Multidimensional PCA-based clustering of molecular dynamics trajectories

Athanasios Baltzis and Nicholas M. Glykos

Department of Molecular Biology and Genetics, Democritus University of Thrace, University Campus, 68100 Alexandroupolis, Greece

Introduction

Molecular dynamics simulations produce a variety of molecular conformations resulting in a huge amount of data. Reducing and meaningfully grouping this data without losing so much biological information is an essential task in order to understand the conformational changes of biomolecules.

Principal Component Analysis (PCA), especially dihedral PCA (dPCA)[1], and Clustering are two effective techniques used to achieve the above goals.[2] On one hand, PCA is a mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set. On the other hand, clustering is a widely-known method for grouping objects into subgroups (clusters) in such a way to minimize intra-cluster and maximize inter-cluster differences.

The clustering analysis of molecular dynamics trajectories is usually based on either two or three principal components due to the availability of graphical representation in two or three dimensions.[3] However, assuming that a great deal of biological information is lost when data compressed in two or three dimensions, we suggest an alternative approach to reveal this information. Here we present the program *cluster5D* that performs multidimensional clustering analysis based on the top five principal components.

cluster5D is written in C programming language and can be used either stand alone via the command line or via the program *grcarma*.[4]

Algorithm

The essence of the algorithm is the following : the PC1-PC2-PC3-PC4-PC5 values are used to calculate a (five-dimensional) density distribution function and to map this distribution on a five dimensional matrix. The higher the value of the matrix at a point, the larger the number of frames with corresponding PC values close to that point. The algorithm, then, performs a peak-picking on this 5D density distribution and identifies clusters as sets of frames with similar PC values. The crucial parameter for the peak-picking step is the density threshold above which peaks are picked and its calculation depends on the dataset and the user's argument.

Program Availability

Cluster5D is free and open source software. It is available for the most recent versions of Windows, GNU/Linux and MACOSX operating systems. The github repository (https://github.com/athbaltzis/cluster5D) contains the source code and ready-made executables for all major architectures as well as a detailed documentation.

Results

cluster5D has been tested using several data sets derived from a wide range of molecular dynamics trajectories. Comparison with the results obtained from 3D-based clustering indicate that this 5D approach may be more sensitive in identifying minor but persistent molecular conformations. Figures 1 and 2 below show representative results from the application of the program on two different trajectories: the first is *lytA* (choline-binding repeat peptide of lytA protein) and the second is *gp41* (a gp-41-derived peptide). Both of them derive from Amber-ILDN force field simulations.

