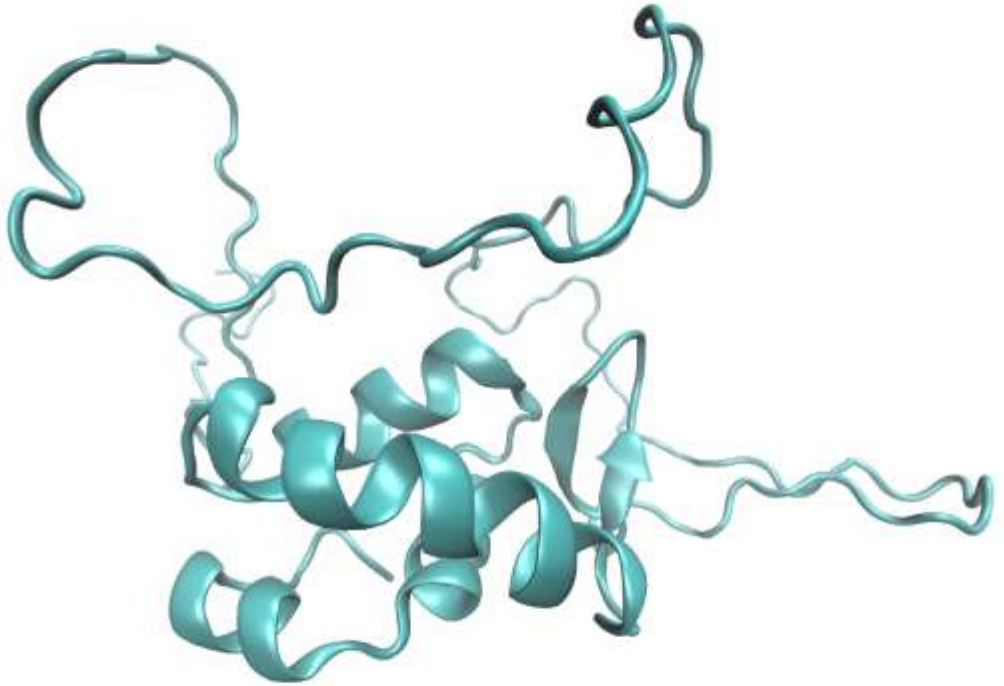
PREDICTION OF DISORDER IN PROTEIN STRUCTURE

Amit Singh
Bioinformatician
Central University of Punjab



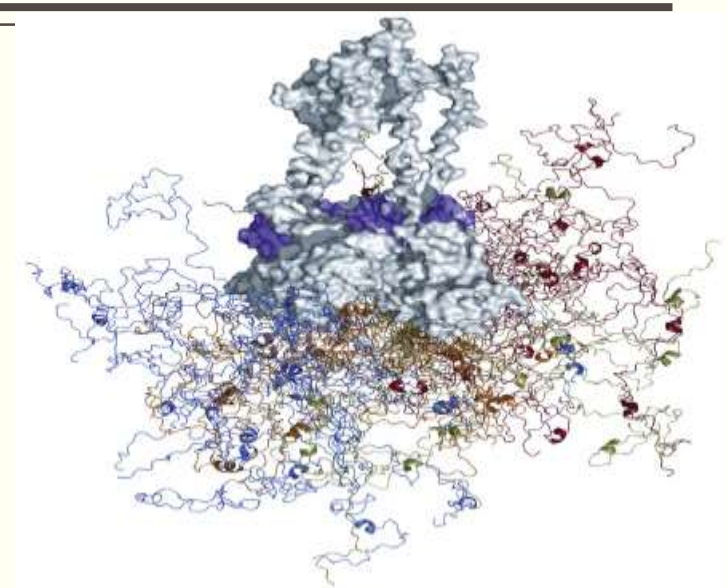
http://www.ceitec.eu/ceitec-mu/protein-structure-and-dynamics/rg

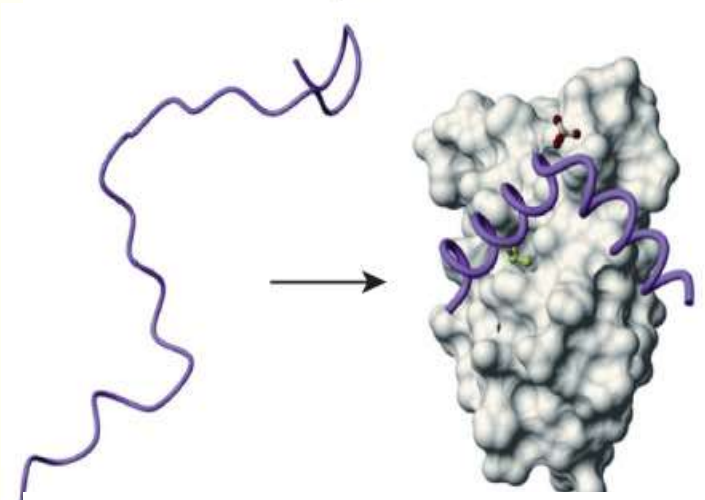
Contents

- What are intrinsically disordered proteins ?
- Why to predict ?
- Bioinformatics approach in prediction of disorder.
- Screenshots of prediction software.

