CSC 2427 - Algorithms in Molecular Biology
Lecture 17 - Protein Secondary Structure

March 15, 2006

Lecturer: Michael Burdno
Scribe Notes by: Rama Natarajan

Lecture Outline

- Protein structures
- Protein folding
- van der Waals Interactions
- Grid Method
- Bounding Volume Hierarchies

1 Introduction

Amino acids are are linked through peptide bonds to form a polymer that is called a polypeptide. The primary structure of a protein is simply the ordering of the amino acids in a polypeptide sequence. It is common practice to represent the structure using only the peptide backbone.

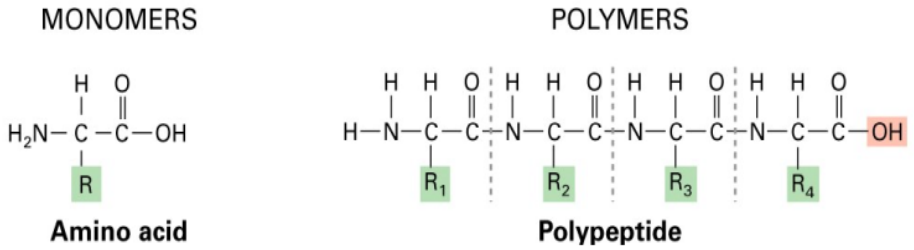

Figure 1: The primary structure of a protein is the amino acid sequence. Adapted from [1]

Secondary structure typically refers to an assembly of alpha helices and beta sheets: regular shapes that depend on hydrogen bonding between elements of the peptide backbone. The secondary structure of a protein gives much more insight into the *function of the protein* than its sequence.

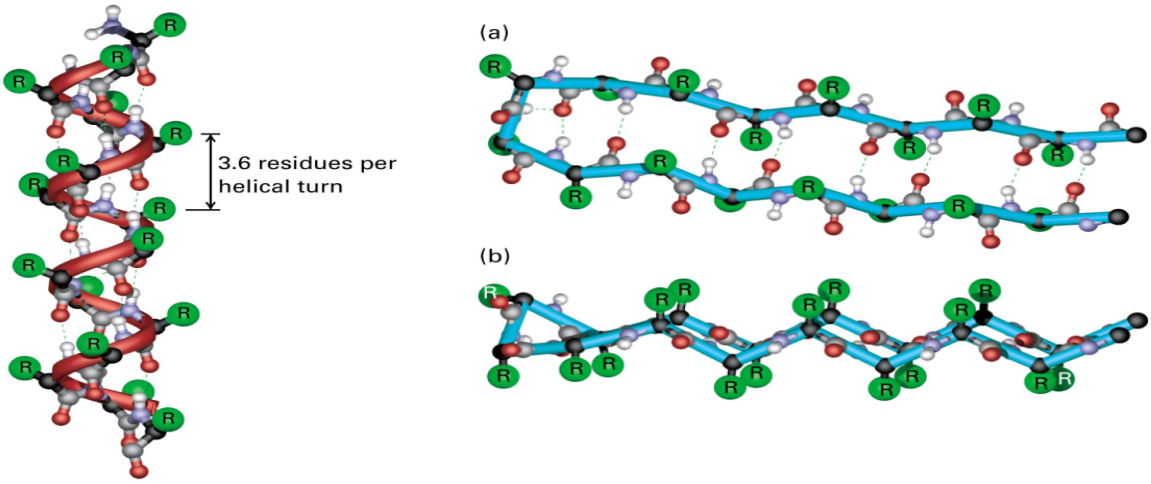

Figure 2: α helices and β sheets are stabilized by hydrogen bonds between backbone oxygen and hydrogen atoms. Adapted from [1]

Protein folding: The process by which the higher structures form is called protein folding and is a consequence of the primary structure. A protein sequence behaves differently at different temperatures: at about 37°c, the protein folds. At higher temperatures, it unfolds. The essential fact of folding is that the amino acid sequence of each protein contains information that specifies both the native structure and the pathway to attain that state. The

