Intrinsically unstructured proteins

Peter Tompa

The recent suggestion that the classical structure–function paradigm should be extended to proteins and protein domains whose native and functional state is intrinsically unstructured has received a great deal of support. There is ample evidence that the unstructured state, common to all living organisms, is essential for basic cellular functions; thus it deserves to be recognized as a separate functional and structural category within the protein kingdom. In this review, recent findings are surveyed to illustrate that this novel but rapidly advancing field has reached a point where these proteins can be comprehensively classified on the basis of structure and function.

Published online: 5 September 2002

Our view of proteins is deeply ingrained with the conviction that a well-defined three-dimensional structure is a prerequisite of their function. Thus, structure classification schemes such as SCOP (for Structural Classification of Proteins)[1], which are based on more than 18 000 three-dimensional structures deposited in the Protein Data Bank (PDB), do not include intrinsically unstructured proteins (IUPs). There are many proteins and protein domains, such as the non-A β component of AD amyloid precursor (NACP) [2], $p21^{Waf1/Cip1/Sdi1}$ [3], anti-sigma-28 factor (FlgM) [4], calpastatin [5], microtubule-associated protein tau [6] and the transactivator domain (TAD) of transcription factors [7], however, that lack a folded structure but display a highly flexible, random-coil-like conformation under physiological conditions. The accumulation of evidence that this is the native and functional state for these and other proteins lead Wright and Dyson to suggest that the classical protein structure–function paradigm needs to be re-assessed [8].

In the past few years, a large number of papers on proteins denoted as 'natively denatured/unfolded' or 'intrinsically unstructured/disordered' have appeared and it has become apparent that this phenomenon is quite general [9–11]. The number of full-length proteins and domains that have been identified as IUPs already exceeds 100 [10,11]. Moreover, the functional importance of the unstructured state is underlined by the fact that most of these proteins have basic regulatory roles in key cellular processes. The focus of recent reviews has been on the possible structural [11] and functional [10] classification of these IUPs.

In this paper, an illustrative set of 21 unstructured proteins and segments has been selected to advance and extend these points. IUPs are suggested to segregate into five general functional categories and the benefits of structural disorder in these are elucidated. Further, it is shown that IUPs are not fully disordered, but exhibit order at the level of primary and secondary/tertiary structure that correlates with their specific functions. Taken together, these examples and concepts demonstrate that this novel class of proteins has 'come of age' and that their recognition, characterization and classification will predictably be one of the most exciting undertakings of structural biology in the years ahead.

Peter Tompa
Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, 1518 Budapest, PO Box 7, Hungary.
e-mail: tompa@enzim.hu

Essential regulatory roles

Recent lists of IUPs are based on experimental evidence obtained by a variety of techniques [2,8–11]. These proteins are characterized by an almost complete lack of folded structure and an extended conformation with high intramolecular flexibility and little secondary structure. These unusual features are apparent with most physicochemical techniques, which makes their detection straightforward (Box 1 and references therein). It should be kept in mind, however, that the deviation from the norm of any physical parameter could have several interpretations and might be misleading in indicating the lack of structure. Verification of an IUP thus needs to be based on several techniques, preferably using different physical principles. The 21 full-length proteins and segments/domains described in Table 1 were confirmed to be unstructured by several techniques.

In a recent review, Dunker *et al.* [10] discussed that the functions carried out by IUPs are manifold, but that certain unifying themes emerge. Unstructured proteins/regions have no enzymatic activity as the proper spatial organization of active site residues requires a rigid fold they cannot provide. Furthermore, their functions are invariably linked to their structural disorder and can be classified into 28 distinct categories. Here, it is suggested that they actually fall into five broad functional classes based on their mode of action (Table 1). The first class is that of entropic chains, with functions that directly stem from disorder and thus fall outside the realm of folded proteins [9,10]. The IUPs in the other four categories function via molecular recognition; as noted [8–10], the target/partner with which they bind with high specificity can be DNA, RNA, another protein, or a range of small ligands. On binding, IUPs can alter the action of their partner in a variety of ways (Table 1). Some of them, the effectors, modify the activity of a single partner protein or assembled proteins; so far, only inhibitors appear to belong to this second class. The third class is that of scavengers, which store and/or neutralize small ligands. The fourth class are the assemblers, which assemble, stabilize and regulate large multiprotein complexes such as the ribosome, cytoskeleton, transcription preinitiation complex, chromatin, and even the extracellular matrix.

