Determining Protein Structure from Sequence using Computational Approaches

M. Saleet Jafri

Program in Bioinformatics and Computational Biology
George Mason University
and
Medical Biotechnology Center
University of Maryland Biotechnology Institute

Protein Structure – Why do we care?

- Structure Function Relation – The shape of a protein molecule directly determines its biological function.
- Proteins with similar function often have similar shape or similar regions or domains.
- Hence, if we find a new protein and know it's shape, we can make a good guess about it's biological function.

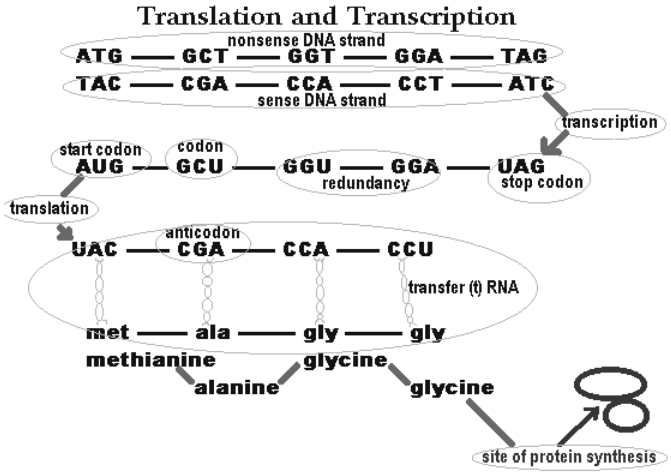
Protein Databases

- As of June 2000, 12,500 protein structures have been deposited into the Protein Data Bank (PDB) and 86,500 protein sequence entries were contained in SwissProt protein sequence database.
- This is a 1:7 ratio – relatively few structures are known.
- The number of sequence will increase much faster than the number of structures due to advances in sequencing.

Protein Basic Structure

- A protein is made of a chain of amino acids.
- There are 20 amino acids found in nature
- Each amino acid is coded in the DNA by one or more codons, i.e. a three base sequence.

Transcription and Translation



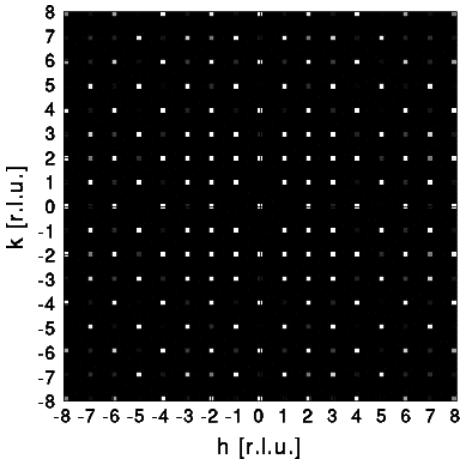
From http://www.agen.ufl.edu/~chyn/age2062/lect/lect_07/of7_1a.GIF

Finding the Protein Sequence

- From DNA sequence
- From protein sequencer
- From mRNA sequence

Measuring Protein Structure

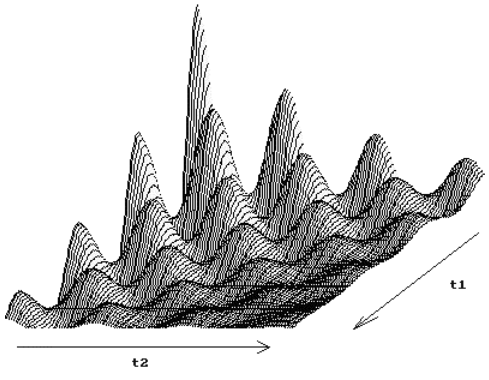
- Determining protein structure directly is difficult
- X-ray diffraction studies – must first be able to crystallize the protein and then calculate its structure by the way it disperses X-rays.



From http://www.uni-wuerzburg.de/mineralogie/crystal/teaching/inv_a.html

Measuring Protein Structure

- NMR – Use nuclear magnetic resonance to predict distances between different functional groups in a protein in solution. Calculate possible structures using these distances.



http://www.cis.rit.edu/htbooks/nmr/inside.htm

Why not stick to these methods?

- X-ray Diffraction –
  - Only a small number of proteins can be made to form crystals.
  - A crystal is not the protein's native environment.
  - Very time consuming.
- NMR Distance Measurement –
  - Not all proteins are found in solution.
  - This method generally looks at isolated proteins rather than protein complexes.
  - Very time consuming.

Four Levels of Protein Structure

- Primary Structure – Sequence of amino acids
- Secondary Structure – Local Structure such as ? -helices and ? -sheets.
- Tertiary Structure – Arrangement of the secondary structural elements to give 3-dimensional structure of a protein
- Quaternary Structure – Arrangement of the subunits to give a protein complex its 3-dimensional structure.

