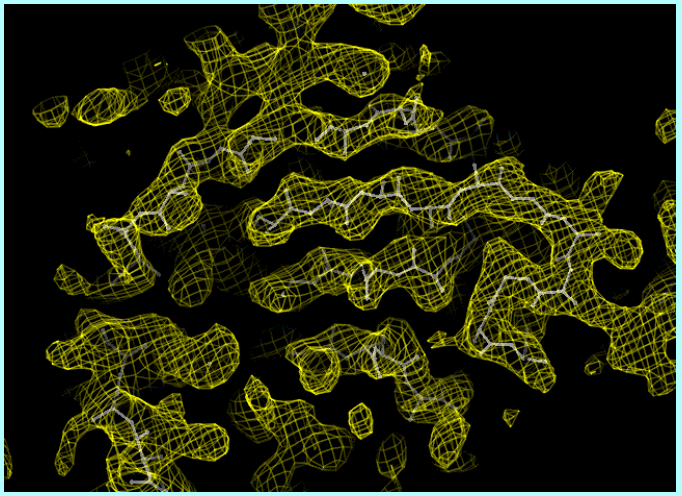
Phenix

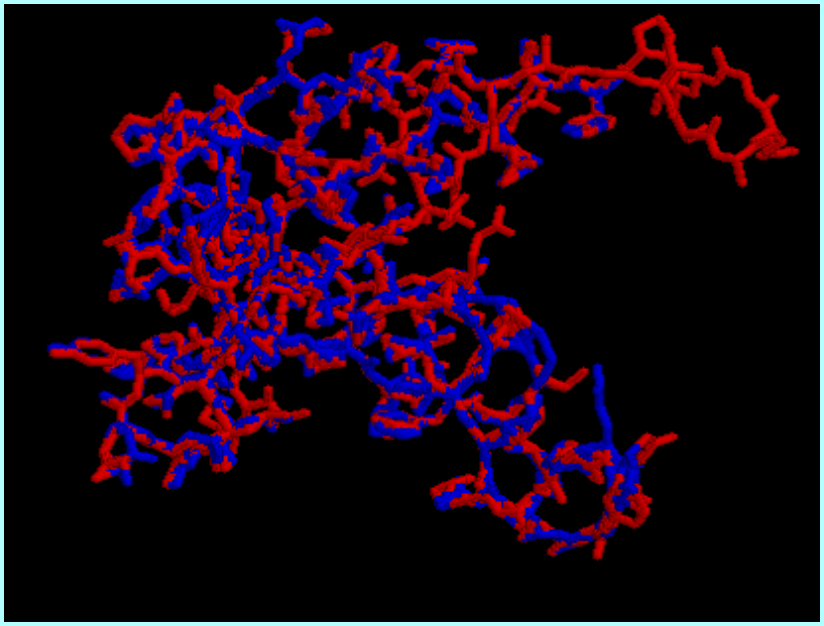
RESOLVE model-building

Tom Terwilliger
Los Alamos National Laboratory

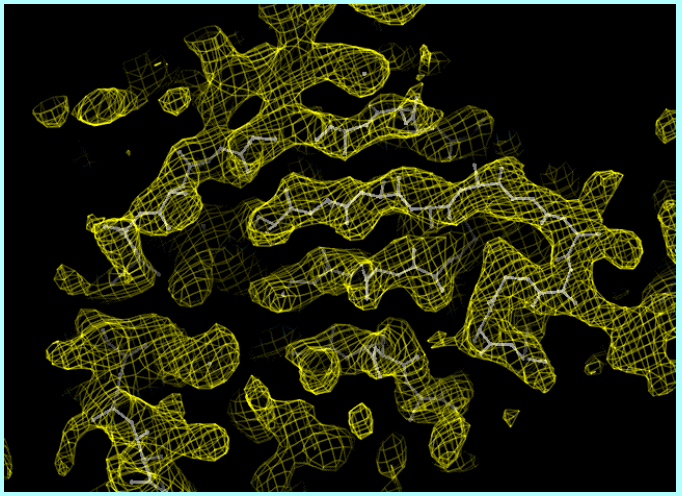
RESOLVE model-building at moderate resolution



- FFT-based identification of helices and strands
- Extension with tripeptide libraries
- Probabilistic sequence alignment
- Automatic molecular assembly

