Protein Peeling 3D: new tools for analyzing protein structures.

Jean-Christophe Gelly, Alexandre de Brevern

Accepted for publication in Bioinformatics (2011)

Protein Peeling 3D: New tools for analyzing protein structures

Jean-Christophe Gelly[*,1,2,3] and Alexandre G. de Brevern[1,2,3]

1 Dynamique des Structures et Interactions des Macromolécules Biologiques, INSERM UMR-S 665, 6, rue
Alexandre Cabanel - 75739 Paris Cedex 15, France
2 Université Paris Diderot - Paris 7, 6, rue Alexandre Cabanel - 75739 Paris Cedex 15, France
3 Institut National de la Transfusion Sanguine, 6, rue Alexandre Cabanel - 75739 Paris Cedex 15, France

*To whom correspondence should be addressed.

ABSTRACT

Summary: We present an improved version of our Protein Peeling web server dedicated to the analysis of protein structure architecture through the identification of Protein Units produced by an iterative splitting algorithm. New features include identification of structural domains, detection of unstructured terminal elements and evaluation of the stability of protein unit structures.

Availability: The website is free and open to all users with no login requirements at http://www.dsimb.inserm.fr/dsimb-tools/peeling3

Contact: jean-christophe.gelly@univ-paris-diderot.fr

1 INTRODUCTION

Analyzing the architecture and organization of protein structures is essential for understanding protein flexibility, folding, functions and interactions. We previously proposed an innovative representation of protein architecture that gives a detailed description of protein structure anatomy. Proteins are split into sets of compact subregions, called protein units (PUs). A PU corresponds to one sequence fragment characterized by a high number of intra-PU contacts and a low number of inter-PU contacts. Contact probabilities between residues are computed as distances between Cα atoms using a logistic function. Using this contact probability matrix and an optimization procedure based on Matthews's correlation coefficient (MCC) between sub-matrices, the algorithm defines optimal cutting points that separate the region examined into two or three PUs. This methodology is called Protein Peeling (PP) (Gelly et al., 2006). The process is iterated until the compactness of the PUs reaches a given limit. An index assesses the compactness quality and relative independence of each PU.

PUs bridge the representation and description gap between secondary structures and structural domains. PP is a useful tool for better understanding and analyzing the organization of protein structures. The Protein Peeling 2 web server (Gelly et al., 2006) has been developed to accommodate advanced parameters. Here, we present a new version of our Protein Peeling web server that offers substantial improvements and new features: recognition of unstructured N- or C-terminal segments, a novel scoring function for PU characterization and identification of structural domains.

2 METHODS

2.1 Unstructured extremity recognition

Unstructured N- or C-terminus segments can be problematic for protein analysis or molecular simulation. Using PP, this type of segment was observed in half of the protein structures found in Protein Data Bank (Faure et al., 2009). We thus included a new assignment method in the web server. A PU is considered unstructured if it is isolated at the first cutting event and never thereafter.

2.2 Energy calculation

PUs show a wide range of shapes and many differences in the type or density of internal contacts. These characteristics are related to internal energy. To evaluate stability and contact energy, we implemented the DOPE statistical potential (Shen and Sali, 2006) to compute the pseudo-energy of PUs. This value gives an approximation of the internal energy of PUs and insight on the structural significance of observed PUs.

2.3 Structural domain identification

One major potential for the Protein Peeling web server was the possibility to use the PP method to identify protein domains. We therefore developed a new algorithm called domain reconstruction (DR). The identification process includes two steps: (1) PUs are isolated using the basic PP top-down induction approach (2) PUs obtained at the final level of PP are assembled using the DR algorithm with a bottom-up strategy, similar to hierarchical agglomerative clustering, as follows:

Starting from the final PU set, the algorithm combines each, unique PU to another PU at each recursive step. Thus, PUs are gradually merged at each step, forming many possible domain delineations. The best merging events are chosen among all the possible events according to the contact ratio (CR) criterion, derived from the Protein Domain Parser method (Alexandrov and Shindyalov, 2003). This criterion is based on contact probability matrix computed in PP and helps estimate the quality of merging events.

The CR criterion measured for PUs i and j is

$$CR = \frac{cp(i,j) / S(i)^{\alpha} \times S(j)^{\alpha}}{c(ij) \times S(ij)}$$

where cp(i,j) is contact probability between PUi and PUj, S(i) and S(j) are the length of PUi and PUj. Finally c(ij) and S(ij) are respectively the contact probability and the size of the whole domain formed by merging PUi and PUj. The α value is 0.43 as in the PDP method.

High CR values indicate a high number of contacts between PUi and PUj; consequently theses PUs are good candidates for merging them into one domain. During the process, a second measure, the average contact probability density (CPD), also used in other method (Holm and Sander, 1994), is computed to estimate the quality of newly formed domains.

