Published in final edited form as:
J Proteome Res. 2007 May ; 6(5): 1882–1898. doi:10.1021/pr060392u.

Functional Anthology of Intrinsic Disorder. I. Biological Processes and Functions of Proteins with Long Disordered Regions

Hongbo Xie[†], **Slobodan Vucetic**[†], **Lilia M. Iakoucheva**[‡], **Christopher J. Oldfield**[#], **A. Keith Dunker**[#], **Vladimir N. Uversky**[#,§*], and **Zoran Obradovic**[†]

†*Center for Information Science and Technology, Temple University, Philadelphia, PA 19122, USA*

‡*Laboratory of Statistical Genetics, The Rockefeller University, New York, NY 10021, USA*

#*Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Indiana University, School of Medicine, Indianapolis, IN 46202, USA*

§*Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia*

Abstract

Identifying relationships between function, amino acid sequence and protein structure represents a major challenge. In this study we propose a bioinformatics approach that identifies functional keywords in the Swiss-Prot database that correlate with intrinsic disorder. A statistical evaluation is employed to rank the significance of these correlations. Protein sequence data redundancy and the relationship between protein length and protein structure were taken into consideration to ensure the quality of the statistical inferences. Over 200,000 proteins from Swiss-Prot database were analyzed using this approach. The predictions of intrinsic disorder were carried out using PONDR VL3E predictor of long disordered regions that achieves an accuracy of above 86%. Overall, out of the 710 Swiss-Prot functional keywords that were each associated with at least 20 proteins, 238 were found to be strongly positively correlated with predicted long intrinsically disordered regions, whereas 302 were strongly negatively correlated with such regions. The remaining 170 keywords were ambiguous without strong positive or negative correlation with the disorder predictions. These functions cover a large variety of biological activities and imply that disordered regions are characterized by a wide functional repertoire. Our results agree well with literature findings, as we were able to find at least one illustrative example of functional disorder or order shown experimentally for the vast majority of keywords showing the strongest positive or negative correlation with intrinsic disorder. This work opens a series of three papers, which enriches the current view of protein structure-function relationships, especially with regards to functionalities of intrinsically disordered proteins and provides researchers with a novel tool that could be used to improve the understanding of the relationships between protein structure and function. The first paper of the series describes our statistical approach, outlines the major findings and provides illustrative examples of biological processes and functions positively and negatively correlated with intrinsic disorder.

Keywords

Intrinsic disorder; protein structure; protein function; intrinsically disordered proteins; bioinformatics; disorder prediction

*CORRESPONDING AUTHOR FOOTNOTE: Correspondence should be addressed to: Vladimir N. Uversky, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, 635 Barnhill Drive, MS#4021, Indianapolis, IN 46202, USA; Phone: 317-278-9194; Fax: 317-274-4686; E-mail: vuversky@iupui.edu.

Introduction

Among other objectives, computational biology aims to enable an understanding of the relationships between the primary sequence, the higher order structure and the function of proteins. Each protein function is generally thought to originate from a specific 3-dimensional (3-D) structure. Formulation of this view began more than 100 years ago with the lock-and-key model proposed by Fischer.[1] More than 70 years ago Wu,[2] and slightly later, Mirsky and Pauling[3] equated denaturation with loss of specific structure. The dependence of function on 3-D structure was accepted by the time of the protein folding studies of Anfinsen and colleagues.[4] The flood of protein 3-D structures determined by X-ray diffraction and by nuclear magnetic resonance (NMR) spectroscopy has overwhelmed alternative concepts.[5]

In contrast to the dominant view given above, proteins for which intrinsic disorder is required for function have been reported in the literature for many years. By “intrinsic disorder” we mean that the protein (or protein region) exists as a structural ensemble, either at the secondary or at the tertiary level. Thus, both extended regions with perhaps some elements of secondary structure and collapsed (molten globule-like) domains with poorly packed side chains are included in our view of intrinsic disorder.[6] More detailed analysis of extended disordered proteins/regions revealed that they can be further divided in two groups, random coil-like and pre-molten globule-like conformations.[7] Recently, more than 150 proteins have been identified as containing functional disordered regions, or being completely disordered, yet performing vital cellular roles.[8, 9] Twenty-eight separate functions were assigned to these disordered regions, including molecular recognition via binding to other proteins, or to nucleic acids.[8, 10] A complementary view is that functional disorder fits into at least five broad classes based on the mode of disordered protein/region action.[10] Obviously, for these proteins, the predominant structure-function paradigm is insufficient, which suggests that a more comprehensive view is needed.[11] In fact, a new paradigm was recently offered to elaborate the sequence-to-structure-to-function scheme in a way that includes the novel functions of disordered proteins.[6, 7, 12] The complex data supporting this revised view were summarized in “The Protein Trinity” hypothesis, which suggested that native proteins can exist in one of three states, the solid-like ordered state, the liquid-like collapsed-disordered state or the gas-like extended-disordered state.[12] Function is then viewed to arise from any one of the three states or from transitions between them. Later this paradigm was extended to “The Protein Quartet” model to include one more extended-disordered conformation, the pre-molten globule state.[7] For structured proteins; i.e., proteins that form crystals without partners or have ordered globular forms without partners in NMR experiments, we will use the terms “structured”, “intrinsically ordered” or just ordered.

Recent studies revealed that many proteins lack rigid 3-D structure under physiological conditions *in vitro*, existing instead as highly dynamic ensembles of interconverting structures. Indeed, the literature on these proteins, known as intrinsically disordered, natively unfolded, or intrinsically unstructured, has virtually exploded during the last decade.[7, 13] This literature explosion is consistent with bioinformatics studies predicting that about 25 to 30% of eukaryotic proteins are mostly disordered,[14] that more than half of eukaryotic proteins have long regions of disorder,[14, 15] and that more than 70% of signaling proteins and the vast majority of cancer-associated proteins have long disordered regions.[16] As it has been already mentioned, despite the fact that intrinsically disordered proteins fail to form fixed 3-D structures by themselves under physiological conditions, they carry out numerous important biological functions.[6–11, 13, 16–22] Intrinsically disordered regions are typically involved in regulation, signaling and control pathways in which interactions with multiple partners, and high-specificity/low-affinity interactions are often involved.[21–23] Furthermore, sites of posttranslational modifications (acetylation, hydroxylation, ubiquitination, methylation, phosphorylation, etc.) and proteolytic attack are frequently associated with regions of intrinsic disorder.[6, 16] Given the high frequency of intrinsically disordered proteins and their crucially important functions, a curated Database of Disordered Protein (DisProt) has been recently initiated.[24] This database provides structure and function information about proteins that lack a fixed 3D-structure under putatively native conditions, either in their entireties or in part.[24] In spite of all of this, the phenomenon of intrinsic disorder in proteins is still severely under-appreciated; not a single biochemistry textbook discusses these proteins.[25]

There is a large gap between the number of proteins with experimentally confirmed disordered regions and actual number of such proteins in nature. Although studies of functional properties of known disordered proteins are helpful in revealing the functional diversity of protein disorder, they are bound to provide only a limited view. In this study, we propose a statistical approach for comprehensive study of functional roles of protein disorder. This approach relies on use of the VL3E[26] predictor that is currently the most accurate predictor of long disordered regions with estimated accuracy of above 86%.[26] The high accuracy of VL3E ensures that most disordered regions could be successfully detected with only a small fraction of ordered regions being incorrectly labeled as disordered. The VL3E predictor was applied to over 200,000 Swiss-Prot[27] proteins, many of which were annotated with one or more functional keywords. Then, the disorder-and order-correlated functions were detected as those that are overrepresented by proteins predicted to have long disordered regions (> 40 amino acid residues) in comparison with a random selection of proteins with the same length distribution. The proposed approach ensures that adverse effects of sequence redundancy and sequence length are eliminated. Disorder predictors were previously used to analyze functions of disordered proteins. For example, it was shown that a large fraction of cancer-related proteins are likely to be disordered.[16] In another study[28] it was demonstrated that many processes in yeast are related to protein disorder. The current study provides a comprehensive analysis of disorder-related functions by using a much larger set of proteins (i.e., the entire Swiss-Prot database).

Given a list of functions positively and negatively correlated with disorder, we performed an extensive literature survey to find experimental evidence supporting the findings. We were able to find at least one illustrative experimentally validated example of functional disorder/order for a large majority of functional keywords. This work opens a series of three papers dedicated to finding and description of protein functions and activities that are positively and negatively correlated with long disordered regions. Being the first in the series, this paper deals with the description of the statistical approach used here and delineates the major results of the application of this tool for the analysis of over 200,000 proteins from Swiss-Prot database. This paper also provides illustrative literature examples related to the Swiss-Prot keywords associated with the biological processes and functions positively and negatively correlated with intrinsic disorder. The second paper of the series portrays keywords related to the cellular components, domains, technical terms, developmental processes and coding sequence diversity associated with long disordered regions,[29] whereas keywords correlated with ligands, postranslational modifications and diseases associated with long disordered regions are the topic for the last paper of the series.[30] The overall result is that this series of papers represents a functional anthology of intrinsic disorder that includes both the results of our bioinformatics analysis and illustrative literature examples for the majority of functional keywords possessing strongest positive or negative correlation with the intrinsic disorder prediction.

Materials and methods

Dataset

The dataset for analysis was constructed using the Swiss-Prot database (release 48, 2005) containing 201,560 proteins.[27] In this study we used the 196,326 proteins with length longer than 40 amino acid residues. Each protein in Swiss-Prot is annotated with keywords that describe its functional or structural properties. Out of the 874 keywords used by Swiss-Prot, 710 were associated with at least 20 proteins. Swiss-Prot is statistically redundant, as it contains a large number of homologous proteins with highly similar sequences.[31] Ignoring the redundancy would significantly bias statistical inference. To reduce redundancy, TribeMCL[32] was applied to cluster the protein sequences from Swiss-Prot into families. TribeMCL uses the Markov clustering algorithm for the assignment of proteins into families based on the similarity matrix generated from the all-against-all BLASTp[33] comparison of sequences. It is able to produce high quality families despite presence of multi-domain proteins, peptide fragments, and promiscuous domains.[32] The obtained BLAST profiles were imported into TribeMCL software package (http://micans.org/mcl/) and clustering was performed with all parameters set at default. As a result of application of this redundancy reduction procedure, the sequences were grouped into 27,217 families.

Predicting long disordered regions in proteins

Previous studies suggested that in comparison with ordered sequences, disordered sequences tend to have lower aromatic content, higher net charge,[17, 34–36] higher values of the flexibility indices, greater hydropathy values,[34, 36] and lower sequence complexity.[37] Following these observations, the VL3E predictor[26] was developed using 162 long (>30 residues) disordered regions from non-redundant set of 152 DisProt proteins[24, 38] and 290 completely ordered proteins. The predictor consists of an ensemble of neural network classifiers and it achieves ~87% cross-validation accuracy on balanced data with equal number of ordered and disordered residues. We used the VL3E predictor to predict Swiss-Prot proteins with long disordered regions. Each of the 196,326 Swiss-Prot proteins was labeled as putatively disordered if it contained a predicted intrinsically disordered region with ≥40 consecutive amino acids and as putatively ordered otherwise. For notational convenience, we introduce disorder operator d such that $d(s_i) = 1$ if sequence s_i is putatively disordered, and $d(s_\mathrm{i}) = 0$ if it is putatively ordered.

Relationship between long disorder prediction and protein length

The likelihood of labeling a protein as putatively disordered increases with its length. To account for this length dependency, we estimated the probability, P_L, that VL3E predicts a disordered region longer than 40 consecutive amino acids in a SwissProt protein sequence of length L. Probability P_L was determined by partitioning all SwissProt proteins into groups based on their length. To reduce the effects of sequence redundancy, each sequence was weighted as the inverse of its family size; if sequence s_i was assigned to TribeMCL cluster $c(s_i)$, we calculated n_i as the total number of SwissProt sequences assigned to this cluster and set its weight to $w(s_i) = 1/n_i$. In this manner, each cluster is given the same influence in estimation of P_L, regardless of its size. To estimate P_L, all SwissProt sequences with length between $L-l$ and $L+l$ were grouped in set $S_L = \{s_i, L-l \leq |s_i| \leq L+l\}$. The probability P_L was estimated as

$$P_L = \frac{\sum_{s_i \in S_L} w(s_i) d(s_i)}{\sum_{s_i \in S_L} w(s_i)}.$$

Window size l allowed us to control the smoothness of P_L function. In this study we used window size equal to 20% of the sequence length, $l = 0.1 \cdot L$. We show the resulting curve in Figure 1 together with the same results when $l = 0$.

Extracting disorder-and order-related Swiss-Prot keywords

For each of the 710 SwissProt keywords occurring in more than 20 SwissProt proteins, we set to determine if it is enriched in putatively disordered or ordered proteins. For a keyword KW_j, $j = 1 \ldots 710$, we first grouped all SwissProt proteins annotated with the keyword to S_j. To take into consideration sequence redundancy, each sequence $s_i \in S_j$ was weighted based on the SwissProt TribeMCL clusters. If sequence s_i was assigned to cluster $c(s_i)$, we calculated n_{ij} as the total number of sequences from S_j that belonged to that cluster and set its weight to $w_j(i) = 1/n_{ij}$. Then, the fraction of putatively disordered proteins from S_j was calculated as

$$F_j = \frac{\sum_{s_i \in S_j} w_j(s_i) d(s_i)}{\sum_{s_i \in S_L} w_j(s_i)}.$$

The question is how well this fraction fits the null model that is based on the length distribution P_L. Let us define random variable Y_j as

$$Y_j = \frac{\sum_{s_i \in S_j} w_j(s_i) X_{|s_i|}}{\sum_{s_i \in S_L} w_j(s_i)},$$

where X_L is a Bernoulli random variable with $P(X_L = 1) = 1 - P(X_L = 0) = P_L$. In other words, Y_j represents a distribution of fraction of putative disorder among randomly chosen SwissProt sequences with the same length distribution as those annotated with KW_j.

If F_j is in the left tail of the Y_j distribution (i.e. the p-value $P(Y_j > F_j)$ is near 1), the keyword is enriched in ordered sequences, while if it is in the right tail (i.e. the p-value $P(Y_j > F_j)$ is near 0) it is enriched in disordered sequences. We denote all keywords with p-value < 0.05 as disorder-related and those with p-value > 0.95 as order-related.

The distribution Y_j is hard to derive analytically, so we randomly generated 1,000 realizations and calculated the empirical p-value as the fraction of times these realizations were larger than F_j. We also calculated the mean μ_j and standard deviation σ_j of the 1,000 realizations. We observed that, when $|KW_j|$ is large, distribution of Y_j resembles a Gaussian distribution with mean μ_j and standard deviation σ_j. Using the Gaussian approximation, we calculated the Z-score of KW_j as $(F_j - \mu_j) / \sigma_j$ and its p-value as $1/2(1 - erf(Z_j/2))$, where $erf()$ is the error function. The Gaussian approximation is useful since using the fraction of 1,000 replicates is not accurate in estimating p-values below 0.01 or above 0.99. We report the Z-scores together with the empirical p-values in the results.

