Natively unfolded proteins: An overview

Ken Nishikawa[1]

[1]*Department of Bioinformatics, Maebashi Institute of Technology, 460-1 Kamisadori, Maebashi, Gunma 371-0816, Japan*

Proteins with wholly or partly denatured structures *in vivo* are called intrinsically disordered or natively unfolded proteins (NUPs). Functional importance of NUPs was revealed by NMR studies as first reviewed by P. Wright in 1999. Since then, computational analyses on NUPs have also been intensively carried out to predict that approximately one third of eukaryotic proteins are NUPs. I will start this overview with the question why it took so long to identify NUPs as an important subject of protein science, and then move on to several issues such as, whether or not NUPs are specific to eukaryotes, what a particularly higher fraction of NUPs existing in the nucleus means, and what evolutionary implications NUPs have.

Key words: intrinsic disorder, structural domain, disorder prediction, transcription factor, eukaryote

Proteins had been believed to adopt specific 3D structures and fulfill unique functions on the basis of the structures. However, research in recent years has revealed that a large number of proteins contain at least a region that does not adopt 3D structures *in vivo*[1]. It is also not uncommon that regions that do not assume globular structures stretch as long as several hundred residues. This sort of region is called "intrinsically disordered" (ID) and proteins that entirely or partly consist of ID regions are called natively unfolded proteins (NUPs). (Although some authors use the term, intrinsically disordered proteins, as a synonym of NUPs, I

exclusively use ID for regions in this article.) These ID regions are considered to be in a state similar to random coils that globular proteins adopt in the unfolded state. If they exist freely in the cytoplasm, how can they avoid aggregation with each other *in vivo* or how can they escape various types of proteases that cleave off such proteins? Such characteristics that apparently elude conventional understanding may be salient features of NUPs.

A large number of NUPs have been found *in vivo*, including many important proteins involved in signaling and regulation of gene expression. On the other hand, there are many unresolved issues about NUPs partly because they have started to attract attention only recently. In the overview below describing the current state of NUP research, I would like to start with the question why the presence of NUPs was recognized only recently despite the long history of protein research.

Why were NUPs discovered so late?

Partly unstructured proteins have been known for a quite long time. In the field of X-ray crystallography, for instance, ribosomal proteins have been known difficult to crystallize, if at all. In other cases, when the structural determination was attempted after successful crystallization, a portion of protein, sometimes involving more than 100 residues, cannot be seen in the X-ray structure. The invisibility is caused either by movement of the regions or by polymorphic structures they assume in crystals. To facilitate crystallization, some researchers removed the invisible regions and replaced them with short linkers. These invisible regions were nothing but a nuisance for X-ray crystallographers. Meanwhile, research in protein informatics in the 1990s uncovered a plethora of proteins having "low-complexity" sequences, i.e., repetition of the same amino acid residues or sequences consisting of only a few kinds of amino acid residues[2]. As

Corresponding author: Ken Nishikawa, Department of Bioinformatics, Maebashi Institute of Technology, 460-1 Kamisadori, Maebashi, Gunma 361-0816, Japan.
e-mail: mit-nishikawa@maebashi-it.ac.jp

low-complexity sequences do not appear in regions that form normal globular structure, they were thought to assume some other shape, although exactly what shape was left as a matter of speculation. In the meanwhile, the existence of low-complexity sequences has become widely recognized as a result of homology searches: a homology search conducted with a protein containing low-complexity sequences as query identifies a large number of false positive hits. To avoid this problem, it became a standard practice in homology search to run a preparatory program to mask low-complexity sequences in the query.

As above, though the presence of unstructured segments in proteins has been known, ID regions had a negative implication and were not considered as a respectable subject of research. It was the article by Wright and Dyson published in 1999[3] that gave a positive meaning to the ID region and reversed its image. This article made a comprehensive survey of experimental works on proteins, including NMR measurements revealing that various proteins interact with other proteins via ID regions. It was observed that an ID segment gives NMR spectra characteristic of the unfolded state with no fixed shape in the isolated free state, but when a partner protein coexists the segment interacts with it to form a complex by assuming a secondary structure such as an alpha helix. Many experimental studies on NUPs are presented in the article that show various kinds of protein-protein and protein-nucleic acid interactions, including a number of proteins such as transcription factors, coactivators, transcription termination factors, cell-cycle regulators, and RNA-binding proteins. Furthermore, a state-of-the-art experiment at the single-molecule level using high-speed AFM (atomic force microscopy)[4] has revealed that typical NUPs are composed of structural domains and long ID regions.

