Electronic Supplementary Material (ESI) for Molecular BioSystems.

Supplementary Materials

Not an Exception to the Rule: The Functional Significance of Intrinsically Disordered Protein Regions in Enzymes

Shelly DeForte[1] and Vladimir N. Uversky[1,2,3,*]

[1]*Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, Florida 33612*

[2]*USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, Florida 33612*

[3]*Laboratory of Structural Dynamics, Stability and Folding of Proteins, Institute of Cytology, Russian Academy of Sciences, St. Petersburg, Russia*

***Correspondence to:** Shelly DeForte: sdeforte@health.usf.edu; Vladimir N. Uversky: vuversky@health.usf.edu

Table S1. Enzymes with Experimentally Characterized Intrinsically Disordered Regions

Enzyme or Enzyme Family with IDPR(s)	**PMIDs**
2,4-dienoyl-CoA reductase	15531764
5-Aminolevulinate synthase (ALAS)	25240868
ABL tyrosine kinase	17211892
acetylcholinesterase variant, AChE-R	20173328, 23908786
acylphosphatase	24893801, 20223823, 18832052, 18451804, 16287076, 14872538, 9790846
adenylate cyclase toxin	19860484, 20096704, 21416544, 24145447
alphavirus capsid protease	25100849, 9094737
aminoglycoside 6'-N-acetyltransferase type Ii	16131761
anhydrin	20805515
apo-DCpS	15769464
ARG tyrosine kinase	17211892
arginine kinase	21075117
bacillus lipase	25001212
bacterial luciferase mobile loop	21156144
BCR-ABL tyrosine kinase	22632137
BTPC carboxylase	22404138, 24266766, 22569262, 21524275
calcineurin, CaN	22100452
cathepsin F	23684953
CCT, phosphocholine cytidylytransferase	21303909, 24397368, 23238251, 22988242
CDK2AP1	22427660
Cel7a, Trichoderma reesi family 7 cellobiohydrolase	21112302, 23959893
colicin E9	15004032, 16114886, 16166265, 17375930, 18573254, 19021565, 22310049, 23672584, 23812713
c-Src kinase	19520085, 25071818, 23744817
Dbp5p, DEAD-box protein 5	19281819, 24045937, 21884706
dihydroorotase, DHO	19128030
Dnmt1	25533200, 20352123, 19923434
E2 enzymes sub-family 3R	22507829
E2-C	10350465
EcoO109I	19348764
EGFR	15840573, 22579287
Eisenia LK	15327979
epidermal growth factor receptor, EGFR, EGFR kinase domain	15840573, 22579287
ErbB2 receptor tyrosine kinase	24815698
firefly luciferase	19119851, 19492113, 20221465
FtsZ	25305578, 23714328, 23692518, 23692518
glucokinase	23271955

glycolate oxidase	8706682, 18215067
hammerhead ribozyme	10802069, 21740954
HDAC4	18614528
Hef	24947516
hemocyanin	19446530
hepatitis B polymerase	23202419
HIV-1 protease	10739910, 15572155, 17243183
hsp90 atpase	22660624
inteins	16288917, 24236406
lombricine kinase, phosphagen kinase family	15327979, 21212263, 20121101, 21212263
Lysyle oxidase	20192271
Metallo beta lactamase	19395380
mGIP/SCP1	24751520, 24925644
multidomain polymerase protein	25297996
myosin II heavy chaine kinase B	20199682
NEIL1	22902625, 23542007
Nickel Superoxide dismutase	25580509
NS3, non structural protein 3 from hepatitis C virus	23803659, 21112306, 9223519, 24752801
Ntrc1	16169010
nuclease colicin	15004032
nudix hydrolase	20657662
p1 protease	24603811
p300 acetyltransferase	17438265, 23133622, 23307074, 24253305
PAM, peptidylglycine alpha-amidating monooxygenase	19635792
PfTIM, triosephosphate isomerase	19914198, 15465054
phototropin	21222437
PKS, modular polyketide synthase	22282160, 9166770
polyphenol oxidase	16332393
protein kinase A	17222345, 23946424, 24192038, 25112875
protein kinase C alpha V5 domain	23762412
protein kinase R, PKR	19232355
PTP1B/PTPN1	17643420, 24845231
RelA/SPoT	24717772
retinal phosphodiesterase gamma subunit	12643535, 18230733, 19075750, 21393250, 21978030, 22514270
Ribonuclease A	9689069
Ribonuclease P	11258888, 11749217, 20476778
RNA polymerase	25261014
Rnase E	12947103, 15236960, 16094605, 16516921, 17447862, 19215771, 20952404, 25432321
Rnase P	11258888, 20476778, 15518563

