# Protein disorder analysis of Histone Octomer complex in Neucleosome

**[*]Bejon Kumar Bhowmick**
Dept. of Biochemistry and Molecular Genetics, School of Medicine, University of Virginia, C-Ville, VA, USA
**Email: bkb4n@virginia.edu*

**Received date: 26th November 2012, Accepted date: 29th November 2012, Published date: 1st March 2013**

**Abstract:**
Phosphoproteins are found to be enriched in intrinsic sequence disorder (ID), and this enrichment is related with both cellular location and phosphorylation status. The majority of phosphorylation sites are located outside the structural protein motifs but were mostly located in regions of ID. ID proteins are also very important in certain epigenetic and evolutionary studies. However, few ID proteins were recognized in detection process which remained experimentally laborious and cost effective. Also due to the sensitivity to molecular size, condensation or hydrodynamic attraction, detection process through experiments required much trials and errors. Hence proper computational prediction is an aid to narrow down the screening in small sample pool by which wet lab can practice for error less result. Though, this could be a handy approach but majority of the methods provided contradictory results with one another. Thus, we considered consensus approach and verified the accuracy through real experiments. The sample used was Histone as its widely bearing disordered properties and is easy to handle empirically. Others proteins also could be used in this process.

***Key words:*** *Intrinsic disorder, Consensus method, Histone Neucleosome, Protein structure, Protein function, Evolutionary trend.*

**Introduction:**

ID proteins exist as dynamic ensembles in which the atom positions and the backbone Ramachandran angles vary significantly over time with no specific equilibrium values and typically involve non-cooperative conformational changes. In reality, it was hard to define ID and there was no bombastic mark of ubiquitously agreement for ID proteins. Moreover, different parts of proteins are probably ordered under different circumstances [1]. Despite the fact that, ID proteins fail to form fixed 3-D structure under physiological conditions; they carry out critically important biological functions [2,3] attest to the growing interest for these proteins. Prediction of such significant protein feature accurately is very important. We have developed a consensus prediction system with regard to standard scale to predict most accurate ID regions. ID is a very common element of protein structure [4]; the strength of ID prediction is correlated with sequence complexity [5]; and eukaryotes evidently have a much larger fraction of proteins with ID than eubacteria or archaeobacteria.

Distinctive amino acid biases in ordered regions, short ID regions, and long ID regions indicate that the sequences are the determinants for these flexibility categories differ from one another [6].

