Published in final edited form as:
J Proteome Res. 2007 May ; 6(5): 1899–1916. doi:10.1021/pr060393m.

Functional Anthology of Intrinsic Disorder. II. Cellular Components, Domains, Technical Terms, Developmental Processes and Coding Sequence Diversities Correlated with Long Disordered Regions

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

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

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

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

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

Abstract

Biologically active proteins without stable ordered structure (i.e., intrinsically disordered proteins) are attracting increased attention. Functional repertoires of ordered and disordered proteins are very different, and the ability to differentiate whether a given function is associated with intrinsic disorder or with a well-folded protein is crucial for modern protein science. However, there is a large gap between the number of proteins experimentally confirmed to be disordered and their actual number in nature. As a result, studies of functional properties of confirmed disordered proteins, while helpful in revealing the functional diversity of protein disorder, provide only a limited view. To overcome this problem, a bioinformatics approach for comprehensive study of functional roles of protein disorder was proposed in the first paper of this series (Xie H., Vucetic S., Iakoucheva L.M., Oldfield C.J., Dunker A.K., Obradovic Z., Uversky V.N. (2006) Functional anthology of intrinsic disorder. I. Biological processes and functions of proteins with long disordered regions. J. Proteome Res.). Applying this novel approach to Swiss-Prot sequences and functional keywords, we found over 238 and 302 keywords to be strongly positively or negatively correlated, respectively, with long intrinsically disordered regions. This paper describes ~90 Swiss-Prot keywords attributed to the cellular components, domains, technical terms, developmental processes and coding sequence diversities possessing strong positive and negative correlation with long disordered regions.

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

Although the number of papers describing the structural features and functional peculiarities of intrinsically disordered proteins growths exponentially[1, 2], there is a significant gap between the number of proteins that have been experimentally characterized as intrinsically disordered and their actual number in nature. For example, bioinfomatics analysis has revealed that about 25 to 30% of eukaryotic proteins are mostly disordered,[3] that more than half of eukaryotic proteins have long regions of disorder,[3, 4] and that more than 70% of signaling proteins and the vast majority of cancer-associated proteins have long disordered regions.[5] The number of Swiss-Prot proteins that are predicted to contain long disordered regions falls between 30% and 50%,[5, 6] with as much as ~10–20% of Swiss-Prot proteins being predicted to be wholly disordered by two binary identifiers of intrinsic disorder, namely by the charge-hydropathy plot (~10%) and the cumulative distribution function (~20%).[3] Thus, the number of Swiss-Prot proteins that are potentially disordered varies from ~20,000 to >100,000. On the other hand, as of July 2006, there were only 458 proteins containing 1096 disordered regions in a curated Database of Disordered Protein (DisProt),[7] which provides structure and function information about proteins that lack a fixed 3D-structure under putatively native conditions, either in their entireties or in part.

Intrinsically disordered proteins, even with their lack of well-defined 3D-structures under physiological conditions, still carry out numerous important biological functions.[1, 2, 5, 8–18] Intrinsically disordered regions play a number of crucial roles in regulation, signaling and controlling processes where interactions with multiple partners and high-specificity/low-affinity binding are involved. Many posttranslational modifications (including acetylation, hydroxylation, ubiquitination, methylation, phosphorylation, etc.) and proteolytic attack frequently occur within the regions of intrinsic disorder.[10] Recently, more than 150 proteins have been identified as containing functional disordered regions, or being completely disordered, yet performing vital cellular roles.[11, 12] Twenty-eight separate functions were assigned to these disordered regions, including molecular recognition via binding to other proteins, or to nucleic acids,[11, 15] which can be grouped into at least five broad classes based on the mode of disordered protein/region action, including entropic chains, effectors, scavengers, assemblers and display sites.[15] The increasing recognition of the importance of these proteins called for the adjustment of the "lock-and-key" structure-function paradigm,[8] and finally resulted in the elaboration of a novel sequence-to-structure-to-function paradigm, The Protein Trinity[19] or The Protein Quartet Hypotheses,[1] which includes the novel functions of disordered proteins, and suggested that native and functional proteins can be in one of three (four) states, the solid-like ordered state, the liquid-like collapsed-disordered state, the gas-like extended-disordered state, or the extended-disordered pre-molten globule state.[1, 10, 19] This new paradigm suggests that function arises from any one of the three (four) states or from transitions between them.[1, 10, 19]

Interestingly, it was the bioinformatics who played the key role in transforming a set of counter examples of "strange" biologically active proteins without rigid 3-D structure into a new subfield of protein science dealing with intrinsically disordered proteins. Similar to this past success, the future development of our understanding of the functional diversity of protein disorder may rely on such bioinformatics tools as data mining to classify the functional repertoires of confirmed disordered proteins. While such obvious studies are helpful, they are bound to provide only a limited view. Thus, novel bioinformatics tools are needed. In the first paper of this series, a bioinformatics approach was proposed to analyze functional roles of protein disorder in a systematic way.[20] This novel tool uses the long disordered regions PONDR VL3E predictor, which has estimated per residue accuracy of above 86%,[21] to find Swiss-Prot[22] proteins that contain long predicted regions of intrinsic disorder. As about 200,000 Swiss-Prot proteins have been annotated with one or more functional keywords, the disorder-(or order-) correlated functions were determined as those that contain significantly larger (or smaller) fraction of proteins predicted to have long disordered regions (> 40 amino acid residues) than would be expected in a random selection of SwissProt proteins of comparable length. We also ensured that the unfavorable effects of sequence redundancy and sequence length were eliminated.[20] This study performed a comprehensive analysis of disorder-related functions in the entire Swiss-Prot and resulted in a list of 238 and 302 Swiss-Prot functional keywords that were strongly positively or negatively correlated with long intrinsically disordered regions, respectively.

Given this list, an extensive literature survey to find experimental evidence supporting the findings was also performed. Illustrative literature examples related to 80 Swiss-Prot keywords associated with disorder-and order-driven biological processes and protein functions were described in the first paper.[20] The second paper of the series is devoted to the presentation of 87 Swiss-Prot keywords attributed to the cellular components, domains, technical terms, developmental processes and coding sequence diversities possessing strong positive and negative correlation with long disordered regions. As with the first 80 keywords, this manual curation of the next 87 finds supportive evidence for the indicated disorder-function relationship.

Materials and methods

The protocol for dataset assembly was described in the first paper of this series together with the description of the bioinformatics approach for determining order- and disorder-related SwissProt keywords.[20] In brief, we have analyzed 196,326 Swiss-Prot[22] proteins with sequences longer than 40 amino acid residues. The known redundancy of Swiss-Prot[23] was reduced applying the Markov Cluster Algorithm[24] to group the Swiss-Prot proteins into 27,217 families according to sequence similarity. Each of the analyzed proteins is annotated with one or several functional or structural keywords. There are 875 keywords in Swiss-Prot, 710 of which are associated with at least 20 proteins.

Long disordered regions in Swiss-Prot proteins were predicted using the PONDR VL3E predictor[21], which is an ensemble of neural network classifiers that achieves ~87% per-residue cross-validation accuracy on balanced data set with equal numbers of ordered and disordered residues. Each of the 196,326 Swiss-Prot proteins was labeled as putatively disordered if it contained at least one region with more than 40 consecutive amino acids predicted by VL3E to be intrinsically disordered; proteins predicted not to contain such long disordered regions were labeled as putatively ordered.

The probability, P_L, that VL3E predicts a disordered region longer than 40 consecutive amino acids in a SwissProt protein sequence of length L was estimated as the fraction of putatively disordered SwissProt proteins with lengths between 0.9L and 1.1L. TribeMCL clustering was used to reduce effects of sequence redundancy in estimation of P_L, as described previously.[20] Swiss-Prot keywords associated with disorder- (or order-) correlated functions were determined as those that contain a significantly larger (or smaller) fraction of putatively disordered proteins than what would be expected by a random selection of SwissProt sequences with the same length distribution.[20]

Results and Discussion

The bioinformatics results indicate correlations between keywords and whether or not whole proteins contain regions of order or disorder. From these results it is unclear whether the various functions are associated with the regions predicted to be disordered or with other parts of the proteins. We therefore carried out manual literature curation to test whether functions are directly associated with disordered regions. The results of this curation are presented below. The data below are organized in the following way: each discussed keyword is placed at the beginning of the corresponding paragraph and *Italized*. If description provided involves other keywords discussed in this and the accompanying papers these keywords are presented using the *Italic* font.

Cellular components associated with intrinsically disordered proteins

Table 1 lists the cellular components associated with intrinsic disorder. The role of disorder in functioning of several *nuclear proteins*, including *chromosomal proteins*, *nucleosome core* proteins (histones) and transcription factors, has been already discussed. High mobility group (HMG) domain proteins are *nuclear proteins* that are known as transcription factors.[25] These proteins are considered as central 'hubs' of nuclear function, as they are able to bind to 18 known protein partners and addition to several specific DNA structures.[26] The HMG proteins are soluble in dilute (5%) acids, possess unusually high content of charged amino acid residues and prolines, have multiple phosphorylation sites, and exhibit atypical electrophoretic mobility.[27] Furthermore, the HMGA proteins have little, if any, regular secondary structure and certainly no rigid tertiary structure;[28–30] i.e., these molecules behave as typical natively unfolded proteins.[9]

Spliceosome—Protein p14 is a subunit of the essential splicing factor 3b (SF3b) which is present in both the major and minor spliceosomes.[31–33] The p14 molecule is located near the catalytic center of the spliceosome and is responsible for the first catalytic step of the splicing reaction.[33, 34] Using NMR analysis it has been established that the flanking N- (residues 1–20) and C-terminal regions (residues 100–125) of p14 are unstructured.[35] Another example of disordered proteins that are involved in spliceosome assembly is serine/arginine-rich (SR) splicing factors. Besides their importance for both constitutive and alternative splicing,[36]SR proteins play key roles in the spliceosome assembly by facilitating recruitment of components of the spliceosome via protein-protein interactions[37] that are potentially mediated by the disordered SR domains of these splicing factors.[38]

Cytoskeleton—An internal network of proteinaceous structures, *microtubules* and filaments, determines the structure and shape of the cell and contributes to the *cytoskeleton*. Protein intrinsic disorder plays multiple crucial roles in the assembly and function of the *cytoskeleton*. We have previously shown that the high disorder content of cytoskeletal proteins is comparable to that of regulatory and cell signaling proteins.[5] Three illustrative examples of disordered cytoskeletal proteins are described below: microtubule-associated protein tau, neurofilament proteins and stathmin.

There are six major isoforms of the microtubule-associated protein tau, ranging in size from 352 to 441 amino acid residues. These isoforms are produced in the human central nervous system as a result of *alternative splicing*.[39–41] The amino acid sequence of tau is dominated by hydrophilic and charged residues. The C-terminal repeat region is flanked upstream by a basic proline-rich region (about 25% proline) and downstream by another basic stretch that also contains several prolines. The C-terminal half of tau (repeats and their flanking regions) constitutes the microtubule binding domain.[42–44] Among numerous functions associated with tau are the stabilization of axonal microtubules, interaction with the actin cytoskeleton and the plasma membrane, the anchoring of enzymes such as protein kinases and phosphatases, the regulation of intracellular vesicle transport, and the participation in signal transduction and neurite outgrowth.[45] Pathological association of tau protein into paired helical filaments (PHFs) is associated with Alzheimer’s disease development.[46] Despite the crucial role in numerous biological processes and its involvement in the Alzheimer’s disease pathogenesis, the structural information about the tau protein is limited. So far, low-resolution techniques (such as circular dichroism, small angle X-ray scattering, hydrodynamic measurements) have yielded data, and furthermore this protein has resisted all crystallization efforts. Finally it is too large for a structural analysis by NMR.[45] Overall, these low resolution structural data indicate that tau is a natively unfolded protein with little α-helical or β-sheet structure.[47–49] In fact, to distinguish the extremely flexible tau protein from 'normal' globular protein with rigid 3-D structure a special term, "natively denatured protein", was introduced in 1994.[48]

Neurofilaments (NF) are composed of three proteins: light (NFL), medium (NFM), and heavy (NFH) that combine to form an intermediate-sized filament that functions to maintain the inner bore of neurons. NFs are formed by the trimerization of the N-terminal regions of these three proteins. Both the NFM and NFH proteins possess C-terminal tails that are disordered extensions from the assembled neurofilament core. These tails are thought to form highly flexible entropic bristle domains that function to keep the NFs separated, thus keeping the neuron bore open.[50] The human NFM is a 915-residue protein that contains a 312-residue intermediate filament oligomerization domain in the N-terminal half and a 504-residue highly charged C-terminal tail. While the NFH tail requires phosphorylation to acquire a charge and can adopt various phosphorylation states, the NFM tail carries a high intrinsic charge.[51]

Stathmin is a key regulator of microtubule dynamics. The soluble cytoplasmic protein destabilizes microtubules by regulation of microtubule dynamics and stimulating microtubule growth to shortening, thereby playing a central role for cell proliferation, cell migration, and mitotic spindle formation.[52] Stathmin is expressed in high amounts in a wide variety of human malignancies, and it is also recently implicated in anxiety states of mental disorders.[53] It has been shown that stathmin lacks a stable 3-D structure in isolation,[54] and its N-terminal moiety adopts little regular secondary structure, whereas the C-terminal domain populates an ensemble of transient helical conformations. Upon binding of stathmin to two head-to-tail aligned α/β-tubulin heterodimers, the N terminus folds into a β-hairpin, and the C-terminal helical domain becomes stabilized.[55] Recently, the important role of phosphorylation in regulating stathmin-tubulin interactions have been demonstrated.[56]

Centromere—This segment, which is the most condensed and constricted region of a *chromosome*, assembles the proteinaceous kinetochore, maintains sister chromatid cohesion, regulates chromosome attachment to the spindle during *mitosis*, and directs chromosome movement during cell division.[57] Kinetochores are multiprotein complexes that assemble on centromeric DNA and mediate attachment of chromosomes to microtubules. In the budding yeast, *Saccharomyces cerevisiae*, kinetochores contain 60 or more different subunits organized into at least 14 multiprotein complexes.[58] The four-protein Ndc80 complex is an essential kinetochore component, which is conserved from yeast to humans and contains Ndc80p, Nuf2p, Spc24p, and Spc25p.[59] Nuf2p and Spc24p are both coiled-coil proteins.[60] Importantly, although Spc24p has a smaller mass than Spc25p, the former migrates slower on SDS/PAGE,[61] a property typical of intrinsically disordered proteins.[62] Furthermore, the production of the *S. cerevisiae* Ndc80 complex in insect cells results in five products: four corresponded to the molecular weights expected for the full-length Ndc80p, Nuf2p, Spc24p, or Spc25p, whereas the fifth polypeptide, with an apparent molecular mass of 70 kDa, contained an N-terminally truncated version of Ndc80p.[61] Full length and truncated Ndc80p are populated almost equally, suggesting low conformational stability of the N-terminal domain of this protein. Comparison of Ndc80p sequences from 29 organisms reveals that the first 100 residues of the *S. cerevisiae* protein are poorly conserved among fungi and largely absent from higher eukaryotes.[61]