 0968-0004/02/$ – see front matter PII: S0968-0004(02)02169-2

Box 1. Methods for detecting and characterizing IUPs

The most direct way of identifying the lack of structure is to look for residues with missing backbone coordinates in three-dimensional structures determined by X-ray crystallography [a]. To study proteins disordered along their entire length, heteronuclear multidimensional nuclear magnetic resonance is the most versatile and powerful tool. The high flexibility of the chain makes the chemical shift dispersion diminish and disorder easily recognizable. Furthermore, recent advances allow complete assignment of resonances in the unfolded state and enable the sequence-specific identification of residual structure [b]. In addition, spin-relaxation and nuclear Overhauser effect measurements provide insight into internal molecular dynamics in the unstructured state.

The other technique that is used most often is far-UV circular dichroism (CD) spectroscopy, which detects the amount (or lack) of secondary structure. The ellipticity spectrum of IUPs has a large negative peak at around 200 nm and a value close to zero at 220 nm, distinct enough from that of ordered conformations to allow identification of partially or fully unstructured proteins. Moreover, residual structure on the order of 10–20% can be determined.

Further spectroscopic techniques, often combined with denaturing conditions, supplement these data. Fourier transform infrared spectroscopy also provides information on the secondary structure, whereas near-UV CD, UV spectroscopy and fluorimetry report on the environment of aromatic residues (i.e. the lack of a tightly packed hydrophobic core). An important recent development is the characterization of secondary structure by Raman optical activity measurements [c]. This technique provides a snapshot of conformations and permits a distinction between dynamic and static disorder.

Unfolded conformations can also be detected and characterized by hydrodynamic techniques. Gel filtration (size-exclusion chromatography), small-angle X-ray scattering, sedimentation analysis and dynamic light scattering provide information on hydrodynamic parameters, such as the Stokes radius (R_S, the apparent radius of a sphere with identical hydrodynamic behaviour to IUP) and the radius of gyration (R_G). In these experiments IUPs resemble the denatured states of globular proteins.

Some independent techniques can add an extra dimension to these studies. Differential scanning calorimetry unveils the absence of cooperative folding transitions characteristic of a well-defined tertiary fold. The extreme proteolytic sensitivity of IUPs results from proteases cleaving substrates at sites that are sterically accessible and flexible enough to make productive contacts with the enzyme. Globular domains thus show significant resistance; IUPs are usually very sensitive along their entire length. Another method for detecting IUPs relies on their heat-stability. As IUPs have no hydrophobic core, they do not lose solubility at elevated temperatures; in fact, they are often purified via an intermittent heat-treatment step. A further method is sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE), which is routinely used to assess the M_w of proteins. Because of their unusual amino acid composition, IUPs bind less SDS than usual and their apparent M_w is often 1.2–1.8 times higher than the real one calculated from sequence data or measured by mass spectrometry.

References

a Dunker, A.K. *et al.* (2001) Intrinsically disordered protein. *J. Mol. Graph. Model.* 19, 26–59

b Dyson, H.J. and Wright, P.E. (2001) Nuclear magnetic resonance methods for elucidation of structure and dynamics in disordered states. *Methods Enzymol.* 339, 258–270

c Syme, C.D. *et al.* (2002) A Raman optical activity study of rheomorphism in caseins, synucleins and tau. *Eur. J. Biochem.* 269, 148–156

A special case of molecular recognition is exhibited by the fifth class (i.e. display sites), which mediate regulatory posttranslational modification, such as phosphorylation or limited proteolysis. As noted [10], such modifications often require intrinsic disorder, which enables transient but specific interaction with the active site of the modifying enzyme (Table 1).

An important aspect of this classification scheme is that the categories represent various functional modes but are in no way exclusive. Two examples in Table 1 illustrate this point. With microtubule-associated protein 2 (MAP2), the projection domain functions as an entropic bristle, whereas the microtubule-binding domain serves both as an assembler for microtubules and as a display site for phosphorylation. The situation is similar for the TAD of cAMP response element binding protein (CREB), which binds TATA-box-associated factors (TAFs) in the assembly of the transcription preinitiation complex and is also subtly regulated by phosphorylation.

