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 D1.4

A software tool for extending IDP annotations via homology transfer through sequence and structure

Work Package:	**WP1**
Lead Beneficiary:	**EMBL**
Due Date:	**31 January 2020**
Submission Date:	**31 January 2020**
Deliverable Type:	**D**
Dissemination Level:	**P**

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 *"D1.4 A software tool for extending IDP annotations via homology transfer through sequence and structure"* for the IDPfun project. The deliverable was a Software for the identification of homologous IDR clusters from sequence databases. The pipeline was completed on time and as described.

List of Acronyms

Acronym	Definition
IDP	Intrinsically Disordered Protein
IDR	Intrinsically Disordered Region

Project overview

Introduction on the aims of the project

Generating structured annotations for disordered proteins (including the fact that a certain region lacks structure, the experimental method used to ascertain this property, and possible functional annotations) is a labour-intensive process involving manual curation from literature. As the fact that a certain protein region is disordered is highly valuable in this sense, approaches that are able to generate this knowledge in an automated way are of high interest to the community.

It has been shown for ordered, globular proteins that a sufficiently high sequence identity/similarity between two sequences means a very high probability of the two proteins adopting the same fold (Rost 1999). Generalizing this idea, it is plausible that a sufficiently high sequence identity between two proteins should indicate the same structural state (ordered or disordered), i.e. if two proteins have highly similar sequences, and one is disordered, the other one is likely to be disordered as well. This can be the basis of extending disorder annotations from known proteins where disorder has been documented to proteins with similar sequences and unknown structural state. Further credibility of such an approach can be achieved by considering the similarity between homologous proteins only (homology transfer), as the disordered character of proteins is an evolutionarily conserved feature (Chen et al. 2006).

While this concept seems very straight-forward, the implementation has several hidrances at the technical level. There are currently very few available alignment algorithms and substitution matrices tailored for the specific sequence characteristics of IDPs (Szalkowski and Anisimova 2011). As these sequences are often more difficult to align due to their overall higher substitution rates (Gitlin et al. 2014; Brown et al. 2011) and occurence of insertions and deletions (Brown et al. 2011) (Light et al. 2013), recognizing similarity is highly non-trivial and requires the development of a specific bioinformatics pipeline. This project is focused on homology transfer of Intrinsically Disordered Regions (IDR). To make the approach as widely applicable as possible, we aim to transfer disorder annotations using both sequence and structural data, to recognize currently non-characterized putative IDPs from both sequence (UniProt) and structural (PDB) databases.

Overview of the approach

The presented implementation of the IDP-homology transfer approach has three distinct steps.

1. First, the algorithm identifies an IDP or IDR from the DisProt database that will serve as the source of the annotations to be transfered. This can either be specified by the user as an input, or will be automatically identified by the algorithm using precomputed alignments within orthologs groups. As annotations are attached to specific IDRs in DisProt, in case a full IDP is used, the algorithm will treat all annotations attached to any IDRs inside the IDP as transferable.
2. Next, the algorithm will find orthologs proteins that share a high degree of similarity with the identified IDP in DisProt. This search uses precomputed alignments inside orthology clusters, and the candidate proteins are selected by respecting several quality criteria (see the detailed description of the algorithm). These hits from the corresponding orthologs will be the recipients of the disorder annotations that are being transferred.
3. Last, similarly to the previous step, protein chains from the PDB will be assessed for homology to identify possible recipients of the disorder annotations, which have structural information. PDB structures - especially for structures containing IDPs - typically do not contain the whole protein, only a short region. Therefore, in the case of structure-based homology transfer, orthology between the query and hit proteins (in DisProt and PDB, respectively) is not enforced, and only more relaxed sequence identity-based criteria are used. Once PDB structures that contain protein regions with a high enough similarity to the query DisProt protein are identified, they are again used as targets for the transfer of disorder annotations identified in the first step.

