Spécialité Bioinformatique M1: Lecture 2-A

P. Derreumaux

Predicting structure from a bioinformatics or biochemists perspective

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:**

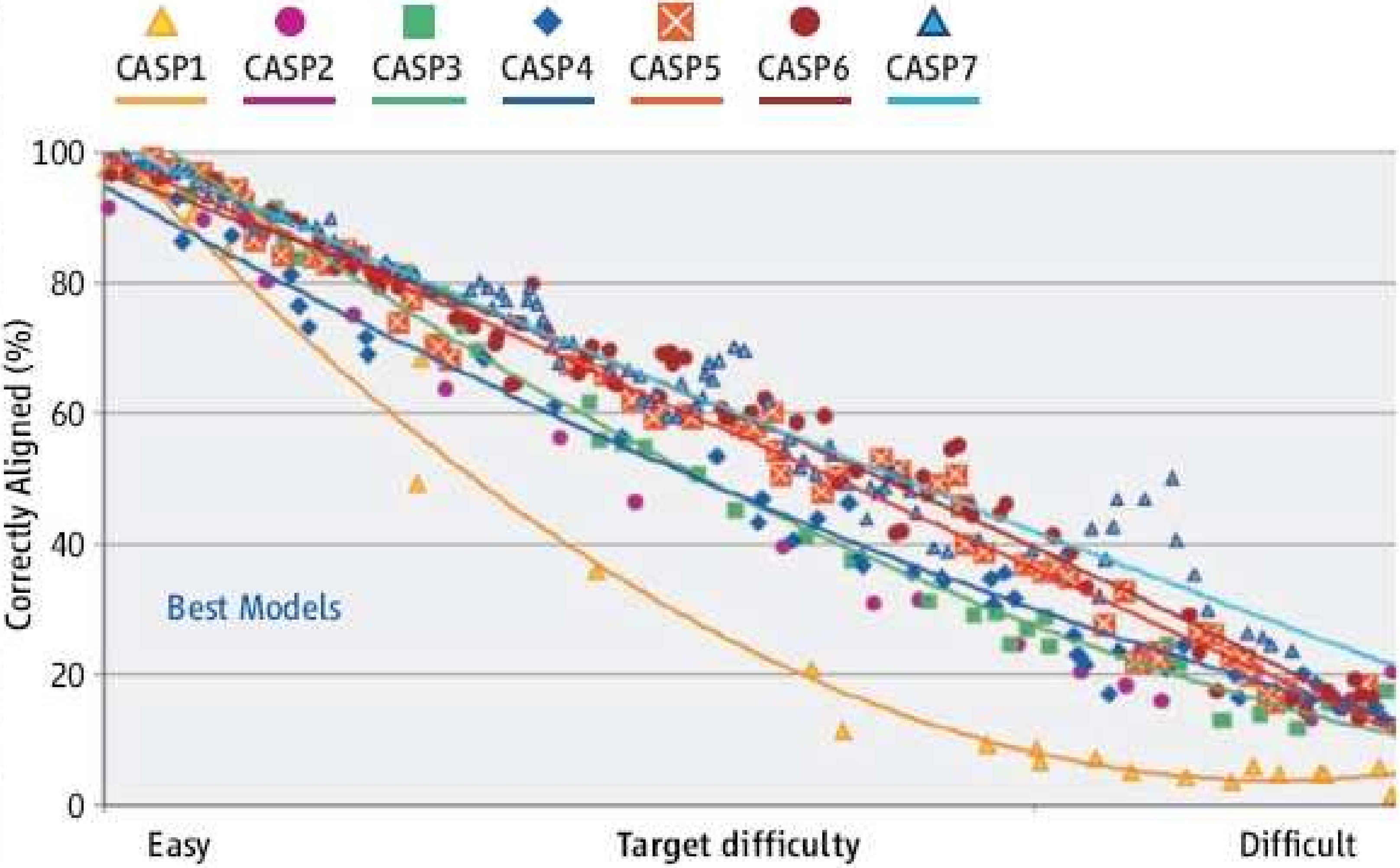
Profiles

Multiple templates

Meta-servers

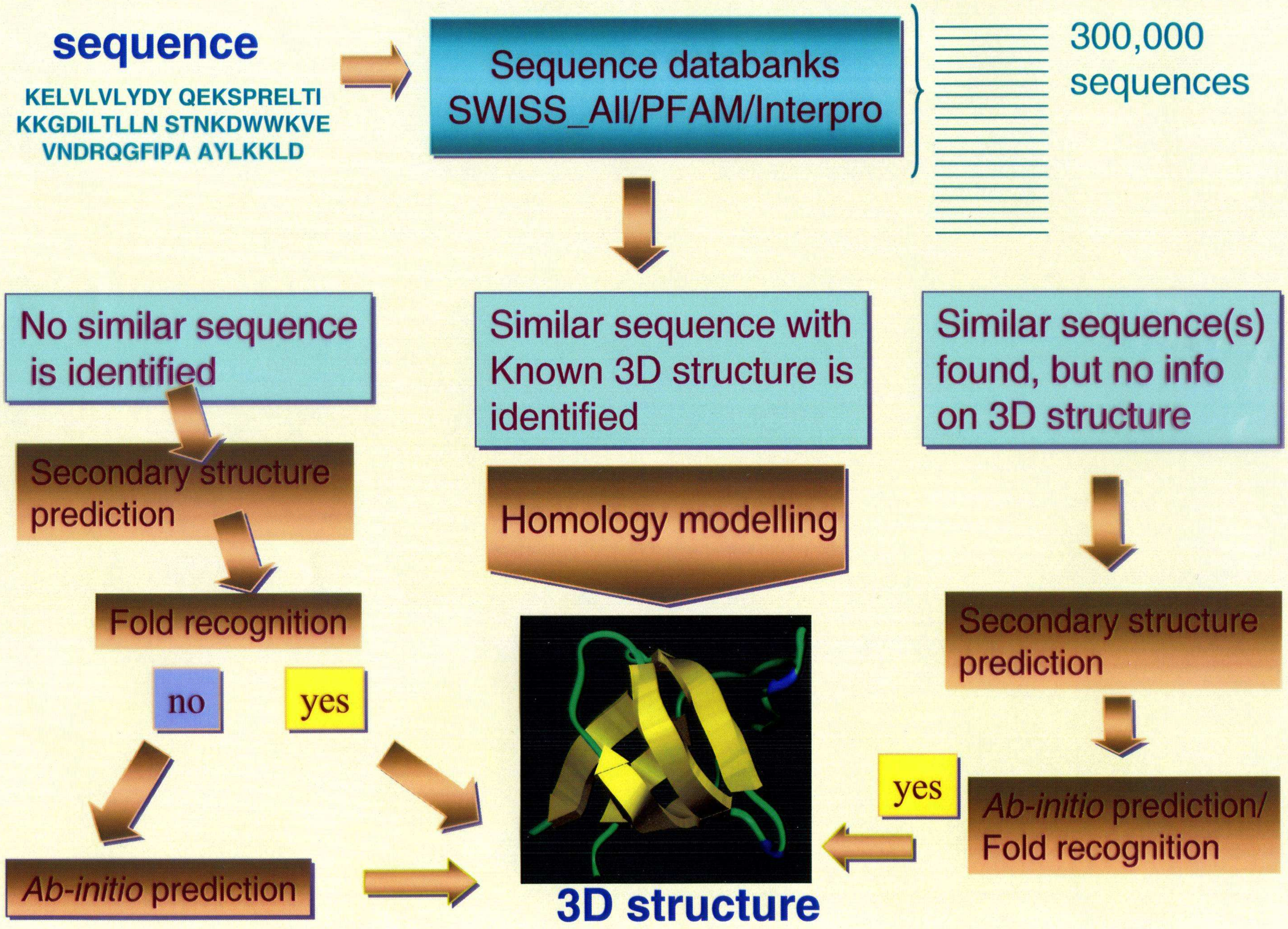
Fragments

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

3D structure

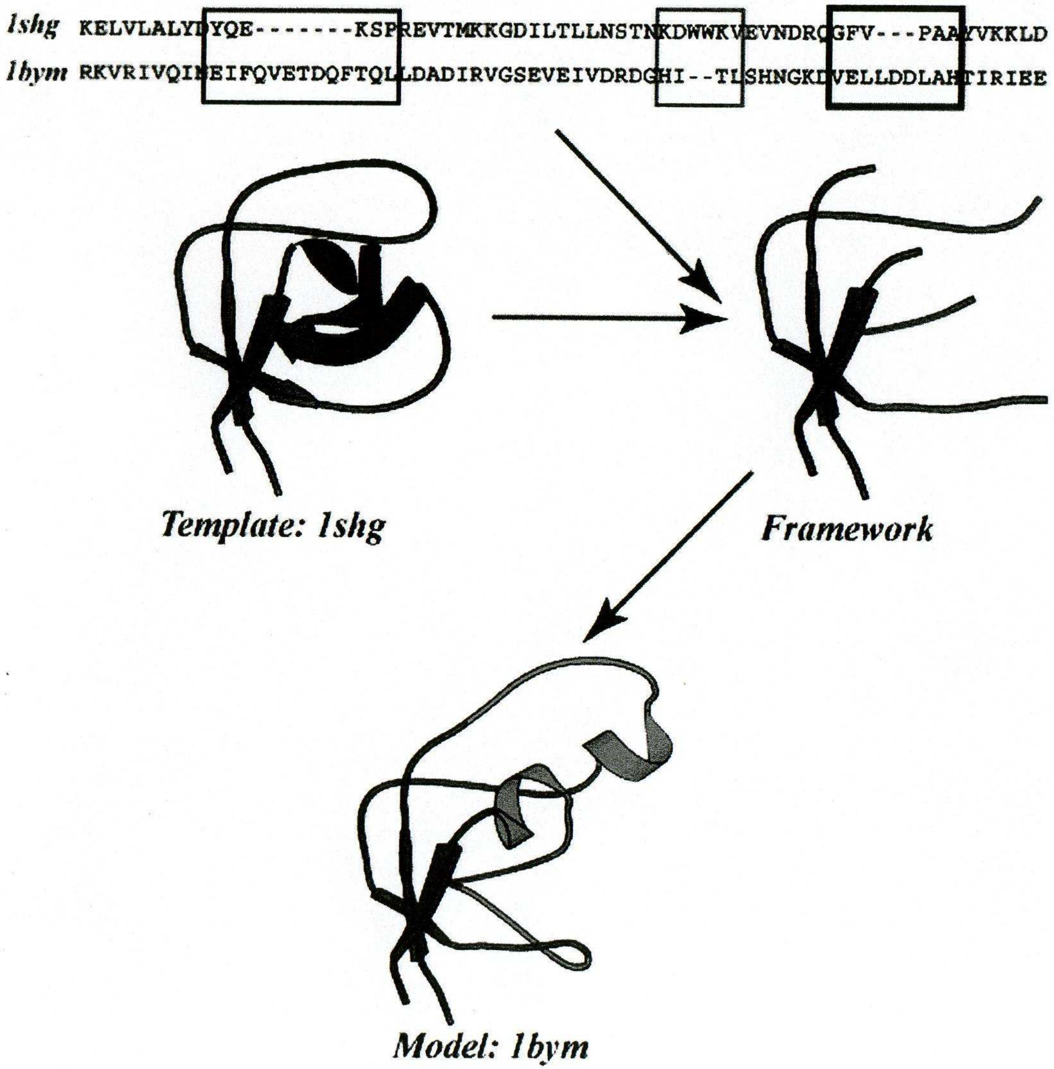
Similar sequence(s)
found, but no info
on 3D structure

