Chemistry, Mathematics, And Coding: Interdisciplinary Research In Estimating The Three-Dimensional Structure of A Protein

Terri M. Grant, Ph.D.
Old Dominion University
Department of Mathematics and Statistics
tgrant@odu.edu
Proteins

- Proteins are large, complex molecules which our bodies and the cells in our bodies need to function properly. Proteins come in many different sizes and shapes; hence, they can perform a variety of important tasks in the body.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Function</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antibodies</td>
<td>Defend</td>
<td>Covid-19</td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>Transport</td>
<td></td>
</tr>
<tr>
<td>Digestive Enzymes</td>
<td>Decompose</td>
<td></td>
</tr>
</tbody>
</table>
The shape of a protein is its three-dimensional structure. By folding into its unique three-dimensional structure:

1. A protein’s biological function can be predicted on the basis of structural similarity of closely related proteins.
2. It is possible to determine which molecules or drugs can efficiently bind and how they will bind to a protein.
3. Gain more information about diseases caused from protein misfolding.
Protein Folding and Levels of Structures

- **Primary structure**
  Sequence of amino acids held together by peptide bonds (polypeptide chain)

- **Secondary structure**
  Arise as H bonds form between local groups of amino acids

- **Tertiary structure** (3D structure)
  Determined by hydrophobic interactions, ionic bonding, etc.

- **Quaternary**
  Two or more polypeptide chains bonded together
3D Structure and Computational Method

Chemistry/Biology

Physics/Mathematics

C/B

P/M

Estimating A Protein’s 3D Structure

CS

Computer Science
3D Structure and Computational Method

- Sequence/Structural Data
- Secondary Structure Prediction
- Homology Modeling
- Experimental Methods (NMR, Electron Microscopy, Crosslinking, Footprinting, Fluorescence Transfer, Hydrodynamics, etc.)

Modeling

- Distances
- Angles
- Volume
- Surface
- Shape
- etc.

Formulation

Algorithm

3D Structure

Biological Function
Elements of Estimating a Protein’s (3D) Structure

“Free” Structural Data

Mathematical Formulation

Algorithm/Validation
“Free” Structural Data
A Protein’s Atomic Representation

Backbone Torsional Angles

Side chain

Bond angles (generally rigid)

Bondlengths (generally fixed)

Local nonbonded distances

Nonlocal nonbonded distances

Bond angles (generally rigid)

Bondlengths (generally fixed)
Experimental Methods and Structural Data

X-ray Crystallography

NMR Spectroscopy
College of Sciences Major Instrumentation Cluster

The College of Sciences Major Instrumentation Cluster (COSMIC) laboratory provides researchers access to state-of-the-art instrumentation in a user-friendly teaching/research environment. COSMIC also provides analytical services for interdisciplinary research applications in chemical, environmental, and biological analyses for ODU, other universities around the world, government agencies, and private industry. It stimulates collaborative research between faculty, students, and staff at ODU and other universities, government agencies, and private companies. It is committed to quality analytical analysis and instrumentation training of researchers.

COSMIC has successfully aided researchers with analytical availability that has brought in millions of dollars' worth of grant awards for ODU faculty and collaborators. Many papers have been published in research journals and proposals have been funded from data obtained in the COSMIC lab. Research groups from around the world recognize and respect ODU as a partner research institution because of the scientific expertise of its faculty, the technical expertise of the COSMIC staff, and the excellent analytical data obtained with state-of-the-art instrumentation in the COSMIC facility.

Instrumentation

- Electron Microprobe
- Mass Spectrometry
- Nuclear Magnetic Resonance (NMR)
- X-Ray Powder Diffractometer

https://www.odu.edu/sci/research/cosmic:
Given a sequence alignment of a set (i.e. family) of reference of proteins, structural data is extracted from conserved regions of the set.
The protein is often represented by its molecular surface to help in:

- Detecting of possible drug binding sites on the surface.
- Identifying of possible proteins for pharmaceutical use.
- Modeling hydrophobic/hydrophilic interactions in proteins.
Mathematical Formulation
### Mathematical Formulation

**Potential Energy**
Biologically active at lowest energy state

**Knowledge-Based**
Statistical potential based on probability of occurrence of structural features in regions

**Fundamental Problem:**
find the coordinates for all atoms in protein (i.e. configuration) that best satisfies a set of structural data

**Nonlinear Least-Squares**
Measure error between the input and the 3D model
Estimation of 3D Structure: Optimization Problem

1. **Objective function**: a function that measures “how well” a given configuration satisfies the structural data, and we want to find the solution by minimizing the value of this function. **Examples: Potential Energy, Least squares Error, Statistical Potential**

2. **Decision variable(s)**: variables for which the objective function reaches its optimal value.