Flagellum—This the locomotion organelle is a long, whip-like extension, which enables certain cells or unicellular organisms to swim. The major component of the bacterial flagellum is a thin filament of 12–25 nm in diameter, which is known as the flagellar filament and is made of protein subunit called "flagellin". It has been established that intrinsic disorder play crucial role in the assembly of the bacterial flagellum.[63] For example, 65 N-terminal and 45 C-terminal residues are disordered in the monomeric forms of flagellin from *E. coli* and *Salmonella*, respectivelly[64, 65] During the flagellum formation, these disordered terminal regions fold to form a concentric double-tubular structure in the filament core, which is mostly made of α-helices aligned parallel to the filament axis.[63, 66] The other flagellar axial proteins including FliE (the putative FliF ring-rod junction), FlgB, FlgC, FlgF, FlgG (four rod proteins), FlgE (hook), FlgK (HAP1), FlgL (HAP3), and FliD (HAP2), which are produced in much smaller copy numbers in wild-type bacterial cells, also have disordered termini in their monomeric forms in solution.[63] Based on these observations it has been concluded that the terminal disorder, being a common motif of the flagellar axial proteins, plays an important role in their assembly process in the same way as flagellin.[63]

Golgi stack—The Golgi apparatus (also known as a dictyosome, Golgi complex, or Golgi body) is an organelle representing a network of stacked membranous vesicles found in most of eukaryotic cells. The central portion of the Golgi apparatus is characterized by a set of thin, flattened membrane-bounded compartments, called cisternae, or *Golgi stack*. The Golgi apparatus is part of the endomembrane system, which functions as a central delivery system for the cell, kind of the cellular post office, processing proteins targeted to the plasma membrane, lysosomes and endosomes and to the secretion. Furthermore, the stability and functional diversity of many newly synthesized proteins in the secretory pathway depend on accurate glycosylation performed by the Golgi apparatus.[67] The Golgi stack receives proteins from the endoplasmic reticulum and transfers them through the different compartments of the stack, which initiates a series of complex sorting events directing various cargo molecules to different subcellular organelles, and the apical and basolateral surfaces of polarized cells. One of the most characteristic features of the Golgi apparatus is the presence of various carbohydrate-processing enzymes that are specifically localized in the Golgi stack.[68] Transport within the Golgi stack is facilitated by the coat protein I (COPI) carrier vesicles that form in response to activation of the small GTPase ARF1.[68] Furthermore, several other small GTPases (members of the GNBP family), including ARL1,[69] Sar1 and Rab1,[68] are associated with the Golgi apparatus function. The role of intrinsic disorder in regulation of the different GNBPs was already discussed in section *GTPase-activating proteins*.

Another example of intrinsically disordered protein that is implicated in several biological functions including *Golgi* membrane tubule fission is a transcriptional corepressor CtBP.[70] CtBP is a moonlighting protein that fulfills different functions in the cell depending on cellular localization.[71] In isolated Golgi apparatus the CtBP3 isoform (previously known as BARS, Brefeldin A-ADP Ribosylated Substrate, and recently renamed as short-CtBP1 or CtBP1-S) has been shown to be a key component of the machinery controlling *Golgi* tubule fission.[72] In recent studies on intact cells, the CtBP fission inducing activity was shown to participate in the fragmentation of the *Golgi* complex during *mitosis*[73] as well as in intracellular membrane traffic.[74] Recently, a combined approach based on bioinformatics, NMR, CD spectroscopy, and small-angle X-ray scattering that was applied to analyze CtBP structure demonstarted that ~90 C-terminal residues of this protein are intrinsically unstructured in the full-length CtBP and in constructs lacking the substrate- and/or the nucleotide-binding domains.[75]

Mitochondrion—The overwhelming information on the mitochondial proteins is accumulated in the MitoRes database.[76] A few illustrative examples of intrinsically disordered *mitochondrial* proteins are described below. Don Juan is a nuclear-encoded, germ-cell specifically expressed protein, which might be involved in the final steps of mitochondrial differentiation within the flagellum.[77] Don Juan is a medium-sized (248 amino acid residues long) basic protein with high lysine content (the protein contains 33% of the amino acid lysine) and multiple hexapeptide-motif DPCKKK repeats.[78] These amino acid biases suggest that

Don Juan is very likely to be a typical intrinsically disordered protein. Mitochondrial transcription factor A (TFAM), a member of an HMG protein family, is essential for maintenance of mitochondrial DNA, potentially playing a histone-like architectural role for maintenance of mitochondrial DNA.[79] As both these functions, transcription regulation and histone, are associated with intrinsic disorder, TFAM potentially behaves as an intrinsically disordered protein as well.

Proteasome—This multicatalytic enzyme complex, which is present in the cytoplasm and nucleus of all eukaryotic cells, has the primary function of degrading proteins. The 20S proteasome (also known as the core particle, CP) is a large, cylinder-shaped protease with the molecular weight of about 700 000 Da, which is formed by 28 protein subunits arranged in 4 stacked rings of 7 subunits each.[80] As with many other multiprotein complexes, CP contains numerous flexible and disordered regions crucial for the assembly of the particle and for its function.[81] Over 80% of all cellular proteins are recycled through the *proteasome*.[82] The list of common substrates of the proteasome includes cell-cycle regulators, signaling molecules, tumor suppressors, transcription factors, and antiapoptotic proteins. Note the vast majority of these proteasome-processed proteins are associated with signaling and regulation. These processed proteins are either completely disordered or contain long disordered regions.[5] Furthermore, it has been recently demonstrated that in cases when a folded protein substrate is directed to proteolysis by the *proteasome*, an unstructured initiation site within such a substrate is absolutely required for a successful degradation.[83] It has been proposed that the *proteasome* degrades a substrate by first binding to its ubiquitin modification and then initiating unfolding at an unstructured region.[83]

Signal recognition particle *(SRP)*—This ribonucleoprotein complex mediates the cotranslational targeting of nascent secretory and membrane proteins to the endoplasmic reticulum.[84–86] SRP is present in all three kingdoms of life. Metazoan SRPs consist of six proteins (SRP54, SRP19, SRP68, SRP72, SRP9, SRP14) and a 300-nucleotide RNA (SRP RNA).[87–89] Analysis of the solution structure of the 104 residue SRP19 from the hyperthermophilic archaeon *Archaeoglobus fulgidus* revealed that residues 53–62 in loop 3 and residues 95–104 in the C-terminal tail exhibit considerable disorder.[90] The SRP located in chloroplasts is different from the cytoplasmic SRPs by lacking RNA and containing instead a 43-kDa subunit cpSRP43. This novel cpSRP43 subunit is composed of repeats of two types of structural motifs: namely, four ankyrin repeats and three chromodomains (chromosome organization modifier).[91–93] NMR analysis revealed that the three chromodomains, CD1, CD2, and CD3, possess significant amounts of disorder. In fact, in CD1, 32 out 47 residues are disordered, including N-terminal residues 1–9, C-terminal residues 36–47 and the loop 23–31. CD2 (which is 56 amino acid residues long) contains a disordered N-terminal region (residues 1–12) and a flexible loop (27–35). There is also a flexible loop (residues 10–19) in CD3.[94]

Synaptosome—The junction between various signal-transmitting cells, e.g., between two neurons or between a neuron and a muscle or gland is the synapse. The *synaptosome* is a saclike structure that results from synapses following homogenization and fractionation of the nerve tissue. The major function of the synapse is to transmit a nerve impulse from an axon terminal to another neuron, a muscle or a gland cell. Many proteins are associated with the synapse and are involved in its function. An illustratative example is a presynaptic natively unfolded protein α-synuclein discussed above (see section *Neurotoxins*). Another example of intrinsically disordered protein involved in the synaptic activity is neuromodulin (also known as growth-associated protein-43 (GAP-43), B50, F1, or P56), which plays a crucial role in several processes in neuron biology, including growth and regeneration, synaptic plasticity and neurotransmitter release.[95] Far-UV CD analysis revealed that the purified neuromodulin is substantially unfolded but undergoes a conformational change from random coil to α-helix as a result of interaction with acidic phospholipids.[96] The intracellular domain of the nicotinic acetylcholine receptor (nAChR) that mediates signal transduction at the postsynaptic membrane of cholinergic synapses is also disordered as determined by several prediction methods, limited proteolysis and NMR.[97]

Gas vesicle—This is a cylindrical shell made of protein enclosing a gas-filled space. The gas vesicles are used by many aquatic microorganisms to regulate their depth in the water column. The structure of the gas vesicle shell is similar to a virus capsid, as it contains no lipids and is built almost exclusively of repeating units of the 7- to 8-kDa gas vesicle protein A (GvpA).[98] This basic shell is strengthened by adhering the different amounts of the larger, ~21-kDa, gas vesicle protein C (GvpC).[99] GvpA is one of the most hydrophobic proteins known.[100, 101] Gas vesicles can be dissolved in highly protic acids, such as 80% formic acid, but solution NMR shows that GvpA is unfolded under these conditions.[102] Contrarily, the GVPc from *Anabaena flos-aquae* is a hydrophilic protein of 193 residues (21,985 Da) containing five highly conserved 33 amino acid repeats,[103] which may interact with the periodic structure provided by GVPa. It is beleived that GVPc provides the hydrophilic outer surface of the gas vesicle wall and stabilizes it.

Keratins—This family of fibrous structural proteins includes the major protein components of hair, wool, nails, horn, hoofs, and the quills of feathers. Keratins comprise the large subgroup of intermediate filament proteins and are differentially expressed as pairs of type I (~48 kDa, acidic) and type II (~61 kDa, basic) intermediate filament proteins in the epithelia.[104] In the early state of hair formation, assemblies of these keratin proteins form a gel network, leading to the production of hair through dehydration. The major structural feature of the hard keratin fibers is a double-stranded α-helical *coiled coil*.[105, 106] According to the 3-D molecular structure of the keratin α-helical coiled coil proposed by Pauling and Corey,[107] the heterodimer molecules (roughly 50 nm long) are characterized by a central domain composed of a double-stranded α-helical coiled coil interrupted by nonhelical segments. Although hard keratin fibers are highly organized structures, the analysis of water-soluble keratin proteins consisting of mixtures of type I and type II revealed that keratins are substantially disordered at pH 8.0, possessing far-UV CD spectra typical of highly unfolded polypeptide chain.[108]

Telomere—Telomeres are physical ends of linear eukaryotic chromosomes, which are specific *nucleoprotein* complexes with numerous important functions, primarily in the protection, replication, and stabilization of the chromosome ends. In human, the ends of chromosomes are characterized by the arrays of TTAGGG repeats. Six telomere-specific proteins, TRF1, TRF2, TIN2, Rap1, TPP1, and POT1, associate to form a complex known as shelterin, which protects chromosome ends.[109] Three of these proteins, TRF1, TRF2 and POT1, bind directly and specifically to telomeric DNA and each bind several proteins that do not interact with telomeric DNA.[110]

Chloroplast—This chlorophyll-containing organelle is found in the cytoplasm of algal and green plant cells. The chloroplast is one of the forms that a plastid, a pigmented cytoplasmic organelle, may take. Chloroplasts convert the light energy into electrochemical potentials of ATP and NADPH through a process called oxygen-evolving photosynthesis. Photosynthesis actually takes place in thylakoids, the sub-organelles, located inside the chloroplasts and stacked in grana. Similar to *mitochondria*, chloroplasts are surrounded by a double lipid-bilayer membrane with an intermembrane space. Although chloroplasts have their own DNA, the majority of their proteins are encoded by genes contained in the cell nucleus, with the protein products trafficked into the chloroplast. Obviously, the chloroplasts house a number of specific proteins involved in energy metabolism. Furhtermore, chloroplasts contain a specialized type of SRP, the chloroplast SRP (cpSRP), the structural and functional peculiarities of which were discussed above. Importantly, besides the post-translational interaction with light-harvesting chlorophyll *a/b*-binding protein, the cpSRP is involved in the co-translational transport of chloroplast-encoded thylakoid proteins. Thus, cpSRP is able to switch between the co-and post-translational modes of interaction with the corresponding substrate proteins.[111]

Surface film—Corneal and conjunctival epithelia of the eye are crucial for light refraction and protection of vision, as well as for holding in place a tear film on the eye surface. Maintenance of the tear film on the ocular surface, lubrication, and provision of a pathogen barrier on this wet surface are controlled and regulated by a class of large, highly glycosylated, hydrophilic glycoproteins – the mucins.[112] Structural analysis of different mucins revealed that in dilute solutions these proteins behave as random coil-like, linear flexible chains.[113]

Tight junction—Tight junction (TJ), also known as impermeable junction, occluding junction or zonula occludens, is formed by closely associated areas of two cells whose junctional membranes join together. TJ constitutes the barrier both to the passage of ions and molecules through the paracellular pathway and to the movement of proteins and lipids between the apical and the basolateral domains of the plasma membrane. More than 40 different proteins have been discovered to be located at the TJs of epithelia, endothelia and myelinated cells.[114] Several integral membrane proteins, including occludin, claudin(s) and the junctional adhesion molecule-A (JAM-A), were localized to functional TJ structures.[115] Human occludin has four *transmembrane* segments, a 65-amino acid cytosolic N terminus, two extracellular loops of 46 and 48 amino acids separated by a 10-amino acid cytosolic loop, and a C-terminal tail of approximately 255 amino acids.[116] Both the N-and C-terminal domains have a large number of serine and threonine residues, and the functionally active form of the protein localizing to the tight junction appears to be hyperphosphorylated at serine and threonine residues.[117] Circular dichroism analysis revealed that extracellular loops of occludin are likely to be disordered.[118]

Extracellular matrix—Extracellular matrix (ECM) is a network of filamentous *glycoproteins* and *proteoglycans*. ECM is attached to the cell surface and provides cells with anchorage, traction for movement, and positional recognition. The most abundant components of ECM are *collagen*, fibrin, elastin, fibrillins, fibrinonectins, laminins and nidogens. Elastin is a massive cross-linked protein network produced by the stepwise association of tropoelastin molecules, which are oxidized at selected lysines and associate in the extracellular matrix.[119] The association of tropoelastin is determined by domain 26 (D26) (residues 515–556), structure of which has been recently determined using high-resolution solution NMR methods.[120] In this study, essentially full chemical shift assignment for D26 at 278 K was obtained using a combination of homonuclear, ^{15}N-separated and triple resonance experiments. A thorough analysis of secondary chemical shifts, NOE and ^{15}N relaxation data revealed that this domain is essentially unstructured in solution.[120]

Cellular components associated with ordered proteins

Table 2 lists the top cellular components associated with ordered proteins. Membrane (*membrane, inner membrane, and outer membrane*) proteins are known to play a number of the critical roles in a myriad of biological and physiological functions. The *transmembrane* fragments, domains and chains of proteins associated with all types membranes can be grouped into two major structural classes, α-helical and β-barrel, both of which are highly organized.[121] In agreement with this statement, the membrane proteins are significantly under-represented in disorder predictions across entire proteomes.[122] This does not rule out the possibility that membrane proteins contain disordered loops and tails, however.

