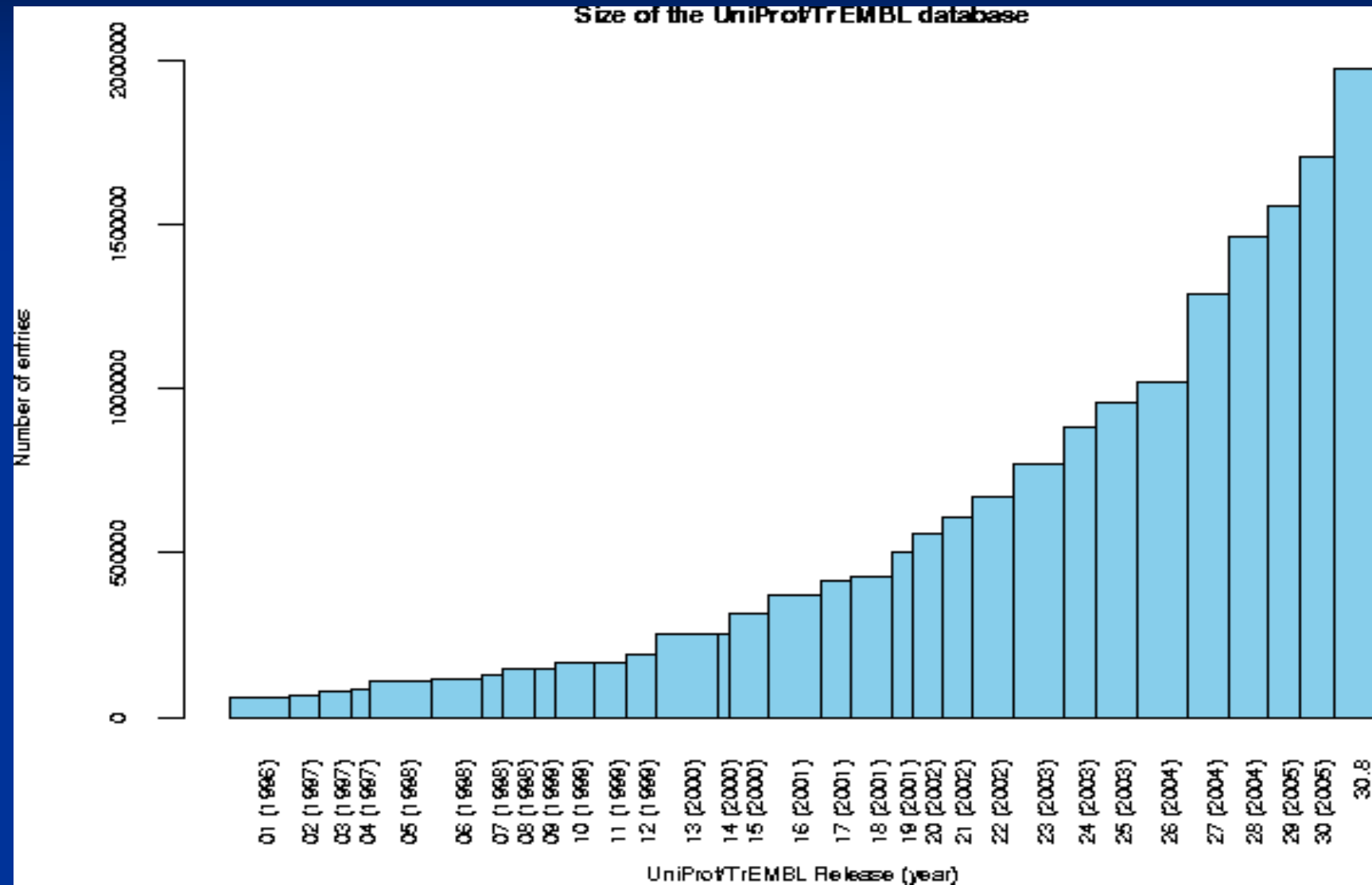
# Motivation

- Experimental function prediction is expensive and time-assuming.
- Protein sequence database nearly doubles every 12 months now.
- Many databases and bioinformatics tools are developed to store/capture protein sequence features.
- Manual curation is slow.