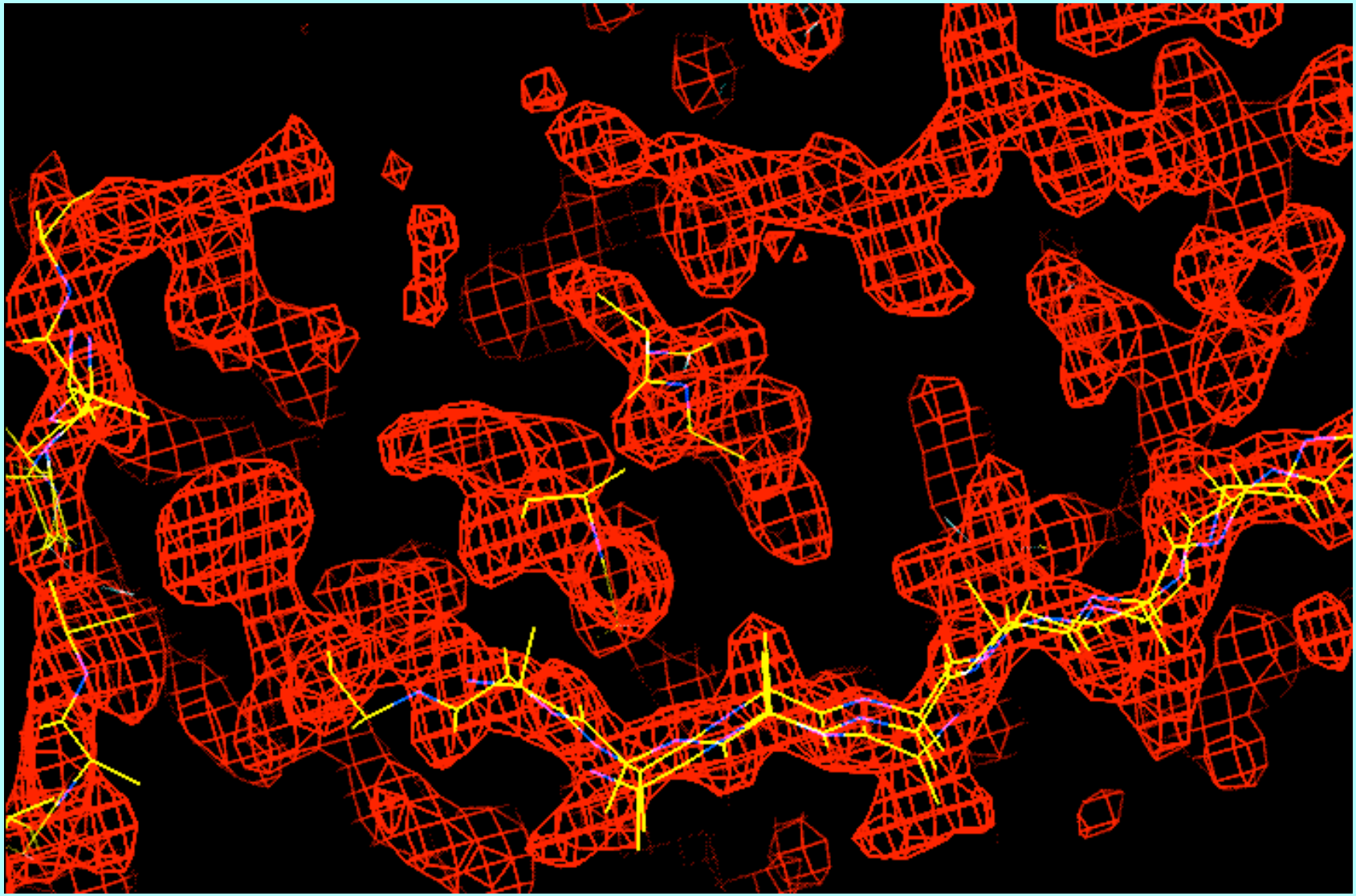


Placement of helical and extended templates

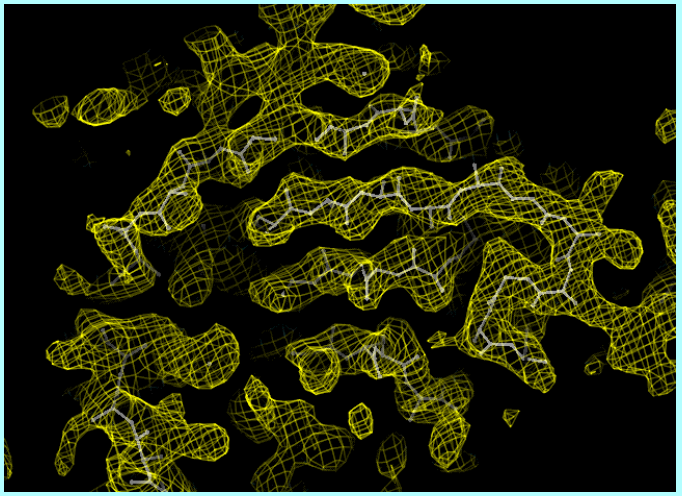


- Identify locations with FFT-based convolution search
- Maximize CC of template with map
- Superimpose each fragment in corresponding library (helix,sheet) on template
- Identify longest segment in good density, score = <density>*sqrt(Natoms)

Initial model-building – strand fragments



Chain extension by placement of tripeptide fragments



- Look-ahead scoring: find fragment that can itself be optimally extended

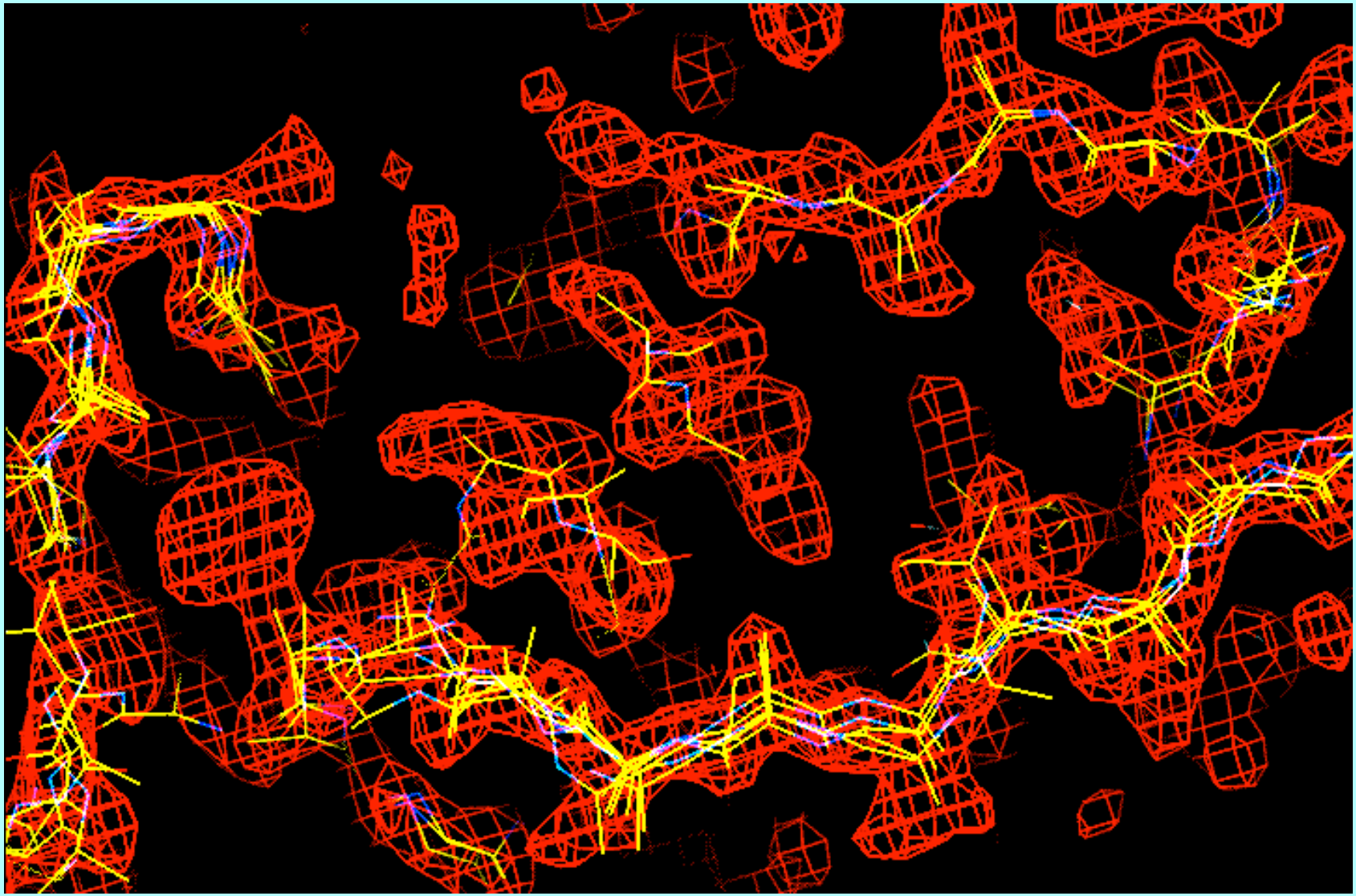
- C-terminal extension. Start at C-terminus of protein

