Progress Report – RISE

1. General Progress of the action

1.1 Please indicate the progress of the action during the period covered by this report:

○ The action has fully achieved its objectives for the period.

● The action has achieved most of its objectives for the period with relatively minor deviations.

○ The action has achieved some of its objectives but corrective action is required.

○ The action has failed to achieve critical objectives and/or is severely delayed.

1.2 Please describe the general scientific progress of the action during the period covered by this report (including by giving qualitative indicators and by describing deliverables and milestones achieved):

The main Action aim is to investigate the topic of intrinsically Disordered Proteins (IDPs), with particular emphasis on the elucidation of their functions in human health and disease. IDPs are characterized by high conformational variability and interaction promiscuity, defying the classic protein structure-function paradigm. IDPs cover almost half of the residues in eukaryotic proteomes.
One of the biggest challenges in this field is to efficiently retrieve information from published articles and subsequently to organize information in a structured and clear way since no strict definition of what is intrinsic disorder and what is not either still exist.
Work performed in the first 12 months of the IDPfun MSCA RISE Action focused on 3 main topics:

- definition of a software package for the automatic extraction of PED entry data from protein ensembles (Deliverable 1.1)
- definition of a software for automatic detection of IDRs and linear motifs from protein structures (Deliverable 1.2)
- definition of a software tool for the automatic extraction of IDR relevant information from literature (Deliverable 1.3)

Definition of a software package for the automatic extraction of PED entry data from protein ensembles - D1.1

Source code is available at: https://github.com/naikymen/PED

IDPfun ambition is to extend the knowledge on IDP, moving from available state of the art computational tools and databases to new levels of IDP characterization. Novel functional characterization of IDPs relies on conformational ensemble data. The Lead Beneficiary (VUB) developed and hosted the Protein Ensemble Database (PED), thus, is involved in multiple tasks related to structural ensembles. IDPfun D1.1 of aims to develop an automated PED pipeline for the automated extraction and processing of the data from the uploaders' entry submissions.

D1.1 relies on “clean” data. Initial efforts tackled the format issues, data inconsistencies, missing pieces of data and unknown unique features of data files. This enabled the identification of new quality control checks besides anticipated data processing tasks. The second domain of D1.1 involved automation of information extraction and visualization of different flavors of protein ensembles that help understand the data better and help the interpretation.

Until now any structural representation of proteins is based on the protein existing in one or a few different and defined states. The PED Database and Pipeline are the first attempts at tackling the problem of IDPs not being easily represented by a single structural state. Having a representation of the ensemble of states an IDP can exist in will render definition and classification of IDP much easier.

Definition of a software for automatic detection of IDRs and linear motifs from protein structures - D1.2

The project source code is available at: https://gitlab.com/idpfun/rise-find-interfaces/

Intrinsically disordered regions contain numerous functional modules that mediate key cellular functions. These function modules are known as Intrinsically Disordered Domains (IDDs) and Short Linear Motifs (SLiMs). Many interactions mediated by the IDDs and SLiMs within the disordered regions of proteomes have been captured in complex with their binding partners. These interactions generally involve linear interactions of the disordered regions with a pocket or pockets on a globular binding partner.

These disorder-globule complexes have two major properties that can be used to identify them from other classes of interaction interface: (i) The disordered binding partner will have a high ratio of intramolecular interactions to intermolecular interactions; (ii) The globular binding partner will have high levels of intramolecular interactions. The tool uses these properties to parse, classify and annotate PDB structures based on the structural properties of the interactors. The tool takes as an input a PDB identifier and outputs the disordered interface-containing protein, disordered interface start and end, disordered interface-binding protein, disordered interface-binding domain. The tool also provides detailed motif interface information including atomic resolution details of the contacts within the interface to generate a binding consensus.

