Voronoi Tesselations of Proteins for Computing Contact Maps

Gregory Zaverucha

Technical Report CS-2005-07

Jun 27, 2005

Faculty of Computer Science
6050 University Ave., Halifax, Nova Scotia, B3H 1W5, Canada

Voronoi Tesselations of Proteins for Computing Contact Maps

Gregory M. Zaverucha
Faculty of Computer Science
Dalhousie University
Halifax NS, B3H 1W5, Canada
gregory@cs.dal.ca

***Abstract*— This paper examines the use of Voronoi tessellations of proteins for generating contact maps, and compares it to existing methods. A Voronoi tesselation is computed on the three dimensional protein data, with each amino acid as a site. If two reigions in the Voronoi tesselation share a face, then the amino acids that created the regions are considered in contact. Amino acids are represented by points at their geometric centers. A simple cutoff method will be used for comparison. The results show that the Voronoi tessellation method can produce accurate, unambiguous contact maps. The new maps have on average 30% fewer contacts, yet retain the patterns found on the maps produced by the cutoff method. More importantly, the data allowed us to tune the cutoff distance used in the reference method, justifying the choice of this parameter.**

***Index Terms*— voronoi tesselation, contact map, protein**

BACKGROUND

Geometric Structures

The *convex hull* of a set of points in a two dimensional plane is the smallest convex polygon that contains all the points. In three dimensions the convex hull is a convex polyhedron containing all of the points.

For n points (sites) in a plane, the two dimensional *Voronoi diagram* divides the plane into n convex polygons, called regions. Each region contains one site. The Voronoi diagram has the property that all of the points in each region, are closest to the region's site than to any other. In three dimensions, the *Voronoi tessellation* (VT) divides the space into a set of convex polyhedra, called *Voronoi cells* (VC). Two sites are said to be neighbors in 2D (3D) if their regions (cells) share an edge (face).

The *Delaunay triangulation* is the dual structure to a Voronoi diagram. Every region in the Voronoi diagram becomes a point in the Delaunay triangulation, an edge is drawn between two points if they are Voronoi neighbors. In three dimensions the triangles are replaced by tetrahedra, and is sometimes called the *Delaunay tetrahedrization*.

The field of computational geometery has provided efficient algorithms for computing all of these structures. The quickhull algorithm computes the three dimensional convex hull with an average case running time of $O(n \log n)$ and a worse case of $O(n^2)$. The Delaunay triangulation, and Voronoi diagrams can also be computed in $O(n \log n)$ time [9]. For this application, the input size is number of amino acids in a protein, ranging from 50-1000. With such a small input set the cost of computing any of these structures (with current hardware) is small.

Contact Maps

A *contact map* (also called a *contact matrix*) of N objects is an $N \times N$ matrix, where entry $a_{ij} = 1$ if element i is in contact with element j, and 0 otherwise. For proteins, the rows and columns represent the amino acids in the chain, making the matrix symmetric. Grayscale images are an excellent way to quickly view the matricies. The images are far from random; identifiable patterns are created based on the structure of the protein.

The simplest way to generate a contact map is the cutoff method. The distance between amino acids is computed and those within the cutoff distance are considered to be in contact. Some people choose the distance between two alpha carbons, others choose the geometric center of the AA. The difficulty comes in choosing the cutoff distance. Different cutoff distances will produce different contact maps, giving ambiguous information about the same protein. In [10] the cutoff method was used to evaluate another method of creating contact maps. The distance was chosen to be $9\mathring{A}$, since this gave a similar number of non-zero contacts. This choice suited the authors' artificial situation, but there is nothing to justify it.

In one software available for generating contact maps [2], the default cutoff value is 15 $\mathring{A}$. There is nothing in the literature to suggest an appropriate cutoff value. Without guidance on what the "correct" value should be, we chose to use a cutoff value of 14 $\mathring{A}$.

Voronoi Tessellation of Proteins

The idea of using the Voronoi tessellation (VT) of a protein seems to have been first suggested as early as 1974. The VT is a common tool to study a variety of materials from from condensed matter physics (random sphere packings, foams, glass, etc ...).

There have bee two main approaches to computing the VT of proteins. One method uses each atom of the protein as the center of a Voronoi cell [7]. The other method creates a Voronoi cell around each amino acid [1], [10], [4], [3].

In [7], the authors are concerned with determining the solvent accessible surface (SAS) and the protein volume. Their Voronoi based algorithm has reduced execution time and greater precision than comparable methods for surfaces and volumes.

