Recent advances in Complexity CIS 6930/CIS 4930	December 10, 2002
Lecture Lecture 29	
Lecturer: Andrew Lomonosov	Scribe: Andrew Lomonosov

1 Molecule

A *molecule* is a group of atoms which remains spatially intact on the time scale of interest. Its *structure* consists of those features which remain invariant on that time scale. The atoms are bound together by *covalent bonds*, a very strong link which can be broken only by high temperatures or in chemical reactions. Generally the *bond lengths* between covalently bonded pairs of atoms have very nearly fixed values on the order of 1 to 2 angstroms (10^{-10} m). The *covalent structure* of a molecule is the way in which its atoms are bonded together. The *conformation* of a molecule is its precise spatial structure at some instant in time. Any conformation can be completely specified by giving Cartesian coordinates of all the atoms. It can also be specified by giving all interatomic distances.

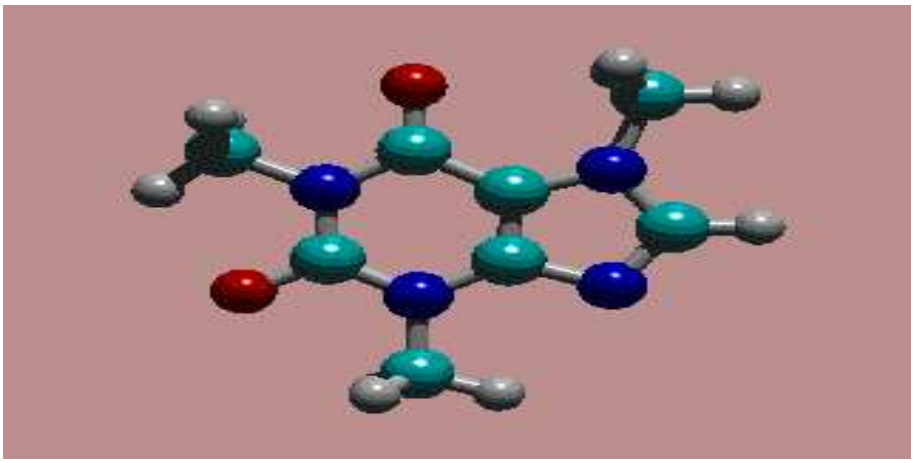

Figure 1: Example of a covalent structure of a $C_8H_{10}N_4O_2$molecule

2 Molecular modeling

Molecular modeling is the science of studying molecular structure and function through model building and computation. It is a multidisciplinary enterprise. It involves biology (cellular picture), chemistry (atomic and molecular details), physics (electronic level and underlying forces), mathematics (numerical models and topology) and computer science (algorithms and implementation for large scale problems).

2.1 Brief overview of computational biology

Currently term *computational biology* encompasses virtually all computational approaches to molecular biology problems. It can be subdivided into several major parts.

One part is *bioinformatics* that refers to a particular branch of computational biology where the emphasis is on "inference" and datamining/learning from known/mapped patterns (Both nucleic acids and proteins), rather than trying to provide an ab-initio biophysically based model. One particular branch of bioinformatics is *chemoinformatics* that uses large existing chemical databases for designing new drugs.

In bioinformatics approaches we infer the tertiary (3d) structure of a protein from its primary structure by looking at similar primary structures whose tertiary structure is known and doing some sort of learning or extrapolation. We are not interested in giving a biophysical explanation of the primary-tertiary correspondence or map.

Contrastingly, in *ab-initio approaches* (structural biology, traditional computational biology, biomolecular computing etc.), we would try to obtain a mathematical model of the function that maps the primary to the tertiary structure, using basic biophysical principles. The function could either be modeled as statically (as the satisfying assignment to a bunch of geometric constraints dictated by bond lengths/angles etc.), or dynamically as a process by which the protein reaches its optimal or stable configuration (a search through an appropriate parameter space).

