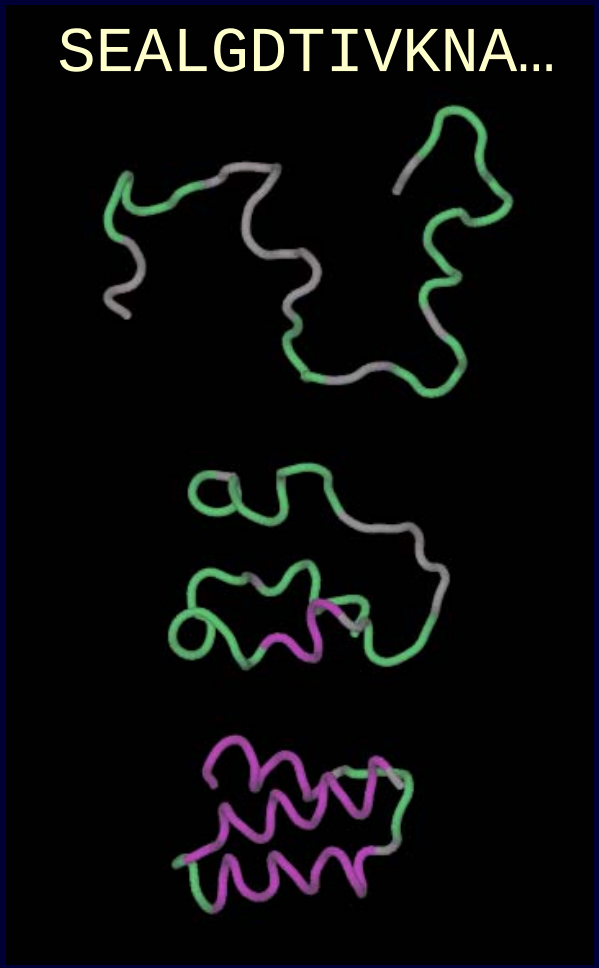
Protein Structure Prediction

Michael Feig
MMTSB/CTBP
2009 Summer Workshop

© Michael Feig

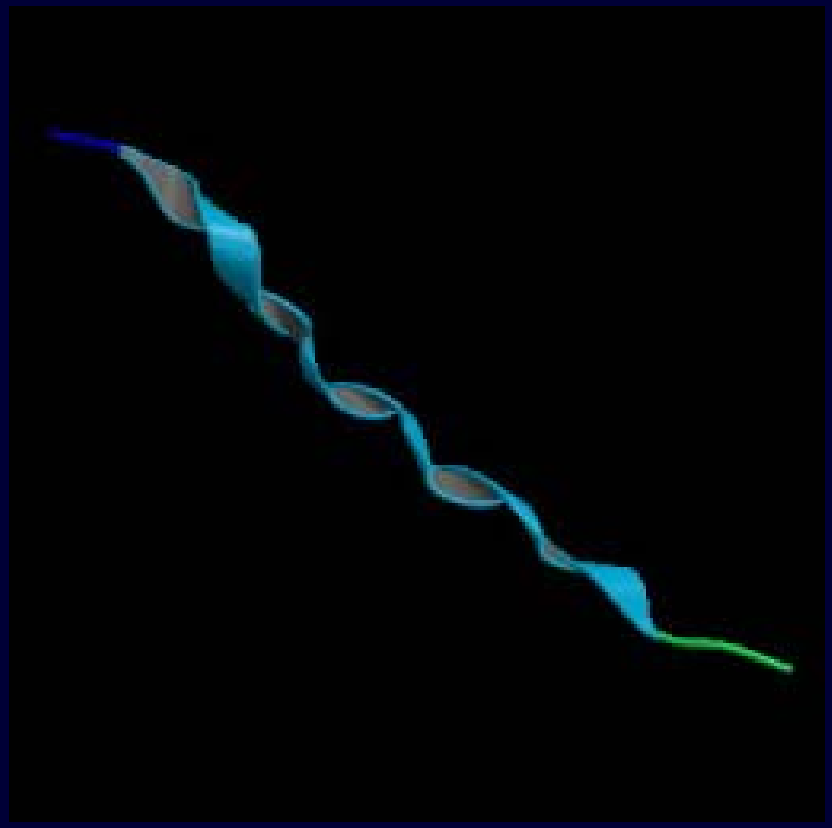
From Sequence to Structure



© Michael Feig

Folding with All-Atom Models

AAQAAAAQAAAAQAA

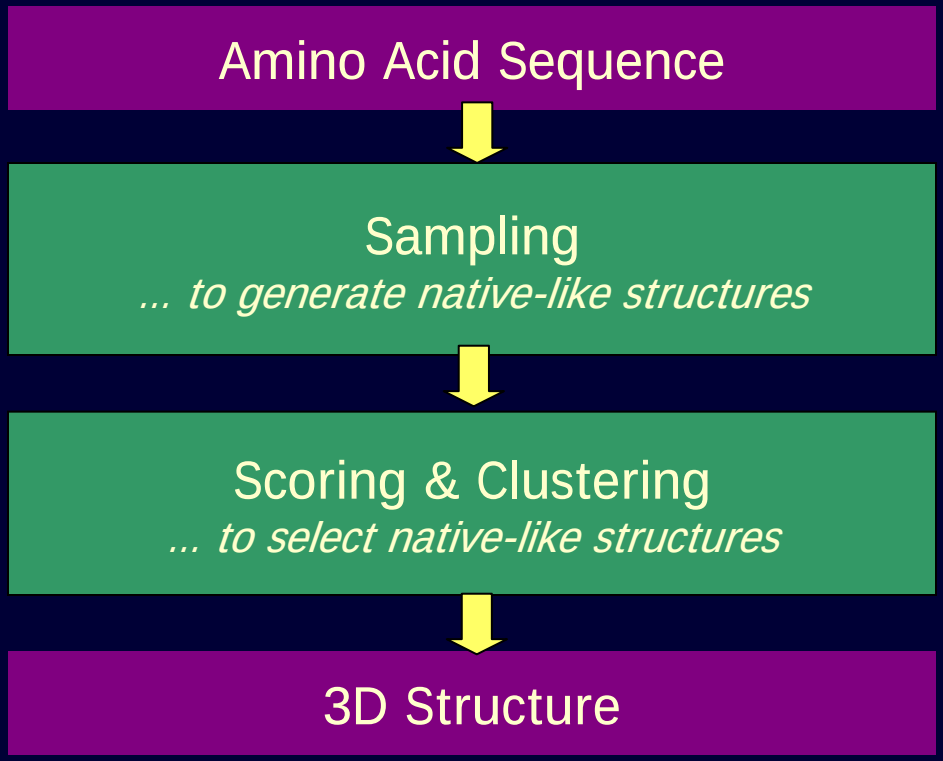


CHARMM force field
Implicit solvent replica exchange simulations
8 replicas, 10 ns/replica

All-atom MD in general not succesful for real proteins because of insufficient sampling but also force field issues.

