Agglomerative clustering of fragment 3D structures based on pairwise RMSD

Antoine Moniot, Isaure Chauvot de Beauchêne, Yann Guermeur

Agglomerative clustering of fragment 3D structures based on pairwise RMSD

Antoine Moniot, Isaure Chauvot de Beauchene and Yann Guermeur
University of Lorraine, CNRS, INRIA, LORIA, Nancy, France
antoine.moniot@loria.fr

Introduction:

In structural biology, fragment-based 3D modeling methods make use of fragment libraries. The library associated with one fragment represents the whole set of possible 3D structures (conformations) that it can adopt (with a chosen precision). Given the computational constraints, deriving libraries of minimal cardinality appears as a strong requirement. This amounts to deriving *ε-nets* whose cardinalities are as close as possible to the corresponding *covering numbers*. A heuristic to derive such ε-nets is to cluster the observed conformations under appropriate constraints, and keep only the *representatives/prototypes*. In the framework of interest, the main difficulty encountered is to implement the clustering with the RMSD as dissimilarity measure. Indeed, the computation of this measure follows a superimposition, the nature of which has both biological and mathematical consequences.

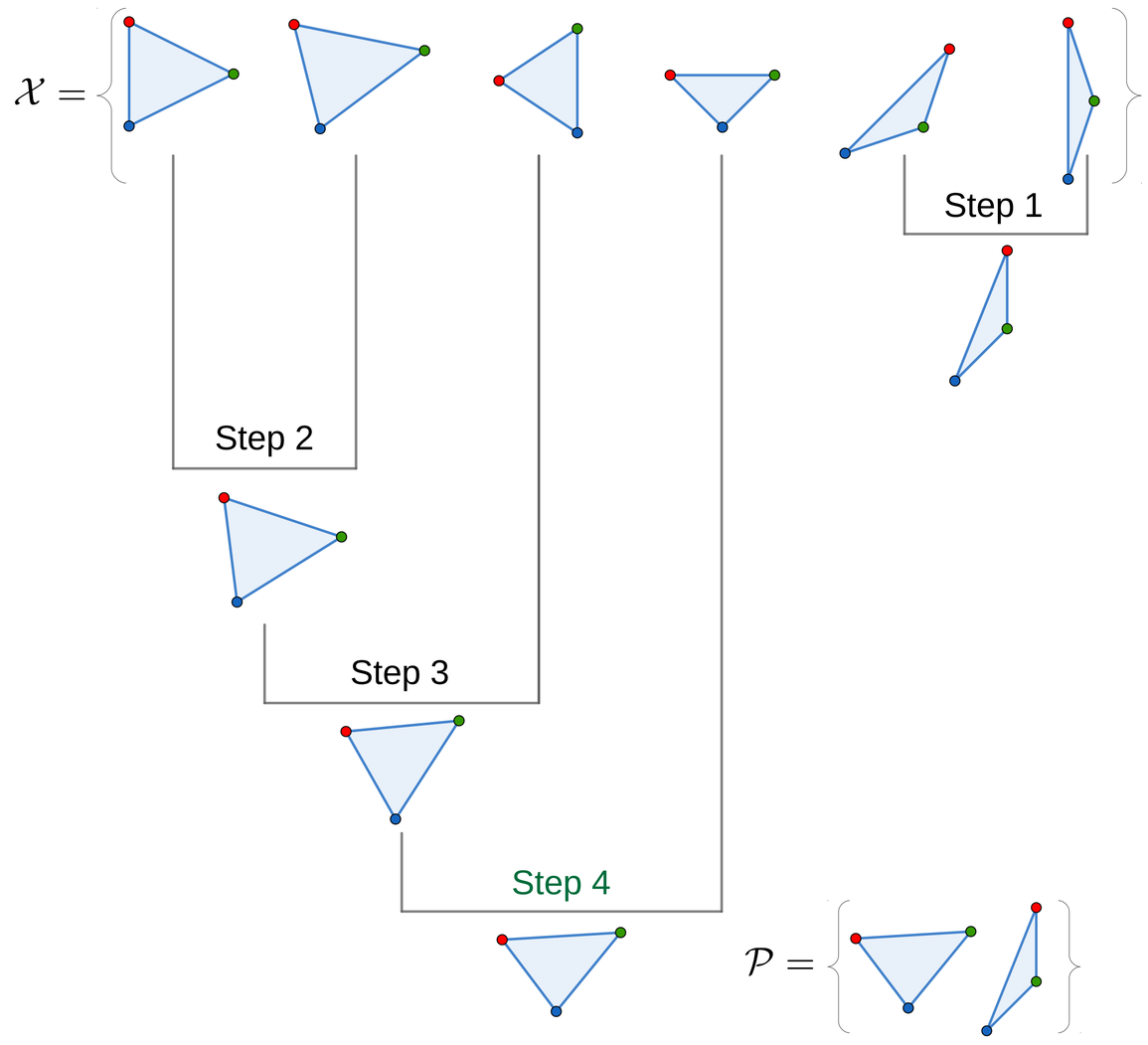
We introduce such a method as a variant of the Hierarchical Agglomerative Clustering (HAC) algorithm. Compared to HAC, it makes it possible to reduce the number of prototypes, while maintaining an acceptable computation time.

