Supplement for an article entitled

RAPID: fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale

Jing Yan[a], Marcin J. Mizianty[a], Paul L. Filipow[a], Vladimir N. Uversky[b,c] and Lukasz Kurgan[a*]

[a] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada

[b] Department of Molecular Medicine and Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA

[c] Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia.

[*] Corresponding author; lkurgan@ece.ualberta.ca; 780-492-5488.

2.3 RAPID predictor

Four groups of features are derived from a given protein chain and physicochemical and structural properties of its amino acids (AAs):

1. *AA composition* (20 features), which is defined as N_r / N, $r = 1,2,\ldots,20$ where N_r is the count of AA of type r in the sequence and N is the length of the sequence. The importance of the AA composition for the prediction of the intrinsic disorder was demonstrated in numerous studies, e.g., (Dunker et al., 2001).
2. *Compositional complexity of sequence* (7 features). Using SEG program (Wootton et al., 1993), a given chain is analyzed to annotate low complexity regions (LCRs), i.e., regions containing relatively small diversity in their AA composition, and high complexity regions, i.e., regions with a diverse set of AA types. The use of low and high complexity regions is motivated by the observation that LCRs are rarely seen in structured proteins or protein segments, while intrinsically disordered proteins or regions exhibit a significant fraction of LCRs (Romero et al., 2001). Seven features were designed to aggregate the annotations of low and high complexity regions over the input protein chain:
   - LCR_ratio: the ratio of low complexity residues to the total number of residues in the input chain.
   - LongestLCR: length of the longest low complexity region divided by the length of the input sequence.
   - LongestHCR: length of the longest high complexity region divided by length of the input sequence.
   - numLCR: the number of low complexity regions divided by length of the input sequence.
   - numHCR: number of high complexity regions divided by length of the input sequence.
   - avgLCR: average length of all low complexity regions divided by length of the input sequence.
   - avgHCR: average length of all high complexity regions divided by length of the input sequence.
3. *Physicochemical and biochemical properties of AAs that are aggregated over the input sequence* (3,717 features). A list of 531 AA indices representing their propensity for formation of certain secondary structures, solvent accessibility, B-factor, hydrophobicity and charge, polarity, size, etc. were collected from the AAindex database (Kawashima et al., 2008) and used to encode the input protein sequence. This is motivated by the fact that these various properties could be used to differentiate between structures and disordered segments in a protein chain. For example, disordered

regions are characterized by lack of secondary structure (Ferron et al., 2006; He et al., 2009), by having a large solvent accessible area (Liu et al., 2003; Schlessinger et al., 2009) and relatively high B-factors (Zhang et al., 2009), and by relatively low amounts of hydrophobic and bulky AAa and higher numbers of charged AAs (Uversky et al., 2000; Uversky 2002; Radivojac et al., 2004). Seven custom-designed features were computed for each AA index, yielding total of 531*7= 3,717 features:

- maxLCR: maximal value among all average residue-based values computed for the low complexity regions.
- minLCR: minimal value among all average residue-based values computed for the low complexity regions.
- avgLCR: average value over all average residue-based values computed for the low complexity regions.
- avgHCR: average value over all average residue-based values computed for the high complexity regions.
- avg: average value over all residues in the input protein chain.
- max5: maximal value among all average residue-based values computed using 5-residues long sliding windows.
- min5: minimal value among all average residue-based values computed using 5-residues long sliding windows.

4. *Predicted propensity for disorder at the residue level aggregated over the input sequence* (14 features). We aggregate the outputs from the fast and relatively accurate IUPred short (Dosztányi et al., 2005) over the input protein chain to calculate 7 custom-designed features using binary (disordered vs. structured) predictions and another 7 features using real-values propensities. For the binary disorder prediction, we remove (too) short disordered segment (size < 4), similar to (Monastyrskyy et al., 2011), and using the remaining predicted disordered segments we calculate:
   - Dis_ratio: fraction of predicted disordered residues in the input chain.
   - longestDis: length of the longest predicted disordered segment in the input chain.
   - longestOrd: length of the longest predicted ordered segment in the input chain.
   - numDisSeg: number of predicted disordered segments in the input chain.
   - numOrdSeg: number of predicted ordered segments in the input chain.
   - avgDisSegL: average length of predicted disordered segments in the input chain.
   - avgOrdSegL: average length of predicted ordered segments in the input chain.
   - For the real-values propensities generated by IUPred we calculate:
   - maxDisProbab: maximal value among all average residue-based values computed for the predicted disordered segments.
   - minDisProbab: minimal value among all average residue-based values computed for the predicted disordered segments.
   - avgAvgDisProbab: average value over all average residue-based values computed for the predicted disordered segments.
   - avgAvgOrdProbab: average value over all average residue-based values computed for the predicted ordered/structured segments.
   - avg: average value over the entire input sequence.
   - max5: maximal value among all average residue-based values computed using 5-residues long sliding windows.
   - min5: minimal value among all average residue-based values computed using 5-residues long sliding windows.

