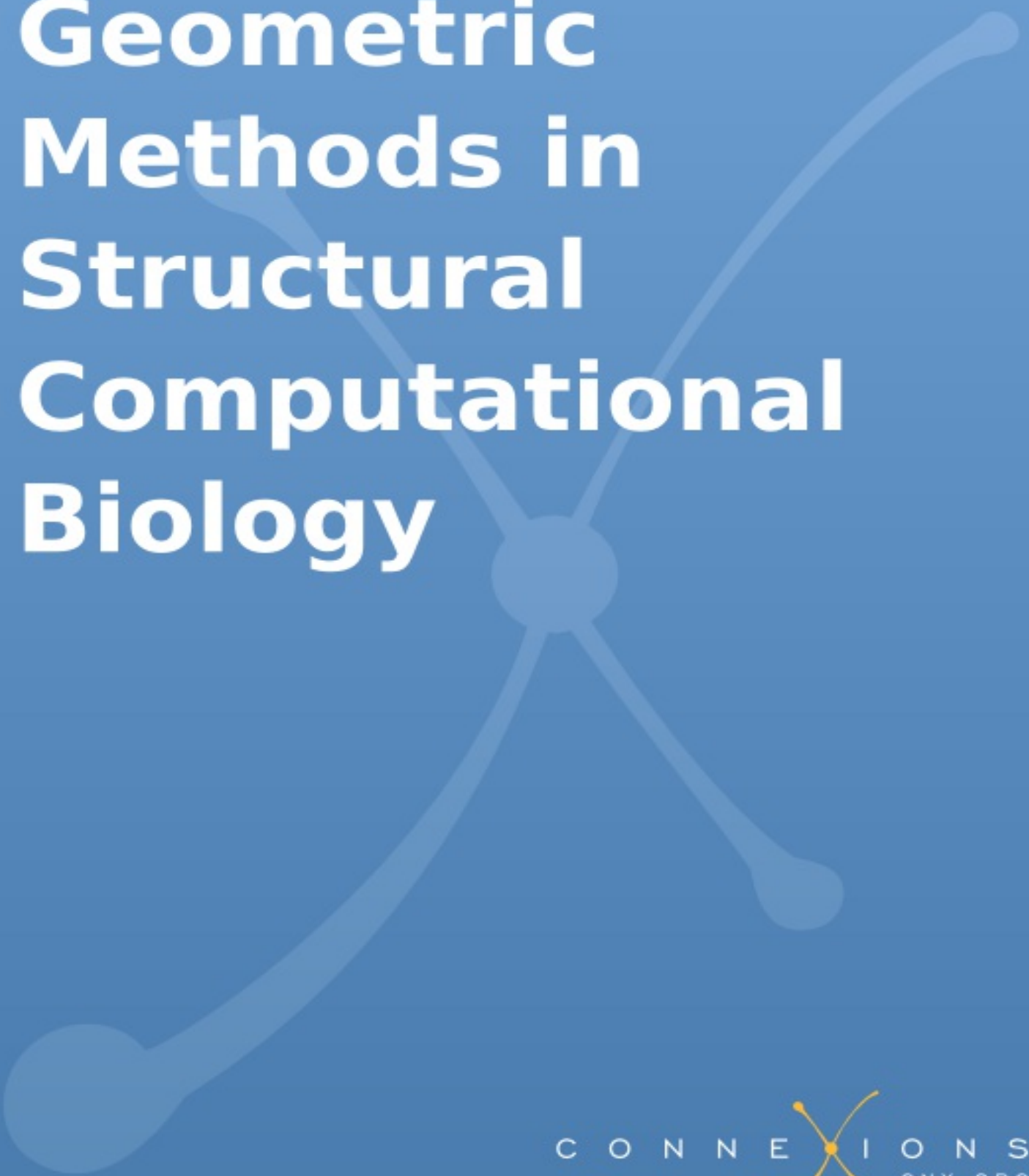


Geometric Methods in Structural Computational Biology



Geometric Methods in Structural Computational Biology

By: Lydia Kavraki

Online: <<http://cnx.org/content/col10344/1.6>>

This selection and arrangement of content as a collection is copyrighted by Lydia Kavraki.

It is licensed under the Creative Commons Attribution License: <http://creativecommons.org/licenses/by/2.0/>

Collection structure revised: 2007/06/11

For copyright and attribution information for the modules contained in this collection, see the "[Attributions](#)" section at the end of the collection.

