Bioinformatique M2: Lecture 4 - part B

P. Derreumaux

III. From protein sequence to 3D structure

The CASP experiment

- *CASP= Critical Assessment of Structure Prediction*

- *Started in 1994, based on an idea from John Moult (Moult, Pederson, Judson, Fidelis, Proteins, 23:2-5 (1995))*

- *First run in 1994; now runs regularly every second year (CASP7 was held last december)*

The CASP experiment: how it works

1) Sequences of target proteins are made available to CASP participants in June-July of a CASP year

- the structure of the target protein is know, but not yet released in the PDB, or even accessible

2) CASP participants have between 2 weeks and 2 months over the summer of a CASP year to generate up to 5 models for each of the target they are interested in.

3) Model structures are assessed against experimental structure

4) CASP participants meet in December to discuss results

CASP

Three categories at CASP

- Homology (or comparative) modeling

- Fold recognition

- Ab initio or de Novo prediction

CASP dynamics:

- Real deadlines; pressure: positive, or negative?

- Competition?

- Influence on science ?

Venclovas, Zemla, Fidelis, Moult. Assessment of progress over the CASP experiments.
Proteins, 53:585-595 (2003)

EVOLVING IDEAS

- **Used to be:**

Secondary structure

Molecular Dynamics

Folding pathways

Fold recognition

- **Now is:**

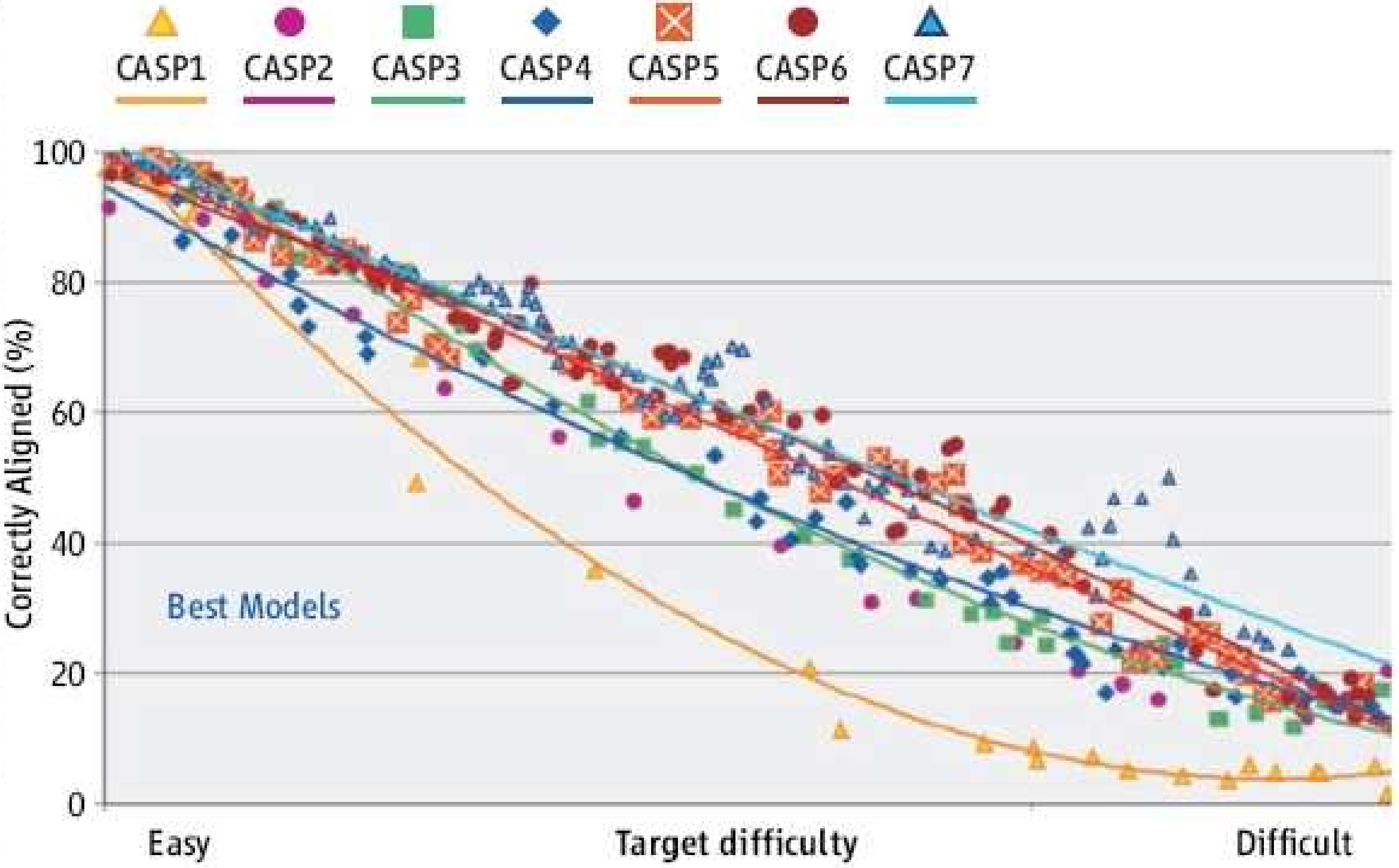
Profiles

Multiple templates

Meta-servers

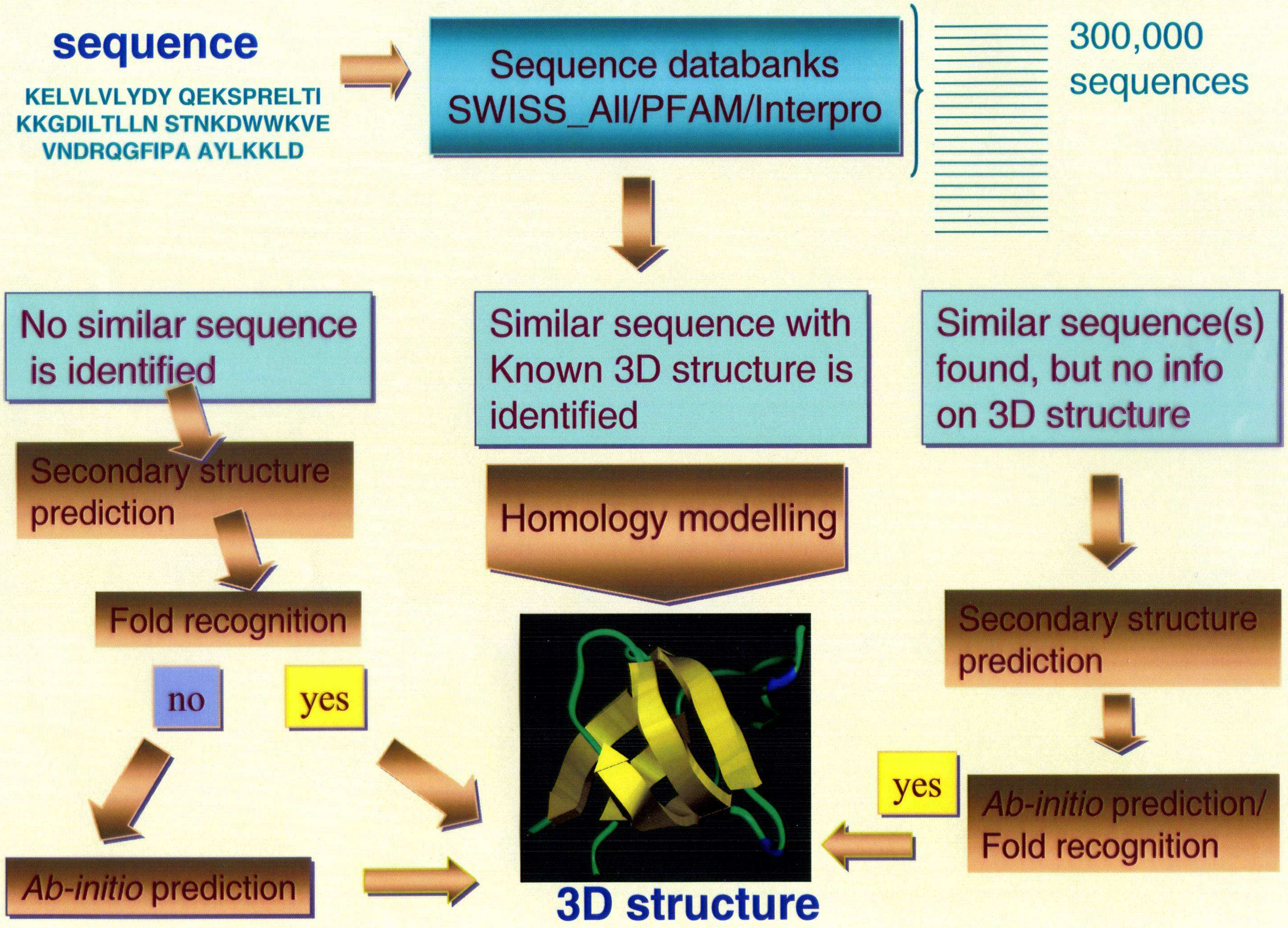
Fragments

Refinement



Steady rise. Computer modelers have slowly but steadily improved the accuracy of the protein-folding models.

