PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Clustering biomolecular complexes by residue contacts similarity

João P. G. L. M. Rodrigues, Mikaël Trellet, Christophe Schmitz, Panagiotis Kastritis, Ezgi Karaca, Adrien S. J. Melquiond, and Alexandre M. J. J. Bonvin*

Bijvoet Center for Biomolecular Research, Faculty of Science, Utrecht University, 3584 CH Utrecht, The Netherlands

ABSTRACT

Inaccuracies in computational molecular modeling methods are often counterweighed by brute-force generation of a plethora of putative solutions. These are then typically sieved via structural clustering based on similarity measures such as the root mean square deviation (RMSD) of atomic positions. Albeit widely used, these measures suffer from several theoretical and technical limitations (e.g., choice of regions for fitting) that impair their application in multicomponent systems ($N > 2$), large-scale studies (e.g., interactomes), and other time-critical scenarios. We present here a simple similarity measure for structural clustering based on atomic contacts—the fraction of common contacts—and compare it with the most used similarity measure of the protein docking community—interface backbone RMSD. We show that this method produces very compact clusters in remarkably short time when applied to a collection of binary and multicomponent protein–protein and protein–DNA complexes. Furthermore, it allows easy clustering of similar conformations of multicomponent symmetrical assemblies in which chain permutations can occur. Simple contact-based metrics should be applicable to other structural biology clustering problems, in particular for time-critical or large-scale endeavors.

Proteins 2012; 80:1810–1817.

Key words: biomolecular clustering; protein docking; fraction of common contacts.

INTRODUCTION

The road to complete comprehension of a biological process inevitably passes through the knowledge of the detailed atomic structures of its participants.[1] Unfortunately, experimental determination of biomolecular structures is often problematic and time consuming, whereas *in silico* molecular modeling methods devised as complementary approaches suffer from chronic inaccuracy because of the simplified physics they are based on.[2] Nevertheless, the relative ease and speed with which the latter yield near-atomic resolution models earned them a spot in the limelight of structural biology methods.

To counterweigh their innate inaccuracy, molecular modeling methods often generate thousands to tens of thousands of possible conformations for a single structure, each representing a discrete point in its energy landscape. A posterior selection process is then necessary to salvage the most native-like conformations. Previous research has shown that, because a native structure is very unlikely to be an isolated event in the energy landscape, it is expected to neighbor similar near-native conformations in a basin with overall low potential energy.[3] This observation hinted at the adoption of clustering techniques, devised to group elements sharing common attributes, to the benefit of the selection process. In fact, it has been shown in both protein structure prediction[3] and protein–protein docking[4] that clustering indeed helps discriminate near-native structures better than energetics alone. Predictably, the most successful algorithms at both CASP (Critical Assessment of Techniques for Protein Structure Prediction)[5] and CAPRI (Critical Assessment of Prediction of Interactions)[6] experiments have incorporated at least one clustering step in their protocols.

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Netherlands Organization for Scientific Research (VICI Grant); Grant number: 700.56.442; Grant sponsor: Focus and Massa Programme of Utrecht University.

*Correspondence to: Alexandre M. J. J. Bonvin, Bijvoet Center for Biomolecular Research, Faculty of Science, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands. E-mail: a.m.j.j.bonvin@uu.nl.

Received 12 December 2011; Revised 14 March 2012; Accepted 30 March 2012
Published online 6 April 2012 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24078

The performance of clustering algorithms is nevertheless dependent on the similarity measure used to determine the similarity between any two elements of a dataset, which for the majority of the state-of-the-art clustering algorithms is the root mean square deviation (RMSD) of atomic coordinates. Yet, previous research has shown that, despite widely adopted, RMSD suffers from several shortcomings. First, it loses sensitivity as the molecular weight of the system increases, because large regions with little deviations become dominant.[7] Second, and more importantly, the necessity of choosing the regions to fit the structures under scrutiny to one another results in biased measurements. Finally, RMSD calculations are CPU intensive and consume a large amount of live memory (RAM), yet another hindrance to the structural comparison of increasingly larger and more complex systems.

This conjecture motivated several studies comparing and assessing similarity measures.[8,9] Metrics such as dihedral angles and variants of RMSD such as distance matrix RMSD have been used to cluster molecular dynamics trajectories. It has also been shown that a metric based on residue contacts—contact matrix distance—accounted for less chaotic clusters. Furthermore, contact-based measures (lDDT[10] and FNAT[11]) are already being used to assess the quality of submitted models in CASP and CAPRI, respectively.

