COMPUTATIONAL STRUCTURAL BIOLOGY

STRUCTURE, SIMULATION, FUNCTION & PREDICTION

Lecture 8

Michael Levitt
Structural Biology, Stanford

http:/csb.stanford.edu/clas

STRUCTURE PREDICTION I

Why Predict Structure?

Knowledge-Based Physics.

Critical Assessment of Structure Prediction (CASP).

Predict Secondary Structure.

Side Chain Prediction.

Homology Modeling.

Why Predict Structure
Concept 8.1

HUMAN GENOME

- Number of genes 25,000–40,000
  (Psi-BLAST E-value < 0.00001)
- Number of Proteins 30,000–100,000
- Number of Proteins Released 29,802
  (As of 14 Feb 03)
- Number Matching SCOP_1.63 19,355
  (Based on SuperFamily HMM Score)

We have structural data on no more than 25% of .

HALF OF PROTEINS ARE NOVEL

Species	Number of identified genes	% Sequence matching SuperFamily
Human	29,800	46
Fly	33,469	42
Worm	18,266	43
Yeast	6,703	42
Mustard	25,581	45

Gough et al. Assignment of Homology to Genome Sequences using a Library of Hidden Markov Models that Represent all Proteins of Known Structure. J. Mol. Biol. 313: 903 (2001). SuperFamily at: http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY.

IS THERE A FINITE NUMBER OF FOLDS?

- Chothia's 1,000 fold hypothesis.
- Each organism has their own unique protein sequences.
  - About 50% in Eukaryotes so far (Human, Worm, Fly and Yeast have been sequenced)
  - About 30% in Bacteria so far (over 80 species have been sequenced).
- How hard is it to evolve a new protein fold?

Chothia, C. One thousand families for the molecular biologist. Nature, 357:543–544 (1992).

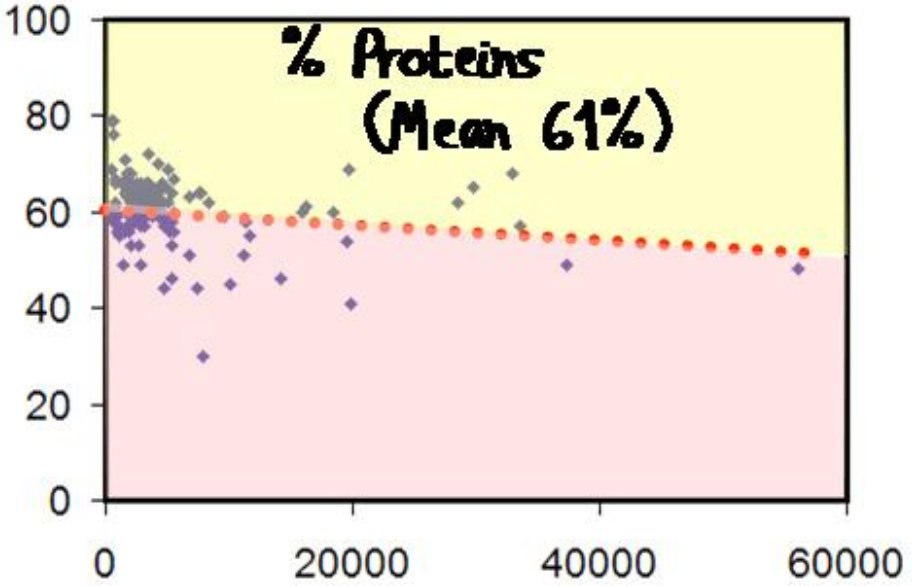
CHOTHIA'S 1,000 FOLD HYPOTHESIS

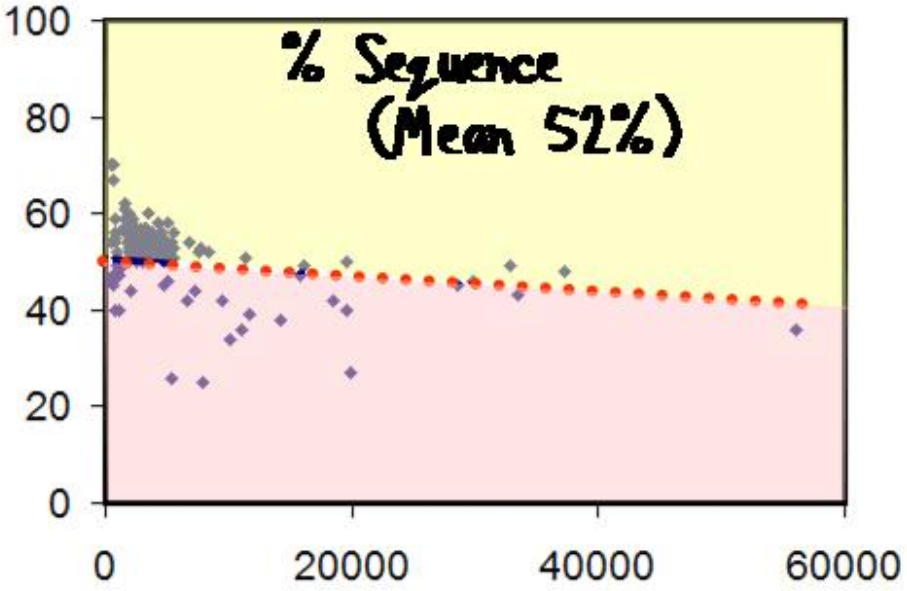
- If new structural families are found and a fraction **f** of these turn out to be one of **M** known superfamilies, the total number of superfamilies is estimated as $N_{SF} = M/f$.
- The results so far indicate that N_{SF} may be much larger.

DATE	N_{SF}
1992	570
1993	615
1994	724
1995	920

- Currently there are 1232 SCOP superfamilies (891 are found in the Human Genome).

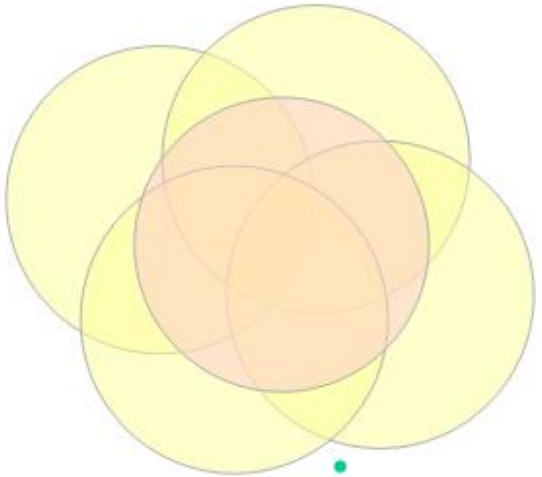
CHOTHIA'S 1,000 FOLD HYPOTHESIS





- 1,232 SCOP super--families can be assigned to about 60% of all the 788,564 proteins in 147 different genomes.

- How many super--families will cover the remaining 40% of the proteins?



Knowledge-Based Physics
Concept 8.2

KNOWLEDGE-BASED PHYSICS

Energy Functions.

Environment.

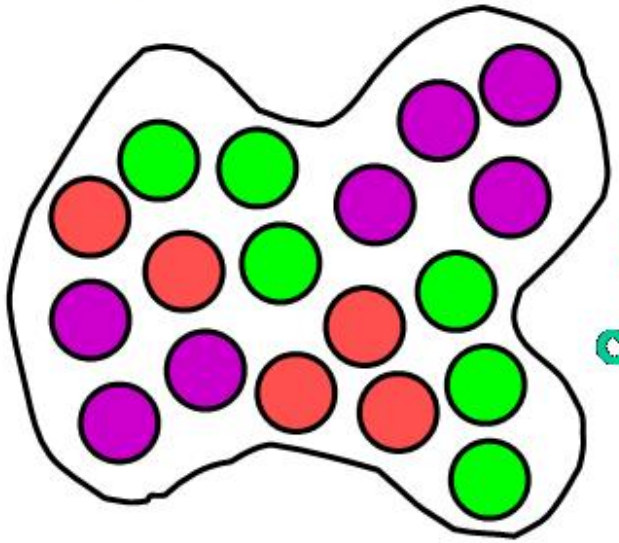
Pairwise

Geometry.

Fragment libraries.

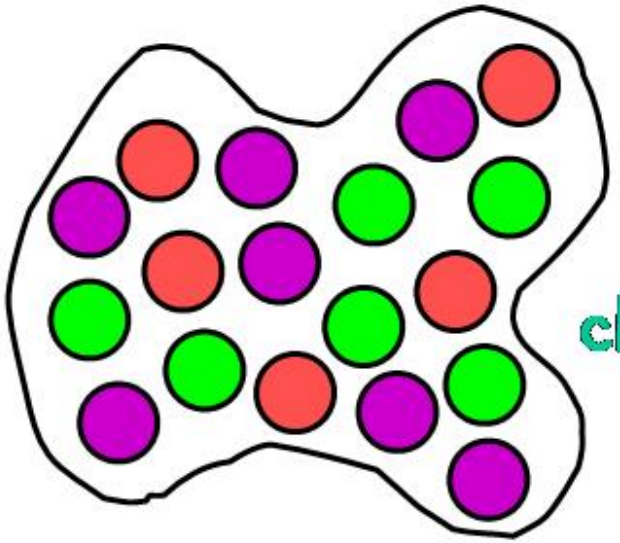
Simulation.

KNOWLEDGE-BASED ENERGIES



Lots of clustering

Native

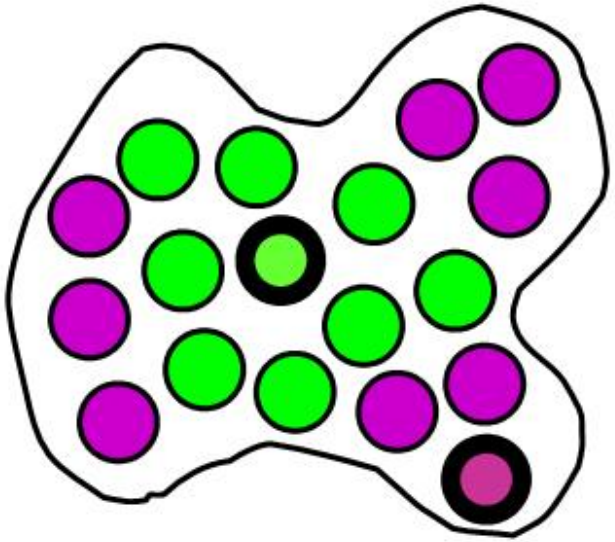


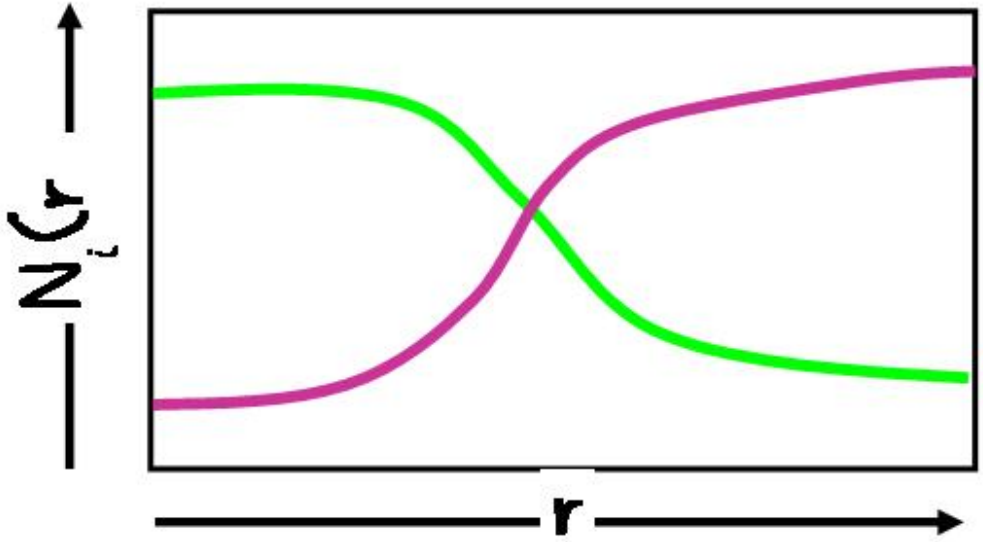
No clustering

Randomized

- Count pairs of centers of each type at different separations, r, to give $N_{ij}(r)$.
- Normalize by the expected count for a random arrangement given by $M_{ij}(r)$.
- Convert to additive score: $E_{ij}(r) = \log(N_{ij}(r)/M_{ij}(r))$.

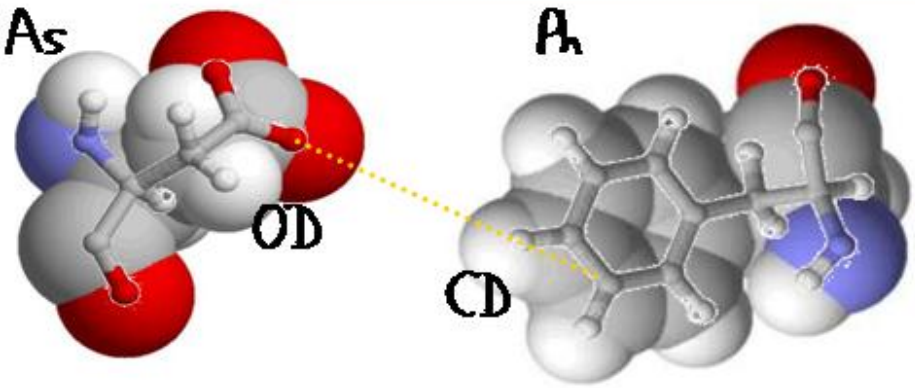
ENVIRONMENTAL ENERGIES

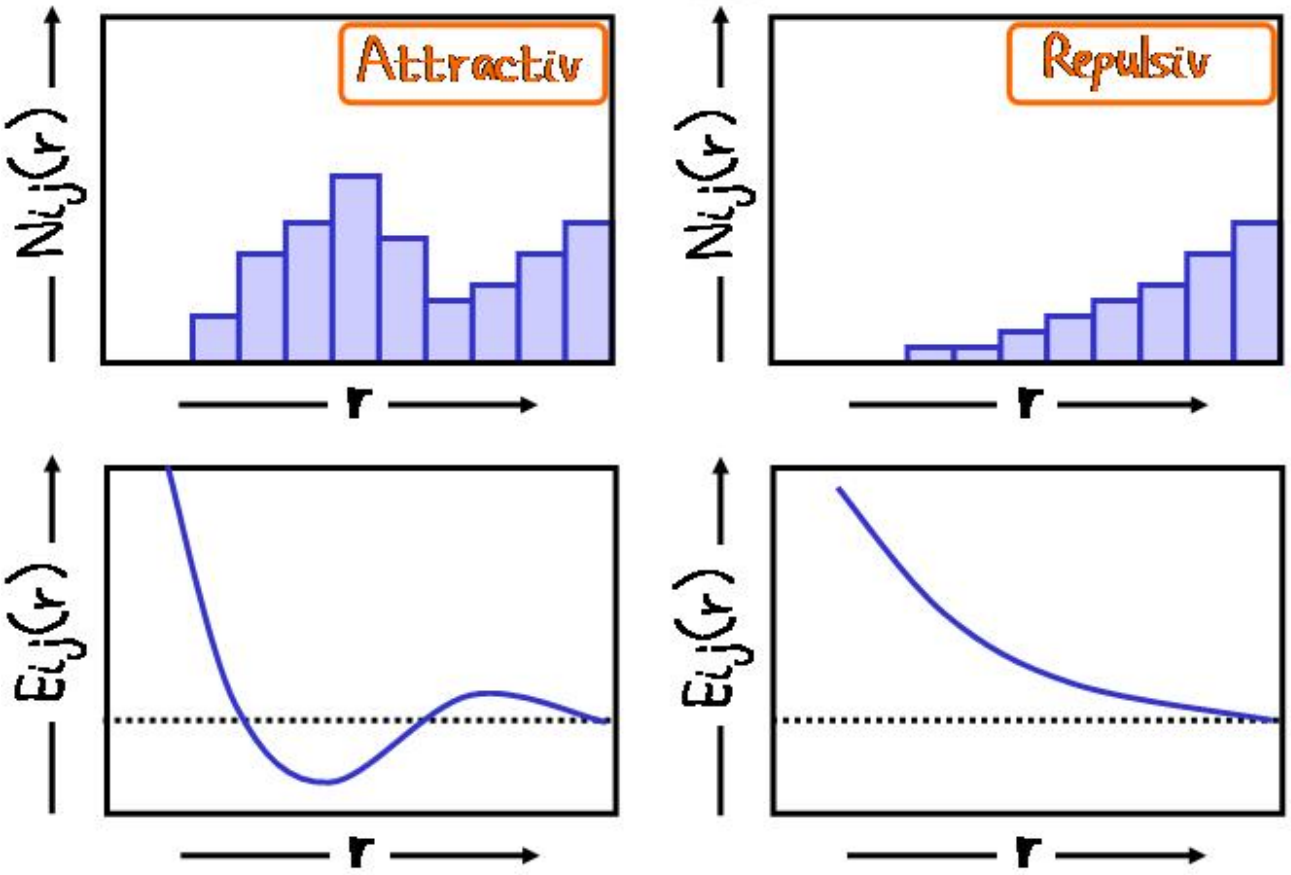




- Count number of neighbors at each distance and for each atom or residue type to give $N_i(r)$.
- Normalize by the expected count for a random arrangement given by $M_i(r)$.
- Convert to additive score: $E_i(r) = \log(N_i(r)/M_i(r))$

PAIR-WISE ENERGIES

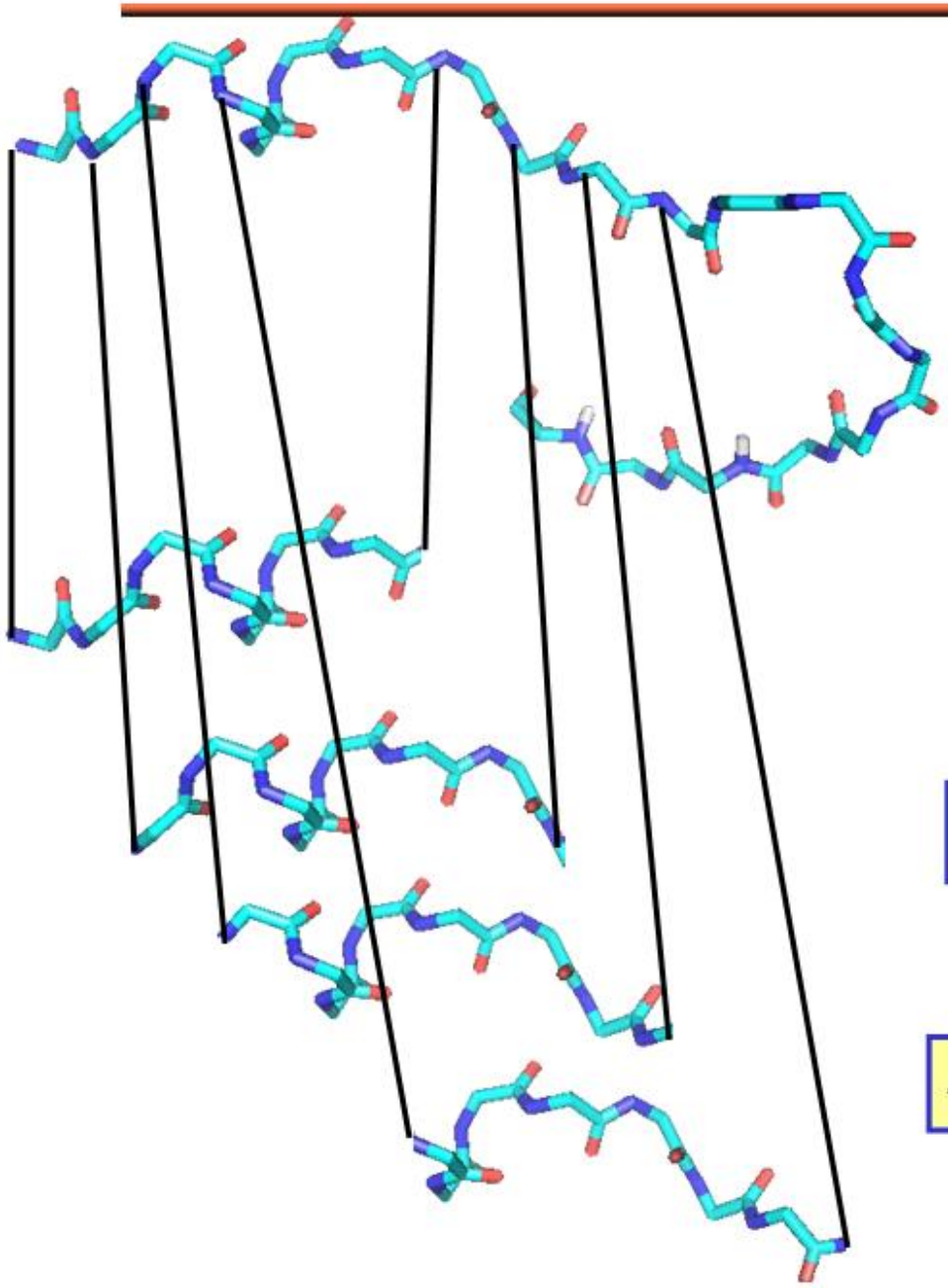




- Get distribution of distances between pairs of atom centers of a particular type, e.g. D-OD1...F-CD2.
- Normalize and take log to get Energy score:

$$E_{ij}(r) = \log(N_{ij}(r)/M_{ij}(r)).$$

KNOWLEDGE-BASED GEOMETRY



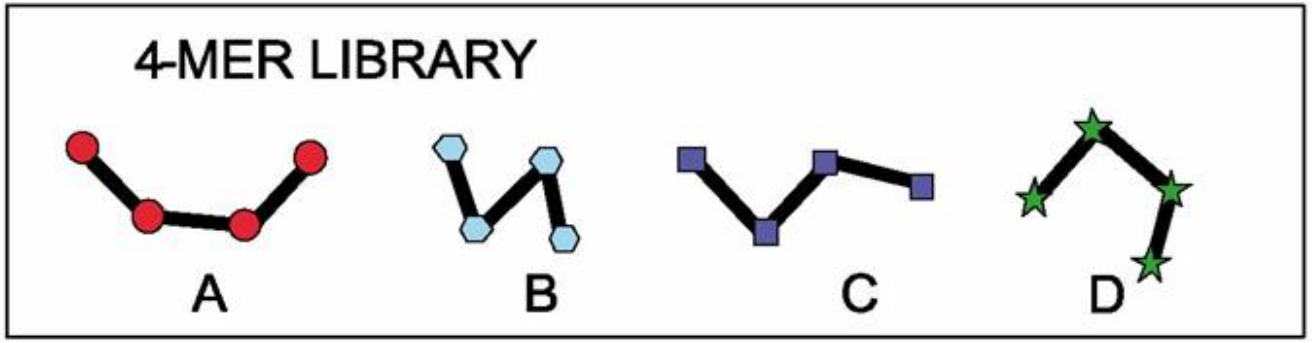
- Cut a protein into over-lapping pieces for re-use.
- May cluster to have less redundancy.

Jones & Thirup, EMBO J. 5, 819 (1986)

Levitt, *J.Mol.Biol.* 226: 507-533 (1992).

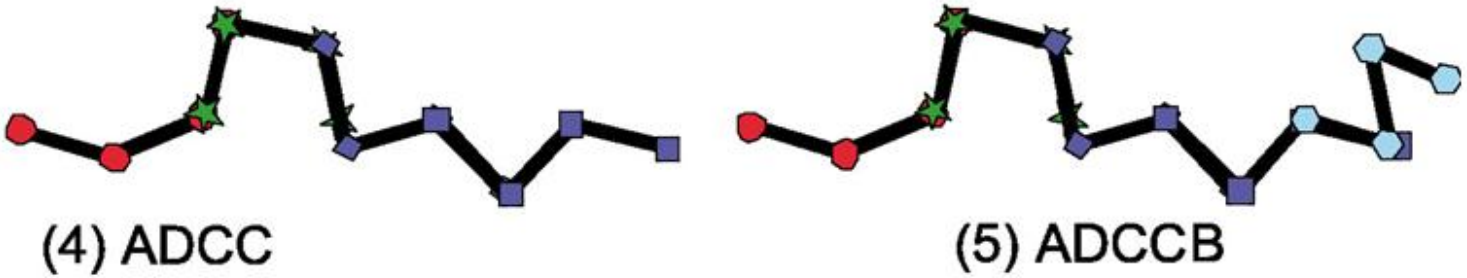
FRAGMENT LIBRARIES

- Build any structure from a library of fragments.

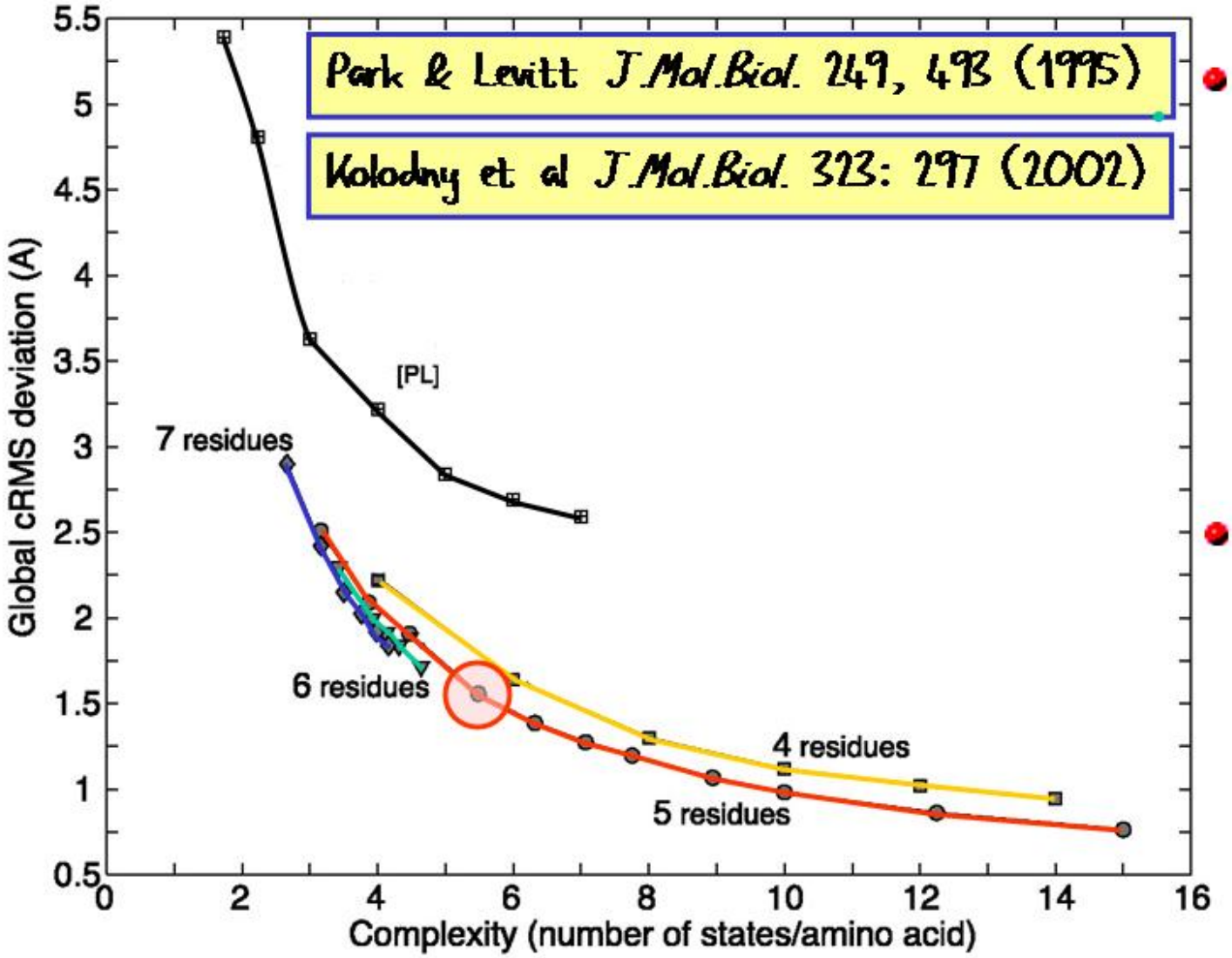


- For 5-residue 20-mer library, a protein of length N will consist of (N-3)/2 fragments.