- Each of 10000 fragments: superimpose CA C O on same atoms of last residue in chain (extending by 2 residues): pick best 10

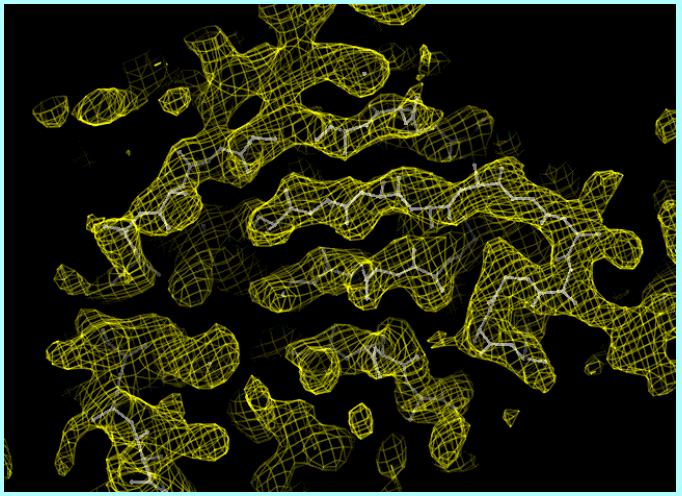
- Each of best 10: extend again by 2 residues and pick best 1; score for 2-residue extension= best <density> for 4-residue extension based on this 2-residue extension

- N-terminal: same, but going in opposite direction

Chain extension
(result: many overlapping fragments)