Geometric Methods in Structural Computational Biology

Table of Contents

- [Structural Computational Biology: Introduction and Background](#)
 - [1.](#)
 -
 - [Proteins and Their Significance to Biology and Medicine](#)
 - [Protein Structure](#)
 - [Experimental Methods for Protein Structure Determination](#)
 - [X-ray Crystallography](#)
 - [NMR](#)
 - [Electron Diffraction](#)
 - [Structure Prediction of Large Complexes](#)
 - [Protein Structure Repositories](#)
 - [Visualizing Protein Structures](#)
 - [Visualizing HLA-AW with VMD](#)
 - [Visualizing HLA-AW with Protein Explorer](#)
- [Representing Proteins in Silico and Protein Forward Kinematics](#)
 - [1.](#)
 -
 - [Modeling Proteins on a Computer](#)
 - [Cartesian Representation of Protein Conformations](#)
 - [The Internal Degrees of Freedom of a Protein](#)
 - [Bonds and Bond Length](#)
 - [Bond Angles](#)
 - [Dihedral Angles](#)
 - [Dihedral Representation of Protein Conformations](#)
 - [Protein Forward Kinematics](#)
 - [Mathematical Background: Matrices and Transformations](#)
 - [Forward Kinematics](#)
 - [A Simple Approach](#)
 - [Denavit-Hartenberg Local Frames](#)
 -
 - [References](#)
- [Protein Inverse Kinematics and the Loop Closure Problem](#)
 - [1.](#)
 -
 - [Background Material](#)

- [Inverse Kinematics and its Relevance to Proteins](#)
- [Solving Inverse Kinematics](#)
 - [Inverse Kinematics Methods](#)
 - [Classic Inverse Kinematics Methods](#)
 - [Inverse Kinematics Methods with Optimization](#)
 - [Cyclic Coordinate Descent and Its Application to Proteins](#)
 -
- [References](#)
- [Molecular Shapes and Surfaces](#)
 - [1.](#)
 - [Introduction](#)
 - [Representing Shape](#)
 - [Alpha-Shapes](#)
 - [Computing the Alpha-Shape: Delaunay Triangulation](#)
 - [Weighted Alpha Shapes](#)
 - [Calculating Molecular Volume Using \$\alpha\$ -Shapes](#)
 - [Software](#)
 - [References](#)
- [Molecular Distance Measures](#)
 - [1.](#)
 - [Comparing Molecular Conformations](#)
 - [RMSD and IRMSD](#)
 - [Optimal Alignment for IRMSD Using Rotation Matrices](#)
 - [Optimal Alignment for IRMSD Using Quaternions](#)
 - [Introduction to Quaternions](#)
 - [Quaternions and Three-Dimensional Rotations](#)
 - [Optimal Alignment with Quaternions](#)
 - [Intramolecular Distance and Related Measures](#)
 - [References](#)
- [Protein Classification, Local Alignment, and Motifs](#)
 - [1.](#)
 - [Protein Classification](#)
 - [Protein Alignment](#)
 - [Protein Classification](#)
 - [Local Matching: Geometric Hashing, Pose Clustering and Match Augmentation](#)
 - [Motifs](#)
 - [Protein Function Prediction](#)
 - [Identification of Matches](#)
 - [Match Augmentation](#)
 - [Seed Matching](#)
 - [Augmentation](#)
 - [Filtering Matches](#)
 - [Designing Effective Motifs](#)

- [References](#)
- [Dimensionality Reduction Methods for Molecular Motion](#)
 - [1.](#)
 - [Introduction](#)
 - [Dimensionality Reduction](#)
 - [Principal Components Analysis](#)
 - [PCA of conformational data](#)
 - [Non-Linear Methods](#)
 - [Isometric Feature Mapping \(Isomap\)](#)
 - [References](#)
- [Robotic Path Planning and Protein Modeling](#)
 - [1.](#)
 - [Proteins as Robotic Manipulators](#)
 - [Robotic Path Planning](#)
 - [Background](#)
 - [The Path Planning Problem](#)
 - [Sampling-Based Path Planning](#)
 - [Sampling Based Planners for Proteins](#)
 - [References](#)
- [Motion Planning for Proteins: Biophysics and Applications](#)
 - [1.](#)
 - [Free Energy and Potential Functions](#)
 - [Free Energy](#)
 - [Potential Functions](#)
 - [Terms of energy functions](#)
 - [Bonds](#)
 - [Bond Angles](#)
 - [Torsions](#)
 - [Van der Waals Interactions and Steric Clash](#)
 - [Electrostatic Interactions](#)
 - [Other Classes of Interactions](#)
 - [Parameters](#)
 - [An Example: The CHARMM All-Atom Empirical Potential](#)
 - [Applications of Roadmap Methods](#)
 - [Kinetics of Protein Folding](#)
 - [A PRM-Based Approach](#)
 - [Stochastic Roadmap Simulations](#)
 - [Markovian State Models](#)
 - [Protein-Ligand Docking Pathways and Kinetics](#)
 - [References](#)
- [Protein-Ligand Docking, Including Flexible Receptor-Flexible Ligand Docking](#)
 - [1.](#)
 - [Background and Motivation](#)