Rnase Y	21803996
San1, PQC ubiquitin ligase	21211726, 23363599, 21551067, 21941105
selenoprotein k	22963794
selenoprotein methionine sulfoxide reductase B1, MsrB1	20605785
selenoprotein s, VIMP reductase	23566202, 23914919, 22700979
sgs1 helicase	24038467, 21111748
sirt1, sirtuin family	23497088, 23811471, 24020004
sortase	22468560
src family kinase	25071818
SRPK1	21600902
Sso Acp	24893801
sulfhydryl oxidase ALR	23207295
TAFI	18722183
Tbsp1	23192346
TgGCN5 family histone acetyltransferases	21055425
thrombin	21782041
thymidylate synthase	16259621, 14967037, 19797058, 20815815, 21878626, 23181752, 23684952
TPPP/p25	21995432
Type IA topoisomerases	18186484
Ube2w E2	25436519
UreG	15542602, 17309280
Vpr, feather degrading minor extracellullar protease	19383694
xylanases	25576604
yck2 kinase	21653825
YopH	16698773

These are enzymes found through manual sorting of the results returned from the PubMed search (intrinsically OR natively OR naturally OR inherently) AND (disordered OR unfolded OR unstructured OR denatured OR flexible) AND (protein OR region OR peptide OR domain) AND (1978/1/1:2014/10/15[dp]) as well as additional enzyme specific searches.

The criterion for inclusion in this list is that the associated published studies for each enzyme must refer to the enzyme using IDP terminology. Due to the size of the body of literature, this list should not be taken as exhaustive.

Table S2. Bray-Curtis similarity coefficient and expected values for the longest predicted stretch of disorder in the reference proteomes for eukaryotes and prokaryotes. *n* is the number of proteins, *S* is the similarity coefficient, *STD* is the standard deviation of the similarity coefficient, *err* is the dissimilarity of the group of non-enzymes with itself for *n*, and *EV* is the expected value for the distribution values.

Eukaryotes				
EC#	**n**	**S ± STD**	**err**	**EV**
1: Oxidoreductases	20285	0.44 ± 0.0042	0.012	17
2: Transferases	43112	0.91 ± 0.0041	0.0082	50
3: Hydrolases	49002	0.85 ± 0.0039	0.0077	44
4: Lyases	3278	0.55 ± 0.012	0.028	24
5: Isomerases	2915	0.74 ± 0.011	0.030	39
6: Ligases	4062	0.66 ± 0.012	0.026	30
Multiple EC#s	6860	0.87 ± 0.010	0.020	47
Non-Enzymes	428126			59
Prokaryotes				
EC#	**n**	**S ± STD**	**err**	**EV**
1: Oxidoreductases	5751	0.75 ± 0.012	0.027	10
2: Transferases	8344	0.82 ± 0.011	0.022	12
3: Hydrolases	8784	0.92 ± 0.0090	0.022	15
4: Lyases	1563	0.72 ± 0.024	0.049	9.5
5: Isomerases	1059	0.77 ± 0.026	0.058	11
6: Ligases	1385	0.73 ± 0.026	0.052	9.7
Multiple EC#s	1704	0.83 ± 0.023	0.047	12
Non-Enzymes	48443			17

Table S3. Bray-Curtis similarity coefficient and expected values for the longest missing region in the PDB dataset for eukaryotes and prokaryotes. *n* is the number of proteins, *S* is the similarity coefficient, *STD* is the standard deviation of the similarity coefficient, *err* is the dissimilarity of the group of non-enzymes with itself for *n*, and *EV* is the expected value for the distribution values.

