Automation of structure determination

Use of scoring procedures to assist in decision-making

Simple procedures for automation choosing the current best path at each decision-point

What is automation?

Procedures (things to do)

Control (deciding what to do)

What is automation?

Automation as a set of linked procedures

Each procedure has clearly-defined…

Inputs
Methods to apply to inputs
Outputs

What is automation?

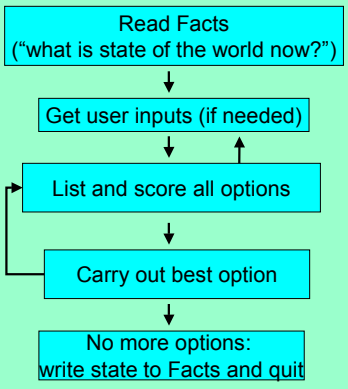
Automation as a set of linked procedures

Control steps have clearly-defined…

Possible decisions to make
Information required to make decisions
Next step(s) to take based on decisions

(Including…what to do if things go wrong)

Simple automation using a scoring scheme for decision-making (as implemented in PHENIX wizards)



Modular PYTHON routines in AutoSol

Create function to do something ("score_all_solutions")

↓

Identify all required antecedents in a list
required_antecedents=['solution_list','scoring_table',…]

↓

Create a method to score the utility of carrying out the procedure "eval_score_all_solutions"

Deciding which solutions to follow up: "COVERAGE"

User sets "coverage" = "the desired confidence of keeping the best solution in consideration"

↓

Score solutions, with confidence intervals

↓

Follow up on any solution that could really be the top one
(i.,e, top solution Z-score = 14, next solution Z=13, "coverage"=95%
-> carry through with BOTH solutions because either could be the best)

Automation of structure determination

↓

Scale data

↓

HYSS heavy-atom search

↓

Score and rank solutions

↓

Phase and quick density modification
Get sites with difference Fourier

↓

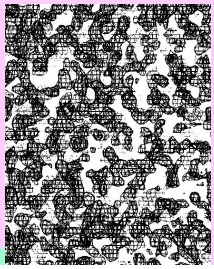
Coverage satisfactory:
go on to full density modification and iterative model-building

PHENIX AutoSol wizard standard sequence

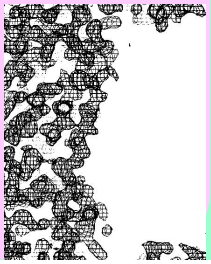
Why we need good measures of the quality of an electron-density map:

Which solution is best?

Are we on the right track?



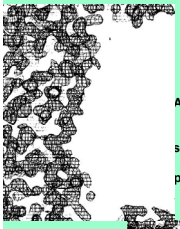
If map is good:
It is easy
(which is correct?)



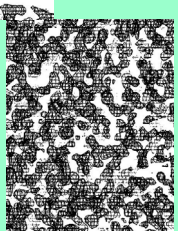
Evaluating electron density maps: Methods examining the map itself

Basis	*Good map*	*Random map*
Skew of density (Podjarny, 1977)	Highly skewed (very positive at positions of atoms, zero elsewhere)	Gaussian histogram
SD of local rms densities (Terwilliger, 1999)	Solvent and protein regions have very different rms densities	Map is uniformly noisy
Connectivity of regions of high density (Baker, Krukowski, & Agard, 1993)	A few connected regions can trace entire molecule	Many very short connected regions
Presence of tubes of density or helices/strands or local patterns in map (Colovos, Toth & Yeates, 2001; Terwilliger, 2004)	CC of map with a filtered version is high	CC with filtered version is low

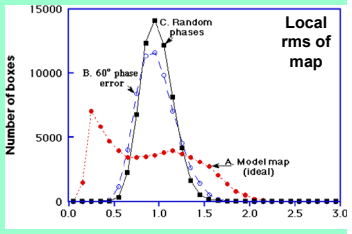
Scoring: does the native Fourier look like a protein?

