BioGeometry NEWS
September 2003
biogeometry.cs.duke.edu

## Research

### Flexible Spanners for Tracking Proximity in Folding Molecules

by Jie Gao, Leonidas Guibas, and An Nguyen of Stanford University

Since most molecular forces act between pairs of atoms but are short-range, it is often important to know which atoms are near which in a molecule. For example, molecular dynamics (MD) computations often maintain so-called "neighbor lists" for each atom; these are lists of all the other atoms in the molecule that are within a certain short distance of the given atom (see Fig. 1).

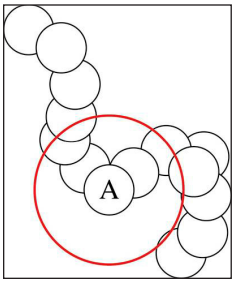

Figure 1

These lists might be used to cut-off Lennard-Jones interactions for pairs of atoms more than 10Å apart, to speed up the computation. Knowledge of atom proximities is also important in Monte-Carlo (MC) simulations, in order to generate collision-free conformational moves that are likely to be of low enough energy to be accepted. Fundamental geometric packing arguments tell us that an atom can have only a few near neighbors (but possibly many far neighbors) --- and thus near-neighbor information can be compactly encoded. What makes the proximity maintenance problem hard, however, is that a deforming molecule is a system with many degrees of freedom, all of which change across every MD or MC step.

Historically, MD codes have used for this purpose a voxel grid into which the atoms of a given conformation of the molecule are apportioned by a simple indexing process. This spatial tessellation allows, under some conditions, efficient extraction of an atom's spatial neighbors: only the tessellation cells intersected by a given atom and some of their tessellation neighbors need to be searched for these neighbors. But a voxel grid has to be tuned carefully to work well. First, since the total number of voxels can be large, a hash table is usually employed that keeps only the tessellation cells intersected by at least one atom --- but this makes finding the neighbors of a cell more cumbersome. Furthermore, the proper cell size is critical for good performance in a voxel grid. Performance degrades rapidly both when the voxel grid cells are too small (because then many and possibly empty neighboring cells have to be searched), and when they are too big (because too many pairs of atoms within a cell have to be examined).

Proximity detection is, of course, closely related to the classical problem of collision detection that has been addressed by many communities, including robotics, graphics/animation, etc. Solutions developed in those fields, however, were mostly aimed at multiple moving rigid objects. Bounding volume hierarchies have been successfully used for such collision checking in modeling complex virtual environments. But for a folding molecule, a bounding volume hierarchy can be expensive to maintain and use. A folding linear object can come into self-proximity in much more complex ways than two-dimensional or three-dimensional deformable objects can. As the molecule changes conformation, the hierarchy has to be recomputed. And even after the hierarchy is available, a self-collision or self-proximity check can be quite expensive for a shape that comes near itself in many places, requiring multiple traversals of the hierarchy. If one tries to make the hierarchy looser fitting, to avoid frequent recomputation, then these self-intersection tests become more inefficient and expensive.

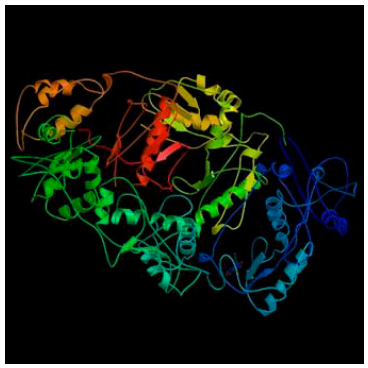
Figure 2

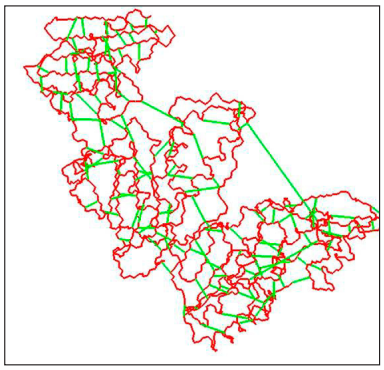
Figure 3

Motivated by a desire to address some of the drawbacks of these earlier methods, we have recently developed a lightweight, combinatorial structure that can be used for collision detection and proximity maintenance in a folding molecule. The new structure is based on the notion of a graph spanner, first developed in the context of communication networks and graph theory. A spanner of a (generally dense) graph G is much sparser subgraph S on the same vertices, so that the lengths of all edges in G are well approximated by the length shortest path between the corresponding vertices in S. Thus a spanner is a way to "compress' a graph, while preserving distance information as much as possible. In our setting, the idea is to treat a molecule as the complete graph on its atoms. The corresponding spanner S will contain all natural edges corresponding to bonds, plus a number of additional edges that we call shortcuts. The shortcuts guarantee that that the distance between any two

## People

**Robert Gdanitz**, Senior Research Associate in the Department of Physics at North Carolina A&T State University, has been named a PI on the BioGeometry project effective June 2003. Robert is a theoretical physical chemist with expertise in developing electronic structures as well as force field methods. He interacts with group members at Chapel Hill, Duke, and Stanford to further develop his method to predict molecular crystal structures. Robert has local as well as national collaborations on projects of biochemical interest, which will further strengthen the bond of NCA&T to the remainder of the BioGeometry group.

pair of atoms in the molecule can be well approximated by the shortest path between these atoms in the spanner S. For example, in Fig. 2 is a picture of a small protein, 3HVT. In Fig. 3, the backbone of 3HVT is shown, together with the shortcuts (green edges) which were added to make it a spanner for the complete Euclidean graph on the backbone atoms.

