IDP *f*(un)

Project Acronym:	**IDPfun**
Project Full Title:	**Driving functional characterization of intrinsically disordered proteins**
Grant Agreement:	**778247**
Project Duration:	**48 months (01/03/2018 - 28/02/2022)**

Deliverable D1.1

Software package for the automatic extraction of PED entry data from protein ensembles

Work Package:	**WP1 – IDP function detection / WP1.1 – PED pipeline**
Lead Beneficiary:	**VUB**
Due Date:	**30 Nov 2018 (M9)**
Submission Date:	**30 Nov 2018 (M9)**
Deliverable Type:	**Demonstrator**
Dissemination Level:	**Public**

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 778247

Table of Content

Executive summary

This document describes the Deliverable 1.1 (D1.1) for the work package 1 (WP1) of the IDPfun project. The report provides an overview of the algorithms and program scripts developed that make up a software package for the automatic extraction and processing of the Protein Ensemble Database (PED) entry data from protein ensembles. This package will be used throughout the WP1 of the IDPfun project by participating project partners, with regard to the existing processed data and also the unprocessed data. These data will be revised, cleaned and fully processed during the duration of WP1.1 in order to provide clear information for the users of the database.

List of Acronyms

Acronym	Definition
CA, CB	Carbon alpha, carbon beta
Dmax	Maximum dimension
ESR	Early stage researcher
IDP	Intrinsically disordered protein
IDR	Intrinsically disordered region
PDB	Protein Data Bank
PED	Protein Ensemble Database
Rg	Radius of gyration
UNIPD	Universita degli Studi di Padova
UNQ	Universidad Nacional de Quilmes
UNSAM	Universidad Nacional de San Martin
VUB	Vrije Universiteit Brussel
WP	Work package

Project overview

Introduction on the aims

The main purpose of this document is to provide relevant information concerning data and software developed within the framework of D1.1 during the implementation phase of WP1 of the IDPfun action.

The main project 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. Growing evidence suggests that IDPs, interacting with multiple partners, are major players in cellular regulation and involved in numerous human diseases. However, functional knowledge for IDPs remains very limited.

The partner’s ambition, in the context of the IDPfun action, is to extend the knowledge on IDP functions, 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 WP1.1 of aims to develop an automated PED pipeline for the automated extraction and processing of the data (D1.1) from the uploaders' entry submissions.

Identification of tasks relevant for the deliverable

IDPfun action is entirely an *in silico* project. As all data science projects, WP1.1 also relies on “clean” data. Thus, initial efforts in this WP 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 third 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.

Description of results

1. Data exploration, revision and curation

1.1. Manual examination and curation of data

a) Downloading all available data

b) Discovery of non-standard features in the data files and minor inconsistencies (e.g. format, naming, numbering)

c) Identification of missing metadata

1.2. **Revision and interpretation of existing tools and program scripts** used for building the earlier version of the database

1.3. Identification and specification of tasks related to the automation of data processing and interpretation

2. Automation of data processing

2.1. **Pipeline** (pipe.py / pedbcall / splitPDB & pdb): splits entries to ensembles and enumerates their conformers, performs Rg and Dmax calculation by CRYSOL (pipe.py / pedbcall / pdb / pipe1_crysol → crysol)

2.2. Quality control checks

a) Sequence compliance (pipe1_json.py / extract_uniprot_sequence & extract_seq_from_pdb): Check if the UniProt region matches the sequence given in the PDB files.

b) PDB format checks (qc.py / qcall / qcheck / pdbcheck): no unknown ('UNK') residues, no missing chains, removal of NMR dummies (Q atoms) and correction of hydrogen naming (Ensembles-Format-to-be-modified/doit_Format_PDB)

c) Stereochemistry check with MolProbity (qc.py / qcall / qcheck / molcheck): reporting steric clashes, dihedral angle outliers, CB deviations and backbone deformations.

3. Automation of data interpretation

3.1. **CA-CA distance distribution** of ensemble conformers (pipe.py/ pedbcall / pdb / n2n)

3.2. **End-to-end distance distributions** for all chains in the ensembles (pipe1_json.py / end_to_end_distance & end_to_end_processing)

3.3. **Visualization of Rg distribution** of ensembles (pipe.py / pedbcall / pdb → Pipe245.R)

3.4. **Secondary structure assignment and propensity visualization** using DISICL assignments (pipe2_disicl_json.py & data_discl_plot.r)

3.5. **Conformer visualization** in PyMol (pipe.py / pedbcall / pdb / buildPymolCalls; / pedbcall / pdb → Pipe5.2.py)

Availability

All scripts, codes, documentation, list of dependencies and other elements of the software package are available in Github: https://github.com/naikymen/PED

External backup of the software package and data is also stored in a local computer in the office of the Lead Beneficiary of D1.1 (VUB).

Self-evaluation and concluding remarks

The three early stage researchers (ESRs) from UNQ and UNSAM (Tadeo E. Saldeno, Julia Marchetti and Nicolas A. Mendez) completed the tasks of D1.1 with joint efforts and developed a software package that will serve as a major component of the new version of the Protein Ensemble Database (PED). Initial testing of the software package was performed by researchers from UNQ, UNSAM, VUB and UNIPD but further testing will be done for the new version of the database that might require refinements and improvements on the software – this will be done by past and new seconded researchers from Argentina.

Future perspectives concerning the continuation of improvements on PED will involve more sustainable front- and back-end, improved user interface with a more comprehensive view on the IDP ensembles stored, furthermore an informative quality report will be made available for each entry. The new database tools will enable the discovery of molecular transitions (D3.1) and different types of IDRs (D3.2); moreover, analysis of fuzzy ensemble complexes (D3.6) will also become much easier, as the researchers will have cleaned and regularized data with a quality report.