The protein docking community is shifting its focus to more intricate systems such as entire interactomes or supramolecular assemblies consisting of a large number of components.[12] As a result, similarity measures that retain a high sensitivity while performing substantially faster than traditional RMSD-based metrics are required. Inspired by the widespread usage of contact information in structure comparison, we theorized that calculating the fraction of common contacts (FCCs) between two structures, akin to the notion of fraction of native contacts used in CAPRI, would first describe the relative orientation of the interacting partners and second provide detailed residue-level information. Such a measure, if applied to structural clustering, should yield sufficient discriminatory power without suffering from any of the theoretical downsides of positional RMSD measures and save a considerable amount of computation time as it discards the structural alignment step. In the following, we introduce the concept of FCC clustering and demonstrate its performance in a set of binary and multimeric complexes, selected not only to reflect typical scenarios in protein docking but also challenging cases including assemblies with internal symmetry and protein–DNA complexes.

MATERIALS AND METHODS

Identifying residue contacts

For each nonhydrogen atom pair (i, j) in a structure, the Euclidean distance between the atoms (r_{ij}) is computed. If this distance is below a threshold r_c and both atoms belong to different polypeptide chains, the pair of residues to which the atoms belong to is considered to be in contact. We defined r_c as 5 Å in accordance with CAPRI criteria, the standard in the docking field.

Calculating the FCCs

We define our similarity measure, FCC$_{AB}$, as the FCCs between structures A and B with respect to the total number of contacts in A:

$$\mathrm{FCC}_{AB} = \frac{|A \cap B|}{|A|} \in [0, 1] \qquad (1)$$

The outcome is a value ranging from zero, when the structures share no contacts, to a maximum of one when all contacts of structure A are present in structure B. The normalization of the number of common contacts over the number of contacts of the first structure brings asymmetry to the similarity measure and consequently to the similarity matrix as FCC$_{AB}$ might not be equal to FCC$_{BA}$. In principle, the matrix could be symmetrized before clustering. However, a comparison of the clustering coverage and entropy of the obtained clusters using the two different matrices revealed that, for the majority of the cases, the symmetric matrix produces larger clusters but also with a larger entropy (Supporting Information Fig. S4). In addition, the averaging of both FCC values reduces the resolution of the matrix, making it harder to optimize the clustering threshold. In light of these observations, the asymmetric matrix approach was chosen for all subsequent work.

In the case of symmetrical complexes, the chain identifier is omitted from the contact string identifier for the FCC calculation. This "chain-agnostic" variant of the FCC computation allows efficient clustering of structures that share the same interface regardless of the permutation of their chains along the symmetry axis.

Clustering algorithm

We adapted a version of the disjoint Taylor–Butina clustering algorithm developed to use asymmetric matrices[13] that can be described in four steps:

1. Create a nearest-neighbor table from the full similarity matrix using a predefined threshold for the FCCs. Structure A is a neighbor of B only if FCC$_{BA}$ is above the threshold, and vice versa.
2. Detect true singletons (structures with an empty nearest-neighbor list, i.e., no neighbors at this threshold) and remove them from the dataset.
3. Find the structure with the largest nearest-neighbor list and define it as the center of the first cluster. Exclude this structure and all its neighbors from the dataset and update all nearest-neighbor lists. This

update step is crucial to have disjointed clusters, or in other words, to ensure that structures belong to one and only one cluster.

4. Repeat step 3 until no structures are left in the dataset or the remaining have nearest-neighbor lists shorter than a predefined minimum cluster size threshold (default 4).

The original algorithm by Prinzie and Van der Poel[13] comprised the inclusion of the remaining structures—false singletons—in the cluster with the largest number of structures neighboring them. Preliminary analysis revealed that it contributed to an increase in the structural variability within the clusters, while being not at all justified with a significant increase of cluster population. Consequently, we forfeited this step in our implementation.

Implementation of the FCC-based clustering algorithm

Our clustering algorithm was implemented in the Python programming language and is freely available upon request. All the calculations were performed on a standard desktop computer with 2.66-GHz CPU and 4-GB RAM.

