GRAPH LOCALIZATION ALGORITHMS FOR MOLECULAR CONFORMATION

Forbes Burkowski
Department of Computer Science
University of Waterloo

PROTEINS: INTRODUCTION

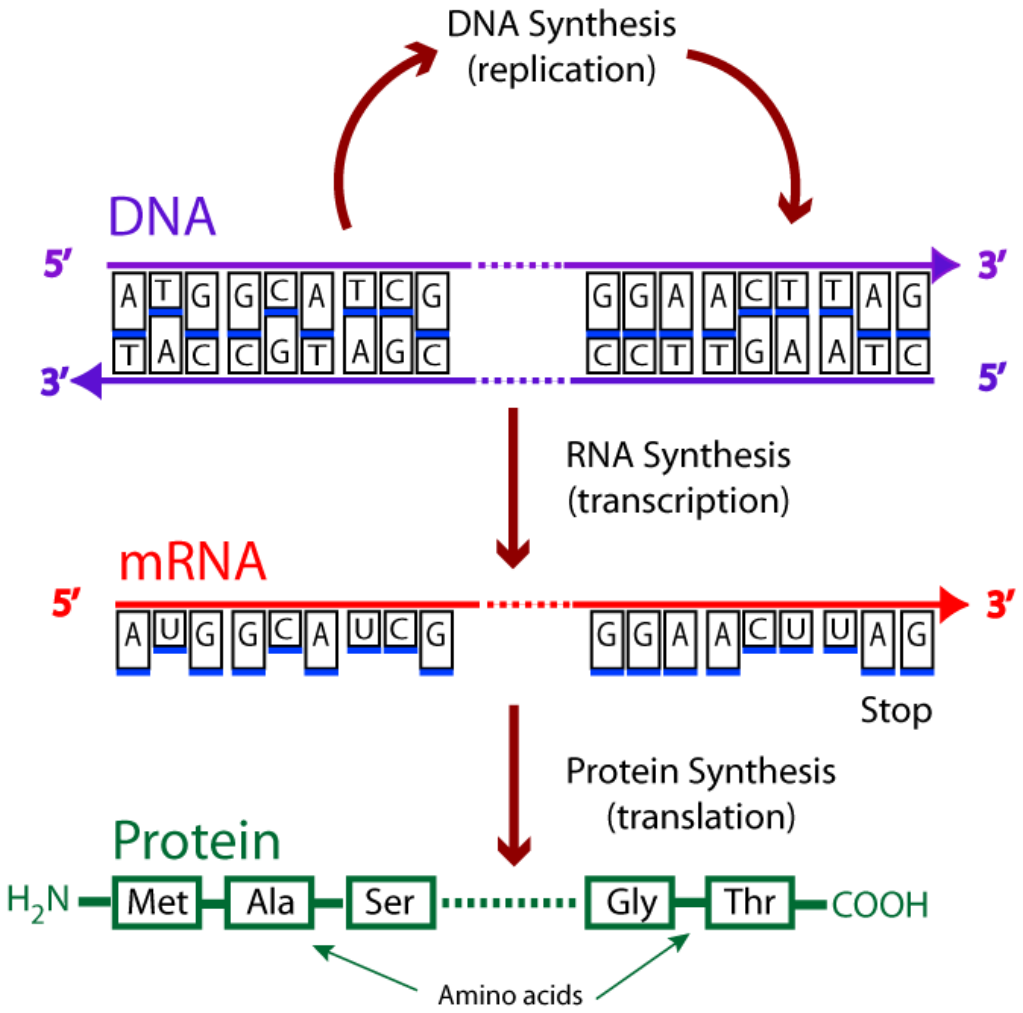
- Proteins are the versatile building blocks and active molecules that form the basis of living systems.
- Function follows structure
  - We study protein structure and its dynamic changes so that we can better understand protein function.
  - These studies will involve mathematical modeling of protein structure using geometric analysis, statistics, machine learning, …

PROTEIN FUNCTION (1)

Protein function includes:

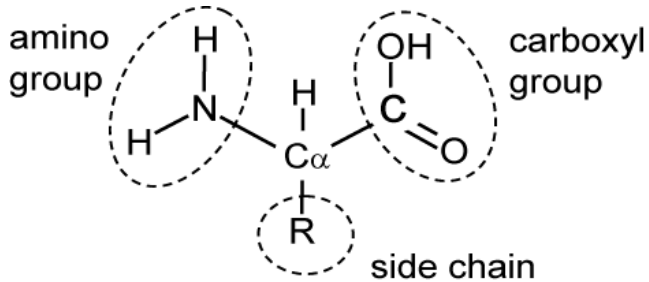
- Molecular movement
- Enzymatic catalysis
- Structural systems
- Signal transmission

CENTRAL DOGMA

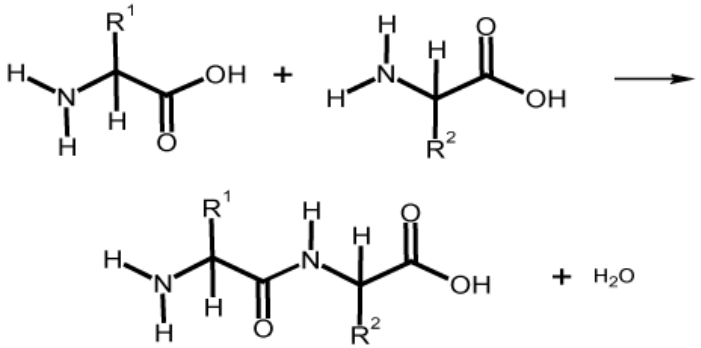


AMINO ACIDS

- The monomers of proteins are amino acids.
  - They have the following general form:



  - A peptide bond is formed by a condensation reaction:

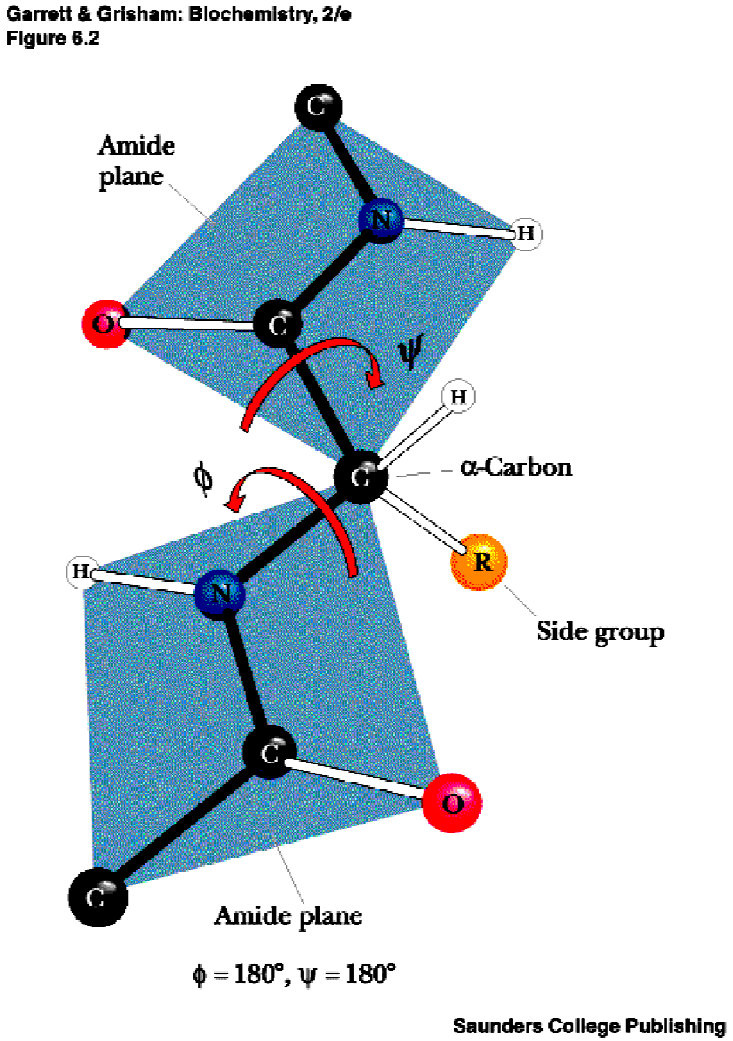


THE TOPIC OF THIS TALK

- Distance Geometry Problems:
  - There are various problems that come under this heading.
  - In general, we are given a set of distances between atoms and we are required to compute the coordinates of all atoms.
  - But first, we will look at protein conformation and its implications for the problem.

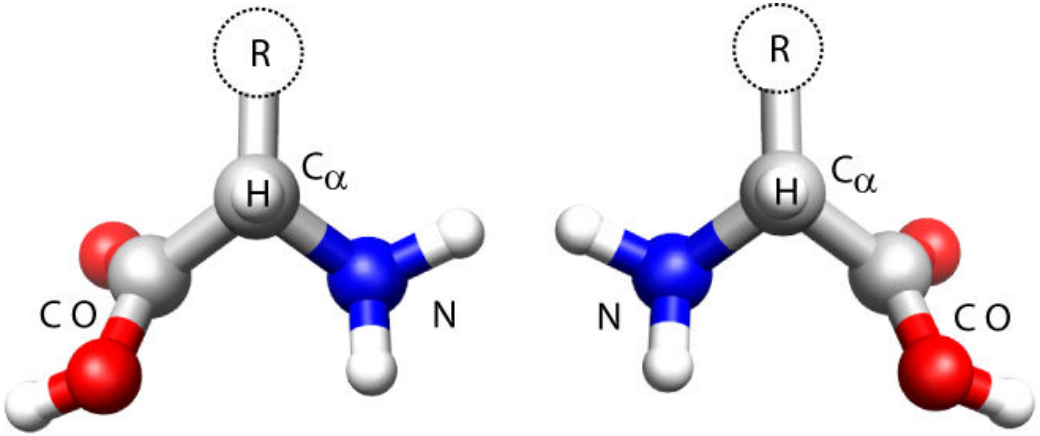
DIHEDRAL ANGLES

- Atoms in the "amide plane" tend to have bond lengths and bond angles with little variation.
  - As a first approximation: Backbone conformation is determined by the phi & psi dihedral angles.



PROTEIN CHIRALITY

- Distance information alone does not determine atomic positions with appropriate chirality.
  - An independent check is needed.
- Looking down the H-C_{alpha} bond (from H) the L-form can be read clockwise as “CORN”.
  - The D-form can be read as “NRCO”.
- Protein amino acids *always* have the L-form.