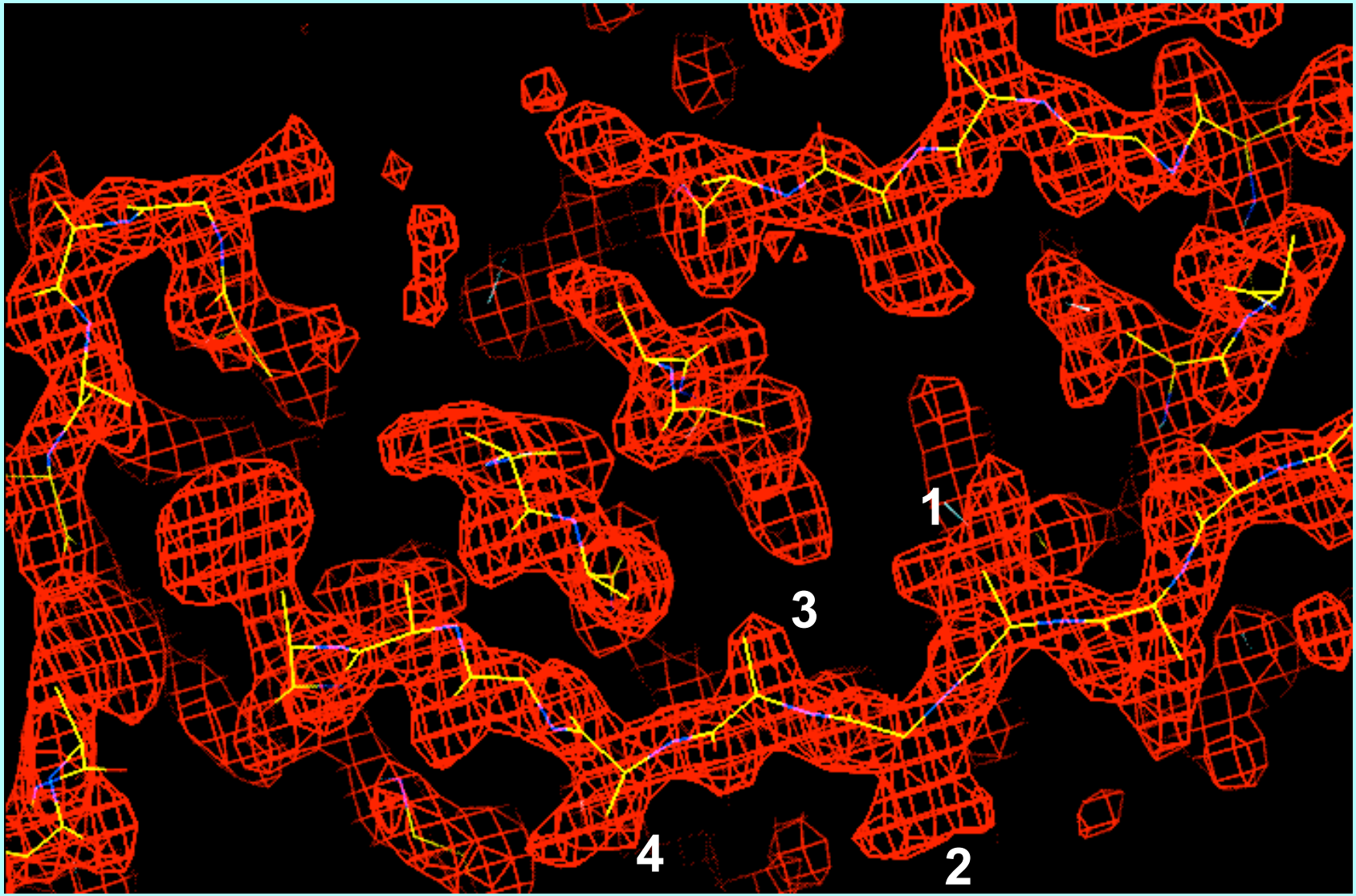


Assembly of main-chain

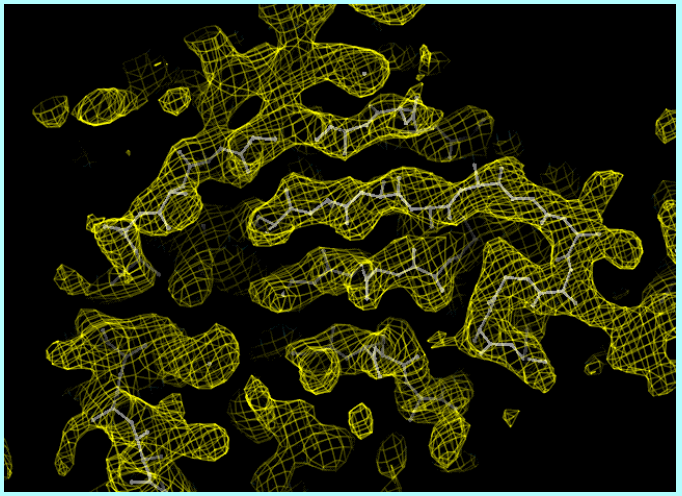


- Choose highest-scoring fragment
- Test all overlapping fragments as possible extensions
- Choose one that maximizes score when put together with current fragment
- When current fragment cannot be extended: remove all overlapping fragments, choose best remaining one, and repeat

Main-chain as a series of fragments
(choosing the best fragment at each location)



Side-chain rotamer templates

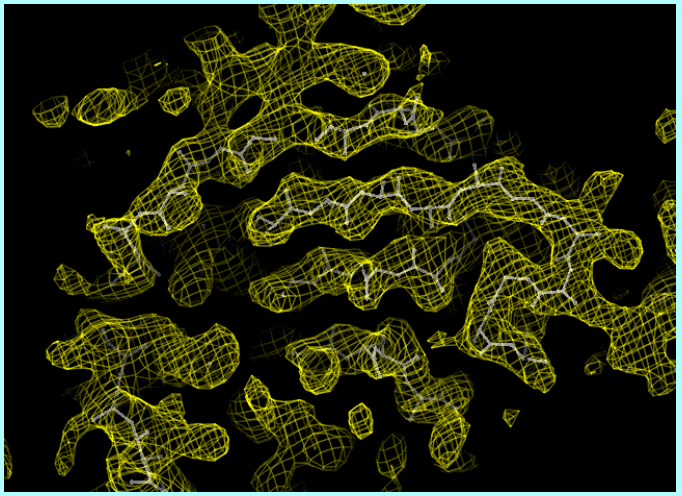


- Define side-chain orientation based on N CA C of main-chain
- Up to 40 rotamers per side chain

- Create template from average calculated electron density based on all occurrences of rotamer in 637 unique proteins