Measures for cluster quality assessment

The quality of the cluster i can be assessed by the conformational variability of its N members. In the case of complexes, it is defined as the mean interface positional RMSD (i-RMSD) of all members from the center of the cluster, clus.ctr:

$$\text{Cluster}_i\,\text{Entropy} = \frac{1}{N}\sum_{s=1}^{n_i} \text{i} - \text{RMSD}(s)_{\text{clus.ctr}} \qquad (2)$$

This measure was further expressed to account for the entropy of a given clustering run consisting of M clusters as the population-weighted average of the individual cluster entropies:

$$\text{Average Cluster Entropy} = \frac{\sum_{i=1}^{M} S(i)\times\text{Cluster}_i\,\text{Entropy}}{\sum_{i=1}^{M} S(i)} \qquad (3)$$

where $S(i)$ represents the number of elements in cluster i. Because of the internal symmetry of some cases of our structure set, calculating their cluster and average cluster entropies required an iterative i-RMSD calculation where all chain combinations were tried. The lowest i-RMSD value was then chosen. This ensured that, despite chain permutations, the entropies truly reflected the conformational variability within the clusters, while free from symmetry-induced artifacts.

Definition of i-RMSD and iL-RMSD

i-RMSD

The interface RMSD is defined following CAPRI standards as the positional RMSD of all interface residues (calculated on the Cα, N, C, and O atoms) that have a heavy atom within 10 Å of any other interacting partner.

L-RMSD

The ligand RMSD is also defined following CAPRI standards. The models are first fit onto the larger chain (receptor) and then the RMSD (on Cα, N, C, and O atoms) is calculated on the smaller chain (ligand).

iL-RMSD

The interface–ligand RMSD (iL-RMSD) used in HADDOCK for clustering purposes[14] is a slight variation of i-RMSD, in which the models are first fit on the interface of the first molecule and the RMSD is then calculated on the interface residues of all other molecules. Interface residues are automatically defined based on all contacts observed over all generated docking solutions. For speed purposes only CA atoms are considered. Depending on the conformation sampling of the docking models, this measure will be somewhere in between i- (sampling only close to the true interface) and L-RMSD (sampling of the entire surface of all molecules).

RESULTS

Evaluating FCC as a similarity descriptor

To evaluate the performance of FCC clustering, we analyzed docking models obtained with HADDOCK[14] for a set of six complexes consisting of two to five components and with various internal symmetries (see Tables I and II) whose structures were experimentally determined. We calculated both the FCCs with the reference native structure (FCC_{NAT}) and the i-RMSD ($\text{i-RMSD}_{\text{NAT}}$) from the native structure (see Materials and Methods section). Additionally, we also calculated the ligand RMSD ($\text{L-RMSD}_{\text{NAT}}$) from the native structure.

Unsurprisingly, for all complexes, near-native models (low $\text{i-RMSD}_{\text{NAT}}$) share a substantial number of contacts (high FCC_{NAT}) with the native structure, whereas those more dissimilar share progressively fewer or none at all (Fig. 1). The same, albeit less obvious for some structures, is observed for $\text{L-RMSD}_{\text{NAT}}$ (Supporting Information Fig. S1). The anomaly observed for the LecB protein, a dimer of dimers, is due to its particular symmetry type (D2), in which the larger intradimer interface accounts for the majority of contacts, causing solutions mirrored across the interdimeric axis to have high FCC values (Supporting Information Fig. S2). Nevertheless,

Table I
Biomolecular Complexes Used to Assess FCC Clustering Performance

Complex	PDB ID	# Components	Type	# Models	Symmetry type
E2A/HPR	1GGR[15]	2	Protein/protein	200	None
Barnase-Barstar	1BRS[16]	2	Protein/protein	200	None
TBEV	1SVB[17]	3	Protein/protein	400	C3
LecB	1OUS[18]	4	Protein/protein	400	D2
VP1	1VPN[19]	5	Protein/protein	400	C5
PVUII/DNA	1EYU[20]	2[a]	Protein/DNA	200	None

The models were taken from previously published datasets. The references of the experimental structure determination protocols are shown in parenthesis after the PDB ID.

[a] The focus in this complex was on the protein–DNA interface. Accordingly, the protein–protein interface was not considered for clustering purposes.

the most native-like structures have a distinctly higher FCC value. These observations indicate that FCC is a good similarity descriptor, suitable for clustering of biomolecular interfaces, regardless of their molecular components and their quaternary arrangement.