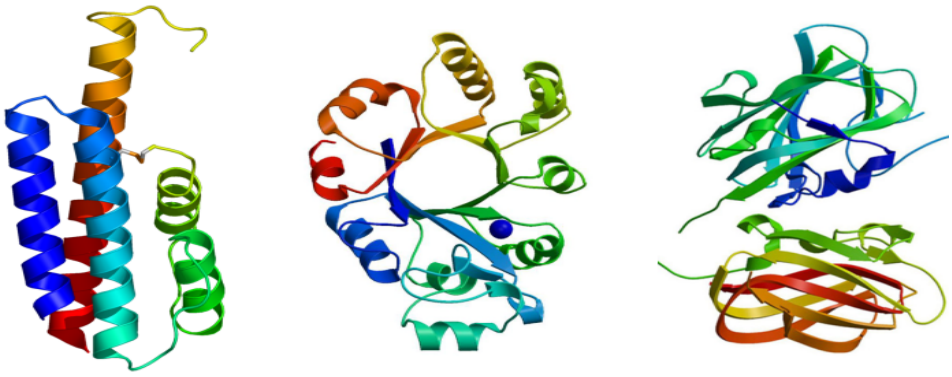

Figure 3: Protein folds composed of α helices and β sheets and others.

process of attaining the folded state is mainly guided by *van der Waals forces*. A unique polypeptide may have more than one stable folded conformation, each of which could have a different biological activity. However, usually, only one conformation is considered to be the active, or native conformation.

The entire duration of the folding process varies dramatically depending on the protein of interest. The slowest folding proteins require many minutes or hours to fold. However, small proteins, with lengths of a hundred or so amino acids, typically fold on time scales of milliseconds.

2 Protein Structure Prediction

It is possible to determine the three-dimensional structure of proteins from their amino acid sequences. A theoretical approach to understanding protein folding has been the calculation of protein energy landscapes. The main physical/chemical interactions among atoms are the basic components considered in energy calculation.

The energy landscape of a protein is the variation of its free energy as a function of its conformation, due to the interactions between the amino acid residues. It seems to be desirable to have a configuration with minimum free energy: it has been proposed that natural proteins have evolved such that this complicated energy surface has a shape which leads towards the native state, which is the lowest-energy conformation available to the protein.

Such a landscape allows the protein to fold to the native state through any of a large number of pathways and intermediates, rather than being restricted to a single mechanism. As mentioned earlier, the process of attaining the folded state is mainly guided by *van der Waals forces*.

van der waals interaction is an energy term, which refers to the strong interactions when two atoms approach to each other closely and their electron clouds start overlapping.

This interaction can exist between any pair of atoms in close vicinity. These forces are negligible between atoms that are far apart. The van der Waals force treats atoms as weakly attracting hard spheres thereby making collisions energetically unfavorable. Van der Waals energy function is specified as follows; it can be defined as a function of preferred bond length, or bond angle.

$$E_{vdw} = k(d - d_0)^2) \tag{1}$$

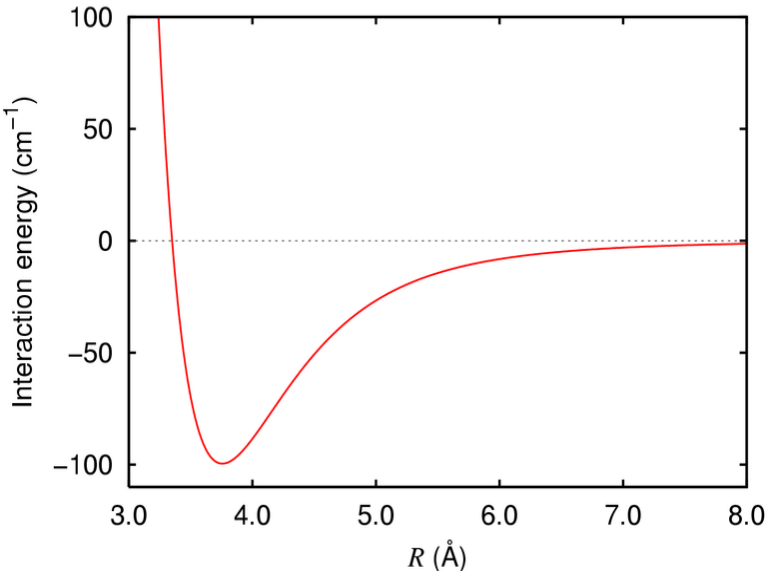

Figure 4: van der Waals Interaction: At small distances, 2 atoms repel eachtoher; at larger distances, there is no interaction. The van der Waal's force asymptotically goes to 0.

A stable state is found by summing the bunch of forces acting on atoms, computing the van der Waals interaction between every pair of atoms and moving the atoms in the direction of force. Sufficient iteration of this procedure should converge to a stable state. Folding@home is a distributed computing project designed to perform computationally intensive simulations of protein folding.

