Two Optimization Problems in X-ray Crystallography

George N. Phillips, Jr.

University of Wisconsin – Madison
Proteins are the machinery of life
Myoglobin and Hemoglobin Deliver Oxygen in Organisms with Vascular Systems
Light microscope

Electron microscope

X-ray Crystallography

2,000 X

1,000,000 x

100,000,000 X
X-ray Laboratory
The phase problem

\[ F(S) = \sum f_j e^{2\pi i r_j \cdot S} \] to calculate diffraction is easy

To get the original structure,

\[ \rho(r) = \sum |F(S)| e^{i\alpha} e^{-2\pi i r \cdot S} \]

which is an inverse problem with unknown phases, as only the magnitude of \( F \) can be measured directly.
Solutions

- More experimental data (heavy atoms)
- Use prior information from a related molecule (molecular replacement)
- Use knowledge about distributions expected in the electron density map
Crystallography reveals locations of electron clouds of atoms
Interpreting an isosurface is tedious
CS ideas have helped

- Computer graphics
- Image analysis
- Computational geometry
- Machine learning
- Databases
- Cheap arithmetic
Semi-immersive VR
Two topics

• The phase problem
• Interpretation of noisy/low resolution electron density maps
Ab initio Solutions to the Phase Problem

• Solved for atomic resolution data (Hauptman and Karle Nobel Prize)

• Unsolved for lower resolution cases
  – Reciprocal space methods
  – Real space filters
  – Iteration with in both spaces
Ideas

• Histogram matching as a real space filter
• Solvent region identification and flattening/flipping
• Reparameterization in lower dimensional space (Zernike descriptors, other period basis functions- Atilla Sit, Julie Mitchell
• “Lego block” model – Steve Wright
Different approaches

• High resolution – atomic locations as delta functions against a vacuum background, i.e. find how to make things pile-up in spikes

• Low resolution – smoothed function with protein in a blob surrounded by less dense, uniform water region (50% of the total volume) find a continuous region
Lego Blocks

• Reduce the problem to the linear problem of assembling blocks (with known scattering contributions)
Wright’s formulation

For a given set of coefficients $\alpha := (\alpha_1, \alpha_2, \ldots, \alpha_m)$ (which will be the variables in the reconstruction), the predicted structure factors are

$$F_H = \int_V \rho(r) \exp(2\pi i H \cdot r) \, dr = \sum_{i=1}^{m} \alpha_i \int_V \rho_i(r) \exp(2\pi i H \cdot r) \, dr. \quad (2.2)$$

If we define the complex vector $A_H \in \mathbb{C}^m$ by

$$(A_H)_i := \int_V \rho_i(r) \exp(2\pi i H \cdot r) \, dr, \quad i = 1, 2, \ldots, m,$$

the structure factor $F_H$ is linear in the variables $\alpha$, that is,

$$F_H = A_H^T \alpha = \overline{A_H^T} \alpha. \quad (2.3)$$

Hence, given a set of measured amplitudes $G_H$ for lattice vectors $H \in \mathcal{H}$, we could identify $\rho$ by solving this optimization problem:

$$\min_{\alpha} f(\alpha) := \frac{1}{2} \sum_{H \in \mathcal{H}} \left[ |A_H^T \alpha|^2 - G_H^2 \right]^2. \quad (2.4)$$
Is it computationally feasible?

“ The optimization problem (2.4) is no doubt nasty and highly nonconvex, but at least it has a simple form.” --S. Wright
The phase problem

\[ F(S) = \sum f_j e^{2\pi i r_j \cdot S} \] to calculate diffraction is easy

To get the original structure,

\[ \rho(r) = \sum |F(S)| e^{ia} e^{-2\pi i r \cdot S} \]

which is an inverse problem with unknown phases, as only the magnitude of \( F \) can be measured directly.
Given: Sequence + Density Map
Resolution & Related Work

Resolution is a property of the protein crystal.

Our Method: ACMI

ARP/wARP

TEXTAL & RESOLVE

Higher Resolution : Better Quality
Automatic Crystallographic Map Interpretation (ACMI)

- Local Match
  - Algorithm searches for sequence-specific 5-mers centered at each amino acid
  - Many false positives in ED search

- Global Consistency
  - Use probabilistic model to filter false positives
  - Find most probable backbone trace
5-mer Lookup and Cluster

...VKHVLVSPEKIEELIKGY...

Cluster 1
wt = 0.67

Cluster 2
wt = 0.33

PDB
5-mer Search

- 6D search (rotation + translation) for representative structures in density map

- Compute “similarity”
  \[ t(x) = \sum_y \varepsilon_{frag}(y) \left( \rho_{frag}(y) - \rho_{map}(x-y) \right)^2 \]

- Computed by Fourier convolution (Cowtan 2001)

- Use tuneset methods to convert similarity score to probability
Convert Scores to Probabilities

5-mer representative

match to tuneset

search density map

Bayes’ rule

NEG

POS

distribution score distributions

probability distribution over unit cell

P(5-mer at $u_i | \text{Map}$)

scores $t_i(u_i)$
Next step...

