Assessing Conservation of Disordered Regions in Proteins

Ágnes Tóth-Petróczy[1], Bálint Mészáros[1], István Simon[1], A. Keith Dunker[2], Vladimir N. Uversky[2,3,4,*] and Monika Fuxreiter[1,*]

[1]*Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, H-1518 Budapest, Hungary,* [2]*Department of Biochemistry and Molecular Biology, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, IN 46202-2111 Indianapolis, USA,* [3]*Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia,* [4]*Institute for Intrinsically Disordered Protein Research, Indiana University School of Medicine, IN 46202-2111 Indianapolis, USA*

Abstract: Intrinsically disordered regions (IDRs) are highly populated in eukaryotic proteomes and serve pivotal, mostly regulatory functions. Many IDRs appear to be functionally conserved and analysis of protein domains indicates high propensity of conserved regions predicted to be disordered. Nevertheless, it is difficult to assess conservation of IDRs in general due to their fast evolution and low sequence similarity. We propose three measures to evaluate conservation of IDRs: i) similarities of the disorder profiles using different prediction conditions; ii) the conservation of amino acids with propensities for promoting either disorder or order; and iii) the overlap between disordered/ordered regions. These measures are computed on multiple sequence alignments that also include low-complexity regions of proteins. Using three subunits of the Mediator complex of transcription regulation from Homo sapiens and Drosophila melanogaster as an example we show that despite of their sequence dissimilarity IDRs can be conserved and likely carry out the same function in different organisms.

INTRODUCTION

The wealth of recent experimental and theoretical evidence indicates that proteins or protein segments may exist as a rapidly fluctuating ensemble of conformations both in *vitro* and in *vivo* conditions [1, 2]. These intrinsically disordered proteins (IDPs) or proteins with long intrinsically disordered regions (IDRs; >30 aa) can adopt a continuum of structural states such as completely unstructured, molten globules or locally disordered tails and linkers [3, 4]. The variety of disordered states can be beneficial, even prerequisite for various biological roles [5-9].

IDRs can act as *entropic chains* (linkers, clocks, bristles) as the Nup2p FG repeat region of the nuclear pore complex for example is responsible for regulation of gating [10]. IDRs often serve as target sites for post-translational modifications (*display sites*) such as the KID domain of CREB, which phosphorylation induces its binding to the KIX domain of CBP [11]. Binding of IDPs can also modulate the effect of the partner (*effectors*). For example, $p27^{Kip1}$ regulates cell-cycle by binding to cyclin dependent kinases and inhibiting their activity [12]. Intriguingly, malleability of IDPs enables binding in different conformations leading to unrelated, even opposite functions [13]. Activation and inhibition of ryanodine receptor can be resulted by the binding of the disordered C fragment of dihydropiridine receptor (DHPR) in two conformations [14]. IDPs frequently participate in folding of proteins (like heat-shock proteins, Hsps [15]) or RNA partly by unfolding the incorrect structures and facilitating formation of new contacts (*chaperones*) [16]. Formation of the scrapie form of *prion*s is also critically dependent on the intermediate disordered state [17]. Large multiprotein complexes also take advantage of IDPs that assist assembly of these organizations (*assemblers*). The RNA polymerase II disordered C-terminal domain provides a platform for the mRNA processing machinery [18]. Alternatively, IDPs can capture and store small ligands (*scavengers*). This underlies the response to dehydration stress in plants achieved by water retention by Desiccation stress protein (Dsp) 16 [19].

IDPs or proteins with long IDRs (> 30 aa) are highly populated in eukaryotic proteomes [20, 21] and are often associated with regulatory functions such as signal transduction or transcription [5]. Analysis of several sets of proteins related to various diseases revealed that IDRs are highly abundant in proteins associated with cancer [22], cardiovascular disease [23], Parkinson's disease and other synucleinopathies, Alzheimer's, prion diseases and diabetes. Additional confirmation of the high prevalence of intrinsically disordered proteins in human diseases came from the functional annotation over the entire Swiss Protein database from a structured-versus-disordered point of view [24]. Thus, intrinsic disorder is very common in disease-associated proteins, giving rise to the disorder in disorders concept, which we are calling the "D^2 concept" [25].

In spite of their biological importance, it is very difficult to assess the conservation of IDRs based on simple sequence comparisons. IDRs in general are located in low-complexity regions, which is depleted in aliphatic (Ile, Leu, and Val) and aromatic amino acid residues (Trp, Tyr, and Phe) [26], which hampers formation of hydrophobic cores that promote folding of globular structures. Instead, they are enriched in charged and polar amino acid residues: Arg, Gln, Ser, Glu,

*Address correspondence to these authors U.V.N. at the Department of Biochemistry and Molecular Biology, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, IN 46202-2111 Indianapolis, USA; E-mail: vuversky@iupui.edu and M.F. at the Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, H-1518 Budapest, Hungary; E-mail: monika@enzim.hu

1875-0397/08

Lys and structure breaking residues (Gly, Pro) designated as disorder-promoting residues [27]. Furthermore, as it has been shown for 26 protein families, intrinsically disordered regions evolve faster ($K_A/K_S = 0.4 – 0.8$) than globular protein domains ($K_A/K_S = 0.1 – 0.2$) [28]. Nevertheless, in multiple sequence alignments that underlie identification of domains a high percentage of positions with conserved predicted disorder was found [29, 30]. These results suggest conservation of IDRs even in the absence of apparent sequence similarity. This calls for a novel measure that can be used to assess conservation of IDRs.