It had been a common sense that proteins function upon forming specific 3D structures. The Wright-Dyson article[3] entitled "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm" changed just that. The difficulty in altering the accepted view is one reason why the discovery of NUPs was delayed: one does not want to accept data contrary to common knowledge and tends to deny them as meaningless. Another possible cause for the delay comes from the fact that molecular biology study started not with eukaryotes, but with bacteria. In general, bacterial proteins are relatively simple and fit the above-mentioned paradigm (common knowledge), whereas eukaryotes have many complex and heterogeneous proteins. In fact, a great majority of NUPs are found in eukaryotes as described below.

According to the Wright-Dyson article, the presence of NUPs was already a matter of common knowledge among people engaged in NMR studies in the mid-1990s. As early as in the 1980s, however, Sigler[5] presciently pointed out the abnormality of transcription factor. He described the functional site of transcription factor called activation domain as being in a state of shapeless "acid blobs" and/or "negative noodles". His prescience notwithstanding, coining an appropriate term seems necessary for a novel concept to become established. By the introduction of the term "intrinsic disorder", the presence of NUPs has been publicly recognized.

Approach from computational analysis

It was Dunker[6] who first brought the study of NUPs into the field of computational analyses. What Dunker and his colleagues utilized as a basis for their analyses is the biased amino acid composition of ID regions. While the globular structure requires a certain amount of hydrophobic residues for the folding, the ID region tends to have very few hydrophobic residues and instead have an excess of charged and polar residues. Using the distinctive characteristics of amino acid composition, one can predict ID regions from sequence data. Since the advent of computational analyses at around 2000, ID region prediction programs have been applied to a large number of proteins from various organisms as the amount of publicly available genome information has exploded, and a number of articles on NUPs were published in rapid succession (see Ref. 7 and citations therein).

As a result, the following have been revealed: 1) ID regions often contain sequence repeats consisting of short sequence patterns[8]. Thus, the aforementioned low complexity sequences are also included in ID regions. 2) Compared to structural domains, ID regions have higher evolution rates[9], resulting in extremely poor sequence conservation. This point can be easily verified by conducting a BLAST homology search using an NUP as query. 3) Our examination of the molecular structure of NUPs that partly consist of structural domains and partly of ID regions revealed that most ID regions exist in linkers connecting domains and/or at terminal tails, but some are inserted in structural domains[10]. In the latter case, ID sequences are inserted on the surface of globular structure (loop region between secondary structures) and forms long, unstructured loops extruding from the molecular surface. 4) According to a comparative genome study, the percentage of NUPs is higher in eukaryotes than in prokaryotes. Although quantitative values differ depending on the prediction programs used, the values reported by Jones et al.[11] (the length-wise fractions of ID regions in eukaryotes and prokaryotes are 16–22% and 3–7%, respectively) are considered reasonable. 5) In eukaryotes, NUPs are found most frequently localized in the nucleus and least in mitochondria[11]. The latter observation is consistent with the previous point, considering that mitochondria were an endosymbiont of bacterial origin.

Meanwhile, as described above concerning the functional aspect of ID regions, complex formation upon interaction with the partner protein, so-called "coupled folding and binding"[12], has been experimentally observed in various NUPs. In normal circumstances, such interactions occur transiently/reversibly[13]. As the interaction sites in ID regions

often indicate a propensity to form secondary structure (e.g., alpha helix), attempts to predict functional sites from sequence characteristics have been made[14]. Another functional feature uniquely attributed to NUPs is a phenomenon called promiscuity interaction: one and the same interaction site interacts with different target proteins[13]. This phenomenon is not observed in interactions between conventional globular proteins. This unique phenomenon of NUPs prompted many researchers to investigate the role hub proteins play in the intracellular interaction network. In fact, it has been statistically shown that hub proteins contain significantly more ID regions[15,16]. Furthermore, functional sites *sensu lato*, e.g., residues subject to posttranslational modifications such as phosphorylation and ubiquitination, and signal sequences for nuclear transport, also exist in the ID regions of NUPs.