Prediction of protein 3D structure



sequence

KELVLVLYDY QEKSPRELTI
KKGDILTLLN STNKDWWKVE
VNDRQGFIPA AYLKKLD

Sequence databanks
SWISS_All/PFAM/Interpro

300,000
sequences

No similar sequence is identified

Secondary structure prediction

Fold recognition

no

yes

Ab-initio prediction

Similar sequence with Known 3D structure is identified

Homology modelling

Similar sequence(s) found, but no info on 3D structure

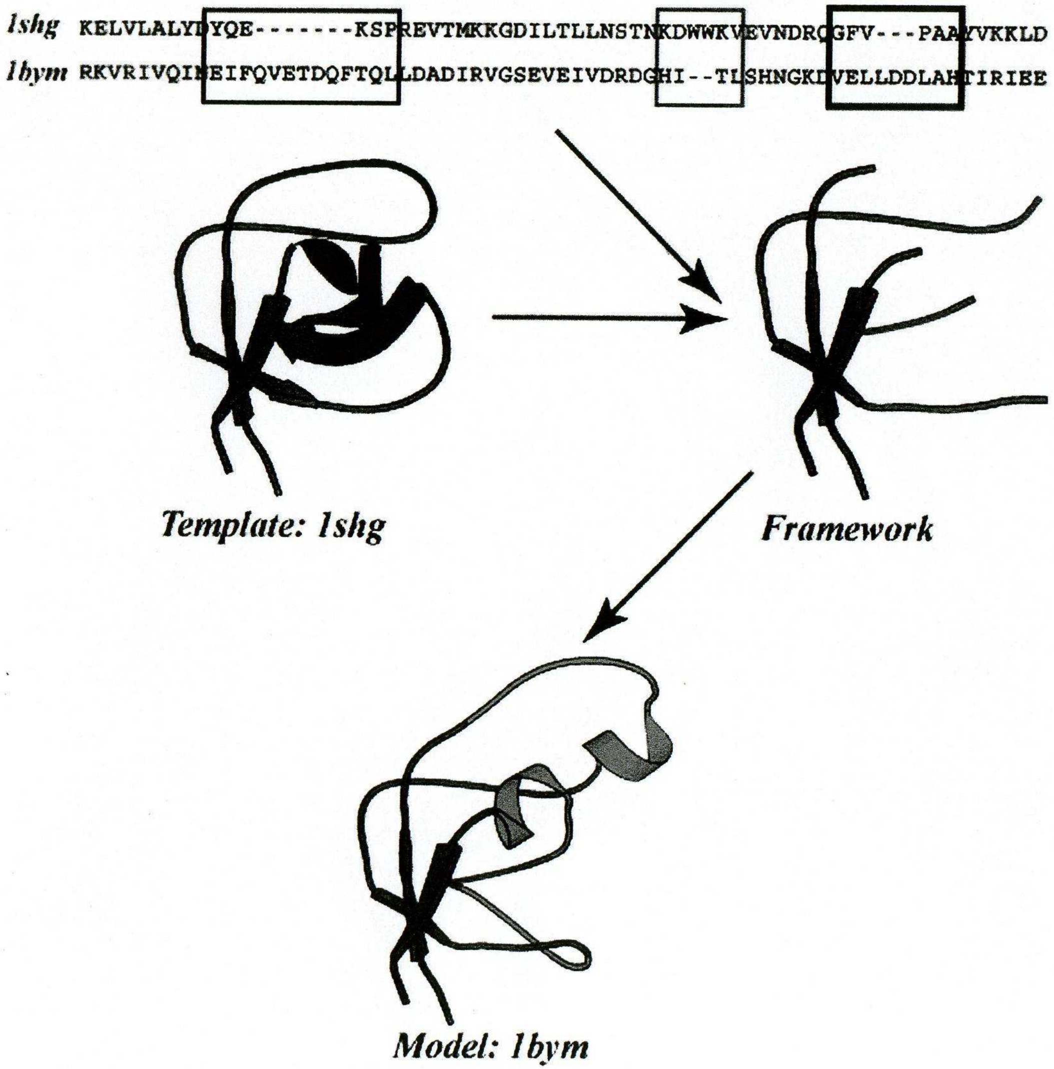
Secondary structure prediction

Ab-initio prediction/ Fold recognition

yes

3D structure

Homology Modeling: How it works

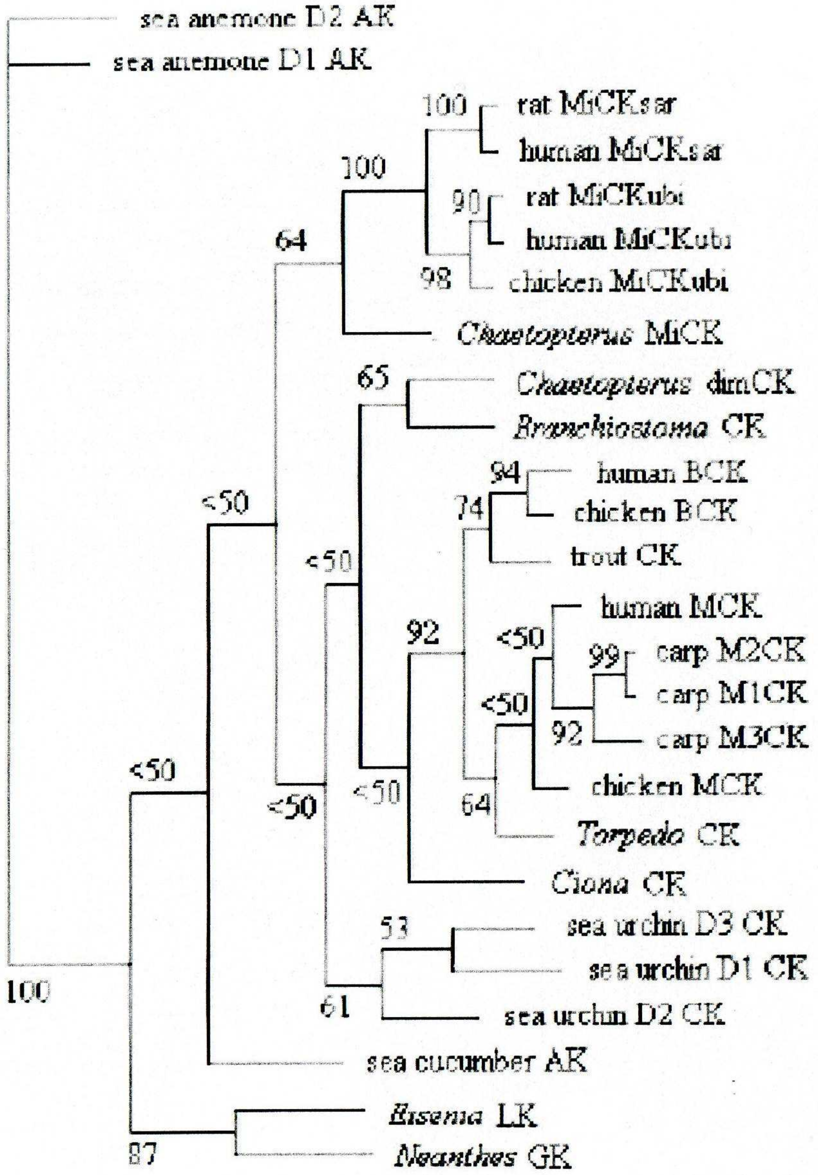


```
1shg KELVLALYDYQE-------KSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFV---PAAYVKKLD
1bym RKVRIVQINEIFQVETDQFTQLLDADIRVGSEVEIVDRDGHI--TLSHNGKDVELLDDLAHTIRIBE
```

