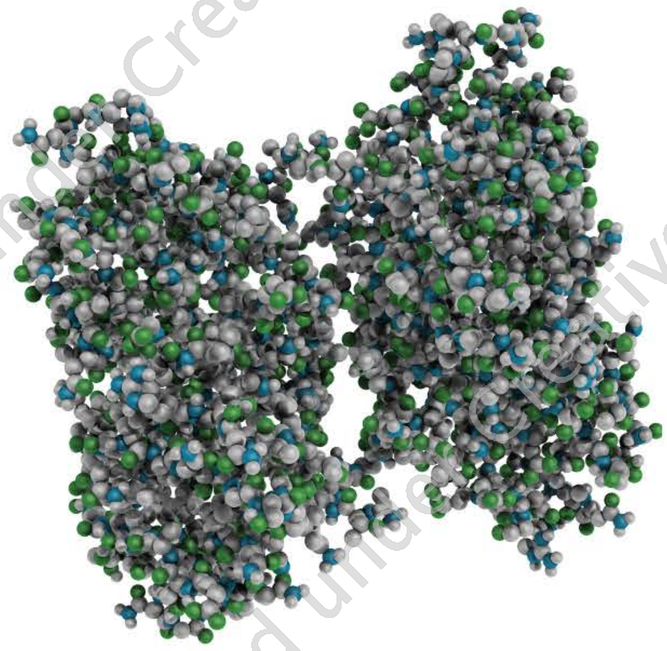
Universiteit Utrecht

Clustering biomolecular structures by residue contact similarity

João Rodrigues, Christophe Schmitz, Mikaël Trellet, Adrien Melquiond, Alexandre Bonvin
Faculty of Science | Chemistry | Bijvoet Center for Biomolecular Research
Contact details: j.rodrigues@uu.nl



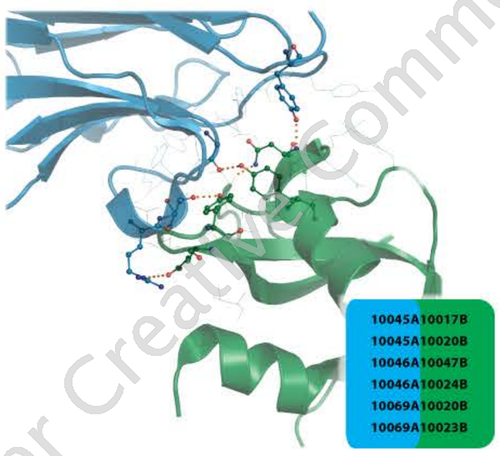
Motivation

Clustering has long been regarded as a helpful tool to retrieve near-native solutions from computational modelling efforts [1]. The majority of clustering methods use positional RMSD to assess similarity between structures at the atomic level, which due to the necessity of computing an alignment/fitting, is both computationally expensive and time consuming. More importantly, an alignment biases which regions of the structures are compared, and by proxy, also the RMSD calculation. We theorize that a naïve calculation of the fraction of common contacts (FCC) between two structures ought yield sufficient discriminatory power while eliminating the need, and consequent bias and performance bottleneck of the alignment/fitting step. In a benchmark of 6 biomolecular complexes of several assembly orders, our FCC clustering method discriminates structures more efficiently, regardless of size or molecule type, and does so dramatically faster than a standard RMSD-based algorithm.

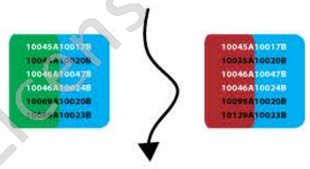
Conclusions

- FCC clustering works on any three-dimensional structure, regardless of molecular types (e.g. DNA/RNA/Protein) or structural complexity (dimers, trimers, etc).
- FCC algorithm eliminates the bias of fitting and alignments of RMSD-based methods,, with the consequence a ~100-fold decrease in computation time.
- FCC-generated clusters are more compact around their center (i.e. lower entropy) due to better discrimination of fringe structures.
- The FCC algorithm significantly outperforms RMSD-based clustering for multimeric assemblies, both in quality of the clusters and in computational resources.
- FCC clustering deals efficiently and effortlessly with symmetrical quarternary structures, bypassing the iterative RMSD calculations needed in current methods.