Homology transfer by sequence (steps 1 and 2)

The main goal of this section was to transfer annotations of disordered regions and ontology terms from Disprot proteins to orthog proteins. If the user specifies an IDR from DisProt as an input, that is directly used to search for candidate orthologs currently lacking disorder annotations. If the user specifies a protein outside of DisProt, the software searches if it is present in any of the alignments done for every DisProt entry. In either case, the specified/identified DisProt proteins were used as a seed to retrieve orthologs proteins. The CD-HIT software (Fu et al. 2012)(Fiser 2006; Li and Godzik 2006)(Huang et al. 2010) was used to cluster the disprot proteins at 60 and 80 percent of identity. The databases OmaDb(Altenhoff et al. 2018) and OrthoInspector (Nevers et al. 2019) were used to obtain the ortholog proteins. The ortholog proteins were added to clusters with 60 and 80 percent of identity. For each cluster on each percent of identity, 3 types of sequence selections were applied: all proteins in the alignment, proteins with at least 50% of coverage and proteins with at least 75% of coverage. For each combination between identity percent, cluster and sequence selection, an alignment with the Clustal Omega algorithm was made. The quality of each alignment was measured with NorMD score(Thompson et al. 2001), on 3 region types: full alignment, disordered regions, and regions with ontology terms. The alignment made with 80% of identity and 75% of coverage yielded the best results for the NorMD score in the full alignment, disordered regions, and regions with ontology terms. For this reason, this data was used in the final software.

Homology transfer by structure (step 3)

The specified/identified DisProt protein is also used as a source of annotations for protein regions with solved structures reported in the PDB. In this case, the source IDR sequence is mapped to all relevant protein sequences in the PDB. This step uses a precomputed alignment between all DisProt IDRs and all protein regions in the PDB, generated by using BLAST. DisProt IDRs are given as coordinates in UniProt sequences. PDB protein chains are mapped to UniProt sequences using the EBI's SIFTS service. All corresponding UniProt sequences (both corresponding to PDB regions or DisProt IDRs) are downloaded via the UniProt API, and these sequences are aligned to each other in an all-against-all, pairwise fashion. The result of these BLAST runs are filtered for minimum values of quality scores (identity, coverage and E-value), and a single map file between DisProt and PDB is generated. This map file contains pairs of sequence regions that could be aligned with reasonable similarity, together with the gapped sequence generated by BLAST. These values can be used by the final software to pinpoint candidates in PDB for the transfer of disorder annotations, and to assess the quality of these matches.

Final software

The final software combines the data obtained from both approaches. The input of the software is the uniprot name of a protein and a region. The region is not mandatory, and by default corresponds to the full length protein. The software provides the following information: disprot annotation for the protein query, data for alignment found for the protein query, disprot annotation for the representative protein of the alignment if it is different to the query protein and structural information for the query protein showing pdbs with regions of the query protein. Figure 1 shows the General Flow Chart for the software.

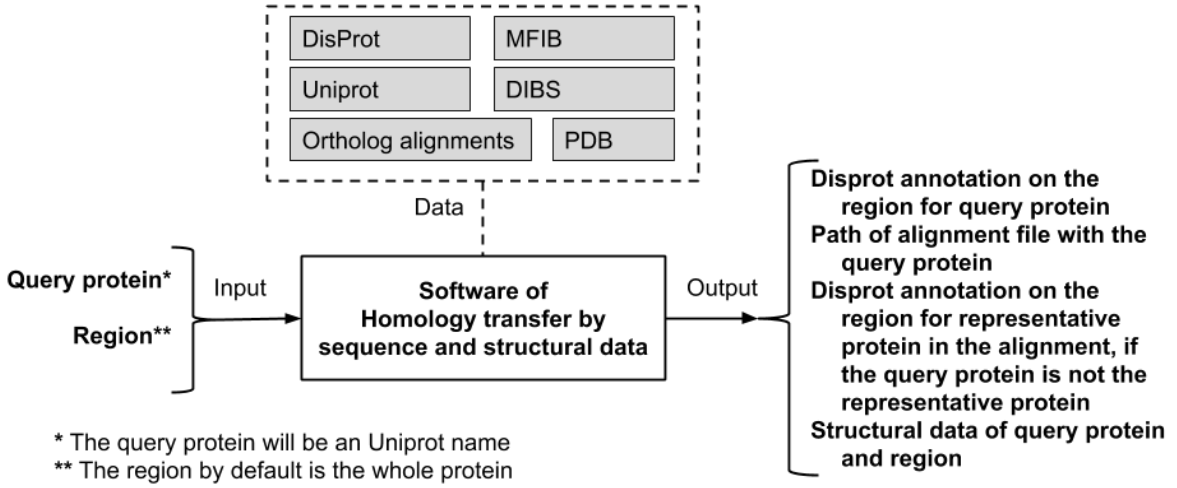

Figure 1. *Flow Chart for the software.*

Description of results

Result of Homology transfer by sequence

Methodology

Figure 2 shows the methodology used for homology transfer by sequence. The methodology was made as a part of task 1.5 by Elizabeth Martínez Pérez supervised by Cristina Marino Buslje and Toby Gibson (see report of the task 1.5).