The above four feature sets total to 20 + 7 + 3,717 + 14 = 3,758 features.

SUPPLEMENTARY FIGURES

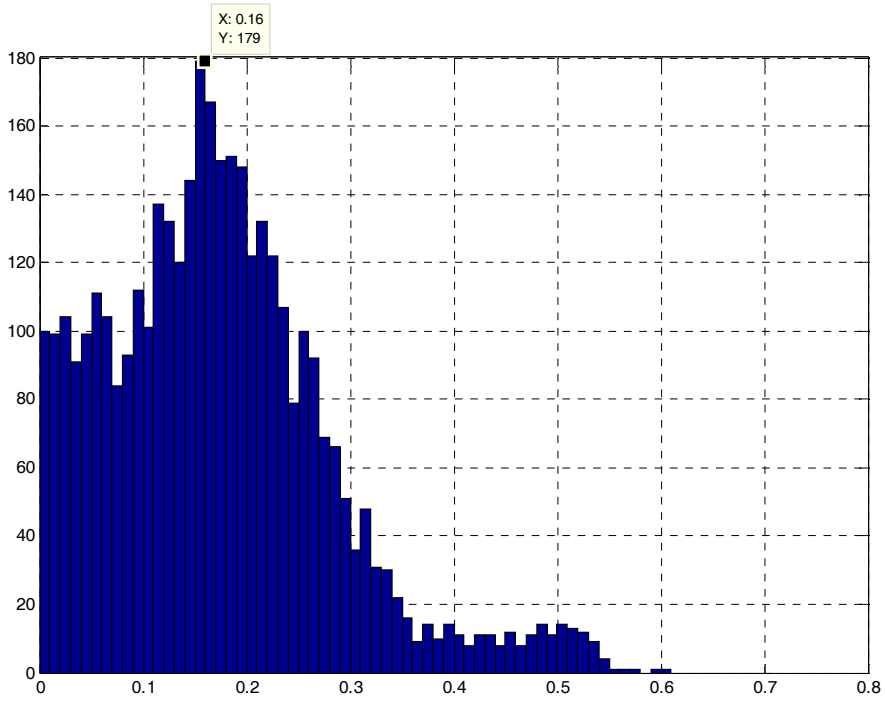

Figure S1. Histogram of the absolute PCC values shown on the *x*-axis (averages over the five training folds based on the 5-fold cross validation on the TRAINING dataset) between the considered 3,758 features and the native disorder content. The *y*-axis shows the count of features for a given value of PCC.

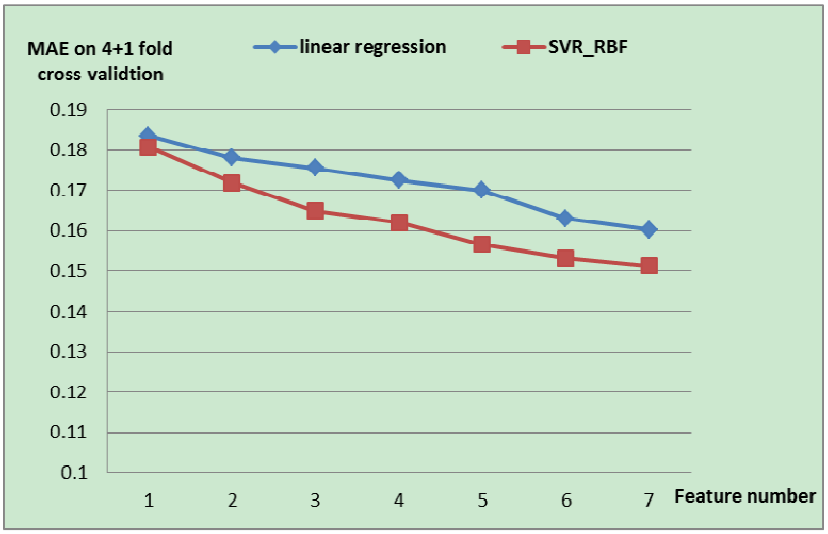

Figure S2. MAE values (*y*-axis) computed using the 4+1-fold cross validation on the TRAINING dataset during the sequential forward selection in the second step of the feature selection for the linear ridge regression (blue line with diamond markers) and support vector regression with RBF kernel (SVR_RBF) (red line with square markers) models. The *x*-axis shows the number of added features.