Recently we have carried out the bioinformatics analysis on proteins of the Mediator complex of transcription regulation (Fig. **1**) (Tóth-Petróczy *et al*., submitted). These proteins play role in transmitting regulatory information from activators/repressors to the basal transcription machinery [31]. They exhibit low sequence conservation and lack globular domains that are usually present in transcriptional proteins. Using two independent predictors, IUPred [32] and PONDR-VSL [33], we have shown the abundance of disordered regions in the Mediator proteins, especially in those that participate in regulatory signal transfer. We have also found that in spite of the low sequence conservation of IDRs in Mediator proteins, they exhibit fairly similar location and distribution in different organisms. Motivated by these observations we propose to assess conservation of intrinsically disordered regions based on i) similarity of the IDRs predicted in different conditions ; ii) conservation of propensities of amino acid residues promoting order and disorder iii) an overlap between ordered/disordered regions. These measures reflect different aspects of intrinsic disorder/order properties of a given system and their combination provides comprehensive characteristics. We demonstrate the application of these measures on Med4, Med9 and Med12 proteins of the Mediator complex of transcription regulation to assess the structural and thus the possible functional conservation of IDRs.

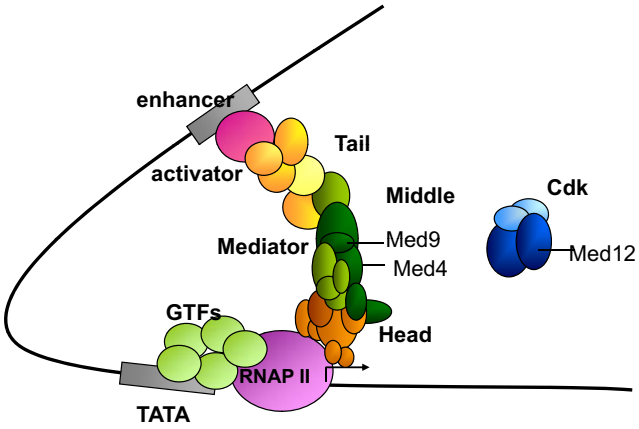

Fig. (1). The Mediator of transcription regulation. The Tail (yellow) interacts with activators bound at enhancers, the Middle (green) transmits regulatory signals, while the Head (orange) interacts with RNA polymerase II. The Cdk module (blue) usually dissociates prior to initiation of transcription. The analyzed subunits: Med4 and Med9 of the Middle and Med12 of Cdk are labelled inividually.

ALIGNMENT OF PROTEINS WITH INTRINSICALLY DISORDERED REGIONS

We designed an iterative, PSI-BLAST-based [34] alignment scheme to align full sequences of Mediator proteins containing low-complexity regions. First we generated profiles based on groups of sequences that fulfilled the $E < 10^{-5}$ threshold (> 30 bits). Then, the PSI-BLAST search is repeated using these profiles and the resulting groups are compared to the previously identified ones, from which the profiles have been extracted. If the two PSI-BLAST searches resulted in identical groups, a multiple sequence alignment can be carried out. Alternatively, the PSI-BLAST searches are continued with profiles updated in each cycle until the sequences in the resulted sequence groups converge. These sequences are subjected to a multiple sequence alignment using the profile, which has been extracted from the last PSI-BLAST run. These multiple sequence alignments were performed using the CLUSTALW algorithm [35].

The alignments for the Med4 and Med9 subunits of the Middle module and for the Med12 of the Cdk module of the Mediator complex from *Drosophila melanogaster* and *Homo sapiens* generated by the iterative alignment algorithm are shown in Fig. (**2**). In case of Med4 and Med9 these alignments show a fair agreement with previous results [36] as indicated by the superposition between the marked ordered segments. Advantages of using an iterative algorithm became apparent when more sequences were considered. For example, in case of Med9, sequences from *Saccharomyces cerevisiae* (*Sc*), *Saccharomyces pombe* (*Sp*), *Caenorhabditis elegans* (*Ce*), *Drosophila melanogaster* (*Dm*), *Homo sapiens* (*Hs*) were found to be homologous based on ordered motifs [36], while by our alignment only *Ce*, *Dm*, *Hs* sequences could be aligned (Tóth-Petróczy *et al*., submitted). A higher value of sequence conservation on *Ce*, *Dm*, *Hs* organisms can also be obtained on the alignment generated by the iterative algorithm as compared to alignment based on ordered motifs.