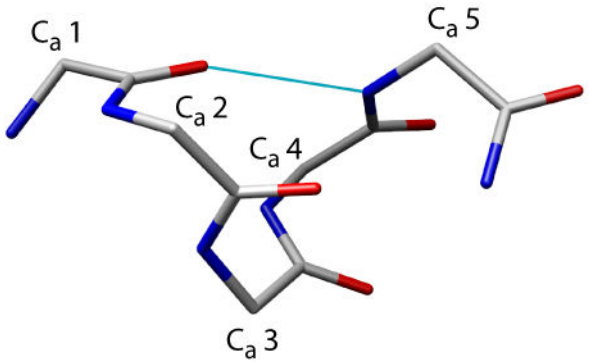


LEVELS OF STRUCTURE FOR PROTEINS

- Primary Structure
  - The primary structure is simply the sequence of amino acids.
- Secondary Structure
  - Secondary structure is described by categorizing the amino acids as being part of alpha helices, beta-sheets, or loops.
- Tertiary Structure
  - The helices and sheets combine to form a definite three dimensional conformation of the molecule.
- Quaternary Structure
  - Quaternary structure is specified by combining multiple tertiary structures (molecules) to form a working unit.
    - Not all proteins do this.

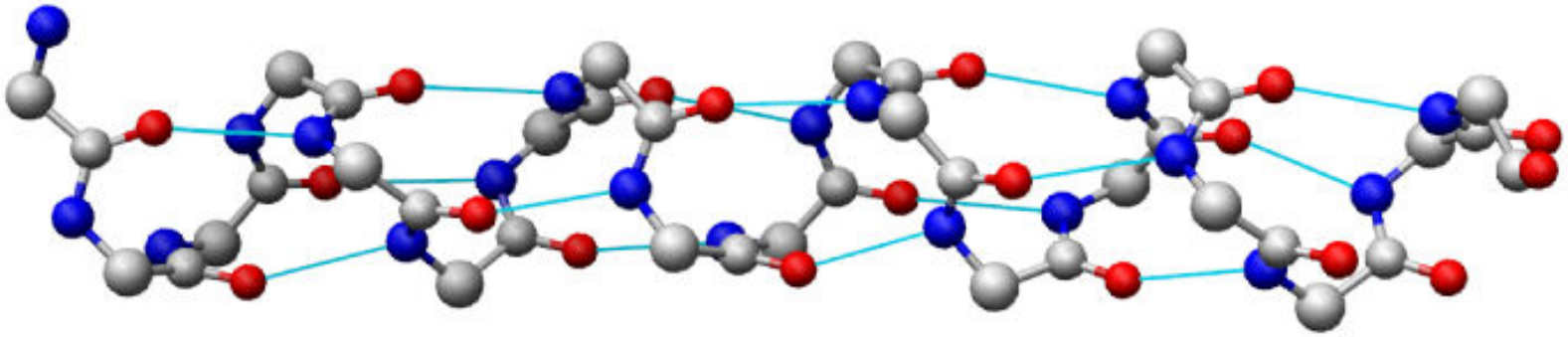
THE ALPHA HELIX

- The consecutive peptide planes twist into a helix.
- The side chains typically point outside the helix.
- The ideal alpha helix has 3.6 residues for every complete turn of the helix.
- Note the hydrogen bond between the H on the nitrogen atom and the double bonded oxygen of the downstream carbon atom.

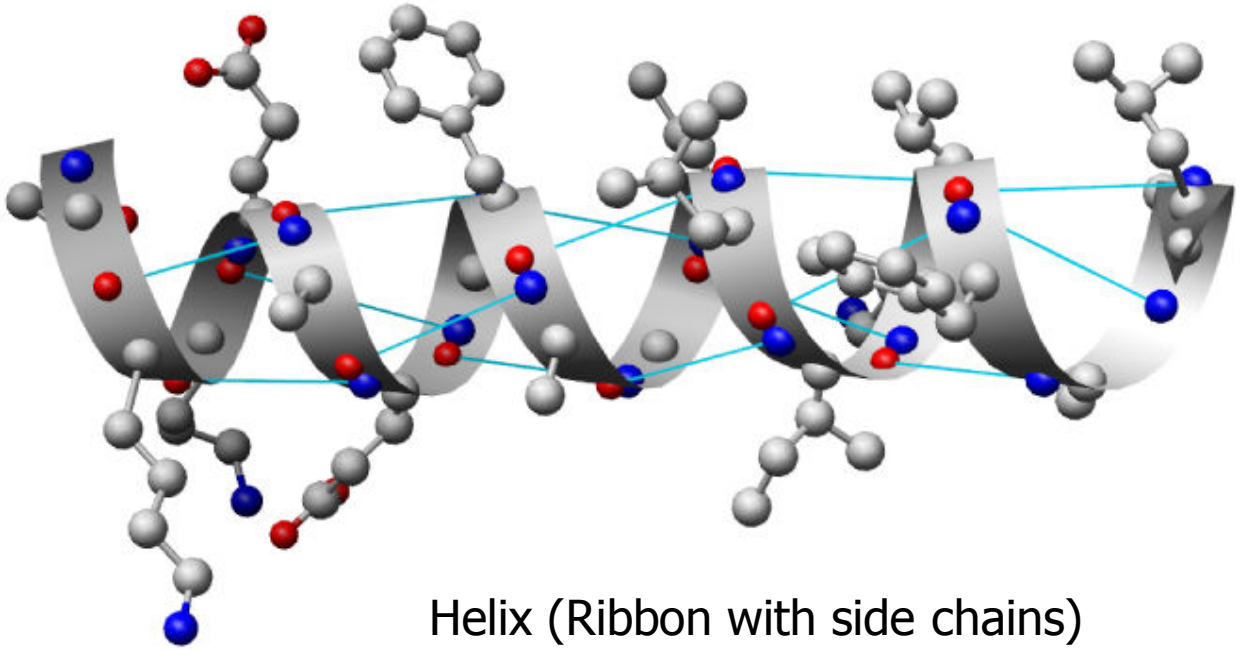


Most of the figures in this presentation were created using USDF Chimera: http://www.cgl.ucsf.edu/chimera/

HYDROGEN BONDS IN HELICES



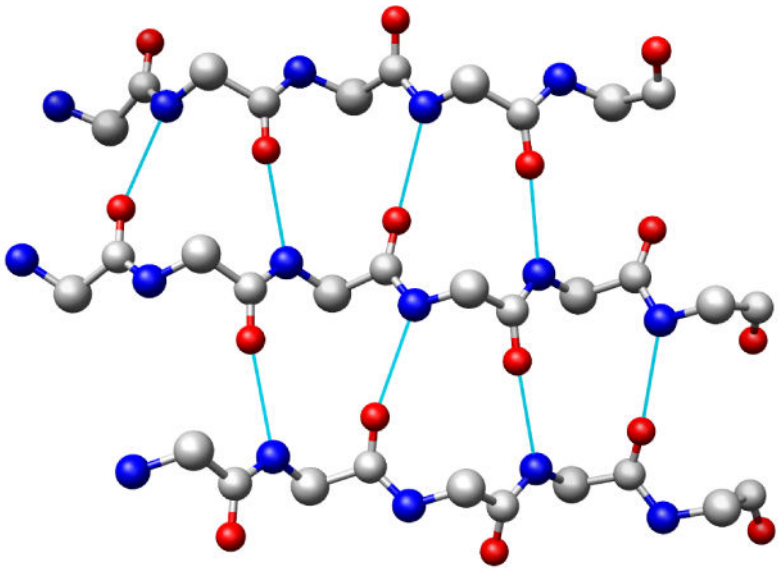
Helix (Backbone atoms only)



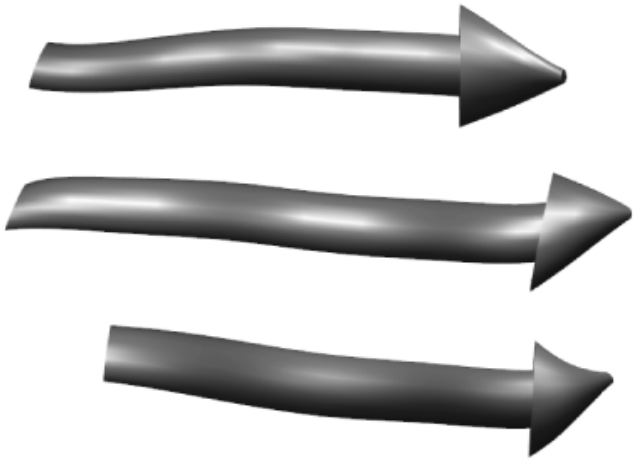
Helix (Ribbon with side chains)

THE BETA SHEET

- The beta-sheet is formed when peptide planes tend to align and form hydrogen bonding.
- Note the typical hydrogen bond between the H on the nitrogen atom and the double bonded oxygen of a carbon atom that is much more distant in the sequence.

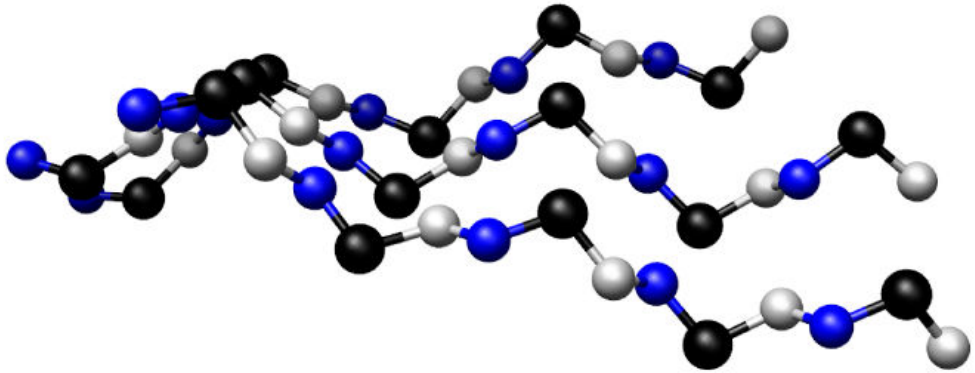


Parallel Beta Sheet

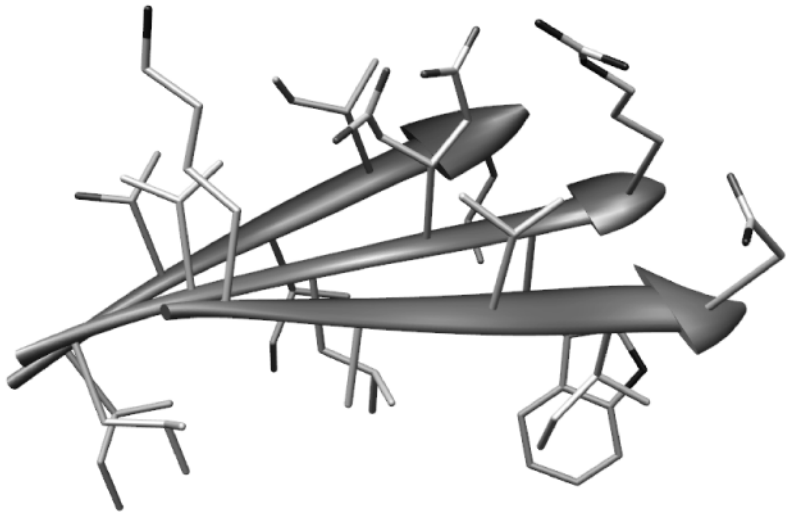


THE BETA-SHEET (CONT.)