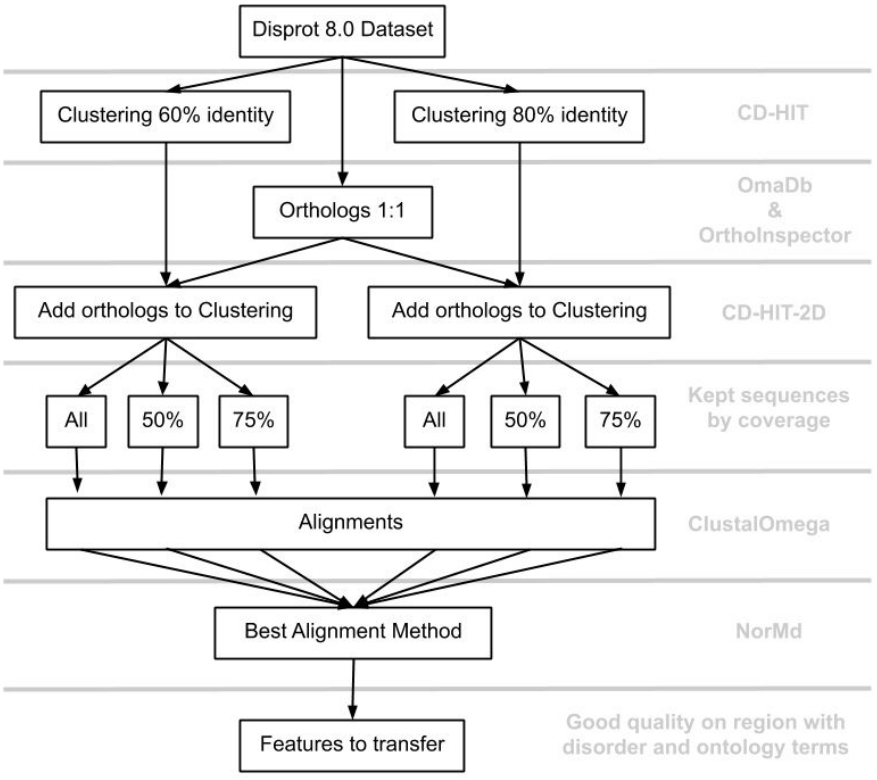

Figure 2. *WorkFlow chart for homology transfer by sequence.*

Data preparation

We downloaded the Disprot version 8.0 (Hatos et al. 2020). The proteins were filtered as follows: i) canonic proteins, ii) full length proteins (non fragments), iii) sequence having no "X" amino acid. We ended up with 1359 proteins. For those proteins, orthologs sequences were found in two databases: OmaDb (Altenhoff et al. 2018) and OrthoInspector (Nevers et al. 2019). We considered proteins to be orthologs if: i) they are orthologs 1:1 with the reference protein, ii) they have a valid uniprot (not an obsolete uniprot), iii) they are not considered as fragment in uniprot databases, iv) they have no amino acid "X" in the

sequence. From each valid Disprot entry, we generated an “orthologs family of proteins”. Figure 3 shows Disprot proteins with orthologs and the orthologs found in each dataset.

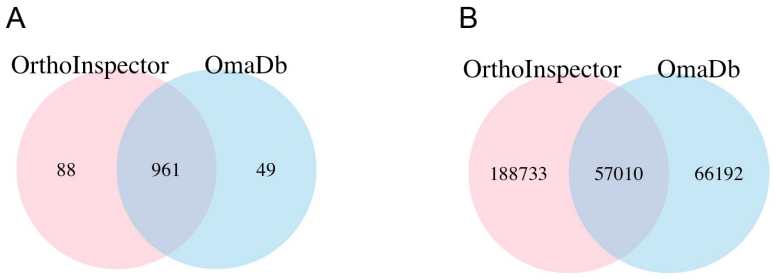

Figure 3. *Disprot proteins with orthologs and the orthologs found in each dataset.* ***A:*** *Venn diagram of disprot proteins that have 1:1 orthologs.* ***B:*** *Venn diagram of ortholog proteins (1:1) found with a valid uniprot.*

Alignments

Software CD-HIT (Fu et al. 2012)(Fiser 2006; Li and Godzik 2006)(Huang et al. 2010) was used for clustering the 1359 selected disprot proteins to 60% and 80% of identity, 1241 and 1286 proteins were selected as representative proteins respectively. We looked for orthologs in two databases and clustered them at 60% or 80% identity with the representative proteins. Also, some proteins were deleted due to a small coverage of the representative protein (see Table 1 for more details). For each protein family, the alignments were made with ClustalOmega using R language and the msa library (Bodenhofer et al. 2015).

