Alpha Complexes in Protein Structure Prediction

Pawel Winter and Rasmus Fonseca
Department of Computer Science, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark

Keywords: Protein Structure Prediction, Force Field, Alpha-complexes, Kinetic Data Structures.

Abstract: Reducing the computational effort and increasing the accuracy of potential energy functions is of utmost importance in modeling biological systems, for instance in protein structure prediction, docking or design. Evaluating interactions between nonbonded atoms is the bottleneck of such computations. It is shown that *local* properties of α-*complexes* (subcomplexes of Delaunay tessellations) make it possible to identify nonbonded pairs of atoms whose contributions to the potential energy are not marginal and cannot be disregarded. Computational experiments indicate that using the local properties of α-complexes, the relative error (when compared to the potential energy contributions of *all* nonbonded pairs of atom) is well within 2%. Furthermore, the computational effort (assuming that α-complexes are given) is comparable to even the simplest and therefore also fastest cutoff approaches.
The determination of α-complexes from scratch for every configuration encountered during the search for the native structure would make this approach hopelessly slow. However, it is argued that *kinetic* α-complexes can be used to reduce the computational effort of determining the potential energy when “moving” from one configuration to a neighboring one. As a consequence, relatively expensive (initial) construction of an α-complex is expected to be compensated by subsequent fast kinetic updates during the search process.
Computational results presented in this paper are limited. However, they suggest that the applicability of α-complexes and kinetic α-complexes in protein related problems (e.g., protein structure prediction and protein-ligand docking) deserves furhter investigation.

1 INTRODUCTION

In protein structure prediction a vast atomic configuration space has to be searched when looking for the *native* configuration minimizing its potential energy. Good potential energy estimators require substantial computational effort. Reducing this effort is therefore important. Furthermore, similar searches and potential energy estimations arise in for example protein-protein docking and in protein design.

Interactions between nonbonded atoms are the computational bottleneck of potential energy estimations. Commonly used cutoff methods compute the distances between all pairs of nonbonded atoms and calculate the contributions of those within some prespecified cutoff distance. Different types of contributions such as van der Waals and Coulomb potentials may require different cutoff values (Schlick, 2010). Hierarchical decompositions of proteins with appropriately chosen bounding volumes have also been used to speed up potential energy estimations (Lotan et al., 2004; Winter and Fonseca, 2012). We show that α-complexes (which are subcomplexes of well-known Delaunay tessellations) for appropriately chosen real values of α, $\alpha \geq 0$, are well-suited for the identification of nonbonded pairs of atoms essential for the estimation of potential energy of proteins. The identification of such pairs involves exploiting the structural properties of α-complexes while making the distance computations for cutoff purposes unnecessary. Computational experiments reported in this paper indicate that the relative error is well within 2% while the computational effort is comparable with even the simplest cutoff approaches.

Searching for a configuration minimizing the potential energy typically involves perturbing one configuration to obtain the next. It can for example be achieved by small dihedral rotations of covalent bonds. As these rotations are carried out, the underlying α-complexes can be appropriately updated. We sketch how these updates can be carried out using the kinetic data structure framework. In particular, bond rotations imply that groups of atoms rotate around the same axis with the same rotational speed on circular orbits in parallel planes. This significantly speeds up the computations needed to update α-complexes. Fast

Winter P. and Fonseca R..
Alpha Complexes in Protein Structure Prediction.
DOI: 10.5220/0005251401780182
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2015), pages 178-182
ISBN: 978-989-758-070-3

updates, in turn, imply fast determination of potential energy for the next neighboring configuration.

2 SYSTEMS AND METHODS

A *force field* is a collection of parameters and mathematical expressions that together define a function approximating the potential energy of a system of atoms. Such a function typically includes *bonded* terms capturing forces between covalently bonded atoms and *nonbonded* terms capturing forces of nonbonded atoms.