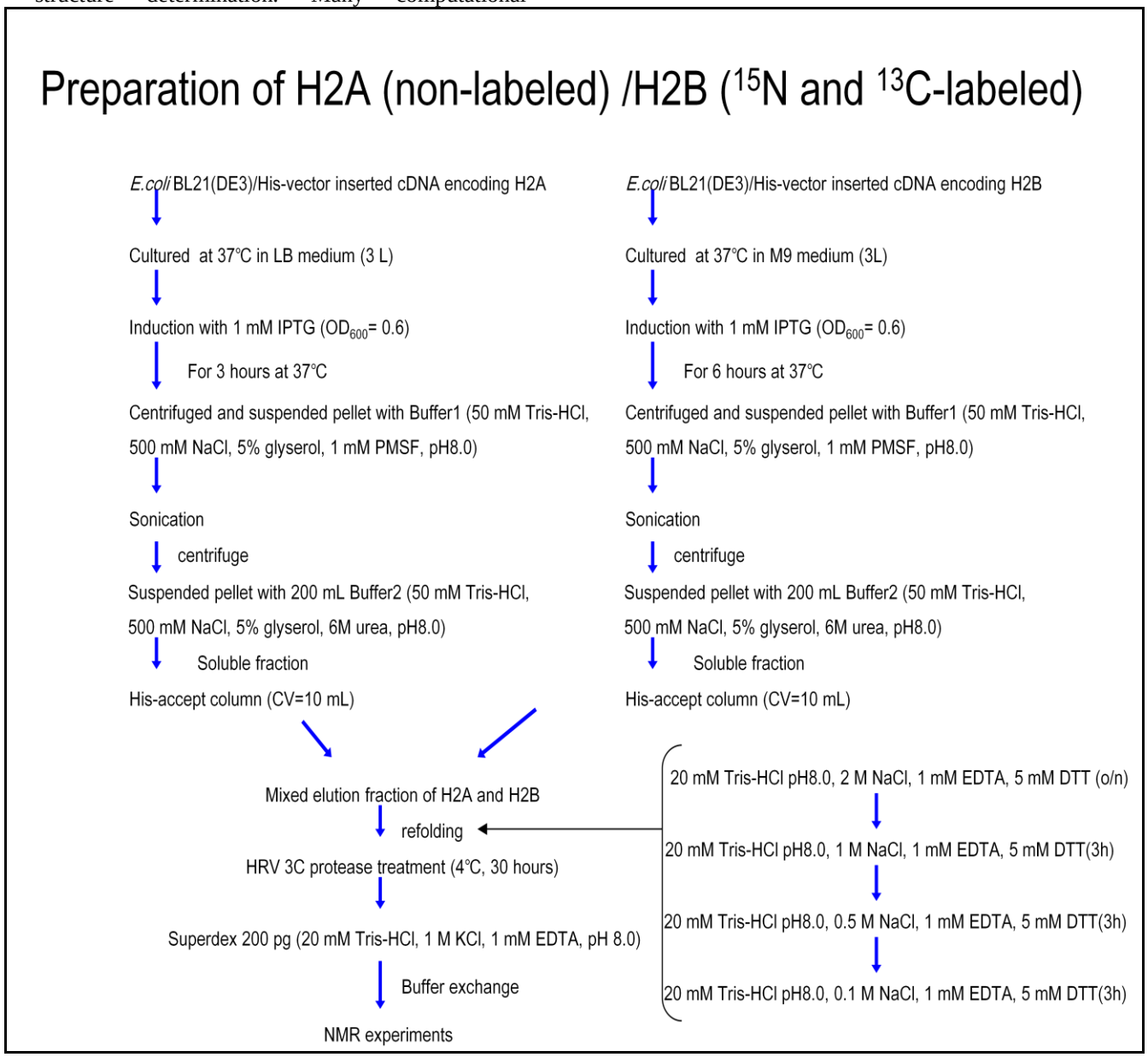
Predicting ID proteins is important because they are thought to carry out various cellular functions even though they have no stable three-dimensional structure. The structure and function of unknown proteins in nature can be inferred by those proteins whose structures have been determined experimentally. By using novel methods, it can accurately predict ID proteins and their functions from a huge amount of structurally-known sequences. [7]. Currently, more than 200 counter examples in which function depends on non folded or incompletely folded regions of protein have been described. It is suggested that the existence of proteins with protein ID calls for a re-assessment of the protein structure-function paradigm [8]. ID protein has the potential to increase significantly the drug discovery rate for new molecule entities [9]. As many as 50% of eukaryotic proteins are likely to contain functionally important long ID regions. Many proteins are wholly ID but still possess numerous biologically important functions. However, the number of experimentally confirmed ID proteins with known biological functions is substantially smaller than their actual number in nature. Therefore, there was a crucial need for novel bioinformatics approach that can be used to combine current knowledge of those (few) experimental results to apply on much larger groups of potential proteins [10].

Structural quality of hub proteins enables them to interact with large numbers of diverse targets. One possibility would be to employ binding regions that have the ability to bind multiple, structurally diverse partners. ID can serve as the structural basis for hub protein promiscuity. They can bind to structured hub proteins; can provide flexible linkers between functional domains with the linkers enabling mechanisms that facilitate binding diversity [11].

Unfolded regions have characteristically high net charge and low hydrophobic traits. The amino acid

