ELIXIR Data Platform Implementation Study 2018 PROPOSAL

Application made in response to the Request for Proposals published 4th August 2017.

Proposal length not to exceed total 5 pages

Integration and standardization of intrinsically disordered protein data

Partners:

Node	Name of PI	Role	Person Months
IT	Silvio Tosatto	Lead	5.5
EBI	Rob Finn (InterPro)	Member	1.5
EBI	Maria Martin (UniProt)	Member	1.5
EBI	Sandra Orchard (UniProt, IntAct)	Member	1.5
EBI	Sameer Velankar (PDBe)	Member	1.5
CH	Alan Bridge (UniProt)	Member	1.5
HU	Zsuzsanna Dosztanyi	Member	2.5
IR	Norman Davey	Co-lead	2.5
		Total	18
Work period	May 2018 - April 2019		

Abstract:

Intrinsically disordered proteins (IDPs), characterized by high conformational variability, cover almost a third of the residues in Eukaryotic proteomes. As major players in cellular regulation, IDPs are involved in numerous diseases. Specialized IDP databases provide a starting point for analysis, yet their integration into core databases remains very limited. Here, we propose to start integrating IDP information into ELIXIR core data resources. This will be achieved with a three pronged approach: (1) Integration of an expanded version of MobiDB-lite into UniProtKB and PDBe including links back to MobiDB via the InterPro infrastructure. (2) Creation of a minimal information about IDP experiments (MIADE) standard and data interchange format will help ensure interoperability between databases with manually curated IDP data. (3) Sustainability of curated IDP data will be enhanced by creating PubDisProt, a new deposition resource for linking curated literature and protein identifiers across databases.

Planned Work:

Background

This study on intrinsically disordered proteins (IDPs) is based on previous work from COST Action NGP-net, a community spanning 30 different countries, plus EMBL Heidelberg and EMBL-EBI. Several ELIXIR nodes (e.g. Italy, Hungary, Ireland) have also included IDP-related resources in their national node roadmap, leading to the recent founding of the international DisProtCentral umbrella consortium.

Over the last two decades, IDPs have developed from being bespoke projects of biophysicists interested in protein (non-)folding to being recognized as a major determinant in cellular regulation (Guharoy, 2015; Chouard, 2011). It is estimated that almost half of the human proteome is made up of residues mostly encoding for IDPs and related phenomena (Perdigão, 2015; Mistry, 2013). Similarly, IDPs are central for protein homeostasis, cellular signaling and implicated in many human diseases. One of the key problems with IDPs, and the main limitation why their existence is not yet translated into the ELIXIR core data resources (e.g. UniProt, PDBe), was the lack of a clear definition of the phenomenon (Dunker, 2001; Uversky, 2002; Wright, 1999). In a way, the concept of intrinsic disorder is itself disordered, as different authors have used it to mean somewhat different things (Orosz, 2011). One key result of the NGP-net has been a comprehensive definition of IDPs. The NGP-net community maintains several major databases describing the structure and function of IDPs. DisProt is a database of manually curated IDPs, established over a decade ago in the USA (Sickmeier, 2007), and recently brought to Europe after years of inactivity and completely re-annotated by NGP-net (Piovesan, 2017). Its sister database, MobiDB (Potenza, 2015), is the central resource for automatic structural prediction of IDPs which has recently joined the InterPro consortium (Finn, 2017). The recently developed DIBS (Schad, 2017) and MFIB (Fichó, 2017) databases provide manually curated examples of binding regions residing in IDPs, mediating interactions with ordered and disordered partner proteins, respectively. All these databases describing IDP function are maintained by DisProtCentral members.

NGP-net has been running a series of thematic workshops on IDPs since 2016 to drive the development of computational resources and community standards. A strategic workshop was organized by NGP-net on June 1-2 2017 at the EBI in Hinxton to discuss the integration of IDP-related computational resources into the ELIXIR framework. A major outcome of the strategic workshop was the realization that IDPs are significantly underrepresented in the Core Data Resources (CDRs) (Figure 1). However, as part of its interaction with NGP-net, InterPro adopted IDP annotation from sequence in late 2016. This could be used as an example for the successful integration of IDP data into additional CDRs, by providing basic annotation for IDPs to be propagated to other databases.