Algorithm:

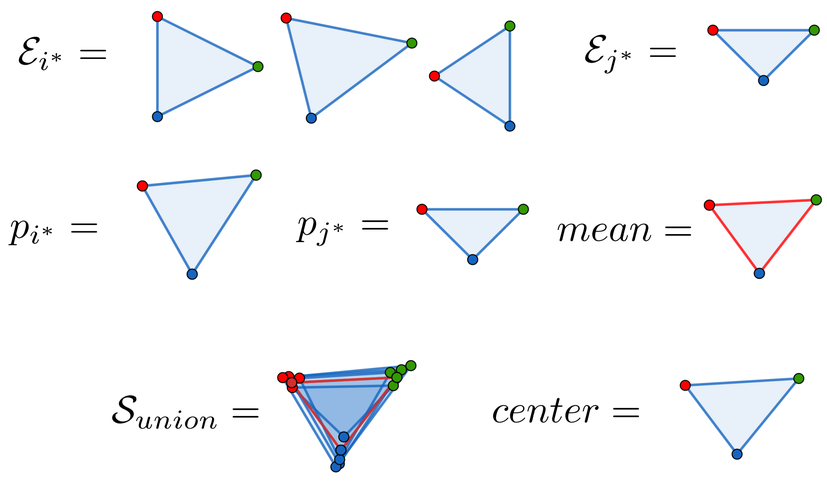
Input : $\mathcal{X} := \{x_i : 1 \leqslant i \leqslant n\}$ # set of conformers
$t > 0$ # threshold value
Output: $\mathcal{P} := \{p_i : 1 \leqslant i \leqslant s\}$ # set of prototypes
Initialisation:
$s := n$ # number of prototypes
$\mathcal{P} := \mathcal{X}$ # set of prototypes
$R := (r_{i,j}) \in \mathbb{R}^{s \times s}$ # matrix of RMSD of prototypes after **one-against-one superimposition**
$R^{mask} := (\mathrm{r}_{i,j}^{mask}) \in \{0,1\}^{s \times s}$
$\mathcal{E} := \{\{x_i\}_{1<j<s} : 1 \leqslant i \leqslant n\}$ # set of clusters

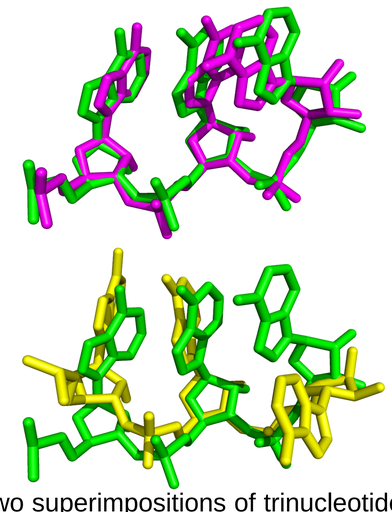
$fusion := True$;
while $fusion$ **do**
  $fusion := False$;
  $R^{mask} := False$; # initialized at $False$
  $tag := False$;
  while *(not tag* ***and*** *not* $fusion$*)* **do**
    $i^*, j^* := \underset{1 \leqslant i < j \leqslant s}{\operatorname{argmin}} (r_{i,j} : r_{i,j}^{mask} == False)$;
    $mean := (p_{i^*} + p_{j^*})/2$;
    $\mathcal{S}_{union} := \mathcal{E}_{i^*} \cup \mathcal{E}_{j^*}$; # **superimposed on** mean
    Compute the *center* of the smallest englobing ball associated with $\mathcal{S}_{union}$;
    if $\left(\max_{x_i \in \mathcal{S}_{union}} (d_{superimposed}(x_i, center)) \leqslant t \right)$ **then**
      $\mathcal{P} \setminus \{p_{i^*}, p_{j^*}\} \cup \{center\}$;
      Update R;
      Update $\mathcal{E}$;
      $s := s - 1$;
      $fusion := True$;
    else
      $r_{i_{min},j_{min}}^{mask} := True$;
      if $(\forall\ 1 \leqslant i < j \leqslant s : r_{i,j}^{mask} == True)$ **then**
        $tag := True$;
      end if
    end if
  end while
end while
return $\mathcal{P}$;

Example:



Application on step 4 of the example





Two superimpositions of trinucleotides, RMSD of 1Å (top) and 3Å (bottom). 1Å corresponds to the fragment libraries giving the best docking results.

Analysis:

The algorithm is based on the HAC [1], the main difference is the linkage method. We defined a linkage method which is the "smallest enclosing balls". The computation of those balls is a quadratic problem and to solve it we are using the Frank-Wolfe algorithm [3].

The algorithm was applied to trinucleotides of RNA. We compared the results obtained, for the sequence AAA, to a classic HAC with the "complete" linkage method. For the "complete" linkage we obtain 2817 clusters, and with our "smallest enclosing balls" we obtain 2421 prototypes (coming from a set of 11,813 conformations). The difference is that the "complete" linkage is not adapted to calculate ε-nets, it is merging clusters when all members are at maximum 1Å from other members, and it does not give any prototypes.

In the fragment-based method for the docking of RNA on protein [2], the docking of the conformers is proportional to the number of prototypes.

Conclusion and ongoing research:

- Our method performs better than the state-of-the-art one;
- On the application of trinucleotides, the cardinality directly translate into a gain in computation time;
- We are planning to apply our method to other biological problems;
- Statistical analysis is in progress to derive generalization error bounds and excess risk bounds.

References:

1. AK. Jain, and RC. Dubes, Algorithms for clustering data. Prentice Hall, Englewood Cliffs, 1998.
2. I. Chauvot de Beauchene, S. de Vries, and M. Zacharias. Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins. NucleicAcidsResearch, 44(10):4565-4580, 2016.
3. M. Frank, and P. Wolfe. An algorithm for quadratic programming. Naval Res. Logist. Quart., 3(1):95-110, 1956.