Protein Structure Prediction

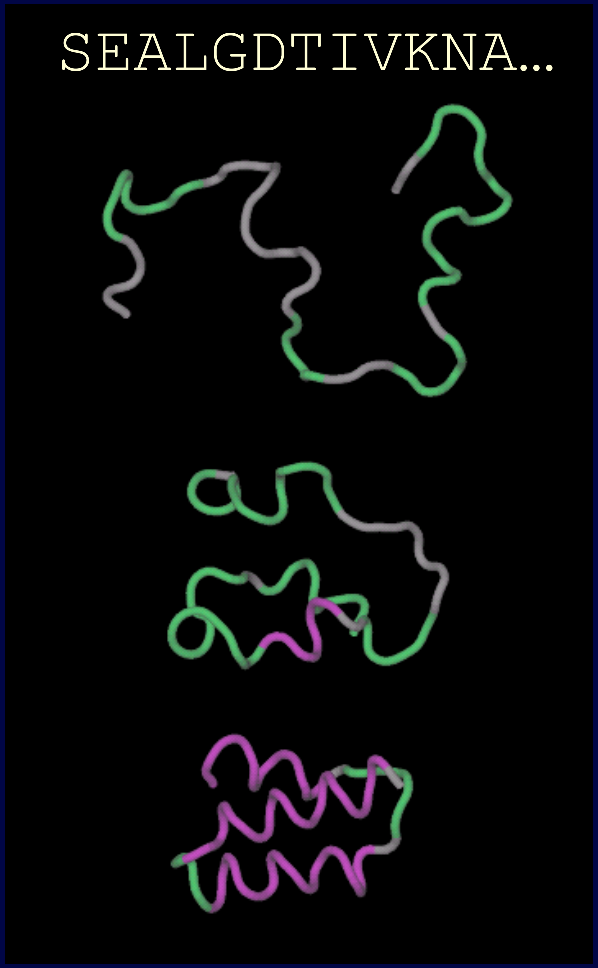
Michael Feig

MMTSB/CTBP

2006 Summer Workshop

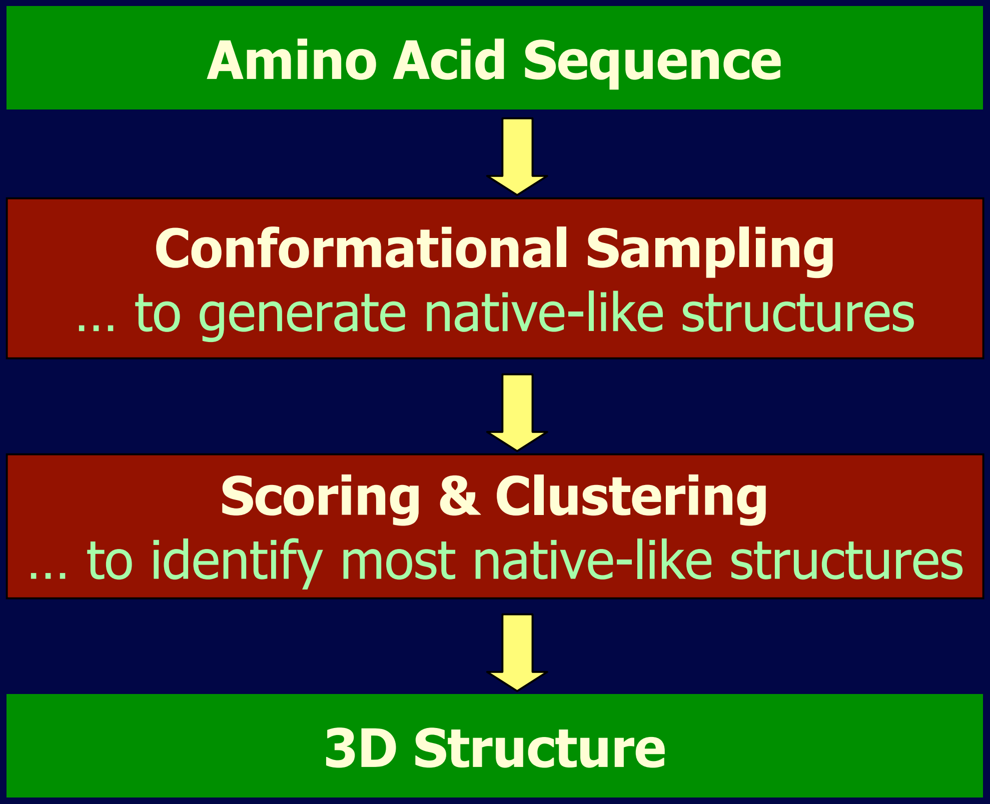
© Michael Feig, 2006.

From Sequence to Structure



© Michael Feig, 2006.

Ab initio Structure Prediction Protocol



Amino Acid Sequence

Conformational Sampling
… to generate native-like structures

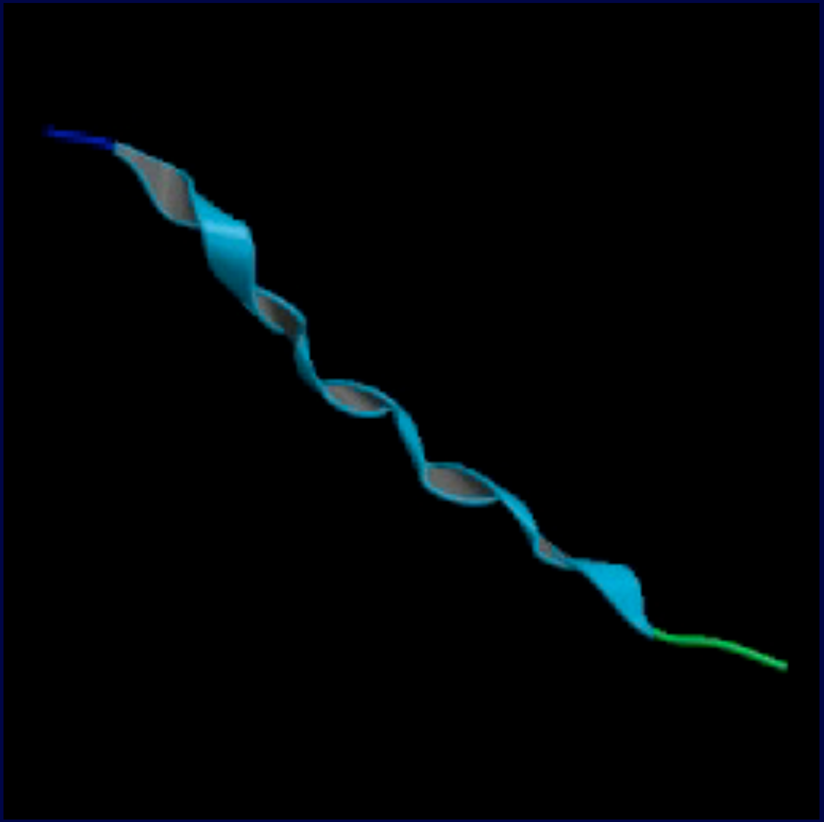
Scoring & Clustering
… to identify most native-like structures

3D Structure

© Michael Feig, 2006.

Folding with All-Atom Models

AAQAAAAQAAAAQAA

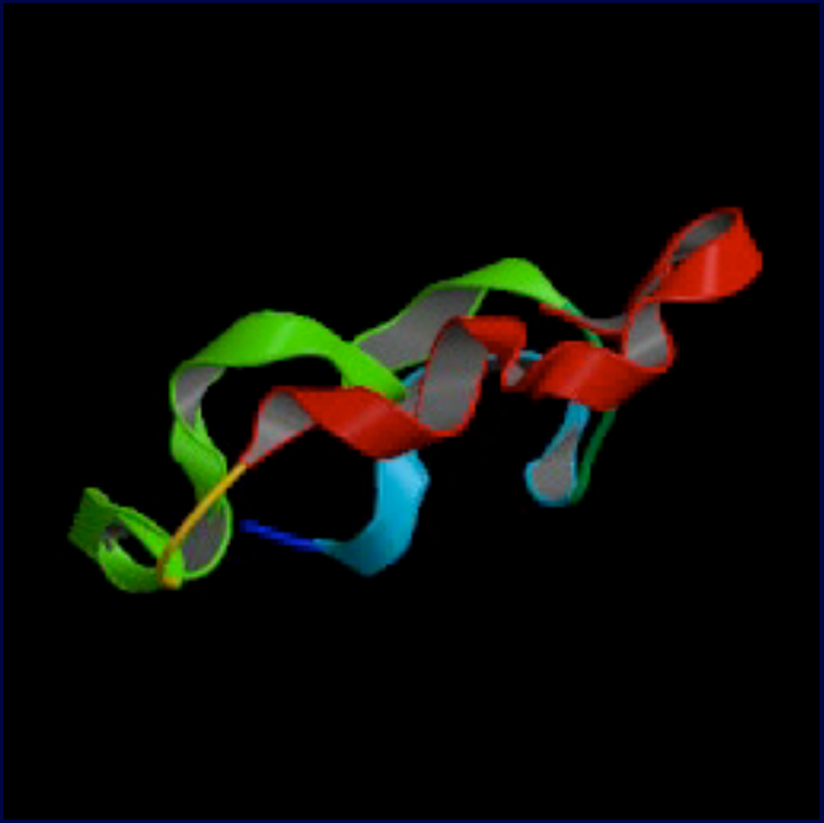


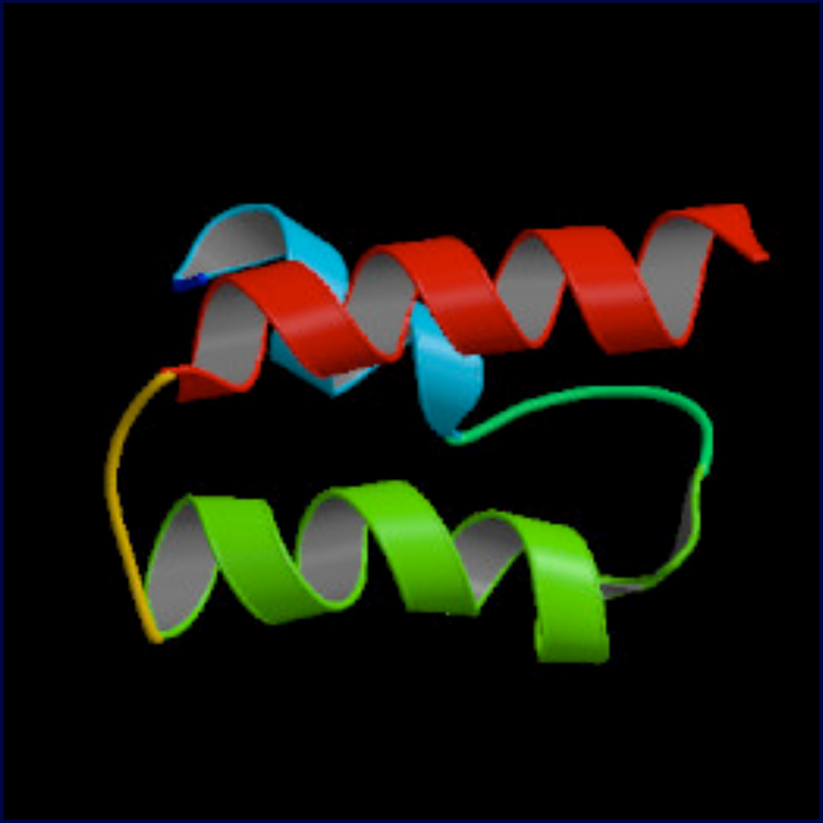
CHARMM force field
Implicit solvent replica exchange simulations
8 replicas, 10 ns/replica

© Michael Feig, 2006.