The tool was benchmarked on the ELM database using a collection of 448 manually curated ELM-containing PDB structures. The tools collections ran without issue on the dataset. The results were as follows:

- 59 structures did not return a motif:

    - 39 structures returned an did not return any interface information. In all cases the output of the tools was correct. 3 Entry is marked as obsolete and 36 contained only a single entity in the structure and were therefore not a complex.
    - 20 structures found the interface but incorrectly classified the in interfaces class as containing no motif. In these cases, the motif-containing chain have a high number of intramolecular contact and fall below the threshold for recognition of a disordered binding partner in a structure.
- 389 structures returned a motifs
    - In total, there are 582 (Motif, ELM instance) pairs to compare (Most of them are 1 to 1)
    - 510 returned the correct ELM instance.
    - 61 returned the correct ELM instance but the protein mapping was different to ELM.
    - 11 Motif region do not overlaps with ELM instance.

One of the most important features of IDPs is their ability to interact with a wide array of partners. Recognition of interactors is often mediated by SLiMs, usually found inside IDPs. Being able to predict SLiMs is very important to infer IDPs functions. This deliverable provides the first method to identify SLiMs from protein atomic coordinates and will help to further clarify function for these proteins.

Definition of a software tool for the automatic extraction of IDR relevant information from literature - D1.3

The project source code is available at: https://gitlab.com/idpfun/annotator-helper

The goal of this IDPfun task is to employ text-mining to retrieve and standardize relevant information from scientific publications. This can be of particular interest for annotators and curators of biological databases, who may struggle to find publications of interest within PubMed. The task is focused on the curation of protein-protein interactions (PPI) mediated by Short Linear Motifs (SLiMs).

Text-mining and classification *(Curation Helper - Classifier tool)*

The *Curation Helper - Classifier tool* allows a user to take a set of labeled (classified) publications and use them to predict the class of unlabeled (unclassified) publications. *Curation Helper - Classifier tool* has 5 major parts:

The user provides a set to publications to create article classifiers. The training set is given as a standalone tab-separated file of publications (defined by PubMed Identifiers) - classification pairs. The tool gathers the data required to train the machine learning classifier from PubMed and UniProt REST services and stores the data locally in XML format. The default information used as training data for the article classifiers is the article title and abstract. This data can be augmented using MESH terms, SciLite data or UniProt-derived gene and protein names.

The training data from each publication is converted to a matrix of TF-IDF (term-frequency times inverse-document-frequency) word occurrences. The TF-IDF matrix encodes the enrichment of a given term in a given document relative to the whole set of documents. The TF-IDF vectors of each article are used to create a classifier based on the publication - classification pairs in the input training set.

The goal of the classifier is to correctly identify the correct class of an unseen articles. *Curation Helper - Classifier tool* accepts articles for classification as either one or a list of PubMed IDs. The article information is retrieved and processed as described for the construction of the classifier.

The *Curation Helper - Classifier tool* software can be easily integrated into curation interfaces like those provided by DisProt and SwissProt/UniProtKB databases. This will improve the selection of articles relevant to disorder and guide the curation process. Moreover, the method is applicable to any text-mining axis, i.e. to aspects different from disorder structural aspects. Particularly relevant will be ranking of the literature about functions discovered very recently like phase separation and droplet formation.

Papers published in the context of the IDPfun Action

A total of 3 papers have been published during the 1st year of the IDPfun Action. All of them resulted from spin-offs or collaborations raised in the context of IDPfun secondments.

All the publications have been reported in the Funding & Tenders portal.

Necci, M., Piovesan, D. & Tosatto, S.C.E., 2018. Where differences resemble: sequence-feature analysis in curated databases of intrinsically disordered proteins. Database: the journal of biological databases and curation, 2018. Available at: http://dx.doi.org/10.1093/database/bay127.

Mier, P. et al., 2019. Disentangling the complexity of low complexity proteins. Briefings in Bioinformatics. Available at: http://dx.doi.org/10.1093/bib/bbz007.

Marchetti, J. et al., 2019. Ensembles from Ordered and Disordered Proteins Reveal Similar Structural Constraints during Evolution. Journal of Molecular Biology, 431(6), pp.1298–1307.

