Statistical Machine Learning Methods for Bioinformatics

VI. Support Vector Machine Applications in Bioinformatics

Jianlin Cheng, PhD
Computer Science Department and Informatics Institute
University of Missouri
2012

SVM Applications in Bioinformatics

- **Cancer Classification using Gene Expression Data**
- Protein Mutation Stability Prediction
- Protein Secondary Structure Prediction
- **Protein Fold Recognition**
- **Protein Contact Map Prediction**
- Protein Structure Classification
- ……

Project 5

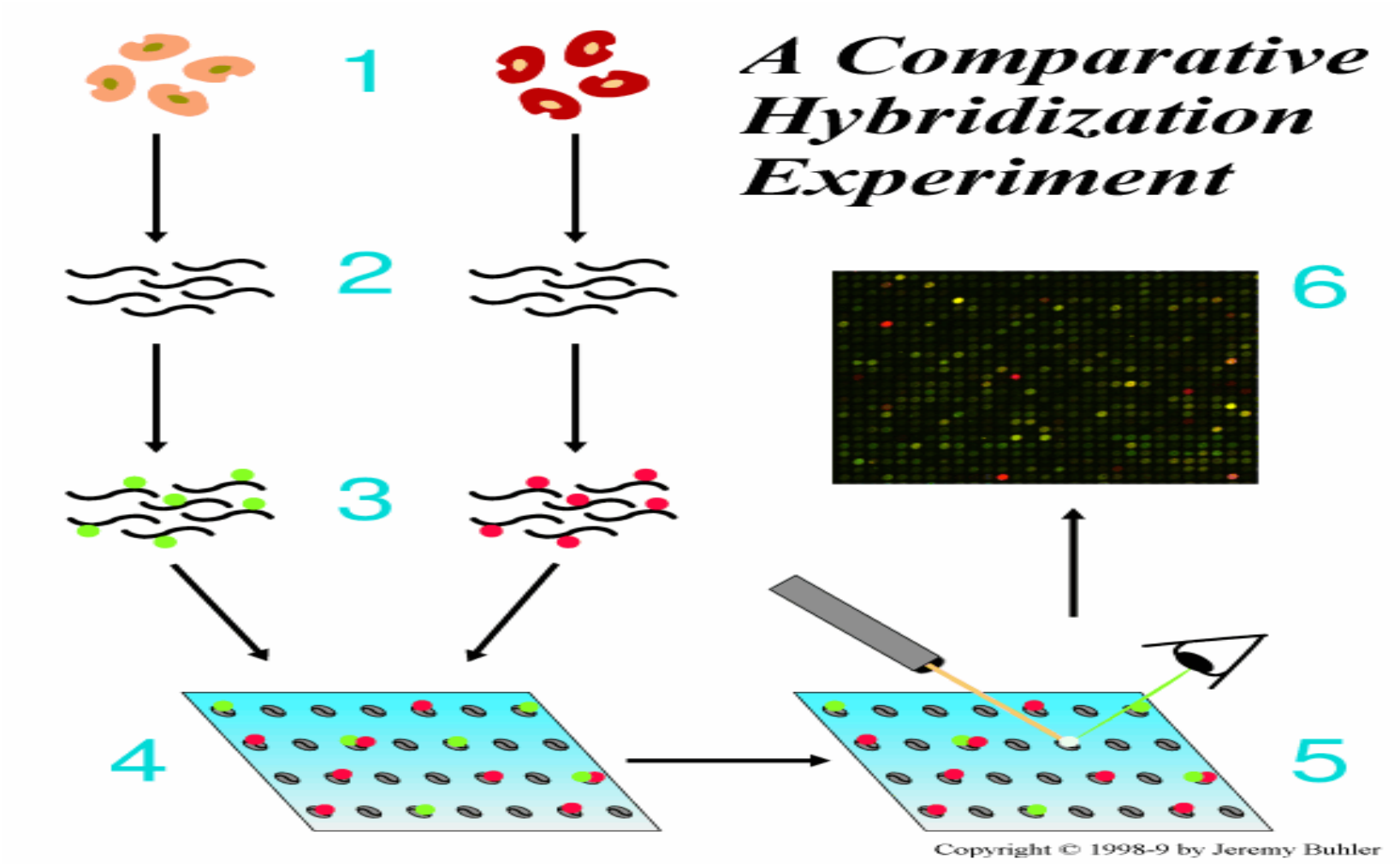
- Classify cancer using gene expression data by SVM
- SVM tools: SVM-light (C++) or Weka (Java)
- Reference: Golub et al, Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science, 1999.

Current Cancer Diagnosis

- A reliable and precise classification of tumors is essential for successful treatment of cancer.
- Current methods relies on the subjective interpretation of both clinical histopathological information with an eye toward placing tumors in currently accepted categories based on the tissue of origin of the tumor.
- However, clinical information can be misleading or incomplete.
- there is a wide spectrum in cancer morphology and many tumors are atypical or lack morphologic features, which results in diagnostic confusion.

Jia Yi, 2005

Typical DNA Microarray Experiment



DNA Microarray-based Cancer Diagnosis

- Molecular diagnostics offer the promise of precise, objective, and systematic cancer classification
- Recently, DNA microarray tumor gene expression profiles have been used for cancer diagnosis.
- By allowing the monitoring of expression levels for thousands of genes simultaneously, such techniques will lead to a more complete understanding of the molecular variations among tumors, hence to a finer and more reliable classification.

Tumor Classification Types

- There are three main types of statistical problems associated with tumor classification:
  - The identification of new tumor classes using gene expression profiles --- unsupervised learning.
  - The classification of malignancies into known classes --- supervised learning.
  - The identifications of “marker” genes that characterize the different tumor classes --- variable selection.

Source of Datasets (cont.)

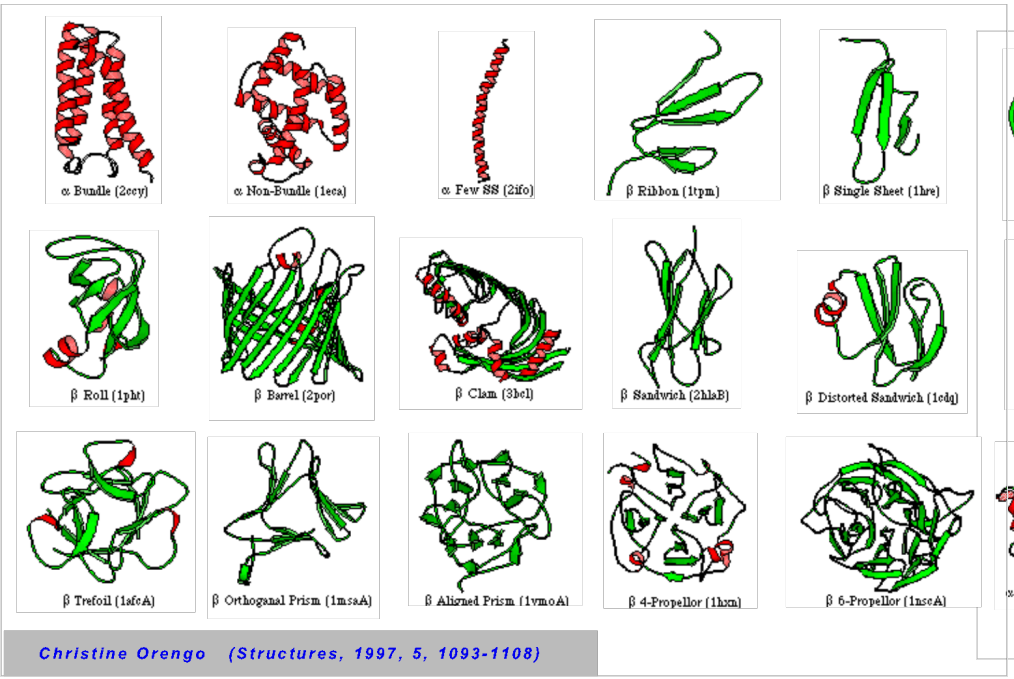
- Leukemia dataset
  - This dataset is the gene expression in two types of acute leukemias: ALL and AML.
  - This study produced gene expression data for p=6,817 genes in n=72 mRNA samples.