Results

Estimating correlation between long disordered regions and Swiss-Prot keywords

We applied the procedure described above to each of the 710 Swiss-Prot keywords occurring each in more than 20 Swiss-Prot proteins. These 710 keywords can be grouped into 11 functional categories, which are listed in Table 1. We denote keywords with p-value > 0.95 as disorder-related and the ones with p-value < 0.05 as order-related. Keywords with p-value between 0.95 and 0.05 are ambiguous. These functions might depend on structured of disordered regions but simply exhibit signals that are too weak. Alternatively these functions might depend on short regions of disorder or might require both ordered and disordered regions.

The number of keywords strongly correlated with disorder and order is significantly larger than expected by the random model. This is evident by observing that, for a p-value threshold of 0.05, a random predictor would result in about 5% (~36) of order and 5% of disorder-related keywords. These results suggest that presence or absence of disordered regions is an important factor in majority of biological functions and processes. Overall, this analysis shows that 238 Swiss-Prot functional keywords are disorder-related, whereas 302 are order-related. Interestingly, only two of the categories, “Biological Process” and “Ligand”, are enriched in order-related keywords, while the remaining 9 are enriched in the disorder-related keywords. This result supports an earlier conjecture that disordered regions have a larger functional repertoire than the ordered regions.[20]

To further understand these function-disorder relationships, we carried out manual literature mining and studied a large number of individual experimental examples. To organize the presentation of these results, the keywords from various functional categories, which are most significantly associated with protein order and disorder arranged into specific groups (Table 2–Table 6). In each table, the disorder-function relationships are ranged by their Z-scores (see Materials and Methods). The Z-scores for all 710 functions are given in Supplementary Materials (see Table S1). One of the major goals here was to determine for each example whether the indicated function was carried out by regions of disorder or regions of structure. After all, the keyword-disorder correlations established by the method of Figure 2 do not determine whether the indicated association implies direct involvement of disorder with function or not.

Biological processes associated with intrinsically disordered proteins

The set of top 20 Swiss-Prot annotated cellular processes associated with predicted disorder are listed in Table 2. Presented below are several illustrative examples of biological processes from this list for which the associations with long disordered regions have been experimentally determined. The data below are organized in the following way: each discussed keyword is placed at the beginning of the corresponding paragraph and *Italized*. If the following description involves other keywords discussed in this and the subsequent papers[29, 30] these keywords are presented using the *Italic* font.

Differentiation—In developmental biology, cellular differentiation describes the process by which different cell types are derived from a single fertilized egg cell. Differentiation is a continuously regulated process, with specific interactions between the cell and its environment playing a major role in maintaining stable expression of differentiation-specific genes.[39] Obviously, numerous intracellular and extracellular proteins are involved in the differentiation control and regulation. For example, extracellular matrix (ECM), which is an important component of the cellular environment, was shown to play a role in regulating differentiation and the differentiated phenotype of cells.[40, 41] An ECM is present within mammalian embryos from the two-cell stage and is a component of the environment of all cell types, although the composition of the ECM and the spatial relationships between cells and ECM differ between tissues. The ECM offers structural support for cells, and can also act as a physical barrier or selective filter to soluble molecules.[40] The ECM is composed of glycoproteins, glycosaminoglycans and proteoglycans that are secreted and assembled locally into an organized network to which cells adhere.[42] Cells interact with the ECM via numerous cell-binding sites located within individual ECM glycoproteins and ECM receptors. For example, in case of fibronectin the primary determinant of cell-binding activity for many cell types resides in the sequence GRGDSP, which occurs in one of the type III repeats that form the central domain of the molecule.[43] Intriguingly, this cell-binding sequence exclusively consists of strongly disorder-promoting amino acid residues,[37] thus it very likely is intrinsically disordered. Importantly, it has been experimentally shown that the fibronectin binding domains from several different species of Gram-positive bacteria[44] as well as the N-terminal domain of BBK32, a fibronectin-binding lipoprotein from *Borrelia burgdorferi*, the causative agent of Lyme disease, are all intrinsically disordered.[45]

Transcription and transcription regulation—Transcription, being one of the key processes in the living cell through which a DNA sequence is enzymatically copied to produce a complementary RNA, is the transfer of genetic information from DNA into RNA. In the case of protein-encoding DNA, transcription initiates the process that ultimately leads to the translation of the genetic code into a functional protein. Transcription is strongly regulated by a number of proteins, especially transcription factors that include activators, repressors and enhancer-binding factors. Transcription factors function through the recognition of specific DNA sequences and the recruitment and assembly of the transcription machinery. Thus, both protein-DNA and protein-protein recognition are central processes in function of transcription factors. Several examples of intrinsically disordered proteins in transcriptional regulation have been reported.[18, 19] For example, the C-terminal activation domain of the bZIP proto-oncoprotein c-Fos unstructured and highly mobile, yet this protein effectively suppresses transcription *in vitro*.[46] The C-terminal domain of the transcriptional corepressor CtBP, which serves as a scaffold in the formation of a multiprotein complex hosting the essential components of both gene targeting and coordinated histone modifications, is also intrinsically disordered, as determined by using several complementary approaches (bioinformatics, NMR, CD, and small-angle X-ray scattering).[47] Recent analysis of high-resolution structures of transcription factors in the Protein Data Bank revealed that these proteins are, on average, largely disordered molecules with over 60% of amino acids residing in 'coiled' configurations.[48] The abundance of intrinsic disorder in transcriptional regulation was further demonstrated using a set of bioinformatics tools, including the Predictor Of Natural Disorder Regions (PONDR). This analysis showed that up to 94% of transcription factors have extended regions of intrinsic disorder. Furthermore, the analysis of the disorder distribution within the transcription factor datasets revealed that the degree of disorder is significantly higher in eukaryotic transcription factors than in prokaryotic transcription factors.[49] The complementary analysis of human transcriptional regulation factors revealed that although their average sequence is more than twice as long as that of prokaryotic proteins, the fraction of human sequences aligned to domains of known structure in PDB is less than half of that found for bacterial transcription factors,[50] suggesting that the increased length of eukaryotic transcription factors results to a significant degree from the addition of disordered regions.

Spermatogenesis—Spermatogenesis is the formation and development of mature spermatozoa from stem cells by *meiosis* and *spermiogenesis*. As *spermatogenesis* progresses, there is a widespread reorganization of the haploid genome followed by the extensive *DNA condensation* suggesting that the dynamic composition of chromatin is crucial for the activities of enzymes that act upon it. Histone variants such as H3.3, H2AX, and macroH2A play important roles at the various stages of *spermiogenesis*. Furthermore, *posttranslational modifications* of different histones, including specifically modulated acetylation of histone H4 (acH4), ubiquitination of histones H2A and H2B (uH2A, uH2B), and phosphorylation of histone H3 (H3p), are also involved in the regulation *spermatogenesis*.[51] Furthermore, during the final stages of *spermatogenesis*, the DNA of sperm in most organisms is compacted due to the replacement of somatic-type histones by *DNA-condensing* sperm nuclear basic proteins (SNBPs), sperm histones (H type), protamine-like (PL type), and protamines (P type).[52] Analysis of amino acid composition of PL-I sperm nuclear protein from *Spisula solidissima* revealed that it contains high amounts of lysine and arginine (24.8 and 23.1%, respectively).[53] Also, the PL-I has been shown to possess a tripartite structure, consisting of N- and C-terminal flexible "tails" flanking a globular, trypsin-resistant core of 75 amino acids.[54–56]

DNA condensation—DNA condensation *in vivo* relies on electrostatics-driven interaction of DNA with small cations and/or a number of abundant proteins including histones. In eukaryotes, the basic unit of chromatin (a condensed form of DNA) is commonly defined as a nucleosome, which is made up of DNA wrapped in two left-handed superhelical turns around a proteinaceous core.[57] The nucleosome core contains eight histone proteins, two dimers of H2A–H2B that serve as molecular caps for the central $(H3–H4)_2$ tetramer.[58] Thus, nucleosome represents the first level of chromatin condensation and is often termed 'beads on a string'.

Other crucial components of chromatine are the linker (H1 family) histones, which bind to the DNA that enters and exits the nucleosome and which facilitate the shift in equilibrium of chromatin towards more condensed, higher order forms.[57] It was established long ago that purified core histones being dissolved in water with no added salt, behave as polypeptides in an "extended loose form".[58–63] Recently, using a combination of bioinformatics tools with several biophysical techniques it has been shown that in low salt all bovine core histones are typical natively unfolded proteins; i.e., they possess exceptionally high level of intrinsic disorder.[64] Importantly, in the presence of high enough salt concentrations, core histones adopt a folded conformation.[58–64] In the crystal structure, histones are highly helical proteins, with α-helices accounting for 65–70% of the total structure. Only 3% of residues can be assigned to short parallel β-sheets, the remainder, approximately 30%, is not ordered.[65, 66] It has been also emphasized that the N-terminal "tail" domains (NTDs)[67] of the core histones and the C-terminal tail domain (CTD) of linker histones are intrinsically disordered, yet they are able to bind to many different macromolecular partners in chromatin.[68] Particularly, histone tails are known to be involved in the conformational changes of the nucleosome core particle (NCP) as well as in the structural phase transitions occurring at the supramolecular level. It is generally accepted that these tails interact with DNA at low salt and are extended outside of the particle at salt concentrations above ~0.2 M monovalent salt.[69] Analysis of the extension process of isolated NCP tails as a function of ionic strength has been reported. The addition of salt simultaneously screens Coulombic repulsive interactions between NCP and Coulombic attractive interactions between tails and DNA inside the NCP.[70]

Cell cycle, cell division, mitosis, meiosis—The *cell cycle* consists of an ordered series of events between the two cell divisions and involve the growth, replication, and division of a eukaryotic cell. Depending on the type of cell, the *cell division* might result in two different outcomes: in the division of somatic cells (*mitosis*), daughter cells are identical to the parent cell and contain a complete copy of the parental chromosomes; in *meiosis* (the division of sex cells), the daughter cells contain a half of the genes of the parent. Progression through the cell cycle is controlled in part by the activity of cyclin-dependent kinases, which are considered to be the major timekeepers of cell division.[71] Cdks are regulated by binding to their *cyclin* protein partners thus forming active heterodimeric complexes. Eight Cdk family members (Cdk1–Cdk8) and nine cyclins (A–I) have been identified so far. Interestingly, each Cdk pairs with a separate cyclin class, most of which have at least two members.[72, 73] For example, Cdk1 together with cyclin B1 directs the G2/M transition. Exit from G1, in contrast, is primarily under the control of cyclin D/Cdk4/6. Finally, two other cyclins (A and E) that pair with Cdk2 are required for the G1/S transition and progression through the S phase.[72, 73] The activity of Cdks throughout the cell cycle is known to be precisely regulated by a combination of several mechanisms, including the control of cycle-dependent variations in the levels of activating partners, *cyclins*; coordination of Cdk phosphorylation and dephosphorylation; and variations in the levels of the Cdk inhibitor proteins, CKIs, which are responsible for the deactivation of the Cdk–cyclin complexes.[71, 74] Five major mammalian CKIs are known: p21$^{Waf1/Cip1/Sdi1}$ and p27^{Kip1} inactivate Cdk2 and Cdk4 cyclin complexes by binding to them, p16^{INK4} and p15^{INK4B} are specific for Cdk4 and Cdk6, whereas p57^{Kip2} is specific for Cdk2.[71, 74] The p21$^{Waf1/Cip1/Sdi1}$,[75] p27^{Kip1},[76–78] and p57^{Kip2} CKIs[79] are all intrinsically disordered proteins that undergo sequential folding upon binding to their functional partners.

mRNA processing and splicing—An average gene in higher eukaryotes is very large due to the interruption of the coding sequence with large noncoding introns. Introns are known to be co-transcriptionally removed with great accuracy by pre-messenger RNA (*mRNA*) *splicing*. A large number of proteins are involved in generating specificity in pre-*mRNA processing*. Among the different pre-mRNA processing possibilities, *alternative splicing* is the most prevalent mechanism to generate proteomic diversity. The role of intrinsic disorder in alternative splicing is discussed in the second paper of this series.[29] Astounding examples of extensively alternatively spliced genes includes the Down syndrome cell adhesion molecule gene (Dscam) from *Drosophila*, the Neurexin and CD44 genes in humans, which can produce as many as about 38,000, 3000 and 1000 different splice forms, respectively.[80–82]. *Splicing* involves the stepwise assembly of five (U1, U2, U4, U5 and U6) small *ribonucleoprotein* particles (snRNPs) and a large number of proteins onto the pre-mRNA to form a large complex called the *spliceosome*.[83] The role of intrinsic disorder in the *spliceosome* function is discussed below in the Section entitled *Cellular components associated with intrinsic disorder*.