Choice of optimal threshold for FCC-based clustering

The clustering threshold defines the rigor with which the clustering algorithm considers two structures similar enough to belong together in the same cluster. Although previous works[4] have attempted to derive an optimal threshold for protein–protein docking using the distribution of values in the similarity matrix, pursuing a similar approach for FCC proved unreasonable. The distribution of the values in the similarity matrix depends on the conformational variability of the generated models. For complexes whose models are widespread over the conformational landscape, such as in the majority of our structure set (Table II), the distribution of similarity matrix values resembles a negative exponential function [Supporting Information Fig. S3(B–E)] and is therefore unsuitable for extracting an optimal clustering threshold as per Ref. 4. Interestingly, the chain-agnostic variant of the algorithm affects the distribution of the matrices of multimeric complexes, producing a shift toward higher FCC values [Supporting Information Fig. S3(C–E)]. In light of these observations, we opted to evaluate several clustering runs at different thresholds (starting at 0.5,

Table II
Structural Characteristics of the Model Set and Clustering Statistics for Both RMSD and FCC Clustering Methods

Complex	Mean i-RMSD of the ensemble (Å)	Similarity measure	Number of clusters	Clustered structures (%)	Average cluster entropy (a.u.)
E2A-HPR	2.4 ± 1.8	RMSD	3	99	1.6
		FCC	7	83	1.7
Barnase-Barstar	8.9 ± 4.0	RMSD	8	92	1.9
		FCC	7	65	1.8
TBEV	27.2 ± 11	RMSD	22	94	1.7
		FCC	17	92	0.9
LecB	8.9 ± 7.7	RMSD	25	93	1.9
		FCC	18	79	0.8
VP1	15.0 ± 5.3	RMSD	32	72	2.7
		FCC	32	50	2.1
PVUII/DNA	5.57 ± 1.3	RMSD	8	93	2.2
		FCC	11	80	2.1

The mean interface RMSD of the ensemble informs on the variability of the conformations present in the model set. The percentage of clustered structures and average cluster entropy refer to two measures we defined to assess the clustering algorithms. FCC clustering is shown to have consistently lower entropy, at a cost of less structures clustered, and performing particularly well for multimeric assemblies (TBEV, LecB, and VP1).

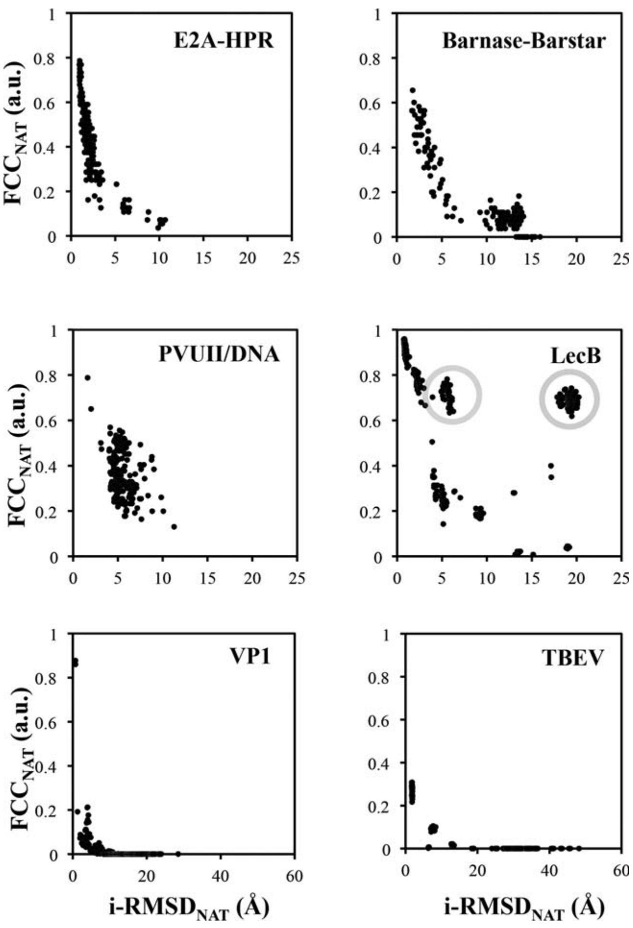

Figure 1
Assessing the FCCs (FCC_{NAT}) as a similarity descriptor by comparison with the i-RMSD ($i\text{-}RMSD_{NAT}$). Both FCC and i-RMSD are calculated with respect to the experimentally determined structure of each complex. Low $i\text{-}RMSD_{NAT}$ values correspond to high FCC_{NAT} values, which supports the hypothesis that FCC is a good similarity descriptor and, hence, a good similarity measure for structural clustering. In the case of LecB, symmetrical solutions that share only the larger of the two dimeric interfaces with the native structure have high FCC values (highlighted by gray circles) (see also Supporting Information Fig. S2 and the main text for explanation).