Intrinsically Disordered/Unfolded Proteins(IDP/IUP)

- They are characterized by the lack of stable secondary and tertiary structure under physiological conditions and in absence of a binding partner.
- Either completely disordered or contain large disordered region in their native state.
- IUP uses 50% of total surface for interaction with partner as compared to only 5-10% for most ordered proteins.
- 70% of the cases of IUP contains a single sequence continuous segment for binding while IOP have number of fragments for binding.





Intrinsically unstructured proteins and their functions
H. Jane Dyson & Peter E. Wright
Nature Reviews Molecular Cell Biology **6**, 197-208 (March 2005)
doi:10.1038/nrm1589

Why to predict ?

- These proteins are difficult to study experimentally because of the lack of unique structure in the isolated form.
- In X-ray crystallography, crystal packing may enforce certain disordered regions to become ordered, and disordered binding segments are often crystallized in complex with their partner and are classified ordered despite their lack of structure in isolation.
- With NMR,disorder often is concluded from poor signal dispersion, which does not differentiate between random coils and molten globules of high potential to fold in the presence of a partner.

Bioinformatics approach in prediction of disorder.

Pairwise energy content of aa residues.

Frequencies of aa residues and hydrophobic cluster.

Mean packing densities of aa residues.

Pairwise energy content of aa residues.

- Pairwise energy of protein is a function of its amino acid sequenc

$$E = \sum_{ij=1}^{20} M_{ij} C_{ij}$$

M_{ij} is the interaction energy between amino acid t

C_{ij} is the number of interactions between residues

- Energy per amino acid is approximated by :-

$$\frac{E_{\text{estimated}}}{L} = \sum_{ij}^{20} n_i P_{ij} n_j$$

OR

$$e_i^k(\text{estimated}) = N_i^k \sum_{j=1}^{20} P_{ij} n_j^k$$

N_i denote the number of amino acid residues of type i in the sequence

$n_i = N_i / L$ its frequency, P=energy predictor matrix for i and j

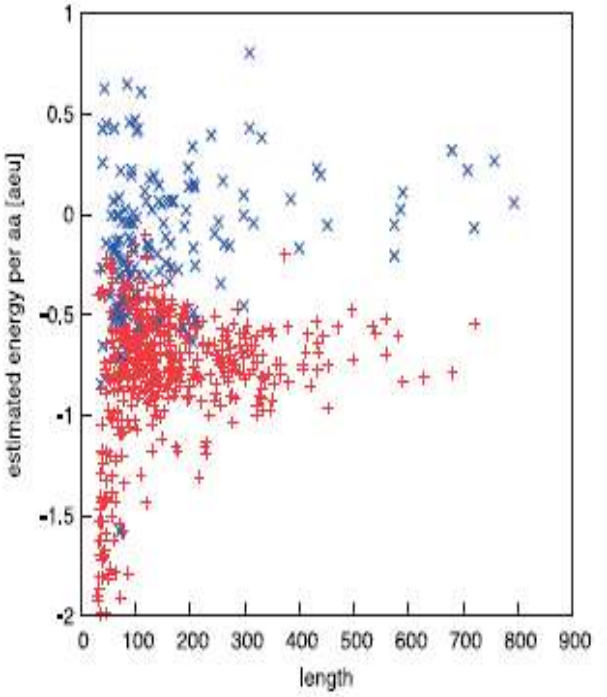


Figure 2. Estimated pairwise interaction energies of globular proteins and IUPs. The total pairwise interaction energy of 559 globular proteins in Filt_Glob_list (red +) and 129 disordered proteins in IUP_list (blue ×) was estimated from their amino acid composition and plotted as a function of their length. Values more negative represent more stabilization due to pairwise amino acid interactions. The average pairwise interaction energy of globular proteins and IUPs are −0.81 and −0.07 [aeu], respectively.

Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 2005;21:3433–3434.

- Total energy of the kth protein into amino acid specific contribution :-

$$E^k = \Sigma e_i^k$$

e_i^k energy of all amino acid residues type *i.*

- e_i^k Depends on the number of contacts this residue makes with other amino acid residues of *j* in the sequence.

$$Z_i = \sum_k (e_i^k - N_i^k \sum_{j=1}^{20} P_{ij} n_j^k)^2$$

- Letting $\partial Z=0$, for all P_{ij} leads to a linear equation which are solved for each amino acid by GSL scientific library.