Eukaryotes				
EC#	**n**	**S ± STD**	**err**	**EV**
1: Oxidoreductases	481	0.91 ± 0.31	0.076	15
2: Transferases	981	0.88 ± 0.016	0.053	20
3: Hydrolases	935	0.94 ± 0.014	0.055	17
4: Lyases	160	0.85 ± 0.051	0.12	14
5: Isomerases	119	0.85 ± 0.051	0.14	14
6: Ligases	63	0.62 ± 0.083	0.18	35
Multiple EC#s	61	0.81 ± 0.067	0.18	21
Non-Enzymes	7888			16
Prokaryotes				
EC#	**n**	**S ± STD**	**err**	**EV**
1: Oxidoreductases	680	0.93 ± 0.014	0.058	11
2: Transferases	994	0.86 ± 0.017	0.049	17
3: Hydrolases	719	0.91 ± 0.018	0.056	12
4: Lyases	386	0.93 ± 0.022	0.076	11
5: Isomerases	233	0.90 ± 0.026	0.095	11
6: Ligases	220	0.76 ± 0.047	0.098	19
Multiple EC#s	53	0.85 ± 0.059	0.17	11
Non-Enzymes	10868			11

Table S4A. Expected values for the fraction of disorder by protein length interval in eukaryotes. *n* is the number of proteins, *EV* is the expected value for the distribution values, and the *Std Err* is the standard error. The last interval contains the length value at the 95th percentile, and all values greater.

1: Oxidoreductases

Interval	<100	100-200	200-400	400-600	600-800	>800
n	493	2012	8301	6583	1792	1104
EV	0.19	0.11	0.056	0.046	0.069	0.13
Std Err	0.0059	0.0028	0.00086	0.00084	0.0021	0.0048

2: Transferases

Interval	<100	100-200	200-400	400-600	600-800	800-1000	1000-1200	>1200
n	461	2593	14070	12791	5445	2888	1655	3209
EV	0.28	0.12	0.080	0.11	0.17	0.20	0.22	0.26
Std Err	0.0086	0.0028	0.0005	0.0011	0.0023	0.0034	0.0049	0.0039

3: Hydrolases

Interval	<100	100-200	200-400	400-600	600-800	800-1000	1000-1200	1200-1400	>1400
n	533	3530	14755	11789	6861	4193	2706	1650	2895
EV	0.21	0.10	0.076	0.093	0.13	0.15	0.18	0.18	0.18
Std Err	0.0093	0.0024	0.000815	0.0011	0.0017	0.0022	0.0032	0.0038	0.0031

4: Lyases

Interval	<100	100-200	200-400	400-600	600-800	800-1000	>1000
n	17	241	1208	1006	439	109	258
EV	0.13	0.086	0.073	0.059	0.076	0.080	0.11
Std Err	0.024	0.0062	0.0027	0.0022	0.0037	0.0087	0.00073

5: Isomerases

Interval	<100	100-200	200-400	400-600	>600
n	28	508	1214	778	387
EV	0.23	0.11	0.080	0.14	0.19
Std Err	0.038	0.0060	0.0028	0.0058	0.011

6: Ligases

Interval	<100	100-200	200-400	400-600	600-800	800-1000	>1000
n	34	359	578	1599	911	269	312
EV	0.21	0.14	0.11	0.067	0.086	0.14	0.10
Std Err	0.033	0.0060	0.0057	0.0022	0.0036	0.0083	0.0072

Multiple EC#s

Interval	<100	100-200	200-400	400-600	600-800	800-1000	1000-1200	1200-1400	1400-1600	1600-1800	1800-2000	>2000
n	29	623	1674	1494	913	667	561	179	145	137	70	368
EV	0.15	0.12	0.10	0.10	0.11	0.12	0.094	0.14	0.17	0.24	0.20	0.13
Std Err	0.021	0.0042	0.0030	0.0040	0.0044	0.0052	0.0052	0.0093	0.012	0.013	0.020	0.0069