Kolodny et al, *J. Mol. Biol.* 323: 297 (2002).



FIT AND COMPLEXITY



- Best cRMS decreases with increasing Complexity
- Choose a 20-state 5 residue model as best compromise. $(20)^{1/2} = 4.47$.

KNOWLEDGE-BASED SIMULATION

- Must use methods that do not require physical continuity:
  - Enumeration. Just try different geometries.
  - Monte Carlo simulated annealing.
  - Mean Field optimization.

CASP
Concept 8.3

WHAT IS CASP?

Critical Assessment of Structure Prediction.

Conceived of by John Moult in 1993.

First run in 1994 with December meeting in Asilomar.

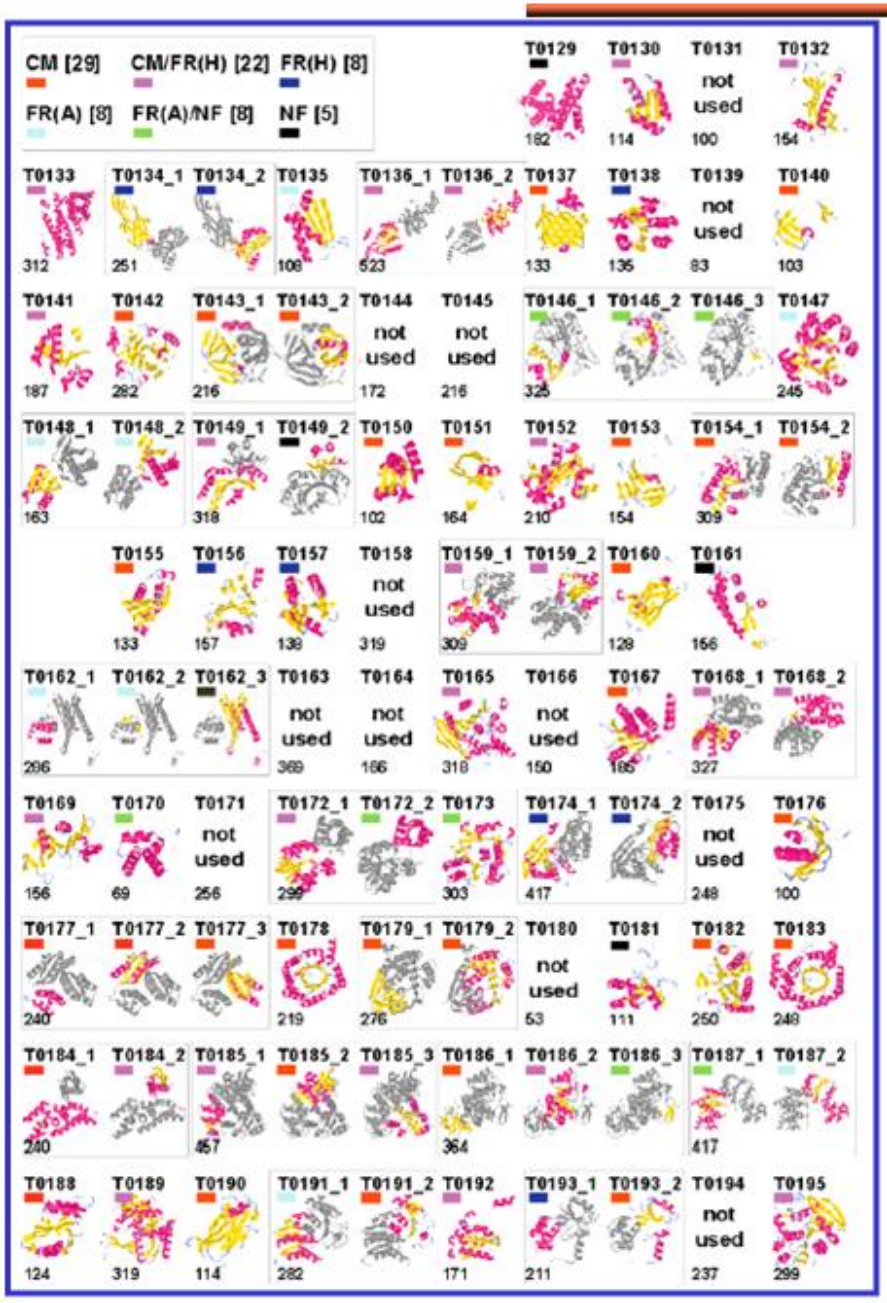
Has been CASP1 through CASP5.

Four CASP categories:

- Secondary Structure.
- Homology or Comparative Modeling (CM).
- Fold Recognition and Threading (FR).
- Ab Initio or New Fold Prediction (NF).

Moult et al. Proteins. 23: ii–v (1995)

CASP TARGETS



- Where do targets come from?
- Always a small number.
- Cannot ensure variety or balance.
- Poor statistics at any meeting.

91 Targets at CASP5.

10 not used?

THE CASP PROCESS

Real deadlines.

Pressure to cut corners.

Resource management.

Group dynamics.

Collaboration vs. Competition.

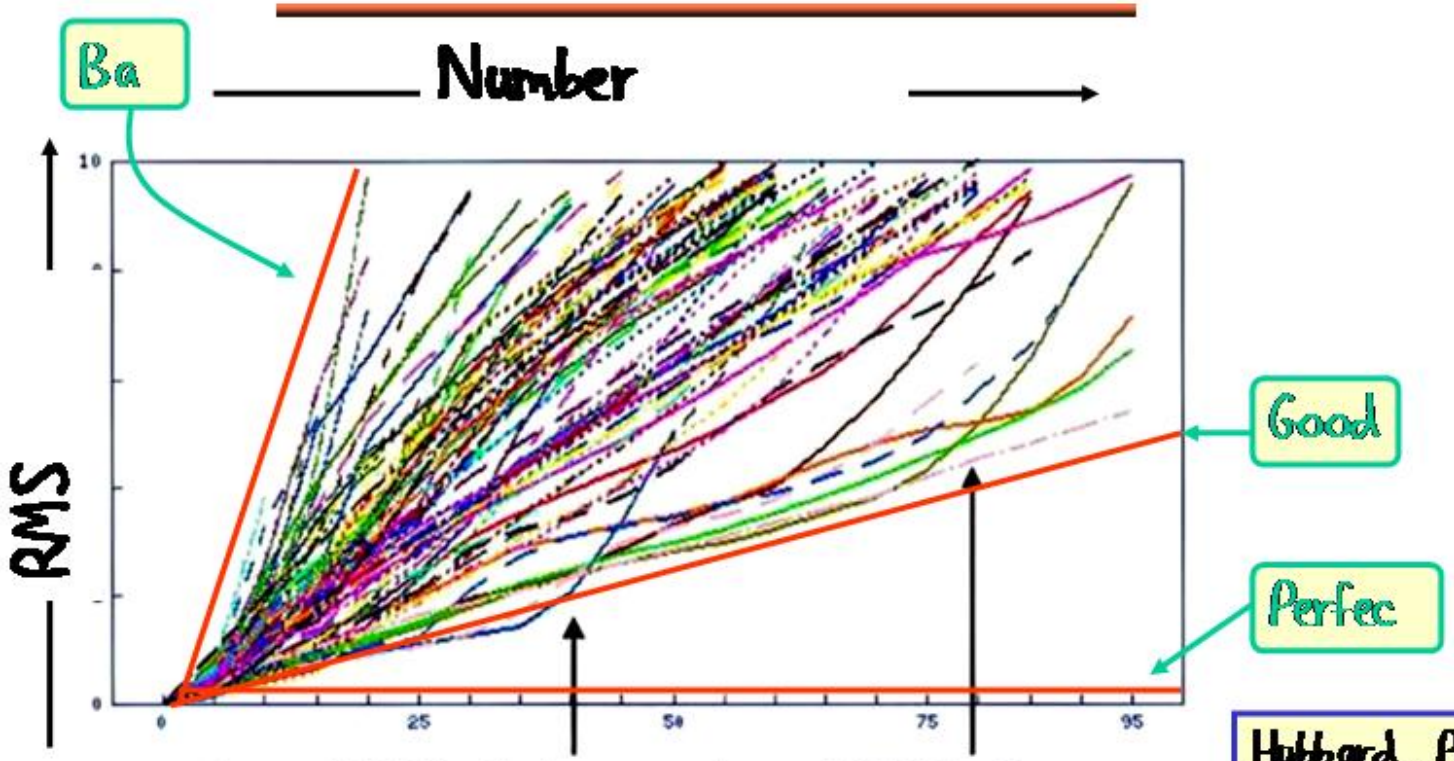
Winners and Losers?

CASP STATISTICS

MEETING	TARGETS	PREDICTORS	SERVERS	PREDICTIONS
CASP 1 94	35	35	0	100
CASP 2 96	42	72	0	947
CASP 3 98	43	98	0	3,807
CASP 4 00	43	133	30	11,136
CASP 5 02	67	216	72	28,728

- The number of targets increased from
- Massive increase in number of
- Servers entered for the first time at

HUBBARD PLOTS



Hubbard. Proteins. 53, 15–21 (1999).

- Find RMS of a non-consecutive subset of atoms that fit "best".
  - Generate all possible superpositions of three adjacent CA atoms.
  - Extend each to include additional points. Re-superimpose.
  - Store lowest RMS for each number of atoms in the subset.

GDT-TS SCORE

- GDT is an abbreviation for "Global Distance Test".

- The Prediction Center has defined a quality index, GDT-TS.

  $GDT\text{-}TS = 100(GDT_1 + GDT_2 + GDT_4 + GDT_8)/(4N_T)$

  where N_T is number of residues in target and GDT_n is maximum number of residues superimposed under n Å distance cutoff.

  (because any residue that is part of the GDT1 subset will also be part of the GDT2, GDT4 and GDT8 subsets, this gives most weight to the residues that fit well)

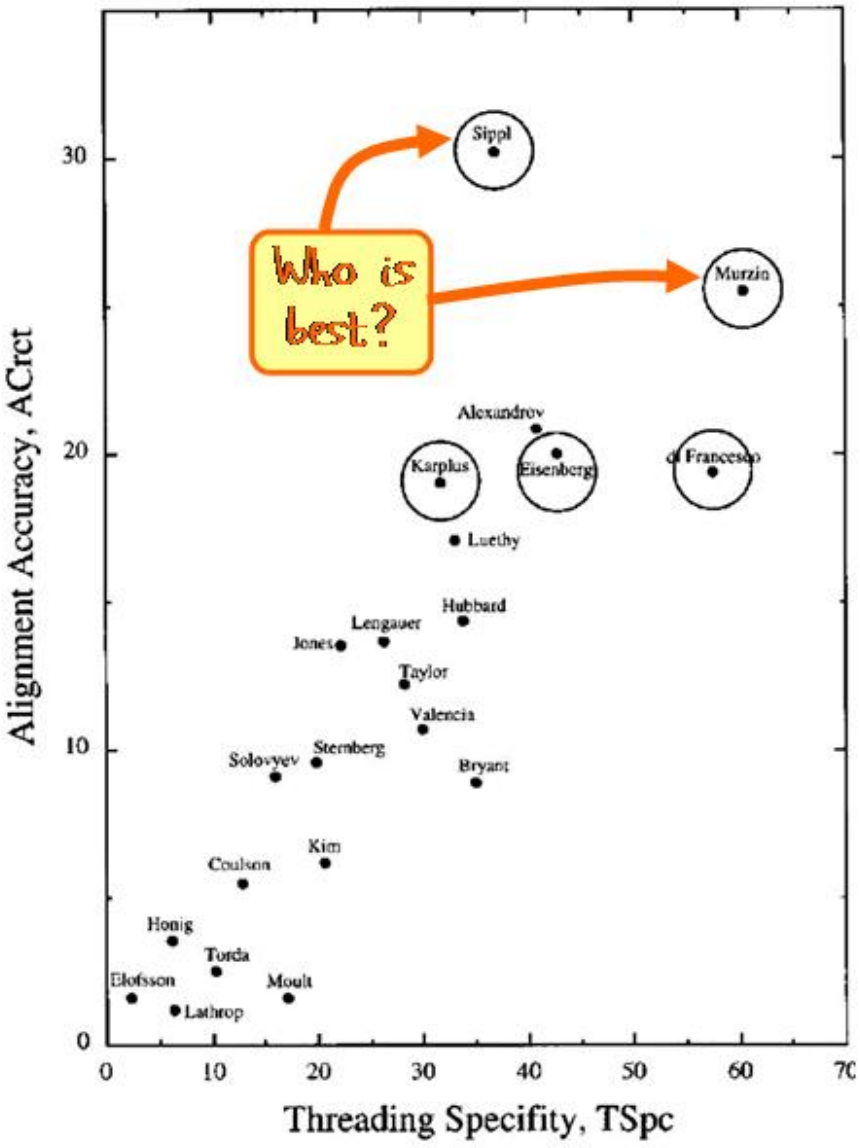
- GDT-TS is ideally suited to automatic assessment.

Zemla, Nucleic Acids Research, 31: 3370. (2003).

ASSESSMENT

- Manual Assessment
  - Subtle and Intuitive.
  - Inconsistent and Biased.
  - Needed it early on to know what was important.
  - Move to totally automated evaluation at CASP6?
- Automatic Assessment
  - Levitt's CASP2 Experiences.
  - Must be one-dimensional measure.
  - Use a simple sum of Z-scores.

CASP2 EXPERIENCES



- Used indices provided by evaluators; did not select a single index.
- This is bad as impossible to rank in 2-dimensions.
- What is more important Accuracy or Specificity?
- Must use 1-Dimensional score for objective ranking.

Levitt, Proteins, S1: 92 (1997).