MHC I—The major histocompatibility complex (MHC) is a group of genes that code for the cell-surface histocompatibility antigens and are the major determinants of tissue type and transplant compatibility. In humans, there are 140 genes in the MHC region on Chromosome 6,[123] which can be divided into three major subgroups, MHC class I (*MHC I*), MHC class II (*MHC II*), and MHC class III (MHC III). The MHC I encodes heterodimeric peptide binding proteins and some antigen processing molecules including transporters associated with antigen processing (TAP) and Tapasin. The MHC II encodes heterodimeric peptide binding proteins and proteins that modulate peptide loading onto MHC class II proteins in the lysosomal compartment such as MHC II DM, MHC II DQ, and MHC II DP. Finally, the MHC III region encodes for other immune components, such as complement components (e.g., C2, C4, factor B), and some of this same region encodes for cytokines as well (e.g., TNF-α). *MHC I* proteins are constitutively expressed at almost all nucleated cells and present small intracellularly generated protein fragments to the CD8-positive T cells and natural killer (NK) cells, providing the basis for immune recognition of pathogen-infected cells.[124] The central feature of the highly ordered MHC I function is its ability to bind a wide spectrum of 8–10-mer peptides with high affinity within a specific groove formed by the αl and α2 domains.[125]

MHC II—proteins are heterodimers expressed on the surface of epithelial cells in the thymus and on professional antigen-presenting cells in the periphery. They display a wide range of peptides for recognition by the T-cell receptors of $CD4^+$ T helper cells. *MHC II* consists of two non-covalently associated polypeptide chains: the α-chain and the β-chain, which in addition to their extracellular regions, both have a single *transmembrane* sequence and a short cytoplasmic tail. The N-terminal α- and β-regions of the chains combine to form a membrane-distal peptide-binding domain, that consists of a groove, with a floor provided by a β-sheet, and two walls, each formed from an α-helix.[126] It has been pointed out that the displayed peptide is bound in the groove in an extended conformation that is imposed by hydrogen bonding from conserved MHC II residues (in the walls of the groove) to the peptide backbone.[127] This mode of interaction ensures that the conformation adopted by the bound peptide is independent of its sequence and bears no relationship to the conformation of the epitope sequence in the context of the native antigenic protein,[128] thus providing highly ordered MHC II with an outstanding binding plasticity and polymorphism.

Periplasmic space—This space is located between the plasma membrane and the outer membrane in the gram-negative bacteria. A smaller periplasmic space between the plasma membrane and the peptidoglycan layer (cell wall) may be observed in gram-positive bacteria. The periplasmic space is composed of a peptidoglycan (or murine) frame and is filled with an aqueous solution containing oligosaccharides, monosaccharides, proteins, and other solutes.[129] The proteins residing in or transiting through the periplasmic space are subjected to frequent environmental changes (e.g., extreme pH values, high salt concentrations and extreme temperatures) that may cause their unfolding and aggregation.[129] In contrast to non-periplasmic proteins, the periplasmic proteins are marginally stable but are exceptionally highly resistant toward aggregation as examined under various denaturing conditions.[129]

Lysosome—This membrane-enclosed organelle is found in plant and animal cells containing different hydrolytic enzymes that function in intracellular digestion. The important feature of lysosomes is their slightly acidic (pH 4.8) reducing interior, which is crucial for the conformational destabilization of proteins targeted for degradation. Lysosomal cysteine proteases, generally known as the cathepsins, are stable well-folded proteins that possess optimal activity in this slightly acidic, reducing milieu.[130] Cathepsins comprise a group of papain-related enzymes, sharing similar amino acid sequences and folds.[130]

Hexon protein—Adenoviruses are double-stranded DNA viruses that are found in all vertebrates.[131] Adenoviruses have an icosahedral capsid, the major components of which are the *hexon*, a trimeric protein with a hexagonal shape at its base, and the penton that is a noncovalent complex between the pentameric penton base and the trimeric fibre protein. In the capsid, 240 hexons form the 20 facets of the icosahedron, whereas the pentons form and project from the 12 vertices.[132] The atomic structures of the hexon of human adenoviruses type 5[133] and type 2[134] were determined.

Fimbriae *(also known as pili)*—These are proteinaceous appendages, which are formed from filamentous protein polymers, protrude from the outer surface of bacteria. *Fimbriae* are present in many gram-negative bacteria. They are thinner and shorter than a flagellum and are used by bacteria to adhere to one another during mating and to adhere to animal cells.[135] For example, type IV pili, being comprised of pilin, are key virulence factors for many important human pathogens including *Vibrio cholerae*, *Pseudomonas aeruginosa*, *Neisseria gonorrhoeae*, *Neisseria meningitidis*, and *Escherichia coli*.[136] Structurally, fimbriae are polymers of a single protein subunit, which is usually of relatively low molecular weight, but also contain small amounts of associated proteins.[137] The high-resolution X-ray crystal structures of the type IVb toxin-coregulated pili subunit, TcpA, which is one of the largest type IV pilins (199 amino acids), and full length PAK pilin, one of the smallest pilins (144 amino acids) have been recently reported.[136]

Bacterial capsule—This layer of surface-associated mucopolysaccharides provides an outer shell enveloping certain bacteria. Biosynthesis of the bacterial capsule is a complex process, which involves polymerization and transmembrane export of capsular polysaccharides followed by the surface assembly of the capsule. In *E.coli*,[138] polymerization involves sequential action of glycosyltransferases, whereas the transport of nascent polysaccharide across the plasma membrane is performed either by the polysaccharide exporter protein Wzx[139] or by the ABC-2 (ATP-binding cassette) transporter.[140] Surface assembly of group 1 capsules requires Wza[139] and Wzc.[141] Thus, proteins involved in the bacterial capsule production are either enzymes or membrane proteins.

Microsome—This small vesicle is derived from fragmented endoplasmic reticulum produced as a result of the cell homogenization and used for the isolation of cytochrome p450 enzymes (P450s or CYPs). The P450s are heme-containing monooxygenases, often catalyzing hydroxylation of hydrophobic substrate molecules. This hydroxylation plays a crucial role in the metabolism of the majority of drug molecules. P450s are also involved in the synthesis and degradation of numerous important endogenous compounds in many species of microorganisms, plants and animals. Crystal structures of several P450s have been solved, and these proteins all show approximately the same highly organized mixed α/β fold.[142]

Peroxisomes—These ubiquitous and essential eukaryotic organelles are characterized by the presence of a proteinaceous matrix surrounded by a single membrane. These organelles are involved in a variety of metabolic pathways, and it is estimated that mammalian peroxisomes contain about 50 different enzyme activities.[143] These many activities require the presence of a large number of peroxisomal proteins. These proteins are involved in peroxisomal alpha oxidation (acyl-CoA oxidases; peroxisomal multifunctional enzymes; thiolases; alpha-methylacyl-CoA racemase; carnitine acetyltransferase and carnitine octanoyltransferase; delta3,5-,delta2,4-dienoyl-CoA isomerase; peroxisomal 2,4-dienoyl-CoA reductase 2; peroxisomal 3,2-trans-enoyl-CoA isomerase; very-long-chain acyl-CoA synthetase; acyl-CoA thioesterases; and peroxisomal *trans*-2-enoyl-CoA reductase); peroxisomal beta oxidation (phytanoyl-CoA 2-hydroxylase and 2-hydroxyphytanoyl-CoA lyase); plasmalogen biosynthesis (dihydroxyacetone phosphate acyltransferase and fatty acyl-CoA reductases);

glyoxylate metabolism (alanine:glyoxylate aminotransferase); lysine metabolism (peroxisomal sarcosine oxidase/L-pipecolate oxidase); polyamine metabolism (N^1-acetylspermine/spermidine oxidase); and oxygen metabolism (catalase; peroxiredoxin V; d-amino acid oxidase; d-aspartate oxidase; glycolate oxidase; hydroxyacid oxidases, epoxide hydrolase; and glutathione S-transferase) among other crucial functions.[143] Importantly, all proteins listed above are enzymes. This implies that they tend to be ordered to carry out their catalytic activities.

Cell wall—The rigid outermost cell layer, the cell wall, is found primarily in plants and is composed of polysaccharides and proteins. Certain proteins are essential components of plant cell walls, participating in modifications of cell wall components, wall structure, signaling and interactions with plasma membrane proteins at the cell surface.[144] Many cell wall proteins (CWPs) are known to possess catalytic functions, and thus are ordered. The list of catalytic CWPs includes numerous glycoside hydrolases, glycosyl transferases, polysaccharide lyases, carbohydrate esterases, expansins, oxido-reductases and proteases.[144]

Cell envelope—The *cell membrane*, *cell wall*, an *outer membrane*, an *S-layer* (which is composed of proteins and glycoproteins directly attached to the outer membrane), and an *apoplast* (the free diffusion space outside the plasma membrane in plants), if one is present, comprise the cell envelope. Obviously, the cell envelope contains all the ordered proteins described above for its different components.

CF_1—F_0F_1-adenosine triphosphate (ATP) synthases are bound to energy-transducing membranes of chloroplasts, mitochondria, and bacteria. These enzymes couple a downhill proton flow to ATP synthesis through adenosine diphosphate (ADP) photophosphorylation.[145, 146] The ATP synthases can be biochemically fragmented into two different subcomplexes, known as F_0 and F_1 (CF_0 and CF_1 in chloroplasts). The CF_0 complex is membranous and acts as the proton channel, whereas the CF_1 moiety is a soluble $\alpha_3\beta_3\gamma\delta\varepsilon$ complex which is capable of ATP hydrolysis.[147] Crystal structures determined for several F_1 ATPases and corresponding subcomplexes revealed that these multichain enzymes are structured (ordered) and highly organized entities.[147, 148]

Reaction center—Photosynthetic *reaction centre* is a membrane proteinaceous machine, which is the site of the light-modulated reactions associated with photosynthesis. The first X-ray structure of the bacterial photochemical reaction centre was reported in 1984.[149] Since the initial work, high resolution structures have been determined for numerous bacterial reaction centres, light-harvesting 2 complexes, and Photosystem I and Photosystem II complexes.[150] All of these proteins are well-structured molecules located mostly within the confines of the bulges.

Antenna complex—Antenna is a complex of proteins in the thylakoid membrane of chloroplasts that captures and transfers light energy to the photochemical reaction center of the photosystem, one of the most fascinating membrane protein complexes.[151] The crystal structure of photosystem I from the thermophilic cyanobacterium *Synechococcus elongatus* provides a picture at atomic detail of 12 protein subunits and 127 cofactors comprising 96 chlorophylls, 2 phylloquinones, 3 Fe_4S_4 clusters, 22 carotenoids, 4 lipids, a putative Ca^{2+} ion and 201 water molecules.[152]

Eye lens—The α-, β- and γ-crystallins are the major protein components of the vertebrate eye lens. The α-crystallin is a molecular chaperone as well as a structural protein, whereas β- and γ-crystallins are structural proteins.[153] The outstanding feature of crystallins is that these proteins are not renewed. Thus, for the lens to be able to retain life-long transparency in the absence of protein turnover, the crystallins must meet not only the requirement of solubility associated with high cellular concentration but that of longevity as well.[153] Crystal structures of representative examples of the β- and γ-crystallins have been determined.[154]

Intrinsic disorder, protein domains and functions

The correlation between disorder and keywords associated with domains provides another important view of protein structure-function relationships. Protein functionality is often correlated with a protein domain, which is usually thought to be structured. Finding and comparing information related to a target protein domain represents a popular method for protein function identification. To our surprise, we discovered that some domains and structural keywords were strongly correlated with predicted disorder, suggesting that the "domains" are likely not structured units but rather are regions of sequence, whereas other domain-related keywords were strongly correlated with predicted order, as expected. Several of the domain-related keywords showing strongest disorder and order association are listed in Table 3 and Table 4, respectively.

These results provide useful information about relationships between protein structural properties, domain structure and disorder. Protein domains listed in Table 3 are frequently associated with intrinsic disorder. Known examples include *coiled-coils*, which are versatile protein domains, supporting a wide range of biological functions. Coiled-coil domains are protein-protein interaction motifs which consist of two or more α-helices that twist around one another to form a supercoil.[155, 156] Among well known two-stranded coiled-coil domains are short coiled-coil domains of six or seven heptad repeats, also called leucine zippers or *leucine-rich repeats* (e.g., jun-fos and GCN4 dimers) and long coiled-coil domains of several hundred amino acids (e.g., structural maintenance of chromosomes proteins, intermediate-filament proteins and nuclear lamins).The stability of the coiled-coil is derived from a characteristic interchain packing of the hydrophobic side chains into a hydrophobic core ('knobs into holes'[157,158]). *Collagen* represents the best known example of the three-stranded coiled-coils,[159] although this protein form is typically called a "triple helix" rather than a coiled-coil. The monomeric forms of coiled-coil and triple helix proteins are often completely disordered,[9] and, indeed, these proteins have sequence complexities lower than those observed for globular, structured proteins but within a range often seen for disordered proteins.[160] The flexibility of the coiled-coil domains and its importance for tropomyosin-actin interactions has been recently emphasized.[161]

Zinc-fingers—Many DNA-binding proteins have multiple copies of small independently folded domains that contain conserved cysteines and histidines coordinated to zinc; such proteins are commonly called *zinc-finger* proteins.[162] Zinc-finger domains are important constituents of transcription factors, including development-related homeodomain transcription factors that are encoded by the *homeobox* genes and that possess these conserved, 60-amino-acid DNA-binding domains.[163] The structure of a typical zinc-finger domain depends dramatically on the metal ion – in the presence of Zn^{2+} zinc-finger domains are well-folded proteins, whereas their apo-forms are often unfolded (e.g., see studies on the C-terminal zinc fingers of human MTF-1[164]). An important subclass of zinc-finger domains is the *LIM domain,* which is a cysteine-rich sequence found in proteins from a wide variety of eukaryotic organisms. In the human genome, there are 135 currently identified LIM-encoding sequences located within 58 genes.[163] LIM-domain containing proteins play important roles in a variety of fundamental biological processes including cytoskeleton organization, cell lineage specification and organ development. Importantly, the LIM domain has been demonstrated to be a protein-protein interaction motif that is critically involved in these processes.[165]

Repeats—The presence of *repeats* among other amino acid sequence biases represents a characteristic feature of many intrinsically disordered proteins,[160, 166] including α-synucelin[167] and protein tau,[47–49] which are both discussed above.