- *Find template*
- *Align target sequence with template*
- *Generate model:*
  - *add loops*
  - *add sidechains*
- *Refine model*

Template choice

1. Higher the sequence identity, the more likely the template will be suitable
2. Most closely related from a phylogenetic point of view
3. Template "environment" (solvent, pH, temperature, quaternary structure)
4. Quality of the template structure (resolution and R factor)



Homology modelling

Building the model

MODELLING THE WHOLE FOLD

1. Rigid-body assembly (COMPOSER)
2. Spare-parts approach
3. Satisfaction of spatial restraints (MODELLER)

MODELLING LOOPS

1. Database search of segments fitting fixed end-points
2. Molecular mechanics, molecular dynamics
3. Combination of 1+2

MODELLING SIDE CHAIN CONFORMATIONS

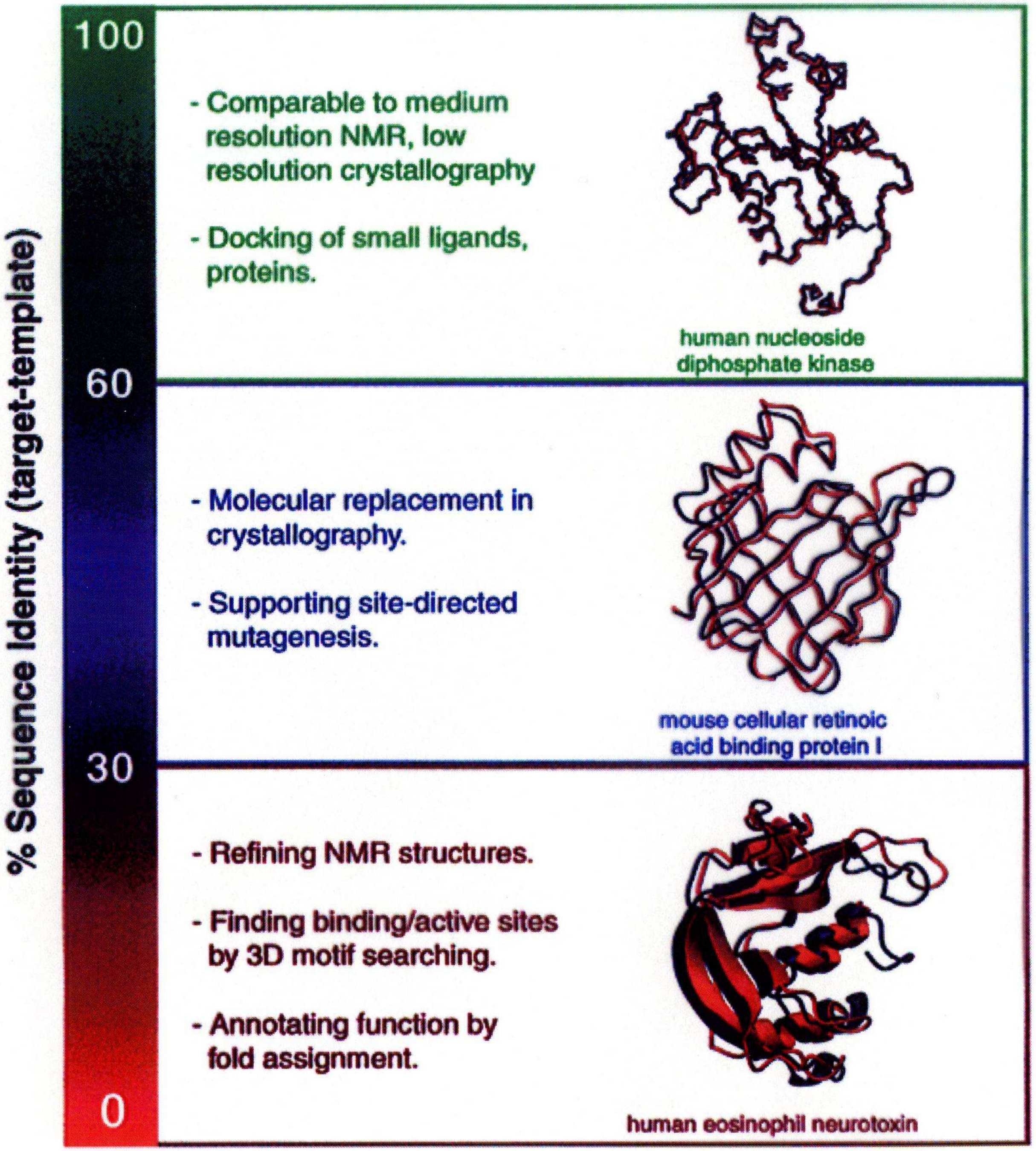
1. Use of rotamer libraries (backbone dependent)
2. Molecular mechanics optimization
3. Mean-field methods

Typical types of errors

- Sequence alignment errors.
- Loops which cannot be rebuilt.
- Inappropriate template selection.
- Subunit displacement.

Structure Modeling by Homology: Limitations

Homology modelling is the method that can be applied to generate reasonable models of protein structure.



- Comparable to medium resolution NMR, low resolution crystallography
- Docking of small ligands, proteins.

human nucleoside diphosphate kinase

- Molecular replacement in crystallography.
- Supporting site-directed mutagenesis.

mouse cellular retinoic acid binding protein I

- Refining NMR structures.
- Finding binding/active sites by 3D motif searching.
- Annotating function by fold assignment.

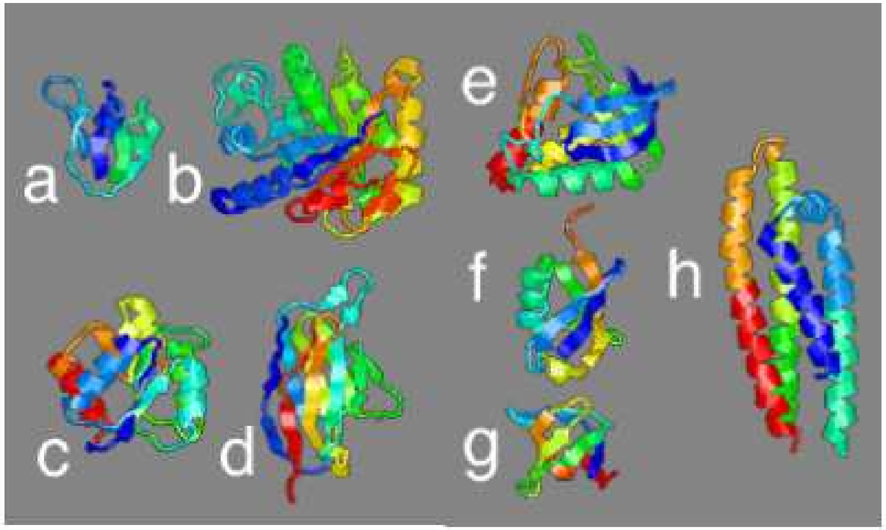
human eosinophil neurotoxin

Fold recognition / Threading

Find a compatible fold for a given sequence ....

```
>Protein XY
MSTLYEKLGGTTAVDLAV
DKFYERVLQDDRIKHFFA
DVDMAKQRAHQKAFLTYA
FGGTDKYDGRYMREAHKE
LVENHGLNGEHFDAVAED
LLATLKEMGVPEDLIAEV
AAVAGAPAHKRDVLNQ
```

$\overset{?}{\approx}$



Number of protein folds that occurs in nature is limited. Fold Recognition can be used to:

- Identify templates for comparative modeling
- Assign Protein Function

5.2. Remote homology modeling = Fold Recognition

- Concept

- 3 families of methods.

(1) Sequence Profiles PSI-BLAST

Ref Dunbrack, Proteins (1999)
Suppl 3 : 81-87.

