Analysis and Prediction of Protein Structure (II)

Jianlin Cheng, PhD

School of Electrical Engineering and Computer Science
University of Central Florida

2006

Outline

I. Sequence, Structure, Function Relation

II. Determination, Storage, Visualization, and Comparison

III. Structure Classification

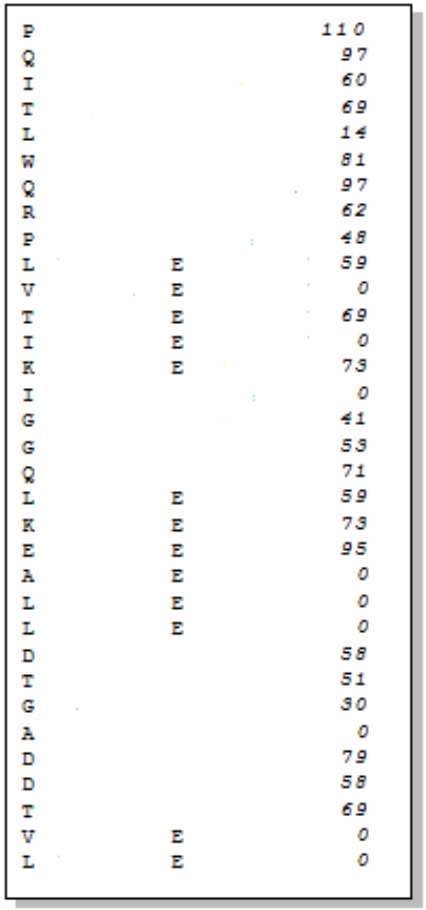
IV. 1D Prediction

V. 2D Prediction

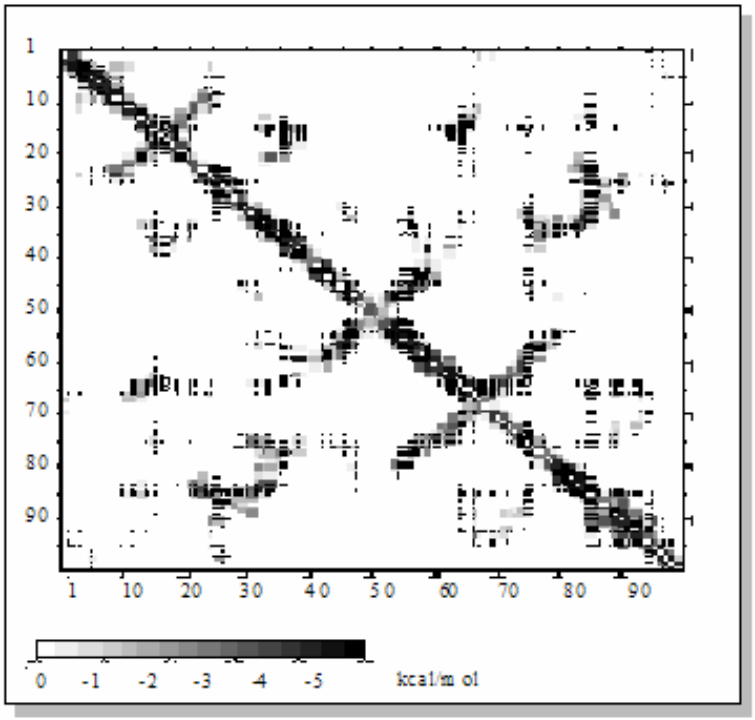
VI. 3D Prediction (emphasis)

VII. Tools and Applications

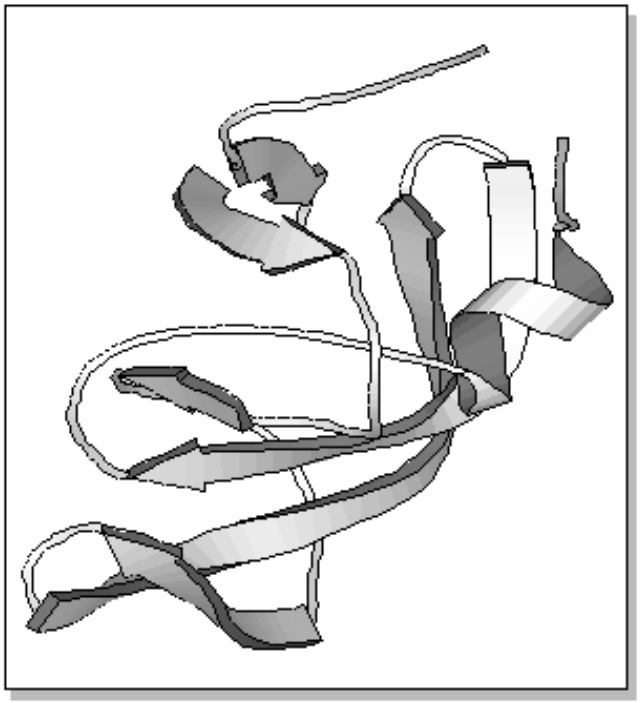
Notation: protein structure 1D, 2D, 3D



1D



2D



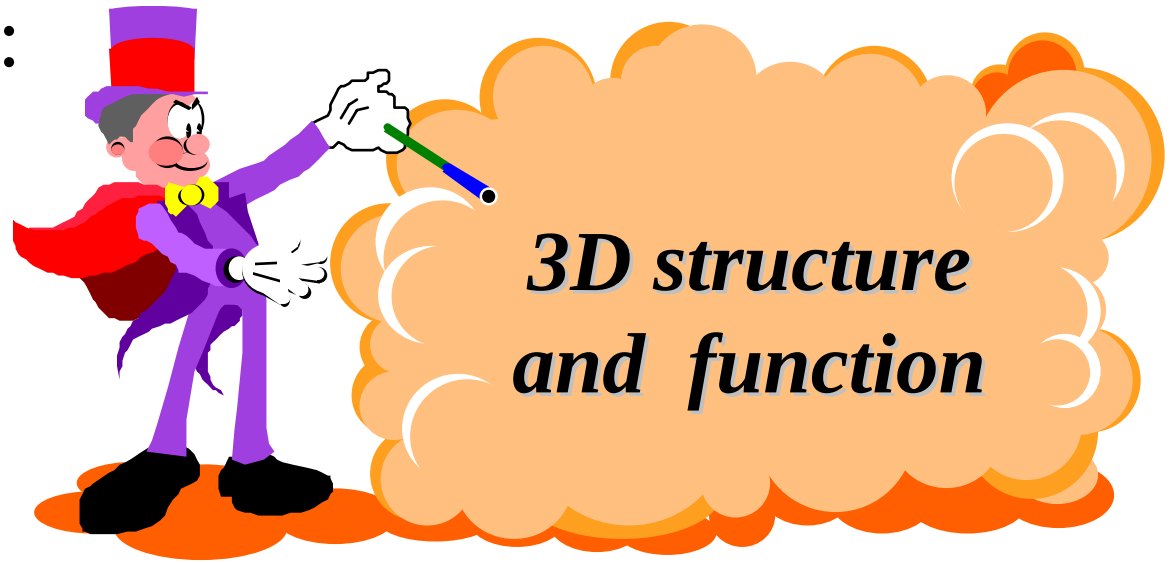
3D

B. Rost, 2005

Goal of structure prediction

- Epstein & Anfinsen, 1961:
  sequence uniquely determines structure

- INPUT: sequence
- OUTPUT:



B. Rost, 2005

CASP – Olympics of Protein Structure Prediction

- Critical Assessment of Techniques of Protein Structure Prediction
- 1994,1996,1998,2000,2002,2004,2006
- Blind Test, Independent Evaluation
- CASP7: 100 targets

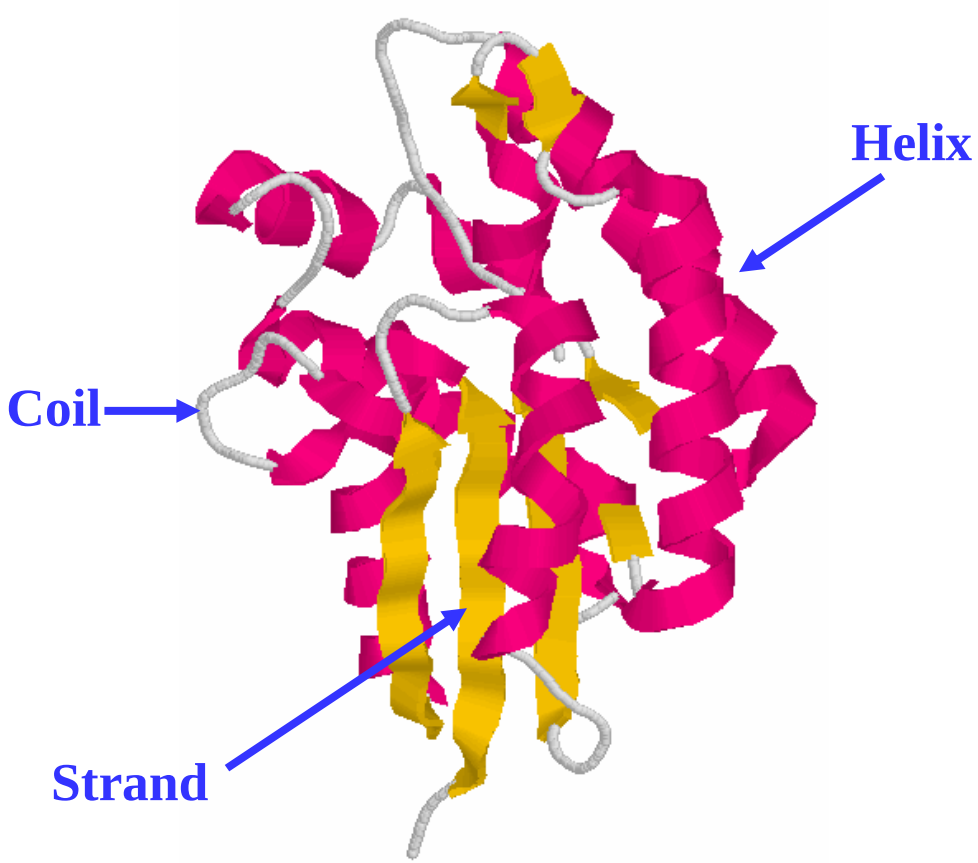
Announcement

- I will travel to CASP7 from Nov 25 to Dec. 1. (no class for the last week)
- Vote for final exam date (**Nov 29**)
- Project due date (**Dec 4**)
- Project organization (up to 3 persons in a group, your own project abstract by Oct 27).

1D Structure Prediction

- Predict the structural features of each residues along one dimensional sequence
- Secondary structure (detailed)
- Solvent accessibility
- Disordered regions
- Domain boundary

1D: Secondary Structure Prediction



MWLKKFGINLLIGQSV…

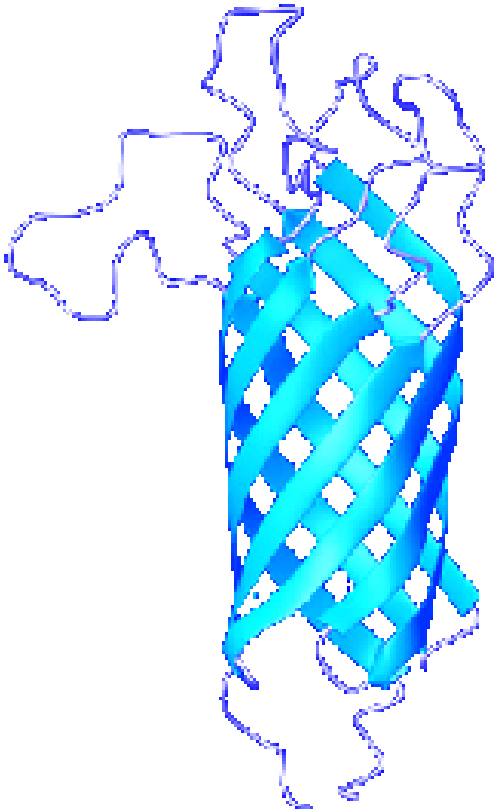
Neural Networks + Alignments

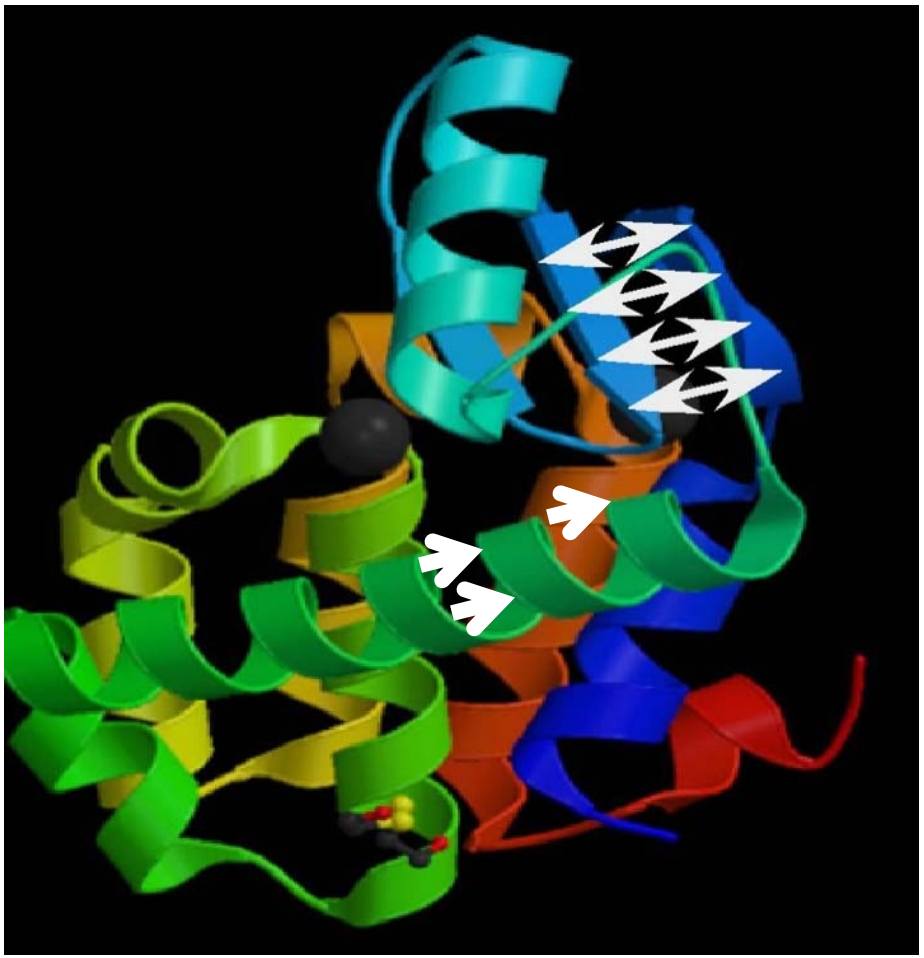
CCCCHHHHHCCCSSSSS…

Accuracy: 78%

Cheng, Randall, Sweredoski, Baldi. *Nucleic Acid Research*, 2005

Secondary structure



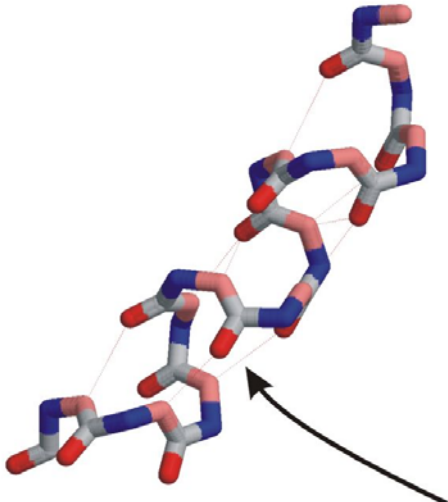


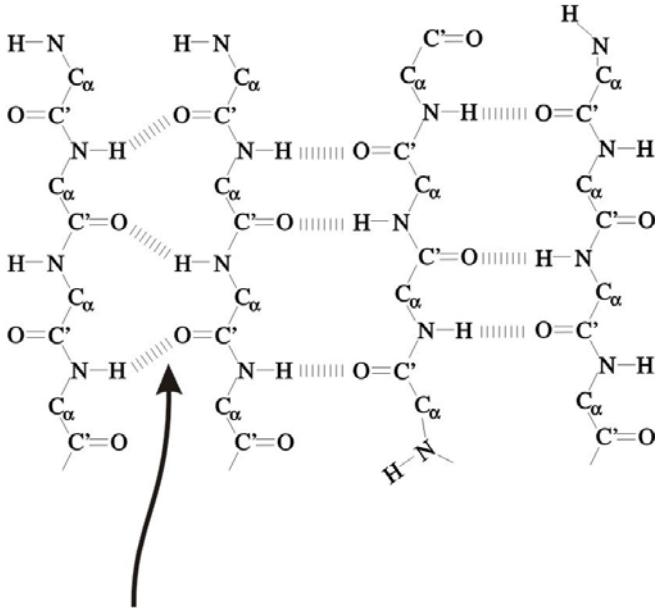
B. Rost, 2005

Linus Pauling's H-bond pattern used in DSSP

α-helix

β-sheets





DSSP $E_{HB} < -0.5$ kcal/mol

Pauling, L. & Corey, R. B. (1953) *PNAS* **39,** 247-252.
Pauling, L., Corey, R. B. & Branson, H. R. (1951) *PNAS.* **37,** 205-234.
DSSP: Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577-2637.

B. Rost, 2005

Homology-Based Secondary Structure Prediction

- Proteins in the same family are homologous proteins.
- Proteins in the same super family are distant homologs.
- Homology-based method tries to identify a homologous protein with known structure for a query protein and generate an alignment for them. Then it copies the secondary structure of homologous protein to the query protein.
- Homology-based method works well when a significant template (protein with known structure) is identified (e.g., for BLAST search, e-value < 10^{-50}).
- An accurate method that uses homology information is SSpro 4.0. (http://contact.ics.uci.edu/sspro4.html).

Ab-Initio Secondary Structure Prediction

- Predict secondary structure from sequence information without using any structural information
- Generation I (statistical methods)
- Generation II (statistical machine learning)
- Generation III (evolutionary information + statistical machine learning)

B. Rost, 2005

Secondary Structure Prediction (Generation 1)

Single residues (1. generation)

– Chou-Fasman, GOR 1957-70/80
50-55% accuracy

Secondary structure propensity score of an amino acid AA

$$Log\ \ P(\text{AA in Helix}) / P(\text{AA})$$
$$Log\ \ P(\text{AA in Sheet}) / P(\text{AA})$$
$$Log\ \ P(\text{AA in Loop}) / P(\text{AA})$$

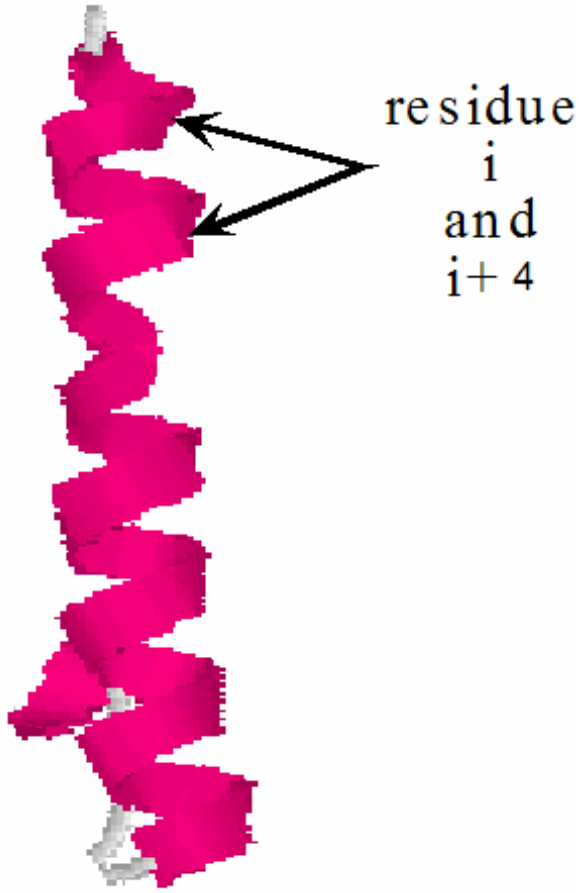
Secondary Structure Prediction (Generation II)

- Segments (window)
- GOR III
- Accuracy: 55-60%
- 1986-1992
- Estimation: max < 65%, Strand: non-local, < 40%