In general, the IUPs in all five functional classes are involved in the regulation of key cellular processes, such as transcription, translation, signal transduction and the cell cycle. Their commonness, however, cannot be assessed from the importance of this limited number of cases. Although the number of proteins so far identified to be unstructured is just above 100 [2,10–12], direct and indirect considerations nevertheless suggest that they are quite common. Kim *et al.* [13], for example, found a large fraction of the proteome (20% in Jurkat cells and 70% in human serum) to be heat-resistant (Box 1). Neural network predictors, developed to recognize sequences that correspond to structurally disordered regions, have shown that 35–51% of eucaryotic proteins have at least one long (>50 residues) disordered region [10,14], and 11% of proteins in Swiss-Prot and 6–17% of proteins encoded by various genomes are probably fully disordered [9,14]. A similar conclusion was reached based on the observation that the unfolded state is ensured by low mean hydrophobicity and high net charge [11,12]. By selecting 'natively unfolded' proteins using these criteria, they were found to represent a substantial fraction of Swiss-Prot [12]. Pertinent to these estimations is that unfolded proteins or regions often show low compositional complexity; as much as 25% of all amino acids in Swiss-Prot are found in such regions and 34% of all proteins have at least one such segment [15]. These regions sometimes correspond to repetitive structural units in fibrillar proteins, but most often they are disordered [16]. A final note on generality is that the average genomially encoded protein is significantly different in terms of size and amino-acid composition from folded proteins in the PDB [17]. This difference would seem to indicate that the PDB does not represent a random selection of proteins and thus folded proteins are not representative of the proteome.

The benefits of being unstructured

Structural disorder is essential for IUPs as their various functions stem either directly from this state or from some local folding/ordering in molecular recognition. The functional benefits of this unconventional behaviour have been elucidated in several recent reviews [8–10,12,18].

The uncontested advantage of the lack of a folded structure is realized in entropic chain functions (Table 1), which depend directly on the disordered state and are thus out of reach of globular proteins [10]. Often, globular domains are connected by flexible linkers/spacers in multidomain proteins; these regions regulate distance and enable unprecedented freedom in orientational search. Other unique functions

Table 1 Intrinsically unstructured proteins (IUPs) and domains[a]

IUP (protein/domain)	Target/partner	Function/action
Entropic chains		
Microtubule-associated protein 2 (MAP2) projection domain[b]	Not applicable[c]	Entropic bristle (spacing in microtubule architecture)
Titin PEVK domain[b]	Not applicable[c]	Entropic spring (passive contractile force in muscle)
SNAP-25 linker region[b]	Not applicable[c]	Flexible spacer/linker of binding domains
Effectors		
Calpastatin	Ca^{2+}-activated protease (calpain)	Inhibitor of calpain in Ca^{2+} signalling
p21/27	Cyclin-dependent kinases	Kip/Cip class inhibitors in cell cycle regulation
4EBP1, 2, 3	Eucaryotic translation initiation factor (eIF4E)	Inhibitor of translation initiation
Securin[d]	Separase	Inhibitor of chromosome separation before anaphase in mitosis
FlgM	Sigma 28 transcription factor	Inhibitor of flagellin-specific gene expression in bacteria
Stathmin	Tubulin dimers	Microtubule disassembly, catastrophe
Scavengers		
Thymosins (proTα)	Zn^{2+}, histone	Not reported
Caseins	Calcium phosphate	Nanocluster formation, inhibition of precipitation in milk
Salivary proline-rich protein (PRP)	Tannin	Binding/neutralization of polyphenolic plant compounds
Desiccation stress protein (Dsp) 16	Water	Retention of water to prevent desiccation of plants
Assemblers		
MAP2 microtubule-binding domain[b]	Tubulin dimers	Microtubule polymerization, bundling
Caldesmon	Ca^{2+}/calmodulin, F-actin, myosin, tropomyosin	Actin polymerization, bundling
Bob1	Oct1 transcription factor, Igκ promoter, $TAF_{II}105$	B-cell-specific expression of immunoglobulin genes
λ phage N protein	mRNA, NusA, RNA Pol II	Translation antitermination
SIBLING proteins	Integrin, complement factor H, CD44, fibronectin	Assembly of bone extracellular matrix
Fibronectin receptor (MSCRAMM) D_1-D_4[b]	Fibronectin	Adherence to extracellular matrix of host in bacterial invasion
CREB transactivator domain (TAD)[b]	TATA-box-associated factors (TAFs), CREB-binding protein	Assembly of transcription preinitiation complex
Display sites		
CREB TAD[b]	Protein kinases (e.g. PKA, CaMKIV)[e]	Regulation by phosphorylation
MAP2 microtubule-binding domain[b]	Protein kinases (e.g. PKA, MARK)[e]	Regulation by phosphorylation
Bcl-2 antiapoptotic protein (24–93)[b]	Proteases (e.g. caspase)[e]	*In vivo* proteolysis site

A illustrative set of IUPs and unstructured domains, for which multiple evidence for the lack of structure exists, have been selected from recent reviews on IUPs (see text for details). The target/partner (if any) and function of the protein is shown, and a broad functional classification of entropic chains, effectors, scavengers, assemblers and display sites is suggested.