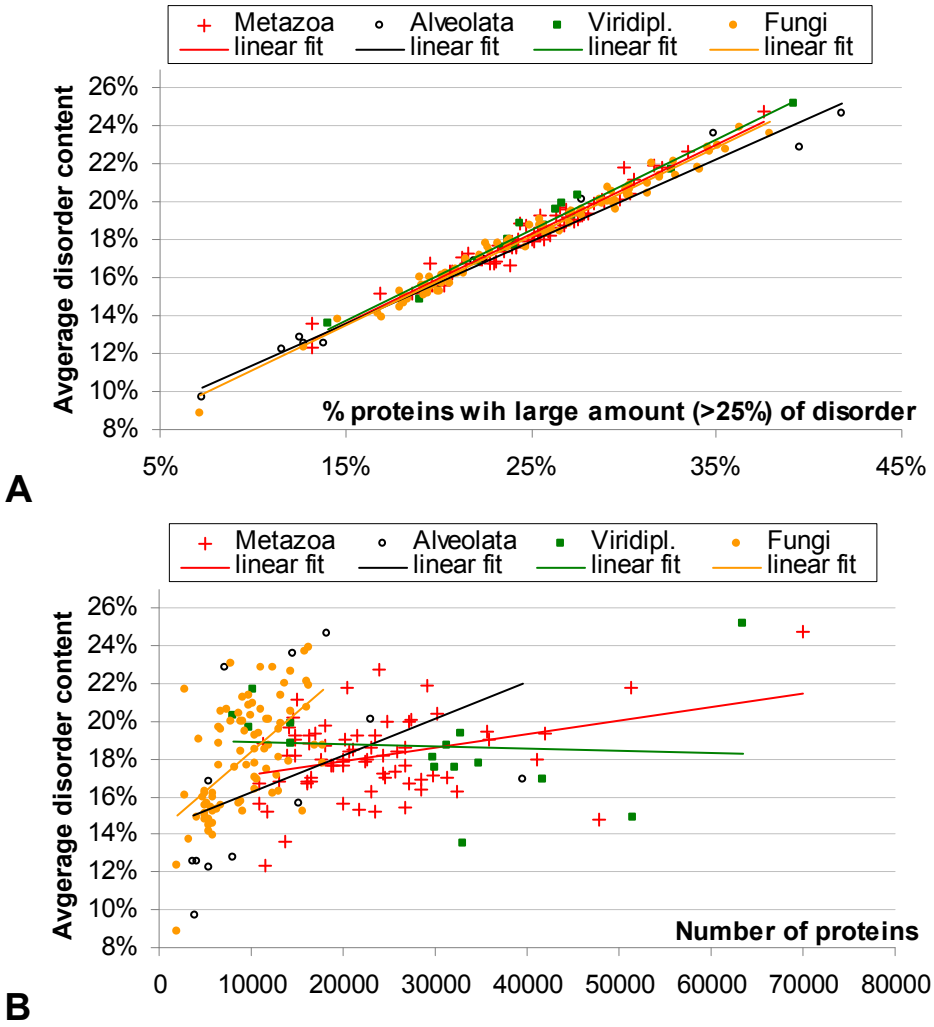

Figure S3. Scatter plots of relation between predicted disorder content and (panel A) fraction of proteins with large (>25%) disordered content, (panel B) number of proteins, for the eukaryotic proteomes. Eukaryotic proteomes are divided by their kingdoms/phyla into *Alveolata*, *Fungi*, *Metazoa*, and *Viridiplantae*.

SUPPLEMENTARY TABLES

Table S1. The MAE values generated utilizing the grid search for parameters corresponding to the lowest MAE value on the 4+1-fold cross validation on the TRAINING dataset. The initial parameterization (the top portion of the table) was performed using the top 5% features selected in the first step of the feature selection. The final parameterization (the bottom portion of the table) was performed on the selected subset of 7 features.