Transit peptides—Several thousand different proteins are targeted to the chloroplast via the *transit peptides* that act as chloroplast targeting sequences. These transit peptides are probably the largest class of targeting sequences in plants.[168] While these peptides are highly divergent in length, amino acid composition and amino acid sequence, they do possess one common structural feature – they are largely unstructured in an aqueous environment.[169–172]

SH3 domain *and* SH3-binding—The Src homology 3 domain (also known as *SH3 domain*) is a small protein domain of about 60 amino acid residues that is present in non-catalytic parts of several cytoplasmic tyrosine kinases (including Abl and Src) as well as in some other protein families, such as phospholypases, PI3 kinases, GAPs, CDC24, CDC25, etc.[173] For example, the cellular form of the Abelson leukemia virus tyrosine kinase (c-Abl) is a large protein that consists of ~1150 residues whose N-terminal half includes an N-terminal "cap" of ~80 residues that is important for autoinhibition followed by an SH3 domain, an SH2 domain, and a tyrosine kinase domain.[174] The analysis of the c-Abl kinase crystal structure revealed that there is no clearly interpretable electron density for the ~80 residue N-terminal cap region.[175] Although SH3 domain possesses a characteristic fold consisting of five or six β-strands arranged in a form of two tightly packed anti-parallel β sheets, with the linker regions containing short helices,[173] it has been pointed out that extensive flexibility is necessary for target recognition.[176] The vast majority of sequences responsible for the *SH3-binding* are short peptides (7–9 residues in length) that contain a XP-X-XP sequence motif: two XP dipeptides separated by a scaffolding residue. This sequence motif is unfolded in the unbound state but adopts a PPII conformation when bound by the SH3 domain.[177]

Signal-anchor—Similar to signal sequences, *signal-anchor* (SA) sequences interact transiently with the endoplasmic reticulum translocase, but are not cleaved and move laterally out of the translocase to become permanent membrane anchors.[178] The orientation of the protein on the membrane was shown to depend on the length of the hydrophobic (H) *transmembrane* segment and on the number and distribution of positively charged residues following the H-segment.[179] Importantly, it has been established that the translocating polypeptide chain forms an extended conformation.[179]

SH2 domain—The Src homology 2 domain (*SH2 domain*) is the prototype for protein–protein interaction modules that control the formation of multiprotein complexes during signaling.[180, 181] Since SH2 domains specifically recognize phosphorylated tyrosines (pTyrs) in binding partners, their functions are related to the protein tyrosine kinase (PTK) pathways. In addition to the pTyr residue, each SH2 domain recognizes several flanking residues, usually three to five amino acids C-terminal to pTyr, thereby acquiring selectivity for specific phosphorylated sites.[182] The high plasticity of SH2 domains has been emphasized, which might provide mechanical grounds for the “adjustable fit” used to accommodate various binding partners and thus this flexibility or disorder is relevant to the SH2 domain’s interactions with physiological ligands.[182]

Paired box—The *paired box* is a conserved 124 amino acid residue domain, which is encoded by the paired box-containing (PAX) gene family.[183] PAX genes encode a family of developmentally regulated transcription factors that have been implicated in a number of human and murine congenital disorders, as well as in tumorigenesis.[184, 185] PAX proteins contain an evolutionarily conserved DNA binding domain, known as the paired domain. Analysis of the purified PAX-6 paired domain by CD and NMR spectroscopies revealed that this protein is mostly unfolded in solution, but gains α-helical structure as a result of the DNA binding.[186]

EGF-like domain—Epidermal growth factor (EGF)-like proteins comprise a group of structurally similar growth factors, that contain a conserved six-cysteine residue motif called the EGF-like domain. Intrinsic disorder in growth factors has been already discussed (see above, section *Hormones and growth factors*). Importantly, many proteins with unrelated functions, have similar EGF-like domains. For example, the EGF-like domain is present in human C1, which is the multimolecular protease triggering the classical pathway of complement, a system that participates in innate immunity against various bacteria, parasites, and retroviruses.[187] In fact, the activation and enzymatic activity of C1 are mediated by two serine proteases, C1r and C1s, respectively, possessing the same type of modular organization, each containing two CUB modules surrounding a single EGF-like module, a pair of complement control protein (CCP) modules, and a serine protease domain.[188] The NMR analysis of the solution structure of the C1r-EGF module (residues 123–175) revealed that this domain is characterized by the well-ordered C-terminal part (residues Cys144-Ala174) and a highly disordered N-terminal part (residues 123–143).[189]

Phorbol-ester binding domain—The members of the protein kinase C (PKC) family play crucial role in regulation of various signaling pathways.[190–192] The activity of PKC is modulated by binding the 1,2-diacyl-*sn*-glycerol (DAG) or phorbol esters. The PKC consists of two functional domains, the "catalytic domain, " which contains the apparatus for protein phosphorylation, and the "regulatory domain", which retains phorbol ester/DAG-binding ability.[193] Solution structures of the peptides B (incorporating residues 36–87 of the first cysteine-rich repeat) and C (incorporating residues 101–151 of the second cysteine-rich repeat) derived from the phorbol-ester binding domain of the rat brain PKCγ were analyzed by NMR. This study revealed that the peptide B becomes ordered only in the presence of phospholipids, suggesting that that PKC might not be preorganized for ligand binding (activation) except when it is in the region of the cell where its endogenous activator is generated.[194]

Sushi domains (also known as Complement control protein (CCP) modules, or short consensus repeats (SCR)). These protein domains are present in a wide variety of complement and adhesion proteins, such as CD21 (C3d receptor), Epstein Barr virus receptor and factor H.[195] Many of these proteins contain tandem arrays of Sushi domains interspersed by short linking sequences. Factor H, for example is made up of 20 Sushi domains.[196] Sushi domains are characterized by a consensus sequence spanning approximately 60 residues and are involved in protein–protein and protein–ligand interactions.[197] NMR analysis of the solution structure of the interleukin-15α receptor sushi domain revealed that that significant part of this module does not have regular secondary structure, with only 18 out of total 84 residues being involved in the formation of five very short β-strands.[198]

Bromodomains—Many chromatin-associated proteins and nearly all known nuclear histone acetyltransferases (HATs) contain these domains.[199] Bromodomains recognize acetylated lysine residues on the N-terminal tails of *histones* and other proteins. This recognition triggers a crucial mechanism for regulating protein–protein interactions in numerous cellular processes including *chromatin remodeling* and *transcriptional activation*.[200] The role of intrinsic disorder in the function of histones and other proteins involved in *chromatin remodeling* and *transcriptional activation* has been already discussed. NMR analysis revealed that a prototypical bromodomain from the transcriptional co-activator p300/CBP-associated factor (P/CAF) adopts an atypical left-handed up-and-down four-helix bundle with two highly flexible loops ZA and BC,[201] which are responsible for the accommodation of numerous bromodomain binding partners.[200]

Kringle—These domains are conserved sequences that are involved in protein-protein interactions and that fold into large loops stabilized by 3 disulfide bridges.[202, 203] For example, it has been shown that the solution structure of human apolypoprotein(a) kringle IV type 6 (119 amino acid residues) contains only a small amount of regular secondary structure elements, including a short piece of antiparallel β-sheet formed by residues Trp62-Tyr64 and Trp72-Tyr74, a short piece of parallel β-sheet formed by the residues Cys1-Tyr2 and Thr78-Gln79, and a small 3_{10}-helix within residues Thr38-Tyr40.[203]

Domains associated with ordered proteins

The protein domains listed in Table 4 are associated with structured proteins.

Transmembrane *proteins*—Using the examples of porins,[204] it has been already pointed out that *transmembrane* protein domains are highly ordered.

Immunoglobulin V regions—Also known as V or variable domains, these regions are complementarity-determining regions (immunoglobulin binding sites) known to possess high amino acid diversity that determines the endless specificity of immunoglobulins.[205] These domains are mostly ordered and possess specific immunoglobulin folds.

Glutamine amidotransferase—Glutamine amidotransferases (GATases) are ubiquitous enzymes that transfer the amide nitrogen of glutamine to a variety of substrates.[206] GATases catalyze two separate reactions at two active sites, which are located either on a single polypeptide chain or on different subunits. In the glutaminase reaction, glutamine is hydrolyzed to glutamate and ammonia, which is added to an acceptor substrate in the synthase reaction. Crystal structures of several GATases were solved. For example, the crystal structure of human gamma-glutamyl hydrolase, a class I glutamine amidotransferase, determined at 1.6-Å resolution, reveals that the protein contains 11 α-helices and 14 β-strands, with a fold in which a central eight-stranded β-sheet is sandwiched by three and five α-helices on each side.[207]

Redox-active center—The heart of numerous well-folded redox-active enzymes, including different oxidoreductases, such as CcmG protein,[208] thioredoxin and thioredoxin-like proteins,[209] DsbA,[210] DsbC,[211] and TlpA,[212] is the redox-active center.

TonB box—This is a conserved N-terminal region of the TonB-dependent outer membrane transporters (TBDTs), which are involved in the import of essential organometallic micronutrients (such as iron-siderophores and vitamin B_{12}) across the outer membrane of Gram-negative bacteria. Crystal structures of five TBDTs have been determined, which illustrate clearly the architecture of the protein where an N-terminal hatch (or plug or cork) domain occludes the lumen of a 22-stranded β barrel.[213]

Kelch repeat—This repeat is a segment of 44–56 amino acids in length, which usually appears as a series of four to seven motifs that form a kelch repeat domain.[214, 215] Kelch repeats collectively form a β-propeller, where each kelch motif forms a four-stranded β-sheet corresponding to a single blade of the propeller, with the series of blades being tilted around a central axis.[216]

Ank repeat—The ankyrin repeat, a 33-residue sequence motif, is one of the most abundant repeat motifs in proteins: the PFAM-A database, as of October 2003, contained 9689 ankyrin repeat sequences in 1871 proteins identified from the Swissprot and SP-TrEMBL databases, whereas the SMART database contained 19,276 ankyrin repeat sequences in 3608 proteins identified from the nonredundant protein database.[217] Ankyrin repeats have been observed to exist by themselves as a single domain protein or in conjunction with other domains in the same protein. The number of ankyrin repeats contained in a single protein varies from one to 33 repeats per protein, with the majority of proteins containing six or fewer repeats.[217] The crystal structures of several ankyrin repeat proteins have been solved, showing that the repeat has a well-defined structure where the polypeptide chain folds into two antiparallel α-helices followed by a β-hairpin or a long loop.[217]

CBS domain—The cystathionine-beta-synthase (CBS) domain is an evolutionarily conserved protein domain of ~60 amino acids that is present in the proteome of archaebacteria, prokaryotes, and eukaryotes and that is usually found in cytosolic and membrane proteins performing different functions (metabolic enzymes, kinases, and channels).[218] The crystal structures of several bacterial proteins containing CBS domains show how two CBS domains associate to form a CBS pair, with a single CBS domain consisting of a conserved β_1-α_1-β_2-β_3-α_2 pattern.[218]

Annexins—This proteins comprise a unique class of Ca^{2+}-effectors that mediate cellular responses to changes in intracellular Ca^{2+} levels and that can bind to certain membrane phospholipids in a Ca^{2+}-dependent manner, thus providing a link between Ca^{2+} signaling and important membrane functions.[219] High-resolution crystal structures of several annexins have been determined.[219]

Technical terms strongly correlated with predicted disorder

Table 5 represents keywords describing technical terms strongly associated with intrinsic disorder.

Pharmaceutical—Some proteins are used as *pharmaceutical* drugs to treat different diseases. Among pharmaceuticals are such proteins as *cytokines*, *protease inhibitors*, *toxins*, *antimicrobial peptides*, and *immunoglobulins*, for which the importance of intrinsic disorder was already discussed.

ERV—Endogenous retroviruses (*ERVs*) are vertically transmitted intragenomic elements derived from integrated retroviruses. They reproduce within the somatic tissues of infected individuals and can proliferate within the genome of their host until they either acquire inactivating mutations or are lost by recombinational deletion. Retroviruses are RNA viruses. To insert themselves into the DNA-based genome of hosts, they encode a unique enzyme, reverse transcriptase (RT) that copies the viral RNA template to its complementary DNA, which is then integrated into the chromosomes. Intriguingly, 8% of the human genome was shown to consist of human endogenous retroviruses, or HERVs. It has been suggested that being extended to HERV fragments and derivatives, the retroviral heritage could account for almost a half of human DNA.[220, 221] Illustrative example of human endogenous retrovirus is HIV, which will be discussed in the last paper of this series.[222]

Technical terms strongly correlated with predicted order

Table 6 represents technical terms which are strongly correlated with predicted order.

Multifuncitonal enzyme—Enzymes (including *multifunctional* and *allosteric enzymes*) illustrate the validity of the standard structure-function paradigm, which states that the function of a given protein is determined by its 3-D structure.[10]

Hypothetical proteins—Proteins corresponding to open reading frames but for which there is no experimental evidence that they are expressed *in vivo* are called hypothetical. Many of the hypothetical proteins are predicted to be enzymes with different catalytic activities (see respective entries in the Swiss-Prot databse).

3D-structure—Importantly, a key-word *3D-structure* is shown to be strongly correlated with proteins predicted to be ordered (see Table 6).

Plasmids—Many proteins is Swiss-Prot are encoded by *plasmids*, which are self-replicating circular DNA molecules that can be transferred from one organism to another. Plasmids often code for different enzymes (e.g., deaminases, DNA glycosylases, acetyltransferases, nucleotidyltransferases, kinases, etc.), functions of which are known to be dependent on rigid highly ordered structures.

Developmental keywords strongly correlated with predicted disorder

Developmental keywords predicted to be strongly associated with intrinsic disorder are listed in Table 7.

Merozoite—This daughter cell of the protozoan parasite is produced during asexual reproduction. One of the merozoite-related proteins is the *Plasmodium falciparum* acidic–basic repeat antigen (ABRA), localized in the parasitophorous vacuole and associated with the merozoite surface. ABRA is 743 amino acid residues long surface protein, which has several heavily charged tandem repeats and lysine-rich C-terminal tail (residues 672–743). Another illustrative example is the apical membrane antigen 1 of the malarial parasite *Plasmodium falciparum* (Pf AMA1), which is considered to be a strong candidate for inclusion in a malaria vaccine. The solution structure of AMA1 domain III, a 14kDa protein, has been determined using NMR spectroscopy. The protein was shown to consist of unstructured the N-and C-terminal regions (4 and 37 residues, respectively), and a well-defined disulphide-stabilized core region containing a long disordered loop of 27 residues.[223]

Early proteins—These viral proteins are produced following entry into the host cell but prior to virus assembly. The expression of viral genes encoding early, non-structural proteins initiates replication of the viral genome and expression of late genes. Among early proteins are transcription activators, transacting transcriptional proteins, DNA-binding proteins, and DNA polymerases, all of which are known to rely on intrinsic disorder in their activities. Infected cell protein 47 (ICP47) is an early protein encoded by the herpes simplex virus (HSV). ICP47 is a crucial factor in the evasion of cellular immune response against HSV-infected cells, acting as a specific inhibitor of the transporter associated with antigen processing (TAP), thus preventing peptide transport into the endoplasmic reticulum, and downregulating the subsequent loading of major MHC class I molecules.[224] CD and NMR spectroscopic analyses of the ICP47 active domain (residues 2–34) revealed that this peptide has no ordered secondary structure in aqueous solution, but adopts an α-helical conformation in the presence of membrane mimetics.[225, 226]

Sporozoites—These protozoan cells infect new hosts and so are cruical for spreading the infection. Circumsporozoite (CS) protein from different *Plasmodium* species display common sequence features, including a *signal peptide*, a central domain composed mostly of amino acids *repeats*, and a C-terminal hydrophobic sequence.[227] Following sporozoite invasion of hepatocytes, CS is also detected on the plasma membrane of early exo-erythrocytic forms and in the cytoplasm of the infected cells.[228] To understand the conformational preferences of the tandemly repeating tetrapeptide unit of the circumsporozoite coat protein of the malaria parasite *Plasmodium falciparum*, peptides based on the Asn-Ala-Asn-Pro and Asn-Pro-Asn-Ala cadences and composed of one to three tetrapeptide units were analyzed by CD and NMR. These peptides are significantly unfolded in aqueous media.[229]

Intrinsic disorder and coding sequence diversity

Keywords associated with coding sequence diversity that potentially relies on protein intrinsic disorder are listed in Table 8.