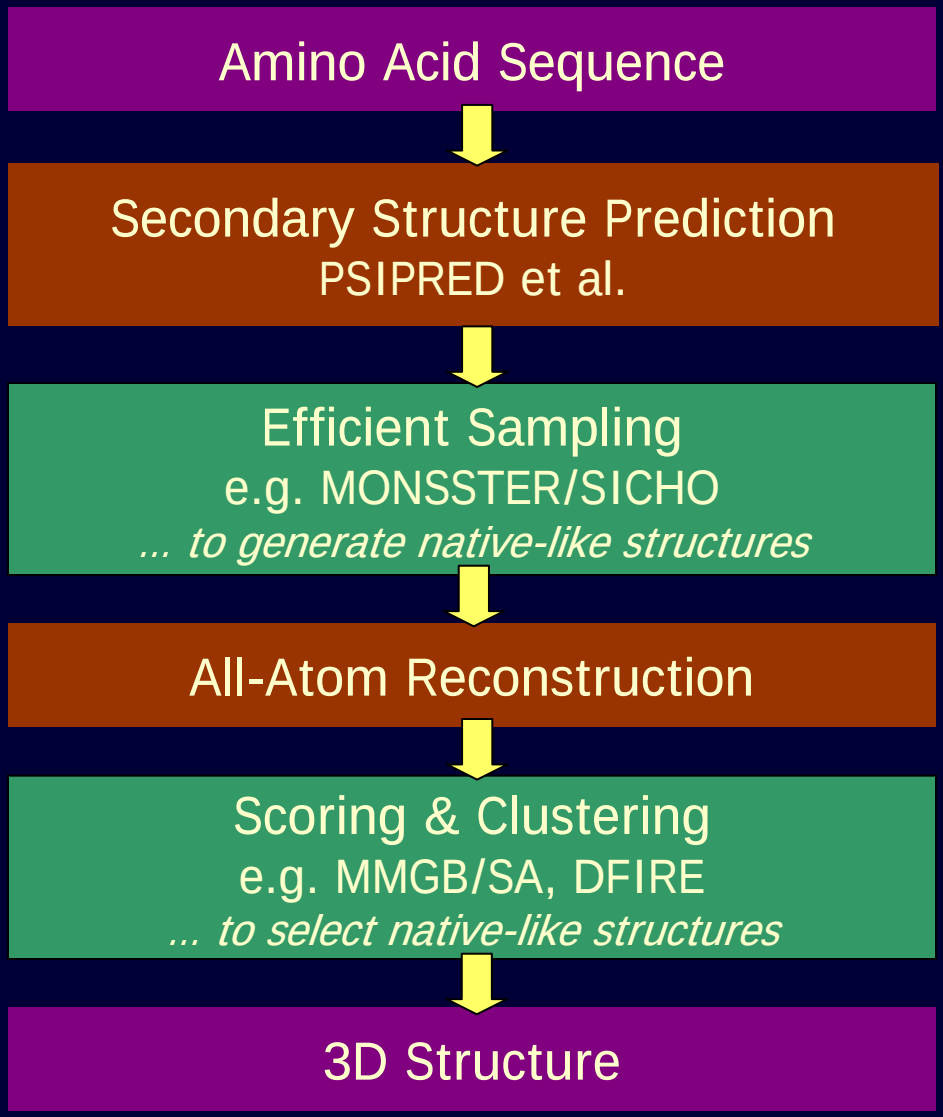
© Michael Feig

Ab initio Structure Prediction Protocol



© Michael Feig

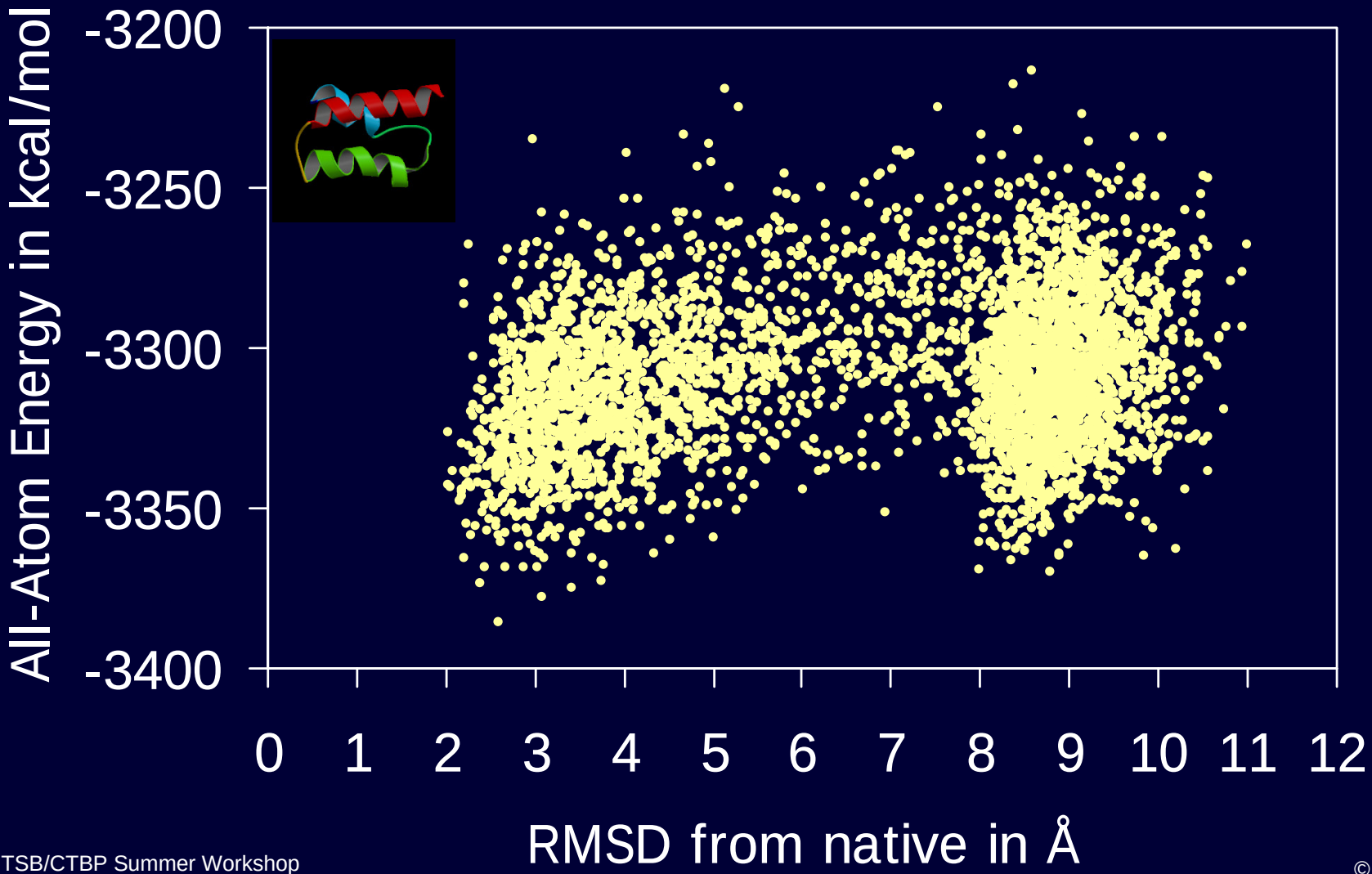
Improved Ab initio Prediction Protocol



© Michael Feig

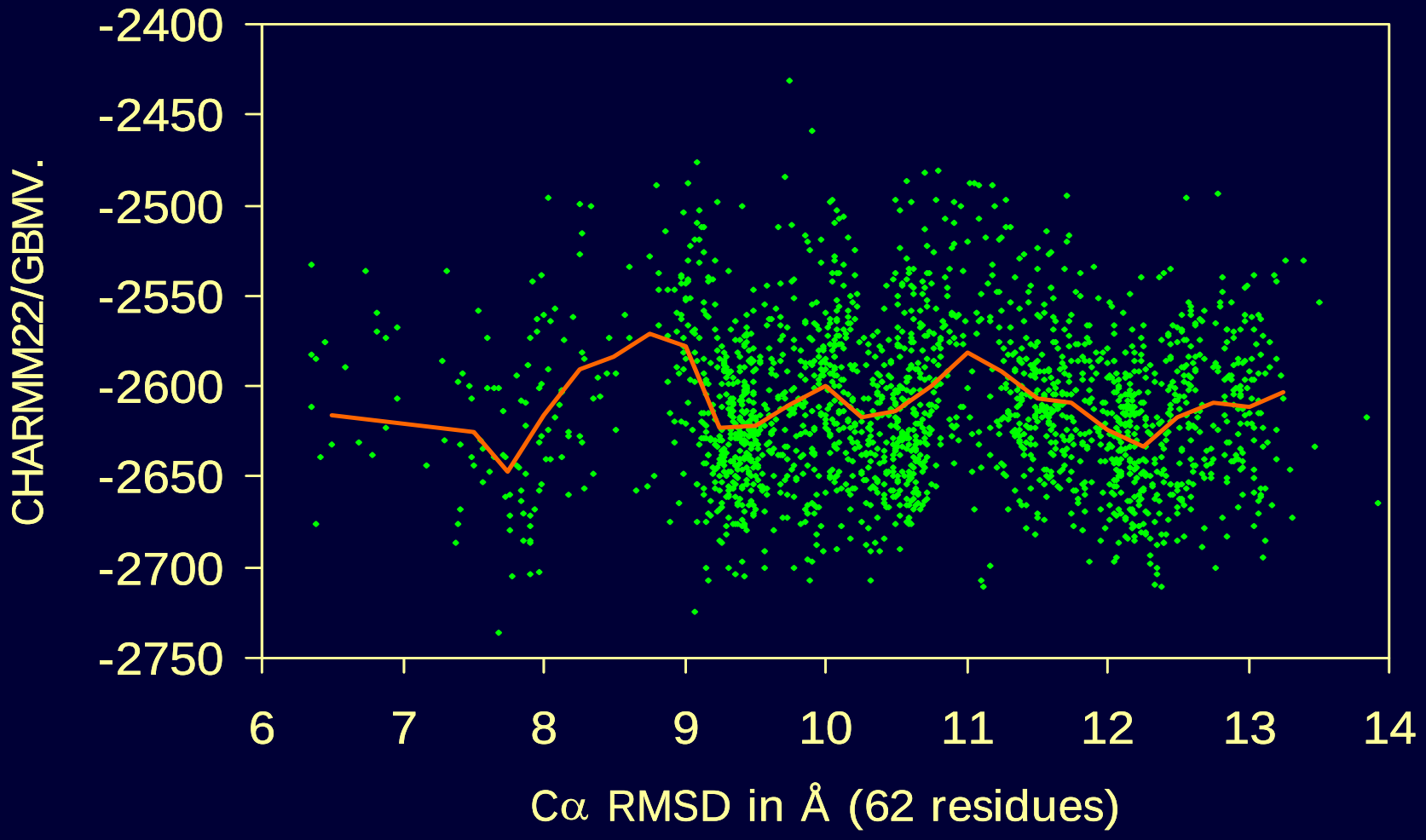
Conformational Sampling vs. Scoring

SICHO sampling of protein A



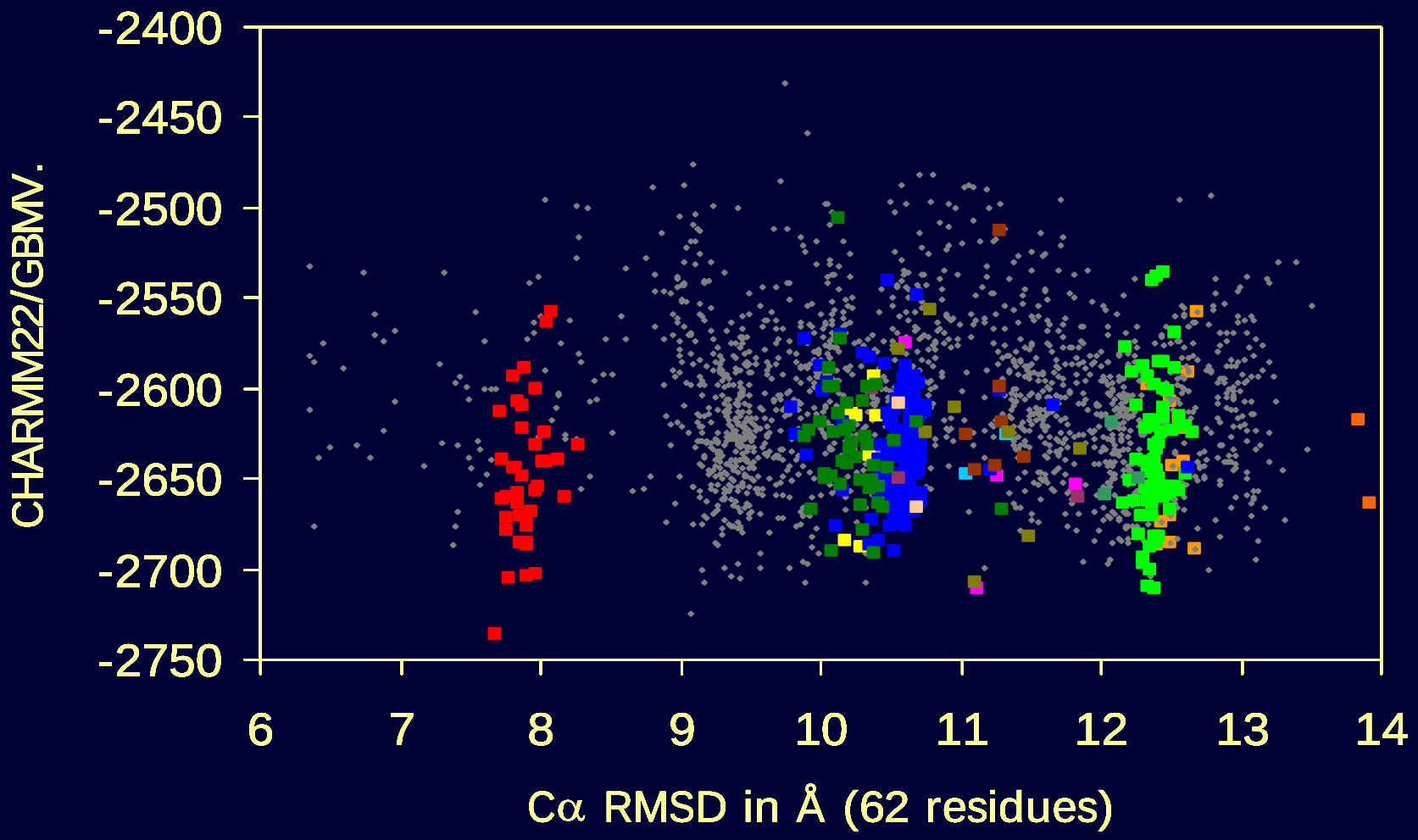
© Michael Feig

Realistic Example from CASP5



© Michael Feig

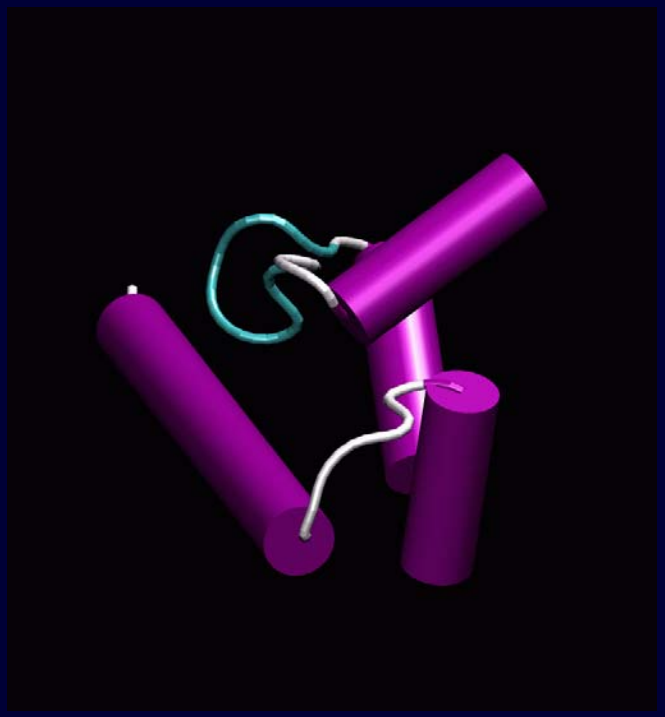
Sampling, Scoring, Clustering



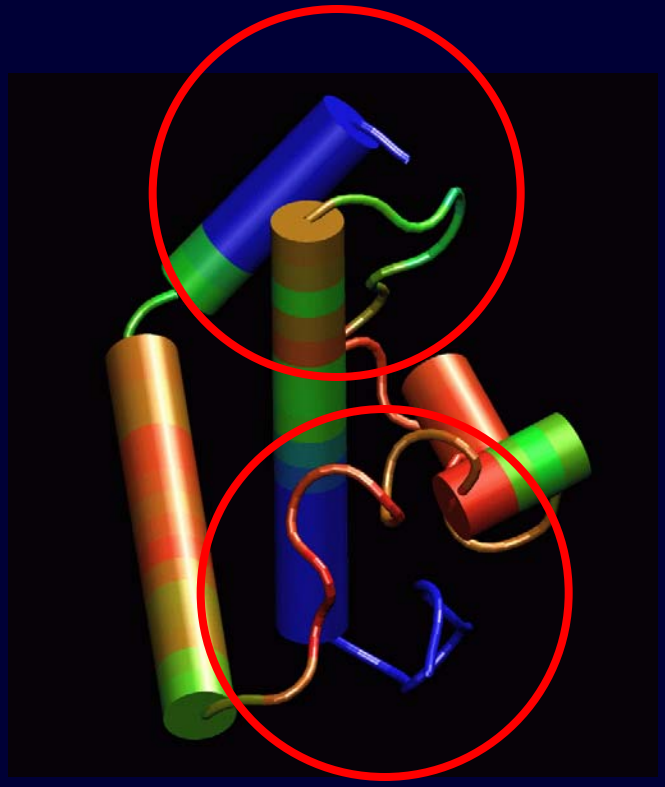
© Michael Feig

Ab initio Predictions

DNase fragmentation factor



NMR structure 1KOY



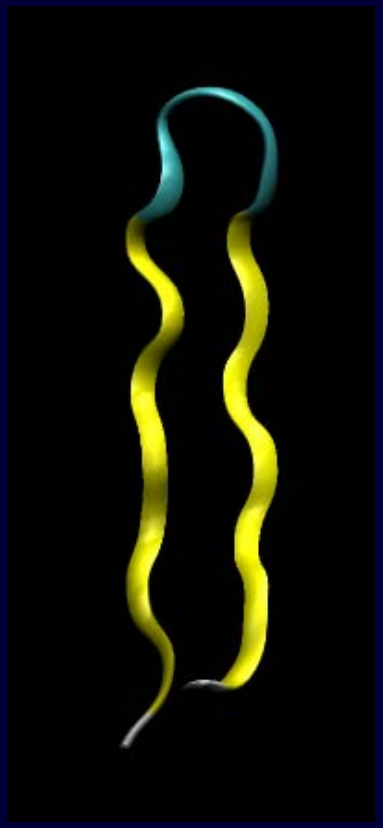
Best-scoring prediction, 7.4 Å