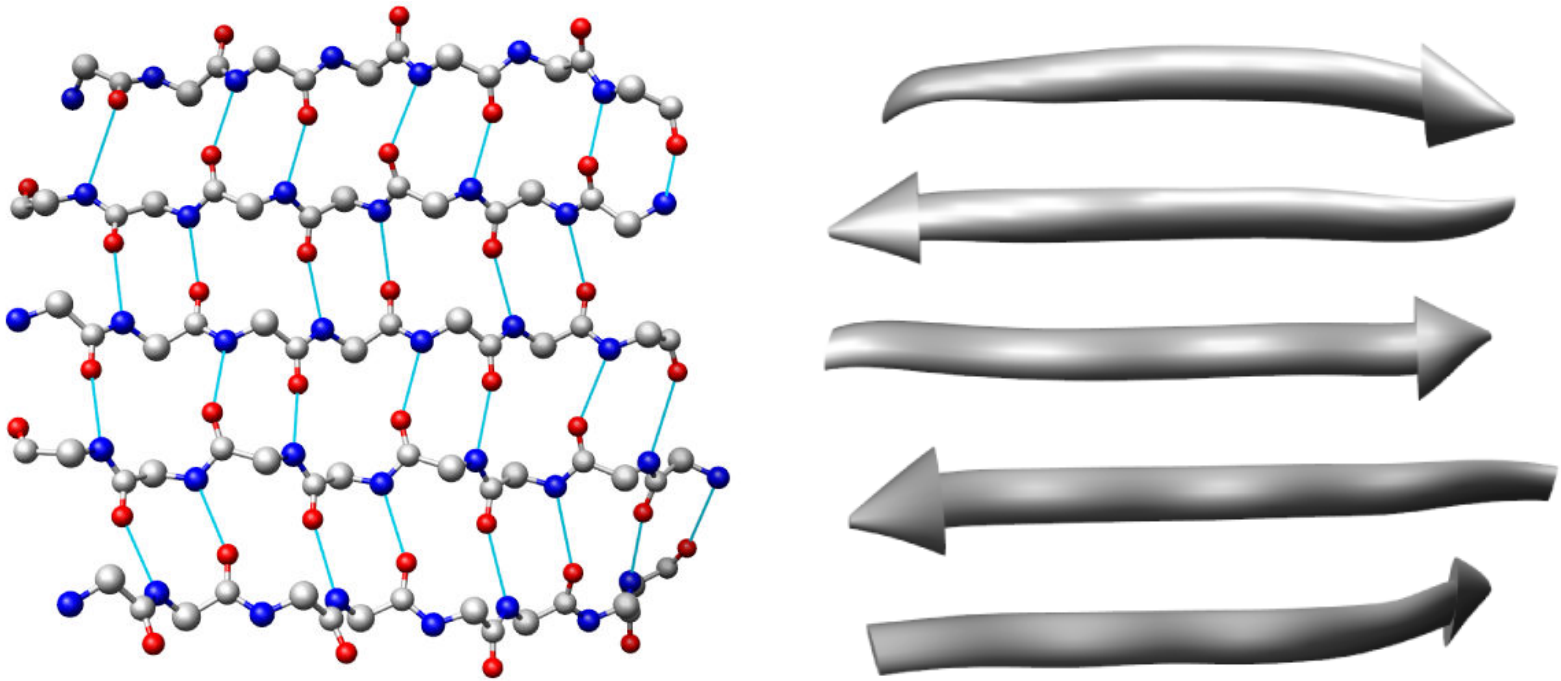
- Pleated conformation of a beta-sheet:



Same sheet
with side chains:



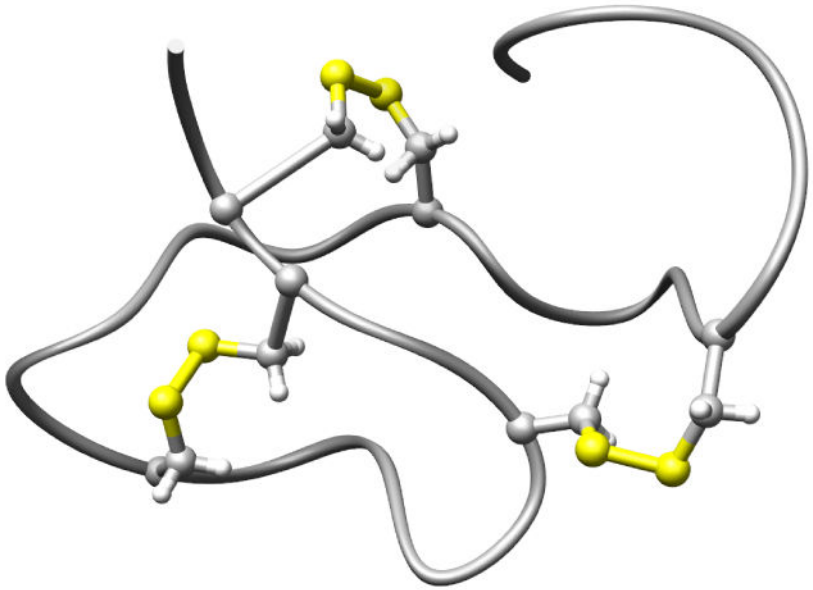
THE BETA-SHEET (CONT.)



Anti-Parallel Beta Sheet

LOOPS

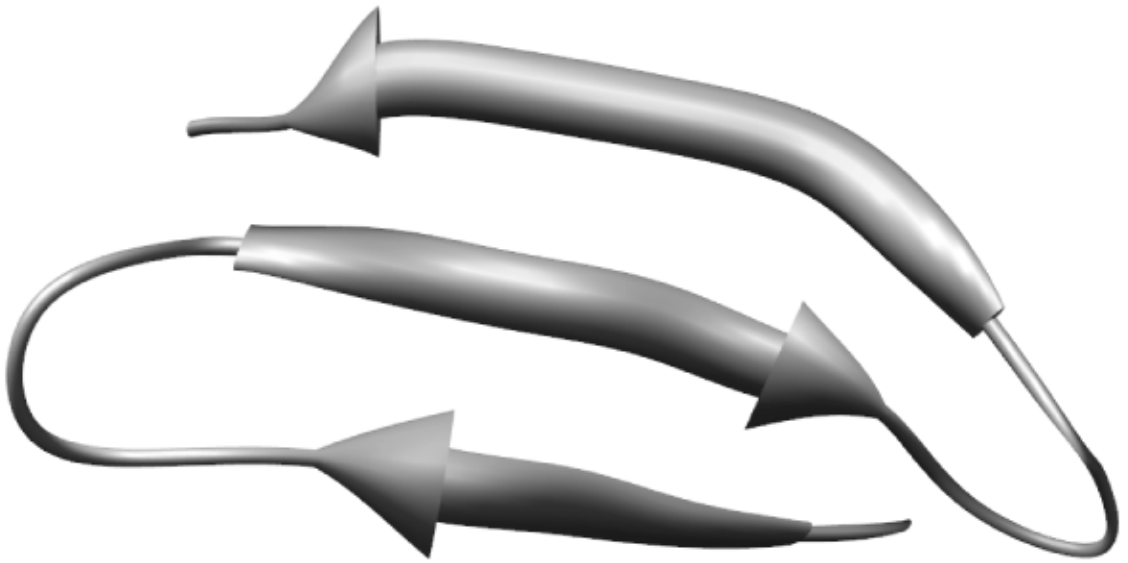
- Loops are chains of amino acids that have no particular hydrogen bonding patterns with other parts of the protein.



1ANS: a neurotoxin from the sea anemone, Anemonia sulcata.

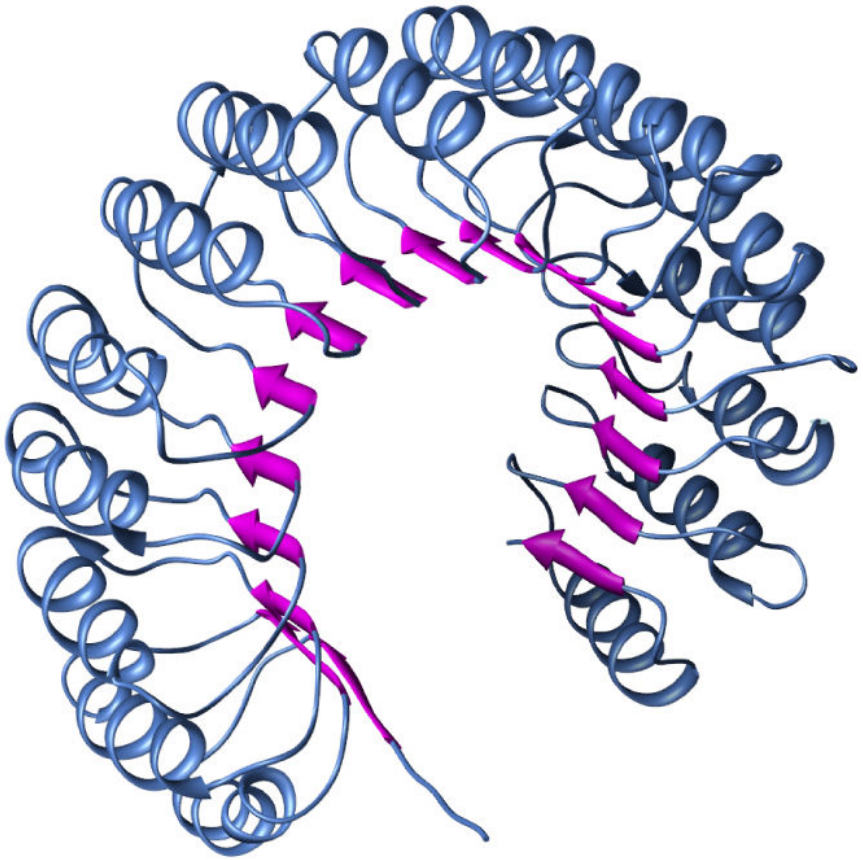
LOOPS (CONT.)

- Proteins that are “all loop” are fairly rare.
  - Most proteins will have beta sheets and helices forming a hydrophobic core and these secondary structures will be interconnected by loop segments.