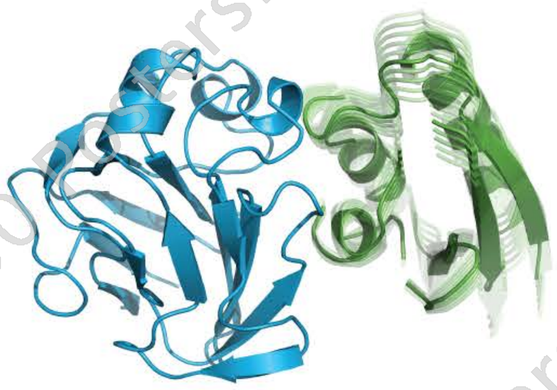
FCC Algorithm



$$FCC(S_a S_b) = \frac{|S_a \cap S_b|}{S_a} = [0,1]$$



$$\begin{bmatrix} - & 0.1 & 0.5 & 0.8 \\ 0.3 & - & 0.2 & 0 \\ 0.5 & 0.4 & - & 0.9 \\ 0.7 & 0.2 & 0.1 & - \end{bmatrix}$$



List residues closer than a distance threshold (5Å) that belong to different chains in each structure.

Pair-wise comparison of residue lists using the Fraction of Common Contacts (FCC) and build a Similarity Matrix.

Cluster structures on the similarity matrix by a pre-defined clustering cutoff (~0.8).

Results: Comparison with RMSD-based Clustering

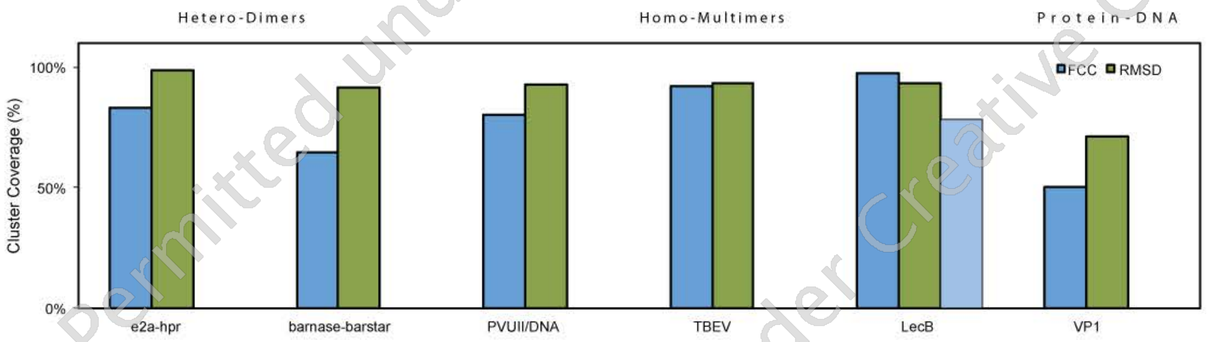


Figure 1. Comparison of the total number of structures included in clusters between the standard protocol (RMSD, green columns) and the FCC-based algorithm (blue columns). LecB has a particular symmetrical arrangement that is only distinguishable at higher thresholds (FCC=0.9, light blue column). The chosen threshold of 0.75 for FCC is generally reasonable, producing slightly smaller clusters than those of the RMSD algorithm.