- [Components of a Docking Program](#)
 - [Ligand placement algorithm](#)
 - [Scoring function](#)
 - [Explicit force field scoring function](#)
 - [Empirical scoring functions](#)
 - [Knowledge-based scoring functions](#)
- [Rigid Receptor Docking](#)
 - [Parameterization of the Problem](#)
 - [Examples of rigid-receptor docking programs](#)
 - [Autodock 3.0](#)
 - [Search technique](#)
 - [Scoring function](#)
 - [FlexX](#)
 - [Search technique](#)
 - [Scoring function](#)
 - [DOCK](#)
 - [Search technique](#)
 - [Scoring function](#)
 - [Links](#)
- [Flexible Receptor Docking](#)
 - [Introduction](#)
 - [Flexibility Representations](#)
 - [Soft Receptors](#)
 - [Selection of Specific Degrees of Freedom](#)
 - [Multiple Receptor Structures](#)
 - [Molecular Simulations](#)
 - [Collective Degrees of Freedom](#)
- [References](#)
- [Chapter 1. Homework assignments](#)
 - [1.1. Assignment 1: Visualization and Ranking of Protein Conformations](#)
 - [Protein Data Bank](#)
 - [Visualizing Protein Conformations](#)
 - [A. Visualizing a Set of Conformations](#)
 - [B. Molecules in Motion](#)
 - [Ranking Conformations](#)
 - [Visualizing Protein Substructures](#)
 - [Structurally Classifying Proteins](#)
 - [For Submission](#)
 - [Appendix: Installing VMD](#)
 - [1.2. Assignment 2: Performing Rotations](#)
 - ["Defining the Connectivity of a Backbone Chain"](#)
 - [Dihedral Rotations](#)

- [Setup with Matlab](#)
 - [Ranking by Energy](#)
 - [For Submission](#)
- [1.3. Assignment 3: Inverse Kinematics](#)
 - [Motivation for Inverse Kinematics in Proteins](#)
 - [Inverse Kinematics for a Polypeptide Chain](#)
 - [Setup with Matlab](#)
 - [For Submission](#)
 -
 - [References](#)
- [Index](#)

Structural Computational Biology: Introduction and Background

Topics in this Module

- [Proteins and Their Significance to Biology and Medicine](#)
- [Protein Structure](#)
- [Experimental Methods for Protein Structure Determination](#)
- [Protein Structure Repositories](#)
- [Visualizing Protein Structures](#)

Proteins and Their Significance to Biology and Medicine

Proteins are the molecular workhorses of all known biological systems. Among other functions, they are the motors that cause muscle contraction, the catalysts that drive life-sustaining chemical processes, and the molecules that hold cells together to form tissues and organs.

The following is a list of a few of the diverse biological processes mediated by proteins:

- Proteins called enzymes catalyse vital reactions, such as those involved in metabolism, cellular reproduction, and gene expression.
- Regulatory proteins control the location and timing of gene expression.
- Cytokines, hormones, and other signalling proteins transmit information between cells.
- Immune system proteins recognize and tag foreign material for attack and removal.
- Structural proteins prevent cells from collapsing on themselves, as well as forming large structures such as hair, nails, and the protective, largely impermeable outer layer of skin. They also provide a framework along which molecules can be transported within cells.

The estimate of the number of genes in the human genome has been changing dramatically since it was annotated (the latest gene count estimates can be found in this [Wikipedia article on the human genome](#)). Each gene encodes one or more distinct proteins. The total number of distinct proteins in the human body is larger than the number of genes due to [alternate splicing](#). Of those, only a small fraction have been isolated and studied to the point that their purpose and mechanism

of activity is well understood. If the functions and relationships between every protein were fully understood, we would most likely have a much better understanding of how our bodies work and what goes wrong in diseases such as cancer, amyotrophic lateral sclerosis, Parkinson's, heart disease and many others. As a result, protein science is a very active field. As the field has progressed, computer-aided modeling and simulation of proteins have found their place among the methods available to researchers.

Protein Structure

An amino acid is a simple organic molecule consisting of a basic (hydrogen-accepting), amine group bound to an acidic (hydrogen-donating) carboxyl group via a single intermediate carbon atom:

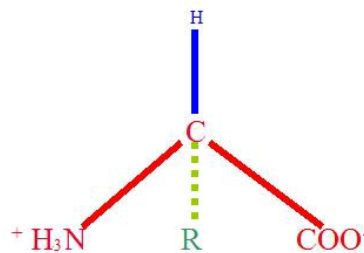


Figure 1. An α -amino acid

A generic α -amino acid. The "R" group is variable, and is the only difference between the 20 common amino acids. This form is called a zwitterion, because it has both positive and negatively charged atoms. The zwitterionic state results from the amine group (NH_2) gaining a hydrogen atom from solution, and the acidic group (COO) losing one.

During the translation of a gene into a protein, the protein is formed by the sequential joining of amino acids end-to-end to form a long chain-like molecule, or **polymer**. A polymer of amino acids is often referred to as a **polypeptide**. The genome is capable of coding for 20 different amino acids whose chemical properties depend on the composition of their **side chains** ("R" in the above figure). Thus, to a first approximation, a protein is nothing more than a sequence of these amino acids (or, more properly, amino acid **residues**, because both the amine and acid groups lose their acid/base properties when they are part of a polypeptide). This sequence is called the **primary structure** of the protein.

