Approximate Point Set Match for Partial Protein Structure Alignments

Arno Formella
Universidad de Vigo
Department of Computer Science
Edificio Politécnico
E–32004 Ourense, Spain
Email: formella@ei.uvigo.es

keywords: structural flexible alignment, parcial pattern recognition

Abstract

We present a practical algorithm and its implementation which is able to find partial and approximative matches of patterns, such as small parts of molecules, in large search spaces, such as complete proteins. The search pattern can be matched in a flexible way including torsion of parts of the pattern. The alignment of the pattern in the search space can be based on distance functions other than RMSD. Being a general method for the approximate point set match problem, the algorithm can be applied to other fields of pattern recognition.

1 Introduction

Approximate point set match is the task to align a search pattern (a 3D point set) in a—usually larger—search space. The challenge is not to find only perfect matches, but allowing certain types of flexibility of the search pattern when embedded in the search space. The match itself might be only approximative allowing some deviation between the pattern and its counterpart in the search space. The measurement of the deviation can be based on the root mean square (RMSD) distance, but other distance functions might be of interest as well. Algorithms for perfect matches can be found for instance in [2, 3, 9].

Structural alignment of molecules taking into account the flexibility of the structure to be aligned is an important task in computational biology as stated for instance in [4, 8]. An overview of different algorithms for aligning molecules is presented in [8]. The methods include geometric hashing, clique detection, distance geometry and general optimization methods such as genetic algorithms and Gaussian overlap optimization. The article contains an extended reference section. The main disadvantage of clique detection is the huge amount of memory needed to hold the necessary data structures. The disadvantages of geometric hashing are that the preprocessing time increases with $O(n^4)$ and that the grid size used to build the hash table has a strong impact on runtime and type of matches found. With geometric hashing flexible patterns cannot be handled.

The approach we present in this article is based on distance geometry and graph matching. The algorithm is able to find partial matches of the search pattern. The search pattern can be flexible to a certain extent including torsion. The alignment of the pattern in the search space can be based on any distance function.

2 Problem description

Let $S = \{s_0, \ldots, s_{n-1}\} \subset \mathbb{R}^3$, $|S| = n > 1$, be a finite set of 3D points, the *search space*. Let $P = \{p_0, \ldots, p_{k-1}\} \subset \mathbb{R}^3$, $n \geq |P| = k > 1$, be a finite set of 3D points, the *search pattern*. We call a surjective function, $\mu : P \longrightarrow S$, a *matching function*. As transformations $\tau : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ we consider only *rigid motion transformations* (or Euclidean transformations) with a possible scaling. With the help of a suitable deviation function $D : (\tau(P), \mu(P)) \longrightarrow \mathbb{R}^+$ we assign

a quality to a given match. Possible deviation functions are, e.g., *root mean square distance*, *maximum distance* or *average distance*.

To compute the optimal rotation matrix for a rigid motion transformation minimizing RMSD the Kabsch algorithm [6, 7] is used. An optimal scaling factor v can be derived as well. For other, possibly not differentiable deviation functions, we use an iterative derivative–free minimization method [5]. To achieve the necessary, unconstrained problem we work in quaternion space.

Given S, P, and a suitable deviation function D, we call (μ, τ) a *match* of P in S. Hence, the objective of the approximate point set match algorithm is to find the match (μ, τ) with smallest deviation. An obvious extension of the problem is the following: find the maximum subset of P that can be matched with at most the given deviation.

3 Complete Matches

Let $G_P = (V_P, E_P)$ be a distance graph over P. We can use the complete graph, an edge reduced graph which maintains the rigidity of the structure, or a user supplied graph which describes a structure with certain degrees of freedom, e.g., hinges in specific areas. We call $G_S(G_P)_\varepsilon$ with node set V_S and edge set E_S the distance graph over S induced by the distance graph G_P and the tolerance ε, if $V_S = S$ and $(s_k, s_l) \in E_S$ whenever there exists an edge $(p_i, p_j) \in E_P$ which has a similar distance $\mathrm{dis}(p_i, p_j)$ compared to the distance between s_k and s_l. Similarity is guided by a suitable function taking ε into account.