Apoptosis—Apoptosis is the programmed death of a cell. Regulation and control of apoptosis is crucial for the normal functioning of the organism. On the other hand, cancer cells avoid apoptosis and continue to multiply in an unregulated manner. The tumor suppressor protein p53 represents an outstanding example of this concept. The p53 molecule regulates expression of genes involved in numerous cellular processes, including cell cycle progression, apoptosis induction, DNA repair, as well as others involved in responding to cellular stress.[84] When p53 function is lost, either directly through mutation or indirectly through several other mechanisms, the cell often undergoes cancerous transformation.[85, 86] Cancers showing mutations in p53 are found in colon, lung, esophagus, breast, liver, brain, reticuloendothelial tissues and hemopoietic tissues.[85] When activated, p53 accumulates in the nucleus and binds to specific DNA sequences.[86, 87] It has been shown to induce or inhibit over 150 genes, including *p21*, *GADD45*, *MDM2*, *IGFBP3*, and *BAX*.[87] The p53 protein can be divided into three functional domains: an amino-terminal transactivation region, a central DNA binding domain, and a carboxy-terminal tetramerization and regulatory region. At the physiological temperature of 37°C and in the absence of modifications or stabilizing partners, wild-type p53 is more than 50% unfolded.[88] According to NMR analysis, the isolated transactivation domain lacks rigid structure,[89, 90] although it does possess an amphipathic helix that forms secondary structure part of the time, which can be stabilized by binding to Mdm2[91] or in the membrane environment.[92] Besides Mdm2, the transactivation domain interacts with numerous other proteins including TFIID, TFIIH, RPA, CBP/p300 and CSN5/Jab1,[84] thus playing a crucial role in the regulation of p53 function. For example, p53 can be inhibited by interaction with E3 ubiquitin ligase Mdm2,[93] which is bound to a short stretch of p53, specifically to residues 13–29.[91] As this region of p53 is within the transactivation domain, p53 cannot activate or inhibit other genes when Mdm2 is bound. Thus, p53-Mdm2 interaction exemplifies a crucial mechanism of vital regulation via the binding-induced folding of one of the interacting partners. The C-terminal regulatory domain is also unstructured.[94, 95] The structures of the core domain bound to DNA[96] and of several oncogenic mutants have been solved.[97] The crystal and NMR structures of the tetramerization domain are also known.[89, 98, 99] The high disorder content of p53 may help to explain its inherent instability and extreme oncogenic potential.[100, 101]

The BH3-only proteins belong to the proapoptotic family of proteins that function as key initiators of the programmed cell death. The BH3-only members of the Bcl-2 family proteins (those with a single Bcl-2 homology (BH) domain), including Bim, Bid, Bmf and Bad, are key initiators of *apotosis*, and they interact specifically with numerous binding partners.[102] In a healthy cell BH3-only molecules are either repressed, or are present in an inactive state.[103] The molecular mechanism of apoptosis initiation involves a stage of BH3-only proteins activation by a death stimulus, followed by their interaction and inactivation of prosurvival Bcl-2 proteins (such as CED-9 in *Caenorhabdhitis elegans*, and Bcl-x_L in humans).[102] Until recently, the structural knowledge about BH3-only proteins has been limited to peptide fragments bound to their targets. For example, in the complex of Bcl-x_L and a 33 residues long peptide corresponding to the BH3 domain of Bim, the peptide forms an α-helix upon binding to the hydrophobic groove of Bcl-x_L.[104] Analysis of crystal structures of BH3 domain peptides bound to the prosurvival proteins CED-9 and Bcl-x_L revealed that the nature of this interaction is highly conserved despite only a low level of shared sequence identity, with the conserved leucine and aspartic acid residues of the LXXXGDE motif, which defines BH3-only proteins, making critical contacts with conserved residues in the hydrophobic binding groove of CED-9 and Bcl-x_L.[104–107] However, in another structure a 9-residue peptide of Bim forms a β-strand upon binding to the component of dynein motor complex, DLC1.[108] Recently, using methods such as CD, NMR, analytical centrifugation, size exclusion chromatography and limited proteolysis, it has been established that the BH3-only proteins Bim, Bad and Bmf are unstructured in the absence of binding partners.[109] Intriguingly, the majority of the Bim residues remains disordered when this protein binds and inactivates prosurvival proteins, with the only the short α-helical molecular recognition element[110] becoming structured.[109] Furthermore, detailed sequence analyses suggest that most BH3-only proteins are unstructured.[109] The disorder of this proapoptotic protein family is likely to be important for several biological functions such as promiscuous binding, extensive splicing, and regulation via phosphorylation.

Ubl conjugation pathway—*Posttranslational modification* via the covalent attachment of *ubiquitin* and different *ubiquitin-like proteins* (*Ubls*, including SUMO, ISG15, Nedd8, and Atg8) is a crucial regulatory cellular mechanism, which plays a number of important roles in controlling *cell division*, signal transduction, embryonic development, endocytic trafficking and the *immune response*.[111] For example, conjugation of *ubiquitin-like proteins* (the *Ubl conjugation pathway*) to components of the *transcriptiol* machinery is an important regulatory mechanism allowing switching between different activity states. While ubiquitination of *transcription factors* is associated with *transcriptional activation*, their SUMOylation is most often connected with *transcriptional repression*.[112] Recent bioinformatic analysis of a limited number of known ubiquitination substrates showed that protein ubiquitination sites are preferentially located within surface exposed, flexible loop regions.[113] In addition, the sequence analysis of ubiquitination sites and regions adjacent to them showed that their properties such as low sequence complexity, high negative net charge and low hydrophobicity, are similar to those of intrinsically disordered regions (Iakoucheva and Radivojac, personal communication). An example of SUMOylation occurring within intrinsically unstructured region is the conjugation of transcription factor Ets-1 with SUMO-1.[114] Using NMR spectroscopy it has been shown that the sumoylation motif of Ets-1 containing Lys15 is located within the unstructured N-terminal segment of Ets-1 preceding its PNT domain.[114] The authors hypothesize that flexibility of the linking polypeptide sequence may be a general feature contributing to the recognition of SUMO-modified proteins by their downstream effectors.[114]

Wnt signaling pathway—Wnt is a critical pathway for embryogenesis, carcinogenesis, and cancer stem cells.[115, 116] Detailed information on this pathway can be found on the Wnt Homepage (http://www.stanford.edu/~rnusse/wntwindow.html). The Wnt pathway shows evolutionary conservation across a wide range of species, ranging from the freshwater polyp *Hydra* to vertebrates.[117] Mammals have 19 Wnt genes that can be grouped into 12 subfamilies.[118] Surprisingly, at least 11 of these subfamilies are present in Cnidaria (specifically, the sea anemone *Nematostella vectensis*) suggesting that they are not the result of any recent evolutionary diversification.[119] This indicates that the acquisition of the Wnt subfamilies was an early development in the evolution of metazoa and likely occurred about 650 million years ago.[119, 120]

The best-understood *Wnt pathway* is often called the Wnt/β-catenin pathway, in which the Wnt signal leads to activation of the nuclear functions of β-catenin. These functions activate expression of a number of genes leading to cell survival, proliferation, or differentiation.[121] A second vertebrate Wnt pathway, the Wnt/Ca^{2+} pathway, promotes intracellular Ca^{2+} release and regulates cell movements in development and in some cancers.[122] A number of Wnt protein isoforms are generated by *alternative splicing*.[123–125]

Wnt proteins comprise a large family of highly conserved secreted *growth factors* that activate target-gene expression in both a short- and long-range manner and regulate cell-to-cell interactions during embryogenesis. Wnt signaling is involved in virtually every aspect of embryonic development and also controls homeostatic self-renewal in a number of adult tissues.[115, 117]

Glycogen synthase kinase 3β (GSK3β) is a Ser/Thr protein kinase, which is one of the major players in the Wnt signaling pathway as GSK3β hyperphosphorylates β-catenin, thus promoting its ubiquitination and targeted destruction.[126] The crystal structure of human GSK3β (420 residues) has been solved at 2.8 Å.[127] Clear electron density was only evident for the 351 residues from Lys35 to Ser386, with the segments of the polypeptide preceding Lys35 and following Ser386 being disordered in the crystal.[127] The structure of the ordered part of GSK3β agrees with the consensus observed for "activation-segment" protein kinases, consisting of an N-terminal β-sheet domain, coupled to a C-terminal α-helical domain. The visible N-terminal domain (35–134) consists of a seven-stranded β-sheet, which folds to a closed orthogonal β-barrel. The core of the C-terminal α-helical domain (152–342) has a similar topology to the equivalent region in such mitogen activated protein kinases, such as MAPK, as ERK2 and p38.[127] It is important to emphasize that the major difference between the C-terminal α-helical domain of GSK3β and MAPK is the absence of the second helix in the hairpin segment from 276–293 in the GSK3β domain. Furthermore, in GSK3β this region represents a highly mobile and poorly defined 285–299 loop.[127]

Neurogenesis—Numerous proteins are involved in *neurogenesis*, the formation and development of nervous tissue. Among these proteins are the *transcription factors* Pax3,[128] Pax6,[129] Glis2,[130] and Erm,[131] which play an important regulatory role in this process. These transcription factors, like transcription factors in general, are highly disordered. For example, Pax3 has a highly flexible linker (53 amino acids) separating two DNA binding domains: a paired domain (128 amino acids) and a paired type homeodomain (60 amino acids).[132] Similar to Pax3, transcription factor PAX6 has two DNA-binding domains, a paired domain and a homeodomain (HD), joined by a glycine-rich linker and followed by a proline-serine-threonine-rich (PST) transactivation region at the C terminus.[133] Structural analysis revealed that the central 250 amino acid residues of the transcription factor Erm has very little (if any) ordered structure.[134]

Chromosome partition—Chromosome partition in two daughter cells is a complex process that involves a number of proteins. For example, proteins such as topoisomerase IV and XerCD recombinase, as well as MukB and FtsZ are related to chromosome partition in *Escherichia coli*.[135] MukB exists as two thin rods (long antiparallel coiled coils) with globular domains at the ends emerging from the very flexible (read disordered) linker domain (123 amino acids).[136] The flexibility of the hinge is crucial for the MukB function, as the arms can open up to 180°, separating the terminal domains by 100 nm, or close to near 0°, bringing the terminal globular domains together.[136]

Immune response—The immune system is capable of generating specific antibodies against an almost infinite diversity of physiological or synthetic antigens. However, the repertoire of antibodies produced in any organism is fixed, suggesting that the immune system is an example of nearly unlimited functional multiplicity based on limited sequence diversity.[137] The high flexibility of antigen-binding sites in the immunoglobulin, which provides the antibody with a unique capability to access a great variety of configurations with similar stabilities, was long ago proposed to be the basis of this binding diversity.[138] In more detail the interplay between the intrinsic disorder, antigenic structure and immunogenicity has been recently overviewed to emphasize the crucial role of intrinsic disorder in the development of immune response.[22] For example, the conformational flexibility of antibodies drives their polyreactivity, thus expanding the antigen-binding capacity of the antibody repertoire. On the other hand, short intrinsically disordered regions are likely important for the antigenicity of continuous determinants. Furthermore, the conformational flexibility and spatial adaptation play important roles in the antigen-antibody recognition and interaction. Finally, short intrinsically disordered regions are good antigens, whereas several long disordered regions and intrinsically disordered proteins initiate weak immune responses or are even completely non-immunogenic.[22] Based on these observations it has been hypothesized that the role of intrinsic disorder in immunogenicity and antigenicity of a protein depends on the length of the disordered segment: short disordered regions (usually five to eight residues) are important for the development of the immune response to continuous epitopes, whereas long disordered regions (longer than 30 amino acids) are less likely to be immunogenic.[22]

The role of intrinsic disorder in autoimmune diseases has also been emphasized recently.[139] The observation that the majority of the nuclear systemic autoantigens are extremely disordered proteins allowed the authors to introduce a new model of autoimmunity, disorder-based epitope spreading.[139]

Another example that illustrates the importance of disorder for *immune response* is the unstructured nature of the interferon tails.[140]

Finally, cytoplasmic domains of several immune receptors members of the family of multichain immune recognition receptors (MIRRs) (e.g., T-cell receptors (TCRs), B-cell receptors (BCRs), and the high-affinity IgE receptor) have signaling subunits carrying so-called immunoreceptor tyrosine-based activation motif (ITAM).[141–143] ITAM-containing cytoplasmic domains of signaling subunits of several MIRRs are intrinsically disordered.[144, 145] An intriguing feature of these signaling subunits is their tendency for the specific homooligomerization, which is not accompanied by their folding.[145, 146]

***Ribosome biogenesis* and *rRNA processing*—**Ribosomes are responsible for the production of the entire complement of proteins required for cellular maintenance, growth, and survival. Eukaryotic ribosomes contain four RNA molecules: 25S, 18S, 5.8S, and 5S. The 5S rRNA is transcribed by RNA polymerase III, while the three other rRNA molecules are transcribed as a long 35S polycistronic precursor by RNA polymerase I.[147] *Ribosome biogenesis* and *rRNA processing* are universal cellular processes, which encompass complicated series of events involving hundreds of transiently interacting components. It has been shown, for example, that in *Saccharomyces cerevisiae* the biogenesis of pre-18S ribosomal RNA is controlled by a large *ribonucleoprotein* (RNP) complex, which contains the U3 small nucleolar RNA (snoRNA) and 28 proteins.[148] The analysis yielded five small subunit ribosomal proteins (Rps4, Rps6, Rps7, Rps14 and Rps28) among other proteins. Intriguingly, in eukaryotic cells, ribosomal protein S6 (Rps6) is the major *phosphorylated* protein on the small ribosomal subunit,[149] suggesting that this protein might contain functionally important intrinsically disordered regions (see below, section dedicated to the *posttranslational modifications*). Furthermore, bioinformatics analysis revealed that 14 of the U three proteins (Utps) bear different *repeats* comprising crucial regions of their *protein-protein interaction* domains (WD repeats, coiled-coil domains, HEAT repeats and a crooked-neck-like (crnlike) tetratrico peptide repeat (TPR)).[148] The crn-like TPR is found in several proteins involved in other RNA processing events including *pre-mRNA splicing* (Prp42, Prp6 and Clf1) and polyadenylation (RNA14).[150] NMR analysis of the solution structure of the cytosolic TPR domain of Tom20 in the complex with the presequence peptide revealed that the C-terminal region of this protein (residues 105–145) is disordered.[151]

Chondrogenesis—This is the earliest phase of skeletal development, the process by which the cartilage is formed. Cartilage is an elastic connective tissue found in such parts of the body as the joints, outer ear, and larynx. Furthermore, cartilage represents the major constituent of the embryonic and young vertebrate skeleton, which is converted largely to bone with maturation. *Chondrogenesis* involves multiple steps, including mesenchymal cell recruitment and migration, condensation of progenitors, chondrocyte differentiation, and maturation, resulting in the formation of cartilage and bone during endochondral ossification.[152] This complex process is precisely controlled by interactions with the surrounding matrix, growth and differentiation factors, as well as other environmental factors responsible for the initiation or suppression of the cellular signaling pathways and for the regulation of transcription of specific genes.[153] For example, the development of vertebrate limb is controlled by the fibroblast growth factor, bone morphogenetic protein (BMP, a secreted signaling molecule, multifunctional growth factor belonging to the transforming growth factor β superfamily), Wnt and hedgehog pathways.[154] Recently, crucial roles of different mediators (including GADD45β, transcription factors of the Dlx, βHLH, leucine zipper, and AP-1 families, and the Wnt/β-catenin pathway) that interact at different stages during chondrogenesis have been revealed.[153] Also, members of the mammalian RUNX protein family, which includes three *transcription factors* RUNX1, RUNX2, and RUNX3, are expressed during chondrogenesis. These *transcription factors* also play active roles in mesenchymal condensation, chondrocyte proliferation, and chondrocyte maturation, and regulate transcription of target genes.[155] Thus, regulation and control of chondrogenesis involve multiple players, many of which possess functional disordered regions. For example, the abundance and functional roles of intrinsic disorder in *transcription factors* were already discussed (see section entitled *Transcription and transcription regulation*), whereas the role of disorder in the *leucine zippers* will be discussed in the second paper of this series.[29]

Growth regulation—Numerous proteins and pathways are implemented in *growth regulation*. For example, *cyclin* G was shown to be highly expressed in regenerating hepatocytes and motoneurons and in rapidly growing cancer cells and to have growth-promoting functions.[156, 157] *Cyclin* G interacts with cyclin-dependent kinase 5 (cdk5) and GAK, a cyclin G-associated kinase,[158] as well as with with the B′ subclass of PP2A phosphatase.[159] In addition, cyclin G directly interacts with Mdm2 and can stimulate the ability of PP2A to dephosphorylate Mdm2.[159] Furthermore, cyclin G was one of the earliest p53 target genes to be identified.[160] This suggests that cyclin G is a key regulator of the p53-Mdm2 network. The role of intrinsic disorder in p53 function was already discussed.