[a]Abbreviations: PEVK, region rich in Pro, Glu, Val and Lys; 4EPB, eukaryotic translation initiation factor 4E binding protein; CREB, cAMP response element binding protein; PKA, cAMP-dependent protein kinase; CaMKIV, Ca^{2+}/calmodulin-dependent protein kinase IV; MARK, microtubule-affinity regulating kinase.

[b]Domain/region.

[c]Target/partner (if any) is not relevant with respect to function.

[d]Unfoldedness (K. Boatright, The Burnham Institute, La Jolla, pers. commun.).

[e]For display sites, the relevant partner is the modifying enzyme.

carried out by such proteins/domains are entropic springs, clocks and bristles, which act via physical forces generated in, or exerted against, global conformational changes. For example, the projection domain of MAP2 exerts a long-range, entropic repulsion on objects approaching microtubules and probably provides a physical mechanism to regulate spacing in the cytoskeletal architecture [19].

The major functional asset of IUPs, which function by molecular recognition, is related to their significant disorder–order transition (i.e. induced local folding upon binding to their target). As first recognized in site-specific DNA binding [20] and treated in depth in recent reviews [18,21], this transition is accompanied by a large decrease in conformational entropy, which uncouples binding strength from specificity and renders highly specific interactions reversible. This is fundamental in regulation, as demonstrated by comparing the interactions of a constitutive (c-Myb) transcription factor and an inducible (CREB) transcription factor with the KIX domain of CREB-binding protein (CBP). It was clearly demonstrated that inducibility stems from the disorder and the accompanying negative entropy change upon binding of the TAD of CREB (Table 1), as opposed to the binding of c-Myb, which is structured in isolation [22]. Disorder–order transition has been directly demonstrated also for FlgM [4], cAMP-dependent protein kinase (PKA) inhibitor [23], eukaryotic translation initiation factor 4E binding protein 1 (4EBP1) [24] and stathmin [25], and this phenomenon could apply to all IUPs involved in molecular recognition [8–10,12,18]. Reversibility is of utmost importance for regulation, which is also underlined by the fact that the action of IUPs is often modulated by phosphorylation, as has been reported