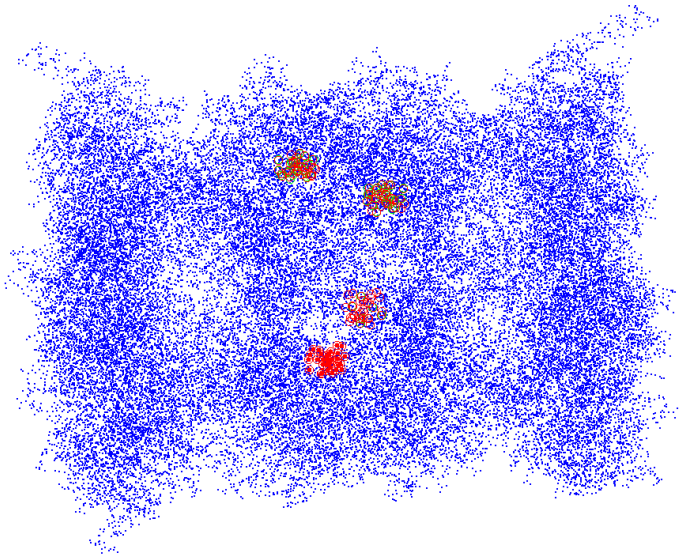
The three main parts of the approximate point set match algorithm are: First, the algorithm generates appropriate and possibly small distance graphs over P and S. Then, it enumerates with the help of a backtracking algorithm all possible matching functions. Finally, for each possible match it computes the rigid motion transformation which exhibits the smallest deviation and selects the best matches.

If the longest edge in G_P is small compared to the farest distance of two points in S—which is the case in the proposed application—, we take advantage of a regular grid over the box containing S to reduce the number of comparisons. Heuristics are used in the backtracking algorithm to order the points in P in such a way that the algorithm early discards 'non profitable' partial matches. If no match is found, the best prefix list is reported.

Additional restrictions can be imposed on the candidates in S to be matched to points in P, e.g., in a search for patterns in a protein, one might want that the element types of the atoms to be matched are equal or at least similar in some sense. Obviously, if we skip the calculation of the rigid motion transformation and consider only the graph matching, deformed graphs can be found. As deviation of the match the RMSD (or any other suitable function) of the differences of the edges of the matched graph can be taken. It is trivial to include so–called *L–matches*, where the pattern is reflected on a plane, as possible solutions. Another extension of the basic algorithm is to use absolute or individual tolerances ϵ for the different edges of the graph.

As described so far, the algorithm finds matches for a search pattern P provided there is a connected (and possibly rigid) graph G_P available. To find a maximum partial match, we apply a reactive tabu search approach adapted from [1].

4 Implementation and Experiments



Everything described in this article is implemented in a program called `psm`. The figure shows a part of the large protein 1IRU with 47562 atoms where `psm` found all matches of a small excerpt of 34 atoms. The excerpt was encountered four times, once perfectly and three times approximately ($\varepsilon = 0.18$). The runtime was less than a

minute on a 900MHz platform. More information can be found on `www.ei.uvigo.es/~formella/inv/aaa/psm_en.html`.

Acknowledgment

We want to thank Thorsten Pöschel, Kristian Rother, and Hans–Peter Lenhof for their help and valuable comments.

References

[1] R. Battiti and M. Protasi. Reactive local search for the maximum clique problem. *Algorithmica*, 29(4):610–637, 2001.

[2] L. Boxer. Faster point set pattern matching in 3-d. *Pattern Recognition Letters*, 19:1235–1240, 1998.

[3] L. Boxer and R. Haralick. Even faster point set pattern matching in 3-d. In *Proceedings of the International Society for Optical Engineering Conference (SPIE)*, volume 3811, pages 168–178, July 1999.

[4] P.W. Finn, L.E. Kavraki, J.-C. Latombe, R. Motwani, Ch.R. Shelton, S. Venkatasubramanian, and A. Yao. RAPID: Randomized pharmacophore identification for drug design. In *Proc. of 13th Annual ACM Symposium on Computational Geometry*, pages 324–333, 1997.

[5] U.M. García-Palomares and J.F. Rodríguez. New sequential and parallel derivative–free algorithms for unconstraint optimization. *SIAM Journal on Optimization*, 13(1):79–96, April 2002.

[6] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica A*, 32:922–923, 1976.

[7] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica A*, 34:827–828, 1978.

[8] Ch. Lemmen and Th. Lengauer. Computational methods for the structural alignment of molecules. *J. of Computer-Aided Molecular Design*, 14:215–232, 2000.

[9] P.B. van Wamelen, Z. Li, and S.S. Iyengar. A fast expected time algorithm for the 2-d point pattern matching problem. *Pattern Recognition Journal*, 37(8):1699–1711, 2004.