A Generic Framework for Geometrically Matching Molecular Shapes

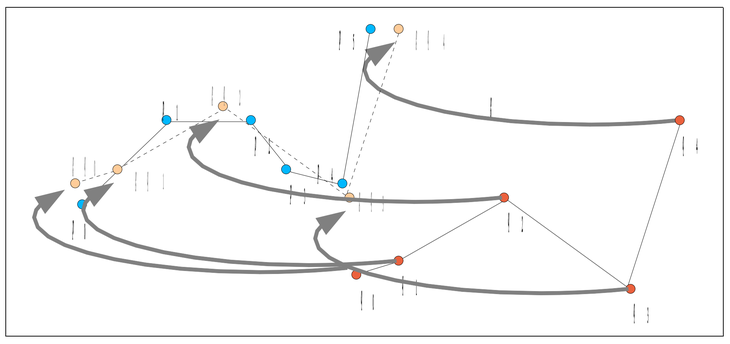
Michael Clausen
Department of Computer Science III
Universität Bonn, Germany
clausen@iai.uni-bonn.de

Axel Mosig
Bioinformatics
Universität Leipzig, Germany
axel@bioinf.uni-leipzig.de

Abstract

Motivated by problem settings such as the determination of motifs in proteins or molecular docking, we present a generic framework for finding geometric similarities between two molecular shapes. Our approach is based on minimizing a distance between the two given shapes, where a problem-specific distance function can be chosen from a certain class of distance measures, the so-called *relational distance measures*.
The setting we investigate is as follows: we are given two molecules, modeled as point sets (or, in some cases, as point sequences) P and Q in $\mathbb{R}^3$, where each point represents a chemical entity such as a single atom or an amino acid of a protein. Furthermore, we are given a distance measure $\mathbf{d}$ between point sets such that $\mathbf{d}(P,Q)$ measures the resemblance of two molecules in a fixed spatial position, with values of $\mathbf{d}(P,Q)$ close to zero indicating large resemblance; the resemblance usually changes when one of the molecules, say Q, is rotated or translated (i.e., transformed by a rigid motion g). In this setting, many typical pattern matching problems involving molecular structures can be stated as either determining the *global resemblance between P and Q* or finding *largest common substructures of P and Q* w.r.t. some suitable distance measure $\mathbf{d}$.

Problem Setting



We are intested in two settings:

- *Global resemblance between P and Q:* It is our goal to find a transformation g that minimizes the distance between P and Q, i.e., $\arg\min_{g\in\mathrm{RM}(3)}\mathbf{d}(P,gQ)$, with RM(3) denoting the set of all rigid motions in three dimensions and gQ denoting Q transformed by $g \in \mathrm{RM}(3)$.
- *Largest common substructures of P and Q:* Given a fault tolerance $\varepsilon \geq 0$, we want to determine largest possible substructures P' of P and Q' of Q such that $\mathbf{d}(P',gQ') \leq \varepsilon$ for some transformation g.

Relational Distance measures

- Given two families of points $P = \langle p_0,\ldots,p_m\rangle$ and $Q = \langle q_0,\ldots,q_n\rangle$ as well as a fault tolerance $\varepsilon \geq 0$, we obtain a relation

$$R(P,Q,\varepsilon) := \{(i,j) \mid \|p_i - q_j\| \leq \varepsilon\} \subseteq [1:m]\times[1:n].$$

- If deciding whether $\mathbf{d}(P,Q) \leq \varepsilon$ can be done by looking at $R(P,Q,\varepsilon)$, we say that the distance measure $\mathbf{d}$ is *relational*.
- *More formally:* We say that a distance measure $\mathbf{d}$ is *relational* iff for all $m,n > 0$ there is a set of relations $\mathbf{R}(\mathbf{d},m,n) \subseteq 2^{[1:m]\times[1:n]}$ so that

$$\mathbf{d}(P,Q) \leq \varepsilon \iff R(P,Q,\varepsilon) \in \mathbf{R}(\mathbf{d},m,n).$$

  for all point sequences P and Q of lengths m and n, respectively.
- Certain chemical and/or physical features at the points p_i and q_j can be taken into account as well.