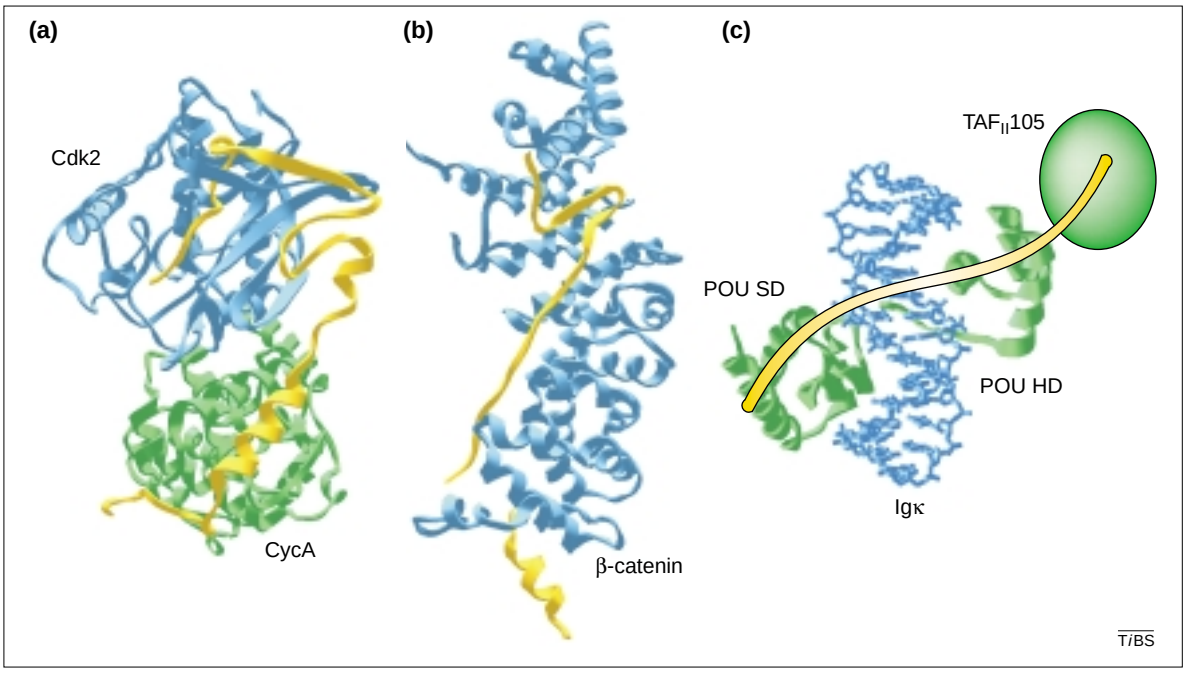

Fig. 1. Intrinsically unstructured proteins (IUPs) contact their target over a large surface area. In some cases, the structure of an IUP bound to its target is known from X-ray crystallography or can be inferred from biochemical data on its interaction with its partner(s). The three examples (IUPs are shown in yellow) demonstrate that IUPs largely preserve their extended, open conformation when they bind to their target. This feature allows multiple contact points over a disproportionately large surface area for a protein of a given size. (a) p27^{Kip1} complexed with cyclin-dependent kinase 2 (Cdk2) and cyclin A [CycA; Protein Data Bank (PDB) number 1JSU]. (b) The transactivator domain CBD of Tcf3 bound to β-catenin (PDB number 1G3J; note that part of the β-catenin has been cleaved off for a clear view of the CBD). (c) Bob1 transcriptional coactivator in contact with TAF$_{II}$105 as well as with the POU specific domain (SD) and homeodomain (HD) of Oct1 and the Igκ promoter; this structure has been drawn from biochemical data on binding of separate Bob1 regions to the four partners (for references, see the text). The structure of Oct1 complexed with Igκ was generated from the PDB file 1OCT. Three-dimensional structures have been visualized by the Swiss-PDB Viewer.

for CREB TAD [26], stathmin [27] and p21/27 [28] (see also the examples in [10]).

Another unique functional faculty of IUPs is that their open structure is largely preserved when they complex with their target, which provides for a disproportionately large binding surface and multiple contact points for a protein of the given size. For effectors, it provides specificity because of interaction with distant regions on the target, as exemplified by the X-ray structure of the cyclin-dependent-kinase inhibitor p27^{Kip1} bound to the cyclin A–cyclin-dependent kinase 2 (Cdk2) complex [29] (Fig. 1a). In this complex, p27 is in contact with its partners along its entire length. On cyclin A, it binds in a groove formed by conserved cyclin box residues; on Cdk2, it binds to the N-terminal lobe and also inserts into the catalytic cleft, mimicking ATP [29]. The situation is similar for scavengers, for which multivalent binding is also essential; good examples are the binding and stabilization of multiple small clusters of calcium phosphate by casein [30] or the precipitation of dietary tannins by salivary proline-rich protein (PRP) [31].

Perhaps most obviously, this requirement is important for assemblers, which by definition function by bringing multiple partners together. Convincing examples are the transactivator domain β-catenin binding domain (CBD) of transcription factor Tcf3 bound to the armadillo repeat domain of β-catenin [32] and the transcriptional coactivator Bob1, which has been shown in biochemical experiments to bind to two domains of Oct1 transcription factor [POU (Pit-1, Oct-1 and Oct-2, unc-86) specific domain and homeodomain] and to the Igκ promoter [33] and TAF$_{II}$105 [34] (Fig. 1b,c). Further examples are stathmin, which binds two tubulin dimers in an extended conformation [25], λ phage N translation antitermination protein, which binds RNA, NusA and RNA polymerase [35], and several ribosomal proteins that interact with distinct domains of rRNA in an extended conformation [36].