Percent identity	Sequences by length (coverage)	Clusters >= 2 sequences	Sequences in clusters >=2 sequences	Sequences in Swiss-prot
60%	All	989	53497	6334
	At least 50%	983	53021	6323
	At least 75%	980	51231	6300
80%	All	954	29069	4085
	At least 50%	950	28851	4079
	At least 75%	947	28045	4063
Alignments with the 6 strategies		5803		

Table 1. *Alignment for each cluster with different percent of identity and coverage.*

Alignment score

The quality of alignments was measured using NorMD software(Thompson et al. 2001). Scores below to 0.6 were considered bad alignments. We measured the quality score into 3 regions types: full alignment, disordered regions and ontology regions, the last 2 are those regions annotated as disordered or with ontology respectively in DisProt. Figure 4 shows the distribution of NorMD score for each strategy, as expected, the alignments made at 80% of identity have better scores than the alignments made at 60% of identity.

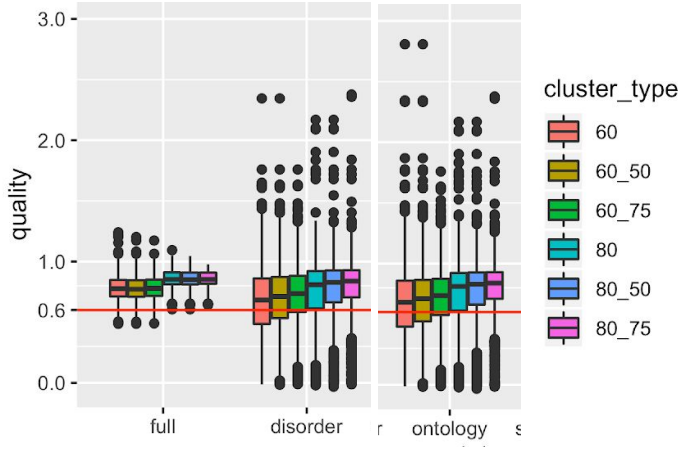

Figure 4. *norMd score for each strategy in different segment types. Axis "x",* ***full*** *the full alignment;* ***disorder****: alignment in the disordered region;* ***ontology****: regions with ontology terms;* ***order****: for order regions.*

The statistics of NorMD score are in Table 2, and show the scores obtained for 80% of identity are better than the scores obtained with 60% of identity over the 3 region types.

A

identity percent	coverage percent	total of alignments	mean	sd	min	Q1	median	Q3	max
60	all	989	0.82	0.7	0.49	0.71	0.78	0.85	16.33
60	50	983	0.78	0.11	0.49	0.71	0.77	0.85	1.2
60	75	980	0.79	0.1	0.49	0.72	0.78	0.85	1.17
80	all	954	0.86	0.07	0.61	0.81	0.85	0.91	1.09
80	50	950	0.85	0.06	0.61	0.81	0.85	0.91	1.05
80	75	947	0.86	0.06	0.64	0.82	0.85	0.91	0.98

B

identity percent	coverage percent	total of regions	mean	sd	min	Q1	median	Q3	max
60	all	1340	0.63	0.28	-0.01	0.46	0.67	0.85	1.63
60	50	1328	0.66	0.26	-0.02	0.5	0.69	0.86	1.63
60	75	1324	0.69	0.24	-0.01	0.56	0.72	0.87	1.48
80	all	1293	0.72	0.26	-0.01	0.59	0.81	0.92	1.9
80	50	1286	0.74	0.24	-0.01	0.66	0.82	0.93	1.9
80	75	1281	0.77	0.21	-0.01	0.7	0.84	0.93	1.68