Problems with **secondary structure prediction** and **sampling**.
Scoring luckily worked here.

© Michael Feig

Secondary Structure Prediction

...GDPIVKNAKLDSRLANKEALRLL...



?

© Michael Feig

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

Secondary Structure Prediction Methods

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

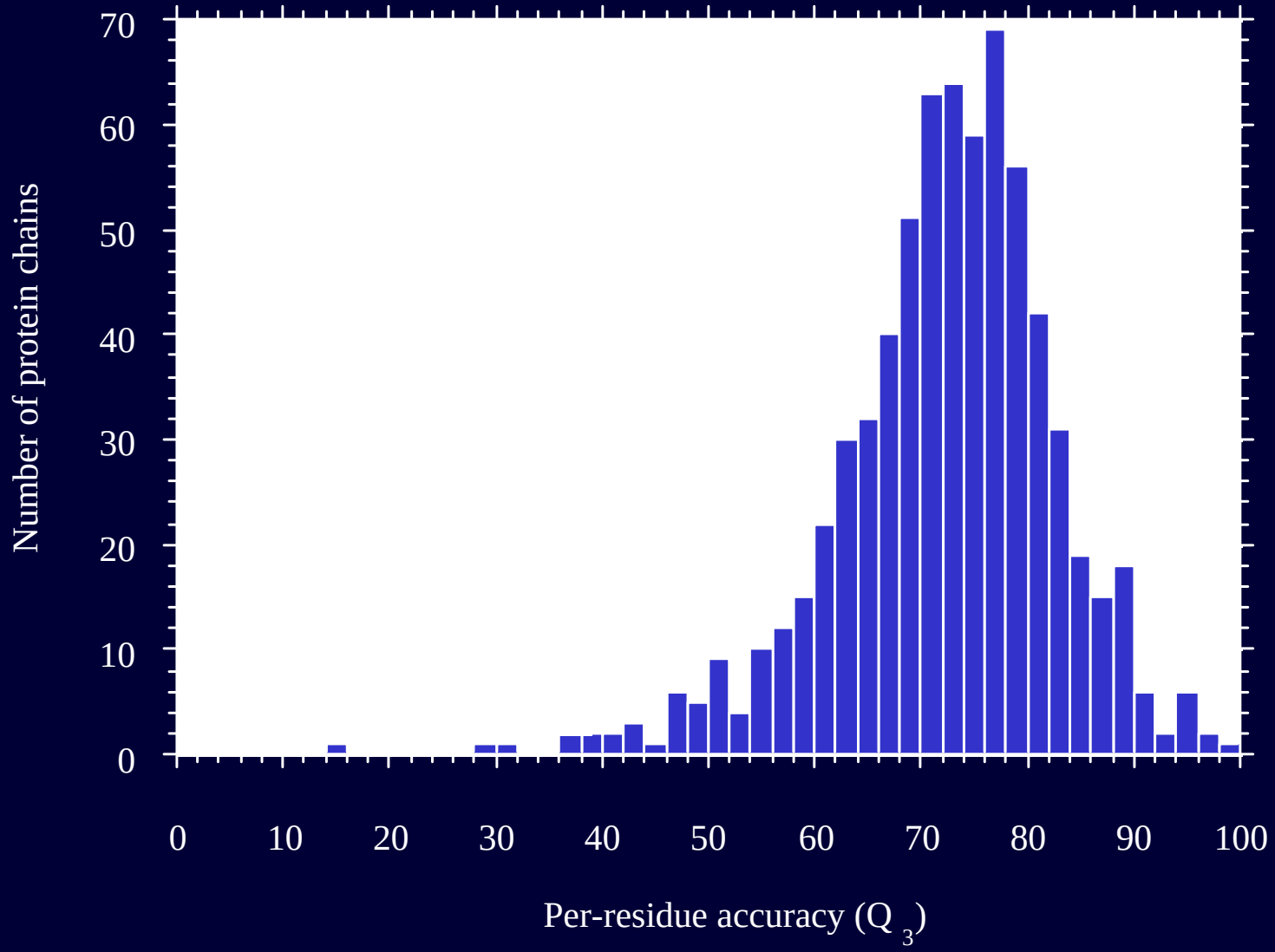
C+F: Chou & Fasman
GOR: Garnier, Osguthorpe, Robson

Improved accuracy is obtained via known structural information from homologous proteins

© Michael Feig

Secondary structure prediction

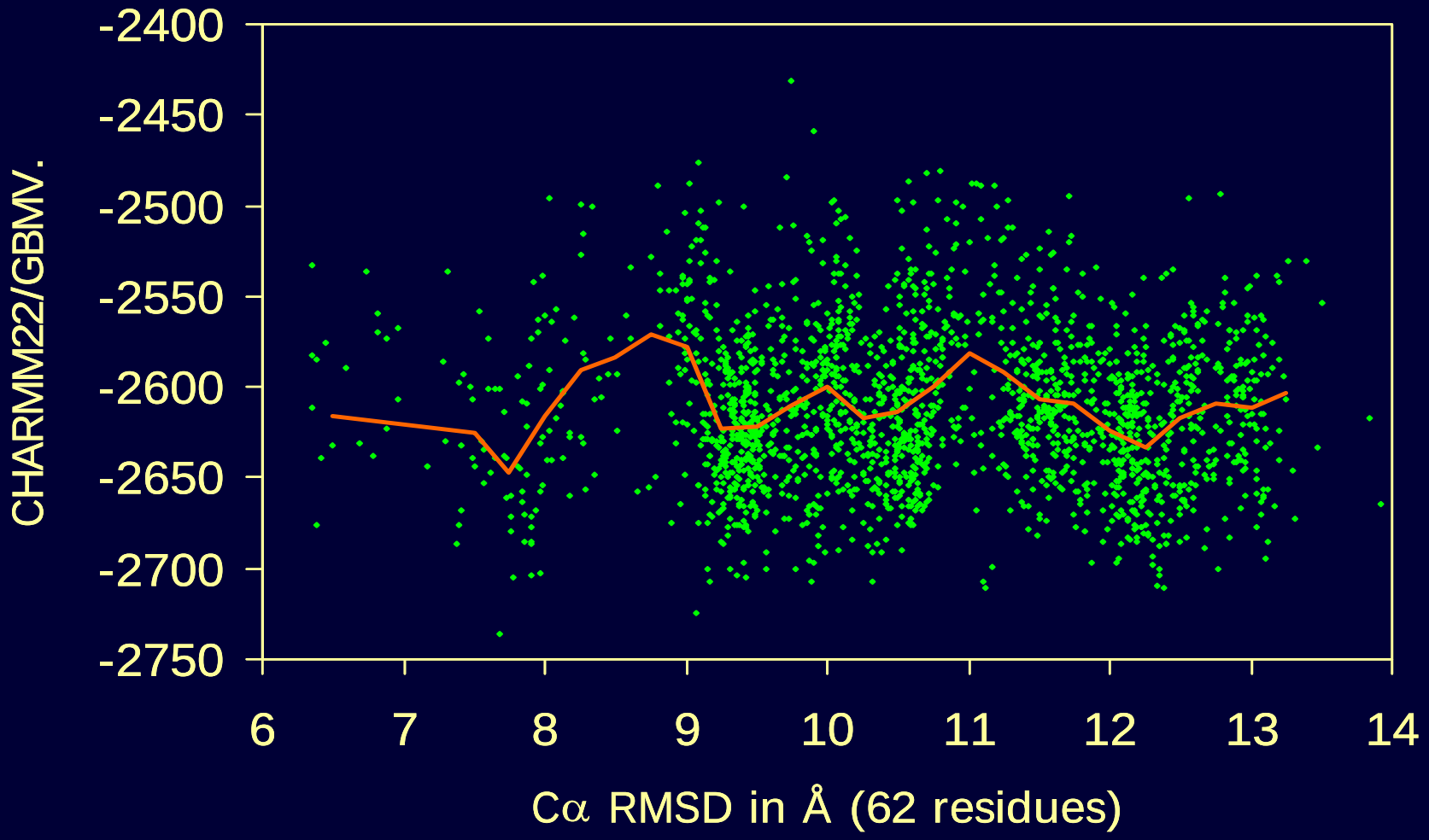
Per-Residue Accuracy



© Michael Feig

Sampling

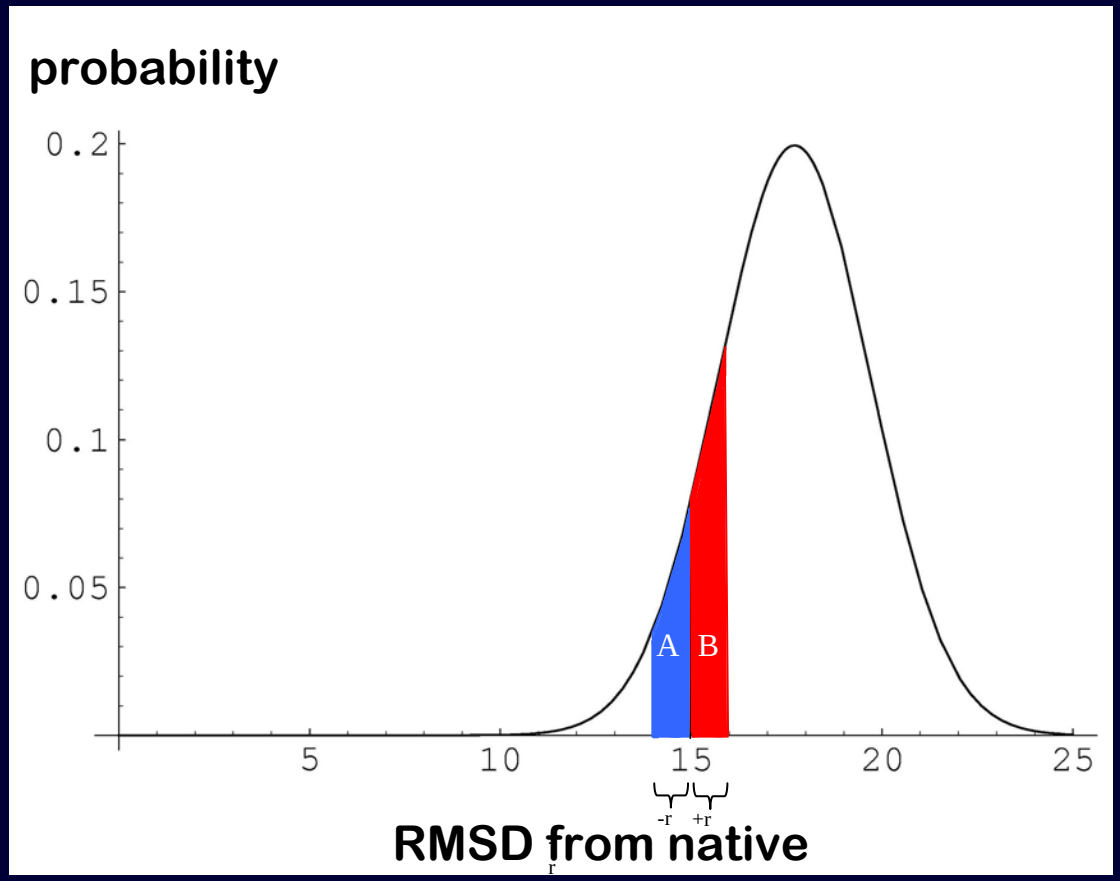
Could we have reached the native state?