A simple but still reasonable force field approximating the potential energy (in kcal/mol) of a particular conformation of a protein (Levitt et al., 1995; Schlick, 2010) is given by $E = E_B + E_N$ where the E_B term comprises the contributions from bonded atoms and the E_N term comprises the potential energy contributions from nonbonded atoms. It is defined by $E_N = E_{NV} + E_{NC}$, where

$$E_{NV} = \sum_{p=(i,j)} (e_{ij}[\frac{r^e_{ij}}{r_{ij}}]^{12} - 2e_{ij}[\frac{r^e_{ij}}{r_{ij}}]^6)$$

and

$$E_{NC} = 332 \sum_{p=(i,j)} \frac{q_i q_j}{r_{ij}}$$

where p is a pair of atoms i and j, $r^e_{ij} = (r^e_i + r^e_j)/2$ with r^e_i and r^e_j being van der Waals radii of interacting atoms, r_{ij} is the actual distance between the atoms i and j, $e_{ij} = \sqrt{e_i e_j}$ with e_i and e_j being partial charge parameters of interacting atoms. Finally, q_i and q_j are partial charges of the atoms i and j.

Nonbonded interactions between atoms separated by less than three bonds along the covalent structure are disregarded (Levitt et al., 1995). Their interactions are assumed to be accounted for by bonded interactions.

It is evident that nonbonded terms depend on the distances between interacting atoms. As the distances grow, the potential energy contributions become negligible. In order to speed up the computations, van der Waals interactions between atoms more than 8-12Å apart can be disregarded. The cutoff distance for E_{NC} is higher (Levitt et al., 1995). Other, more sophisticated cutoff techniques have been suggested (Schlick, 2010).

3 ALGORITHMS

An *α-ball* b_p centered at a point p with radius α, $\alpha \geq 0$, is the set of points at most α away from p. Let S be a set of n points in the 3-dimensional Euclidean space E^3. Let T be a subset of S of size $|T| = k+1$, $0 \leq k \leq 3$. The convex hull σ_T of T is also referred to as a *k-simplex*. Given a k-simplex σ_T, any k'-simplex $\sigma_{T'}$, $T' \subset T$, is a (proper) *face* of σ_T. A k-simplex σ_T, $0 \leq k \leq 3$, belongs to the *Delaunay complex*, ($\mathcal{DC}$) iff there exists a sphere that passes through the points of T and all other points of S are strictly outside this sphere. Assuming the general position (i.e., no five points lie on a common sphere), the Delaunay complex is a simplicial complex.

Let b_T denote the smallest ball that passes through the points of a k-simplex T, $0 \leq k \leq 3$. Let ρ_T denote the radius of b_T. If b_T contains no point of S in its interior, σ_T is called *Gabriel*. Let α be a positive real number. T is said to be α-short if $\rho_T \leq \alpha$. The *α-complex* S_α consists of Delaunay simplices that are Gabriel and α-short together with their proper faces (Edelsbrunner, 1992).

Since $S_{\alpha_1} \subseteq S_{\alpha_2}$ for $\alpha_1 \leq \alpha_2$, α-complexes provide a *filtration* of a growing family of simplicial complexes, beginning with S_0 (= S) and ending with S_∞ (= $\mathcal{DC}$). A filtration has $O(n^2)$ α-complexes since there are $O(n^2)$ k-simplices, $0 \leq k \leq 3$ (Seidel, 1995). α-complexes for the protein 2JZC with 224 amino acids and 3170 atoms and for selected values of α are shown in Fig. 1.

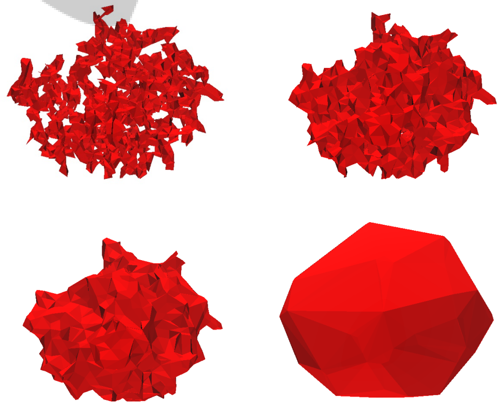

Figure 1: α-complexes for the protein 2JZC with $\alpha = 1.5, 2.0, 3.0$ and ∞.