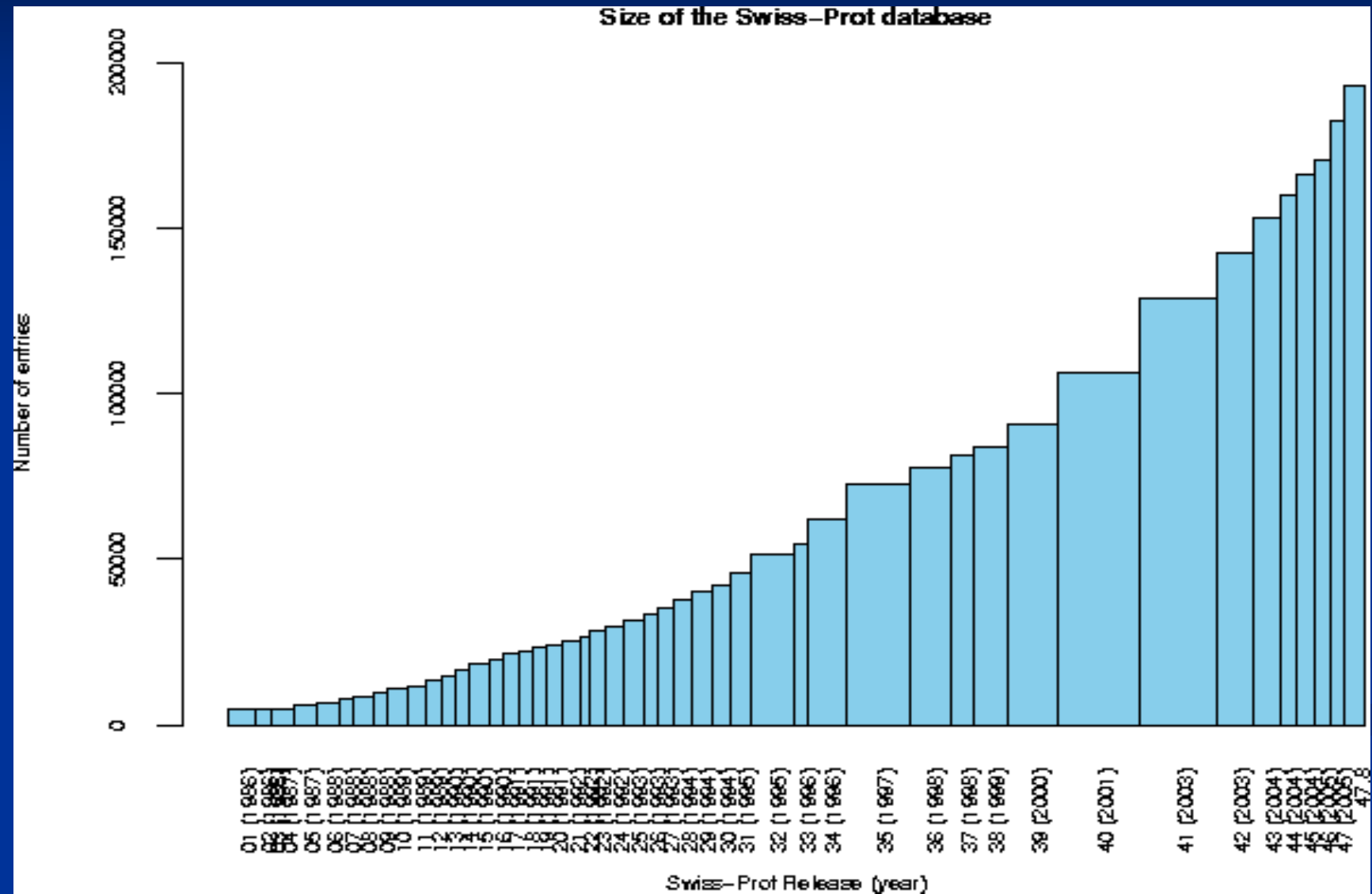


(Figure courtesy of <http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html> )