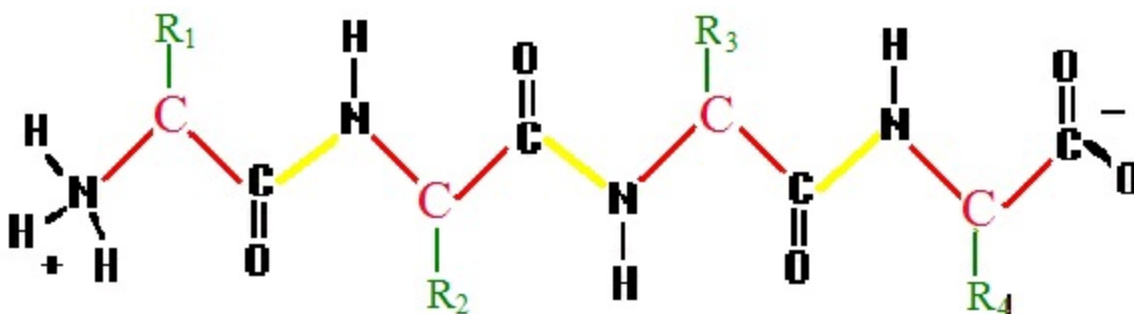


Figure 2. A polypeptide

A generic polypeptide chain. The bonds shown in yellow, which connect separate amino acid residues, are called **peptide bonds**.

The [Wikipedia entry on amino acids](#) provides a more detailed background, including the structure, properties, abbreviations, and genetic codes for each of the 20 common amino acids.

The primary structure of a protein is easily obtainable from its corresponding gene sequence, as well as by experimental manipulation. Unfortunately, the primary structure is only indirectly related to the protein's function. In order to work properly, a protein must fold to form a specific three-dimensional shape, called its **native structure** or **native conformation**. The three-dimensional structure of a protein is usually understood in a hierarchical manner. **Secondary structure** refers to folding in a small part of the protein that forms a characteristic shape. The most common secondary structure elements are **α -helices** and **β -sheets**, one or both of which are present in almost all natural proteins.

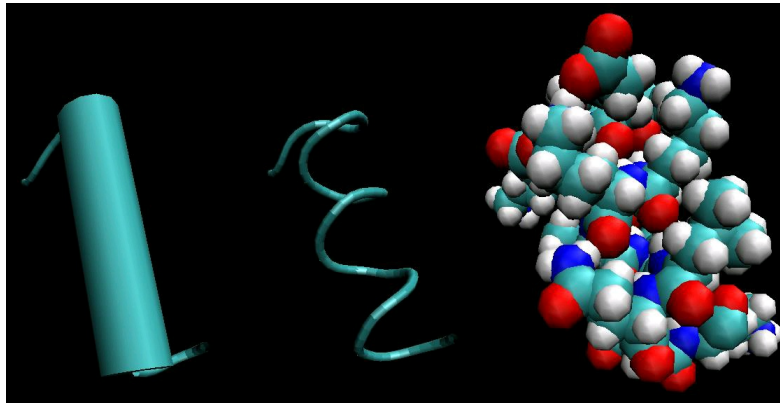
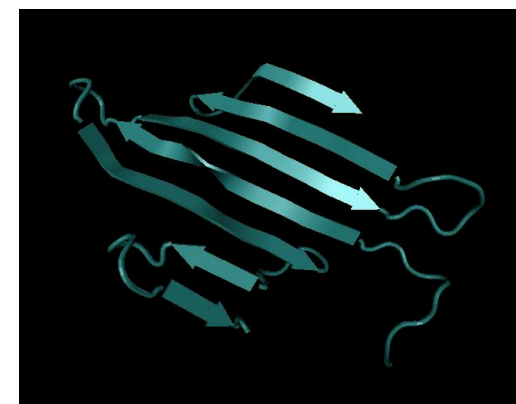


Figure 3. Secondary Structure: α -helix

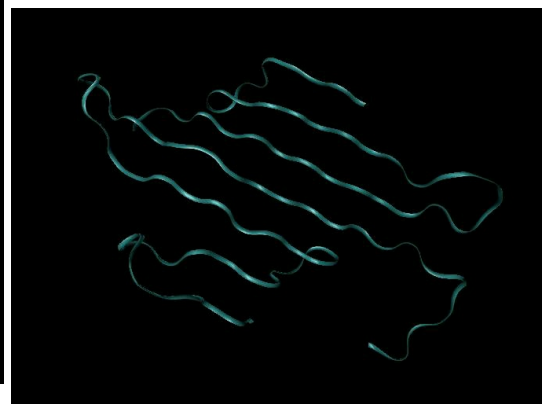
α -helices, rendered three different ways. Left is a typical cartoon rendering, in which the helix is depicted as a cylinder. Center shows a trace of the backbone of the protein. Right shows a space-filling model of the helix, and is the only rendering that shows all atoms (including those on side chains).

<db:title> Cartoon
representation </db:title>



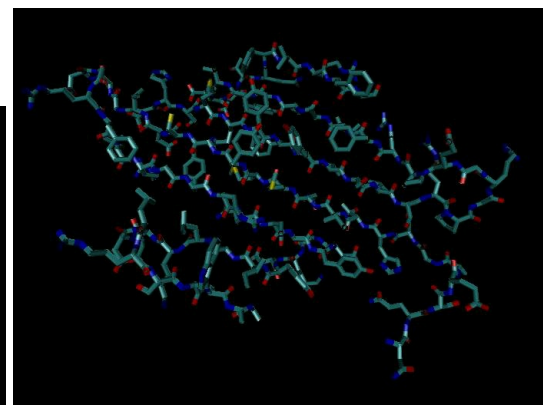
(a) Different parts of the polypeptide strand align with each other to form a β -sheet. This β -sheet is **anti-parallel**, because adjacent segments of the protein run in opposite directions.