Non-Enzymes

Interval	<100	100-200	200-400	400-600	600-800	800-1000	1000-1200
n	31685	89538	138700	76154	38085	21338	32626
EV	0.29	0.23	0.20	0.23	0.26	0.26	0.30
Std Err	0.0016	0.000848	0.00062	0.00083	0.0012	0.0016	0.0014

Table S4B. Expected values for the fraction of disorder by protein length interval in prokaryotes. *n* is the number of proteins, *EV* is the expected value for the distribution values, and the *Std Err* is the standard error. The last interval contains the length value at the 95th percentile, and all values greater.

1: Oxidoreductases

Interval	<100	100-200	200-400	400-600	>600
n	69	860	2769	1579	1474
EV	0.16	0.078	0.045	0.039	0.044
Std Err	0.018	0.0027	0.00099	0.00123	0.0027

2: Transferases

Interval	<100	100-200	200-400	400-600	600-800	>800
n	84	997	4361	1893	563	446
EV	0.15	0.054	0.042	0.054	0.054	0.053
Std Err	0.013	0.0018	0.00081	0.0018	0.0033	0.0038

3: Hydrolases

Interval	<100	100-200	200-400	400-600	600-800	>800
n	152	1187	4023	1831	967	624
EV	0.23	0.084	0.051	0.060	0.053	0.07
Std Err	0.018	0.0030	0.00099	0.0019	0.0022	0.0042

4: Lyases

Interval	<100	100-200	200-400	400-600	>600
n	11	259	718	419	156
EV	0.14	0.060	0.042	0.041	0.032
Std Err	0.036	0.0035	0.0015	0.0021	0.0033

5: Isomerases

Interval	<100	100-200	200-400	400-600	>600
n	20	154	642	186	57
EV	0.19	0.077	0.040	0.050	0.090
Std Err	0.031	0.0056	0.0020	0.0052	0.014

6: Ligases

Interval	<100	100-200	200-400	400-600	>600	>800
n	21	64	360	680	168	92
EV	0.15	0.12	0.036	0.030	0.038	0.020
Std Err	0.030	0.0017	0.0021	0.0013	0.0038	0.0028

Multiple EC#s

Interval	<100	100-200	200-400	400-600	600-800	800-1000	>1000
n	7	105	737	412	188	145	110
EV	0.33	0.050	0.038	0.033	0.043	0.044	0.047
Std Err	0.082	0.0050	0.0017	0.0020	0.0032	0.0048	0.0061

Non-Enzymes

Interval	<100	100-200	200-400	400-600	>600
n	7347	13883	17104	6795	3314
EV	0.28	0.14	0.084	0.077	0.080
Std Err	0.0031	0.0015	0.00094	0.0014	0.0022

Table S5. Disordered enzymes at the second level of the EC hierarchy. These are eukaryotic enzymes from the reference proteomes dataset at the second level of the EC hierarchy which have a similarity coefficient greater than or equal to 0.85 when compared to the group of non-enzymes. *n* is the number of proteins, *S* is the similarity coefficient, *STD* is the standard deviation of the similarity coefficient, *err* is the dissimilarity of the group of non-enzymes with itself for *n*, and *EV* is the expected value for the distribution values.

EC#	**n**	**S ± STD**	**err**	**EV**
2.1: single carbon transferases	4629	0.90 ± 0.0087	0.024	53
2.7: phosphorus transferases	21253	0.96 ± 0.0048	0.011	63
3.1: ester bond hydrolases	2190	0.85 ± 0.018	0.034	46
3.4: peptide bond hydrolases	3934	0.92 ± 0.013	0.026	52
3.6: acid anhydride hydrolases	14410	0.95 ± 0.0048	0.014	60
5.2: cis-trans isomerases	1002	0.85 ± 0.017	0.049	67
6.5: phosophoric ester bond ligase	244	0.91 ± 0.027	0.094	60
Non-Enzymes	428126			59

