ArchDB 2014

Bonet, Jaume; Planas-Iglesias, Joan; Garcia-Garcia, Javier; Marín-López, Manuel A.; Fernandez-Fuentes, Narcis; Oliva, Baldo

Published in:
Nucleic Acids Research

DOI:
10.1093/nar/gkt1189

Publication date:
2014

Citation for published version (APA):
Bonet, J., Planas-Iglesias, J., Garcia-Garcia, J., Marín-López, M. A., Fernandez-Fuentes, N., & Oliva, B. (2014). ArchDB 2014: Structural classification of loops in proteins. *Nucleic Acids Research*, *42*(D1), D315-D319. https://doi.org/10.1093/nar/gkt1189

ArchDB 2014: structural classification of loops in proteins. Supplementary Material

Supplementary Methods 1. Density Search

The DS clustering method is based on the density or mode-seeking technique (searching for regions containing a relatively dense concentration of loops around a centroid), a version of single-linkage clustering (33). Basically, the DS algorithm detects regions with a high density of loops around a centroid defined by the loop length, the internal coordinates of the bracing secondary structures, and the conformation in (ϕ,ψ) space of the aperiodic region. In this clustering, loops belonging in the same cluster have the same length, similar bracing secondary structures falling under the same interval (see (34) for details) and high percentage of residues with identical (ϕ,ψ) angles in the aperiodic fragment (identified by a consensus conformation). A cluster is required to have at least 3 loops. This type of clustering was used in the previous release of the database (16).

Supplementary Methods 2. Markov Clustering Algorithm

The MCL is a graph-based clustering algorithm. The MCL algorithm simulates a flow of information within the graph, enhancing the flow where the current is strong and hindering it where the current is weak. In MCL, the flow is controlled expanding and inflating the stochastic (Markov) matrix that represents the graph (17). One of the major reasons to use MCL was to account for some variability in the loop length, feature that was computationally unaffordable using DS. Hence, we initially grouped loops according to their length in different categories: short loops (length between 0 and 3); short-medium loops (length between 4 and 6); long-medium loops (length between 7 and 13); long loops (between 14 and 20) and very-long loops (between 21 and 30). Then, we clustered loops in each of such categories with the MCL algorithm (as obtained from http://micans.org/mcl/) using default parameters. The program performs the clustering in three steps:

i) A disconnected graph is built where nodes represent different loops.

ii) Edges are placed according to the similarity between loops, which is defined as a binary function (B). Two loops are similar (B=1) if:

   a. their geometry falls within the same interval (2): distance +/- 1 Å, hoist angle +/- 15º, packing angle +/- 15º, meridian angle +/- 25º, and

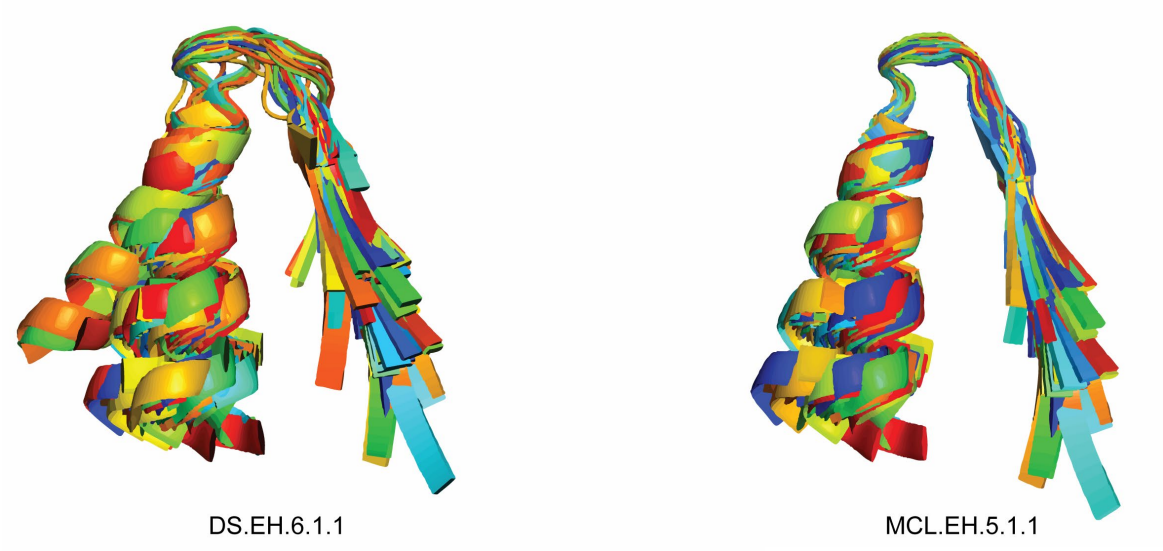
   b. the (ϕ,ψ) angles in the region defined by the loop plus two residues on the flanking secondary structures are nearly identical (the minimum percentage of required identity ranges from 95% to 98%).