Folding with Low-Resolution Model

EQQNAFYEILHLPNLNEEQRNGFIQSLKDDPSQSANLLAEAKKLNDAQA

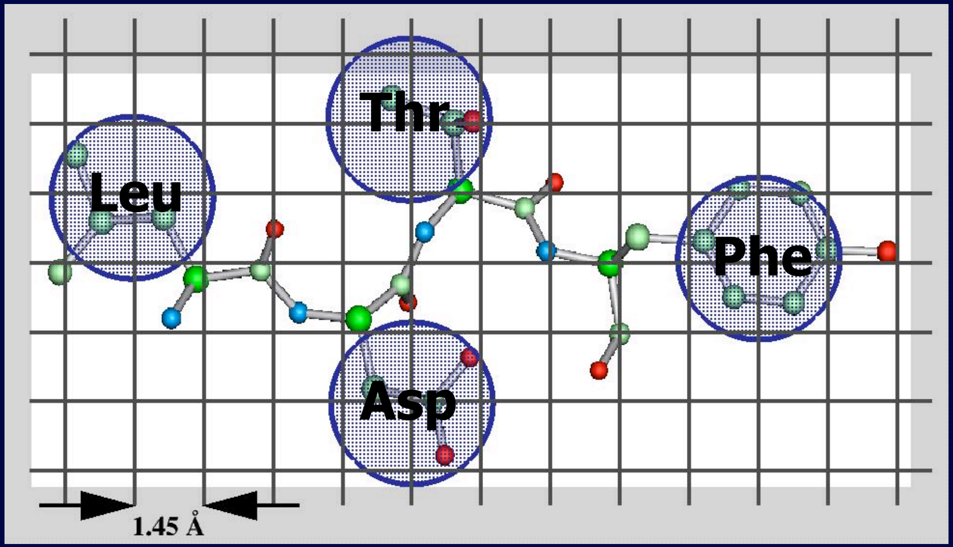




SICHO model, MONSSTER simulated annealing run

© Michael Feig, 2006.

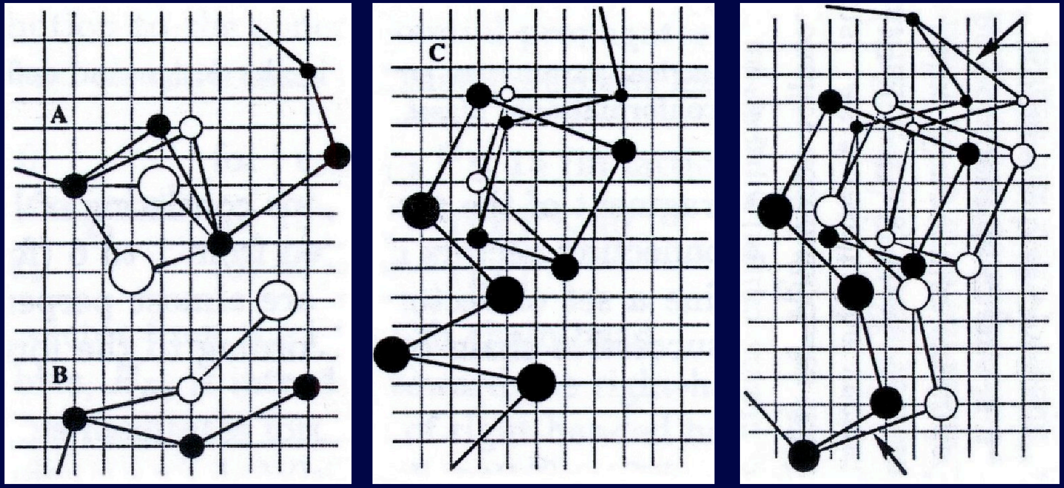
SICHO Lattice Model



Monte Carlo simulations:

- > Attempt move
- > Compute ΔE
- > Accept with probability p:

$$p = \begin{cases} 1 & \Delta E \leq 0 \\ \exp(-\Delta E / k_B T) & \Delta E > 0 \end{cases}$$



Simulated annealing

Constant Temperature

Replica Exchange Sampling

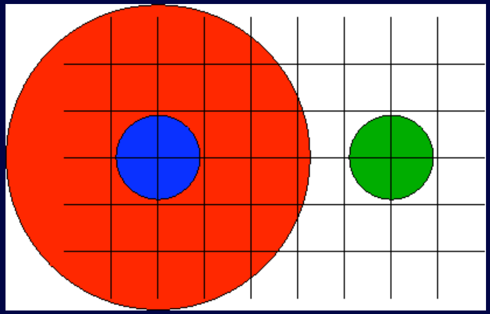
Kolinski & Skolnick: Proteins *32*, 475 (1998)

© Michael Feig, 2006.

SICHO Energy Function

Knowledge-Based Terms

□ Excluded volume



□ Side chain burial propensity

follows Kyte-Doolittle scale

Ala	1.8
Arg	-4.5
Asp	-3.5
Ile	4.5

□ Centrosymmetric bias

$$r_g = 2.2\ N_{res}^{0.38}$$

© Michael Feig, 2006.

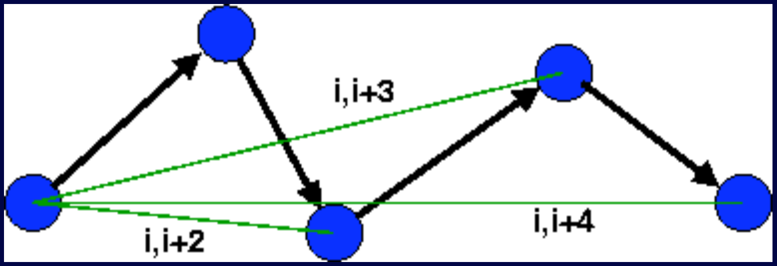
SICHO Energy Function

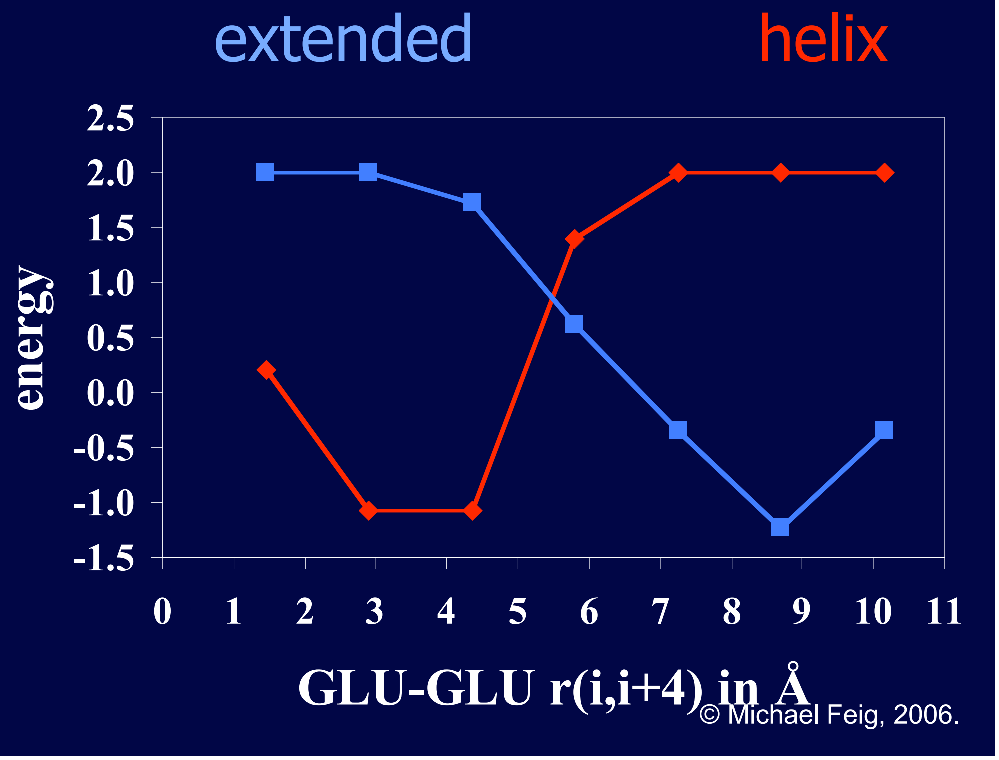
Statistical Terms

Potential of mean force (PMF):