Predicting Protein Structure from the Amino Acid Sequence

- Goal: Predict the 3-dimensional structure of a protein from the sequence of amino acids (primary structure).
- Sequence similarity methods predict secondary and tertiary structure based on homology to know proteins.
- Secondary structure predictions methods include Chou-Fasman, GOR, neural network, and nearest neighbor methods.
- Tertiary structure prediction methods include energy minimization, molecular dynamics, and stochastic searches of conformational space.

Sequence similarity methods

- These methods can be very accurate if there is > 50% sequence similarity.
- They are rarely accurate if the sequence similarity < 30%.
- They use similar methods as used for sequence alignment such as the dynamic programming algorithm, hidden markov models, and clustering algorithms.

Secondary Structure Prediction Algorithms

- These methods are 70-75% accurate at predicting secondary structure.
- A few examples are
  - Chou Fasman Algorithm
  - Garnier-Osguthorpe-Robson (GOR) method
  - Neural network models
  - Nearest-neighbor method

Chou-Fasman Algorithm

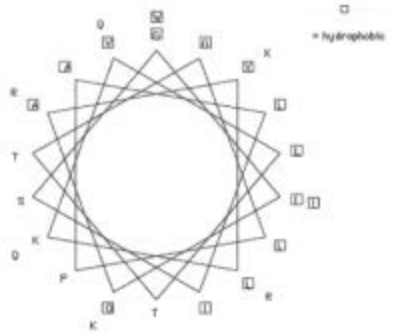
- Analyzed the frequency of the 20 amino acids in ? helices, ? sheets and turns.
- Ala (A), Glu (E), Leu (L), and Met (M) are strong predictors of ? helices.
- Pro (P) and Gly (G) break ? helices.
- When 4 of 5 amino acids have a high probability of being in an ? helix, it predicts a ? helix.
- When 3 of 5 amino acids have a high probability of being in a ? strand, it predicts a ? strand.
- 4 amino acids are used to predict turns.

Garnier-Osguthorpe-Robson Method

- Chou-Fasman assumes that each individual amino acid influences secondary structure.
- GOR assumes the the amino acids flanking the central amino acid also influence the secondary structure.
- Hence, it uses a window of 17 amino acids (8 on each side of the central amino acid).
- Each amino acid in the window acts independently on influencing structure (to save computational time).
- Certain pair-wise combinations of amino acids in the window also contribute to influencing structure.

Hydrophobicity/Hydrophilicity Plots

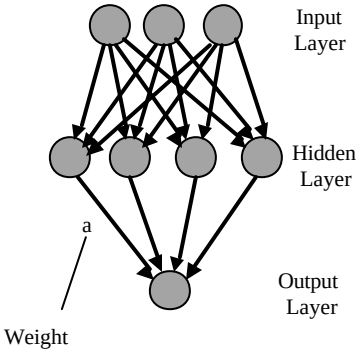
- Charge amino acids are hydrophilic, i.e. Asp (D), Glu (E), Lys (K), Arg (R).
- Uncharged amino acids are hydrophobic, i.e. Ala (A), Leu (L) Ile (I), Val (V), Phe (F), Trp (W), Met (M), Pro (P).
- In an ? helix, hydrophobic amino acids might line up on one side, which suggests that that side is on the interior of a protein or protein complex.



From *Bioinformatics: Sequence and Genome Analysis* by David Mount – Helicalwheel plot by GCG

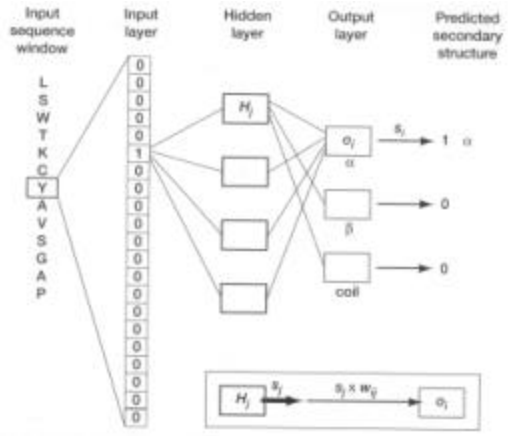
Neural Network Model

- A neural network is trained to recognize amino acid patterns that are located in know secondary structure.
- This fits weights on the connections of the connections between the nodes.
- The neural network can be used to predict the secondary structure on test proteins.
- These methods are over 70% accurate.



Neural Network

Rost and Sander Neural Network Model



From *Bioinformatics: Sequence and Genome Analysis*
by David Mount

Nearest Neighbor Method

- Like neural networks, this is another machine learning approach to secondary structure prediction.
- A very large list of short sequence fragments is made by sliding a window (n=16) along a set of 100-400 training sequences of know structure but with minimal similarity.
- A same-size window is selected from the query sequence and the 50 best matching sequences are found.
- The frequencies of the of the secondary structure of the middle amino acid in each of the matching fragments is used to predict the secondary structure of the middle amino acid in the query window.
- Can be very accurate (up to 86%).