Type of parameterization	Type of test / parameters	Linear ridge regression		Support vector regression	
		Default	Best selected	Default	Best selected
Initial parameterization	parameter values	r = 1.0E-8	r = 1	C = 1, γ= 0.01	C = 2, γ= 0.25
	results for 4-fold cross validation	0.168	0.165	0.160	0.154
	results on the 5th fold	0.182	0.182	0.211	0.170
	results for 4+1-fold cross validation (average 4-fold cross-validation & 5th fold)	0.175	0.174	0.186	0.162
Final parameterization	parameter values	r = 1	r = 0.1	C = 2, γ = 0.25	C = 2, γ = 0.25
	results for 4-fold cross validation	0.159	0.159	0.140	0.140
	results on the 5th fold	0.162	0.161	0.162	0.162
	results for 4+1-fold cross validation (average 4-fold cross-validation & 5th fold)	0.160	0.160	0.151	0.151

Table S2. Features used by RAPID including feature name, brief description, average absolute PCC with native disorder content on the TRAINING dataset, and MAE value when they are used individually to perform prediction on the TRAINING dataset. Features are sorted by their MAE value.

Feature name	Brief description	PCC	MAE
IUPred_p_avg	Average value, over the entire input sequence, of the real-values propensities predicted by IUPred	0.601	0.181
IUPred_b_avgDisSegL	Average length of predicted disordered segments in the input chain predicted by IUPred	0.485	0.188
VINM940104_avg	Average value, over the entire input sequence, of the VINM940104 AA index, which quantifies normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbours	0.492	0.199
CORJ870108_avgLCR	Average value, over all average residue-based values computed for the low complexity regions extracted with SEG, of the VINM940104 AA index, which quantifies hydrophobicity	0.396	0.206
RACS820102_avg	Average value, over the entire input sequence, of the RACS820102 AA index, which quantifies relative fractional occurrence in right-handed helix/bend regions	0.309	0.208
IUPred_b_numOrdSeg	Number of ordered segments in the input chain predicted by IUPred	0.407	0.214
OOBM850105_avg	Average value, over the entire input sequence, of the OOBM850105 AA index, which quantifies optimized side chain interactions	0.359	0.214

Table S3. Summary of the 201 eukaryotic species sorted alphabetically by their phyla following by the name of species. The proteomes are assigned to their taxonomic lineage based on NCBI (Geer et al., 2010) where the lowest taxonomic level, which we refer to as "species", could be the genus, family or species.