Molecular simulations involve sampling adoptable conformations and searching for low energy states. Monte Carlo simulations are a popular approach to study the thermodynamic kinetic properties of proteins, to generate meaningful distribution and motion pathways of protein. The steps involved are:

- Make a random walk through the conformation space
- At each cycle, perturb the current conformation at random
- If the energy of the new conformation decreases, accept
- If the energy of the new conformation *increases*, then accept that conformation with a certain acceptance probability
- Otherwise, reject

However, a quadratically large number of interactions $O(n^2)$ need to be considered in the energy calculation. Therefore, it is important to try to make this calculation as efficient as possible. The algorithms that try to solve this problem, exploit two features:

- Energy terms go to 0 as distance increases
- Van der Waals forces prevent atoms from getting too close together, so only a subset of the interacting pairs are possible. (considering that each sphere can only be close to K other spheres)

This leads to two questions:

- How to find the interacting pairs without enumerating all atom pairs?
- How to detect atom overlaps as soon as possible to reject the unstable conformation early?

Two widely applied methods towards the quick detection of interacting pairs are Grid method and Boundary-Volume Hiearchies.

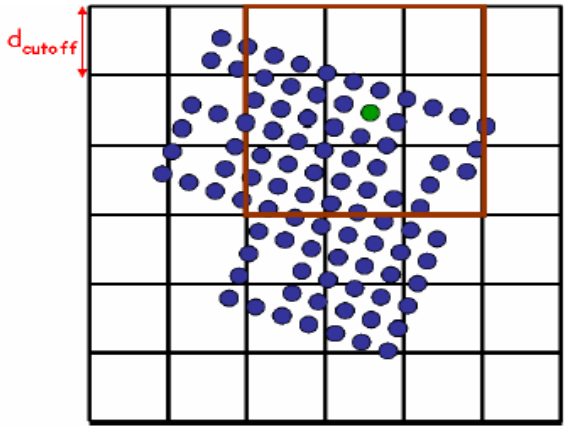

Figure 5: The 2D view of Grid Method. All atoms within the distance d-cutoff from the green atom are contained in the red rectangle. Adapted from [1]

3 Grid Method

- Subdivide the space into cubic cells.
- The length of each cell is approximately the cut-off distance for energy interactions. This is because it is desirable to have all interacting atoms to hash to nearby (or preferably the same) cell.
- For each atom center, mark the cell that contains it.
- The size of the cell for energy calculations is on the order of the cut-off distance for energy calculations.

- It is also possible to have a more fine-grained grid where each length of each cell is on the order of the atom radius. This allows overlaps to be detected quickly.

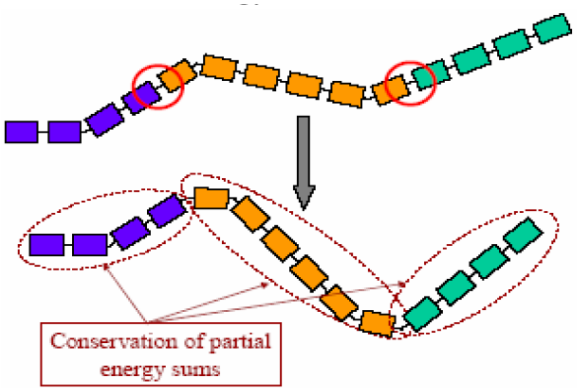

Figure 6: Energy of the sub-chains stay the same. Any change in energy comes only from interactions between 2 sub-chains. Adapted from [1]

The number of cells required to uniformly cover space increases exponentially with dimension. But the number of atoms is far fewer than that. Also, when updating energy calculations, the energy of each new interacting pair needs to be added to the previous energy value. The energy of all pairs that were broken then needs to be subtracted. This means that the grid needs to be recomputed after each time step, even if very few pairs were broken or created. Grid methods are also not invariant to rotation of the atomic coordinates.

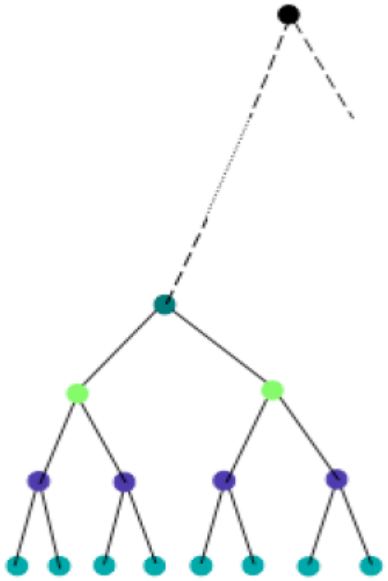

Figure 7: Bounding volume tree. Adapted from [1]

4 Bounding Volume Hierarchies

A bounding volume hierarchy (BVH) is a tree of bounding volumes. The bounding volume at a given node encloses the bounding volumes of its children. The bounding volume of a leaf encloses an atom (a primitive).