ARSKQPCTWYRE**S**QTCEQRKRTPA

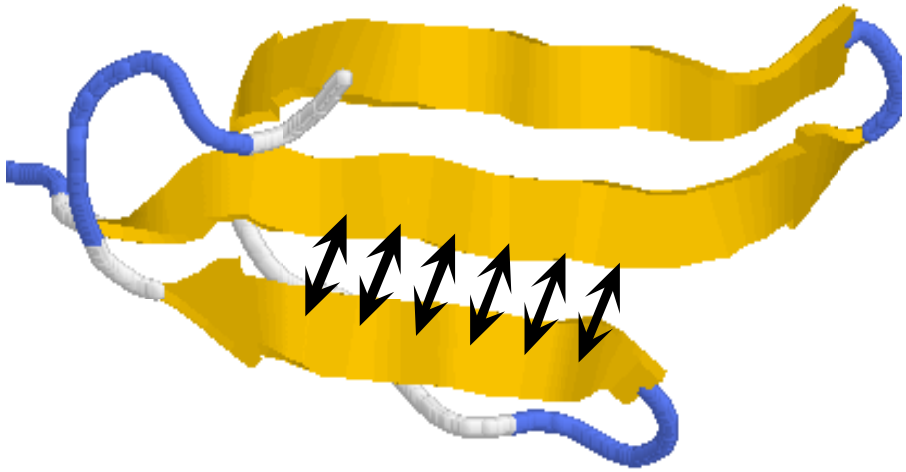
Average propensity scores of a window

Helix formation is local



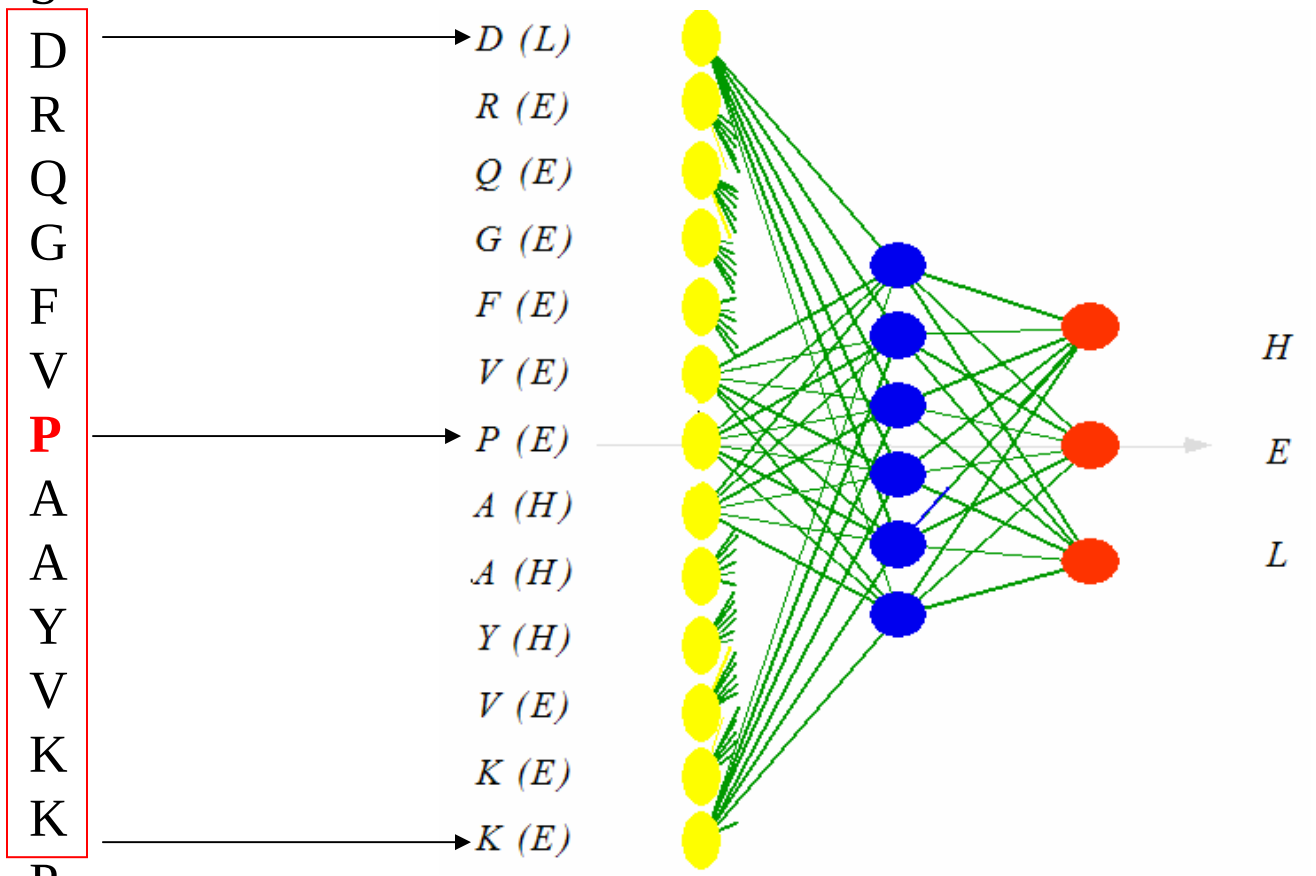
B. Rost, 2005

β-sheet formation is NOT local



Erabutoxinβ (3ebx)

Secondary Structure Prediction (Generation III – Neural Network)

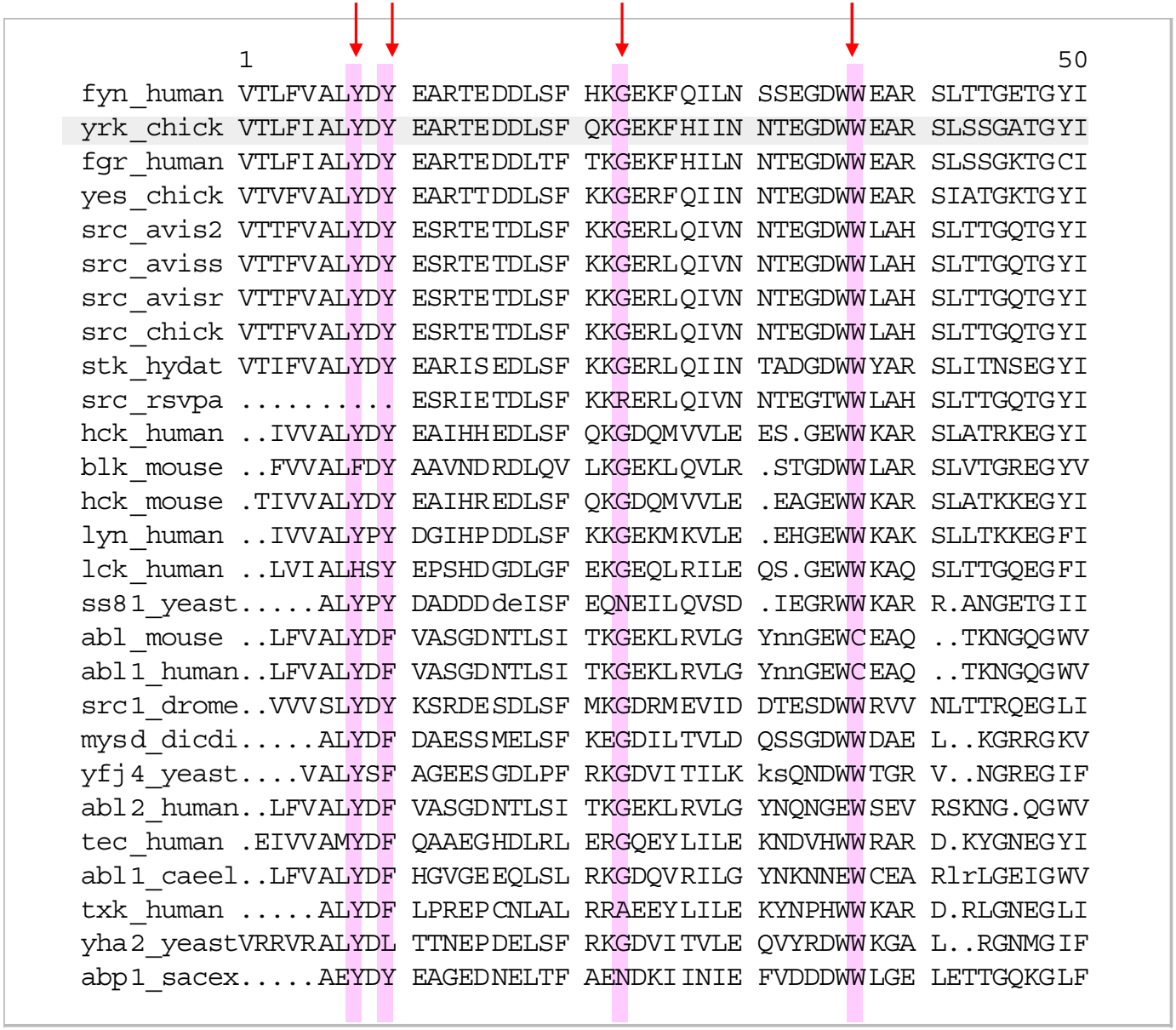


Output:
Prob (H)
Prob (E)
Prob (L)

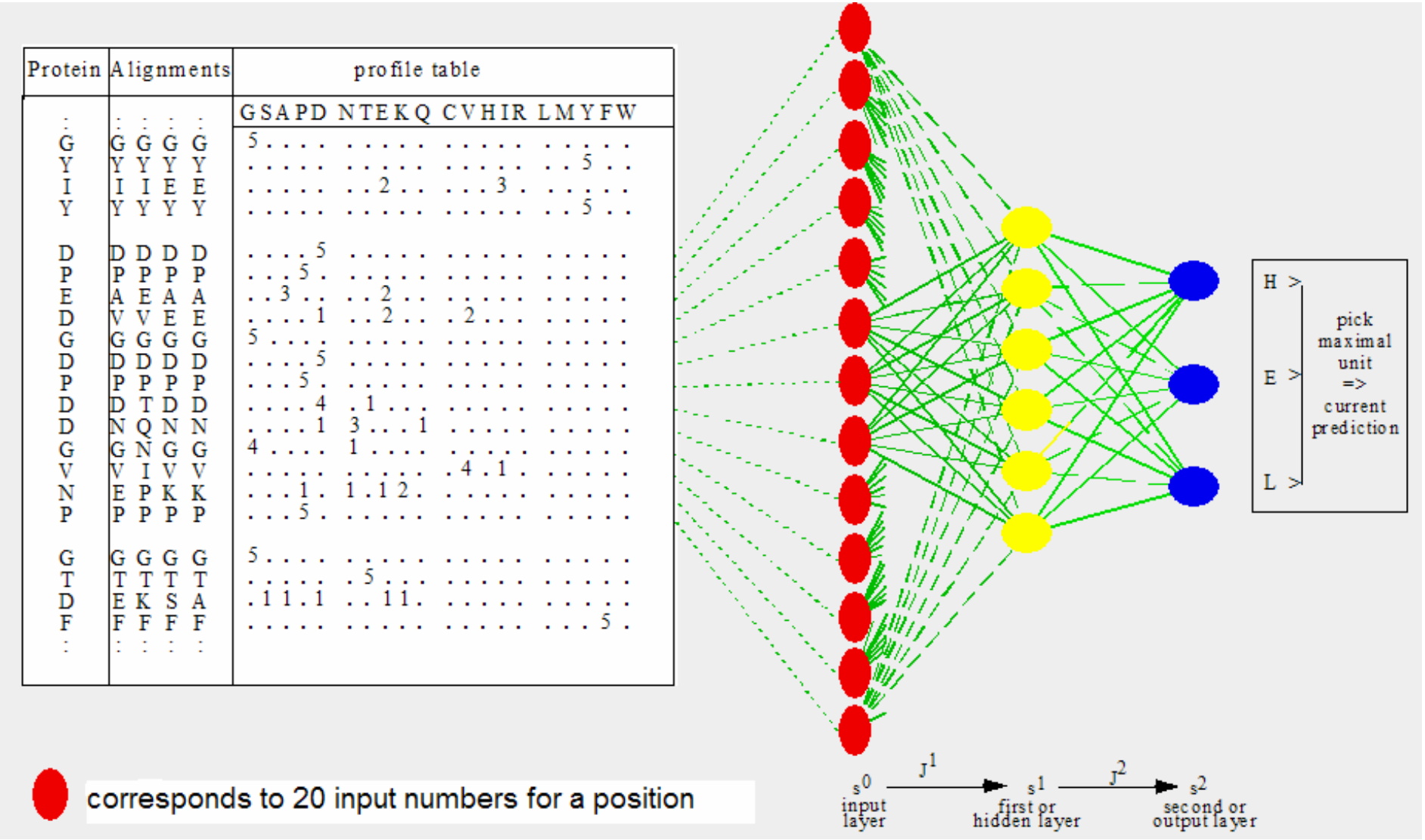
Question 1: how to encode an amino acid?
Question 2: how to train neural networks?
Parameter to decide: window size (51-101)

B. Rost, 2005

Second Breakthrough: **Evolutionary Information - Profile**



B. Rost, 2005



Comments: frequency is often normalized into probability

B. Rost, 2005

Secondary Structure Prediction Protocol

- Use PSI-BLAST to search the query sequence against a large sequence database
- Generate a multiple alignments for the query and sequences found by PSI-BLAST
- Create a profile from MSA
- Feed profile into neural network to predict secondary structure

Improve Segment Prediction Using Second Neural Network

PHDsec

local alignment 13 adjacent residues

:::
AAA
AA.
LLL
LII
AAG
CCS
GVV
:::

global statist. whole protein

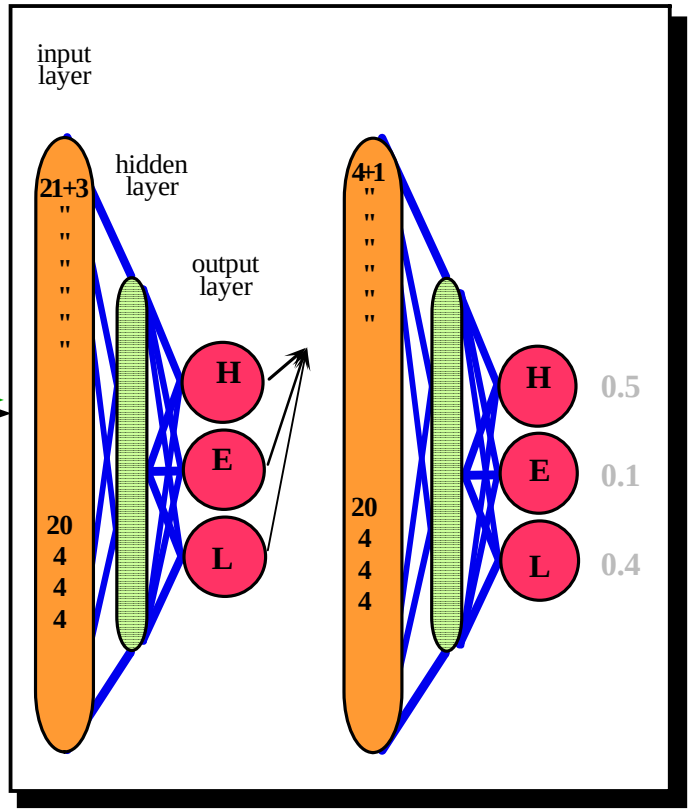
%AA
Length
² N-term
² C-term

input local in sequence

A	C	L	I	G	S	V	ins	del	cons
100	0	0	0	0	0	0	0	0	1.17
100	0	0	0	0	0	0	33	0	0.42
0	0	100	0	0	0	0	0	33	0.92
0	0	33	66	0	0	0	0	0	0.74
66	0	0	0	33	0	0	0	0	1.17
0	66	0	0	0	33	0	0	0	0.74
0	0	0	33	0	0	66	0	0	0.48

input global in sequence

percentage of each amino acid in protein
length of protein (Š60, Š120, Š240, >240)
distance: centre, N-term (Š40,Š30,Š20,Š10)
distance: centre, C-term (Š40,Š30,Š20,Š10)

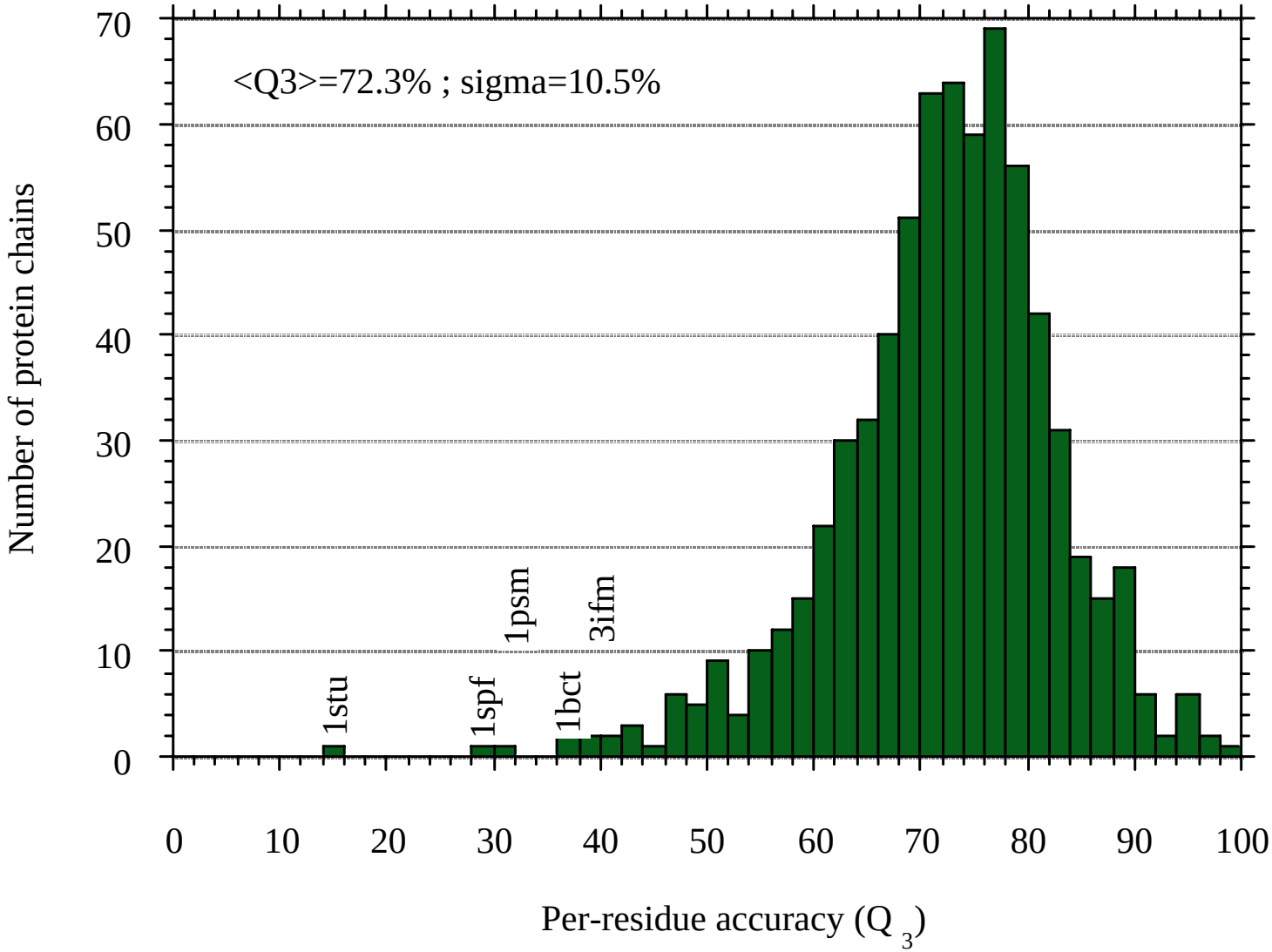


first level sequence-to-structure

second level structure-to-structure

B. Rost, 2005

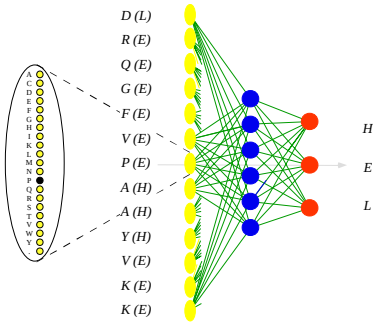
Prediction accuracy varies!



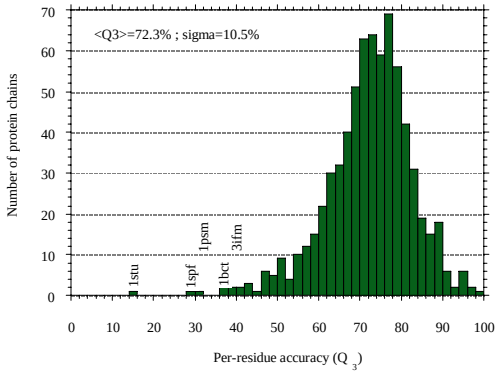
Accuracy is determined by profile quality.

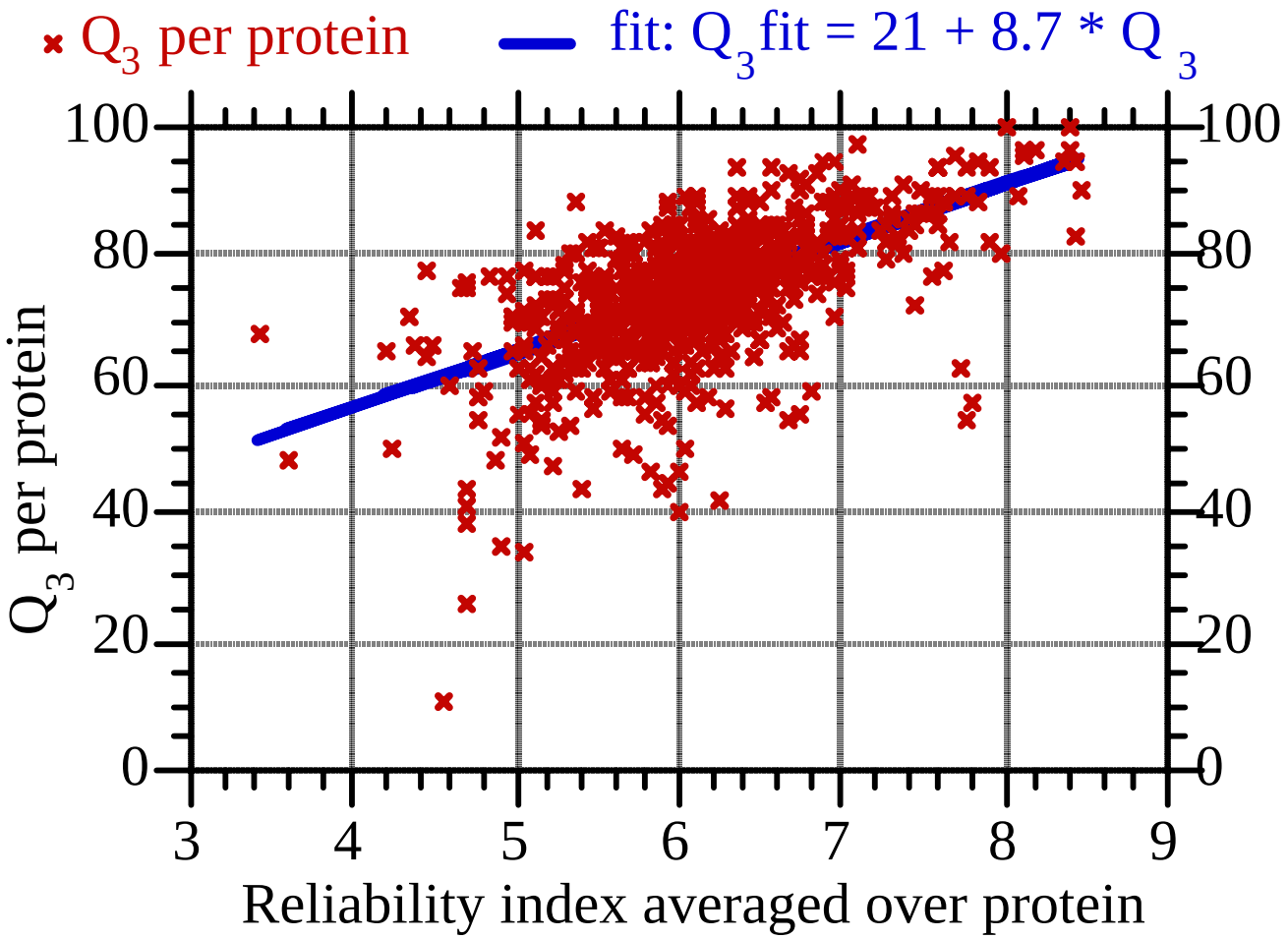
B. Rost, 2005

Stronger predictions more accurate!