Functions associated with intrinsically disordered proteins

Table 3 presents a list of the top 20 SwissProt functional keywords associated with intrinsic disorder.

Ribonucleoproteins—Numerous facts have been accumulated to demonstrate intrinsic disorder is crucial for function of different *ribonucleoproteins*. In fact, ribonucleoprotein assembly is nearly always accompanied by changes in the conformation of the interacting RNA or protein, or both.[161–163] For example, the inter-domain linkers of sex-lethal protein (SXL) possess significant disorder, which provides the RNA recognition motifs (RRMs) with a possibility to be flexibly tethered in solution.[164] Another example is ribonuclease P (RNase P), a ribonucleoprotein complex containing one RNA subunit and at least one protein subunit. RNase P is involved in pre-tRNA processing.[165] In *E.coli*, RNase P consists of a small (119 amino acid residue) C5 protein bound to the much larger (377 nucleotide) P RNA subunit.[166] The C5 protein of *E.coli* is essentially disordered in buffer alone, but gains significant amount of ordered secondary structure in an anion-dependent manner.[167] A similar behavior was also described for the *Bacillus subtilis* RNase P.[168]

Ribosomal proteins—The assembly of the ribosome, which involves the sequential binding of numerous proteins *via* multiple pathways leading to large-scale changes in the conformation of the associated RNA and proteins, represents an extreme case involving dramatic structural changes induced by protein-RNA interaction.[169–172] In fact, many ribosomal proteins have been shown to be significantly disordered prior binding to rRNA and to acquire ordered structure during ribosome formation.[7, 16, 17]

Developmental proteins—α-Fetoprotein (AFP), a member of the family of albumin-like proteins, is a serum glycoprotein belonging to the intriguing class of onco-developmental polypeptides. AFP is homologous to human serum albumin (HSA).[173] Similarly to HAS, AFP is able to bind a number of small molecules, including metal ions, estrogens and different fatty acids (reviewed in [174]). Importantly, the removal of all ligands from AFP is accompanied by a complete loss of rigid 3-D structure (reviewed in [174]).

Hormones and growth factors—Growth hormone (GH), prolactin (PRL) and placental lactogen (PL), being the pituitary hormones, are members of an extensive cytokine superfamily of hormones and receptors that share many of the same general structure-function relationships in expressing their biological activities.[175] These hormones were shown to have two receptor-binding sites that have different topographies and electrostatic character. This feature is crucial for the regulation of these systems by producing binding surfaces with dramatically different binding affinities to the receptor extracellular domains. The receptor evidently possesses an exceptional conformational plasticity to be able to bind the topographically dissimilar sites on the hormone.[175] Human parathyroid-hormone-related protein (hPTHrP) is a hormone that is over-expressed by a large number of tumors and is produced by a variety of normal cells. The N-terminal fragment (1–34) of hPTHrP is responsible for the major biological functions of this hormone. Furthermore, this fragment is mostly unstructured possessing only a small content of α-helical secondary structure.[176] Secretin, a gut hormone consisting of 27 amino acid residues, was shown to be completely unfolded in aqueous solution but gain a fully ordered structure in the presence of 40% trifluoroethanol.[177]

Activators, repressors, cytokines, protease inhibitors, antimicrobial peptides and amphibian defense peptides—Several other function-related keywords, that are associated with intrinsically disordered proteins or regions are illustrated below. Protein AphA is a homodimeric member of a family of transcriptional *activators*. Transcriptional *activators*, including nuclear receptors, activate target genes through two broad classes of co-activators – those that remodel/modify chromatin and those that directly interfere with the general transcription machinery to facilitate formation and/or function of the preinitiation complexes. The AphA monomer is highly unstable by itself (i.e., likely it is highly disordered) and the dimer is formed in such a way that the two AphA chains wrap around each other,[178] suggesting that the dimmer arose via a disorder-to-order transition. The first 61 amino acid residues of the DNA-free lac *repressor* (i.e., a fragment which includes headpiece and the hinge region) are disordered, and thus are unobserved in the crystal structure.[179] Lymphotactin (Ltn) is unique member of family of pro-inflammatory activation-inducible *cytokines*. Ltn possesses a unique C-terminal extension, which is required for biological activity[180, 181] and which is disordered and highly mobile.[182] Analysis of the Bowman-Birk *protease inhibitor* by Raman optical activity revealed that it possesses a "static" type of disorder similar to that in disordered states of poly(L-lysine) and poly(L-glutamic acid).[183] The pediocin-like class IIa bacteriocins, which are *antimicrobial peptides* from lactic acid bacteria,[184, 185] were shown to be significantly unstructured in water.[186] Similarly, hylaseptin P1, an *amphibian defense peptide*, is in a random coil conformation in aqueous solutions.[187]

Neuropeptides—Pituitary adenylate cyclase activating polypeptide (PACAP),[188] which occurs naturally in two forms consisting of a 38 amino acid peptide amide (PACAP38) and its 27 amino acid N-terminus (PACAP27), belongs to the secretin/glucagons/*vasoactive* intestinal peptide (VIP) family.[189] Structural analysis of PACAP38 and PACAP27 revealed that these two neuropeptides are mostly disordered and retain only small transitory amounts of stable structure in aqueous solution.[190] Other opioid peptides are the enkephalins. The term enkephalin mainly refers to two peptides, [Met]-enkephalin and [Leu]-enkephalin, that both are products of the proenkephalin gene. [Met]-enkephalin is Tyr-Gly-Gly-Phe-Met; [Leu]-enkephalin has Leu in place of Met. Recently performed structural characterization of methionine and leucine enkephalins by hydrogen/deuterium exchange and electrospray ionization tandem mass spectrometry revealed that the monomer forms of both peptides adopt an unfolded conformation in aqueous solvent, whereas they prefer β-turn secondary structure under the membranemimetic environment.[191]

GTPase activation and GTPase-activating proteins (GAPs)—The GTP-GDP conversion by guanine nucleotide binding proteins (GNBPs) represents an important timer in intracellular signaling and transport processes. GNBPs are highly abundant in different genomes. For example, there are at least 140 small GTPases encoded in human (including the Ras, Rho, Arf, Rab and Ran GTPases), with various subclasses of this protein superfamily being implicated in almost all aspects of cell biology, including proliferation, nucleocytoplasmic transport, differentiation, vesicle trafficking, cytoskeletal organization and gene expression.[192] These small GTPases are considered to be molecular switches, the cycling of which between active and inactive forms is regulated by cellular factors.[192] There are two major classes of GNBP regulators, the guanine nucleotide exchange factors (GEFs), which promote the formation of active GTP-bound GTPases and the GTPase activating proteins (GAPs), which promote GTPase inactivation by stimulating GTP-hydrolysis activity.[193] In fact, the natural rate of GNBP-mediated GTP hydrolysis is slow but the reaction is accelerated by up to five orders of magnitude by the interaction of GNBPs with GAPs.[194] At least 160 human genes have been recently predicted to encode proteins that resemble GAPs for various members of the Ras GPTase superfamily.[195] Furthermore, ~ 0.5% of all predicted human genes likely encode GAPs suggesting that these proteins have widespread and important roles in GTPase regulation. Finally, such famous domains as ankyrin, BAR, BTK, CH, CNH, PDZ, PTB, RUN, SAM, SH2, SH3, WW and many others are all GAPs.[196]

Chromatin regulator—Several nuclear proteins serve as *chromatin regulators*, being involved in modulation of chromosome structure, chromatin and nucleosome remodeling and therefore playing a role in the controlling of gene transcription. Members of the HMGA family of non-histone chromatin proteins (formerly known as HMGI/Y proteins) serve as an illustrative example of such *chromatin regulators*.[197] HMGA proteins are the founding members of a new class of regulatory elements called ‘architectural transcription factors’ that participate in a wide variety of cellular processes including regulation of inducible gene transcription, integration of retroviruses into chromosomes, the induction of neoplastic transformation and promotion of metastatic progression of cancer cells.[198] HMGA proteins are highly flexible and are characterized by the total lack of ordered structure.[199–201]

Pyrogen—Substances that can cause a rise in body temperature are known as *pyrogens*. Fever is the multiphasic response of elevation and decline of the body core temperature regulated by central thermoregulatory mechanisms localized in the preoptic area of the hypothalamus. Some *cytokines* (which are highly inducible, secreted proteins mediating intercellular communication in the nervous and immune system), including interleukin 1 (IL-1), interleukin 6 (IL-6) and the tumor necrosis factor alpha (TNFα), act as endogenous pyrogens.[202] The role of intrinsic disorder in cytokine function was already discussed.

Opioid peptides and endorphin—*Opioid peptides* are short natural peptides that mimic the effect of opiates in the brain and therefore are potent pain suppressants. Some opioid peptides (e.g., endorphin, dynorphin and endorphin) are produced endogenously, some are produced by microbes (deltorphins and dermorphine), whereas others are absorbed from partially digested food (casomorphins, exorphins, and rubiscolins). Opioid peptides mediate their physiological and pharmacological effects through three major opioid receptor types (μ, δ, κ),[203] which are the members of the G protein-coupled receptor (GPCR) family.[204] The ^{1}H-NMR spectra of human β-*endorphin* (the largest natural *opioid peptide* of 31 amino acid residues) indicate that the peptide exists in random-coil conformation in aqueous solution but becomes helical in mixed solvent.[205]

Inhibitors—The activity of many important proteins is regulated by specific proteins-inhibitors. It has been already mentioned that the functionality of *Cdk inhibitor proteins* relies on intrinsic disorder. *Serine protease inhibitor* elafin is a 57 amino acid residue peptide inhibiting human leukocyte elastase, porcine pancreatic elastase and proteinase-3.[206] Elafin was shown to be almost completely unfolded in aqueous solutions.[207]

Protein phosphatase inhibitors—Calcineurin (CaN) is a calcium- and calmodulin-dependent protein serine/threonine phosphate, which is critical for several important cellular processes, including T-cell activation.[208] CaN is a heterodimer composed of subunits A and B (CaNA and CaN, respectively).[208] CaN phosphatase activity is regulated by Ca^{2+} binding to CaNB and by Ca^{2+}-induced binding of calmodulin (CaM) to CaNA.[209] Activity of CaN is modulated by a number of factors, including an autoinhibitory domain (residues 457–479), which binds in the active site cleft in the absence of Ca^{2+}/CaM and inhibits the enzyme, acting in concert with the CaM binding domain to confer CaM regulation.[209] Analysis of CaN crystal structure revealed that CaNA residues 1–13, 374–468, and 487–521 are not visible in the electron density map; i.e., disordered.[209] Importantly, long disordered region 374–468 includes the CaM-binding helix.[209] Therefore, this transient helix in CaN becomes bound and surrounded by CaM, turning on the CaN's serine/threonine phosphatase activity. Locating the CaN helix within the disordered region is essential for enabling CaM to surround its target upon binding.[6]

Cyclin—The progression of cells through the *cell cycle* is regulated by several specific proteins, known as *cyclins* and cyclin-dependent kinases. The concentrations of cyclins vary in a cyclical fashion during the cell cycle. They are produced or degraded as needed in order to drive the cell through the different stages of the cell cycle. The crucial role of intrinsic disorder in function and regulation of Cdk–cyclin complexes was already discussed.

Biological processes and functions associated with ordered proteins

Table 4 and Table 5 list the top 20 biological processes and functions that are significantly associated with predicted order. An examination of order-correlated keywords suggests the presence of 5 major functional categories: (1) catalysis (this category includes all functions listed in Table 5; i.e., *oxidoreductase, transferase, hydrolase, lyase, glycosidase, kinase, isomerase, ligase, decarboxylase, glycosyltransferase, protease, acyltransferase, monooxygenase, aminotransferase, metalloprotease, methyltransferase, aminoacyl-tRNA synthetase, aminopeptidase*, and *dioxygenase*); (2) transport (*electron transport, sugar transport*, and *transport*), (3) biosynthesis (*amino-acid biosynthesis, GMP biosynthesis, gluconeogenesis, amino-acid biosynthesis, pyrimidine biosynthesis, peptidoglycan synthesis, lipopolysaccharide biosynthesis, aromatic amino acid biosynthesis, branched-chain amino acid biosynthesis, purine biosynthesis, lipid a biosynthesis*, and *lipid synthesis*) (4) metabolism (*carbohydrate metabolism, tricarboxylic acid cycle, aromatic hydrocarbons catabolism*, and *one-carbon metabolism*) (5) trans-membrane proteins (porins). Note that catalysis overlaps strongly with biosynthesis and metabolism in that all of the proteins associated with these keywords are enzymes. The proteins associated with transport are often membrane proteins and are necessarily structured so that their backbone hydrogen bonds are formed in the low dielectric environment of the membrane. Other transport-associated proteins (which are not the membrane proteins) often need to bind very small molecules (or a single atom such as metals, or even electrons), which requires a precise coordination, and therefore, a well-defined structure. Proteins from the fifth category, *porins*, represent are transmembrane proteins that are large enough to facilitate passive diffusion. They are prevalent in the outer membrane of the mitochondria and Gram-negative bacteria. Porins are almost entirely composed of beta sheets and control the diffusion of small metabolites like sugars, ions, and amino acids.[210] They have been shown to have a highly stable structure using various characterization methods.[211] For example, porin from *Paracoccus denitrificans* is extremely stable toward heat, pH, and chemical denaturants.[212] Thus, current result agrees with our previous observations that proteins involved in catalysis, transport, biosynthesis and metabolism are less disordered than regulatory proteins.

Interestingly, the smaller disorder content as well as the greater amount of structural information (*e.g*. a greater PDB coverage) has previously been reported for proteins from the first four functional categories as compared to highly disordered signaling and cancer-associated proteins.[16] Thus, the current result agrees with our previous observations that proteins involved in catalysis, transport, biosynthesis and metabolism are less disordered than regulatory proteins.

Finally, one noticeable exception should be mentioned here. Although glycosidases are among the top 20 proteins with predicted functional order (Table 5), many of them in fact possess large disordered regions, even though their catalytic function requires a well defined structure. This is especially true for cellulases (Biological process: cellulose degradation, strong correlation with predicted order, see Table S1) for which protein disorder has been experimentally determined.[213, 214] These cellulases are composed of a catalytic domain, linked to a cellulose binding domain through a long disordered linker (109 amino acid residues in Cel5G, an endoglucanase from *Pseudoalteromonas haloplanktis*), which could be considered as an entropic spring. In fact, the SAXS analysis of dimensions, shape, and conformation of Cel5G full length in solution and especially of the linker between the catalytic module and the cellulose-binding module revealed that the linker is unstructured, and unusually long and flexible.[213] This modular organization and the presence of a disordered linker are crucial to optimize the biphasic process of crystalline cellulose degradation.