Table 1. M matrix

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	−0.20	−0.44	0.16	0.26	−0.46	−0.26	0.50	−0.57	0.10	−0.36	−0.22	0.07	0.14	0.01	0.20	−0.09	−0.05	−0.42	0.05	−0.50
C	−0.44	−2.99	0.21	0.19	−0.88	−0.34	−1.11	−0.36	−0.09	−0.53	−0.43	−0.52	−0.14	−0.43	−0.24	0.13	−0.22	−0.62	0.24	−0.79
D	0.16	0.21	0.17	0.55	0.38	0.35	−0.23	0.44	−0.39	0.28	0.35	−0.02	1.03	0.49	−0.37	0.19	−0.12	0.69	0.04	0.43
E	0.26	0.19	0.55	0.60	0.55	0.65	0.18	0.37	−0.47	0.33	0.29	0.01	0.69	0.04	−0.52	0.18	0.37	0.39	0.03	0.17
F	−0.46	−0.88	0.38	0.55	−0.94	0.17	−0.40	−0.88	0.01	−1.08	−0.78	0.22	0.20	0.26	−0.19	−0.22	0.02	−1.15	−0.60	−0.88
G	−0.26	−0.34	0.35	0.65	0.17	−0.12	0.18	0.24	0.19	0.24	0.02	−0.04	0.60	0.46	0.50	0.28	0.28	0.27	0.51	−0.35
H	0.50	−1.11	−0.23	0.18	−0.40	0.18	0.42	−0.00	0.79	−0.24	−0.07	0.20	0.25	0.69	0.24	0.21	0.11	0.16	−0.85	−0.26
I	−0.57	−0.36	0.44	0.37	−0.88	0.24	−0.00	−1.16	0.15	−1.25	−0.58	−0.09	0.36	−0.08	0.14	0.32	−0.27	−1.06	−0.68	−0.85
K	0.10	−0.09	−0.39	−0.47	0.01	0.19	0.79	0.15	0.42	0.13	0.48	0.26	0.50	0.15	0.53	0.10	−0.19	0.10	0.10	0.04
L	−0.36	−0.53	0.28	0.33	−1.08	0.24	−0.24	−1.25	0.13	−1.30	−0.50	0.21	0.42	−0.01	−0.07	0.17	0.07	−0.97	−0.95	−0.63
M	−0.22	−0.43	0.35	0.29	−0.78	0.02	−0.07	−0.58	0.48	−0.50	−0.74	0.32	0.01	0.26	0.15	0.48	0.16	−0.73	−0.56	−1.02
N	0.07	−0.52	−0.02	0.01	0.22	−0.04	0.20	−0.09	0.26	0.21	0.32	0.14	0.27	0.37	0.13	0.15	0.10	0.40	−0.12	0.32
P	0.14	−0.14	1.03	0.69	0.20	0.60	0.25	0.36	0.50	0.42	0.01	0.27	0.27	1.02	0.47	0.54	0.88	−0.02	−0.37	−0.12
Q	0.01	−0.43	0.49	0.04	0.26	0.46	0.69	−0.08	0.15	−0.01	0.26	0.37	1.02	−0.12	0.24	0.29	0.04	−0.11	0.18	0.11
R	0.20	−0.24	−0.37	−0.52	−0.19	0.50	0.24	0.14	0.53	−0.07	0.15	0.13	0.47	0.24	0.17	0.27	0.45	0.01	−0.73	0.01
S	−0.09	0.13	0.19	0.18	−0.22	0.28	0.21	0.32	0.10	0.17	0.48	0.15	0.54	0.29	0.27	−0.06	0.08	0.12	−0.22	−0.14
T	−0.05	−0.22	−0.12	0.37	0.02	0.28	0.11	−0.27	−0.19	0.07	0.16	0.10	0.88	0.04	0.45	0.08	−0.03	−0.01	0.11	−0.32
V	−0.42	−0.62	0.69	0.39	−1.15	0.27	0.16	−1.06	0.10	−0.97	−0.73	0.40	−0.02	−0.11	0.01	0.12	−0.01	−0.89	−0.56	−0.71
W	0.05	0.24	0.04	0.03	−0.60	0.51	−0.85	−0.68	0.10	−0.95	−0.56	−0.12	−0.37	0.18	−0.73	−0.22	0.11	−0.56	−0.05	−1.41
Y	−0.30	−0.79	0.43	0.17	−0.88	−0.35	−0.26	−0.85	0.04	−0.63	−1.02	0.32	−0.12	0.11	0.01	−0.14	−0.32	−0.71	−1.41	−0.76

Contact potential derived from 785 proteins using the approach of Thomas & Dill.[20]