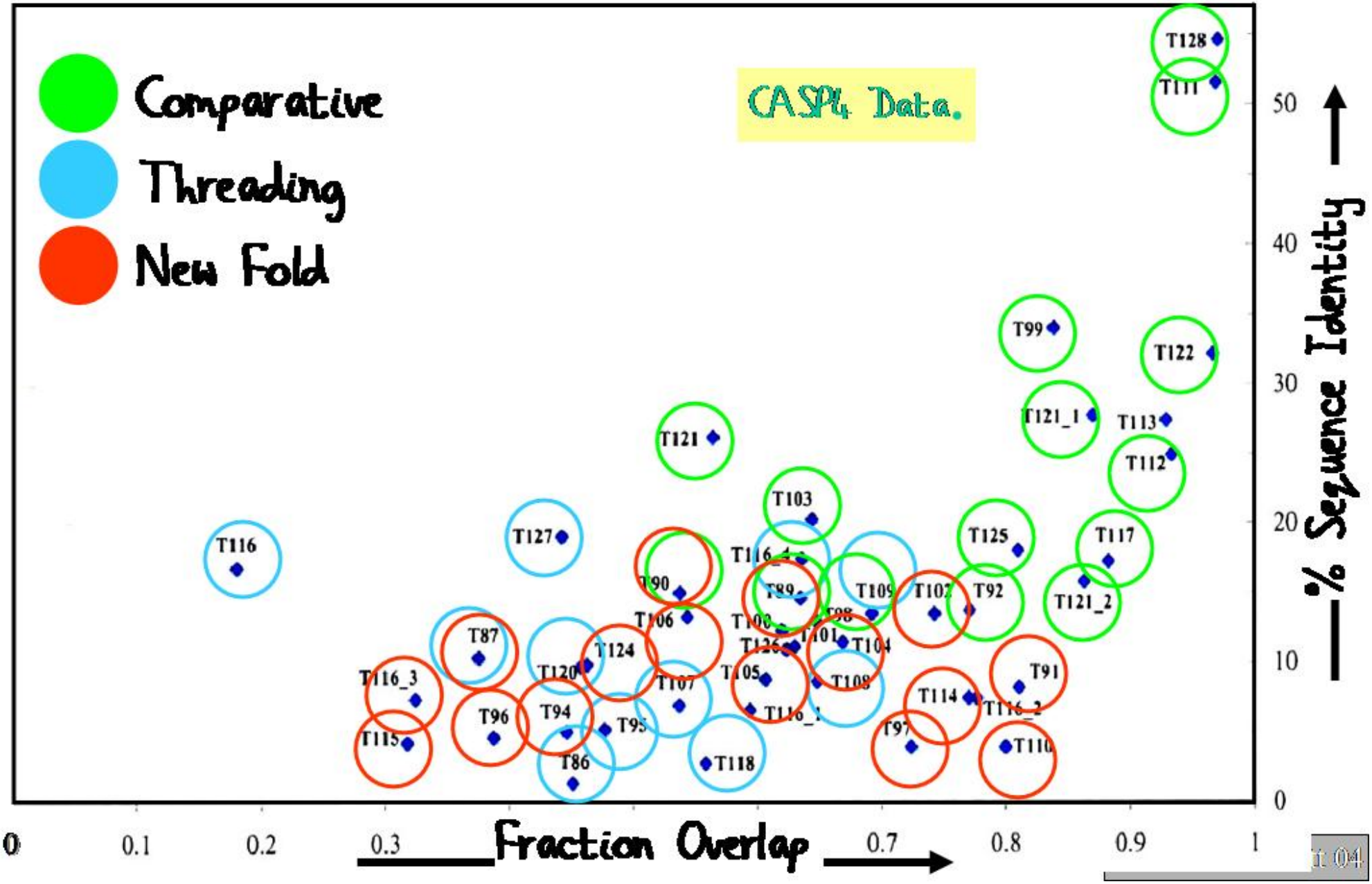
SIMPLE Z-SCORE SUM

- Choose some index like Zemla's GDT-TS score.
- For each target protein, get the mean score of all predictions as well as standard deviation.
- Express each individual score as: Z-score = (score-mean)/SD.
- If the Z-score is above a threshold, sum the Z-scores (do not penalize for bad predictions).
- Threshold will be between Z = 0 (lenient) to Z = 2 (strict).

Sum or average?

Still subjective!!!

WHICH TARGETS ARE WHICH?



Predict Secondary Structure
Concept 8.4

PREDICT SECONDARY STRUCTURE

Predicting Secondary Structure?

Early History.

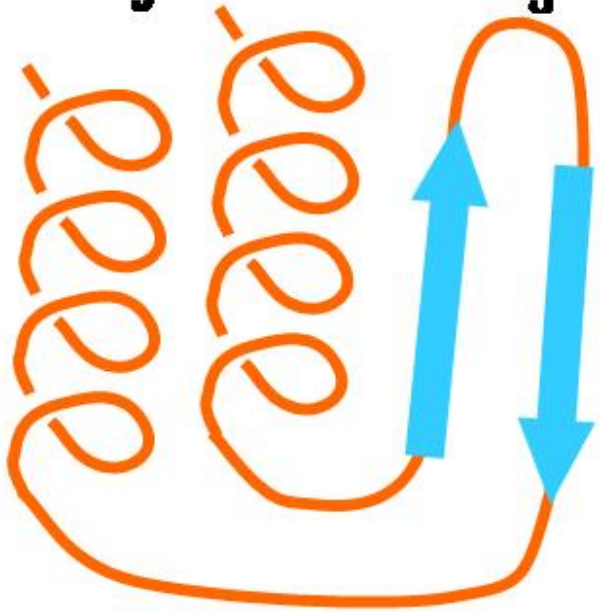
Neural Networks.

Multiple Sequence Methods.

Secondary Structure Prediction at CASP.

PREDICTING SECONDARY STRUCTURE

- Assign the secondary structure of every residue from protein structure.



Amino Acid

ALHEASGPSVILFGSDVTVPPASNAEQAK

αααααttttββββtttββββttttαααααα

Secondary Structure

- Now predict this secondary structure from the amino acid sequence.
- Two general schemes relate secondary structure to sequence.
  - Statistical: Count how often each type of residue occurs in each type of secondary structure.
  - Patterns: Look for characteristic sequence patterns that define the ends of helices, β--turns, etc.

SECONDARY STRUCTURE STATES

Normally three states are defined:

STATE	ABBREVIATIONS		
Alpha Helix	α	H	
Beta Strand	β	E	
Other residues	t	C	L

E is for Extended chain.

t is for turn.

C for Coil.

L for Loop.

EARLY HISTORY

- Chou-Fasman Secondary Structure (Chou & Fasman, 1974).
- Early CASP-like Test (Adenylate Kinase, Schulz et al., 1974).
- Amino Acid Classification (Levitt 1978).

CHOU-FASMAN SECONDARY STRUCTURE

Name	P(α)	P(β)	P(turn)	
Ala	1.42	0.83	0.66	
Arg	0.98	0.93	0.95	
Asp	1.01	0.54	1.46	←
Asn	0.67	0.89	1.56	
Cys	0.70	1.19	1.19	←
Glu	1.39	1.17	0.74	←
Gln	1.11	1.10	0.98	←
Gly	0.57	0.75	1.56	
His	1.00	0.87	0.95	
Ile	1.08	1.60	0.47	←
Leu	1.41	1.30	0.59	←
Lys	1.14	0.74	1.01	←
Met	1.45	1.05	0.60	
Phe	1.13	1.38	0.60	←
Pro	0.57	0.55	1.52	
Ser	0.77	0.75	1.43	
Thr	0.83	1.19	0.96	
Trp	1.08	1.37	0.96	←
Tyr	0.69	1.47	1.14	←
Val	1.06	1.70	0.50	←

- Find 4 out of 6 contiguous residues with $P(\alpha) > 1$. Extend the helix in both directions until a set of 4 contiguous residues with an average $P(\alpha)<1$.
- Find 3 out of 5 contiguous residues have $P(\beta) > 1$. Extend the strand in both directions until a set of 4 contiguous residues with an average $P(\beta)<1$..

Chou & Fasman. Prediction of Protein Conformation. Biochemistry, 13: 211-245 (1974).

Problem is that 11 of 20 amino acids favor two secondary structures.

EARLY BLIND EXPERIMENT

- First truly blind test. It was organized by the crystallographer about to solve a structure (George Schulz, adenylate kinase).
- About 60% correct (more about this later).
- Consensus better than any one prediction.

Schulz, G. E., C. D. Barry, J. Friedman, P. Y. Chou, G. D. Fasman, A. V. Finkelstein, V. I. Lim, O. B. Ptitsyn, E. A. Kabat, T. T. Wu, M. Levitt, B. Robson and K. Nagano. Comparison of Predicted and Experimentally Determined Secondary Structure of Adenylate Kinase. Nature 250, 140–142 (1974).

STATISTICAL PREFERENCES

- Statistical secondary structure preferences have changed little since 1978.
- The preferences are weak but significant.

1978

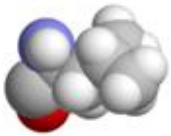
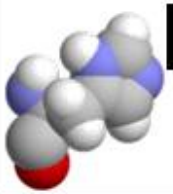
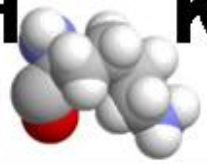
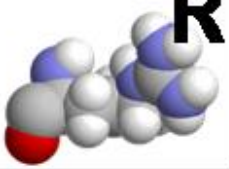
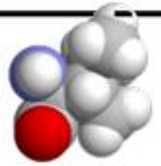
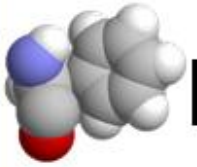
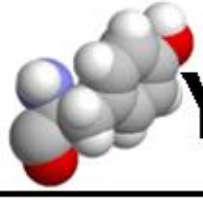
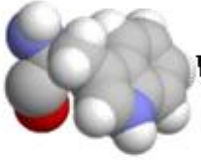
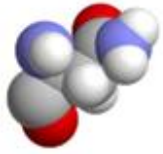
AMINO ACID	α-HELIX (P_α)	β-SHEET (P_β)	REVERSE TURN (P_t)
Ala	1.29	0.90	0.78
Cys	1.11	0.74	0.80
Leu	1.30	1.02	0.59
Met	1.47	0.97	0.39
Glu	1.44	0.75	1.00
Gln	1.27	0.80	0.97
His	1.22	1.08	0.69
Lys	1.23	0.77	0.96
Val	0.91	1.49	0.47
Ile	0.97	1.45	0.51
Phe	1.07	1.32	0.58
Tyr	0.72	1.25	1.05
Trp	0.99	1.14	0.75
Thr	0.82	1.21	1.03
Gly	0.56	0.92	1.64
Ser	0.82	0.95	1.33
Asp	1.04	0.72	1.41
Asn	0.90	0.76	1.28
Pro	0.52	0.64	1.91
Arg	0.96	0.99	0.88

2001

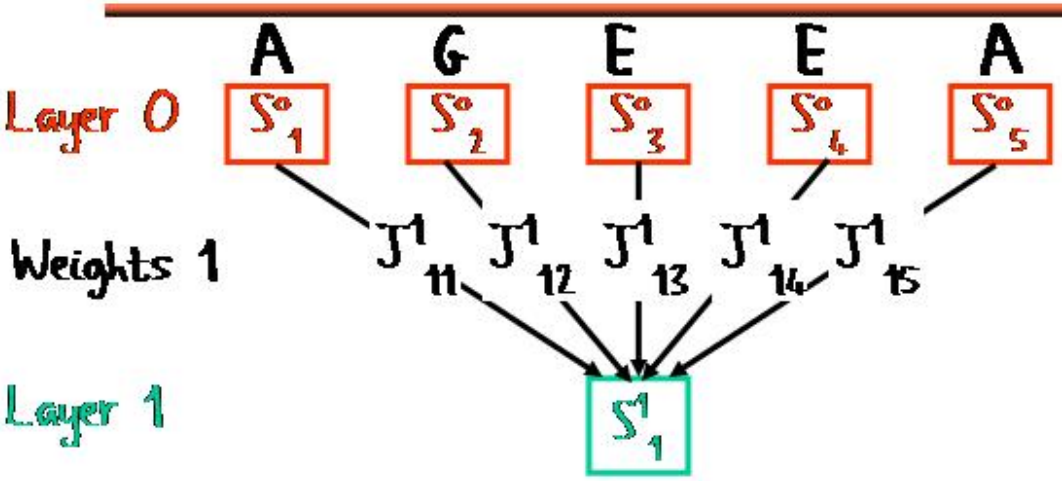
Amino Acid	α	β	other
ALA	1.472	.780	.784
GLU	1.385	.745	.862
LEU	1.352	1.123	.696
GLN	1.332	.789	.877
MET	1.290	.978	.811
ARG	1.245	.892	.885
LYS	1.161	.828	.975
VAL	.894	1.806	.672
ILE	1.020	1.712	.632
TYR	.974	1.466	.786
PHE	.962	1.417	.819
TRP	.989	1.271	.873
THR	.759	1.245	1.044
CYS	.748	1.209	1.070
PRO	.409	.455	1.678
GLY	.444	.644	1.560
ASP	.862	.547	1.320
ASN	.799	.671	1.302
SER	.771	.866	1.225
HIS	.922	1.035	1.037

Levitt, Biochemistry, 17, 4277 (1978).

AMINO ACID CLASSIFICATION

	Non-Polar	Polar
Alpha Helix	 A  C  L  M	 N  E  H  K  R
Beta Strand	 V  I  F  Y  W	T  Bulky
Turn	 G  P Short.	S  N  D 

ELEMENTARY NEURAL NETWORK



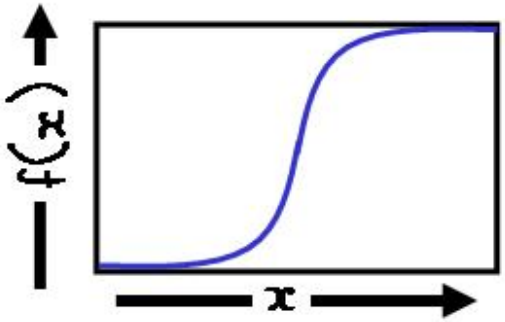
$S^1_1 = f(h^1_1)$, f( ) is sigmoid function

$h^1_1 = \sum J^1_{1i} S^0_i$, as weighted sum of values on Layer 0.