In this particular case, the spanning ratio is 3, which means that, for every pair of atoms, the length of the shortest path between them in the spanner S is at most 3 times the true Euclidean distance of the atoms. In our construction, the spanning ratio can be made as close to 1 as we like, by increasing the number of shortcuts. In general, if we have a molecule consisting of n atoms, then we can construct a spanner for that molecule that uses only O(n) edges of total length O(n log n). Furthermore, each atom has only O(log n) edges in the spanner. Thus a spanner is a lightweight encoding of all proximities between the atoms.

Once we have this spanner, we can use it to get all atoms within, say, 10Å of a given atom, by doing an easy breath-first search in the spanner graph from the given atom out to a total distance of 3 times 10Å = 30Å. This is guaranteed to capture all atoms within a Euclidean distance of 10Å (and possibly some others as well). We can even show that, when amortized over all atoms, the number of additional atoms pairs found this way is proportional to those that belong to the neighbor lists. Furthermore, before two atoms can collide, there has to be a spanner edge between them. For otherwise there would be two graph nodes that are arbitrarily close in space but for which the shortest path in the spanner cannot be made arbitrarily small.

A big part of our research has been to study how to define such spanners which are stable under molecular deformations. This is possible because spanners are highly non-canonical structures --- there are many possible spanners of roughly the same quality for a given conformation of the molecule. As a result, when the molecule deforms somewhat, there is likely to be be a spanner for the new conformation that shares many edges (shortcuts) with the old conformation. Using ideas from our earlier work with Kinetic Data Structures (KDS), we have developed stable spanners that are efficiently updateable after an external agent (MD or MC in this case) moves the atoms. Our experiments show that in typical MD simulations, only 4% of the spanner edges change after 10-100 MD steps. As mentioned, we have efficient techniques for discovering what these changes in the spanner are. In fact, the closer we can couple the physical process moving the atoms to the spanner construction, the more we can reduce the spanner maintenance cost.

It turns out that, from the spanner, we can also derive a well-separated pair decomposition of the atoms. This is a list of O(n log n) pairs of atom groups, so that each of the quadratic atom pairs is covered by some group. Furthermore, each pair is well separated, in the sense that the diameter of each group is small compared to the separation of the groups. This decomposition makes it possible to treat the electrostatic interactions between the atoms across each group pair via a dipole-dipole or other less expensive approximation, just as in the classical n-body case. But note that, unlike the classical methods of voxel grids and n-body approximations which are structures affixed to the ambient space, the spanner is a structure affixed to the molecule itself and invariant under rigid motions.

To summarize, the key properties of the spanner are:

- it is a purely combinatorial structure --- just a list of a small number of pairs of atoms
- it replaces an expensive spatial search for neighbors by an inexpensive graph traversal that is equally efficient at all distances
- it enables the efficient approximation of distant electrostatic interactions
- it is quite stable under motion/deformation and can be updated efficiently
- it suggests a “one stop” solution for all spatial proximity needs of molecular conformational search or dynamics

## Student Projects

Kevin Wedderburn, graduate student, and Sewyalew Taddele, undergraduate student, both from NC A&T, spent three weeks this summer at the group of Weitao Yang. Dr. Yang is a pioneer of density functional theory (DFT). At the time being, DFT is the most important electronic structure theory and it is widely applicable to a vast range of atoms, molecules, and solids. In 1998, DFT has been awarded with the Nobel Prize. Kevin and Sewyalew learned how to apply DFT on the notoriously difficult examples of Fe atom and organic Fe complexes. At a later stage, Kevin and Sewyalew will compute accurate atomic charges of molecules and molecular fragments using a fit to the electrostatic potential (ESP). These charges will be used to enhance the quality of biological simulations with common force field approximations.

Ibrahima “Ibou” Mbaye, undergraduate student from NC A&T, spent two months this summer with Pankaj Agarwal and Herbert Edelsbrunner at Duke. Ibou is working on the further development and computer programming of algorithms to treat rigid body orientations with so-called “quaternions.” Quaternions are 4-dimensional unit vectors that can be regarded as points on a so-called unit “3-sphere” (a 3-dimensional sphere of unit radius in 4-dimensional space). Quaternions provide a most compact representation of these orientations and they avoid the notorious biases and discontinuities that are associated with the usage of Euler angles. Ibou’s code will be used for the enhancement of an algorithm to predict molecular crystal structures.

The BioGeometry project is funded by the National Science Foundation under grant CCR-00-86013. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.