BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 19 2013, pages 2517–2518
doi:10.1093/bioinformatics/btt427

Databases and ontologies

Advance Access publication July 26, 2013

mpMoRFsDB: a database of molecular recognition features in membrane proteins

Foivos Gypas, Georgios N. Tsaousis and Stavros J. Hamodrakas*

Faculty of Biology, Department of Cell Biology and Biophysics, University of Athens, Panepistimiopolis, Athens 157 01, Greece

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Molecular recognition features (MoRFs) are small, intrinsically disordered regions in proteins that undergo a disorder-to-order transition on binding to their partners. MoRFs are involved in protein–protein interactions and may function as the initial step in molecular recognition. The aim of this work was to collect, organize and store all membrane proteins that contain MoRFs. Membrane proteins constitute ~30% of fully sequenced proteomes and are responsible for a wide variety of cellular functions. MoRFs were classified according to their secondary structure, after interacting with their partners. We identified MoRFs in transmembrane and peripheral membrane proteins. The position of transmembrane protein MoRFs was determined in relation to a protein's topology. All information was stored in a publicly available mySQL database with a user-friendly web interface. A Jmol applet is integrated for visualization of the structures. mpMoRFsDB provides valuable information related to disorder-based protein–protein interactions in membrane proteins.

Availability: http://bioinformatics.biol.uoa.gr/mpMoRFsDB

Contact: shamodr@biol.uoa.gr

Received on June 3, 2013; revised on July 2, 2013; accepted on July 18, 2013

1 INTRODUCTION

Intrinsically disordered proteins (IDPs) possess no rigid three-dimensional structure under physiological conditions, yet they are functionally active (Uversky, 2011). IDPs are separated in fully disordered proteins and partially disordered proteins (Oldfield *et al.*, 2005a). Partially disordered proteins contain intrinsically disordered regions (IDRs). IDRs are found in both prokaryotes and eukaryotes. In all, 20–30% of prokaryotic proteins (Dunker *et al.*, 2000) and more than half of eukaryotic proteins contain IDRs (Oldfield *et al.*, 2005a). Vast abundance and functional importance characterize these proteins. For a deeper understanding of IDPs and IDRs, several databases have been developed: DisProt (Vucetic *et al.*, 2005), (Sickmeier *et al.*, 2007), MobiDB (Di Domenico *et al.*, 2012), IDEAL (Fukuchi *et al.*, 2012), ComSin (Lobanov *et al.*, 2010) and D(2)P(2) (Oates *et al.*, 2013) provide information about experimentally determined or theoretically predicted IDPs and IDRs. Moreover, a variety of predictors have been developed for the prediction of IDRs from protein sequence (He *et al.*, 2009).

*To whom correspondence should be addressed.

Special cases of IDRs are molecular recognition features (MoRFs) or molecular recognition elements (Mohan *et al.*, 2006). MoRFs are small regions (between 10 and 70 residues) in proteins that undergo a disorder-to-order transition on binding to their partners (Tompa, 2002; Uversky *et al.*, 2000; Wright and Dyson, 1999). Proteins containing MoRFs play an important role in molecular recognition. When they are bound to their partners, MoRFs can take various shapes according to their secondary structure. They can form alpha-helices (α-MoRFs), beta-strands (β-MoRFs), irregular structures (i-MoRFs) or a combination of the previous elements (complex-MoRFs). A number of predictors are available for the prediction of MoRFs from protein sequences (Cheng *et al.*, 2007; Disfani *et al.*, 2012; Dosztanyi *et al.*, 2009; Mooney *et al.*, 2012; Oldfield *et al.*, 2005b).