Secondary structure
prediction

Ab-initio prediction/
Fold recognition

yes

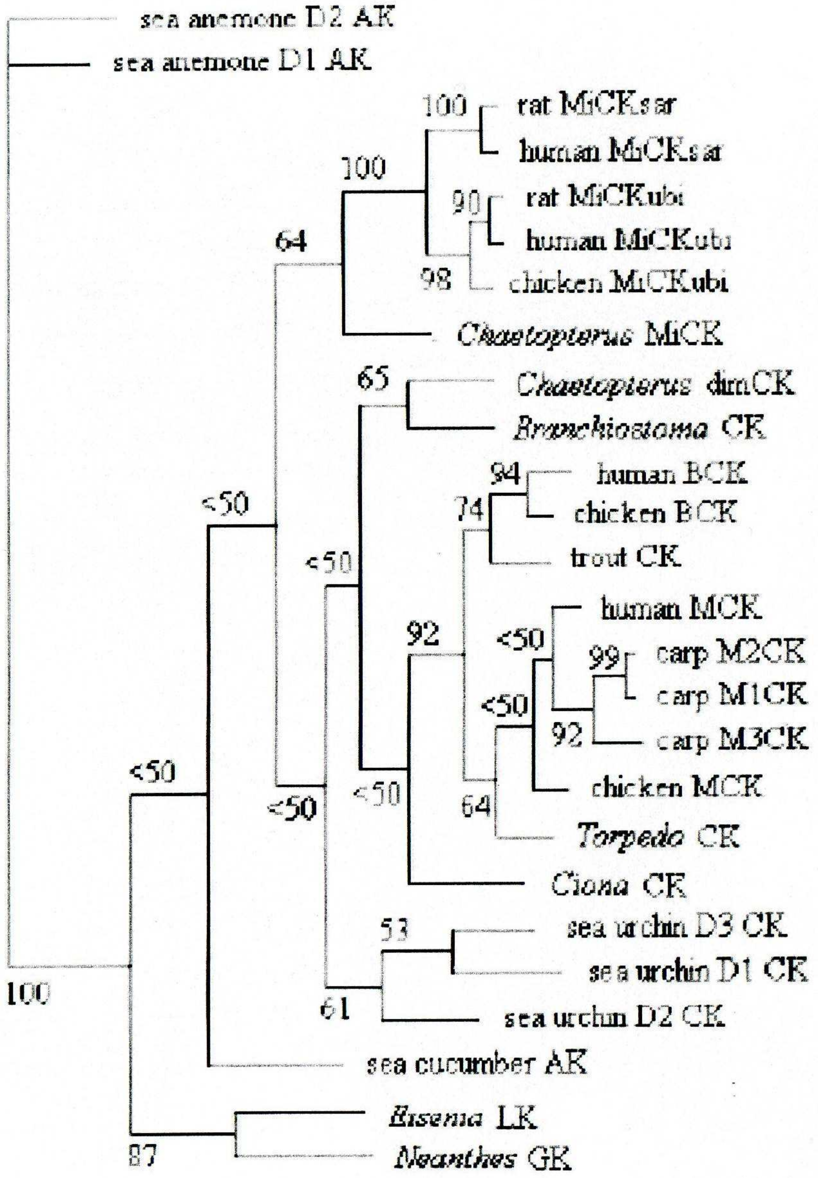
Homology Modeling: How it works


1shg KELVLALYDYQE-------KSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFV---PAAYVKKLD
1bym RKVRIVQINEIFQVETDQFTQLLDADIRVGSEVEIVDRDGHI--TLSHNGKDVELLDDLAHTIRIEE
Template: 1shg
Framework
Model: 1bym

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

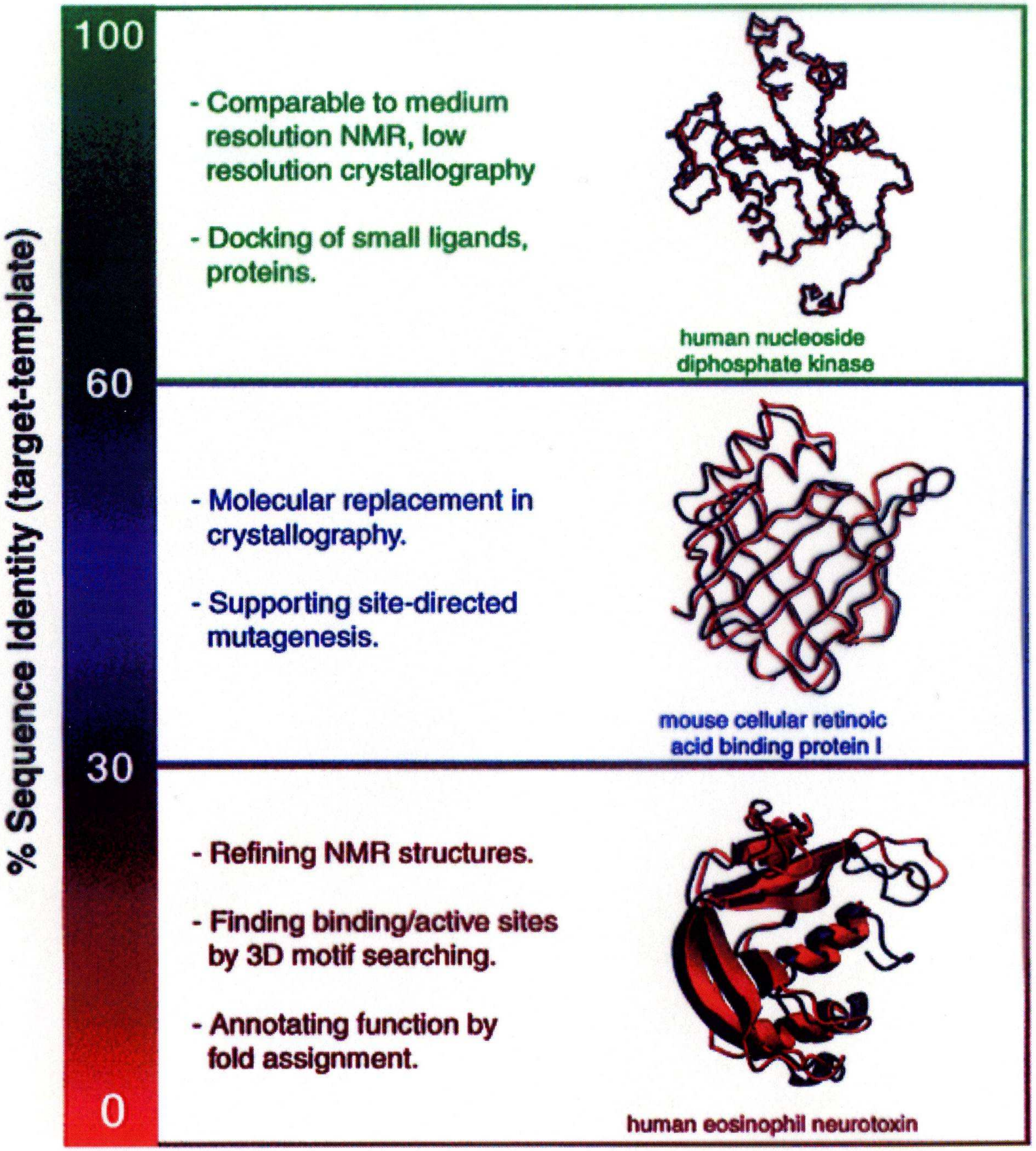
1.Use of rotamer libraries (backbone dependent)
2. Molecular mechanics optimization
3. Mean-field methods

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.


% Sequence Identity (target-template)
100
60
30
0
human nucleoside diphosphate kinase
mouse cellular retinoic acid binding protein I
human eosinophil neurotoxin

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

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

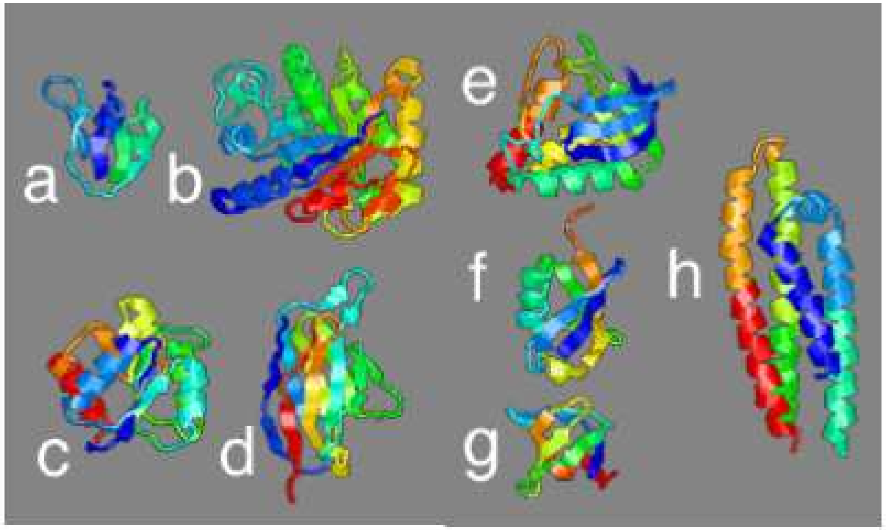
- Refining NMR structures.
- Finding binding/active sites by 3D motif searching.
- Annotating function by fold assignment.

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

- <u>Concept</u>

- 3 families of methods.

(1) Sequence Profiles PSI-BLAST

<u>Ref</u> Dunbrack, Proteins (1999)
Suppl 3 : 81-87.
(very close to comparative modelling)

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

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

(3) Threading = Fold Recognition

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

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

TOPITS server (Predict Protein Server)

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

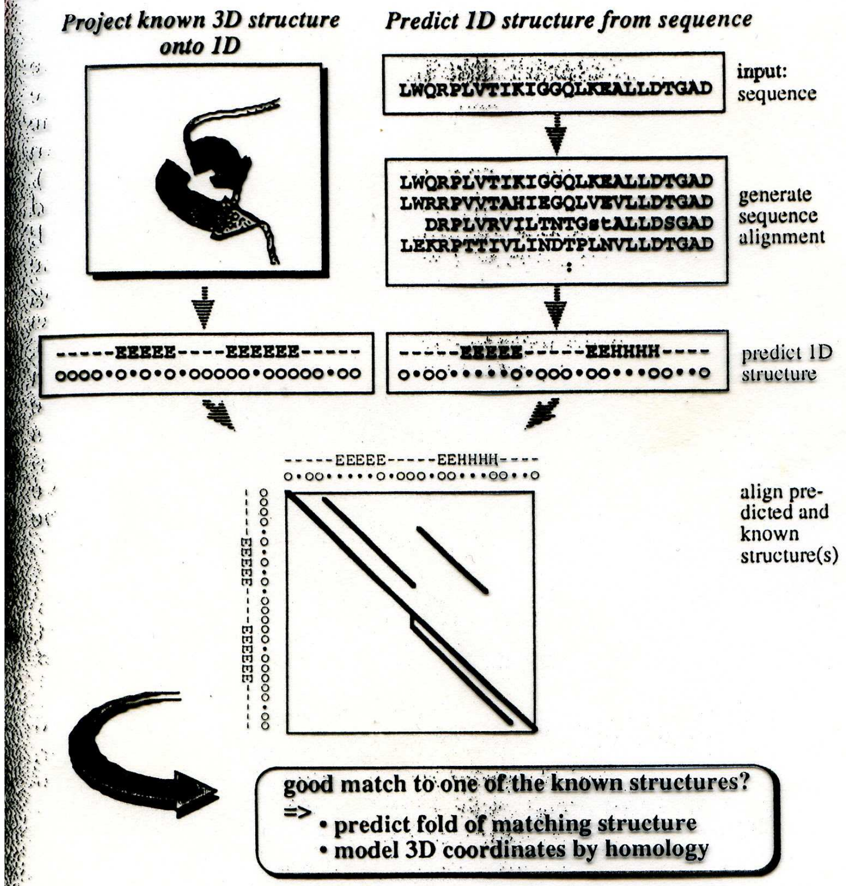

Project known 3D structure onto 1D
Predict 1D structure from sequence
LWQRPLVTIKIGGQLKEALLDTGAD
input: sequence
LWQRPLVTIKIGGQLKEALLDTGAD
LWRRPVVTAHIEGQLVEVLLDTGAD
DRPLVRVILTNTGstALLDSGAD
LEKRPTTIVLINDTPLNVLLDTGAD
generate sequence alignment
-----EEEEE----EEEEEE-----
-----EEEEE-----EEHHHH----
predict 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).