3. **Configuration Space**: the space of all possible solutions for which function is minimized
   - all possible configurations

4. **Constraint(s)**: limitations on configuration space that are determined by the physical nature of the problem
   - Minimum Separation Distance, Van der Waals, Chirality, etc.
## Examples of Terms in Objective Function

<table>
<thead>
<tr>
<th>Name</th>
<th>Function</th>
<th>Constant of Proportionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bond stretching</td>
<td>( E_{bl} = \sum_{i,j \in P} s_{ij} (r_{ij} - d_{ij})^2 )</td>
<td>( s_{ij} ) is stretching constant or reciprocal of variance</td>
</tr>
<tr>
<td>Or Least Squares Error function</td>
<td></td>
<td>( r_{ij} ) is the known distance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( d_{ij} ) is the calculated distance</td>
</tr>
<tr>
<td>Bond bending</td>
<td>( E_{bl} = \sum_{ij \text{ bonded}} \frac{c_{ijk}}{2} (\theta_{ijk} - \tilde{\theta}_{ijk})^2 )</td>
<td>( c_{ij} ) is stretching constant or reciprocal of variance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \theta_{ij} ) is the known torsional angle</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \tilde{\theta}_{ijkl} ) is the calculated torsional angle</td>
</tr>
<tr>
<td>Electrostatic Potential</td>
<td>( E_{elec} = \sum_{i,j \in P} q_i q_j / \varepsilon_{ij} r_{ij} )</td>
<td>( q_i ) Charge of atom i</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( q_j ) Charge of atom j</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \varepsilon_{ij} ) Dielectric constant</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( r_{ij} ) is the known distance</td>
</tr>
<tr>
<td>Van der Waals Potential</td>
<td>( E_{vdw} = \sum_{i,j \in P} \varepsilon_{ij} \left[ \left( \frac{1}{r_{ij}} \right)^{12} - \left( \frac{1}{r_{ij}} \right)^{6} \right] )</td>
<td>( \varepsilon_{ij} ) Dielectric constant</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( r_{ij} ) is the known distance</td>
</tr>
<tr>
<td>Torsional (Dihedral) Angle</td>
<td>( E_{dih} = \sum_{ijkl \text{ angle}} c_{ijkl} [1 + \cos(n\theta_{ijkl} - \tilde{\theta}_{ijkl})] )</td>
<td>( c_{ijkl} ) torsional constant</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \theta_{ijkl} ) is the known torsional angle</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \tilde{\theta}_{ijkl} ) is the calculated torsional angle</td>
</tr>
</tbody>
</table>
Penalty function – is an additional term included in the objective function that assigns very large function to the objective function for any atoms placed in violation regions.

- MSD/VDW constraint
- Chirality constraint
- Torsional Angles Constraint
- Surface Constraint
3D Structure Estimation as An Optimization Problem

Unconstrained Optimization Formulation:

Find \( \{x_0, y_0, z_0, x_1, y_1, z_1, \ldots, x_n, y_n, z_n\} \) that minimizes:

\[
E_{bl} + E_{ba} + E_{vdw} + E_{dihe} + E_{chirality} + \ldots
\]

Constrained Optimization Formulation:

Find \( \{x_0, y_0, z_0, x_1, y_1, z_1, \ldots, x_n, y_n, z_n\} \) that minimizes:

\[
f = E_{distances}
\]

subject to:
- MSD/VDW constraint
- Torsional Angle constraint
- Chirality constraint
Algorithm/Validation
Types of Minima

- **Local Minimum**: Shoulder
- **"Flat" Local Minimum**
- **Global Minimum**

**Goal**: Identify all local minima and from this set select the global minimum

**Local Minima Occurrence**
- Nonlinear Multi-term Objective Function
- Function of $3n$ variables were $n$ is number of atoms
- Violation of Physical Constraints
Local Optimization Algorithm