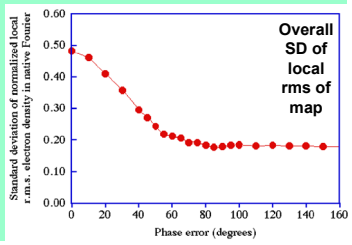


A good map: clear solvent vs protein



A poor map: it looks the same all over





Evaluating electron density maps: Methods examining the map itself

Basis	*Good map*	*Random map*
Skew of density (Podjarny, 1977)	Highly skewed (very positive at positions of atoms, zero elsewhere)	Gaussian histogram
SD of local rms densities (Terwilliger, 1999)	Solvent and protein regions have very different rms densities	Map is uniformly noisy
Connectivity of regions of high density (Baker, Krukowski, & Agard, 1993)	A few connected regions can trace entire molecule	Many very short connected regions
Presence of tubes of density or helices/strands or local patterns in map (Colovos, Toth & Yeates, 2001; Terwilliger, 2004)	CC of map with a filtered version is high	CC with filtered version is low

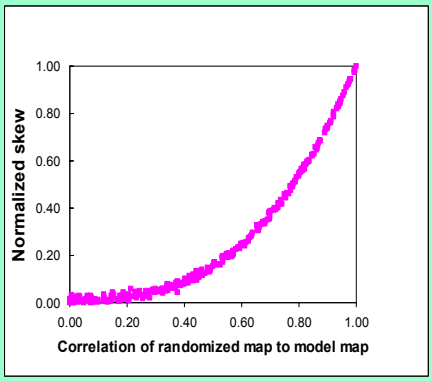
Evaluating electron density maps

Methods based on density-modification and R-factors

Basis	*Good map*	*Random map*
R-factor in 1st cycle of density modification (Cowtan, 1996)	Low R-factor	High R-factor
Correlation of map made with map-probability phases with original map (Terwilliger, 2001) (map-probability from solvent flattening or from truncation at high density level)	High correlation	Lower correlation

Skew of electron density in maps of varying quality

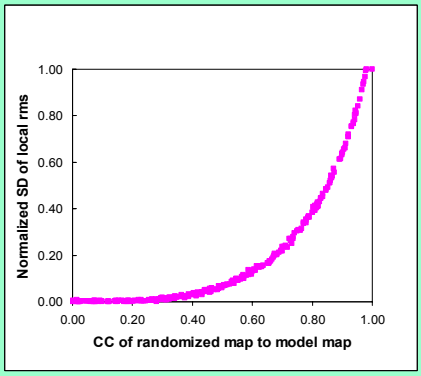
IF5A (*P. aerophilum,* 60% solvent; randomized maps)



(High electron density at positions of atoms; near zero everywhere else => high skew for good map)

SD of local rms of electron density in maps of varying quality

IF5A (*P. aerophilum,* 60% solvent; randomized maps)

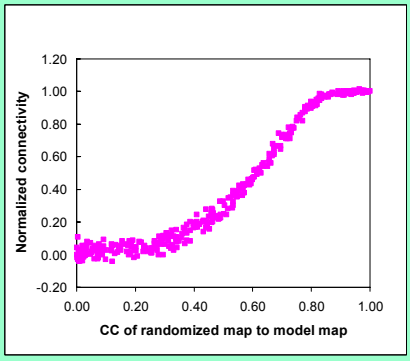


(Large rms in protein region; low in solvent => high SD for good map)

Connectivity of maps of varying quality

IF5A (*P. aerophilum,* 60% solvent; randomized maps;

Number of contiguous regions required to enclose top 5% of density)



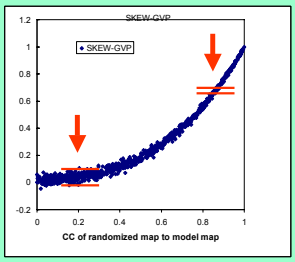
(Most of high density is connected in a good map)

Bayesian estimation of map quality from skew measurement on map

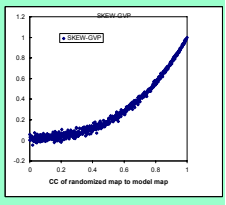
Start with database of randomized model data:

What values of skew do I measure if the actual map correlation is CC?