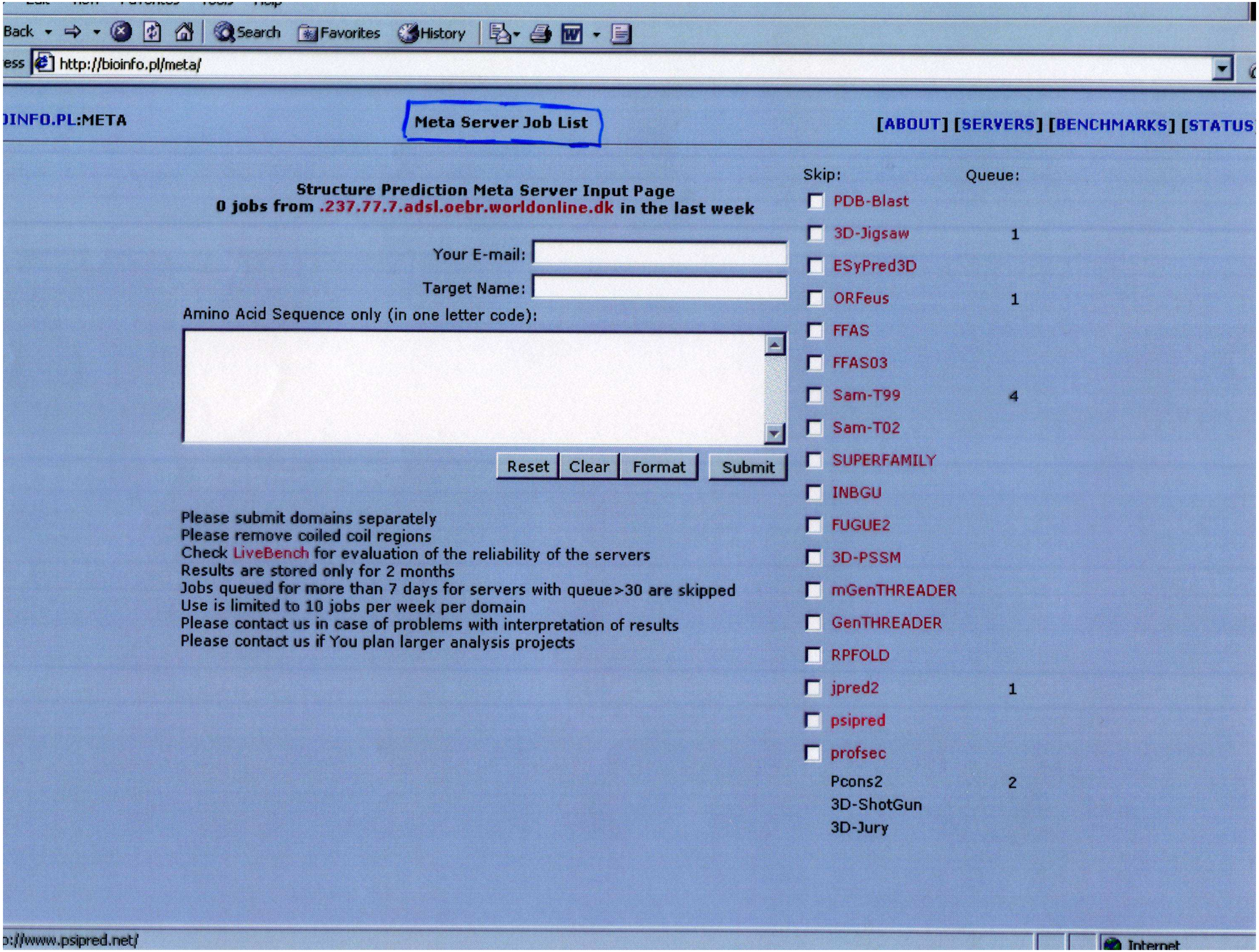
Free parameters for dynamic programming

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 (Sander & Schneider, 1991; Schneider, 1994). The following free parameters had to be adjusted:

or a Blosum62 (Henikoff & Henikoff, 1992) exchange matrix:

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

where M_{ij} determined the score for a match at a given position between state *i* in the first string and state *j* in the second string, and $\mu = 0$ to 100



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

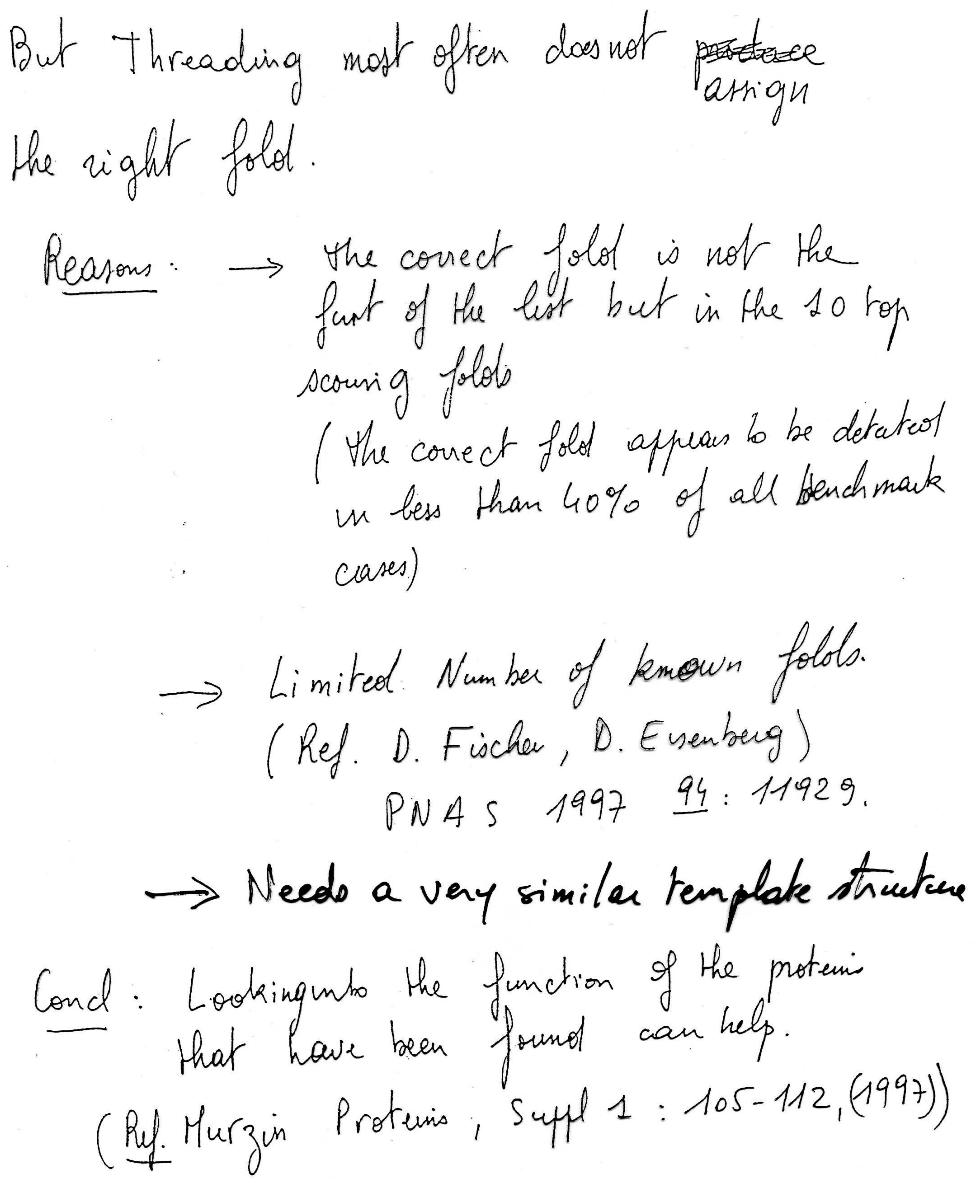
Your E-mail:

Target Name:

Amino Acid Sequence only (in one letter code):

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))

→ **Errors in the scoring functions**

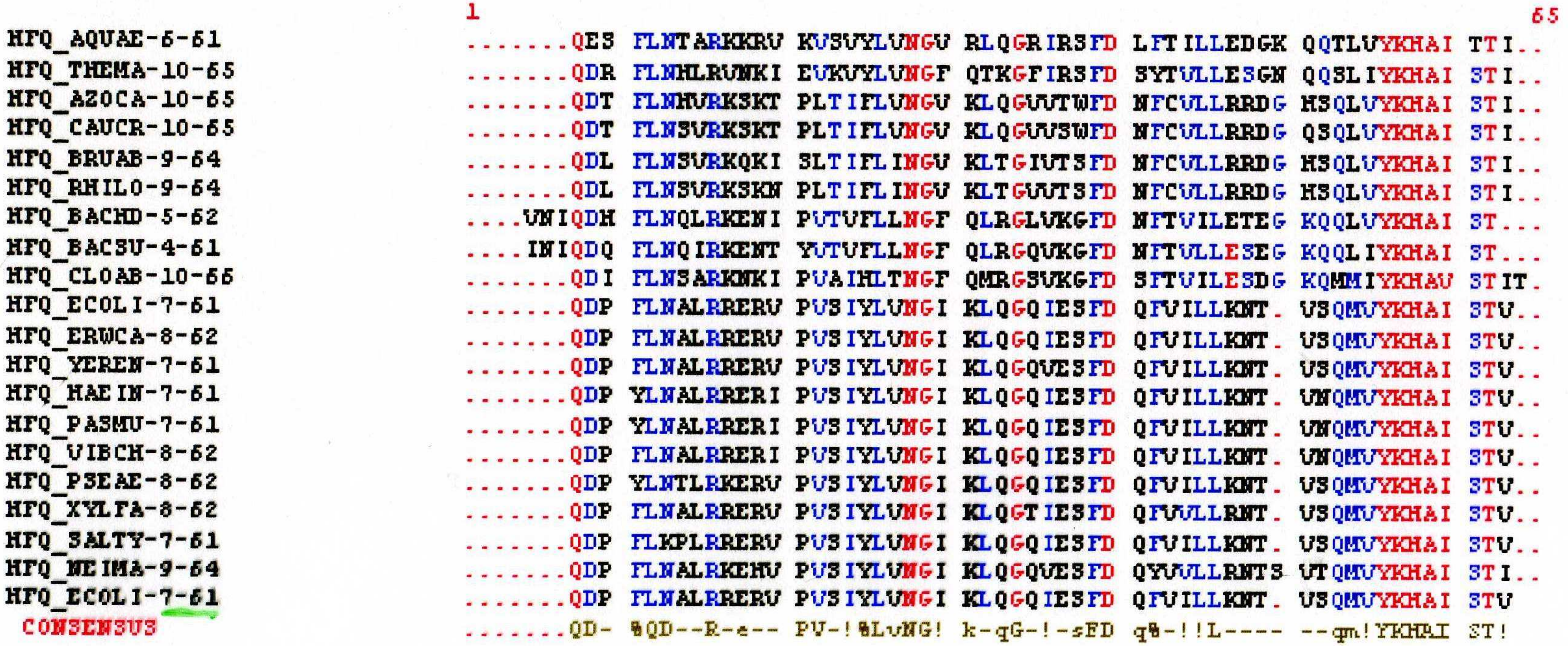
V. Arluison et al., J. Mol. Biol. (2002) 320, 705-712.

La protéine HFq

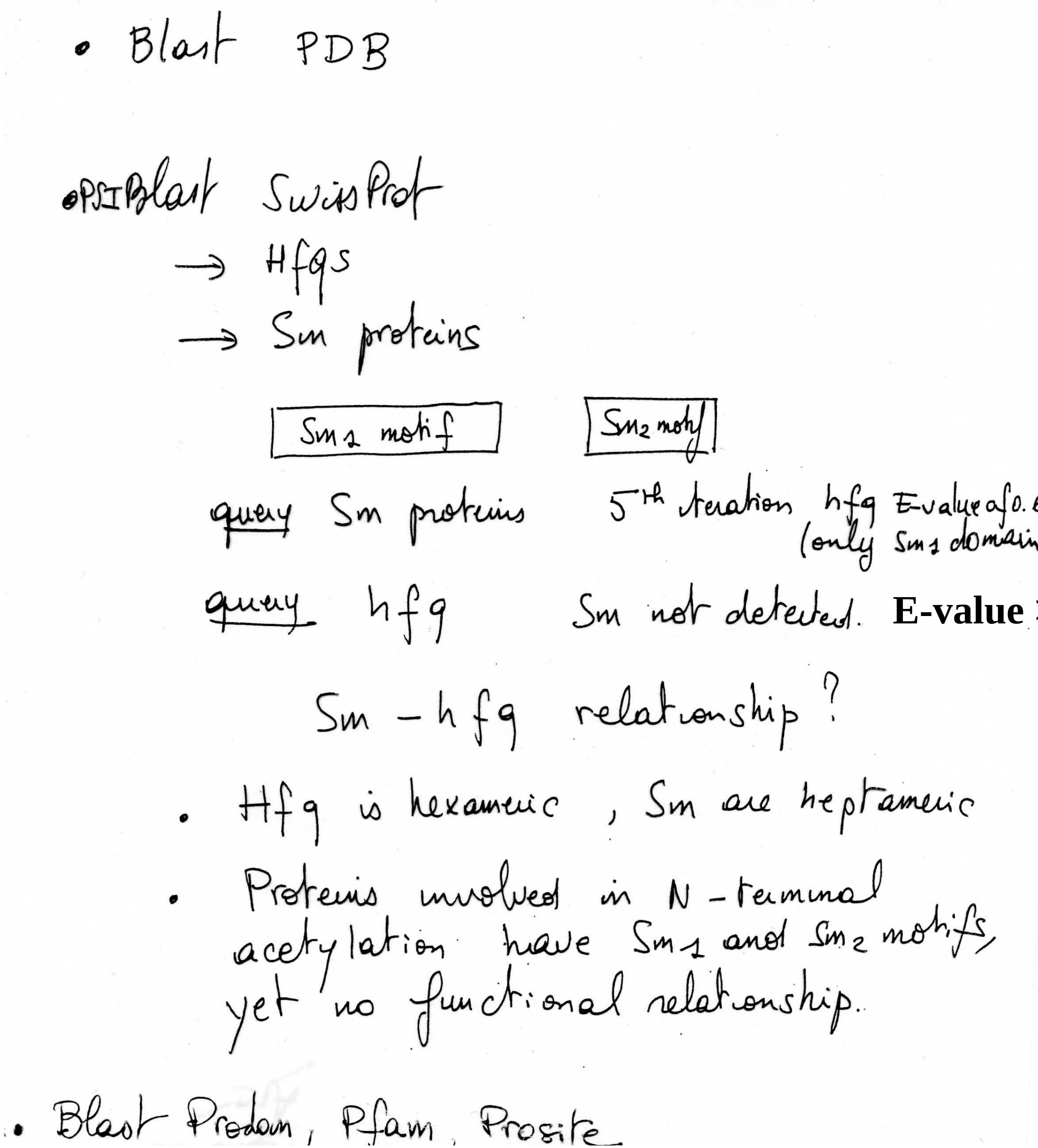
HFq est une protéine conservée chez les bactéries