Another example of an enzyme that possesses functional disordered regions is retinaldehyde dehydrogenase II (RalDH2).[215] This enzyme converts retinal to the transcriptional regulator retinoic acid in the developing embryo. It has been shown that a 20-amino acid span in the substrate access channel is disordered, but folds during the course of catalysis and provides a means for an enzyme that requires a large substrate access channel to restrict access to the catalytic machinery by smaller compounds that might potentially enter the active site and be metabolized.[215] Therefore, RalDH2 represents a unique example of a protein that exhibits a catalytic activity in which a large disordered region folds upon catalysis.

Comparing the identified disorder functions with literature findings

Recently, literature analysis identified 28 functions associated with 98 confirmed disordered regions containing 30 or longer contiguous disorder residues.[8, 9] These functions were grouped into four broad categories: molecular recognition, molecular assembly, protein modification, and entropic chains. Entropic chains carry out functions that depend directly on the disordered state, and so such functions are simply outside the capabilities of fully folded structures.[8, 9] The use of partially folded subunits for molecular assembly appears to have significant advantages compared to the use of ordered subunits.[21, 22] Molecular recognition appears to be a common function for both ordered and disordered proteins: molecular recognition by disordered proteins may be primarily used for signaling whereas recognition by ordered proteins may be primarily used for catalysis,[8, 9] or for the assembly of functional complexes. Finally, sites of some types of posttranslational modification frequently occur within the regions with very strong preference for disorder.[8–11, 18, 19, 22, 216]

Out of 28 functions associated with the confirmed disordered regions, 13 were also found in the Swiss-Prot keyword list. Eleven of these functions are strongly correlated with predicted disorder with p-value above 0.95. Table 6 lists these 11 functions together with their z-scores. Furthermore, previously reported[16] strong functional correlations to intrinsic disorder were also found by using the method proposed in this study. These results strongly support the validity of the proposed statistical methodology for finding disorder-correlated functions.

Conclusions

We proposed a statistical approach that estimates the correlations between protein structure and protein function. As an application, we studied the relationship between intrinsic protein disorder and function in the Swiss-Prot database. Overall, 238 Swiss-Prot functional keywords were discovered to be strongly associated with predicted intrinsic disorder, and 302 function keywords were shown to be strongly correlated with predicted order. We validated a significant fraction of these findings by comparing them to literature data. The numerous correlations between the known experimental data and the results of the analysis demonstrate the general validity of our approach. However, a more thorough comparison with literature data will be needed to determine the frequencies with which exceptions occur as compared to the observed trends. In our original, manual curation of disorder-function correlations, we determined that each function was unambiguously associated with a specific region of disorder.[8, 9] While the current methodology does not determine whether each function is directly associated with a disordered segment, the literature data verify that these functions are indeed, for the most part, associated with regions of disorder. Of special interest is that disease related proteins were shown to have the high correlation with disordered regions of proteins (see the last paper of this series[30]). Overall, this approach provides an innovative and relevant method to examine protein structure-function relationships.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

The authors express their deepest gratitude to Celeste Brown and Predrag Radivojac for numerous valuable discussions. This work was supported by grants from the National Institutes of Health LM007688-0A1 (A. K. D. and Z. O.) and GM071714-01A2 (A.K.D and V.N.U.), and by the Indiana Genomics Initiative (INGEN) (A. K. D.). INGEN is supported in part by Lilly Endowment Inc. The Programs of the Russian Academy of Sciences for the "Molecular and cellular biology" and "Fundamental science for medicine" provided partial support to V. N. U., and L. M. I. was supported by the NSF grant MCB 0444818.

References

1. Fischer E. Einfluss der Configuration auf die Wirkung der Enzyme. Ber. Dt. Chem. Ges 1894;27:2985–2993.
2. Wu H. Studies on denaturation of proteins XIII A theory of denaturation. Chin. J. Physiol 1931;1:219–234.
3. Mirsky AE, Pauling L. On the structure of native, denatured, and coagulated proteins. Proc Natl Acad Sci U S A 1936;22:439–447. [PubMed: 16577722]

4. Sela M, White FH Jr, Anfinsen CB. Reductive cleavage of disulfide bridges in ribonuclease. Science 1957;125:691–692. [PubMed: 13421663]

5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242. [PubMed: 10592235]

6. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. Intrinsically disordered protein. J Mol Graph Model 2001;19:26–59. [PubMed: 11381529]

7. Uversky VN. Natively unfolded proteins: a point where biology waits for physics. Protein Sci 2002;11:739–756. [PubMed: 11910019]

8. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry 2002;41:6573–6582. [PubMed: 12022860]

9. Dunker AK, Brown CJ, Obradovic Z. Identification and functions of usefully disordered proteins. Adv Protein Chem 2002;62:25–49. [PubMed: 12418100]

10. Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci 2002;27:527–533. [PubMed: 12368089]

11. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 1999;293:321–331. [PubMed: 10550212]

12. Dunker AK, Obradovic Z. The protein trinity--linking function and disorder. Nat Biotechnol 2001;19:805–806. [PubMed: 11533628]

13. Uversky VN. What does it mean to be natively unfolded? Eur J Biochem 2002;269:2–12. [PubMed: 11784292]

14. Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK. Comparing and combining predictors of mostly disordered proteins. Biochemistry 2005;44:1989–2000. [PubMed: 15697224]

15. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform 2000;11:161–171.

16. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 2002;323:573–584. [PubMed: 12381310]

17. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 2000;41:415–427. [PubMed: 11025552]

18. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. Curr Opin Struct Biol 2002;12:54–60. [PubMed: 11839490]

19. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208. [PubMed: 15738986]

20. Daughdrill, GW.; Pielak, GJ.; Uversky, VN.; Cortese, MS.; Dunker, AK. Natively disordered proteins. In: Buchner, J.; Kiefhaber, T., editors. Handbook of Protein Folding. Weinheim, Germany: Wiley-VCH, Verlag GmbH & Co. KGaA; 2005. p. 271-353.

21. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets. The roles of intrinsic disorder in protein interaction networks. Febs J 2005;272:5129–5148. [PubMed: 16218947]

22. Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J Mol Recognit 2005;18:343–384. [PubMed: 16094605]

23. Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, Uversky VN, Vidal M, Iakoucheva LM. Intrinsic Disorder is a Common Feature of Hub Proteins from Four Eukaryotic Interactomes. PLoS Comput Biol. 2006in press

24. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, Cortese MS, Lawson JD, Brown CJ, Sikes JG, Newton CD, Dunker AK. DisProt: a database of protein disorder. Bioinformatics 2005;21:137–140. [PubMed: 15310560]

25. Dyson HJ, Wright PE. According to current textbooks, a well-defined three-dimensional structure is a prerequisite for the function of a protein. Is this correct? IUBMB Life 2006;58:107–109. [PubMed: 16608823]

26. Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK, Obradovic Z. Optimizing long intrinsic disorder predictors with protein evolutionary information. J Bioinform Comput Biol 2005;3:35–60. [PubMed: 15751111]

27. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365–370. [PubMed: 12520024]

28. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 2004;337:635–645. [PubMed: 15019783]

29. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. Functional anthology of intrinsic disorder. II. Cellular components, domains, technical terms, developmental processes and coding sequence diversity associated with long disordered regions. J Proteome Res. 2006

30. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. Functional anthology of intrinsic disorder. III. Ligands, postranslational modifications and diseases associated with long disordered regions. J Proteome Res. 2006

31. O'Donovan C, Martin MJ, Glemet E, Codani JJ, Apweiler R. Removing redundancy in SWISS-PROT and TrEMBL. Bioinformatics 1999;15:258–259. [PubMed: 10222414]

32. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 2002;30:1575–1584. [PubMed: 11917018]

33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410. [PubMed: 2231712]

34. Romero P, Obradovic Z, Dunker AK. Sequence data analysis for long disordered regions prediction in the calcineurin family. Genome Informatics 1997;8:110–124. [PubMed: 11072311]

35. Romero, P.; Obradovic, Z.; Kissinger, C.; Villafranca, JE.; Dunker, AK. Identifying disordered regions in proteins from amino acid sequence; 1997 Proceedings of International Conference on Neural Networks; 1997. p. 90-95.

36. Xie Q, Arnold GE, Romero P, Obradovic Z, Garner E, Dunker AK. The Sequence Attribute Method for Determining Relationships Between Sequence and Protein Disorder. Genome Inform Ser Workshop Genome Inform 1998;9:193–200.

37. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins 2001;42:38–48. [PubMed: 11093259]

38. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK. DisProt: the Database of Disordered Proteins. Nucleic Acids Res 2007;35:D786–D793. [PubMed: 17145717]

39. Blau HM, Baltimore D. Differentiation requires continuous regulation. J Cell Biol 1991;112:781–783. [PubMed: 1999456]

40. Adams JC, Watt FM. Regulation of development and differentiation by the extracellular matrix. Development 1993;117:1183–1198. [PubMed: 8404525]

41. Watt FM. The extracellular matrix and cell shape. Trends Biochem Sci 1986;11:482–485.

42. Hay, ED. Cell Biology of Extracellular Matrix. New York: Plenum Press; 1981.

43. Ruoslahti E, Pierschbacher MD. New perspectives in cell adhesion: RGD and integrins. Science 1987;238:491–497. [PubMed: 2821619]

44. House-Pompeo K, Xu Y, Joh D, Speziale P, Hook M. Conformational changes in the fibronectin binding MSCRAMMs are induced by ligand binding. J Biol Chem 1996;271:1379–1384. [PubMed: 8576127]

45. Kim JH, Singvall J, Schwarz-Linek U, Johnson BJ, Potts JR, Hook M. BBK32, a fibronectin binding MSCRAMM from Borrelia burgdorferi, contains a disordered region that undergoes a conformational change on ligand binding. J Biol Chem 2004;279:41706–41714. [PubMed: 15292204]

46. Campbell KM, Terrell AR, Laybourn PJ, Lumb KJ. Intrinsic structural disorder of the C-terminal activation domain from the bZIP transcription factor Fos. Biochemistry 2000;39:2708–2713. [PubMed: 10704222]

47. Haynes C, Iakoucheva LM. Serine/arginine-rich splicing factors belong to a class of intrinsically disordered proteins. Nucleic Acids Res 2006;34:305–312. [PubMed: 16407336]

48. Bhalla J, Storchan GB, Maccarthy CM, Uversky VN, Tcherkasskaya O. Local flexibility in molecular function paradigm. Mol Cell Proteomics. 2006

49. Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK. Intrinsic Disorder in Transcription Factors. Biochemistry 2006;45:6873–6888. [PubMed: 16734424]

50. Minezaki Y, Homma K, Kinjo AR, Nishikawa K. Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. J Mol Biol 2006;359:1137–1149. [PubMed: 16697407]

51. Lewis JD, Abbott DW, Ausio J. A haploid affair: core histone transitions during spermatogenesis. Biochem Cell Biol 2003;81:131–140. [PubMed: 12897846]

52. Ausio, J. Histone H1 and the evolution of the nuclear sperm-specific proteins. In: Jamieson, BGM.; Ausio, J.; Justine, JL., editors. Advances in Spermatozoal Phylogeny and Taxonomy. Paris, France: Memoires du Museum National d'Histoire Naturelle; 1995.

53. Ausio J, Subirana JA. A high molecular weight nuclear basic protein from the bivalve mollusc Spisula solidissima. J Biol Chem 1982;257:2802–2805. [PubMed: 7061451]

54. Ausio J, Toumadje A, McParland R, Becker RR, Johnson WC Jr, van Holde KE. Structural characterization of the trypsin-resistant core in the nuclear sperm-specific protein from Spisula solidissima. Biochemistry 1987;26:975–982. [PubMed: 3567164]

55. Lewis JD, Ausio J. Protamine-like proteins: evidence for a novel chromatin structure. Biochem Cell Biol 2002;80:353–361. [PubMed: 12123288]

56. Lewis JD, McParland R, Ausio J. PL-I of Spisula solidissima, a highly elongated sperm-specific histone H1. Biochemistry 2004;43:7766–7775. [PubMed: 15196019]

57. Harvey AC, Downs JA. What functions do linker histones provide? Mol Microbiol 2004;53:771–775. [PubMed: 15255891]

58. Isenberg I. Histones. Annu Rev Biochem 1979;48:159–191. [PubMed: 382983]

59. Boublik M, Bradbury EM, Crane-Robinson C. An investigation of the conformational changes in histones F1 and F2a1 by proton magnetic resonance spectroscopy. Eur J Biochem 1970;14:486–497. [PubMed: 5530727]

60. Li HJ, Wickett R, Craig AM, Isenberg I. Conformational changes in histone IV. Biopolymers 1972;11:375–397. [PubMed: 5016554]

61. Wickett RR, Li HJ, Isenberg I. Salt effects on histone IV conformation. Biochemistry 1972;11:2952–2957. [PubMed: 4339475]

62. D'Anna JA Jr, Isenberg I. Conformational changes of histone ARE(F3, III). Biochemistry 1974;13:4987–4992. [PubMed: 4474008]

63. D'Anna JA Jr, Isenberg I. Conformational changes of histone LAK (f2a2). Biochemistry 1974;13:2093–2098. [PubMed: 4857059]

64. Munishkina LA, Fink AL, Uversky VN. Conformational prerequisites for formation of amyloid fibrils from histones. J Mol Biol 2004;342:1305–1324. [PubMed: 15351653]

65. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 1997;389:251–260. [PubMed: 9305837]

66. Arents G, Burlingame RW, Wang BC, Love WE, Moudrianakis EN. The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. Proc Natl Acad Sci U S A 1991;88:10148–10152. [PubMed: 1946434]

67. Hansen JC. Conformational dynamics of the chromatin fiber in solution: determinants, mechanisms, and functions. Annu Rev Biophys Biomol Struct 2002;31:361–392. [PubMed: 11988475]

68. Hansen JC, Lu X, Ross ED, Woody RW. Intrinsic protein disorder, amino acid composition, and histone terminal domains. J Biol Chem 2006;281:1853–1856. [PubMed: 16301309]

69. Walker IO. Differential dissociation of histone tails from core chromatin. Biochemistry 1984;23:5622–5628. [PubMed: 6509040]

70. Mangenot S, Leforestier A, Vachette P, Durand D, Livolant F. Salt-induced conformation and interaction changes of nucleosome core particles. Biophys J 2002;82:345–356. [PubMed: 11751321]

71. Morgan DO. Principles of CDK regulation. Nature 1995;374:131–134. [PubMed: 7877684]

72. Morgan DO. Cyclin-dependent kinases: engines, clocks, and microprocessors. Annu Rev Cell Dev Biol 1997;13:261–291. [PubMed: 9442875]

73. Nigg EA. Mitotic kinases as regulators of cell division and its checkpoints. Nat Rev Mol Cell Biol 2001;2:21–32. [PubMed: 11413462]

74. Pavletich NP. Mechanisms of cyclin-dependent kinase regulation: structures of Cdks, their cyclin activators, and Cip and INK4 inhibitors. J Mol Biol 1999;287:821–828. [PubMed: 10222191]

75. Kriwacki RW, Hengst L, Tennant L, Reed SI, Wright PE. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. Proc Natl Acad Sci U S A 1996;93:11504–11509. [PubMed: 8876165]

76. Flaugh SL, Lumb KJ. Effects of macromolecular crowding on the intrinsically disordered proteins c-Fos and p27(Kip1). Biomacromolecules 2001;2:538–540. [PubMed: 11749217]

77. Bienkiewicz EA, Adkins JN, Lumb KJ. Functional consequences of preorganized helical structure in the intrinsically disordered cell-cycle inhibitor p27(Kip1). Biochemistry 2002;41:752–759. [PubMed: 11790096]

78. Lacy ER, Filippov I, Lewis WS, Otieno S, Xiao L, Weiss S, Hengst L, Kriwacki R. W. p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding. Nat Struct Mol Biol 2004;11:358–364. [PubMed: 15024385]

79. Adkins JN, Lumb KJ. Intrinsic structural disorder and sequence features of the cell cycle inhibitor p57Kip2. Proteins 2002;46:1–7. [PubMed: 11746698]

80. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell 2000;101:671–684. [PubMed: 10892653]

81. Zhu J, Shendure J, Mitra RD, Church GM. Single molecule profiling of alternative pre-mRNA splicing. Science 2003;301:836–838. [PubMed: 12907803]

82. Tabuchi K, Sudhof TC. Structure and evolution of neurexin genes: insight into the mechanism of alternative splicing. Genomics 2002;79:849–859. [PubMed: 12036300]

83. Jurica MS, Moore MJ. Pre-mRNA splicing: awash in a sea of proteins. Mol Cell 2003;12:5–14. [PubMed: 12887888]

84. Anderson, CW.; Appella, E. Signaling to the p53 tumor suppressor through pathways activated by genotoxic and nongenotoxic stress. In: Bradshaw, RA.; Dennis, EA., editors. Handbook of Cell Signaling. New York: Academic Press; 2004. p. 237-247.