S^0_i depends on type of amino acid at i.

$f(x) = 1/(1+\exp(-x))$

Sigmoid Trigger



- Connect all points in one layer to each point in next layer.
- The value at this point is a weighted sum of its inputs.
- The output of this point is thresholded to introduce non-linearity (like a real neuron).

TRAINING A NEURAL NETWORK

Calculated state of central residue is:

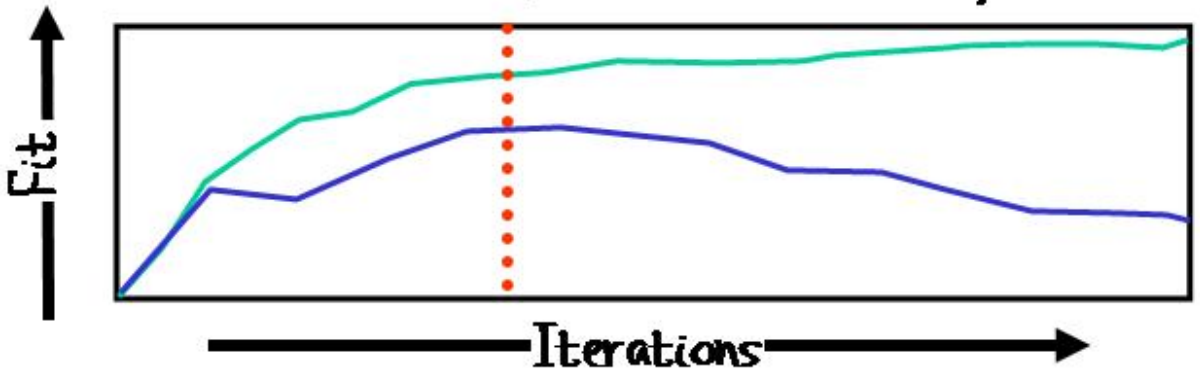
$$S^1_1 = f(h^1_1)$$

$$h^1_1 = \sum J^1_{1i} S^0_i$$

Observed state of central residue is 1 if helical, 0 if not.

Need to adjust the parameters J^1_{1i} and S^0_i to get calculated state to be same as observed state.

Use a form of steepest descent minimization



- A Neural net has very many adjustable weights.
- For this reason, it is easy to over-learn.
- We want to extract the generalizable information and so need to stop the learning process early.

SEQUENCE PROFILE

```
1    1bpi     ..RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA.
2    1bpi     ..RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA.
3    1bzxI    ..RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA.
4    1fakI    ..APDFCLEPPYDGPCRALHLRYFYNAKAGLCQTFYYGGCLAKRNNFESAEDCMRTC....
5    1bunB    ..RHPDCDKPPDTKICQTVVRAFYYKPSAKRCVQFRYGGCNGNGNHFKSDHLCRCECLEY.
6    1bf0     ..PPWYCKEPVRIGSCKKQFSSFYFKWTAKKCLPFLFSGCGGNANRFQTIGECRKKCLGK.
```

		F	C	L	E	P	P	Y	T	G
1	1bpi									
2	1bpi									
3	1bzxI									
4	1fakI									
5	1bunB									
6	1bf0									

Number of	A	0	0	0	0	0	0	0	0	0
Number of	C	0	6	0	0	0	0	0	0	0
Number of	D	1	0	1	0	0	0	1	1	0
Number of	E	0	0	0	5	0	0	0	0	0
Number of	F	4	0	0	0	0	0	0	0	0
Number of	G	0	0	0	0	0	0	0	0	5
Number of	H	0	0	0	0	0	0	0	0	0
Number of	I	0	0	0	0	0	0	0	1	0
Number of	K	0	0	1	1	0	0	0	0	1
Number of	L	0	0	4	0	0	0	0	0	0
Number of	M	0	0	0	0	0	0	0	0	0
Number of	N	0	0	0	0	0	0	0	0	0
Number of	P	0	0	0	0	6	5	0	0	0
Number of	Q	0	0	0	0	0	0	0	0	0
Number of	R	0	0	0	0	0	0	1	0	0
Number of	S	0	0	0	0	0	0	0	0	0
Number of	T	0	0	0	0	0	0	0	4	0
Number of	V	0	0	0	0	0	0	0	0	0
Number of	W	0	0	0	0	0	0	0	0	0
Number of	Y	1	0	0	0	0	1	4	0	0
Number of	.	0	0	0	0	0	0	0	0	0

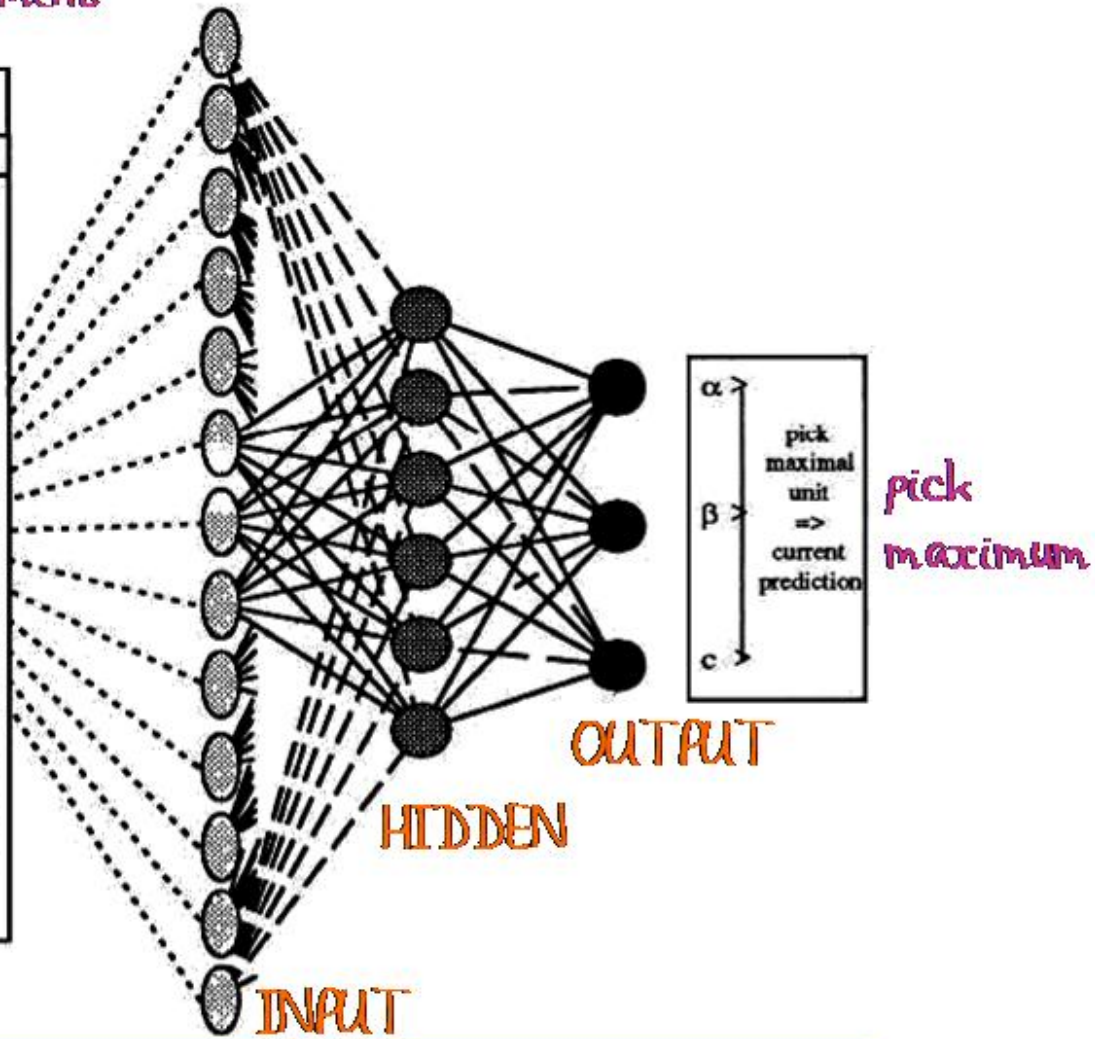
- For each position along the sequence, tabulate how often each of the 20 amino acids occur (also count the gap denoted ".").
- Convert to frequencies.
- A profile is always 21 by N no matter how many sequences are compared.

Also called Position Sensitive Scoring Matrix.

ROST 1993 NEURAL NETWORK

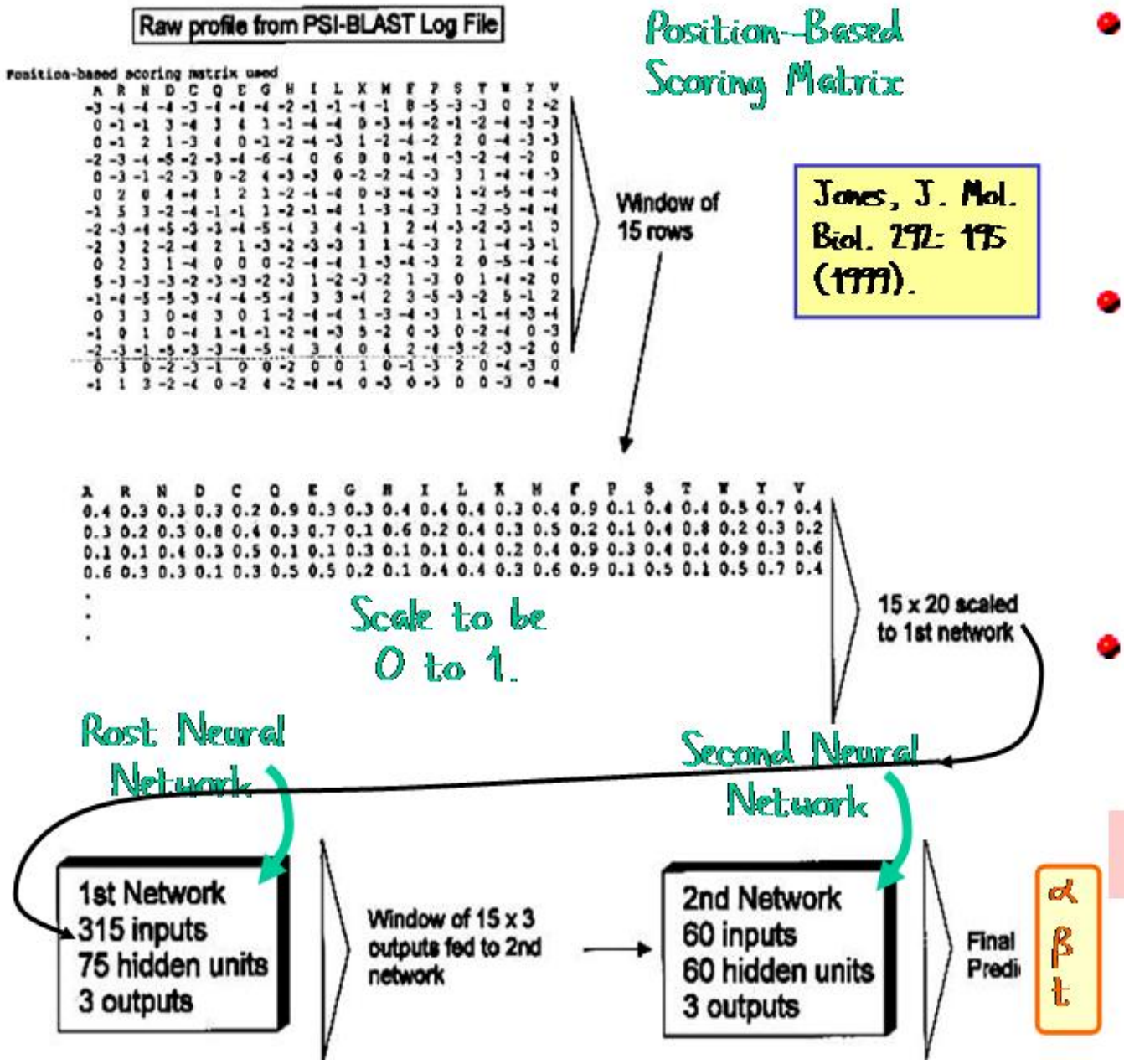
Start with Multiple Sequence Alignment

Protein	Alignments	profile table
		GSA PD NT EKQ C VH IR LM YF W
:	: : : :	
G	G G G G	5 . . . . .
Y	Y Y Y Y	. . . . . 5 . .
I	I I E E	. . . . . 2 . . . 3 . . . . .
Y	Y Y Y Y	. . . . . 5 . .
D	D D D D	. . . . 5 . . . . .
P	P P P P	. . . 5 . . . . .
E	A E A A	. . 3 . . . 2 . . . . .
D	V V E E	. . . . 1 . . 2 . . 2 . . . . .
G	G G G G	5 . . . . .
D	D D D D	. . . . 5 . . . . .
P	P P P P	. . . 5 . . . . .
D	D T D D	. . . . 4 . 1 . . . . .
D	N Q N N	. . . . 1 3 . . . 1 . . . . .
G	G N G G	4 . . . . 1 . . . . .
V	V I V V	. . . . . 4 . 1 . . . . .
N	E P K K	. . . 1 . 1 . 1 2 . . . . .
P	P P P P	. . . 5 . . . . .
G	G G G G	5 . . . . .
T	T T T T	. . . . . 5 . . . . .
D	E K S A	. 1 1 . 1 . . 1 1 . . . . .
F	F F F F	. . . . . 5 .
:	: : : :	



Rost & Sander. Improved Prediction of Protein Secondary Structure by Use of Sequence Profiles and Neural Networks. PNAS. 90: 7558–7562 (1993).

JONES 1999 NEURAL NETWORK



- Neural net is almost identical to that used by Rost & Sander.
- Use PSI-BLAST to both find similar sequences and to align them. (QUICK)
- Train on a much bigger set of proteins.

Clear winner at CASP4.

SECONDARY STRUCTURE PREDICTION QUALITY

- Determine secondary structure of solved protein by DSSP or Stride.

Levitt & Greer. Automatic Identification of Secondary Structure in Globular Proteins. J Mol Biol. 114: 181–239 (1977).