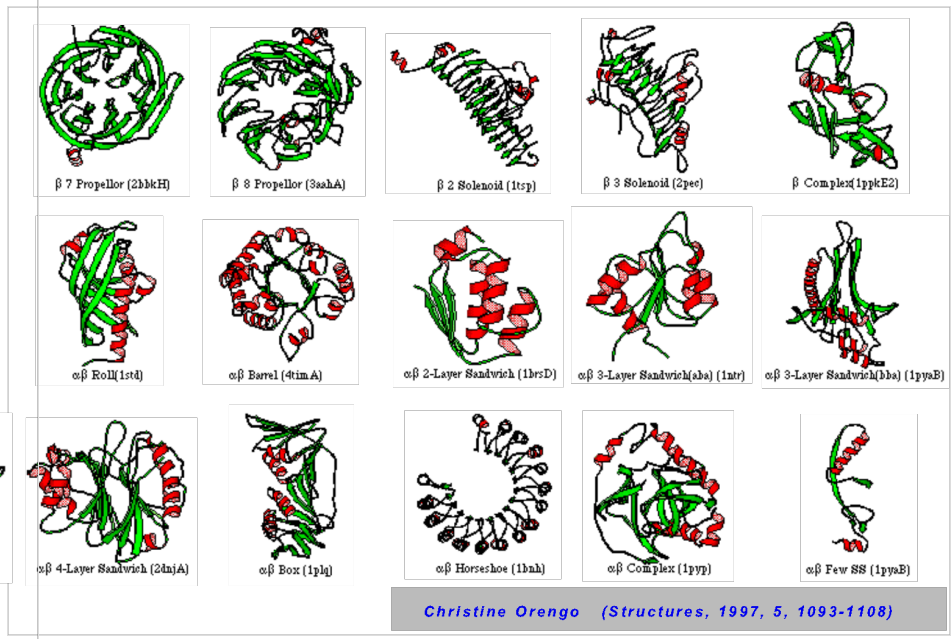
    47 × ALL (38 B-cell All,9 T-cell All)

    25 × AML

http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

The Universe of Protein Structures



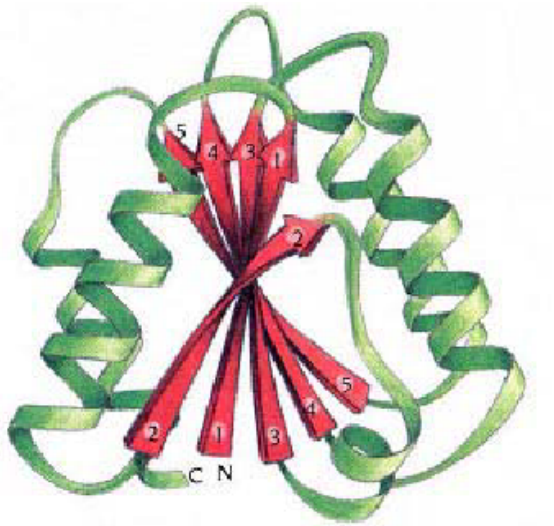


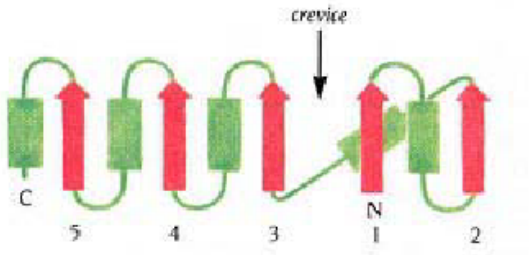
Christine Orengo **1997** *Structures* **5** 1093-1108

B. Rost, 2005

Typical Folds

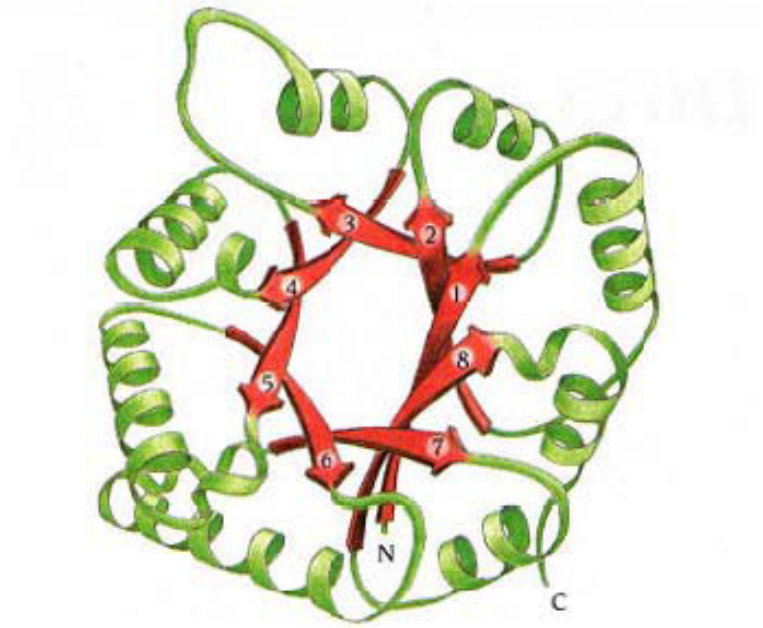
- Fold: connectivity or arrangement of secondary structure elements.
- NAD-binding Rossman fold
- 3 layers, a/b/a, parallel beta-sheet of 3 strands. Order: 321456



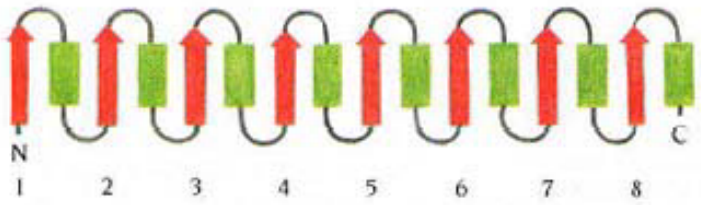


http://scop.berkeley.edu/rsgen.cgi?chime=1;pd=1b16;pc=a

Fold: TIM Beta-Alpha Barrel

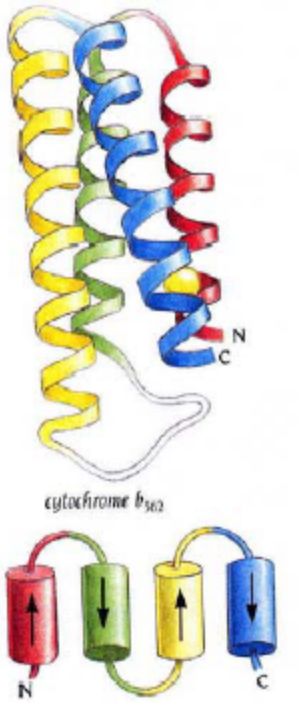


Contains parallel beta-sheet Barrel, closed. 8 strands. Order 1,2,3,4,5,6,7,8.



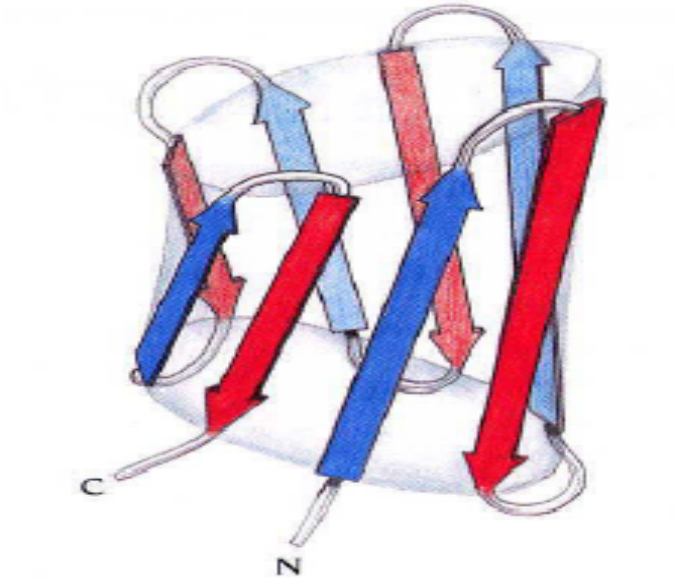
http://scop.berkeley.edu/rsgen.cgi?chime=1;pd=1hti;pc=a

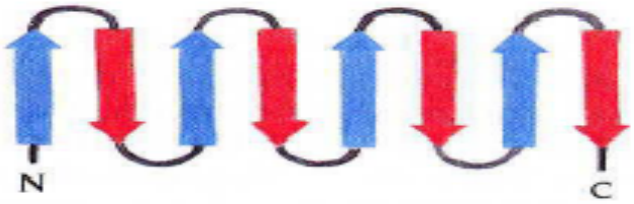
Helix Bundle (Human Growth Factor)