Abbreviated taxonomy (domain_phyla_species)	Phyla	Taxonomy ID	Number of proteins
Eukaryota_Alveolata_Apicomplexa_Theileria	Alveolata	5875	4071
Eukaryota_Alveolata_Ciliophora_Tetrahymena	Alveolata	312017	15283
Eukaryota_Alveolata_Apicomplexa_Plasmodium (Laverania)	Alveolata	36329	5354
Eukaryota_Alveolata_Apicomplexa_Cryptosporidium	Alveolata	441375	3930
Eukaryota_Alveolata_Apicomplexa_Plasmodium (Vinckeia)	Alveolata	5825	14533
Eukaryota_Alveolata_Ciliophora_Ichthyophthirius	Alveolata	857967	8049
Eukaryota_Alveolata_Apicomplexa_Babesia	Alveolata	5865	3687
Eukaryota_Alveolata_Apicomplexa_Plasmodium (Plasmodium)	Alveolata	126793	5389
Eukaryota_Alveolata_Ciliophora_Paramecium	Alveolata	5888	39461
Eukaryota_Alveolata_Perkinsea_Perkinsus	Alveolata	423536	23114
Eukaryota_Alveolata_Apicomplexa_Neospora	Alveolata	572307	7110
Eukaryota_Alveolata_Apicomplexa_Toxoplasma	Alveolata	5811	18246
Eukaryota_Amoebozoa_Mycetozoa_Polysphondylium	Amoebozoa	13642	12351
Eukaryota_Amoebozoa_Mycetozoa_Dictyostelium	Amoebozoa	44689	12744
Eukaryota_Amoebozoa_Archamoebae_Entamoeba	Amoebozoa	370354	8606
Eukaryota_Choanoflagellida_Salpingoecidae_Salpingoeca	Choanoflagellida	946362	11698
Eukaryota_Choanoflagellida_Codonosigidae_Monosiga	Choanoflagellida	81824	9188
Eukaryota_Diplomonadida_Hexamitidae_Giardia	Diplomonadida	184922	7154
Eukaryota_Euglenozoa_Kinetoplastida_Schizotrypanum	Euglenozoa	353153	19242
Eukaryota_Euglenozoa_Kinetoplastida_Nannomonas	Euglenozoa	1068625	5904
Eukaryota_Euglenozoa_Kinetoplastida_Trypanosoma	Euglenozoa	679716	9668
Eukaryota_Euglenozoa_Kinetoplastida_Duttonella	Euglenozoa	1055687	3771
Eukaryota_Euglenozoa_Kinetoplastida_Leishmania braziliensis species complex	Euglenozoa	5660	8101
Eukaryota_Euglenozoa_Kinetoplastida_Leishmania	Euglenozoa	5671	8048
Eukaryota_Fungi_Microsporidia_Nosema	Fungi	578460	2060
Eukaryota_Fungi_Microsporidia_Enterocytozoon	Fungi	481877	3311
Eukaryota_Fungi_Dikarya_Candida	Fungi	237561	9118
Eukaryota_Fungi_Dikarya_Scheffersomyces	Fungi	322104	5797
Eukaryota_Fungi_Microsporidia_Encephalitozoon	Fungi	284813	1909
Eukaryota_Fungi_Dikarya_Schizosaccharomyces	Fungi	284812	5090
Eukaryota_Fungi_Dikarya_Spathaspora	Fungi	619300	5973
Eukaryota_Fungi_Dikarya_Penicillium	Fungi	441960	10422
Eukaryota_Fungi_Dikarya_Botryotinia	Fungi	332648	16365
Eukaryota_Fungi_Dikarya_Vanderwaltozyma	Fungi	436907	5352
Eukaryota_Fungi_Dikarya_Kazachstania	Fungi	1071382	5354
Eukaryota_Fungi_Dikarya_Sclerotinia	Fungi	665079	14400
Eukaryota_Fungi_Dikarya_Meyerozyma	Fungi	294746	5919
Eukaryota_Fungi_Dikarya_Debaryomyces	Fungi	284592	6242
Eukaryota_Fungi_Microsporidia_Nematocida	Fungi	944018	2769
Eukaryota_Fungi_Dikarya_Serpula	Fungi	936435	14339
Eukaryota_Fungi_Dikarya_Clavispora	Fungi	306902	5932
Eukaryota_Fungi_Dikarya_Talaromyces	Fungi	441959	12996
Eukaryota_Fungi_Dikarya_Tetrapisispora	Fungi	1071380	5385
Eukaryota_Fungi_Dikarya_Neosartorya	Fungi	331117	10392
Eukaryota_Fungi_Dikarya_Hypocrea	Fungi	413071	12389
Eukaryota_Fungi_Dikarya_Orbilia	Fungi	756982	11478
Eukaryota_Fungi_Dikarya_Podospora	Fungi	515849	10523
Eukaryota_Fungi_Chytridiomycota_Batrachochytrium	Fungi	684364	8610
Eukaryota_Fungi_Dikarya_Glarea	Fungi	1104152	7904
Eukaryota_Fungi_Dikarya_Yarrowia	Fungi	284591	6414
Eukaryota_Fungi_Dikarya_Naumovozyma	Fungi	1064592	5565
Eukaryota_Fungi_Dikarya_Aspergillus	Fungi	425011	14072
Eukaryota_Fungi_Dikarya_Melampsora	Fungi	747676	16237
Eukaryota_Fungi_Dikarya_Saccharomyces	Fungi	559292	6652
Eukaryota_Fungi_Dikarya_Glomerella	Fungi	645133	12019
Eukaryota_Fungi_Dikarya_Lodderomyces	Fungi	379508	5779
Eukaryota_Fungi_Dikarya_Kluyveromyces	Fungi	284590	5071
Eukaryota_Fungi_Dikarya_Piriformospora	Fungi	1109443	11765
Eukaryota_Fungi_Dikarya_Laccaria	Fungi	486041	17895
Eukaryota_Fungi_Dikarya_Millerozyma	Fungi	559304	8851
Eukaryota_Fungi_Dikarya_Tuber	Fungi	656061	7476
Eukaryota_Fungi_Dikarya_Trichoderma	Fungi	452589	11815
Eukaryota_Fungi_Dikarya_Ogataea	Fungi	871575	4171