$$\frac{p_i}{p_j} = e^{-\frac{\Delta E_{ij}}{k_B T}}$$

$$\Delta E = -kT\ln(p)$$

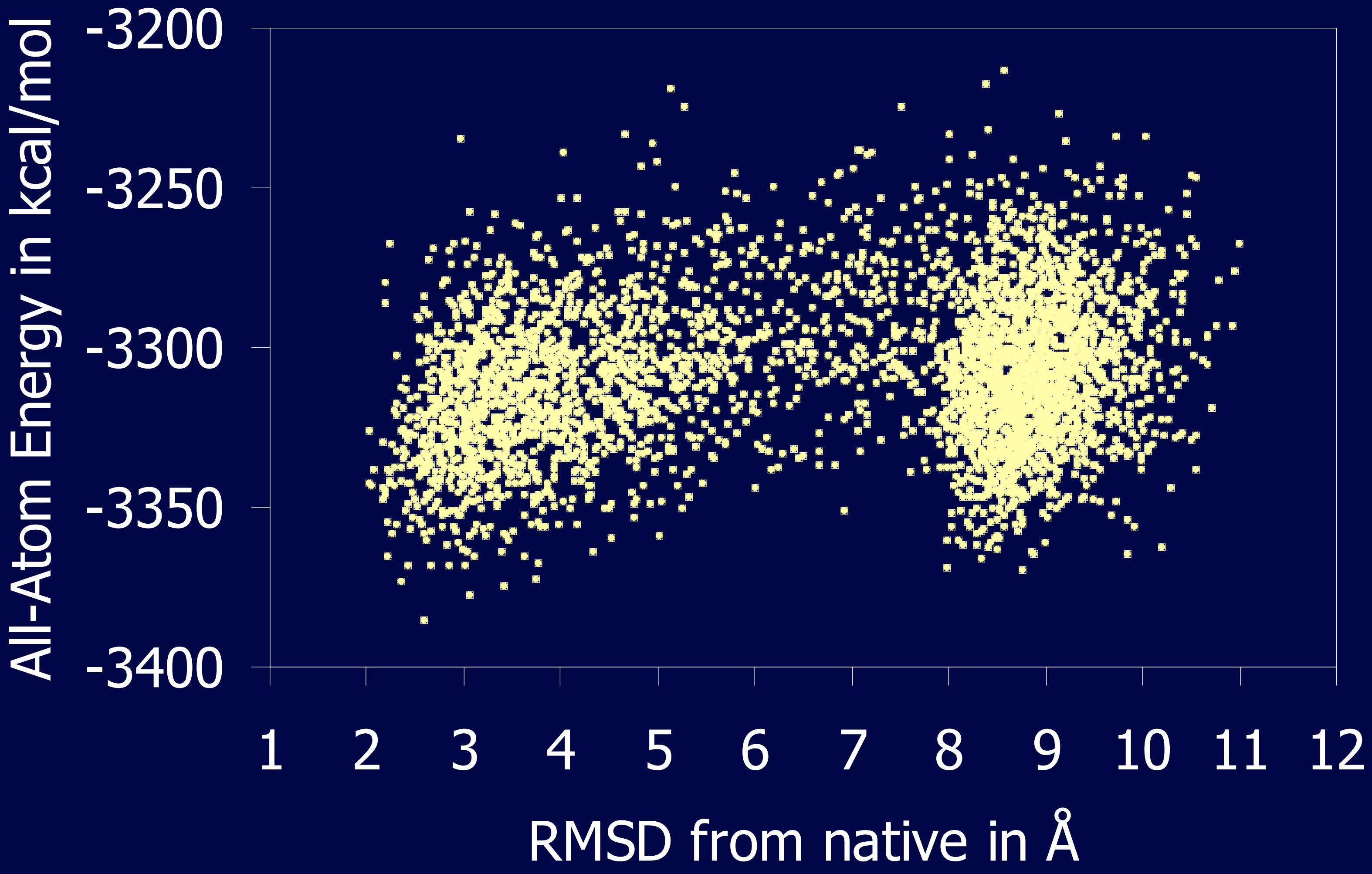




© Michael Feig, 2006.

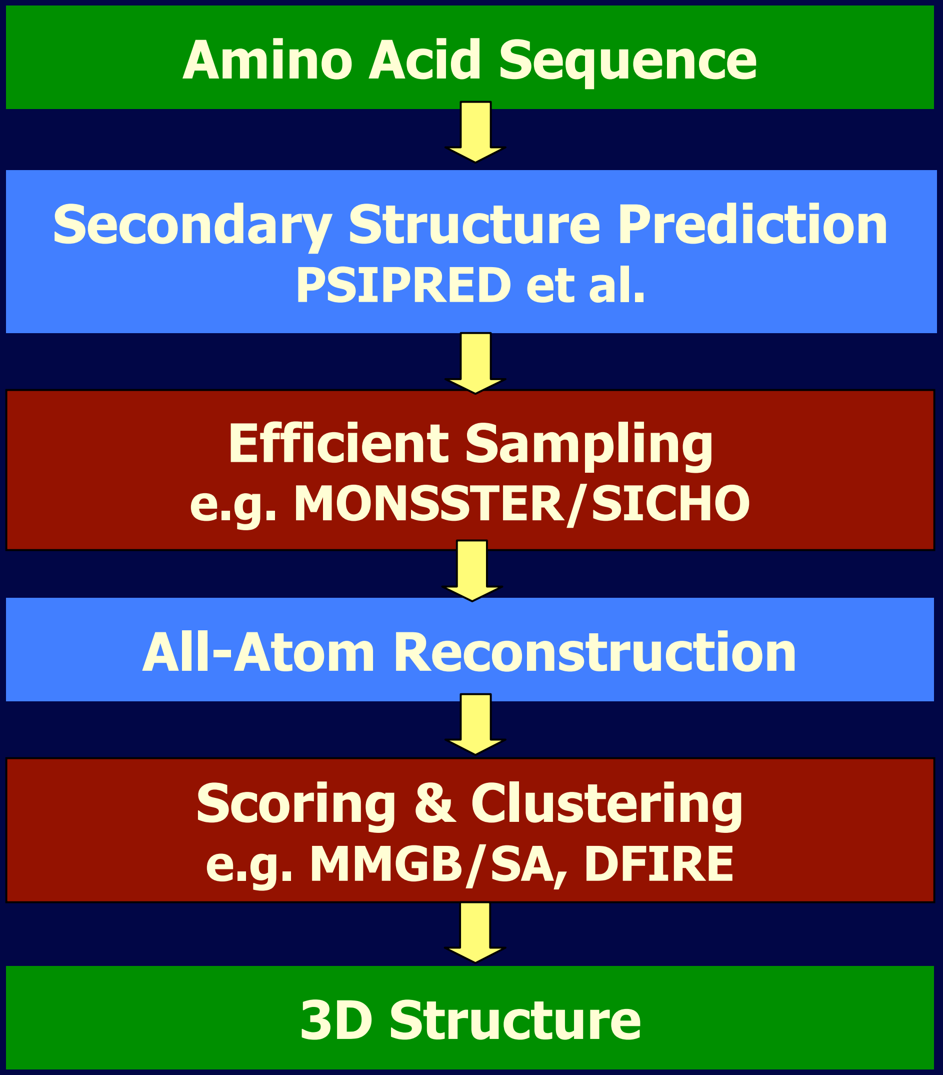
Conformational Sampling with SICHO

Protein A



© Michael Feig, 2006.

Ab initio Structure Prediction Protocol



Amino Acid Sequence

Secondary Structure Prediction
PSIPRED et al.

Efficient Sampling
e.g. MONSSTER/SICHO

All-Atom Reconstruction

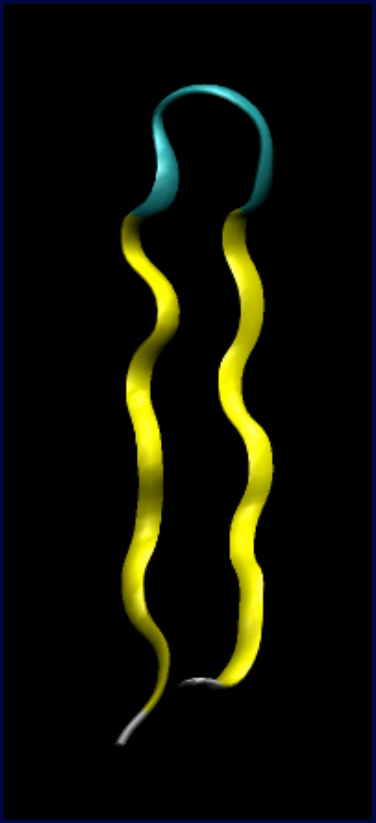
Scoring & Clustering
e.g. MMGB/SA, DFIRE

3D Structure

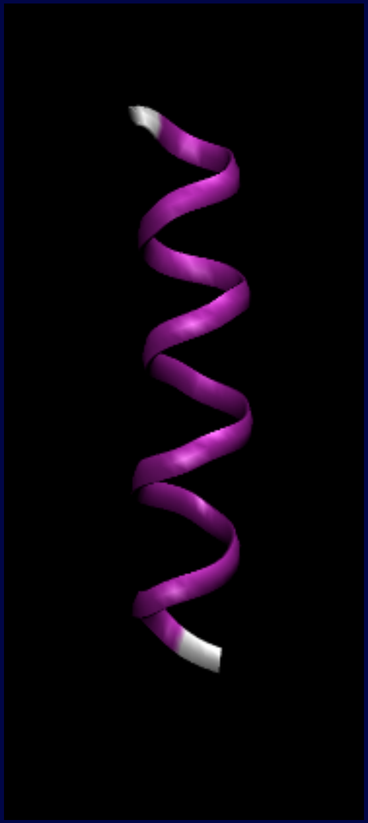
© Michael Feig, 2006.

Secondary Structure Prediction

…GDPIVKNAKLDSRLANKEALRLL…

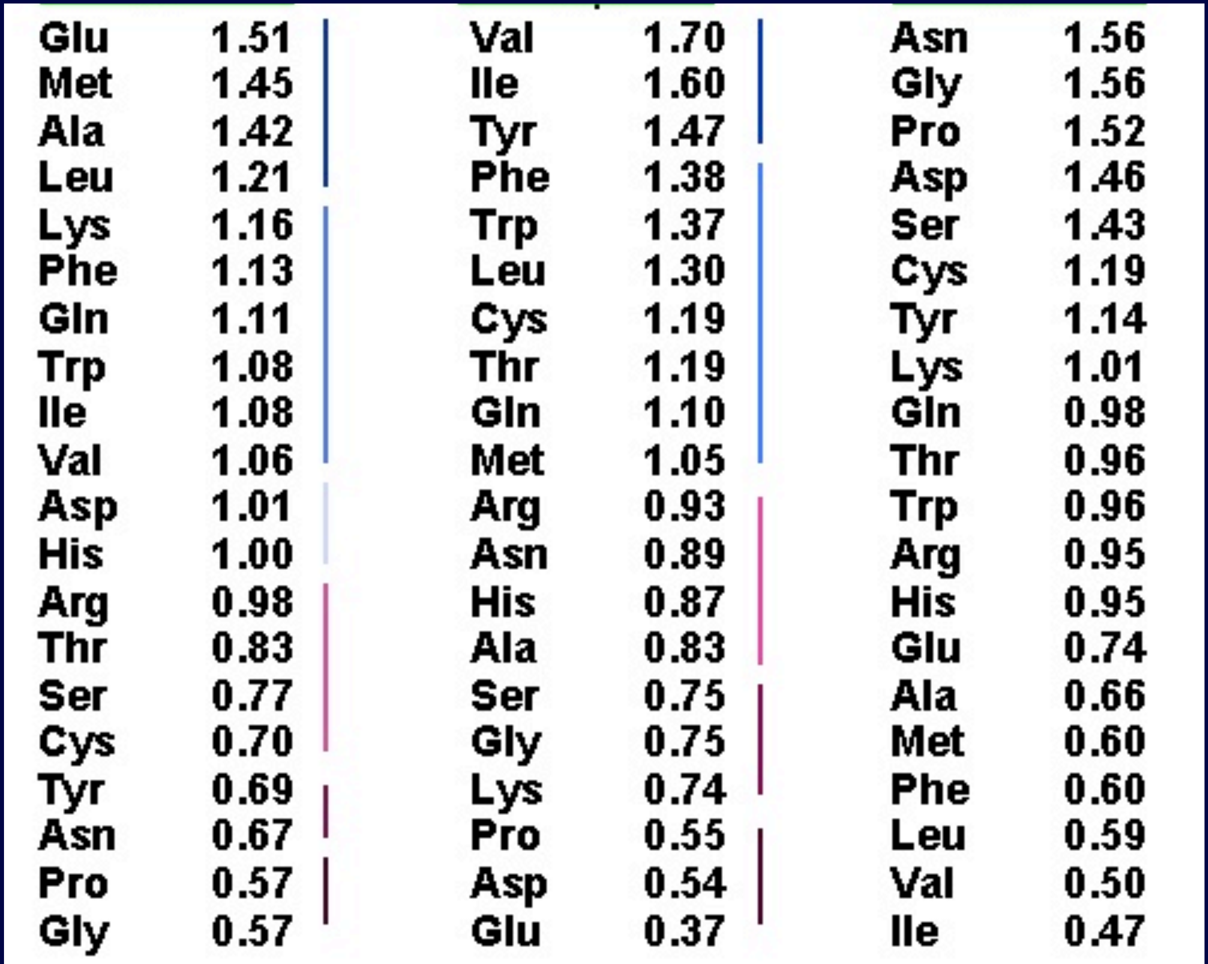


?