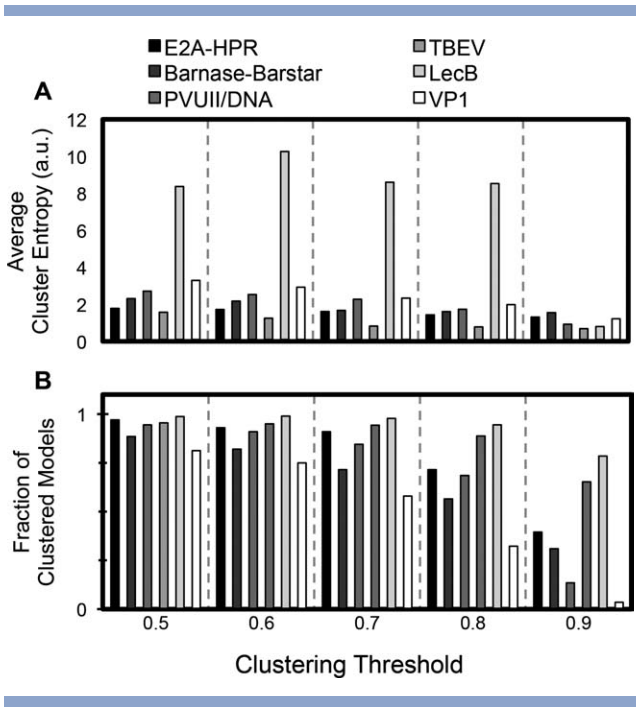

Figure 2

Definition of an optimal clustering threshold from an analysis of different runs at different clustering thresholds. A value of 0.75 was selected based on the observation that the entropy of the clusters declines with increasing values of threshold, whereas the number of structures included in the clusters only drops sharply at 0.9. This threshold is appropriate for good clustering in all cases but LecB, which requires a higher value (0.9) because of its particular symmetric arrangement.

with increasing steps of 0.1) monitoring the conformational entropy of the resulting clusters and the total percentage of models included in clusters (Fig. 2).

As expected, raising the clustering threshold enhances structure discrimination. This increasingly isolates structures and consequently reduces the size of the resulting clusters. Eventually, these clusters fail to meet the minimum size requirement (four members) and their members are considered isolated events in the conformational landscape [Fig. 2(B)]. This effect is particularly evident at very highly discriminative thresholds (0.9) where the fraction of clustered structures drops below 0.5 for most cases. The average cluster entropy depends on the quality of the docking prediction and on the dispersion of the models over the conformational landscape of the molecule. Well-defined model sets such as E2A-HPR produce clusters through the FCC algorithm with an entropy comparable to those of RMSD clustering at thresholds as low as 0.5 (50% of the interface contacts in common) [Fig. 2(A)]. Stricter discrimination has little effect on the structural variability in each cluster, as seen by the slow decrease in average cluster entropy of E2A-HPR (1.78–1.32). On the other hand, clustering more chaotically distributed model sets (e.g., VP1) clearly benefits from higher thresholds, because the entropy of the resulting clusters steadily drops (from 3.30 to 1.22) as the threshold increases. Notably, LecB deviates from the rest of the complexes because of its particular symmetrical arrangement (Supporting Information Fig. S2). Up to a threshold of 0.8, most clusters include several mirror-like symmetrical conformations and have consequently very high entropy values (>10). Increasing the threshold to 0.9 allows the discrimination of both interdimer and intradimer interfaces, splitting the very large cluster obtained at 0.8 (Cluster #1, entropy 10.55, $N = 297$) into smaller but extremely compact subclusters. This brings the average cluster entropy sharply down (0.91) while retaining the large majority of the structures (78.5%; Fig. 2). These observations suggest that a threshold between 0.7 and 0.8—empirically, 0.75—is the most suitable for generic application of FCC clustering. However, this might require adaptation in particular cases, such as LecB.