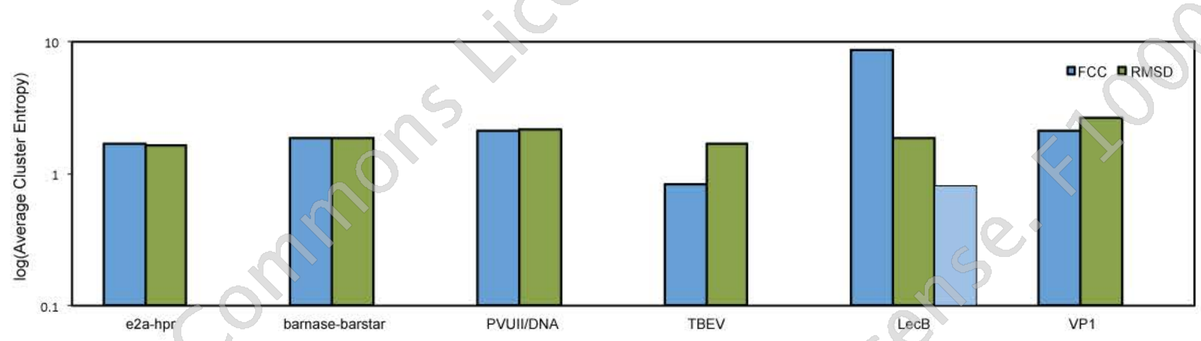


Figure 2. Comparison of the average cluster entropy between the standard protocol (RMSD) and the FCC-based algorithm. The entropy is a measure of how similar are the structures within a cluster. Good clusters should have a low entropy. FCC shows similar entropies to RMSD, except in the case of symmetrical multimers where it is clearly outperforming.

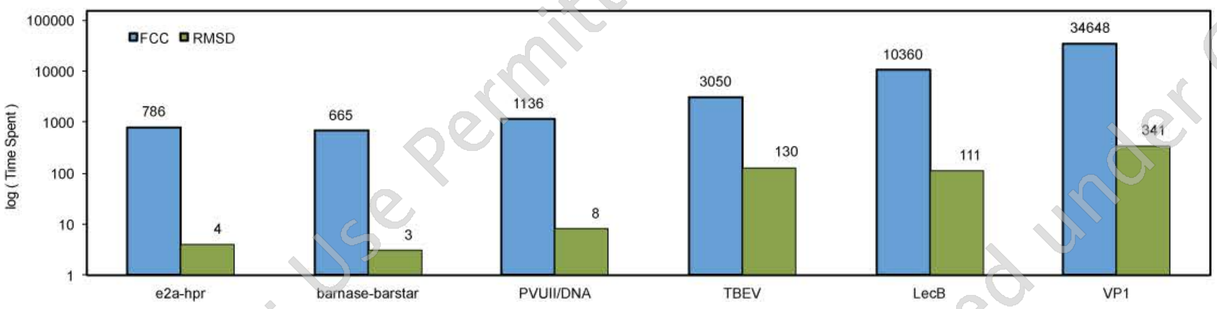


Figure 3. Comparison of the time spent building the pairwise similarity matrix - the longest step in clustering - for both RMSD and FCC algorithms. A logarithmic scale was used to ease the visualization of the difference in performance between both methods, while the real values in seconds are shown over each column. The FCC algorithm is generally much faster than RMSD, even when dealing with symmetrical complexes (slower chain agnostic FCC algorithm), making it very suitable for large scale or large systems clustering.

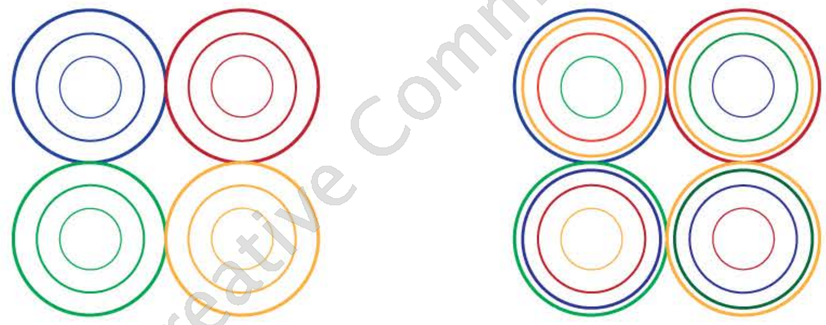


Figure 4. Schematic representation of clusters of the tetrameric LecB complex produced by the regular FCC algorithm (left) and its chain-agnostic version (right). The regular FCC algorithm, much like RMSD clustering, clusters together structures taking into account chain identifiers. Therefore, structures that are structurally similar but have different chain arrangements are placed in different clusters. Our chain-agnostic version of the FCC algorithm ignores the chain identifiers and thus correctly places structurally similar structures in the same cluster, regardless of their chain arrangement. This version is somewhat slower than the regular FCC version, but still orders of magnitude faster than RMSD (see Fig. 3).

References

Kozakov et al. **Optimal clustering for detecting near-native conformations in protein docking.** Biophysical Journal (2005) vol. 89 (2) pp. 867-875

Prinzie and Van den Poel. **Incorporating sequential information into traditional classification models by using an element/position-sensitive SAM**. Decision Support Systems (2006) vol. 42 (2) pp. 508-526