http://scop.berkeley.edu/rsgen.cgi?chime=1;pd=1hgu

Fold: Beta Barrel



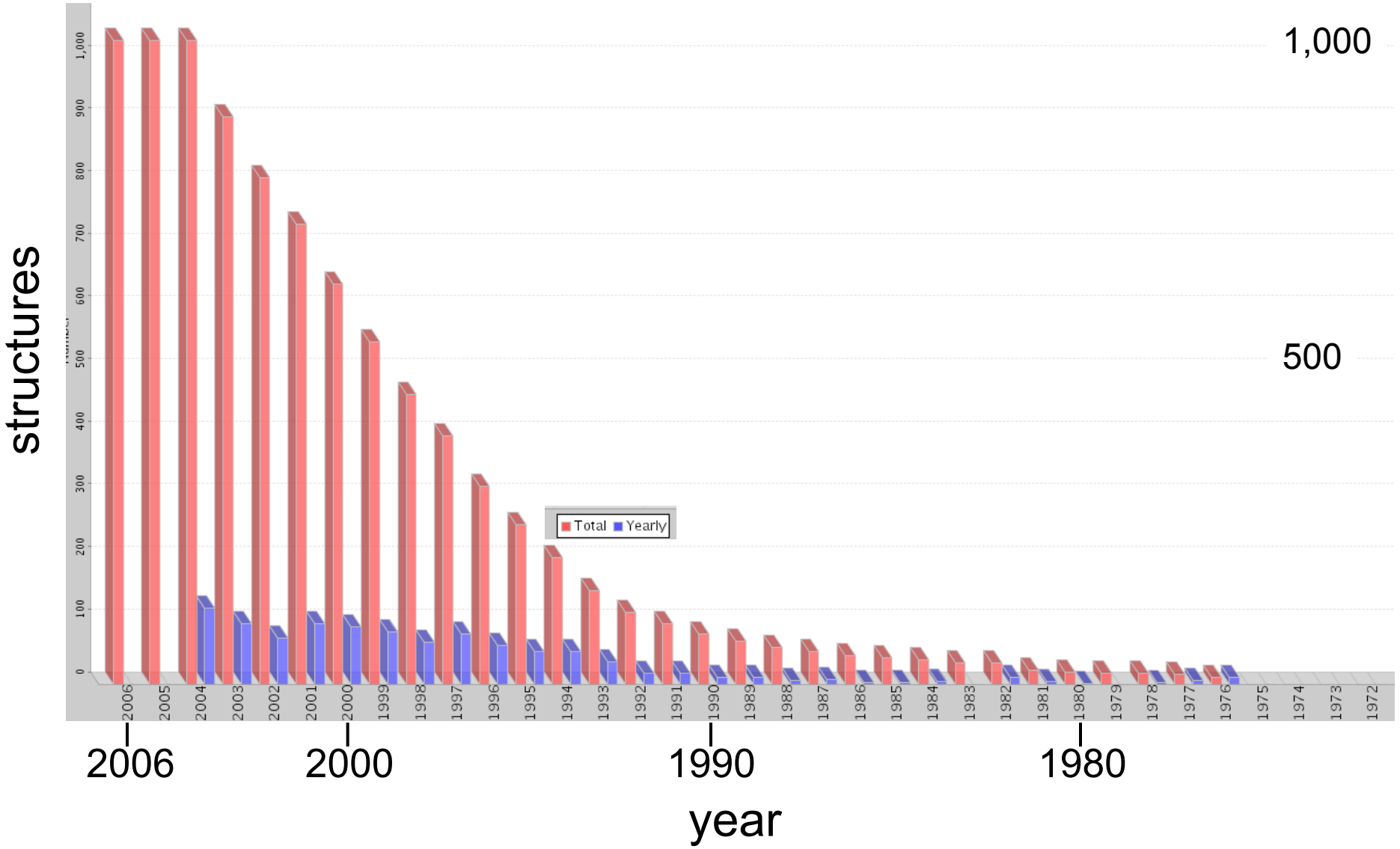


http://scop.berkeley.edu/rsgen.cgi?chime=1;pd=1rbp

Fold: Lamda Repressor DNA Binding

http://scop.berkeley.edu/rsgen.cgi?chime=1;pd=5cro;pc=0

Number of Unique Folds (defined by SCOP) in PDB

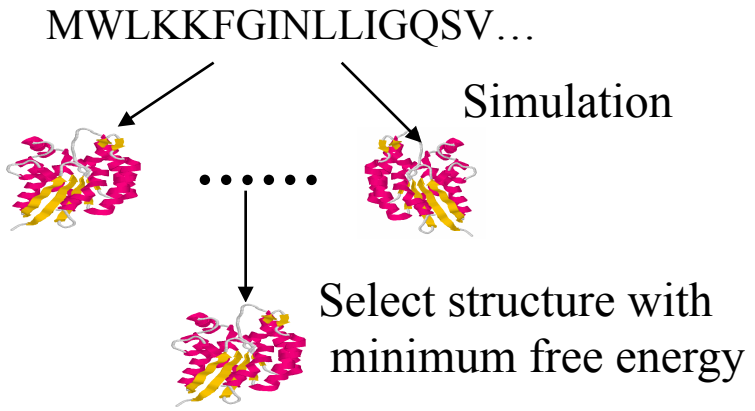


J. Pevsner, 2005

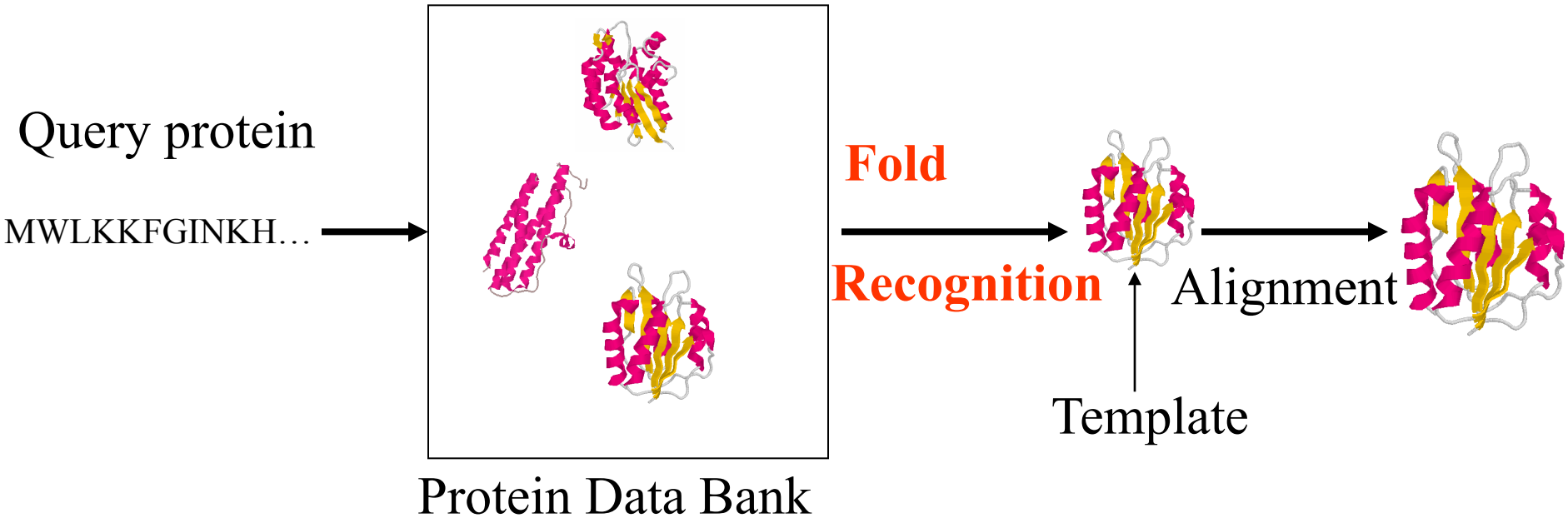
3D Structure Prediction

- **Ab-Initio Structure Prediction**

  Physical force field – protein folding
  Contact map - reconstruction



- **Template-Based Structure Prediction**



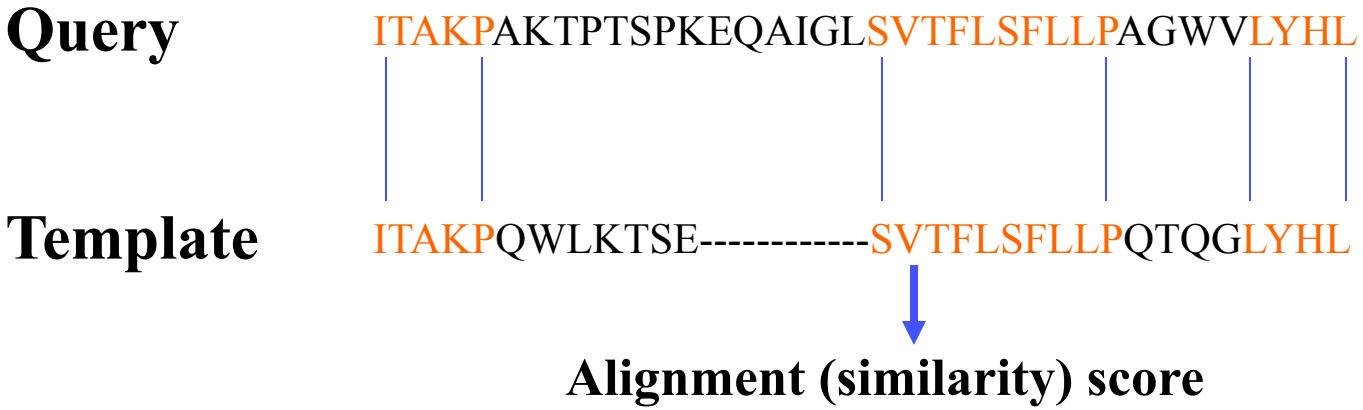
Template-Based Structure Prediction

1. **Template Identification**
2. **Query-Template Alignment**
3. Model Generation (Modeller, Sali and Blundell, 1993)
4. Model Evaluation
5. Model Refinement

Classic Fold Recognition Approaches

Sequence - Sequence Alignment

(Needleman and Wunsch, 1970. Smith and Waterman, 1981)