Quantitative assessment of FCC clustering

To cement the quality of FCC as a valid similarity measure for structural clustering, we performed a direct comparison with the protocol integrated in HADDOCK, which uses iL-RMSD of atomic coordinates (see Materials and Methods section) and the clustering algorithm implemented by Daura *et al.*[21] with a default clustering threshold of 7.5 Å (Fig. 3). iL-RMSD clustering at this threshold collects a larger number of structures at an expected cost of higher entropy clusters [Fig. 3(A,B)]. Although for the heterodimers and PVUII/DNA this is acceptable, analysis of clustering of symmetric multicomponent complexes reveals an important limitation of iL-RMSD, and by extension all positional RMSD-based metrics, as clustering similarity measures: recognizing similar conformations with different symmetrical chain arrangements is not trivially possible and results in several clusters that should, in truth, be merged. This happens because these methods are bound to the chain identifiers of the PDB file format, which in turn results in high RMSD values for structures that share very similar structural features but whose chain identifiers are swapped, placing them in separate clusters. Detailed analysis of the centers of iL-RMSD-generated clusters corroborates this hypothesis, showing little conformation differences between several models, indicating that these should belong in the same cluster. By contrast, the chain-agnostic variant of the FCC clustering algorithm agglomerates the several chain permutations (i.e., for a three-chain complex: ABC and ACB) in one single and larger cluster ($N = 87$). Because these structures are nevertheless very similar, the entropy of the clusters remains extremely low [Fig. 3(B)].

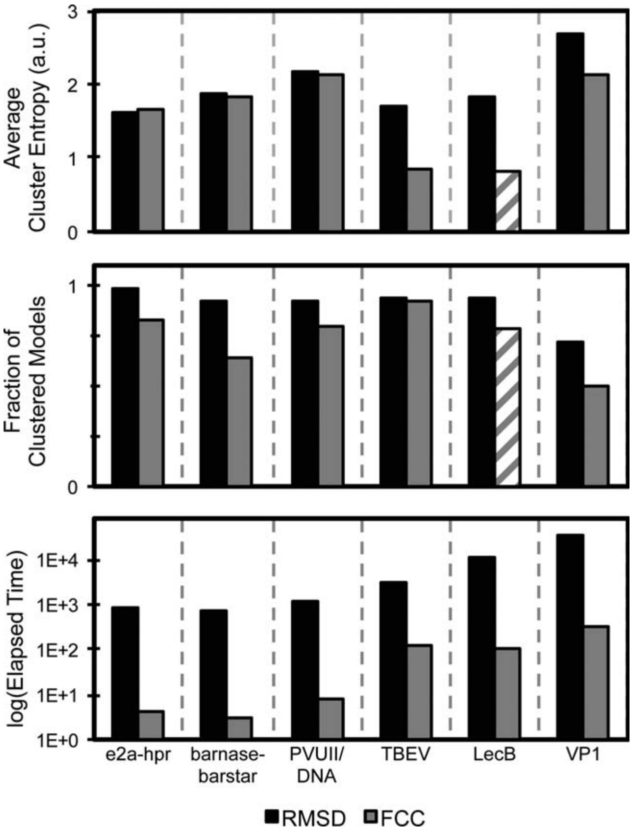

Figure 3

Average cluster entropy, cluster coverage, and computational performance for both the FCC clustering (gray bars) and the iL-RMSD clustering (black bars). Clusters were generated using the default threshold for iL-RMSD of 7.5 Å in all cases and 0.75 for FCC, except LecB, which was clustered at 0.9 (striped bar). FCC clustering leads to smaller but more compact clusters, which indicates a better discrimination of fringe structures. For multimeric structures with internal symmetry, in particular TBEV and VP1, the advantage of clustering based on FCC is evident. Performance-wise, avoiding structural fitting reduces the computation time required for FCC clustering by a factor of 100 on an average.

Finally, because structural alignment and fitting is absent in FCC clustering, computational efficiency is greatly enhanced [Fig. 3(C)]: iL-RMSD clustering takes several minutes to several hours to build similarity matrices for the complexes, depending on the interface size. Using FCC as a similarity measure reduces this computation time by a factor of, on an average, 100, smoothing the path for structural clustering of intricate multicomponent systems such as those described before.

DISCUSSION

Residue contacts are enough to differentiate binding poses

We have developed a new clustering approach for macromolecular complexes based on the premise that residue contacts alone are enough to discriminate binding poses between interacting partners. Although already used by the docking community to assess the accuracy of the docking results, the application of the FCCs in structural clustering of docking solutions is novel and shows good results. Direct comparison with the commonly used i-RMSD reveals that FCC is a good descriptor of structural similarity (Fig. 1). We have shown that a high value of FCC unequivocally corresponds to low i-RMSD and L-RMSD values, and therefore similar structures, independently of the number of components in the complex or its symmetry type.

