

CS 273
April 5, 2006

CS 273 Algorithms for Structure and Motion in Biology

Scribe for today: Eric Willgohs

Class Introduction

Instructors: Serafim Batzoglou and Jean-Claude Latombe
TA: Sam Gross

Schedule: Today's class slides include a class schedule, still tentative, especially regarding the project presentation dates.

Scribed Notes: Scribed notes should be written to augment the class slides with the additional significant material presented in class. They may include some slides but should not primarily consist of just the class slides. They should generally not, however, include absolutely every thread of discussion that surfaces in class, as such would impede a clear presentation of the primary material. Well written notes will generally run from five to seven pages, though for some classes they may be as short as three pages and other classes may require ten pages for adequate coverage. Notes should be submitted within one week of class; the instructors will review them, give comments and perhaps ask for changes.

Coursework: The coursework for the class will consist of the scribed notes and a project; there will be no other homework or tests. The structure of the project will be to use some existing software, develop some new software to expand the current capabilities, and to present the project. It is important for the students to be excited about the project (so long as it is at least loosely associated with the class). Two to three people are recommended for the class, though four would be fine and single person projects have also been successful. The sooner milestone 1 is accomplished, the better.

Meeting Instructors: to meet with instructors contact via email to schedule.

Range of Bio-CS Interaction

Class slide 3 shows a number of categories/levels of Bio-CS interactions; there are a great many potential such interactions. The example levels were Gene, Molecules, Cells, Tissue/Organs, Body System. These categories vary tremendously in the typical unit of time and/or space used in their analysis.

An example interaction at the body system level is robotic surgery, which includes simulation. There is more and more surgery being done today with robotics, and being assisted by simulations, including brain surgery. Such simulations are generally at the tissue/organ level; an example would be the simulation of severing an artery during surgery.

CS 273 will focus on Bio-CS problems at the molecular level, in particular the molecular structure, similarities and motions of proteins.

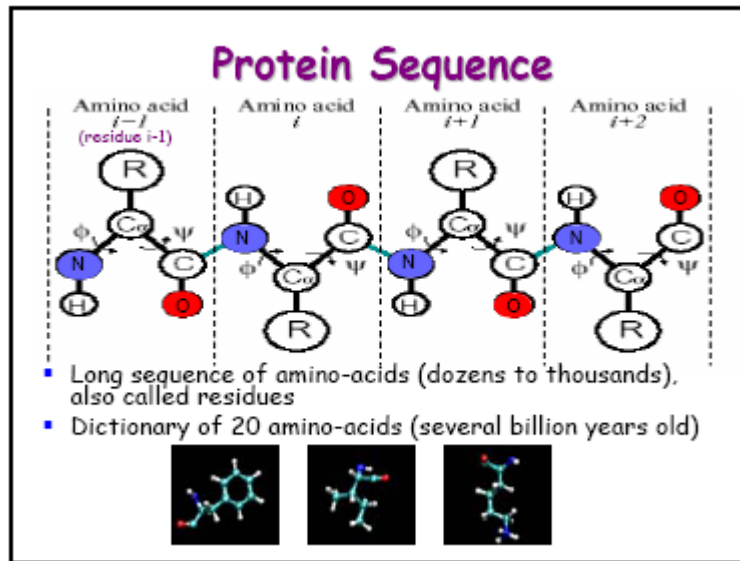
Focus on Proteins

The class presentation then turned to the class' focus on proteins, noting some of the very important functions performed by proteins, and noting aspects of particular interest from a computational point of view. One of the characteristics of particular interest is the unusual kinematic structure: the long serial linkages of proteins have backbones with typically hundreds of degrees of freedom, and each of the side chains also generally have a few degrees of freedom. Luckily, techniques have been developed in the field of robotics which can be used to model the motion of such linkages.

Many challenging problems involving proteins are being tackled in the Bio-CS arena today; a number of these are listed in slide 6. A common thread to many of these problems relates to the fact that proteins generally adopt a compact structure called the folded or native state.

Protein Sequence

(slides 9-11)



The amino acids all share a common structure but have different side chains. The structural template is an amine group (NH_2) connected to a carbon atom (denoted C_α) in turn connected to another carbon atom (denoted C_β), in turn connected to an oxygen and a hydroxyl group (OH). When joined together in a chain the hydroxyl group from the first amino acid (AA) joins with a hydrogen from the amine group of the second AA, forming a water that leaves the molecule and leaving the C_β carbon of the first AA bonded to the nitrogen of the second AA (this is called a peptide bond), as shown above. Some of the electron cloud from the $\text{C}=\text{O}$ bond migrates to the $\text{C}-\text{N}$ bond, making it a partial double bond, very strong.

Levels of Structure

(slides 12-13)

Proteins are generally considered to have four levels of structure: primary (linear sequence); secondary (formation of alpha helices and beta sheets); tertiary (how the secondary structures arrange into a compact overall structure); and quaternary (how multiple protein chains can combine into one macromolecule).

Loops (any secondary structure segments that are not alpha helices or beta sheets) do not have a defined shape but give flexibility to the protein, allowing it to adopt a compact conformation. Usually there is little or no water inside the folded protein; the water solvent surrounds the folded protein.