C

identity percent	coverage percent	total of regions	mean	sd	min	Q1	median	Q3	max
60	all	1638	0.62	0.28	-0.01	0.45	0.67	0.85	1.87
60	50	1621	0.65	0.26	-0.02	0.49	0.69	0.86	1.87
60	75	1617	0.68	0.24	-0.01	0.55	0.72	0.87	1.48

80	all	1587	0.72	0.26	-0.01	0.59	0.81	0.92	1.9
80	50	1577	0.74	0.25	-0.01	0.65	0.82	0.93	1.9
80	75	1570	0.77	0.21	-0.01	0.7	0.83	0.93	1.68

Table 2. *Statistics of 3 regions types on the 6 strategies. In red the best values of mean and median for each region type.* ***A:*** *Statistics of NorMD score of full alignment.* ***B:*** *Statistics of NorMD score of disorder regions.* ***C:*** *Statistics of NorMD score of regions with ontology terms.*

To compare the strategies a kruskal-wallis test was applied for each region type (p.value of 9.726658e-190, 1.566173e-69 and 2.464107e-93 respectively), and evaluated the significant differences with a Dunn test. The Table 2 has the p.values of Dunn test. For scores of full alignment (Table 3) we can see the alignments made with 80% of identity have significant differences with the alignments made with 60% of identity, but the alignments made with 80-all, 80-50 and 80-75 don't have significant differences, that happened because the median of this 3 strategies are the same (Table 2A). For disordered regions and regions with ontology terms (Table 3B-C), the dunn test shows all the p.values have significant differences, supporting the strategy 80-75 present the best scores for those region types, because the strategy 80-75 have the best median for those region types (Table 2B-C).

A

	60-all	60-50	60-75	80-all	80-50
60-50	0.60				
60-75	0.82	0.46			
80-all	0.00	0.00	0.00		
80-50	0.00	0.00	0.00	0.82	
80-75	0.00	0.00	0.00	0.82	0.76

B

	60-all	60-50	60-75	80-all	80-50
60-50	0.03				
60-75	0.00	0.03			
80-all	0.00	0.00	0.00		
80-50	0.00	0.00	0.00	0.03	
80-75	0.00	0.00	0.00	0.00	0.04

C

	60-all	60-50	60-75	80-all	80-50
60-50	0.02				
60-75	0.00	0.01			
80-all	0.00	0.00	0.00		
80-50	0.00	0.00	0.00	0.02	
80-75	0.00	0.00	0.00	0.00	0.01

Table 3. *P.value of Dunn test comparing the 6 strategies on the 3 region types. P.values on red are significative differences, and blue for non-significant differences.* ***A:*** *P.values of Dunn test for full alignment.* ***B:*** *P.values of Dunn test for disorder regions.* ***C:*** *P.values of Dunn*

test for regions with ontology terms.

Data for final software

We decided to use the data generated with alignments made with 80% of identity and 75% of coverage (80-75) to the final software, because this was the strategy with the best alignment scores for disordered regions and regions with ontology terms, and also was one of the best strategies for full alignment scores. The data includes all the alignments and a file with the mapping of proteins and clusters.

Result of Homology transfer by structure

Methodology

The methodology for transferring disorder annotation via structures was done as a part of task 1.5 by Bálint Mészáros supervised by Lucía Beatriz Chemes (see secondment report).

Data preparation

The structure-based homology transfer module of the software relies on four datasets, of which the following versions have been used:

1. **DisProt 8 v0.1.0** for the source of IDP annotations
2. **PDB version 11-10-2019** for structural data
3. **UniProt version 11-10-2019** for protein sequence data
4. **SIFTS version 11-10-2019** for region-level mapping of PDB chain sequences to UniProt sequences

First, SIFTS was used to obtain all UniProt accessions that represent proteins that have at least one region as part of any PDB structure. These UniProt accessions, together with the ones referenced in DisProt, were used to download all corresponding UniProt sequences via the UniProt API service.

Next, each UniProt sequence referenced in DisProt (‘*source sequences*’, as these are the sources of disorder annotation to be transferred) were aligned to each UniProt sequence referenced in PDB (‘*target sequences*’). BLAST version 2.2.18 was used and in each pairwise alignment, full length canonical isoform UniProt sequences were used. In each case where BLAST produced a reasonable alignment using an e-value cutoff of 10^{-4}, the BLAST report defining the alignment was saved. These files were processed and further filtered by keeping hits only where the aligned region covers at least 10% of both the source and the target sequences, the e-value is less than 10^{-6}, and the sequence identity over the aligned region is at least 25% (counting gaps as mismatches).