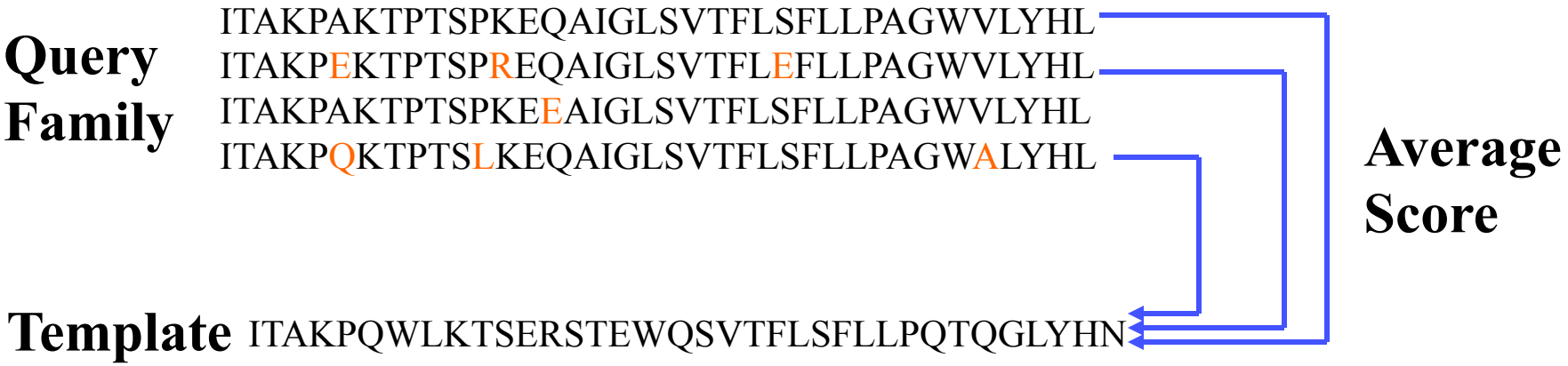


Works for >40% sequence identity
(Close homologs in protein family)

Classic Fold Recognition Approaches

Profile - Sequence Alignment

(Altschul et al., 1997)



More sensitive for distant homologs in superfamily.
(> 25% identity)

Classic Fold Recognition Approaches

Profile - Sequence Alignment

(Altschul et al., 1997)

Query Family

```
12.....n
ITAKPAKTPTSPKEQAIGLSVTFLSFLLPAGWVLYHL
ITAKPEKTPTSPREQAIGLSVTFLEFLLPAGWVLYHL
ITAKPAKTPTSPKEEAIGLSVTFLSFLLPAGWVLYHL
ITAKPQKTPTSLKEQAIGLSVTFLSFLLPAGWALYHL
```

	1	2	…	n
A	0.4			
C	0.1			
…				
W	0.5			

Position Specific Scoring Matrix Or Hidden Markov Model

Template ITAKPQWLKTSERSTEWQSVTFLSFLLPQTQGLYHN

More sensitive for distant homologs in superfamily.
(> 25% identity)

Classic Fold Recognition Approaches

Profile - Profile Alignment

(Rychlewski et al., 2000)

Query Family

ITAKPAKTPTSPKEQAIGLSVTFLSFLLPAGWVLYHL
ITAKPEKTPTSPREQAIGLSVTFLEFLLPAGWVLYHL
ILAKPAKTPTSPKEEAIGLSVTFLSFLLPAGWVLYHL
ITAKPQKTPTSLKEQAIGLSVTFLSFLLPAGWALYHL

	1	2	…	n
A	0.1			
C	0.4			
…				
W	0.5			

Template Family

ITAKPQWLKTSERSTEWQSVTFLSFLLPQTQGLYHN
IPARPQWLKTSKRSTEWQSVTFLSFLLPYTQGLYHN
IGAKPQWLWTSERSTEWHSVTFLSFLLPQTQGLYHM

	1	2	…	m
A	0.3			
C	0.5			
…				
W	0.2			

More sensitive for very distant homologs.
(> 15% identity)

Classic Fold Recognition Approaches

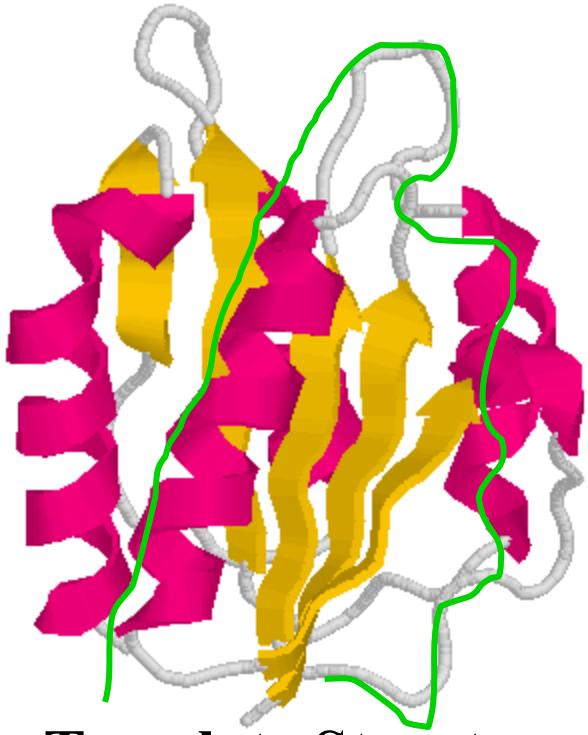
Sequence - Structure Alignment (Threading)

(Bowie et al., 1991. Jones et al., 1992. Godzik, Skolnick, 1992. Lathrop, 1994)

Query

MWLKKFGINLLIGQS….

Fit

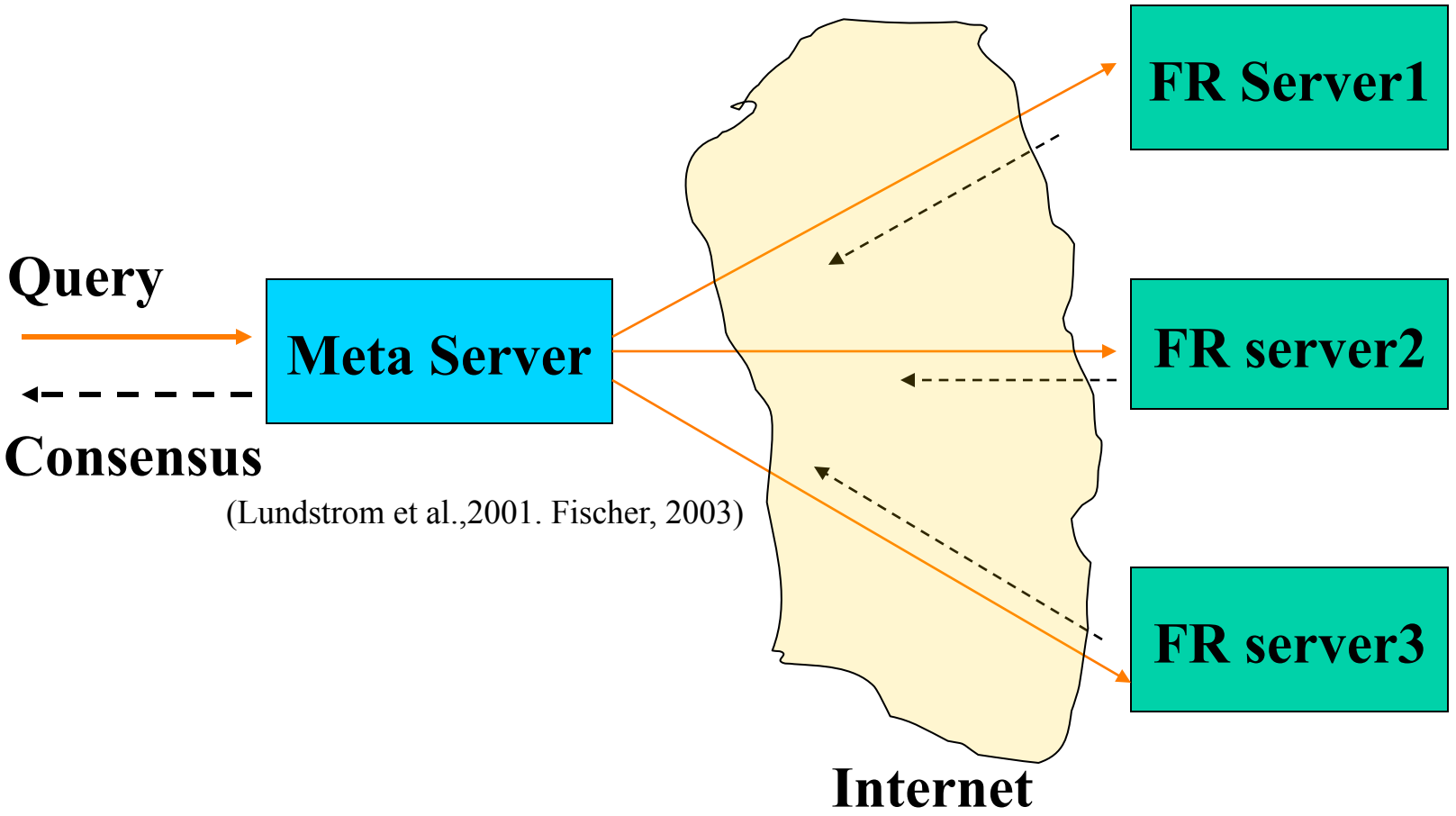


Template Structure

Fitness Score

Useful for recognizing similar folds without sequence similarity. (no evolutionary relationship)