Many distance measures considered in related work are relational distance measures

- *Directed Hausdorff distance:* [6, 10, 5]: $d_{\mathrm{H}}(Q,P) := \max_{q\in Q}\min_{p\in P} d(q,p)$.
- *Undirected Hausdorff distance:* [10, 5]: $\mathbf{d}_{\mathrm{H}}(P,Q) := \max\{d_{\mathrm{H}}(P,Q), d_{\mathrm{H}}(Q,P)\}$.
- *Botteleneck Distance:* [3, 4]: $\mathbf{d}_{\mathrm{B}}(P,Q) = \min_{\pi\in S_n}\max_{i\in[1:n]}\|p_{\pi(i)} - q_i\|$, where S_n denotes the set of all permutations of $[1:n]$.
- *Discrete Fréchet distance:* [9, 8, 7]: $\mathbf{d}_{\mathrm{F}}(P,Q) = \min_{(\kappa,\lambda)}\|P\circ\kappa - Q\circ\lambda\|_\infty$, where κ and λ range over the set of all increasing and surjective mappings from $[0:m+n]$ to $[0:m]$ and $[0:m+n]$ to $[0:n]$, repsectively.

Candidate Transformations

- Candidate transformations are the building blocks for our pattern matching algorithms.
- Let $V = \mathbb{R}^3$, and let $A,B \in V^3$, where $A = (a_1,a_2,a_3)$ and $B = (b_1,b_2,b_3)$.
- Two points a_1,a_2 define a ray $[a_1;a_2]$.
- Three points a_1,a_2,a_3 define a half plane $[a_1;a_2;a_3]$.
- We say that $g \in \mathrm{RM}(3)$ is an (A,B)-*candidate transformation* iff g establishes the following three conditions of coincidence, collinearity and coplanarity between A and gB:

  (Coincidence) $a_1 = gb_1$ and
  (Collinearity) $[a_1;a_2] = [gb_1;gb_2]$ and
  (Coplanarity) If neither a_1,a_2,a_3 nor b_1,b_2,b_3 are collinear, we have $[a_1;a_2;a_3] = [gb_1;gb_2;gb_3]$.
- Some care needs to be taken for degenerate cases (i.e., if either the points in A or the points in B are collinear)
- If neither the three points of A nor the three points of B are collinear, the candidate transformation is uniquely defined.

Global resemblance between P and Q

- *Goal:* Find a transformation $g \in \mathrm{RM}(3)$ that minimizes $\mathbf{d}(P,gQ)$.
- Obtain more efficient algorithms for *approximate* answer: compute $g \in \mathrm{RM}(3)$ such that $\mathbf{d}(P,gQ) \leq c\varepsilon$, for some fixed $c > 0$ and any $\varepsilon > \inf_{g\in G}\mathbf{d}(P,gQ)$
- Generalizing results from [5, 4, 1], we obtain the following algorithm:

Algorithm 1 (Global Resemblance)

Input: $P \in V^{[1:m]}$ *and* $Q \in V^{[1:n]}$*; relational distance measure* $\mathbf{d}$.
Output: $g \in \mathrm{RM}(3)$ *such that* $\mathbf{d}(P,gQ) \leq 16\varepsilon$, *for any* $\varepsilon > \inf_{g\in G}\mathbf{d}(P,gQ)$.

```
Candidate-Match(P, Q, d)
    D := ∞;
    for (i1,i2,i3) ∈ {(μ1,μ2,μ3) ∈ [1:m]^3 | μ1≠μ2, μ1≠μ3, μ2≠μ3}
       for (j1,j2,j3) ∈ {(ν1,ν2,ν3) ∈ [1:n]^3 | ν1≠ν2, ν1≠ν3, ν2≠ν3}
          A := (p_{i1}, p_{i2}, p_{i3});
          B := (q_{j1}, q_{j2}, q_{j3});
          Compute an (A,B)-candidate transformation g;
          d := d(P, gQ);
          if (d < D) then D := d; h := g;
    return h;
end.
```

- *Complexity:* $O(m^3n^3T(\mathbf{d},m,n))$, where $T(\mathbf{d},m,n)$ denotes the time required for computing $\mathbf{d}(P,Q)$.
- Time complexity can be reduced to $O(m^2nT(\mathbf{d},m,n))$ if $\mathbf{d}$ has a *reference point* [2] or is *right-complete*, see [7] for details.
- *Ratio of approximation:* In practice, the ratio of approximation can be expected to be lower than 16, see [5].