Alternative splicing—While splicing of pre-mRNA by joining exons and discarding introns occurs in all eukaryotes, alternative splicing takes place commonly only in muticellular eukaryotes and so is one of the fundamental components of gene regulation associated with cell differentiation. Alternative splicing occurs when different mRNAs are assembled from a single gene by joining exons in different ways. Thus, this process is proposed to generate complexity in multicellular eukaryotes by increasing protein diversity, and thus proteome size, from a relatively small number of genes.[230] Alternative splicing can modulate organism complexity, not only by effectively increasing proteome size and regulatory and signaling network complexity, but also by doing so in a time- and tissue-specific manner, supporting cell differentiation, developmental pathways, and other processes associated with multicellular organisms.[231] Recently we have shown that a large majority (~80%) of alternatively spliced fragments in a set of experimentally characterized proteins is associated with fully or partially disordered regions. This suggests that polypeptide segments affected by alternative splicing are most often intrinsically disordered. Therefore, alternative splicing enables functional and regulatory diversity while avoiding structural complications associated with the removal of segments from well-folded proteins.[232] In addition, serine/arginine-rich splicing factors that play an important role in alternative splicing have been shown to belong to a class of intrinsically disordered proteins.[38]

Polymorphism—The ability to appear in many forms is denoted as polymorphism. DNA polymorphism occurring in coding regions might result in the development of numerous pathological conditions, including Huntington’s disease and other triplet repeat expansion diseases (see below), cystic fibrosis, inherited muscular dystrophy, cancers and many other diseases.[233] Genetic polymorphism is responsible for the appearance of different protein isoforms, which might be different in their functionality and structural organization. For example, it has been shown that human apolipoprotein E (apoE), which plays a key role in cholesterol transport and lipoprotein metabolism, has three major isoforms in human, apoE2, apoE3 and apoE4, with apoE3 being the most common isoform. The isoforms differ at residues 112 and 158; whereas apoE3 has cysteine at position 112 and arginine at 158, apoE2 has cysteine and apoE4 arginine at both positions. The functional and structural consequences of these point mutations are significant.[234] For example, based on the detailed analysis of the conformational behavior of these three isoforms, it has been suggested that apoE3 and apoE4, but not apoE2, may be partially unfolded *in vivo*.[235] A recent analysis of the genetic polymorphism in calmodulin superfamily, which is a major class of Ca^{2+} sensor proteins, represents another illustrative example.[236] While the calmodulin amino acid sequence is highly conserved in all eukaryotes,[237] there is slight variability among the nearly 600 members in the calmodulin superfamily.[238] In contrast, the troponin C family has just two isoforms in humans (skeletal and cardiac muscles) but many isoforms in invertebrates.[239] Similarly, the neuronal calcium sensor (NCS) and S100 proteins are highly diverse in sequence and function.[236] Intrinsic disorder plays a crucial role in the function of all these Ca^{2+} sensor proteins. Recently, using bioinformatics approaches, we have established that calmodulin-binding targets are intrinsically disordered.[240]

Chromosomal translocation—This involves the interchange of parts between nonhomologous chromosomes. It occurs in leukemias, lymphomas, sarcomas and some epithelial tumors. Some chromosomal translocations are known to generate unique fusion proteins.[241] For example, in Ewing's sarcoma family of tumors, which are highly malignant tumors of bone and soft tissue that occur in children, adolescents, and young adults, the EWS-

FLI1 protein is produced as a result of the translocation-generated fusion.[242] EWS-FLI1 combines the amino-terminus of EWS (residues 1–264) from chromosome 22 with the carboxy-terminus of FLI1 (232 carboxy-terminal residues) from chromosome 11. EWS-FLI1 retains the conserved Ets DNA binding domain from FLI1 in the fusion protein suggesting function as a transcription factor.[242, 243] The EWS-FLI1 protein has been shown recently to be highly disordered.[244]

Triplet repeat expansions—These DNA sequence aberrations occur in both coding and non-coding regions. Such expansions are the cause for several neurogenetic disorders. The list of coding trinucleotide expansion disorders includes eight polyglutamine diseases (Huntington's disease, dentatorubropallidoluysian atrophy, spinobulbar muscle atrophy and spinocerebellar ataxia types 1, 2, 3, 6 and 7). All of these diseases are associated with the same repeated codon, CAG that codes for glutamine. There are also numerous pathological conditions associated with polyalanine tract expansions.[245, 246] The polyalanine tract disorders are based on the repetitive expansion of the GCG codon. Non-coding trinucleotide expansion disorders are the fragile X syndrome caused by the expansion of CGG trinuleotide; fragile XE mental retardation due to GCC expansion; Friedreich ataxia is induced by the multiplication of GAA triplet; myotonic dystrophy and spinocerebellar ataxia type 8 are caused by the CTG expansion, and finally spinocerebellar ataxia type 12 is caused by CAG expansion in the non-coding region. For these various triplet repeats, the normal alleles contain 5–35 copies of corresponding trinucleotides, whereas the expanded repeats range from small expansions of 20–100 copies to larger expansions of up to several thousand units.[246, 247]

The mechanisms of the non-coding trinucleotide expansion disorders are unique for each disease. For example, in the fragile X syndrome, the CGG repeating codon is expanded to such a degree that the corresponding portion of DNA is easily methylated, effectively silencing the expression of the FMR1 protein. As for the coding repeats, the deases states evidently result from the conformational changes brought about by the codon extensions. For example, the abnormal and destabilized protein products with extended polyglutamine tracts and altered function are thought to be the major factors in pathogenesis.[248, 249] Recently, computational analysis of glutamine-stretch embedded domains in the respective proteins predicted these regions to be "natively unfolded " when they extend beyond a threshold of 40 glutamines.[250] In another bioinformatic study polyglutamine repeats derived from nine disease proteins have been found to be unfolded independently of their lengths.[251] Chen and colleagues have found that monomeric polyglutamine, which is a disordered statistical coil in solution, is the critical nucleus for aggregation.[252] Multiple molecular dynamics simulations confirmed the disordered nature of the polyQ motifs and showed that the effective concentration of side chain primary amides around backbone units is inherently high, and that peptide units are solvated either by hydrogen bonds to side chains or surrounding water molecules.[253] Furthermore, NMR experiments also show that monomers of such long polyglutamines adopt random coil conformations.[254]

Alternative initiation *and* alternative promoter usage—The genes for *p63*, *p73* and *p53* have a dual structure conserved in drosophila, zebrafish and man. They encode for multiple p63, p73 or p53 proteins (a family of transcription factors involved in cell response to stress and development) containing different protein domains (isoforms) due to multiple alternative splicing, *alternative promoter usage* and *alternative initiation* of translation.[255] The crucial role of intrinsic disorder in the functioning of the tumor suppressor protein p53 has been already discussed (see apoptosis-related section in the first paper of this series[20]). One of the p53 isoforms, Δ40p53 (also named p47 or ΔNp53) is an amino-terminally truncated p53 protein with the first 40 amino-acids being deleted, which likely results from alternative initiation of translation.[255] Note that the deleted residues comprise the intrinsically disordered Mdm2 binding domain. Thus, together with the alternative splicing, alternative initiation might be commonly associated with the intrinsically disordered regions for similar reasons, namely to avoid structural complications that arise from the removal of segments of structured proteins.

Conclusions

As suggested previously (see the first paper in this series[20]), our bioinformatics approach reveals high positive or negative correlations between putative disorder and many functional keywords. To supplement the bioinformatics analysis, manual literature mining was carried out. The literature suggests over and over that, for the disorder-related SwissProt keywords, the functions are indeed carried out or enabled by the regions of protein disorder. For the order-related keywords, the functions, which very often involve enzymatic catalysis, are carried out by structured regions of proteins.

Acknowledgment

The authors express their deepest gratitude to Celeste Brown and Predrag Radovojac 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. Uversky VN. Natively unfolded proteins: a point where biology waits for physics. Protein Sci 2002;11:739–756. [PubMed: 11910019]
2. Uversky VN. What does it mean to be natively unfolded? Eur J Biochem 2002;269:2–12. [PubMed: 11784292]
3. 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]
4. 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.
5. 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]
6. Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Garner E, Guilliot S, Dunker AK. Thousands of proteins likely to have long disordered regions. Pac Symp Biocomput 1998:437–448. [PubMed: 9697202]
7. 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]
8. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 1999;293:321–331. [PubMed: 10550212]
9. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 2000;41:415–427. [PubMed: 11025552]
10. 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]
11. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry 2002;41:6573–6582. [PubMed: 12022860]
12. Dunker AK, Brown CJ, Obradovic Z. Identification and functions of usefully disordered proteins. Adv Protein Chem 2002;62:25–49. [PubMed: 12418100]
13. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. Curr Opin Struct Biol 2002;12:54–60. [PubMed: 11839490]

14. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208. [PubMed: 15738986]
15. Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci 2002;27:527–533. [PubMed: 12368089]
16. 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.
17. 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]
18. 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]
19. Dunker AK, Obradovic Z. The protein trinity--linking function and disorder. Nat Biotechnol 2001;19:805–806. [PubMed: 11533628]
20. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. Functional anthology of intrinsic disorder. I. Biological processes and functions of proteins with long disordered regions. J Proteome Res. 2006
21. 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]
22. 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]
23. 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]
24. 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]
25. 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]
26. Reeves R. Molecular biology of HMGA proteins: hubs of nuclear function. Gene 2001;277:63–81. [PubMed: 11602345]
27. Lund T, Holtlund J, Fredriksen M, Laland SG. On the presence of two new high mobility group-like proteins in HeLa S3 cells. FEBS Lett 1983;152:163–167. [PubMed: 6297996]
28. 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]
29. 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.
30. 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]
31. Will CL, Schneider C, Reed R, Luhrmann R. Identification of both shared and distinct proteins in the major and minor spliceosomes. Science 1999;284:2003–2005. [PubMed: 10373121]
32. Will CL, Schneider C, MacMillan AM, Katopodis NF, Neubauer G, Wilm M, Luhrmann R, Query CC. A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site. Embo J 2001;20:4536–4546. [PubMed: 11500380]
33. Will CL, Schneider C, Hossbach M, Urlaub H, Rauhut R, Elbashir S, Tuschl T, Luhrmann R. The human 18S U11/U12 snRNP contains a set of novel proteins not found in the U2-dependent spliceosome. Rna 2004;10:929–941. [PubMed: 15146077]
34. Query CC, Strobel SA, Sharp PA. Three recognition events at the branch-site adenine. Embo J 1996;15:1392–1402. [PubMed: 8635472]
35. Spadaccini R, Reidt U, Dybkov O, Will C, Frank R, Stier G, Corsini L, Wahl MC, Luhrmann R, Sattler M. Biochemical and NMR analyses of an SF3b155-p14-U2AF-RNA interaction network involved in branch point definition during pre-mRNA splicing. Rna 2006;12:410–425. [PubMed: 16495236]

36. Zahler AM, Lane WS, Stolk JA, Roth MB. SR proteins: a conserved family of pre-mRNA splicing factors. Genes Dev 1992;6:837–847. [PubMed: 1577277]

37. Roscigno RF, Garcia-Blanco MA. SR proteins escort the U4/U6.U5 tri-snRNP to the spliceosome. RNA 1995;1:692–706. [PubMed: 7585254]

38. 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]

39. Lee G, Cowan N, Kirschner M. The primary structure and heterogeneity of tau protein from mouse brain. Science 1988;239:285–288. [PubMed: 3122323]

40. Himmler A, Drechsel D, Kirschner MW, Martin DW Jr. Tau consists of a set of proteins with repeated C-terminal microtubule-binding domains and variable N-terminal domains. Mol Cell Biol 1989;9:1381–1388. [PubMed: 2498649]

41. Goedert M, Spillantini MG, Jakes R, Rutherford D, Crowther RA. Multiple isoforms of human microtubule-associated protein tau: sequences and localization in neurofibrillary tangles of Alzheimer's disease. Neuron 1989;3:519–526. [PubMed: 2484340]

42. Butner KA, Kirschner MW. Tau protein binds to microtubules through a flexible array of distributed weak sites. J Cell Biol 1991;115:717–730. [PubMed: 1918161]

43. Gustke N, Trinczek B, Biernat J, Mandelkow EM, Mandelkow E. Domains of tau protein and interactions with microtubules. Biochemistry 1994;33:9511–9522. [PubMed: 8068626]

44. Goode BL, Denis PE, Panda D, Radeke MJ, Miller HP, Wilson L, Feinstein SC. Functional interactions between the proline-rich and repeat regions of tau enhance microtubule binding and assembly. Mol Biol Cell 1997;8:353–365. [PubMed: 9190213]

45. Friedhoff P, von Bergen M, Mandelkow EM, Mandelkow E. Structure of tau protein and assembly into paired helical filaments. Biochim Biophys Acta 2000;1502:122–132. [PubMed: 10899437]

46. Mandelkow EM, Mandelkow E. Tau in Alzheimer's disease. Trends Cell Biol 1998;8:425–427. [PubMed: 9854307]

47. Wille H, Drewes G, Biernat J, Mandelkow EM, Mandelkow E. Alzheimer-like paired helical filaments and antiparallel dimers formed from microtubule-associated protein tau in vitro. J Cell Biol 1992;118:573–584. [PubMed: 1639844]

48. Schweers O, Schonbrunn-Hanebeck E, Marx A, Mandelkow E. Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. J Biol Chem 1994;269:24290–24297. [PubMed: 7929085]

49. von Bergen M, Barghorn S, Biernat J, Mandelkow EM, Mandelkow E. Tau aggregation is driven by a transition from random coil to beta sheet structure. Biochim Biophys Acta 2005;1739:158–166. [PubMed: 15615635]

50. Brown HG, Hoh JH. Entropic exclusion by neurofilament sidearms: a mechanism for maintaining interfilament spacing. Biochemistry 1997;36:15035–15040. [PubMed: 9424114]

51. Geisler N, Vandekerckhove J, Weber K. Location and sequence characterization of the major phosphorylation sites of the high molecular mass neurofilament proteins M and H. FEBS Lett 1987;221:403–407. [PubMed: 3114005]

52. Cassimeris L. The oncoprotein 18/stathmin family of microtubule destabilizers. Curr Opin Cell Biol 2002;14:18–24. [PubMed: 11792540]

53. Shumyatsky GP, Malleret G, Shin RM, Takizawa S, Tully K, Tsvetkov E, Zakharenko SS, Joseph J, Vronskaya S, Yin D, Schubart UK, Kandel ER, Bolshakov VY. stathmin, a gene enriched in the amygdala, controls both learned and innate fear. Cell 2005;123:697–709. [PubMed: 16286011]

54. Steinmetz MO, Kammerer RA, Jahnke W, Goldie KN, Lustig A, van Oostrum J. Op18/stathmin caps a kinked protofilament-like tubulin tetramer. Embo J 2000;19:572–580. [PubMed: 10675326]

55. Ravelli RB, Gigant B, Curmi PA, Jourdain I, Lachkar S, Sobel A, Knossow M. Insight into tubulin regulation from a complex with colchicine and a stathmin-like domain. Nature 2004;428:198–202. [PubMed: 15014504]

56. Honnappa S, Jahnke W, Seelig J, Steinmetz MO. Control of intrinsically disordered stathmin by multisite phosphorylation. J Biol Chem 2006;281:16078–16083. [PubMed: 16554300]

57. Schueler MG, Sullivan BA. Structural and Functional Dynamics of Human Centromeric Chromatin. Annu Rev Genomics Hum Genet. 2006

58. McAinsh AD, Tytell JD, Sorger PK. Structure, function, and regulation of budding yeast kinetochores. Annu Rev Cell Dev Biol 2003;19:519–539. [PubMed: 14570580]

59. Maresca TJ, Heald R. The long and the short of it: linker histone H1 is required for metaphase chromosome compaction. Cell Cycle 2006;5:589–591. [PubMed: 16582611]