As α grows, more and more simplices are included in S_α. A simplex σ_T becomes a member of S_α when $\alpha = \rho_T$ and b_T is exposed. As α continues to grow, σ_T can become a face of another simplex (for example a triangle can become a face of a tetrahedron). Later on, it can become an interior simplex (for example a face of tetrahedron can become a face of another tetrahedron). The α-values of these events can be easily computed from the DC (Edelsbrunner et al., 1998). This is important since in many applications one is not really interested in S_α but in the simplices on the boundary of S_α. For a survey of the applications of α-complexes to various protein-related problems, see (Zhou and Yan, 2014; Winter et al., 2009).

We make a simplifying assumption that all atoms have the same size. Consequently, we focus on α-complexes of points (or balls with the same radius). *Weighted* α*-complexes* (Edelsbrunner and Mücke, 1994) and β*-complexes* (Kim et al., 2006) for spheres of different sizes have also been developed. However, these more complicated structures require more computational effort to be constructed and updated.

Consider an α-complex S_α for some protein and for a fixed value of α, $\alpha > 0$. S_α can also be viewed as an undirected graph $G(\alpha)$ with the vertices corresponding to the atom centers and the edges corresponding to 1-simplices of S_α. Since the lenghts of covalent bonds in proteins are assumed to be fixed and are between 1Å and 1.5Å, $G(\alpha)$ will be connected already for $\alpha \approx 0.80$. For each vertex i, let $N_d(i)$ denote the vertices of $G(\alpha)$ that can be reached from i by traversing at most d edges. Let $N^*_d(i)$ denote the subset of $N_d(i)$ containing vertices representing atoms that are at least three covalent bonds away from i. Let $E^*_N = E^*_{NV} + E^*_{NC}$ where

$$E^*_{NV} = \frac{1}{2} \sum_{i \in G(\alpha)} \sum_{j \in N^*_d(i)} \left(e_{ij}\left[\frac{r^e_{ij}}{r_{ij}}\right]^{12} - 2e_{ij}\left[\frac{r^e_{ij}}{r_{ij}}\right]^{6}\right)$$

and

$$E^*_{NC} = \frac{332}{2} \sum_{i \in G(\alpha)} \sum_{j \in N^*_d(i)} \left[\frac{q_i q_j}{r_{ij}}\right]$$

We are interested in estimating relative errors

$$\varepsilon_{NV} = |E^*_{NV} - E_{NV}|/|E_{NV}|$$

and

$$\varepsilon_{NC} = |E^*_{NC} - E_{NC}|/|E_{NC}|$$

at different values of α and d.

4 IMPLEMENTATION

The applicability of α-complexes to the determination of potential energy of proteins was carried out as follows. In the first stage, 5 proteins (1X5R, 1X0O, 1XDX, 1AKP, 1Y6D) with 110-120 amino acids were tested. This was done to verify the stability of the approach as well as to estimate the quality of the solutions obtained. In the second stage, two bigger proteins (2ZJC, 224 amino acids and 3WCZ, 308 amino acids) were investigated to check if the approach remains robust as the size of proteins increases.