<db:title> Ribbon
representation </db:title>



(b) β -sheets are sometimes referred to as β pleated sheets, because of the regular zig-zag of the strands evident in this representation.

<db:title> Bond representation
</db:title>



(c) Each segment in this representation represents a bond. Unlike the other two representations, side chains are illustrated. Note the alignment of oxygen atoms (red) toward nitrogen atoms (blue) on adjacent strands. This alignment is due to hydrogen bonding, the primary interaction involved in

Figure 4. Secondary Structure: β -sheet

Beta-sheets represented in three different rendering modes: cartoon, ribbon, and bond representations.

Tertiary structure refers to structural elements formed by bringing more distant parts of a chain together into structural **domains**. The spatial arrangement of these domains with respect to each other is also considered part of the tertiary structure. Finally, many proteins consist of more than one polypeptide folded together, and the spatial relationship between these separate polypeptide chains is called the **quaternary structure**. It is important to note that the native conformation of a protein is a direct consequence of its primary sequence and its chemical environment, which for most proteins is either aqueous solution with a biological pH (roughly neutral) or the oily interior of a cell membrane. Nevertheless, no reliable computational method exists to predict the native structure from the amino acid sequence, and this is a topic of ongoing research. Thus, in order to find the native structure of a protein, experimental techniques are deployed. The most common approaches are outlined in the next section.

Experimental Methods for Protein Structure Determination

A **structure** of a protein is a three-dimensional arrangement of the atoms such that the integrity of the molecule (its connectivity) is maintained. The goal of a protein structure determination experiment is to find a set of three-dimensional (x, y, z) coordinates for each atom of the molecule in some natural state. Of particular interest is the native structure, that is, the structure assumed by the protein under its biological conditions, as well as structures assumed by the protein when in the process of interacting with other molecules. Brief sketches of the major structure determination methods follow:

X-ray Crystallography

The most commonly used and usually highest-resolution method of structure determination is **x-ray crystallography**. To obtain structures by this method, laboratory biochemists obtain a very pure, crystalline sample of a protein. X-rays are then passed through the sample, in which they are diffracted by the electrons of each atom of the protein. The diffraction pattern is recorded, and can be used to reconstruct the three-dimensional pattern of electron density, and therefore, within some error, the location of each atom. A high-resolution **crystal structure** has a resolution on the order of 1 to 2 **Angstroms** (Å). One Angstrom is the diameter of a hydrogen atom (10^{-10} meter, or one hundred-millionth of a centimeter).

Unlike other structure determination methods, with x-ray crystallography, there is no fundamental limit on the size of the molecule or complex to be studied. However, in order for the method to work, a pure, crystalline sample of the protein must be obtained. For many proteins, including many membrane-bound receptors, this is not possible. In addition, a single x-ray diffraction

experiment provides only static information - that is, it provides only information about the native structure of the protein under the particular experimental conditions used. As we will see later, proteins are often flexible, dynamic objects when in their natural state in solution, so a single structure, while useful, may not tell the full story. More information on X-ray Crystallography is available at [Crystallography 101](#) and in the [Wikipedia](#).

NMR

Nuclear Magnetic Resonance (NMR) spectroscopy has recently come into its own as a protein structure determination method. In an NMR experiment, a very strong magnetic field is transiently applied to a sample of the protein being studied, forcing any magnetic atomic nuclei into alignment. The signal given off by a nucleus as it returns to an unaligned state is characteristic of its chemical environment. Information about the atoms within two chemical bonds of the resonating nucleus can be deduced, and, more importantly, information about which atoms are spatially near each other can also be found. The latter information leads to a large system of distance constraints between the atoms of the protein, which can then be solved to find a three-dimensional structure. Resolution of NMR structures is variable and depends strongly on the flexibility of the protein. Because NMR is performed on proteins in solution, they are free to undergo spatial rearrangements, so for flexible parts of the protein, there may be many more than one detectable structures. In fact, NMR structures are generally reported as **ensembles** of 20-50 distinct structures. This makes NMR the only structure determination technique suited to elucidating the behavior of **intrinsically unstructured proteins**, that is, proteins that lack a well-defined tertiary structure. The reported ensemble may also provide insight into the dynamics of the protein, that is, the ways in which it tends to move.

NMR structure determination is generally limited to proteins smaller than 25-30 kilodaltons (kDa), because the signals from different atoms start to overlap and become difficult to resolve in that range. Additionally, the proteins must be soluble in concentrations of 0.2-0.5 mM without aggregation or precipitation. For more information on how NMR is used to find molecular structures, please see [NMR Basics](#) and [The World of NMR: Magnets, Radio Waves, and Detective Work](#) at the National Institutes of Health's [The Structures of Life](#) website.

Electron Diffraction

Electron diffraction works under the same principle as x-ray crystallography, but instead of x-rays, electrons are used to probe the structure. Because of difficulties in obtaining and interpreting electron diffraction data, it is rarely used for protein structure determination. Nevertheless, ED structures do exist in the PDB. For more on ED, see this [Wikipedia article](#).