Eukaryota_Fungi_Dikarya_Grosmannia	Fungi	655863	8311
Eukaryota_Fungi_Dikarya_mitosporic Nakaseomyces	Fungi	284593	5197
Eukaryota_Fungi_Dikarya_Penicillium chrysogenum complex	Fungi	500485	12771
Eukaryota_Fungi_Dikarya_Torulaspora	Fungi	1076872	4966
Eukaryota_Fungi_Fungi incertae sedis_Rhizopus	Fungi	246409	16968
Eukaryota_Fungi_Dikarya_Colletotrichum	Fungi	759273	16110
Eukaryota_Fungi_Dikarya_Mixia	Fungi	764103	6724
Eukaryota_Fungi_Dikarya_Moniliophthora	Fungi	554373	13648
Eukaryota_Fungi_Dikarya_Emericella	Fungi	227321	10523
Eukaryota_Fungi_Dikarya_Komagataella	Fungi	644223	5072
Eukaryota_Fungi_Dikarya_Coprinopsis	Fungi	240176	13335
Eukaryota_Fungi_Dikarya_Pyrenophora	Fungi	426418.00	12062
Eukaryota_Fungi_Dikarya_Thielavia	Fungi	578455	9760
Eukaryota_Fungi_Dikarya_Nectria haematococca complex	Fungi	660122	15709
Eukaryota_Fungi_Dikarya_Postia	Fungi	561896	8983
Eukaryota_Fungi_Dikarya_Zymoseptoria	Fungi	336722	10972
Eukaryota_Fungi_Dikarya_Verticillium	Fungi	498257	10530
Eukaryota_Fungi_Dikarya_Zygosaccharomyces	Fungi	559307	4987
Eukaryota_Fungi_Dikarya_Filobasidiella or Cryptococcus neoformans species complex	Fungi	283643	6569
Eukaryota_Fungi_Dikarya_Schizophyllum	Fungi	578458	13128
Eukaryota_Fungi_Dikarya_Lachancea	Fungi	559295	5093
Eukaryota_Fungi_Dikarya_Sordaria	Fungi	771870	9890
Eukaryota_Fungi_Dikarya_Gibberella	Fungi	229533	13139
Eukaryota_Fungi_Dikarya_Leptosphaeria maculans complex	Fungi	985895	12468
Eukaryota_Fungi_Dikarya_Metarhizium	Fungi	655844	10581
Eukaryota_Fungi_Dikarya_Phaeosphaeria	Fungi	321614	15998
Eukaryota_Fungi_Dikarya_Fusarium oxysporum species complex	Fungi	660025	17725
Eukaryota_Fungi_Dikarya_Neurospora	Fungi	510952	11177
Eukaryota_Fungi_Dikarya_Exophiala	Fungi	858893	9391
Eukaryota_Fungi_Dikarya_Chaetomium	Fungi	306901	11041
Eukaryota_Fungi_Dikarya_Ajellomyces	Fungi	653446	10089
Eukaryota_Fungi_Dikarya_Cordyceps	Fungi	983644	9651
Eukaryota_Fungi_Dikarya_Magnaporthe	Fungi	242507	13290
Eukaryota_Fungi_Dikarya_Arthroderma	Fungi	535722	8907
Eukaryota_Fungi_Dikarya_Myceliophthora	Fungi	573729	9081
Eukaryota_Fungi_Dikarya_Uncinocarpus	Fungi	336963	7760
Eukaryota_Fungi_Dikarya_Paracoccidioides	Fungi	502779	9114
Eukaryota_Fungi_Dikarya_Puccinia	Fungi	418459	15808
Eukaryota_Fungi_Dikarya_Sporisorium	Fungi	999809.00	6673
Eukaryota_Fungi_Dikarya_Coccidioides	Fungi	443226	10212
Eukaryota_Fungi_Dikarya_Trichophyton	Fungi	559305	8705
Eukaryota_Fungi_Dikarya_Eremothecium	Fungi	284811	4761
Eukaryota_Fungi_Dikarya_Rhodotorula	Fungi	1001064	2831
Eukaryota_Fungi_Dikarya_Ustilago	Fungi	237631	6548
Eukaryota_Fungi_Dikarya_Malassezia	Fungi	425265	4272
Eukaryota_Heterolobosea_Schizopyrenida_Naegleria	Heterolobosea	5762	15636
Eukaryota_Ichthyosporea_Capsaspora	Ichthyosporea	595528	8374
Eukaryota_Metazoa_Porifera_Amphimedon	Metazoa	400682	29678
Eukaryota_Metazoa_Placozoa_Trichoplax	Metazoa	10228	11519
Eukaryota_Metazoa_Cnidaria_Nematostella	Metazoa	45351	24435
Eukaryota_Metazoa_Chordata_Ciona	Metazoa	51511	20004
Eukaryota_Metazoa_Arthropoda_Pediculus	Metazoa	121224	10763
Eukaryota_Metazoa_Arthropoda_Tribolium	Metazoa	7070	16501
Eukaryota_Metazoa_Nematoda_Caenorhabditis	Metazoa	31234	31232
Eukaryota_Metazoa_Chordata_Silurana	Metazoa	8364	23528
Eukaryota_Metazoa_Arthropoda_Danaus	Metazoa	13037	16251
Eukaryota_Metazoa_Arthropoda_Stegomyia	Metazoa	7159	16046
Eukaryota_Metazoa_Chordata_Branchiostoma	Metazoa	7739	28544
Eukaryota_Metazoa_Arthropoda_Apis	Metazoa	7460	10953
Eukaryota_Metazoa_Arthropoda_Bombyx	Metazoa	7091	14760
Eukaryota_Metazoa_Echinodermata_Strongylocentrotus	Metazoa	7668	28505
Eukaryota_Metazoa_Arthropoda_Culex	Metazoa	7176	18703
Eukaryota_Metazoa_Arthropoda_Sophophora	Metazoa	7227	17534
Eukaryota_Metazoa_Arthropoda_Anopheles	Metazoa	7165	13072
Eukaryota_Metazoa_Chordata_Danio	Metazoa	7955	41030
Eukaryota_Metazoa_Chordata_Oreochromis	Metazoa	8128	26733
Eukaryota_Metazoa_Arthropoda_Daphnia	Metazoa	6669	30137
Eukaryota_Metazoa_Platyhelminthes_Schistosoma	Metazoa	6183	11677
Eukaryota_Metazoa_Chordata_Latimeria	Metazoa	7897	21734
Eukaryota_Metazoa_Chordata_Oryzias	Metazoa	8090	24625
Eukaryota_Metazoa_Arthropoda_Atta	Metazoa	12957	18068
Eukaryota_Metazoa_Arthropoda_Harpegnathos	Metazoa	610380	15024