Transcription factors as typical NUPs

Transcription of the genetic information encoded by DNA is the first step of gene expression. The major molecular machinery to carry out transcription in eukaryotes is RNA polymerase II (RNAPII). Transcription is initiated when a transcription initiation complex consisting of general transcription factors guiding RNAPII to a correct location (promoter) on DNA as well as a large number of protein cofactors (mediators) receives activation signals from transcription factors. Transcription factors that turn on and off gene expression are different depending on genes and the number of transcription factor families per genome in higher animals and plants has been known to exceed 1,000. As transcription factors bind to double-stranded DNA, DNA-binding domains they possess characterize them.

The above-mentioned NMR studies of Wright and Dyson implied that many transcription factors contain ID regions. Computational analyses supported the suggestion as two publications independently reported prediction results of all human transcription factors[17,18]. Although the two groups used different prediction programs, they both arrived at the conclusion that ID regions lengthwise occupy as much as half of transcription factors, a very high fraction as an average of a large number of proteins. From the graphic presentation of the molecular structure of transcription factors, the reader can readily see the relation between ID regions and structural regions (domains) (see Fig. 3 in Ref. 17). 60% of human transcription factors are proteins consisting exclusively of ID regions except for DNA-binding domains (termed "type I" transcription factors)[17]. Considering the function of transcription factors, this finding has an important implication. As transcription factors bind to double-stranded DNA, they transmit transcription activation signals to the transcription initiation complex. Given the molecular structure of the type I transcription factor, a second functional site has to exist in the ID region. In fact, the above-mentioned papers[17,18] predicted transcription activation sites to be in ID
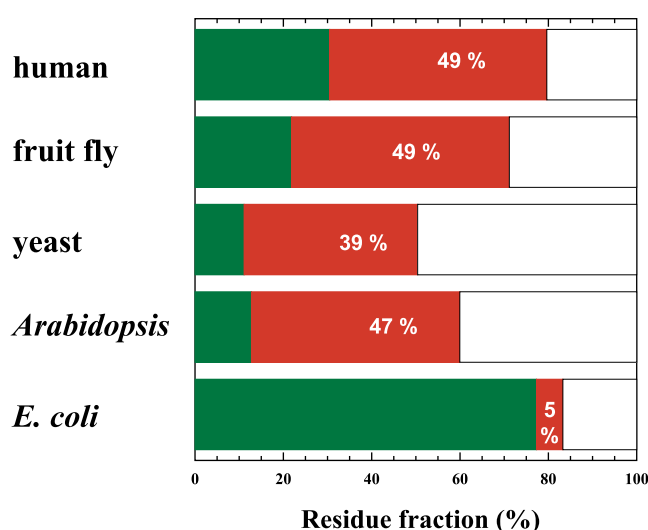
Figure 1 Comparison of transcription factors from various organisms. Length-wise fractions of structural domains (green), ID regions (red) and unassigned regions (blank) were estimated from a large number of transcription factors of each organism indicated. Note the clear difference between the ID fractions of eukaryote (human, fruit fly, yeast, and *Arabidopsis*) and prokaryote (*E. coli*).

regions. As Sigler pointed out previously, the transcription activation site resides in a noodle-like amorphous region. Furthermore, the transcription activation site forms a complex with a target protein by the "coupled folding and binding" mechanism[12].

Due to the high fraction of ID regions and the functional importance, human transcription factors can be regarded as typical NUPs. Although such characteristics of transcription factors are shared by eukaryotes in general, they are rare in prokaryotes (Fig. 1): prokaryotic transcription factors almost completely lack ID regions[19].

Are NUPs unique to eukaryotes?

In transcription factors, ID regions are found more frequently in eukaryotes than in prokaryotes (Fig. 1). The presented results were obtained by automatically applying a prediction program. The "disordered" regions constituting a mere ~5% in transcription factors in *E. coli* are almost entirely composed of unstructured linkers connecting domains and tails on both ends of sequence. In other words, *E. coli* transcription factors have almost no long ID regions. Can the finding extended to prokaryotic proteins in general?