Structure Prediction of Large Complexes

Large macromolecular complexes and molecular machines present a particular challenge in

structure determination. Generally too large to be crystallized, and too complex to solve by NMR, determining the structure of these objects usually requires the combination of high-resolution microscopy combined with computational refinement and analysis. The main techniques used are [cryo-electron microscopy \(Cryo-EM\)](#) and standard light microscopy.

Protein Structure Repositories

Most of the protein structures discovered to date can be found in a large protein repository called the [RCSB Protein DataBank \(PDB\)](#). The **Protein Data Bank (PDB)** is a public domain repository that contains experimentally determined structures of three-dimensional proteins. The majority of the proteins in the PDB have been determined by x-ray crystallography, but the number of proteins determined using NMR methods has been increasing as efficient computational techniques to derive structures from NMR data have been developed. A few electron diffraction structures are also available. The PDB was originally established at Brookhaven National Laboratory in October, 1971, with 7 structures. Currently, the database is maintained by Rutgers University, the State University of New Jersey, the San Diego Supercomputer Center at the University of California, San Diego, and the National Institute of Standards and Technology. The current number of proteins (and/or nucleic acids) in the PDB database is displayed at the top-right corner of the main PDB page. The imaging method statistics of these structures (i.e., which methods were used for what fraction of the structures), as well as other classifications, can be found [here](#). The European Bioinformatics Institute Macromolecular Structure Database group (UK) and the Institute for Protein Research at Osaka University (Japan) are international contributors to the contents of the PDB.

Visualizing Protein Structures

Numerous tools are available for visualizing the structures stored in the PDB and other repositories. Most such tools allow a detailed examination of the molecule in a variety of rendering modes. For example, sometimes it may be useful to have a detailed image of the surface of the molecule as experienced by a molecule of water. For other purposes, a simple, cartoonish representation of the major structural features may be sufficient.

A Few Molecular Visualization Programs

- [Visual Molecular Dynamics \(VMD\)](#) was originally developed for viewing molecular simulation trajectories. It is a very powerful, full-featured, and customizable molecular viewing package. Customization is available using Tcl/Tk scripting. Information on Tcl/Tk scripting can be found at this [Tcl/Tk](#) website.
- [PyMol](#) is an open-source molecular viewer that can be used to generate professional-looking images. PyMol is highly customizable through the Python scripting language.
- [Protein Explorer](#) is an easy-to-use, web browser-based visualization tool. Protein explorer is

built using the [MDL Chime](#) browser plugin, which in turn is based on the [RasMol](#) viewer. Because Chime only works under Windows and Macintosh OS, the use of Protein Explorer is restricted to those platforms.

- [J Mol](#) is a Java-based molecular viewer. In applet form, it can be downloaded on-the-fly to view structures from the web. A stand-alone version also exists, which can be used independently of a web browser.
- [Chimera](#) is a powerful visualizer and analysis tool that can be comfortably used with very large molecular complexes. It can also produce very high-quality images for use in presentations and publications.

Visualizing HLA-AW with VMD

What follows will be a very brief introduction to what can be done with VMD. Only the most basic viewing functionality will be discussed. For a complete description of the capabilities of VMD and how to use them, please refer to the [VMD web site](#).

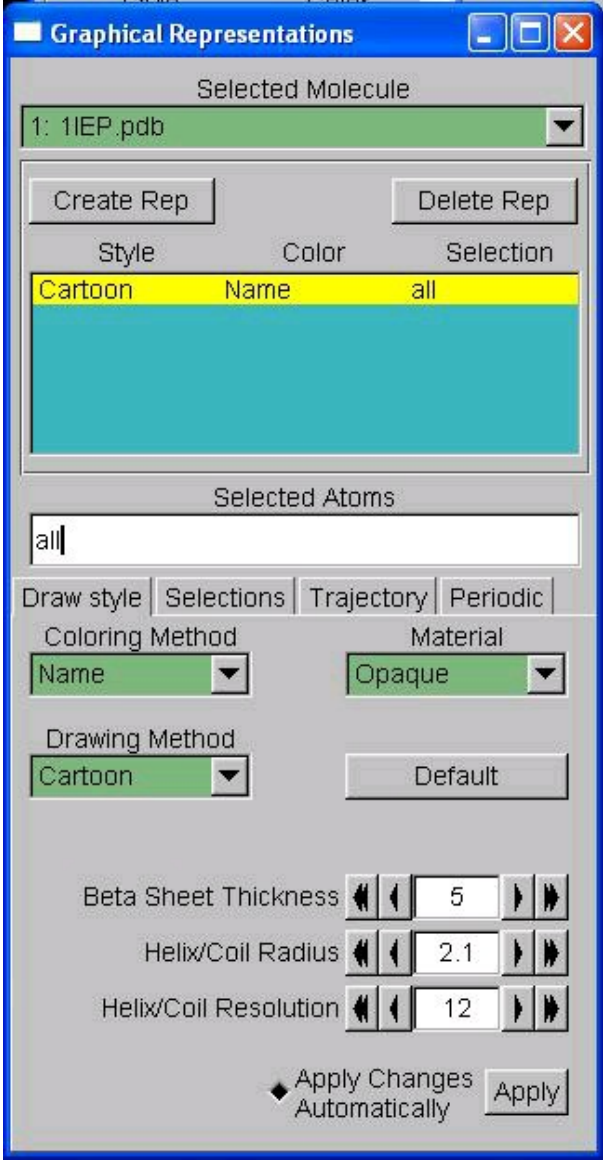
In this section, a human leukocyte-associated antigen, HLA-AW (PDB structure ID 2HLA), will be shown under various rendering methods in VMD. This section is intended to convey, first, a general idea of the types of visual representations that are available for protein structures, and second, what information is and is not conveyed by each representation.