Largest common substructures of P and Q

- *Goal:* Determine $\mathrm{LCSC}(P,Q,\mathbf{C}) := \max_{g\in\mathrm{RM}(3)}\mathbf{C}(P,gQ)$, where $\mathbf{C}$ is a function measuring the size of a common substructure of two families of points.
- *Requirements for $\mathbf{C}$:* $\mathbf{C}$ needs to be *relational* in slighlty different sense, see below.
- $\mathrm{LCSC}(P,Q,C)$ can be computed using Algorithm 1 by exchanging

```
d := d(P, gQ);
if (d < D) then D := d; h := g;
```

→

```
c := C(P, gQ);
if (c > C) then C := c; h := g;
```

- *Example:* Fix some $\varepsilon \geq 0$ and choose $\mathbf{C} := C_\varepsilon(P,Q)$ as the longest possible length $|P'| + |Q'|$ of common subcurves P' of P and Q' of Q such that $\mathbf{d}_{\mathrm{F}}(P',Q') \leq \varepsilon$
  ⇝ suitable distance measure for protein backbones;
  ⇝ C_ε is relational in the sense that $C_\varepsilon(P,Q)$ can be determined from $R(P,Q,\varepsilon)$.
- *Quality of approximation:* Let g denote the transformation computed by the above algorithm. Then, we have

$$C_\varepsilon(P,gQ) \geq \max_{h\in\mathrm{RM}(3)} C_{\varepsilon/16}(P,hQ),$$

- The longest common subcurve of P and Q can be seen as a motive shared by the two proteins.

Conclusion and Perspective

- The approach presented here works for minimizing arbitrary relational distance measures as well as maximizing relational target functions.
- Using $\mathbf{d}_{\mathrm{F}}$, the approach is suitable for aligning protein backbones.
- Generalizes to *multiple structure alignments*.

References

[1] T. Akutsu. Protein structure alignment using dynamic programming and iterative improvement. *TIEICE: IEICE Transactions on Communications/Electronics/Information and Systems*, 1996.

[2] H. Alt, O. Aichholzer, and G. Rote. Matching shapes with a reference point. *Internat. J. Comput. Geom. Appl.*, 7:349–363, 1997.

[3] H. Alt, K. Mehlhorn, H. Wagener, and E. Welzl. Congruence, similarity, and symmetries of geometric objects. J. on Discr. Comp. Geom., 1988, pp. 237-256.

[4] S. Chakraborty and S. Biswas. Approximation algorithms for 3-d commom substructure identification in drug and protein molecules. In *Workshop on Algorithms and Data Structures*, pages 253–264, 1999.

[5] M. T. Goodrich, J. S. B. Mitchell, and M. W. Orletsky. Approximate geometric pattern matching under rigid motions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):371–379, 1999.

[6] P. Indyk, R. Motwani, and S. Venkatasubramanian. Geometric matching under noise: Combinatorial bounds and algorithms. In *SODA: ACM-SIAM Symposium on Discrete Algorithms (A Conference on Theoretical and Experimental Analysis of Discrete Algorithms)*, 1999.

[7] A. Mosig. *Efficient Algorithms for Shape and Pattern Matching.* PhD thesis, Institut für Informatik III, Universität Bonn, 2004.

[8] A. Mosig and M. Clausen. Approximately matching polygonal curves with respect to the Fréchet distance. *Accepted for Special Issue of Computational Geometry: Theory and Applications*, 2004.

[9] M. E. Munich and P. Perona. Continuous dynamic time warping for translation-invariant curve alignment with applications to signature verification. In *IEEE International Conference on Computer Vision*, volume 1, pages 108–, Corfu, Greece, 1999.

[10] R. C. Veltkamp and M. Hagedoorn. Shape similarity measures, properties, and constructions. In *Advances in Visual Information Systems*, volume 1929 of *Lecture Notes in Computer Science*, pages 467–476, 2000.

Printed at the TBI Vienna