Membrane proteins constitute 30% of fully sequenced proteomes and are responsible for a wide variety of crucial cellular functions, such as binding and signaling (Krogh *et al.*, 2001). Membrane proteins are separated in transmembrane proteins, peripheral membrane proteins and lipid-anchored proteins. Transmembrane proteins are divided into single-spanning and multi-spanning proteins, according to the number of transmembrane segments. An important number of MoRFs can be found in membrane proteins (Mohan *et al.*, 2006) and especially in transmembrane proteins (Kotta-Loizou *et al.*, 2013). IDRs are included in both alpha-helical and beta-barrel transmembrane proteins (Xue *et al.*, 2009) and occur mostly on the cytoplasmic side of human plasma transmembrane proteins (Minezaki *et al.*, 2007; Stavropoulos *et al.*, 2012).

mpMoRFsDB is the first publicly available database that collects and provides information about MoRFs found in membrane proteins.

2 METHODS

An initial dataset was constructed from the Protein Data Bank (PDB) (Berman *et al.*, 2000), following the methodology proposed by Mohan *et al.* (2006). We retrieved protein complexes containing at least two entities, with one chain varying from 10 to 70 residues and a second one having a length >100 residues (until May 2013). We further removed proteins where the MoRF's sequence contained errors or not valid amino acid residues, ending up with 2458 PDB entries mapping to 785 unique Uniprot Accession numbers (Uniprot_Consortium, 2012). Membrane proteins were selected using Uniprot's annotation. Moreover, we used the secondary structure assignment and the accessible surface area values inferred by DSSP (Kabsch and Sander, 1983) to categorize MoRFs and to evaluate whether a MoRF can interact with its possible partner, respectively. The position of transmembrane protein MoRFs in relation to a protein's topology was determined. Transmembrane protein topology was determined based on experimentally derived data from ExTopoDB (Tsaousis *et al.*, 2010) and Uniprot.

The process is automated so that new MoRFs can be collected from membrane proteins, as novel structures are deposited in PDB. Finally, we organized all data in a publicly available mySQL database, with a user-friendly web interface based on HTML, CSS, PHP and Javascript. Protein information can be accessed through three different file formats (Fasta, Text and XML), apart from the classic web view. Moreover, the entire database can be downloaded locally for further analysis.

3 RESULTS

The database includes 173 membrane proteins containing 244 MoRFs. Membrane proteins are divided in 102 transmembrane proteins (70 single-spanning and 32 multi-spanning) and 71 peripheral membrane proteins. MoRFs were classified in categories according to their secondary structure when bound to their partners, with 33.47% categorized as α-MoRFs, 3.83% as β-MoRFs, 60.48% as i-MoRFs and 2.22% as complex-MoRFs.

In the main page of mpMoRFsDB, a user may find links to the following tools: Browse, Search, Blast Search and Download. Through the Browse page, the user has the ability to browse all the entries. Moreover, there is an option for browsing by membrane protein type (transmembrane or peripheral) or by the secondary structure of MoRFs (α-MoRFs, β-MoRFs, i-MoRFs and complex-MoRFs). Through Search, the user may submit advanced queries, whereas through Blast Search, we provide an interface for running Blast searches against the database. Each entry contains information about the respective membrane protein and related MoRFs. A Jmol (Hanson, 2010) applet is integrated for visualization of the structures, and cross-references to many publicly available databases are included, providing information for protein domains, molecular interactions and diseases. In the case of transmembrane proteins, we determined whether the MoRFs are positioned in the cytoplasmic or the extracellular space. We observed that the majority of MoRFs in transmembrane proteins are found in the cytoplasmic side.

4 DISCUSSION

A database containing MoRFs in membrane proteins was constructed. Data were collected with automated Perl scripts and verified manually. The whole process can easily be repeated, and we intend to update the database every 6 months. The proteins in our database are highly connected nodes in protein interaction networks (52% of mpMoRFsDB's entries have more than five interactions in molecular interaction databases) and are essential to cell survival (Jeong *et al.*, 2001). mpMoRFsDB provides an up-to-date dataset, which can be used for the design and evaluation of methods predicting MoRFs in membrane proteins. The database will contribute to the emerging 'protein non-folding problem' (Dill and MacCallum, 2012) and provide insights in disorder-based interactions in membrane proteins.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers and the handling associate editor for their valuable comments and constructive criticism.

