SHELX Workshop 2006

Automated Tracing of Nucleic Acids

Tim Grüne
University of Göttingen
Dept. of Structural Chemistry
September 2006

http://shelx.uni-ac.gwdg.de

tg@shelx.uni-ac.gwdg.de
Automated Building of Proteins

Well established automated programs for proteins include

- Arp/wArp (S. Cohen, V. Lamzin, A. Perrakis, R. Morris et al.)
- Resolve (T. Terwilliger)
- Buccaneer (K. Cowtan)

Semi–automated procedures implemented in (probably) all graphics programs, e. g.

- Main (D. Turk)
- O (A. Jones)
- coot (P. Emsley)
Automated Building of Nucleic Acids?

- XFP suite (F. Pavelcik and B. Schneider, ECM 2006) According to documentation, builds 65% – 100% of RNA structures. Difficult to set up, no well known (I dare say).

With more popular programs, restricted fitting of nucleic acids if considered as ligands, *e.g.*

Resolve: Treat as ligand, up 5–10 bases (according to T. Terwilliger)

Graphical programs support fitting or building of nucleic acids

- Coot (P. Emsley): as ligand
- X–Ligand (T. Oldfield): Monte Carlo search of torsion angle space — large molecules probably exceeds time limits
Building Strategies — a Quick Overview

Building starts with an electron density map and ends with complete a model as possible.

Four steps:

1. Finding starting points ($C_\alpha$, $P$, secondary structure elements)
2. Combination into larger fragments
3. Sequence assembly from fragments
4. Addition of side chains
Fragment search

- Arp/wArp (A. Perrakis, V. Lamzin, et al.): from single peaks via main chain to model with side chains.

- Resolve (T. Terwilliger): $\alpha$–helices and $\beta$–strands, search in reciprocal space, statistical pattern matching

- Buccaneer (K. Cowtan): oriented $N - C_\alpha C_\beta - C$ template fragment

All methods use characteristic features of proteins for evaluation and selection.
Protein vs. Nucleic Acids — the backbone

$N_i \rightarrow C_\alpha \rightarrow C \rightarrow N_{i+1}$

$P_i \rightarrow O_5' \rightarrow C_{5'} \rightarrow C_{4'} \rightarrow C_{3'} \rightarrow O_{3'} \rightarrow P_{i+1}$
Proteins — The Ramachandran plot

Structure of protein backbone described by $\Phi$ and $\Psi$ angles: Strong correlation, represented by the Ramachandran plot.

Coot, P. Emsley

Protein vs. Nucleic Acids
Conformational Diversity for Nucleic Acids


Schneider, B. et al., *Nucleic Acid Research*, 2004

Protein vs. Nucleic Acids
Distance Distributions between Building Blocks

Structures at the PDB by August 2006. Proteins: 73935 samples ($d < 1.1\text{Å}$), DNA: 1638 samples ($d < 1.5\text{Å}$), RNA: 678 samples ($d < 1.5\text{Å}$).
Protein vs. Nucleic Acids — buddies and baddies

<table>
<thead>
<tr>
<th>Proteins</th>
<th>Nucleic Acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>⊕ 3.8Å distance</td>
<td>⊗ broad $P - P$ distance distribution, conformational flexibility</td>
</tr>
<tr>
<td>⊕ secondary structure elements ($\alpha$-helices, $\beta$-strands)</td>
<td>⊗ 70% helical (Pavelcik)</td>
</tr>
<tr>
<td>⊗ 20 very different side chains, partially very flexible</td>
<td>⊗ helices often superhelical, not straight</td>
</tr>
<tr>
<td></td>
<td>⊗ A-, B-, Z-[DR]NA</td>
</tr>
<tr>
<td></td>
<td>⊗ 4 bases (A, C, G, T/U) in two groups (Purines, Pyrimidines): rigid and large</td>
</tr>
<tr>
<td></td>
<td>⊗ Protein—Nucleic Acid complexes: density of protein often much better than for nucleic acid</td>
</tr>
</tbody>
</table>
Nucleic Acids — Think Positive

Bases

- Only two types, Purines and Pyrimidines
- Prominent features, even at moderate resolution: flat and large
Nucleic Acids — Think Positive

Phosphates

- Strong scattering power \((15e^- \text{ compared to } 6 - 8e^- \text{ for } C, N, O) \Rightarrow \text{good for peak search}\)

- Tetrahedrally shaped
Finding Bases

A flat solid body: one momentum of inertia much larger than the other two

1. Peak Search through electron density map
2. Calculation of principal “Momenta of Inertia” of 2Å sphere around “centre of density”

3. Sort by Momenta

\[ Q \ast \frac{J_{\text{max}}}{J_{\text{min}} + J_{\text{mid}}} \]
Excursion: Tensor of Inertia

Definition:

\[ \Theta_{ij} = \sum_{\nu} \rho(x^\nu) \left( \delta_{ij} x_i^\nu x_i^\nu - x_i^\nu x_j^\nu \right) \]

Properties:

- Expressed as symmetric \( 3 \times 3 \) matrix
- Calculation independent of orientation
- Eigenvalues are momenta of inertia along Eigenvectors — principal axes
- (A body rotates stably about either the axis with largest \( J_{\text{max}} \) or smallest \( J_{\text{min}} \) momentum)
Finding Phosphates

1. Peak Search through electron density map
2. Calculation of principal “Momentum of Inertia” of 2Å sphere around peak

\[ Q \ast \exp\left(\frac{J_{\text{min}} - J_{\text{max}}}{J_{\text{mid}}}\right) \]

3. Sort by Momenta

4. Alternatively, exploit tetrahedrally coordinated oxygens: calculate correlation coefficient of opposite points. Spherical: \( \approx 1 \), Phosphate: \( \approx -1 \)
Putting it all together

- some bases occupy the same space — overlaps
- phosphate peaks at correct positions, but there are extra peaks

(from 1.75 Å data)
Sequencing for Computers

- Need to find sequence through P–Base–P triplets

- Select amongst overlapping bases

Problem: relatively large conformational freedom

Required: several independent quantities to achieve reliable sequencing

Target function: quantify a quality, e.g. the difference to the 3.8Å distance between adjacent $C_\alpha$ atoms.

We have: putative phosphate and base positions $\Rightarrow$ use $5'P, 3'P, C1', N9$ (purines), $N1$ (pyrimidines)
Target Functions — Candidates

Scatter plots

$\Delta(N9/1, 5'P) \text{ vs. } \angle(C1', 3'P, C1')$

$\angle(N9/1, C1', 5'P) \text{ vs. } \angle(N9/1, C1', 3'P)$

$\angle(N9/1, C1', 3'P) \text{ vs. } \Delta(C1', 3'P)$

Strategies
Does this work?

- Considering only Bases and Base–neighbours turns out not to be sufficient for reliable selection of bases, phosphates and for assignment of sequence.
- Larger Fragments required
  - helical templates (aka resolve, hel (XFP))
  - templates cannot be too large: deviation from linearity
  - flexible search: time consuming
  - characteristics of $P PPP$ and $PPPPP$, e.g. $\Delta((PPP), P)$ (distance to $PPP$–plane)?
Acknowledgments

G. M. Sheldrick
S. Cohen
K. Cowtan
P. Emsley
M. Runnwerth
B. Moolenaar

*et al. mult.*