- -Protéobacteries: subdivisions α, β, et γ
- -Firmicutes: groupes Bacillus et Clostridium
- -Thermotogales
- -Aquificales

Alignement ProDom



```
                   1                                                                       65
HFQ_AQUAE-6-61     ........QES FLNTARKKRV KVSVYLVNGV RLQGRIRSFD LFTILLEDGK QQTLVYKHAI TTI..
HFQ_THEMA-10-65    ........QDR FLNHLRVNKI EVKVYLVNGF QTKGFIRSFD SYTVLLESGN QQSLIYKHAI STI..
HFQ_AZOCA-10-65    ........QDT FLNHVRKSKT PLTIFLVNGV KLQGVVTWFD NFCVLLRRDG HSQLVYKHAI STI..
HFQ_CAUCR-10-65    ........QDT FLNSVRKSKT PLTIFLVNGV KLQGVVSWFD NFCVLLRRDG QSQLVYKHAI STI..
HFQ_BRUAB-9-64     ........QDL FLNSVRKQKI SLTIFLINGV KLTGIVTSFD NFCVLLRRDG HSQLVYKHAI STI..
HFQ_RHILO-9-64     ........QDL FLNSVRKSKN PLTIFLINGV KLTGVVTSFD NFCVLLRRDG HSQLVYKHAI STI..
HFQ_BACHD-5-62     ....VNIQDH FLNQLRKENI PVTVFLLNGF QLRGLVKGFD NFTVILETEG KQQLVYKHAI ST...
HFQ_BACSU-4-61     ....INIQDQ FLNQIRKENT YVTVFLLNGF QLRGQVKGFD NFTVLLESEG KQQLIYKHAI ST...
HFQ_CLOAB-10-66    ........QDI FLNSARKNKI PVAIHLTNGF QMRGSVKGFD SFTVILESDG KQMMIYKHAV STIT.
HFQ_ECOLI-7-61     ........QDP FLNALRRERV PVSIYLVNGI KLQGQIESFD QFVILLKNT. VSQMVYKHAI STV..
HFQ_ERWCA-8-62     ........QDP FLNALRRERV PVSIYLVNGI KLQGQIESFD QFVILLKNT. VSQMVYKHAI STV..
HFQ_YEREN-7-61     ........QDP FLNALRRERV PVSIYLVNGI KLQGQVESFD QFVILLKNT. VSQMVYKHAI STV..
HFQ_HAEIN-7-61     ........QDP YLNALRRERI PVSIYLVNGI KLQGQIESFD QFVILLKNT. VNQMVYKHAI STV..
HFQ_PASMU-7-61     ........QDP YLNALRRERI PVSIYLVNGI KLQGQIESFD QFVILLKNT. VNQMVYKHAI STV..
HFQ_VIBCH-8-62     ........QDP FLNALRRERI PVSIYLVNGI KLQGQIESFD QFVILLKNT. VNQMVYKHAI STV..
HFQ_PSEAE-8-62     ........QDP YLNTLRKERV PVSIYLVNGI KLQGQIESFD QFVILLKNT. VSQMVYKHAI STV..
HFQ_XYLFA-8-62     ........QDP FLNALRRERV PVSIYLVNGI KLQGTIESFD QFVVLLRNT. VSQMVYKHAI STV..
HFQ_SALTY-7-61     ........QDP FLKPLRRERV PVSIYLVNGI KLQGQIESFD QFVILLKNT. VSQMVYKHAI STV..
HFQ_NEIMA-9-64     ........QDP FLNALRKEHV PVSIYLVNGI KLQGQVESFD QYVVLLRNTS VTQMVYKHAI STI..
HFQ_ECOLI-7-61     ........QDP FLNALRRERV PVSIYLVNGI KLQGQIESFD QFVILLKNT. VSQMVYKHAI STV
CONSENSUS          ........QD- %QD--R-e-- PV-!%LvNG! k-qG-!-sFD q%-!!L---- --qm!YKHAI ST!
```



- Blast PDB

- PSIBlast SwissProt
  - → Hfqs
  - → Sm proteins

$\boxed{Sm_1 \text{ motif}}$ $\boxed{Sm_2 \text{ motif}}$

<u>query</u> Sm proteins 5th iteration hfq E-value of 0.
(only Sm_1 domain

<u>query</u> hfq Sm not detected. **E-value > 1**

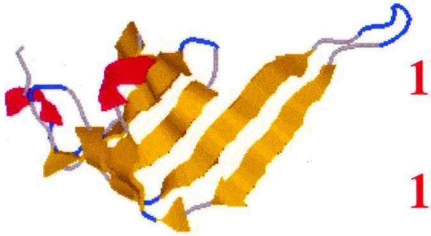
Sm - hfq relationship ?

- Hfq is hexameric , Sm are heptameric
- Proteins involved in N-terminal acetylation have Sm_1 and Sm_2 motifs, yet no functional relationship.

- Blast Prodom, Pfam, Prosite

Identification de la topologie de HFq

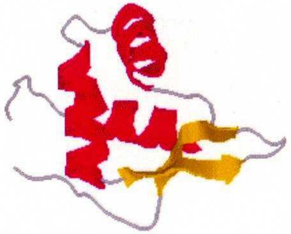
Topologies prédites par des méthodes de reconnaissance de “fold”
(Topits, 3D-PSSM, GenThreader, 123D, Méthode de Fischer)



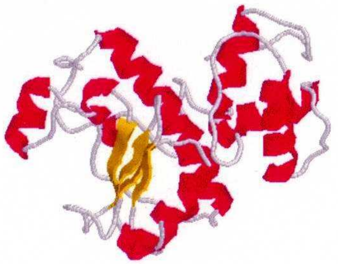
Topologie Sm

Code PDB (type de protéine)	% identité avec HFq
1B34b (Sm)	19
1B34a (Sm)	10
1D3b (Sm)	8
1MJC (CspA)	22
1AOY (Arc)	22
1LBU (peptidase)	11

Topologie CspA



Topologie Arc



Topologie 1LBU

Table 1. The predicted topologies of the three Hfq proteins using fold-recognition methods

PDB entry	*E. coli* Hfq	*A. aeolicus* Hfq	*A. caulinodans* Hfq
1B34b (Sm)	3 (0.1, 19)	2 (1.6, 19)	2 (0.5, 16)
1B34a (Sm)	2 (1.4, 10)	1 (0.7, 20)	1 (0.6, 18)
1D3b (Sm)	1 (0.8, 8)	2 (1.3, 13)	3 (0.8, 10)
1MJC (CspA)	2 (1.1, 22)	3 (3.0, 17)	3 (4.0, 20)
1AOY (Arc)	1 (4.7, 22)	0	1 (4.5, 24)
1DIV (L9)	0	1 (1.5, 15)	1 (0.6, 18)
1LBU (peptidase)	1 (1.3, 11)	0	0

Each suggested fold, defined by its PDB entry number, is characterised by three values. The first value is the number of times the fold is suggested in the top positions. The *E*-value and the percentage of identity between each Hfq protein and the fold identified are then given in parentheses.

HFq semble être une protéine Sm-like

1. HFq présente des similarités fonctionnelles avec les protéines Sm: elles sont impliquées dans le métabolisme de l'ARNm

Protéine Sm:

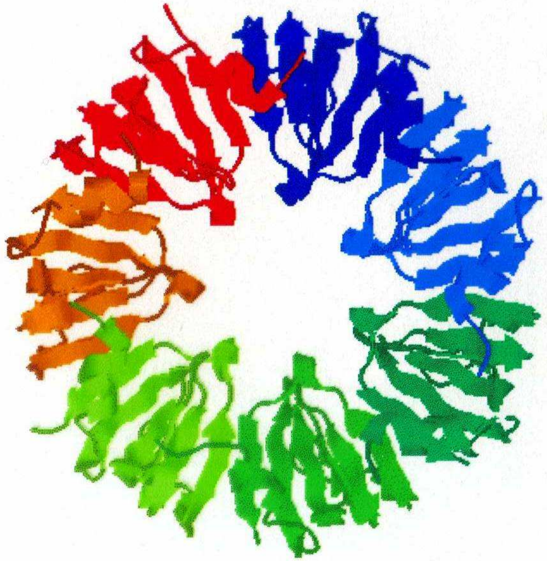
Composant du spliceosome eucaryote impliqué dans l'épissage des ARNm

Protéines retrouvées également chez les archaebactéries

Protéine heptamérique (homo- ou hétéroheptamère)

Forme un anneau avec un trou central (EM)

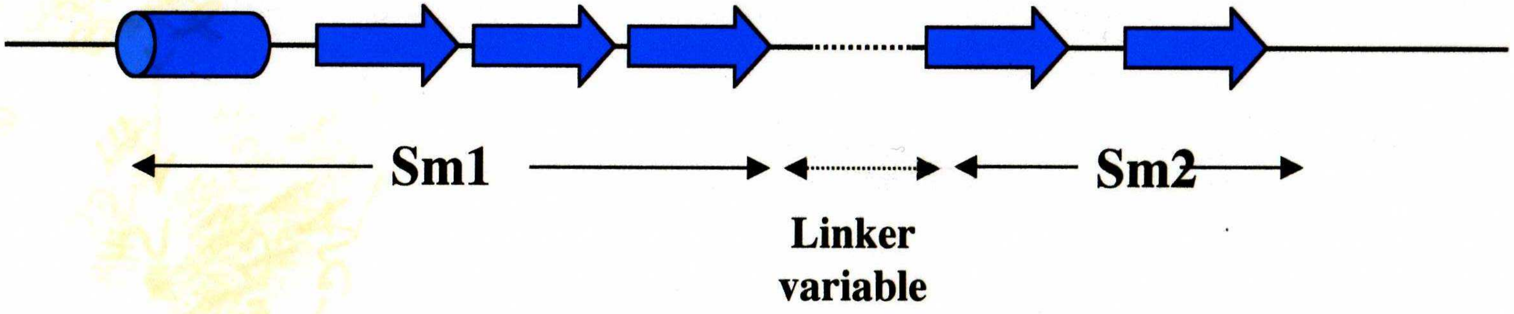
Essentiellement structurée en feuillet β



2a La prédiction de structure secondaire est en accord avec une topologie Sm

PHD et PSIPRED

```
              10        20        30        40        50        60        70        80        90        100
  AA: AKGQSLQDPFLNALRRERVPVSIYLVNGIKLQGQIESFDQFVILLKNTVSQMVYKHAISTVVPSRPVSHHSNNAGGGTSSNYHHGSSAQNTSAQQDSEETE
PSI : 9888730788999876572589999738887679998601899997898279999520101478863233157877777776536526778865466445 79
rel :        HHHHHHHHH  EEEEEEEE  EEEEEEEEEE  EEEEEE   EEEEEE HHEE

PHD :        HHHHHHHHH    EEEEEE   EEEEEEEEE EEEEEEE   EEEEEEEEEEEEE
rel : 99998781998899744893699983263479999742337999973991699883138983487421269999999999999999887787775667899
```



2b Infrared Spectroscopy (FTIR)

UV-CD spectrum analysis.

37 (± 3)% β → Sm

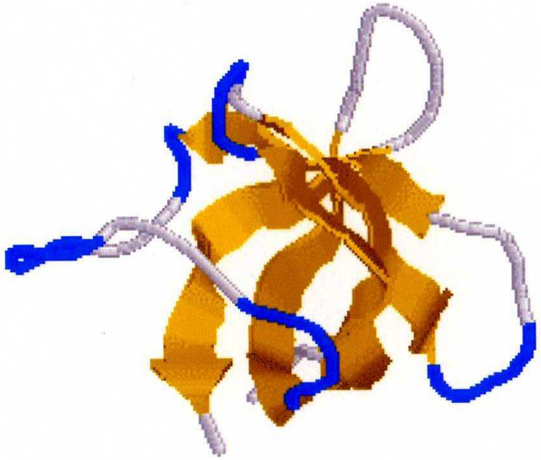
15 (±3)% α

~~Peptidase~~

~~Arc~~

↳ CspA remains

3. HFq ne présente pas la signature du domaine "cold shok" de CspA (code Prosite PS00352)



Prosite has true positives and false positives

Incertitudes:

➡ Sm1 n'est pas détécté dans toutes les HFq

➡ Sm2 n'est jamais détecté

➡ Il existe des protéines qui présentent Sm1 et Sm2 qui ne sont pas des protéines Sm (Acetyltransferase NatC de levure)

→ % β et α compatible avec topologie CspA

Modélisation de la structure de HFq

La modélisation moléculaire a été effectuée avec une protéine matrice Sm humaine (~~1D3B~~, structure résolue sous forme héxamérique)

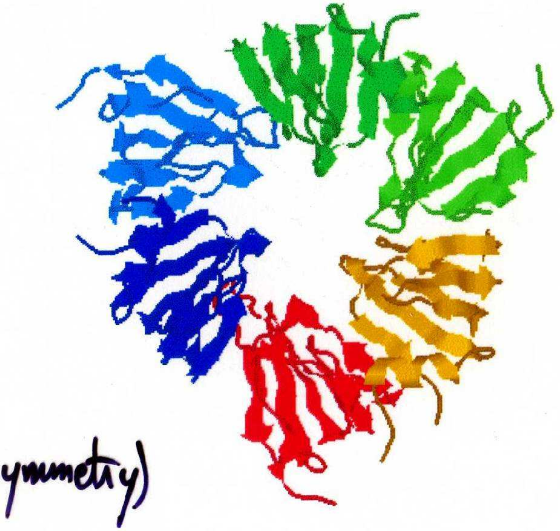
1B34

Swiss model + ...