© Michael Feig

Random Sampling

Random sampling of protein-like conformations:



$$P_{r \le R} = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{RMSD} e^{\frac{-(r-<RMSD>)^2}{2\sigma^2}} dr$$

$$< RMSD > \approx 3.333 N_{res}^{1/3}$$

For N=62:
<RMSD> = 13.2 Å

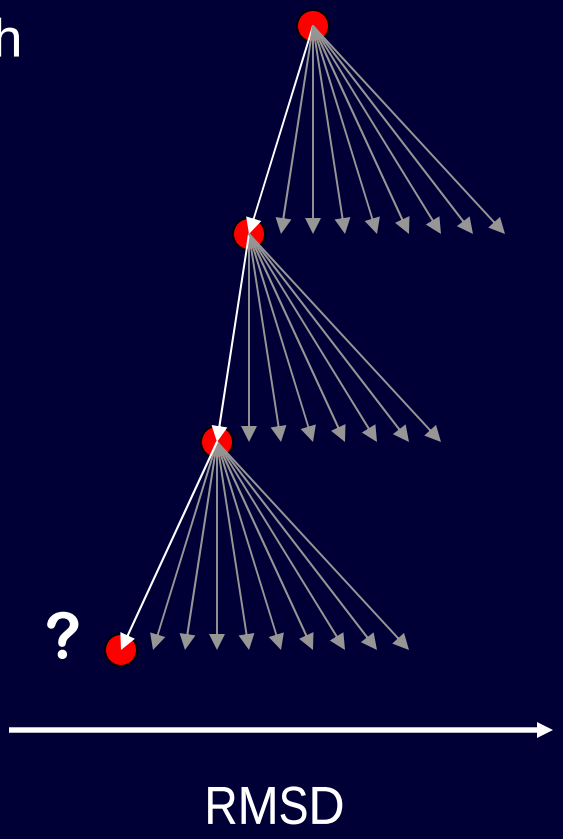
so where are doing better than random sampling!

Reva BA, Finkelstein AV, Skolnick J. Folding & Design (1998) 3:141-147

© Michael Feig

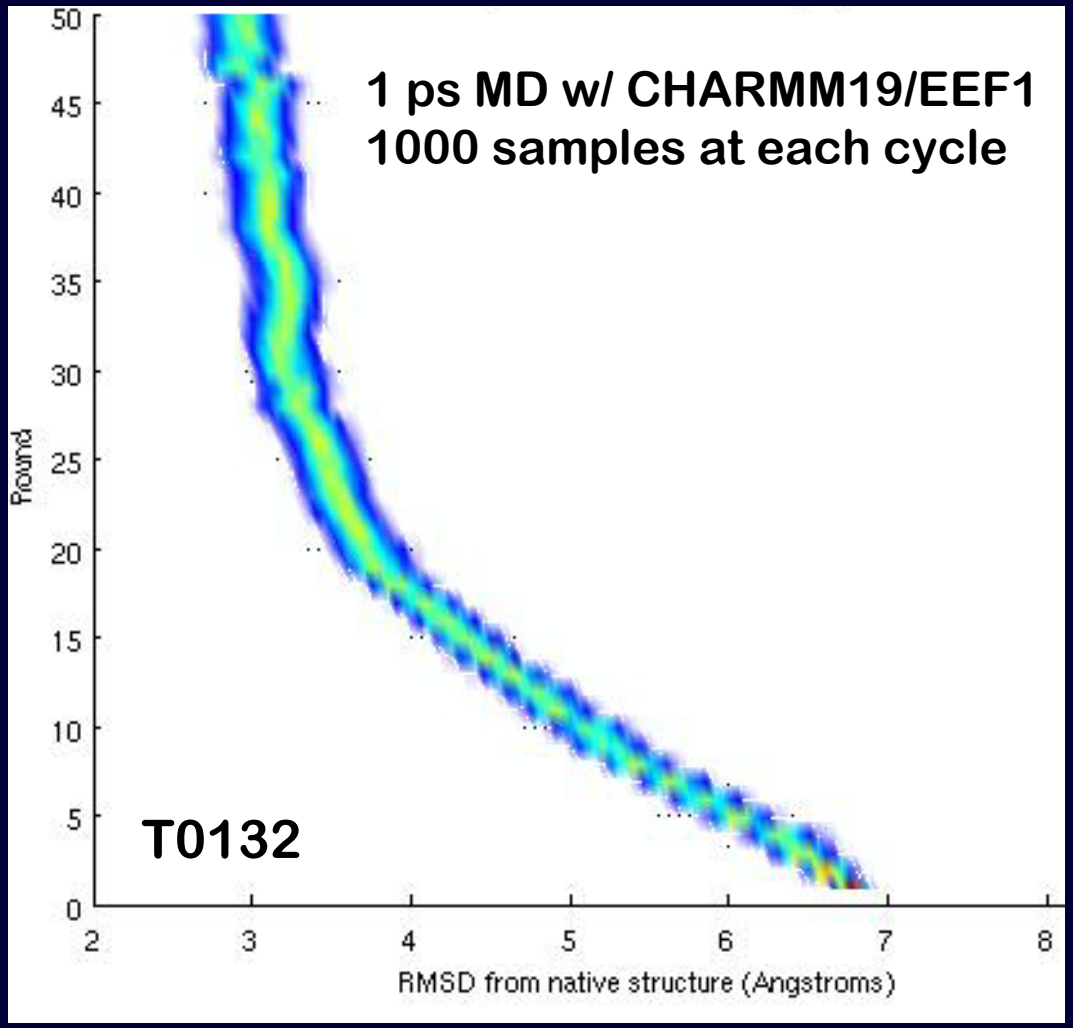
Can we Reach the Native State?

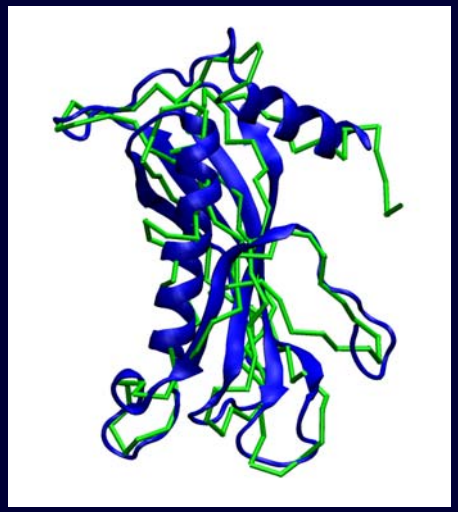
Iterative refinement with
'ideal' scoring function:
RMSD from native



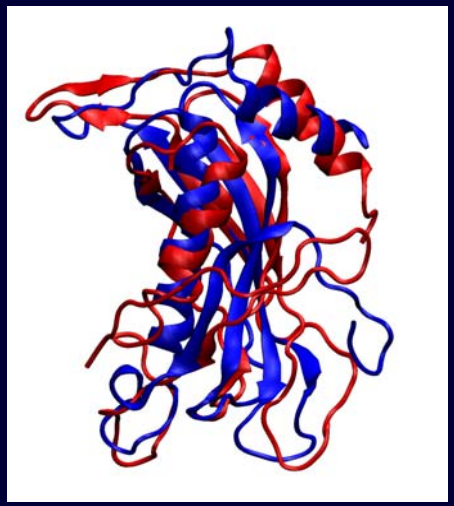
© Michael Feig

Ideal Sampling with MD





@50 cycles
2.6 Å

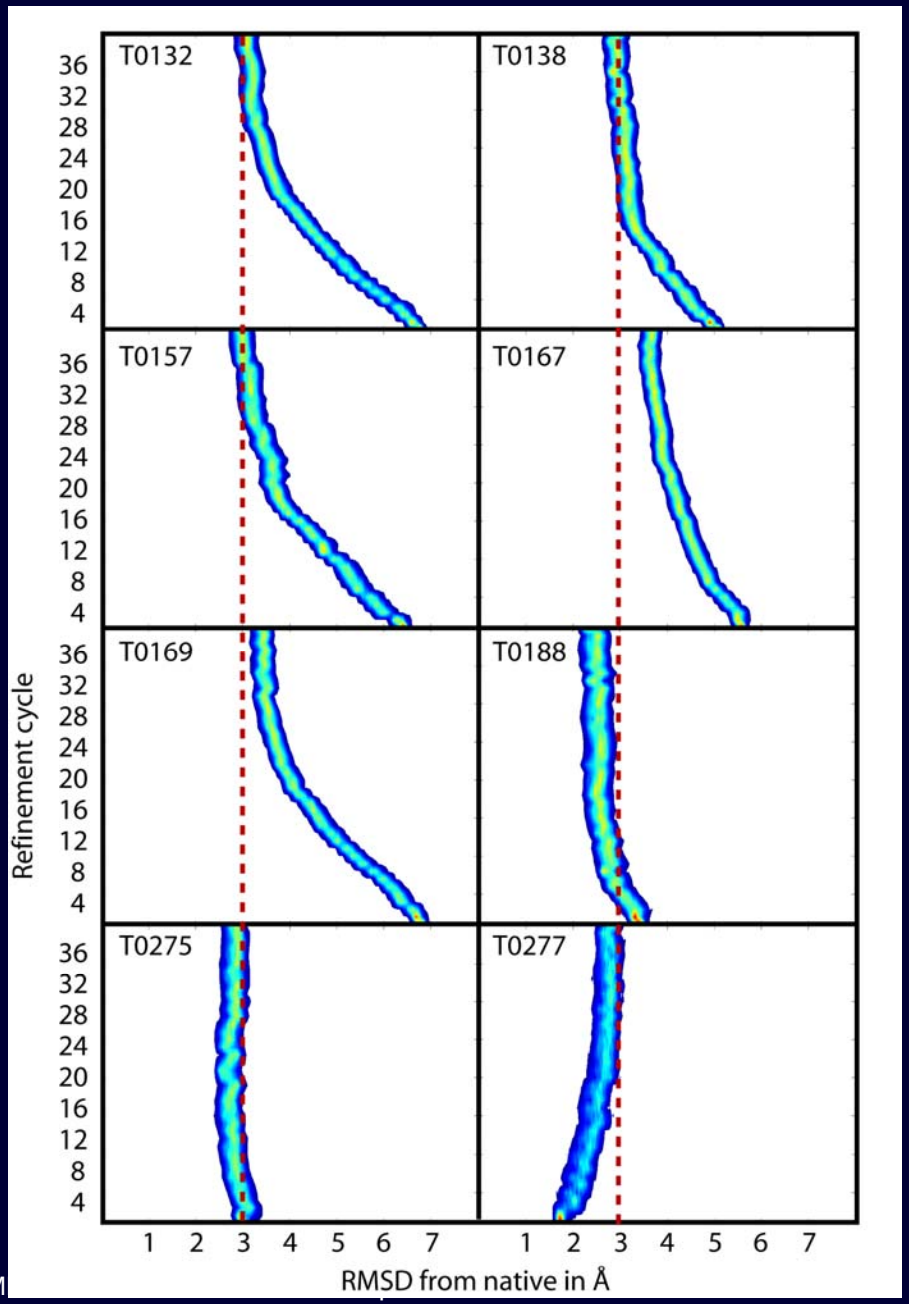


initial:
6.7 Å

Stumpff-Kane, Maksimiak, Lee, Feig: Proteins (2008) 70, 1345-1356

© Michael Feig

Can we reach the native?