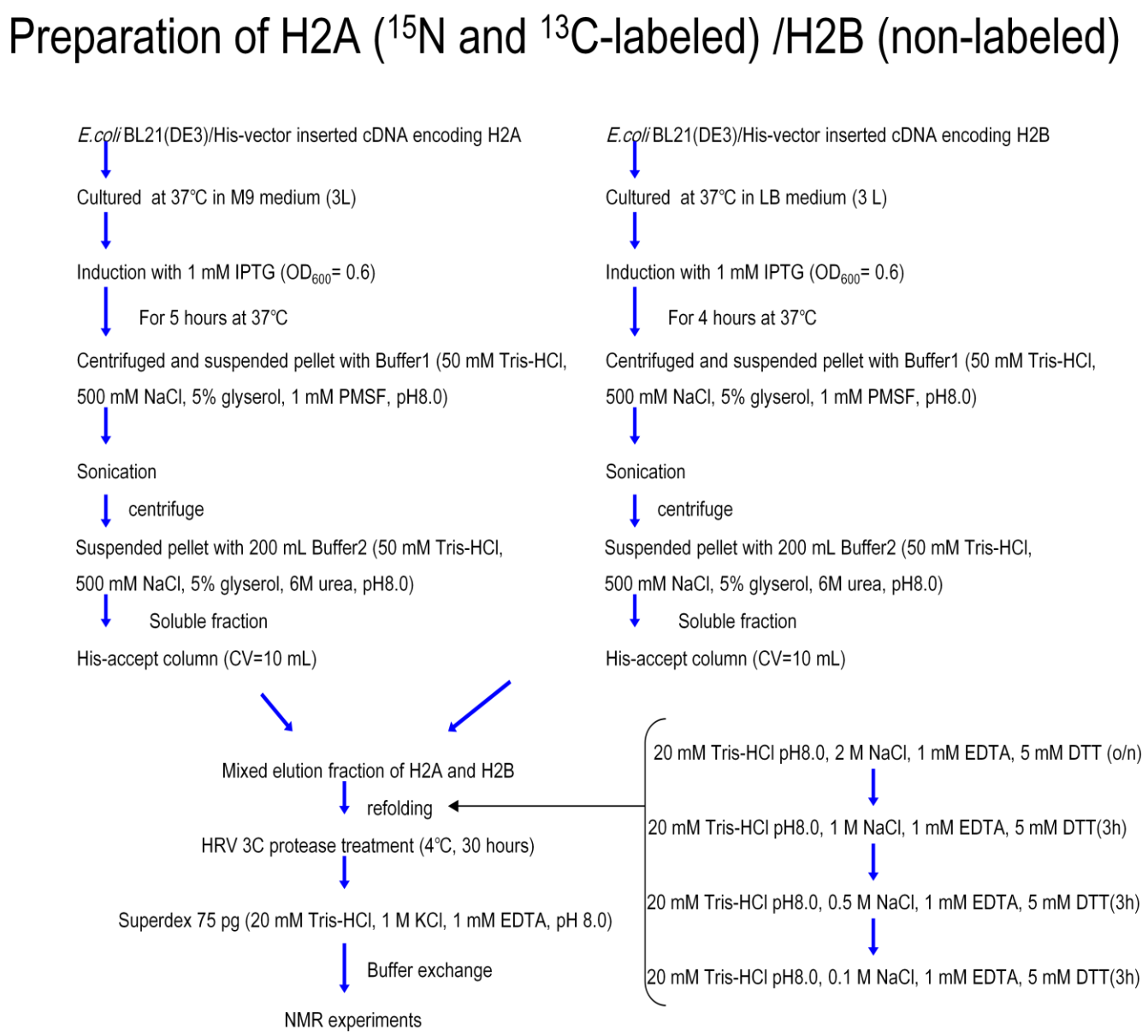
sequence determines the ID of a region and therefore, efforts are ongoing to delineate the sequence domains, which might contribute to protein ID [12].ID has been shown to be responsible for a wide variety of biological functions and to be common in nature [13]

Natively unstructured or ID protein regions are particularly abundant in eukaryotes and often evade structure determination. Many computational methods predict unstructured regions by training on outliers [14]. Histone protein is a family of protein consisting of neucleosome and chromosomes with its variants. Present study focused on Histone complex which is located in nucleosome. We tried to see the evaluation of our predictive system and a comparative analysis of histone disorders with regard to localization, interaction as well.




Unfolded proteins have backbone peptide groups exposed to the solvent, so that they easily break by enzymes (proteases), accelerate hydrogen-deuterium exchange and expose to dispersion (<1 ppm) in their 1H amide chemical shifts as measured by NMR. (Folded proteins typically show higher dispersions >3 ppm - amide protons.) ID proteins are characterized by a low content of hydrophobic amino acids and a high proportion of polar and charged amino acids. Thus, disordered sequences cannot merge its

hydrophobic core to fold like stable globular proteins. Sometimes, hydrophobic clusters in ID are identified for folding and binding. Such phenomena are the basis for the prediction methods. Many disordered proteins also show repeats of a few residues.
Thus repeating is an indication of disorder. However, not all disordered proteins have such trait. Disordered proteins are disorganized to make secondary structure. Once purified, Proteins can be identified by various experimental methods. Folded proteins are usually condensed. Unfolded proteins can be detected by methods that are sensitive to molecular size, condensation or hydrodynamic attraction, such as circular dichroism in infrared spectroscopy , Gel filtration (SP) chromatography, analytical ultracentrifugation, \X-ray crystallography scattering (SAXS), and measurements of the diffusion constant , Fast parallel proteolysis (FASTpp) to determine the fraction folded/disordered without purification etc