An additional prominent feature is that IUPs can adopt different structures upon different stimuli or with different partners, which enables their versatile interaction with various targets. This phenomenon, termed binding promiscuity [3] or one-to-many signalling [9], enables an exceptionally plastic behaviour in response to the needs of the cell. Such is the case for p21$^{Waf1/Cip1/Sdi1}$, which is able to interact with cycA–Cdk2, cycE–Cdk2, and cycD–Cdk4 [3], the SIBLING proteins (e.g. osteopontin or dentin matrix protein I), which can interact with integrin, CD44 and complement factor H [37], the TAD of cFOS, which can alternatively bind TATA box binding protein, CBP and Smad3 [38], or the proline-rich C-terminal domain of RNA polymerase II, which recruits different complexes for the sequential, highly coordinated capping, polyadenylation and splicing of pre-mRNA in transcription [39].

A further advantage is increased speed of interaction. Macromolecular association rates can be greatly enhanced by increasing the lifetime of the

Table 2 Amino acid frequencies of ordered and disordered proteins

Amino acid	Intrinsically unstructured proteins	Swiss-Prot[a]	Globular[b]
Ala (A)	7.15	7.62	8.15
Arg (R)	4.21	5.19	4.61
Asn (N)	2.06*	4.36	4.66*
Asp (D)	5.05	5.25	5.78
Cys (C)	0.61*	1.62	1.64*
Gln (Q)	4.46	3.93	3.69
Glu (E)	14.26*	6.47	5.98*
Gly (G)	4.31	6.85	7.99
His (H)	1.51	2.25	2.33
Ile (I)	3.67	5.85	5.43
Leu (L)	5.44	9.54	8.37
Lys (K)	10.43	5.97	6.05
Met (M)	1.30	2.37	2.03
Phe (F)	1.66*	4.10	3.95*
Pro (P)	12.07*	4.89	4.61*
Ser (S)	6.91	7.08	6.31
Thr (T)	5.14	5.57	6.15
Trp (W)	0.32*	1.21	1.55*
Tyr (Y)	1.42*	3.16	3.64*
Val (V)	8.02	6.61	7.00

Amino acid frequencies in % are given for globular proteins, Swiss-Prot proteins and for the intrinsically unstructured proteins and regions listed in Table 1. Asterisks identify amino acids that are at least two times more or less frequent than in an average globular protein in the Protein Data Bank.
[a]Data from http://us.expasy.org/sprot.
[b]Data 'globular-3D' from http://disorder.chem.wsu.edu.

encounter complex by an initial, relatively nonspecific association (reviewed in [40]). This effect can be of several orders of magnitude and the extended conformation of IUPs is well suited to accomplish such rate enhancements. A striking example is casein, which prevents precipitation of calcium phosphate in milk by binding and neutralizing small clusters very rapidly as they form. From kinetic data the binding-site activity of casein is calculated to be 10^6–10^7 s^{-1}, equalling the turnover rate of the fastest enzymes [30].

Flexibility is also instrumental in the assembly of large complexes. As found by examining the structural details of the ribosome [36], viral capsids and bacterial flagella [41], such large complexes cannot be assembled from rigid components due to steric hindrance: domain rearrangement and binding-induced folding must aid these to achieve the final functional state.

A final point is that the extreme proteolytic sensitivity of IUPs seen *in vitro* could provide for their effective control via rapid and perhaps regulated turnover *in vivo* [8,9]. This possible mechanism brings to mind the classical PEST hypothesis, which states that the presence of PEST regions (i.e. segments rich in P, E(D) and S/T flanked by R/K residues in proteins) correlates with a short lifetime in the cell [42]. As IUPs are enriched with the same amino acids (Table 2), this offers a plausible interpretation of the PEST hypothesis by associating proteolytic sensitivity with the extended conformation of IUPs.

A primary structure showing order

IUPs are often referred to as fully unstructured, featureless proteins, which is hard to reconcile with their functional perfection. In fact, evidence is growing against this oversimplified view and demonstrates order at the levels of amino acid composition (Table 2), sequence (Fig. 2) and secondary/tertiary structure (see below).

The proteins listed in Table 1 have a distinctive amino acid composition (Table 2), as has also been described for 'intrinsically disordered' protein regions [9,43]. In short, they are significantly enriched in P, E, K, S and Q, and depleted in W, Y, F, C, I, L and N, compared with the average folded protein in the PDB. Dunker and colleagues term the first group disorder-promoting amino acids, and the second group order-promoting amino acids [9,43]. The characteristic primary structure corresponding to this composition bears recognizable marks, as demonstrated by the performance of neural network algorithms trained to recognize protein disorder [9,14,43]. The persistence of this sequential trait through evolution [44] argues that the open structure is evolutionarily stable, which offers a rationale for this amino acid composition.