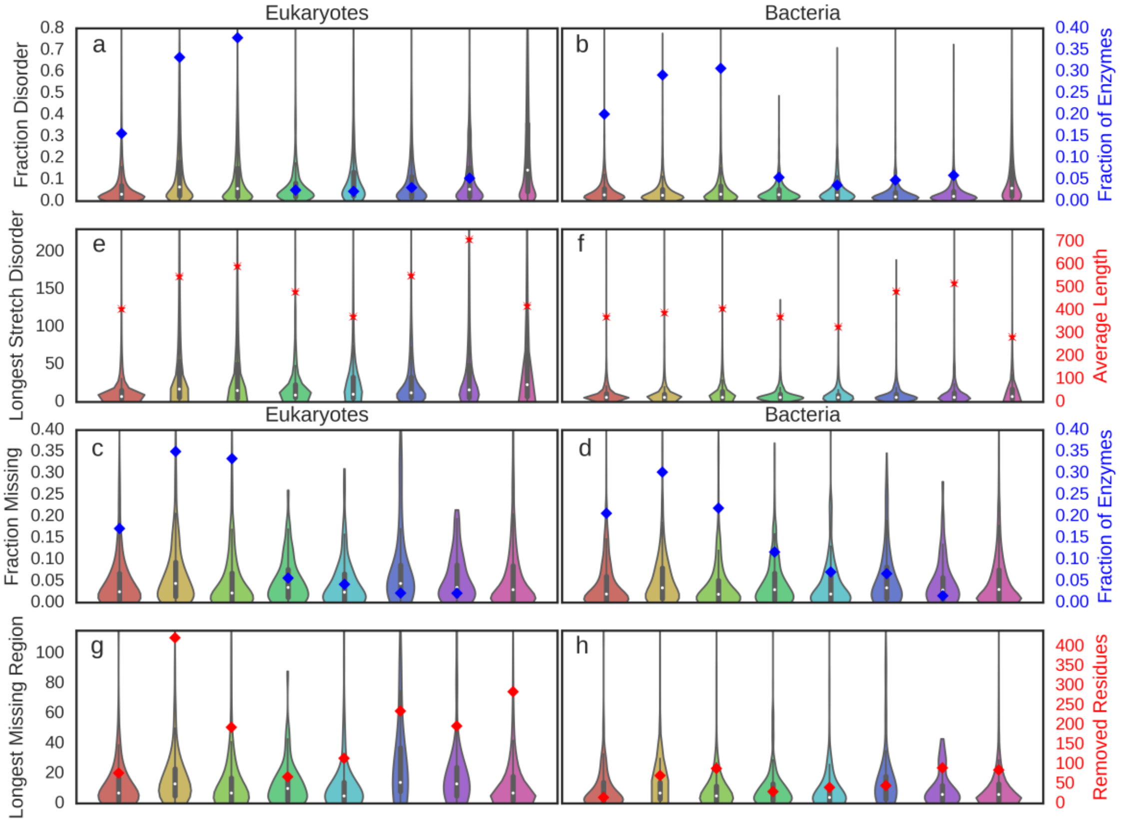

Figure S1. Violin plots corresponding to Figure 2. These plots correspond to the box plots displayed in Figure 2 and show a smoothed distribution of scores via a violin plot. **a-b**. The fraction of predicted disorder per-protein in reference proteomes in eukaryotes (a) and bacteria (b). **e-f** The longest predicted missing region per-protein in reference proteomes in eukaryotes (e) and bacteria (f). The red star shows the average protein length. **c-d** The fraction of missing residues per-protein in the PDB in eukaryotes (c) and bacteria (d). **g-h** The longest missing region per-protein in eukaryotes (g) and bacteria (h). **a-d** The blue diamond shows the relative fraction of the full set of enzymes that each enzyme type occupies. **e-h** The red diamond shows the average number of removed residues per-protein before crystallization. The white dot corresponds to the median, and the thick black inner bar corresponds to the interquartile range.