© Michael Feig, 2006.

Secondary Structure Propensities

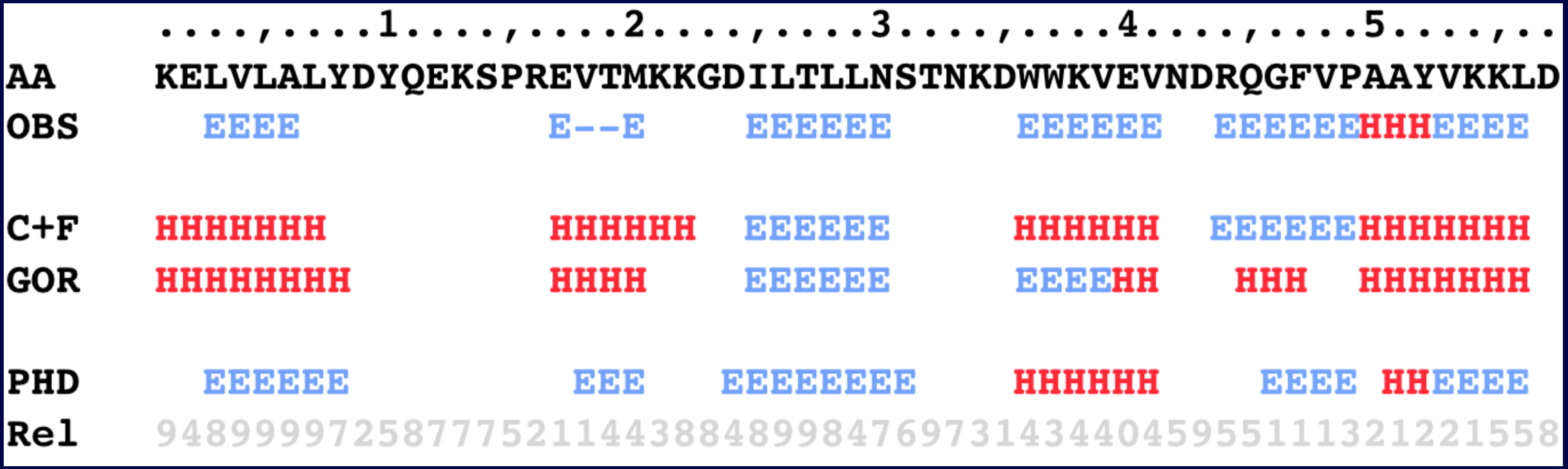


α-helix		β-sheet		turn	
Glu	1.51	Val	1.70	Asn	1.56
Met	1.45	Ile	1.60	Gly	1.56
Ala	1.42	Tyr	1.47	Pro	1.52
Leu	1.21	Phe	1.38	Asp	1.46
Lys	1.16	Trp	1.37	Ser	1.43
Phe	1.13	Leu	1.30	Cys	1.19
Gln	1.11	Cys	1.19	Tyr	1.14
Trp	1.08	Thr	1.19	Lys	1.01
Ile	1.08	Gln	1.10	Gln	0.98
Val	1.06	Met	1.05	Thr	0.96
Asp	1.01	Arg	0.93	Trp	0.96
His	1.00	Asn	0.89	Arg	0.95
Arg	0.98	His	0.87	His	0.95
Thr	0.83	Ala	0.83	Glu	0.74
Ser	0.77	Ser	0.75	Ala	0.66
Cys	0.70	Gly	0.75	Met	0.60
Tyr	0.69	Lys	0.74	Phe	0.60
Asn	0.67	Pro	0.55	Leu	0.59
Pro	0.57	Asp	0.54	Val	0.50
Gly	0.57	Glu	0.37	Ile	0.47

Chou & Fasman (1974)

© Michael Feig, 2006.

Secondary Structure Prediction Methods



```
      ....,....1....,....2....,....3....,....4....,....5....,..
AA    KELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD
OBS     EEEE          E--E    EEEEEE     EEEEEE  EEEEEEHHHEEEE

C+F   HHHHHHH         HHHHHH  EEEEEE     HHHHHH  EEEEEEHHHHHHH
GOR   HHHHHHHH        HHHH    EEEEEE     EEEEHH   HHH  HHHHHHH

PHD     EEEEEE         EEE   EEEEEEEE    HHHHHH    EEEE HHEEEE
Rel   948999972587775211443884899847697314344045955111321221558
```

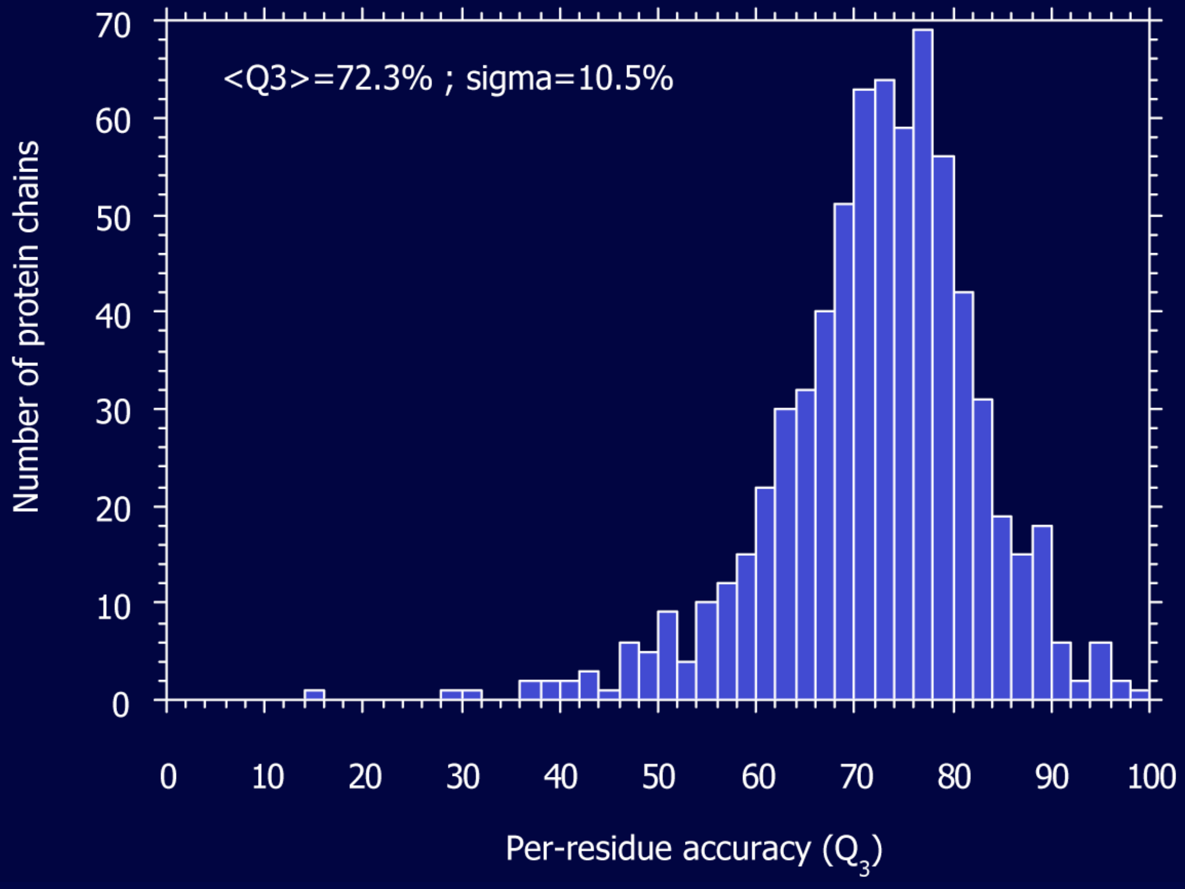
SABLE	77.6%
PSIPRED	76.2%
PSSP	75.1%
SAM-T99-sec	76.1%
PHD	72.3%
C+F	50-60%

C+F: Chou & Fasman
GOR: Garnier, Osguthorpe, Robson

© Michael Feig, 2006.