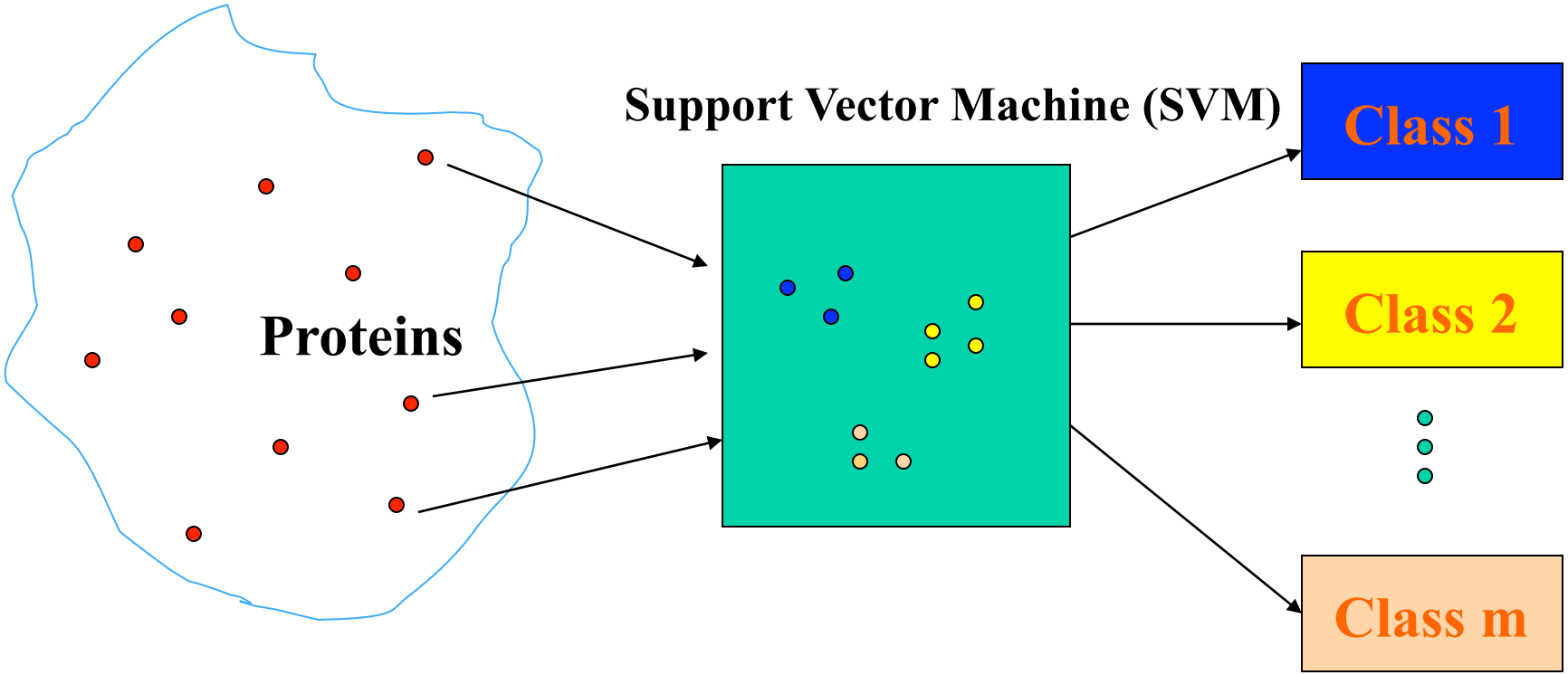
Integration of Complementary Approaches



(Lundstrom et al.,2001. Fischer, 2003)

1. **Reliability depends on availability of external servers**
2. **Make decisions on a handful candidates**

Machine Learning Classification Approach



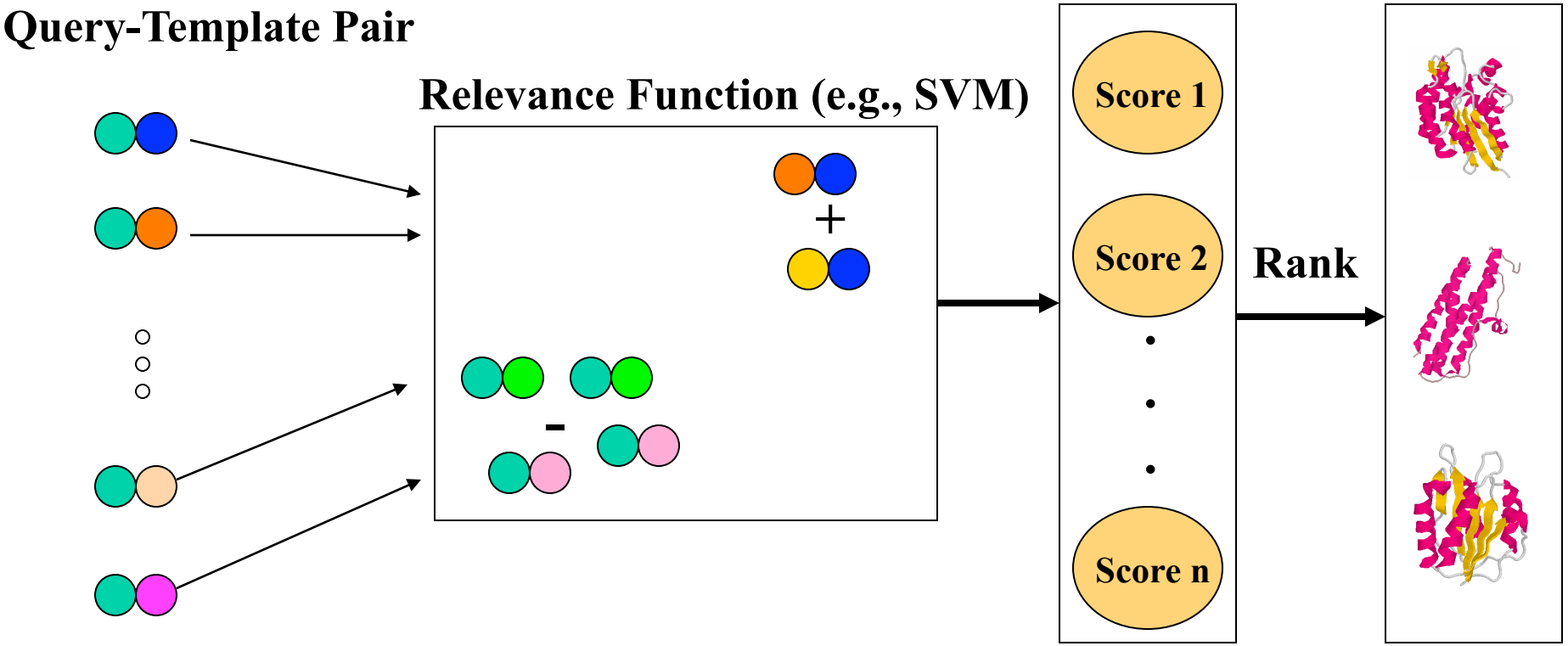
Classify individual proteins to several or dozens of structure classes

(Ding and Dubchk, 2001, Jaakkola et al., 2000. Leslie et al., 2002. Saigo et al., 2004, Rangwala and Karypis,2005)

Problem 1: can't scale up to thousands of protein classes

Problem 2: doesn't provide templates for structure modeling

Machine Learning Information Retrieval Framework



- **Extract pairwise features**
- **Comparison of two pairs (four proteins)**
- **Relevant or not (one score) vs. many classes**
- **Ranking of templates (retrieval)**

Cheng and Baldi, Bioinformatics, 2006

Pairwise Feature Extraction

- **Sequence / Family Information Features**

  Cosine, correlation, and Gaussian kernel
- **Sequence – Sequence Alignment Features**

