BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 19 2013, pages 2512–2514
doi:10.1093/bioinformatics/btt405

Databases and ontologies

Advance Access publication July 11, 2013

CoDNaS: a database of conformational diversity in the native state of proteins

Alexander Miguel Monzon, Ezequiel Juritz, María Silvina Fornasari and Gustavo Parisi*

Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, B1876BXD, Buenos Aires, Argentina

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Conformational diversity is a key concept in the understanding of different issues related with protein function such as the study of catalytic processes in enzymes, protein-protein recognition, protein evolution and the origins of new biological functions. Here, we present a database of proteins with different degrees of conformational diversity. Conformational Diversity of Native State (CoDNaS) is a redundant collection of three-dimensional structures for the same protein derived from protein data bank. Structures for the same protein obtained under different crystallographic conditions have been associated with snapshots of protein dynamism and consequently could characterize protein conformers. CoDNaS allows the user to explore global and local structural differences among conformers as a function of different parameters such as presence of ligand, post-translational modifications, changes in oligomeric states and differences in pH and temperature. Additionally, CoDNaS contains information about protein taxonomy and function, disorder level and structural classification offering useful information to explore the underlying mechanism of conformational diversity and its close relationship with protein function. Currently, CoDNaS has 122 122 structures integrating 12 684 entries, with an average of 9.63 conformers per protein.

Availability: The database is freely available at http://www.codnas.com.ar/.

Contact: gusparisi@gmail.com

Received on April 25, 2013; revised on June 18, 2013; accepted on July 5, 2013

1 INTRODUCTION

It is well established that the native state of a protein is represented by an ensemble of conformers in equilibrium (Tsai *et al.*, 1999). Different factors, such as ligands, post-translational modifications or pH variations could shift this equilibrium toward a given conformer (Kumar *et al.*, 2000). Following conformational selection theory, the change in the dynamic landscape and the re-distribution of conformer population is a key feature to understand protein function (Nussinov and Ma, 2012). It has been shown that different structures for the same protein obtained under different conditions represent snapshots of protein dynamism and then characterize putative structural conformers (Best *et al.*, 2006). Here, we describe the design, development and use of a database of proteins showing different degrees of conformational diversity. Conformational Diversity of Native State (CoDNaS) is a collection of redundant structures, obtained from different experimental protocols, for the same protein. CoDNaS is a database of complete proteins, linked with Protein Conformational DataBase (Juritz *et al.*, 2011), a previous development from our group oriented to the analysis of conformational diversity in protein domains. The database is extensively linked with physicochemical and biological information, and it is designed to allow the user to explore how different parameters could modulate protein conformational diversity.

2 IMPLEMENTATION

Using the BlastClust application (Altschul *et al.*, 1990) on protein data bank database (Berman, 2000), for every protein deposited, we obtained all available clusters with 100% of sequence identity. We only considered clusters with more than two structures for each protein and structures with resolutions <4 Å. Structural similarity measures were performed using MAMMOTH (Ortiz *et al.*, 2002), ProFit (McLachlan, 1982) and TM-score (Zhang and Skolnick, 2004) for all the structures retrieved for each protein. The maximum C-alpha root-mean-square deviation (RMSD) value for every entry was registered as a measure of the conformational diversity extension. Each structure in CoDNaS is characterized by the conditions of its experimental determination: presence of ligands, mutations, post-translational modifications, oligomeric state, pH, loop regions and temperature. According to the conformational selection theory, these factors could be used to study conformational changes and to correlate them with biological information. In addition, proteins were linked with other databases, such as UniProt (Jain *et al.*, 2009), gene ontology (Ashburner *et al.*, 2000), enzyme commission (Kotera *et al.*, 2004), CATH (Greene *et al.*, 2007), SIFTS (Velankar *et al.*, 2005) and MobiDB (Di Domenico *et al.*, 2012), to obtain a broad spectrum of biological and physicochemical information such as taxonomy, source organism, protein function, degree of protein disorder and structural class, among others. Structures (putative conformers) for a given protein were clustered using two algorithms: hierarchical clustering and affinity propagation clustering. C-alpha RMSD per position values were also derived to complement global similarity measures together with additional parameters characterizing conformational change (such as degree of protein disorder, loopy residues content and solvent-exposed area).

The web application server is implemented on HTML, PHP and Java languages, connected with a MySQL-based database. CoDNaS search is based on dynamics SQL queries generated in PHP.

*To whom correspondence should be addressed.

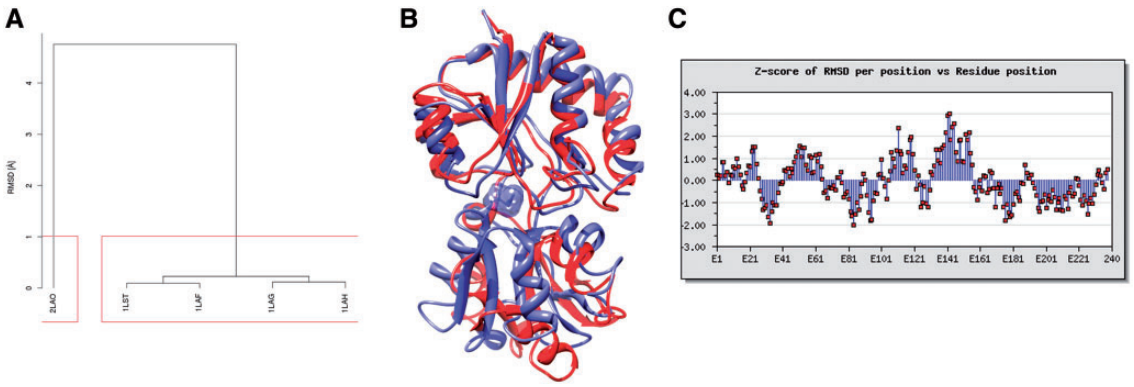

Fig. 1. (**A**) Example of a protein entry retrieved from CoDNaS showing conformational diversity. In this case, we show the LAO-binding protein exhibiting five different putative conformers. Hierarchical clustering using C-alpha RMSD as distance measure is shown. (**B**) Maximum conformational diversity is obtained when bound and unbound conformers are compared [C-alpha RMSD = 4.71 1LAH with bound ligand represented as space-filled model and 2LAO_A]. (**C**) Per position C-alpha RMSD z-score showing local structural differences between conformers displayed in (B)