Eukaryota_Metazoa_Arthropoda_Solenopsis	Metazoa	13686	14177
Eukaryota_Metazoa_Arthropoda_Camponotus	Metazoa	104421	14782
Eukaryota_Metazoa_Arthropoda_Hawaiian Drosophila	Metazoa	7222	14754
Eukaryota_Metazoa_Arthropoda_Drosophila	Metazoa	7230	14525
Eukaryota_Metazoa_Chordata_Bos	Metazoa	9913	26659
Eukaryota_Metazoa_Chordata_Takifugu	Metazoa	31033	47821
Eukaryota_Metazoa_Chordata_Gasterosteus	Metazoa	69293	27249
Eukaryota_Metazoa_Arthropoda_Acromyrmex	Metazoa	103372	13962
Eukaryota_Metazoa_Nematoda_Loa	Metazoa	7209	16271
Eukaryota_Metazoa_Arthropoda_Ixodes	Metazoa	6945	20473
Eukaryota_Metazoa_Chordata_Oikopleura	Metazoa	34765	17020
Eukaryota_Metazoa_Chordata_Sarcophilus	Metazoa	9305	22392
Eukaryota_Metazoa_Nematoda_Brugia	Metazoa	6279	11338
Eukaryota_Metazoa_Chordata_Rattus	Metazoa	10116	35788
Eukaryota_Metazoa_Chordata_Tetraodon	Metazoa	99883	23134
Eukaryota_Metazoa_Nematoda_Pristionchus	Metazoa	54126	29079
Eukaryota_Metazoa_Chordata_Canis	Metazoa	9615	25839
Eukaryota_Metazoa_Chordata_Meleagris	Metazoa	9103	16534
Eukaryota_Metazoa_Nematoda_Trichinella	Metazoa	6334	15988
Eukaryota_Metazoa_Chordata_Gallus	Metazoa	9031	23429
Eukaryota_Metazoa_Chordata_Monodelphis	Metazoa	13616	32441
Eukaryota_Metazoa_Chordata_Oryctolagus	Metazoa	9986	24391
Eukaryota_Metazoa_Chordata_Ailuropoda	Metazoa	9646	21142
Eukaryota_Metazoa_Chordata_Anolis	Metazoa	28377	18867
Eukaryota_Metazoa_Chordata_Cavia	Metazoa	10141	19906
Eukaryota_Metazoa_Chordata_Ictidomys	Metazoa	43179	19945
Eukaryota_Metazoa_Chordata_Otolemur	Metazoa	30611	19932
Eukaryota_Metazoa_Chordata_Loxodonta	Metazoa	9785	25624
Eukaryota_Metazoa_Chordata_Sus	Metazoa	9823	27365
Eukaryota_Metazoa_Chordata_Taeniopygia	Metazoa	59729	18142
Eukaryota_Metazoa_Chordata_Gorilla	Metazoa	9595	27240
Eukaryota_Metazoa_Chordata_Macaca	Metazoa	9544	35611
Eukaryota_Metazoa_Chordata_Myotis	Metazoa	59463	20647
Eukaryota_Metazoa_Chordata_Equus	Metazoa	9796	22678
Eukaryota_Metazoa_Chordata_Ornithorhynchus	Metazoa	9258	26783
Eukaryota_Metazoa_Chordata_Pan	Metazoa	9598	20128
Eukaryota_Metazoa_Chordata_Callithrix	Metazoa	9483	42009
Eukaryota_Metazoa_Chordata_Cricetulus	Metazoa	10029	23872
Eukaryota_Metazoa_Chordata_Pongo	Metazoa	9601	24705
Eukaryota_Metazoa_Chordata_Homo	Metazoa	9606	69917
Eukaryota_Metazoa_Chordata_Mus	Metazoa	10090	51398
Eukaryota_Metazoa_Chordata_Nomascus	Metazoa	61853	23146
Eukaryota_Metazoa_Chordata_Heterocephalus	Metazoa	10181	21421
Eukaryota_Metazoa_Platyhelminthes_Clonorchis	Metazoa	79923	13606
Eukaryota_Parabasalia_Trichomonadida_Trichomonas	Parabasalia	5722	50191
Eukaryota_stramenopiles_Blastocystis	Stramenopiles	12968	5820
Eukaryota_stramenopiles_Oomycetes_Phytophthora	Stramenopiles	1094619	25721
Eukaryota_stramenopiles_Bacillariophyta_Thalassiosira	Stramenopiles	35128	11718
Eukaryota_stramenopiles_Bacillariophyta_Phaeodactylum	Stramenopiles	556484	10465
Eukaryota_stramenopiles_Pelagophyceae_Aureococcus	Stramenopiles	44056	11501
Eukaryota_stramenopiles_PX_clade_Ectocarpus	Stramenopiles	2880	16334
Eukaryota_Viridiplantae_Streptophyta_Glycine	Viridiplantae	3847	51558
Eukaryota_Viridiplantae_Streptophyta_Arabidopsis	Viridiplantae	81972	32113
Eukaryota_Viridiplantae_Streptophyta_Vitis	Viridiplantae	29760	29729
Eukaryota_Viridiplantae_Streptophyta_Populus	Viridiplantae	3694	41794
Eukaryota_Viridiplantae_Streptophyta_Ricinus	Viridiplantae	3988	31255
Eukaryota_Viridiplantae_Streptophyta_Selaginella	Viridiplantae	88036	33150
Eukaryota_Viridiplantae_Streptophyta_Sorghum	Viridiplantae	4558	32812
Eukaryota_Viridiplantae_Streptophyta_Physcomitrella	Viridiplantae	145481	34837
Eukaryota_Viridiplantae_Streptophyta_Brachypodium	Viridiplantae	15368	29933
Eukaryota_Viridiplantae_Streptophyta_Oryza	Viridiplantae	39947	63544
Eukaryota_Viridiplantae_Chlorophyta_Volvox	Viridiplantae	3067	14335
Eukaryota_Viridiplantae_Chlorophyta_Chlamydomonas	Viridiplantae	3055	14336
Eukaryota_Viridiplantae_Chlorophyta_Chlorella	Viridiplantae	554065	9831
Eukaryota_Viridiplantae_Chlorophyta_Micromonas	Viridiplantae	564608	10250
Eukaryota_Viridiplantae_Chlorophyta_Ostreococcus	Viridiplantae	70448	7966