ASSIGNMENT OF INTRINSICALLY DISORDERED REGIONS

Intrinsically disordered regions were predicted based on the unfavorable contact energies according by the IUPred algorithm (http://iupred.enzim.hu) that exploits the principle that IDPs cannot fold as their amino acids cannot form sufficient inter-residue interactions to overcome the entropic penalty of folding [32, 37]. The overall interaction energy is estimated by pairwise interresidue potentials, approached by a low-resolution force field. In general, we distinguish two types of disorder: short and long disorder. Short disorder is associated with loops that are missing from crystal structures and usually are 5-12 residues in length. In contrast, long disordered regions can span several hundred residues in length, and may be functional on their own, even separated from the rest of the sequence. The sequence window along which the inter-residue potentials are computed can discriminate between these two types of disorder. In general, long disordered regions are identified using a 100-residue window, while for a short disorder a 25-residue window is used. Changes in the predicted disorder profile upon altering the sequence window size can discriminate far-lying and proximal sequence effects and may provide a general, inherent feature of the sequence. In Fig. (**2A** and **B**) we indicated the predicted short and long disorder (predicted with 25 and 100-residue windows, respectively) for human and drosophila Med4 and Med9 proteins. Disordered regions (IDRs) were defined as continuous segments of residues with score


A Med 4
B Med 9
Hs_25
Hs_100
Dm_25
Dm_100
Hs_ave
Dm_ave
Hs
Dm

(Fig. 2) contd….

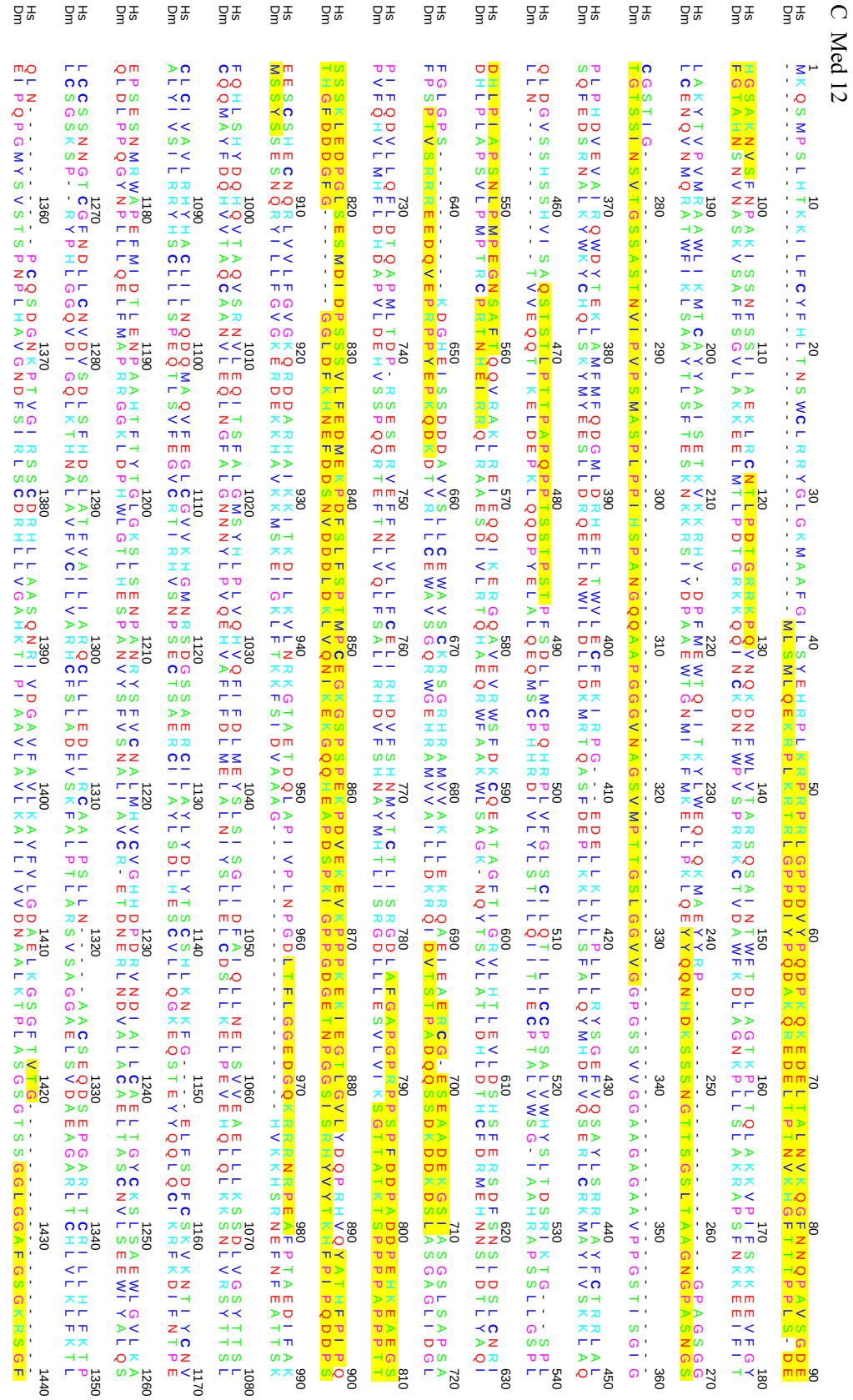

Fig. (2). Alignments of sequences of **A**) Med4, **B**) Med9 and **C**) Med12 proteins from Homo sapiens (Hs) and Drosophila melanogaster (Dm) generated by the PSI-BLAST based iterative algorithm. Intrinsically disordered regions predicted by the IUPred algorithm using 25 and 100 residues windows are shown by green and brown bars above the sequence. Consensus predictions are marked on the corresponding sequences by yellow. Conserved ordered regions predicted by previous alignments are shown by black boxes. Groups of similar amino acid residues are shown by cyan (K/R/H), green (A/S/T), blue (I/L/V/M/C/F/Y/W), magenta (G/P), red (E/D/N/Q) [36]. Plots were generated by the Alscript program [48].