In [4] contact maps are created, and used to assign secondary structure. Using the area of the shared face, the contacts are classified as strong or normal. A contact is called "strong" when the area of the shared face is 2 $\mathring{A}^2$ larger than the mean area of faces in the tessellation. This definition of a strong contact is misleading, since the area of the face is

not representative of the physical contact area between the amino acids. In general and in our observations, there is no correlation between the distance between two neighboring sites and the area of their shared face.

METHODS

In this section we will outline the procedure used to create a contact map from a PDB file. Our software was written in Perl, and used the `qhull` program [8] to compute the geometric structures.

The Environment

The Voronoi cells at the proteins' surface will have infinite volume. This occurs because the faces of the regions are created from the bisecting plane between two sites. The regions' of surface sites will not form closed polyhedron; they will be open on the surface-facing side.

In order to prevent this from happening, we must embed the protein in an environment. The environment will consist of a random packing of spheres, each with a diameter of 7 $\mathring{A}$, the mean diameter of an amino acid [1].

This will also prevent extra contacts from appearing in the "dents" of the protein. When a protein has a crater in it, the Voronoi cells on each side of this crater will touch. The distance is too far for any actual interaction and this contact should not reported. By filling any craters with environment spheres, the size of these regions is kept closer to the size of the amino acid sites.

To create the environment, the Jodrey-Torey algorithm [6], was used to simulate a random packing of spheres. This algorithm is somewhat inefficient, so the environment was created in parts. Initially, 1200 spheres were packed into a $75 \times 75 \times 75$ cube, which was then translated 8 times to create a $150 \times 150 \times 150$ cube containing 9600 environment spheres. This environment was computed once and re-used for all tests.

The Protein

The coordinates of the atoms of each amino-acid were removed from the PDB file. The geometric center of the points was computed as the mean of the x, y and z values. These centers were used as representatives for each amino acid.

Embeding the Protein in the Environment

The geometric center of the entire protein was also computed. To embed the protein, it was translated so that it's center was at the center of the environment. All of the environment spheres that intersected the protein were removed.

Relaxing the Environment

Before reporting the contact map, successive Voronoi tessellations are computed. After each is computed, a shell of environment spheres are determined. Any environment sphere in contact with an amino acid is in the protein's shell. For each sphere in the shell, we compute the geometric center of it's region, and move it to this center. After each iteration this distance becomes smaller and smaller. This "regularizes" the regions, giving them a more uniform volume and shape. The authors in [1] relax the environment 9 times. The spheres remained fixed after 6, but an extra 3 were computed. We found that 5 iterations were sufficient, on the fifth iteration spheres were moving less than .01 of an Angstrom, an insignificant distance in our cube.

Computing the Voronoi Tesselation

After the last iteration relaxing the environment, we are left with a tessellation of the protein and the environment. The regions were output by `qhull` as a set of points, making it necessary to compute their convex hull. The convex hull gave the individual faces of the regions, allowing us to compute their area. Finally, the Delaunay tetrahedrization is computed to reveal neighboring sites. From it, we stored information about the distance and area of contacts and output the contact map.

RESULTS

Contact maps were computed for 325 PDB files. The sample set was chosen from a larger list [1] that was culled from the Protein Data Bank [5].

The most obvious result was that the VT contact maps had far fewer contacts than those generated by the cutoff method. On average, the VT maps showed 30% fewer contacts than the cutoff maps for the corresponding proteins. The lowest percentage of removed contacts was 12%, while the highest was 53%. Visually, the removed contacts left only the 'skeletons' of the patterns and the resulting were images much clearer (Figures 1 and 2).

We inspected the distances of the contacts that were removed to try and determine *which* contacts where removed. Over all 325 samples, the mean distance of the removed contacts was in the range 9.44 - 11.64 $\mathring{A}$. The median distances were all in the range 10.22 - 11.79 $\mathring{A}$, and the closest removed contacts were in the range 5.33 - 7.33 $\mathring{A}$.

This suggested that a cutoff value of $\approx 10\mathring{A}$ would give similar results. A few preliminary trials and visual inspections confirm this, however time did not allow for a complete analysis. Figures 1 and 2 show the contact maps made using all 3 methods described.

In some cases contacts were added by the VT method. That is, there were positions that did not have a contact with

[1] cullpdb_pc20_res1.6_R0.25_d050320_chains833.gz available at [11]

the cutoff method, that did have one with the VT method. However, the percentage added was always below 1% and these maps also had roughly 30% removed. The additions were deemed insignificant.

Since "strong" contacts were defined in [4] in terms of larger area, it was suspected that there may be some correlation between area and distance. For two neighboring sites, it was expected that the area of their shared face would be larger for closer neighbors and smaller for distant neighbors. This was not the case, no correlation was found whatsoever. In the general case, this is to be expected, since the size of a bisecting plane between two sites is determined by the surrounding sites.