4.1 Building a BVH

The BVH is a balanced binary tree that is constructed bottom up as follows:

- Start with the smallest object.
- Enclose it in a bounding volume (sphere or box).
- At the lowest level, one BV bounds each link.
- Then, pairs of neighboring BVs at each level are bounded by new BVs to form the next level
- Continue this process iteratively till the root BV encloses the entire chain

So, at each level, we have a chain with half the number of BVs in the chain at the level below it. This chain of BVs encloses the geometry of the chains of BVs at all lower levels [2].

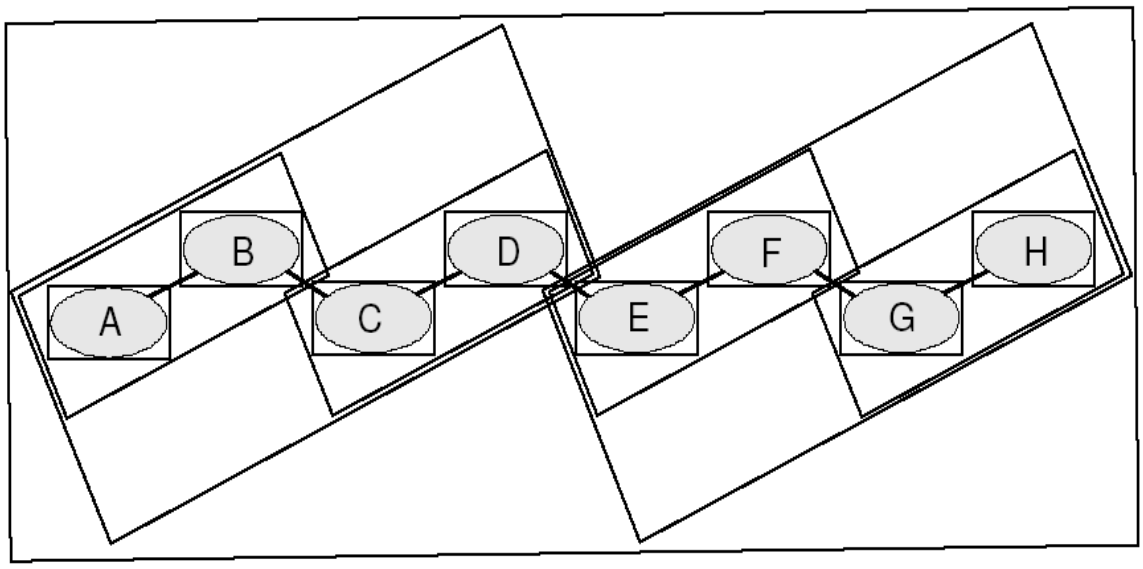

Figure 8: Bounding Volume Hierarchy. Adapted from [2]: Each box approximates the geometry of a sequence of links.

4.2 Testing for collisions

BVHs can be used to detect collision between pairs of rigid objects, each described by its hierarchy. The idea behind this is to first encapsulate the atoms into some geometrically

simple bounding volumes and then check if they overlap. If they do, then the encapsulated objects are decomposed into smaller pieces and checked again for overlap. A hierarchy tree of bounding volumes of objects is first built. The leaves are the boundary volumes that are not bigger than a set threshold and can be explicitly examined.

- Given the hierarchies of two objects, the test for collision proceeds as follows:
- First check whether the root boxes overlap.
- If they do not, safely return that the two objects do not collide.
- If they do overlap, then descend one level in both hierarchies and test all four pairs of children.
- Continue this process iteratively.
- When the lowest level of one hierarchy is reached, continue descent through the other hierarchy
- Test the leaf boxes of one hierarchy against boxes at the newly reached level in the other hierarchy
- On reaching the leaves in both hierarchies, test the actual components of the objects for overlap
- Returns a collision when an overlap is detected
- Terminate when the two objects are well separated

It is thus possible to encapsulate the objects of interest into some geometrically simple bounding volumes, and check if they collide by iterating through the bounding volume hierarchy tree.

References

[1] Lecture notes for CS273 Algorithms for Structure and Motion in Biology, Stanford Computer Science Department

[2] I. Lotan, F,Schwarzer, D.Halperin and J.Latombe, Algorithm and Data Structures for Efficient Energy Maintenance during Monte Carlo Simulation of Proteins.