Secondary structure prediction

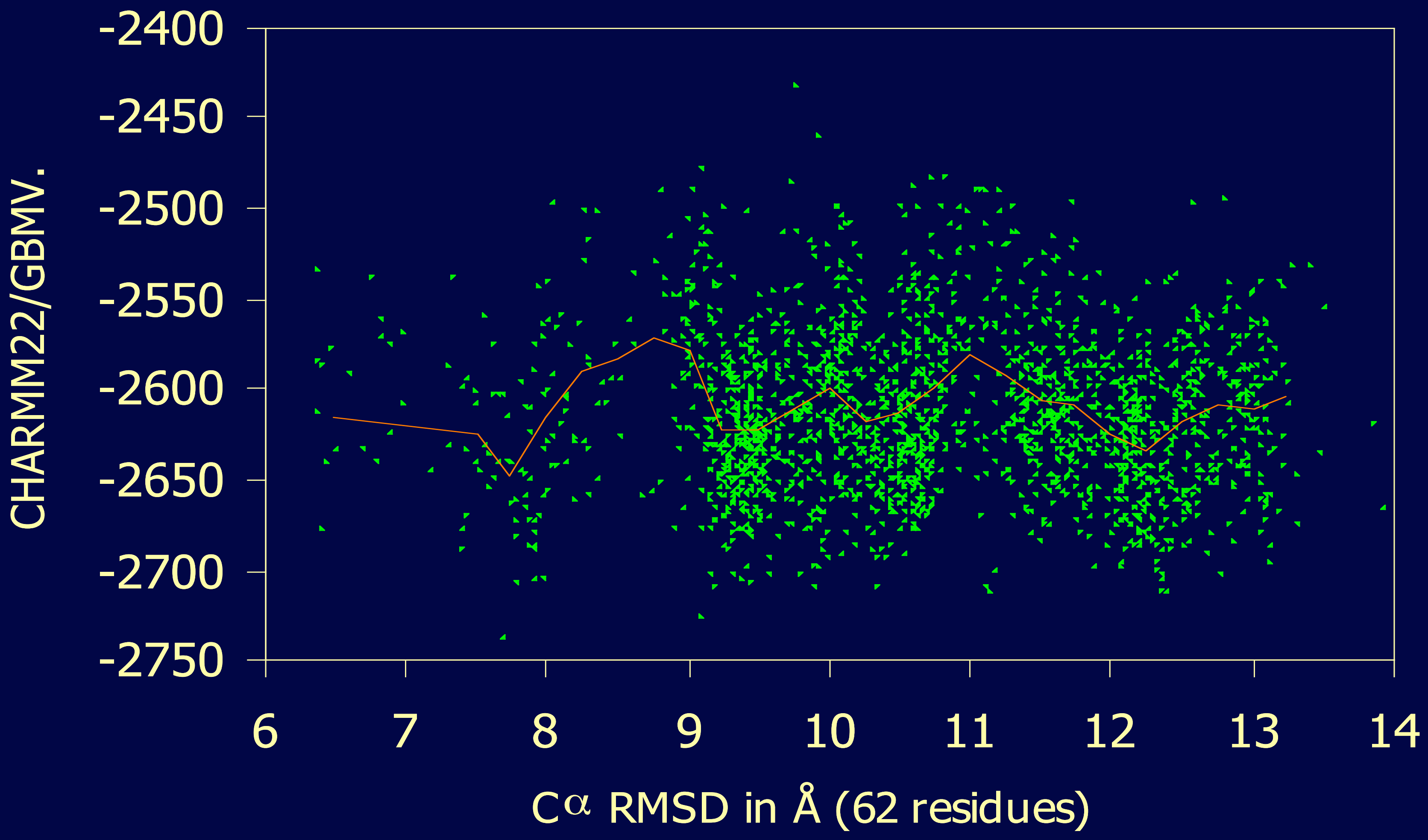
Per-Residue Accuracy



MMTSB/CTBP Summer Workshop

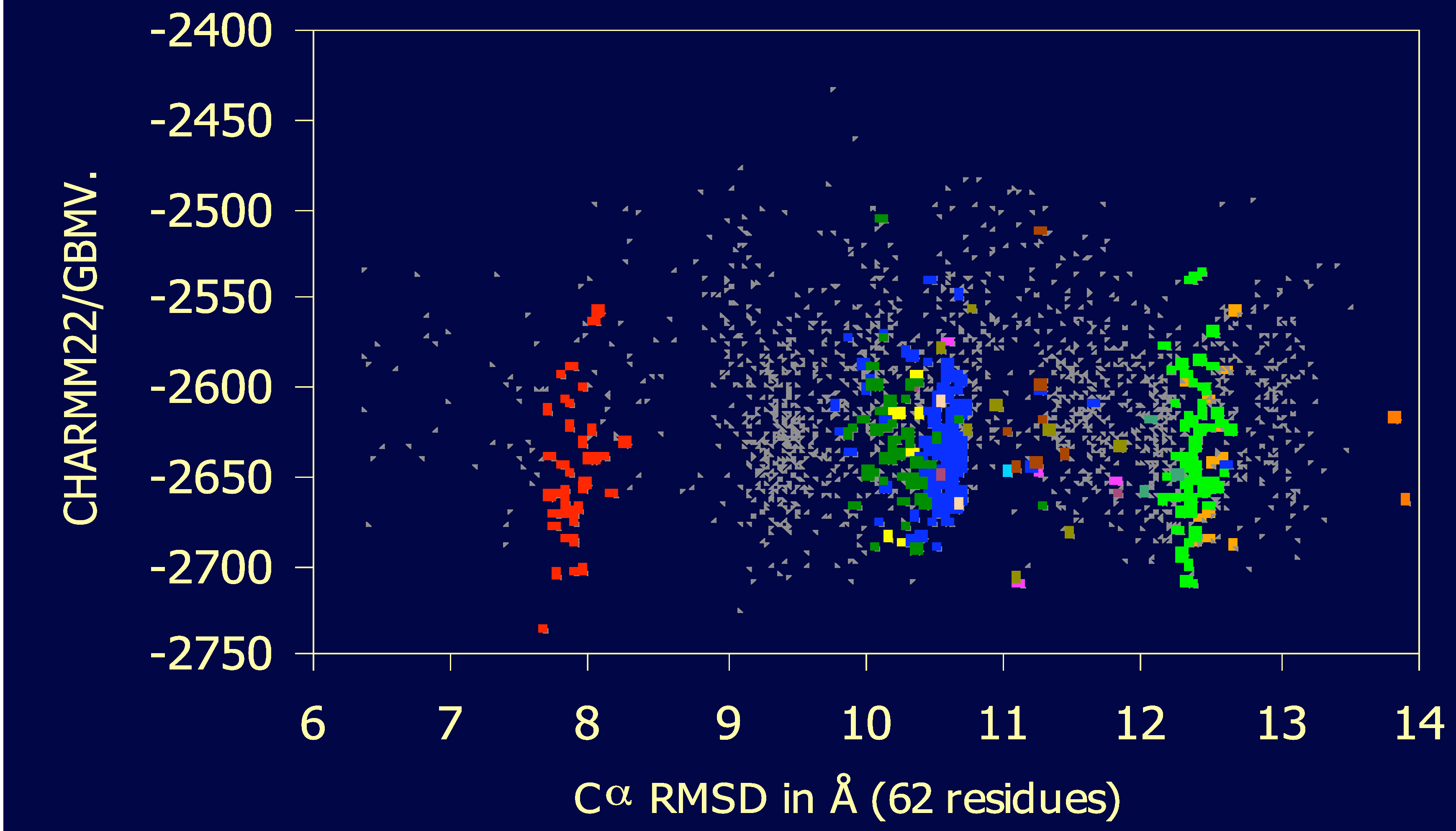
© Michael Feig, 2006.

Realistic ab initio Structure Prediction



© Michael Feig, 2006.

Sampling, Scoring, Clustering

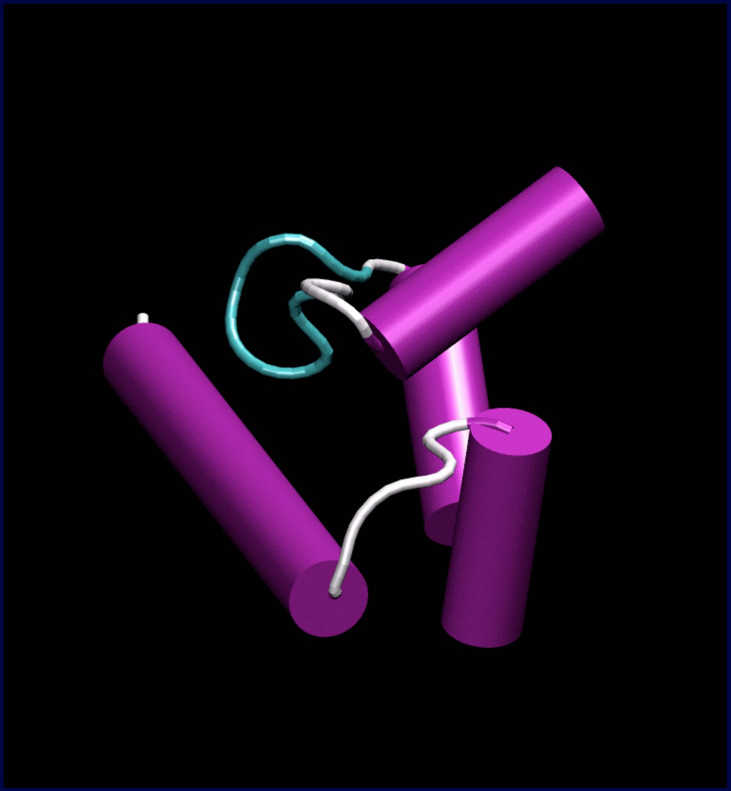


MMTSB/CTBP Summer Workshop

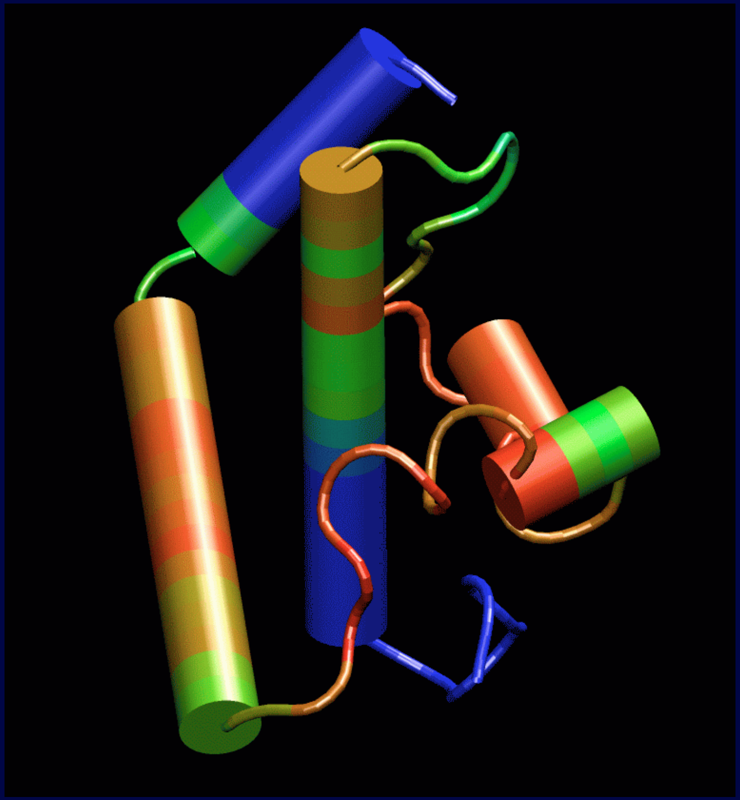
© Michael Feig, 2006.

Ab initio Predictions

DNase fragmentation factor



NMR structure 1KOY



Best-scoring prediction
7.4 Å RMSD

© Michael Feig, 2006.