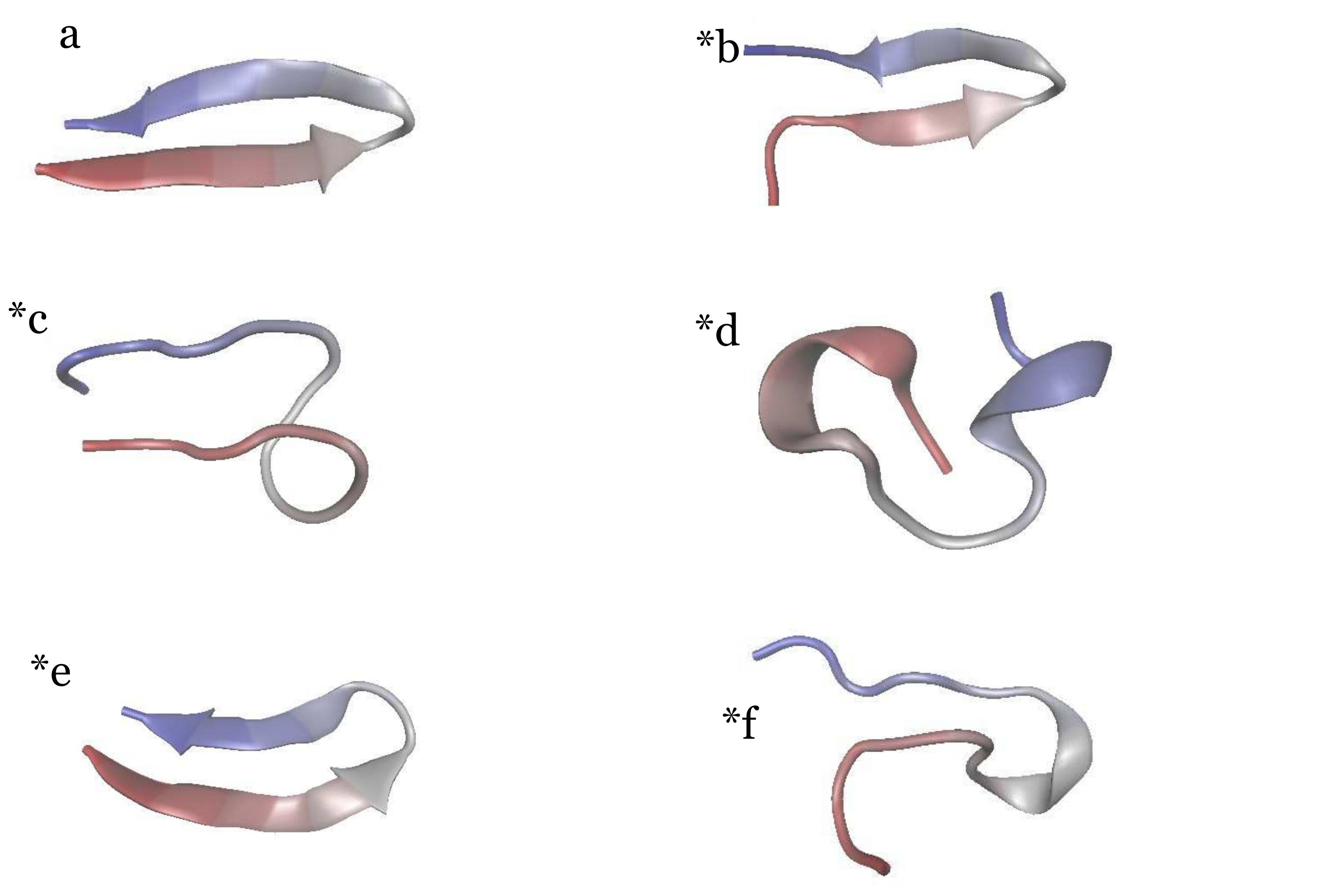

Figure 1. Representative structures for 6 clusters of lytA peptide produced by cluster5D. The structures with an asterisk (*) were not found with 3D-based clustering. a) cluster 1 (824,097 frames), b) cluster 4 (26,221 frames), c) cluster 10 (10,050 frames), d) cluster 11 (9487 frames), e) cluster 17 (3862 frames), f) cluster 18 (2933 frames).