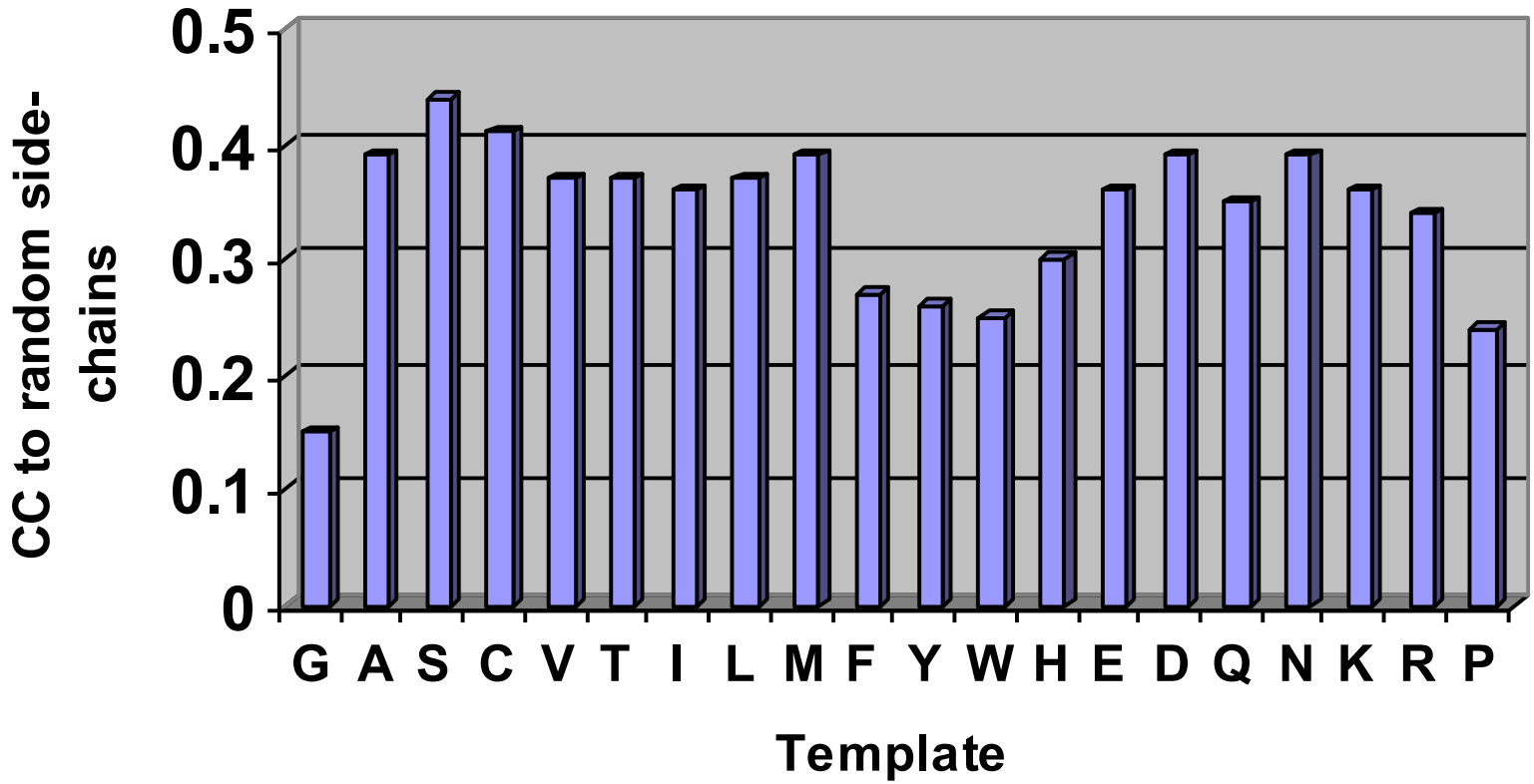
- Total of 400 side-chain templates

Scoring side-chain templates at each position



- Identify side-chain orientation from N CA C of main-chain
- Get CC of template with density -> Z-score
- (Compare CC with mean, SD of all side chain density with this template)
- P(this side-chain/rotamer is correct)= Po(this side-chain/rotamer)*P(Z)

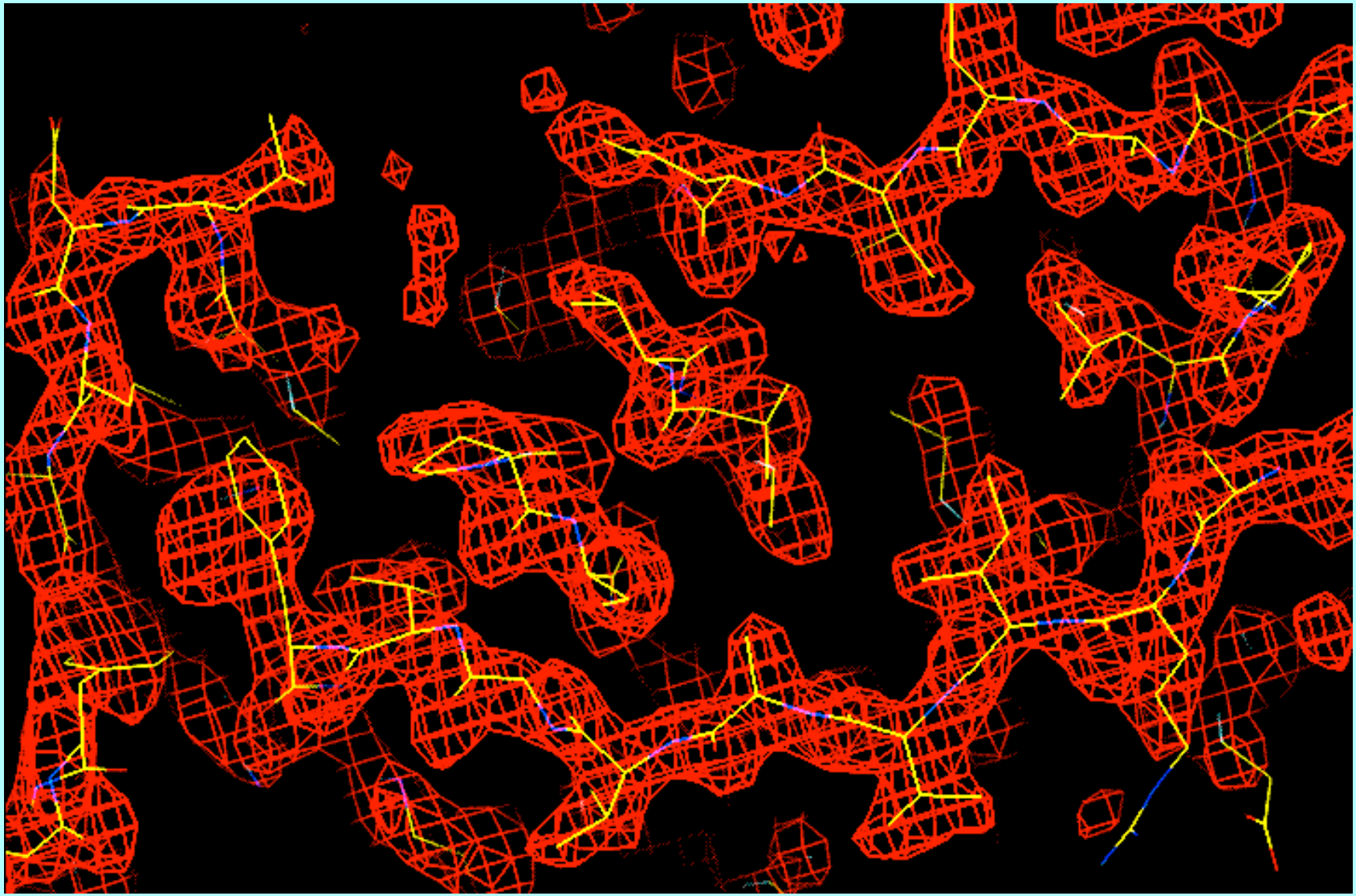
Evaluating which side-chain template is best matched by a pattern of density:
A good match to a glycine means more than a good match to an alanine