# Motivation

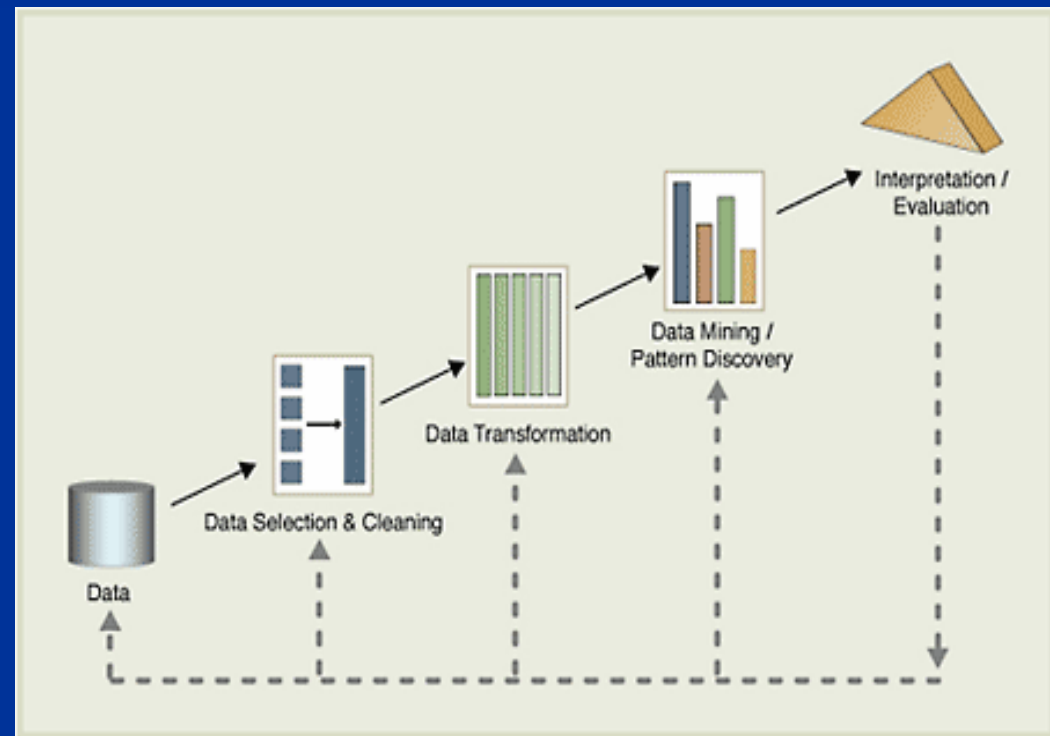


# Motivation



# Data Mining

- Data mining is – “extraction of useful *patterns* from data sources, e.g., databases, texts, web, image”.
- Patterns:
  - Equations
  - Decision trees
  - Predictive rules
  - Association rules
  - Probabilistic models
  - Distances and partitions
- Task
  - Classification
  - Clustering
  - Association analysis
- Process



# Problem Description

- Goal: prediction of protein function accurately and quickly.
- Data
  - One-per sequence value
    - Protein\_ID, taxonomy, and phenotype
  - Structured value
    - Sequence chemical/physical properties
    - Domain/motif/functional sites
    - Protein structures
    - Similarity data
- Class
  - EC number(4 numbers hierarchy)

# Challenge1: Data Representation

Protein_ID	Taxonomy	Organism	Function
P07024	Bacteria	E.Coli	3.6.1.45
P05079	Virus	Tobacco rattle virus	unknown

Protein_ID	Feature	Begin Position	End Position
P07024	Signal peptide	1	25
P07024	Helix	116	119
P07024	Helix	56	72

Protein_ID	Domain_ID	Begin Position	End Position	Significance
P07024	IPB00001	1	50	1.5E-20
P07024	IPB00002	55	86	1.0E-90

# Challenge 1: Data Representation

## ■ Propositionalization

- Ideal for conventional data mining programs.
- Simply join all tables
  - Time and space?
- Find frequent patterns and compress patterns
  - Space efficient
  - Frequent patterns search in multi-tables

## ■ Leave it

- Space efficient
- Need multi-table mining tools.

# Challenge 2: Large Data

- Over 2,000,000 proteins in protein databases right now.
- Over 200,000 proteins with function annotation.
- Number of features: ?
- Number of functions: 4442

# Related Work

- “Machine learning and data mining for yeast functional genomics” Ph.D thesis of Amanda Clare, Univ of Wales Aberystwyth.
- Various tools in multi-relational data mining, such as FOIL, TILDE, WARMR, and etc.

# Amanda Clare's PhD Thesis

- Principle: “expert propositionalization”. Use knowledge about biology in combination with various machine learning techniques for feature construction and finally uses C4.5 for classification.
- Data:
  - Chemical/physical properties of protein sequences
  - Phenotype
  - Data from microarray experiments
  - Predicted protein secondary structures
  - Similarity data
  - Size
    - initial sets: 3924 and 4290 proteins
    - Extended sets: 6300 proteins

# Amanda Clare's PhD Thesis

## ■ Methods

- Multi-label extension of C4.5.
- Clustering methods to investigate the relationship between microarray data and known biology.
- PolyFARM to mine large volumes of relational data in a distributed hardware.
- Use of hierarchically structured data and classes.

## ■ Results

- Initial sets average accuracy between 61-76% for level 1-3.
- Extended sets average accuracy 49-62% for level 1-2.

# Proposed Work

- Better coverage
  - All proteins in current databases
  - Most features
    - Sequence chemical/physical properties
    - Domain/motif/functional sites
    - Protein structures
    - Similarity data
- Data retrieval/preparation
- High-throughput computational environment using GRID (done)
- Data representation
- New mining algorithm for these needs

# References

- “Machine learning and data mining for yeast functional genomics”, PhD thesis, Amanda Clare.
- “Multi-relational learning for genetic data: issues and challenges”, C.Perlich and S.Merugu.2005
- “Relational Data Mining”, S.Dzeroski and N.Lavrac. Springer,2001.
- “Limits of homology detection by pairwise sequence comparison”, R.Spang and M.Vingron, Bioinformatics, Vol.17 no.4, 2001 pg.338-342.