**Materials and Methods:**

All Histones were collected from Swiss prot and Uniprot databases. Control set (Human 3D protein set) was used from Swiss port. By using five highly accurate prediction systems: Fold Unfold [15], Glob Plot, PrDOS [16], IUPred [17] and Dip Pred [18], standard ID levels were

made from ordered protein set. By using that standard scale, we have considered whether a protein is ID or not. From Fig-2, finally we selected a set of ID proteins among the entire Histone family. We verified by using other methodological properties like Depth Index for Buried Atoms (DPX) [19] and Hydrophobicity assessments [20] (See Fig -3). Long ID and short ID were identified using POODLE [21]. As short ID is less informative, these were excluded from the study

Then we tried to see whether predicted results are consistent with experimental outcomes or not. With that regard, we pursued a series of empirical works for Histone (H2A and H2B only). As considering all proteins is not easy. Experimental parts were shown in the flow chart.

Analytical gel filtration (Size exclusion) of purified histone was performed using a Sucrose analytical column equilibrated with buffer, pH 7.4 and 50 mM NaCl. Histone eluted with a flow rate of 1 ml/min at 4°C and the elution profile was recorded by continuously monitoring the UV absorbance. The eluted peaks were analyzed for their molecular weight using mass spectrometry.

MS fragmentation in a MALDI-TOF/TOF instrument is mediated by collision induced dissociation, which causes peptides to preferentially fragment at their peptide bonds.

**Results and discussion:**

Disordered patterns are a bit different in necleosomes than that of chromosoms. Sequence differences could be due to that location orientation, though, sequence differences are only few percents. But in disordered fractions, they are highly different (fig 1&4). We did the work with N terminal and without N terminal regions and found that not only N terminal regions responsible for disorder ness but also others' sporadic regions. In our check up for computational predictions, we expressed directly H2A and H2B genes and their proteins (fig 2 &3) and refolded those by using NMR. In Fig -3: Structure prediction from sequences by using consensus method is depicted [22]. Red colored residues were predicted to be disordered, but, there may be some short helical segments in our prediction. Residues included in H2A: 11–118, H2A0: 15–118, H2B: 30–125, H2B0: 31–124. The remaining portions of the histone tails were disordered (Fig -4). Analytical size exclusion chromatography was observed. In the absence of a reducing agent a second peak of histone dimer was present, it confirms oligomeric state. The identities of all the peaks were verified by mass spectrometry. Histone eluted from the gel filtration column as a single peak that corresponded to the mass of the monomer. In the absence of DTT an additional peak of mixed dimer was observed. (Fig-5) In a MALDI-TOF/TOF, most of the times, there is just one fragmentation event per peptide molecule, so the resulting fragments are pieces of the peptide from either end, and the difference in mass between them corresponds to one or several amino acid residues, in such way providing sequence information about the peptide. However, fragmentation is not limited to the peptide bond, and it is normally not possible to read off' the peptide sequence or parts in MS-MS spectrum. (Fig -6)

We found that empirical results are supporting our Insilico results even though some times in low scale. We headed to navigate how and with which agents are interacting during disordering, we used prevalent results and our empirical results together in this aspect. As the H2A, H2B separately are good but in the neucleosome, they are highly disordered which may be due to several reasons like disorder itself and /or epigenetic action. But some previous results, it showed that epigenetic actions could not keep role for disorderness. We also checked whether they are disordered or not or staying only in neucleosome. We found in other primate organisms where Histones are complex but localizations are not in neucleosome (data not shown). It means that the reason is not for neucleosome location rather can be due to complex formation. Histone octamers are not well studied in other primates. It could be a study of chimpanzee Histone octomers to see how different it is from human as Chimpanzee database is rapidly growing. Again, as Histone modification is keeping key role epigenetically it can reveal some interactions within human and non human primates. Histone is conserved zone and much important for evolutionary study which made it a signature for study. Whatever it is, functional analysis is not the focus of the present study and thus we skipped