(very close to comparative modelling)

(2) Profile Searches
or
Fold Recognition with
sequence-derived properties

$3D \xrightarrow{\text{projection}} 1D$, - alignt in seq space (NW)
- Complex Substitution Matrices

(3) Threading = Fold Recognition

3D , - alignt in coord space
- pairwise potentials of mean space.

Rost et al. (1997) J. Mol. Biol. 270:471-480

TOPITS server (PredictProtein Server)

main factor limiting performance:
— Sec. str. pattern degeneracy pb

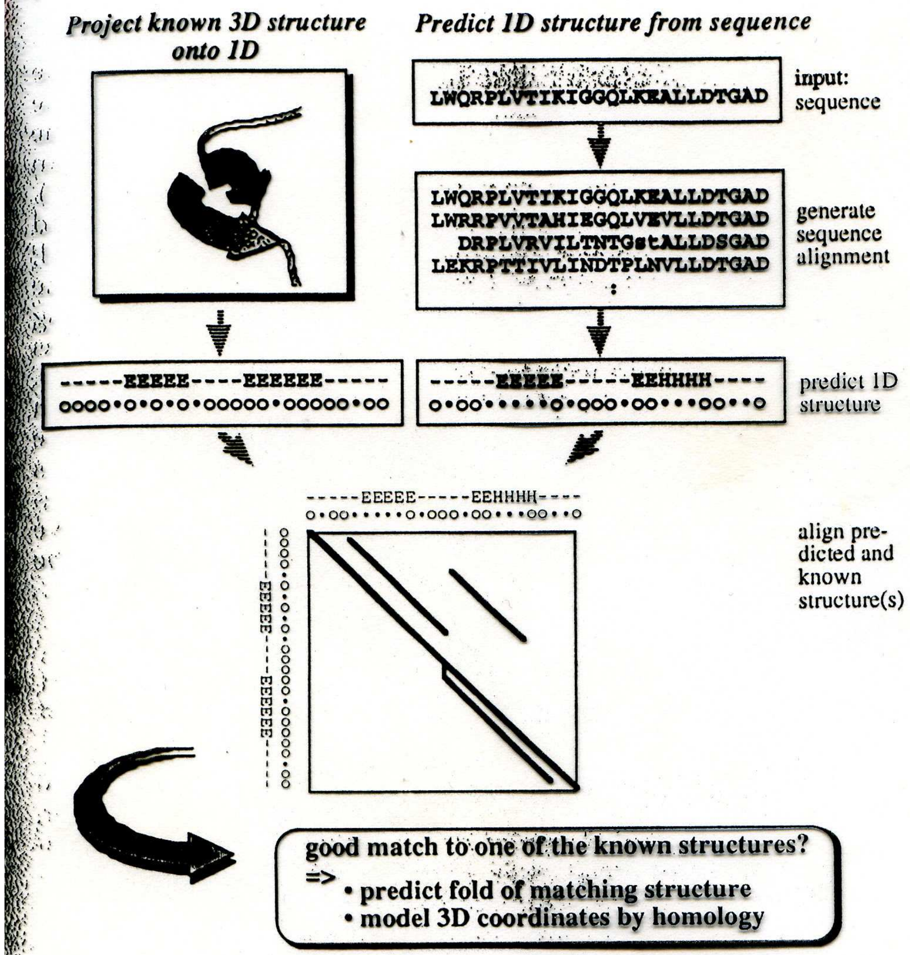

Project known 3D structure onto 1D
Predict 1D structure from sequence
LWQRPLVTIKIGGQLKEALLDTGAD
input: sequence
LWQRPLVTIKIGGQLKEALLDTGAD
LWRRPVVTAHIEGQLVEVLLDTGAD
DRPLVRVILTNTGstALLDSGAD
LEKRPTTIVLINDTPLNVLLDTGAD
generate sequence alignment
-----EEEEE----EEEEEE-----
-----EEEEE-----EEHHHH----
predict 1D structure
align predicted and known structure(s)
good match to one of the known structures?
=>
• predict fold of matching structure
• model 3D coordinates by homology

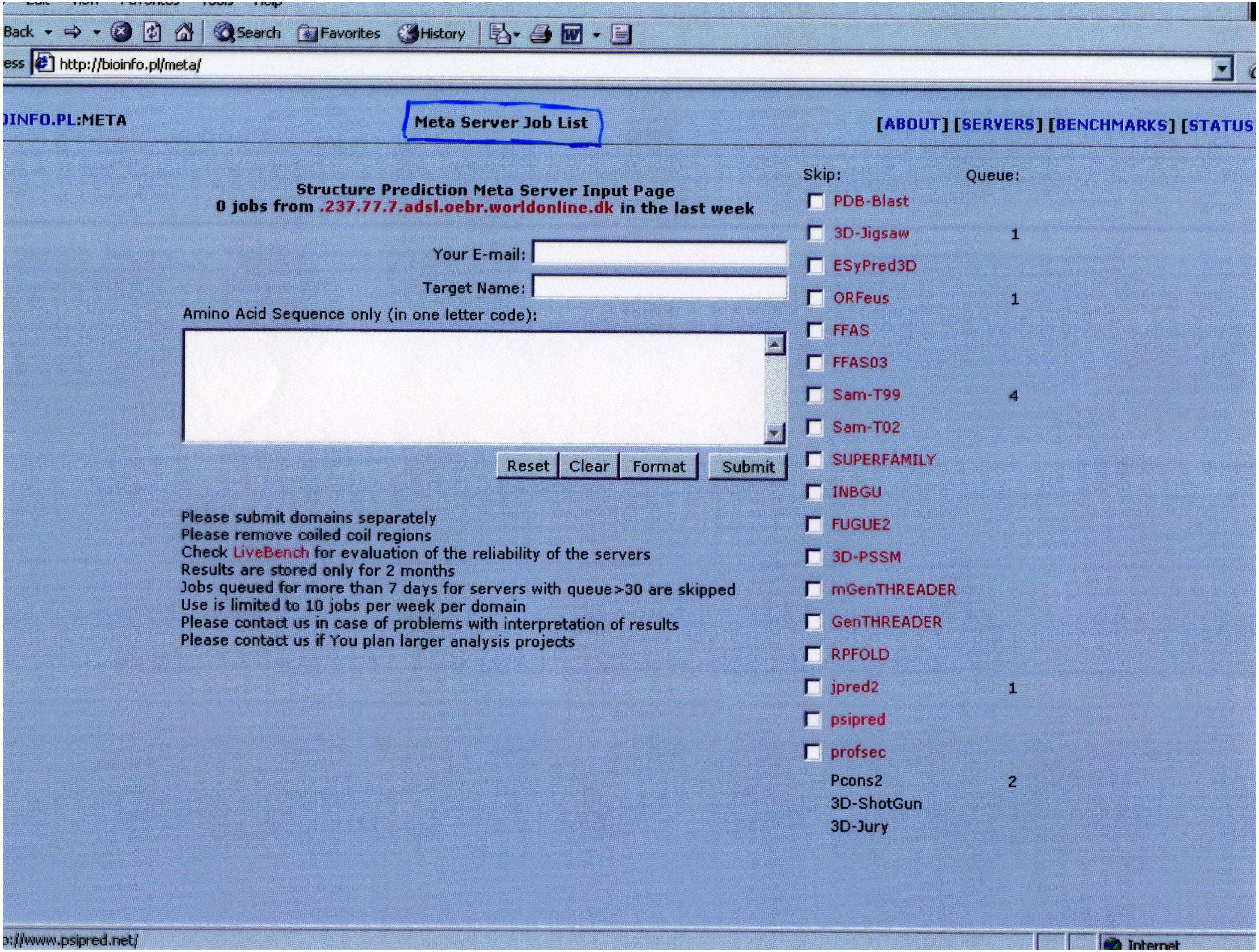
Figure 1. Threading predicted 1D structure profiles into known 3D structures. (1) A multiple sequence alignment is generated for a given sequence of unknown structure (*U*). (2) The alignment profile of *U* is used as the input to a neural network system (PHD) that predicts secondary structure and relative solvent accessibility. (3) The resulting predicted 1D structure profile for *U* is aligned by dynamic programming (program MaxHom; Sander & Schneider, 1991) to 1D structure strings assigned from known structures by the program DSSP (Kabsch & Sander, 1983). Abbreviations: H, helix; E, strand; L, rest; ●, buried (<15% solvent accessible); ○, exposed (≥15% solvent accessible).