H=0.5 H=0.8
E=0.4 E=0.1
L=0.1 L=0.1





B. Rost, 2005

Prediction of protein secondary structure

- 1980: 55% simple
- 1990: 60% less simple
- 1993: 70% evolution
- 2000: 76% more evolution
- what is the limit?
  - 88% for proteins of similar structure
  - missing through:
    better definition of secondary structure
    including long-range interactions
  - structural switches
  - chameleon / folding

B. Rost, 2005

EVA: secondary structure

76%

Method [B]	Q_3 [C]	Q_3 Claim [D]	SOV [E]	Info [F]	CorrH [G]	CorrE [H]	CorrL [I]	Class [K]	BAD [L]
PROF	76.0		72	0.35	0.67	0.63	0.55	82	2.7
PSIPRED	76.0	76.5-78.3 [M]	72	0.36	0.65	0.62	0.55	78	2.8
SSpro	76.0	76	71	0.35	0.67	0.63	0.56	83	2.8
JPred2	75.0	76.4	69	0.34	0.65	0.60	0.54	76	2.6
PHDpsi	75.0		71	0.33	0.65	0.60	0.54	81	3.0
PHD	71.4	71.6	68	0.28	0.59	0.58	0.49	77	4.3
Copenhagen	78 [N]	77.8							
Wang/Yuan								*53* [O]	

EVA is an automatic evaluation system of protein structure prediction.

B. Rost, 2005

Secondary structure prediction 2005

- history
  - 1st generation 50-55%
  - 2nd generation 55-62%
  - 3rd generation 1992 70-72%
    2000 > 76%
- what improves?
  - database growth +3
  - PSI-BLAST +0.5
  - new training +1
  - ‘clever method’ +1
- limit?
  - max 88% -> 12% to go
  - and from there?

B. Rost, 2005

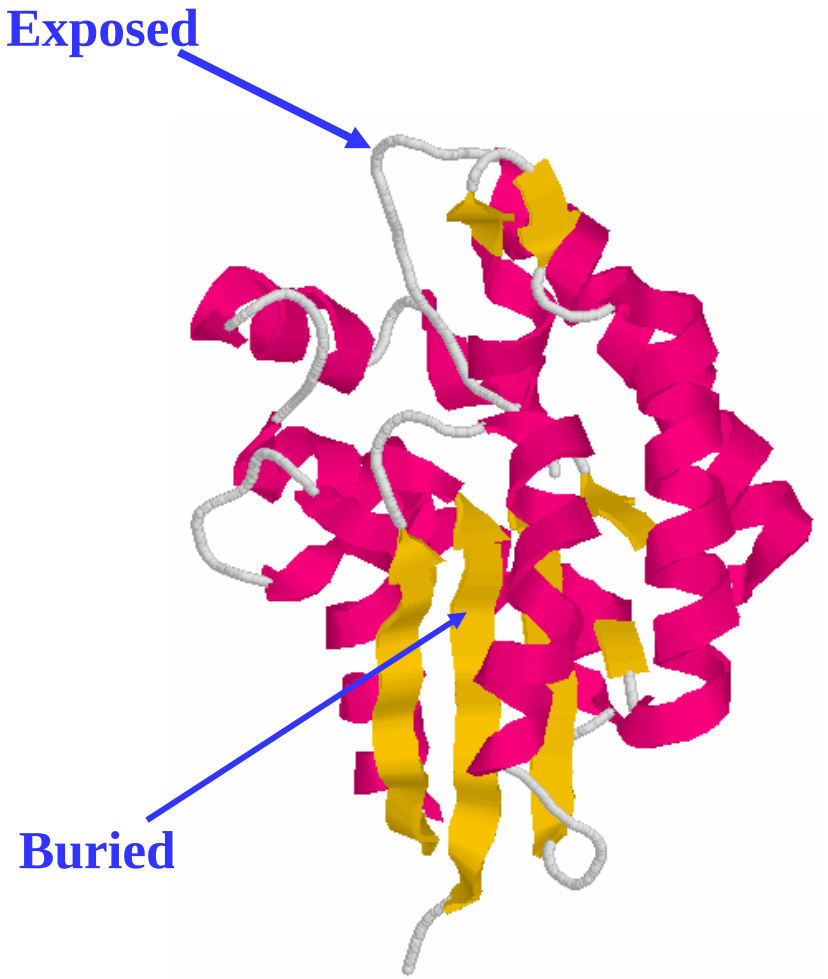
Conclusion: secondary structure prediction

- big gain through using evolutionary information
- are we going to reach above 80%? How high?
- continuous secondary structure
- better methods
- other features

Useful Tools

- PSI-PRED (http://bioinf.cs.ucl.ac.uk/psipred/)
- SSpro (http://contact.ics.uci.edu/sspro4.html)
- Porter (http://distill.ucd.ie/porter/)
- JPred (http://www.compbio.dundee.ac.uk/~www-jpred/submit.html)
- Prof_PHD (http://cubic.bioc.columbia.edu/predictprotein/)
- SAM (http://www.cse.ucsc.edu/research/compbio/sam.html)

1D: Solvent Accessibility Prediction



MWLKKFGINLLIGQSV…

Neural Networks + Alignments

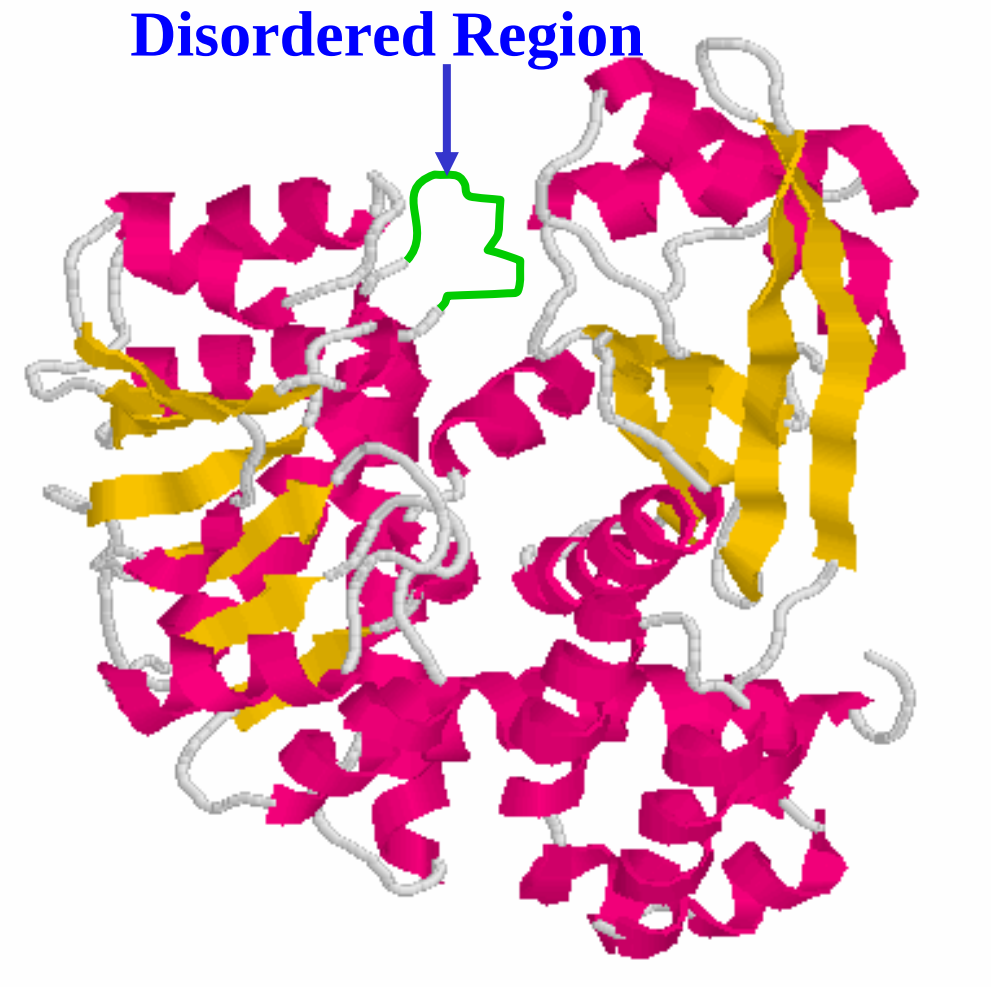
eeeeeeebbbbbbbbeeeebbb…

Accuracy: 79% at 25% threshold

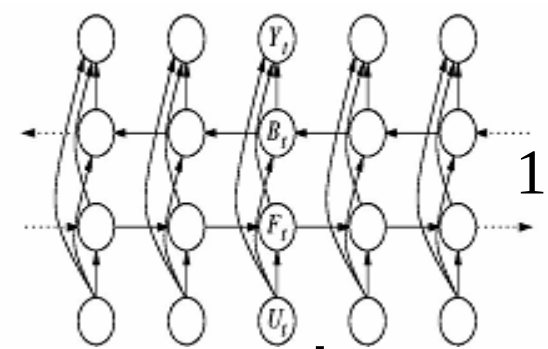
ACCpro: http://contact.ics.uci.edu/download.html

Cheng, Randall, Sweredoski, Baldi. *Nucleic Acid Research*, 2005

1D: Disordered Region Prediction Using Neural Networks



MWLKKFGINLLIGQSV…



1D-RNN

OOOOODDDDOOOOO…

93% TP at 5% FP

Cheng, Sweredoski, Baldi. *Data Mining and Knowledge Discovery*, 2005

Self-Evaluation of DISpro in CASP7 (93 targets)

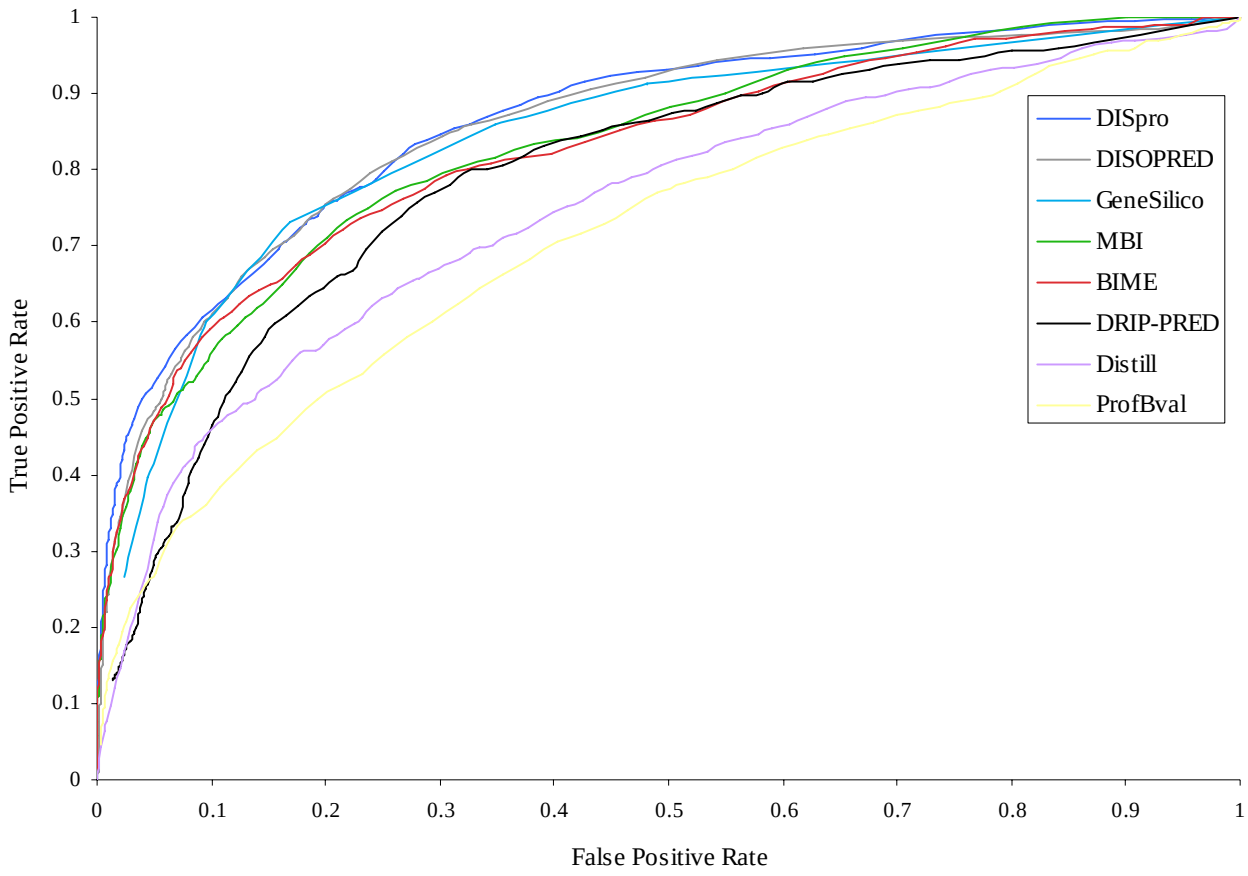


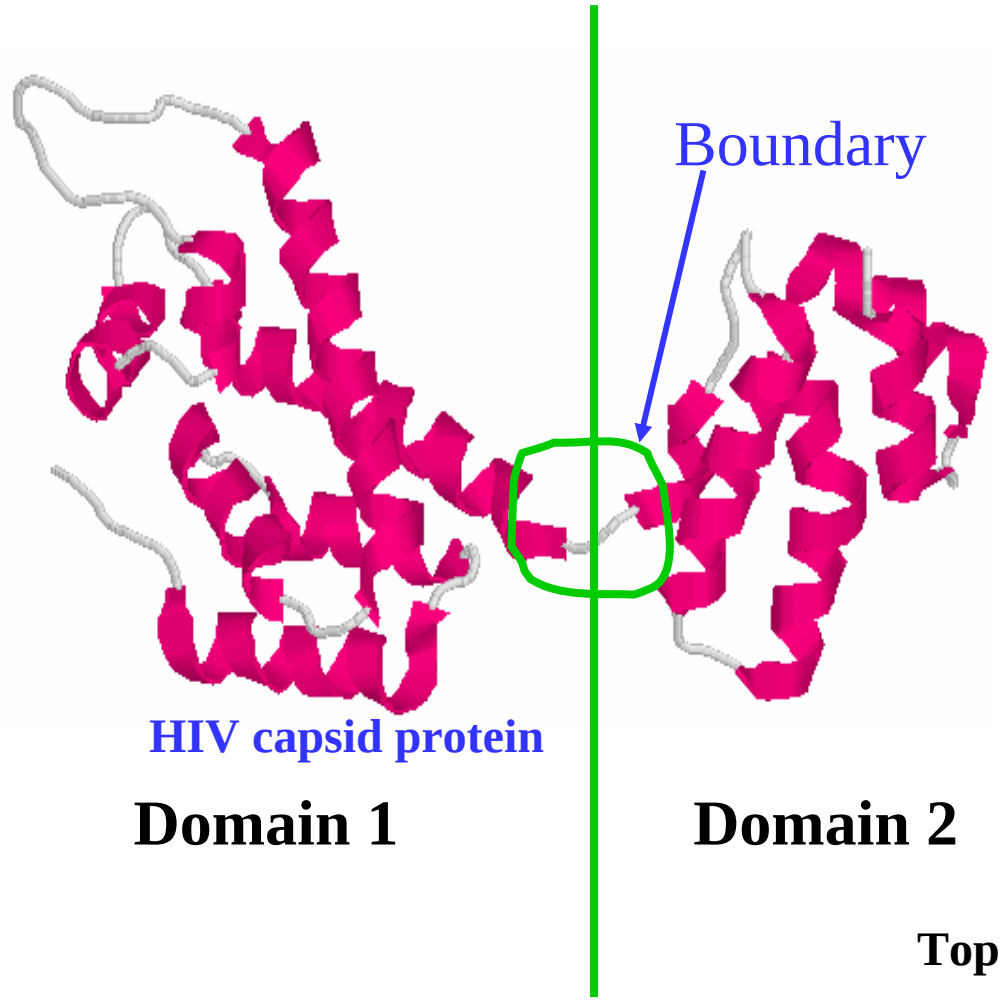
Figure 5: ROC curves of eight predictors on the CASP7 dataset. The false positive rates and true positives rates are computed on 93 targets used in CASP7

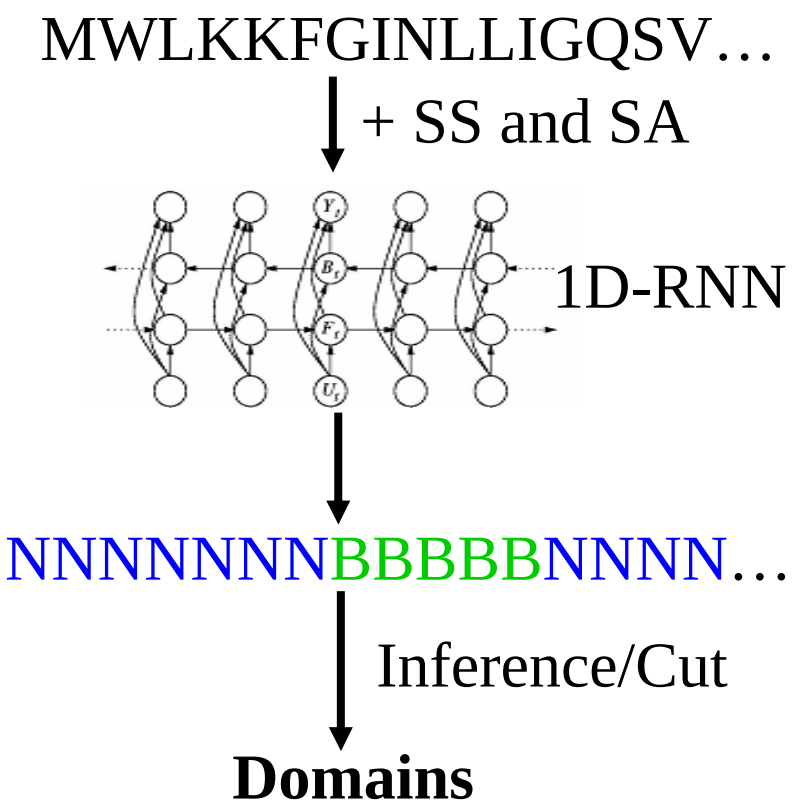
Predictor	ROC score
DISpro	**0.863**
DISOPRED	0.857
GeneSilico	0.838
MBI	0.831
BIME	0.825
DRIP-PRED	0.792
Distill	0.748
ProfBval	0.710

Results on Dec 30, 2006

Notice: According to the official CASP7 evaluation, **DISOPRED** is Ranked 1st. **DISpro** is ranked 2nd among servers. The difference is probably due to the minor discrepancy in the classification of disorder regions.

1D: Protein Domain Prediction Using Neural Networks





Top *ab-initio* domain predictor in CAFASP4

Cheng, Sweredoski, Baldi. *Data Mining and Knowledge Discovery*, 2006.

Comments on Domain Prediction

- Domain prediction for homologous proteins is more accurate.
- Ab-initio domain prediction is still very hard and an open problem.
- It is important to integrate homology and ab-initio method. Our domain predictor FOLDpro of combining homology and ab-initio method is one of the best method in the latest **CASP7** competition.

Self-Evaluation of FOLDpro-DOMpro in CASP7 (95 targets)

Method	Target Num	Domain Num Acc. (%)	CASP7 Score
FOLDpro-DOMpro	**95**	**93.7**	**0.963**
Baker-RosettaDom	94	86.2	0.940
Ma-OPUS-DOM	94	87.2	0.933
ROBETTA-GINZU	94	84.0	0.932
DomSSEA	94	78.7	0.910
HHpred3	95	75.8	0.910
Meta-DP	95	74.7	0.907
HHpred1	93	75.3	0.902
DomFOLD	95	75.8	0.898
DPS	93	75.3	0.889
Chop	83	56.6	0.827
Distill	95	70.5	0.819
NN_PUT-Lab	92	58.7	0.795

According to the official CASP7 results, FOLDpro-DOMpro is ranked 1st.