  Palign, ClustalW
- **Sequence – Profile Alignment Features**

  PSI-BLAST, IMPALA, HMMer, RPS-BLAST
- **Profile – Profile Alignment Features**

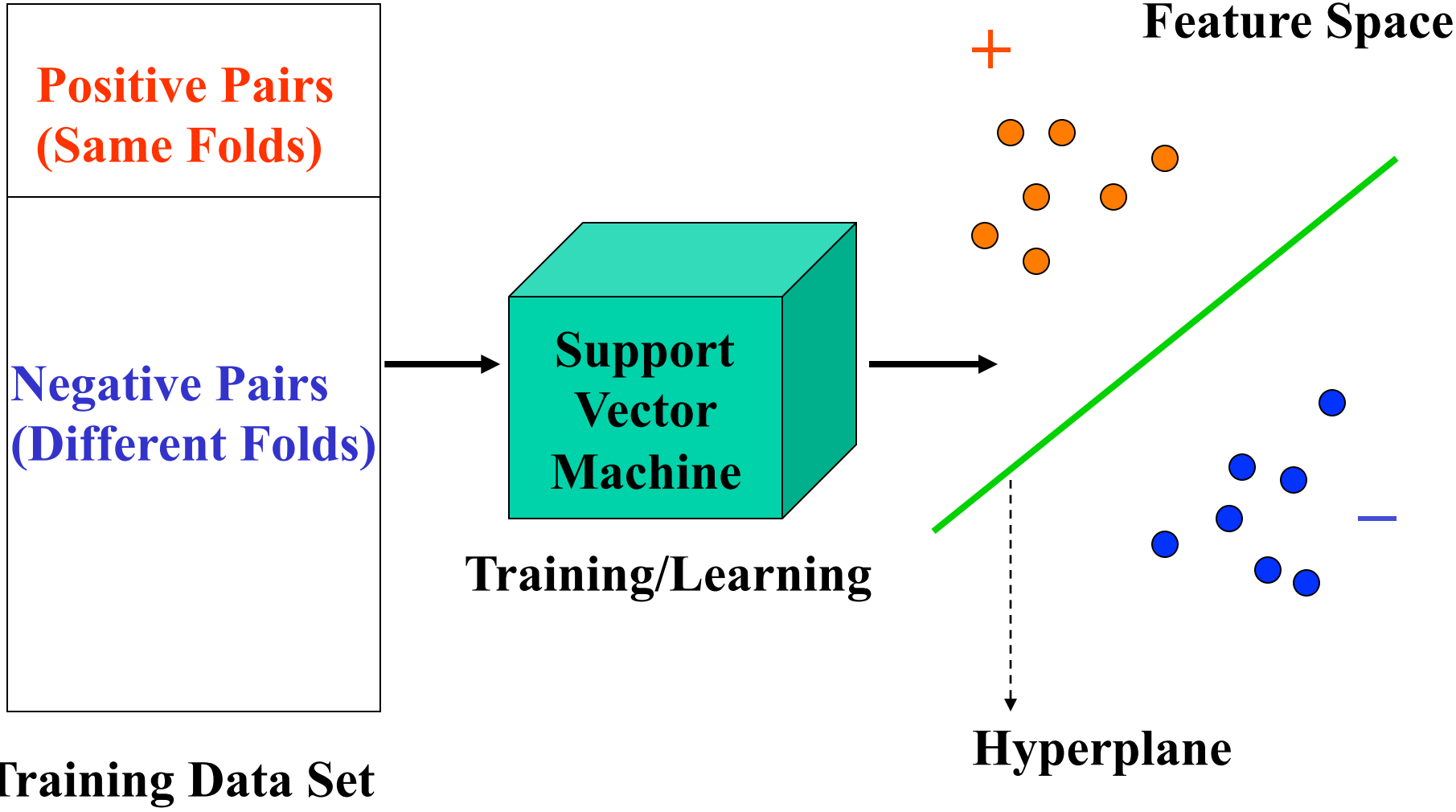
  ClustalW, HHSearch, Lobster, Compass, PRC-HMM
- **Structural Features**

  Secondary structure, solvent accessibility, contact map, beta-sheet topology

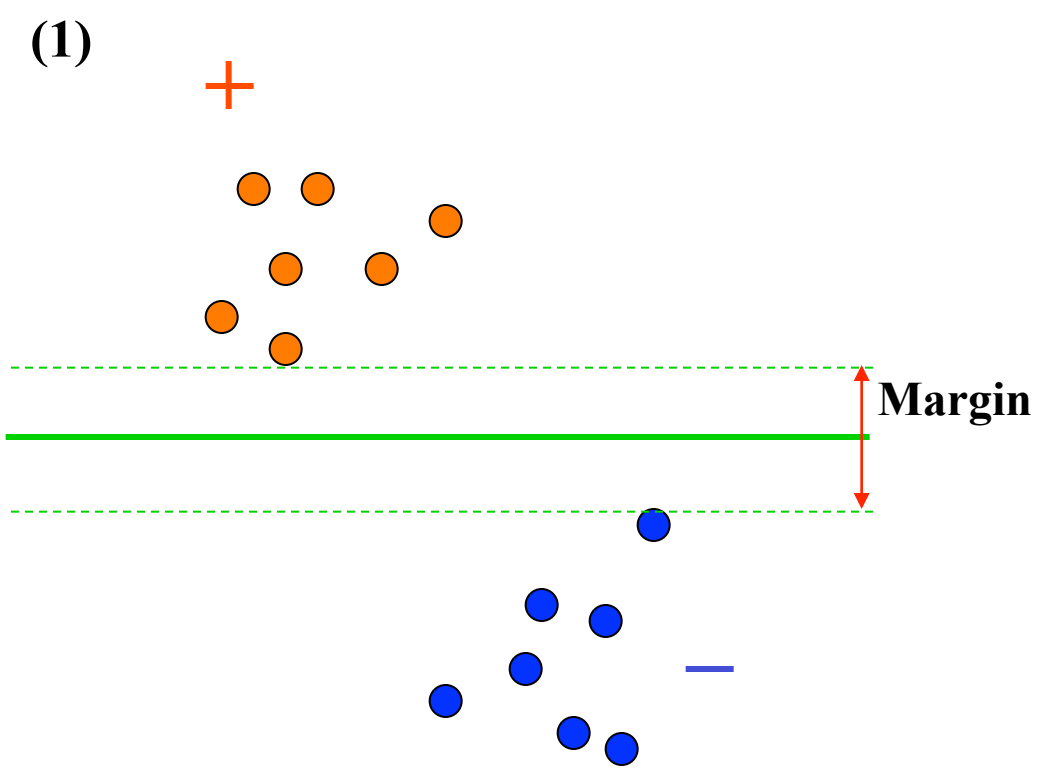
Top Ranked Features

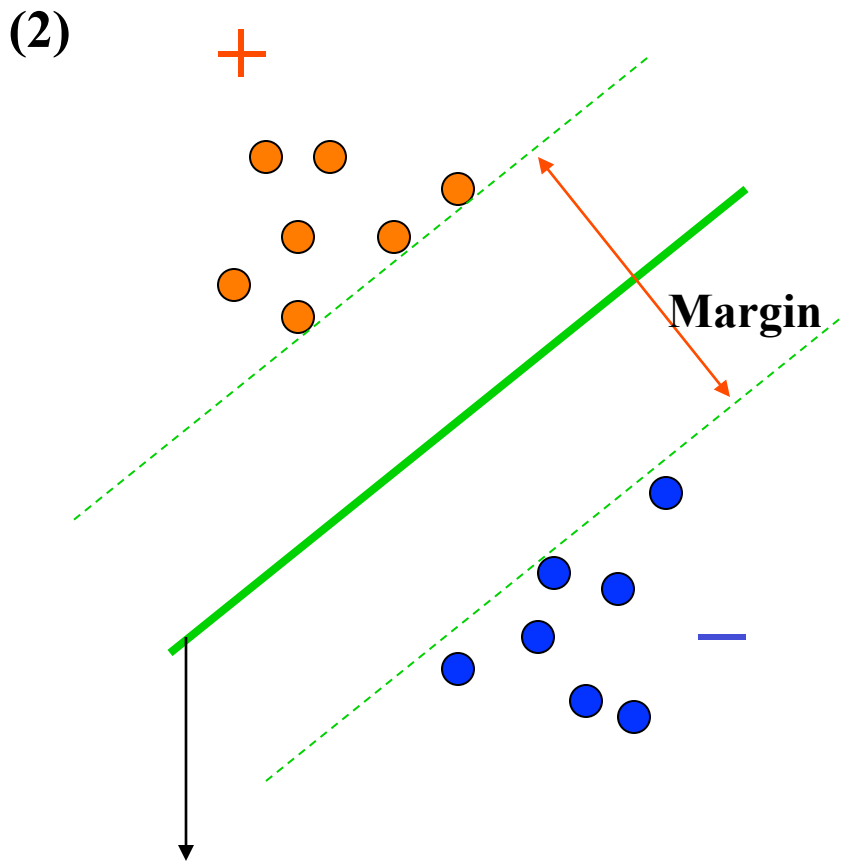
Feature	Information gain
HHSearch score	0.0375
COMPASS *e*-value	0.0370
PRC reverse score on chk profile	0.0354
PRC reverse score on HMM profile	0.0341
HMMer pfam *e*-value	0.0287
Dot product of SS and RSA vectors	0.0266
HMMer search *e*-value	0.0264
SS match ratio	0.0263
Correlation of SS and RSA vectors	0.0263
PRC simple score on HMM profile	0.0248
Cosine of SS and RSA vectors	0.0246
Gaussian kernel on SS and RSA vectors	0.0237
COMPASS score	0.0235
PRC coemis score on HMM profile	0.022
PSI-BLAST *e*-value	0.0205
IMPALA *e*-value	0.0181
RPS-BLAST *e*-value	0.0180
SA match ratio	0.0154
Cosine of residue contact num (8 Å)	0.0150
HMMer search score	0.0142