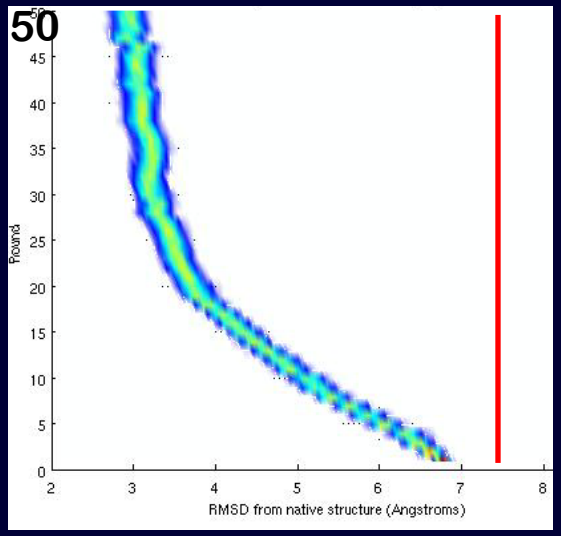
Resolution limit of about 2.5-3 Å

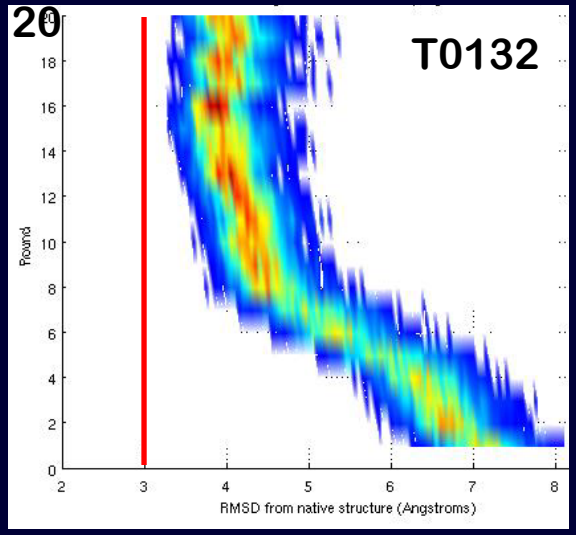
Sampling?
Force field?

Stumpff-Kane, Maksimiak, Lee, Feig:
Proteins (2008) 70, 1345-1356

© Michael Feig

How about coarse-grained sampling?

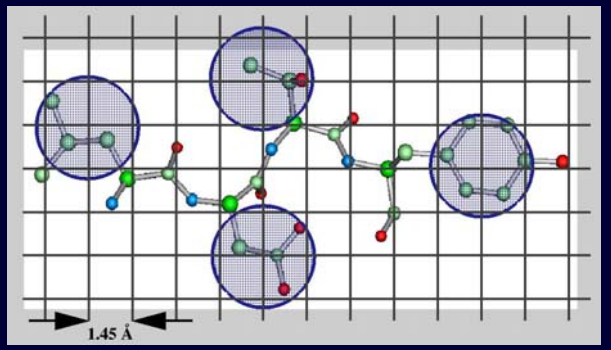




MD
all-atom/impl. solvent
1 ps MD @ 300K
1000 samples/cycle

SICHO
coarse-grained/lattice
MC sampling
1000 samples/cycle

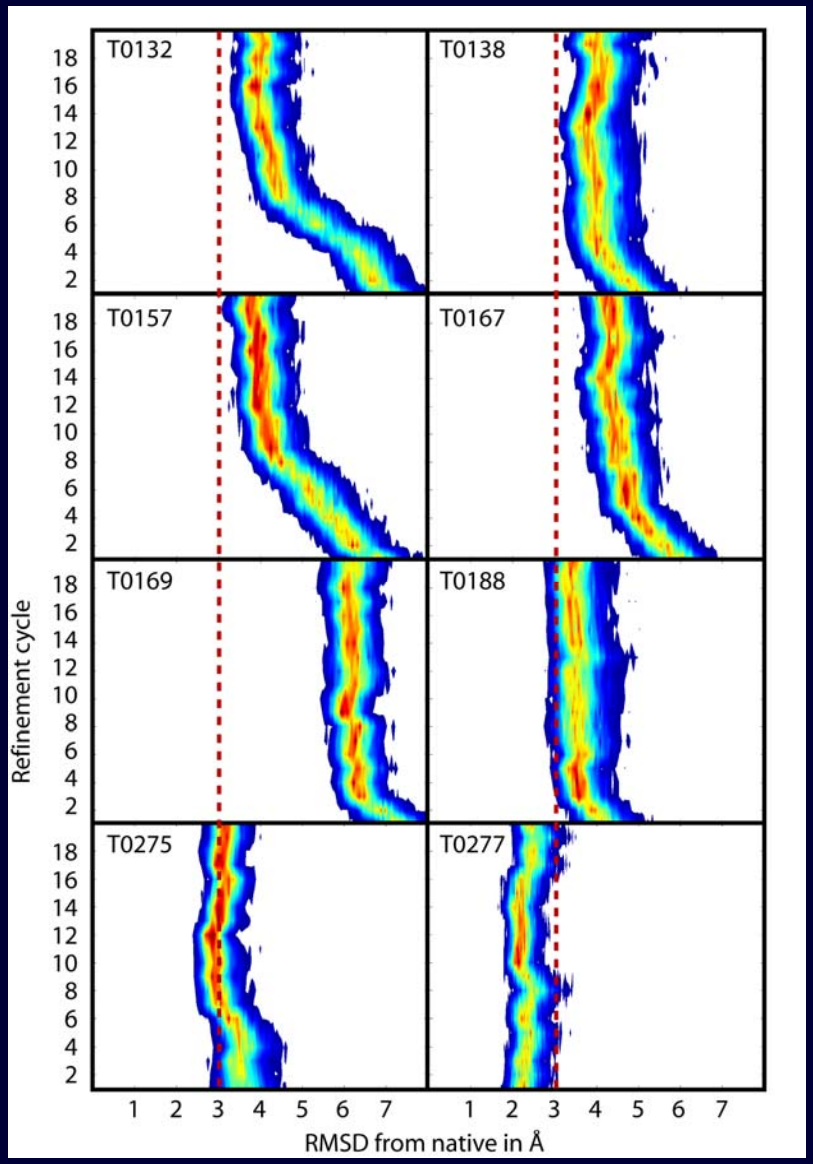
$$-\nabla V = F = ma$$



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

© Michael Feig

Can we reach the native?



Resolution limit of about >3 Å !

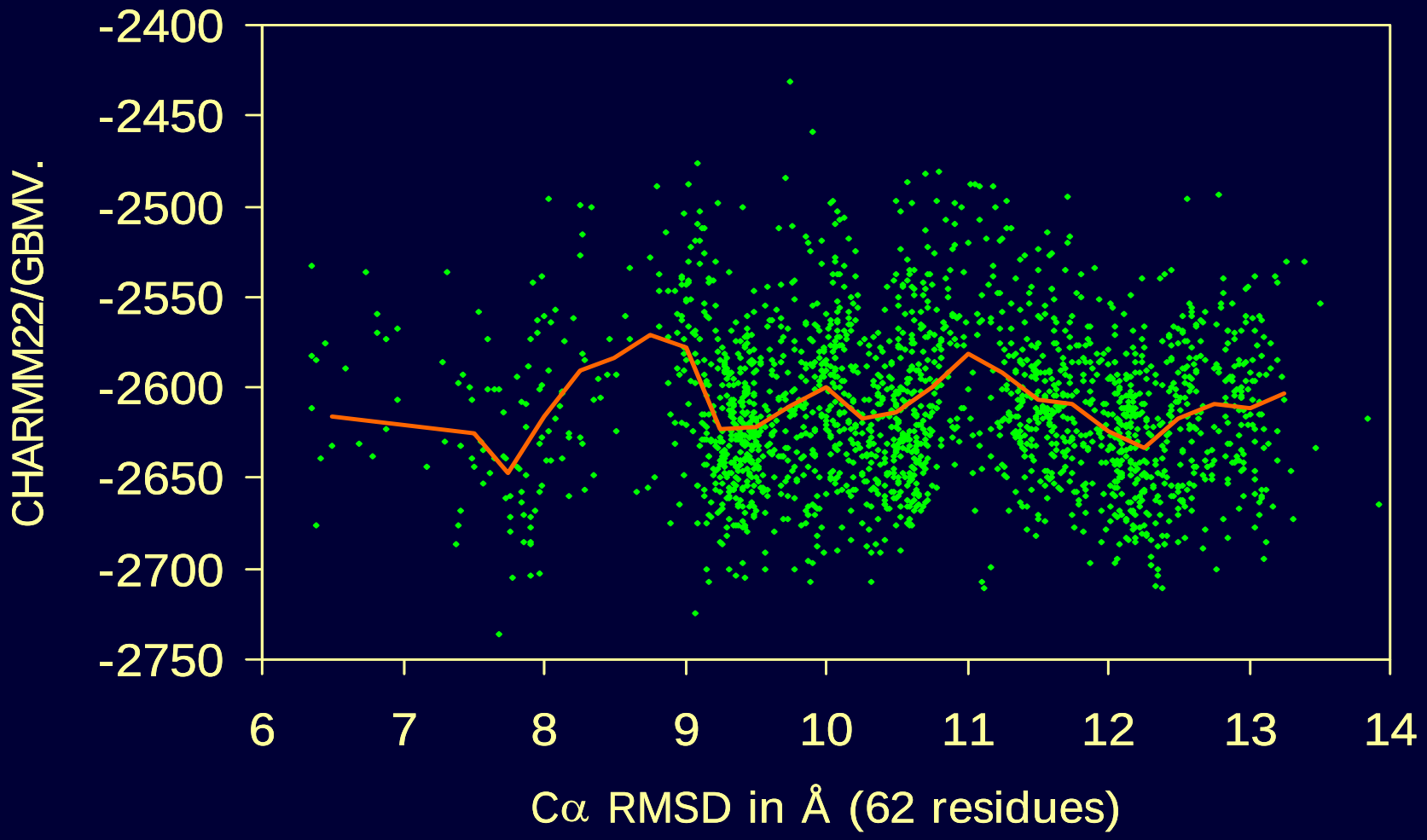
SICHO

Stumpff-Kane, Maksimiak, Lee, Feig: Proteins (2008) 70, 1345-1356

© Michael Feig

Scoring

Can we pick out the native state?