Table 2. P energy predictor matrix

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	−1.65	−2.83	1.16	1.80	−3.73	−0.41	1.90	−3.69	0.49	−3.01	−2.08	0.66	1.54	1.20	0.98	−0.08	0.46	−2.31	0.32	−4.62
C	−2.83	−39.58	−0.82	−0.53	−3.07	−2.96	−4.98	0.34	−1.38	−2.15	1.43	−4.18	−2.13	−2.91	−0.41	−2.33	−1.84	−0.16	4.26	−4.46
D	1.16	−0.82	0.84	1.97	−0.92	0.88	−1.07	0.68	−1.93	0.23	0.61	0.32	3.31	2.67	−2.02	0.91	−0.65	0.94	−0.71	0.90
E	1.80	−0.53	1.97	1.45	0.94	1.31	0.61	1.30	−2.51	1.14	2.53	0.20	1.44	0.10	−3.13	0.81	1.54	0.12	−1.07	1.29
F	−3.73	−3.07	−0.92	0.94	−11.25	0.35	−3.57	−5.88	−0.82	−8.59	−5.34	0.73	0.32	0.77	−0.40	−2.22	0.11	−7.05	−7.09	−8.80
G	−0.41	−2.96	0.88	1.31	0.35	−0.20	1.09	−0.65	−0.16	−0.55	−0.52	−0.32	2.25	1.11	0.84	0.71	0.59	−0.38	1.69	−1.90
H	1.90	−4.98	−1.07	0.61	−3.57	1.09	1.97	−0.71	2.89	−0.86	−0.75	1.84	0.35	2.64	2.05	0.82	−0.01	0.27	−7.58	−3.20
I	−3.69	0.34	0.68	1.30	−5.88	−0.65	−0.71	−6.74	−0.01	−9.01	−3.62	−0.07	0.12	−0.18	0.19	−0.15	0.63	−6.54	−3.78	−5.26
K	0.49	−1.38	−1.93	−2.51	−0.82	−0.16	2.89	−0.01	1.24	0.49	1.61	1.12	0.51	0.43	2.34	0.19	−1.11	0.19	0.02	−1.19
L	−3.01	−2.15	0.23	1.14	−8.59	−0.55	−0.86	−9.01	0.49	−6.37	−2.88	0.97	1.81	−0.58	−0.60	−0.41	0.72	−5.43	−8.31	−4.90
M	−2.08	1.43	0.61	2.53	−5.34	−0.52	−0.75	−3.62	1.61	−2.88	−6.49	0.21	0.75	1.90	2.09	1.39	0.63	−2.59	−6.88	−9.73
N	0.66	−4.18	0.32	0.20	0.73	−0.32	1.84	−0.07	1.12	0.97	0.21	0.61	1.15	1.28	1.08	0.29	0.46	0.93	−0.74	0.93
P	1.54	−2.13	3.31	1.44	0.32	2.25	0.35	0.12	0.51	1.81	0.75	1.15	−0.42	2.97	1.06	1.12	1.65	0.38	−2.06	−2.09
Q	1.20	−2.91	2.67	0.10	0.77	1.11	2.64	−0.18	0.43	−0.58	1.90	1.28	2.97	−1.54	0.91	0.85	−0.07	−1.91	−0.76	0.01
R	0.98	−0.41	−2.02	−3.13	−0.40	0.84	2.05	0.19	2.34	−0.60	2.09	1.08	1.06	0.91	0.21	0.95	0.98	0.08	−5.89	0.36
S	−0.08	−2.33	0.91	0.81	−2.22	0.71	0.82	−0.15	0.19	−0.41	1.39	0.29	1.12	0.85	0.95	−0.48	−0.06	0.13	−3.03	−0.82
T	0.46	−1.84	−0.65	1.54	0.11	0.59	−0.01	0.63	−1.11	0.72	0.63	0.46	1.65	−0.07	0.98	−0.06	−0.96	1.14	−0.65	−0.37
V	−2.31	−0.16	0.94	0.12	−7.05	−0.38	0.27	−6.54	0.19	−5.43	−2.59	0.93	0.38	−1.91	0.08	0.13	1.14	−4.82	−2.13	−3.59
W	0.32	4.26	−0.71	−1.07	−7.09	1.69	−7.58	−3.78	0.02	−8.31	−6.88	−0.74	−2.06	−0.76	−5.89	−3.03	−0.65	−2.13	−1.73	−12.39
Y	−4.62	−4.46	0.90	1.29	−8.80	−1.90	−3.20	−5.26	−1.19	−4.90	−9.73	0.93	−2.09	0.01	0.36	−0.82	−0.37	−3.59	−12.39	−2.68

The pairwise energy per amino acid is estimated as a quadratic form in the amino acid composition vector using the elements of this matrix.

Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 2005;21:3433–3434.

Software based on pairwise energy estimation **IUpred**

- Predict regions that lack a well-defined 3D structure under native conditions.
- The energy and amino acid composition for each position was calculated only by considering interaction partners 2 to 100 residues apart.
- The choice of this range represents the intention of covering most structured domains, but separating distinct domains in multi-domain proteins.
- This procedure yields an estimated energy at position p of type i:

$$e_i^p = \sum_{j=1}^{20} P_{ij}^p n_j^p$$

- where P^p is the position specific energy predictor matrix.
- Software is written in C and interface is PHP.
- Available at http://iupred.enzim.hu/.

Frequencies of aa residues and hydrophobic cluster.