Free parameters for dynamic programming

The predicted strings were aligned based on a Smith-Waterman type dynamic programming algorithm (Smith & Waterman, 1981). This algorithm was implemented in the program MaxHom or a Blosum62 (Henikoff & Henikoff, 1992) exchange matrix:

$$M_{ij} = \alpha \times M_{ij}^{\text{1D structure}} + (100 - \mu) \times M_{ij}^{\text{sequence}} \quad (1)$$

where M_{ij} determined the score for a match at a



)INFO.PL:META

Meta Server Job List

[ABOUT] [SERVERS] [BENCHMARKS] [STATUS

Structure Prediction Meta Server Input Page
0 jobs from .237.77.7.adsl.oebr.worldonline.dk in the last week

Your E-mail:

Target Name:

Amino Acid Sequence only (in one letter code):

Reset | Clear | Format | Submit

Please submit domains separately
Please remove coiled coil regions
Check LiveBench for evaluation of the reliability of the servers
Results are stored only for 2 months
Jobs queued for more than 7 days for servers with queue>30 are skipped
Use is limited to 10 jobs per week per domain
Please contact us in case of problems with interpretation of results
Please contact us if You plan larger analysis projects

Skip:	Queue:
□ PDB-Blast	
□ 3D-Jigsaw	1
□ ESyPred3D	
□ ORFeus	1
□ FFAS	
□ FFAS03	
□ Sam-T99	4
□ Sam-T02	
□ SUPERFAMILY	
□ INBGU	
□ FUGUE2	
□ 3D-PSSM	
□ mGenTHREADER	
□ GenTHREADER	
□ RPFOLD	
□ jpred2	1
□ psipred	
□ profsec	
Pcons2	2
3D-ShotGun	
3D-Jury	

But Threading most often does not ~~produce~~ assign the right fold.

Reasons: → the correct fold is not the first of the list but in the 10 top scoring folds
(the correct fold appears to be detected in less than 40% of all benchmark cases)

→ Limited Number of known folds.
(Ref. D. Fischer, D. Eisenberg)
PNAS 1997 94: 11929.

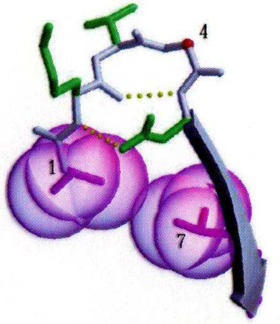
→ **Needs a very similar template structure**

Concl: Looking into the function of the proteins that have been found can help.
(Ref. Murzin Proteins, Suppl 1: 105-112, (1997))

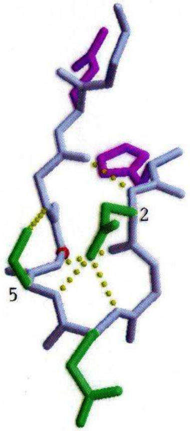
Free modelling: De novo or ab initio

Protein Structure Prediction: Rosetta

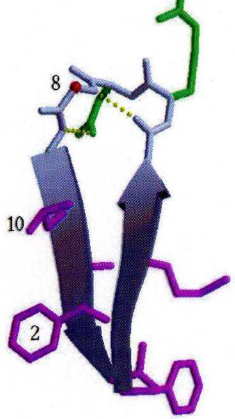
I-sites Library = a catalog of local sequence-structure correlations



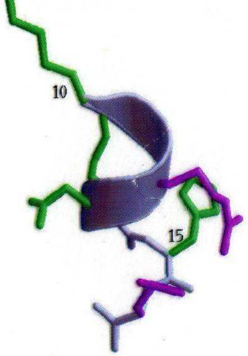
diverging type-2 turn



Serine hairpin

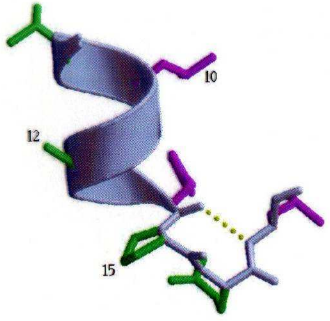


Type-I hairpin

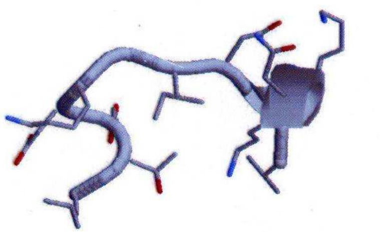


Frayed helix

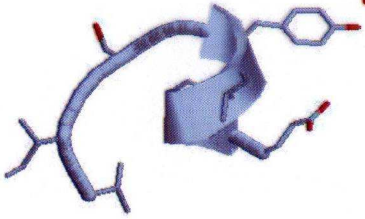
Local
structur
e motifs



Proline helix C-cap

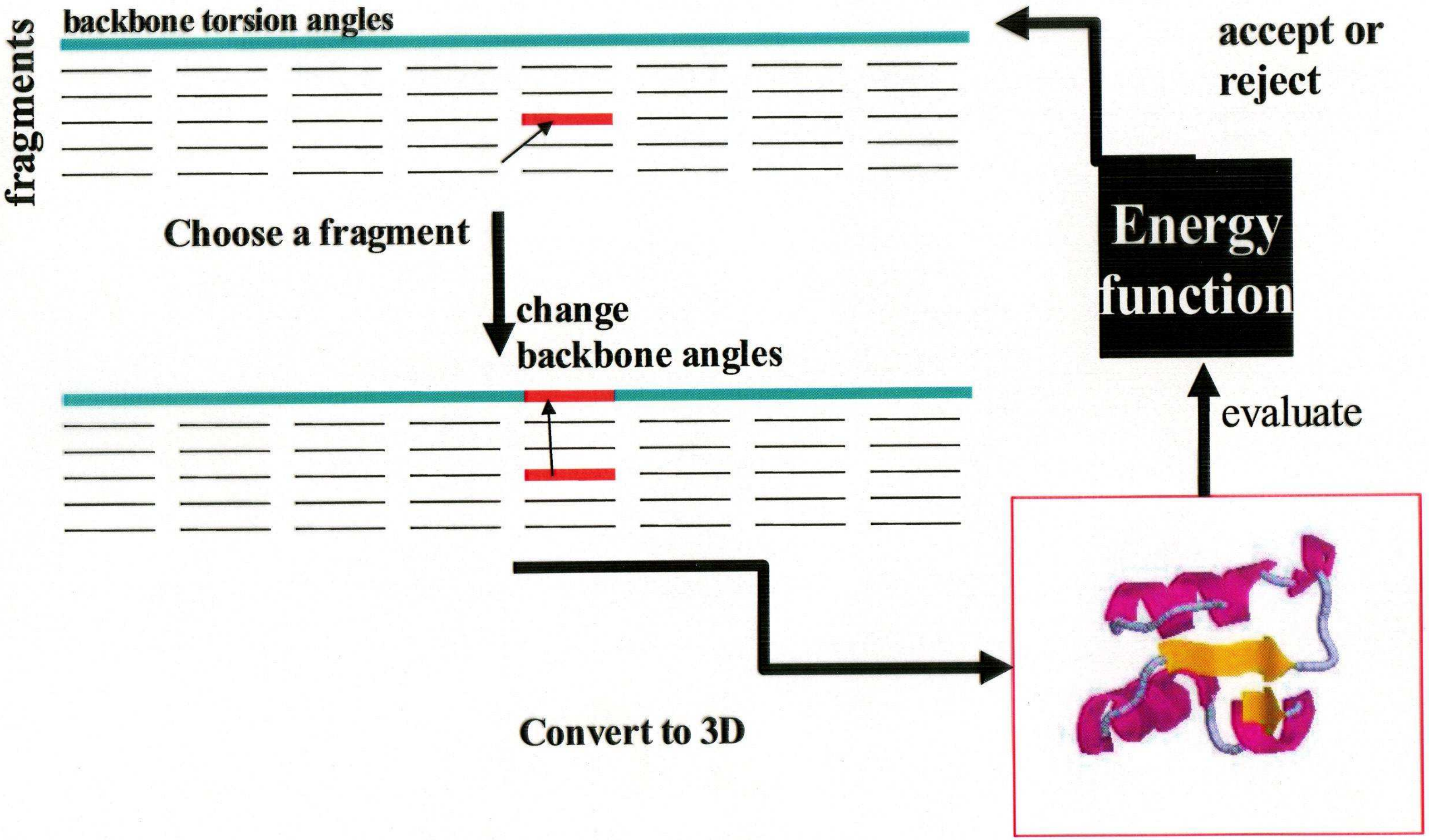


alpha-alpha corner



glycine helix N-cap

Rosetta: a folding simulation program



backbone torsion angles

fragments

Choose a fragment

change backbone angles

Convert to 3D

evaluate

Energy function

accept or reject

Fragment insertion Monte Carlo

Rosetta (Baker) in CASP4

Improvements of the method.