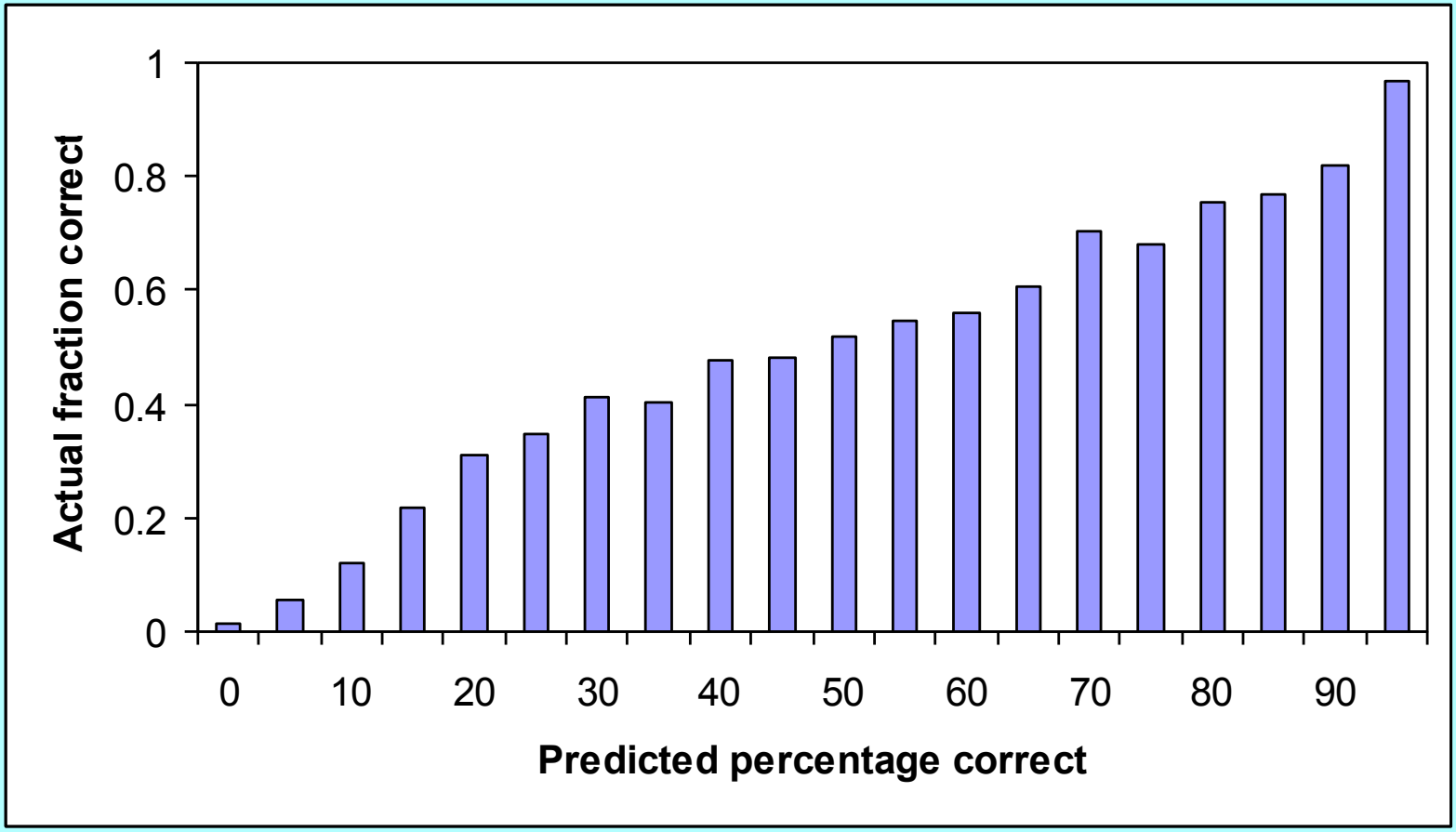
Side-chain template matching to identify sequence alignment to map (IF5A data)
Relative probability for each amino acid at each position
(Correct amino acids in bold)

#	G	A	S	V	I	L	M	C	F	Y	K	R	W	H	E	D	Q	N	P	T
1	6	5	4	18	**18**	6	1	1	1	2	6	2	2	1	9	6	1	0	1	4
2	4	11	14	**37**	5	2	0	2	0	0	2	3	0	0	1	2	0	0	0	6
3	11	**23**	5	12	5	3	2	0	1	3	7	3	1	0	5	3	2	0	2	2
4	7	9	6	**16**	8	5	2	0	1	3	8	4	1	0	7	6	2	0	3	4
5	**31**	7	3	7	4	2	1	0	1	3	5	4	1	0	6	2	2	0	11	1
6	1	3	3	**41**	14	8	0	0	0	0	2	1	0	0	2	4	0	0	1	9
7	0	0	0	0	0	0	0	0	**15**	**63**	1	0	**17**	1	0	0	0	0	0	0
8	2	3	6	23	10	6	2	1	0	1	4	3	0	0	5	**16**	1	0	1	6
9	**96**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