2. Corrective Measures

2.1 Please explain any delays accumulated in the secondments / activities / deliverables foreseen in the Grant Agreement and the measures taken to oversee them.

Secondment delays

At the time of this report, 26 secondments have been concluded or are on duty, out of the 30 originally planned (87% of the original plan).

The remaining 4 secondments have been proposed to be postponed to the first part of the 2nd IDPfun year.

Physiological variations (few days up to one month) on secondment starting dates have been communicated to the Project Officer before the submission of this report. None of these changes affected in any way the scientific aims of the project and have been authorized to accommodate secondees familiar duties and flight fares.

Some discrepancies with respect to the original secondment plan have been noticed in the secondment schedule, very likely introduced during the negotiation phase. Most of these discrepancies affect Work Packages assigned to secondments. In agreement with the the IDPfun Executive Board, it has been decided to revert the WPs to the original plan in order to meet Deliverable/Milestone deadlines.

These discrepancies have been sent to the Project Officer before the submission of this report and reverting the WPs to the original plan does not produce any negative effect on the IDPfun scientific program, conversely the reversion better allows the achievement of project deadlines on time.

Deliverable/Milestone delays

All deliverables due in the first year of the Action have been submitted on time.

Milestone 1 - IDR detection from literature - EMBL (due date 1/03/2019 - postponed)

IDPfun Milestone 1 has not been submitted so far (due date 1/03/2019). All technical work is completed and the software deliverables are available for download from the IDPfun repository. The manuscript is currently being written and will be finalised for submission by the end of April with a likely publication date in summer. The nature of the milestone (scientific publication) does not allow full control on the delivery date, as it hinges on the peer-review and publication process which, by definition, has no fixed duration and depends on many factors. For this reason, the dynamics of paper publication could clash with the hard deadlines for milestone submissions also in the future. Project Officer has been informed about this delay before the deadline expired.

This risk has been acknowledged by the IDPfun Executive Board and the consortium will strive to submit the future papers linked to milestones well ahead of the deadline.

2.2 Please indicate any potential risks identified and suggested approaches to mitigate them.

The Davey group at University College Dublin (UCD) has moved to the Institute of Cancer Research, London, in the United Kingdom. Given the current confusing situation surrounding Brexit and the time sensitive nature of the organisation of 2nd year secondments, the IDPFun grant will remain at University College Dublin for 2019. Professor Denis Shields will become the lead PI. It has been agreed that Dr. Norman Davey will be invited to join the yearly IDPfun all-hands meeting in Europe regardless of the Brexit status in order to keep him involved. We will revisit the situation before the start of year three.

Considering the remaining 3 years of the action, 8 secondments from Partner Countries to UCD and 4 from UCD to Partner Countries would potentially be affected. For this reason, we will consider later in this year whether we would need to add a new consortium member to compensate for the possible problems caused by Brexit.

3. Ethical Issues

Please indicate how the ethical issues have been addressed during the period covered by this report and mention all the approvals/authorisations already provided to the REA (if applicable).

Not applicable.

4. Additional information

Please indicate any additional information which you may consider useful to assess the project implementation during the period covered by this report, including management issues.

The consortium agreement was finalised in January 2019. This lengthy process has been mostly caused by the lack of responsiveness from the Hungarian partner (ELTE), apparently due to some internal dynamics outside control of the consortium. A partnership agreement is being prepared and will be signed by the Argentinean partners in the current year.

In general, it should be noted that the Coordinator had to overcome the initial inertia of several partners regarding the non-scientific tasks of the project (e.g. meeting participation can be improved). This situation is being addressed by holding monthly Executive Board teleconferences, which are helping to overcome the problem. Communication, both internal and external, is also being strengthened by hiring a dedicated manager working closely with the Coordinator. It is envisaged that these measures should be sufficient to improve the responsiveness of all partners in the current year.