CC -> P(skew$_{obs}$ |CC)



Bayesian estimation of map quality from skew measurement on map



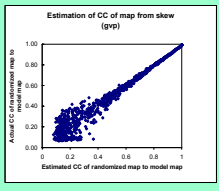
Given measurement of skew : estimate CC…

For each possible value of CC:

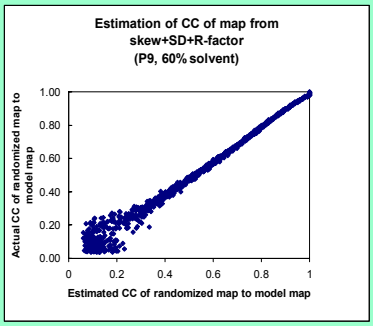
"probability that CC is correct is proportional to probability of measuring skew$_{obs}$ given this CC"

P(CC) = α P(skew$_{obs}$ |CC)

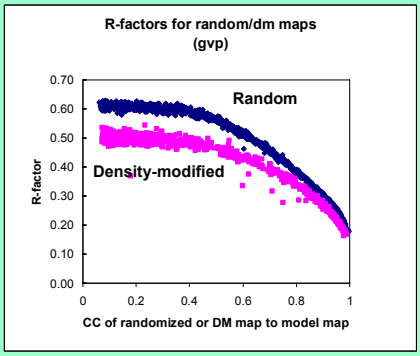
Combine all independent sources of information



Bayesian estimation of map quality using Skew, SD of local rms density, R-factor



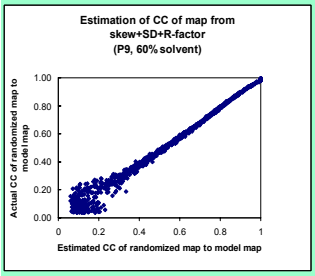
R-factors for density-modified maps are systematically lower than those of randomized maps of same quality



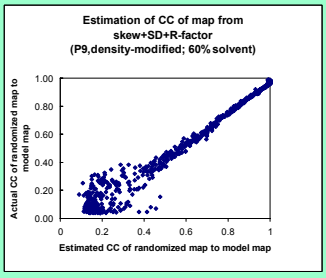
Bayesian estimation of map quality

Estimates for randomized maps are much better than those of density-modified maps

Randomized maps

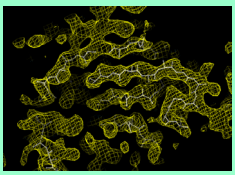


Density-modified maps

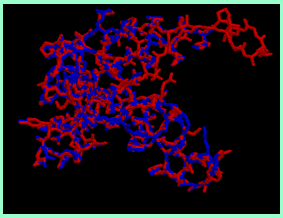


Model-building at moderate resolution using scoring methods for decision-making

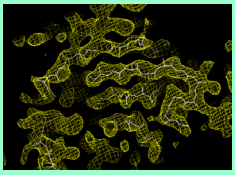
(following ideas of T.A. Jones, Cowtan, Oldfield, McRee, Levitt, Perrakis, Lamzin)