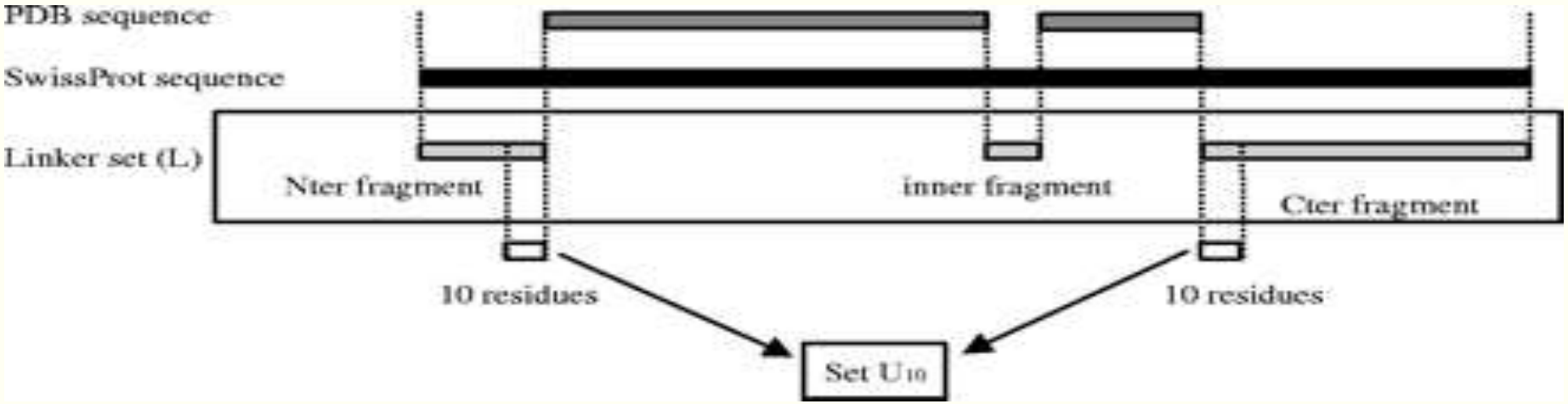
Based on two properties:-

- 1.Disordered regions have a biased composition
- They usually contain either small or no hydrophobic clusters.

System and Methods:-

- Constitution of reference set.
- Ratio and probabilities of aa occurrence
- Cluster distance

Constitution of reference set



- A subset named U10 is extracted from (L), Containing last ten residue of N-terminal fragments and first ten residues of C-terminal fragments.
- Amino acid frequencies in structured and linker region were computed using the two sets S and U10.

Ratio and probabilities of AA occurrence

- The probabilities of occurrence PL and PS of a given sequence in linker and structured regions, respectively, are calculated using a multinomial law:

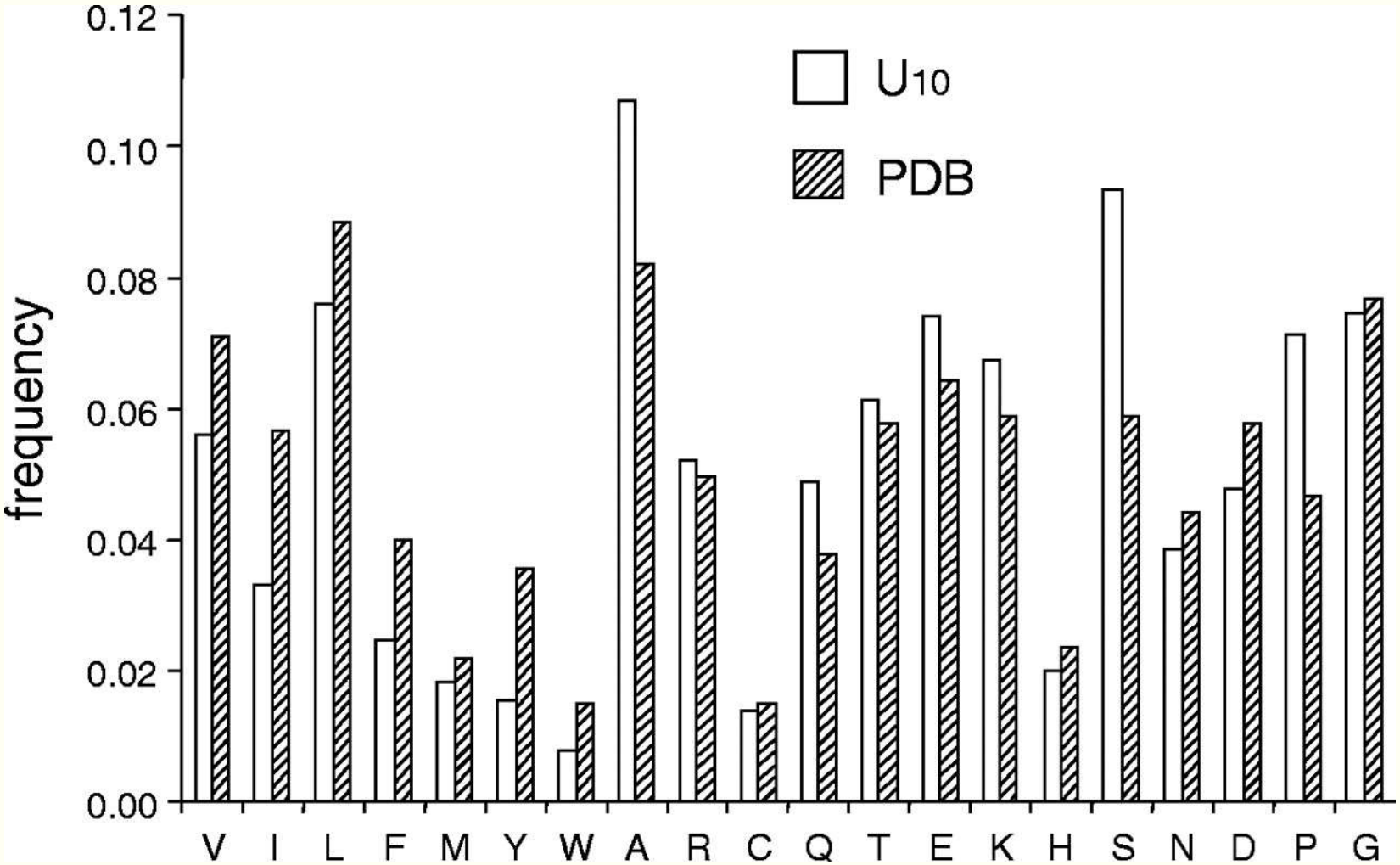
$$\mathrm{PL}\ (\ \text{Linker}\ ) = P(N_{\mathrm{V}} = N_{\mathrm{V}}, \ldots, n_{\mathrm{G}} = n_{\mathrm{G}}) = \frac{n!}{n_{\mathrm{V}}! \ldots n_{\mathrm{G}}!}\, \mathrm{pl}_{\mathrm{V}}^{n_{\mathrm{V}}} \cdots \mathrm{pl}_{\mathrm{G}}^{n_{\mathrm{G}}}$$