- Combine alternative 2D prediction methods (PSIPRED, SAMT99, PHD) to bias the fragment picking method.

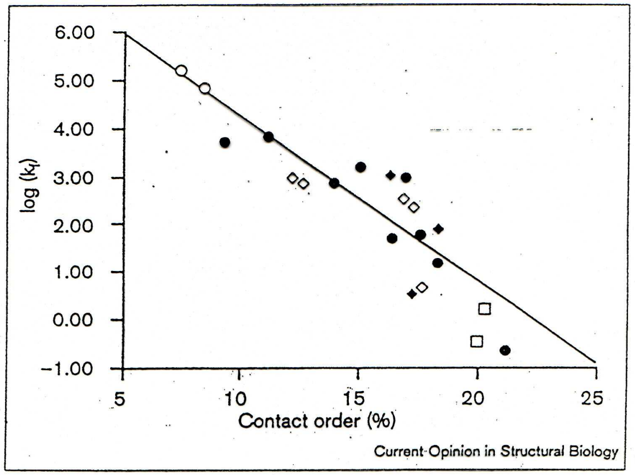
- Filters to eliminate non protein-like structure

  a. poorly formed β-sheets

  b. poorly packed interiors using LJ, Hb and solvation terms

  c. low contact orders.

Plaxco et al. J. Mol. Biol. 277, 985-994 (1998)



Current Opinion in Structural Biology

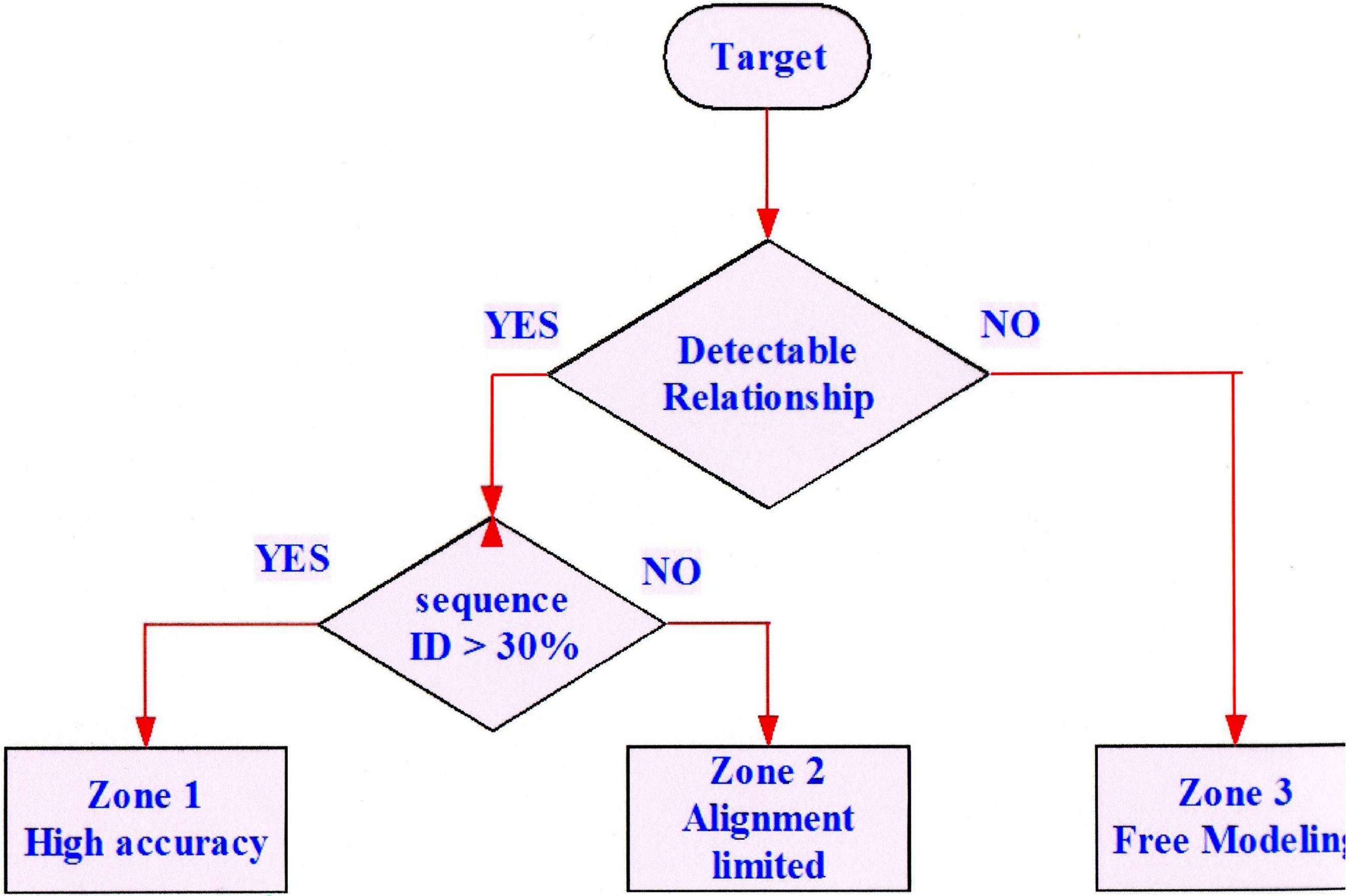
Updated correlation between contact order and the logarithm of the folding rate (log[k$_f$]). Contact order is defined as the average sequence separation between residues that make contact in the native structure divided by the sequence length [13••]. Thus, a contact order of 10% indicates that residue pairs that make contact in the three-dimensional structure are separated by 10% of the length of the protein on average. Circles represent all-helical proteins, squares represent sheet proteins and diamonds represent proteins comprised of both helix and sheet structures. Open points represent proteins characterized after the publication of [13••]. The best-fit line for the original 12-protein data set (filled points) is shown.

$$\% \mathrm{Co} = \frac{100}{LN} \sum^{N} \Delta S_{ij}$$

- Clustering of conformations generated independently for several homologs.

→ In most cases, the largest 5 unique clusters were submitted.

CASP 7 Conclusions.