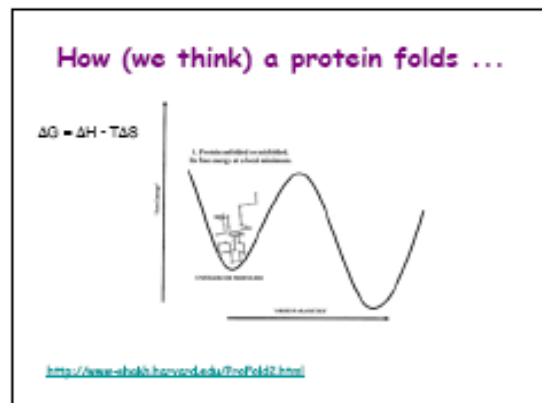
Slide 12 shows example structures in ribbon format (which shows the carbon backbone but not any of the side chain residues). These examples included one primarily alpha helices, one primarily beta sheets, and one mixed.

Folding

(slides 14-19)

Protein folding is not well understood. There are many theories of how it can progress, and it may well be that different methods of folding are employed by different proteins. Any given protein could have multiple pathways to its final state, which is generally considered unique for each protein (i.e., one to one correspondence between sequence and final conformation).

Slides 15-19 show a 2-D graph of an energy landscape for a protein as it folds. The change in total energy is equal to the change in enthalpy (free energy) plus the temperature times the change in entropy (based on the number of equivalent states).



The energy barrier shown in the graph arises because as initial contacts are made during the folding process, entropy decreases faster than the free energy, yielding the barrier with an associated intermediate state. As the barrier is overcome, the energy decreases faster than the entropy.

Motion in the Folded State

(slides 20-21)

The left side of slide 20 shows the high frequency motion of the sidechains; the right side shows the low frequency movement of the backbone (in a “breathing” type movement). Low frequency backbone movement is probably necessary for many protein functions.

Slide 21 shows a representation of protein movement that emphasizes structural variability—the central beta sheets move only small amounts, but the loops move quite a

bit. It is being accepted more and more that loop movement is critical to many protein functions. The movement represented in this slide was produced by a combination of NMR and computational techniques (an approach becoming common today). NMR can be used to observe certain properties of molecules in solution, whereas x-ray crystallography requires crystallization (which likely deforms protein conformation somewhat). Because it is only *some* properties that can be observed by NMR, computational methods are used to augment the analysis.

These analyses also try to determine the likelihood of the various conformations. This likelihood depends not only on the energy of the conformation, but also on the number of states that can transition to it.

Flexible Loops & Binding (slides 22-25)

Although hydrogen bonds are individually somewhat weak and flexible, when combined such as in helices and sheets they make such structures have a very stable/rigid shape. Loops, however, can deform by a large amount with just a small variation in energy. As our understanding of induced fit as developed, people first realized that ligands could deform to improve fit, and later realized that the proteins could deform also. Deformation of surgical sutures was noted as potentially analogous to loop deformations—twisting one end of the suture can cause significant deformation in the suture loop, which is often relied upon by surgeons.

Modeling protein-ligand binding is very important for medicine. Modeling protein-protein binding is important to tissue engineering, for example, and is more difficult than protein-ligand binding. People study the conformations of both proteins, but it is very difficult to assess the deformations that will arise as they bind. Regarding the slide on pyruvate binding, it is the deformation of a loop, moving outwards, that allows the pyruvate to come in and bind to the enzyme.

More Info re CS 273 Focus and Structure (slides 26-37)

CS 273 will focus on computational algorithms for molecular biology problems, and in particular will focus on how to design efficient data structures and algorithms for biology. One thing that CS273 is NOT going to get into is “how the energy should be modeled” – that topic is for chemists and biologists.

Prof. Latombe noted that biologists often prefer to accumulate large data sets, and categorize the underlying observations, whereas computer scientists often prefer to model functions describing the relationships between data without caring so much about the actual values or biological significance of the data. One challenge of computational biology is to fuse these cultures.

CS 273

April 5, 2006

The broadbrush main ideas behind CS 273 include the classic idea that a great deal of information is provided by the sequence, and that *some* function can be predicted from sequence alone, but that in most contexts a satisfactory analysis will require some simulation of the molecular motion needed for the process of binding.

Class Vision: assisting the transition from high cost, low bandwidth, in-vivo biology to low cost, high bandwidth, in-silico biology. An exemplar of this kind of transition would be the aviation transition from designing airplanes with heavy reliance on wind-tunnel experiments, to designing airplanes with almost complete reliance on simulations and little to no wind-tunnel experiments. Highlighting the need for this transition in biology is the fact that the drug development cycle typically takes about ten years, and even then drugs still hit the market with unexpected side effects. We will design more efficient algorithms by exploiting physical constraints: for example, since the van der Waals forces prevent atoms from overlapping, we know that there will be a limited number of atoms in any given space, and we can use this constraint to make our algorithms more efficient. Also part of the class vision is that to contribute the most to furthering biology, we need to maintain a high level view, not just engage in detailed simulations.