Results on Dec 30, 2006

Outline

I. Sequence, Structure, Function Relation

II. Determination, Storage, Visualization, and Comparison

III. Structure Classification

IV. 1D Prediction

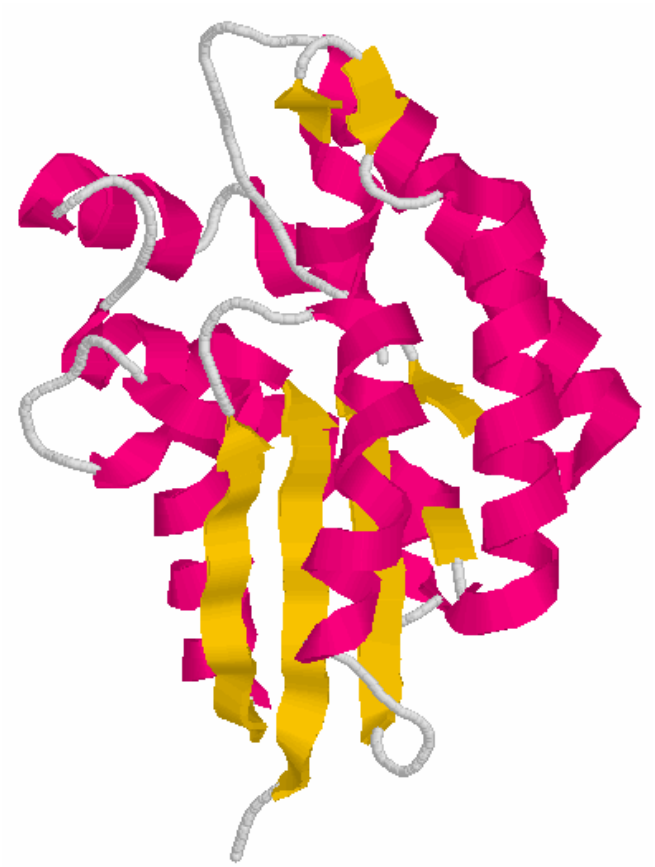
V. 2D Prediction

VI. 3D Prediction

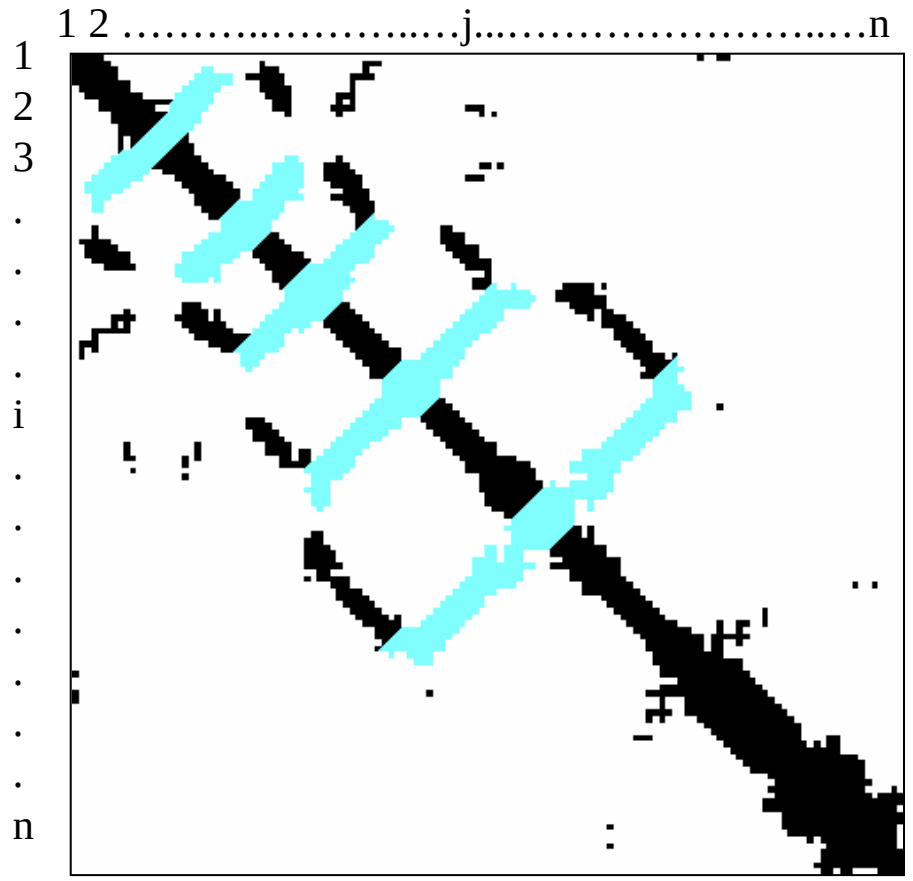
VII. Tools and Projects

2D: Contact Map Prediction

3D Structure



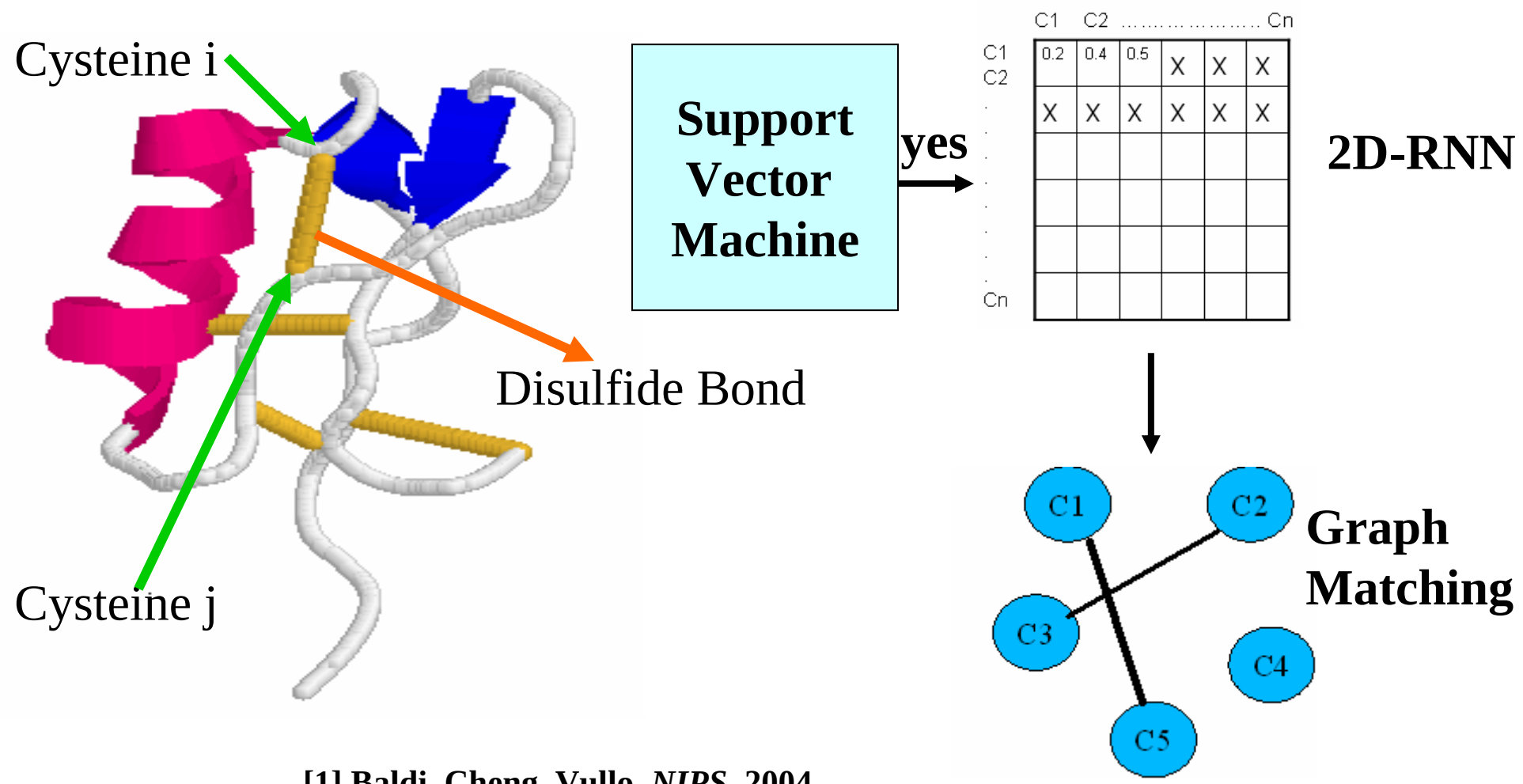
2D Contact Map



Distance Threshold = 8A°

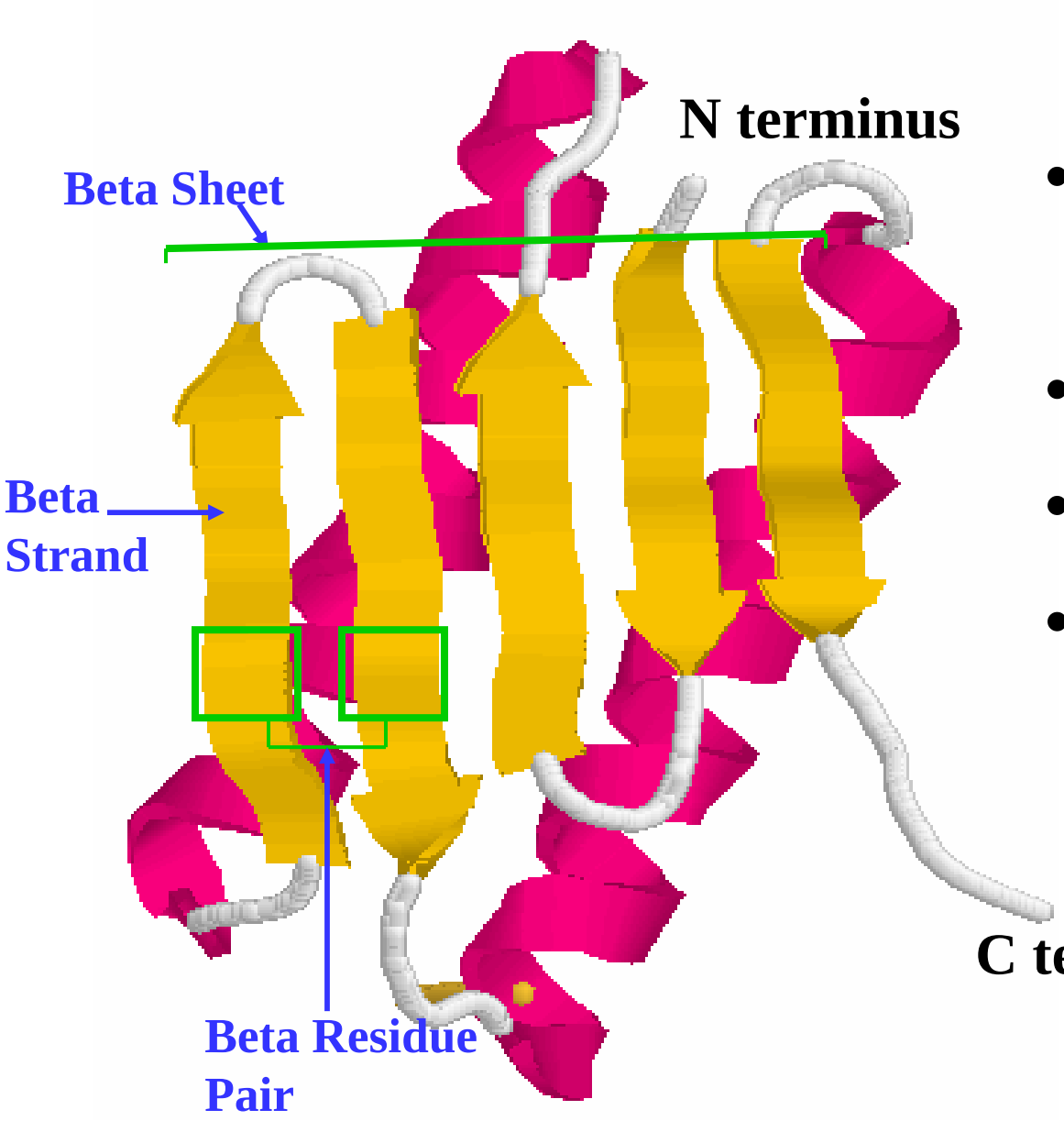
Cheng, Randall, Sweredoski, Baldi. *Nucleic Acid Research*, 2005

2D: Disulfide Bond Prediction



[1] Baldi, Cheng, Vullo. *NIPS*, 2004.
[2] Cheng, Saigo, Baldi. *Proteins*, 2005

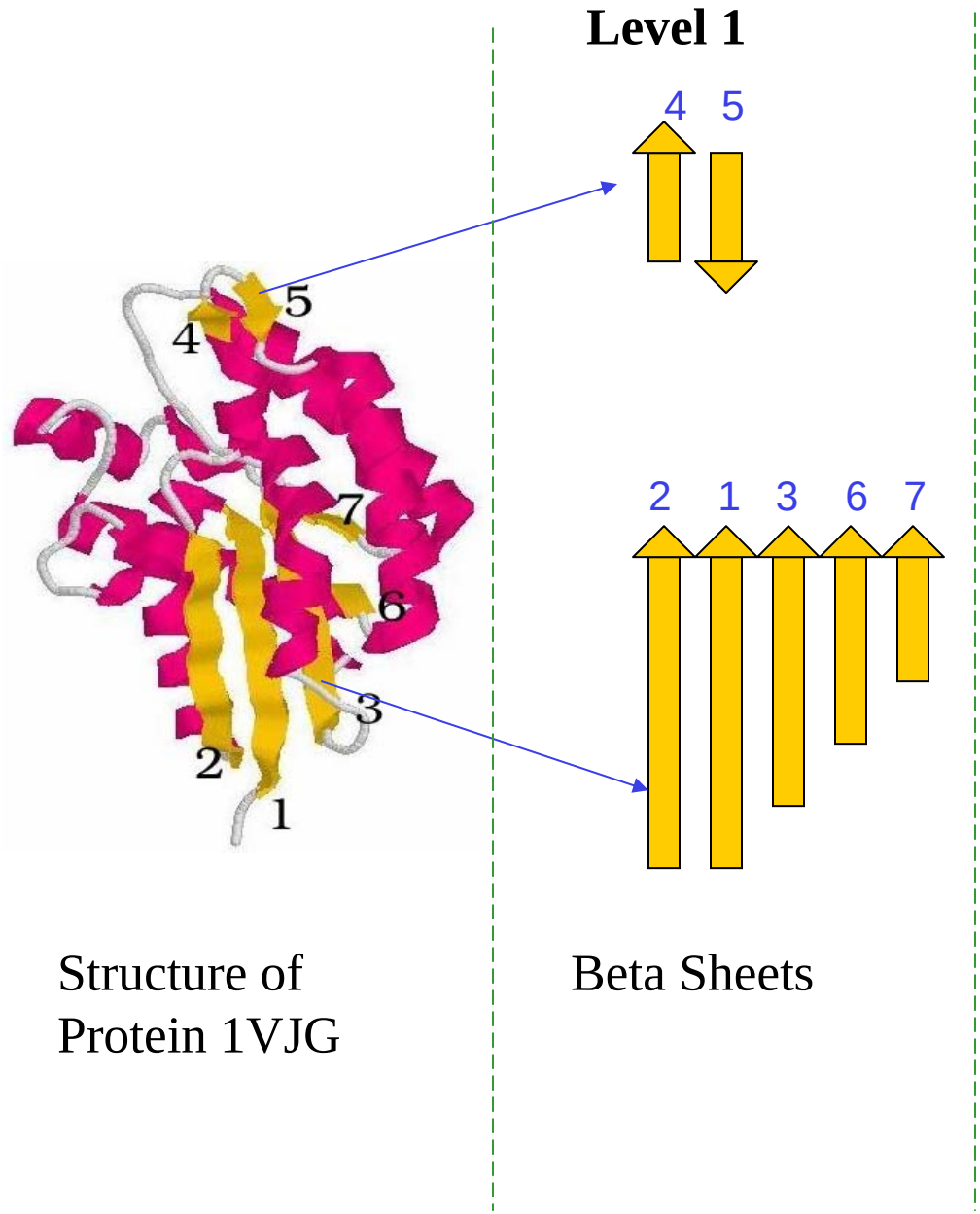
2D: Prediction of Beta-Sheet Topology



- Ab-Initio Structure Prediction
- Fold Recognition
- Protein Design
- Protein Folding

Cheng and Baldi, *Bioinformatics*, 2005

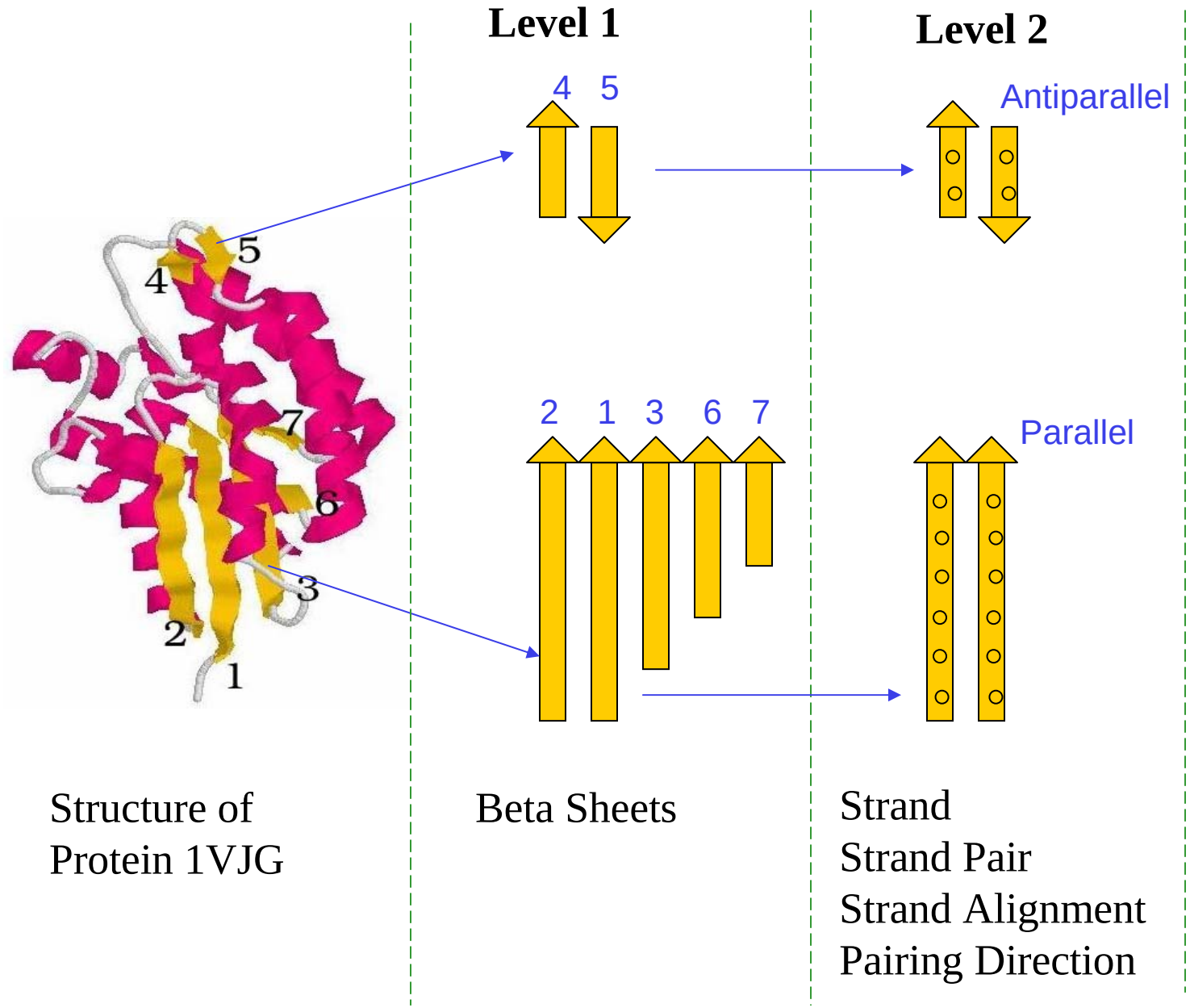
An Example of Beta-Sheet Topology



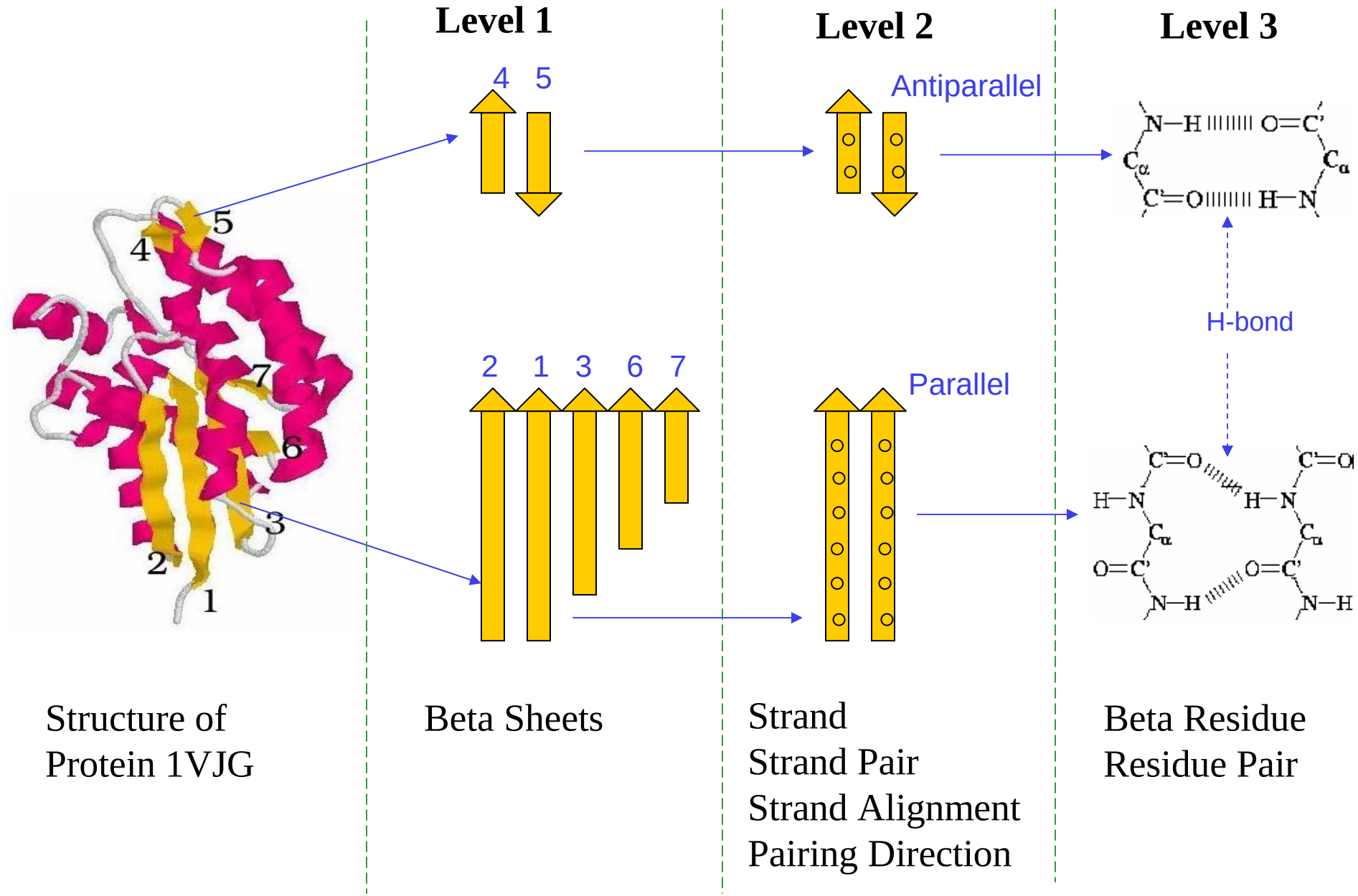
Structure of Protein 1VJG

Beta Sheets

An Example of Beta-Sheet Topology



An Example of Beta-Sheet Topology



Three-Stage Prediction of Beta-Sheets

- Stage 1

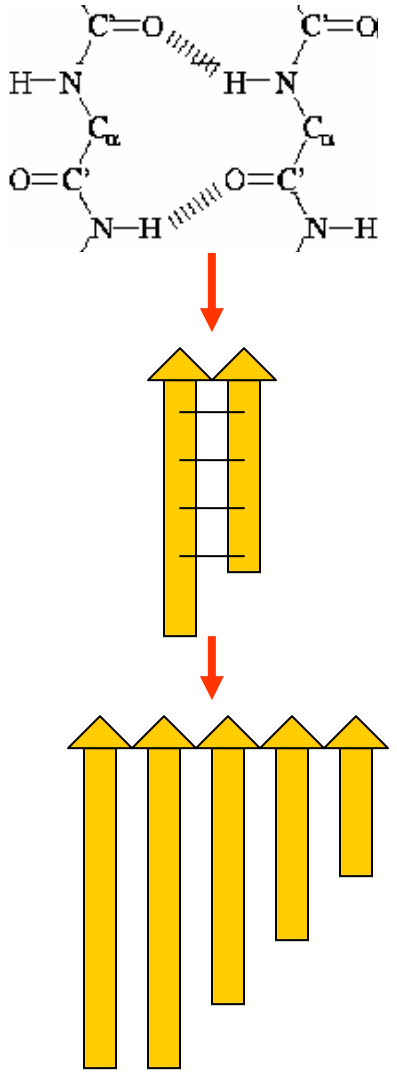
  Predict beta-residue pairing probabilities using 2D-Recursive Neural Networks (2D-RNN, Baldi and Pollastri, 2003)