TERTIARY STRUCTURE

- Helices, Sheets, and Loops combine to give a complete molecule in a three dimensional conformation:

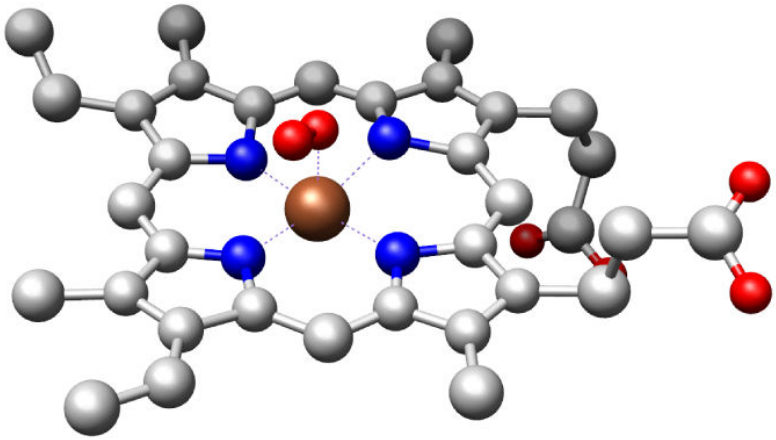


Chain D of 1A4Y

TERTIARY STRUCTURE: MYOGLOBIN (1)

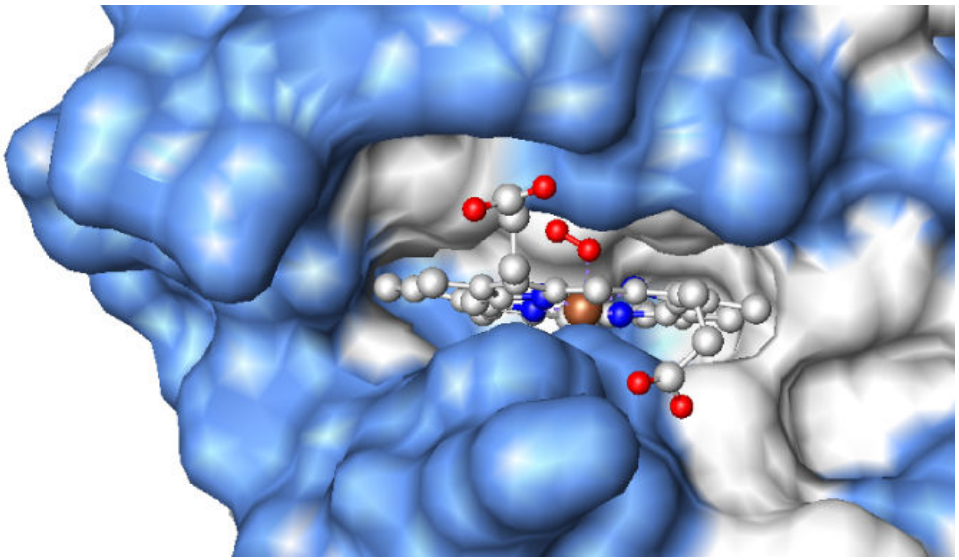
- A globin fold:
  - 1MBN: 153 residues in 8 helices with short loops forming a hydrophobic pocket containing a heme group.

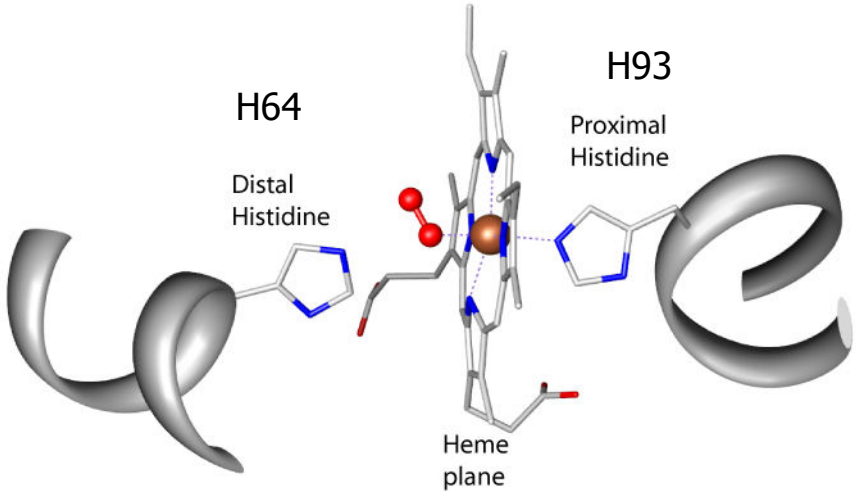
(A good example of structure supporting function).



TERTIARY STRUCTURE: MYOGLOBIN (2)

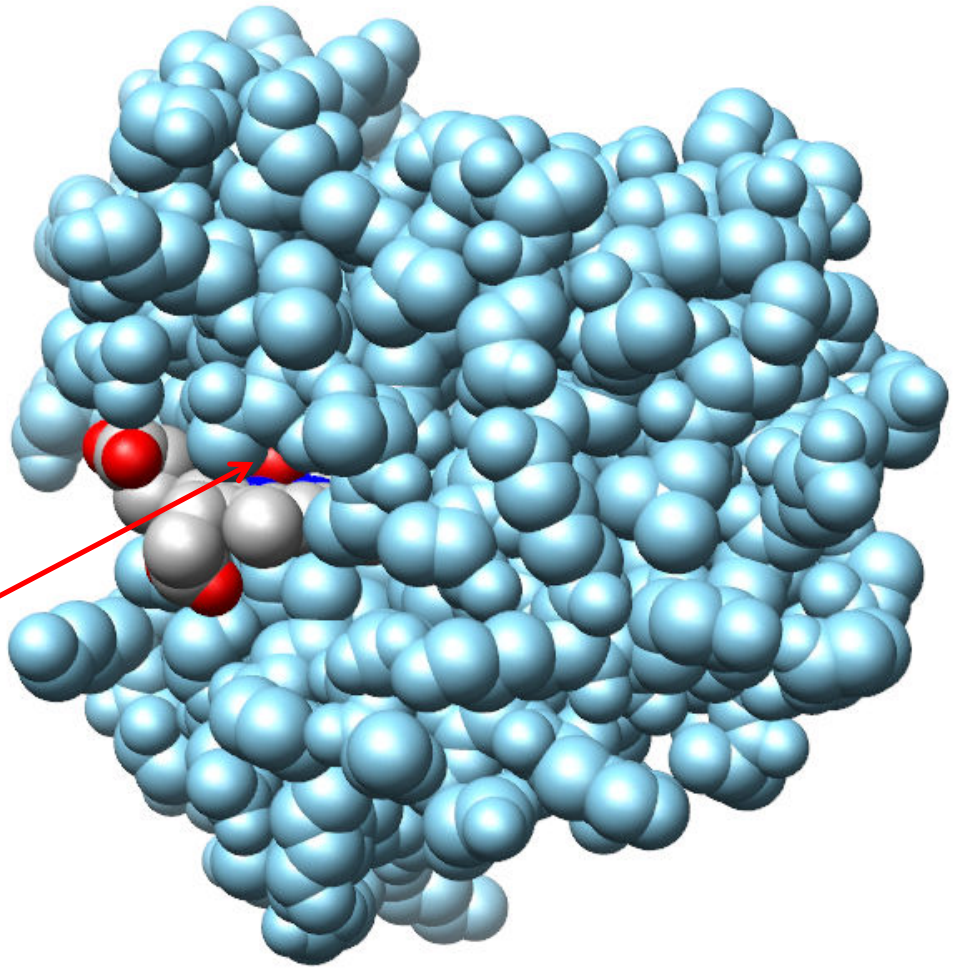
- Heme group in the globin pocket:





FLEXIBILITY & FUNCTIONALITY (1)

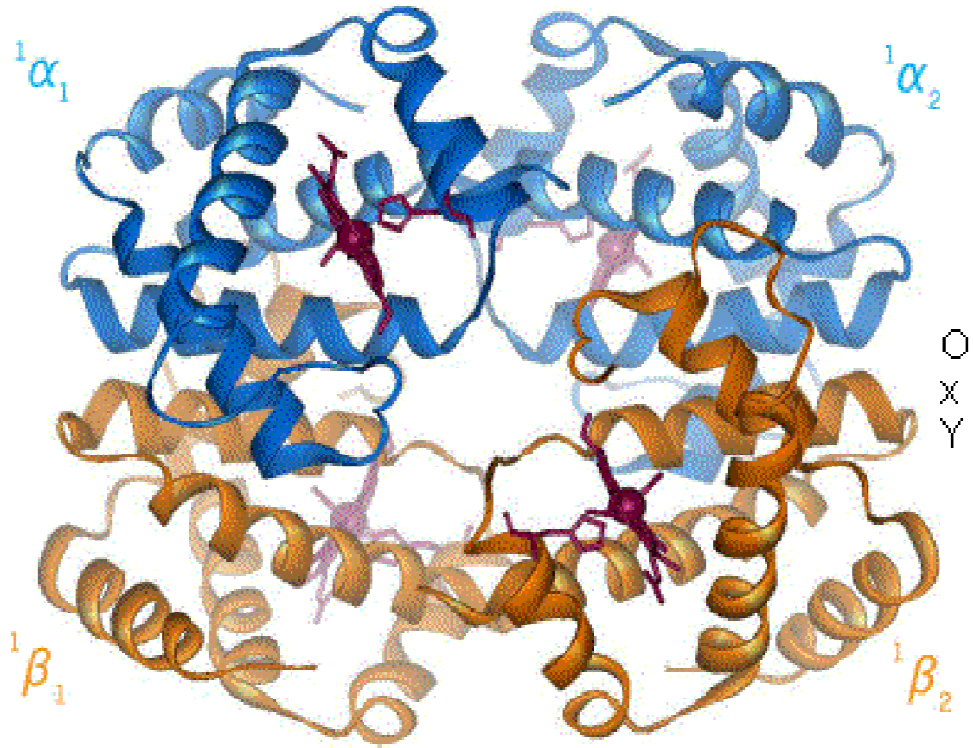
- Myoglobin will slightly change conformation when accepting or donating the O_2 molecule.