   Otherwise, they are not similar (B=0).

iii) The algorithm extracts natural groups from the graph, clustering loops with similar geometry and conformation. We enforced MCL to perform this step clustering loops in disjoint (non-overlapping) groups, requiring a minimum of three loops to form a cluster. Thus, clusters formed less than three loops were disregarded and such loops were not classified.

The final classification was obtained by joining the clusters of the length-based categories of loops (See Supplementary Example 1 and Supplementary Figure S2).

Supplementary Example 1. The phosphate-binding loop (P-loop)



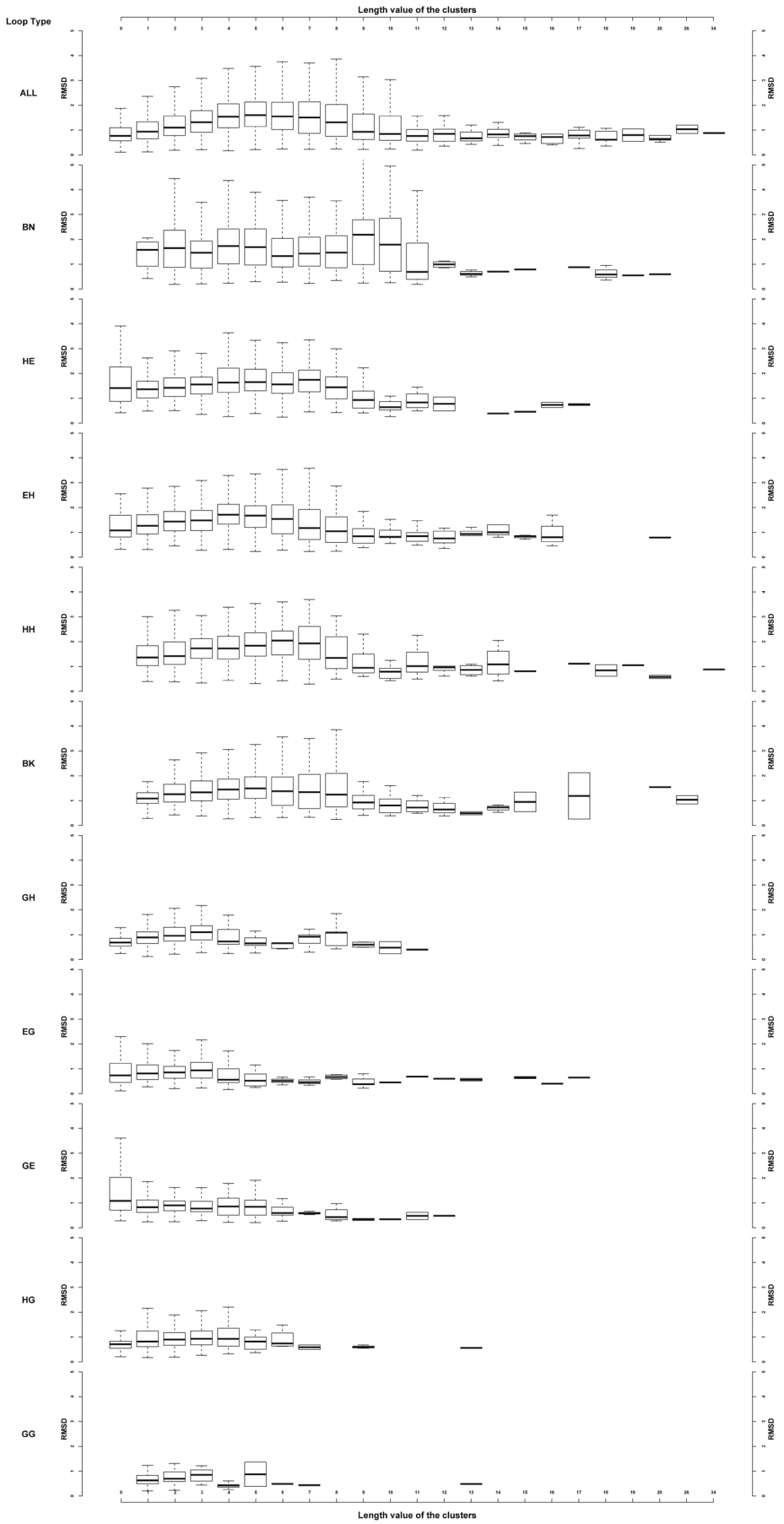
DS.EH.6.1.1 MCL.EH.5.1.1

The phosphate binding-loop or P-loop is a well-described protein motif. It commonly appears in adenine and guanine nucleotide-binding proteins, and shows a consensus sequence motif such as: GXXXXGK[TS] (29).

When looking to it in ArchDB, we can see that the loop belongs to one subclass for each clustering method: DS.EH.6.1.1 (right) and MCL.EH.5.1.1 (left). The fact that in both cases it belongs to the class code 1 and subclass code 1 means that, in both classifications, it belongs to the most populated cluster for that given type/length combination, as class and subclass identifiers are assigned by cluster-size (197 loops in the DS subclass and 146 in the MCL subclass). When looking into the subclass information, in both cases the sequence pattern is almost always conserved and correctly aligned. The SCOP (3) classification is “P-loop containing nucleoside triphosphate hydrolases” for virtually all those loops with a given SCOP assignation, as well as the ATP-binding (and, to a lesser extend, GTP-binding) from GO (30) molecular function. Both clusters also show a high number of known contacts with heteroatoms (including ATP, ADP, GTP, GDP) and known sites. The full information for each cluster can be browsed directly from the database at http://sbi.imim.es/archdb/browse/DS.EH.6.1.1, http://sbi.imim.es/archdb/browse/MCL.EH.5.1.1.