The relative error ε_{NV} of the van der Waals contributions to the potential energy of 1X0O remains the same already for $\alpha > 1.6$Å, $d = 1,2,...,5$. Furthermore, $\varepsilon_{NV} \approx 0\%$, $d = 3,4,5$ while $\varepsilon_{NV} \approx 0.22\%$ for $d = 2$ and $\varepsilon_{NV} \approx 1.86\%$ for $d = 1$. Similar relative errors were observed for the other four proteins with 110-120 amino acids. For the larger proteins 2JZC and 3WCZ, ε_{NV} was also stable for $\alpha > 1.6$Å. Furthermore, for 2JZC, $\varepsilon_{NV} \approx 0\%$ for $d = 3,4,5$ while $\varepsilon_{NV} \approx 0.19\%$ for $d = 2$ and $\varepsilon_{NV} \approx 1.74\%$ for $d = 1$. For 3WCZ, $\varepsilon_{NV} \approx 0\%$ for $d = 3,4,5$ while $\varepsilon_{NV} \approx 0.41\%$ for $d = 2$ and $\varepsilon_{NV} \approx 3.16\%$ for $d = 1$.

Fig. 2 shows the relative error ε_{NC} of the Coulomb contributions to the potential energy of 1X0O for the values of $\alpha = 1.0, 1.1., ..., 5.9$. It can be seen that $\varepsilon_{NC} \leq 2\%$ already for $\alpha > 1.6$Å and $d = 3,4,5$. Also, $\varepsilon_{NC} \leq 2.5\%$ for $d = 2$. For $d = 1$ and $\alpha > 1.6$, $\varepsilon_{NC} \approx 6\%$. Similar relative errors were observed for the other four smaller proteins with 110-120 amino acids. For 2JZC with 224 amino acids, $\varepsilon_{NC} \leq 1.5\%$ for $\alpha > 1.6$ and for *all* $d = 1,2,...,5$. It is perhaps somewhat surprising that this was the case for $d \leq 2$. For 3WCZ with 308 amino acids, $\varepsilon_{NC} \leq 1.5\%$ for $\alpha > 1.6$ and $d = 3,4,5$. For $d = 2$ and $\alpha > 1.6$, $\varepsilon_{NC} \leq 4\%$ while for $d = 1$ and $\alpha > 1.6$, $\varepsilon_{NC} \leq 7\%$.

Fig. 3 shows the time (in ms) needed to compute van der Waals and Coulomb contributions of non-bonded atom pairs of 1X0O (using Mac OS X with 1.7 Ghz Intel Core i5 processor). The graph for $d = 1$ is not shown as it overlaps with the x-axis and is also covered by the graph for $d = 2$. Similar computational times were observed for the other 4 proteins with 110-120 amino acids. The computational time does not include the construction of the α-complex but it includes the determination of $N^*_d(i)$ for each vertex i. Not surprisingly, the computational time increases with d as well as with α. However, for $d = 1$ and $\alpha < 2$Å, the computational time is below 2.1 ms. More interestingly, for $d = 2$ and $\alpha < 2$Å (where ε_{NV} and ε_{NC} are reasonably small), the computational time is below 28 ms. For $d = 3$ and $\alpha < 2$Å, the computational time is below 230 ms. For comparison, computational time with cutoff = 8Å is below 27 ms, with cutoff = 20Å is below 52 ms, and without cutoff is below 216 ms. Hence, picking $d = 2$ and $\alpha \approx 1.8$Å seems to be a good choice when computing nonbonded contributions to the potential energies of protein conformations.

Computational times for 2JZC and 3WCZ are of course higher. However, for $d = 2$ and $\alpha < 2$Å, the computational time for 2JZC is below 75 ms and it is below 110-120 ms for 3WCZ.