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

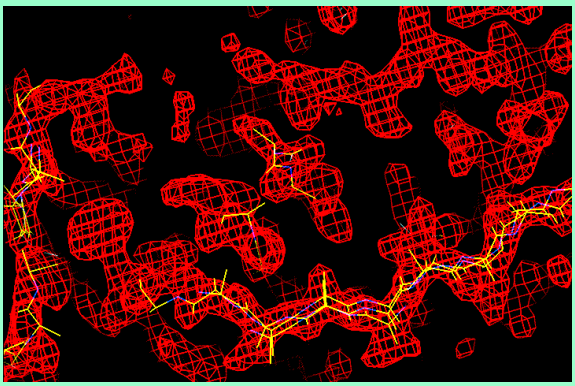


Placement of helical and extended templates

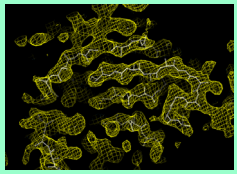


- ***Identify locations with FFT-based convolution search***
- ***Maximize CC of template with map***
- ***Superimpose each fragment in corresponding library (helix,sheet) on template***
- ***Identify longest segment in good density, score = <density>*sqrt(Natoms)***

Initial model-building – strand fragments

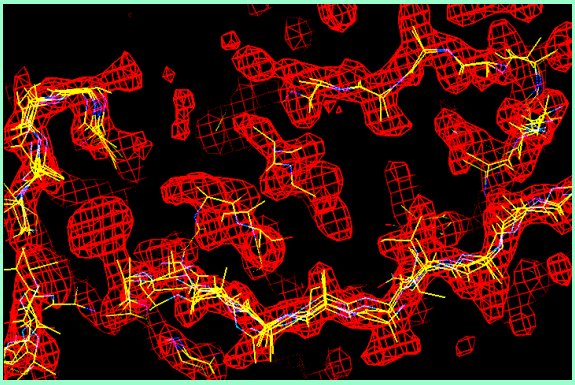


Chain extension by placement of tripeptide fragments

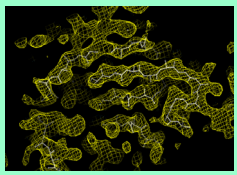


- *Look-ahead scoring: find fragment that can itself be optimally extended*
- *C-terminal extension. Start at C-terminus of protein*
- *Each of 10000 fragments: superimpose CA C O on same atoms of last residue in chain (extending by 2 residues): pick best 10*
- *Each of best 10: extend again by 2 residues 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 with tripeptide libraries (result: many overlapping fragments)

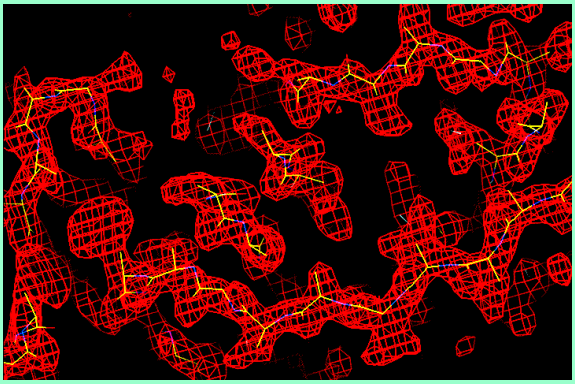


Assembly of main-chain

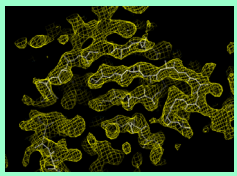


- *Choose highest-scoring fragment*
- *Test all overlapping fragments as possible extensions*
- *Choose one that maximizes score when put together with current fragment*
- *When current fragment cannot be extended: remove all overlapping fragments, choose best remaining one, and repeat*

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



Scoring side-chain templates at each position

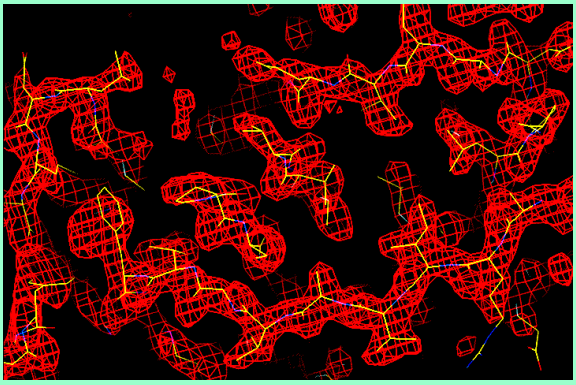


- *Identify side-chain orientation from N CA C of main-chain*
- *Get CC of template with density -> Z-score*
- *(Compare CC with mean, SD of all side chain density with this template)*
- *P(this side-chain/rotamer is correct)= Po(this side-chain/rotamer)*P(Z)*