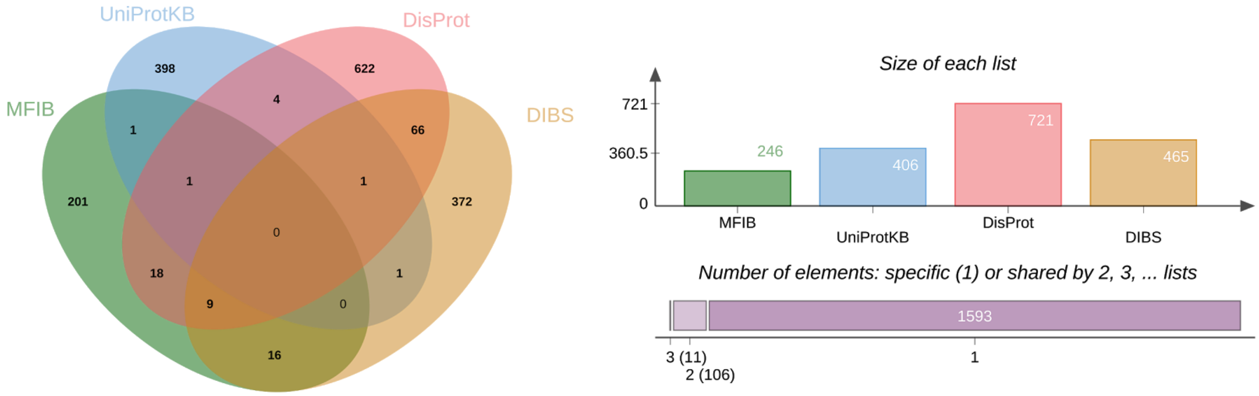

Figure 1. *Comparison between manually curated IDP datasets from different databases (UniProtKB, DisProt, DIBS, MFIB). The regions being annotated are similar, but not identical (data not shown). Notice the total lack of overlap, with only 117 proteins (out of 1,710) annotated in at least two databases.*

The development of curation guidelines and standards in line with the requirements of the CDRs will streamline the integration process. Annotation of bona fide IDPs, especially for their function, is a labour-intensive process run by the NGP-net community. A key requirement of the IDP community is the definition of curation guidelines and standards to improve the reproducibility, interpretation, and dissemination of experimental data. These developments will result in Minimum Information About Disorder Experiments (MIADE) standards focusing on promoting effortless integration into the CDRs. For structural data, in the simplest cases, experimental evidence for disordered regions in IDPs can be mapped to protein sequences as sequence features. However, structural descriptions of protein ensembles (e.g. based on NMR or SAXS data) will require more complex standards. Molecular interaction interchange standards based on the guidelines of the HUPO-PSI-MI molecular standards can be reused to describe IDP interaction data. IDPs will require some expansion to controlled vocabularies describing protein structure to include concepts such as conditional disorder, where structural changes regulate functions. These will drive the *FAIRification* of IDP databases and result in the dissemination of IDP data to the larger biological user community.

Description of work

The ultimate goal of the implementation study proposal is to drive the inclusion of IDP-related annotation into the CDRs UniProt and PDBe, engaging actively with ELIXIR while guaranteeing long-term sustainability and impact for IDP community resources. The first part of the project is about the propagation of predicted disorder annotation into the CDRs through InterPro and the MobiDB-lite software package (thereby ensuring the calculations are performed once and disseminated to many CDRs). This study also aims to develop a Minimal Information About Disorder Experiments (MIADE) format to represent disorder evidences unambiguously and the definition of APIs for exchanging this information across different databases. Finally, we propose a new deposition resource (PubDisProt) linking literature PubMed IDs with the corresponding identifiers in curated IDP databases to improve curation workload and favor annotation exchange across different databases.

Task 1 *- Integration of IDP annotation into UniProtKB and PDBe by propagating InterPro data*

MobiDB-lite provides disorder prediction by combining eight different methods and applying a post-processing in order to filter out non functional short regions. MobiDB-lite has been recently incorporated into InterProScan software (Finn, 2017) for large-scale annotation. The output from this software is made available for all UniProtKB proteins through the InterPro and MobiDB web interfaces. However, the former lacks associated functional annotations, *e.g.* the IDP is a recognition site or the conformational preference (swollen coil, globule, etc.). In task 1 MobiDB-lite will be updated by incorporating ANCHOR software (maintained by HU) (Dosztányi, 2009) for prediction of binding sites and a structural classification algorithm based on the pattern of charged residues (Das and Pappu, 2013). The task involves both the IT and HU nodes as MobiDB-lite maintainers and the EMBL-EBI for the definition of the new output format. While MobiDB-lite is already incorporated within InterProScan, the existing implementation needs to be reorganised to incorporate the additional annotations and structured to follow appropriately into UniProt, enabling the annotations to appear in UniProt records (The UniProt Consortium, 2016). Thereafter, these annotations will be propagated to the SIFTS (Velankar, 2013), to enable the propagation between UniProtKB and PDBe sequences. In all three CDR sites (InterPro, UniProt, PDBe), websites will be updated accordingly and made available using the new protein Bioschema profiles. Ensuring that these protein CDRs contain IDPs will ensure broad dissemination to other resources that use either protein sequences or structures. IntAct curation guidelines will be updated to inform curators on the procedures for adding MobiDB-lite cross-references where a protein binding regions maps to a region of disorder, as identified via InterPro.