85. Hollstein M, Sidransky D, Vogelstein B, Harris CC. p53 mutations in human cancers. Science 1991;253:49–53. [PubMed: 1905840]

86. Balint EE, Vousden KH. Activation and activities of the p53 tumour suppressor protein. Br J Cancer 2001;85:1813–1823. [PubMed: 11747320]

87. Zhao R, Gish K, Murphy M, Yin Y, Notterman D, Hoffman WH, Tom E, Mack DH, Levine AJ. Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. Genes Dev 2000;14:981–993. [PubMed: 10783169]

88. Bell S, Klein C, Muller L, Hansen S, Buchner J. p53 contains large unstructured regions in its native state. J. Mol. Biol 2002;322:917–927. [PubMed: 12367518]

89. Lee H, Mok KH, Muhandiram R, Park KH, Suk JE, Kim DH, Chang J, Sung YC, Choi KY, Han KH. Local structural elements in the mostly unstructured transcriptional activation domain of human p53. J Biol Chem 2000;275:29426–29432. [PubMed: 10884388]

90. Dawson R, Muller L, Dehner A, Klein C, Kessler H, Buchner J. The N-terminal domain of p53 is natively unfolded. J Mol Biol 2003;332:1131–1141. [PubMed: 14499615]

91. Kussie PH, Gorina S, Marechal V, Elenbaas B, Moreau J, Levine AJ, Pavletich NP. Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. Science 1996;274:948–953. [PubMed: 8875929]

92. Rosal R, Pincus MR, Brandt-Rauf PW, Fine RL, Michl J, Wang H. NMR solution structure of a peptide from the mdm-2 binding domain of the p53 protein that is selectively cytotoxic to cancer cells. Biochemistry 2004;43:1854–1861. [PubMed: 14967026]

93. Vargas DA, Takahashi S, Ronai Z. Mdm2: A regulator of cell growth and death. Adv Cancer Res 2003;89:1–34. [PubMed: 14587869]

94. Ayed A, Mulder FA, Yi GS, Lu Y, Kay LE, Arrowsmith CH. Latent and active p53 are identical in conformation. Nat Struct Biol 2001;8:756–760. [PubMed: 11524676]

95. Weinberg RL, Freund SM, Veprintsev DB, Bycroft M, Fersht AR. Regulation of DNA binding of p53 by its C-terminal domain. J Mol Biol 2004;342:801–811. [PubMed: 15342238]

96. Cho Y, Gorina S, Jeffrey PD, Pavletich NP. Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science 1994;265:346–355. [PubMed: 8023157]

97. Joerger AC, Ang HC, Veprintsev DB, Blair CM, Fersht AR. Structures of p53 cancer mutants and mechanism of rescue by second-site suppressor mutations. J Biol Chem 2005;280:16030–16037. [PubMed: 15703170]

98. Clore GM, Ernst J, Clubb R, Omichinski JG, Kennedy WM, Sakaguchi K, Appella E, Gronenborn AM. Refined solution structure of the oligomerization domain of the tumour suppressor p53. Nat Struct Biol 1995;2:321–333. [PubMed: 7796267]

99. Lee W, Harvey TS, Yin Y, Yau P, Litchfield D, Arrowsmith CH. Solution structure of the tetrameric minimum transforming domain of p53. Nat Struct Biol 1994;1:877–890. [PubMed: 7773777]

100. Canadillas JM, Tidow H, Freund SM, Rutherford TJ, Ang HC, Fersht AR. Solution structure of p53 core domain: structural basis for its instability. Proc Natl Acad Sci U S A 2006;103:2109–2114. [PubMed: 16461916]

101. Veprintsev DB, Freund SM, Andreeva A, Rutledge SE, Tidow H, Canadillas JM, Blair CM, Fersht AR. Core domain interactions in full-length p53 in solution. Proc Natl Acad Sci U S A 2006;103:2115–2119. [PubMed: 16461914]

102. Hinds MG, Day CL. Regulation of apoptosis: uncovering the binding determinants. Curr Opin Struct Biol 2005;15:690–699. [PubMed: 16263267]

103. Puthalakath H, Strasser A. Keeping killers on a tight leash: transcriptional and post-translational control of the pro-apoptotic activity of BH3-only proteins. Cell Death Differ 2002;9:505–512. [PubMed: 11973609]

104. Liu X, Dai S, Zhu Y, Marrack P, Kappler JW. The structure of a Bcl-xL/Bim fragment complex: implications for Bim function. Immunity 2003;19:341–352. [PubMed: 14499110]

105. Sattler M, Liang H, Nettesheim D, Meadows RP, Harlan JE, Eberstadt M, Yoon HS, Shuker SB, Chang BS, Minn AJ, Thompson CB, Fesik SW. Structure of Bcl-xL-Bak peptide complex: recognition between regulators of apoptosis. Science 1997;275:983–986. [PubMed: 9020082]

106. Petros AM, Nettesheim DG, Wang Y, Olejniczak ET, Meadows RP, Mack J, Swift K, Matayoshi ED, Zhang H, Thompson CB, Fesik SW. Rationale for Bcl-xL/Bad peptide complex formation from structure, mutagenesis, and biophysical studies. Protein Sci 2000;9:2528–2534. [PubMed: 11206074]

107. Yan N, Gu L, Kokel D, Chai J, Li W, Han A, Chen L, Xue D, Shi Y. Structural, biochemical, and functional analyses of CED-9 recognition by the proapoptotic proteins EGL-1 and CED-4. Mol Cell 2004;15:999–1006. [PubMed: 15383288]

108. Fan J, Zhang Q, Tochio H, Li M, Zhang M. Structural basis of diverse sequence-dependent target recognition by the 8 kDa dynein light chain. J Mol Biol 2001;306:97–108. [PubMed: 11178896]

109. Hinds MG, Smits C, Fredericks-Short R, Risk JM, Bailey M, Huang DC, Day CL. Bim, Bad and Bmf: intrinsically unstructured BH3-only proteins that undergo a localized conformational change upon binding to prosurvival Bcl-2 targets. Cell Death Differ 2006;14:128–136. [PubMed: 16645638]

110. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK. Coupled folding and binding with alpha-helix-forming molecular recognition elements. Biochemistry 2005;44:12454–12470. [PubMed: 16156658]

111. Huang DT, Walden H, Duda D, Schulman BA. Ubiquitin-like protein activation. Oncogene 2004;23:1958–1971. [PubMed: 15021884]

112. Hay RT. Role of ubiquitin-like proteins in transcriptional regulation. Ernst Schering Res Found Workshop 2006:173–192. [PubMed: 16568955]

113. Catic A, Collins C, Church GM, Ploegh HL. Preferred in vivo ubiquitination sites. Bioinformatics 2004;20:3302–3307. [PubMed: 15256413]

114. Macauley MS, Errington WJ, Scharpf M, Mackereth CD, Blaszczak AG, Graves BJ, McIntosh LP. Beads-on-a-string, characterization of ETS-1 sumoylated within its flexible N-terminal sequence. J Biol Chem 2006;281:4164–4172. [PubMed: 16319071]

115. Logan CY, Nusse R. The Wnt signaling pathway in development and disease. Annu Rev Cell Dev Biol 2004;20:781–810. [PubMed: 15473860]

116. Kelleher FC, Fennelly D, Rafferty M. Common critical pathways in embryogenesis and cancer. Acta Oncol 2006;45:375–388. [PubMed: 16760173]

117. Clevers H. Wnt/beta-catenin signaling in development and disease. Cell 2006;127:469–480. [PubMed: 17081971]

118. Prud'homme B, Lartillot N, Balavoine G, Adoutte A, Vervoort M. Phylogenetic analysis of the Wnt gene family. Insights from lophotrochozoan members. Curr Biol 2002;12:1395. [PubMed: 12194820]

119. Kusserow A, Pang K, Sturm C, Hrouda M, Lentfer J, Schmidt HA, Technau U, von Haeseler A, Hobmayer B, Martindale MQ, Holstein TW. Unexpected complexity of the Wnt gene family in a sea anemone. Nature 2005;433:156–160. [PubMed: 15650739]

120. Gordon MD, Nusse R. Wnt signaling: multiple pathways, multiple receptors, and multiple transcription factors. J Biol Chem 2006;281:22429–22433. [PubMed: 16793760]

121. Moon RT, Bowerman B, Boutros M, Perrimon N. The promise and perils of Wnt signaling through beta-catenin. Science 2002;296:1644–1646. [PubMed: 12040179]

122. Kohn AD, Moon RT. Wnt and calcium signaling: beta-catenin-independent pathways. Cell Calcium 2005;38:439–446. [PubMed: 16099039]

123. Katoh M, Kirikoshi H, Saitoh T, Sagara N, Koike J. Alternative splicing of the WNT-2B/WNT-13 gene. Biochem Biophys Res Commun 2000;275:209–216. [PubMed: 10944466]

124. Pospisil H, Herrmann A, Butherus K, Pirson S, Reich JG, Kemmner W. Verification of predicted alternatively spliced Wnt genes reveals two new splice variants (CTNNB1 and LRP5) and altered Axin-1 expression during tumour progression. BMC Genomics 2006;7:148. [PubMed: 16772034]

125. Struewing IT, Toborek A, Mao CD. Mitochondrial and nuclear forms of Wnt13 are generated via alternative promoters, alternative RNA splicing, and alternative translation start sites. J Biol Chem 2006;281:7282–7293. [PubMed: 16407296]

126. Ding Y, Dale T. Wnt signal transduction: kinase cogs in a nano-machine? Trends Biochem Sci 2002;27:327–329. [PubMed: 12114015]

127. Dajani R, Fraser E, Roe SM, Young N, Good V, Dale TC, Pearl LH. Crystal structure of glycogen synthase kinase 3 beta: structural basis for phosphate-primed substrate specificity and autoinhibition. Cell 2001;105:721–732. [PubMed: 11440715]

128. Apuzzo S, Gros P. The paired domain of pax3 contains a putative homeodomain interaction pocket defined by cysteine scanning mutagenesis. Biochemistry 2006;45:7154–7161. [PubMed: 16752906]

129. Haubst N, Berger J, Radjendirane V, Graw J, Favor J, Saunders GF, Stoykova A, Gotz M. Molecular dissection of Pax6 function: the specific roles of the paired domain and homeodomain in brain development. Development 2004;131:6131–6140. [PubMed: 15548580]

130. Zhang F, Nakanishi G, Kurebayashi S, Yoshino K, Perantoni A, Kim YS, Jetten AM. Characterization of Glis2, a novel gene encoding a Gli-related, Kruppel-like transcription factor with transactivation and repressor functions. Roles in kidney development and neurogenesis. J Biol Chem 2002;277:10139–10149. [PubMed: 11741991]

131. Paratore C, Brugnoli G, Lee HY, Suter U, Sommer L. The role of the Ets domain transcription factor Erm in modulating differentiation of neural crest stem cells. Dev Biol 2002;250:168–180. [PubMed: 12297104]

132. Chalepakis G, Wijnholds J, Gruss P. Pax-3-DNA interaction: flexibility in the DNA binding and induction of DNA conformational changes by paired domains. Nucleic Acids Res 1994;22:3131–3137. [PubMed: 8065927]

133. Mishra R, Gorlov IP, Chao LY, Singh S, Saunders GF. PAX6, paired domain influences sequence recognition by the homeodomain. J Biol Chem 2002;277:49488–49494. [PubMed: 12388550]

134. Mauen S, Huvent I, Raussens V, Demonte D, Baert JL, Tricot C, Ruysschaert JM, Van Lint C, Moguilevsky N, de Launoit Y. Expression, purification, and structural prediction of the Ets transcription factor ERM. Biochim Biophys Acta 2006;1760:1192–1201. [PubMed: 16730909]

135. Hiraga S. Chromosome partition in Escherichia coli. Curr Opin Genet Dev 1993;3:789–801. [PubMed: 8274864]

136. Melby TE, Ciampaglio CN, Briscoe G, Erickson HP. The symmetrical structure of structural maintenance of chromosomes (SMC) and MukB proteins: long, antiparallel coiled coils, folded at a flexible hinge. J Cell Biol 1998;142:1595–1604. [PubMed: 9744887]

137. James LC, Tawfik DS. Conformational diversity and protein evolution--a 60-year-old hypothesis revisited. Trends Biochem Sci 2003;28:361–368. [PubMed: 12878003]

138. Pauling L. A theory of the structure and process of formation of antibodies. J Am Chem Soc 1940;62:2643–2657.

139. Carl PL, Temple BR, Cohen PL. Most nuclear systemic autoantigens are extremely disordered proteins: implications for the etiology of systemic autoimmunity. Arthritis Res Ther 2005;7:R1360–R1374. [PubMed: 16277689]

140. Slutzki M, Jaitin DA, Yehezkel TB, Schreiber G. Variations in the Unstructured C-terminal Tail of Interferons Contribute to Differential Receptor Binding and Biological Activity. J Mol Biol. 2006

141. Sigalov AB. Multichain immune recognition receptor signaling: different players, same game? Trends Immunol 2004;25:583–589. [PubMed: 15489186]

142. Sigalov AB. Immune cell signaling: a novel mechanistic model reveals new therapeutic targets. Trends Pharmacol Sci 2006;27:518–524. [PubMed: 16908074]

