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)

Read Facts
("what is state of the world now?")

Get user inputs (if needed)

List and score all options

Carry out best option

No more options:
write state to Facts and quit

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)

PHENIX AutoSol wizard standard sequence

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

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

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

Evaluating electron density maps: Methods examining the map itself

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

Evaluating electron density maps: Methods examining the map itself

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

Scoring: does the native Fourier look like a protein?

A good map: clear solvent vs protein

A poor map: looks the same all over

Overall SD of local rms of map
Evaluating electron density maps
Methods based on density-modification and R-factors

<table>
<thead>
<tr>
<th>Basis</th>
<th>Good map</th>
<th>Random map</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-factor in 1st cycle of density modification</td>
<td>Low R-factor</td>
<td>High R-factor</td>
</tr>
<tr>
<td>(Cowtan, 1996)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation of map made</td>
<td>High correlation</td>
<td>Lower correlation</td>
</tr>
<tr>
<td>with map-probability phases with original map</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Terwilliger, 2001)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(map-probability from solvent flattening or from</td>
<td></td>
<td></td>
</tr>
<tr>
<td>truncation at high density level)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Skew of electron density in maps of varying quality
IF5A (P. aerophilum, 60% solvent; randomized maps)

SD of local rms of electron density in maps of varying quality
IF5A (P. aerophilum, 60% solvent; randomized maps)

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?

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"

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

Estimates for randomized maps are much better than those of 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

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

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)= P(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 | 13| 18| 16| 11| 12| 2 | 7 | 2 | 2 | 2 | 2 | 0 | 0 | 1 | 4 | 3 | 3 | 1 |
| 2    | 4 | 11| 14| 37| 5 | 2 | 0 | 2 | 0 | 0 | 2 | 3 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 6 | 3 | 3 |
| 3    | 11| 23| 5 | 12| 5 | 3 | 2 | 5 | 2 | 1 | 3 | 3 | 3 | 3 | 5 | 1 | 3 | 2 | 0 | 2 | 2 | 3 |
| 4    | 7 | 9 | 6 | 16| 5 | 2 | 0 | 1 | 3 | 0 | 4 | 1 | 0 | 7 | 0 | 2 | 0 | 3 | 4 | 0 | 3 | 4 |
| 5    | 31| 7 | 3 | 7 | 4 | 2 | 5 | 9 | 1 | 3 | 3 | 1 | 0 | 4 | 1 | 0 | 8 | 2 | 2 | 0 | 11| 1 |
| 6    | 1 | 3 | 3 | 41| 14| 8 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 4 | 0 | 0 | 0 | 1 | 9 |
| 7    | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15| 0 | 15| 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8    | 2 | 3 | 6 | 23| 10| 6 | 2 | 1 | 0 | 1 | 4 | 3 | 0 | 0 | 8 | 1 | 6 | 0 | 0 | 1 | 6 |
| 9    | 95| 0 | 0 | 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)

- Fp, phases, HL coefficients
- Density modify (with NCS, density histograms, solvent flattening, fragment ID, local pattern ID)
- Build and score models
- Refine with phenix.refine
- Density modify including model information
- Evaluate final model

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

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

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)

<table>
<thead>
<tr>
<th>Link</th>
<th>Mol</th>
<th>NCS</th>
<th>NCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>48</td>
<td>6.6</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>74</td>
<td>24.5</td>
</tr>
<tr>
<td>3</td>
<td>115</td>
<td>23</td>
<td>14.4</td>
</tr>
<tr>
<td>4</td>
<td>166</td>
<td>21</td>
<td>5.2</td>
</tr>
</tbody>
</table>

Residues placed for this molecule: 98
Residues placed: 309 of 588 or 52%
Residues built without side chains: 45
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 Reddy

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, Hamsapriye (Cambridge)

Tom Ioerger, Jim Sacchettini, Kreshna 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