The remaining BLAST reports were processed into a single map file that contains the UniProt accessions of both the source and the target sequences, the respective region boundaries of both proteins, and the aligned sequences containing the gaps introduced by

BLAST. This mapping file is used by the final software to map DisProt regions to candidate PDB protein chains.

Structure module in the final software

Disorder annotations from the source IDR are mapped to PDB structures/chains using the above detailed SIFTS/BLAST mapping file by respecting several cutoffs to guarantee the quality of the annotation transfer. Disorder annotations are transferred from the source region (in DisProt) to the target region (in the PDB) if all of the following criteria are met:

- The alignment between the two proteins covers at least 80% of the target region (as defined in the SEQRES record), to ensure proper coverage of the target
- The source region covers at least 80% of the target region, to ensure that the target region is (almost) fully disordered
- The sequence identity in the aligned region covering the target region is at least 50%, to ensure a high enough similarity

All of the above cutoff values are defined in the script, and can easily be modified to give the program more flexibility.

If there is a target region that satisfies all the above criteria, the structure module of the software prints the hit by specifying the following information:

- PDB ID and chain ID of the target region
- The source and target regions using the appropriate UniProt accessions and region boundaries
- The two coverage values and the identity value used in the above detailed filtering step

In addition, the program also gives information about the novelty of the identified PDB structure, by specifying if the identified PDB hit is already included in any of the two IDP interaction databases DIBS and MFIB. At this step, the algorithm parses the PDB IDs referenced in these databases - at the alignment level this corresponds to a 100% sequence identity between the target and the source sequences. This step takes into account the ‘related structures’ (additional structures that weren’t selected as the representative structures for the DIBS/MFIB entries, but share a high degree of similarity to them) in both databases in addition to the representative structures (the main structures for each of the DIBS/MFIB entries) used for creating DIBS/MFIB entries. If the identified PDB structure is absent from both databases, the software marks the hit as ‘novel’. This does not necessarily mean that the identified IDR isn’t already present in these databases, but even if it is, it is present bound to a different partner, counted as a different interaction. The presence of a novel hit is an indication that these structures can be considered as candidates for extending DIBS or MFIB.

Examples of final software

Call the program as,

./IDP_homology_transfer.pl <uniprot_name> [start_region end_region]

- uniprot_name is mandatory
- start and end region aren't mandatory. By default (omitted this parameters), start_region take value of 1, and end_region take the length of protein.

Example of a Disprot protein

Uniprot P06179 (Flagellin from Salmonella typhimurium) is in the Disprot database and has a disordered region from residues 455 to 494. The disordered regions from P06179 can be transferred by homology to other proteins in the alignment.

Software call

./IDP_homology_transfer.pl P06179 455 494

Annotation for query protein of disorder region from Disprot

#hit_type regionID region PubMedID annotation_type experimental_technique_used
ANNOTATION DP00026r001 (455 - 494) 2810365 Disorder (Structural state - DO:00076) Circular dichroism spectroscopy far-UV (Detection method - DO:00096)
ANNOTATION DP00026r002 (455 - 494) 2810365 Prion (Disorder function - DO:00047) Circular dichroism spectroscopy far-UV (Detection method - DO:00096)
ANNOTATION DP00026r004 (455 - 494) 2810365 Protein binding (Interaction partner - DO:00063) Circular dichroism spectroscopy far-UV (Detection method - DO:00096)
ANNOTATION DP00026r003 (455 - 494) 2810365 Disorder to order (Structural transition - DO:00050) Circular dichroism spectroscopy far-UV (Detection method - DO:00096)
ANNOTATION DP00026r009 (455 - 494) 2810365 Disorder (Structural state - DO:00076) Sensitivity to proteolysis (Detection method - DO:00085)
ANNOTATION DP00026r010 (455 - 494) 2810365 Prion (Disorder function - DO:00047) Sensitivity to proteolysis (Detection method - DO:00085)
ANNOTATION DP00026r012 (455 - 494) 2810365 Protein binding (Interaction partner - DO:00063) Sensitivity to proteolysis (Detection method - DO:00085)
ANNOTATION DP00026r011 (455 - 494) 2810365 Disorder to order (Structural transition - DO:00050) Sensitivity to proteolysis (Detection method - DO:00085)