143. Sigalov A. Multi-chain immune recognition receptors: spatial organization and signal transduction. Semin Immunol 2005;17:51–64. [PubMed: 15582488]

144. Sigalov AB, Aivazian DA, Uversky VN, Stern LJ. Lipid-Binding Activity of Intrinsically Unstructured Cytoplasmic Domains of Multichain Immune Recognition Receptor Signaling Subunits. Biochemistry 2006;45:15731–15739. [PubMed: 17176095]

145. Sigalov A, Aivazian D, Stern L. Homooligomerization of the cytoplasmic domain of the T cell receptor zeta chain and of other proteins containing the immunoreceptor tyrosine-based activation motif. Biochemistry 2004;43:2049–2061. [PubMed: 14967045]

146. Sigalov AB, Zhuravleva AV, Orekhov VY. Binding of intrinsically disordered proteins is not necessarily accompanied by a structural transition to a folded form. Biochimie. 2006

147. Wehner KA, Baserga SJ. The sigma(70)-like motif: a eukaryotic RNA binding domain unique to a superfamily of proteins required for ribosome biogenesis. Mol Cell 2002;9:329–339. [PubMed: 11864606]

148. Dragon F, Gallagher JE, Compagnone-Post PA, Mitchell BM, Porwancher KA, Wehner KA, Wormsley S, Settlage RE, Shabanowitz J, Osheim Y, Beyer AL, Hunt DF, Baserga SJ. A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. Nature 2002;417:967–970. [PubMed: 12068309]

149. Hernandez VP, Higgins L, Schwientek MS, Fallon AM. The histone-like C-terminal extension in ribosomal protein S6 in Aedes and Anopheles mosquitoes is encoded within the distal portion of exon 3. Insect Biochem Mol Biol 2003;33:901–910. [PubMed: 12915181]

150. Chung S, McLean MR, Rymond BC. Yeast ortholog of the Drosophila crooked neck protein promotes spliceosome assembly through stable U4/U6.U5 snRNP addition. Rna 1999;5:1042–1054. [PubMed: 10445879]

151. Abe Y, Shodai T, Muto T, Mihara K, Torii H, Nishikawa S, Endo T, Kohda D. Structural basis of presequence recognition by the mitochondrial protein import receptor Tom20. Cell 2000;100:551–560. [PubMed: 10721992]

152. Olsen BR, Reginato AM, Wang W. Bone development. Annu Rev Cell Dev Biol 2000;16:191–220. [PubMed: 11031235]

153. Goldring MB, Tsuchimochi K, Ijiri K. The control of chondrogenesis. J Cell Biochem 2006;97:33–44. [PubMed: 16215986]

154. Tickle C. Molecular basis of vertebrate limb patterning. Am J Med Genet 2002;112:250–255. [PubMed: 12357468]

155. Yoshida CA, Komori T. Role of Runx proteins in chondrogenesis. Crit Rev Eukaryot Gene Expr 2005;15:243–254. [PubMed: 16390320]

156. Morita N, Kiryu S, Kiyama H. p53-independent cyclin G expression in a group of mature neurons and its enhanced expression during nerve regeneration. J Neurosci 1996;16:5961–5966. [PubMed: 8815878]

157. Skotzko M, Wu L, Anderson WF, Gordon EM, Hall FL. Retroviral vector-mediated gene transfer of antisense cyclin G1 (CYCG1) inhibits proliferation of human osteogenic sarcoma cells. Cancer Res 1995;55:5493–5498. [PubMed: 7585620]

158. Kanaoka Y, Kimura SH, Okazaki I, Ikeda M, Nojima H. GAK: a cyclin G associated kinase contains a tensin/auxilin-like domain. FEBS Lett 1997;402:73–80. [PubMed: 9013862]

159. Okamoto K, Kamibayashi C, Serrano M, Prives C, Mumby MC, Beach D. p53-dependent association between cyclin G and the B' subunit of protein phosphatase 2A. Mol Cell Biol 1996;16:6593–6602. [PubMed: 8887688]

160. Okamoto K, Beach D. Cyclin G is a transcriptional target of the p53 tumor suppressor protein. Embo J 1994;13:4816–4822. [PubMed: 7957050]

161. Williamson JR. Induced fit in RNA-protein recognition. Nat Struct Biol 2000;7:834–837. [PubMed: 11017187]

162. Leulliot N, Varani G. Current topics in RNA-protein recognition: control of specificity and biological function through induced fit and conformational capture. Biochemistry 2001;40:7947–7956. [PubMed: 11434763]

163. Chen Y, Varani G. Protein families and RNA recognition. Febs J 2005;272:2088–2097. [PubMed: 15853794]

164. Crowder SM, Kanaar R, Rio DC, Alber T. Absence of interdomain contacts in the crystal structure of the RNA recognition motifs of Sex-lethal. Proc Natl Acad Sci U S A 1999;96:4892–4897. [PubMed: 10220389]

165. Torres-Larios A, Swinger KK, Pan T, Mondragon A. Structure of ribonuclease P - a universal ribozyme. Curr Opin Struct Biol 2006;16:327–335. [PubMed: 16650980]

166. Harris ME, Christian EL. Recent insights into the structure and function of the ribonucleoprotein enzyme ribonuclease P. Curr Opin Struct Biol 2003;13:325–333. [PubMed: 12831883]

167. Guo X, Campbell FE, Sun L, Christian EL, Anderson VE, Harris ME. RNA-dependent Folding and Stabilization of C5 Protein During Assembly of the E. coli RNase P Holoenzyme. J Mol Biol. 2006

168. Henkels CH, Kurz JC, Fierke CA, Oas TG. Linked folding and anion binding of the Bacillus subtilis ribonuclease P protein. Biochemistry 2001;40:2777–2789. [PubMed: 11258888]

169. Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. The structural basis of ribosome activity in peptide bond synthesis. Science 2000;289:920–930. [PubMed: 10937990]

170. Recht MI, Williamson JR. RNA tertiary structure and cooperative assembly of a large ribonucleoprotein complex. J Mol Biol 2004;344:395–407. [PubMed: 15522293]

171. Schroeder R, Barta A, Semrad K. Strategies for RNA folding and assembly. Nat Rev Mol Cell Biol 2004;5:908–919. [PubMed: 15520810]

172. Klein DJ, Moore PB, Steitz TA. The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. J Mol Biol 2004;340:141–177. [PubMed: 15184028]

173. Deutsch HF. Chemistry and biology of alpha-fetoprotein. Adv Cancer Res 1991;56:253–312. [PubMed: 1709334]

174. Gillespie JR, Uversky VN. Structure and function of alpha-fetoprotein: a biophysical overview. Biochim Biophys Acta 2000;1480:41–56. [PubMed: 11004554]

175. Kossiakoff AA. The structural basis for biological signaling, regulation, and specificity in the growth hormone-prolactin system of hormones and receptors. Adv Protein Chem 2004;68:147–169. [PubMed: 15500861]

176. Gronwald W, Schomburg D, Tegge W, Wray V. Assessment by 1H NMR spectroscopy of the structural behaviour of human parathyroid-hormone-related protein(1–34) and its close relationship with the N-terminal fragments of human parathyroid hormone in solution. Biol Chem 1997;378:1501–1508. [PubMed: 9461349]

177. Gronenborn AM, Bovermann G, Clore GM. A 1H-NMR study of the solution conformation of secretin. Resonance assignment and secondary structure. FEBS Lett 1987;215:88–94. [PubMed: 2883029]

178. De Silva RS, Kovacikova G, Lin W, Taylor RK, Skorupski K, Kull FJ. Crystal structure of the virulence gene activator AphA from Vibrio cholerae reveals it is a novel member of the winged helix transcription factor superfamily. J Biol Chem 2005;280:13779–13783. [PubMed: 15647287]

179. Lewis M, Chang G, Horton NC, Kercher MA, Pace HC, Schumacher MA, Brennan RG, Lu P. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. Science 1996;271:1247–1254. [PubMed: 8638105]

180. Hedrick JA, Zlotnik A. Lymphotactin: a new class of chemokine. Methods Enzymol 1997;287:206–215. [PubMed: 9330324]

181. Marcaurelle LA, Mizoue LS, Wilken J, Oldham L, Kent SB, Handel TM, Bertozzi CR. Chemical synthesis of lymphotactin: a glycosylated chemokine with a C-terminal mucin-like domain. Chemistry 2001;7:1129–1132. [PubMed: 11303872]

182. Kuloglu ES, McCaslin DR, Markley JL, Volkman BF. Structural rearrangement of human lymphotactin, a C chemokine, under physiological solution conditions. J Biol Chem 2002;277:17863–17870. [PubMed: 11889129]

183. Smyth E, Syme CD, Blanch EW, Hecht L, Vasak M, Barron LD. Solution structure of native proteins with irregular folds from Raman optical activity. Biopolymers 2001;58:138–151. [PubMed: 11093113]

184. Ennahar S, Sashihara T, Sonomoto K, Ishizaki A. Class IIa bacteriocins: biosynthesis, structure and activity. FEMS Microbiol Rev 2000;24:85–106. [PubMed: 10640600]

185. Eijsink VG, Axelsson L, Diep DB, Havarstein LS, Holo H, Nes IF. Production of class II bacteriocins by lactic acid bacteria; an example of biological warfare and communication. Antonie Van Leeuwenhoek 2002;81:639–654. [PubMed: 12448760]

186. Kaur K, Andrew LC, Wishart DS, Vederas JC. Dynamic relationships among type IIa bacteriocins: temperature effects on antimicrobial activity and on structure of the C-terminal amphipathic alpha helix as a receptor-binding region. Biochemistry 2004;43:9009–9020. [PubMed: 15248758]

187. Prates MV, Sforca ML, Regis WC, Leite JR, Silva LP, Pertinhez TA, Araujo AL, Azevedo RB, Spisni A, Bloch C Jr. The NMR-derived solution structure of a new cationic antimicrobial peptide from the skin secretion of the anuran Hyla punctata. J Biol Chem 2004;279:13018–13026. [PubMed: 14715660]

188. Miyata A, Arimura A, Dahl RR, Minamino N, Uehara A, Jiang L, Culler MD, Coy DH. Isolation of a novel 38 residue-hypothalamic polypeptide which stimulates adenylate cyclase in pituitary cells. Biochem Biophys Res Commun 1989;164:567–574. [PubMed: 2803320]

189. Arimura A. Pituitary adenylate cyclase activating polypeptide (PACAP): discovery and current status of research. Regul Pept 1992;37:287–303. [PubMed: 1313597]

190. Wray V, Kakoschke C, Nokihara K, Naruse S. Solution structure of pituitary adenylate cyclase activating polypeptide by nuclear magnetic resonance spectroscopy. Biochemistry 1993;32:5832–5841. [PubMed: 8504103]

191. Cai X, Dass C. Structural characterization of methionine and leucine enkephalins by hydrogen/deuterium exchange and electrospray ionization tandem mass spectrometry. Rapid Commun Mass Spectrom 2005;19:1–8. [PubMed: 15568184]

192. Takai Y, Sasaki T, Matozaki T. Small GTP-binding proteins. Physiol Rev 2001;81:153–208. [PubMed: 11152757]

193. Symons M, Settleman J. Rho family GTPases: more than simple switches. Trends Cell Biol 2000;10:415–419. [PubMed: 10998597]

194. Scheffzek K, Ahmadian MR. GTPase activating proteins: structural and functional insights 18 years after discovery. Cell Mol Life Sci 2005;62:3014–3038. [PubMed: 16314935]

195. Bernards A. GAPs galore! A survey of putative Ras superfamily GTPase activating proteins in man and Drosophila. Biochim Biophys Acta 2003;1603:47–82. [PubMed: 12618308]

196. Bernards A, Settleman J. GAP control: regulating the regulators of small GTPases. Trends Cell Biol 2004;14:377–385. [PubMed: 15246431]

197. Grosschedl R, Giese K, Pagel J. HMG domain proteins: architectural elements in the assembly of nucleoprotein structures. Trends Genet 1994;10:94–100. [PubMed: 8178371]

198. Reeves R, Beckerbauer L. HMGI/Y proteins: flexible regulators of transcription and chromatin structure. Biochim Biophys Acta 2001;1519:13–29. [PubMed: 11406267]

199. Lehn DA, Elton TS, Johnson KR, Reeves R. A conformational study of the sequence specific binding of HMG-I (Y) with the bovine interleukin-2 cDNA. Biochem Int 1988;16:963–971. [PubMed: 3262346]

200. Evans JN, Nissen MS, R R. Assignment of the 1H NMR spectrum of a consensus DNA-binding peptide from the HMG-I protein. Bull Mag Reson 1992;14:171–174.

201. Evans JN, Zajicek J, Nissen MS, Munske G, Smith V, Reeves R. 1H and 13C NMR assignments and molecular modelling of a minor groove DNA-binding peptide from the HMG-I protein. Int J Pept Protein Res 1995;45:554–560. [PubMed: 7558586]

202. Conti B, Tabarean I, Andrei C, Bartfai T. Cytokines and fever. Front Biosci 2004;9:1433–1449. [PubMed: 14977558]

203. Janecka A, Kruszynski R. Conformationally restricted peptides as tools in opioid receptor studies. Curr Med Chem 2005;12:471–481. [PubMed: 15720255]

204. Chaturvedi K, Christoffers KH, Singh K, Howells RD. Structure and regulation of opioid receptors. Biopolymers 2000;55:334–346. [PubMed: 11169924]

205. Lichtarge O, Jardetzky O, Li CH. Secondary structure determination of human beta-endorphin by 1H NMR spectroscopy. Biochemistry 1987;26:5916–5925. [PubMed: 2960378]

206. Tsunemi M, Kato H, Nishiuchi Y, Kumagaye S, Sakakibara S. Synthesis and structure-activity relationships of elafin, an elastase-specific inhibitor. Biochem Biophys Res Commun 1992;185:967–973. [PubMed: 1627147]

207. Francart C, Dauchez M, Alix AJ, Lippens G. Solution structure of R-elafin, a specific inhibitor of elastase. J Mol Biol 1997;268:666–677. [PubMed: 9171290]

208. Rusnak F, Mertz P. Calcineurin: form and function. Physiol Rev 2000;80:1483–1521. [PubMed: 11015619]

209. Kissinger CR, Parge HE, Knighton DR, Lewis CT, Pelletier LA, Tempczyk A, Kalish VJ, Tucker KD, Showalter RE, Moomaw EW, et al. Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex. Nature 1995;378:641–644. [PubMed: 8524402]

210. Delcour AH. Structure and function of pore-forming beta-barrels from bacteria. J Mol Microbiol Biotechnol 2002;4:1–10. [PubMed: 11763966]

211. Sukumaran S, Hauser K, Maier E, Benz R, Mantele W. Structure-function correlation of outer membrane protein porin from Paracoccus denitrificans. Biopolymers 2006;82:344–348. [PubMed: 16345000]