Funding: The present work was funded by the SYNERGASIA 2009 PROGRAMME. This Programme is co-funded by the European Regional Development Fund and National resources (Project Code 09SYN-13-999).

Conflict of interest: none declared.

REFERENCES

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
Cheng,Y. *et al.* (2007) Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry*, **46**, 13468–13477.
Di Domenico,T. *et al.* (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics*, **28**, 2080–2081.
Dill,K.A. and MacCallum,J.L. (2012) The protein-folding problem, 50 years on. *Science*, **338**, 1042–1046.
Disfani,F.M. *et al.* (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics*, **28**, i75–i83.
Dosztanyi,Z. *et al.* (2009) ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*, **25**, 2745–2746.
Dunker,A.K. *et al.* (2000) Intrinsic protein disorder in complete genomes. *Genome Inform.*, **11**, 161–171.
Fukuchi,S. *et al.* (2012) IDEAL: intrinsically disordered proteins with extensive annotations and literature. *Nucleic Acids Res.*, **40**, D507–D511.
Hanson,R.M. (2010) Jmol – a paradigm shift in crystallographic visualization. *J. Appl. Crystallogr.*, **43**, 1250–1260.
He,B. *et al.* (2009) Predicting intrinsic disorder in proteins: an overview. *Cell Res.*, **19**, 929–949.
Jeong,H. *et al.* (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
Kotta-Loizou,I. *et al.* (2013) Analysis of molecular recognition features (MoRFs) in membrane proteins. *Biochim. Biophys. Acta*, **1834**, 798–807.
Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
Lobanov,M.Y. *et al.* (2010) ComSin: database of protein structures in bound (complex) and unbound (single) states in relation to their intrinsic disorder. *Nucleic Acids Res.*, **38**, D283–D287.
Minezaki,Y. *et al.* (2007) Intrinsically disordered regions of human plasma membrane proteins preferentially occur in the cytoplasmic segment. *J. Mol. Biol.*, **368**, 902–913.
Mohan,A. *et al.* (2006) Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.*, **362**, 1043–1059.
Mooney,C. *et al.* (2012) Prediction of short linear protein binding regions. *J. Mol. Biol.*, **415**, 193–204.
Oates,M.E. *et al.* (2013) D(2)P(2): database of disordered protein predictions. *Nucleic Acids Res.*, **41**, D508–D516.
Oldfield,C.J. *et al.* (2005a) Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, **44**, 1989–2000.
Oldfield,C.J. *et al.* (2005b) Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry*, **44**, 12454–12470.
Sickmeier,M. *et al.* (2007) DisProt: the database of disordered proteins. *Nucleic Acids Res.*, **35**, D786–D793.
Stavropoulos,I. *et al.* (2012) Protein disorder and short conserved motifs in disordered regions are enriched near the cytoplasmic side of single-pass transmembrane proteins. *PLoS One*, **7**, e44389.
Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.
Tsaousis,G.N. *et al.* (2010) ExTopoDB: a database of experimentally derived topological models of transmembrane proteins. *Bioinformatics*, **26**, 2490–2492.
Uniprot_Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
Uversky,V.N. (2011) Intrinsically disordered proteins from A to Z. *Int. J. Biochem. Cell Biol.*, **43**, 1090–1103.
Uversky,V.N. *et al.* (2000) Why are ''natively unfolded'' proteins unstructured under physiologic conditions? *Proteins*, **41**, 415–427.
Vucetic,S. *et al.* (2005) DisProt: a database of protein disorder. *Bioinformatics*, **21**, 137–140.
Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.
Xue,B. *et al.* (2009) Analysis of structured and intrinsically disordered regions of transmembrane proteins. *Mol. Biosyst.*, **5**, 1688–1702.