VMD allows the user to load and view molecule description files in a wide variety of common formats, including trajectory files with multiple structures of the same molecule, such as might be generated by a simulation. Once the molecules are loaded, the way each molecule is rendered may be controlled using the Graphical Representations menu:

<db:title>VMD Graphical
Representations menu </db:title>

<db:title>VMD atom
coloring methods

<db:title>VMD molecule
drawing methods



(a) This menu allows the user to control in detail how each molecule is rendered.

Name
Type
Element
ResName
ResType
ResID
Chain
SegName
Molecule
Structure
ColorID
Beta
Occupancy
Mass
Charge
Pos
User
Index
Backbone
Throb
Timestep
Volume

</db:title>

(b) Coloring schemes to highlight features of interest.

Lines
Bonds
DynamicBonds
HBonds
Points
VDW
CPK
Licorice
Trace
Tube
Ribbons
NewRibbons
Cartoon
NewCartoon
MSMS
Surf
VolumeSlice
Isosurface
Beads
Dotted
Solvent

</db:title>

(c) Rendering methods in VMD. Which one to use depends on the features to highlight.

Figure 5.

The built-in rendering options of VMD.

Molecules may be displayed by various rendering modes:

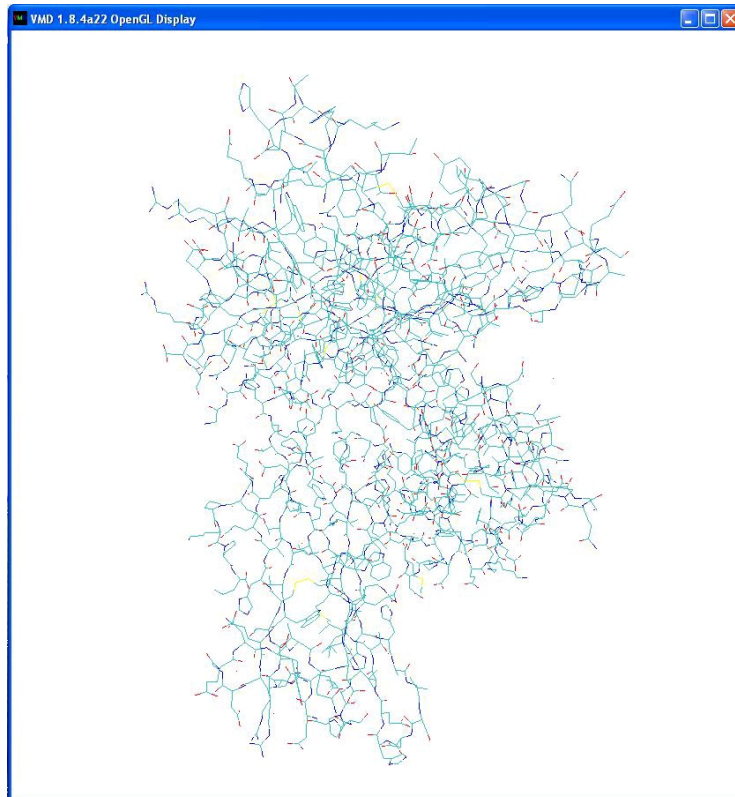


Figure 6. HLA-AW. Drawing method: LINES. Coloring method: NAME

In this representation, each line represents a bond between two atoms. The color of each half-bond corresponds to the element of the atom at the corresponding end of the bond (red for oxygen, blue for nitrogen, yellow for sulfur, and teal for carbon). Line representation gives a clear idea of the molecule's connectivity, but for large molecules it can be difficult to isolate protein sub-structures.