Path of alignment file

data_alignment_UniProt/7_align80_75clustalomega/clustalO_cluster75_452.fasta.aln

Ortholog proteins from the alignment

#hit_type orthologue_region identitiy_over_full_protein identity_over_aligned_region coverage_of_query_region_by_orthologs_region current_status_in_DisProt
SEQUENCE A0A0F6B2U2(455-494) 1.000 1.000 1.000 novel hit
SEQUENCE A0A0H3NMJ6(455-494) 1.000 1.000 1.000 novel hit
SEQUENCE E8XAA1(455-494) 1.000 1.000 1.000 novel hit

Structural data

#hit_type PDB_ID chainID region_in_PDB region_in_query PDB_coverage_by_alignemnt PDB_coverage_by_DisProt identity_over_alignment current_status_in_DIBS/MFIB
STRUCTURE 5yud C P52616(463-506) P06179(452-495) 1.000 0.909 1.000 novel hit

Example of non Disprot protein

Uniprot E8XAA1 isn't in the Disprot database. Uniprot E8XAA1 is in the same orthologs cluster than P06179. The disordered regions from P06179 can be transferred by homology to E8XAA1.

Software call

./IDP_homology_transfer.pl E8XAA1

Path of alignment file

data_alignment_UniProt/7_align80_75clustalomega/clustalO_cluster75_452.fasta.aln

Ortholog proteins from the alignment

#hit_type orthologue_region identitiy_over_full_protein identity_over_aligned_region coverage_of_query_region_by_orthologs_region current_status_in_DisProt
SEQUENCE A0A0F6B2U2(1-495) 1.000 1.000 1.000 novel hit
SEQUENCE A0A0H3NMJ6(1-495) 1.000 1.000 1.000 novel hit
SEQUENCE P06179(1-495) 1.000 1.000 1.000 already in DisProt

Representative ortholog protein form Disprot

P06179

Annotation for representative protein of disorder region from Disprot

#hit_type regionID region PubMedID annotation_type experimental_technique_used
ANNOTATION DP00026r001 P06179 (455 - 494) 2810365 Disorder (Structural state - DO:00076) Circular dichroism spectroscopy far-UV (Detection method - DO:00096)

ANNOTATION DP00026r002 P06179 (455 - 494) 2810365 Prion (Disorder function - DO:00047) Circular dichroism spectroscopy far-UV (Detection method - DO:00096)
ANNOTATION DP00026r004 P06179 (455 - 494) 2810365 Protein binding (Interaction partner - DO:00063) Circular dichroism spectroscopy far-UV (Detection method - DO:00096)
ANNOTATION DP00026r003 P06179 (455 - 494) 2810365 Disorder to order (Structural transition - DO:00050) Circular dichroism spectroscopy far-UV (Detection method - DO:00096)
ANNOTATION DP00026r005 P06179 (1 - 65) 2810365 Disorder (Structural state - DO:00076) Sensitivity to proteolysis (Detection method - DO:00085)
ANNOTATION DP00026r006 P06179 (1 - 65) 2810365 Prion (Disorder function - DO:00047) Sensitivity to proteolysis (Detection method - DO:00085)
ANNOTATION DP00026r008 P06179 (1 - 65) 2810365 Protein binding (Interaction partner - DO:00063) Sensitivity to proteolysis (Detection method - DO:00085)
ANNOTATION DP00026r007 P06179 (1 - 65) 2810365 Disorder to order (Structural transition - DO:00050) Sensitivity to proteolysis (Detection method - DO:00085)
ANNOTATION DP00026r009 P06179 (455 - 494) 2810365 Disorder (Structural state - DO:00076) Sensitivity to proteolysis (Detection method - DO:00085)
ANNOTATION DP00026r010 P06179 (455 - 494) 2810365 Prion (Disorder function - DO:00047) Sensitivity to proteolysis (Detection method - DO:00085)
ANNOTATION DP00026r012 P06179 (455 - 494) 2810365 Protein binding (Interaction partner - DO:00063) Sensitivity to proteolysis (Detection method - DO:00085)
ANNOTATION DP00026r011 P06179 (455 - 494) 2810365 Disorder to order (Structural transition - DO:00050) Sensitivity to proteolysis (Detection method - DO:00085)
ANNOTATION DP00026r013 P06179 (1 - 65) 2810365 Disorder (Structural state - DO:00076) Circular dichroism spectroscopy far-UV (Detection method - DO:00096)
ANNOTATION DP00026r014 P06179 (1 - 65) 2810365 Prion (Disorder function - DO:00047) Circular dichroism spectroscopy far-UV (Detection method - DO:00096)
ANNOTATION DP00026r016 P06179 (1 - 65) 2810365 Protein binding (Interaction partner - DO:00063) Circular dichroism spectroscopy far-UV (Detection method - DO:00096)
ANNOTATION DP00026r015 P06179 (1 - 65) 2810365 Disorder to order (Structural transition - DO:00050) Circular dichroism spectroscopy far-UV (Detection method - DO:00096)