PDB ID: IMBN

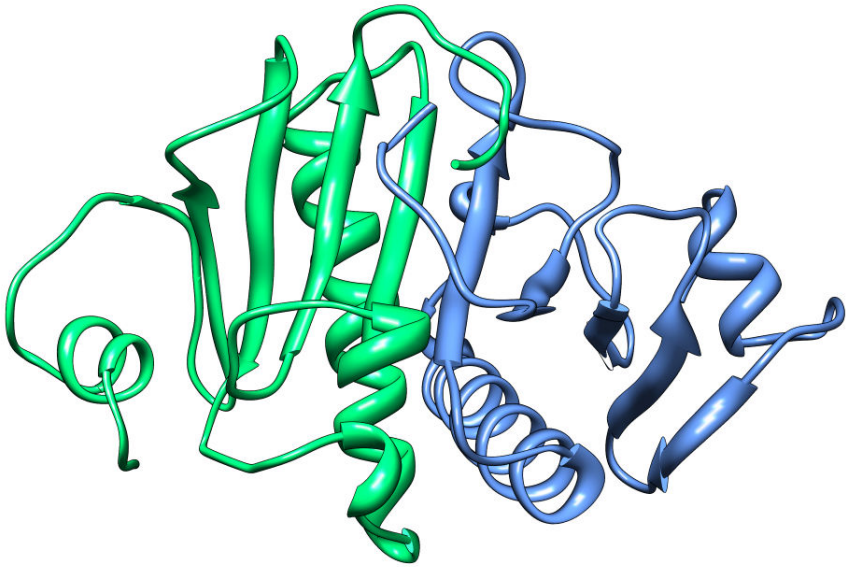
FLEXIBILITY & FUNCTIONALITY (2)

- Protein flexibility is important for protein functionality:



QUATERNARY STRUCTURE

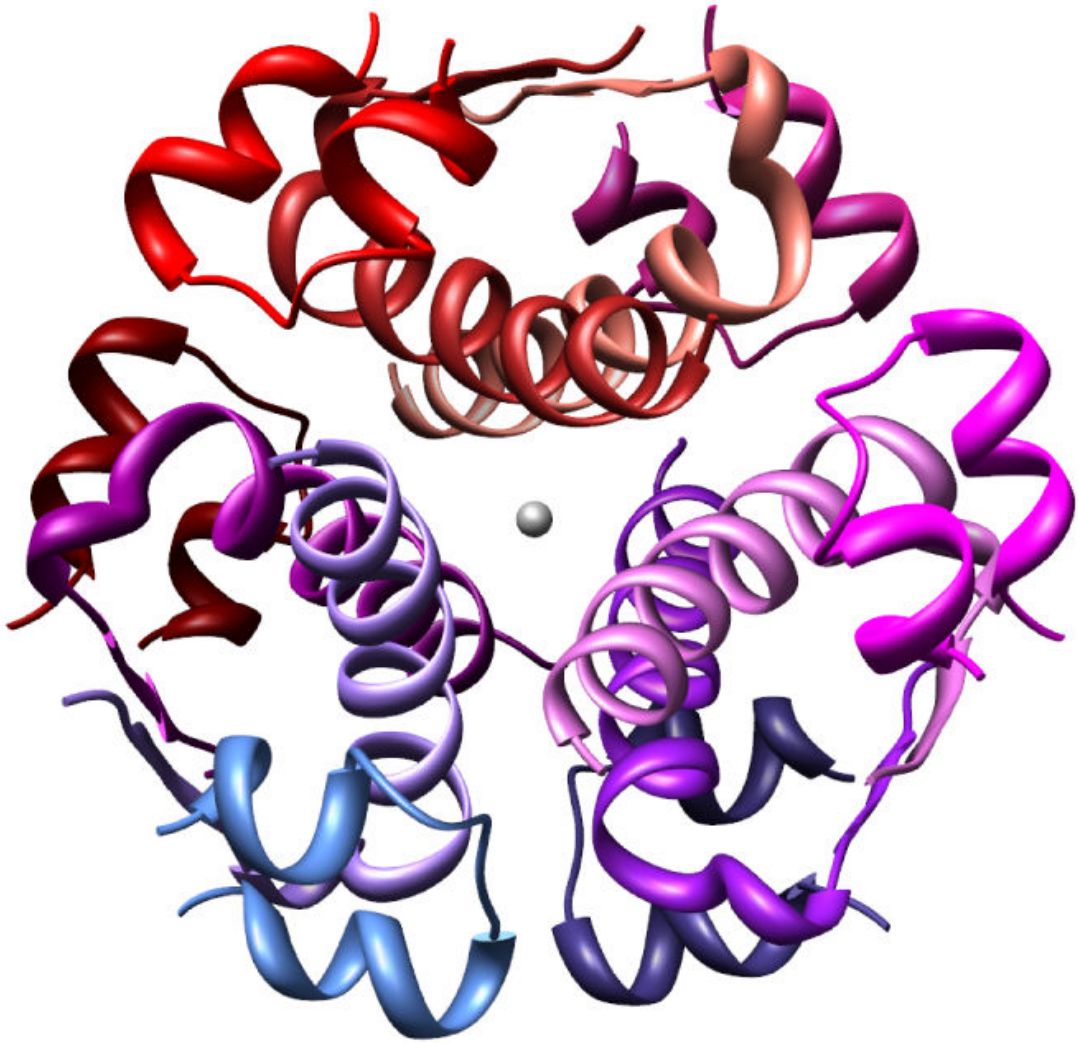
- Chains may combine to give a higher level structure.
  - Here we have the complete protein: Kinase C Interacting Protein (both chain A and chain B).
    - Often the tertiary components are replicates (as shown here).



1KPA

QUATERNARY STRUCTURE: INSULIN (4)

- Hexamer structure of insulin.
  PDB ID: 1znj



DIGRESSION (1)

- A Cyclotide protein as Sculpture
  - Kalata, 2004
    - Stainless steel, length 50" (1.30 m)
    - The protein Kalata is a small cyclic protein that has been recently found to be the utero-active component in a traditional African herbal medicine used to accelerate labor in childbirth.

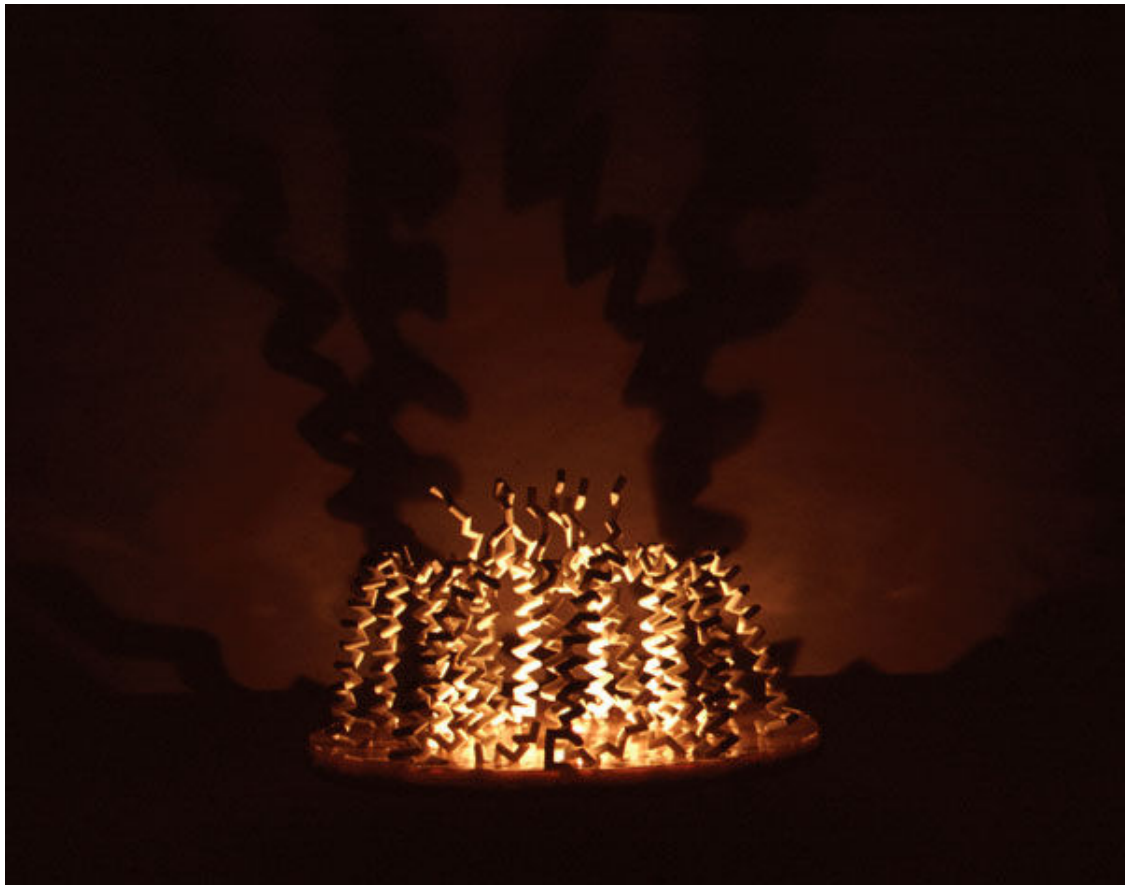
http://www.julianvossandreae.com/Work/protein6gallery/pages/Kalata.html

DIGRESSION (2)

- Voss-Andreae (quantum physicist & sculptor):
  - Heart of Steel (Hemoglobin) (2005)

DIGRESSION (3)

- Voss-Andreae :



Light-Harvesting Complex (2003)

Unravelling Collagen (2005)

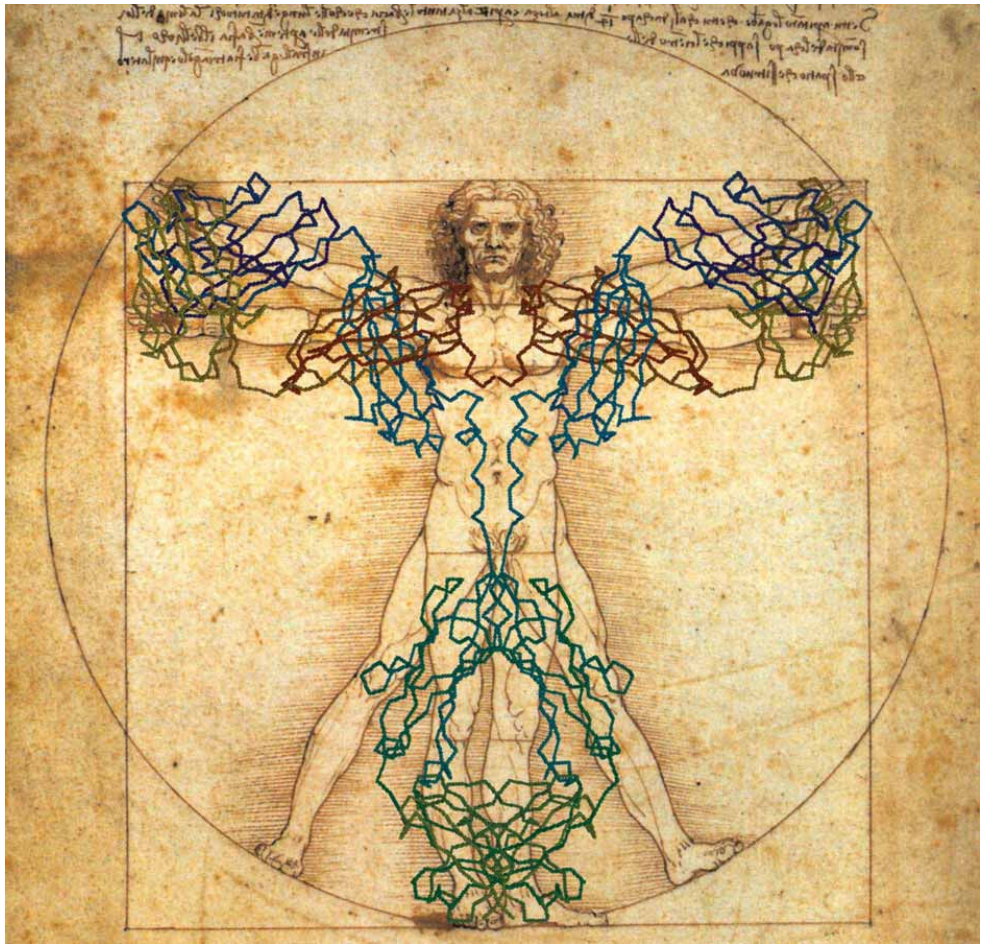
DIGRESSION (4)