Scoring Functions

- □ **Knowledge-based/statistical**

  derived from known protein structures
  limited by training data
  usually fast
  e.g. DFIRE, RAPDF, prosaII

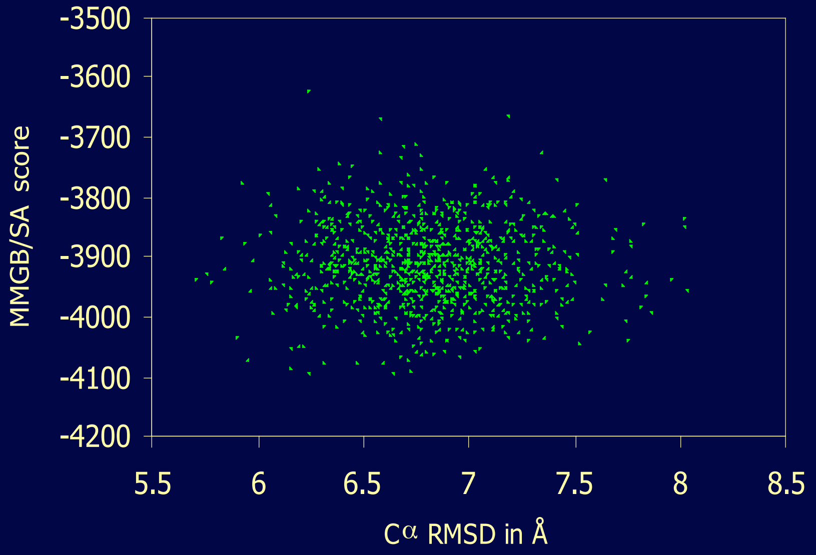
- □ **Force field based**

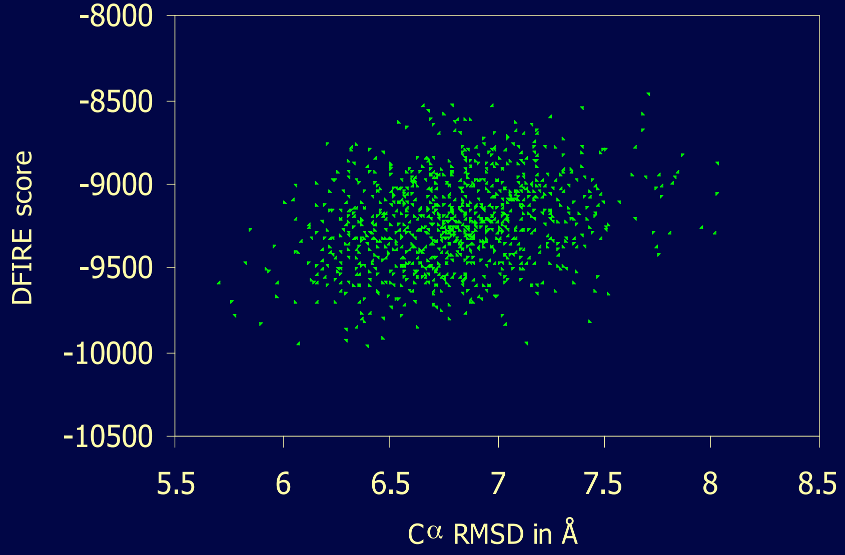
  model physical energy landscape
  more robust and transferable
  often expensive (require minimization)
  e.g. MMPB(GB)/SA, UNRES

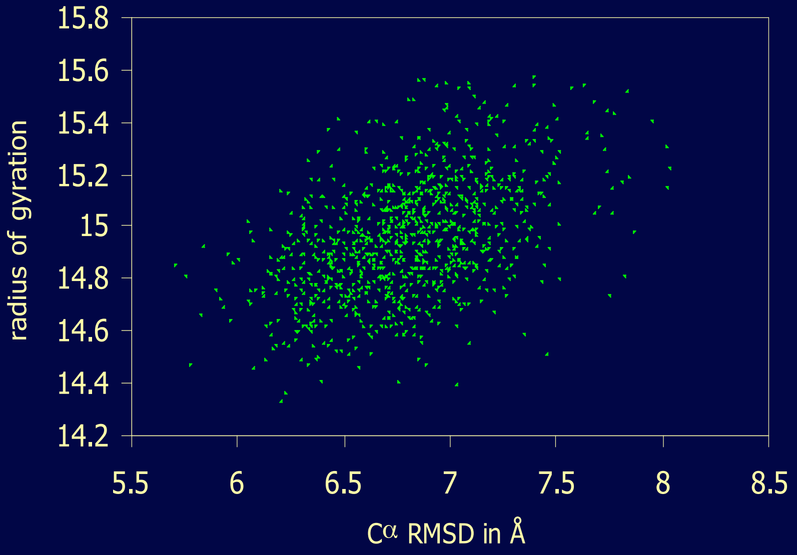
© Michael Feig, 2006.

Scoring Function Comparison

MMGB/SA vs. DFIRE







© Michael Feig, 2006.

Sampling with Restraints

□ **Secondary structure bias**

- Secondary structure prediction
- NMR shift data

□ **Distance restraints**

- Experimental restraints (disulfides, NMR, EPR)
- Side chain contacts from analogous structures

□ **Shape restraints**

- cryoEM data, small-angle X-ray scattering

© Michael Feig, 2006.

… but the solution may lie elsewhere.

© Michael Feig, 2006.

Sequence Homology

Human thioredoxin (1AUC)

```
SDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGKLTV
A
  : :.      ..    .:  ..::: : :::::::: :.. .....:.. . .
MVKQIESKTAFQEALDAAGDKLVVVDFSATWCGPCKMIKPFFHSLSEKYSNVIFL
-


KLNIDQNPGTAPKYGIRGIPTLLLFKNGEVAATKVGALSKGQLKEFLDAN---LA
....:.   .: .  .. .::. .::.:.    ::: .: :  :: :.:.   :.
EVDVDDCQDVASECEVKCMPTFQFFKKGQ----KVGEFS-GANKEKLEATINELV
```

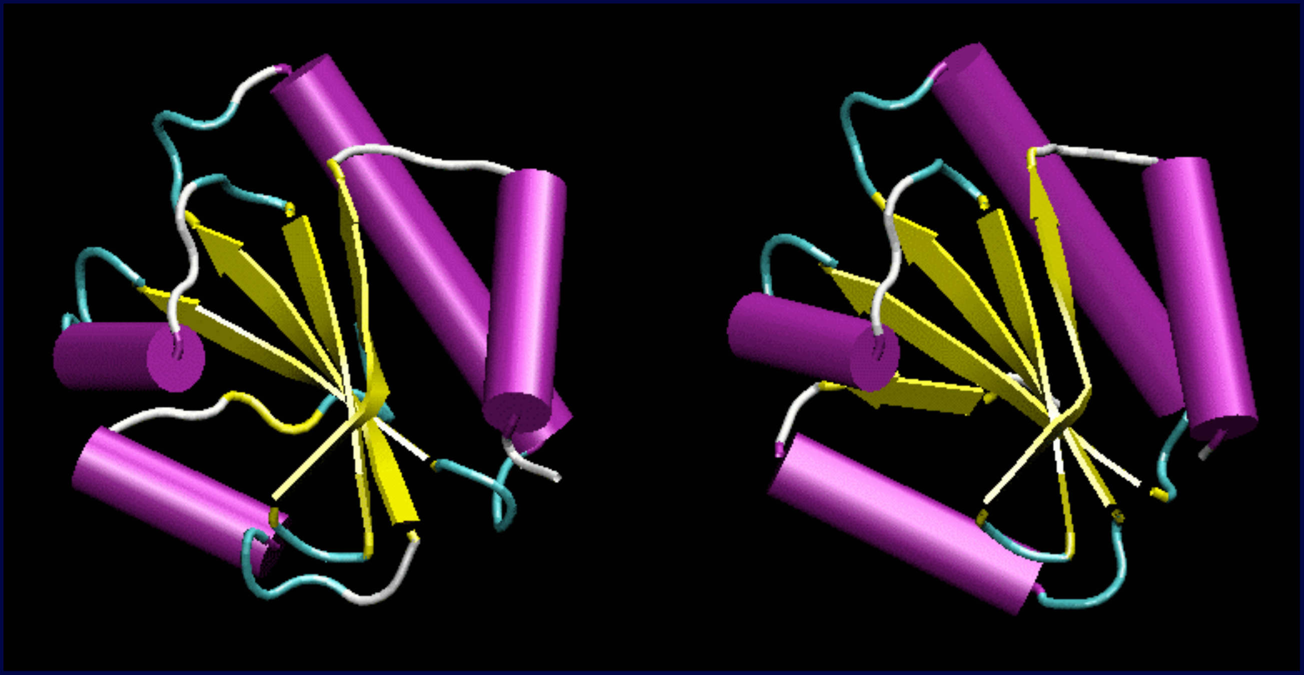
E. Coli thioredoxin (1THO)

© Michael Feig, 2006.

Comparative Modeling

Assumption:

Proteins with similar sequence have similar structure

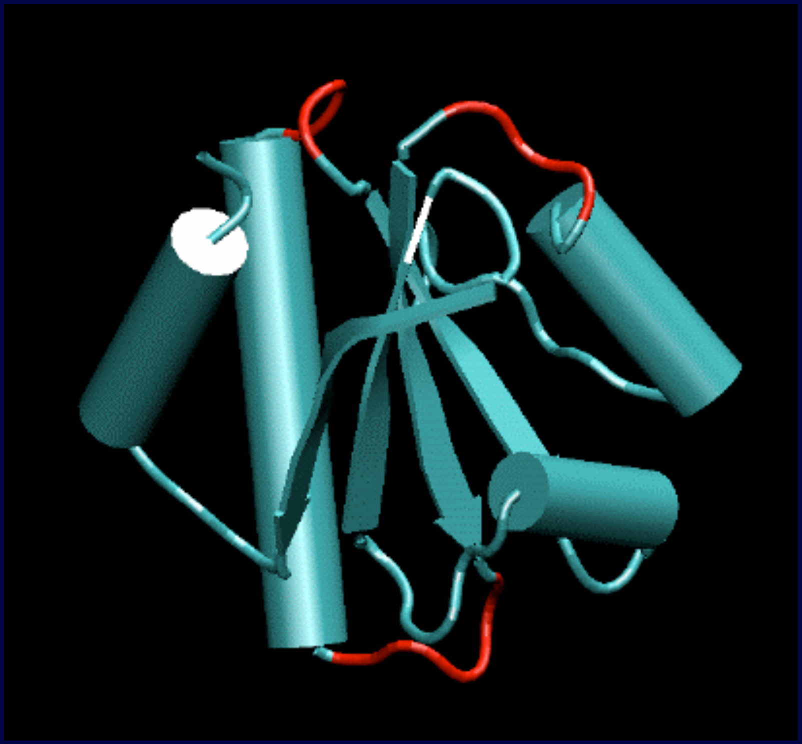


Human thioredoxin
(1AUC)

E. Coli thioredoxin
(1THO)

© Michael Feig, 2006.