- Initialize \texttt{current} to starting configuration
- Loop:
  - Let \texttt{next} = update of the \texttt{current}
  - If value(\texttt{next}) > value(\texttt{current}) return \texttt{current}
  - Else let \texttt{current} = \texttt{next}
- Variants: choose first better configurations, randomly choose among better configurations
Local Optimization and Update Equation

Let

\[ \vec{x} = \{x_0, y_0, z_0, \ldots, x_n, y_n, z_n\} \]

Update Equation:

\[ \vec{x}_{new} = \vec{x}_{old} + \alpha \Delta \vec{x} \]

- Determination of \( \alpha \) and \( \Delta x \) will be referred to as \textit{local optimization} methods.

- “Standard” local optimization methods involve simultaneous solution of \( n \) linear equations, where \( n = \# \) of elements in \( \Delta x \).
Local Optimization Algorithms

- **Gradient-based updates** (uses only first derivative information)
  -- Steepest descent
  -- Conjugate gradient

- **Newton’s method updates** (uses first and second derivatives)
  -- Improved convergence rate
  -- Numerically unstable near solution

- **Quasi-Newton updates** (use approximations to second derivatives)
  -- BFGS, Gauss-Newton, SR1, ...

- Cubic/quadratic **line search** to determine $\alpha$
Global Optimization Approaches

- **Deterministic methods**
  - Branch and bound, interval methods
  - Very reliable, deterministic guarantees
  - Too expensive for more than 20-50 variables

- **Stochastic methods**
  - Random steps or sampling
  - Probabilistic guarantees
  - Practical for <300 variables

- **Heuristic Methods**
  - e.g. Simulated annealing, Tabu search, Genetic algorithms, local optimization with backtracking line search
  - Effective on some very large problems
  - No practical guarantees
Challenges in Estimating 3D Structure

• Estimating 3D structures of complex proteins from sparse and noisy input data leads to difficult nonlinear optimization problems.

• Difficulty in including specific interactions (in an objective function) that can determine the detail features of substructures.

• Optimization methods often yield poor quality structures due to local minima and/or placement of atoms into positions that violate known physical constraints (van der Waals, chirality, etc.).
What are the Error Metrics? How are they defined?

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Implications</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low Function Error and High Root Mean Square Deviation (RMSD)</strong></td>
<td>function minimized; (estimated and some structural similarities); need more information about the protein</td>
<td><img src="image1.png" alt="Example Image" /></td>
</tr>
<tr>
<td><strong>Low Function Error and Low Root Mean Square Deviation (RMSD)</strong></td>
<td>function minimized; (estimated and true structures nearly identical in conformation);</td>
<td><img src="image2.png" alt="Example Image" /></td>
</tr>
<tr>
<td><strong>High Function Error and High Root Mean Square Deviation (RMSD)</strong></td>
<td>function not minimized; (estimated and true structures are dissimilar);</td>
<td><img src="image3.png" alt="Example Image" /></td>
</tr>
</tbody>
</table>
My Research: GNOMAD
GNOMAD Algorithm

Fundamental Problem:
find the coordinates for a set of atoms (i.e. configuration) that best satisfies a set of structural data

Nonlinear Least-Squares
• Simple Objective function
• Atom based - approach
• Quasi-Newton local optimization
• Domain-specific, non-random perturbation method to aid in global convergence
• Physical constraint enforcement via line search from local optimization

Constrained Formulation:

minimize: \[ f = \frac{1}{2} \sum_{i=1}^{m} \frac{(d_{ci} - d_{mi})^2}{\sigma_i} \]

subject to:

- MSD/VDW constraint
- Chirality constraint
- Torsional Angle constraint
- Surface Constraint

\[ m = \# \text{ of input distances} \]
\[ d_{ci} = \text{distance } i \text{ calculated from model} \]
\[ d_{mi} = \text{distance } i \text{ input mean} \]
\[ \sigma_i = \text{distance } i \text{ input variance} \]
## Modeling with Atomic Distances