$$\mathrm{PS}\ (\ \text{Sequence}\ ) = P(N_{\mathrm{V}} = n_{\mathrm{V}}, \ldots, N_{\mathrm{G}} = n_{\mathrm{G}}) = \frac{n!}{n_{\mathrm{V}}! \ldots n_{\mathrm{G}}!}\, \mathrm{ps}_{\mathrm{V}}^{n_{\mathrm{V}}} \cdots \mathrm{ps}_{\mathrm{G}}^{n_{\mathrm{G}}},$$

(NV,...,NG) are the variables taking as values the numbers (nV, nI,...,nG), of valines, isoleucines,..., glycines in the sequence, $\mathrm{pl}_{\mathrm{V}}^{n_{\mathrm{V}}}$ and $\mathrm{ps}_{\mathrm{V}}^{n_{\mathrm{V}}}$ are the probabilities of occurrence of nv valines in a linker sequence and in a structured sequence, respectively.

- For each sequence if it is more likely to be structured or unstructured, took the ratio of these two probabilities, R = PL/PS.

Amino acids frequencies in the PDB and in the hydrophilic set U10.



Karen Coeytaux, and Anne Poupon Bioinformatics 2005;21:1891-1900

Bioinformatics

Cluster Distance

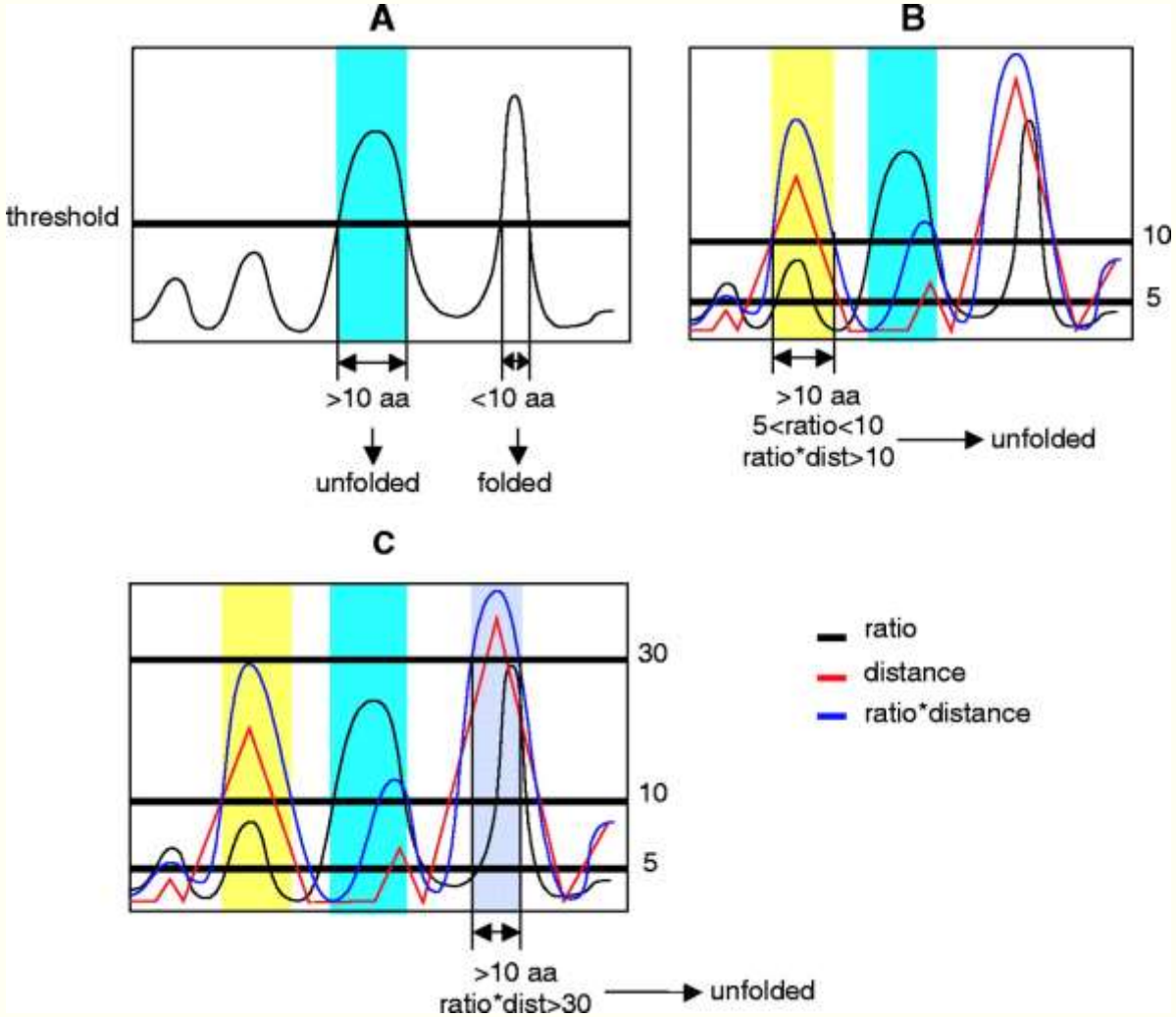
- Sequences were coded into ternary code