Target

Detectable Relationship

YES

NO

sequence ID > 30%

YES

NO

Zone 1 High accuracy

Zone 2 Alignment limited

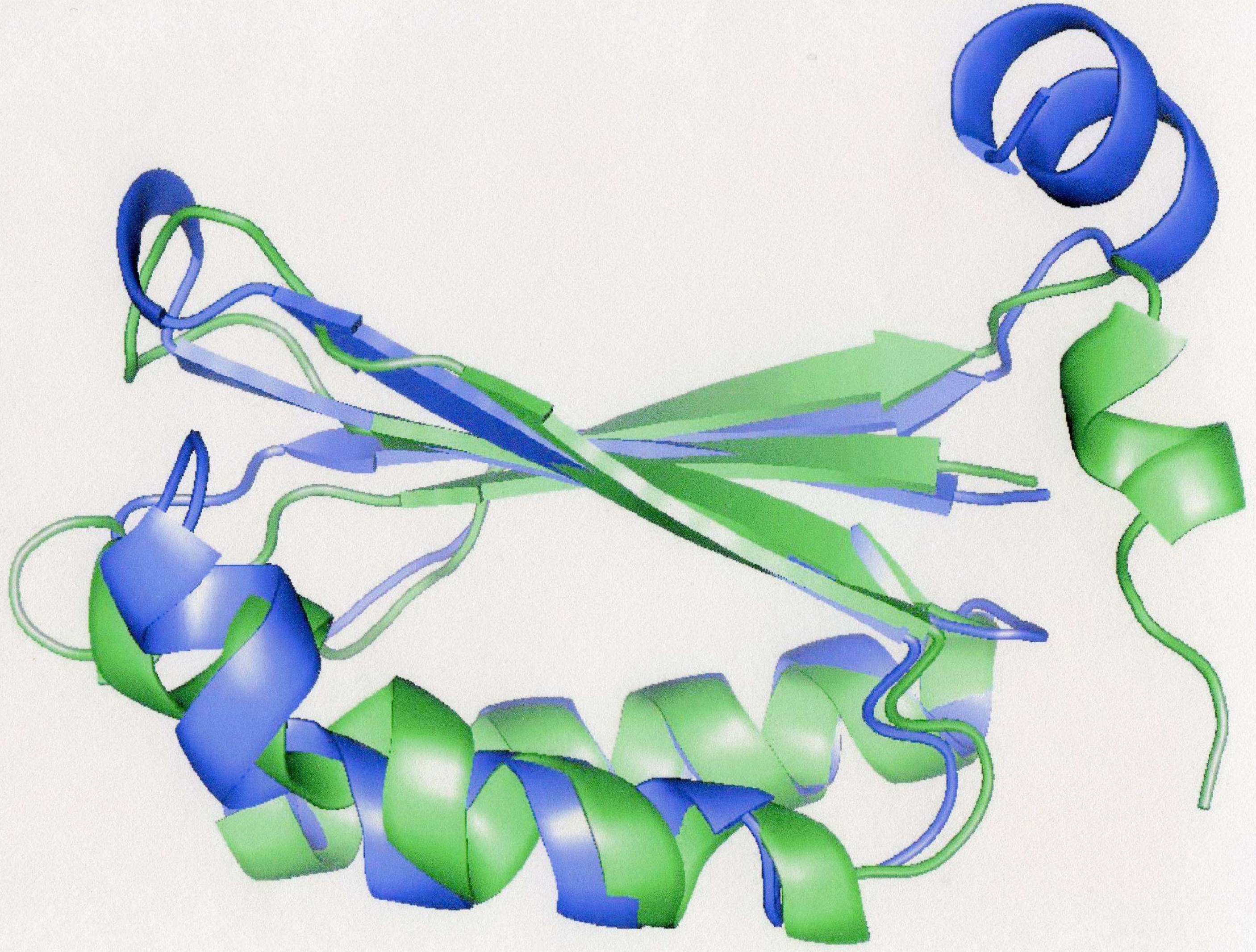
Zone 3 Free Modelin

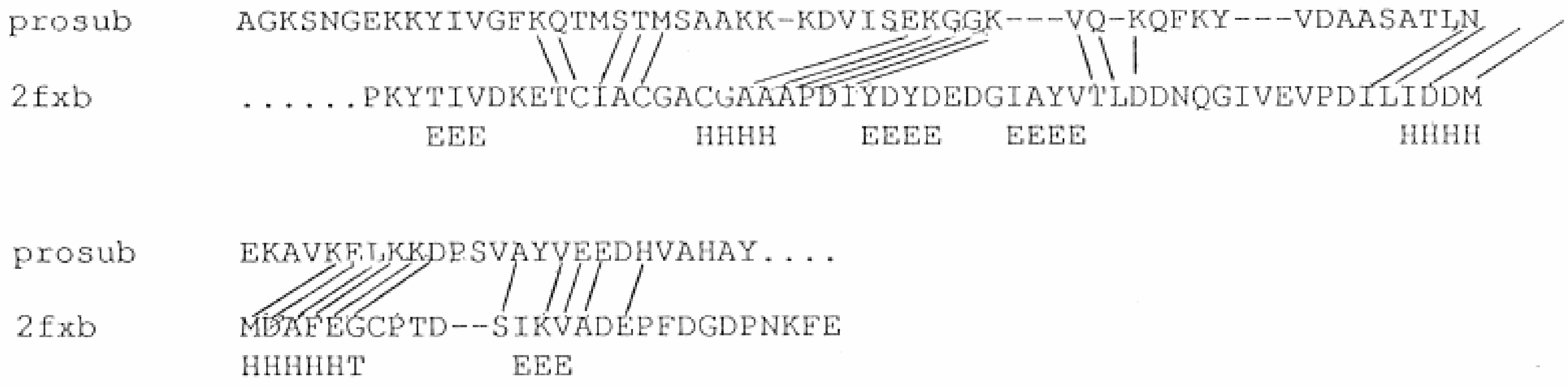
Homology Modeling

Ab Initio. De n

Zone 1: Good models, but not as good as high Resolution models.