3 USAGE

CoDNaS database can be searched and browsed using different global structural measures between conformers. This structural information is linked with physicochemical and biological parameters. The user can define and combine different parameters to limit the search such as C-alpha RMSD extension, causes of conformational diversity or taxonomy, among others. A typical query, for example, could involve the search for those proteins showing a conformational diversity >1 Å of C-alpha RMSD owing to the presence of ligands. This query will retrieve all entries containing at least one pair of conformers presenting a C-alpha RMSD >1 Å where the unique difference in the structure estimation (NMR or XRD) is the presence or absence of ligands. An example of that search is shown in Figure 1, where five conformational states are observed for the protein lysine, arginine and ornithine-binding protein (LAO protein) (Fig. 1A). One of these structures (2LAO) has no ligand bound, whereas the other four (1LAH, 1LAF, 1LAG and 1LST) have arginine, lysine or ornithine bound. Maximum conformational diversity is obtained when the conformer with unbound ligand is compared with any of the other conformers containing a ligand. The conformers containing a ligand are structurally similar to each other as derived from the clustering analysis shown in Figure 1A. Global and local structural differences are shown in Figure 1B and C. The pair of conformers with maximum structural diversity is shown in Figure 1B with a C-alpha RMSD of 4.71 Å. Figure 1C represents the *z*-scores derived from per position C-alpha RMSD showing the local differences between bound and unbound forms of the protein. Similar information is provided for all the aligned pairs of conformers for each protein at the web server. For each entry in CoDNaS, structural alignments, sequence-derived alignment and structural and physicochemical comparisons between conformers are available to download. Other parameters besides the presence of ligands can be used to search CoDNaS database, such as differences in pH and temperature, change in oligomeric structure (e.g. the occurrence of transient oligomers for a given protein), presence of mutations, presence of disordered and/or loopy regions and the presence of post-translational modifications.

4 DISCUSSION

Conformational diversity is a key feature to understand protein function and evolution (Juritz *et al*., 2013), enzyme catalysis and molecular recognition (Boehr *et al*., 2010).

Recently, it has been shown that explicit consideration of conformational diversity improves prediction of disease-related mutations (Juritz *et al*., 2012) and protein-protein interactions and molecular docking (Kuzu *et al*., 2013; Osguthorpe *et al*., 2012). As computational prediction of large conformational changes could be challenging [e.g. using molecular dynamics (Petruk *et al*., 2013)], CoDNaS database could provide a useful dataset to test new tools to predict and explore possible causes of conformational diversity in proteins.

ACKNOWLEDGEMENTS

G.P. and M.S.F. are grateful for receiving the following grants: PIP CONICET (112-200801-02849) and UNQ (1004/11). A.M. and E.J. have CONICET fellowships.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Berman,H.M. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Best,R.B. *et al.* (2006) Relation between native ensembles and experimental structures of proteins. *Proc. Natl Acad. Sci. USA*, **103**, 10901–10906.

Boehr,D.D. *et al.* (2010) The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.*, **5**, 789–796.

Di Domenico,T. *et al.* (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics*, **28**, 2080–2081.

Greene,L.H. *et al.* (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.*, **35**, D291–D297.

Jain,E. *et al.* (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics*, **10**, 136.

Juritz,E. *et al.* (2012) On the effect of protein conformation diversity in discriminating among neutral and disease related single amino acid substitutions. *BMC Genomics*, **13** (**Suppl. 4**), S5.

Juritz,E. *et al.* (2013) Protein conformational diversity modulates sequence divergence. *Mol. Biol. Evol.*, **30**, 79–87.
Juritz,E.I. *et al.* (2011) PCDB: a database of protein conformational diversity. *Nucleic Acids Res.*, **39**, D475–D479.
Kotera,M. *et al.* (2004) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.*, **126**, 16487–16498.
Kumar,S. *et al.* (2000) Folding and binding cascades: dynamic landscapes and population shifts. *Protein Sci.*, **9**, 10–19.
Kuzu,G. *et al.* (2013) Exploiting conformational ensembles in modeling protein-protein interactions on the proteome scale. *J. Proteome Res.*, **12**, 2641–2653.
McLachlan,A.D. (1982) Rapid comparison of protein structures. *Acta Crystallogr. A*, **38**, 871–873.
Nussinov,R. and Ma,B. (2012) Protein dynamics and conformational selection in bidirectional signal transduction. *BMC Biol.*, **10**, 2.
Ortiz,A.R. *et al.* (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
Osguthorpe,D.J. *et al.* (2012) Generation of receptor structural ensembles for virtual screening using binding site shape analysis and clustering. *Chem Biol Drug Des.*, **80**, 182–193.
Petruk,A.A. *et al.* (2013) Molecular dynamics simulations provide atomistic insight into hydrogen exchange mass spectrometry experiments. *J. Chem. Theory Comput.*, **9**, 658–669.
Tsai,C.J. *et al.* (1999) Folding funnels, binding funnels, and protein function. *Protein Sci.*, **8**, 1181–1190.
Velankar,S. *et al.* (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, D262–D265.
Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.