Where we are now

For each amino acid in the protein, we have a probability distribution over the unit cell

\[ P(u_i \mid \text{Map}) \]

Where we are headed

Find the backbone layout maximizing

\[
\left[ \prod_{\text{AAs } i} P(u_i \mid \text{Map}) \right] \times \left[ \prod_{\text{AA-pairs } i,j} P(\text{conformation } \{u_i, u_j\}) \right]
\]
Pairwise Markov Field Models

- A type of undirected graphical model
- Represent joint probabilities as a product of vertex and edge potentials
- Similar to (but more general than) Bayesian networks

\[ p(U | y) \propto \prod_{\text{edges } s \rightarrow t} \psi_{st}(u_s, u_t) \prod_{\text{vertices } s} \psi_s(u_s | y) \]
Two types of structural (edge) potentials

- Adjacency constraints ensure adjacent amino acids are \(~3.8\text{Å}~\) apart \textit{and} in the proper orientation
- Occupancy constraints ensure nonadjacent amino acids do not occupy same 3D space
Backbone Model Potential

$$p(u | \text{Map}) \propto \prod_{\text{adjacent AAs} \atop i \leftrightarrow j} \psi_{adj}(u_i, u_j) \times \prod_{\text{nonadjacent AAs} \atop i \leftrightarrow j} \psi_{occ}(u_i, u_j) \times \prod_{\text{amino acid } i} \psi_i(u_i | \text{Map})$$

Constraints between adjacent amino acids:

$$\psi_{adj}(u_i, u_j) = p_x(|| x_i - x_j ||) \times p_\theta(u_i, u_j)$$
Backbone Model Potential

\[ p(u \mid \text{Map}) \propto \prod_{i \leftrightarrow j} \psi_{\text{adj}}(u_i, u_j) \times \prod_{i \leftrightarrow j} \psi_{\text{occ}}(u_i, u_j) \times \prod_{\text{amino acid } i} \psi_i(u_i \mid \text{Map}) \]

Constraints between nonadjacent amino acids:

\[ \psi_{\text{occ}}(u_i, u_j) = \begin{cases} 0 & \text{if } \| x_i - x_j \| < K \\ 1 & \text{otherwise} \end{cases} \]
Backbone Model Potential

\[
p(U | \text{Map}) \propto \prod_{\text{adjacent AAs} \ i \leftrightarrow j} \psi_{\text{adj}}(u_i, u_j) \times \prod_{\text{nonadjacent AAs} \ i \leftrightarrow j} \psi_{\text{occ}}(u_i, u_j) \times \prod_{\text{amino acid } i} \psi_i(u_i | \text{Map})
\]

Observational ("amino-acid-finder") probabilities

\[
\psi_i(u_i | \text{Map}) = \Pr(5\text{mer } s_{i-2} \ldots s_{i+2} \text{ at } u_i)
\]
Probabilistic Inference

- Want to find backbone layout that maximizes

\[
\prod_{i \leftrightarrow j} \psi_{adj}(u_i, u_j) \times \prod_{i \leftrightarrow j} \psi_{occ}(u_i, u_j) \times \prod_{\text{amino acid } i} \psi_i(u_i \mid \text{Map})
\]

- Exact methods are intractable
- Use belief propagation (BP) to approximate marginal distributions

\[
p_i(u_i \mid \text{Map}) = \sum_{u_k, k \neq i} \cdots \sum_p(U \mid \text{Map})
\]
Belief Propagation (BP)

- Iterative, message-passing method (Pearl 1988)

- A message, $m_{i 	o j}^n$, from amino acid $i$ to amino acid $j$ indicates where $i$ expects to find $j$

- An approximation to the marginal (or belief), $b_i^n$, is given as the product of incoming messages
Belief Propagation Example

\[ b_{ALA}^0(x_{ALA} \mid y) \quad m_{GLY \rightarrow GLY}^p(x_{GLY}) \quad b_{GLY}^0(x_{GLY} \mid y) \]
Phase 2: Standard ACMI

\[ P(b_k) \]
Phase 2: Ensemble ACMI

Protocol 1

Protocol 2

Protocol C

MRF

$P_1(b_k)$

$P_2(b_k)$

$P_C(b_k)$

Ameet Soni – CS Thesis work
1YDH at 3.5Å Resolution

prob(AA at location)

1.5Å RMSd
90% coverage
Experiments

• Tested ACMI against other map interpretation algorithms: TEXTAL and Resolve

• Used ten model-phased maps

• Smoothly diminished reflection intensities yielding 2.5, 3.0, 3.5, 4.0 Å resolution maps
RMS Deviation

![Graph showing the relationship between Density Map Resolution and Ca RMS Deviation with lines for ACMI, RMS, and Deviation.]
Model Completeness

Density Map Resolution

% chain traced

% residues identified correctly

ACMI

Textual Resolve

% completeness

Density

ResoluVon

ACMI

Resolve

% chain traced

% residues identified correctly

ACMI

Textual Resolve

% completeness

Density

ResoluVon

ACMI

Resolve
Per-protein RMS Deviation

ACMI RMS Error

Resolve RMS Error

TEXTAL RMS Error
Inclusion of More Probablistic “Prior” Information Underway

- Disorder Prediction
- Info from Fast ID of helices and sheets (Phenix)
- Contact map predictions from homologs
- Solvent exposure
- Rosetta repacking of side chains
Acknowledgements

- Frank DiMaio
- Ameet Soni
- Jude Shavlik
- Craig Bingman
- Siddarth Puthur
- Atilla Sit
- Julie Mitchell
- Stephen Wright
- NIH-PSI grant U54 GM074901
- NLM grants R01 LM008796 and T15 LM007359