- Stage 2

  Use beta-residue pairing probabilities to align beta-strands

- Stage 3

  Predict beta-strand pairs and beta-sheet topology using graph algorithms



Take home: 2D prediction

- Prediction hard, but stakes are high
- inter-residue
  - Correlated mutations can imply spatial proximity
  - Distinction between different models, no accurate prediction of 3D, yet
- inter-strand
  - Sometimes good enough for approximate modelling of 3D
- Don’t freak out when accuracy is low
    - 1) how accurate are these prediction methods on average
    - 2) are all important contacts predicted?
- for 5% of the best-predicted contacts prediction accuracy about 50% (sequence separation >= 6)

B. Rost, 2005

Self-evaluation in CASP7

(13 de novo domains, long-range contacts, sequence separation >= 12 and 24, respectively)

	Separation >= 12		Separation >= 24	
Method	Acuracy (%)	Coverage (%)	Accuracy (%)	Coverage (%)
SVMcon	27.7	4.7	13.1	2.8
BETApro	35.4	5.1	19.7	3.2
SAM-T06	20.7	3.5	18.5	3.9
Distill	26.4	2.9	13.7	1.4
Possum	15.0	2.3	21.4	2.6
GPCPRED	12.2	2.1	10.5	2.0
GajdaPairings	9.8	1.5	10.4	1.9

Results on Dec 30, 2006

According to the official CASP7 results on all de novo protein targets, SAM-T06 is ranked 1st and BETApro 2nd for separation >= 24.

Outline

I. Sequence, Structure, Function Relation

II. Determination, Storage, Visualization, and Comparison

III. Structure Classification

IV. 1D Prediction

V. 2D Prediction

VI. 3D Prediction

VII. Useful Tools

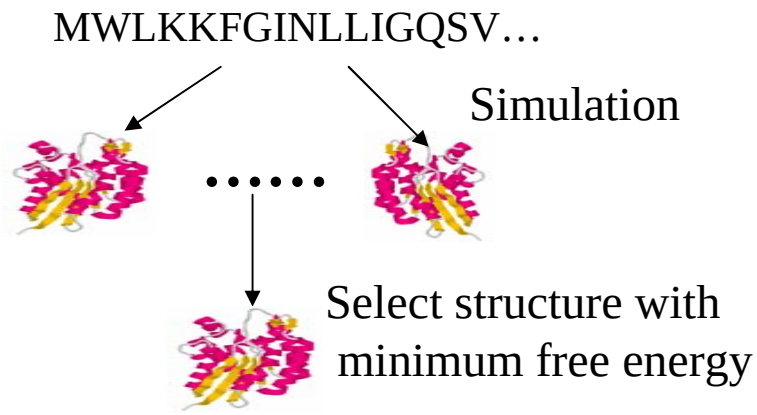
Two Methodologies for 3D Structure Prediction

- AB Initio Method (physical-chemical principles / molecular dynamics, knowledge-based approaches)
- Template-Based Method (knowledge-based approaches)

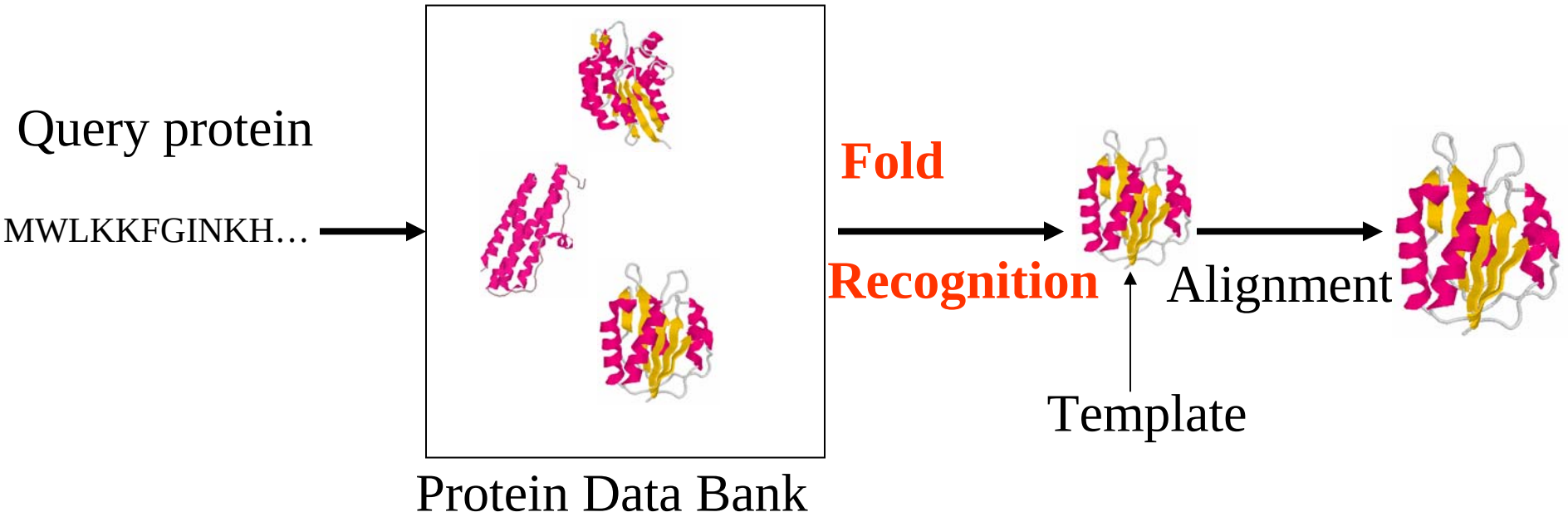
Two Approaches

- **Ab Initio Structure Prediction**

  Physical force field – protein folding
  Contact map - reconstruction



- **Template-Based Structure Prediction**



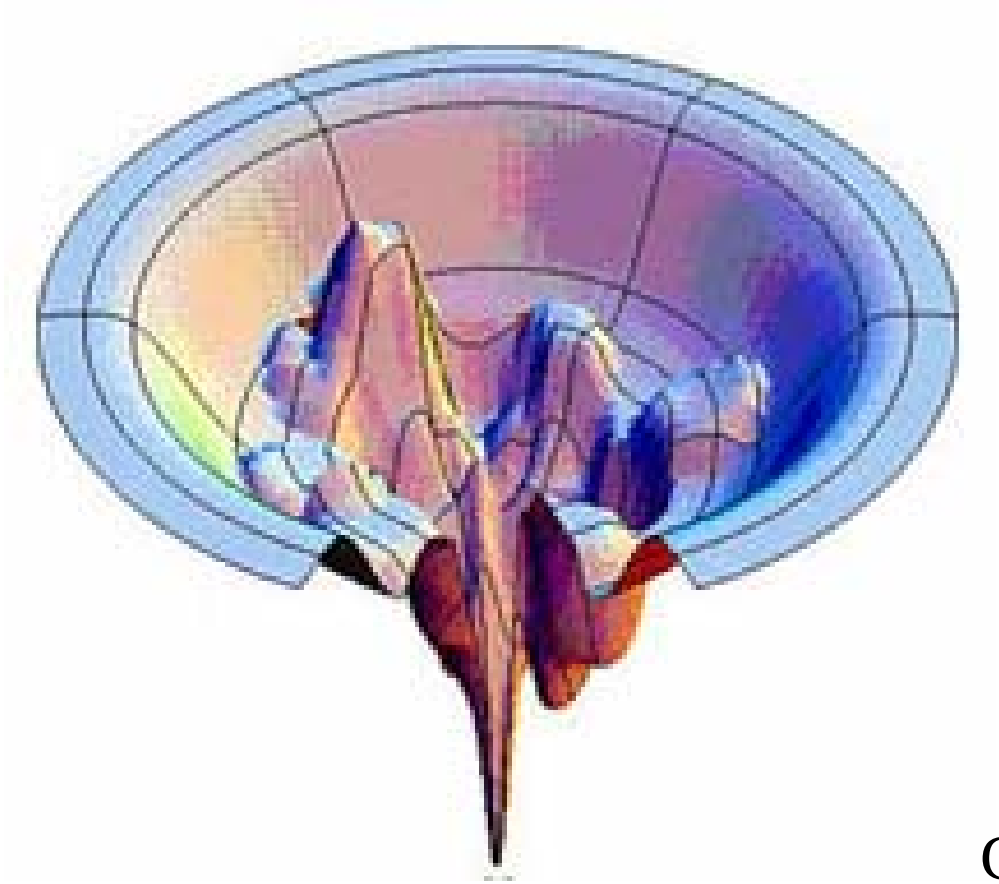
Protein Folding and Structure Prediction is Different

- Folding is to understand how protein folds into a unique, well-determined structure in a short time (e.g., 1 minute), which otherwise takes millions of years by random sampling.
- Protein structure prediction is to predict the final 3D shape of protein.
- Two problems are related and enrich each other.

Ab Initio Structure Prediction

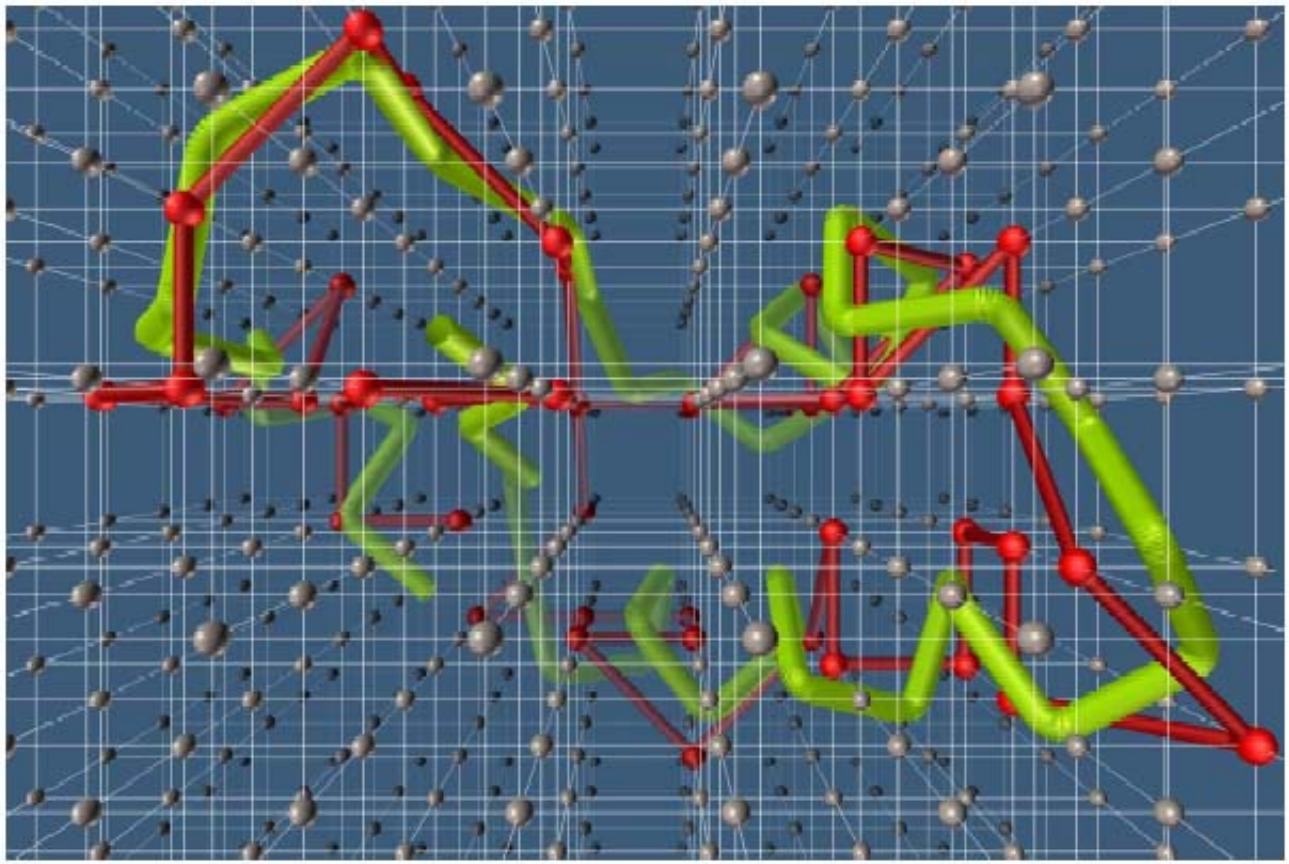
- Energy Function
- Structure Sampling
- Currently, the accuracy of *Ab-Initio* method is low.
- Energy function is not accurate.
- Sampling method can’t find the structure with lowest energy in majority of cases.
- Sampling takes a long time and only is able to sample a small structure space.

Protein Energy Landscape



C. Park, 2005

Markov Chain Monte Carlo Simulation



picture from http://www.scripps.edu/pub/olson-web/people/sanner/html/lat_gallery.html

Some *Ab Initio* Tools

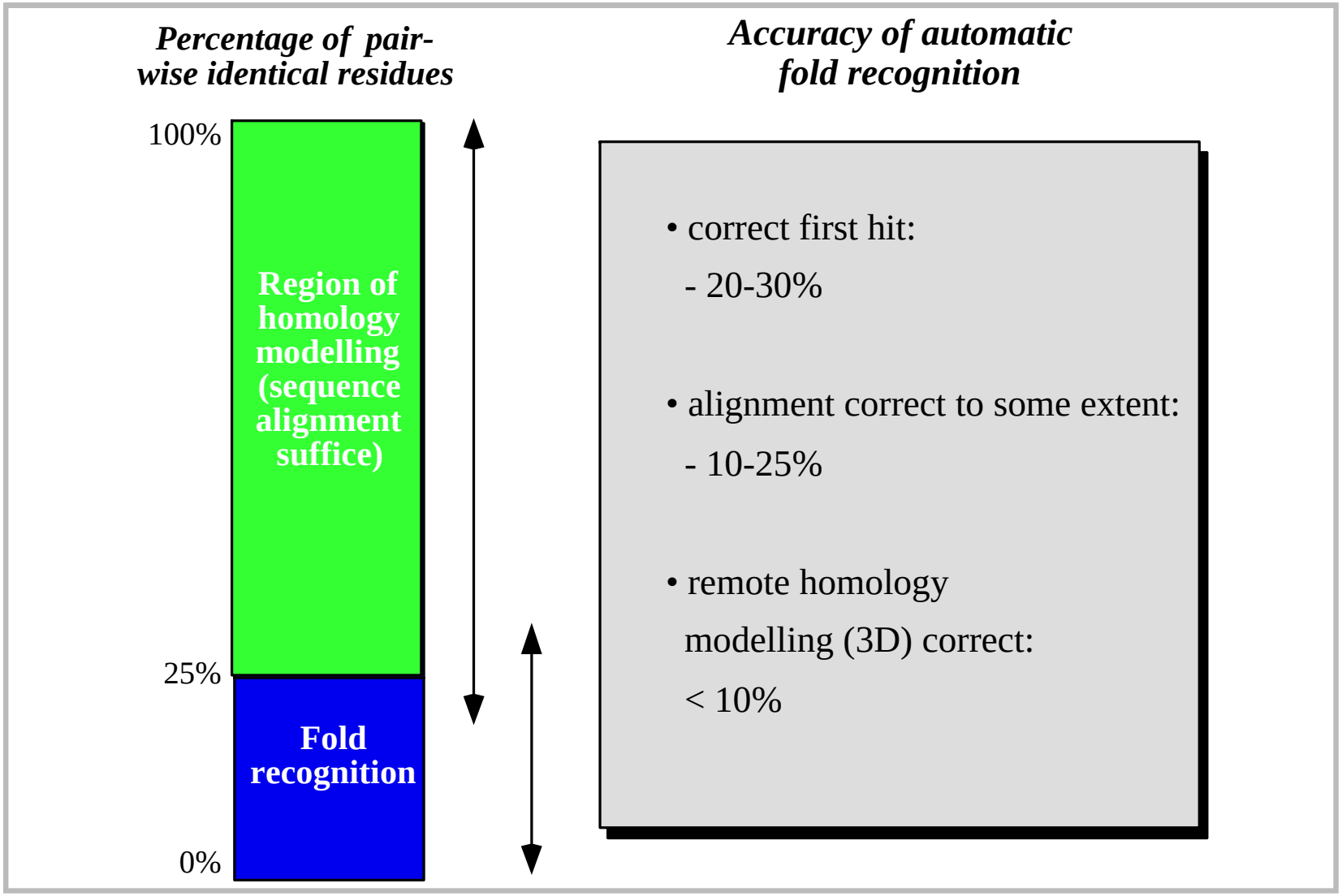
- David Baker's Rosetta (http://depts.washington.edu/ventures/UW_Technology/Express_Licenses/Rosetta/)
- Jeffrey Skolnick's TouchStone

Template-Based Structure Prediction

1. Template identification
2. Query-template alignment
3. Model generation
4. Model evaluation
5. Model refinement

Notes: if template is easy to identify, it is often called **comparative Modeling or homology** modeling. If template is hard to identify, it is often called **fold recognition**.

