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 \( \varepsilon \)-nets whose cardinalities are as close as possible to the corresponding covering numbers. A heuristic to derive such \( \varepsilon \)-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:

\[
\begin{align*}
\text{Input} & : \mathcal{X} := \{x_i : 1 \leq i \leq n\} & \text{# set of conformers} \\
& t > 0 & \text{# threshold value} \\
\text{Output} & : \mathcal{P} := \{p_i : 1 \leq i \leq s\} & \text{# set of prototypes} \\
\end{align*}
\]

Initialization:
\[
\begin{align*}
\mathcal{P} & := \mathcal{X} & \text{# number of prototypes} \\
R & := (r_{i,j}) \in \mathbb{R}^{s \times s} & \text{# set of prototypes} \\
R_{\max} & := \{(r_{i,j}) \in \{0, 1\}^{s \times s} \text{ one-against-one superimposition}\} \\
\mathcal{E} & := \{x_i : 1 \leq i \leq n\} & \text{# set of clusters} \\
\text{fusion} & := \text{True}; \\
\end{align*}
\]

while fusion do
\[
\begin{align*}
\text{R}_{\text{mask}} & := \text{False}; & \text{# initialized at False} \\
\text{tag} & := \text{False}; \\
\end{align*}
\]

while (not tag and not fusion) do
\[
\begin{align*}
t^* & := \text{argmin}_t \{r_{i,j} : \text{R}_{\text{mask}} \Rightarrow \text{False}\}; \\
& \text{mean} := (p_{t^*} + p_j)/2; \\
& S_{\text{union}} := \mathcal{E}_{t^*} \cup \mathcal{E}_j; \\
& \text{Compute the center of the smallest enclosing ball associated with } S_{\text{union}}; \\
& \text{if (} \max_{x \in S_{\text{union}}} d_{\text{superimposed}}(x, \text{center}) \leq t) \text{ then} \\
& \text{P} \setminus (p_{t^*}, p_j) \cup \{\text{center}\}; \\
& \text{Update } R; \\
& \text{Update } \mathcal{E}; \\
& \text{tag} := \text{False}; \\
& \text{fusion} := \text{True}; \\
& \end{align*}
\]

else
\[
\begin{align*}
& \text{R}_{\text{mask}} := \text{True}; \\
& \text{if (} \forall 1 \leq i < j \leq s : r_{i,j} := \text{True}) \text{ then} \\
& \text{tag} := \text{True}; \\
& \text{end if} \\
& \text{end if} \\
& \end{align*}
\]

end while

return \( \mathcal{P} \).

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 \( \varepsilon \)-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.

Example:

Step 1

Step 2

Step 3

Step 4

Application on step 4 of the example

Conclusion and ongoing research:
● Our method performs better than the state-of-the-art one;
● On the application of trinucleotides, the cardinality directly translates 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: