Bioinformatics, 33(22), 2017, 3682–3684
doi: 10.1093/bioinformatics/btx486
Advance Access Publication Date: 3 August 2017
Applications Note

OXFORD

Databases and ontologies

MFIB: a repository of protein complexes with mutual folding induced by binding

Erzsébet Fichó[1], István Reményi[2], István Simon[1,*] and Bálint Mészáros[1,*]

[1]Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest H-1117, Hungary and [2]Institute of Enzymology, RCNS, Hungarian Academy of Sciences, 'Momentum' Membrane Protein Bioinformatics Research Group, Budapest H-1117, Hungary

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on March 22, 2017; revised on June 26, 2017; editorial decision on July 26, 2017; accepted on August 2, 2017

Abstract

Motivation: It is commonplace that intrinsically disordered proteins (IDPs) are involved in crucial interactions in the living cell. However, the study of protein complexes formed exclusively by IDPs is hindered by the lack of data and such analyses remain sporadic. Systematic studies benefited other types of protein–protein interactions paving a way from basic science to therapeutics; yet these efforts require reliable datasets that are currently lacking for synergistically folding complexes of IDPs.

Results: Here we present the Mutual Folding Induced by Binding (MFIB) database, the first systematic collection of complexes formed exclusively by IDPs. MFIB contains an order of magnitude more data than any dataset used in corresponding studies and offers a wide coverage of known IDP complexes in terms of flexibility, oligomeric composition and protein function from all domains of life. The included complexes are grouped using a hierarchical classification and are complemented with structural and functional annotations. MFIB is backed by a firm development team and infrastructure, and together with possible future community collaboration it will provide the cornerstone for structural and functional studies of IDP complexes.

Availability and implementation: MFIB is freely accessible at http://mfib.enzim.ttk.mta.hu/. The MFIB application is hosted by Apache web server and was implemented in PHP. To enrich querying features and to enhance backend performance a MySQL database was also created.

Contact: simon.istvan@ttk.mta.hu, meszaros.balint@ttk.mta.hu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Intrinsically disordered proteins (IDPs) do not have a stable structure under native conditions (Wright and Dyson, 1999), yet they perform crucial biological roles being deeply embedded in regulatory and signaling pathways, amongst others (Dyson and Wright, 2005; Wright and Dyson, 2015). Despite the lack of intrinsic tertiary structure of IDPs, many critical biological processes require them to interact with molecular partners, most often other proteins. During the vast majority of these interactions IDPs do adopt a stable bound structure—hence their folding is coupled to binding (Sugase *et al.*, 2007) giving rise to weak, transient, yet highly specific interactions. In accord, IDPs often represent hubs of protein–protein interaction networks (Haynes *et al.*, 2006) presenting promising therapeutic targets (Joshi and Vendruscolo, 2015).

In line with their biological importance, IDPs are heavily studied. The resulting information are collected in disorder-specific databases (such as DisProt, Piovesan *et al.*, 2016 or IDEAL, Fukuchi *et al.*, 2014) and are disseminated as various levels of annotation in core biology databases, such as UniProt (Pundir *et al.*, 2017). The majority of these information pertains to the establishment of which

protein regions are disordered and which have intrinsic structure, with some additional information about the detailed structural properties of IDPs (Varadi *et al.*, 2014). These data are in turn used to develop prediction algorithms that enable the *in silico* identification of IDP regions (Oates *et al.*, 2013) and functional sites (Dosztanyi *et al.*, 2010; Malhis *et al.*, 2016), which aids experimental verification, creating an iterative synergistic workflow.

This targeted research and synergy can be seen in the identification of IDPs; other areas of unstructural biology still lack this kind of focus. The identification of the interactions of IDPs in structural detail seems to be much more sporadic, lacking systematic targeted efforts. While no specific IDP interaction database exists, a subset of such interactions have been studied in detail (Mészáros *et al.*, 2007; Mohan *et al.*, 2006). The interaction between IDPs and ordered proteins are often mediated by short linear motifs (SLiMs) residing in the IDP partner (Fuxreiter *et al.*, 2007), and in accord, SLiM databases—such as the Eukaryotic Linear Motif database (Dinkel *et al.*, 2016)—can provide a starting point for structural studies of IDP–ordered protein interactions.

In contrast to the study of IDP–ordered protein interactions, protein complexes formed exclusively by IDPs are far less understood from both structural and functional points of view. The primary reason behind the lack of systematic research of IDP-only complexes is the lack of well-organized and accessible data. While several such complexes are known (and some have been studied in detail, see for example, Demarest *et al.*, 2002), no specific database exists, and the majority of corresponding data are scattered in various databases. Yet, a targeted database often proves to be not only beneficial, but vital for the development of research areas in biology (Baxevanis and Bateman, 2015).

Our current work lays this missing foundation of the systematic structural/functional studies of IDP complexes by assembling Mutual Folding Induced by Binding (MFIB). MFIB is constructed by integrating information from a range of databases and a wealth of literature to assemble by far the largest repository of protein complexes, where the interacting chains mutually fold as a result of the interaction.

2 Database assembly

MFIB aims to serve as a starting point for the functional and structural analysis of interactions between IDPs. In accord, the existence of a solved complex structure of the interacting protein partners was a prerequisite for inclusion in the dataset. The existence of a solved structure also serves as verification of the interaction and proof that the proteins involved in fact adopt a stable structure upon interacting. Accordingly, the PDB (version March 28, 2017) was taken as a starting point, and was filtered and annotated using various criteria and information from other databases to derive a high-quality set of interacting IDPs.

Structures that contain at least two protein chains in interaction were selected and were filtered for structure quality (keeping only nuclear magnetic resonance structures, and X-ray structures with a resolution better than 5 Å to discard poor quality structures) and biological relevance (discarding chimeras and other structures containing non-biological polypeptide chains). Complexes where non-protein chains—typically DNA and RNA—participate in the interaction were also discarded. The remaining set of candidate complexes were annotated based on experimental evidence in various annotation databases (see Fig. 1). Disorder annotations were taken from DisProt (version 7 v0.4) (Piovesan *et al.*, 2016) and IDEAL (version March 29, 2017) (Fukuchi *et al.*, 2014). Using these manually curated information, protein chains in the candidate PDB complexes were annotated using three different approaches.

First, some candidate protein chains had direct disorder annotations, meaning that they cover the same region in the corresponding UniProt protein sequence as referenced in disorder databases. Second, annotations were transferred to close homologues, considering proteins that share at least 90% sequence identity (i.e. they belong to the same UniRef90 sequence cluster). As the third level of annotations, disorder information was transferred through Pfam (release 31.0, Bateman, 2000) objects (families, domains, motifs or repeats). If a Pfam object covered at least 70% of both an interacting chain and a disorder annotation, then the disordered status was also assigned to the interacting chain.

Taking all three types of annotations (direct, UniRef90-transferred and Pfam-transferred) into account, all candidate complexes were categorized. Complexes containing only disordered chains were kept; and complexes with both disordered chains and chains without annotations were further inspected. If evidence uncovered using literature searches indicated that the unknown chains were in fact disordered, the complex was also kept. The database-based annotations coupled with information from the literature resulted in a set of 1406 complexes that all exclusively contain protein chains that are disordered in their monomeric form. Each complex is manually inspected by database curators with a focus on the validity of the experimental evidence for disorder to assure the reliability of the database. Curators also check the true biological assemblies of the complexes using PISA (Proteins, Interfaces, Structures and Assemblies) to avoid the inclusion of non-biological contacts due to crystallization. These manually curated protein complexes together comprise MFIB.

To reduce redundancy, complexes in MFIB were clustered based on sequence similarities of their constituent chains. Protein chains were considered to be similar if they belong to the same UniRef90 cluster and show at least 70% overlap. Two complexes are deemed related if they contain the same number of proteins, and the proteins from the two structures show pairwise similarity. Related complexes were grouped into clusters forming the entries in MFIB. This clustering grouped the 1406 structures into 205 MFIB entries. Furthermore, each entry in MFIB is assigned a class and a subclass during the manual annotation and curation step. Supplementary Table S1 shows the 8 classes and 33 subclasses currently defined in MFIB.

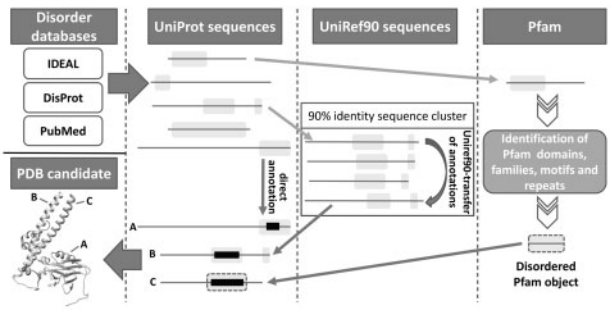

Fig. 1. Workflow of the construction of MFIB. The figure shows the annotation steps of a hypothetical example of three interacting disordered protein regions, where the three chains are annotated through direct, UniRef90-transfer and Pfam-transfer of annotations (marked A, B and C, respectively). Light grey boxes represent disordered protein regions. Smaller black boxes mark regions that are present in the candidate PDB structure. Boxes with dashed outline represent Pfam objects. Arrows show the transfer of annotations either with direct sequence comparisons (direct annotations between UniProt sequences) or with mapping (using Pfam, UniRef90 clusters, or BLAST in the case of transfer between UniRef90 sequences and between UniProt and the PDB candidate proteins)

3 Web interface

MFIB is made available through a dedicated website at http://mfib.enzim.ttk.mta.hu/. The 205 entries representing interactions of IDPs form the core of MFIB. Accordingly, each entry is assigned a unique accession and has a separate page that details information about the given complex. Furthermore, the MFIB server also includes features to ease searching and navigating through the database.

The 'Home' page describes the basis and purpose of the database for users unfamiliar with MFIB. The 'Statistics' page shows basic statistics about MFIB. The 'Help' page answers questions connected to the conception, assembly, design and usability of the database and the server. MFIB also offers several ways of structured access to the database including browsing, searching and multiple ways of downloading data in XML and text formats for local use.

4 Discussion

The construction of MFIB presents the first systematic collection of data concerning complexes formed by IDPs. It is based on the integration of structural and sequence annotation databases coupled with the results of an extensive manual literature survey. Previous studies of complexes of mutually folding IDPs were typically based on 10–35 structures (Gunasekaran *et al.*, 2004; Nussinov *et al.*, 1998; Rumfeldt *et al.*, 2008). In contrast, MFIB contains over 1400 complex structures organized into 205 entries. These data provide the missing cornerstone of future structural and functional studies of the synergistic folding of IDPs.

The data contained in MFIB not only far surpasses the number of complexes used in previous analyses but also provides a wide coverage of possible IDP–IDP interactions in many ways. Entries in MFIB cover all three domains of life and also include complexes from viral proteins shedding light on the importance of synergistic folding in host–pathogen interactions. MFIB entries also cover the majority of possible oligomeric compositions from dimers to hexamers, including both hetero- and homo-oligomers. Most importantly, entries in MFIB also cover the known spectrum of protein disorder. Protein disorder is a highly heterogeneous property with various IDPs exhibiting markedly different levels of flexibility in their unbound form. MFIB contains complexes of IDP regions from near random coil proteins (such as the CBP (CREB Binding Protein)-interacting region of ACTR, Demarest *et al.*, 2002), through molten globules (such as the Arc repressor, Peng *et al.*, 1993) to near-ordered structures, where a monomeric structure can be stabilized with a limited number of mutations (such as the nucleoside diphosphate kinase, Giartosio *et al.*, 1996).

The presented MFIB database currently presents the far largest collection of interactions between IDPs; yet there are undoubtedly many more information scattered in the PDB and the literature that are not currently incorporated. In accord, we consider the present version of MFIB as a stepping stone and plan to constantly update, expand and revise the database. This process will rely on the past experience of the authors in database-maintenance, the firm technical and infrastructural background of the initiative, and the encouragement of a community effort to contribute to MFIB.

Acknowledgements

The authors would like to thank Zsófia Béky for her help in the MFIB graphical design and Gábor E. Tusnády for his help with setting up the MFIB server. The critical comments of Katalin Paréj, László Dobson and Karolina Fichó concerning server functionality are greatly appreciated.

Funding

This is work was supported by the postdoctoral fellowship of the Hungarian Academy of Sciences, the Hungarian Research and Developments Fund [OTKA K115698 and OTKA K104586], and the Momentum Grant of the Hungarian Academy of Sciences [LP2012-35]. Project no. FIEK_16-1-2016-0005 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the FIEK_16 funding scheme.

Conflict of Interest: none declared.

References

Bateman,A. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

Baxevanis,A.D. and Bateman,A. (2015) The importance of biological databases in biological discovery. *Curr. Protoc. Bioinformatics*, **50**, 1–8, Unit 1.1.

Demarest,S.J. *et al.* (2002) Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature*, **415**, 549–553.

Dinkel,H. *et al.* (2016) ELM 2016–data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.*, **44**, D294–D300.

Dosztanyi,Z. *et al.* (2010) Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief. Bioinform.*, **11**, 225–243.

Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **6**, 197–208.

Fukuchi,S. *et al.* (2014) IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res.*, **42**, D320–D325.

Fuxreiter,M. *et al.* (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, **23**, 950–956.

Giartosio,A. *et al.* (1996) Thermal stability of hexameric and tetrameric nucleoside diphosphate kinases. Effect of subunit interaction. *J. Biol. Chem.*, **271**, 17845–17851.

Gunasekaran,K. *et al.* (2004) Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J. Mol. Biol.*, **341**, 1327–1341.

Haynes,C. *et al.* (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput. Biol.*, **2**, e100.

Joshi,P. and Vendruscolo,M. (2015) Druggability of intrinsically disordered proteins. *Adv. Exp. Med. Biol.*, **870**, 383–400.

Malhis,N. *et al.* (2016) MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. *Nucleic Acids Res.*, **44**, W488–W493.

Mészáros,B. *et al.* (2007) Molecular principles of the interactions of disordered proteins. *J. Mol. Biol.*, **372**, 549–561.

Mohan,A. *et al.* (2006) Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.*, **362**, 1043–1059.

Nussinov,R. *et al.* (1998) Mechanism and evolution of protein dimerization. *Protein Sci.*, **7**, 533–544.

Oates,M.E. *et al.* (2013) D^2P^2: database of disordered protein predictions. *Nucleic Acids Res.*, **41**, D508–D516.

Peng,X. *et al.* (1993) Molten-globule conformation of Arc repressor monomers determined by high-pressure 1H NMR spectroscopy. *Proc. Natl. Acad. Sci., USA*, **90**, 1776–1780.

Piovesan,D. *et al.* (2016) DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.*, **45**, D219–D227.

Pundir,S. *et al.* (2017) UniProt protein knowledgebase. *Methods Mol. Biol.*, **1558**, 41–55.

Rumfeldt,J.A.O. *et al.* (2008) Conformational stability and folding mechanisms of dimeric proteins. *Prog. Biophys. Mol. Biol.*, **98**, 61–84.

Sugase,K. *et al.* (2007) Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature*, **447**, 1021–1025.

Varadi,M. *et al.* (2014) pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.*, **42**, D326–D335.

Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.

Wright,P.E. and Dyson,H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.*, **16**, 18–29.