The assembly routine is reiterated and the number of domains gradually decreases until all PUs are merged into a unique domain corresponding to the whole protein. After producing different levels of potential domain delineation, a post-processing step is launched to examine and choose the best domain delineations. First, domain partitions at each level are sorted according to the mean CPD of domain delineation. Second, starting from the highest level, *i.e.*, the full protein, each level is examined and accepted or not, according to a criterion that combines both the minimum CPD measured between domains and the maximum CR measured among all domains. Domains with a size of less than 30 are discarded.

3 SERVER

Given the 3D coordinates of a protein, the server is able to identify PUs, domain delineations, and, if they exist, unstructured extremities. The server also provides some analysis on isolated PUs.

3.1 Input

Protein Peeling 3D only accepts protein structure in PDB file format (Berman et al., 2000). The file is checked (to verify consistency of the format), cleaned (alternative positions are removed and non-standard residues are renamed) and renumbered. In multiprotein chain files, only the first chain is treated. Various parameters can be adjusted, such as parameters that direct cutting, those used for calculating the contact probability map, and that only regular secondary structures should be taken into account.

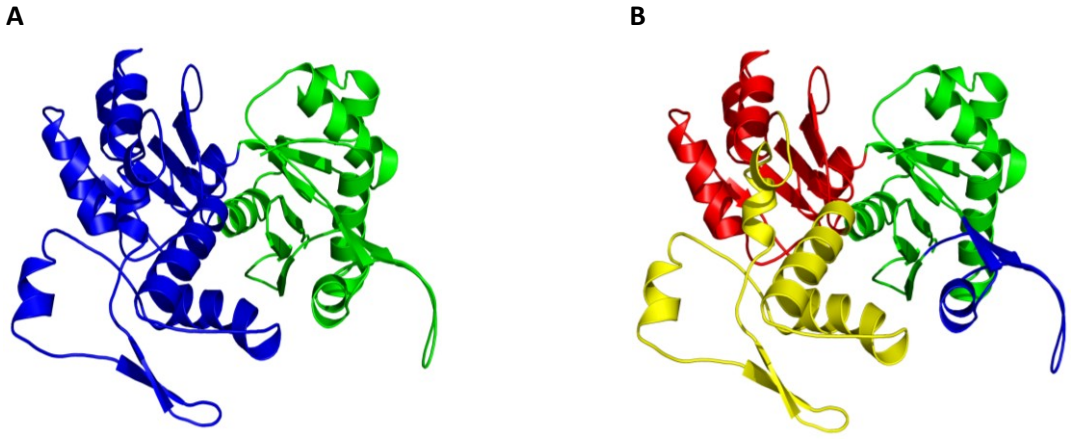

Fig. 1. Comparison between two delineations (**A)** and (**B**) of actin (PDB code: 1atnA) using DR. Delineation (**A)** is identical to SCOP (Murzin, et al., 1995) and (**B)** is identical to the one determined in Jones et al., (1998) dataset.

3.2 Output and graphics

Many graphics and raw output files are available through the web server at the end of process. The graphical results include the contact probability map, different representations of protein structures in which PUs or domains can be easily identified by color, and representations of primary structure, in which PUs and domains are colored and clearly delimited.

ACKNOWLEDGEMENTS

This work was supported by grants from the French Ministry of Research, Université Paris Diderot – Paris 7, National Institute for Blood Transfusion (INTS), National Institute for Health and Medical Research (INSERM) and Orchid Partenariat Hubert Curien.

REFERENCES

Alexandrov, N. and Shindyalov, I. (2003) PDP: protein domain parser, *Bioinformatics*, **19**, 429-430.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*. 235- 242.

Day, R., Beck, D.A., Armen, R.S. and Daggett, V. (2003) A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary, *Protein Sci*, **12**, 2150-2160.

Faure, G., Bornot, A. and de Brevern, A.G. (2009) Analysis of protein contacts into Protein Units, *Biochimie*, **91**, 876-887.

Gelly, J.C., de Brevern, A.G. and Hazout, S. (2006) 'Protein Peeling': an approach for splitting a 3D protein structure into compact fragments, *Bioinformatics*, **22**, 129-133.

Gelly, J.C., Etchebest, C., Hazout, S. and de Brevern, A.G. (2006) Protein Peeling 2: a web server to convert protein structures into series of protein units, *Nucleic Acids Res*, **34**, W75-78.

Holm, L. and Sander, C. (1994) Parser for protein folding units. *Proteins*. 256-268.

Jones, S., Stewart, M., Michie, A., Swindells, M.B., Orengo, C. and Thornton, J.M. (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis, *Protein Sci*, **7**, 233-242.

Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 536-540.

Shen, M.Y. and Sali, A. (2006) Statistical potential for assessment and prediction of protein structures, *Protein Sci*, **15**, 2507-2524.

Veretnik, S. and Shindyalov, I. (2007) Computational Methods for Domain Partitioning of Protein Structures. In Xu, Y., Xu, D. and Liang, J. (eds), *Computational Methods for Protein Structure Prediction and Modeling*. Springer, New York, 125-145.