212. Sukumaran S, Hauser K, Maier E, Benz R, Mantele W. Tracking the unfolding and refolding pathways of outer membrane protein porin from Paracoccus denitrificans. Biochemistry 2006;45:3972–3980. [PubMed: 16548524]

213. Violot S, Aghajari N, Czjzek M, Feller G, Sonan GK, Gouet P, Gerday C, Haser R, Receveur-Brechot V. Structure of a full length psychrophilic cellulase from Pseudoalteromonas haloplanktis revealed by X-ray diffraction and small angle X-ray scattering. J Mol Biol 2005;348:1211–1224. [PubMed: 15854656]

214. von Ossowski I, Eaton JT, Czjzek M, Perkins SJ, Frandsen TP, Schulein M, Panine P, Henrissat B, Receveur-Brechot V. Protein disorder: conformational distribution of the flexible linker in a chimeric double cellulase. Biophys J 2005;88:2823–2832. [PubMed: 15653742]

215. Bordelon T, Montegudo SK, Pakhomova S, Oldham ML, Newcomer ME. A disorder to order transition accompanies catalysis in retinaldehyde dehydrogenase type II. J Biol Chem 2004;279:43085–43091. [PubMed: 15299009]

216. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 2004;32:1037–1049. [PubMed: 14960716]

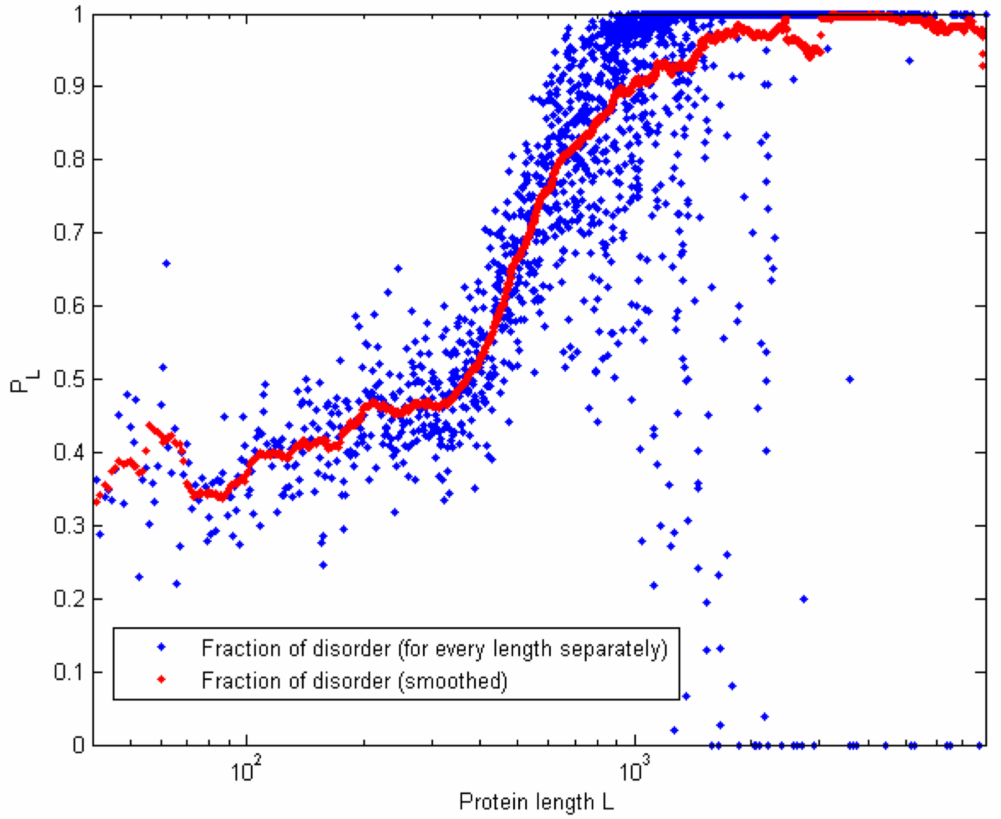

Figure 1.

Fraction of putative disorder as a function of sequence length. The smoothed curve uses averaging window of size equal to 20% of the sequence length.

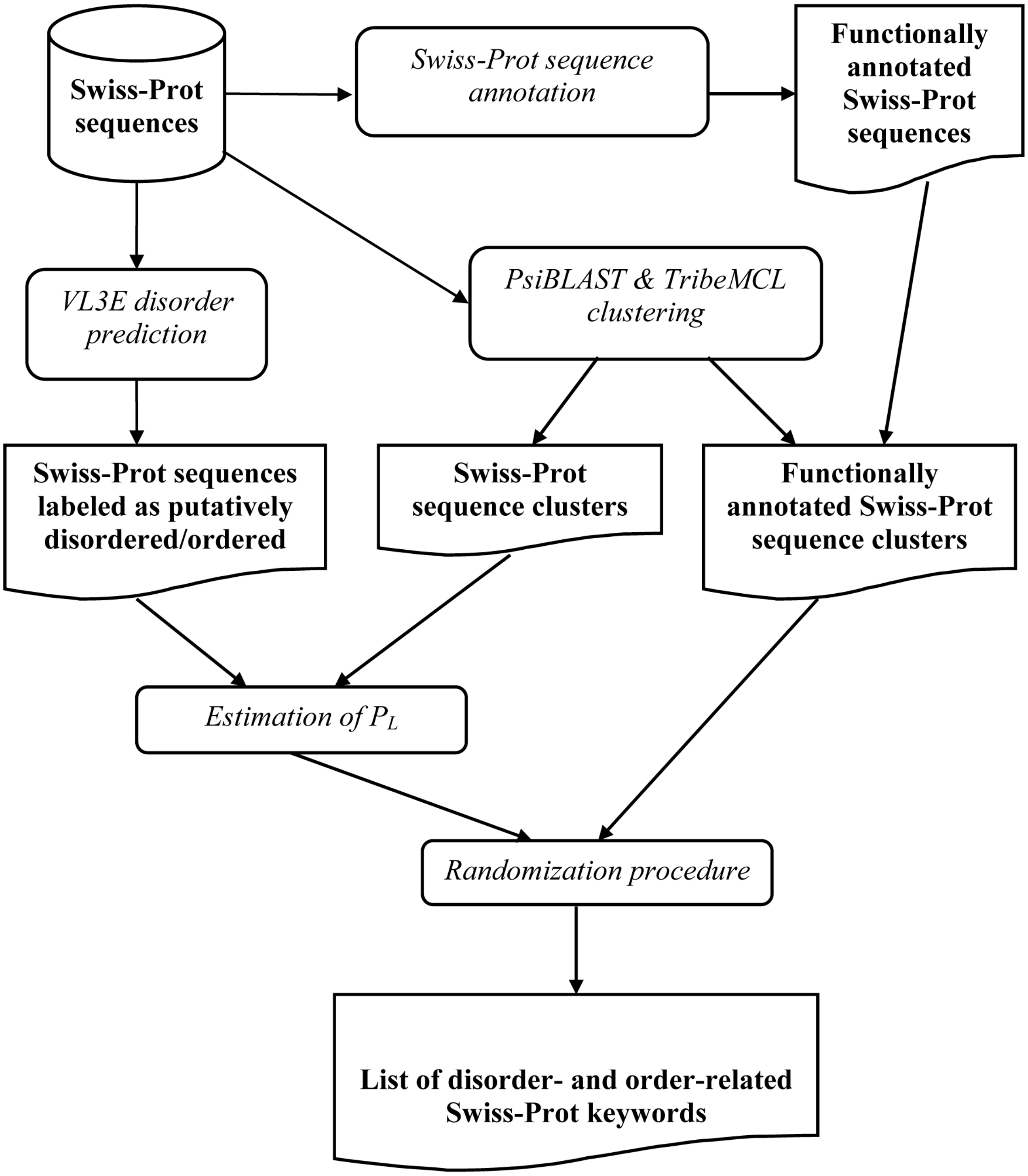
Swiss-Prot sequences
Swiss-Prot sequence annotation
Functionally annotated Swiss-Prot sequences
VL3E disorder prediction
PsiBLAST & TribeMCL clustering
Swiss-Prot sequences labeled as putatively disordered/ordered
Swiss-Prot sequence clusters
Functionally annotated Swiss-Prot sequence clusters
Estimation of P_L
Randomization procedure
List of disorder- and order-related Swiss-Prot keywords

Figure 2.

Schematic representation of the algorithm for extracting disorder- and order-related keywords.

Table 1

Summary of association between the prediction of long disordered regions and keywords for each of the 11 functional categories. For each category, the table lists the total number of keywords associated with it (out of the 710 Swiss-Prot functional keywords), as well as number of keywords associated with predicted order and disorder.

Functional category	# Keywords	# Keywords (p-value < 0.05)	# Keywords (p-value > 0.95)
Biological process	301	174	73
Cellular component	77	23	33
Coding sequence diversity	9	0	6
Developmental stage	4	0	3
Disease	17	0	11
Domain	34	9	21
Ligand	72	41	17
Molecular function	143	37	51
PTM	37	11	18
Technical term	12	7	2
Tissue	4	0	3
TOTAL	710	302	238

Table 2

Top 20 of processes that have strongest correlation with predicted disorder

Keywords	Number of proteins	Number of families	Average sequence length	Z-score	P-value
Differentiation	1406	422	439.25	18.81	1
Transcription	11223	1653	442.64	14.62	1
Transcription regulation	9758	1554	413.31	14.33	1
Spermatogenesis	332	189	280.49	13.9	1
DNA condensation	317	130	300.06	13.34	1
Cell cycle	4278	612	494.17	12.17	1
mRNA processing	1575	249	515.55	10.92	1
mRNA splicing	716	180	459.06	10.13	1
Mitosis	718	215	620.43	9.42	1
Apoptosis	810	211	465.48	9.35	1
Protein transport	3081	579	421.73	8.77	1
Meiosis	284	170	639.16	8.7	1
Cell division	3466	385	451.63	8.51	1
Ubl conjugation pathway	1254	244	525.99	8.13	1
Wnt signaling pathway	417	41	476.84	6.58	1
Neurogenesis	322	74	667.4	6.56	1
Chromosome partition	556	67	495.39	6.39	1
Ribosome biogenesis	319	71	391.79	5.9	1
Chondrogenesis	64	6	332.85	5.58	1
Growth regulation	155	45	354.9	5.14	1

Table 3

Top 20 of functions that have strongest correlation with predicted disorder

Keywords	Number of proteins	Number of families	Average sequence length	Z-score	P-value
Ribonucleoprotein	12236	412	150.55	22.13	1
Ribosomal protein	11692	330	140.58	20.63	1
Developmental protein	3260	721	477.93	19.28	1
Hormone	1187	161	141.13	15.58	1
Growth factor	785	84	255.7	11.16	1
Cytokine	899	110	213.28	10.21	1
Neuropeptide	268	209	95.08	9.65	1
Activator	3086	573	428.47	9.04	1
GAP protein	47	2	232.96	7.42	1
Antigen	1113	455	437.48	6.99	1
Repressor	2309	449	374.46	6.92	1
Chromatin regulator	334	100	801.24	6.7	1
Pyrogen	37	2	262.59	6.44	1
Vasoactive	125	39	160.39	5.56	1
Amphibian defense peptide	123	148	50.64	5.44	1
GTPase activation	311	70	831.03	5.36	1
Endorphin	42	4	226.68	5.35	1
Opioid peptide	24	4	216.96	5.14	1
Protein phosphatase inhibitor	47	8	366.51	5.07	1
Cyclin	182	25	430.58	4.88	1

Table 4

Top 20 of processes that have strongest correlation with predicted order

Keywords	Number of proteins	Number of families	Average sequence length	Z-score	P-value
GMP biosynthesis	225	3	473.11	−17.62	0
Amino-acid biosynthesis	7098	212	361.5	−17.11	0
Transport	19888	2199	378.13	−14.87	0
Electron transport	4633	346	272	−13.72	0
Lipid A biosynthesis	533	13	291.25	−13.22	0
Aromatic hydrocarbons catabolism	320	105	300.36	−12.37	0
Glycolysis	2255	50	390.64	−12.14	0
Purine biosynthesis	1208	28	445.46	−11.89	0
Pyrimidine biosynthesis	1310	27	383.27	−11.7	0
Carbohydrate metabolism	1797	180	404.2	−11.68	0
Branched-chain amino acid biosynthesis	963	26	404.12	−11.11	0
Lipopolysaccharide biosynthesis	481	102	335.93	−11.09	0
Sugar transport	903	109	387.37	−11	0
Antibiotic resistance	1203	177	354.24	−10.66	0
Lipid synthesis	2184	122	328.02	−10.17	0
Tricarboxylic acid cycle	1013	54	460.88	−10.04	0
Arginine biosynthesis	1353	17	414.06	−9.53	0
Ion transport	5275	459	464.46	−9.37	0
Rhamnose metabolism	85	4	372.84	−9.12	0
Peptidoglycan synthesis	1839	38	372.73	−9.03	0

Table 5

Top 20 of functions that have strongest correlation with predicted order

Keywords	**Number of proteins**	**Number of families**	**Average sequence length**	**Z-score**	**P-value**
Oxidoreductase	14995	992	376.63	−29.54	0
Transferase	26525	1606	445.17	−24.25	0
Lyase	7262	347	377.92	−22.64	0
Hydrolase	20464	1995	430.68	−21.75	0
Isomerase	4487	220	383.98	−14.18	0
Glycosidase	1826	244	444.73	−13.98	0
Glycosyltransferase	2950	261	437.53	−12.51	0
Acyltransferase	2239	179	402.83	−10.85	0
Methyltransferase	3524	224	349.6	−10.53	0
Kinase	7017	322	448.29	−10.22	0
Ligase	8010	230	529.41	−10.06	0
Decarboxylase	1293	63	345.26	−9.66	0
Monooxygenase	1668	73	444.87	−9.26	0
Metalloprotease	1100	109	553.73	−7.89	0
Aminopeptidase	452	39	509.17	−7.55	0
Dioxygenase	360	66	433.2	−7.32	0
Aminoacyl-tRNA synthetase	3402	37	571.83	−7.15	0
Protease	4423	380	549.7	−7.1	0
Aminotransferase	955	28	420.27	−6.02	0

Table 6

All (11) Swiss-Prot keywords associated with at least one of the 98 confirmed long disordered protein regions[8,9]. For each function, number of the associated regions (out of 98) and z-ratio are listed.

Function	**Number of regions associate with function in literatures (Dunker, et al 2002)**	**Z-ratio in SwissProt database**
Protein-DNA interaction	19	18.2
Phosphorylation	16	27.1
Structural mortar	>10	3.0
Ubiquitination	7	8.7
Protein-rRNA interaction	5	11.5
Fatty acylation	4	6.0
Protein-genomic RNA binding	3	17.7
Glycosylation	3	6.8
Methylation	1	2.9
ADP-ribosylation	1	2.0
Protein-tRNA interaction	1	1.8