> 0.5 with the minimum length of 5 residues. Above disordered segment lengths of 10 residues a 3 residue ordered gap was allowed. Interestingly, using a larger window in disorder prediction does not necessarily increase the size of the predicted IDRs. In case of Med4 long disorder prediction identifies shorter IDRs in both organisms than short disorder prediction, while in Med9 an opposite trend can be observed (longer disordered segments were predicted in both sequences using a 100 residue window). Overall, the propensities of residues belonging to disordered segments vary with the size of sequence window as shown in Table **1** (for 25 and 100-residue windows, respectively). The trends are rather similar for the same protein from different organisms, as they mostly depend on the preference for a certain type of disorder. This observation suggests that variation in disordered segments upon altering the sequence context (over which the prediction was performed) is an inherent property of the overall sequence.

Based on propensities of residues predicted to be disordered using different windows for prediction, Med9 can be considered as a disordered protein and Med12 as an ordered protein that contains some long IDRs. The level of disorder in Med4 is between Med9 and Med12 with an ordered N-terminal and a disordered C-terminal region (Fig. **2C**).

Table 1. Propensities of Amino Acid Residues in Intrinsically Disordered Segments of Med4, Med9 and Med12 Proteins from Homo Sapiens (*Hs*) and Drosophila Melanogaster (*Dm*). The Disordered Regions were Predicted by the IUPred Algorithm [32], with 100 Residue and 25 Residue Windows, Respectively. O/D Ratio Designates Ratio of the Propensities of Disorder- and Order Promoting Amino Acid Residues in Exclusively Disordered Regions

	Disorder (w=100)	Disorder (w=25)	O/D Ratio
Med4 *Hs*	0.44	0.47	0.43
Med4 *Dm*	0.35	0.35	0.42
Med9 *Hs*	0.57	0.48	0.25
Med9 *Dm*	0.66	0.63	0.31
Med12 *Hs*	0.32	0.37	0.30
Med12 *Dm*	0.40	0.43	0.32

SEQUENCE CONSERVATION OF INTRINSICALLY DISORDERED REGIONS

Sequence conservations of human and drosophila Med4, Med9 and Med12 proteins were computed separately for disordered and ordered regions, over individual amino acid residues (Table **2A**). As expected, the sequence similarity of disordered regions is considerably lower than that of the ordered regions. It especially holds for Med9 and Med12 that are equipped with over 400-residue-long IDRs. Please note that since conservation in disordered and ordered regions refer only to overlapping segments (cf., Fig. **4**), the total conservation is not an average of the conservations obtained separately for these two regions. Conservation of the groups of similar amino acid residues (defined as R/K/H, A/S/T, I/L/V/M/C/F/Y/W, G/P and E/D/N/Q [36]) exhibits higher values, in agreement with the low-complexity of IDR sequences (Table **2B**).

Table 2. Conservation of A) Individual Amino Acid Residues (AA_CONS) and B) Groups of Similar Amino acid Residues (GAA_CONS; Defined as R/K/H, A/S/T, I/L/V/M/C/F/Y/W, G/P and E/D/N/Q [36]) in the Disordered Segments (DIS), Ordered Segments (ORD) and in the Whole Med4, Med9 and Med12 Proteins (TOT) in Homo Sapiens and Drosophila Melanogaster. For Disordered and Ordered Segments Only the Overlapping Regions were Considered, Whereas for the Total Conservations Only the Gaps have been Excluded. The Conservation Scores have been Computed by Using a Simple Sum-of-Pairs Formula on the Alignment Generated by the Iterative Algorithm Described in the Text

A	AA_CONS		
	DIS	ORD	TOT
Med4	36.90	42.07	42.26
Med9	13.9	51.36	26.99
Med12	14.50	43.00	33.43

B	GAA_CONS		
	DIS	ORD	TOT
Med4	58.33	62.07	61.92
Med9	27.19	69.55	42.05
Med12	31.19	63.96	52.37

CONSERVATION OF AMINO ACID COMPOSITION OF INTRINSICALLY DISORDERED REGIONS

Amino acid composition of disordered regions in Med4, Med9 and Med12 proteins as compared to an average composition of globular proteins [6] are shown in Fig. (**3**). All proteins are enriched in Arg, Gly, Gln, Ser, Pro, Glu, and Lys residues that are generally abundant in IDPs [27] (referred as disorder-promoting residues) and are depleted in hydrophobic amino acids (Ile, Leu, Val, Trp, Tyr, and Phe) (referred to order-promoting residues) in similar manner [26] as inferred from the analysis of the DisProt database [38]. Although compositions in human and drosophila proteins are biased for intrinsic disorder, remarkable deviations can be observed (e.g. in propensities of Gly in Med9 and Pro in Med4 that show an opposite deviation from the composition of globular proteins), and also in the composition of charged residues (K, E). Compositions of order-promoting and disorder-promoting residues in disordered regions however, are considerably more stable as shown in column three of Table **1** by the ratios of the percentages of the two types of residues

(referred as O/D ratio). The O/D ratios exhibit a good agreement between the two organisms and in accord with the propensities of residues involved in disordered segments the O/D ratios reflect a similar level of disorder of the protein in different organisms. These results are further corroborated by previous studies on Med4 from 15 organisms that exhibit negligible variation in the composition of disorder- and order-promoting amino acid residues (Tóth-Petróczy *et al.*, submitted).