- Voss-Andreae
  - Angel of the West (2008)

Scripps Research Institute

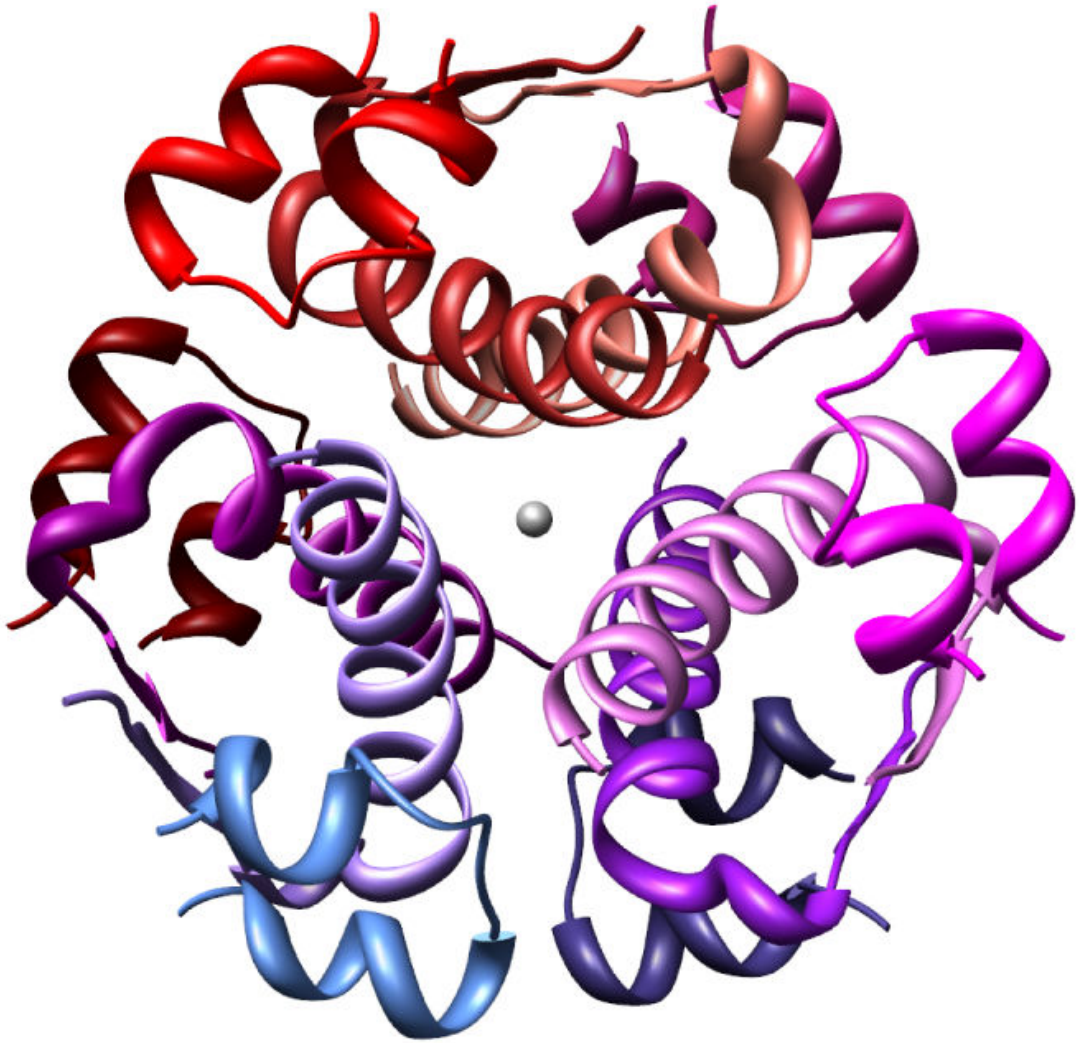
DIGRESSION (5)

- Voss-Andreae
  - Vitruvian Man & Antibody



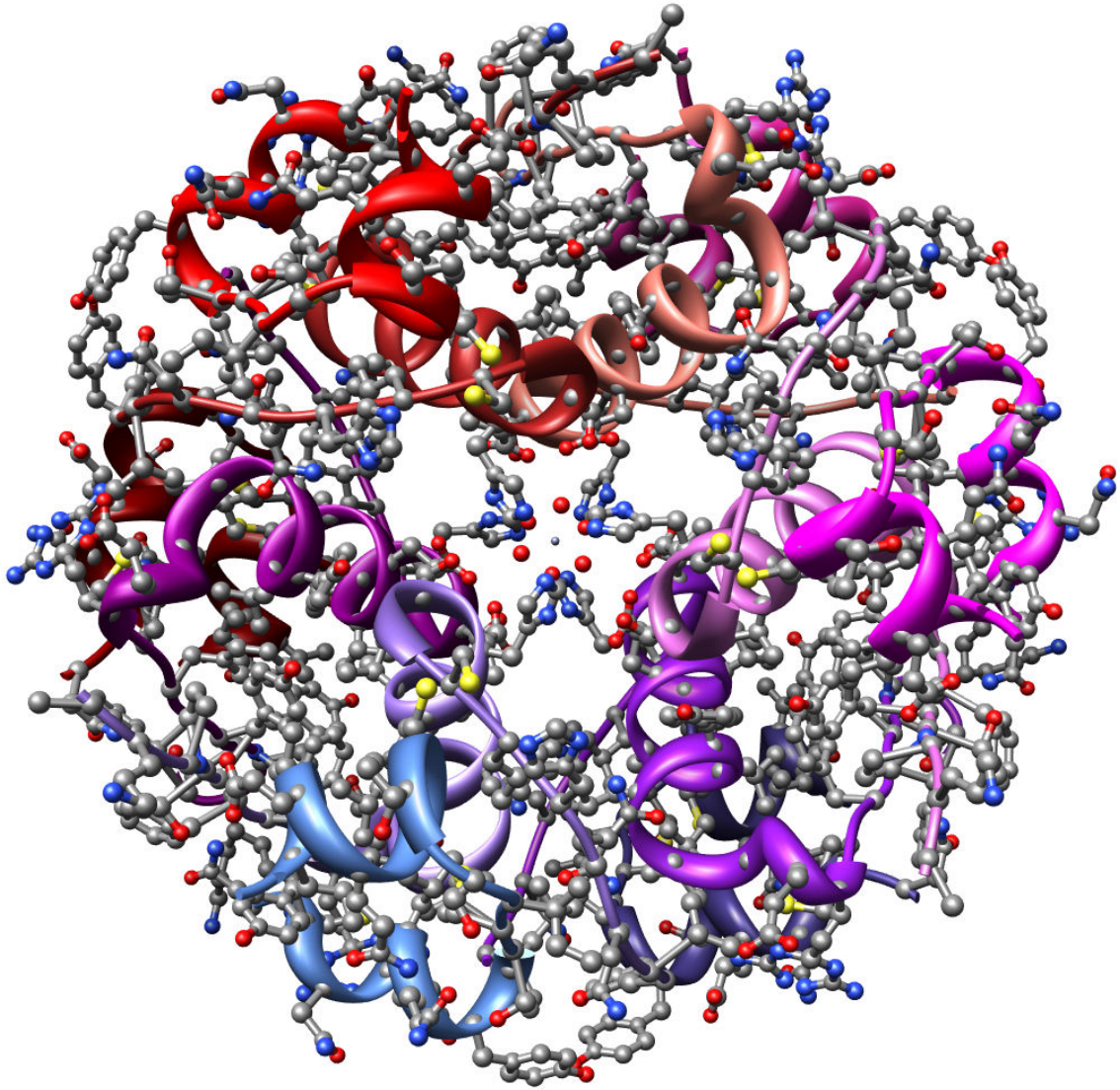
QUATERNARY STRUCTURE: INSULIN (4)

- Hexamer structure of insulin.
  PDB ID: 1znj



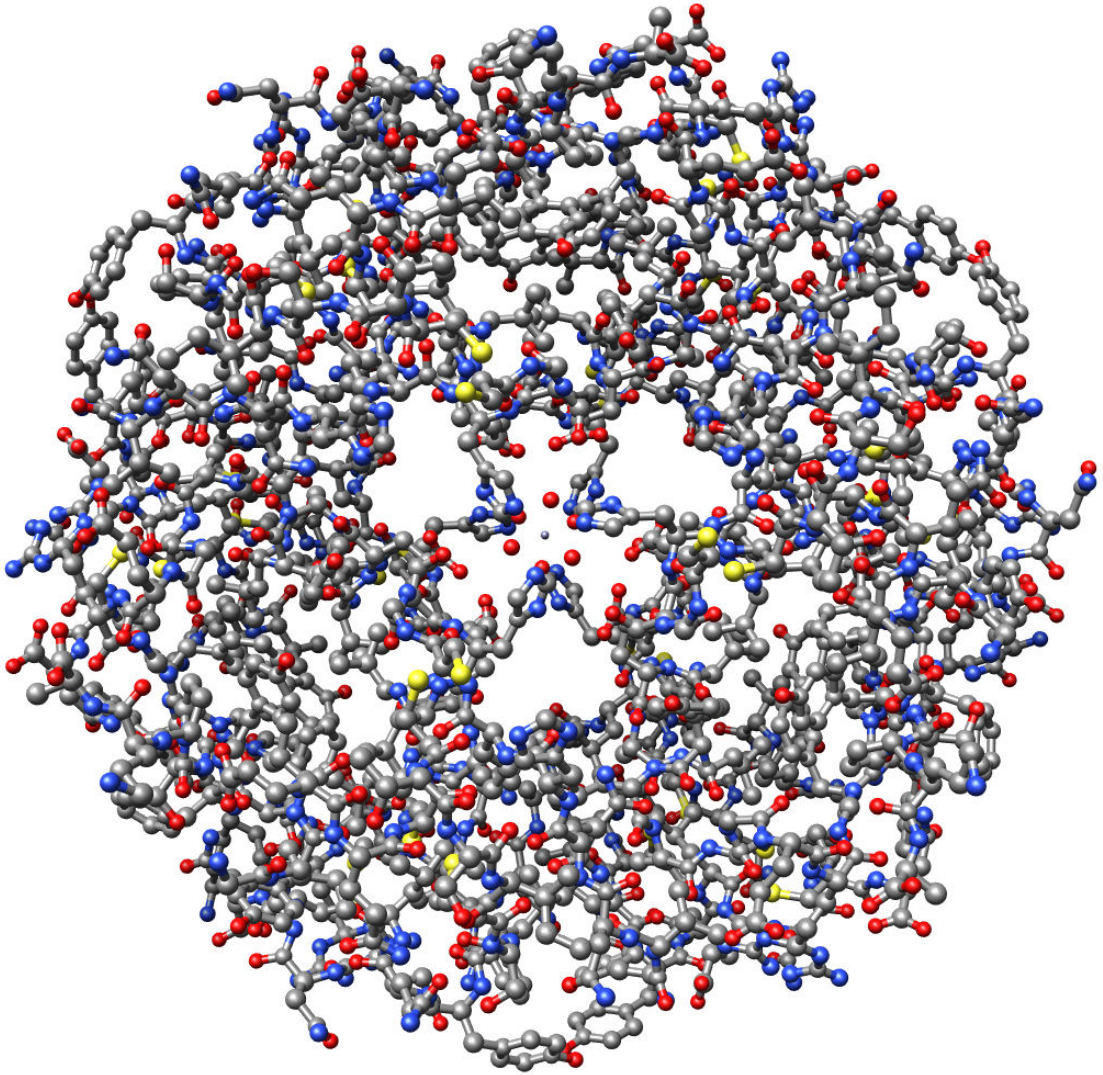
INSULIN: ALL ATOMS VISIBLE

- Hexamer structure of insulin.
  PDB ID: 1znj



QUATERNARY STRUCTURE: INSULIN (4)