~~**Sm 1D3B**~~:

1B34

trimer of dimer
(pseudo 6-fold symmetry)



8 % seq identity
with Hfq from E. coli

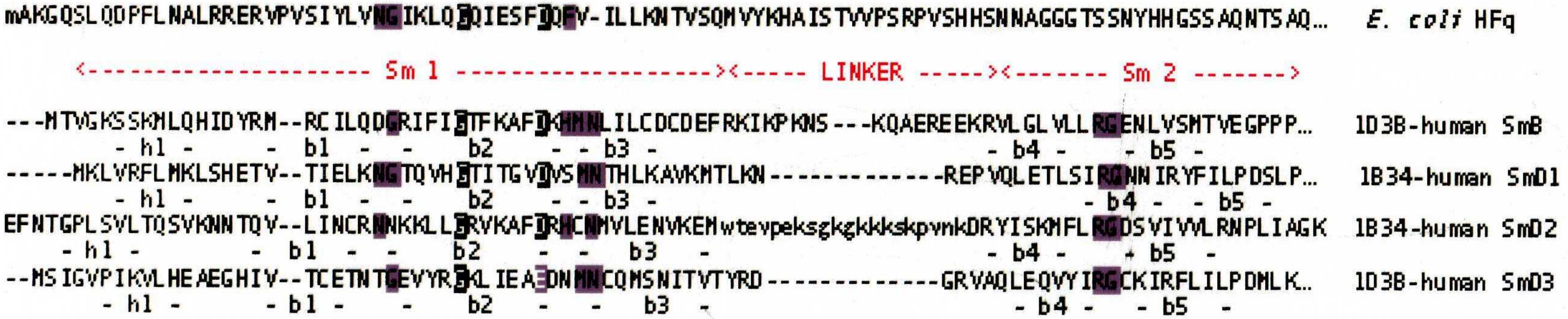
Alignement entre HFq et les protéines Sm

➭ Crucial pour la génération du modèle

Problème: Sm2 peu conservé et linker de taille variable

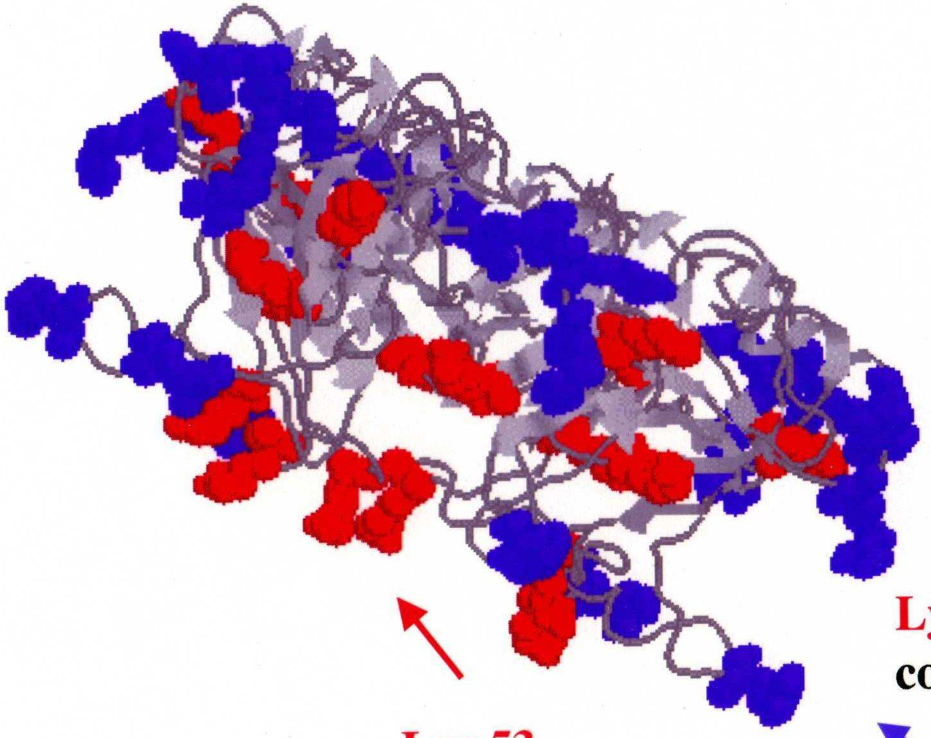
2 alignements possibles

1er alignement



```
mAKGQSLQDPFLNALRRERVPVSIYLVNGIKLQGQIESFDQFV-ILLKNTVSQMVYKHAISTVVPSRPVSHHSNNAGGGTSSNYHHGSSAQNTSAQ...   E. coli HFq

    <-------------------- Sm 1 ------------------><----- LINKER ----><------- Sm 2 ------->

---MTVGKSSKMLQHIDYRM--RCILQDGRIFIGTFKAFDKHMNLILCDCDEFRKIKPKNS---KQAEREEKRVLGLVLLRGENLVSMTVEGPPP...   1D3B-human SmB
       - h1 -     -  b1  -  -   b2     - - b3 -                       - b4 -  - b5 -
-----MKLVRFLMKLSHETV--TIELKNGTQVHGTITGVDVSMNTHLKAVKMTLKN-------------REPVQLETLSIRGNNIRYFILPDSLP...   1B34-human SmD1
       - h1 -     -  b1  -  -   b2     - - b3 -                              - b4 -  - b5 -
EFNTGPLSVLTQSVKNNTQV--LINCRNNKKLLGRVKAFDRHCNMVLENVKEMwtevpeksgkgkkkskpvnkDRYISKMFLRGDSVIVVLRNPLIAGK   1B34-human SmD2
    - h1 -       -  b1  -  -   b2    -  -   b3   -                          - b4 -  + b5  -
--MSIGVPIKVLHEAEGHIV--TCETNTGEVYRGKLIEAEDNMNCQMSNITVTYRD-------------GRVAQLEQVYIRGCKIRFLILPDMLK...   1D3B-human SmD3
      - h1 -     - b1 -    -    b2    -   -   b3  -                           - b4 -   - b5  -
```

Structure résultant du premier alignement:



Problème:

Lys 53 et Arg 63 non protégées contre la trypsine

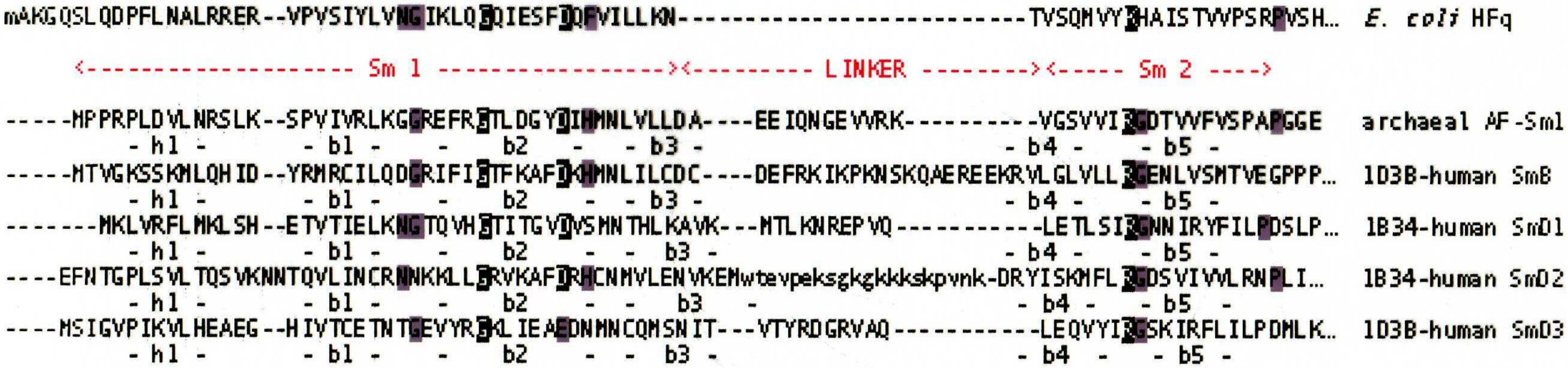
linker

Alignement entre HFq et les protéines Sm

➪ Crucial pour la génération du modèle

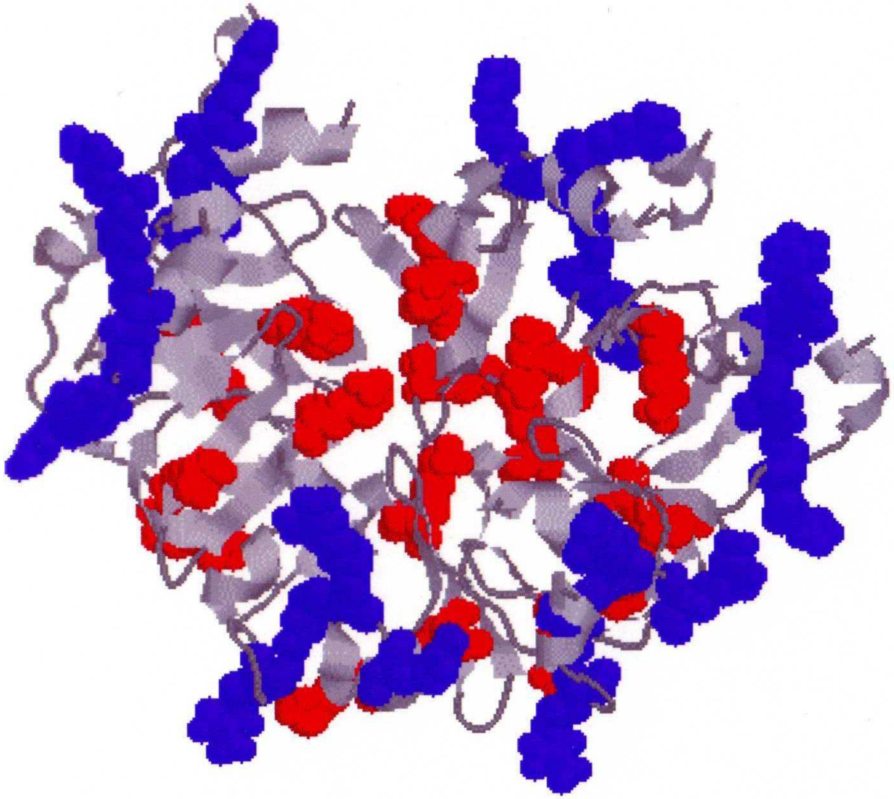
2ème alignement

➡ tient compte aussi de la prédiction de structure secondaire de HFq



```
mAKGQSLQDPFLNALRRER--VPVSIYLVNGIKLQGQIESFDQFVILLKN-------------------------TVSQMVYKHAISTVVPSRPVSH...   E. coli HFq

    <------------------- Sm 1 ----------------><--------- LINKER -------->-<----- Sm 2 ---->

-----MPPRPLDVLNRSLK--SPVIVRLKGGREFRGTLDGYDIHMNLVLLDA----EEIQNGEVVRK-----------VGSVVIRGDTVVFVSPAPGGE    archaeal AF-Sm1
        - h1 -      - b1 -    -     b2     -    - b3 -                   - b4 -    - b5 -
-----MTVGKSSKMLQHID--YRMRCILQDGRIFIGTFKAFDKHMNLILCDC----DEFRKIKPKNSKQAEREEKRVLGLVLLRGENLVSMTVEGPPP...  1D3B-human SmB
        - h1 -      - b1 -    -     b2     -    - b3 -                   - b4 -    - b5 -
-------MKLVRFLMKLSH--ETVTIELKNGTQVHGTITGVDVSMNTHLKAVK---MTLKNREPVQ-----------LETLSIRGNNIRYFILPDSLP...  1B34-human SmD1
        - h1 -      - b1 -    -     b2     -    -   b3 -                 - b4 -    - b5 -
----EFNTGPLSVLTQSVKNNTQVLINCRNNKKLLGRVKAFDRHCNMVLENVKEMwtevpeksgkgkkkskpvnk-DRYISKMFLRGDSVIVVLRNPLI...  1B34-human SmD2
        - h1 -      - b1 -    -     b2     -    -    b3    -                 - b4 -    - b5 -
----MSIGVPIKVLHEAEG--HIVTCETNTGEVYRGKLIEAEDNMNCQMSNIT---VTYRDGRVAQ-----------LEQVYIRGSKIRFLILPDMLK..   1D3B-human SmD3
        - h1 -      - b1 -    -     b2     -    -   b3 -                  - b4 -    - b5 -
```