© Michael Feig

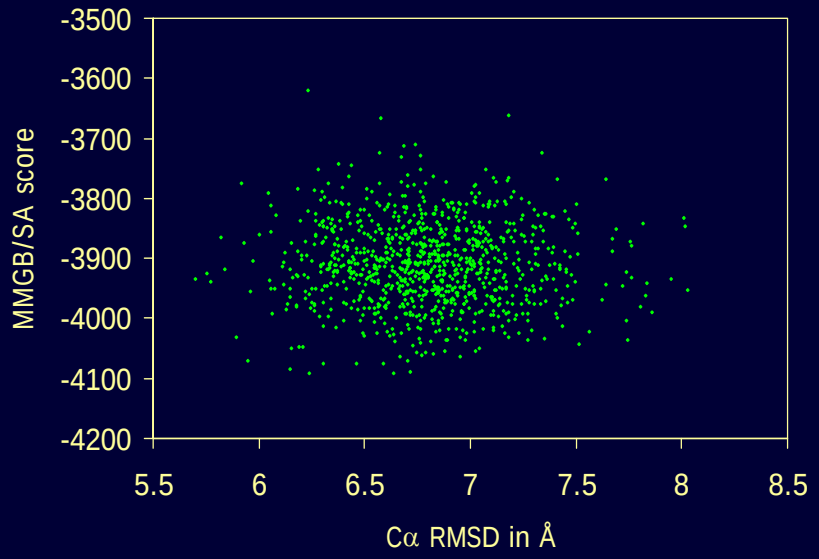
Scoring Functions

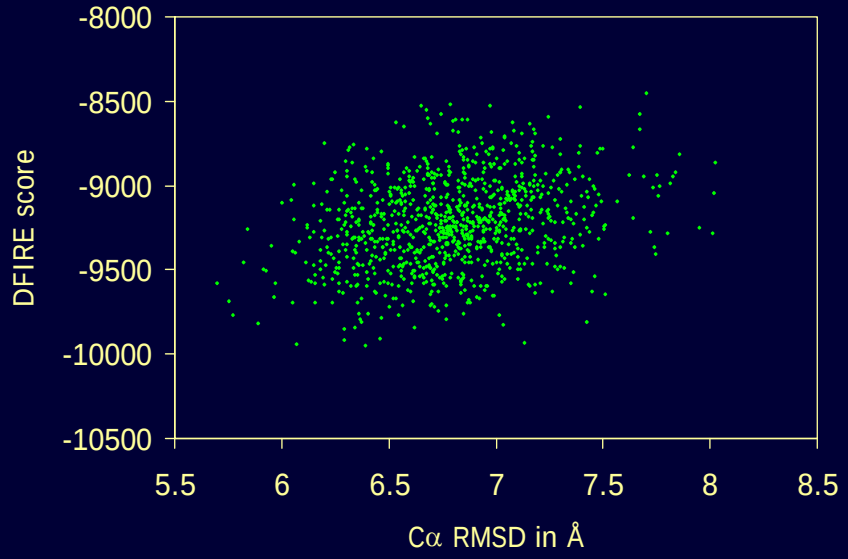
- **Knowledge-based/statistical**
  - derived from known protein structures
  - limited by training data
  - usually fast
  - e.g. DOPE, DFIRE, OPUS-PSP, RAPDF, prosaII
- **Force field based**
  - model physical energy landscape
  - more robust and transferable
  - often expensive (require minimization)
  - e.g. MMPB(GB)/SA, UNRES

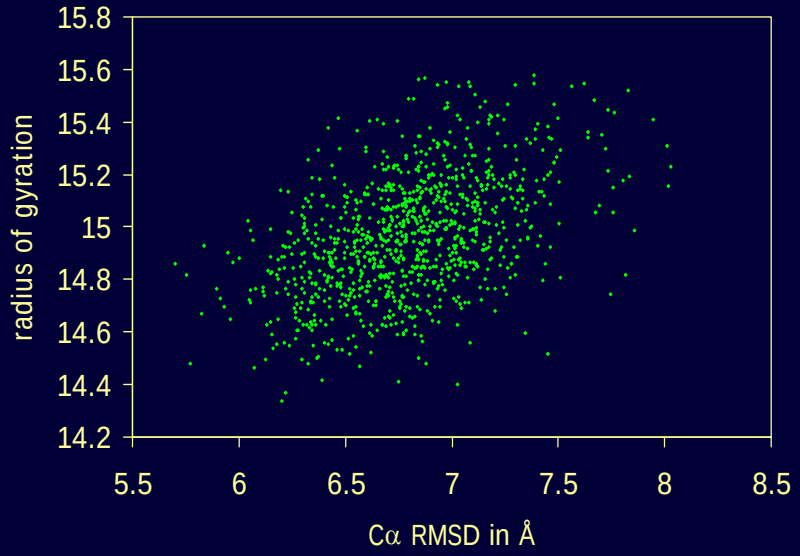
© Michael Feig

Scoring Function Comparison

MMGB/SA vs. DFIRE







© Michael Feig

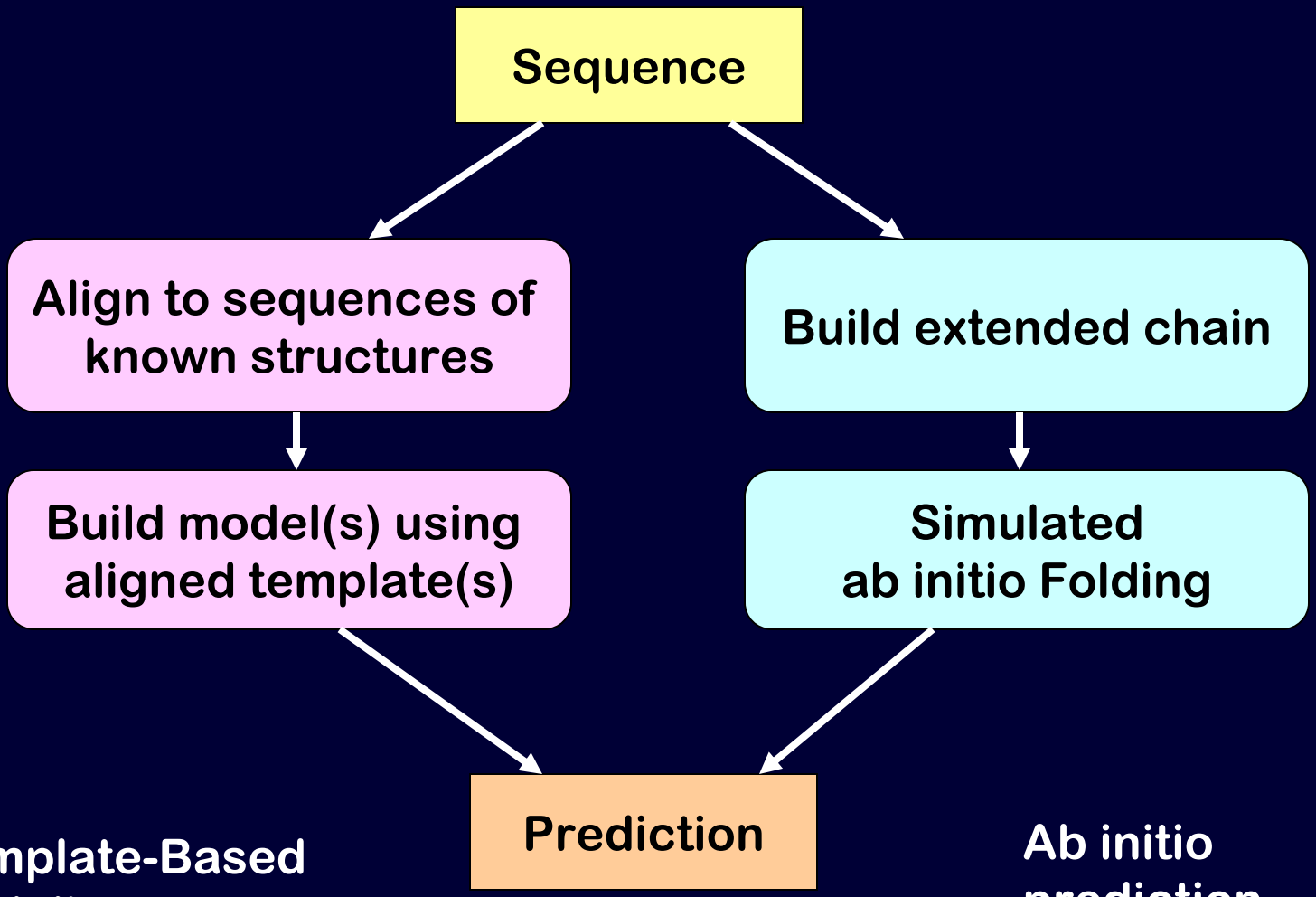
Reliable and accurate ab initio structure prediction may eventually succeed …

© Michael Feig

… but the solution may lie elsewhere.