Jones *et al.* were the first to demonstrate clear differences of ID regions between eukaryotes and prokaryotes[11]. As already described, they have shown that the percentage of NUPs is high in eukaryotes and low in prokaryotes, with a deep chasm between them. Meanwhile, the groups of Dunker and Tompa[6,20], which pioneered computational analysis, both reported that NUPs also exist in prokaryotes and the proportion is no different from that in eukaryotes. How-

ever, their expression has recently been modified possibly due to the influence of the Jones *et al.* article: they now state that NUPs are detected "relatively more" in eukaryotes than in prokaryotes. Different prediction programs can give rise to such different results. Dunker *et al.* developed a prediction program called PONDR, while Jones *et al.* wrote DISOPRED. As we pointed out before[21], PONDR tends to over-predict ID regions more frequently than DISOPRED.

It is of interest to note that the above issue of whether or not NUPs also exist in prokaryotes relates to how the concept of "intrinsic disorder" is interpreted. For example, ribosomal proteins after purification are notorious for being unamenable to crystallization. Should ribosomal proteins be considered as NUPs because of this? Ribosomal proteins are not in intrinsically free state in the cytoplasm, but are bound to gaps of a giant ribosomal RNA structure to serve as stabilizing materials. From the crystal structure of ribosomal particle as a whole, one can see that individual ribosomal proteins are immobilized in complex with RNA. While structural flexibility of ribosomal proteins in free state is necessary for the formation of complexes with complicated RNA molecules, ribosomal proteins adopt fixed structures in the end. Therefore, it is inappropriate to call ribosomal proteins as NUPs as the state in complex with RNA is regarded as innate. The same can be said of flagellin, flagellar protein of bacteria, 6% of which is predicted to be ID regions by DISOPRED. The isolated flagellin is in a flexible and unstructured state[22] but becomes integrated in flagellum as flagellin polymer. The flexibility of flagellin is considered necessary to pass through the narrow space inside the flagellum at the time of flagellar polymerization. Thus, newly synthesized flagellin and ribosomal proteins are similar in that both transiently experience a quasi-unfolded state. If the subsequent integrated state is its native state, the view that flagellin is an NUP[1] is mistaken.

It is quite likely that NUPs reported in prokaryotes so far are either results of over-prediction or transient quasi-unfolded states as in the case of flagellins and ribosomal proteins. Proteins having very long ID regions probably exist only in eukaryotes but not in prokaryotes.

Subcellular localization

Although there is at present no conclusive demonstration that NUPs are limited to eukaryotes, most researchers agree that NUPs exist more in eukaryotes than in prokaryotes. Do ID regions exist uniformly in eukaryotic proteins? Jones *et al.*[11] analyzed the fraction of ID regions in yeast proteins after classifying them according to the GO (gene ontology) categories of subcellular localization and found an uneven subcellular distribution: the fraction of ID regions is highest in nuclear proteins, while it is the lowest in mitochondrial proteins. However, the paper merely gave frequencies relative to the expected values.

Table 1   Subcellular localization of human NUPs

Category	No. proteins	Length-wise fraction of ID regions
Nucleus	1374	42%
Nucleus/Cytoplasm	257	32
Cytoplasm	750	25
Plasma membrane	2161	14
Extracellular milieu	813	13
ER/Golgi apparatus	124	13
Mitochondria	276	11
Mitochondrial membrane	131	9
Others	371	–
Total/Average	6257	23

To remedy this shortcoming, I present similar results obtained for human proteins (Minezaki, Y. and Nishikawa, K., unpublished data) using the same prediction program (DISOPRED) in Table 1, where ID fractions are indicated in absolute numbers rather than in relative values. We analyzed all human proteins in the Swiss-Prot database, following the Swiss-Prot annotation for the judgment of subcellular localization of individual proteins. In Table 1, the term "Nucleus/Cytoplasm" indicates proteins that are present in both subcellular compartments, while "Others" means those that are localized to multiple organelles other than the Nucleus/Cytoplasm pair. It is clear from Table 1 that nuclear proteins contain on average a conspicuously high fraction of ID regions. Considering that aforementioned transcription factors are classified as nuclear proteins and constitute approximately 30% of the entire nuclear proteins, the remaining 70% of nuclear proteins must also contain ID regions at high frequency. Why NUPs are so abundant in the nucleus is an interesting unresolved issue.

Another look at Table 1 from a different angle, with proteins classified as intracellular (in the cytoplasm and the nucleus) and extracellular (including the ER and Golgi apparatus), reveals that ID regions exist more in intracellular than in extracellular proteins. Membrane proteins have both intracellular and extracellular domains and thus were expected to have an intermediate frequency of ID regions. In fact, when membrane proteins are divided into intracellular and extracellular domains, the former domains have more ID regions than the latter[23]. What factors give rise to the asymmetry between intracellular and extracellular proteins is another problem to be solved.