60. Wigge PA, Kilmartin JV. The Ndc80p complex from Saccharomyces cerevisiae contains conserved centromere components and has a function in chromosome segregation. J Cell Biol 2001;152:349–360. [PubMed: 11266451]

61. Wei RR, Sorger PK, Harrison SC. Molecular organization of the Ndc80 complex, an essential kinetochore component. Proc Natl Acad Sci U S A 2005;102:5363–5367. [PubMed: 15809444]

62. Iakoucheva LM, Kimzey AL, Masselon CD, Smith RD, Dunker AK, Ackerman EJ. Aberrant mobility phenomena of the DNA repair protein XPA. Protein Sci 2001;10:1353–1362. [PubMed: 11420437]

63. Namba K. Roles of partly unfolded conformations in macromolecular self-assembly. Genes Cells 2001;6:1–12. [PubMed: 11168592]

64. Kostyukova AS, Pyatibratov MG, Filimonov VV, Fedorov OV. Flagellin parts acquiring a regular structure during polymerization are disposed on the molecule ends. FEBS Lett 1988;241:141–144. [PubMed: 3058510]

65. Vonderviszt F, Kanto S, Aizawa S, Namba K. Terminal regions of flagellin are disordered in solution. J Mol Biol 1989;209:127–133. [PubMed: 2810365]

66. Mimori-Kiyosue Y, Vonderviszt F, Namba K. Locations of terminal segments of flagellin in the filament structure and their roles in polymerization and polymorphism. J Mol Biol 1997;270:222–237. [PubMed: 9236124]

67. Puthenveedu MA, Linstedt AD. Subcompartmentalizing the Golgi apparatus. Curr Opin Cell Biol 2005;17:369–375. [PubMed: 15975779]

68. Bannykh SI, Plutner H, Matteson J, Balch WE. The role of ARF1 and rab GTPases in polarization of the Golgi stack. Traffic 2005;6:803–819. [PubMed: 16101683]

69. Latijnhouwers M, Hawes C, Carvalho C, Oparka K, Gillingham AK, Boevink P. An Arabidopsis GRIP domain protein locates to the trans-Golgi and binds the small GTPase ARL1. Plant J 2005;44:459–470. [PubMed: 16236155]

70. Nardini M, Spano S, Cericola C, Pesce A, Massaro A, Millo E, Luini A, Corda D, Bolognesi M. CtBP/BARS: a dual-function protein involved in transcription co-repression and Golgi membrane fission. Embo J 2003;22:3122–3130. [PubMed: 12805226]

71. Turner J, Crossley M. The CtBP family: enigmatic and enzymatic transcriptional co-repressors. Bioessays 2001;23:683–690. [PubMed: 11494316]

72. Weigert R, Silletta MG, Spano S, Turacchio G, Cericola C, Colanzi A, Senatore S, Mancini R, Polishchuk EV, Salmona M, Facchiano F, Burger KN, Mironov A, Luini A, Corda D. CtBP/BARS induces fission of Golgi membranes by acylating lysophosphatidic acid. Nature 1999;402:429–433. [PubMed: 10586885]

73. Hidalgo Carcedo C, Bonazzi M, Spano S, Turacchio G, Colanzi A, Luini A, Corda D. Mitotic Golgi partitioning is driven by the membrane-fissioning protein CtBP3/BARS. Science 2004;305:93–96. [PubMed: 15232108]

74. Bonazzi M, Spano S, Turacchio G, Cericola C, Valente C, Colanzi A, Kweon HS, Hsu VW, Polishchuck EV, Polishchuck RS, Sallese M, Pulvirenti T, Corda D, Luini A. CtBP3/BARS drives membrane fission in dynamin-independent transport pathways. Nat Cell Biol 2005;7:570–580. [PubMed: 15880102]

75. Nardini M, Svergun D, Konarev PV, Spano S, Fasano M, Bracco C, Pesce A, Donadini A, Cericola C, Secundo F, Luini A, Corda D, Bolognesi M. The C-terminal domain of the transcriptional corepressor CtBP is intrinsically unstructured. Protein Sci 2006;15:1042–1050. [PubMed: 16597837]

76. Catalano D, Licciulli F, Turi A, Grillo G, Saccone C, D'Elia D. MitoRes: a resource of nuclear-encoded mitochondrial genes and their products in Metazoa. BMC Bioinformatics 2006;7:36. [PubMed: 16433928]

77. Santel A, Blumer N, Kampfer M, Renkawitz-Pohl R. Flagellar mitochondrial association of the male-specific Don Juan protein in Drosophila spermatozoa. J Cell Sci 1998;111(Pt 22):3299–3309. [PubMed: 9788872]

78. Santel A, Winhauer T, Blumer N, Renkawitz-Pohl R. The Drosophila don juan (dj) gene encodes a novel sperm specific protein component characterized by an unusual domain of a repetitive amino acid motif. Mech Dev 1997;64:19–30. [PubMed: 9232593]

79. Kanki T, Nakayama H, Sasaki N, Takio K, Alam TI, Hamasaki N, Kang D. Mitochondrial nucleoid and transcription factor A. Ann N Y Acad Sci 2004;1011:61–68. [PubMed: 15126284]

80. Hegerl R, Pfeifer G, Puhler G, Dahlmann B, Baumeister W. The three-dimensional structure of proteasomes from Thermoplasma acidophilum as determined by electron microscopy using random conical tilting. FEBS Lett 1991;283:117–121. [PubMed: 2037064]

81. Groll M, Bochtler M, Brandstetter H, Clausen T, Huber R. Molecular machines for protein degradation. Chembiochem 2005;6:222–256. [PubMed: 15678420]

82. Dalton WS. The proteasome. Semin Oncol 2004;31:3–9. [PubMed: 15799238]discussion 33.

83. Prakash S, Tian L, Ratliff KS, Lehotzky RE, Matouschek A. An unstructured initiation site is required for efficient proteasome-mediated degradation. Nat Struct Mol Biol 2004;11:830–837. [PubMed: 15311270]

84. Walter P, Johnson AE. Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane. Annu Rev Cell Biol 1994;10:87–119. [PubMed: 7888184]

85. Rapoport TA, Matlack KE, Plath K, Misselwitz B, Staeck O. Posttranslational protein translocation across the membrane of the endoplasmic reticulum. Biol Chem 1999;380:1143–1150. [PubMed: 10595576]

86. Rapoport TA, Jungnickel B, Kutay U. Protein transport across the eukaryotic endoplasmic reticulum and bacterial inner membranes. Annu Rev Biochem 1996;65:271–303. [PubMed: 8811181]

87. Nagai K, Oubridge C, Kuglstatter A, Menichelli E, Isel C, Jovine L. Structure, function and evolution of the signal recognition particle. Embo J 2003;22:3479–3485. [PubMed: 12853463]

88. Wild K, Weichenrieder O, Strub K, Sinning I, Cusack S. Towards the structure of the mammalian signal recognition particle. Curr Opin Struct Biol 2002;12:72–81. [PubMed: 11839493]

89. Sauer-Eriksson AE, Hainzl T. S-domain assembly of the signal recognition particle. Curr Opin Struct Biol 2003;13:64–70. [PubMed: 12581661]

90. Pakhomova ON, Deep S, Huang Q, Zwieb C, Hinck AP. Solution structure of protein SRP19 of Archaeoglobus fulgidus signal recognition particle. J Mol Biol 2002;317:145–158. [PubMed: 11916385]

91. Eichacker LA, Henry R. Function of a chloroplast SRP in thylakoid protein export. Biochim Biophys Acta 2001;1541:120–134. [PubMed: 11750668]

92. Jonas-Straube E, Hutin C, Hoffman NE, Schunemann D. Functional analysis of the protein-interacting domains of chloroplast SRP43. J Biol Chem 2001;276:24654–24660. [PubMed: 11306572]

93. Goforth RL, Peterson EC, Yuan J, Moore MJ, Kight AD, Lohse MB, Sakon J, Henry RL. Regulation of the GTPase cycle in post-translational signal recognition particle-based protein targeting involves cpSRP43. J Biol Chem 2004;279:43077–43084. [PubMed: 15292240]

94. Sivaraja V, Kumar TK, Leena PS, Chang AN, Vidya C, Goforth RL, Rajalingam D, Arvind K, Ye JL, Chou J, Henry R, Yu C. Three-dimensional solution structures of the chromodomains of cpSRP43. J Biol Chem 2005;280:41465–41471. [PubMed: 16183644]

95. Apel ED, Storm DR. Functional domains of neuromodulin (GAP-43). Perspect Dev Neurobiol 1992;1:3–11. [PubMed: 1345683]

96. Hayashi N, Matsubara M, Titani K, Taniguchi H. Circular dichroism and 1H nuclear magnetic resonance studies on the solution and membrane structures of GAP-43 calmodulin-binding domain. J Biol Chem 1997;272:7639–7645. [PubMed: 9065419]

97. Kukhtina V, Kottwitz D, Strauss H, Heise B, Chebotareva N, Tsetlin V, Hucho F. Intracellular domain of nicotinic acetylcholine receptor: the importance of being unfolded. J Neurochem. 2005

98. Hayes PK, Walsby AE, Walker JE. Complete amino acid sequence of cyanobacterial gas-vesicle protein indicates a 70-residue molecule that corresponds in size to the crystallographic unit cell. Biochem J 1986;236:31–36. [PubMed: 3098234]

99. Hayes PK, Buchholz B, Walsby AE. Gas vesicles are strengthened by the outer-surface protein, GvpC. Arch Microbiol 1992;157:229–234. [PubMed: 1510555]

100. Walsby AE, Hayes PK. Gas vesicle proteins. Biochem J 1989;264:313–322. [PubMed: 2513809]

101. Walsby AE. Gas vesicles. Microbiol Rev 1994;58:94–144. [PubMed: 8177173]

102. Belenky M, Meyers R, Herzfeld J. Subunit structure of gas vesicles: a MALDI-TOF mass spectrometry study. Biophys J 2004;86:499–505. [PubMed: 14695294]

103. Hayes PK, Lazarus CM, Bees A, Walker JE, Walsby AE. The protein encoded by gvpC is a minor component of gas vesicles isolated from the cyanobacteria Anabaena flos-aquae and Microcystis sp. Mol Microbiol 1988;2:545–552. [PubMed: 3141741]

104. Kirfel J, Magin TM, Reichelt J. Keratins: a structural scaffold with emerging functions. Cell Mol Life Sci 2003;60:56–71. [PubMed: 12613658]

105. Pauling L, Corey RB. The structure of hair, muscle, and related proteins. Proc Natl Acad Sci U S A 1951;37:261–271. [PubMed: 14834149]

106. Kasapi MA, Gosline JM. Micromechanics of the equine hoof wall: optimizing crack control and material stiffness through modulation of the properties of keratin. J Exp Biol 1999;202:377–391. [PubMed: 9914146]

107. Pauling L, Corey RB. Compound helical configurations of polypeptide chains: structure of proteins of the alpha-keratin type. Nature 1953;171:59–61. [PubMed: 13025480]

108. Ikkai F, Naito S. Dynamic light scattering and circular dichroism studies on heat-induced gelation of hard-keratin protein aqueous solutions. Biomacromolecules 2002;3:482–487. [PubMed: 12005518]

109. de Lange T. Shelterin: the protein complex that shapes and safeguards human telomeres. Genes Dev 2005;19:2100–2110. [PubMed: 16166375]

110. Rodier F, Kim SH, Nijjar T, Yaswen P, Campisi J. Cancer and aging: the importance of telomeres in genome maintenance. Int J Biochem Cell Biol 2005;37:977–990. [PubMed: 15743672]

111. Schunemann D. Structure and function of the chloroplast signal recognition particle. Curr Genet 2004;44:295–304. [PubMed: 14569414]

112. Gipson IK, Argueso P. Role of mucins in the function of the corneal and conjunctival epithelia. Int Rev Cytol 2003;231:1–49. [PubMed: 14713002]

113. Carlstedt I, Sheehan JK. Macromolecular properties and polymeric structure of mucus glycoproteins. Ciba Found Symp 1984;109:157–172. [PubMed: 6083849]

114. Gonzalez-Mariscal L, Betanzos A, Nava P, Jaramillo BE. Tight junction proteins. Prog Biophys Mol Biol 2003;81:1–44. [PubMed: 12475568]

115. Utech M, Bruwer M, Nusrat A. Tight junctions and cell-cell interactions. Methods Mol Biol 2006;341:185–195. [PubMed: 16799199]

116. Ando-Akatsuka Y, Saitou M, Hirase T, Kishi M, Sakakibara A, Itoh M, Yonemura S, Furuse M, Tsukita S. Interspecies diversity of the occludin sequence: cDNA cloning of human, mouse, dog, and rat-kangaroo homologues. J Cell Biol 1996;133:43–47. [PubMed: 8601611]

117. Sakakibara A, Furuse M, Saitou M, Ando-Akatsuka Y, Tsukita S. Possible involvement of phosphorylation of occludin in tight junction formation. J Cell Biol 1997;137:1393–1401. [PubMed: 9182670]

118. Nusrat A, Brown GT, Tom J, Drake A, Bui TT, Quan C, Mrsny RJ. Multiple protein interactions involving proposed extracellular loop domains of the tight junction protein occludin. Mol Biol Cell 2005;16:1725–1734. [PubMed: 15659655]

119. Vrhovski B, Weiss AS. Biochemistry of tropoelastin. Eur J Biochem 1998;258:1–18. [PubMed: 9851686]

120. Mackay JP, Muiznieks LD, Toonkool P, Weiss AS. The hydrophobic domain 26 of human tropoelastin is unstructured in solution. J Struct Biol 2005;150:154–162. [PubMed: 15866738]

121. Minetti CA, Remeta DP. Energetics of membrane protein folding and stability. Arch Biochem Biophys. 2006

122. 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]

123. Consortium MS. Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. Nature 1999;401:921–923. [PubMed: 10553908]

124. Falk K, Rotzschke O, Rammensee HG. Cellular peptide composition governed by major histocompatibility complex class I molecules. Nature 1990;348:248–251. [PubMed: 2234092]

125. Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC. Structure of the human class I histocompatibility antigen, HLA-A2. Nature 1987;329:506–512. [PubMed: 3309677]

126. Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, Strominger JL, Wiley DC. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. Nature 1993;364:33–39. [PubMed: 8316295]

127. Jones EY, Fugger L, Strominger JL, Siebold C. MHC class II proteins and disease: a structural perspective. Nat Rev Immunol 2006;6:271–282. [PubMed: 16557259]

128. Schueler-Furman O, Altuvia Y, Margalit H. Examination of possible structural constraints of MHC-binding peptides by assessment of their native structure within their source proteins. Proteins 2001;45:47–54. [PubMed: 11536359]

129. Markiewicz Z. Structure and functions of the periplasmic space. Acta Microbiol Pol 1989;38:199–206. [PubMed: 2484738]

130. Turk V, Turk B, Turk D. Lysosomal cysteine proteases: facts and opportunities. Embo J 2001;20:4629–4633. [PubMed: 11532926]

131. Davison AJ, Benko M, Harrach B. Genetic content and evolution of adenoviruses. J Gen Virol 2003;84:2895–2908. [PubMed: 14573794]

132. Fabry CM, Rosa-Calatrava M, Conway JF, Zubieta C, Cusack S, Ruigrok RW, Schoehn G. A quasi-atomic model of human adenovirus type 5 capsid. Embo J 2005;24:1645–1654. [PubMed: 15861131]

133. Rux JJ, Burnett RM. Type-specific epitope locations revealed by X-ray crystallographic study of adenovirus type 5 hexon. Mol Ther 2000;1:18–30. [PubMed: 10933908]

134. Athappilly FK, Murali R, Rux JJ, Cai Z, Burnett RM. The refined crystal structure of hexon, the major coat protein of adenovirus type 2, at 2.9 A resolution. J Mol Biol 1994;242:430–455. [PubMed: 7932702]

135. Paranchych W, Frost LS. The physiology and biochemistry of pili. Adv Microb Physiol 1988;29:53–114. [PubMed: 2898203]

136. Craig L, Taylor RK, Pique ME, Adair BD, Arvai AS, Singh M, Lloyd SJ, Shin DS, Getzoff ED, Yeager M, Forest KT, Tainer JA. Type IV pilin structure and assembly: X-ray and EM analyses of Vibrio cholerae toxin-coregulated pilus and Pseudomonas aeruginosa PAK pilin. Mol Cell 2003;11:1139–1150. [PubMed: 12769840]

137. Lindberg F, Lund B, Johansson L, Normark S. Localization of the receptor-binding protein adhesin at the tip of the bacterial pilus. Nature 1987;328:84–87. [PubMed: 2885755]

138. Whitfield C, Roberts IS. Structure, assembly and regulation of expression of capsules in Escherichia coli. Mol Microbiol 1999;31:1307–1319. [PubMed: 10200953]

139. Drummelsmith J, Whitfield C. Gene products required for surface expression of the capsular form of the group 1 K antigen in Escherichia coli (O9a:K30). Mol Microbiol 1999;31:1321–1332. [PubMed: 10200954]

140. Bliss JM, Silver RP. Coating the surface: a model for expression of capsular polysialic acid in Escherichia coli K1. Mol Microbiol 1996;21:221–231. [PubMed: 8858578]

141. Russel M. Macromolecular assembly and secretion across the bacterial cell envelope: type II protein secretion systems. J Mol Biol 1998;279:485–499. [PubMed: 9641973]

142. Ortiz de Montellano, P. Cytochrome P450: Structure, Mechanism, and Biochemistry. 3rd ed.. New York: Kluwer Academic/Plenum Publishers; 2005.