We have taken an approach to determine disordered parts of Histone family in the Histone Octomer Complex. Then we tried to purify H2A/H2B proteins in order to observe structures by NMR. The result was compared to observe how accurately our approaches perform empirically. For experimental part, until now we have results for purified samples and did 2D and 3D NMR. Present study solely focused on H2A and H2B. As study on all proteins through experimental procedure is indeed a hectic and time consuming. NMR is used to detect ID proteins. The processes of DNA replication, gene transcription, and mRNA translation, all necessary procedures to prepare a polypeptide chain, are extremely convoluted, and involve layers of control that are only now being explained. However, the procedure of polypeptide remains the same, no matter what the

primary sequence of the protein. The details of the final step in the procedure, the folding of the protein, depend strongly on the exact composition and primary sequences although the folding process are the same for all proteins. Folding pathways may differ significantly for different proteins. Some proteins require extra help, such as pro sequences and chap-erones, to fold to the correct conformation [23]. NMR has an important role, but there is a disharmony. This mismatch is due to the different time scales of the folding process, which complete in milliseconds for many proteins, and of the NMR experiment (the fastest 2D spectra can only be accumulated in minutes). NMR experiments to study aspects of the folding process, therefore give valid information. Valuable information on the kinetics of protein folding can be obtained from quench-flow hydrogen exchange methods, detected by NMR or using mass spectrometry [24] .Magnetization transfer methods have been successfully used to study fast-folding proteins [25]. Recently, equilibrium approaches, in which stable unfolded or partly folded states can be studied in solution over a relatively long period, have been extremely fruitful for a number of protein systems.. The importance of the composition of the denatured state in the study of protein folding processes was first recognized by Tanford [26].

NMR has been instrumental in identifying and characterizing unfolded and partly folded protein domains that are functional. Intrinsically unstructured, functional proteins show different degrees of disorder in their native states. The individual domains of a DNA-binding protein are connected by flexible linkers and tumble independently in the absence of DNA, but adopt a rigid, ordered structure in complex with DNA, providing a mechanism for high-affinity, sequence-specific binding. Another case (common one) in which a protein is unfolded in isolation but folds when bound to its (folded) interaction partner. Sometimes it shows the case of two unstructured proteins that are mutually folded when they interact.

The lack of electron density in X-ray crystallographic studies may also be a sign of disorder. Due to fragility issue, we did not take this approach. Figure 7 illustrates the [$^{1}$H–$^{15}$N]-HSQC spectrum and its assignments of the amide resonances; its unfolded nature is evident from the small dispersion in the amide proton resonances. The construct contains histone residues, the first couple of which are His-tag, followed by short linker sequence. A standard sequential assignment has the observable backbone $^{1}$H, $^{15}$N and $^{13}$C resonances were assigned (excluding the tag). The observable $^{1}$H/$^{15}$N are displayed by labeled one-letter amino acid code. Here we did not show all labeled codes to make it visible. Urea and temperature dependence of the CD (circular dichroism) spectrum of its proline, Valine, Leusine, Tyrosine -rich region reveals the presence of the extended and rather stiff polyproline II helix conformation that keeps the interaction site exposed. These data suggest that functionally significant residual structure exists in both of these ID proteins. (See fig 7)

**Conclusion:**

ID regions are implicated with various regulations, but a few ID proteins were experimentally done and huge amounts of such proteins are still undone. Computational methods can be novel ways to determine such ID regions rapidly and efficiently. Thereafter, experimental verifications reduce labor and cost intensity. ID proteins are actively crucial for bio functions and cellular locations. Experimental data are insufficient. And, there is no method solely sufficient to predict ID proteins accurately. Thus, our consensus approach takes the benefit to predict ID proteins more accurately and it will help to analyze functions more pertinently. Histone is a group of proteins present in all cells in all life forms. This family is very important for cellular functions and regulations. With that regard, we identified and analyzed all of the ID parts of Histone complex by our efficient prediction system and interpreted their significant molecular relations. We found that ID regions are conserved, included with various functional relations. The above computational approaches eased to analyze ID regions and to navigate significant biological roles. Just because, even singular processes are undertaken to see the process through, but, hopefully it will be applicable in broad spectra.

**Acknowledgement:**

Highly gratitude to Prof. K. Shimizu $^{\beta}$ ( for JSPS fund) and Prof Y. Nishimura $^{\delta}$ for various helps and to use their laboratory-facilities, Dr. Aritaka Nagadoi $^{\delta,}$ Masahiko Sato $^{\delta}$ and UVA Bio Physics facility for their technical helps.

