Basic Computing with Biological Structures in 3D

Russ B. Altman
BMI 214
CS 274

© Copyright Russ Altman 2000

General Issues
• How do we represent structure for computation?
• How do we compare structures?
• How can we summarize structural families?

© Copyright Russ Altman 2000

Outline
• General Issues in Representing Structure
• Bond lengths, bond angles, dihedral (torsion) angles
• Coordinate Systems
  • Cartesian Coordinates
  • Internal Coordinates
  • Object-based representations
• Comparing structures, the RMSD (root mean squared distance between two structures).
• Summarizing structures as mean, variance.

© Copyright Russ Altman 2000

Basic Measurements
• Bond lengths
• Bond angles
• Dihedral (torsion) angles

© Copyright Russ Altman 2000

Bond Length
• the distance between bonded atoms is constant
• depends on the “type” of the bond (single: C-C, double: C=C, triple: C≡C)
• varies from 1.0 Å (C—H) to 1.5 Å (C—C) some others are slightly longer
• BOND LENGTH IS A FUNCTION OF THE POSITION OF TWO ATOMS.

© Copyright Russ Altman 2000

Basic Measurements
• Bond lengths
• Bond angles
• Dihedral (torsion) angles

© Copyright Russ Altman 2000

Computing Bond Length = computing distance
For two points (x1, y1, z1) and (x2, y2, z2)
Distance = \sqrt{(x_1-x_2)^2 + (y_1-y_2)^2 + (z_1-z_2)^2}

© Copyright Russ Altman 2000

Some non-covalent distances are virtually constant in the protein peptide backbone:
The Cα—Cα distance for consecutive peptides is 3.8 Å (we will see why in a moment).
Bond Angles

- All bond angles are determined by chemical makeup of the atoms involved, and are constant.
- Depends on the type of atom, and the number of electrons available for bonding.
- Ranges from 100° to 180°
- BOND ANGLE IS A FUNCTION OF THE POSITION OF THREE ATOMS.

Computing Bond Angle

\[
\theta = \arccos \left( \frac{\mathbf{X} \cdot \mathbf{Y}}{||\mathbf{X}|| ||\mathbf{Y}||} \right)
\]

Angle can be computed by computing the arccosine of the dot product between unit vectors \( \mathbf{BA} \) and \( \mathbf{BC} \).

Dihedral Angle

- These are usually variable.
- Range from 0—360° in molecules
- Most famous are \( \phi \), \( \psi \), \( \omega \) and \( \chi \) (see next figure)
- DIHEDRAL ANGLES ARE A FUNCTION OF THE POSITION OF FOUR ATOMS.

Computing Dihedral Angle

Compute cross-product of \( \mathbf{BA} \) and \( \mathbf{CB} \)
Compute cross-product of \( \mathbf{CB} \) and \( \mathbf{DC} \)
This produces two vectors perpendicular to the ABC plane and BCD plane.
Angle between those vectors (ala bond angle) is dihedral angle. Need to check if it is positive or negative.
Omega is constant = 180 (C-N doesn’t rotate)
Phi, Psi have range of values (Ca-N, N-C rotate)
The range is restricted by having things not bump into each other.

Typical values for secondary structures you have known.
• Alpha helices phi = –57, psi = –47
• Parallel Beta strands phi = –119, psi = 113
• Antiparallel beta strands phi = –139, psi = 135
• 3-10 Helix phi = –49, psi = –26

Rotation Matrices
• Rotation matrix is 3 x 3 matrix of unit vectors that are orthogonal.
• Represents how to take point in coordinate system (100, 010, 001) and find its coordinates in the coordinate system given in the Rotation matrix.

How to rotate around an axis by theta degrees.
Always multiply vector on the right as a column vector!

\[
R(X, \theta) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{pmatrix}
\]

\[
R(Y, \theta) = \begin{pmatrix} \cos \theta & 0 & \sin \theta \\ 0 & 1 & 0 \\ -\sin \theta & 0 & \cos \theta \end{pmatrix}
\]

\[
R(Z, \theta) = \begin{pmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix}
\]
Rotation around x-axis does not change x coordinate.