Extending comparative modelling: fold recognition / threading



B. Rost, 2005

Basic Idea of Template-Based 3D Structure Prediction

- **Assumption: Query and Template have similar structure (due to homology or convergent evolution)**
- **Strategy: modelling of query based on template**

B. Rost, 2005

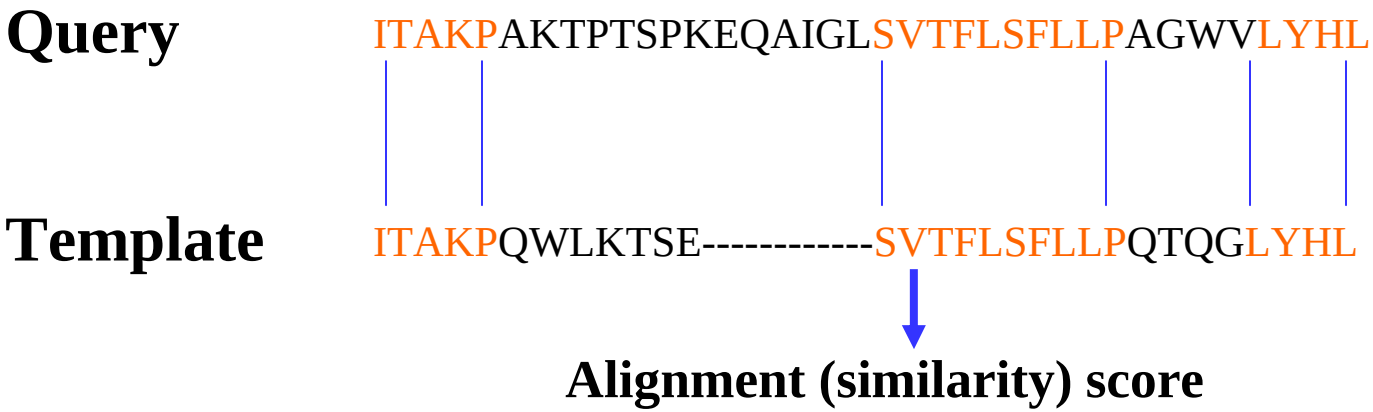
Template Identification

- Sequence alignment (BLAST, Smith-Waterman algorithm)
- Sequence profile alignment (PSI-BLAST, HMM (SAM-T02, Hammer)
- Profile-Profile alignment (HHSearch, Sparks, Compass, FFAS, BASIC)
- Sequence-Structure Alignment (3D-1D, FUGUE, mGenThreader, Raptor)
- Consensus (Pcons, 3D-Jury)
- Machine Learning Information Retrieval Approach (FOLDpro)

Classic Fold Recognition Approaches

Sequence - Sequence Alignment

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

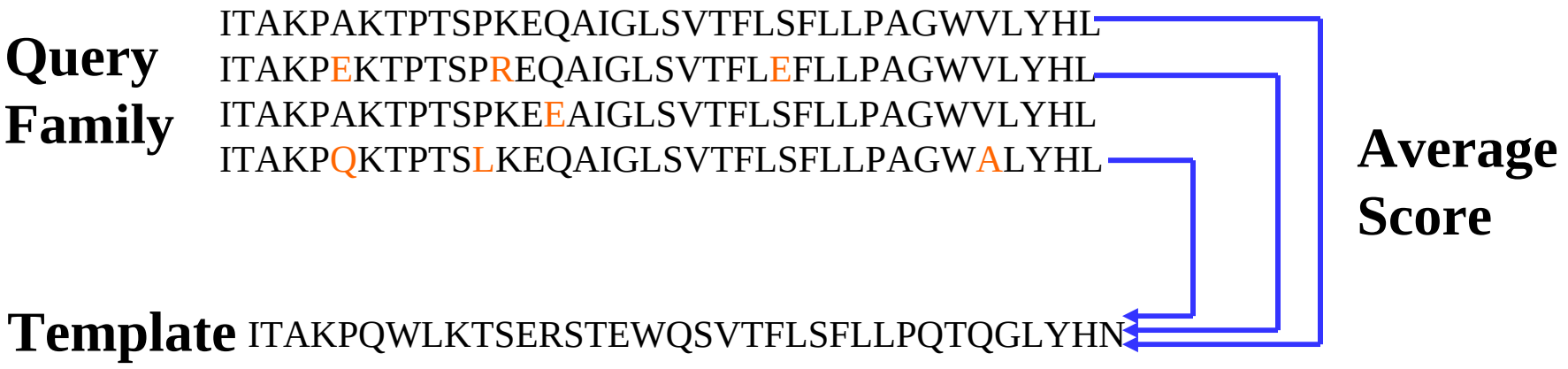


Works for >40% sequence identity
(Close homologs in protein family)

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

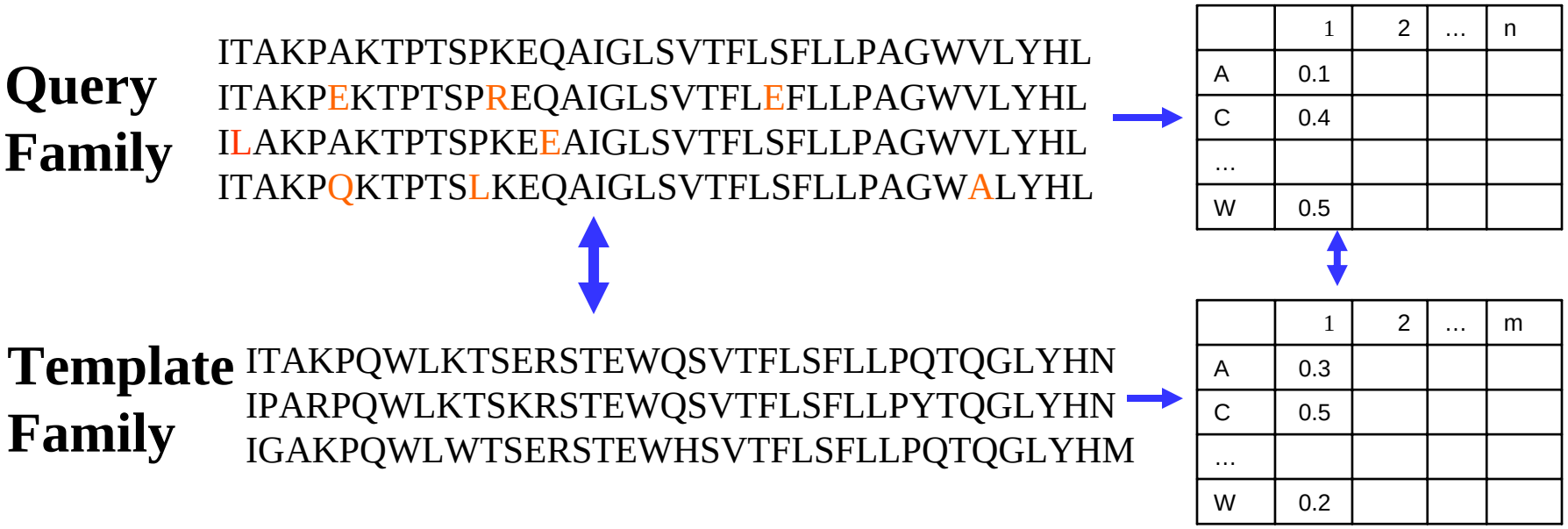
Template ITAKPQWLKTSERSTEWQSVTFLSFLLPQTQGLYHN

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

Classic Fold Recognition Approaches

Profile - Profile Alignment

(Rychlewski et al., 2000)



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

Classic Fold Recognition Approaches

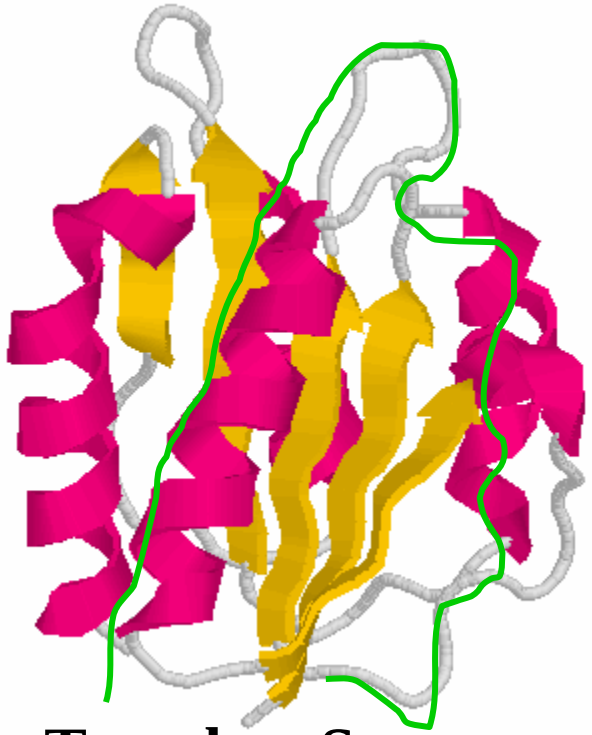
Sequence - Structure Alignment (Threading)

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

Query

MWLKKFGINLLIGQS….

Fit

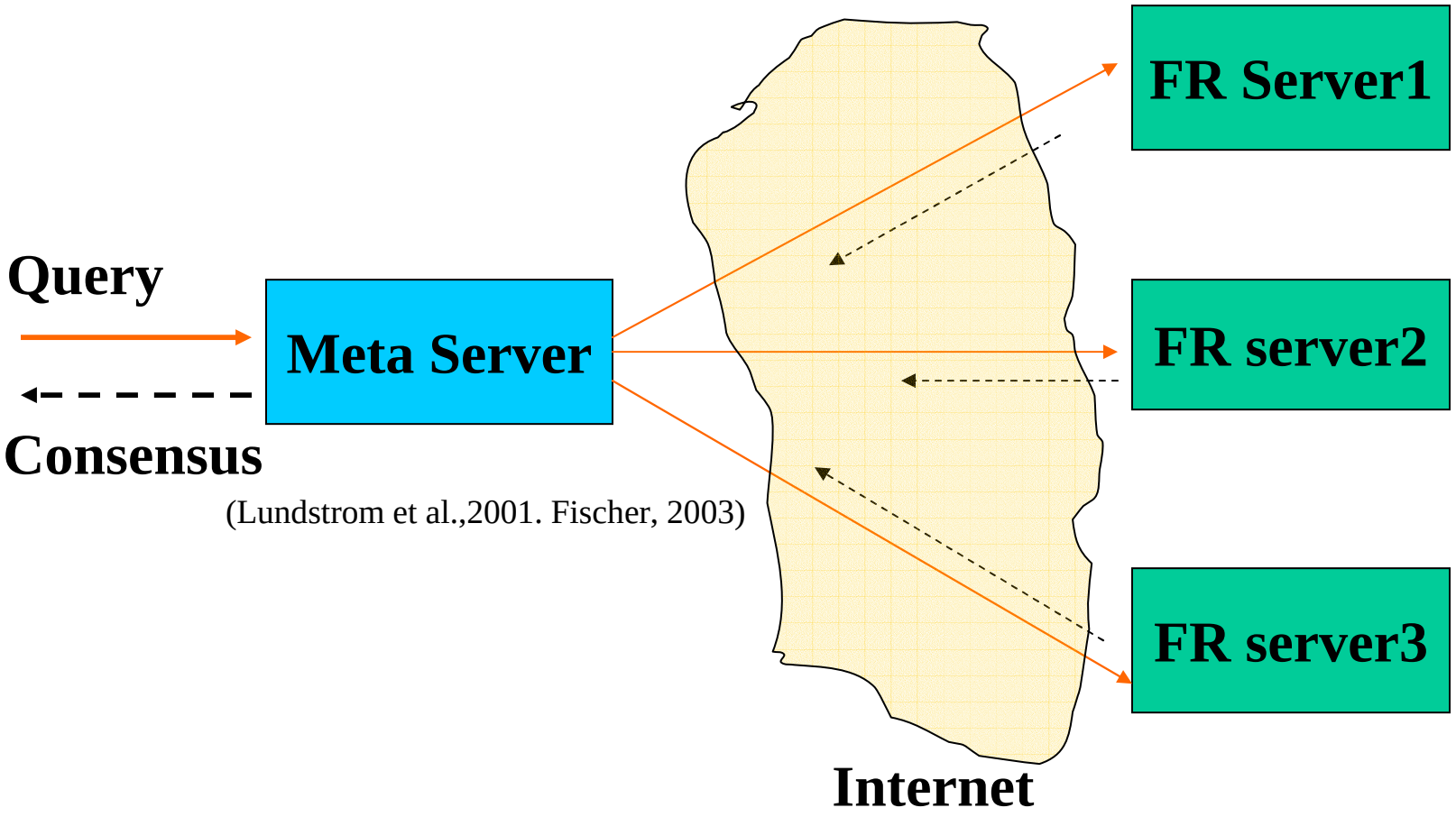


Template Structure

Fitness Score

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

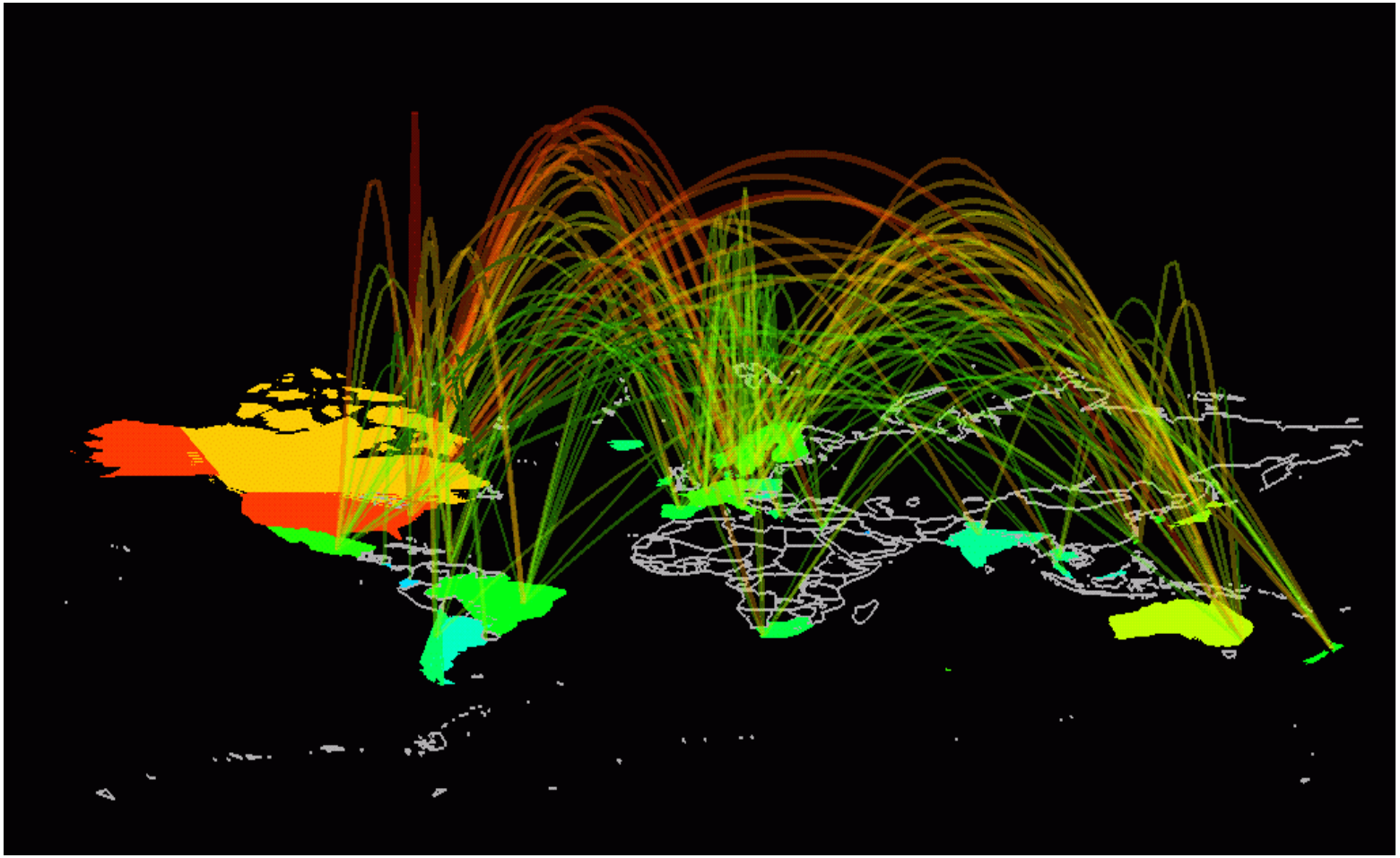
Integration of Complementary Approaches



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

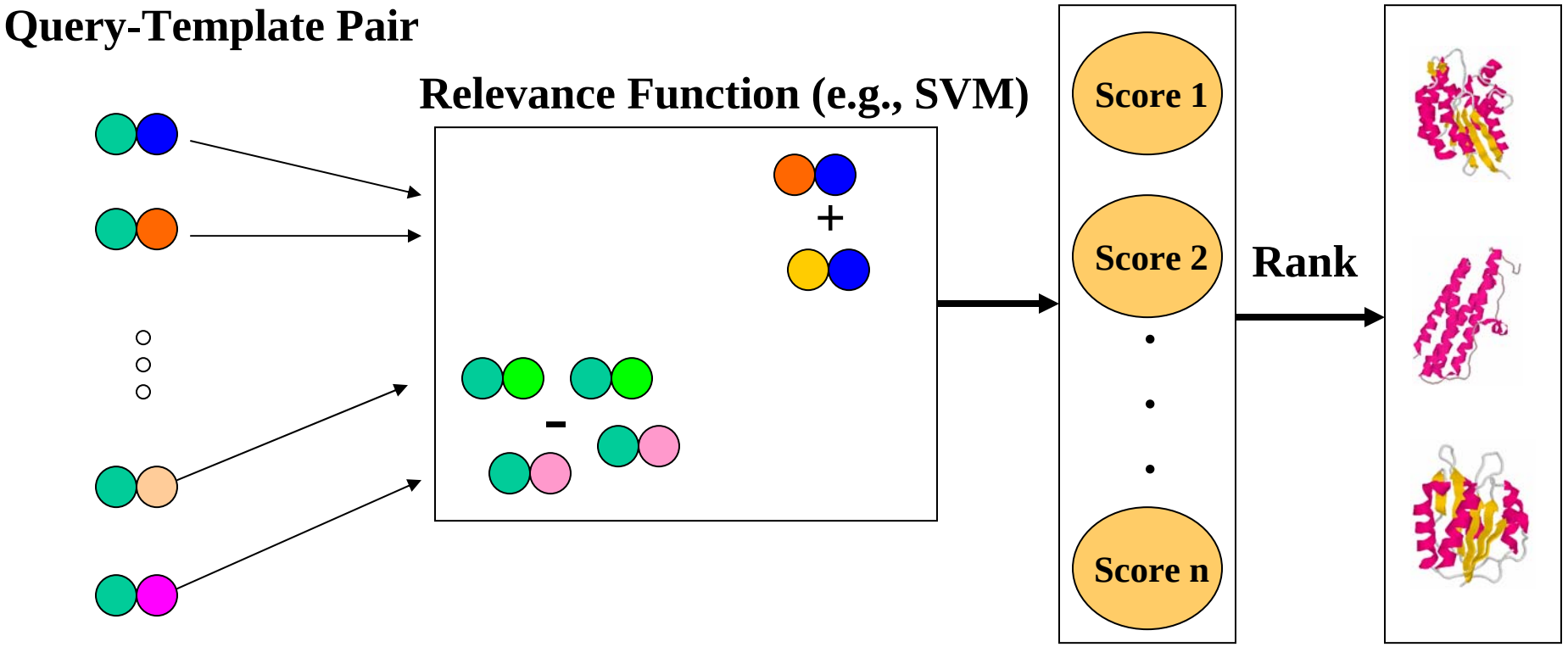
1. **Reliability depends on availability of external servers**
2. **Make decisions on a handful candidates**

Servers, META-servers, META-META, …



B. Rost, 2005

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)**

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**

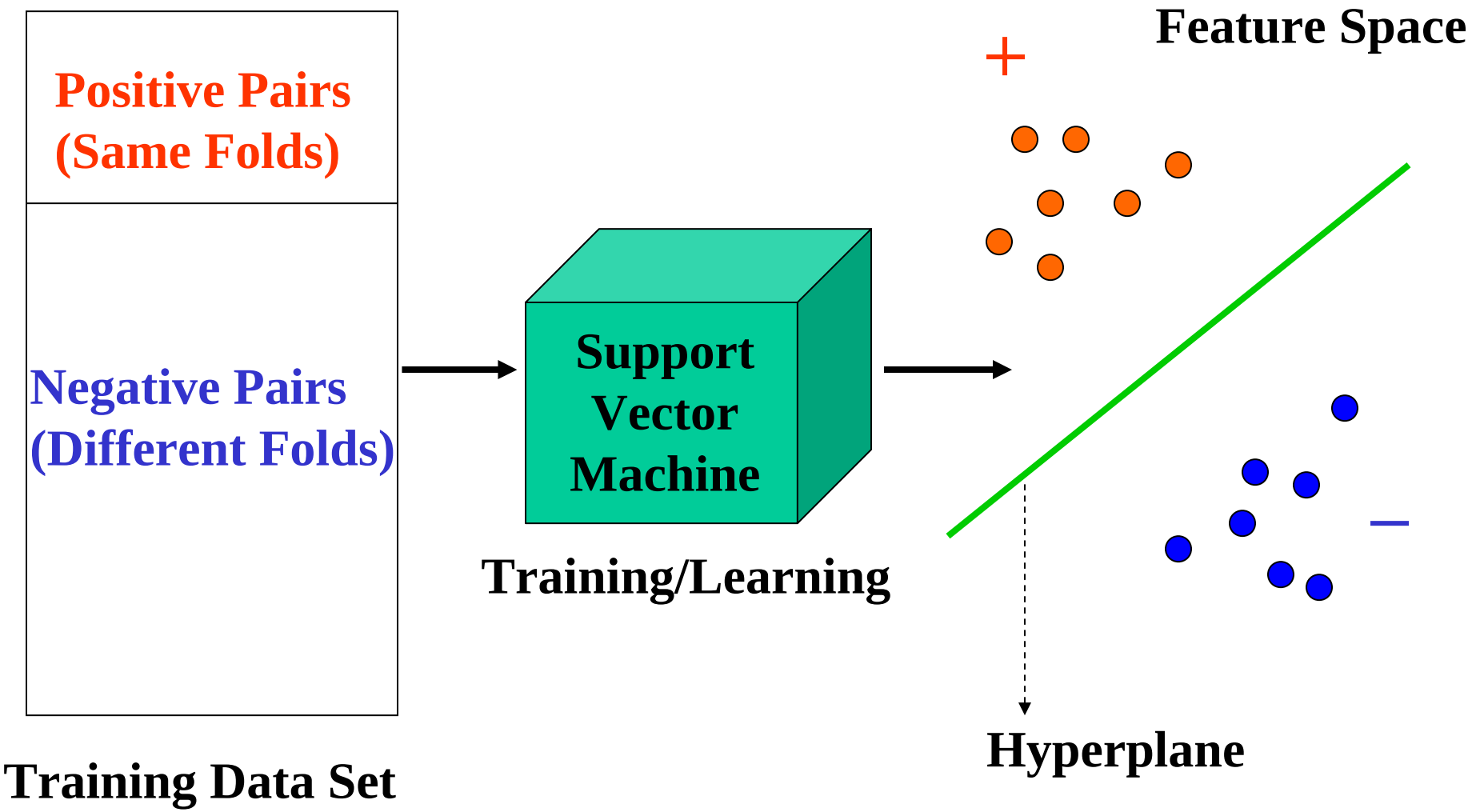
  ClustalW, HHSearch, Lobster, Compass, PRC-HMM
- **Structural Features**