Discussion

The usefulness of the VT method for contact maps will depend for a large part, on the applications that use contact maps. With the cutoff method using a value of 14 $\mathring{A}$, the resulting map may be up to 53% inaccurate. For some applications, these inaccuracies may be intolerable. On the other hand, some applications may favor extra contacts, to emphasize patterns in the map.

The computational cost of computing a contact map using the VT method is roughly one minute [2]. When compared to the cutoff method, which runs in a few seconds, this may be unacceptable for some real-time applications or studies that must process thousands of PDB files. If the contact map must be created quickly, then the cutoff method using a value of 10 $\mathring{A}$ is the best option. Alternatively, since the contact map of a protein is static, it may be computed once and reused when needed.

Future Work

There are a few other approaches that may lead to slightly more accurate contact maps. In this work, we have worked under the simplification that all amino acids can be represented by spheres $7\mathring{A}$ in diameter. In [10] the Laguerre polyhedral decomposition is presented as a tool for analyzing protein folds. It is similar to the VT method, however, the regions are weighted to reflect the varying sizes of AA. Larger AA would form a larger region in the tessellation. In the example given, the Laguerre contact map differs from the VT contact map slightly, suggesting it may be an improvement.

Ultimately, we believe the most accurate contact map could be created by inspecting atom-atom contacts. If an atom of residue A is in contact with an atom of residue B, then A and B are considered to be in contact. We will represent each atom as a sphere, using the VanderWaals radii of each atom. A cutoff method that measures the distance between two atom surfaces could be used to determine atom-atom contacts. Again, the question of choosing the cutoff distance arises. The Laguerre polyhedral decomposition could be used in place of the cutoff method, or it may be used to determine a practical cutoff distance.

With the contact map generated by atom-atom contacts in hand, the accuracy of all other methods can be evaluated against it.

[2] On modern hardware, with an Intel P4 3GHZ CPU.

Conclusions

We see a qualitative improvement in the contact maps generated by the Voronoi tessellation method. The data has also suggested an ideal cutoff distance for the existing cutoff method. This experiment has justified the choice of this parameter, which had previously been guessed or chosen to suit the situation. Applications that use contact maps will benefit from these maps and from these new insights about contacts maps.

Acknowledgements

Thanks to Christian Blouin, Chris Hamilton and Glen Hickey for their discussions and suggestions.

References

[1] B. Angelov, J.F. Sadoc, R. Jullien, A. Soyer, J.P. Mormon, and J. Chomilier. Nonatomic solvent-driven voronoi tesselation of proteins: An open tool to analyze protein folds. *Proteins*, 49(4):446–456, 2002.

[2] C. Blouin, D.J. Butt, and A.J. Roger. *libcov a c++ mini library for phylogenetics*. World Wide Web, http://www.cs.dal.ca/∼cblouin/libcov/, Visited March 2005.

[3] F. Dupuis, J.F. Sadoc, R. Jullien, B. Angelov, and J.P. Mormon. Voro3d: 3d voronoi tessellations applied to protein structures. *Bioinformatics*, 2002. Online edition only: http://bioinformatics.oupjournals.org/cgi/content/abstract/bth365v1.

[4] F. Dupuis, J.F. Sadoc, and J.P. Mormon. Protein secondary structure assignment through voronoi tessellation. *Proteins*, 55(3):519–528, 2004.

[5] Research Collaboratory for Structural Bioinformatics. *RCSB Protein Data Bank*. World Wide Web, http://www.pdb.org, Visited March 2005.

[6] W.S. Jodrey and E.M. Tory. Computer simulation of close random packing of equal spheres. *Physical Review A (General Physics)*, 32:2347–2351, October 1985.

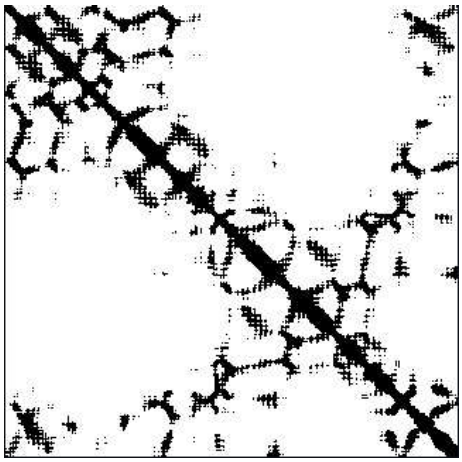
[7] B.J. McConkey, V.Sobolev, and M. Edelman. Quantification of protein surfaces, volumes and atom-atom contacts using a constrained voronoi procedure. *Bioinformatics*, 18(10):1365–1373, 2002.

[8] University of Minnesota Geometry Center. *Qhull*. World Wide Web, http://www.qhull.org, Visited March 2005.