Kabsch & Sander. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-bonded and Geometrical Features. Biopolymers, 22: 2577–2637 (1983).

Frishman & Argos. Knowledge-Based Secondary Structure Assignment. Proteins, 23: 566–579 (1995).

ALHEASGPSVILFGSDVTVPPASNAEQAK Amino Acid Sequence

αααααttttβββββtttβββββttttttααααα Actual Secondary Structure

---

tααααttttβββββttttttβββββtttααααααα $Q_3 = 22/29 = 76\%$

This is a very useful prediction

αααααattttααααattttααααattttttααααα $Q_3 = 22/29 = 76\%$

This is a terrible prediction

- Due to minor coordinates shifts, the secondary structure is uncertain to about 10%.
- Thus an essentially perfect prediction would have Q3 of 90%, provided the error is uniform over the sequence.

HISTORICAL RECORD OF BEST PREDICTIONS AT CASP

CASP & YEAR		NUMBER TARGETS	BEST RESULT <Q3>	GROUP
CASP1	1994	6	63	Rost
CASP2	1996	24	70	Rost
CASP3	1998	18	75	Jones
CASP4	2000	28	80	Jones

- Q3 is the percentage correct for a three-state model ($\alpha\beta t$) averaged over all the non-Comparative Modeling targets.
- Steady improvement of about 5% per CASP (every two years).

Sidechain Prediction
Concept 8.5

SIDE CHAIN MODELING

Basic Idea.

Simulated Annealing.

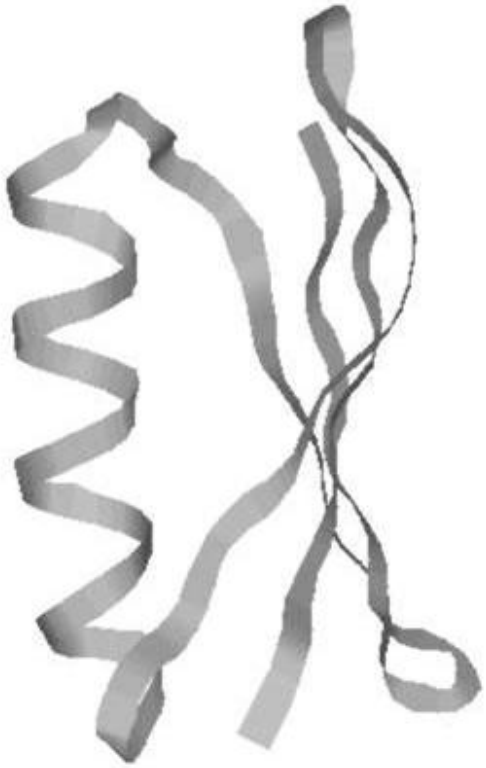
Segment match modeling.

Mean Field.

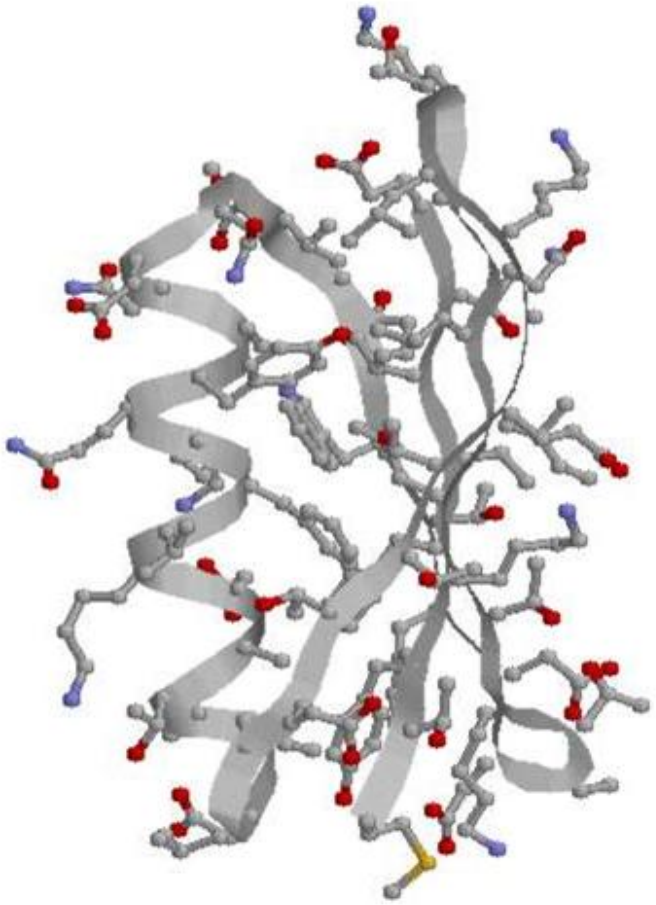
Other Methods: Dead-End Elimination.

Best Methods.

BASIC IDEA



CA Ribbon.

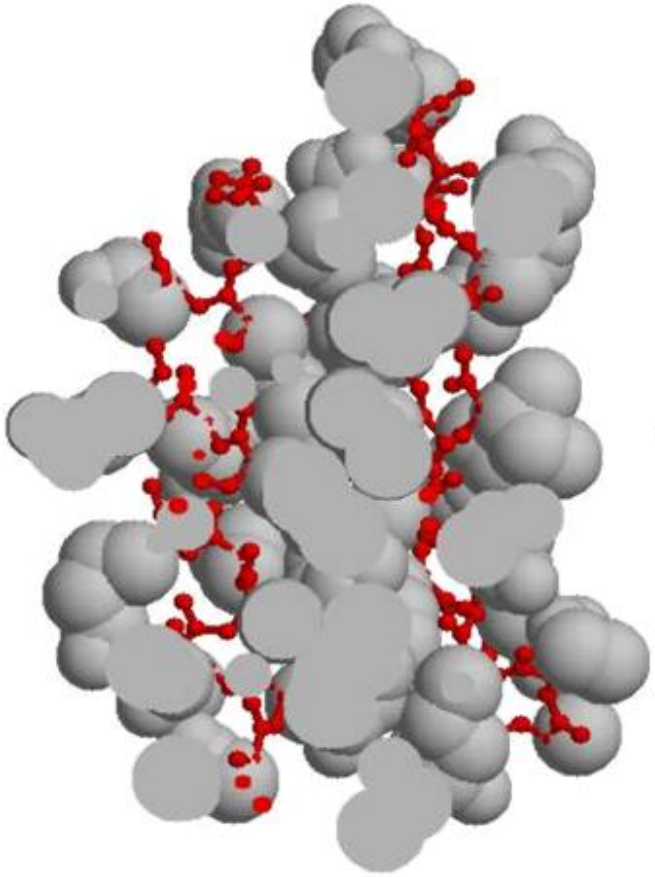


CA Ribbon with sidechains.

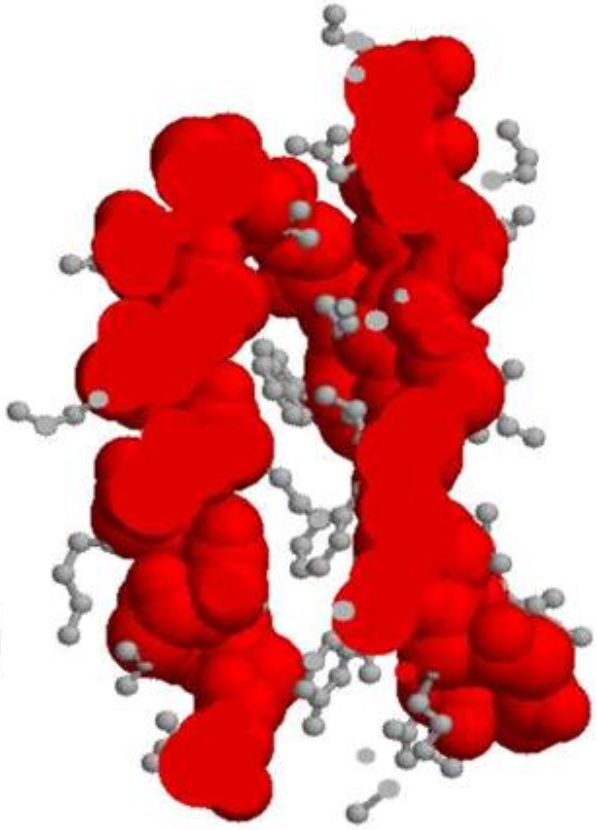
This is 56 residue protein G (1pgp.pdb)

- Sidechains pack well in protein interior.
- If we delete all the sidechains, can we rebuild them?

BASIC IDEA



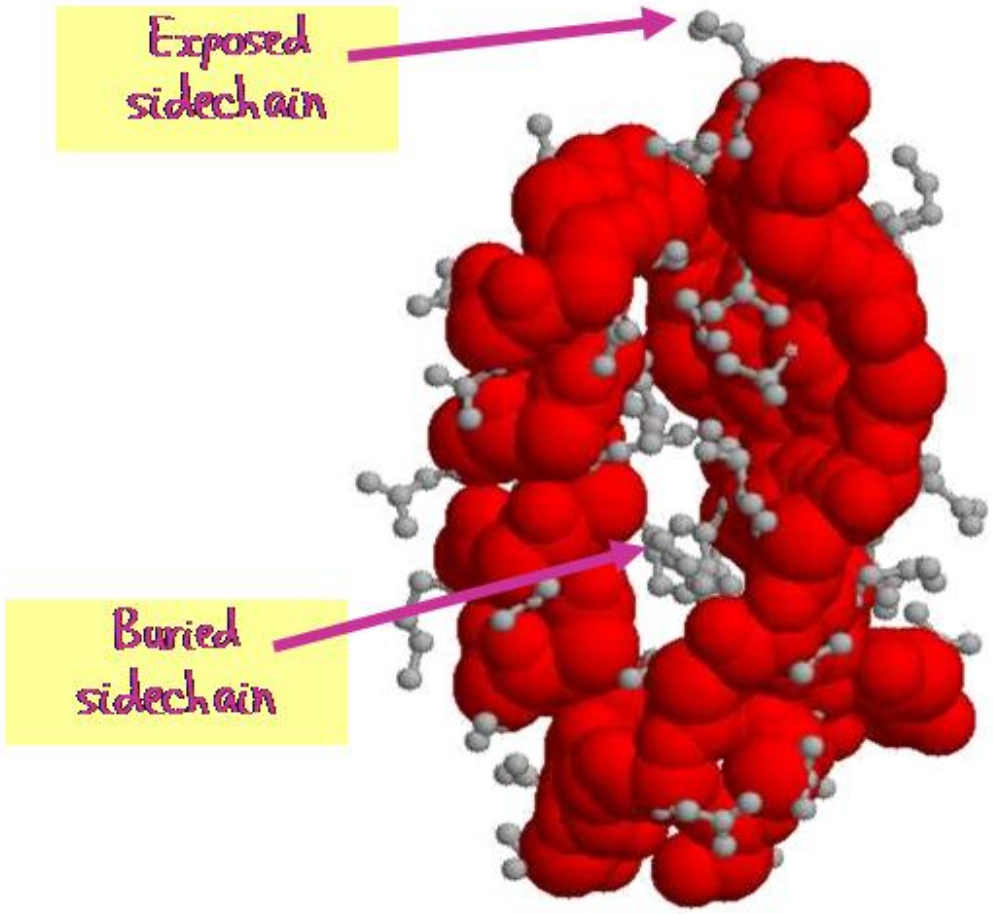
Sidechain atoms are gray.
Mainchain atoms are red.



Protein is like a sandwich

- Sidechains interact with one another.
- They also interact with the mainchain.
- Is rebuilding all the sidechains a hard problem?

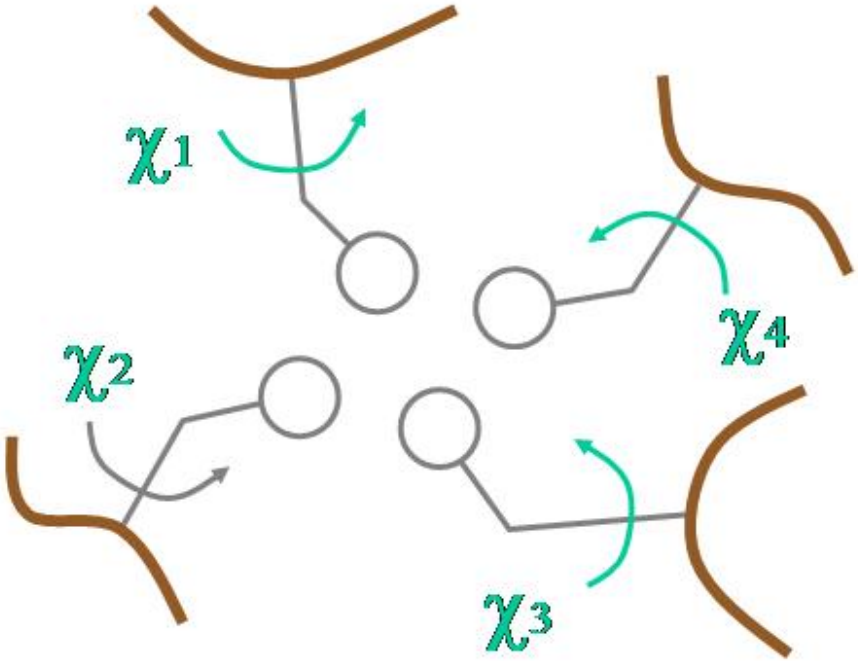
BASIC IDEA



- Sidechains are either buried or exposed.
- Which will be easier to rebuild?

Sidechain atoms are gray.
Mainchain atoms are red.

SIMULATED ANNEALING



- Need to solve the sidechain packing problem.
- Treat it as an optimization problem and solve with Monte Carlo moves.
- Vary all the χ torsion angles until the packing is best.
- For as small protein, there are over 100 χ angles.

This converges easily.

Many of the χ angles of the buried sidechains are predicted to within 30°.

Lee & Subbiah, *J. Mol. Biol.* 217: 373 (1991).