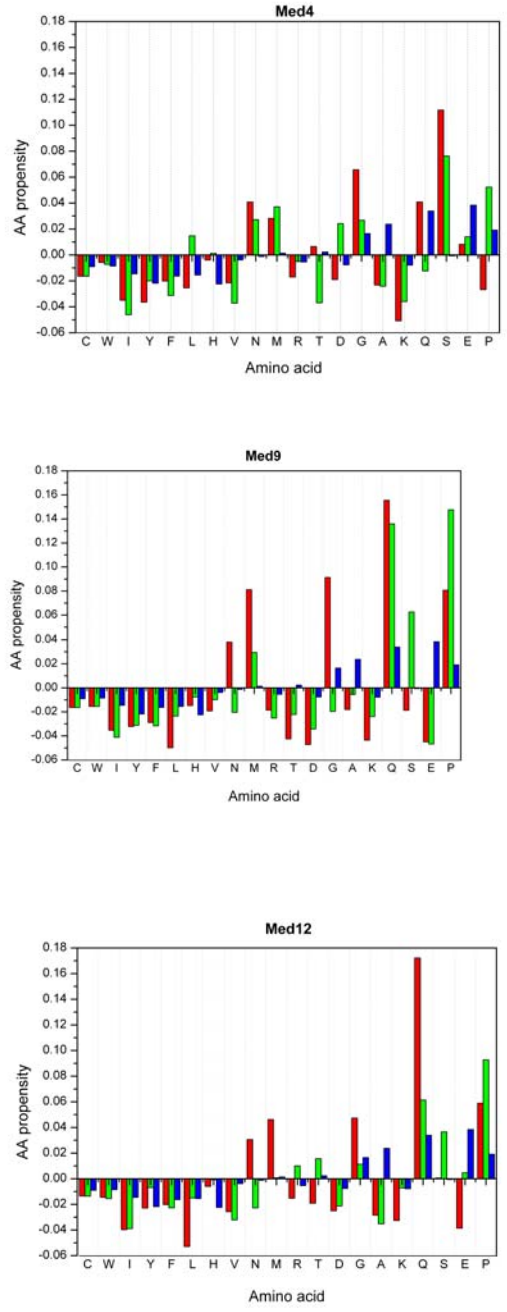

Fig. (3). Amino acid compositions in disordered segments of A) Med4, B) Med9 and C) Med12 from Homo sapiens (red) and Drosophila melanogaster (green), relative to the set of globular proteins. Composition of IDPs of the DisProt database [38] is shown by blue bars. The amino acids are arranged from left to right in order of their increasing propensity to promote disorder.

SEGMENTAL OVERLAP BETWEEN INTRINSICALLY DISORDERED REGIONS

In spite of their low sequence conservation, IDRs align well in human and drosophila Med4, Med9 and Med12 sequences even by visual inspection (cf. yellow regions in Fig. **2**). To assess the similarity of the IDR patterns, we quantified the overlap between ordered/disordered regions predicted in different sequences. Multiple alignments generated by Clustalw algorithm [35] were converted into a binary code, with two states (D and O, for disordered and ordered regions, respectively) defined based on the IUPred predictions. Gaps were excluded as they reflect variations in the size of a given disordered/ordered segment. Estimating the disorder/order properties in gapped regions of the alignment is in progress in our laboratory. Similarly to the assessment of the quality of secondary structure predictions, we compared the overlap between residues predicted to be ordered/disordered in different sequence pairs.

The accuracy matrix M was built from the number of residues that were predicted to conform to identical disordered (D) or ordered (O) states. The two-state overall accuracy is defined as [39]:

$$Q_2 = \frac{100}{N}\sum_{i=1}^{2} M_{ii} \qquad i \in \{D,O\} \qquad (1)$$

where N is the total number of residues and *i* runs over the two conformational states.

In addition to the per-residue based evaluation the actual overlap between patterns of disordered and ordered segments can be computed using the so-called segmental overlap (SOV) measure [40]. The advantage of using SOV is that it effectively captures the segmental characteristics of the sequence that is schematically illustrated in Fig. (**4**). For M conformational states, SOV is defined as:

$$SOV = \frac{100}{N}\sum_{i=1}^{M}\sum_{S_i}\frac{\min ov(S_1;S_2)+\delta(S_1;S_2)}{\max ov(S_1;S_2)}\times len(S_1) \qquad (2)$$

where S_1 and S_2 stand for segments in two distinct sequences, respectively, min*ov*(S_1; S_2) is the length of the overlap between S_1 and S_2, max*ov*(S_1; S_2) is the total extent of S_1 and S_2 in the given conformational state and len(S_1) is the length of the segment in the reference sequence. $\delta(S_1; S_2)$ is the minimum of [(max*ov*(S_1; S_2) – min*ov*(S_1; S_2); min*ov*(S_1; S_2); int(len(S_1)/2); int(len(S_2)/2)]. The normalization factor N is given by the number of residues in conformational state *i* and the second summation runs over all M conformational states. We have to note that computing SOV separately for disordered and ordered segments is also meaningful.