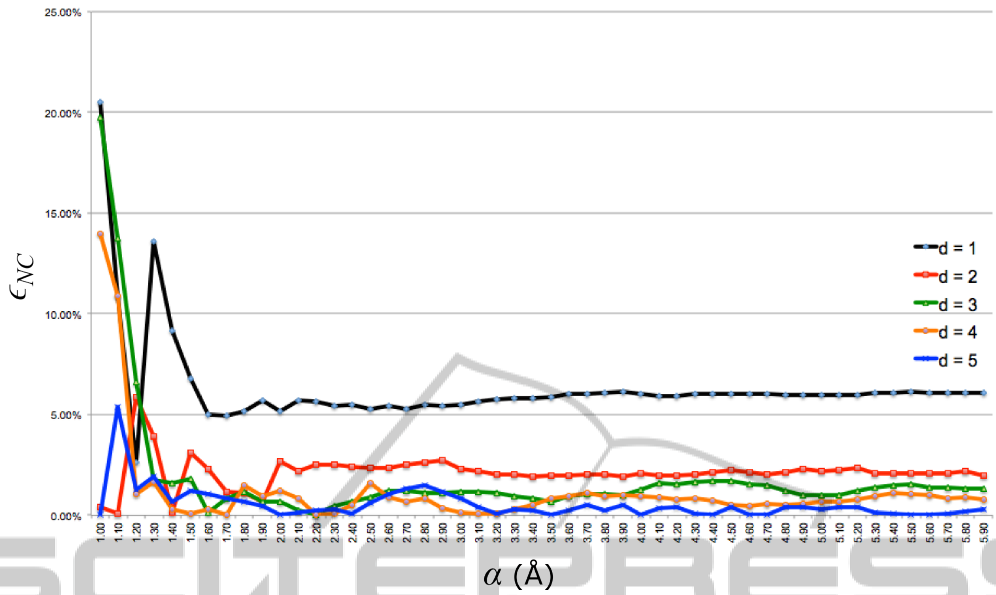

Figure 2: Relative errors for Coulomb interactions in 1X0O.

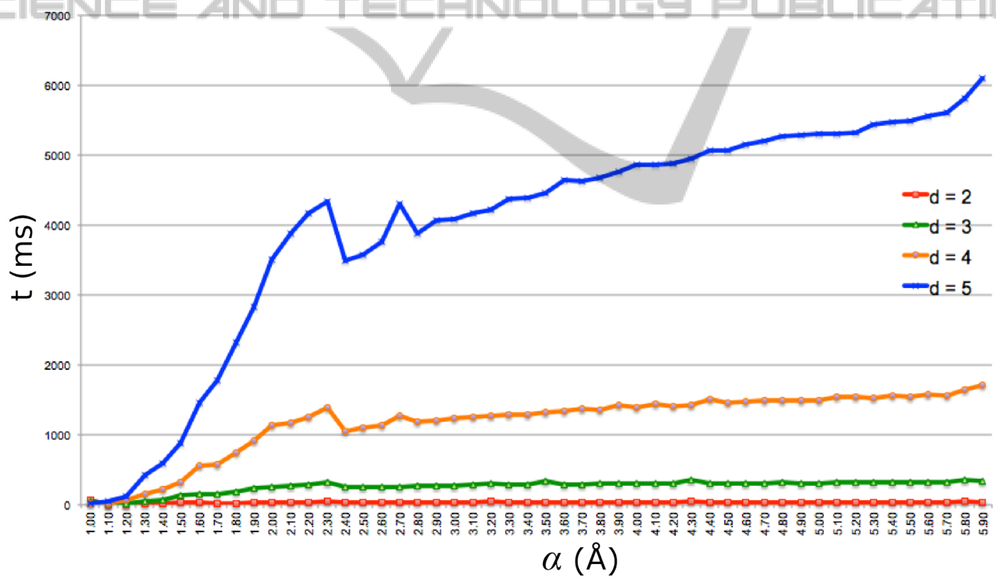

Figure 3: Computational times for 1X0O (in ms).

5 DISCUSSION

The results show that $E_{NV} \approx E^*_{NV}$ and $E_{NC} \approx E^*_{NC}$ for various proteins, already for small d (= 2), and small α ($\approx$ 2Å). Hence, 2-complexes seem to provide a useful discrete structure that can be used to speed-up the determination of the potential energy of protein configurations. This conclusion is of course only valid if 2-complexes are given beforehand. Otherwise, the determination of a 2-complexes with 2000 or more atoms would be computationally much more expensive than when the potential energy is determined by any cutoff approach.