Task 2 *- Definition of MIADE (Minimal Information About Disorder Experiments) standard*

Different databases provide information about different aspects of disorder. For example DisProt enriches disorder annotations including functional aspects like disorder-to-order transitions upon binding and/or ambiguities in the experimental evidences whereas, DIBS reports experimental data such as the dissociation constant (K_d) for interacting disordered regions. Agreement on which are the minimal fundamental parameters to describe disorder and a format to exchange these data are still missing. A tentative standardization process has been started by the NGP-Net community, however, further community input (including a dedicated workshop) is required to finalise this work. Task 2.a will be to produce a first version of an interchange format that will enable the exchange and merging of this data, and to generate a guidelines document (MIADE, Minimal Information About Disorder Experiments). A controlled vocabulary, or addition of required terms to an existing resource such as the Sequence Ontology will provide boundaries for disorder descriptors. All of these will be submitted for review to the HUPO-PSI formal document review process (Vizcaíno, 2007). This group will also work with the Molecular Interaction workgroup of the PSI (PSI-MI) to ensure that the current PSI-MI interchange formats and IMEx Consortium curation guidelines are appropriate for the exchange of interaction data pertaining to IDPs. Task 2.b will be the adoption and implementation of the standards by the community databases (DisProt, DIBS and MFIB). The adoption of standard formats will boost the integration of manually curated annotations into CDRs.

Task 3 *- Establishing the PubDisProt data repository*

Intrinsic disorder is often described in the literature using different ambiguous names. It is very difficult to retrieve relevant papers querying PubMed and an effective automatic text-mining method calibrated for disorder does not yet exist. Task 3 will be the implementation of a new resource, PubDisProt, for the collection of manually curated literature about disorder. The database will provide links between curated papers and the corresponding annotated entries available in member databases (UniProtKB/Swiss-Prot, DisProt, DIBS and MFIB). In task 3.a the database schema and server will be implemented starting from the minimal format (PMID, Database name and entry). Task 3.b will establish an automatic update system to populate the database. UniProtKB/Swiss-Prot and DisProt will be queried periodically by exploiting the existing APIs. DIBS and MFIB, instead, will be connected through a minimal API implemented ad hoc by the HU node. PubDisProt will serve as a reference literature corpus for curators, improving sustainability of the member databases. It will provide a body for text-mining software training and facilitate the generation benchmarking experiments for CASP-like experiments.

Alignment with Evaluation Criteria

Scientific focus, scope, need: This study aims to integrate a novel feature, IDPs, still largely missing from CDRs. IDPs are a hot topic in protein science, affecting about 1/3 of proteins in higher organisms.
Community served: The proposal serves both the community interested in IDPs and all UniProt users who may discover that their protein of interest is an IDP.
Quality of service: Several participating resources are CDRs, e.g. UniProt, PDBe, InterPro. The IDP-specific resources are on the national node roadmaps.
Towards supporting the mission of the ELIXIR Data Platform and sustainability and impact of the implementation: The study increases the sustainability of IDP resources by providing new exchange formats and a common repository to avoid work duplication. Impact will be achieved by propagating novel IDP information into CDRs.

Funding statement

The proposed work is not being supported by alternative funding sources.

Expected outcome

There are three main outcomes associated to the implementation study, corresponding to the three main tasks:

1. **Integration**. MobiDB-lite data will be made available in InterPro, UniProtKB and PDBe as additional annotation for all entries.
2. **Standards.** A report describing the MIADE standard and interchange format will be published and submitted to HUPO-PSI, allowing interested parties to share IDP data.
3. **Improved sustainability**. The PubDisProt repository will be made available for interested curators and developers (e.g. of text mining methods) to exchange curated IDP papers.

Dissemination plans

The implementation study will be disseminated by dedicated communications at the ELIXIR All Hands meeting as well as the HUPO PSI workshop as well as through publications describing the work.

Project Timeline:

Task		1	2	3	4	5	6	7	8	9	10	11	12	PM	Participants
MobiDB-lite update	1.a	▨ M1	▨											1.5	IT, HU
Integration InterPro	1.b		▨	▨ M2	D1									1.5	EBI
Integration UniProt	1.c				▨	▨ D2								1.5	EBI
Integration PDBe, SIFTS	1.d					▨	▨	D3						1.5	EBI
Interchange format definition (MIADE)	2.a		▨	▨	▨	▨	▨	▨ M3	▨	D4				6.5	IR, EBI, SIB, HU, IT
MIADE implemented DisProt, DIBS, MFIB	2.b								▨	▨	D5			2	IT, HU
Implementation of PubDisProt database	3.a										▨	▨ M11		2	IT
Populating PubDisProt	3.b											▨	▨ D6	1.5	IT, HU
Total														18	

Delivery Schedule - by Node:

Milestones:

1. Define MobiDB-lite output format (M1, IT)
2. MobiDB-lite integrated in InterProScan (M3, EBI)
3. MIADE interchange format defined (M7, IR)
4. PubDisProt automatic update system deployed (M11, IT)

Deliverables:

1. Release of InterPro with updated version of MobiDB-lite (M4, EBI)
2. MobiDB-lite integrated in UniProtKB (M6, EBI)
3. MobiDB-lite integrated in PDBe (M7, EBI)
4. Report on MIADE standard and format. Submission to HUPO-PSI (M9, IR)
5. MIADE demonstrator implemented in DisProt and DIBS (M10, IT)
6. PubDisProt web server deployed and populated with data from DisProt, UniProtKB/Swiss-Prot, DIBS and MFIB (M12, IT)