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

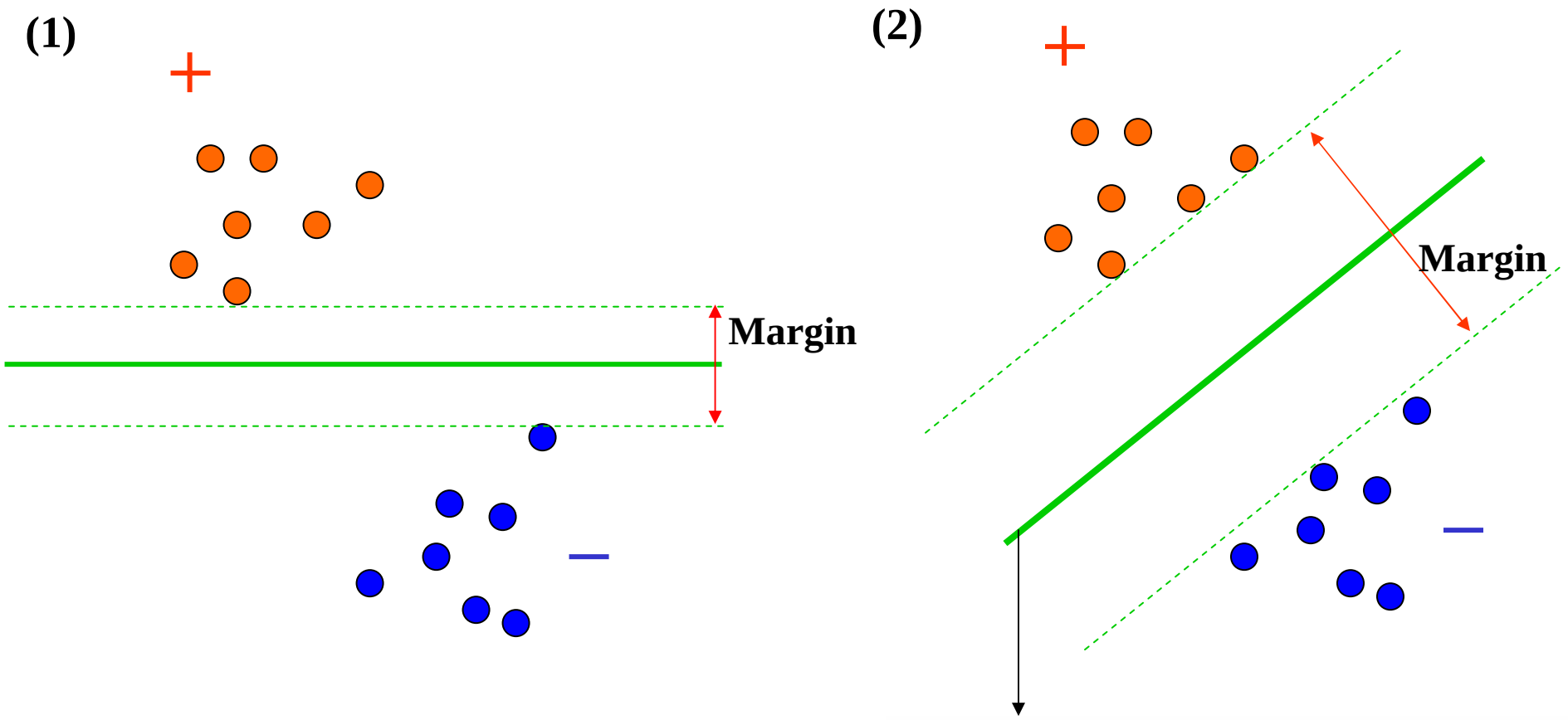
Pairwise Feature Extraction

Category	Feature	Method	Num
Seq and Family Info.	Seq monomer compo	cos/corr/Gauss	3
	Seq dimer compo	cos/corr/Gauss	3
	Fam monomer compo	cos/corr/Gauss	3
	Fam dimer compo	cos/corr/Gauss	3
Seq–Seq Align.	Local alignment	PALIGN	2
	Global alignment	CLUSTALW	1
Seq–Prof Align.	Prof versus seq	PSI-BLAST	3
	Prof versus seq	IMPALA	3
	Prof versus seq	HHMER	2
	Seq versus prof	RPS-BLAST	3
	Seq versus prof	HMMER	1
Prof–Prof Align.	Multiple alignment	CLUSTALW	1
	PSSM	COMPASS	2
	HMM prof	PRC	6
	HMM prof	HHSearch	1
Structural Info.	SS and RSA match	ratio	2
	SS and RSA compo	cos/corr/Gauss	4
	Contact probability	average	2
	Residue contact order	cos/corr	4
	Residue contact num	cos/corr	4
	Beta-sheet pair prob.	average	1
Total	-	-	54

Relevance Function: Support Vector Machine Learning



Relevance Function: Support Vector Machine Learning

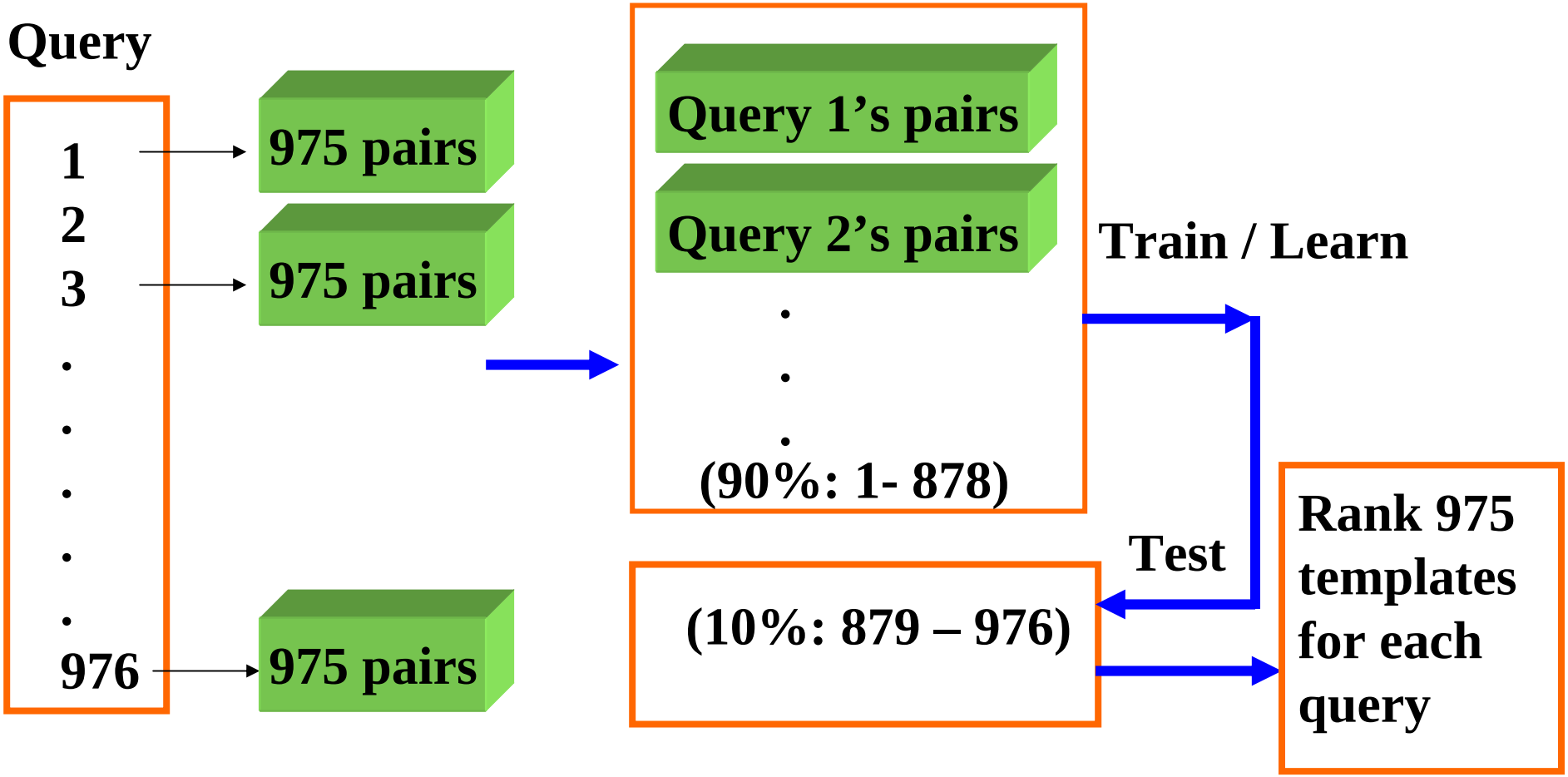


$$f(x) = \sum_{x_i \in S} \alpha_i y_i K(x, x_i) + b$$

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

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

Advantages of MLIR Framework

- Integration
- Accuracy
- Extensibility
- Simplicity
- Reliability
- Completeness
- Potentials

Disadvantages

Slower than some alignment methods

Does not generate alignments

Query-Template Alignment

- Most fold recognition methods are some kind of specialized alignment methods. So they generate alignments.
- Consensus method or machine learning methods need to re-select or re-regenerate alignments.
- For similar sequences, PSI-BLAST alignment is ok. For distantly related sequences, profile-profile alignment methods seem to be better (HHSearch, COMPASS, LOBSTER (COACH), CLUSTALW, T-Coffee, and so on).

Model Generation

- Modeller (http://www.salilab.org/modeller/download_installation.html)
- Swiss-Model (http://swissmodel.expasy.org//SWISS-MODEL.html)
- **Segmod/ENCAD (csb.stanford.edu/levitt/segmod/)**

TARGET

TEMPLATE

ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDD
VPERLIPERASFQWMNDK

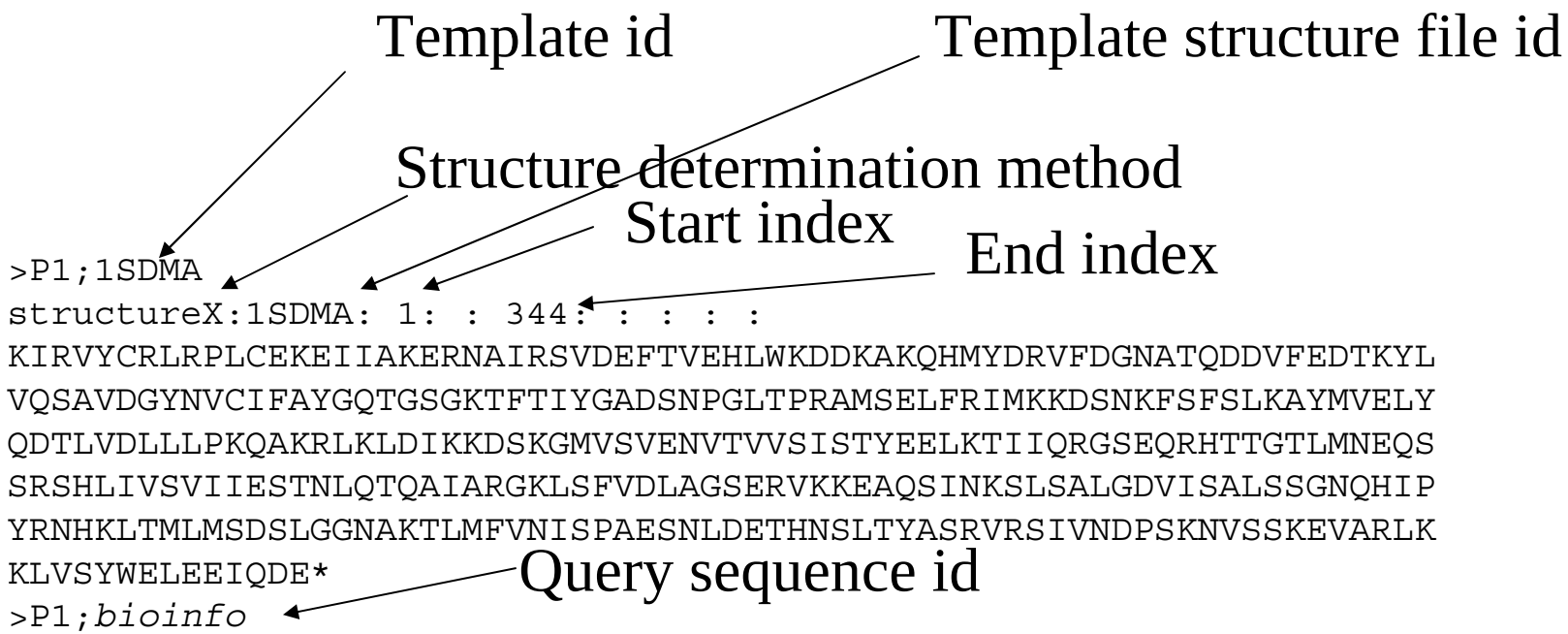
ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE

A. Fisher, 2005

How to use Modeller

- Need an alignment file between query and template sequence in the PIR format
- Need the structure (atom coordinates) file of template protein
- You need to write a simple script (Python for version 8.2) to tell how to generate the model and where to find the alignment file and template structure file.
- Run Modeller on the script. Modeller will automatically copy coordinates and make necessary adjustments to generate a model.
- See project step 5-8 for more details.

An PIR Alignment Example



Structure File Example (1SDMA.atm)

```
ATOM      1  N   LYS     1      -3.978  26.298 113.043  1.00 31.75           N
ATOM      2  CA  LYS     1      -4.532  25.067 113.678  1.00 31.58           C
ATOM      3  C   LYS     1      -5.805  25.389 114.448  1.00 30.38           C
ATOM      4  O   LYS     1      -6.887  24.945 114.072  1.00 32.68           O
ATOM      5  CB  LYS     1      -3.507  24.446 114.631  1.00 34.97           C
ATOM      6  CG  LYS     1      -3.743  22.970 114.942  1.00 36.49           C
ATOM      7  CD  LYS     1      -3.886  22.172 113.644  1.00 39.52           C
ATOM      8  CE  LYS     1      -3.318  20.766 113.761  1.00 41.58           C
ATOM      9  NZ  LYS     1      -1.817  20.761 113.756  1.00 43.48           N
ATOM     10  N   ILE     2      -5.687  26.161 115.522  1.00 26.16           N
ATOM     11  CA  ILE     2      -6.867  26.500 116.302  1.00 22.75           C
ATOM     12  C   ILE     2      -7.887  27.226 115.439  1.00 21.35           C
ATOM     13  O   ILE     2      -7.565  28.200 114.770  1.00 20.95           O
ATOM     14  CB  ILE     2      -6.513  27.377 117.523  1.00 21.68           C
ATOM     15  CG1 ILE     2      -5.701  26.563 118.526  1.00 21.13           C
ATOM     16  CG2 ILE     2      -7.782  27.875 118.200  1.00 18.96           C
ATOM     17  CD1 ILE     2      -5.368  27.325 119.787  1.00 21.39           C
ATOM     18  N   ARG     3      -9.120  26.737 115.461  1.00 22.04           N
ATOM     19  CA  ARG     3     -10.214  27.327 114.693  1.00 23.95           C
ATOM     20  C   ARG     3     -10.783  28.563 115.400  1.00 22.82           C
ATOM     21  O   ARG     3     -10.771  28.645 116.629  1.00 22.62           O
ATOM     22  CB  ARG     3     -11.327  26.290 114.510  1.00 26.34           C
ATOM     23  CG  ARG     3     -11.351  25.586 113.161  1.00 30.68           C
ATOM     24  CD  ARG     3     -10.004  25.034 112.771  1.00 35.43           C
ATOM     25  NE  ARG     3     -10.104  24.072 111.672  1.00 43.37           N
ATOM     26  CZ  ARG     3     -10.575  24.350 110.458  1.00 46.04           C
ATOM     27  NH1 ARG     3     -10.997  25.572 110.168  1.00 48.68           N
ATOM     28  NH2 ARG     3     -10.627  23.400 109.532  1.00 48.37           N
ATOM     29  N   VAL     4     -11.278  29.524 114.630  1.00 20.49           N
ATOM     30  CA  VAL     4     -11.853  30.724 115.225  1.00 17.59           C
ATOM     31  C   VAL     4     -13.082  31.211 114.471  1.00 18.31           C
ATOM     32  O   VAL     4     -13.030  31.446 113.264  1.00 16.37           O
ATOM     33  CB  VAL     4     -10.834  31.872 115.272  1.00 19.94           C
ATOM     34  CG1 VAL     4     -11.512  33.168 115.759  1.00 15.64           C
ATOM     35  CG2 VAL     4      -9.668  31.489 116.168  1.00 15.45           C
```

Modeller Python Script
(bioinfo.py)

```
# Homology modelling by the automodel class

from modeller.automodel import *    # Load the automodel class

log.verbose()    # request verbose output
env = environ()  # create a new MODELLER environment to build this model in

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

a = automodel(env,
              alnfile  = 'bioinfo.pir',  # alignment filename
              knowns   = '1SDMA',        # codes of the templates
              sequence = 'bioinfo')      # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 1                 # index of the last model
                                    # (determines how many models to calculate)
a.make()                            # do the actual homology modelling
```

Where to find structure file

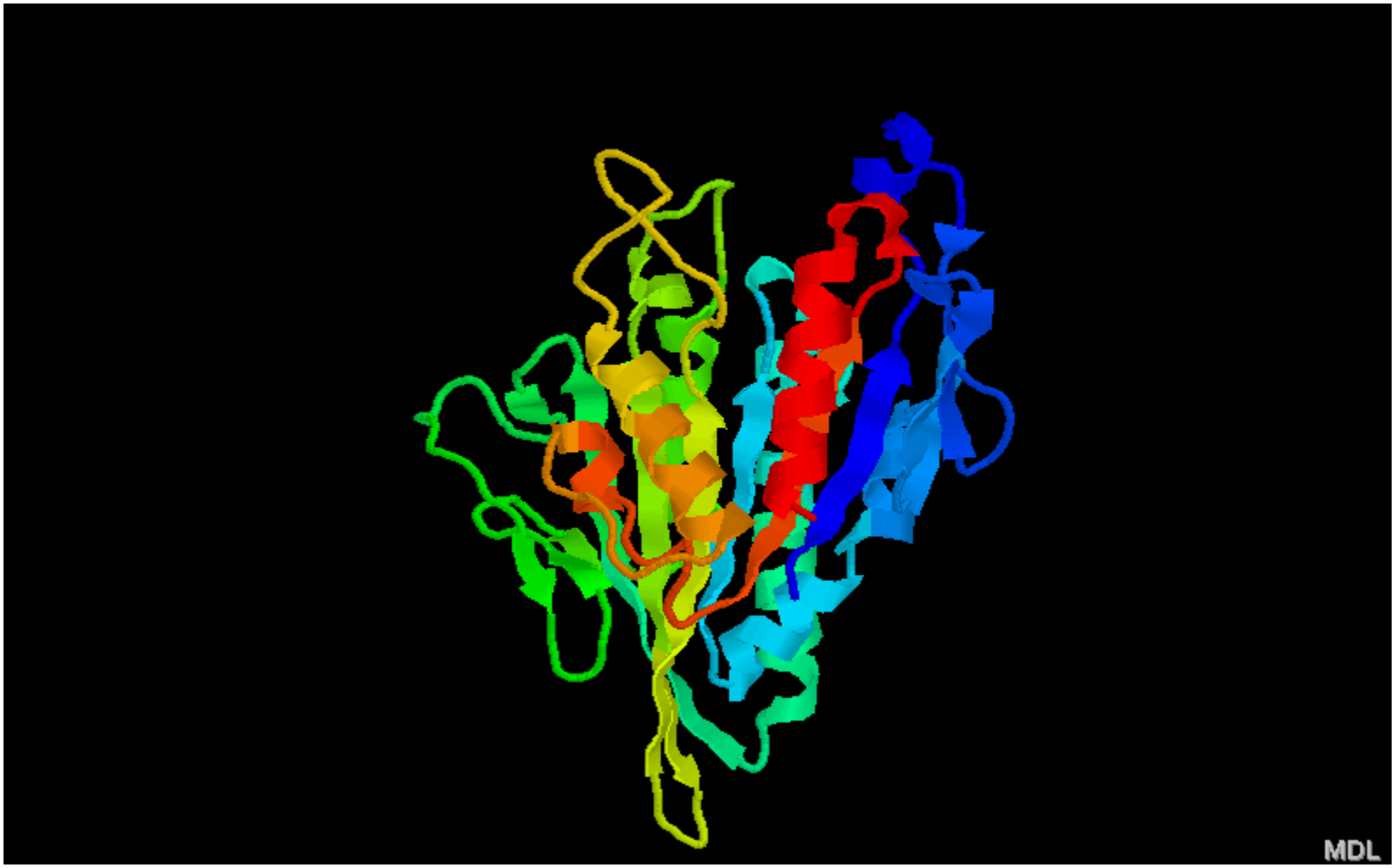
PIR alignment file name

Template structure file id

Query sequence id

Output Example

Command: mod8v2 bioinfo.py



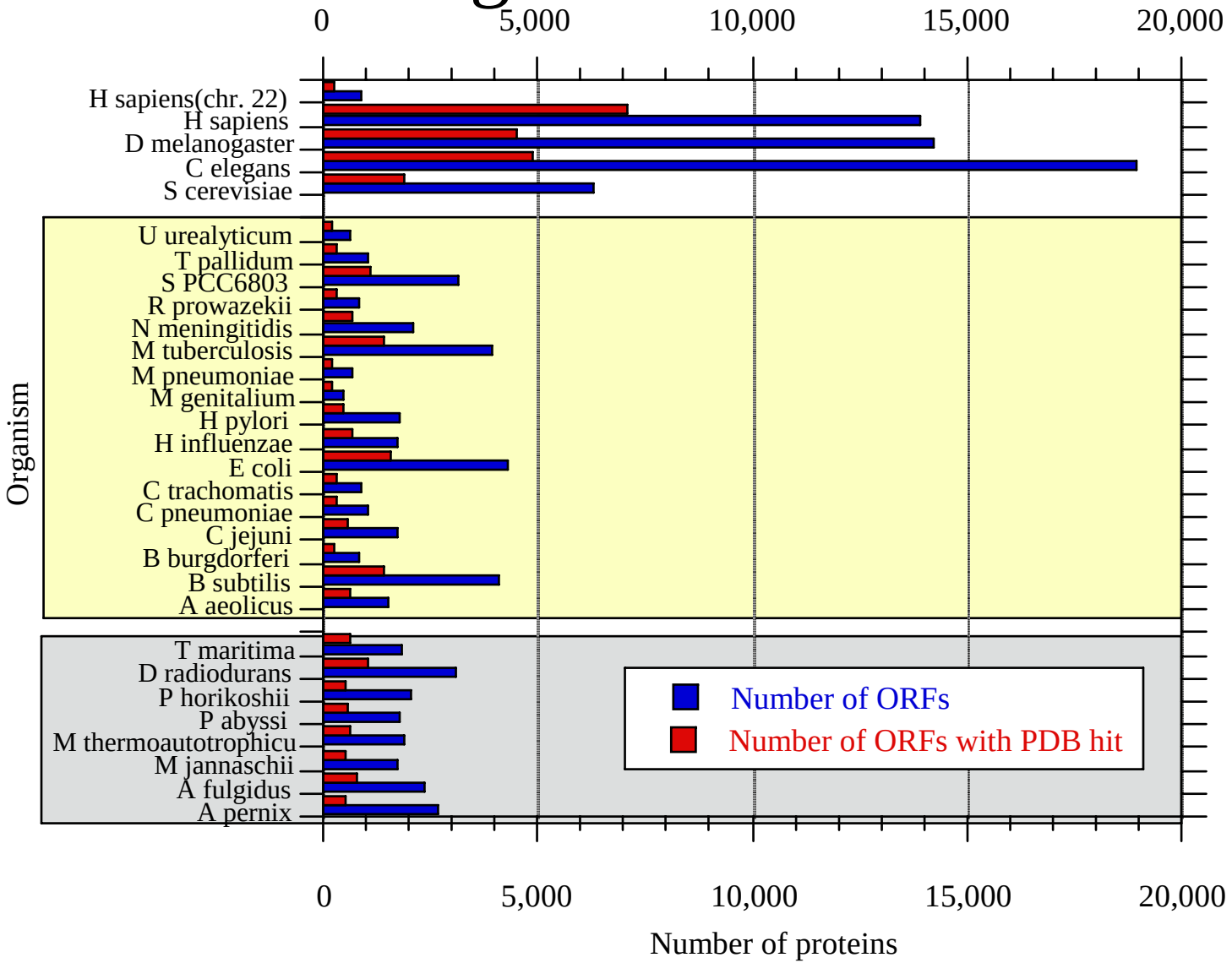
Model Evaluation

- Prosa (http://www.came.sbg.ac.at/Services/prosa.html)
- Verify-3D (http://nihserver.mbi.ucla.edu/Verify_3D/)
- ProCheck (http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html)
- Machine learning approach
- Increasingly important area. A new category in CASP7 – Quality Assurance)