Ex. Zone 2

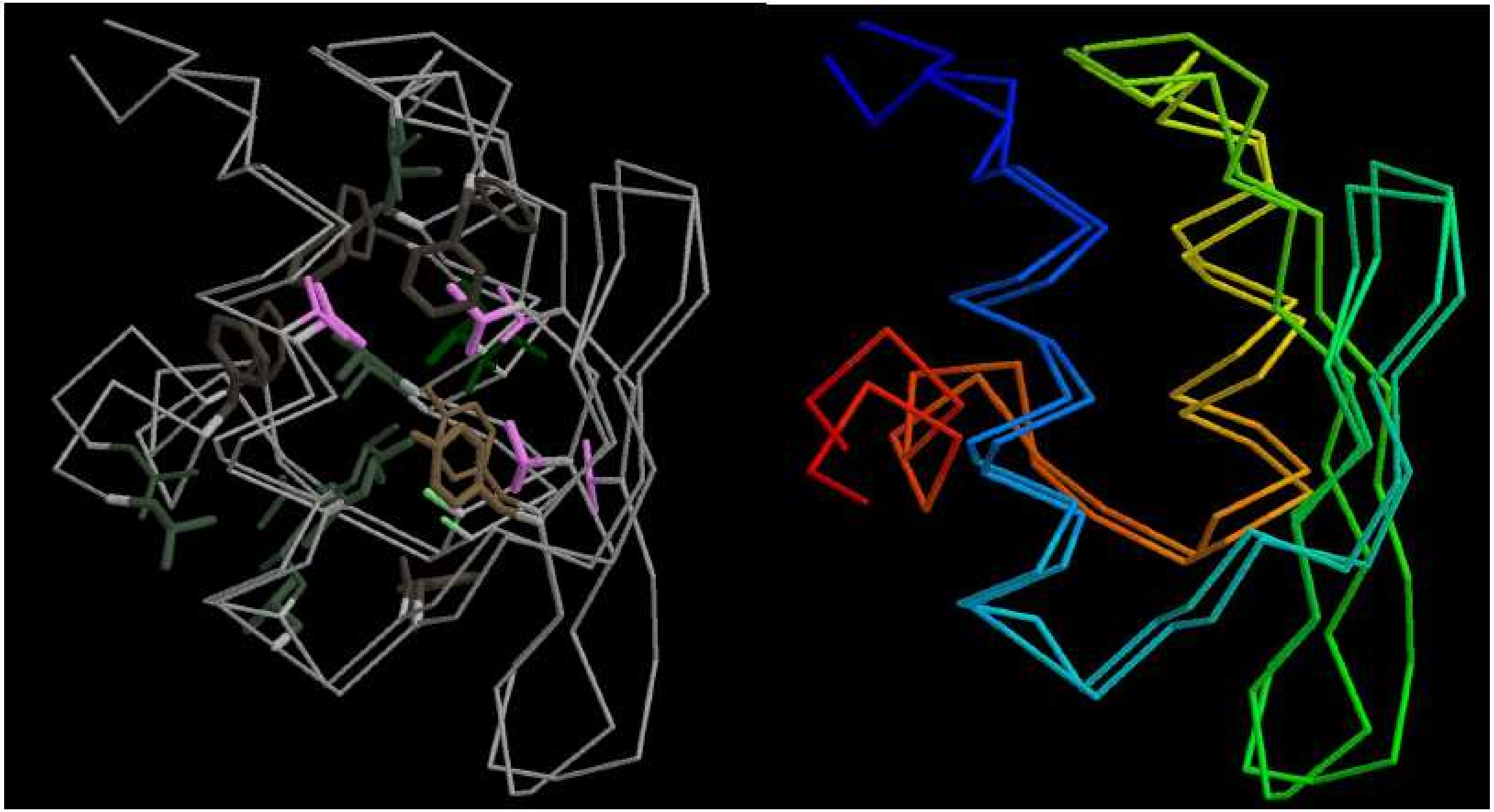


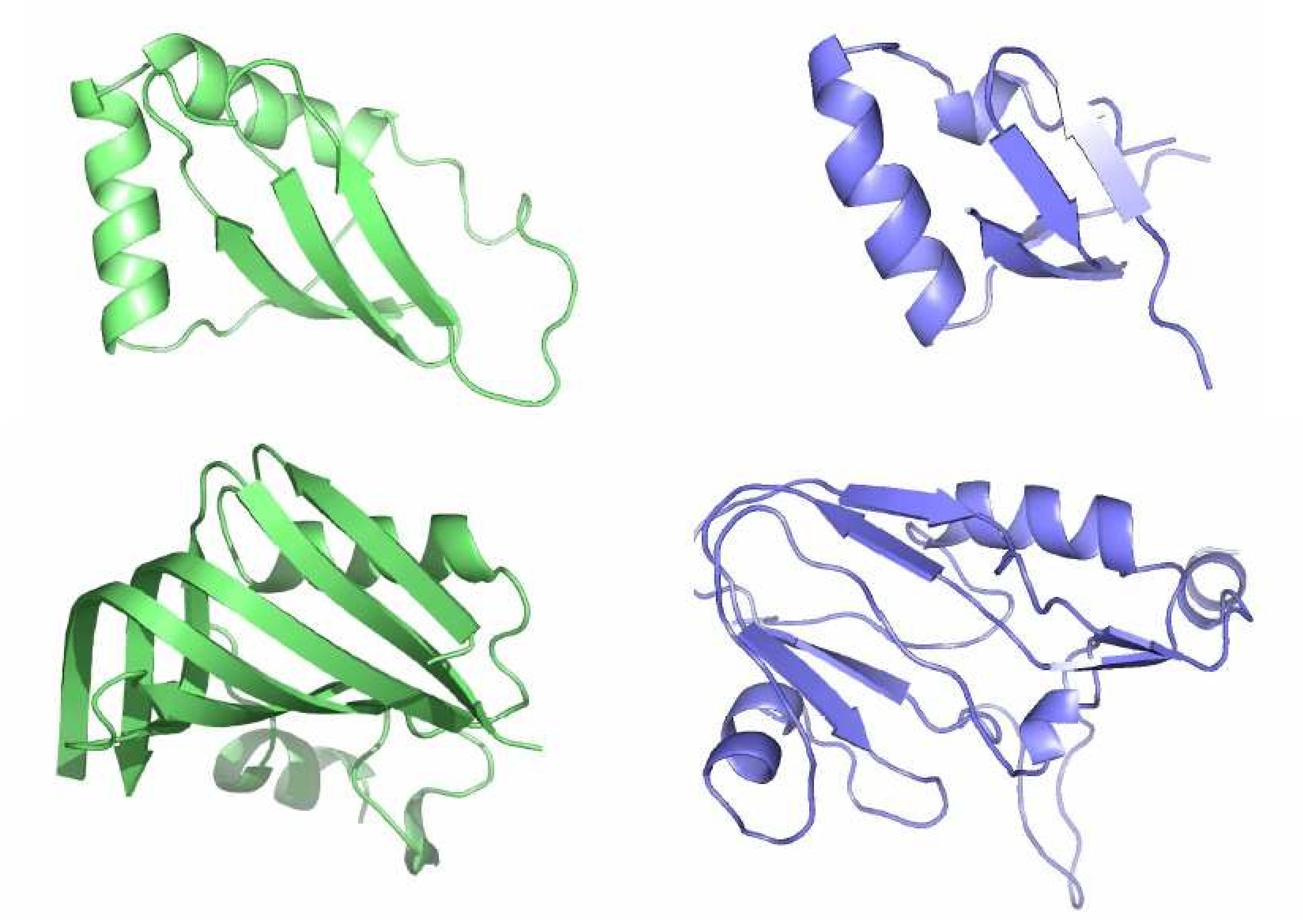

prosub
AGKSNGEKKYIVGFKQTMSTMSAAKK-KDVISEKGGK---VQ-KQFKY---VDAASATLN
2fxb
......PKYTIVDKETCIACGACGAAAPDIYDYDEDGIAYVTLDDNQGIVEVPDILIDDM
EEE
HHHH
EEEE
EEEE
HHHH
prosub
EKAVKELKKDPSVAYVEEDHVAHAY....
2fxb
MDAFEGCPTD--SIKVADEPFDGDPNKFE
HHHHHT
EEE

Zone 2 Conclusions

- Approximate models, but never-the-less valuable.
- Alignment has improved, but still a way to go.
- Further improvement probably requires an all atom description and refinement.
- ‘Free modeling’ needed for non-template parts.

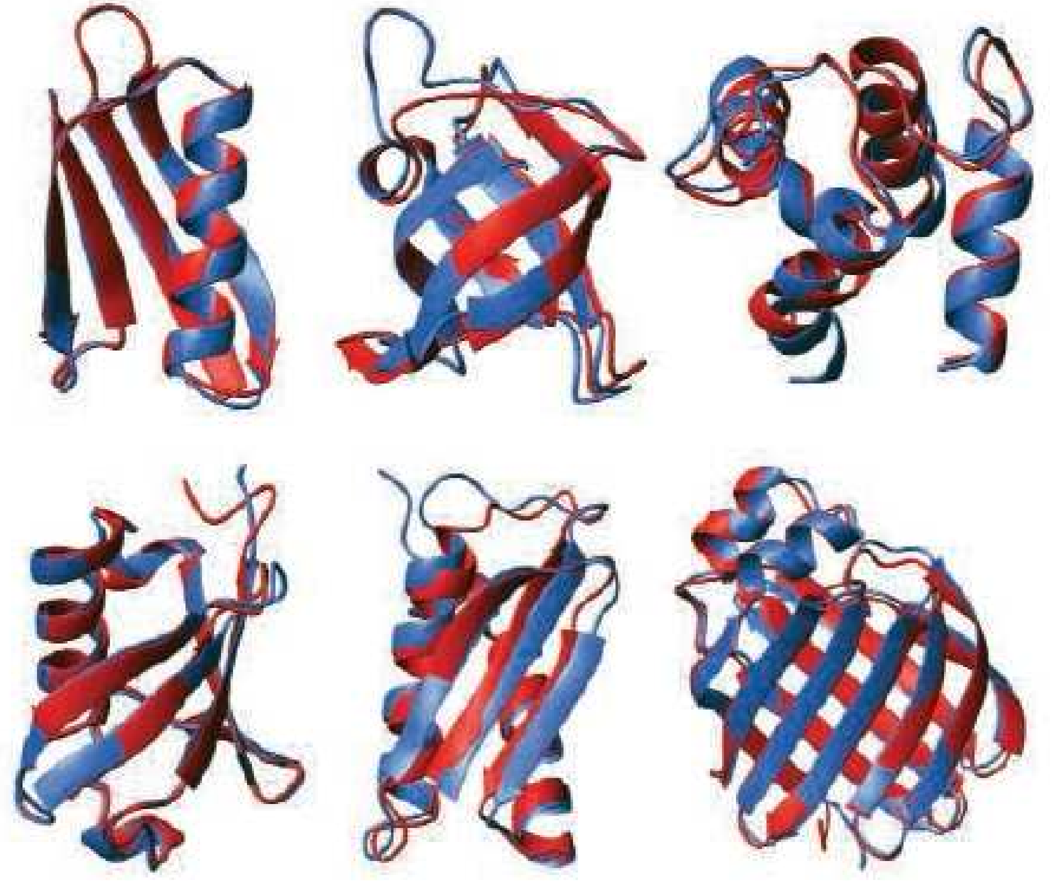
T0281 *ab initio* prediction (1.59Å)





Zone 3 Conclusions

- A lot of progress over the CASPs.
- A long way to go still.
- Knowledge integration, multiple trajectories key.
- Discrimination remains a bottleneck.
- All atom description and refinement probably necessary.



Tight fit. Adding data from nuclear magnetic resonance experiments improves the accuracy of computer models of how proteins fold.