FCC clustering can accommodate various levels of biomolecular complexity

We have shown that FCC clustering deals effortlessly with large assemblies, greatly reducing the computation time while generating clusters of similar quality with the current state-of-the-art methods (Fig. 3). This leap in performance is due to the avoidance of pair-wise structural alignments, which has the added value of removing the bias stemming from the choice of regions on which to perform the alignment. Another problem tied to structural alignment lies in the handling of symmetrical solutions. Although structural biologists artificially name molecular chains to distinguish them from one another, for structural comparison and by proxy, structural clustering purposes, the chain arrangement does not matter as long as the molecular architecture is similar. Because RMSD calculations are bound to the chain identifiers, clustering based on such measures often produces very similar clusters whose structures differ only in the symmetrical arrangement of their chains. This is evident in all the cases with internal symmetry presented above (TBEV, LecB, and VP1) and poses a problem for postclustering analysis. Avoiding this problem in RMSD-based methods requires an iterative calculation of all the several permutations of the chain arrangements (e.g., ABC and ACB), which further aggravates computational performance. FCC clustering sidesteps all these issues by considering each complex a whole entity free from chain identifiers—the chain-agnostic variant. Although simplistic, this solution successfully merges the several clusters that share the same conformation, which not only accounts for larger clusters but also facilitates posterior analysis, namely in determining the lowest energy cluster, likely to contain the best representative structure. This advantage will be crucial in case where only few similar conformations are present in the model set. Furthermore, as the calculation of the FCCs, as per the current algorithm, reads only the residue index within the structure, FCC has a wide range of applications regarding different molecular representation scales (coarse-grained to all-atom). It also allows for clustering of point mutants of the same structure, or even gapped models, given that the numbering is preserved and consistent across all models.

Ranking of clusters is largely independent of the clustering method

The discriminative power of FCC clustering for the chosen general threshold of 0.75 (75% of the interface in common) is superior to that of iL-RMSD clustering, reducing the entropy of the clusters, but also the size of the clusters. Analysis of which structures are effectively discarded through FCC clustering showed that these are largely fringe structures, the furthest away from the cluster center, and that in most cases do not impact the overall quality of the clusters when compared with the native structure. An analysis on the average i-RMSD$_{NAT}$ of all clusters generated with both FCC and iL-RMSD algorithms for an extended dataset composed of 20 real-case scenarios (previous CAPRI experiment targets) showed that the ranking of the clusters is largely unaffected by the clustering method (Supporting Information Table S1). Comparison of the top ranking clusters reveals in a majority of cases a good agreement between both clustering algorithms and for a number of cases, for similar ranking performance, the resulting clusters show an increased accuracy as measured by i-RMSD$_{NAT}$. Therefore, this corroborates that FCC clustering is not discarding important native-like structures and is therefore suitable for large-scale application.

FCC clustering accommodates current and future needs in biomolecular docking

Both the increased computational efficiency and the overall performance of our FCC clustering algorithm are encouraging. Efficient methods that allow for rapid RMSD calculation of protein complexes exist but are, however, mostly based on simple rigid body transformations (i.e., rotations and translations over the center of mass of the complex) and thus do not account for internal flexibility of the system. Because the models were generated with HADDOCK, which includes a semiflexible refinement step, rigid body-based clustering algorithms are inappropriate. In contrast, FCC clustering worked effectively on these models, meaning that the method is suited for flexible docking approaches, without degrading performance. Furthermore, considering the shift toward the modeling of entire interactomes or very large systems (e.g., nuclear pore[1]) to fill in the gaps left by low-resolution or high-throughput experimental techniques, fast and accurate clustering methods will be critical in the near future. We have demonstrated here that our FCC algorithm is well suited for this task as it performs well in diverse environments, from traditional protein–protein complexes to more complicated multicomponent assemblies and heterogeneous biomolecular systems like protein–DNA complexes, while being computationally efficient.