Model Refinement

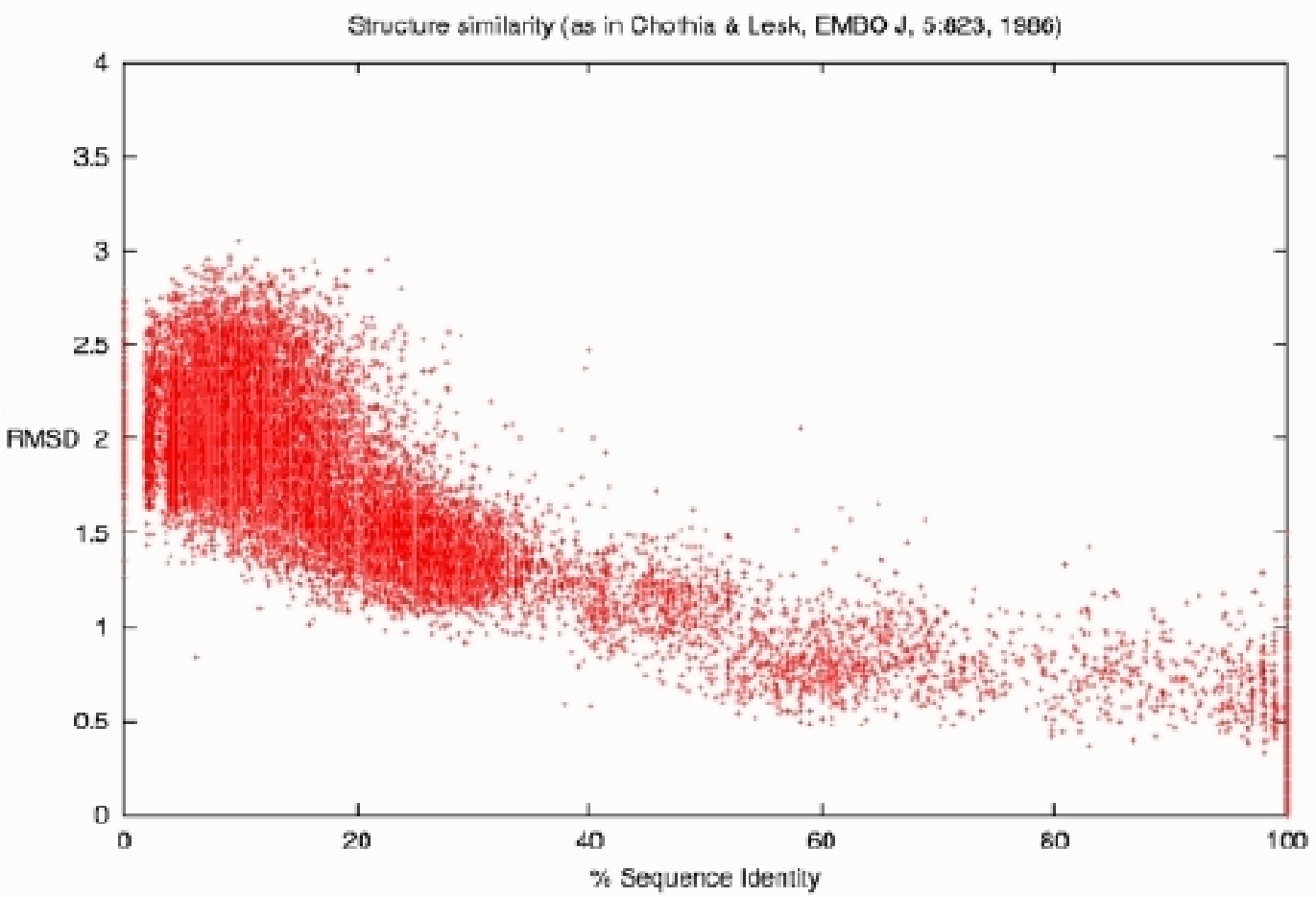
- A new emerging area
- How to refine a moderately accurate model (3-4 Angstrom) to close to native (1-2 Angstrom)?
- A new competition / evaluation category in CASP7

Homology modelling for entire genomes



B. Rost, 2005

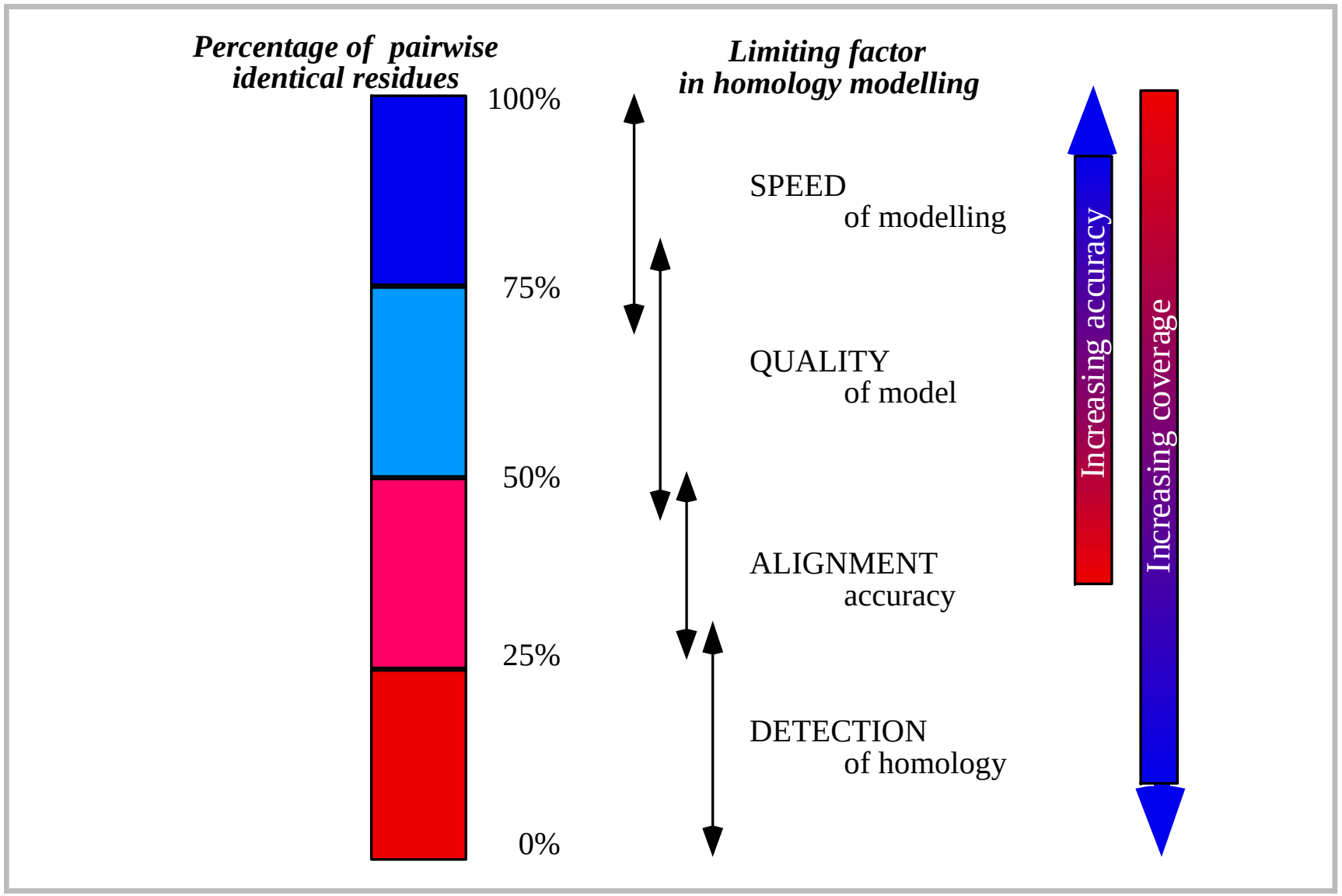
Sequence Identity and Alignment Quality in Structure Prediction



Superimpose
-> RMSD

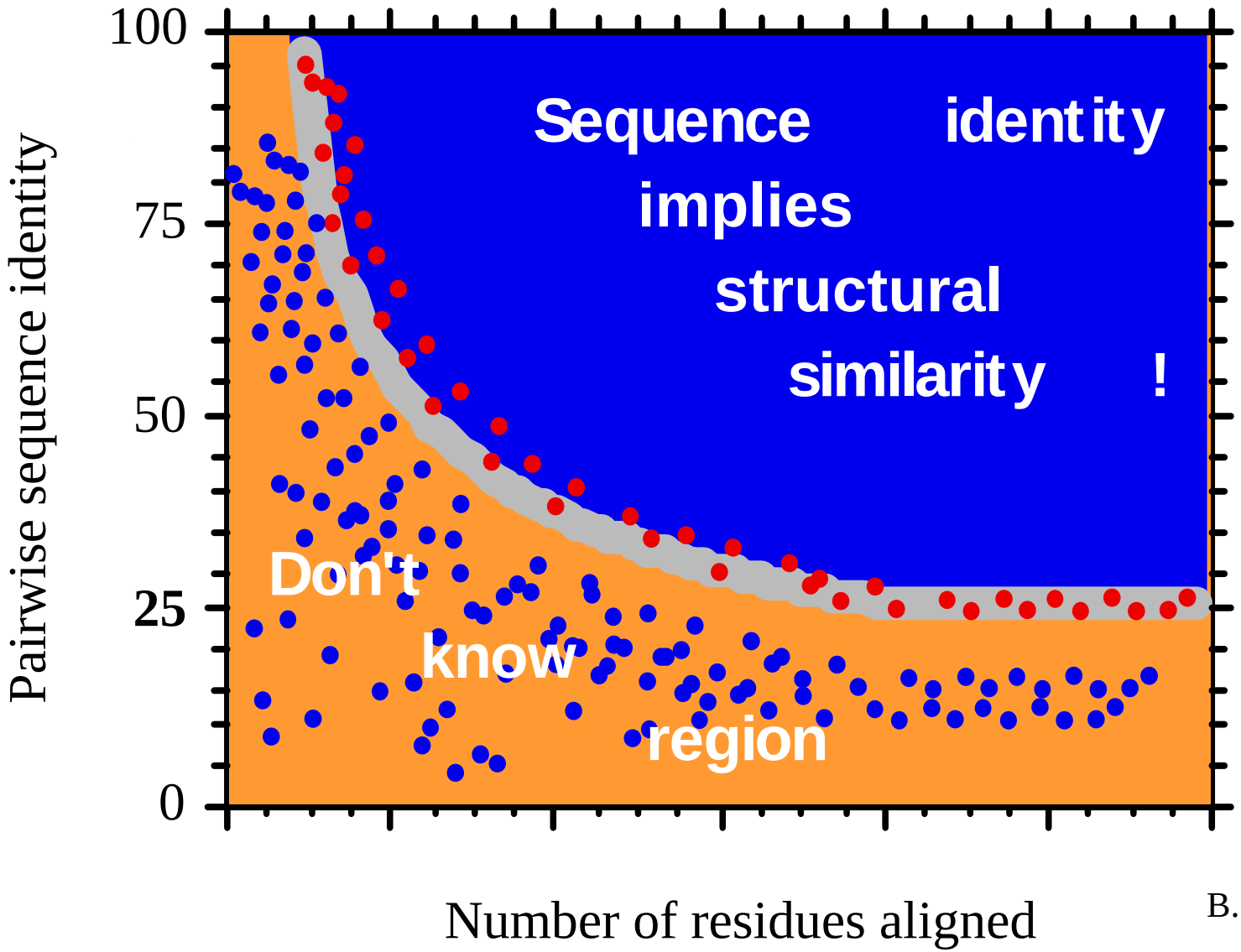
%Sequence Identity: percent of identical residues in alignment
RMSD: square root of average distance between predicted structure and native structure.

Comparative modelling: quality



B. Rost, 2005

What we are threading for

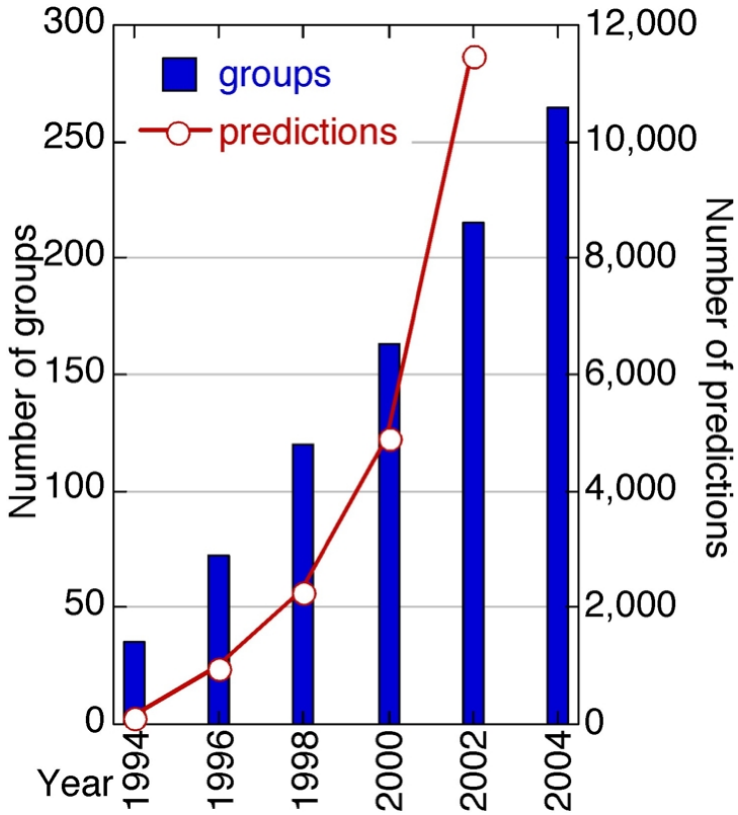


B. Rost, 2005

Protein Structure Prediction

- Only homology /comparative modelling good
- No general prediction of 3D from sequence, yet
- Important improvements in many fields!

CASP



B. Rost, 2005

On 28 high accuracy modeling targets, my servers FOLDpro and 3Dpro are ranked 2nd and 3rd in the seventh edition of Critical Assessment of Techniques for Protein Structure Prediction. (CASP7)

Top **20** servers out of **67** servers that participated in CASP7 worldwide.

Official CASP7 Results

http://zhang.bioinformatics.ku.edu/casp7/22.html

Rank	Predictor	Score
1	Zhang-server	24.44
2	**FOLDpro**	**24.20**
3	**3Dpro**	**24.04**
4	UNI-EID_expm	23.97
5	CIRCLE	23.79
6	RAPTOR	23.75
7	ROBETTA	23.66
8	FAMS	23.63
9	Beautshotbase	23.62
10	Pcons6	23.61
11	Beautshot	23.60
12	RAPTOR-ACE	23.57
13	HHpred1	23.54
14	SPARKS2	23.52
15	SP3	23.51
16	FAMSD	23.51
17	SP4	23.51
18	FUNCTION	23.51
19	Shub	23.47
20	RAPTORESS	23.42

Outline

I. Sequence, Structure, Function Relation

II. Determination, Storage, Visualization, and Comparison

III. Structure Classification

IV. 1D Prediction

V. 2D Prediction

VI. 3D Prediction (emphasis)

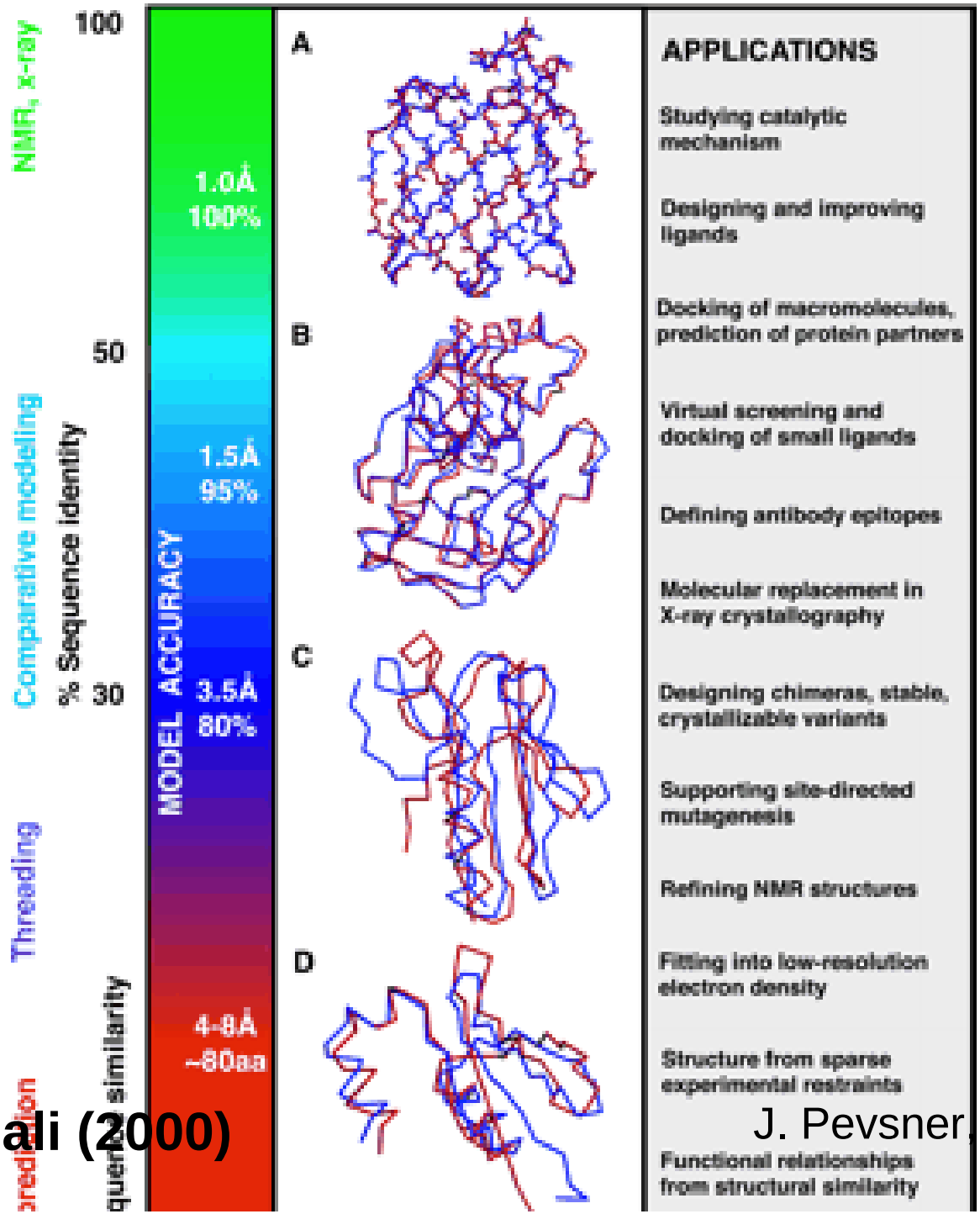
VII. Tools and Applications

3D Structure Prediction Tools

- 3D-Jury (http://bioinfo.pl/Meta/)
- FFAS (http://ffas.ljcrf.edu/ffas-cgi/cgi/ffas.pl)
- Sparks (http://phyyz4.med.buffalo.edu/hzhou/anonymous-fold-sp3.html)
- FUGUE (http://www-cryst.bioc.cam.ac.uk/%7Efugue/prfsearch.html)
- HHpred (http://protevo.eb.tuebingen.mpg.de/toolkit/index.php?view=hhpred)
- FOLDpro (http://mine5.ics.uci.edu:1026/foldpro.html)
- Robetta (http://robetta.bakerlab.org/)
- SAM (http://www.cse.ucsc.edu/research/compbio/sam.html)
- Phyre (http://www.sbg.bio.ic.ac.uk/~phyre/)
- 3D-PSSM (http://www.sbg.bio.ic.ac.uk/3dpssm/)
- mGenThreader (http://bioinf.cs.ucl.ac.uk/psipred/psiform.html)

Application of Structure Prediction

- Structure prediction is improving, but not very fast
- Template-based structure become more and more practical. Particularly, comparative / homology modeling is pretty accurate in many cases.
- Comparative modeling has been widely used in drug design.
- Protein structure prediction (both secondary and tertiary) has become an indispensable tool of investigating function of proteins and mechanisms of biological processes.



Baker and Sali (2000)

J. Pevsner, 2005

Ten Topics

- 1. Introduction to Molecular Biology and Bioinformatics
- 2. Pairwise Sequence Alignment Using Dynamic Programming
- 3. Practical Sequence/Profile Alignment Using Fast Heuristic Methods (BLAST and PSI-BLAST)
- 4. Multiple Sequence Alignment
- 5. Gene Identification
- 6. Phylogenetic Analysis
- **7. Protein Structure Analysis and Prediction**
- 8. RNA Secondary Structure Prediction
- 9. Clustering and Classification of Gene Expression Data
- 10. Search and Mining of Biological Databases, Databanks, and Literature