REFERENCES

Dosztányi Z, Csizmók V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* 2005; 347(4):827-39.

Dunker AK, et al. Intrinsically disordered protein. *J Mol Graph Model*. 2001; 19(1):26-59

Ferron F, Longhi S, Canard B, Karlin D. A practical overview of protein disorder prediction methods. *Proteins* 2006; 65(1):1-14.

He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK. Predicting intrinsic disorder in proteins: an overview. *Cell Research* 2009; 19:929-949.

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, and Kanehisa M. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*. 2008; 36, D202-05.

Liu J, Rost B. NORSp: predictions of long regions without regular secondary structure. *Nucleic Acids Res* 2003; 31(13):3833-5.

Monastyrskyy B, Fidelis K, Moult J, Tramontano A, Kryshtafovych A Evaluation of disorder predictions in CASP9. Proteins 2011; 79(Sup10):107-18.

Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK. Protein flexibility and intrinsic disorder. *Protein Sci* 2004; 13(1):71-80.

Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. *Proteins* 2001; 42(1):38-48.

Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B. Improved disorder prediction by combination of orthogonal approaches. *PLoS One* 2009; 4(2):e4433.

Uversky VN. What does it mean to be natively unfolded? *Eur J Biochem* 2002; 269(1):2-12.

Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* 2000; 41(3):415-27.

Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* 1993; 17:149-163.

Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L. On the relation between residue flexibility and local solvent accessibility in proteins. *Proteins* 2009; 76(3):617-36.