Although the finding that mitochondrial proteins contain the least amount of ID regions is generally consistent with the results of Jones *et al.*, the absolute values (9–11% in Table 1) were not as low as expected: considering that mitochondria are endosymbionts of bacterial origin, an average value as low as 5% was anticipated. We need to compare mitochondrial and bacterial proteins carefully to see what accounts for this quantitative discrepancy.

Remaining problems

Eukaryotes clearly differ from prokaryotes in that eukaryotes generally have gene organization consisting of exons/introns, while prokaryotes do not. In eukaryotic transcription, pre-mRNAs are spliced so as that all introns are excised and only the exons are joined, and mature mRNAs then proceed to translation (protein synthesis). However, splicing often occurs with different combinations of exons, that is, alternative splicing takes place. As alternative splicing produces multiple splicing products (i.e., proteins) called splice variants, it has attracted attention as an ingenious mechanism to produce multiple proteins from a single gene. A cursory glance at this phenomenon from the viewpoint of genes (DNA) may fail to detect any problems. However, the consideration of alternative splicing from the angle of proteins reveals its unbelievably reckless nature. This is because proteins adopt 3D structures composed of domains, and if changes in the amino acid sequence such as deletion of a stretch of sequence encoded by an exon unit were to occur, proteins could no longer maintain the original 3D structures[24]. However, the recognition of NUPs has imposed the view that proteins consist not only of structural domains, but also of ID regions. If alternative splicing occurs within ID regions, any changes in sequence would not affect the protein structures.

The conjecture above was actually supported by a statistical analysis performed by Dunker *et al.*[25]. According to their study, the portions of protein sequences affected by alternative splicing contain a higher percentage of ID regions than the random expectation, implying that alternative splicing and ID regions "get along well". However, since structural domains were not taken into consideration in their analysis, it was unclear what fraction of alternative splicing events involved globular structures. This point can be clarified only if protein molecules are first divided accurately into domain and disorder regions by a dichotomous method[26] and then the results are compared with alternative splicing data. This issue remains to be elucidated in the future.

Both alternative splicing and NUPs are specific to eukaryotes, and NUPs provide favorable soil for alternative splicing as suggested by Dunker *et al.*[25]. Then, is it likely that they coevolved at an early stage of eukaryotes? Although this supposition cannot be easily tested, the skewed subcellular localization provides a clue. Despite the tentative nature of the results in Table 1, the biased localization of NUPs with a particularly high frequency in the nucleus is beyond doubt. On the other hand, it is unknown whether or not a similar bias in subcellular localization is observed in alternative splicing. For instance, a higher frequency of alternative splicing in nuclear proteins than in mitochondrial proteins will give credence to the coevolution idea. Conversely, if alternative splicing is found to occur independently of subcellular localization, then evolutionary relationship between the two will be judged unlikely.

Concluding remarks

While nearly a decade has passed since the concept of NUPs was first introduced, publicity among researchers appears to be still limited. This situation seems rather peculiar considering that NUPs having long ID regions (more than 30 consecutive residues) constitute as much as one third of all eukaryotic proteins[11]. An association with some important phenomena may generate wider publicity. One such phenomenon may be alternative splicing described in the preceding section. If the existence of ID regions is shown to be a prerequisite for alternative splicing to occur, NUPs will have a considerable impact on molecular biologists. Another possibility is related to the fact that NUPs are especially concentrated in the nucleus. The percentage of ID regions in nuclear proteins remains considerably high even if the contribution of transcription factors is excluded as already mentioned. I conjecture that NUPs would be related to non-coding RNAs, which form a hot topic of genome science. If various non-coding RNAs are involved in the regulation of gene expression, proteins that assist their function are likely to be required. In addition, assuming that protein-RNA interactions are quite different from protein-protein interactions and NUPs are especially suited for interactions with RNA (as evidenced by ribosomal proteins), why NUPs exist preferentially in the nucleus would become clear and the popularity of NUPs would grow by virtue of non-coding RNAs. It is exciting to see whether or not future research will prove the above conjectures.