CONCLUSION

The current perspectives for the field of biomolecular docking call for methods able to deal with large datasets, both in number of molecules and molecular size. RMSD-based clustering methods are computationally expensive and their sensitivity decreases with the molecular size of the system. Yet, suggested alternatives so far, although useful in particular scenarios, fail at reproducing both their quality and performance when applied generically. Although the concept of contact-based molecular comparison is known and used in both CASP and CAPRI, it is limited to the assessment of results. The inclusion of FCC clustering in docking algorithms, as shown here with HADDOCK, has the potential to greatly enhance their computational performance. In addition, FCC clustering is able to deal with symmetry and multicomponent complexes with negligible performance degradation. Furthermore, given its sole dependence on residue numbering, it allows for the clustering of mutants and gapped structures, broadening even more its usefulness to the clustering of structures coming from different trajectories or simulations. All these, allied to the simplicity of the algorithm and its flexibility in dealing with several molecule types, tailor FCC clustering for the upcoming challenges in the docking field and offer an effective alternative to traditional RMSD-based clustering methods and their inherent shortcomings.

ACKNOWLEDGMENT

The authors thank M. van Dijk (Utrecht University) for helpful discussions.

REFERENCES

1. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, Rout MP, Sali A. Determining the architectures of macromolecular assemblies. Nature 2007;450:683–694.
2. Piana S, Lindorff-Larsen K, Shaw DE. How robust are protein folding simulations with respect to force field parameterization? Biophys J 2011;100:L47–L49.
3. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci USA 1998;95:11158–11162.
4. Kozakov D, Clodfelter K, Vajda S, Camacho C. Optimal clustering for detecting near-native conformations in protein docking. Biophys J 2005;89:867–875.
5. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction-Round VII. Proteins 2007;69:3–9.
6. Janin J. Protein–protein docking tested in blind predictions: the CAPRI experiment. Mol Biosyst 2010;6:2351–2362.
7. Reva BA, Finkelstein AV, Skolnick J. What is the probability of a chance prediction of a protein structure with an rmsd of 6 A? Fold Des 1998;3:141–147.

8. Cossio P, Laio A, Pietrucci F. Which similarity measure is better for analyzing protein structures in a molecular dynamics trajectory? Phys Chem Chem Phys 2011;13:10421–10425.
9. Wallin S, Farwer J, Bastolla U. Testing similarity measures with continuous and discrete protein models. Proteins 2003;50:144–157.
10. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T. Assessment of template based protein structure predictions in CASP9. Proteins 2011;79 (Suppl 10):37–58.
11. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJE, Vajda S, Vakser I, Wodak SJ. Critical Assessment of PRedicted Interactions. CAPRI: a critical assessment of predicted interactions. Proteins 2003;52:2–9.
12. Melquiond ASJ, Karaca E, Kastritis PL, Bonvin AMJJ. Next challenges in protein-protein docking: from proteome to interactome and beyond. WIREs Comput Mol Sci, DOI: 10.1002/wcms.91.
13. Prinzie A, Van den Poel D. Incorporating sequential information into traditional classification models by using an element/position-sensitive SAM. Decision Support Syst 2006;42:508–526.
14. de Vries SJ, Melquiond ASJ, Kastritis PL, Karaca E, Bordogna A, van Dijk M, Rodrigues JPGLM, Bonvin AMJJ. Strengths and weaknesses of data-driven docking in critical assessment of prediction of interactions. Proteins 2010;78:3242–3249.
15. Wang G, Louis JM, Sondej M, Seok YJ, Peterkofsky A, Clore GM. Solution structure of the phosphoryl transfer complex between the signal transducing proteins HPr and IIA(glucose) of the *Escherichia coli* phosphoenolpyruvate:sugar phosphotransferase system. EMBO J 2000;19:5635–5649.
16. Buckle AM, Schreiber G, Fersht AR. Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution. Biochemistry 1994;33:8878–8889.
17. Rey FA, Heinz FX, Mandl C, Kunz C, Harrison SC. The envelope glycoprotein from tick-borne encephalitis virus at 2 A resolution. Nature 1995;375:291–298.
18. Loris R, Tielker D, Jaeger K-E, Wyns L. Structural basis of carbohydrate recognition by the lectin LecB from *Pseudomonas aeruginosa*. J Mol Biol 2003;331:861–870.
19. Stehle T, Harrison SC. High-resolution structure of a polyomavirus VP1-oligosaccharide complex: implications for assembly and receptor binding. EMBO J 1997;16:5139–5148.
20. Horton JR, Cheng X. PvuII endonuclease contains two calcium ions in active sites. J Mol Biol 2000;300:1049–1056.
21. Daura X, Gademann K, Jaun B, Seebach D, van Gunsteren W, Mark A. Peptide folding: when simulation meets experiment. Angew Chem Int Ed 1999;38:236–240.