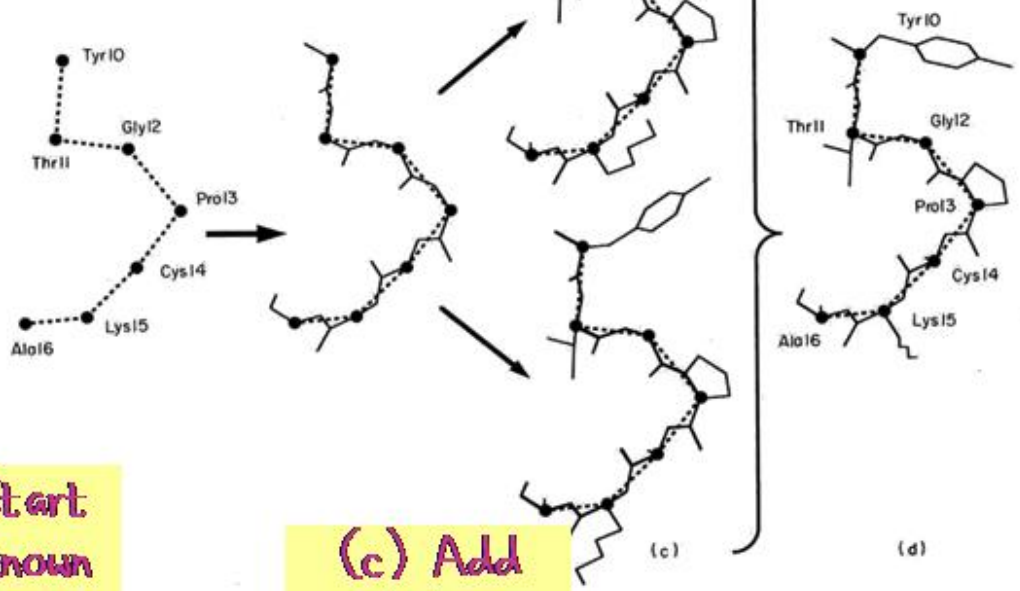
SEGMENT MATCH MODELING

(a) Start with known atoms in a short segment.

(b) Add mainchain atoms from a library.

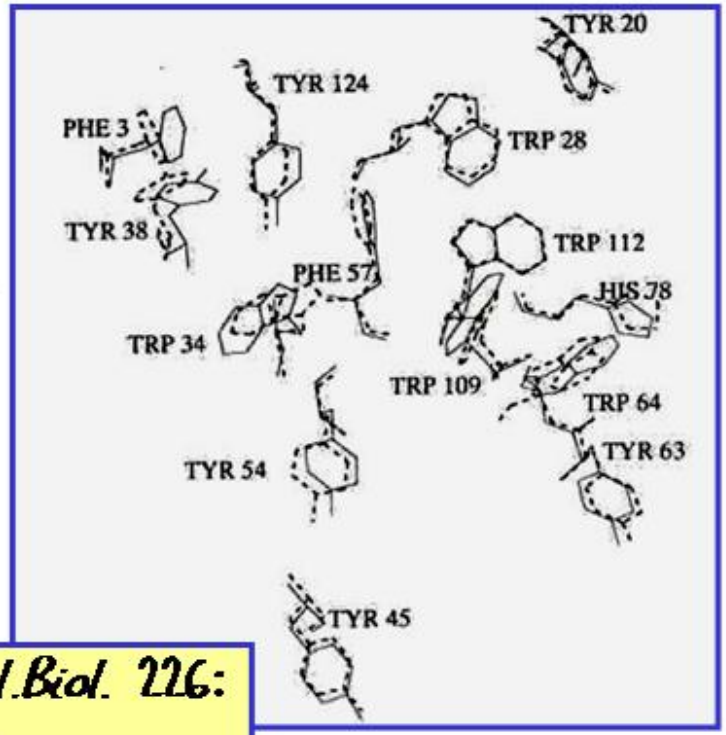
(c) Add sidechain atoms from a library.

(d) Repeat this ten time and average the models.



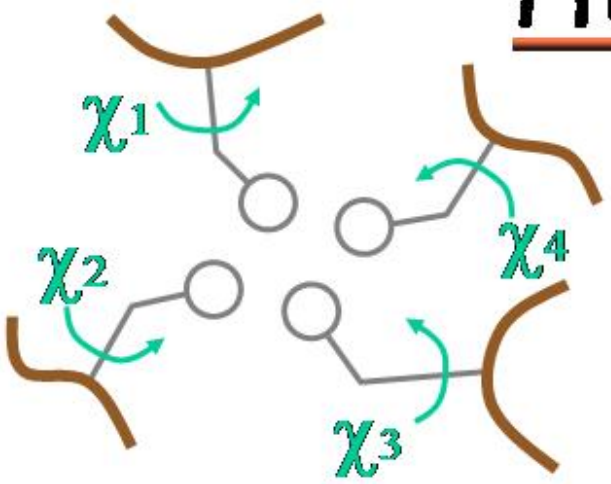
Test this building all the sidechains from the CA atoms.

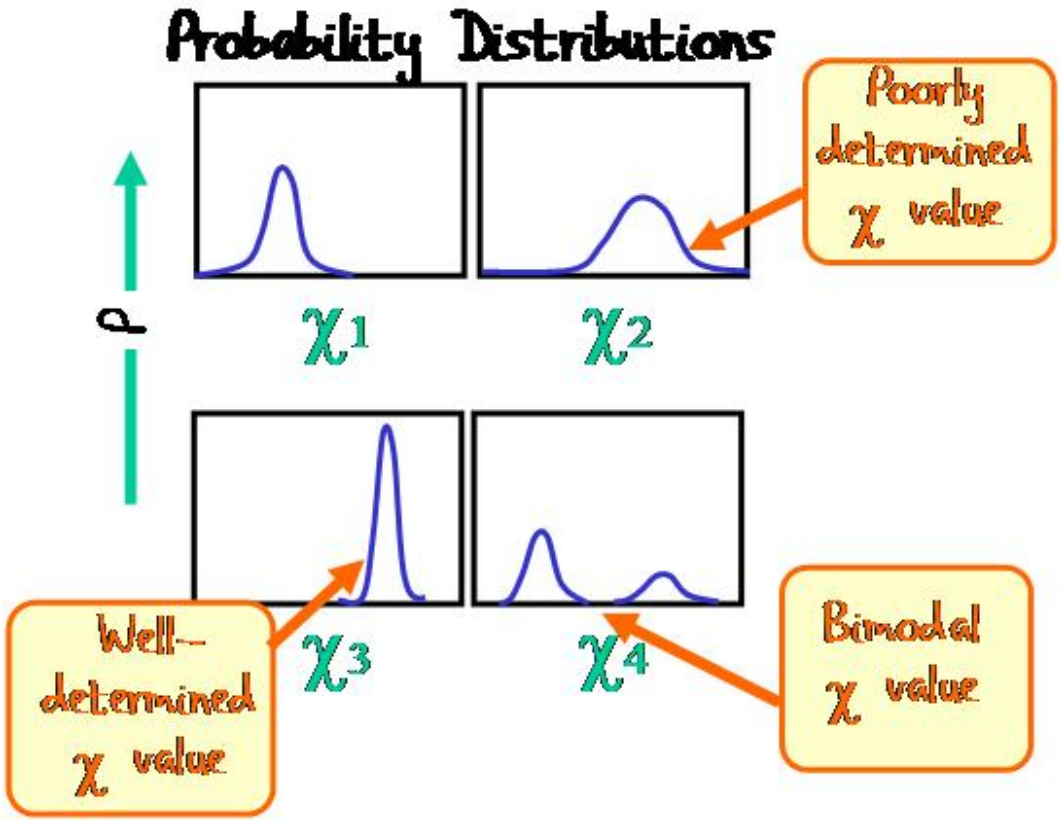
It works!



Levitt, J Mol Biol. 226: 507 (1992).

MEAN FIELD



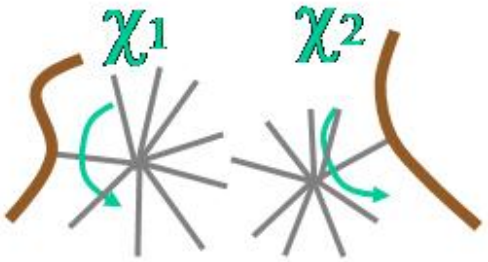


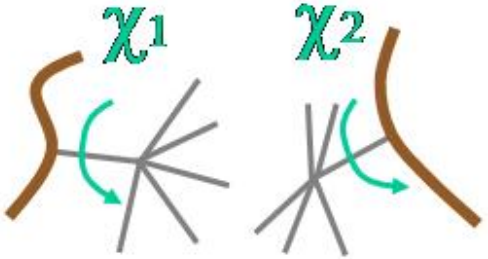
- Do not work with the χ torsion angles directly.
- Rather have a probability distribution for each variable.
- Let the sidechain interact and then lower the effective temperature to get a precise χ value.

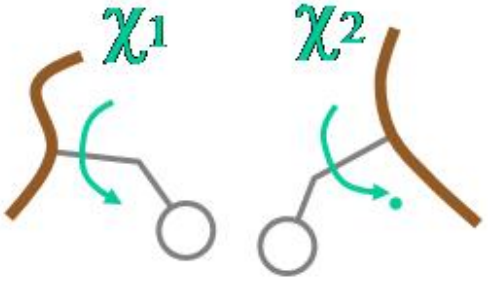
Lee, *J.Mol.Biol.* 236: 918 (1994).

MEAN FIELD

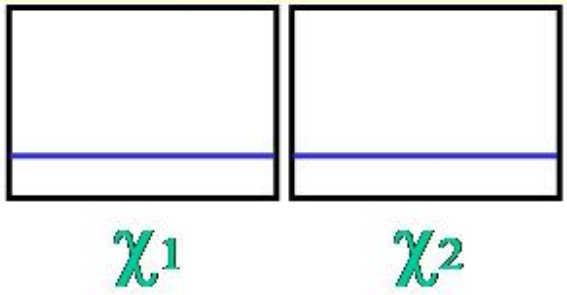
Probability Distributions

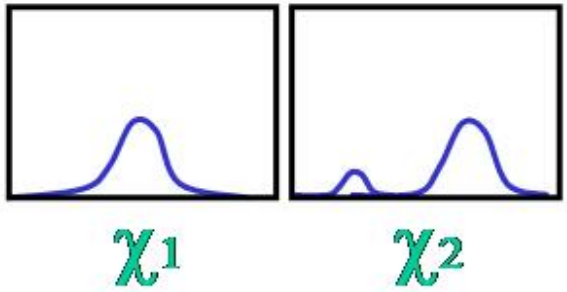


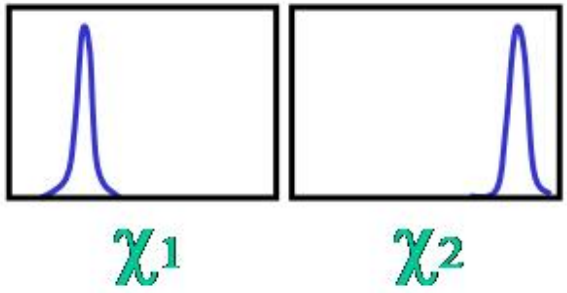




ρ

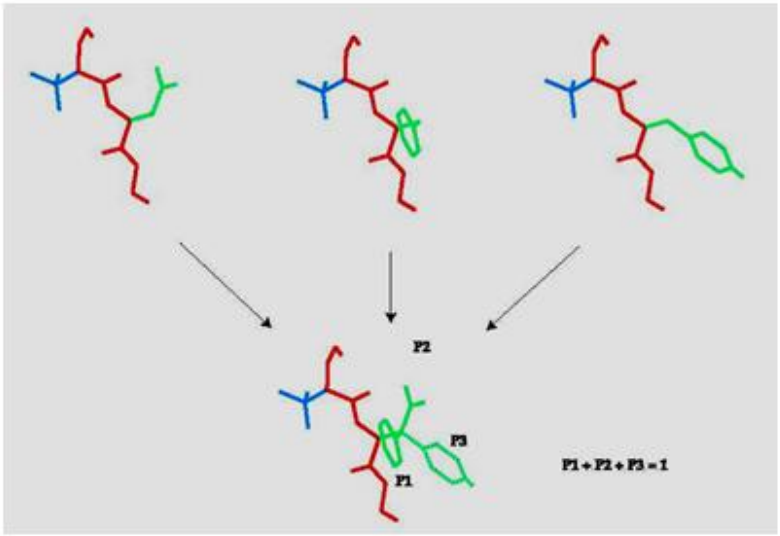






- At high Temperature, T, all arrangements are equally probable.
- As T drops, the distributions get some shape.
- At low T, we have convergence to well-defined χ values.