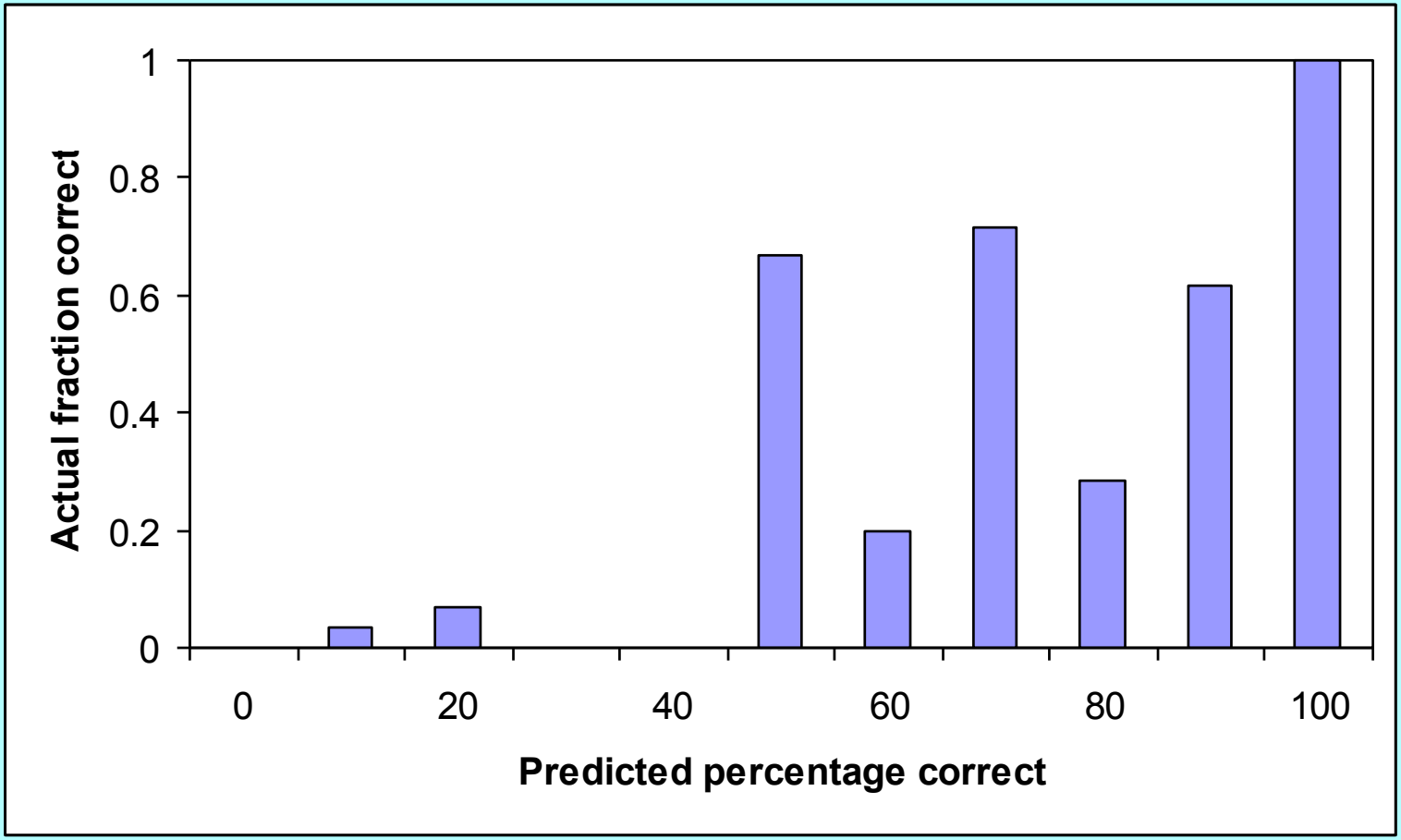
Addition of side-chains to fixed main-chain positions



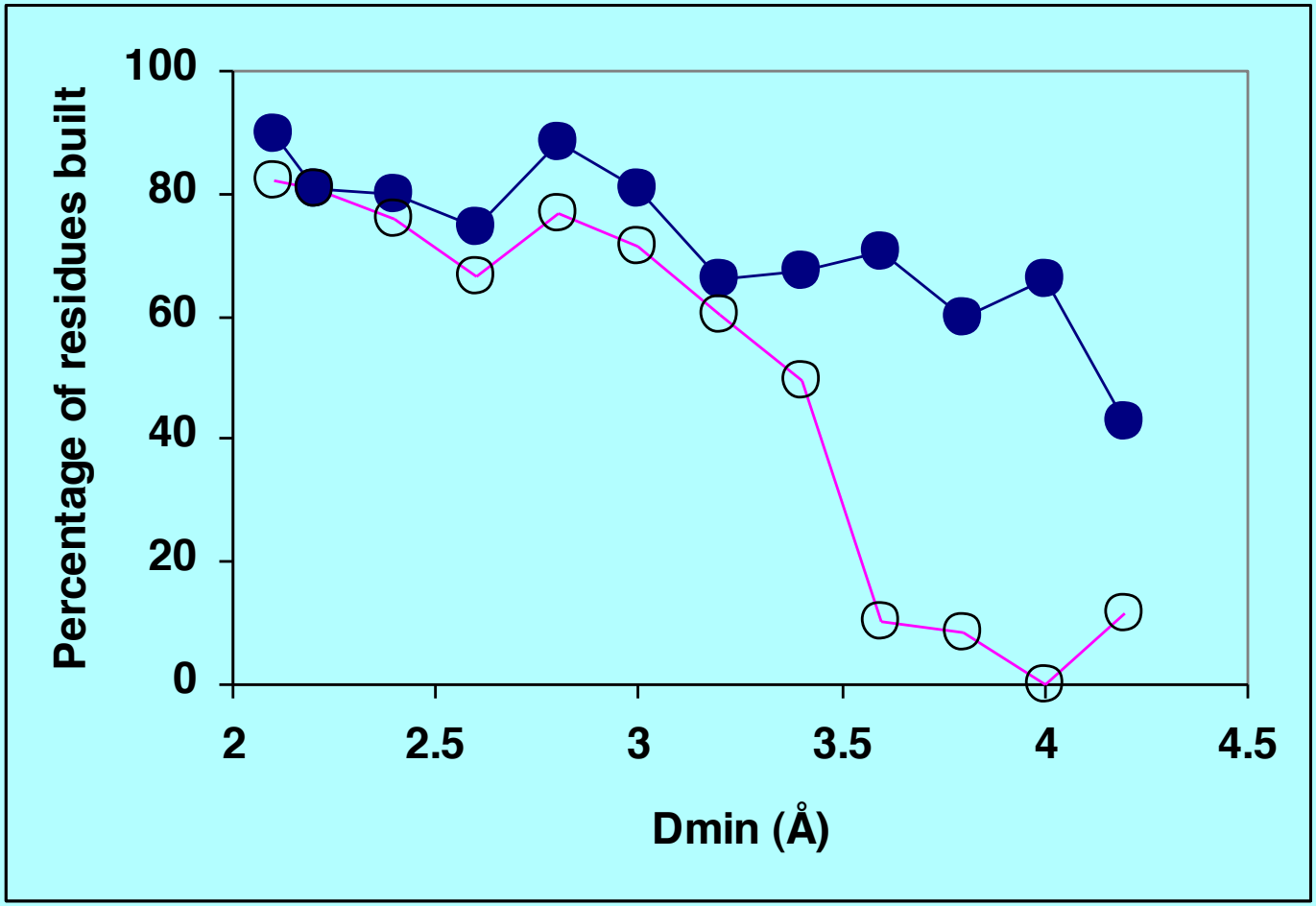
Accuracy of side-chain identification probabilities