General rotation around an axis.

\[ \text{ROT}(K, \theta) = \begin{pmatrix} kxxv\theta + c\theta & kxky\theta + kzs\theta & kxzv\theta + kys\theta \\ kskxv\theta + c\theta & ksyv\theta + kzs\theta & kyszv\theta + kxs\theta \\ kskzv\theta + kxs\theta & kyszv\theta + kxs\theta & kxzv\theta + c\theta \end{pmatrix} \]

\[ K = \begin{bmatrix} kx & ky & k\theta^T \end{bmatrix} \]

\[ c\theta = \cos \theta \]
\[ s\theta = \sin \theta \]
\[ v\theta = 1 - \cos \theta \]

Think about advantages

3 Peptide Units = 12 atoms = 36 coordinates OR 6 dihedral angles
3 Sidechains = 12 atoms = 36 coordinates OR 5 dihedral angles
72 Cartesian Coordinates vs. 11 internal coordinates

Non-Cartesian Coordinates

- Cartesian coordinates assume orthogonal (xyz) axes, and then provide coordinate values along each axis (as in PDB).
- But if bond lengths and bond angles are virtually constant, why not reduce the number of variables by specifying only the dihedral angles?
- This is one of the internal coordinate schemes used.

Object-based Coordinate Systems

- What if we know that certain parts of the protein are in standard alpha-helical conformation?
  We can treat an entire helical backbone as a rigid body.

Think about disadvantages

- Some basic computations become much more difficult.
  What is distance between two points?
  What are the N points closest to point X?
- Difficult to relate objects that are not connected together.
- Much more nonlinear relationships between coordinates, which may make some kinds of optimization computations difficult.
Object-based coordinates

- We can build a helix in an arbitrary coordinate system, and then represent its location in a particular protein as a TRANSLATION and ROTATION.

Translation & Rotation

- Translation = (x y z) vector of some reference point
- Rotation = many different representations
  - Rotation matrix (3 x 3)
  - Euler Angles
    --rotate around z, then around x, then z
    --and others
  - Direction cosines
  - Quaternions
  - Many others

Criteria for comparing structures

- All atoms being equal, we would like to know how closely the two structures can be superimposed
- If they are exactly the same, then they should superimpose perfectly and have a “distance” of 0
- As they diverge, the distance should increase.

Comparing Structures

In order to compare two structures, A and B, we need some basic information:

1. We need to know which atoms in A correspond to which atoms in B.
   This is why we do alignments of sequence....
2. We need to know where atoms are.
   This is what we do with PDB files...
3. We need some criteria for comparing.

The RMSD

- RMSD = root mean squared deviation
  \[ RMSD = \sqrt{\frac{\sum d_i^2}{N}} \]

where N is the number of atoms
d_i is the distance between two atoms with index i from the two structures

We want the minimum RMSD
Computing the RMSD

• Can be formulated as an inefficient search around superimposed centers of mass (Huang, Blostein, Margerum).
• Method based on quaternions (Faugeras & Hebert).
• Method based on singular value analysis of specially formed matrix (Arun, Huang, Blostein).

Algorithm

• See Arun et al paper...
• Basically,
  1. Compute centroid of points for each object
  2. Subtract off centroid, so both objects at origin
  3. Create special matrix as sum of vector products
  4. Decompose matrix using SVD (singular value decomposition) and use resulting matrices to form optimal ROTATION
  5. Compute TRANSLATION necessary to get rotated points in proper location.

Algorithm is guaranteed to be optimal, under very broad set of conditions.

Disadvantages of RMSD

• Weights all atoms equally
• Unclear upperbound (what is RMSD of two random sets of points subject to bond constraints of proteins???)
• Significance of values changes as function of size of protein.

Benefits of RMSD

• Has appropriate behavior (identical structures have RMSD = 0.0, and degrades continuously)
• Relatively easy to compute
• Units are natural ones (e.g. Ångstroms)
• Lots of experience with it.
  --Similar structures usually 1—3 Å RMSD

Case Study: Myoglobin family

• Eight structures taken:
  sperm whale myoglobin
  sea hare myoglobin
  plant leghemoglobin
  sea lampry hemoglobin
  human hemoglobin A & B chains (cf. next lecture)
  chironomous hemoglobin
  bloodworm hemoglobin

• Aligned by hand, because very low amino acid identity (~ 20%).
• 115 common positions
RMS values (for alpha carbons)

- Average RMSD 2.19 Å
- Maximum RMSD 3.16 Å
- Minimum RMSD 1.22 Å

Conclusions

1. We know how to compute basic structural parameters (bond length, bond angle, and torsion angle)
2. Different coordinate systems have different advantages.
3. RMSD is a basic measure of similarity, and can be computed relatively easily.
4. We can superimpose related structures and summarize commonalities/differences.
Major Methods for Structure Determination

1. Xray crystallography
   - grow crystals
   - shoot xrays through the crystals
   - collect pattern of deflected xrays
   - structure = FFT(pattern)

2. NMR spectroscopy
   - concentrate protein in solution
   - expose to high magnetic field
   - watch magnetization transfer between close atoms
   - interpret transfer as close distance
   - use distances to build model

Interpret pattern to electron density

Xray crystallography: grow crystals

Explain NMR a bit.