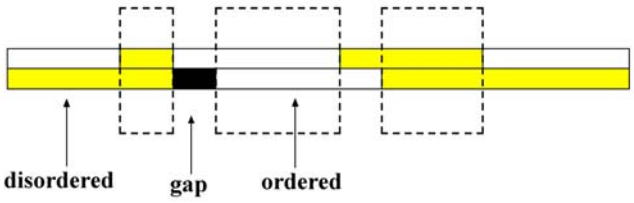

Fig. (4). Schematic representation of the segmental overlap analysis. Disordered regions are colored by yellow, ordered regions by white, and gaps by black. The actual overlap is marked by dashed boxes.

The overall accuracy values obtained for Med4, Med9 and Med12 proteins are shown in Table **3** for the full sequences as well as separately for disordered and ordered regions. The high Q values indicate that more than 70% of the residues in total belong to identical regions in human and drosophila in sequences of the same protein. In the more ordered Med4 and Med12 proteins the match between disordered residues is lower than that of ordered residues, while in Med9 a contrary behavior can be observed, likely due to the abundance of disordered regions. Please note that by definition, the total Q is not an average of the values obtained in disordered and ordered regions separately (cf., Fig. **4**). The SOV values are expected to be smaller than the accuracy measures due to the variations in the length of IDRs. This is indeed the case for Med9 and Med12, where the predicted IDR in the drosophila sequence is significantly longer than in human. For Med4, the SOV value is 100% reflecting a very small deviation between the predicted regions, which is compensated by the δ term in eq. 2. The statistical significance of the accuracy measures and segmental overlap values has been assessed by comparing them to the corresponding Q and SOV values obtained on shuffled sequences (randomizing the sequence 50 times). The resulting RQ and RSOV values are considerably lower than the Q and SOV obtained on the actual sequences corroborating the significance of the results. Differences between the SOV and RSOV values are lower than those between the Q and RQ values reflecting that disordered regions are primarily determined by the amino acid composition rather than the actual sequence.

Table 3. Overlap between Ordered (ORD) and Disordered (DIS) Regions in Med4, Med9 and Med12 *Drosophila* and *Human* Proteins. For Disordered and Ordered Segments Only the Overlapping Regions were Considered, Whereas for the Full Proteins (TOT) Only the Gaps have been Excluded. A) Q is the Overall Two-State Accuracy Computed Based on the Number of Residues in Identical Ordered or Disordered States. B) SOV is the Segmental Overlap [40] that is Obtained Based on the Actual Correspondance between the Disordered/Ordered Segments in Different Sequences. The Reference Values RQ and RSOV were Obtained Using Shuffling the Sequences (50 Times). For the Total Conservations Only the Gaps have been Excluded (so it is not an Average of the Q and SOV Values Obtained on Disordered and Ordered Segments, Separately)

A	Q			RQ		
	DIS	ORD	TOT	DIS	ORD	TOT
Med4	94.7	96.8	95.8	43.5	57.1	43.5
Med9	79.8	72.9	75.5	70.9	28.8	54.8
Med12	79.1	89.9	86.4	34.9	66.5	55.8

B	SOV			RSOV		
	DIS	ORD	TOT	DIS	ORD	TOT
Med4	100.0	100.0	100.0	32.2	29.9	29.9
Med9	55.1	52.4	54.2	33.5	17.9	29.1
Med12	44.4	52.3	49.4	16.8	28.7	25.1

DISCUSSION

Proteins with intrinsically disordered regions are ubiquitous in eukaryotic proteomes: almost 33% of proteins have long (>30 aa) disordered regions [20, 21]. Depending on their actual amino acid compositions, IDRs may conform to different categories of structural disorder such as a conformational ensemble, a pre-molten or molten globule or a disordered tail or linker of an otherwise globular protein [3, 4, 8, 41]. These properties are generally determined by the amino acid composition rather than the actual sequence indicated also by their separation in the charge-hydropathy space [3]. Hence, despite their low sequence conservation (as shown in Table **2** for Med4, Med9 and Med12 proteins) IDRs can have similar structural characteristics and thus can carry out homologous functions in different organisms.

Thus, even in the absence of apparent sequence homology functional information can be inferred from conserved IDRs. Therefore, instead of the conventional sequence conservation we propose three measures to assess conservation of IDRs in different organisms: i) the similarities between disorder patterns using different prediction conditions, ii) the conservation of the propensities of disorder- and order promoting residues and iii) the overlap between ordered/disordered patterns.

We demonstrated the application of these measures for three Mediator proteins. All were shown to contain long IDRs that in case of Med9 and Med12 span several hundred residues. Studies on these three proteins illustrate that despite of sequence dissimilarities, IDRs in Mediator proteins can be aligned well. A similar level of disordered in different organisms was witnessed by the similar disorder/order promoting amino acid ratios in different organisms. The agreement between patterns of disordered/ordered regions was quantified using the segmental overlap measure (SOV) adopted for IDRs. High values of SOV and overall accuracy (Q) and their significant deviation from the corresponding measures obtained on shuffled sequences corroborate the conservation of IDRs in Med4, Med9 and Med12 proteins.