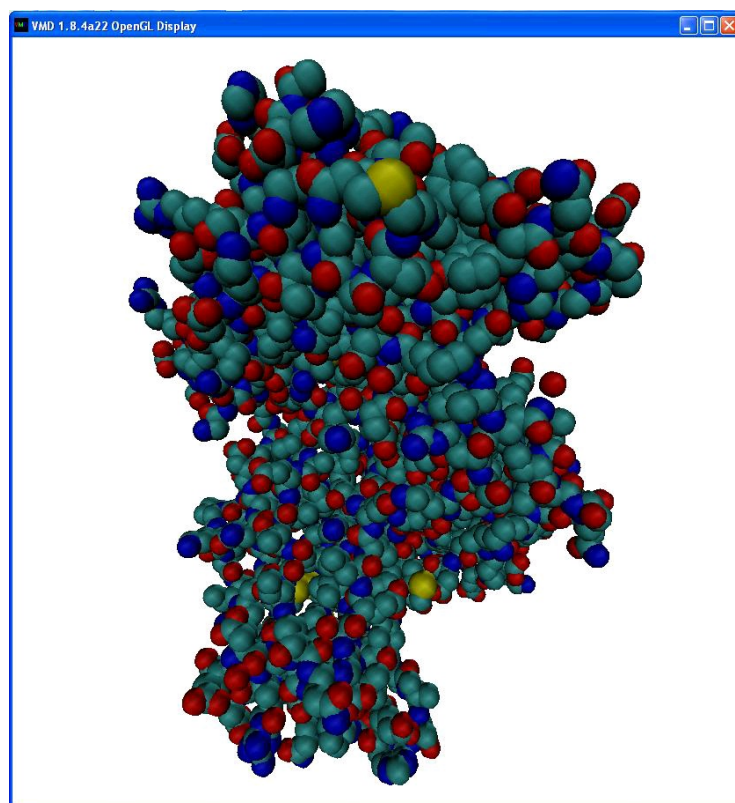


Figure 7. HLA-AW. Drawing method: VDW. Coloring method: NAME

Here each atom is represented by a sphere whose radius is the **Van der Waals radius** of the atom. The Van der Waals radius is half the separation of unbonded atoms packed as tightly as possible, and provides a rough notion of a collision radius, although it is not a firm barrier. This representation of the molecule gives a rough sense of its shape, and is sometimes called a **space-filling** model.

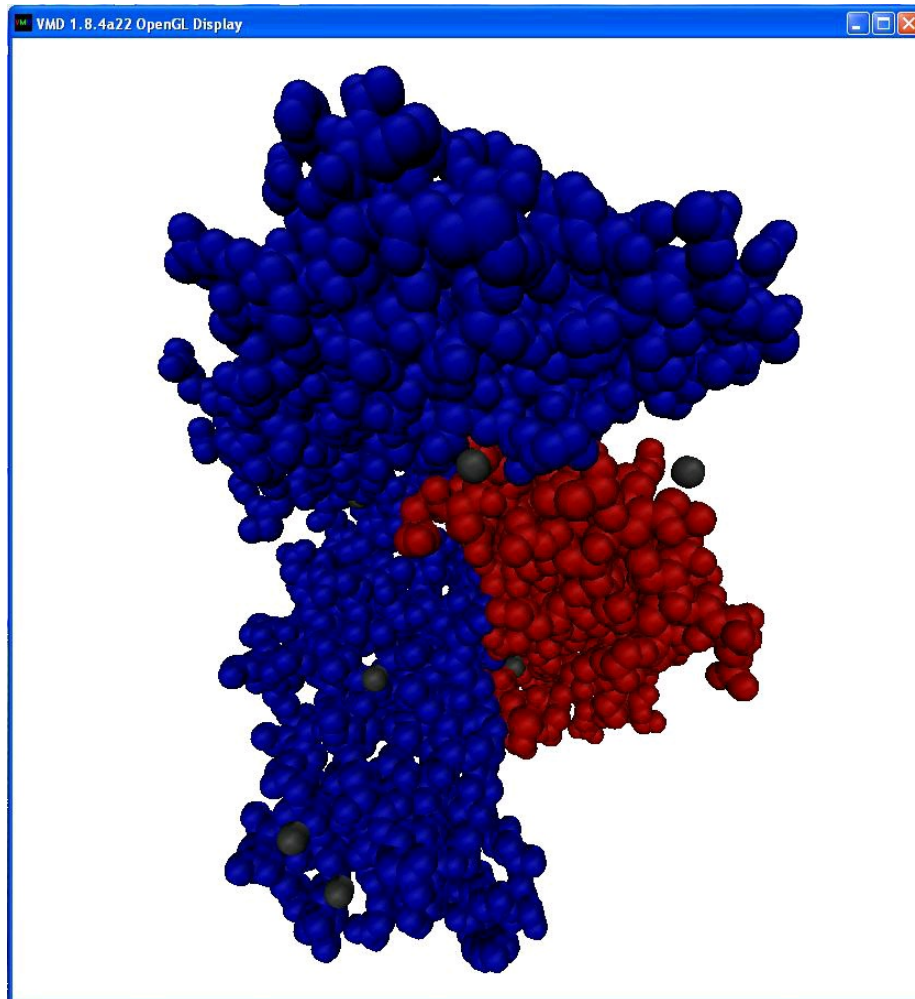


Figure 8. HLA-AW. Drawing method: VDW. Coloring method: CHAIN

This rendering is the same as in the previous figure, except that now the atoms are colored based on which polypeptide chain they belong to. HLA-AW consists of two chains, the alpha chain (blue), which folds into three domains and the smaller $\beta 2$ microglobulin (red), which is a component of a whole class of HLA proteins. Coloring by chain allows an inspection of how the polypeptide subunits come together to form the whole quaternary structure of the protein. The black balls are water molecules near the surface of the protein that always appear in the same place in crystal structures, and may therefore be considered part of the structure for some applications.

Thank You for previewing this eBook

You can read the full version of this eBook in different formats:

- HTML (Free /Available to everyone)
- PDF / TXT (Available to V.I.P. members. Free Standard members can access up to 5 PDF/TXT eBooks per month each month)
- Epub & Mobipocket (Exclusive to V.I.P. members)

To download this full book, simply select the format you desire below