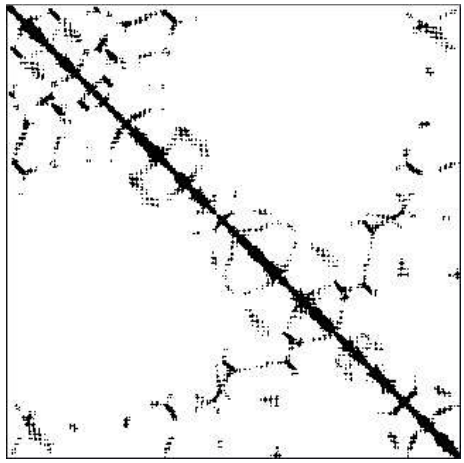
[9] Joseph O'Rourke. *Computational Geometry in C*. Cambridge University Press, New York, NY, USA, 1998.

[10] J.F. Sadoc, R. Jullien, and N. Rivier. The laguerre polyhedral decomposition: application to protein folds. *The European Physical Journal B - Condensed Matter*, 33(3):355 – 363, 2003.

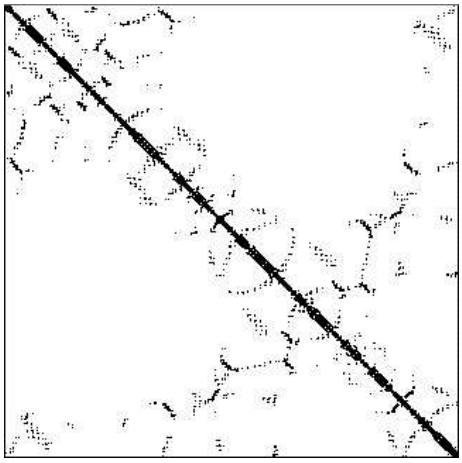
[11] G. Wang and R. L. Dunbrack Jr. *Culling the PDB by Resolution and Sequence Identity*. Wold Wide Web, http://dunbrack.fccc.edu/Guoli/pisces_download.php, Visited March 2005.



(a) 14 $\mathring{A}$ cutoff method.

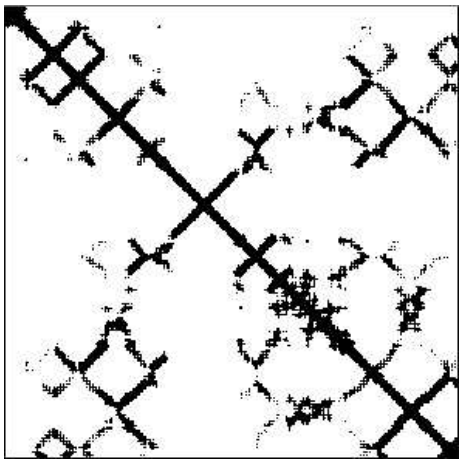


(b) 10 $\mathring{A}$ cutoff method.

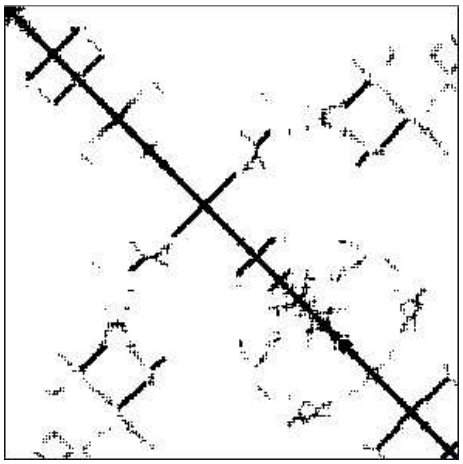


(c) VT method.

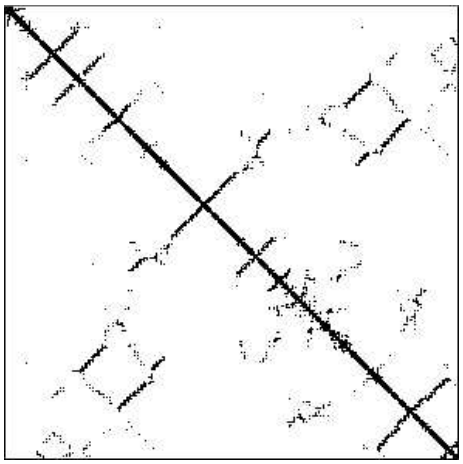
Fig. 1. Comparison of contact maps for protein 1PTM



(a) 14 $\mathring{A}$ cutoff method.



(b) 10 $\mathring{A}$ cutoff method.



(c) VT method.

Fig. 2. Comparison of contact maps for protein 7AHL