Structure résultant du deuxième alignement:



Lys et Arg protégées contre la trypsine

26 aa C-terminaux non modélisés:

Pas d'identité de séquence avec d'autres protéines

Région de faible compléxité

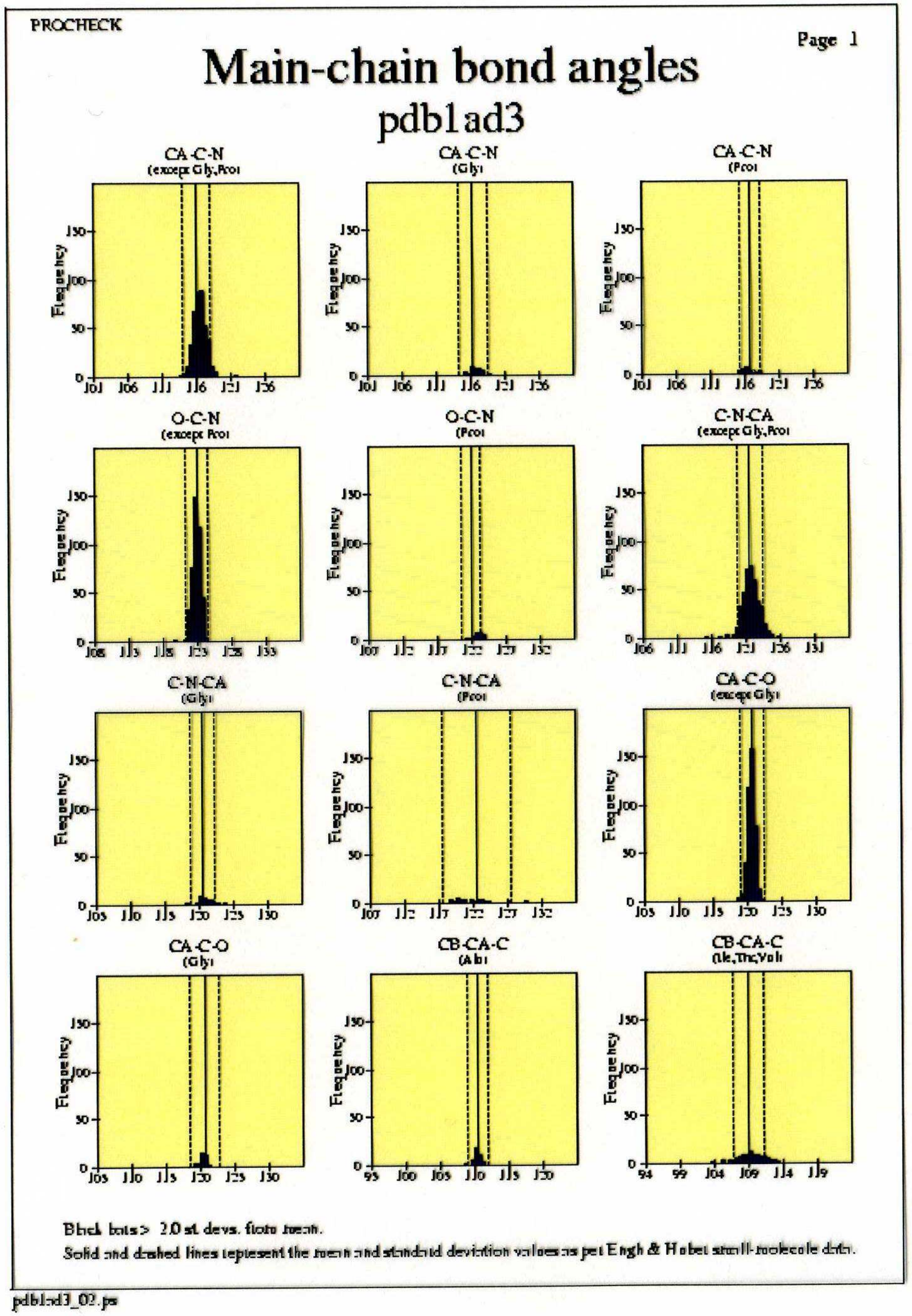
Région flexible (coil prédit par PHD et PSIPRED)

Refinement by energy minimization and short MD simulations in aqueous solution

Detection of Errors

First check should include a stereochemical check on the modeled structure—PROCHECK, WHATCHECK, DISTAN- which will show deviations from normal bond lengths, dihedrals, etc.

Visualization- follow the backbone trace and then subsequently move out to Cα-Cβ orientation.



PROCHECK

Page 1

Main-chain bond angles

pdb1ad3

Black bars > 2.0 st. devs. from mean.

Solid and dashed lines represent the mean and standard deviation values as per Engh & Huber small-molecule data.

pdb1ad3_02.ps

PROCHECK

http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html

Verification of Ramachandran plot (allowed and forbidden regions)

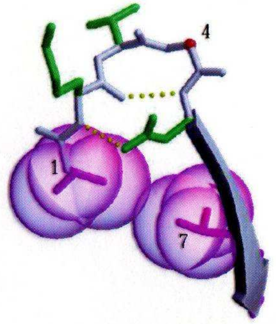
Comp with X-ray (2002), 1.5 Ǻ RMSD

Free modelling: De novo or ab initio

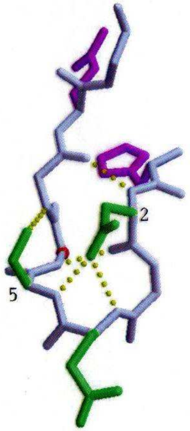
We focus here on fragment assembly approach

Protein Structure Prediction: Rosetta

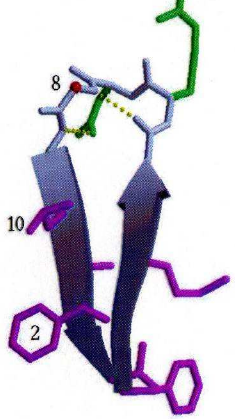
I-sites Library = a catalog of local sequence-structure correlations



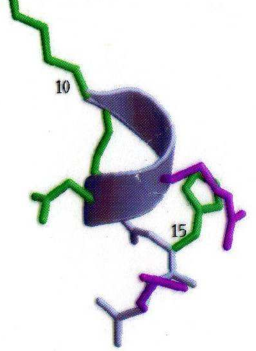
diverging type-2 turn



Serine hairpin

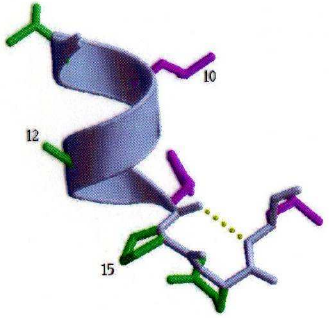


Type-I hairpin

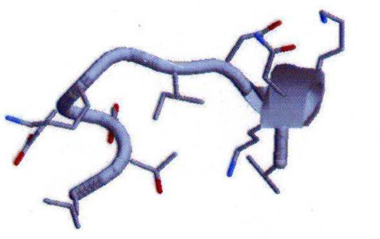


Frayed helix

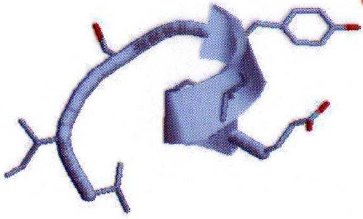
Local structure motifs



Proline helix C-cap

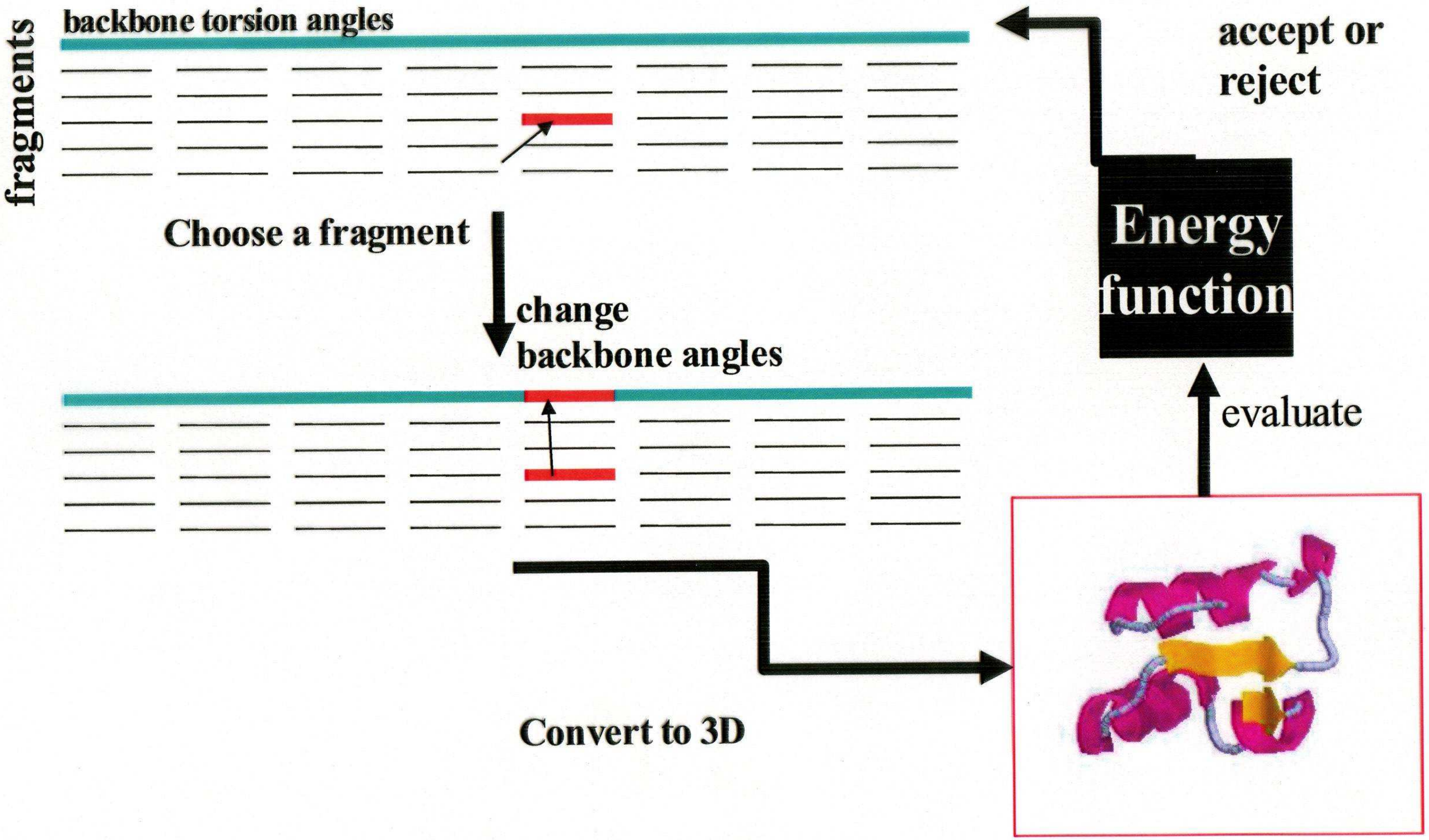


alpha-alpha corner



glycine helix N-cap

Rosetta: a folding simulation program



Fragment insertion Monte Carlo

Monte Carlo: Metropolis criterion.

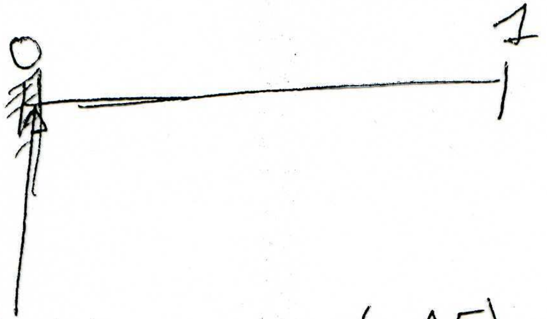
$X \rightarrow E$

$X' \rightarrow E'$

$\exp \frac{-\Delta E}{K_B T} = \frac{-(E'-E)}{K_B T}$

- $\Delta E < 0$ Always Accept
- $\Delta E > 0$ Accept $\begin{cases} \text{with } p = \exp \frac{-\Delta E}{kT} \\ \text{or} \\ r < \exp\left(\frac{-\Delta E}{kT}\right) \end{cases}$

$r \in [0, 1]$



$\Delta E = 10kT \quad \exp\left(-\frac{\Delta E}{kT}\right) \longrightarrow 0$

Rosetta (Baker) in CASP4

Improvements of the method.