GENERALIZED SELF-CONSISTENT MEAN-FIELD



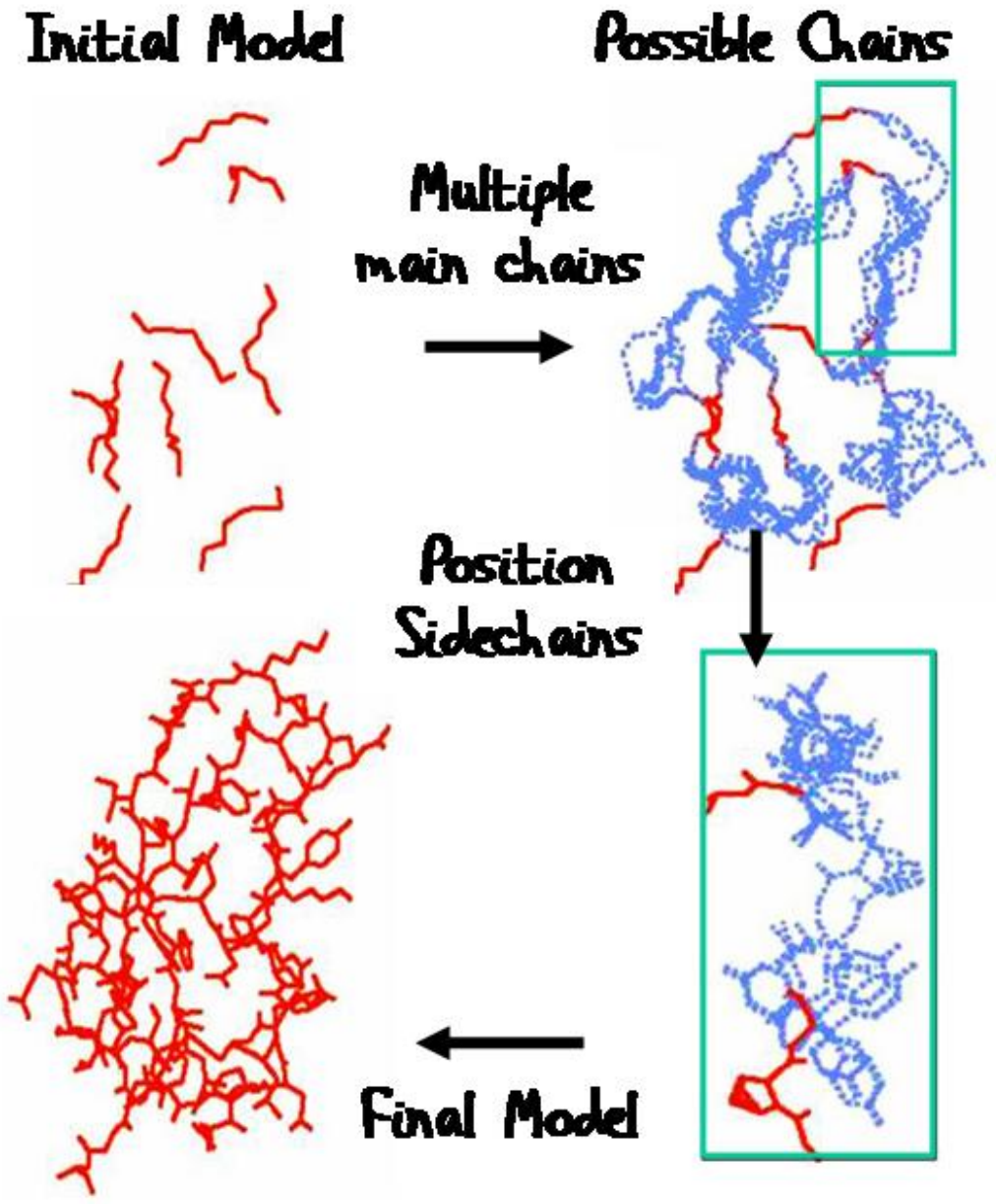
- Any part of the molecule can have multiple states with probabilities P(i,j).
- Use a Boltzmann-like formula to update the probabilities.
- The sum of P(i,j) is always 1.

$$E(i,k) = U(i,k) + U(i,k,Backbone) + \sum_{j=1, j\neq i}^{N} \sum_{l=1}^{Nrot(j)} P(j,l)U(i,k,j,l)$$

$$P_{new}(i,j) = \frac{\exp\left[-\frac{E(i,j)}{kT}\right]}{\sum_{l=1}^{Nrot(i)} \exp\left[-\frac{E(i,l)}{kT}\right]}$$

Koehl & Delarue, J Mol Biol, 239: 249 (1994).

GENERALIZED SELF-CONSISTENT MEAN-FIELD



- This scheme can work for different side chain conformations, for different main chain conformations and even for different side chain sequences.

- The key idea is that the mean energy at any stage can set the probabilities of any state at the next stage.

Koehl & Delarue, *Curr Opin Struct Biol.* 6: 222 (1996).

SOME OTHER METHODS

Dead End Elimination

Evaluate all possible arrangements that can possibly be part of the low energy packing.

Desmet et al. *Nature*, 356: 539 (1992).

Backbone Dependent Rotamers

SCWRL (Side Chains With a Rotamer Library)

Bower et al. *J. Mol. Biol.* 267: 1268 (1997).

Homology Modeling
Concept 8.6

WHAT IS HOMOLOGY MODELING

Sequence#1 YHWSGPHVVIMGRL

Sequence#2 YHLSGVIIVINGKL

Know Structure#1

?

What is Structure#2

- Rely on evolution's re-use of successful components.
- If Sequence#1 is similar to Sequence#2, then Structure#1 looks like Structure#2.

LOOP METHODS ARE VARIED

- Segment matching.

Levitt, J. Mol. Biol. 226: 507 (1992).

- Graph Theoretical approach.

Samudrala & Moult, J. Mol. Biol. 279: 287 (1998).

- Exhaustive enumeration approach.

Moult & James, Proteins, 1: 2146 (1986).

- Ab Initio approaches.

CASP has not given much guidance about works best for loops.

Not enough data? Need to look more carefully?

A MULTIPLE SEQUENCE ALIGNMENT

```
1cpq     ..ADTKEVLEAREAYFKSLGGSMKAMTGVA...K...AF...DAEAAK.VEAAKLEKILAT...DVAPLF..PAGTSSTDL.PG..QTEAKAAIWA.NMDDFG
1cpq     ..ADTKEVLEAREAYFKSLGGSMKAMTGVA...K...AF...DAEAAK.VEAAKLEKILAT...DVAPLF..PAGTSSTDL.PG..QTEAKAAIWA.NMDDFG
1cpr     ..ADTKEVLEAREAYFKSLGGSMKAMTGVA...K...AF...DAEAAK.VEAAKLEKILAT...DVAPLF..PAGTSSTDL.PG..QTEAKAAIWA.NMDDFG
1rcpA    ..ADTKEVLEAREAYFKSLGGSMKAMTGVA...K...AF...DAEAAK.VEAAKLEKILAT...DVAPLF..PAGTSSTDL.PG..QTEAKAAIWA.NMDDFG
1nbbA    ..ADTKEVLEAREAYFKSLGGSMKAMTGVA...K...AF...DAEAAK.VEAAKLEKILAT...DVAPLF..PAGTSSTDL.PG..QTEAKAAIWA.NMDDFG
1rcpB    ..ADTKEVLEAREAYFKSLGGSMKAMTGVA...K...AF...DAEAAK.VEAAKLEKILAT...DVAPLF..PAGTSSTDL.PG..QTEAKAAIWA.NMDDFG
1nbbB    ..ADTKEVLEAREAYFKSLGGSMKAMTGVA...K...AF...DAEAAK.VEAAKLEKILAT...DVAPLF..PAGTSSTDL.PG..QTEAKAAIWA.NMDDFG
1bbhA    AGLSPEEQIETRQAGYEFMGWNMGKIKANL...E..GEY...NAAQVE.AAANVIAAIANS...GMGALY..GPGTD..KN.VGDVKTRVKPEFFQ.NMEDVG
1cgo     XFAKPEDAVKYRQSALTLMASHFGRMTPVV...KGQAPY...DAAQIK.ANVEVLKTLTAL....PWAAF..GPGTEG.........GDARPEIWS.DAASFK
1a7vA    ....QTDVIAQRKAILKQMGEATKPIAAML...K..GEA...KF.DQA.VVQKSLAAIADD.SKKLPALF..PADSK...T.GG..DTAALPKIWE.DKAKFD
1a7vB    ....QTDVIAQRKAILKQMGEATKPIAAML...K..GEA...KF.DQA.VVQKSLAAIADD.SKKLPALF..PADSK...T.GG..DTAALPKIWE.DKAKFD
2ccyA    .QSKPEDLLKLRQGLMQTLKSQWVPIAGFAAGKA...DL...P.ADAA.QRAENMAMVAKL....APIGW..AKGTE..AL.PN...GETKPEAFGSKSAEFL
1ekyA    ..ADTKEVLEAREAYFKSLGGSMKAMTGVA...K...AF...DAEAAK.VEAAKLEKILAT...DVAPLF..PAGTSSTDLPGQ..TEAKAAIW..ANMDDFG
256bA    .........ADLEDNMETLNDNLKVIEKA.........D...NAAQVK.DALTKMRAAALD..AQKATPP.KLEDK..............SPDSP.EMKDFR
1nfn     ...........GQRWELALGRFWDYLRWVQ...T...LSEQVQEELLSSQVTQELRALMDETMKELKAYKSELEEQ...................R.LSKELQ
```

The same methods used to align a pair of sequences can be used to align many sequences.
This is either very slow or approximate.

FSSP STRUCTURAL ALIGNMENTS

FSSP: structural neighbors of 1bpi

Same family as 5pti

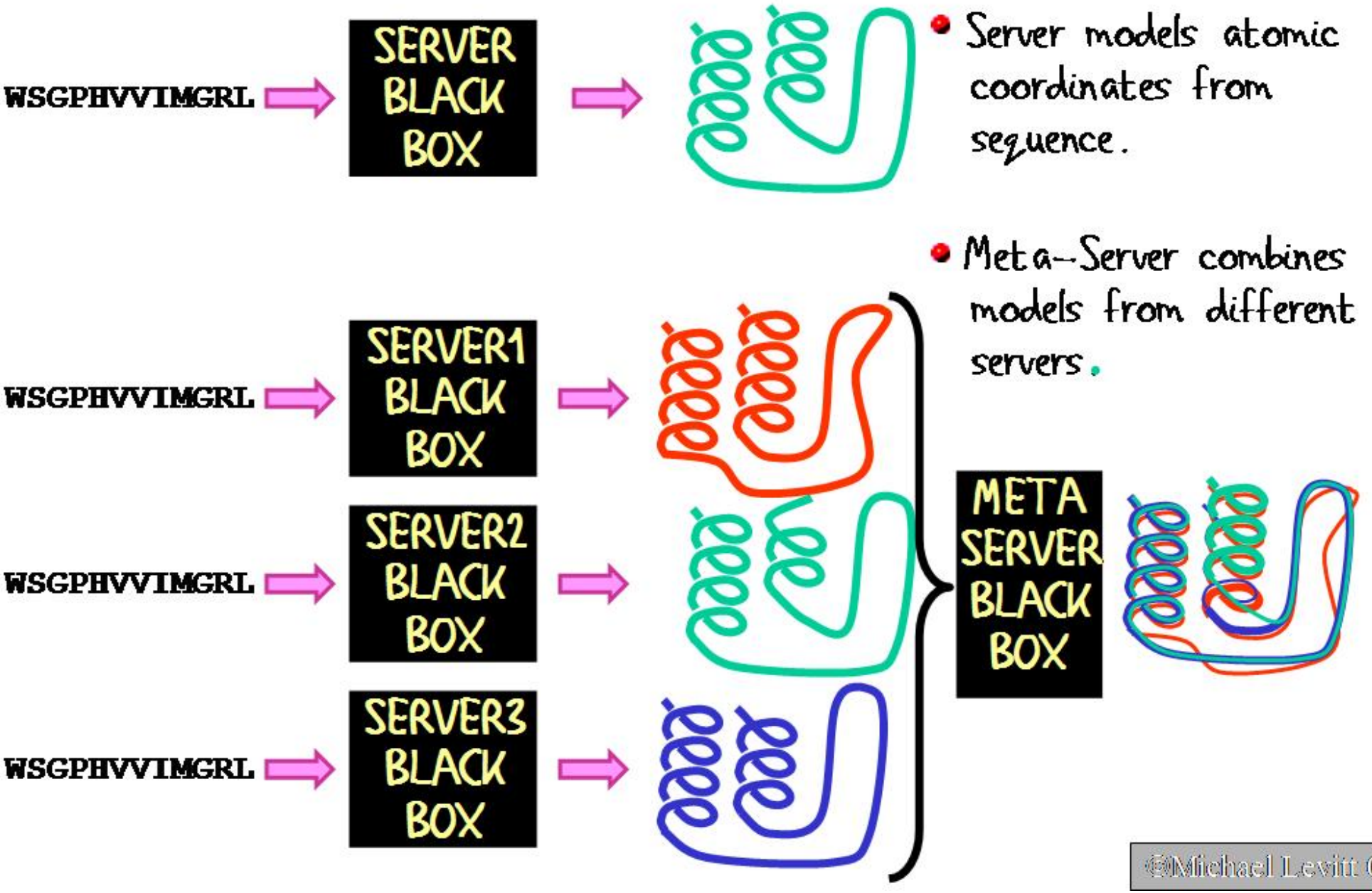
Please cite: L. Holm and C. Sander (1996) Science 273(5275):595-60.
Structural alignment by Dali
Notation: Uppercase: structurally equivalent with 1bpi ; lowercase: structurally non-equivalent with 1bpi

```
1        1bpi     ..RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA.
2        1bpi     ..RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA.
3        1bzxI    ..RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA.
4        1fakI    ..APDFCLEPPYDGPCRALHLRYFYNAKAGLCQTFYYGGCLAKRNNFESAEDCMRTC....
5        1bunB    rkRHPDCDKPPDTKICQTVVRAFYYKPSAKRCVQFRYGGCNGNGNHFKSDHLCRCECLEYr
6        1bf0     wqPPWYCKEPVRIGSCKKQFSSFYFKWTAKKCLPFLFSGCGGNANRFQTIGECRKKCLGK.
```

- Cys are conserved, also Asn 43 (special) and aromatic residues even in distant neighbors.
- Gray residues are not structurally matched.

http://www2.embl-ebi.ac.uk/dali/fssp/fssp.html

SERVERS AND META SERVERS



COMPARATIVE MODELING: SERVERS & CREATORS

Comparative Modeling Targets

RANK	GROUP	N	ZSUM
4	STERNBERG	6	5.84
25	3D-PSSM	4	3.21
12	GODZIK	5	4.28
31	FFAS	2	2.59
36	PDB-BLAST	3	2.32
17	JONES	5	3.84
11	MGENTHREADER	5	4.36
24	GENTHREADER	4	3.29
19	FISCHER	5	3.63
18	INBGU	5	3.77
26	KARPLUS	4	3.13
22	SAM-T99	4	3.44
28	BLUNDELL	3	2.83
23	FUGUE-CAM	4	3.43

- At CASP4, the good servers generally come from people who do well.
- Sometimes a server does better than its creator.

META SERVERS ARE WINNERS

- Meta-Servers help their groups win at CASP5

RANK	GROUP	N	ZSCORE
1	Bujnicki-Janusz	39	47.1
3	GeneSilico	40	42.8
55	GENESILICO.PL	34	7.4
2	Ginalski	42	46.5
5	Bionfo.pl	40	31.2
73	BIOINFO.PL-BASIC	17	4.4
7	Fischer	39	27.2
22	SHGU	26	17.3
71	3DSN-INBGU	30	4.6

- Meta-servers do better than individual servers.
- Experts do better than meta-servers if they have the meta-server output.