Shoot Xrays through crystal

A Note on Computing with Distances

Russ B. Altman
BMI 214
cs 274
Sources of distance information

- NMR experiments
  - measures distances between H atoms
- Cross-linking experiments
  - measures ability of two pieces of structure to be connected by linker.
- Sequence analysis (!)
  - predict proximity based on features of sequence.
- Others...

Distance Numerology

- A molecule with N atoms has \((N^2-N)/2\) total distances (n choose 2).
- How many degrees of freedom?
- How many distances do we need to get the structure back?

History of Distances

GEODESY is the field concerned with measuring the large scale distances on earth (circumference, oblateness, etc.).

- One of the first fields to process distance information using least squares criteria.

ASTRONOMY is also very concerned with measuring distances between points...

Structural Biology

If you know a structure in detail, then you also know all the distances between atoms.

THUS, if you know all the distances, you know the structure.

How many distances?

Minimum number of exact Distances, in general.

We have access to all exact distances, but want to use the minimum number...

We can position the first 4 points (P1, P2, P3 and P4) with 6 distances

- P1 can go arbitrarily at the origin (no distances needed) \(\text{SUM} = 0\)
- P2 can be placed on x-axis arbitrarily at distance D12 from origin. (1 distance needed) \(\text{SUM} = 1\)

Set up coordinate system with 6 distances

\[ \begin{align*}
&x \quad P1 \\
&y \quad P2 \\
&z \quad P3 
\end{align*} \]
Set up coordinate system with 6 distances

Assume Exact Distances

- P3 can be placed in xy-plane arbitrarily at distance D13 from p1 and D23 from P2 (2 distances needed) \( \text{SUM} = 3 \)
- P4 can be placed in positive z hemisphere at distances, D14, D24, and D25 (3 distances needed) \( \text{SUM} = 6 \)

Now, how to add more points?

- We now have \( N-4 \) points to position in the newly established coordinate system.
- We can pick any of the remaining \( (N^2-N)/2 - 6 \) distances.
- How many do we need?
- Pick one distance at a time, and see how much you learn.

Set up coordinate system with 6 distances

First distance = surface of sphere

E.g. Distance to P1
Second distance = intersect spheres

Intersection of two spheres = circle.

E.g. Distances to P1 and P3...

Third distance

Intersection of sphere and circle = 2 points

Add distance to P4

Fourth distance chooses point

Intersection of sphere and point = point

E.g. distance to P2

So, add them all up

6 distances to establish coordinate system for first four points

4 distances for each additional point.

For N points:

\[ 6 + 4(N-4) \]

\[ = 6 + 4N - 16 \]

\[ = 4N - 10 \text{ exact distances} \]

But we don’t have exact distances

• Lacking exact distances, and lacking exactly the right distances needed to do the minimum 4N-10 calculation, we need more distances.

• Empirically, we often need at least twice the minimum number of distances (depending on their precision) to get a good structure.

\[ \sim 10N \]

Degrees of Freedom

For N atoms there are 3N coordinates.

But the coordinates of P1(3), P2(2), P3 (1) can be set arbitrarily, so really

3N-6 independent coordinates.

(Alternatively, given the 3N-6 coordinates, you could derive the positions of P1, P2, P3 knowing just the coordinate system).
4N-10 vs. 3N-6
Minimum number of exact distances needed to fully specify the positions of the atoms is not the same as the degrees of freedom.

WHY?
Coordinates # Distances
Distances don’t specify coordinates 1:1
Distance = \( f(6 \text{ coordinates, square root}) \)

Why do we care about all this?
• To compute biological structure, we often need to gather distance information.
• We now know how much we need to collect
• Where do we get information from?
NOT ALL DISTANCES ARE EQUAL!!

Expensive Distances vs. Cheap
• Distances that confirm primary structure of protein are less useful.
• If we know where the secondary structure is, then distances that confirm this are less useful.
• The real pearls are distances that bring together pieces of the protein that are distant in primary sequence and are not in the same secondary structure.

All distances not equally useful.

Distance Geometry
Based on rigorous mathematical theorem for case of exact distances.
If, for N points, you have all the distances between them, then can construct matrix, \( G \), whose eigenvectors are the vectors of \( x, y \) and \( z \) coordinates for the N points.
\[
G_{ij} = \frac{1}{2} (d(i0)^2 + d(j0)^2 + d(ij)^2)
\]
where \( d(ij) \) is distance from \( i \) to \( j \) (o = origin).
http://www.scripps.edu/case/nab5/NAB-sh-5.1.html

Distance Geometry can fail...
• when you don’t have exact distances.

Advantages:
--beautiful theory for exact case
--computationally fast
--mathematical basis

Disadvantages:
--convergence problems
--only distances
--not well characterized for inexact...