Relevance Function: Support Vector Machine Learning



Relevance Function: Support Vector Machine Learning



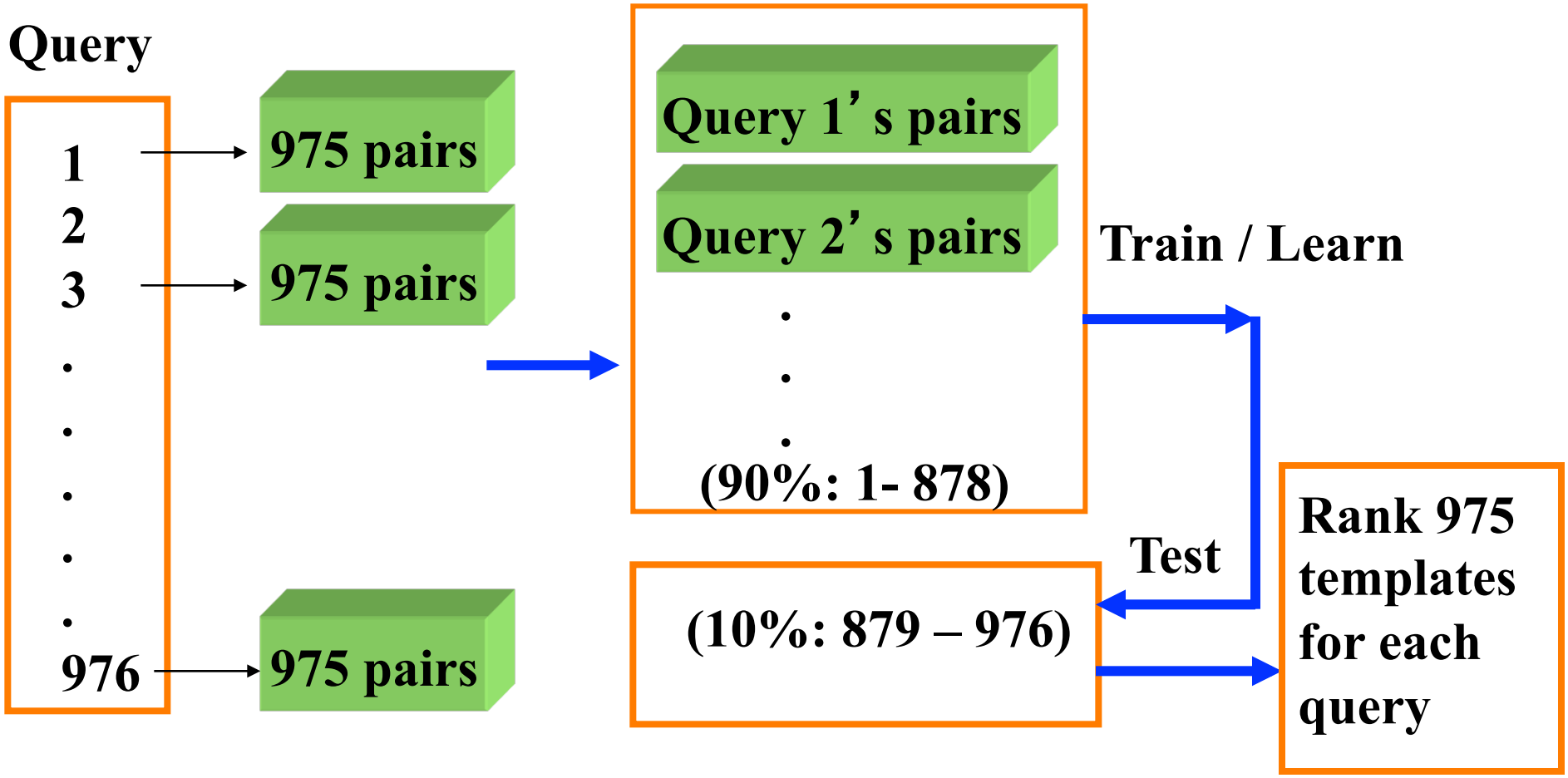


$$f(x) = \sum_{x_i \in S} \alpha_i y_i K(x, x_i) + b$$

K is Gaussian Kernel: $e^{-\gamma ||x-y||^2}$

Training and Cross-Validation

- Standard benchmark (Lindahl's dataset, 976 proteins)
- 976 x 975 query-template pairs (about 7,468 positives)



Results for Top Five Ranked Templates

Method	Family	Superfamily	Fold
PSI-BLAST	72.3	27.9	4.7
HMMER	73.5	31.3	14.6
SAM-T98	75.4	38.9	18.7
BLASTLINK	78.9	4.06	16.5
SSEARCH	75.5	32.5	15.6
SSHMM	71.7	31.6	24
THREADER	58.9	24.7	37.7
FUGUE	85.8	53.2	26.8
RAPTOR	77.8	50	45.1
SPARKS3	86.8	67.7	47.4
FOLDpro	**89.9**	**70.0**	**48.3**

- **Family**: close homologs, more identity
- **Superfamily**: distant homologs, less identity
- **Fold**: no evolutionary relation, no identity

Advantages of MLIR Framework

- Integration, Accuracy, Extensibility
- Simplicity, Completeness, Potentials

Disadvantages

- Slower than some alignment methods

Challenge: analogous fold recognition using machine learning ranking techniques

Project 6: Fold Recognition

- Data splits: fold1, fold2, …, fold10 (www.cs.missouri.edu/~chengji/mlbioinfo/folds.tar.gz)
- Do 10 fold cross-validation using SVM (report the classification accuracy)
- Report the fold recognition results (check if a positive template is ranked as top one, when there is a positive template exists for a query)

TARGET

TEMPLATE

ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDD
VPERLIPERASFQWMNDK

ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE

A. Fisher, 2005

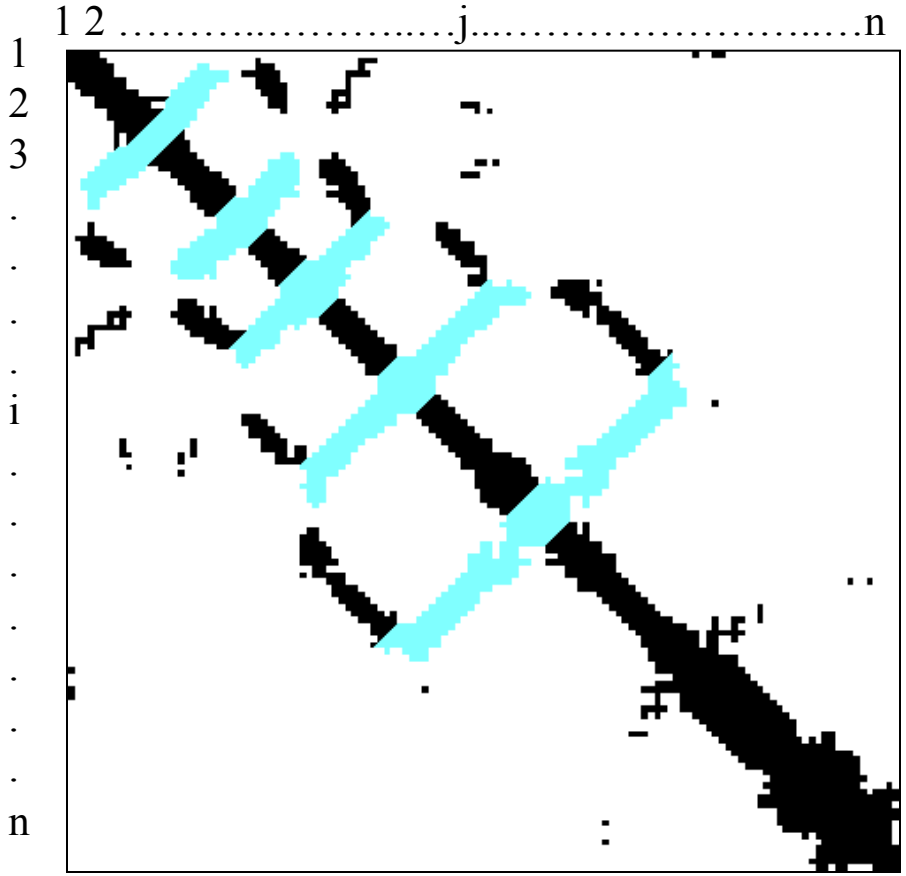
2D: Contact Map Prediction

3D Structure

2D-Recursive Neural Network

Support Vector Machine

2D Contact Map



Distance Threshold = 8A°

Cheng, Randall, Sweredoski, Baldi. *Nucleic Acid Research*, 2005
Cheng and Baldi. *BMC Bioinformatics*, 2007.

Definition of Contact Prediction

- Predict if any two residues i, j are in contact or not according to a distance threshold (8 Angstrom)
- Interested in short to long range contacts (|i-j| >= 6)
- Use a window (size = 9) to encode the information about residue i and j, respectively
- Train on a training dataset and test on a test dataset.

Feature Extraction

- Local window features (20 * 9 * 2)
- Pairwise information feature (cosine, correlation, mutual information)
- Residue type feature (non-polar, polar, acidic, and basic)
- Central segment window features
- Protein information features (global composition, sequence length)

Kernel and Feature Selection

- Gaussion kernel seems to work well. However, we haven't tested other kernels thoroughly
- Feature selection should be able to improve the performance. However, we haven't conducted a thorough feature selection yet due to the limited computing power.