All three proteins are involved in complex regulatory pathways of the Mediator complex. Med4 and Med9 belong to the Middle module of the Mediator complex that transmits regulatory signals for transcription from the Tail module to the Head [42] that in turn interacts directly with the RNA polymerase II-TFIIF for pre-initiation complex formation [43]. The Middle module also receives repression signals from the CDK module, which dissociates prior to transcription [44]. The Med9 was shown to physically and genetically interact with Cdk8 and CycC of the CDK module. Based on the abundance and conservation of IDRs in Med9 we propose that these IDRs play critical role in mediating these interactions and induce large-scale conformational rearrangements of the complex that accompany transcription [45]. The conserved IDR in Med4 contains a phosphorylation site (T237 in yeast) that plays a role in enhancement of

CTD phosphorylation by TFIIH [46]. Med12 belongs to the CDK module, which can inhibit, but also activate transcription [47]. Such complex functioning with opposite outcomes, often termed as moonlighting is also facilitated by the malleability of IDRs [13]. Thus it is likely that such complex functioning of Med12 is also linked to the conserved, long IDRs in this protein.

In conclusion, we find that important functional information can be inferred from identifying conserved IDRs. As sequence similarities of IDRs are generally low, we propose to apply alternative measures such as disorder pattern similarity and segmental overlap between disordered regions to evaluate the conservation of these regions.

ACKNOWLEDGEMENTS

This work was supported in part by the grants MCRTN 2005-09566 of European FP6 (to M.F.), R01 LM007688-01A1 (to A.K.D and V.N.U.) and GM071714-01A2 (to A.K.D and V.N.U.) from the National Institutes of Health and the Bolyai fellowship (M.F.). We gratefully acknowledge the support of the IUPUI Signature Centers Initiative.