Structural Templates from Homology



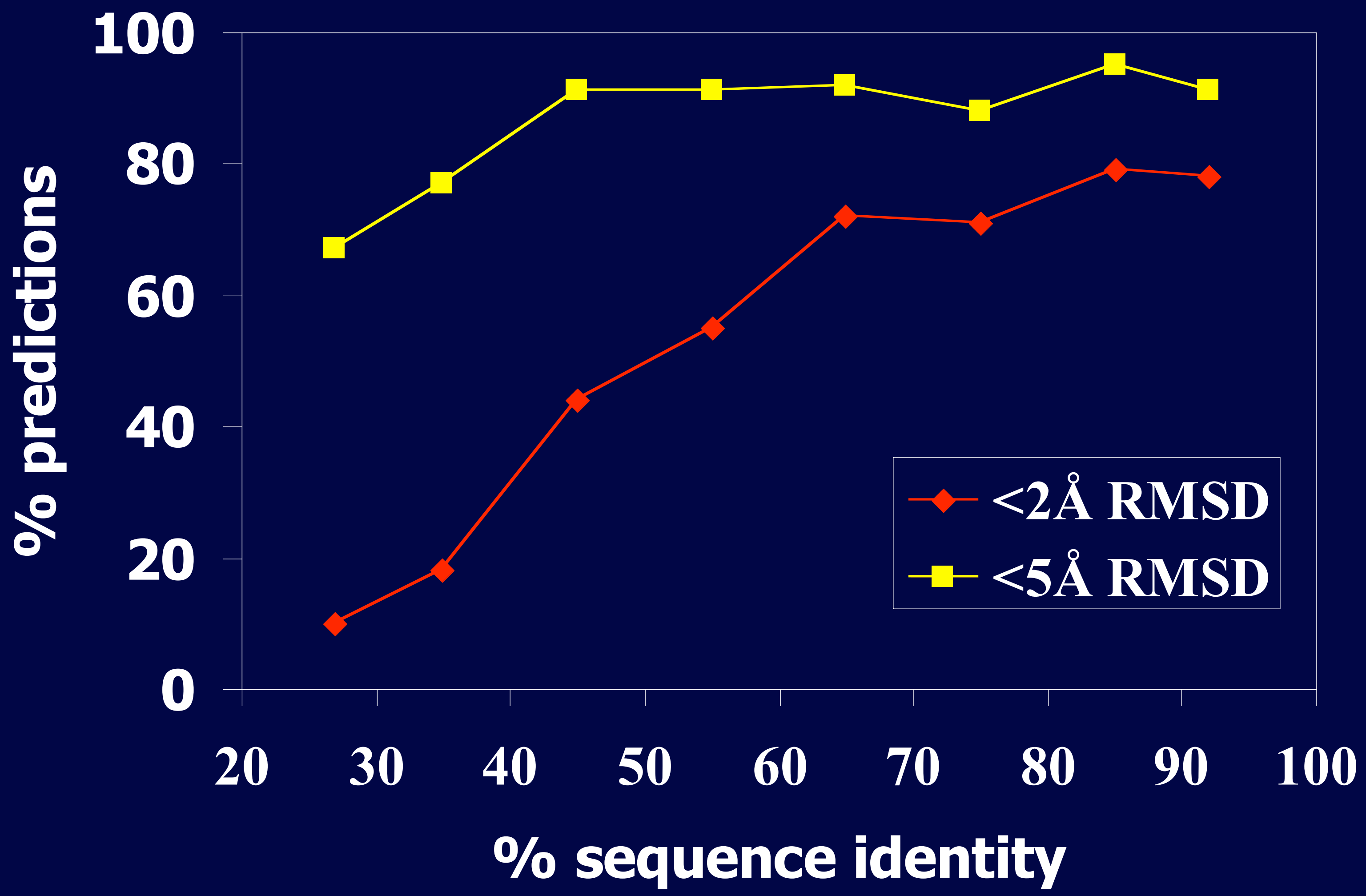
Challenges:

- □ Correct alignment
- □ Loop modeling
- □ Side chain rebuilding

```
PGTAPKYGIRGIPTLLLFKNGEVAATKVGALSKGQLKEFLDAN---LA
  .: .  .. .::. .::.:.    ::: .: :  :: :.:.   :.
QDVASECEVKCMPTFQFFKKGQ----KVGEFS-GANKEKLEATINELV
```

© Michael Feig, 2006.

Accuracy of Predictions by Homology



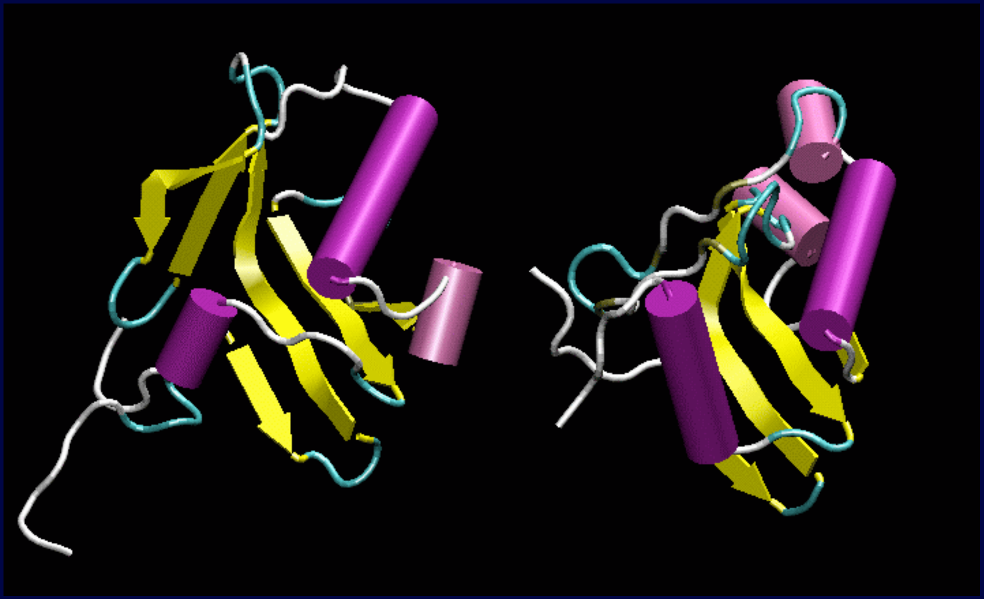
MMTSB/CTBP Summer Workshop

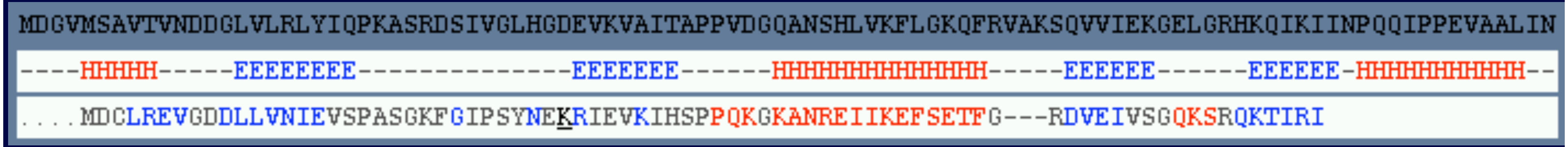
© Michael Feig, 2006.

Prediction through Fold Recognition

Assumption:

Proteins with similar secondary structure share fold





```
MDGVMSAVTVNDDGLVLRLYIQPKASRDSIVGLHGDEVKVAITAPPVDGQANSHLVKFLGKQFRVAKSQVVIEKGELGRHKQIKIINPQQIPPEVAALIN
----HHHHH-----EEEEEEEE--------------EEEEEEE------HHHHHHHHHHHHHH-----EEEEEE------EEEEEE-HHHHHHHHHHH--
....MDCLREVGDDLLVNIEVSPASGKFGIPSYNEKRIEVKIHSPPQKGKANREIIKEFSETFG---RDVEIVSGQKSRQKTIRI
```

© Michael Feig, 2006.

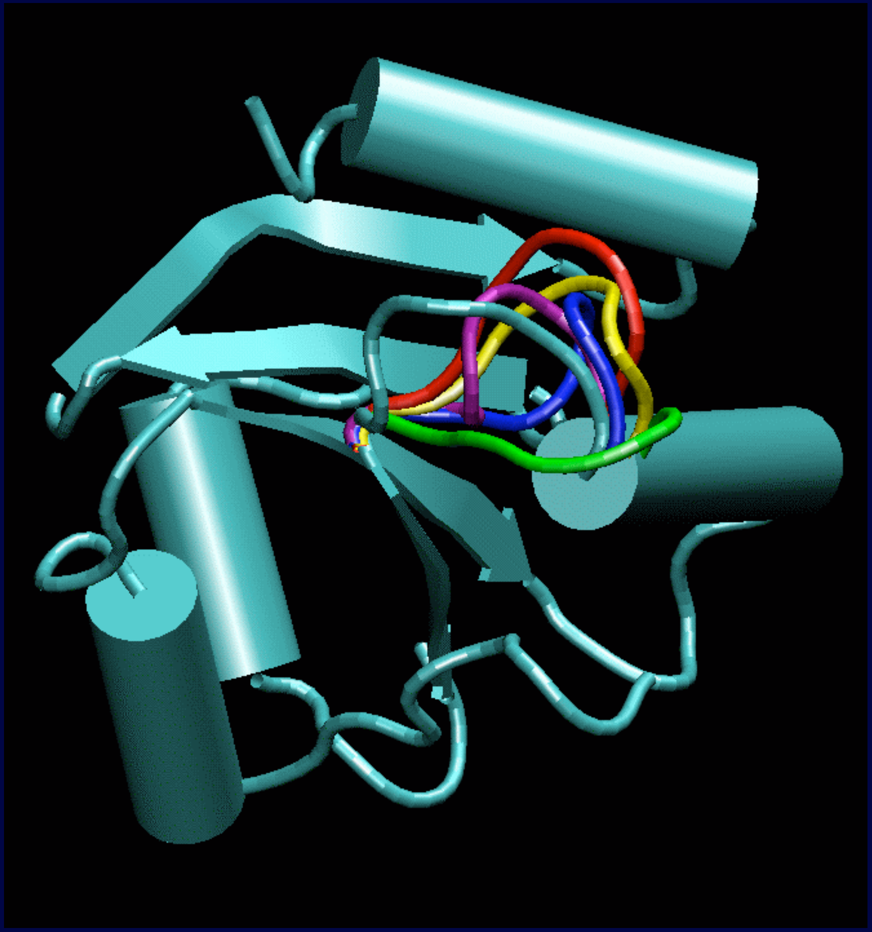
Templates through Fold Recognition

Challenges:

- □ Wrong templates
- □ Alignment uncertain
- □ Fragment modeling
- □ Refinement needed

© Michael Feig, 2006.