  1 for hydrophobic residues (VILFMYW)

  2 for proline and 0 for other amino acids.

- For amino acid in position *i*, we define the cluster distance as being the distance to the closest cluster; the cluster distance is set to 0.5, when *i is* inside a cluster.
- For example, the sequence AGEKTISVVLQLEKEEQ corresponds to the current binary code 0000010111010000O.
- The identification of 1011101 as a hydrophobic cluster corresponding to the sequence ISVVLQL

Rules for prediction of unfolded regions based on the probabilities ratio and the cluster distance.



Karen Coeytaux, and Anne Poupon Bioinformatics 2005;21:1891-1900

Software based on Frequencies of aa residues and hydrophobic cluster. **PreLink.**

- Software is written in C and interface is PHP.
- Available at http://genomics.eu.org.

Mean packing densities of aa residues

Based on two properties:-

Low overall hydrophobicity and a large net charge represents a structural feature of unfolded proteins

The expected average number of contacts per residue for folded and unfolded proteins.

System and Methods

- Construction of a protein database.

of proteins

A database of 90 natively unfolded proteins (http://phys.protres.ru/resources/unfolded_90.html) was based on a published list
A database of 559 globular proteins (http://phys.protres.ru/resources/folded_559.html) was constructed using the PDB codes.

- Average number of close (heavy atoms is less than 8.0 A° apart) residues in the globular state.

close residues (according to Table 1)

The expected average number of close residues was obtained as the total expected number of
divided by the total number of amino acid residues in the protein.

Table 1. Average number of close residues in the globular state as estimated for each of the 20 amino acids

Amino acid	G	P	A	D	E	K	S	N	Q	T
Average number of close residues	17.11	17.43	19.89	17.41	17.46	17.67	18.19	18.49	19.23	19.81
Amino acid	R	H	C	V	M	L	I	Y	F	W
Average number of close residues	21.03	21.72	23.52	23.93	24.82	25.36	25.71	25.93	27.18	28.48

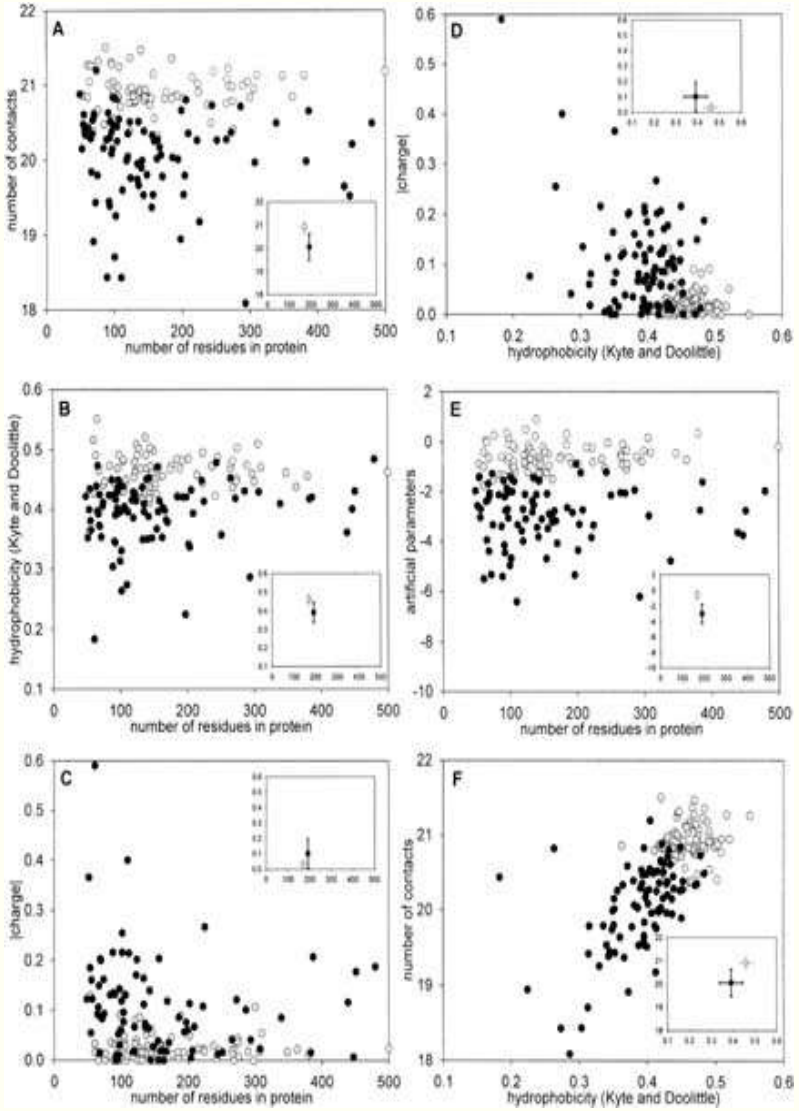
- Hydrophobicity.