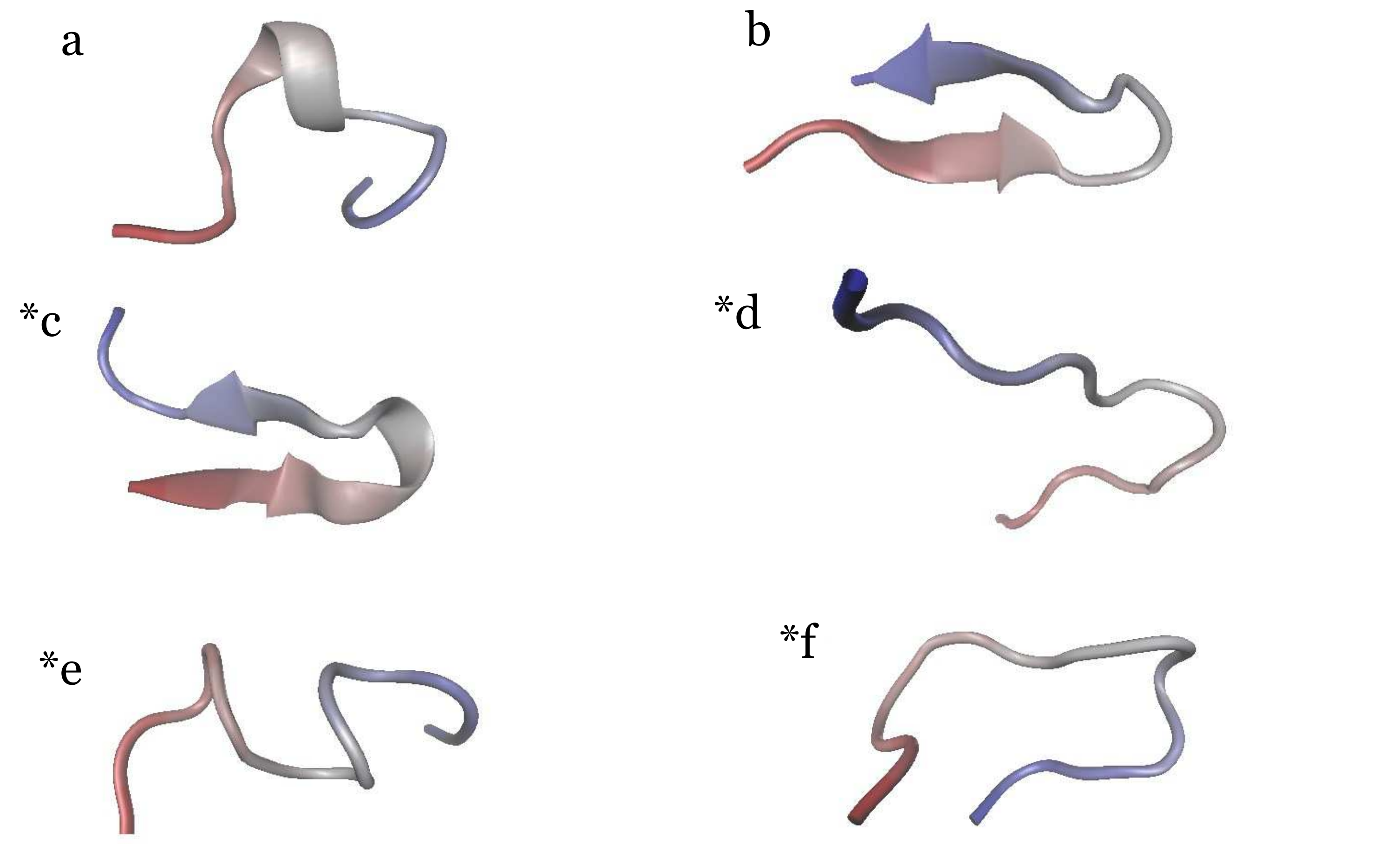

Figure 2. Representative structures for 6 clusters produced by cluster5D of gp41 peptide. The structures with an asterisk (*) were not found with 3D-based clustering a) cluster 1 (570,174 frames), b) cluster 3 (93,134 frames), c) cluster 6 (36,104 frames), d) cluster 13 (6882 frames), e) cluster 14 (9434 frames), f) cluster 17 (7419 frames).

References

1. Y. Mu, P. H. Nguyen, G. Stock. Proteins 2005, 58, 45-52. DOI:10.1002/prot.20310
2. A. Wolf, K. N. Kirschner. J. Mol. Model 2013, 19, 539-549.DOI: 10.1007/s00894-012-1563-4
3. A. Altis, M. Otten, P. H. Nguyen, R. Hegger, G. Stock. The Journal of Chemical Physics 2008, 128, 245102. DOI: 10.1063/1.2945165
4. P. I. Koukos, N. M. Glykos. J. Comput. Chem. 2013, 34,2310-2312.DOI:10.1002/jcc.23381