Accuracy of sequence alignment probabilities



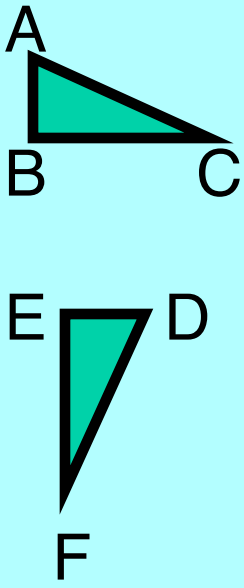
Model-building vs resolution for nearly-perfect data (IF5A)



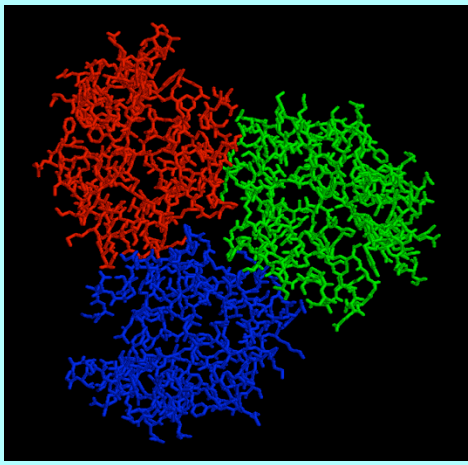
Automated NCS identification with RESOLVE

- Expand heavy-atom sites within radius R of origin
- Make list of all pairs of sites, sorted by distance between sites d

- Choose any 3 HA sites – a triangle ABC
- Find all other sets of 3 HA sites that form the same triangle
  - If some exist (DEF) -> this might correspond to NCS
  - If none…try another set of 3 HA sites



- Testing NCS: Sites ABC match sites DEF
- Does density near ABC match (after rotation/translation) density near DEF?



Automated NCS identification with RESOLVE

Structure	Number of sites found by SOLVE	NCS	NCS (found from heavy-atom sites)	NCS (electron-density map)
NDP Kinase	9	3-fold	3-fold	3-fold
Hypothetical	16	2-fold	2-fold	2-fold
Red Fluorescent Protein	26	4 copies	4 copies	4 copies
AEP Transaminase	66	6 copies	6 copies	6 copies
Formate dehydrogenase	12	2-fold	2-fold*	2-fold
Gene 5 protein	2	None	None	None
Armadillo repeat from β-catenin	15	None	2 copies	None
Dehalogenase	13	None	3 copies	None
Initiation Factor 5A	4	None	None	None

Molecular assembly in RESOLVE

List all chains assigned to sequence (anywhere in space)

A possible arrangement consists of:

- Each chain assigned to a molecule
- Each chain assigned to a symmetry-related position

Score a possible arrangement based on:

- Plausibility of gap distances between position of C of residue i and N of residue j
- RMS distance of chains from molecular center
- RMSD of NCS symmetry for corresponding atoms

- Try a reasonable starting arrangement (each chain assigned to the center of an NCS copy)
- Adjust by moving chains and groups of chains randomly from one symmetry-related position to another. Choose based on score.

Molecular assembly in RESOLVE

Summary of molecular assembly results (NDP-kinase)

NCS copies: 3

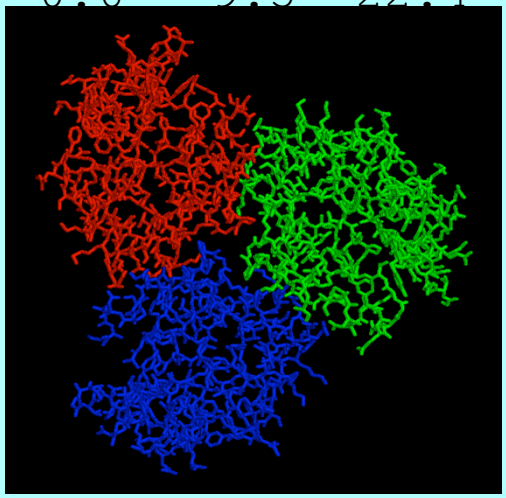
Molecule: 1 Chain: 1 Score for molecular location: 0.83

Frag	Start	End	N	Overlap	Link Length	Mol Radius	NCS RMSD	NCS <N>	Score
1	17	64	48	0	6.6	4.5	0.7	31.0	51.0
2	69	74	6	0	24.5	19.6	0.5	3.0	3.7
3	115	137	23	0	14.4	5.2	0.8	20.5	22.7
4	166	186	21	0		5.2	0.6	9.5	22.4

Residues placed for this molecule: 98

Total residues placed: 309 of 588 or 52%
Residues built without side chains: 65
Total residues built: 374 or 63%

Total score for this arrangement: 314.4



Initial automated structure solution, density modification, NCS-identification, and model-building

Structure	Res. (Å)	% of main-chain built	% of side chains built
Granulocyte stimulating factor (Rozwarski et al., 1996)	3.5	50%	0%
β-catenin (Huber et al., 1997)	2.7	81%	62%
Gene 5 protein	2.6	61%	11%
NDP Kinase (Pédelacq et al, 2002)	2.6	56%	37%
Hypothetical (P. aerophilum ORF, NCBI accession number AAL64711)	2.6	79%	75%
2-Aminoethylphosphonate (AEP) Transaminase (Chen et al., 2000)	2.6	85%	81%
Red Fluorescent Protein (Yarbrough et al, 2001)	2.5	88%	88%
Initiation factor 5A (Peat et al., 1998)	2.1	84%	84%

The PHENIX Project

Lawrence Berkeley Laboratory

Paul Adams, Ralf Grosse-Kunstleve, Pavel Afonine, Nat Echols, Nigel Moriarty, Jeff Headd, Nicholas Sauter, Peter Zwart

Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung

Cambridge University

Randy Read, Airlie McCoy, Gabor Bunkoczi, Rob Oeffner

Duke University

Jane & David Richardson, Vincent Chen, Chris Williams, Bryan Arendall, Laura Murray

An NIH/NIGMS funded Program Project