We used a published hydrophobicity scale. Average hydrophobicity was computed as the total hydrophobicity of all amino acid residues divided by the total number of residues in the protein.

- Charge.

To compute the net charge of a protein, assumed the charge +1 for Lys and Arg, –1 for Glu and Asp, and 0 for the other residues. The average charge per residue was obtained as the net charge divided by the total number of amino acid residues in the protein.

To be folded or to be unfolded?



Protein Science
Volume 13, Issue 11, pages 2871-2877, 29 DEC 2008 DOI: 10.1110/ps.04881304
http://onlinelibrary.wiley.com/doi/10.1110/ps.04881304/full#fig1

Figure 2.1 Comparison of the mean values of different parameters computed from sequence alone for the set of 90 “natively unfolded” proteins (black circles) and for the set of 80 “ideally” folded proteins (gray circles).

Software based on mean packing densities of aa residues **FoldUnfold**

- Software is written in C and interface is PHP.
- Available at http://skuld.protres.ru/~mlobanov/ogu/ogu.cgi.

Prediction of Intrinsically Unstructured Proteins

- IUPred
- Theory
- How to use
- NEW ANCHOR
- Related links
- Downloads

Intrinsically unstructured/disordered proteins have no single well-defined tertiary structure in their native, functional state. Our server recognizes such regions from the amino acid sequence based on the estimated pairwise energy content. The underlying assumption is that globular proteins are composed of amino acids which have the potential to form a large number of favorable interactions, whereas intrinsically unstructured proteins (IUPs) adopt no stable structure because their amino acid composition does not allow sufficient favorable interactions to form.

Protein Sequence

Enter SWISS-PROT/TrEMBL identifier or accession number:

P01308

or paste the amino acid sequence:

Prediction type:

- long disorder
- short disorder
  (e.g. missing residues of X-ray structures)
- structured regions

Output type:

- raw data only
- generate plot

500 plot window size

SUBMIT CLEAR

>sp|P04637|P53_HUMAN Cellular tumor antigen p53

Critical Assessment of Techniques for Protein Structure Prediction(CASP)

- The performance of various disorder prediction methods was critically assessed in the CASP experiments.
- Evaluation of performance of various predictor was first started during CASP5 on *December 1 - 5th, 2002*
- The broad goals of the CASP5 experiment are to address the following questions about the current state of the art in protein structure prediction:
  - Are the models produced similar to the corresponding experimental structure
  - Is the mapping of the target sequence onto the proposed structure (i.e. the alignment) correct?
  - Have similar structures that a model can be based on been identified?
  - Are the details of the models correct?
  - Has there been progress from the earlier CASPs?
  - What methods are most effective?
  - Where can future effort be most productively focused?

References

Coeytaux K., Poupon A. Prediction of unfolded segments in a protein sequence based on amino acid composition. Bioinformatics 2005;21:1891-1900.

Dosztanyi Z., et al. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J. Mol. Biol. 2005;347:827-839.

Galzitskaya O.V., et al. Optimal region of average side-chain entropy for fast protein folding. Protein Sci. 2000;9:580-586.

Galzitskaya O.V., et al. Prediction of natively unfolded regions in protein chains. Mol. Biol. (Moscow) 2006;40:341-348.

Garbuzynskiy S.O., et al. To be folded or to be unfolded? Protein Sci. 2004;13:2871-2877.

Linding R., et al. Protein disorder prediction: implications for structural proteomics. Structure 2003;11:1453-1459.

Obradovic Z., et al. Predicting intrinsic disorder from amino acid sequence. Proteins 2003;53:566-572.

Obradovic Z., et al. Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins 2005;61:176-182.

Radivojac P., et al. Protein flexibility and intrinsic disorder. Protein Sci. 2004;13:71-80.

Romero P., et al. Thousands of proteins likely to have long disordered regions. Pac. Symp. Biocomput. 1998:437-448.