143. Wanders RJ, Waterham HR. Biochemistry of Mammalian peroxisomes revisited. Annu Rev Biochem 2006;75:295–332. [PubMed: 16756494]

144. Jamet E, Canut H, Boudart G, Pont-Lezica RF. Cell wall proteins: a new insight through proteomics. Trends Plant Sci 2006;11:33–39. [PubMed: 16356755]

145. Boyer PD, Cross RL, Momsen W. A new concept for energy coupling in oxidative phosphorylation based on a molecular explanation of the oxygen exchange reactions. Proc Natl Acad Sci U S A 1973;70:2837–2839. [PubMed: 4517936]

146. Boyer PD. The binding change mechanism for ATP synthase--some probabilities and possibilities. Biochim Biophys Acta 1993;1140:215–250. [PubMed: 8417777]

147. Minoletti C, Santolini J, Haraux F, Pothier J, Andre F. Rebuilt 3D structure of the chloroplast f1 ATPase-tentoxin complex. Proteins 2002;49:302–320. [PubMed: 12360520]

148. Abrahams JP, Leslie AG, Lutter R, Walker JE. Structure at 2.8 A resolution of F1-ATPase from bovine heart mitochondria. Nature 1994;370:621–628. [PubMed: 8065448]

149. Deisenhofer J, Epp O, Miki K, Huber R, Michel H. X-ray structure analysis of a membrane protein complex. Electron density map at 3 A resolution and a model of the chromophores of the photosynthetic reaction center from Rhodopseudomonas viridis. J Mol Biol 1984;180:385–398. [PubMed: 6392571]

150. Fyfe PK, Hughes AV, Heathcote P, Jones MR. Proteins, chlorophylls and lipids: X-ray analysis of a three-way relationship. Trends Plant Sci 2005;10:275–282. [PubMed: 15949761]

151. Grotjohann I, Fromme P. Structure of cyanobacterial photosystem I. Photosynth Res 2005;85:51–72. [PubMed: 15977059]

152. Jordan P, Fromme P, Witt HT, Klukas O, Saenger W, Krauss N. Three-dimensional structure of cyanobacterial photosystem I at 2.5 A resolution. Nature 2001;411:909–917. [PubMed: 11418848]

153. Bloemendal H, de Jong W, Jaenicke R, Lubsen NH, Slingsby C, Tardieu A. Ageing and vision: structure, stability and function of lens crystallins. Prog Biophys Mol Biol 2004;86:407–485. [PubMed: 15302206]

154. Slingsby C, Norledge B, Simpson A, Bateman OA, Wright G, Driessen HPC, Lindley PF, Moss DS, Bax B. X-ray diffraction and structure of crystallins. Prog Ret Eye Res 1997;16:3–29.

155. Burkhard P, Stetefeld J, Strelkov SV. Coiled coils: a highly versatile protein folding motif. Trends Cell Biol 2001;11:82–88. [PubMed: 11166216]

156. Lupas AN, Gruber M. The structure of alpha-helical coiled coils. Adv Protein Chem 2005;70:37–78. [PubMed: 15837513]

157. Crick FH. The packing of alpha-helices: simple coiled-coils. Acta Cryst 1953;6:689–697.

158. Walshaw J, Woolfson DN. Extended knobs-into-holes packing in classical and complex coiled-coil assemblies. J Struct Biol 2003;144:349–361. [PubMed: 14643203]

159. Rose A, Meier I. Scaffolds, levers, rods and springs: diverse cellular functions of long coiled-coil proteins. Cell Mol Life Sci 2004;61:1996–2009. [PubMed: 15316650]

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

161. Singh A. Hitchcock-DeGregori, S. E. Dual requirement for flexibility and specificity for binding of the coiled-coil tropomyosin to its target, actin. Structure 2006;14:43–50. [PubMed: 16407064]

162. Iuchi, S.; Kudell, N. Georgetown, TX: Landes Bioscience; 2004.

163. Kadrmas JL, Beckerle MC. The LIM domain: from the cytoskeleton to the nucleus. Nat Rev Mol Cell Biol 2004;5:920–931. [PubMed: 15520811]

164. Giedroc DP, Chen X, Pennella MA, LiWang AC. Conformational heterogeneity in the C-terminal zinc fingers of human MTF-1: an NMR and zinc-binding study. J Biol Chem 2001;276:42322–42332. [PubMed: 11524427]

165. Bach I. The LIM domain: regulation by association. Mech Dev 2000;91:5–17. [PubMed: 10704826]

166. Tompa P. Intrinsically unstructured proteins evolve by repeat expansion. Bioessays 2003;25:847–855. [PubMed: 12938174]

167. Uversky VN. A protein-chameleon: conformational plasticity of alpha-synuclein, a disordered protein involved in neurodegenerative disorders. J Biomol Struct Dyn 2003;21:211–234. [PubMed: 12956606]

168. Bruce BD. Chloroplast transit peptides: structure, function and evolution. Trends Cell Biol 2000;10:440–447. [PubMed: 10998602]

169. Bruce BD. The role of lipids in plastid protein transport. Plant Mol Biol 1998;38:223–246. [PubMed: 9738969]

170. Wienk HL, Czisch M, de Kruijff B. The structural flexibility of the preferredoxin transit peptide. FEBS Lett 1999;453:318–326. [PubMed: 10405168]

171. Krimm I, Gans P, Hernandez JF, Arlaud GJ, Lancelin JM. A coil-helix instead of a helix-coil motif can be induced in a chloroplast transit peptide from Chlamydomonas reinhardtii. Eur J Biochem 1999;265:171–180. [PubMed: 10491171]

172. von Heijne G, Nishikawa K. Chloroplast transit peptides. The perfect random coil? FEBS Lett 1991;278:1–3. [PubMed: 1993469]

173. Harrison SC. Variation on an Src-like theme. Cell 2003;112:737–740. [PubMed: 12654240]

174. Pluk H, Dorey K, Superti-Furga G. Autoinhibition of c-Abl. Cell 2002;108:247–259. [PubMed: 11832214]

175. Nagar B, Hantschel O, Young MA, Scheffzek K, Veach D, Bornmann W, Clarkson B, Superti-Furga G, Kuriyan J. Structural basis for the autoinhibition of c-Abl tyrosine kinase. Cell 2003;112:859–871. [PubMed: 12654251]

176. Yuzawa S, Yokochi M, Hatanaka H, Ogura K, Kataoka M, Miura K, Mandiyan V, Schlessinger J, Inagaki F. Solution structure of Grb2 reveals extensive flexibility necessary for target recognition. J Mol Biol 2001;306:527–537. [PubMed: 11178911]

177. Rath A, Davidson AR, Deber CM. The structure of "unstructured" regions in peptides and proteins: role of the polyproline II helix in protein folding and recognition. Biopolymers 2005;80:179–185. [PubMed: 15700296]

178. Nilsson I, Whitley P, von Heijne G. The COOH-terminal ends of internal signal and signal-anchor sequences are positioned differently in the ER translocase. J Cell Biol 1994;126:1127–1132. [PubMed: 8063852]

179. Kida Y, Morimoto F, Mihara K, Sakaguchi M. Function of positive charges following signal-anchor sequences during translocation of the N-terminal domain. J Biol Chem 2006;281:1152–1158. [PubMed: 16291756]

180. Koch CA, Anderson D, Moran MF, Ellis C, Pawson T. SH2 and SH3 domains: elements that control interactions of cytoplasmic signaling proteins. Science 1991;252:668–674. [PubMed: 1708916]

181. Pawson T, Gish GD, Nash P. SH2 domains, interaction modules and cellular wiring. Trends Cell Biol 2001;11:504–511. [PubMed: 11719057]

182. Machida K, Mayer BJ. The SH2 domain: versatile signaling module and pharmaceutical target. Biochim Biophys Acta 2005;1747:1–25. [PubMed: 15680235]

183. Stapleton P, Weith A, Urbanek P, Kozmik Z, Busslinger M. Chromosomal localization of seven PAX genes and cloning of a novel family member, PAX-9. Nat Genet 1993;3:292–298. [PubMed: 7981748]

184. Gruss P, Walther C. Pax in development. Cell 1992;69:719–722. [PubMed: 1591773]

185. Chalepakis G, Tremblay P, Gruss P. Pax genes, mutants and molecular function. J Cell Sci Suppl 1992;16:61–67. [PubMed: 1363663]

186. Epstein J, Cai J, Glaser T, Jepeal L, Maas R. Identification of a Pax paired domain recognition sequence and evidence for DNA-dependent conformational changes. J Biol Chem 1994;269:8355–8361. [PubMed: 8132558]

187. Cooper NR. The classical complement pathway: activation and regulation of the first complement component. Adv Immunol 1985;37:151–216. [PubMed: 3890478]

188. Schumaker VN, Zavodszky P, Poon PH. Activation of the first component of complement. Annu Rev Immunol 1987;5:21–42. [PubMed: 3036181]

189. Bersch B, Hernandez JF, Marion D, Arlaud GJ. Solution structure of the epidermal growth factor (EGF)-like module of human complement protease C1r, an atypical member of the EGF family. Biochemistry 1998;37:1204–1214. [PubMed: 9477945]

190. Blumberg PM. Complexities of the protein kinase C pathway. Mol Carcinog 1991;4:339–344. [PubMed: 1910477]

191. Nishizuka Y. Protein kinase C and lipid signaling for sustained cellular responses. Faseb J 1995;9:484–496. [PubMed: 7737456]

192. Newton AC. Protein kinase C: structure, function, and regulation. J Biol Chem 1995;270:28495–28498. [PubMed: 7499357]

193. Kishimoto A, Kajikawa N, Shiota M, Nishizuka Y. Proteolytic activation of calcium-activated, phospholipid-dependent protein kinase by calcium-dependent neutral protease. J Biol Chem 1983;258:1156–1164. [PubMed: 6296071]

194. Wender PA, Irie K, Miller BL. Identification, activity, and structural studies of peptides incorporating the phorbol ester-binding domain of protein kinase C. Proc Natl Acad Sci U S A 1995;92:239–243. [PubMed: 7816824]

195. O'Keeffe AH, Green JL, Grainger M, Holder AA. A novel Sushi domain-containing protein of Plasmodium falciparum. Mol Biochem Parasitol 2005;140:61–68. [PubMed: 15694487]

196. Wiles AP, Shaw G, Bright J, Perczel A, Campbell ID, Barlow PN. NMR studies of a viral protein that mimics the regulators of complement activation. J Mol Biol 1997;272:253–265. [PubMed: 9299352]

197. Kirkitadze MD, Barlow PN. Structure and flexibility of the multiple domain proteins that regulate complement activation. Immunol Rev 2001;180:146–161. [PubMed: 11414356]

198. Lorenzen I, Dingley AJ, Jacques Y, Grotzinger J. The structure of the interleukin-15 alpha receptor and its implications for ligand binding. J Biol Chem 2006;281:6642–6647. [PubMed: 16377614]

199. Jeanmougin F, Wurtz JM, Le Douarin B, Chambon P, Losson R. The bromodomain revisited. Trends Biochem Sci 1997;22:151–153. [PubMed: 9175470]

200. Zeng L, Zhou MM. Bromodomain: an acetyl-lysine binding domain. FEBS Lett 2002;513:124–128. [PubMed: 11911891]

201. Dhalluin C, Carlson JE, Zeng L, He C, Aggarwal AK, Zhou MM. Structure and ligand of a histone acetyltransferase bromodomain. Nature 1999;399:491–496. [PubMed: 10365964]

202. Trexler M, Patthy L. Residues Cys-1 and Cys-79 are not essential for refolding of reduced-denatured kringle 4 fragment of human plasminogen. Biochim Biophys Acta 1984;787:275–280. [PubMed: 6329306]

203. Maderegger B, Bermel W, Hrzenjak A, Kostner GM, Sterk H. Solution structure of human apolipoprotein(a) kringle IV type 6. Biochemistry 2002;41:660–668. [PubMed: 11781107]

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

205. Pommie C, Levadoux S, Sabatier R, Lefranc G, Lefranc MP. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. J Mol Recognit 2004;17:17–32. [PubMed: 14872534]

206. Zalkin H, Smith JL. Enzymes utilizing glutamine as an amide donor. Adv Enzymol Relat Areas Mol Biol 1998;72:87–144. [PubMed: 9559052]

207. Li H, Ryan TJ, Chave KJ, Van Roey P. Three-dimensional structure of human gamma - glutamyl hydrolase. A class I glatamine amidotransferase adapted for a complex substate. J Biol Chem 2002;277:24522–24529. [PubMed: 11953431]

208. Edeling MA, Guddat LW, Fabianek RA, Thony-Meyer L, Martin JL. Structure of CcmG/DsbE at 1.14 A resolution: high-fidelity reducing activity in an indiscriminately oxidizing environment. Structure 2002;10:973–979. [PubMed: 12121652]

209. Katti SK, LeMaster DM, Eklund H. Crystal structure of thioredoxin from Escherichia coli at 1.68 A resolution. J Mol Biol 1990;212:167–184. [PubMed: 2181145]