Biomolecular computing, in spirit is the same as structural biology, but could be at a slightly coarser scale. The latter emphasizes the development of mathematical/computational models for biological phenomena occurring in nature, by using basic biophysical concepts as primitives. The former tries to create *new* biological phenomena, either from the same basic primitives (same scale), or in some cases by building on more complex phenomena occurring in nature – using them as oracles or subroutines (coarser scale). In either case, the new biological phenomena thus constructed are then used to compute other difficult functions or a whole new type of computer, an alternative model of computing (this is the spirit of DNA computing, biogeometric models of computation etc..).

3 Structure calculation of biological molecules

In this talk we will be looking at ab-initio approaches of structural biology. Currently there are three major types of techniques that calculate molecular structure.

First technique is based on *distance geometry*. The basic idea of distance geometry is to formulate the problem not in Cartesian space of the atom positions but in the space of all interatomic distances. After a solution has been found in this space, it is embedded in the Cartesian space. The approach generally gives a rough approximation of the real solution, however this approximation has it uses, since it can be given as a starting conformation to the third technique below.

Second technique uses the fact that the problem of finding molecular conformations that are in agreement with certain geometric constraints can always be formulated as one of *minimization of a suitable objective or target function*. The global minimum of the target function is sought, while local minima are to be avoided. The function is often optimized in the space of torsion angle space, and optimization is carried out by using method of conjugate gradients. The problem with this method is that it tends to get stuck at the local minima.

Third technique is based on classical mechanics and proceeds by numerically solving Newton's equation of motion in order to obtain a trajectory for the molecular system. The Cartesian coordinates of the atoms are degrees of freedom.

4 Fundamental Problem of Distance Geometry

Given the distance constraints which define (our state of knowledge of) a molecule, find one or more

conformations which satisfy them, or else prove that no such conformations exist.

Some upper bounds on bond length can be obtained from NMR observations, some lower bounds can be derived from van der Waals forces.

We will denote by A a set of atoms in the molecule, by $u(a,b)$ upper bound on bond length between atoms a and b, and by $l(a,b)$ the corresponding lower bound.

4.1 Importance of the Fundamental problem

Testing experimental data for errors.

Testing hypothesis. For example if we want to check whether it is possible for a to be within δ of b we can simply reset $u(a,b) = \delta$ and test the revised bound for consistency.

Identifying redundancies. *Redundant constraints* do not tell us anything about the conformational state of the molecule above and beyond what the rest of the constraints imply. Suppose that we want to test upper bound $u(a,b)$ for redundancy. This can be done by resetting $l(a,b) = u(a,b) + \epsilon, u(a,b) = \infty$. Then $u(a,b)$ was redundant iff the revised constraints are inconsistent. I.e other constraints do not permit any conformations in which (a,b) distance exceeds its upper bound. Similarly, redundancy for lower bounds can be tested by setting $u(a,b) = l(a,b) - \epsilon, l(a,b) = 0$.

Equivalence of constraints. Given two sets of constraints (l_1, u_1) and (l_2, u_2), do they determine the same set of possible embeddings? This can be done by resetting $l_1(a,b) = u_2(a,b), l_2(a,b) = u_1(a,b), u_1(a,b) = l_2(a,b), u_2 = l_1(a,b)$ and testing resulting bounds for consistency for every pair $a, b \in A$.

Common embeddings. Given two sets of constraints (l_1, u_1) and (l_2, u_2), do they have any embeddings in common? This can be solved by defining new bounds (l_3, u_3) as $l_3(a,b) = \max(l_1(a,b), l_2(a,b))$ and $u_3(a,b) = \min(u_1(a,b), u_2(a,b))$ for all $a, b \in A$. Then there are common conformations iff the bounds (l_3, u_3) are consistent.

5 Triangle inequality bound smoothing

Problem: given a set of lower and upper bounds $l, u \in F(A) = \{f : A \times A \to R^+ | f(a,b) = f(b,a) \& f(a,a) = 0\}$, we want to compute the corresponding *triangle ineqaility limits.*

Function $f \in F(A)$ is *(l,u)-admissible* if $l(a,b) \leq f(a,b) \leq u(a,b)$ holds for all $a, b \in A$.