**References:**

[1] Linding R, Russell RB, Neduva V and Gibson TJ. Glob Plot: exploring protein sequences for globularity and disorder. NAR.31, 3701-3708; 2003

[2] Wright PE. And Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. 293, 321–331; 1999

[3] Tompa P. Intrinsically unstructured proteins. Trends Biochem. Sci, 27, 527–33; 2002

[4]. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Biochemistry 41, 6573–6582 ; 2002

[5] Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. ID in cell-signaling and cancer-associated proteins. J. Mol. Biol 323, 573–584 ; 2002

[6] Radivojac P,Obradovic Z, Smith DK, Zhu, Vucetic G, Brown CJ, Lawson JD and Dunker AK . Protein flexibility and ID. Protein Science, 13, 71-80. ; 2004

[7] Shimizu K, Muraoka Y, Hirose S, Tomii K, Noguchi T. Predicting mostly disordered proteins by using structure-unknown protein data. BMC Bioinformatics. 6, 8:78. ; 2007

[8] Dyson HJ and Wright PE. Coupling of folding and binding for unstructured proteins.Curr. Opin. Struct. Biol 12,54–60 ; 2002

[9.] Cheng Y, LeGall T, Oldfield CJ, Mueller JP, Van YYJ, Romero P, Cortese MS, Uversky VN *et al*. Rational drug design via intrinsically disordered protein. Trends in Biotechnology. 24, 435-442; 2006

[10]. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z and Uversky VN. J. . Functional Anthology of ID.1.Biological processes and Functions of Proteins with Long Dsiorder regions. Proteome Res. 6, 1882-1898 ; 2007

[11]. Dunker AK, Cortese MS, Romero P, Iakoucheva LM and Uversky VN. FEBS J. 272:5129-48.; 2005

[12]. Singh GP, Ganapathi M, Sandhu KS, and Dash D. Intrinsic Unstructuredness and Abundance of PEST Motifs in Eukaryotic Proteomes. Proteins: Structure, Function, and Bioinformatics 62, 309–315 ; 2006

[13]. Peng K, Vucetici S, Radivojac P, Brown CJ, Dunker AK and Obradovici Z. Optimizing long IDs with protein evolutionary information. J. of Bioinformatics and computational Biology. 3, 35-60. ; 2005

[14]. Schlessinger A, Liu J, Rost B . Natively Unstructured Loops Differ from Other Loops. PLoS Comput Biol. 20:e140 17658943 ; 2007

[15]. Galzitskaya OV , Garbuzynskiy SO and Lobanov MY. FoldUnfold: web server for the prediction of disordered regions in protein chain. Bioinformatics 22, 2948-2949 ; 2006

[16]. .Ishida T and Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. NAR. 12 , 17567614 ; 2007

[17.] Dosztányi 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. 1521,3433-4; 2005

[18]. MacCallum RM: Order/disorder prediction with self organizing maps. CASP6 Online Paper, http://www.forcasp.org/paper2127.html

[19]. Mihel J**,** Šikić M**,** Tomić S**,** Jeren B and Vlahoviček K . PSAIA – Protein Structure and Interaction Analyzer. BMC Structural Biology 8**,** 21doi:10.1186/1472-6807-8-21; 2008

[20]. Kyte J and Doolittle RF: A simple method for displaying the hydropathic character of a protien. J Mol Biol.157, 105; 1982.

[21]. Hirose S, Shimizu K, Kuroda Y and Noguchi T. POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. Bioinformatics.23, 2046-53 ; 2007

[22.] Bhowmick, B.K, Dawson, W., Majumder P, Shimizu K., A Consensus Approach for Intrinsic Disorder Analysis for Heat Shock Protein Family. Biotechnology, 8(3): 306-315 ; 2009

[23]. Grantcharova, V, Alm, E. J, Baker, D., Horwich, A. L. Mechanism of Protein folding . Curr. Opin. Struct. Biol,11: 70 ; 2001

[24].Roder, H, Elo¨ve, G. A., Englander, S. W. Structural characterization of folding intermediates in cytochrome c by H-exchange labeling and proton NMR. Nature, 335: 700 ; 1988

[25]. Vugmeyster, L., Kroenke, C. D., Picart, F., Palmer, A. G.,Raleigh, D. P. N15 R1rho measurements allow the determination of ultrafast protein folding rates J. Am. Chem. Soc., 122: 5387. ; 2000

[26.] Tanford, C. Adv. Protein Chem.Protein Denaturation., 23: 122. , 1968

*****