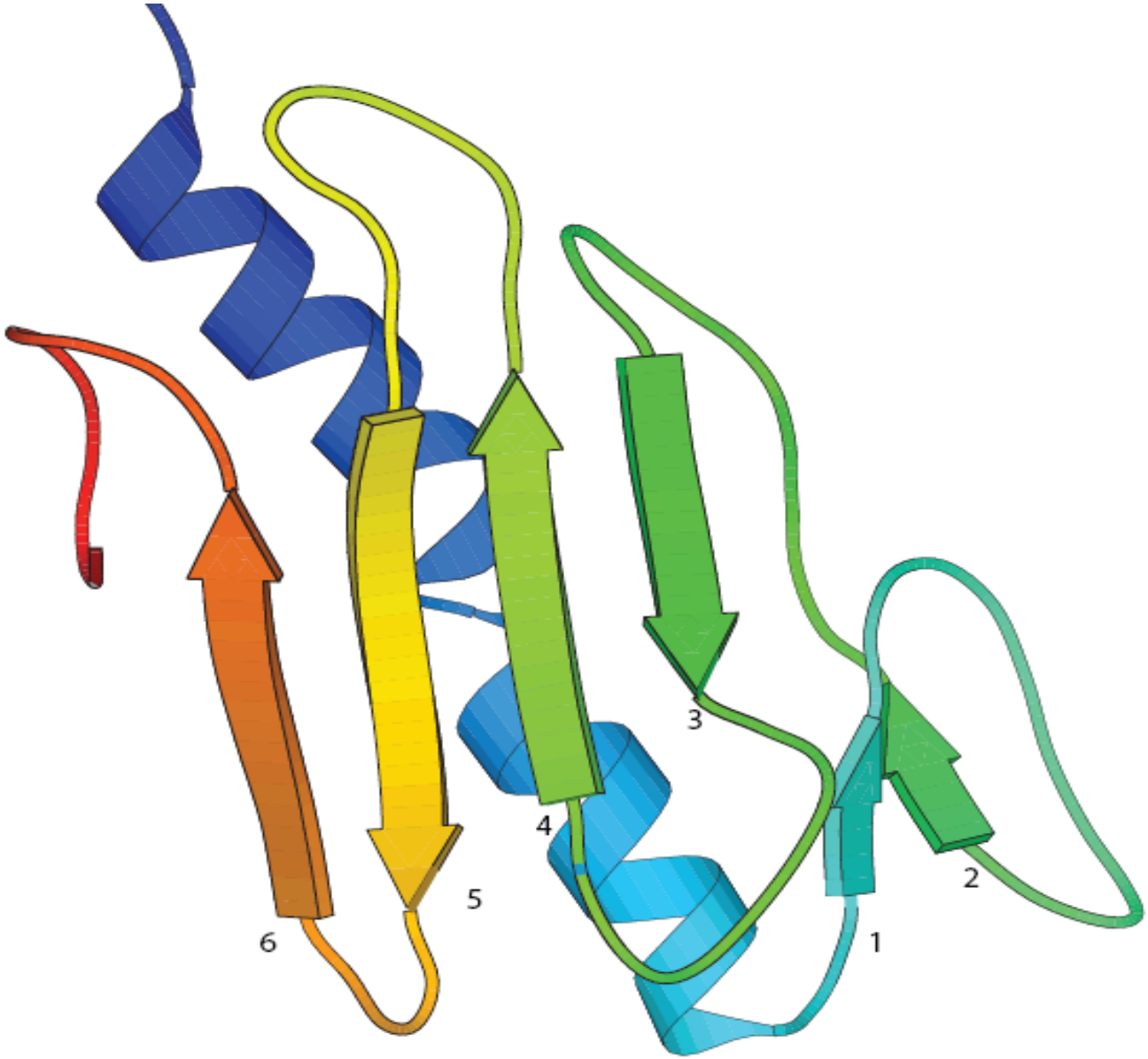
Results

- At break-even point, the sensitivity = specificity = **28%**
- However, the accuracy varies according to the property of the individual proteins significantly.
- Contacts within beta-sheet is predicted with higher accuracy than that in alpha helices or between alpha helix and beta-sheet.

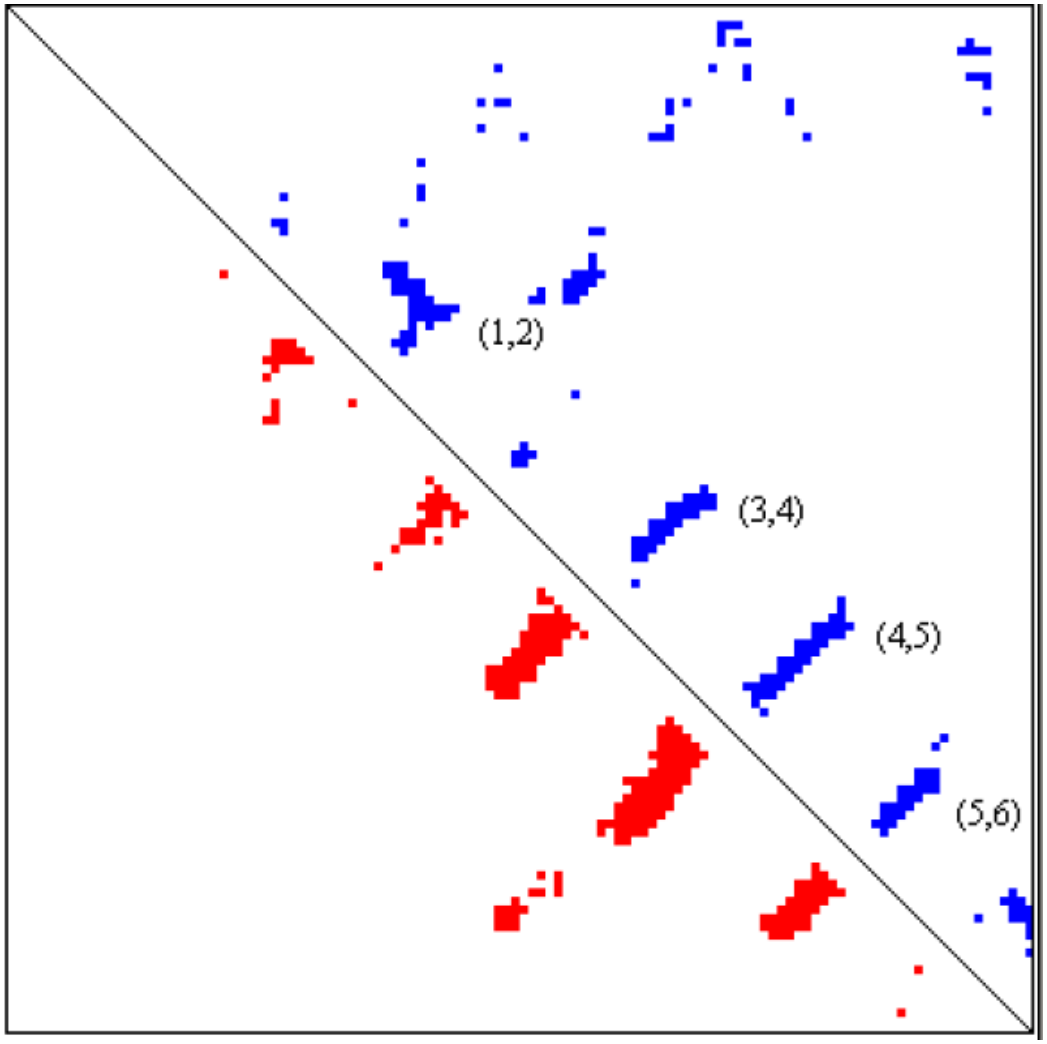
SCOP Class	Num	Separation >= 6		Separation >= 12		Separation >= 24	
		Accuracy	Coverage	Accuracy	Coverage	Accuracy	Coverage
alpha	11	0.24	0.24	0.17	0.18	0.11	0.09
beta	10	0.38	0.17	0.32	0.17	0.22	0.17
a+b	15	0.45	0.25	0.35	0.25	0.21	0.23
a/b	7	0.37	0.19	0.33	0.19	0.28	0.20
small	4	0.36	0.18	0.28	0.19	0.11	0.15
coil-coil	1	0.22	0.40	0.03	0.16	0.00	—
average	48	0.37	0.21	0.30	0.20	0.21	0.19

Results on 48 test proteins

One Example



Predicted VS True Contacts



How to Use Contacts to Reconstruct 3D Structure

- 3D structure prediction problem can be defined as a constrained optimization problem.
- Generate a 3D structure with minimum free energy subject to contact restraints and intrinsic biophysical constraints such as bond length.

Optimization Techniques

- Gradient Descent (Modeller)
- Lattice Monte Carlo Sampling (TASSER)
- Simulated Annealing (Rosetta)
- Multi-Dimensional Scaling
- Many others….

Contact Prediction Software

- www.cs.missouri.edu/~chengji/cheng_software.html (svmcon 1.0, source code and executable)
- SVMcon paper: www.cs.missouri.edu/~chengji/cheng_publication.html
- Reference: J. Cheng and Baldi. Improved residue contact prediction using support vector machine and a large feature set. *BMC Bioinformatics*, 2007.