<table>
<thead>
<tr>
<th>atom</th>
<th>atom</th>
<th>mean</th>
<th>var</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1.5</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>2.0</td>
<td>0.5</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>1.0</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1.7</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**INPUT**

$$(x_0, y_0, z_0)$$

**3D MODEL**

$$(x_1, y_1, z_1)$$

$$(x_2, y_2, z_2)$$

$$(x_3, y_3, z_3)$$

**Objective:**

Minimize differences (residuals) between input data and 3D model.
GNOMAD Algorithm

For $n_{\text{atoms}} = 2 \leq \text{total}_n_{\text{atoms}}$, OPTIMIZE \{A_i: i = 1, \ldots, n_{\text{atoms}}\}

Set up list of atoms ($a_1, \ldots, a_{n_{\text{atoms}}}$) in decreasing order of error

move atom $a_1$

move atom $a_{n_{\text{atoms}}}$

(Note: $a_i$ not necessarily $= A_i$)

“Cycle”

Repeat for $n_{\text{cycles}}$ or until some termination criteria is met

Set up list of atoms ($a_1, \ldots, a_{n_{\text{atoms}}}$) in decreasing order of error

move atom $a_1$

move atom $a_{n_{\text{atoms}}}$

(Note: $a_i$ not necessarily $= A_i$)
**MOVE ATOM Procedure**

1. Determine step direction and length using BFGS method.
2. Adjust step length using backtracking line search.
3. Check MSD and local perturbation constraints along final search line.
4. Continue moving until either:
   - line search fails,
   - convergence, or
   - max # of iterations reached.
Local Perturbation

Initial position

Move to Position A or Position B, depending on which has a lower objective function value.
Physical Constraint Checking

For final iteration only

Move to Position A or Position B, depending on which has a lower objective function value

Position at start of final iteration

Position A

Position at end of final iteration (prior to constraint checking)

Position B
Validation/Test Structure

- **Experiments**: Construct synthetic data sets using known pairwise atomic distances, chirality, torsional angles, and surface data from crystal structure. Investigate code’s ability to recreate the 3D crystal structure from subsets or entire synthetic data sets.

- **Protein Test Structures**:
  - 1ozz: A. Demophon antifungal protein
    - 43 residues, 587/170 atoms
  - 1aw0: M. Copper-Transporting Atpase
    - 72 residues, 936/282 atoms
  - 1bta: Ribonuclease Inhibitor protein
    - 89 residues, 1154/384 atoms
Quantitative Results

![Graphs showing quantitative results for different structures.](image)
Qualitative Effects of Torsional/Surface Information Satisfaction

Crystal Structure

Not Surface Constrained
(max. distance constraint error = 1.07 Å
RMSD = 4.22 Å)

1ozz

Surface Constrained
(max. distance constraint error = 0.86 Å
RMSD = 1.14 Å)
Qualitative Effects of Torsional Information Satisfaction

Crystal Structure

Not Surface Constrained
(max. distance constraint error = 1.85 Å
RMSD = 6.20 Å)

Surface Constrained
(max. distance constraint error = 0.97 Å
RMSD = 2.59 Å)
Qualitative Effects of Torsional Information Satisfaction

Crystal Structure

Not Surface Constrained
(max. distance constraint error = 1.41 Å
RMSD = 5.75 Å)

Surface Constrained
(max. distance constraint error = 0.92 Å
RMSD = 3.54 Å)
Protein Modeling and Estimation at ODU

Chemistry
Dr. L. Greene
Dr. J. Poutsmsas

Mathematics
Dr. T. Grant

Computer Science
Dr. J. He

Estimating A Protein’s 3D Structure
Current Research in Protein Modeling and Estimation at ODU

- **The Greene Group:** [https://fs.wp.odu.edu/lgreene/dr-greene/](https://fs.wp.odu.edu/lgreene/dr-greene/)
- **Dr. Poutsmas' Lab:** [https://ww2.odu.edu/~jradkiew/research/research.shtml](https://ww2.odu.edu/~jradkiew/research/research.shtml)
- **Dr. He’s Group:** [https://www.cs.odu.edu/~jhe/](https://www.cs.odu.edu/~jhe/)
- **ODU Major Instrumentation Cluster:** [https://www.odu.edu/sci/research/cosmic](https://www.odu.edu/sci/research/cosmic)
References

Questions?

Thank you to the REYES Organizers