Triangle inequality (lower/upper) limits are $(l,u)-$admissible functions

$$l_3(a,b) = \inf(f(a,b) | f \in F_3(A))$$

$$u_3(a,b) = \sup(f(a,b) | f \in F_3(A))$$

where

$$F_3(A) = \{f \in F(A) | 0 \leq f(a,b) \leq f(a,c) + f(b,c), \forall a, b, c \in A\}$$

The following theorem provides us with necessary conditions for the geometric consistency of any given distance bounds l, u

Theorem 1 *For any pair of functions $l, u \in F(A)$ the following two conditions are equivalent*

- *The functions l, u are triangle inequality lower/upper limits*
- *The function u satisfies the triangle inequality, i.e $u(a,b) \leq u(a,c) + u(b,c), \forall a, b, c \in A$ while l and u together satisfy $l(a,b) \leq l(a,c) + u(b,c), \forall a, b, c \in A$*

This characterization allows us to compute triangle inequality limits. First we will compute the upper triangle inequality limits. These are given by

$$u_3(a,b) = \sup_{f \in F(A)} (f(a,b) | f(x,y) \leq \min(u(x,y), f(x,z) + f(y,z)) \forall x,y,z \in A\}$$

This shows that upper triangle inequality limits are independent of the lower bounds l.

Shortest paths

Let G be the complete weighted graph on vertices A, weight $w(a,b) = u(a,b)$. Let $s(a,b)$ be the shortest distance in G between a and b.

Lemma 2 *The upper triangle inequality limits are equal to the shortest paths in G.*

Proof 1 *For any shortest path $s(a,b) = \inf(s(a,x) + u(x,b) | \forall x \in A)$ (Bellman equations). Let $v \in F(A)$ then $v(a,b) \geq \inf(v(a,x) + u_3(x,b) | x \in A)$. For the other direction, since $u_3 \leq u, u_3(a,b) \leq u_3(a,x) + u(x,b)$ for all $x \in A$, so that $u_3(a,b) \leq \inf(u_3(a,x) + u(x,b) | x \in A)$.*

■

Similarly, lower triangle inequality limits are given by

$$l_3(a,b) = \inf_{f \in F(A)} (f(a,b) | l(x,y) \leq f(x,y) \leq f(x,z) + u_3(y,z)) \forall x,y,z \in A\}$$

or equivalently

$$-l_3(a,b) = \sup_{-f \in F(A)} (f(a,b) | f(x,y) \leq \min(-l(x,y), f(x,z) + u_3(y,z)) \forall x,y,z \in A\}$$

This supremum is characterized by the following lemma

Lemma 3 *For all $a, b \in A$*

$$-l_3(a,b) = \inf(u_3(a,x) + u_3(b,y) - l(x,y) | x,y \in A)$$

Proof 2 *Let $f(a,b) = \inf(u_3(a,x) + u_3(b,y) - l(x,y) | x,y \in A)$. We have $-l_3(a,b) \leq -l_3(b,x) + u_3(a,x)$ as well as $-l_3(b,x) \leq -l_3(x,y) + u_3(b,y) \leq -l(x,y) + u_3(b,y)$ so that $-l_3(a,b) \leq f(a,b)$. Since $f(a,b) \leq -l(a,b)$, to prove the opposite inequality it suffices to show that $f(b,c) \leq f(a,b) + u_3(a,c)$. Because that upper bounds obey the triangle inequality, we have*

$$u_3(a,c) + f(a,b) = \inf(u_3(a,c) + u_3(a,x) + u_3(b,y) - l(x,y) | x,y \in A) \geq$$

$$\geq \inf(u_3(c,x) + u_3(b,y) - l(x,y) | x,y \in A) = f(b,c)$$

■