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

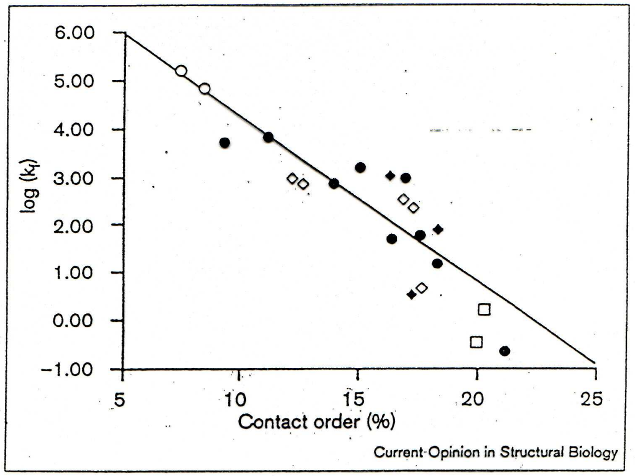
- Filters to eliminate non protein-like structure

  a. poorly formed β-sheets

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

  c. low contract 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.

$$\% Co = \frac{100}{L\,N}\sum^{N} \Delta S_{ij}$$

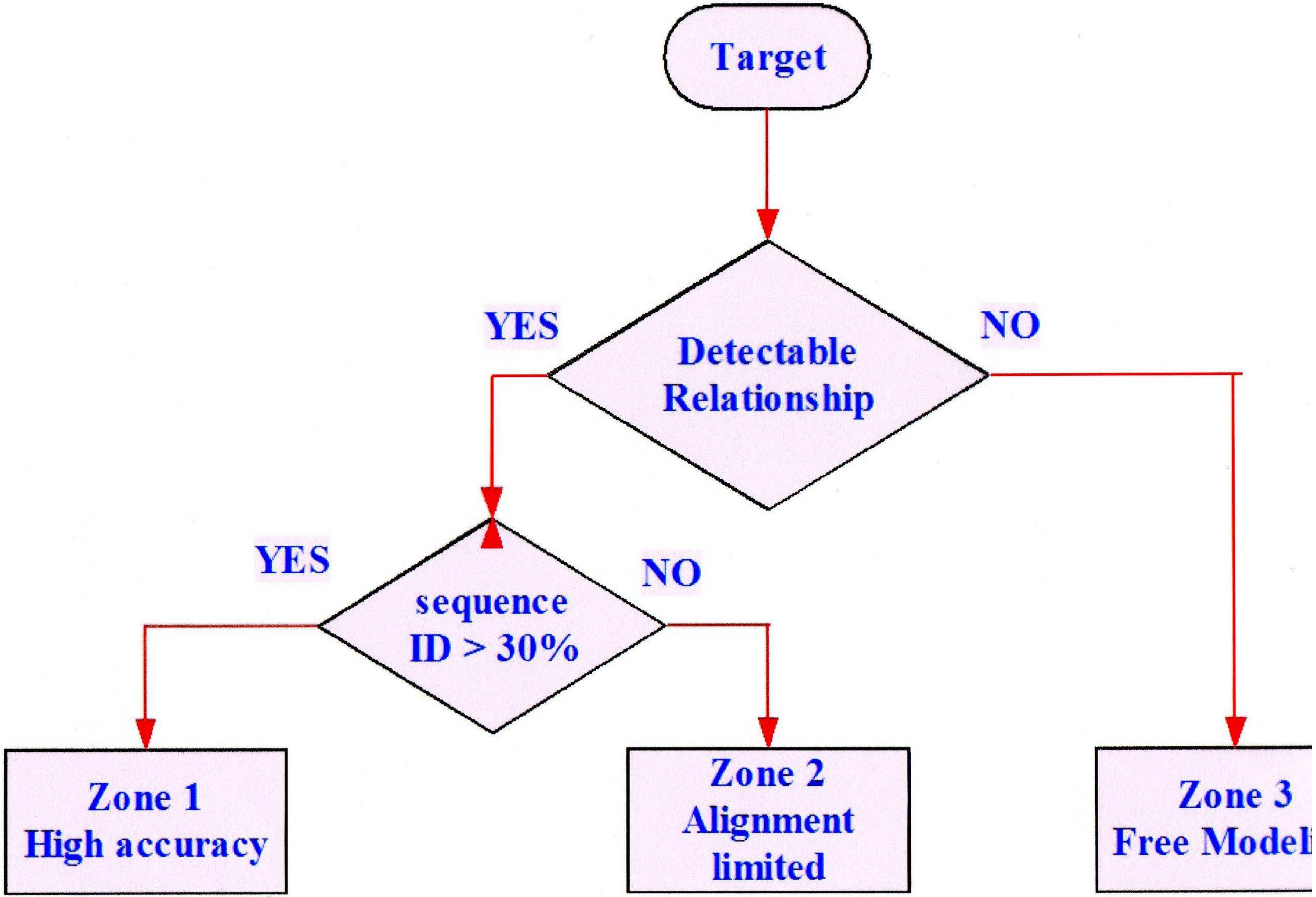
L = lenght AA, normalization factor

N= number of native contacts

- 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

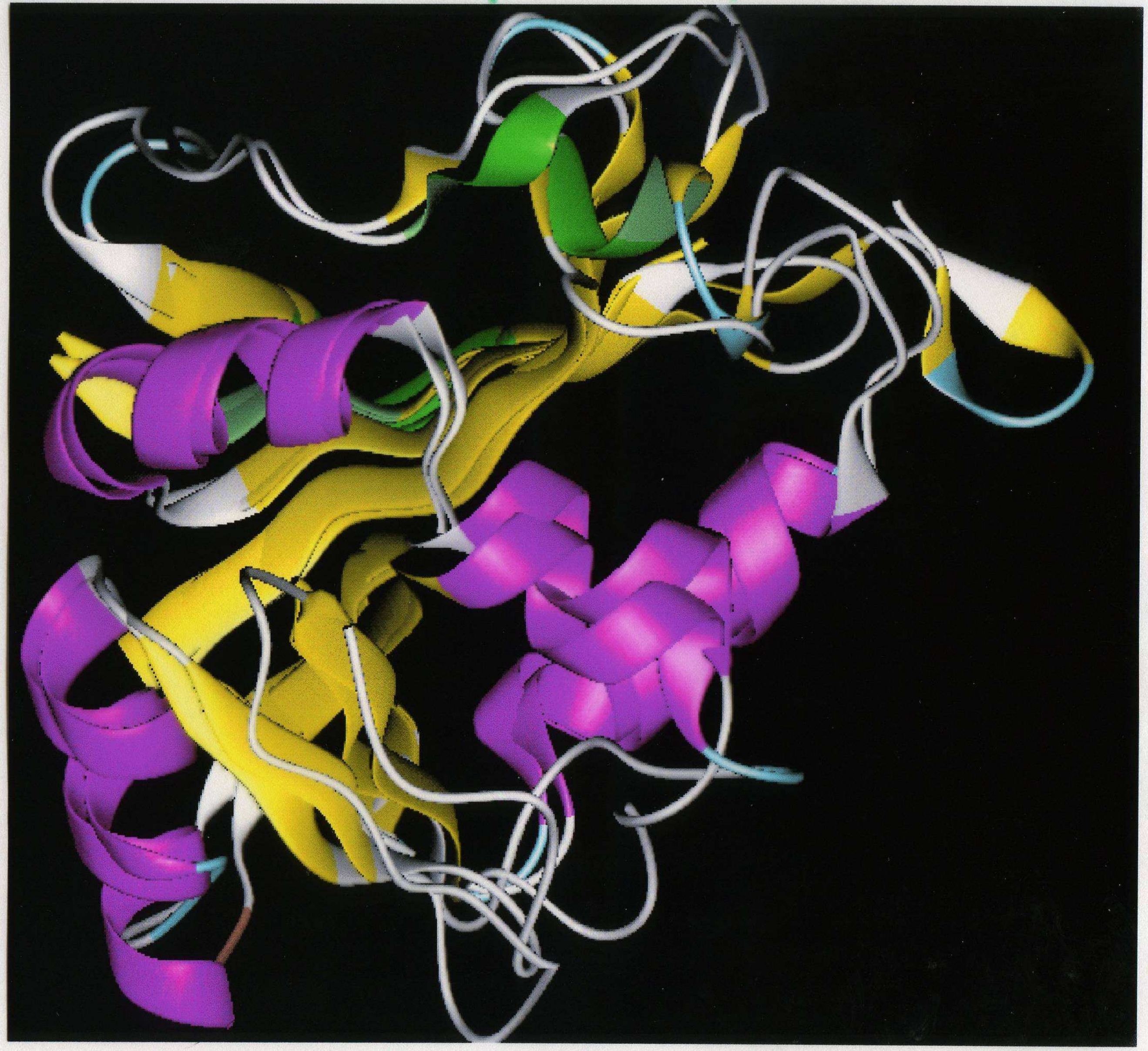
Zone 2 Alignment limited

Zone 3 Free Modeli

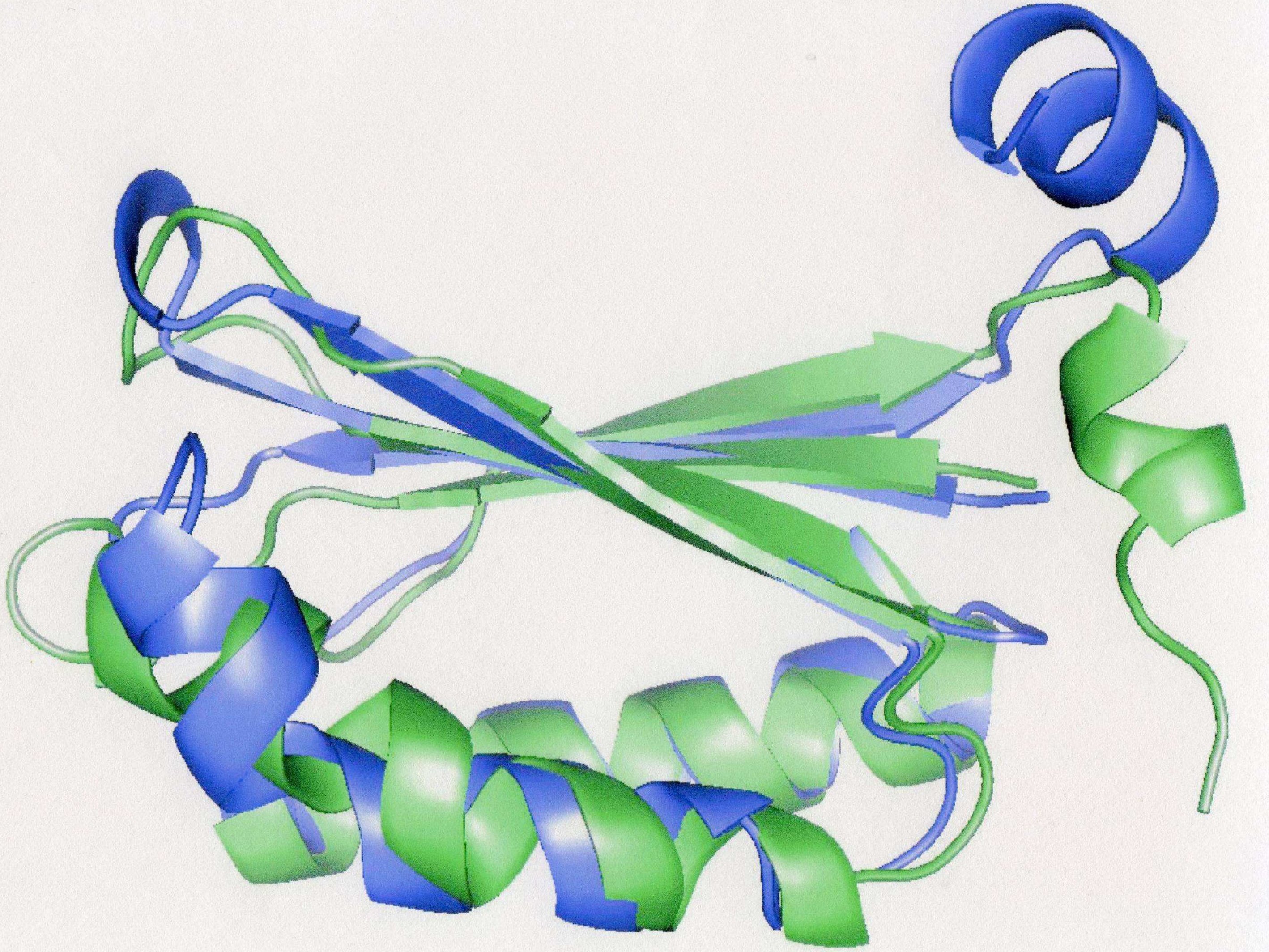
Homology Modeling

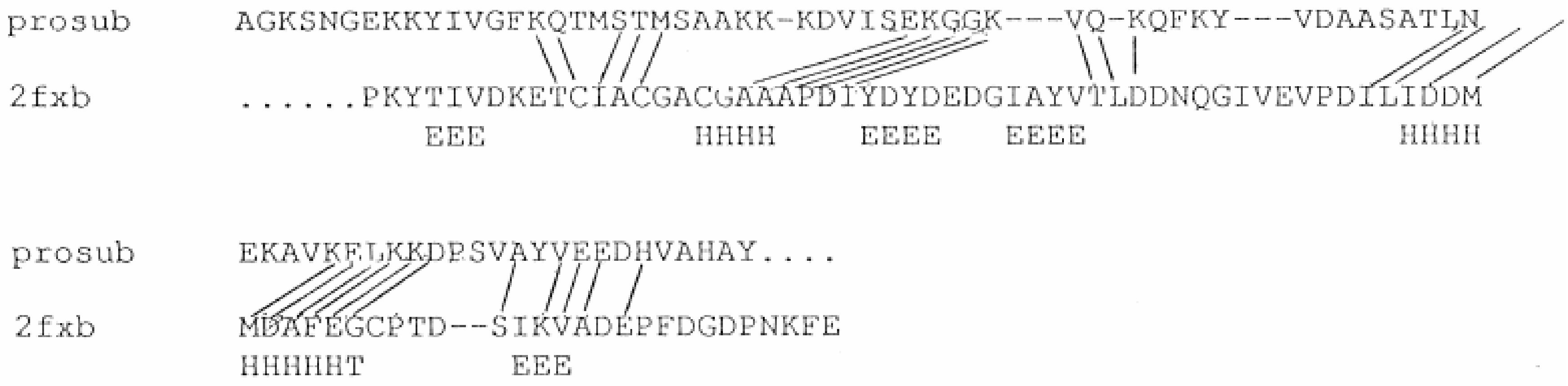
Ab Initio. D