© Michael Feig

Structure Prediction Approaches



© Michael Feig

Sequence Homology

Human thioredoxin (1AUC)

```
SDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGKLTVA
  : :.       ..    .:  ..::: : :::::::: :.. ......:. . .
MVKQIESKTAFQEALDAAGDKLVVVDFSATWCGPCKMIKPFFHSLSEKYSNVIFL-


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

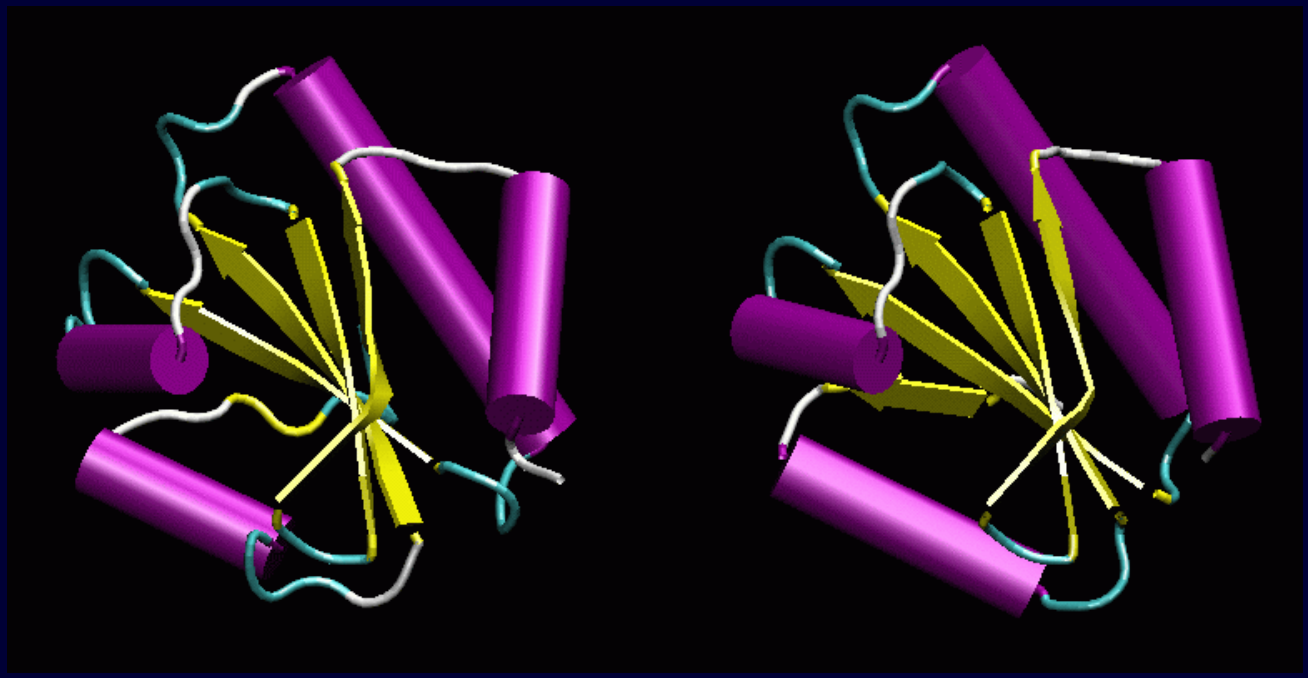
E. Coli thioredoxin (1THO)

© Michael Feig

Comparative Modeling

Assumption:

Proteins with similar sequence have similar structure

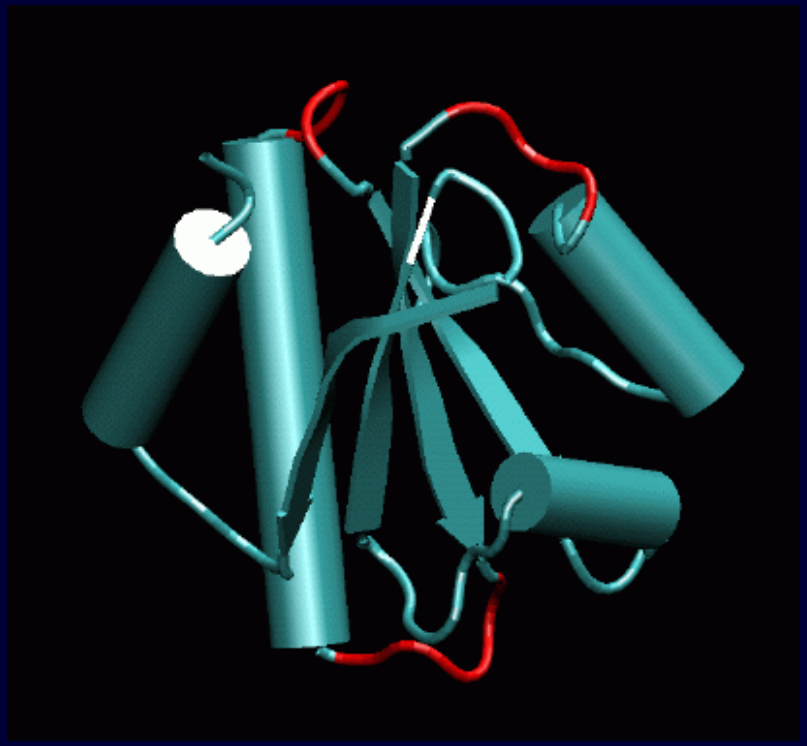


Human thioredoxin
(1AUC)

E. Coli thioredoxin
(1THO)

© Michael Feig

Structural Templates from Homology



Challenges:

- Template identification
- Correct alignment
- Loop modeling
- All-atom reconstruction (side chain rebuilding)

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


© Michael Feig

Template-Based Model Building

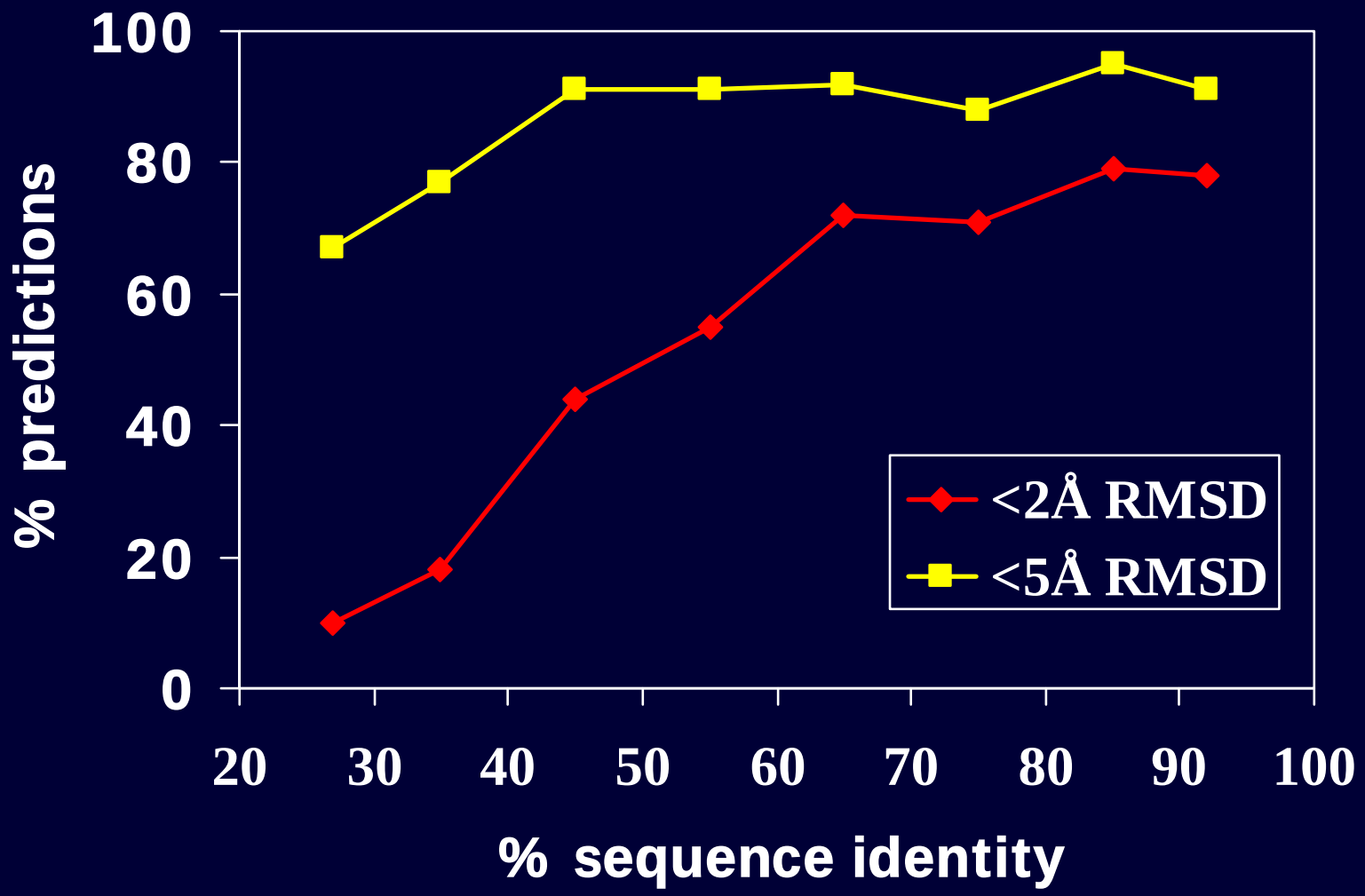
- Direct template-based modeling:
  - Use Cα-trace from template to build backbone
  - Add sidechains to backbone (using rotamer libraries)

- Restraint-based modeling:
  - Use one or more templates to setup list of restraints
  - Sample Cα-trace, CG model (SICHO), or all-atom model (Modeller) from extended chain with force field under template-derived restraints
  - Reconstruct all-atom model

Restraint-based modeling allows use of multiple templates and can reflect structural variations for different sequences.

© Michael Feig

Accuracy of Predictions by Homology



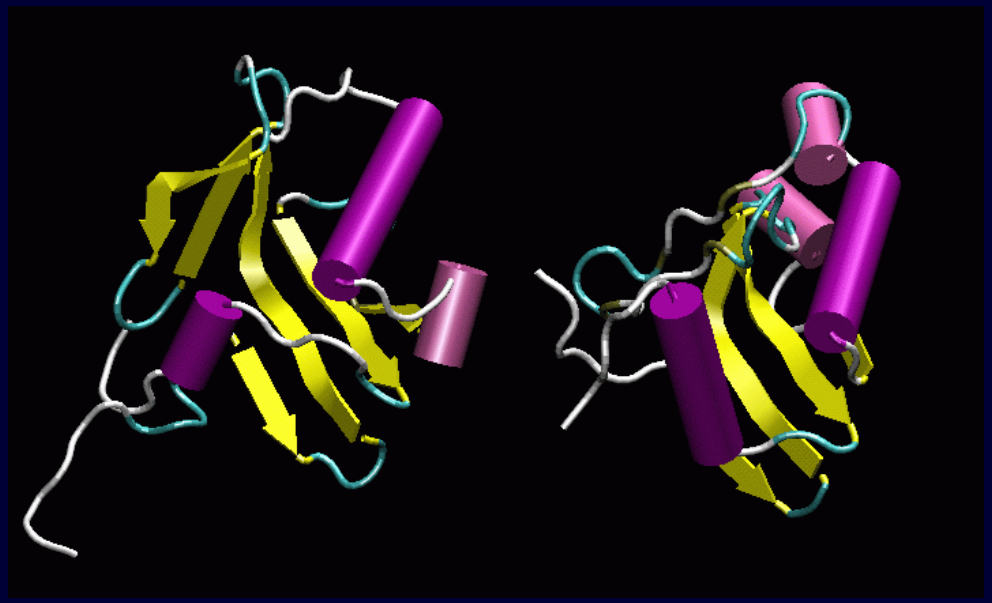
© Michael Feig

Prediction through Fold Recognition

Assumption:

Proteins with similar secondary structure share fold

1N91



1JRM

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


© Michael Feig

Templates through Fold Recognition

Challenges:

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

© Michael Feig

Extreme Homology Modeling: TASSER

- Find fragments of varying sizes (typically 20-100 residues) based on sequence alignment or fold recognition
- “Jam” fragments together in coarse-grained sampling procedure (Monte Carlo enhanced with replica exchange)
- Score resulting structures based on clustering

Method very successful, especially for “hard cases” when no overall template is available.

J. Skolnick, Y. Zhang

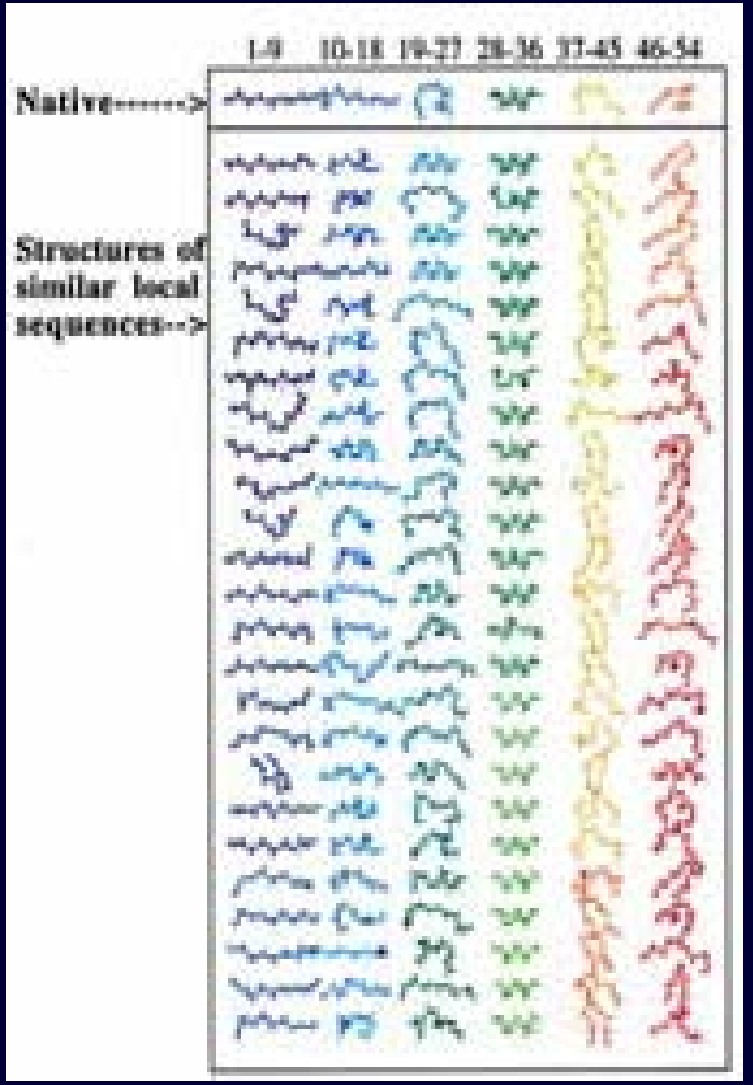
© Michael Feig

Extreme Homology Modeling: Rosetta

- Find short fragments (9 and 3 residues) based on sequence alignment or fold recognition
- Sample fragment assembly at Cα–level with Monte Carlo procedure
- Generate very large number of structures (10,000-1,000,000) and score with series of filtering functions

Method also very successful for “hard” cases.

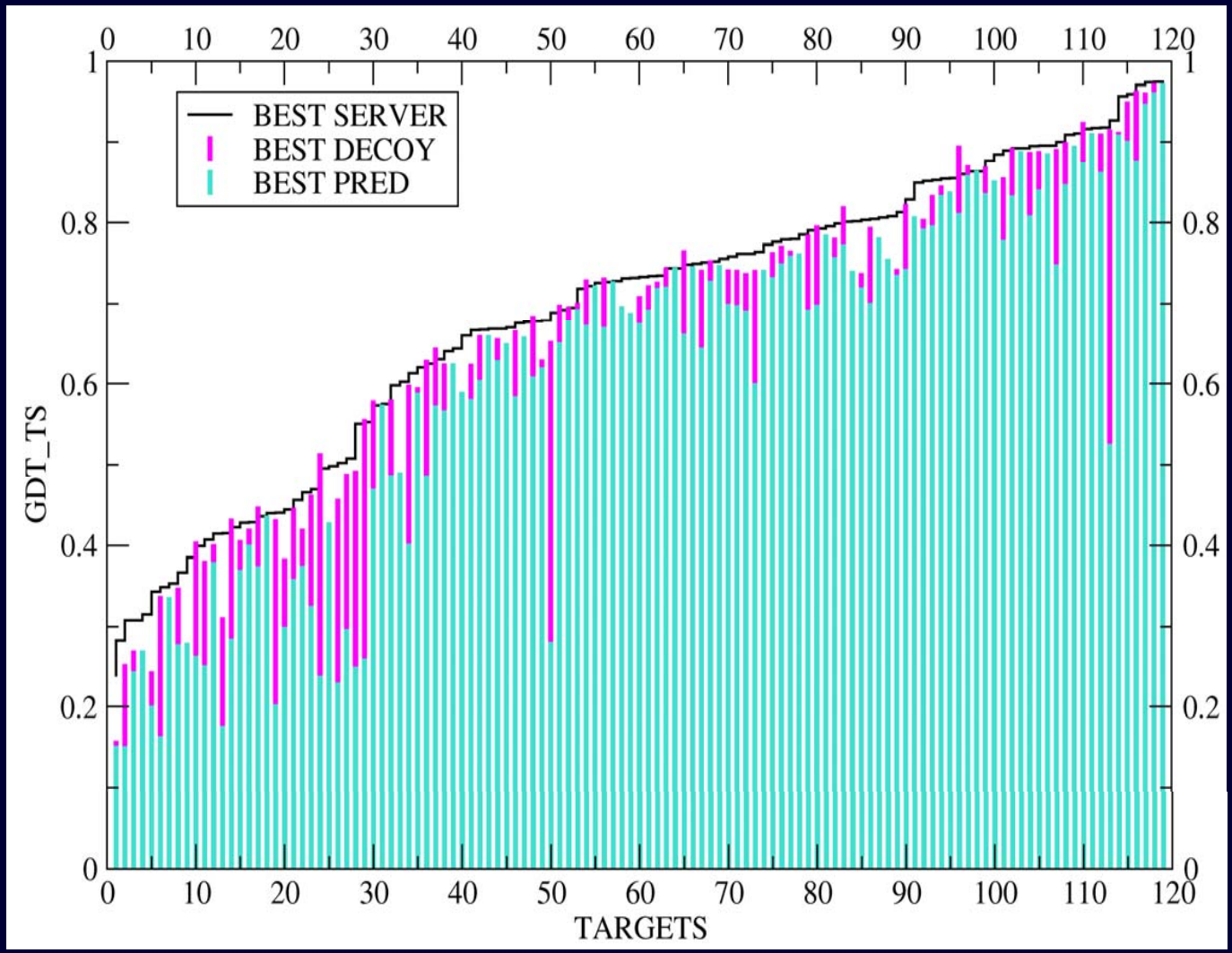
D. Baker



© Michael Feig

State of the Art in Structure Prediction

CASP8 Results



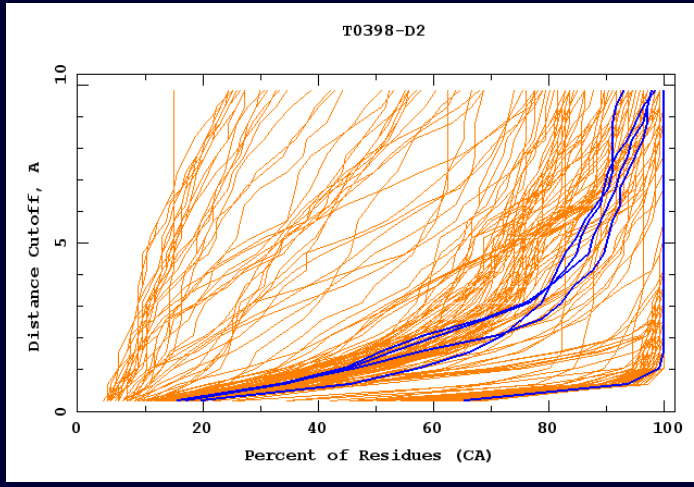
© Michael Feig

CASP8: Human/Server TBM Targets

1	DBAKER	489	1.39468	62	Rosetta+expert/Baker
2	Zhang	071	1.12254	63	TASSER+servers/Zhang
3	SAM-T08-human	046	1.11230	61	SAM+expert/Karplus
4	fams-ace2	434	1.10969	64	Meta/Umeyama
5	TASSER	057	1.08156	64	TASSER+expert?/Skolnick