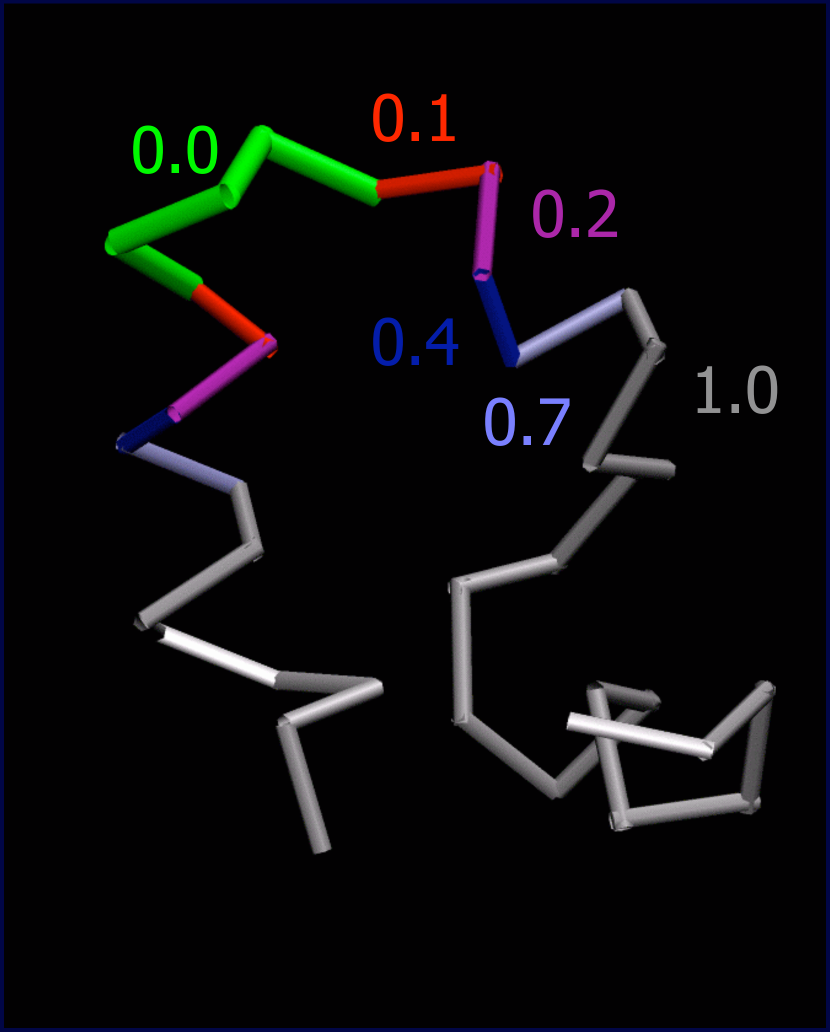
Ab initio Sampling in Template-based Structure Prediction



- □ **Template** provides known protein structure
- □ **Ab initio sampling** of unknown fragments in the context of template

© Michael Feig, 2006.

Template Restraints Near Flexible Part



Restraint potential:

$$U = f \cdot k(r - r_0)^2$$

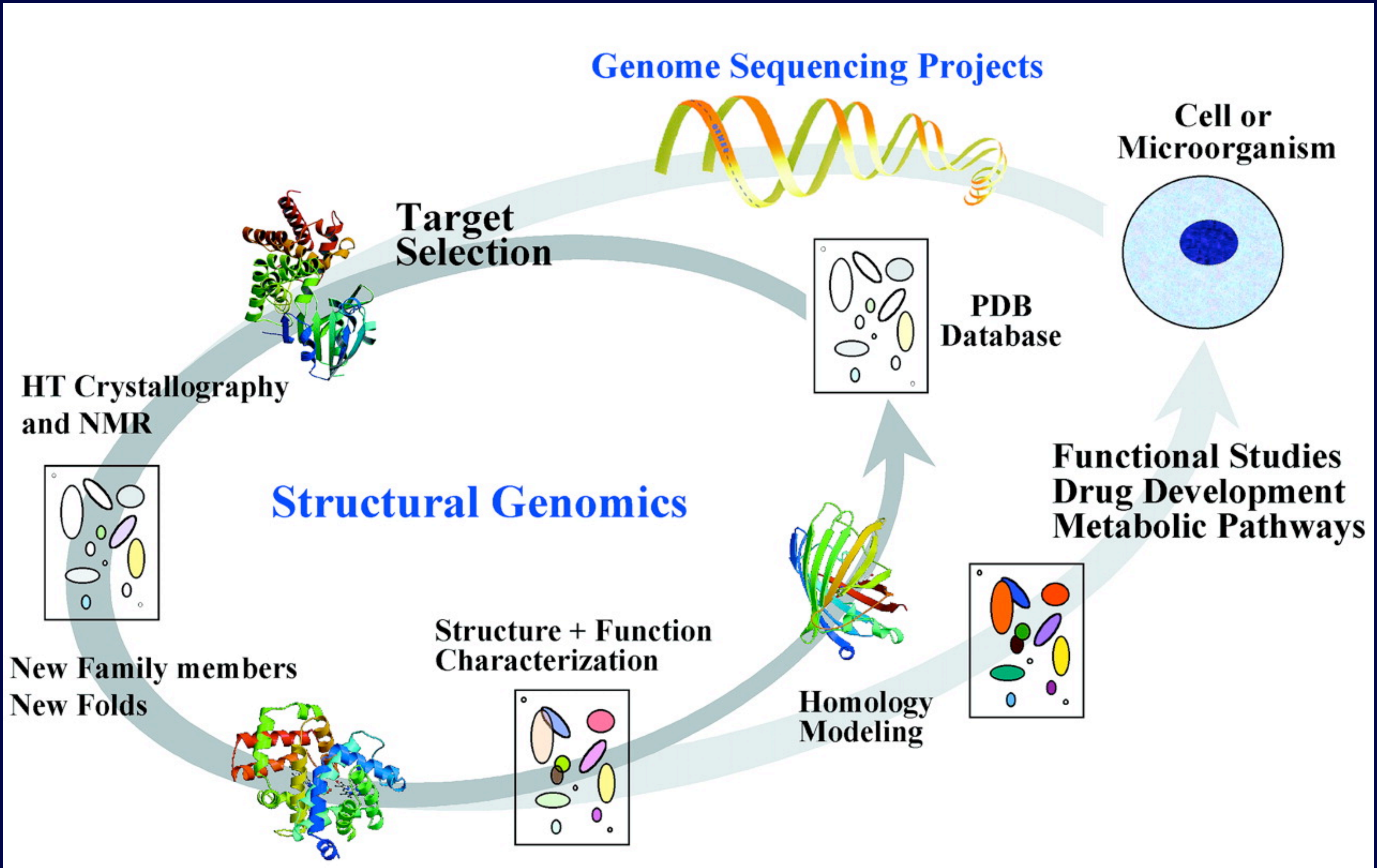
© Michael Feig, 2006.

Loop Sampling Methods

# Residues	Sampling	Program
1-2	All-Atom Reconstruction	MMTSB Tool Set
1-3	Exhaustive Search	
2-12	Torsional Space MC/MD	Modeller (Sali)
5-30	Multi-Scale	MMTSB Tool Set
2-100	Fragment-based	Rosetta (Baker)

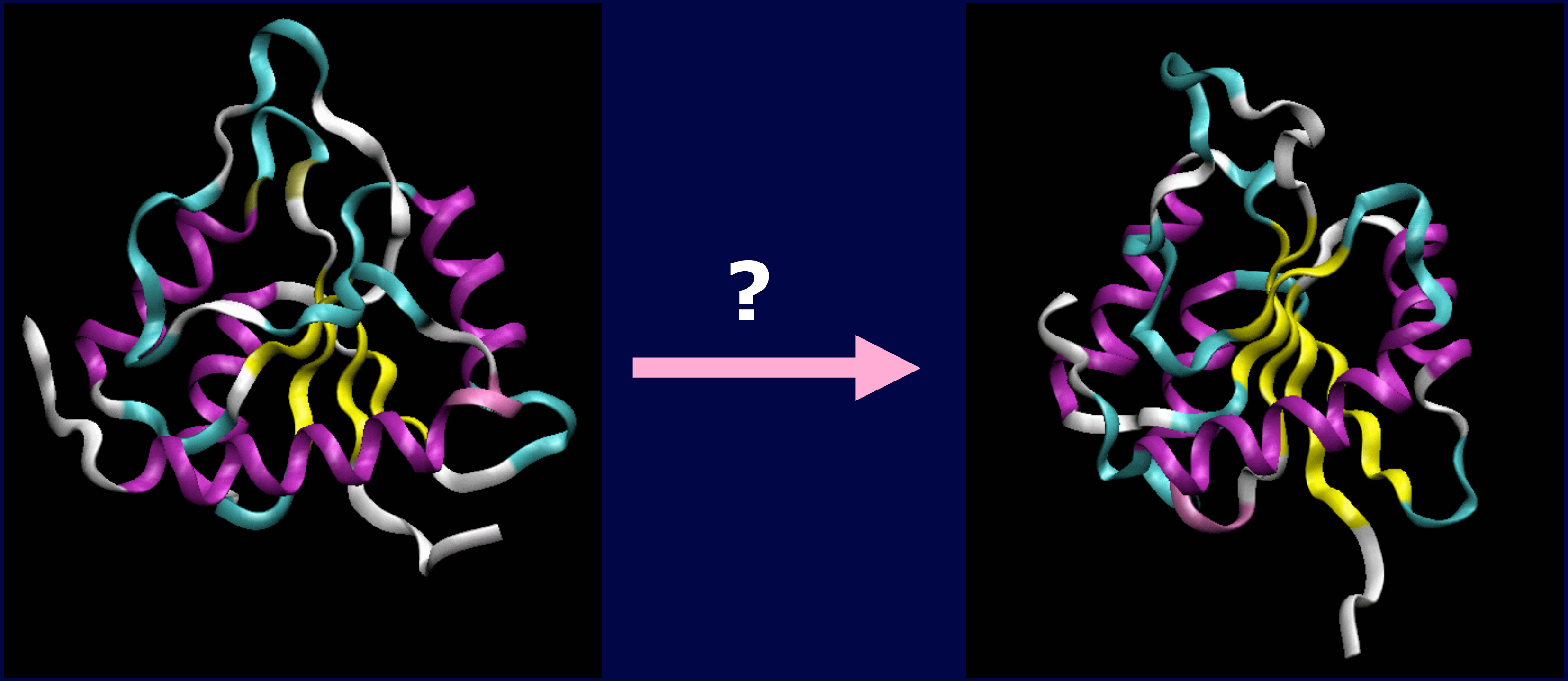
© Michael Feig, 2006.

Structural Genomics Efforts



© Michael Feig, 2006.

Structure Refinement



predicted

native (NMR)

© Michael Feig, 2006.