- Hexamer structure of insulin.
  PDB ID: 1znj



INTRODUCTION (1)

- Distance Geometry Problems:
  - There are various problems that come under this heading.
  - In general, we are given a set of distances between atoms and we are required to compute the coordinates of all atoms.
  - Computed coordinates are typically with respect to a frame of reference that has its origin at the centroid of the atoms.
- Notation:
  - The coordinates of the atoms are $\left\{ x^{(i)} \middle| i = 1, 2, \ldots, n \right\}$.
    - So, we are dealing with a molecule that has n atoms.
    - We want to calculate these $x^{(i)}$, when given all or some subset of:
  - $d_{i,j} = \left\| x^{(i)} - x^{(j)} \right\| \quad 1 \le i, j \le n.$

INTRODUCTION (2)

- Variants of the problem differ with respect to:
  - Number of distances given:
    - The full set of "n choose 2" distances makes the problem fairly easy to solve.
    - In most practical applications we are only given an $O(n)$ sparse set of distances.
  - Accuracy of distances given:
    - Distances may be considered as exact, or
    - The problem may specify upper and lower bounds on distances.
    - Distance may be given as probability distributions.

INTRODUCTION (3)

- So, there are four types of problems:

  P1: A complete set of exact distances

  P2: A complete set of approximate distances

  P3: A sparse set of exact distances

  P4: A sparse set of approximate distances.

MOTIVATION (1)

- NMR
  - While most of the protein structures in the PDB have been computed using X-ray analysis, about 15% have been determined by NMR (Nuclear Magnetic Resonance).
  - Advantages:
    - Unlike X-ray analysis, NMR does not require crystals – the proteins may be in solution.
    - Consequently, we can have more confidence that their conformations are close to that present in the cytosolic environment.

MOTIVATION (2)

- NMR
  - Disadvantages:
    - NMR experiments only report distances between atoms.
      - The atoms must be close to one another (typically within 5 Å of each other).
        - This means the number of distances is much less than $\binom{n}{2}$.
      - These distances have some experimental error.
        - This means we have reduced accuracy.
    - NMR can only be used for shorter proteins.
      - "Short proteins" means less than a few hundred amino acids.

NOTATION

- We work in a 3D Euclidean vector space:
  - The coordinates of the atoms are $\left\{ x^{(i)} \middle| i = 1, 2, \ldots, n \right\}$.
    - So, we are dealing with a molecule that has n atoms.
    - We want to calculate these $x^{(i)}$, when given all or some subset of the inter-atomic distances:

$$d_{i,j} = \left\| x^{(i)} - x^{(j)} \right\| \quad 1 \le i, j \le n.$$

  - An n by n matrix X is used to store the $x^{(i)}$ vectors in a column by column fashion.
    - The n by n symmetric matrix D^2 holds the squares of distances:

$$\left\{ D^2 \right\}_{ij} = d_{ij}^2.$$

The Gram matrix G is defined by: $\left\{ G \right\}_{ij} = \left\langle x^{(i)}, x^{(j)} \right\rangle$.

We use $\hat{d}_{ij}$ to represent an approximation of the distance d_{ij} and we set

$$\left\{ \hat{D}^2 \right\}_{ij} = \hat{d}_{ij}^2.$$

RELATED WORK

- Given all the d_{ij}^2 values, we can compute a Gram matrix using the "double centering formula":

$$\frac{1}{2}\left[\frac{1}{n}\left(\sum_{i=1}^{n} d_{ij}^2 + \sum_{j=1}^{n} d_{ij}^2\right) - d_{ij}^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2\right] = \left\langle x^{(i)}, x^{(j)} \right\rangle = G_{ij}.$$

which can be rewritten as: $G = -\frac{1}{2} H D^2 H$

where: $H = I - \frac{1}{n}\bar{1}\bar{1}^T$

and $\bar{1}$ is an n-dimensional vector with each component equal to 1.

P1: DERIVING COORDINATES GIVEN G (1)

- Theorem:
  - Given matrix D^2, there exists a set of points $\left\{ x^{(i)} \middle| i = 1, 2, \ldots, n \right\}$ in $\mathbb{R}^k$ such that $d_{i,j} = \left\| x^{(i)} - x^{(j)} \right\| \quad 1 \le i, j \le n$ if and only if G is positive semidefinite with rank at most k. In this case $G = X^{\mathrm{T}} X$.
- To calculate X we start with the spectral decomposition of G:

$$G = \sum_{m=1}^{n} \lambda_m u^{(m)} u^{(m)T} \quad \Rightarrow \quad G_{ij} = \sum_{m=1}^{n} \lambda_m u_i^{(m)} u_j^{(m)}.$$

P1: DERIVING COORDINATES GIVEN G (2)

- Using the spectral decomposition of G:
  - For any fixed integer $k \in \{1, 2, \ldots, n\}$ we form the k by n matrix $Y(k)$ defined as:

$$Y(k) = \sqrt{\Lambda(k)}\left[U(k)\right]^{\mathrm{T}} \quad \Rightarrow \quad i^{\text{th}} \text{ column of } Y(k) \text{ is:}$$

$$\left[y(k)\right]^{(i)} = \left[\sqrt{\lambda_1}u_i^{(1)}, \sqrt{\lambda_2}u_i^{(2)}, \cdots, \sqrt{\lambda_k}u_i^{(k)}\right]^{\mathrm{T}}$$

where $\sqrt{\Lambda(k)}$ is the k by k matrix $diag\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_k}\right)$ and $U(k)$ is the n by k matrix with column entry equal to eigenvector $u(i)$.

P1: DERIVING COORDINATES GIVEN G (3)

- Then it can be shown that:

$$\left[y(k)\right]^{(i)\mathrm{T}}\left[y(k)\right]^{(j)} = \sum_{m=1}^{k} \lambda_m u_i^{(m)} u_j^{(m)} \quad \Rightarrow$$

$$G_{ij} - \left[y(k)\right]^{(i)T}\left[y(k)\right]^{(j)} = \sum_{m=k+1}^{n} \lambda_m u_i^{(m)} u_j^{(m)}.$$

  - If the given inter-atomic distances are all exact and consistent with a set of ***n*** atoms in a 3D Euclidean space then the rank of the Gram matrix will be 3 and all eigenvalues beyond λ_3 will be zero.
  - In this case, the last sum is zero.
  - Consequently, we can take X to be the 3 by n matrix $Y(3)$ and this becomes our solution for problem P1.

P2: DERIVING COORDS GIVEN NOISY G (1)

- When given a complete set of distances that have been subjected to a *small* amount of noise, we follow the strategy used by Trosset (1998), which reformulates the problem as follows:

$$\text{minimize } \left\| C - \hat{G} \right\|_F^2 \text{ subject to } C \in S_n^+(k)$$

where $\hat{G} = -\frac{1}{2} H \hat{D}^2 H$ and $S_n^+(k)$

is the set of non-negative n by n semidefinite matrices of rank k.

P2: DERIVING COORDS GIVEN NOISY G (2)

- They solve the minimization problem by using the following theorem:

  Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_n$ denote the eigenvalues of $\hat{G}$ with spectral decomposition $\hat{G} = \hat{U}\hat{\Lambda}\hat{U}^{\mathrm{T}}$ where

$$\hat{\Lambda} = diag\left(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_n\right).$$

  Define:

$$\tilde{\lambda}_i = \max\left\{0, \hat{\lambda}_i\right\} \text{ for } i = 1,2,\ldots, k \text{ and}$$
$$\tilde{\lambda}_i = 0 \qquad \text{for } i = k+1,,\ldots, n.$$

  Let $\tilde{\Lambda} = diag\left(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_n\right).$

  Then $C^* = \hat{U}\tilde{\Lambda}\hat{U}^{\mathrm{T}}$ is a global minimiser (and we can use the strategy for P1 with C^* replacing G).

P3: SPARSE BUT EXACT DISTANCES (1)

- For problem P3 we are given exact distances but only for atoms that are closer than some upper threshold.
- Geometric Buildup (Wu et al. 2003, 2007, 2008)
  - Find four atoms, not in the same plane, such that *all* inter-atomic distances are known.
  - Using the P1 strategy described earlier, derive the coordinates of all four atoms.
  - While there are atoms with undetermined positions repeat:
    - Find an atom with undetermined position but with known distances to four other non-coplanar atoms whose positions are known.
    - Determine the position of the undetermined atom using a triangulation strategy.

P3: SPARSE BUT EXACT DISTANCES (2)

- Triangulation:
  - Let $x^{(k_i)}\ i=1,2,3,4$ represent the coordinates of four non-coplanar atoms with known positions and $x^{(j)}$ holds undetermined coordinates of a nearby atom. Then:

$$\left\|x^{(k_i)}\right\|^2 - 2x^{(k_i)\mathrm{T}}x^{(j)} + \left\|x^{(j)}\right\|^2 = d^2_{k_i,j} \quad i=1,2,3,4$$

  - For each i = 1,2,3 we subtract equation i from equation i+1:

$$\left\|x^{(k_{i+1})}\right\|^2 - 2x^{(k_{i+1})\mathrm{T}}x^{(j)} + \left\|x^{(j)}\right\|^2 = d^2_{k_{i+1},j}$$

$$\left\|x^{(k_i)}\right\|^2 - 2x^{(k_i)\mathrm{T}}x^{(j)} + \left\|x^{(j)}\right\|^2 = d^2_{k_i,j}$$

$$\left\|x^{(k_{i+1})}\right\|^2 - \left\|x^{(k_i)}\right\|^2 - 2\left(x^{(k_{i+1})} - x^{(k_i)}\right)^{\mathrm{T}} x^{(j)} = d^2_{k_{i+1},j} - d^2_{k_i,j}$$

P3: SPARSE BUT EXACT DISTANCES (3)

- Rearranging terms:

$$\left(x^{(k_{i+1})} - x^{(k_i)}\right)^{\mathrm{T}} x^{(j)} = \frac{1}{2}\left(\left\|x^{(k_{i+1})}\right\|^2 - \left\|x^{(k_i)}\right\|^2 - d^2_{k_{i+1},j} + d^2_{k_i,j}\right) \quad i = 1,2,3.$$

- The three equations in matrix form:

$$Bx^{(j)} = c$$

where:

$$B = \begin{bmatrix} \left(x^{(k_2)} - x^{(k_1)}\right)^{\mathrm{T}} \\ \left(x^{(k_3)} - x^{(k_2)}\right)^{\mathrm{T}} \\ \left(x^{(k_4)} - x^{(k_3)}\right)^{\mathrm{T}} \end{bmatrix} \qquad c = \frac{1}{2}\begin{bmatrix} \left\|x^{(k_2)}\right\|^2 - \left\|x^{(k_1)}\right\|^2 - d^2_{k_2,j} + d^2_{k_1,j} \\ \left\|x^{(k_3)}\right\|^2 - \left\|x^{(k_2)}\right\|^2 - d^2_{k_3,j} + d^2_{k_2,j} \\ \left\|x^{(k_4)}\right\|^2 - \left\|x^{(k_3)}\right\|^2 - d^2_{k_4,j} + d^2_{k_3,j} \end{bmatrix}$$

P3: SPARSE BUT EXACT DISTANCES (4)

- Triangulation issues:
  - If only three atoms with known positions were used then the position of the undetermined atom would be ambiguous (Fig. 1(a) or Fig. 1(b)?)
  - We need four atoms but they cannot be coplanar.
  - Even if they are “almost coplanar” we have a problem. An ill-conditioned system is very susceptible to noise.