REFERENCES

[1] Romero, P.; Obradovic, Z.; Kissinger, C.R.; Villafranca, J.E.; Garner, E.; Guillot, S.; Dunker, A.K. *Pac. Symp. Biocomputing*, **1998**, 3437.
[2] Wright, P.E.; Dyson, H.J. *J. Mol. Biol*., **1999**, *2932*, 321.
[3] Uversky, V.N.; Gillespie, J.R.; Fink, A.L. *Proteins*, **2000**, *413*, 415.
[4] Dyson, H.J.; Wright, P.E. *Nat. Rev. Mol. Cell Biol*., **2005**, *63*, 197.
[5] Xie, H.; Vucetic, S.; Iakoucheva, L.M.; Oldfield, C.J.; Dunker, A.K.; Uversky, V.N.; Obradovic, Z. *J. Proteome Res*., **2007**, *65*, 1882.
[6] Tompa, P. *Trends Biochem. Sci*., **2002**, *2710*, 527.
[7] Tompa, P. *FEBS Lett*., **2005**, *57915*, 3346.
[8] Dunker, A.K.; Obradovic, Z. Nat Biotechnol., **2001**, *199*, 805.
[9] Dunker, A.K.; Brown, C.J.; Lawson, J.D.; Iakoucheva, L.M.; Obradovic, Z. *Biochemistry*, **2002**, *4121*, 6573.
[10] Denning, D.P.; Patel, S.S.; Uversky, V.N.; Fink, A.L.; Rexach, M. *Proc. Natl. Acad. Sci. U. S. A*., **2003**, *1005*, 2450.
[11] Parker, D.; Ferreri, K.; Nakajima, T.; LaMorte, V.J.; Evans, R.; Koerber, S.C.; Hoeger, C.; Montminy, M.R. *Mol. Cell Biol.*, **1996**, *162*, 694.
[12] Kriwacki, R.W.; Hengst, L.; Tennant, L.; Reed, S.I.; Wright, P.E. *Proc. Natl. Acad. Sci. U. S. A*., **1996**, *9321*, 11504.
[13] Tompa, P.; Szasz, C.; Buday, L. *Trends Biochem. Sci*., **2005**, *309*, 484.
[14] Haarmann, C.S.; Green, D.; Casarotto, M.G.; Laver, D.R.; Dulhunty, A.F. *Biochem. J*., **2003**, *372*, 305.
[15] van Montfort, R.L.; Basha, E.; Friedrich, K.L.; Slingsby, C.; Vierling, E. *Nat. Struct. Biol*., **2001**, *812*, 1025.
[16] Tompa, P.; Csermely, P. *FASEB J*., **2004**, *1811*, 1169.
[17] Pierce, M.M.; Baxa, U.; Steven, A.C.; Bax, A.; Wickner, R.B. *Biochemistry*, **2005**, *441*, 321.
[18] Proudfoot, N.J.; Furger, A.; Dye, M.J. Cell 2002, 1084, 501.
[19] Chakrabortee, S.; Boschetti, C.; Walton, L.J.; Sarkar, S.; Rubinsztein, D.C.; Tunnacliffe, A. *Proc. Natl. Acad. Sci. U. S. A*., **2007**, *10446*, 18073.
[20] Dunker, A.K.; Obradovic, Z.; Romero, P.; Garner, E.C.; Brown, C.J. *Genome Inform. Ser. Workshop Genome Inform*., **2000**, 11161.
[21] Tompa, P.; Dosztanyi, Z.; Simon, I. *J. Proteome Res*., **2006**, *58*, 1996.
[22] Iakoucheva, L.; Brown, C.; Lawson, J.; Obradovic, Z.; Dunker, A.K. *J. Mol. Biol*., **2002**, *3233*, 573.
[23] Cheng, Y.; LeGall, T.; Oldfield, C.J.; Dunker, A.K.; Uversky, V.N. Biochemistry, **2006**, *4535*, 10448.
[24] Xie, H.; Vucetic, S.; Iakoucheva, L.M.; Oldfield, C.J.; Dunker, A.K.; Obradovic, Z.; Uversky, V.N. *J. Proteome Res*., **2007**, *65*, 1917.
[25] Uversky, V.N.; Oldfield, C.J.; Dunker, A.K. *Ann. Rev. Biophys. Mol. Biol.*, **2008**, *37*, 215.
[26] Romero, P.; Obradovic, Z.; Li, X.; Garner, E.C.; Brown, C.J.; Dunker, A.K. *Proteins*, **2001**, *421*, 38.
[27] Vacic, V.; Uversky, V.N.; Dunker, A.K.; Lonardi, S. BMC *Bioinformatics*, **2007**, 8211.
[28] Brown, C. J.; Takayama, S.; Campen, A.M.; Vise, P.; Marshall, T.W.; Oldfield, C.J.; Williams, C.J.; Dunker, A.K. *J. Mol. Evol*., **2002**, *551*, 104.
[29] Chen, J.W.; Romero, P.; Uversky, V.N.; Dunker, A.K. *J. Proteome Res*., **2006**, *54*, 888.
[30] Chen, J.W.; Romero, P.; Uversky, V.N.; Dunker, A.K. *J. Proteome Res*., **2006**, *54*, 879.
[31] Kornberg, R.D. *Trends Biochem. Sci*., **2005**, *305*, 235.
[32] Dosztanyi, Z.; Csizmok, V.; Tompa, P.; Simon, I. *J. Mol. Biol*., **2005**, *3474*, 827.
[33] Obradovic, Z.; Peng, K.; Vucetic, S.; Radivojac, P.; Brown, C.J.; Dunker, A.K. *Proteins*, **2003**, *53*, 6566.
[34] Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. *Nucleic Acids Res*., **1997**, *2517*, 3389.
[35] Chenna, R.; Sugawara, H.; Koike, T.; Lopez, R.; Gibson, T.J.; Higgins, D.G.; Thompson, J.D. *Nucleic Acids Res*., **2003**, *3113*, 3497.
[36] Boube, M.; Joulia, L.; Cribbs, D.L.; Bourbon, H.M. *Cell*, **2002**, *1102*, 143.
[37] Dosztanyi, Z.; Csizmok, V.; Tompa, P.; Simon, I. *Bioinformatics*, **2005**, *21*, 3433.
[38] Sickmeier, M.; Hamilton, J.A.; LeGall, T.; Vacic, V.; Cortese, M.S.; Tantos, A.; Szabo, B.; Tompa, P.; Chen, J.; Uversky, V.N.; Obradovic, Z.; Dunker, A.K. *Nucleic Acids Res*., **2007**, *35*, D786.
[39] Rost, B.; Sander, C. *Proc. Natl. Acad. Sci. U. S. A*., **1993**, *9016*, 7558.
[40] Zemla, A.; Venclovas, C.; Fidelis, K.; Rost, B. *Proteins*, **1999**, *342*, 220.
[41] Uversky, V.N.; Oldfield, C.J.; Dunker, A.K. *J. Mol. Recognit*., **2005**, *185*, 343.
[42] Kang, J.S.; Kim, S.H.; Hwang, M.S.; Han, S.J.; Lee, Y.C.; Kim, Y.J. *J. Biol. Chem*., **2001**, *27645*, 42003.
[43] Takagi, Y.; Calero, G.; Komori, H.; Brown, J.A.; Ehrensberger, A.H.; Hudmon, A.; Asturias, F.; Kornberg, R.D. *Mol. Cell*, **2006**, *233*, 355.
[44] Elmlund, H.; Baraznenok, V.; Lindahl, M.; Samuelsen, C.O.; Koeck, P.J.; Holmberg, S.; Hebert, H.; Gustafsson, C.M. *Proc. Natl. Acad. Sci. U. S. A*., **2006**, *10343*, 15788.
[45] Davis, J.A.; Takagi, Y.; Kornberg, R.D.; Asturias, F.A. *Mol. Cell*, **2002**, *102*, 409.
[46] Guidi, B.W.; Bjornsdottir, G.; Hopkins, D.C.; Lacomis, L.; Erdjument-Bromage, H.; Tempst, P.; Myers, L.C. *J. Biol. Chem*., **2004**, *27928*, 29114.
[47] Andrau, J.C.; van de Pasch, L.; Lijnzaad, P.; Bijma, T.; Koerkamp, M.G.; van de Peppel, J.; Werner, M.; Holstege, F.C. *Mol. Cell*, **2006**, *222*, 179.
[48] Barton, G.J. *Protein Eng*., **1993**, *61*, 37.

Received: March 28, 2008 Revised: April 15, 2008 Accepted: April 15, 2008