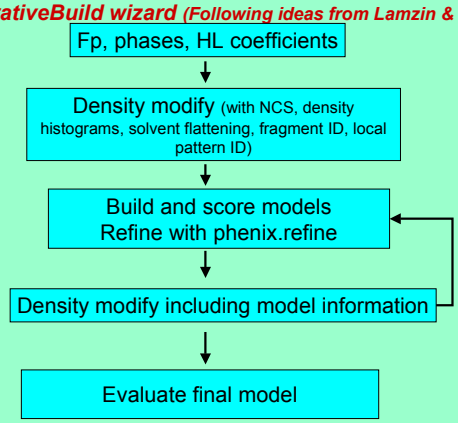
Side-chain template matching to identify sequence alignment to map (IF5A data)
Relative probability for each amino acid at each position
(Correct amino acids in bold)

#	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



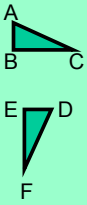
Iterative model-building, density modification and refinement at moderate resolution using the PHENIX IterativeBuild wizard *(Following ideas from Lamzin & Perrakis)*



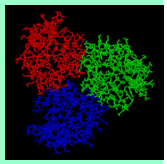
Automated NCS identification from heavy-atom sites

- 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 using heavy-atom sites and analysis of the electron-density map

Structure	Number of sites	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**

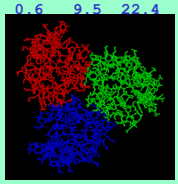
Score a possible arrangement based on:

- **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

                                  Link      Mol    NCS    NCS
Frag  Start End   N  Overlap     Length    Radius  RMSD   <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
```



Automation of structure determination

Use of scoring procedures to assist in decision-making

Simple procedures for automation choosing the current best path at each decision-point

The *PHENIX* project

Crystallographic software for automated structure determination

Computational Crystallography Initiative (LBNL)
-***Paul Adams***, Ralf Grosse-Kunstleve, Peter Zwart, Nigel Moriarty, Nicholas Sauter, Pavel Afonine

Los Alamos National Lab (LANL)
-***Tom Terwilliger***, Li-Wei Hung,Thiru Radhakannan

Cambridge University
-***Randy Read***, Airlie McCoy, Laurent Storoni,
-Hamsapriye

Texas A&M University
-***Tom Ioerger***, Jim Sacchettini, Kreshna Gopal, Lalji Kanbi,
-Erik McKee, Tod Romo, Reetal Pai, Kevin Childs, Vinod Re

Acknowledgements

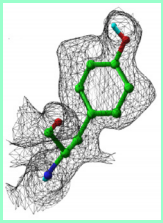
PHENIX: www.phenix-online.org

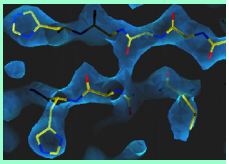
Paul Adams, Ralf Grosse-Kunstleve, Nigel Moriarty, Nick Sauter, Pavel Afonine, Peter Zwart (LBNL Computational Crystallography Initiative)

Randy Read, Airlie McCoy, Laurent Storoni, Hamsaprie (Cambridge)

Tom Ioerger, Jim Sacchettini, Kresna Gopal, Lalji Kanbi, Erik McKee Tod Romo, Reetal Pai, Kevin Childs, Vinod Reddy (Texas A&M)

Li-wei Hung,Thiru Radhakannan (Los Alamos)





Generous support for PHENIX from the NIGMS Protein Structure Initiative