6	MULTICOM	453	1.06672	64	HHSearch+MultiMod+Meta/Cheng
7	McGuffin	379	1.06355	62	Meta/McGuffin
8	ZicoFullSTPFullData	138	1.05079	63	Meta/Girgis
9	ZicoFullSTP	196	1.04905	63	Meta/Girgis
10	**Zhang-Server**	**426s**	**1.04062**	**64**	**TASSER/Zhang**
11	Zico	299	1.03919	62	Meta/Girgis
12	IBT_LT	283	1.03000	54	COMA+expert/Venclovas
13	Chicken_George	081	1.02079	63	SimFold+use servers/Chikenji
14	mufold	310	1.01067	60	Rosetta/Xu
15	**pro-sp3-TASSER**	**409s**	**0.96219**	**64**	**TASSER/Skolnick**
16	GeneSilico	371	0.90565	62	Meta/Bujnicki
17	A-TASSER	149	0.89703	64	TASSER/Skolnick
18	Sternberg	202	0.88222	54	HHSearch+MultiMod+expert?/Sternberg
19	Jones-UCL	387	0.86397	58	Jones
20	**BAKER-ROBETTA**	**425s**	**0.85422**	**64**	**Rosetta/Baker**
21	fais@hgc	198	0.84538	52	Shirota
22	**RAPTOR**	**438s**	**0.84459**	**61**	**Xu**
23	LevittGroup	442	0.83579	57	Levitt
24	LEE	407	0.82484	64	MultiMod+CSA?/Lee
25	keasar	114	0.82034	59	?Keasar
26	Bates_BMM	178	0.81492	61	Bates
27	**Phyre_de_novo**	**322s**	**0.81193**	**57**	**HHSearch+MultiMod+Poing/Sternberg**
28	3DShot1	282	0.78625	64	Seth
29	POEMQA	124	0.76109	64	Wenzel
30	MidwayFolding	208	0.76065	62	Sosnick
31	**METATASSER**	**182s**	**0.76047**	**64**	**TASSER/Skolnick**
32	Elofsson	200	0.76016	63	Elofsson
33	**MULTICOM-CLUSTER**	**020s**	**0.70844**	**64**	**HHSearch+MultiMod/Cheng**
34	**MUSTER**	**408s**	**0.70141**	**64**	**/Zhang**
35	3DShotMQ	419	0.70125	64	Seth
36	**Phyre2**	**235s**	**0.66549**	**51**	**HHSearch+MultiMod/Sternberg**
37	**MULTICOM-REFINE**	**013s**	**0.65078**	**64**	**HHSearch+COMPASS+MultiModCheng**
38	SAMUDRALA	034	0.60904	52	Samudrala
39	**FEIG**	**166s**	**0.60726**	**62**	**TASSER+Rosetta+MultiMod/Feig**
40	PS2-manual	023	0.60649	57	Chen

Rosetta
TASSER
Meta-Method
Human expert

Targets where template(s) could be (easily) found in PDB



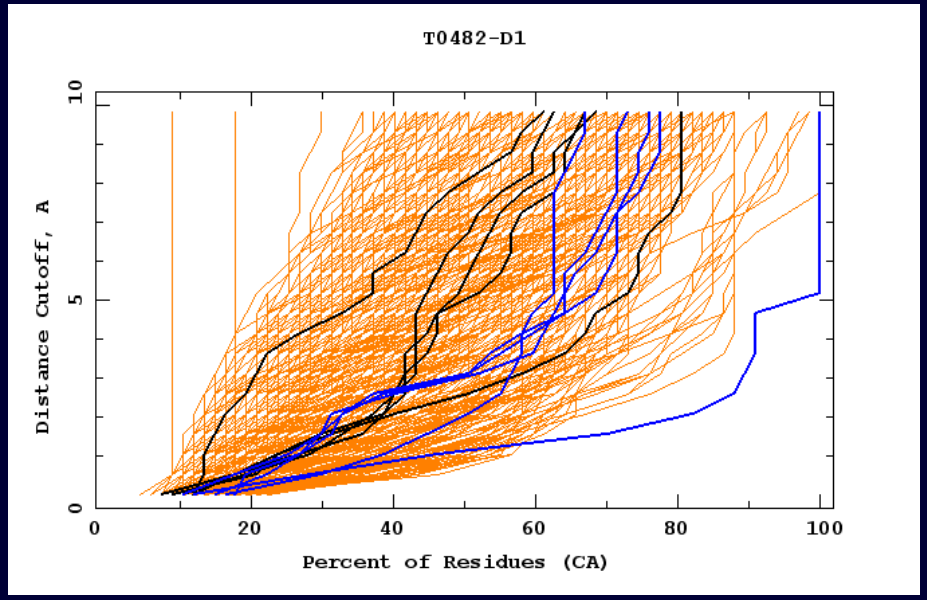
© Michael Feig

CASP8: Human/Server FM Results

1	DBAKER	489	2.22333	9
2	keasar	114	1.99333	9
3	MULTICOM	453	1.99222	9
4	MUFOLD-MD	404s	1.96667	9
5	BAKER-ROBETTA	425s	1.89889	9
6	Zico	299	1.88889	9
7	ZicoFullSTPFullData	138	1.86889	9
8	ZicoFullSTP	196	1.86111	9
9	LevittGroup	442	1.84222	9
10	mufold	310	1.79222	9
11	Zhang	071	1.75667	9
12	POEMQA	124	1.74111	9
13	fams-ace2	434	1.73111	9
14	SAM-T08-human	046	1.63667	9
15	Zhang-Server	426s	1.61000	9
16	Jones-UCL	387	1.51444	9
17	Bates_BMM	178	1.51222	9
18	McGuffin	379	1.50556	9
19	PSI	385s	1.48444	9
20	POEM	207	1.47222	9
21	ABIpro	340	1.42778	9
22	FALCON	351s	1.39667	9
23	A-TASSER	149	1.39000	9
24	RBO-Proteus	479s	1.37889	9
25	TASSER	057	1.34444	9
26	Sternberg	202	1.33556	9
27	pro-sp3-TASSER	409s	1.22222	9
28	Chicken_George	081	1.20556	9
29	FALCON_CONSENSUS	220s	1.16222	9
30	Bilab-UT	325	1.13667	9
31	Sasaki-Cetin-Sasai	461	1.12556	9
32	TJ_Jiang	384	0.98667	9
33	METATASSER	182s	0.95222	9
34	fais-server	116s	0.85111	9
35	3DShotMQ	419	0.77778	9
36	MULTICOM-CLUSTER	020s	0.71222	9
37	3DShot1	282	0.61111	9
38	MUProt	443s	0.60222	9
39	LEE	407	0.59000	9
40	MULTICOM-REFINE	013s	0.55778	9
41	FEIG	166s	0.49222	9

Rosetta
TASSER
Meta-Method
Human expert

Targets where template could not (easily) be found in PDB



© Michael Feig

Questions?

© Michael Feig