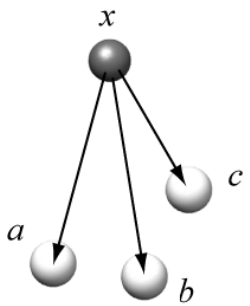


Fig. 1(a)

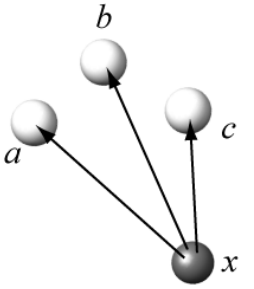


Fig. 1(b)

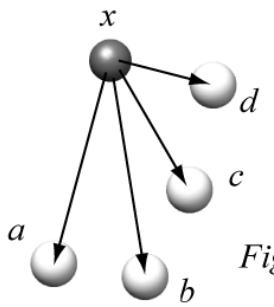


Fig. 1(c)

P4: SPARSE AND NOISY DISTANCES (1)

- Problem P4: we are given approximate distances for atoms that are closer than some upper threshold.
  - As long as the distances are positive there is always a solution, but it may reside in a space with dimension higher than 3.
  - If the given distances are true 3D distances with a small amount of error then the first three eigenvalues will be quite different from 0 and the fourth and later eigenvalues will be quite close to 0.
  - Simply ignoring these values (in effect, projecting from a high dimension space down to 3D) will produce an approximate solution typically characterized as “crowded” since contributions to a distance are being ignored.

P4: SPARSE AND NOISY DISTANCES (2)

- Related strategies:
  - To avoid the crowding issue, Biswas, Toh, and Ye (2008) formulate the problem as an optimization problem:

$$\text{maximize} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\left\|x^{(i)}-x^{(j)}\right\|^2$$

$$\text{s.t.} \quad \underline{d}_{ij}^2 \le \left\|x^{(i)}-x^{(j)}\right\|^2 \le \bar{d}_{ij}^2 \qquad \forall (i,j)\in[1..n]$$

$$\sum_{i=1}^{n} x^{(i)} = 0$$

P4: SPARSE AND NOISY DISTANCES (3)

- Related strategies:
  - They then cast the problem as semidefinite programming (SDP) relaxation:

$$\begin{aligned} \text{maximize} \quad & \left\langle I - \left(ee^{\mathrm{T}}/n\right), Y \right\rangle \\ \text{s.t.} \quad & \underline{d}_{ij}^2 \le e_{ij}^{\mathrm{T}} Y e_{ij} \le \bar{d}_{ij}^2 \qquad \forall (i,j) \in [1..n] \\ & Ye = 0, \quad Y \succeq 0 \end{aligned}$$

where $\boldsymbol{e}$ is the vector of all ones, and e_{ij} is the zero vector with entry $\boldsymbol{i}$ changed to +1 and entry $\boldsymbol{j}$ changed to -1.

P4: SPARSE AND NOISY DISTANCES (4)

- Biswas, Toh and Ye (continued):
  - The SDP strategy, just described, is a heuristic that tries to deal with the noise issue.
    How do they handle the sparseness of the data?
    - Atoms are grouped into clusters such that all inter-atomic distances are known for all possible atom pairs in a cluster.
      - They then use the SDP approach to derive the coordinates of atoms in each cluster.
    - Stitching: Since each cluster will have its own frame of reference, it is necessary to do translate and rotate operations that will bring one set of atoms into the same frame of reference as the neighbouring cluster of atoms.
      - This process is continued until all the atoms are in the same frame of reference.

P4: MDS & OVERLAPPING CLIQUES (1)

- We now describe our strategy for handling the P4 problem (sparse and noisy distances).
- Clique formation:
  - Start by grouping neighbouring atoms to form “cliques”:
    - All inter-atomic distances are known for the atoms in a clique.
    - Recall that we are given all inter-atomic distances less that some threshold (say, τ Angstroms).
    - Each clique will surround a particular atom called the clique center.

P4: MDS & OVERLAPPING CLIQUES (2)

- A clique formation heuristic
- Suppose atom A is to be a clique center.
  - Place all atoms that are within $\tau / 2$ Angstroms of *A* into an initially empty clique set.
  - Collect all the atoms that are within τ Angstroms but beyond $\tau / 2$ Angstroms from atom *A*, and sort them in ascending order with respect to their distance from *A*.
  - Go through the sorted list formed in step (b) and add an atom to the clique set if the input data includes all inter-atomic distances between that atom and every current member of the clique set .

P4: MDS & OVERLAPPING CLIQUES (3)

- Choosing clique centers
- Clique centers are chosen so that the clique has biological relevance. Each amino acid provides centers for two cliques:
  - A clique centered on the alpha carbon atom:
    - Since the clique center is also a chiral center, we can be sure that the computed coordinates have the appropriate chirality.
  - A clique centered on the carbonyl oxygen atom.
    - This clique overlaps the alpha carbon clique and also includes the hydrogen bonds responsible for helix and strand formation.

P4: MDS & OVERLAPPING CLIQUES (4)

- Calculating atomic positions
- For each clique there is a full set of distance values and so we can use the P2 algorithm discussed earlier.
  - Sometimes called an MDS (Multidimensional Scaling) strategy.
- The coordinates of all atoms in a clique will be relative to a frame of reference that is only suitable for that clique.
  - We need to modify coordinates so that all atoms are in the same frame of reference.

P4: MDS & OVERLAPPING CLIQUES (5)

- Combining cliques
- When cliques overlap, with at least 4 atoms in their intersection, we may combine them:
  - This involves a translate and rotate of the second atom set so that both sets have the frame of reference used by the first atom set.
  - The atoms in the intersection will define the appropriate translate and rotate operations.
  - The intersection atoms cannot be coplanar.

P4: MDS & OVERLAPPING CLIQUES (6)

- Combining cliques (continued)
- Recall that we are using $C^* = \hat{U}\tilde{\Lambda}\hat{U}^{\mathrm{T}}$ where $\tilde{\Lambda} = diag\left(\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{\lambda}_3, 0, \ldots, 0\right)$ and $\tilde{\lambda}_i = \max\left\{0, \hat{\lambda}_i\right\} \;\; i = 1, 2, 3.$
  - Noise in the given distance data will increase the magnitude of $\hat{\lambda}_4$ and thus compromise the ability of C^* to yield the "true" 3D coordinates.
  - Consequently, the positions of atoms in the intersection of two cliques will not be precise.

P4: MDS & OVERLAPPING CLIQUES (7)

- Combining cliques (continued)
  - Repeating: Because of noise in the distance data, the computed positions of atoms in the intersection of two cliques are not exact:
    - The translate and rotate operations when doing clique combining can still proceed because the super-positioning of atoms in the intersection is done in the least squares sense.
    - However, it is still necessary to determine the final positions of atoms in the intersection.

P4: MDS & OVERLAPPING CLIQUES (8)

- Modifying distance estimates:
- Before getting both cliques into the same frame of reference we can try to reduce distance errors by averaging:
  - We can recalculate the squares of distances by using the C^* matrix:

$$\left(d^*_{i,j}\right)^2 = C^*_{i,i} - 2C^*_{i,j} + C^*_{j,j}.$$

P4: MDS & OVERLAPPING CLIQUES (9)

- Modifying distance estimates (continued):
- For any pair of atoms (indexed by i, j) within the intersection, each clique will have a different value of $\left(d^{*}_{i,j}\right)^{2}$.
  - We can try to reduce distance errors by computing an average and then replacing all such distances with the average distance.
  - After this is done, coordinates are computed and one clique can be brought into the frame of reference of the other clique.

P4: MDS & OVERLAPPING CLIQUES (10)

- Averaging distances:
  - Using this strategy we can continue to combine cliques until all atoms are in the same frame of reference. Some results:

PDB ID	Atom Count	Residue Count	RMSD σ = 0.05	RMSD σ = 0.10	RMSD σ = 0.15	RMSD σ = 0.20	RMSD σ = 0.25
1HOE	581	97	0.041	0.139	0.307	0.543	1.123
1LFB	641	78	0.045	0.150	0.321	0.544	0.817
1POA	1067	271	0.096	0.412	0.705	0.993	1.638
1HSG	1677	317	0.056	0.228	0.553	1.080	1.741
1RGS	2059	266	0.141	0.444	1.022	1.985	3.693
1BPM	3673	483	0.124	0.404	0.901	1.596	2.514
1TIM	3740	494	0.197	0.662	1.100	3.723	9.878
1HQQ	4116	700	0.279	0.308	0.905	4.267	6.506

Here: $\hat{d}_{ij} = d_{ij} \max\left(0, 1 + N\left(0, \sigma^2\right)\right).$

P4: MDS & OVERLAPPING CLIQUES (11)

- Weighted averages:
- "Not all cliques are created equal":
  - We can use available information to roughly assess the validity of the C^* matrix for a clique (for example, using the magnitude of $\hat{\lambda}_4$).
  - Our current experiments involve weighting schemes to compute a weighted averaging of $\left(d^*_{i,j}\right)^2$ values (the clique with a smaller $\hat{\lambda}_4$ is given more weight in the calculation).

P4: MDS & OVERLAPPING CLIQUES (12)

- Calculating positions of atoms in the intersection:
- After getting both cliques into the same frame of reference:
  - Each clique will determine a particular position for any atom in the intersection. Which one do we use?
    - Assuming that both cliques have equal validity in the calculated positions, we can simply compute the midpoint of the two positions.
    - Or as with distances we can choose a weighted average to get the final position.