Structural data

No structural data was found for query protein

Availability

The project source code is available at the IDPfun GitLab repository: https://gitlab.com/idpfun/idp_homology_transfer_deliverable.

References

Altenhoff, Adrian M., Natasha M. Glover, Clément-Marie Train, Klara Kaleb, Alex Warwick Vesztrocy, David Dylus, Tarcisio M. de Farias, et al. 2018. “The OMA Orthology Database in 2018: Retrieving Evolutionary Relationships among All Domains of Life through Richer Web and Programmatic Interfaces.” *Nucleic Acids Research* 46 (D1): D477–85.

Bodenhofer, Ulrich, Enrico Bonatesta, Christoph Horejš-Kainrath, and Sepp Hochreiter. 2015. “Msa: An R Package for Multiple Sequence Alignment.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv494.

Brown, Celeste J., Audra K. Johnson, A. Keith Dunker, and Gary W. Daughdrill. 2011. “Evolution and Disorder.” *Current Opinion in Structural Biology* 21 (3): 441–46.

Chen, Jessica Walton, Pedro Romero, Vladimir N. Uversky, and A. Keith Dunker. 2006. “Conservation of Intrinsic Disorder in Protein Domains and Families: II. Functions of Conserved Disorder.” *Journal of Proteome Research* 5 (4): 888–98.

Fiser, Andras. 2006. “Faculty of 1000 Evaluation for Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” *F1000 - Post-Publication Peer Review of the Biomedical Literature*. https://doi.org/10.3410/f.1033342.375839.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. “CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data.” *Bioinformatics* 28 (23): 3150–52.

Gitlin, Leonid, Tzachi Hagai, Anthony LaBarbera, Mark Solovey, and Raul Andino. 2014. “Rapid Evolution of Virus Sequences in Intrinsically Disordered Protein Regions.” *PLoS Pathogens* 10 (12): e1004529.

Hatos, András, Borbála Hajdu-Soltész, Alexander M. Monzon, Nicolas Palopoli, Lucía Álvarez, Burcu Aykac-Fas, Claudio Bassot, et al. 2020. “DisProt: Intrinsic Protein Disorder Annotation in 2020.” *Nucleic Acids Research* 48 (D1): D269–76.

Huang, Ying, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. 2010. “CD-HIT Suite: A Web Server for Clustering and Comparing Biological Sequences.” *Bioinformatics* 26 (5): 680–82.

Light, Sara, Rauan Sagit, Oxana Sachenkova, Diana Ekman, and Arne Elofsson. 2013. “Protein Expansion Is Primarily due to Indels in Intrinsically Disordered Regions.” *Molecular Biology and Evolution* 30 (12): 2645–53.

Li, W., and A. Godzik. 2006. “Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btl158.

Nevers, Yannis, Arnaud Kress, Audrey Defosset, Raymond Ripp, Benjamin Linard, Julie D. Thompson, Olivier Poch, and Odile Lecompte. 2019. “OrthoInspector 3.0: Open Portal for Comparative Genomics.” *Nucleic Acids Research* 47 (D1): D411–18.

Rost, B. 1999. “Twilight Zone of Protein Sequence Alignments.” *Protein Engineering* 12 (2): 85–94.

Szalkowski, Adam M., and Maria Anisimova. 2011. “Markov Models of Amino Acid Substitution to Study Proteins with Intrinsically Disordered Regions.” *PloS One* 6 (5): e20488.

Thompson, Julie D., Frédéric Plewniak, Raymond Ripp, Jean-Claude Thierry, and Olivier Poch. 2001. “Towards a Reliable Objective Function for Multiple Sequence Alignments 1

1Edited by J. Karn.” *Journal of Molecular Biology*. https://doi.org/10.1006/jmbi.2001.5187.