Energy Potential Functions

- Contains terms for electrostatic interatction, van der Wals forces, hydrogen bonding, bond angle and bond length energies.
- Common software packages have their own implementation: Charmm, ECEPP, Amber, Gromos, and CVF.
- Structural predictions only as good as the assumptions upon which it is based (mainly the energy potential function).

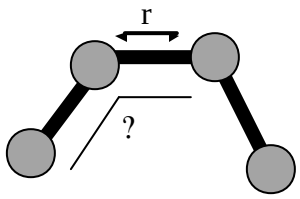
Bonded Terms

Bond Length

$$E_{bond\text{-}length} = ?_{bonds}\ k_b(r - r_0)^2$$

Bond Angle

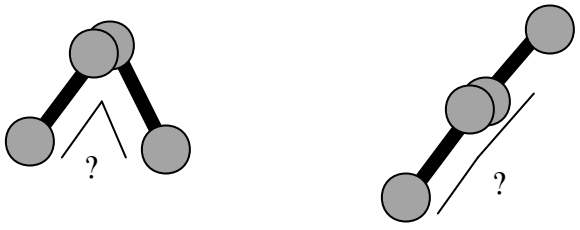
$$E_{bond\text{-}angle} = ?_{angle}\ k_? (? - ?_0)^2$$



Bonded Terms

Dihedral Angle

$$E_{dihedral\text{-}angle} = ?_{dihedrals}\ K_? (1 + \cos [n? ?R)-??$$



Non-Bonded Terms

Lennard-Jones potential (van der Waals force)

$$E_{vdW} = ?_{i,j} \underbrace{A_{ij}/r_{ij}^{12}}_{\text{repulsive}} - \underbrace{B_{ij}/r_{ij}^{6}}_{\text{dispersion}}$$

Electrostatic interactions

$$E_{elec} = ?^{r}_{i,j} q_i q_j/(4? ?_? ?_r r_{ij})$$

$?_??$ permittivity of free space

$?_r$ = dielectric constant of medium around charges

Non-Bonded Terms

Hydrogen Bonding – Some atoms (O, N, and to a lesser degree S) are electronegative, i.e. the attact electrons to fill their valence shells. Hydrogen tends to donate electrons to these atoms forming hydrogen bonds. This is common in water.

Salt Bridges – A positively charged lysine or arginine residue can form a strong interaction with a negatively charged aspartic acid or glutamic acid residue.

Energy Minimization

- Assumes that proteins are found at or near the lowest energy conformation.
- Uses a empirical function that describes the interaction of different parts of the protein with each other (energy potential function).
- Searches conformation space to find the global minimum using optimization techniques such as steepest descents and conjugate gradients.
- To avoid the multiple-minima problem, approaches such as dynamic programming, or simulated annealing have been used.

Molecular Dynamics

$F_i = m_i a_i$	force by Newton's Second Law of Motion
$a_i = dv_i/dt$	acceleration
$v_i = dr_i/dt$	velocity
$-dE/dr_i = F_i$	Work = force x distance
$-dE/dr_i = m_i\, d^2r_i/dt^2$	put it all together

Molecular Dynamics

- Model System – Choose protein model, energy potential function, ensemble, and boundary conditions.
- Initial Conditions – Need initial positions of the atoms, an initial distribution of the velocities (assume no momentum i.e. $\sum_i m_i v_i = 0$), and the acceleration which is determined by the potential energy function.
- Boundary Conditions – If water molecules are not being explicitly included in the potential function, the solvent boundary conditions must be imposed. The water molecules must not diffuse away from the protein. Also, usually a limited number of solvent molecules are included.

Molecular Dynamics

- Integration Algorithm – Solve the equations of motion with an algorithm that conserves energy and momentum, is computationally efficient, and allows a large time step. Examples:
  - Verlet Algorithm
  - Leap-frog Algorithm
  - Velocity Verlet
  - Beeman's algorithm
- Constraints

Molecular Dynamics

- Ensemble – a collection of all possible systems which have different microscopic states but have an identical thermodynamics state.
  - Microcanonical ensemble has fixed number of atoms, volume and energy
  - Canonical ensemble has fixed number of atoms, pressure, and temperature
  - Isobaric-isothermal ensemble has fixed number of atoms, pressure and temperature
  - Grand canonical ensemble has fixed chemical potential, volume and temperature.

Molecular Dynamics

- Result
  - The result of the simulation is a time series of the trajectories (path) followed by the atoms governed by Newton's law of motion.
  - The time scales are usually very small (picoseconds).
  - The motion of the molecule can be seen.
  - The motion will move the atoms into the near-equilibrium conformation of the protein.