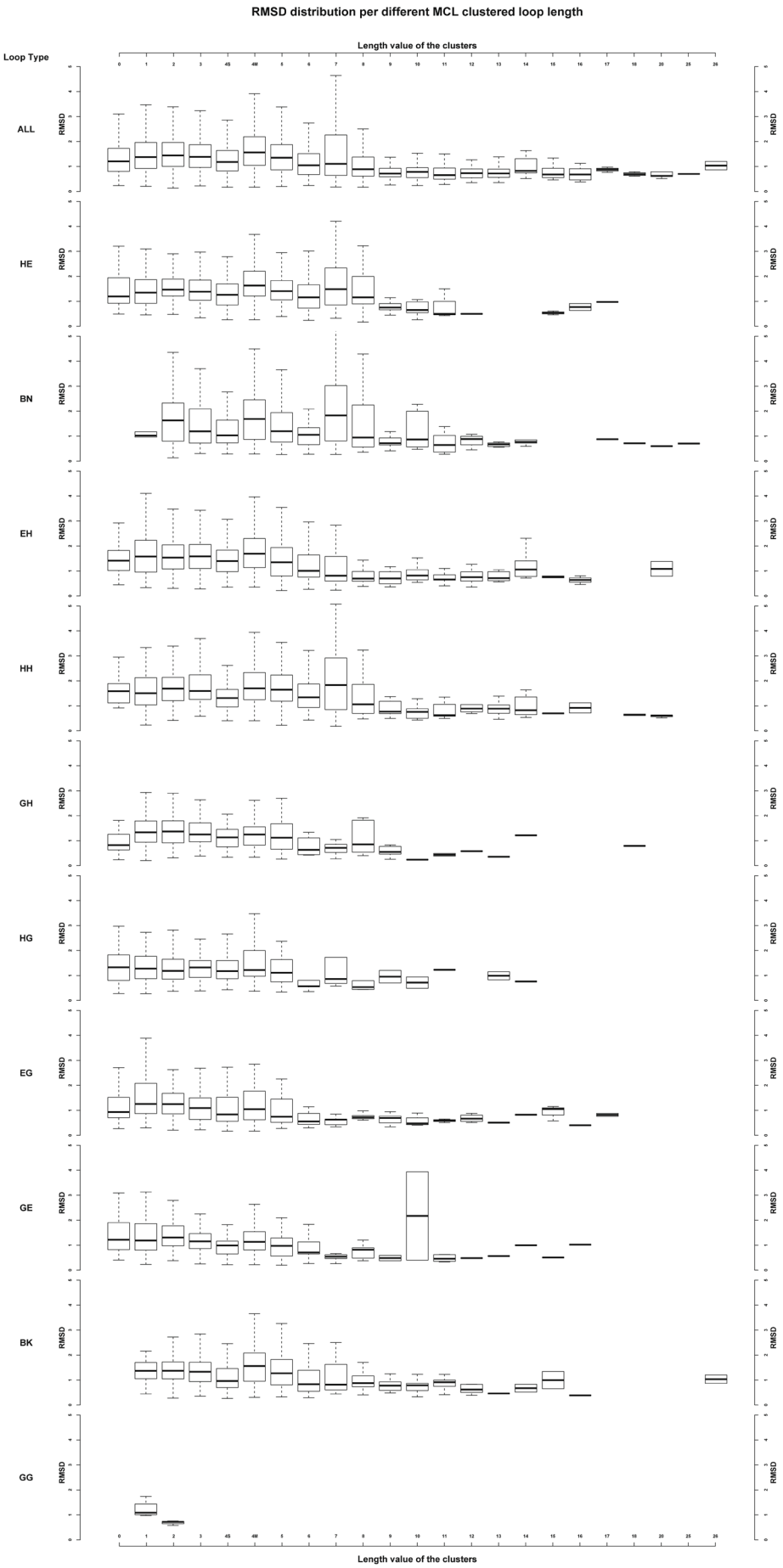
Supplementary Figure S1. Density Search (DS) clusters RMSD variation.

Distribution of the RMSD obtained from a pre-aligned structural superimposition with STAMP (24). The given alignment is the one obtained through the clustering process. Each line represents a loop type (according to its bracing secondary structures) while the first line represents the combination of all. Clusters are grouped according to their length (number of amino acids in the aperiodic region of the loop).


RMSD distribution per different DS clustered loop length
Loop Type
Length value of the clusters
ALL
BN
HE
EH
HH
BK
GH
EG
GE
HG
GG
RMSD
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 26 34

Supplementary Figure S2. Markov CLustering (MCL) clusters RMSD variation.

Distribution of the RMSD obtained from a pre-aligned structural superimposition with STAMP (24). The given alignment is the one obtained through the clustering process. Each line represents a loop type (according to its bracing secondary structures) while the first line represents the combination of all. Clusters are grouped according to their length (number of amino acids in the aperiodic region of the loop).


RMSD distribution per different MCL clustered loop length
Loop Type
Length value of the clusters
ALL
HE
BN
EH
HH
GH
HG
EG
GE
BK
GG
RMSD
0 1 2 3 4S 4M 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 25 26

Supplementary References

3. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–9.
16. Espadaler,J., Fernandez-Fuentes,N., Hermoso,A., Querol,E., Avilés,F.X., Sternberg,M.J.E. and Oliva,B. (2004) ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic Acids Res.*, **32**, D185–8.
17. Van Dongen,S. (2008) Graph Clustering Via a Discrete Uncoupling Process. *SIAM. J. Matrix Anal. & Appl.*, **30**, 121–141.
24. Russell,R.B. and Barton,G.J. (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309–323.
29. Saraste,M., Sibbald,P.R. and Wittinghofer,A. (1990) The P-loop--a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.*, **15**, 430–434.
30. The Gene Ontology Consortium (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.*, **40**, D559–D564.
33. Everitt,B. (1974) Chapter 3. In *Cluster Analysis*. London, pp. 45–60.
34. Oliva,B., Bates,P.A., Querol,E., Avilés,F.X. and Sternberg,M.J. (1997) An automated classification of the structure of protein loops. *J. Mol. Biol.*, **266**, 814–830.