However, the use of 2-complexes in potential energy estimations deserves further investigation. In the protein structure prediction, a vast number of possible protein configurations is examined when searching for the native one minimizing the potential energy. The search typically involves moving from one configuration to the next. One way to define neighborhood of a configuration is by dihedral rotations of one or several covalent bonds by some, usually small, angle. Atoms on one side of the rotating bond remain stationary while the others rotate on orbits in parallel planes and with centers on a common axis. Similarly, if a covalent bond on a side chain is rotated, only a very limited number of atoms on the side chain rotates while all other atoms remain stationary.

Kinetic data structures for objects moving on piecewise continuous trajectories are far from trivial. The determination of how and when such data structures must be updated typically involves finding roots of high-degree polynomials. For $\mathcal{DC}$s, deciding when a k-simplex T, $k \leq 3$, becemes (seizes to be) Delaunay involves finding roots of polynomials of 8-th degree (Russel, 2007). In α-complexes, it is necessary to determine when a k-simplex T, $k \leq 3$, becomes (seizes to be) Gabriel and when it becomes (seizes to be) short (Kerber and Edelsbrunner, 2013).

Fortunately, when rotating covalent bonds of proteins, the computational effort of updating kinetic data structures can be significantly reduced. It can be shown that kinetic $\mathcal{DC}$s and kinetic α-complexes for this kind of restricted and coordinated movement of objects (with the same rotational velocity) involve finding roots of polynomials of degree at most 4. Furthermore, as the results presented in this paper indicate, the depth d of a neighborhood $N^*_d(i)$ of vertex i does not need to be greater than 2 or 3. Hence, these neighborhoods can be updated efficiently along with the α-complexes.

In conclusion, α-complexes of proteins with relatively low α do capture the potential energy contributions of nonbonded atoms. Furthermore, kinetic α-complexes for restricted types of motion can prove useful in protein structure prediction when searching through the vast atomic configuration space.

REFERENCES

Edelsbrunner, H. (1992). *Weighted alpha shapes*. University of Illinois at Urbana-Champaign, Department of Computer Science.

Edelsbrunner, H., Facello, M., and Liang, J. (1998). On the definition and the construction of pockets in macromolecules. *Discrete Applied Mathematics*, 88(1):83–102.

Edelsbrunner, H. and Mücke, E. P. (1994). Three-dimensional alpha shapes. *ACM Transactions on Graphics*, 13(1):43–72.

Kerber, M. and Edelsbrunner, H. (2013). 3d kinetic alpha complexes and their implementation. In *ALENEX*, pages 70–77. SIAM.

Kim, D.-S., Seo, J., Kim, D., Ryu, J., and Cho, C.-H. (2006). Three-dimensional beta shapes. *Computer-Aided Design*, 38(11):1179–1191.

Levitt, M., Hirshberg, M., Sharon, R., and Daggett, V. (1995). Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Computer Physics Communications*, 91(1):215–231.

Lotan, I., Schwarzer, F., Halperin, D., and Latombe, J.-C. (2004). Algorithm and data structures for efficient energy maintenance during monte carlo simulation of proteins. *Journal of Computational Biology*, 11(5):902–932.

Russel, D. (2007). *Kinetic Data Structures in Practice*. PhD thesis, Stanford, CA, USA.

Schlick, T. (2010). *Molecular Modeling and Simulation: An Interdisciplinary Guide*, volume 21. Springer.

Seidel, R. (1995). The upper bound theorem for polytopes: an easy proof of its asymptotic version. *Computational Geometry*, 5(2):115–116.

Winter, P. and Fonseca, R. (2012). Adjustable chain trees for proteins. *Journal of Computational Biology*, 19(1):83–99.

Winter, P., Sterner, H., and Sterner, P. (2009). Alpha shapes and proteins. In *ISVD'09. Sixth International Symposium on Voronoi Diagrams*, pages 217–224. IEEE.

Zhou, W. and Yan, H. (2014). Alpha shape and delaunay triangulation in studies of protein-related interactions. *Briefings in Bioinformatics*, 15(1):54–64.