Where differences resemble: Sequence-feature analysis in curated databases of intrinsically disordered proteins

Marco Necci^{1,2,3}, Damiano Piovesan¹, Silvio C.E. Tosatto^{1,4,*}

 ¹ Dept. of Biomedical Sciences, University of Padua, Padua, Italy
² Dept. of Agricultural Sciences, University of Udine, via Palladio 8, 33100, Udine, Italy
³ Fondazione Edmund Mach, Via E. Mach 1, 38010, S. Michele all'Adige, Italy
⁴ CNR Institute of Neuroscience, Padua, Italy
* To whom correspondence should be addressed at: via U. Bassi 58/b, 35131 Padova, Italy
Email: silvio.tosatto@unipd.it; phone: +39 049 827 6269; fax: +39 049 827 6269

Supplementary Material



Figure S1 – **Merge strategy for overlapping annotation regions**. Overlapping annotations in the origin datasets are converted to a series of adjacent annotations in MobiDB 3.0. In this analysis, adjacent annotations are considered as a single annotation.

DIBS	12770	2852	2249	428	3405	413	98	3325
DisProt	2852	81659	884	2998	8907	1235	276	64507
ELM	2249	884	19325	195	1074	124	36	14763
FuzDB	428	2998	195	8803	742	72	178	4190
IDEAL	3405	8907	1074	742	47315	932	165	32090
MFIB	413	1235	124	72	932	22487	0	19711
UniProt-Dis	98	276	36	178	165	0	38224	37471
	DIBS	DisProt	ELM	FuzDB	IDEAL	MFIB	JniProt-Dis	Unique res

Figure S2 – Number of overlapping residues. Each cell represents the ID regions overlap between a database pair. Unique residues are those annotated uniquely by a single database. The count is by row, e.g. DIBS annotates 3,325 residues.

DIBS	100.00	22.33	17.61	3.35	26.66	3.23	0.77	26.04
DisProt	3.49	100.00	1.08	3.67	10.91	1.51	0.34	79.00
ELM	11.64	4.57	100.00	1.01	5.56	0.64	0.19	76.39
FuzDB	4.86	34.06	2.22	100.00	8.43	0.82	2.02	47.60
IDEAL	7.20	18.82	2.27	1.57	100.00	1.97	0.35	67.82
MFIB	1.84	5.49	0.55	0.32	4.14	100.00	0.00	87.66
UniProt-Dis	0.26	0.72	0.09	0.47	0.43	0.00	100.00	98.03
	DIBS	DisProt	ELM	FuzDB	IDEAL	MFIB	IniProt-Dis	Unique res

Figure S3 – Fraction of overlapping residues. As in Figure S2 but numbers represent percentages calculated over the database size, e.g. DIBS uniquely annotates 26.04% of its ID residues.



Figure S4 – **Taxonomic diversity significance.** Significance is calculated with the Chi-square test and represent diversity of the taxonomic distribution, P-value < 0.05 represents significantly different distributions. The Chi-square is directional, but the test is consistent in both directions.



Figure S5 – **Heat map of statistical significance of differences in region lengths**. The P-value represents the statistical significance of the difference between database length distribution and it is calculate with a T-test.



Figure S6 – Hierarchically clustered heat map of amino acid absolute frequencies. Clustering is based on Euclidean distance between frequency vectors. Each value represent amino acid frequencies of each database.



Figure S7 – **Correlation matrix of absolute amino acid frequencies**. Correlation coefficients (in red) and correlation p-values (in blue) of absolute amino acid.



Figure S8 – Correlation matrix of fold increase amino acid frequencies. Correlation coefficients (in red) and correlation p-values (in blue) of fold increase compared to TrEMBL reference frequencies.



Figure S9 –Heat map of statistical significance of differences in conformational propensity. Significance is calculated with the Chi-square test and represent diversity in the distribution of the five different Pappu's classes. P-value < 0.05 represents significantly different distributions. The Chi-square is directional, but the test is consistent in both directions.



Figure S10 –**Heat map of statistical significance of differences in region low complexity content**. The P-value represents the statistical significance of the difference between region low complexity content distribution and it is calculate with a T-test.