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



Ex. Zone 2





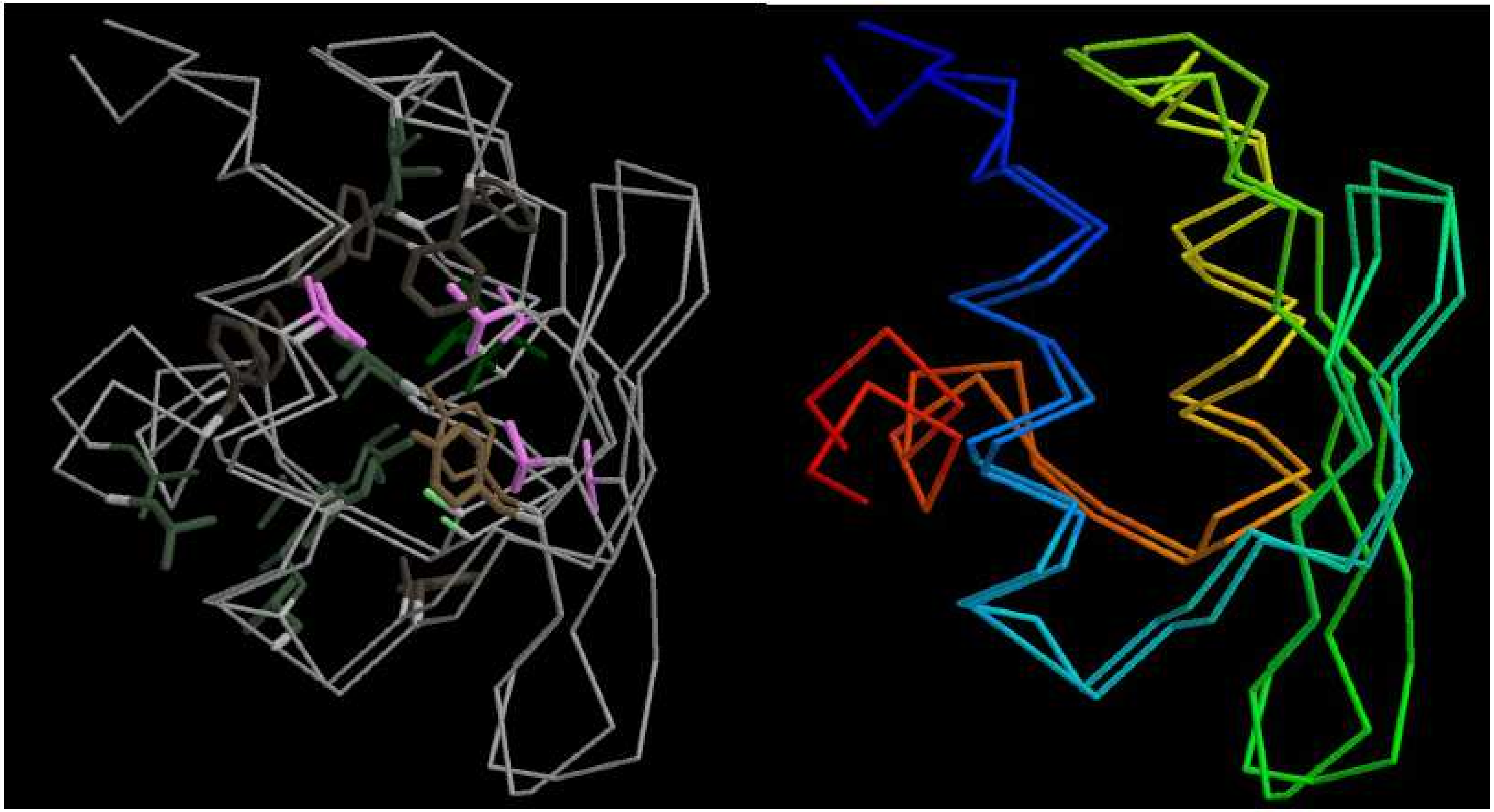
```
prosub     AGKSNGEKKYIVGFKQTMSTMSAAKK-KDVISEKGGK---VQ-KQFKY---VDAASATLN

2fxb       ......PKYTIVDKETCIACGACGAAAPDIYDYDEDGIAYVTLDDNQGIVEVPDILIDDM
                    EEE          HHHH    EEEE      EEEE               HHHH


prosub     EKAVKELKKDPSVAYVEEDHVAHAY....

2fxb       MDAFEGCPTD--SIKVADEPFDGDPNKFE
           HHHHHT          EEE
```

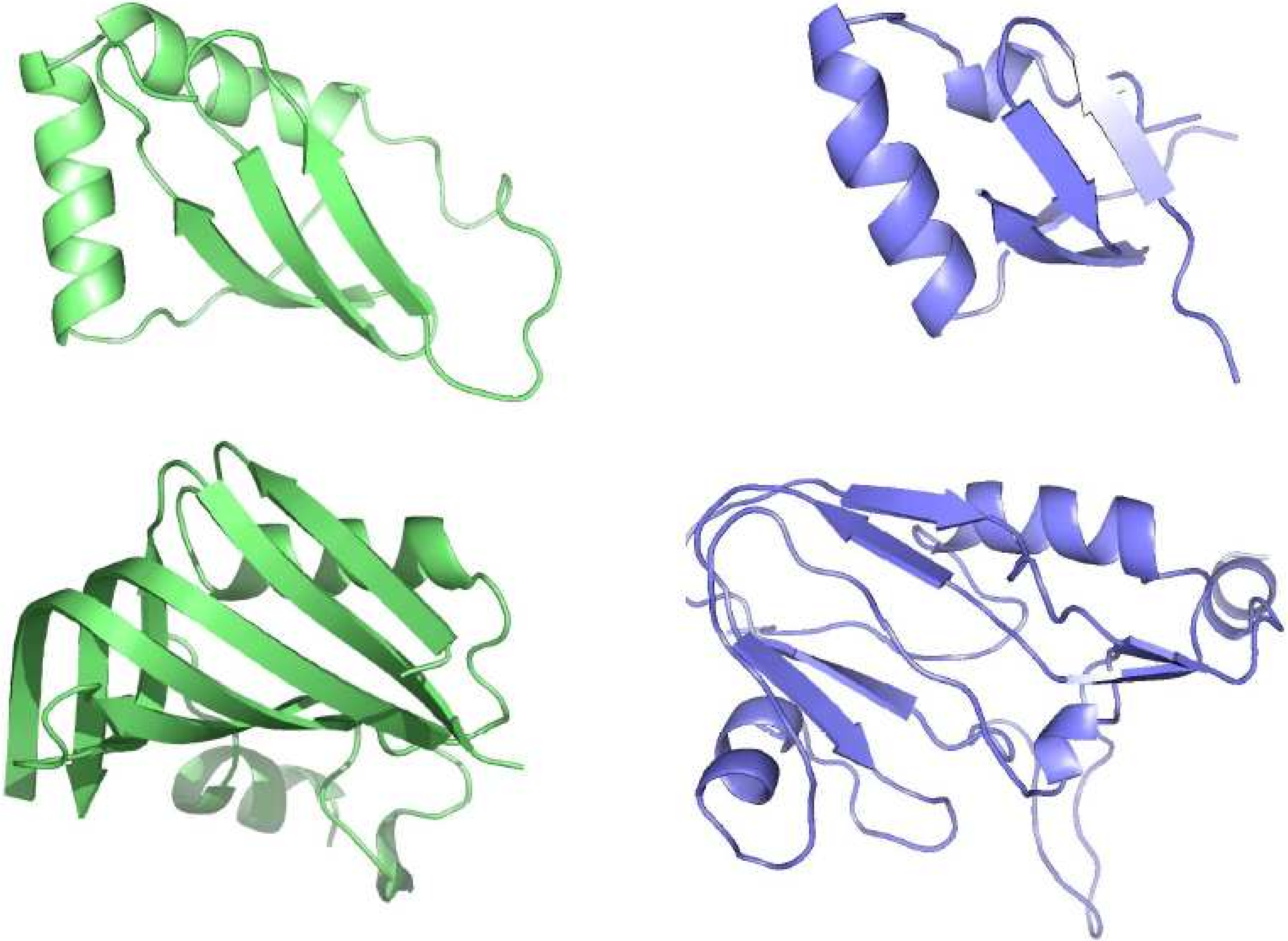
Zone 2 Conclusions

- 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.

ZONE 3

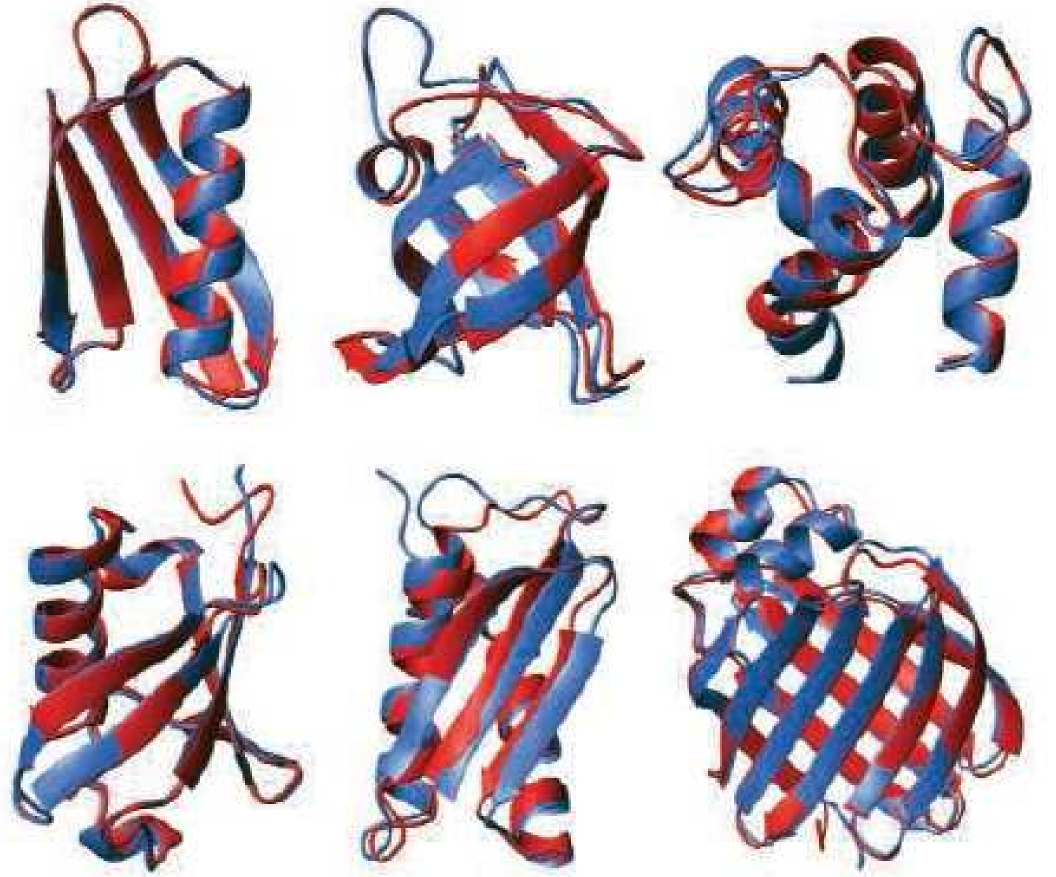
T0281 *ab initio* prediction (1.59Å)





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.