Acknowledgements

References

1. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nature Rev. Mol. Cell Biol.* **6**, 197–208 (2005).
2. Wootton, J. C. & Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571 (1996).
3. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999).
4. Miyagi, A., Tsunaka, Y., Uchihashi, T., Mayanagi, K., Hirose, S., Morikawa, K. & Ando, T. Visualization of intrinsically disordered regions of proteins by high-speed atomic force microscopy. *Chem. Phys. Chem.* **9**, 1859–1866 (2008).
5. Sigler, P. B. Transcription activation: Acid blobs and negative noodles. *Nature* **333**, 210–212 (1988).
6. Dunker, A. K., Brown, C. J. & Obradovic, Z. Identification and functions of usefully disordered proteins. *Adv. Protein Chem.* **62**, 25–49 (2002).
7. Dunker, A. K., Silman, I., Uversky, V. N. & Sussman, J. L. Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.* **18**, 756–764 (2008).

8. Tompa, P. Intrinsically unstructured proteins evolve by repeat expansion. *BioEssays* **25**, 847–855 (2003).

9. Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., Williams, C. J. & Dunker, A. K. Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* **55**, 104–110 (2002).

10. Fukuchi, S., Homma, K., Minezaki, Y. & Nishikawa, K. Intrinsically disordered loops inserted into the structural domains of human proteins. *J. Mol. Biol.* **355**, 845–857 (2006).

11. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645 (2004).

12. Write, P. E. & Dyson, H. J. Linking folding and binding. *Curr. Opin. Struct. Biol.* **19**, 31–38 (2009).

13. Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* **579**, 3346–3354 (2005).

14. Oldfield, C. J., Cheng, Y., Cortese, M. S., Romero, P., Uversky, V. N. & Dunker, A. K. Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* **44**, 12454–12470 (2005).

15. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. Flexible nets: The roles of intrinsic disorder in protein interaction networks. *FEBS J.* **272**, 5129–5148 (2005).

16. Haynes, C., Oldfield, C. J., Klitgord, N., Cusick, M. E., Radivojac, P., Uversky, V. N., Vidal, M. & Iakoucheva, L. M. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comp. Biol.* **2**, e100 (2006).

17. Minezaki, Y., Homma, K., Kinjo, A. R. & Nishikawa, K. Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcription regulation. *J. Mol. Biol.* **359**, 1137–1149 (2006).

18. Liu, J., Perumal, N. B., Oldfield, C. J., Su, E. W., Uversky, V. N. & Dunker, A. K. Intrinsic disorder in transcription factors. *Biochemistry* **45**, 6873–6888 (2006).

19. Minezaki, Y., Homma, K. & Nishikawa, K. Genome-wide survey of transcription factors in prokaryotes reveals many bacteria-specific families not found in archaea. *DNA Res.* **12**, 269–280 (2005).

20. Tompa, P. & Csermely, P. The role of structural disorder in the function of RNA and protein chaperones. *FASEB J.* **18**, 1169–1175 (2004).

21. Nishikawa, K., Minezaki, Y. & Fukuchi, S. Human transcription factors as typical examples of natively unfolded proteins. *Tanpakushitsu Kakusan Koso* (in Japanese), **51**, 1827–1835 (2006).

22. Namba, K. Roles of partly unfolded conformations in macromolecular self-assembly. *Genes Cells* **6**, 1–12 (2001).

23. Minezaki, Y., Homma, K. & Nishikawa, K. Intrinsically disordered regions of human plasma membrane proteins preferentially occur in the cytoplasmic segment. *J. Mol. Biol.* **368**, 902–913 (2007).

24. Homma, K., Kikuno, R. F., Nagase, T., Ohara, O. & Nishikawa, K. Alternative splice variants encoding unstable protein domains exist in the human brain. *J. Mol. Biol.* **343**, 1207–1220 (2004).

25. Romero, P. R., Zaidi, S., Fang, Y. Y., Uversky, V. N., Radivojac, P., Oldfield, C. J., Cortese, M. S., Sickmeier, M., LeGall, T., Obradovic, Z. & Dunker. A. K. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc. Natl. Acad. Sci. USA* **103**, 8390–8395 (2006).

26. Fukuchi, S., Homma, K., Minezaki, Y., Gojobori, T. & Nishikawa, K. Development of an accurate classification system of proteins into structured and unstructured regions that uncovers novel structural domains: Its application to human transcription factors. *BMC Struct. Biology* **9**, 26 (2009).