210. Martin JL, Bardwell JC, Kuriyan J. Crystal structure of the DsbA protein required for disulphide bond formation in vivo. Nature 1993;365:464–468. [PubMed: 8413591]

211. McCarthy AA, Haebel PW, Torronen A, Rybin V, Baker EN, Metcalf P. Crystal structure of the protein disulfide bond isomerase, DsbC, from Escherichia coli. Nat Struct Biol 2000;7:196–199. [PubMed: 10700276]

212. Capitani G, Rossmann R, Sargent DF, Grutter MG, Richmond TJ, Hennecke H. Structure of the soluble domain of a membrane-anchored thioredoxin-like protein from Bradyrhizobium japonicum reveals unusual properties. J Mol Biol 2001;311:1037–1048. [PubMed: 11531338]

213. Wiener MC. TonB-dependent outer membrane transport: going for Baroque? Curr Opin Struct Biol 2005;15:394–400. [PubMed: 16039843]

214. Xue F, Cooley L. kelch encodes a component of intercellular bridges in Drosophila egg chambers. Cell 1993;72:681–693. [PubMed: 8453663]

215. Bork P, Doolittle RF. Drosophila kelch motif is derived from a common enzyme fold. J Mol Biol 1994;236:1277–1282. [PubMed: 8126718]

216. Adams J, Kelso R, Cooley L. The kelch repeat superfamily of proteins: propellers of cell function. Trends Cell Biol 2000;10:17–24. [PubMed: 10603472]

217. Mosavi LK, Cammett TJ, Desrosiers DC, Peng ZY. The ankyrin repeat as molecular architecture for protein recognition. Protein Sci 2004;13:1435–1448. [PubMed: 15152081]

218. Ignoul S, Eggermont J. CBS domains: structure, function, and pathology in human proteins. Am J Physiol Cell Physiol 2005;289:C1369–C1378. [PubMed: 16275737]

219. Gerke V, Creutz CE, Moss SE. Annexins: linking Ca2+ signalling to membrane dynamics. Nat Rev Mol Cell Biol 2005;6:449–461. [PubMed: 15928709]

220. Bannert N, Kurth R. Retroelements and the human genome: new perspectives on an old relation. Proc Natl Acad Sci U S A 2004;101:14572–14579. [PubMed: 15310846]

221. Medstrand P, van de Lagemaat LN, Mager DL. Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res 2002;12:1483–1495. [PubMed: 12368240]

222. 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

223. Nair M, Hinds MG, Coley AM, Hodder AN, Foley M, Anders RF, Norton RS. Structure of domain III of the blood-stage malaria vaccine candidate, Plasmodium falciparum apical membrane antigen 1 (AMA1). J Mol Biol 2002;322:741–753. [PubMed: 12270711]

224. York IA, Roop C, Andrews DW, Riddell SR, Graham FL, Johnson DC. A cytosolic herpes simplex virus protein inhibits antigen presentation to CD8+ T lymphocytes. Cell 1994;77:525–535. [PubMed: 8187174]

225. Beinert D, Neumann L, Uebel S, Tampe R. Structure of the viral TAP-inhibitor ICP47 induced by membrane association. Biochemistry 1997;36:4694–4700. [PubMed: 9109681]

226. Pfander R, Neumann L, Zweckstetter M, Seger C, Holak TA, Tampe R. Structure of the active domain of the herpes simplex virus protein ICP47 in water/sodium dodecyl sulfate solution determined by nuclear magnetic resonance spectroscopy. Biochemistry 1999;38:13692–13698. [PubMed: 10521276]

227. McCutchan TF, Kissinger JC, Touray MG, Rogers MJ, Li J, Sullivan M, Braga EM, Krettli AU, Miller LH. Comparison of circumsporozoite proteins from avian and mammalian malarias: biological and phylogenetic implications. Proc Natl Acad Sci U S A 1996;93:11889–11894. [PubMed: 8876233]

228. Hamilton AJ, Suhrbier A, Nicholas J, Sinden RE. Immunoelectron microscopic localization of circumsporozoite antigen in the differentiating exoerythrocytic trophozoite of Plasmodium berghei. Cell Biol Int Rep 1988;12:123–129. [PubMed: 3293804]

229. Dyson HJ, Satterthwait AC, Lerner RA, Wright PE. Conformational preferences of synthetic peptides derived from the immunodominant site of the circumsporozoite protein of Plasmodium falciparum by 1H NMR. Biochemistry 1990;29:7828–7837. [PubMed: 2261440]

230. Lareau LF, Green RE, Bhatnagar RS, Brenner SE. The evolving roles of alternative splicing. Curr Opin Struct Biol 2004;14:273–282. [PubMed: 15193306]

231. Ast G. How did alternative splicing evolve? Nat Rev Genet 2004;5:773–782. [PubMed: 15510168]

232. Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, Legall T, Obradovic Z, Dunker AK. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc Natl Acad Sci U S A 2006;103:8390–8395. [PubMed: 16717195]

233. Housman D. Human DNA polymorphism. N Engl J Med 1995;332:318–320. [PubMed: 7816069]

234. Weisgraber KH. Apolipoprotein E: structure-function relationships. Adv Protein Chem 1994;45:249–302. [PubMed: 8154371]

235. Acharya P, Segall ML, Zaiou M, Morrow J, Weisgraber KH, Phillips MC, Lund-Katz S, Snow J. Comparison of the stabilities and unfolding pathways of human apolipoprotein E isoforms by differential scanning calorimetry and circular dichroism. Biochim Biophys Acta 2002;1584:9–19. [PubMed: 12213488]

236. Ikura M, Ames JB. Genetic polymorphism and protein conformational plasticity in the calmodulin superfamily: two ways to promote multifunctionality. Proc Natl Acad Sci U S A 2006;103:1159–1164. [PubMed: 16432210]

237. Copley RR, Schultz J, Ponting CP, Bork P. Protein families in multicellular organisms. Curr Opin Struct Biol 1999;9:408–415. [PubMed: 10361098]

238. Carafoli E, Santella L, Branca D, Brini M. Generation, control, and processing of cellular calcium signals. Crit Rev Biochem Mol Biol 2001;36:107–260. [PubMed: 11370791]

239. Kawasaki H, Nakayama S, Kretsinger RH. Classification and evolution of EF-hand proteins. Biometals 1998;11:277–295. [PubMed: 10191494]

240. Radivojac P, Vucetic S, O'Connor TR, Uversky VN, Obradovic Z, Dunker AK. Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. Proteins 2006;63:398–410. [PubMed: 16493654]

241. Rabbitts TH, Stocks MR. Chromosomal translocation products engender new intracellular therapeutic technologies. Nat Med 2003;9:383–386. [PubMed: 12669051]

242. Delattre O, Zucman J, Plougastel B, Desmaze C, Melot T, Peter M, Kovar H, Joubert I, de Jong P, Rouleau G, et al. Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours. Nature 1992;359:162–165. [PubMed: 1522903]

243. Karim FD, Urness LD, Thummel CS, Klemsz MJ, McKercher SR, Celada A, Van Beveren C, Maki RA, Gunther CV, Nye JA, et al. The ETS-domain: a new DNA-binding motif that recognizes a purine-rich core DNA sequence. Genes Dev 1990;4:1451–1453. [PubMed: 2253872]

244. Uren A, Tcherkasskaya O, Toretsky JA. Recombinant EWS-FLI1 oncoprotein activates transcription. Biochemistry 2004;43:13579–13589. [PubMed: 15491164]

245. Brown LY, Brown SA. Alanine tracts: the expanding story of human illness and trinucleotide repeats. Trends Genet 2004;20:51–58. [PubMed: 14698619]

246. Albrecht A, Mundlos S. The other trinucleotide repeat: polyalanine expansion disorders. Curr Opin Genet Dev 2005;15:285–293. [PubMed: 15917204]

247. Cummings CJ, Zoghbi HY. Fourteen and counting: unraveling trinucleotide repeat diseases. Hum Mol Genet 2000;9:909–916. [PubMed: 10767314]

248. La Spada AR, Taylor JP. Polyglutamines placed into context. Neuron 2003;38:681–684. [PubMed: 12797953]

249. Perutz MF. Glutamine repeats and neurodegenerative diseases: molecular aspects. Trends Biochem Sci 1999;24:58–63. [PubMed: 10098399]

250. Pandey N, Mittal U, Srivastava AK, Mukerji M. SMARCA2 and THAP11: potential candidates for polyglutamine disorders as evidenced from polymorphism and protein-folding simulation studies. J Hum Genet 2004;49:596–602. [PubMed: 15368101]

251. Chen YW. Local protein unfolding and pathogenesis of polyglutamine-expansion diseases. Proteins 2003;51:68–73. [PubMed: 12596264]

252. Chen S, Berthelier V, Hamilton JB, O'Nuallain B, Wetzel R. Amyloid-like features of polyglutamine aggregates and their assembly kinetics. Biochemistry 2002;41:7391–7399. [PubMed: 12044172]

253. Wang X, Vitalis A, Wyczalkowski MA, Pappu RV. Characterizing the conformational ensemble of monomeric polyglutamine. Proteins 2006;63:297–311. [PubMed: 16299774]

254. Masino L, Kelly G, Leonard K, Trottier Y, Pastore A. Solution structure of polyglutamine tracts in GST-polyglutamine fusion proteins. FEBS Lett 2002;513:267–272. [PubMed: 11904162]

255. Murray-Zmijewski F, Lane DP, Bourdon JC. p53/p63/p73 isoforms: an orchestra of isoforms to harmonise cell differentiation and response to stress. Cell Death Differ 2006;13:962–972. [PubMed: 16601753]

Table 1

Top 20 of cellular components keywords strongly correlated with predicted disorder

Keywords	Number of proteins	Number of families	Average sequence length	Z-score	P-value
Nuclear protein	13726	2670	504.3	38.36	1
Chromosomal protein	894	227	225.74	16.14	1
Nucleosome core	601	134	106.53	14.7	1
Spliceosome	392	99	501.34	11.93	1
Microtubule	953	122	643.24	9.65	1
Cytoskeleton	1361	186	657.23	8.61	1
Centromere	131	69	398.69	8.01	1
Flagellum	801	157	328.96	7.7	1
Golgi stack	1554	318	490.17	7.18	1
Mitochondrion	6946	1069	322.73	6.89	1
Proteasome	444	40	336.21	5.68	1
Signal recognition particle	153	18	374.71	5.35	1
Synaptosome	77	13	568.87	5.05	1
Gas vesicle	73	18	169.01	4.89	1
Keratin	210	11	317.43	4.43	1
Telomere	85	48	829.79	4.19	1
Chloroplast	5859	646	300.09	4.18	1
Surface film	29	3	234.66	3.96	1
Tight junction	138	19	529.73	3.75	1
Extracellular matrix	594	63	920.42	3.47	1

Table 2

Top 20 of cellular components keywords strongly correlated with predicted order

keywords	number of proteins	number of families	average sequence length	z-score	p-value
Inner membrane	6321	851	345.85	−20.64	0
Membrane	37216	6187	430.59	−14.41	0
MHC I	218	2	268.52	−11.68	0
Periplasmic	1248	311	340.51	−11.18	0
Lysosome	291	69	467.82	−8.18	0
Hexon protein	21	1	784.59	−7.94	0
Fimbria	318	93	277.74	−7.5	0
Bacterial capsule	123	48	332.68	−5.7	0
Microsome	594	41	434.59	−5.37	0
Peroxisome	484	107	475.86	−5.24	0
Cell wall	2689	337	405.14	−4.87	0
S-layer	41	20	966.63	−4.82	0
Outer membrane	1094	258	428.42	−4.44	0
Apoplast	98	9	266.43	−3.68	0
MHC II	78	3	254.24	−3.61	0
Envelope protein	635	72	641.67	−3.53	0
CF_1	814	25	299.91	−3.15	0
Reaction center	127	10	70.52	−2.98	0
Antenna complex	68	12	67.35	−2.95	0
Eye lens protein	204	17	213.51	−2.27	0

Table 3

Top 20 of domain keywords strongly correlated with predicted disorder

Keywords	Number of proteins	Number of families	Average sequence length	Z-score	P-value
Coiled coil	4172	991	711.94	27.97	1
Zinc-finger	5572	706	604.67	21.33	1
Repeat	16015	2258	691.73	19.41	1
Transit peptide	3523	717	390.65	16.24	1
Homeobox	1092	51	356.97	10.55	1
SH3 domain	496	83	841.33	7.81	1
Leucine-rich repeat	598	82	719.41	7.1	1
SH3-binding	157	42	710.46	6.37	1
Signal-anchor	1466	208	455.7	6.3	1
SH2 domain	291	32	612.48	6.15	1
Lim domain	216	16	447.75	4.96	1
Paired box	35	1	430.09	4.91	1
EFG-like domain	669	66	1020.51	4.43	1
Collagen	250	8	767.98	3.7	1
Phorbol-ester binding	156	15	917.27	2.99	1
Sushi	171	18	832.48	2.81	1
Immunoglobulin domain	1519	167	469.66	2.57	1
Immunoglobulin c region	71	2	284.49	2.5	1
Bromodomain	76	20	1252.38	2.37	1
Kringle	72	4	699.38	2.33	1

Table 4

All (9) domain keywords strongly correlated with predicted order

Keywords	Number of proteins	Number of families	Average sequence length	Z-score	P-value
Transmembrane	30651	5120	434.12	−17.13	0
Immunoglobulin V region	293	2	115.8	−14.7	0
Glutamine amidotransferase	1039	16	432.35	−13.71	0
Redox-active center	987	73	270.72	−5.68	0
Tonb box	74	15	755.32	−4.91	0
Kelch repeat	138	27	747.65	−3.81	0
Ank repeat	582	97	700.13	−3.77	0
Cbs domain	209	28	495.6	−2.99	0
Annexin	66	2	362.74	−2.29	0

Table 5

All (2) technical terms keywords strongly correlated with predicted disorder

Keywords	Number of proteins	Number of families	Average sequence length	Z-score	P-value
Pharmaceutical	47	36	298.09	2.98	1
Erv	105	12	436.31	2.09	1

Table 6

All (7) technical terms keywords strongly correlated with predicted order

Keywords	Number of proteins	Number of families	Average sequence length	Z-score	P-value
Multifunctional enzyme	1886	206	671.67	−17.48	0
Complete proteome	101569	16100	333.1	−16.01	0
Hypothetical protein	22983	11683	275.6	−15.53	0
3D-structure	8305	3846	403.54	−13.73	0
Plasmid	3037	1674	309.67	−12.79	0
Allosteric enzyme	492	57	585.04	−7.29	0
Direct protein sequencing	16187	6544	290.21	−6.39	0

Table 7

All (3) developmental stage keywords strongly correlated with predicted disorder

Keywords	Number of proteins	Number of families	Average sequence length	Z-score	P-value
Merozoite	29	6	701.1	3.09	1
Early protein	894	278	296.64	3.06	1
Sporozoite	25	5	412.32	2.11	1

Table 8

All (6) coding sequence diversity keywords strongly correlated with predicted disorder

Keywords	Number of proteins	Number of families	Average sequence length	Z-score	P-value
Alternative splicing	8041	2824	687.17	25.34	1
Polymorphism	4259	2020	621.15	13.39	1
Chromosomal translocation	188	138	741.01	7.2	1
Triplet repeat expansion	23	21	1024.3	2.84	1
Alternative promoter usage	56	30	834.51	2.1	0.99
Alternative initiation	292	149	538.83	2.35	0.98