Low mean hydrophobicity and high net charge, associated with the disordered state [11,12], preclude the formation of a hydrophobic cluster and promote an extended conformation by electrostatic repulsion. The depletion in cysteine fits into this picture, as in globular proteins this amino acid often occurs in active sites or stabilizing disulfide bonds which are not required in IUPs. A high proline content is also linked with the lack of structure. Proline is known to disfavour a rigid secondary structure but has a strong preference for an open conformational motif, the left-handed polyproline II (PP II) helix [45,46], observed by circular dichroism (CD) or Raman optical activity (ROA) for tau protein [47], casein [47], stathmin [25], Bob1 [33] and NACP [47]. The significance of the open PP II conformation is that it is favoured in molecular recognition as it lacks intrachain hydrogen bonds and progresses easily to other conformational states [46]. Furthermore, proline is often actively involved in protein–protein contacts [46], which lends a functional dimension to the prevalence of this amino acid in IUPs, which rely heavily on target recognition.

The ease of the structural conversion of a PP II helix to a β-strand, however, also poses the danger of amyloidosis [8], as seen for NACP and tau protein [48,49]. The amino acid composition shown in Table 2, however, seems to counter the threat, as follows. First, high net charge and low mean hydrophobicity are inhibitory to aggregation [12]. Further, proline and glutamine, which abound in IUPs and prefer the open PP II conformation over all other secondary structural classes [45], are rather infrequent in β-strands: proline is the poorest β-strand-forming residue and glutamine also has a low Chou–Fasman parameter for this conformation [45]. This also applies to glutamic acid, the most frequent amino acid

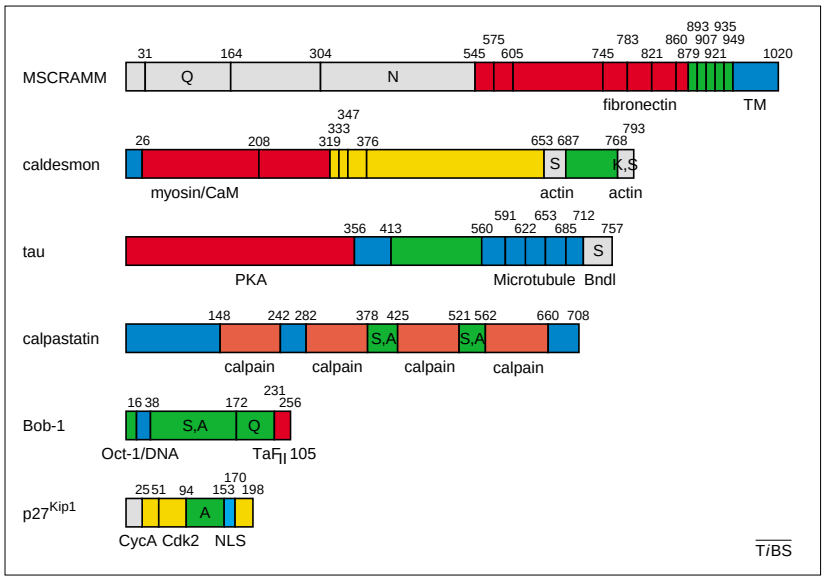

Fig. 2. Domain organization of intrinsically unstructured proteins (IUPs). The schemes are drawn of some IUPs for which a specific enrichment pattern of certain amino acids is apparent. Acidic (red), basic (blue), proline-rich (green) and charged domains/regions (yellow) are marked, or the one-letter code of the amino acid that is particularly prevalent is given. The region known to bind a certain interacting partner is marked by the name of that protein. For other functional regions, the following abbreviations are used: TM, transmembrane; BndI, bundling; NLS, nuclear localization signal. CaM, calmodulin; CycA, cyclin A; PKA, cAMP-dependent protein kinase; CaM, calmodulin; Cdk2, cyclin-dependent kinase 2; TAF$_{II}$105, TATA-box-associated factor II 105.

in IUPs. Conceivably, negative selection against β-strand-forming residues has been of high priority in the evolution of IUPs, as evidenced by the correlation of β-strand-forming potential with order–disorder discrimination [43]. This may well explain the imbalance of certain functionally equivalent amino acid pairs, such as Q>N and S>T (Table 2).

Structural order of IUPs is also apparent at the level of the amino acid sequence, as shown by their low compositional complexity compared with the almost random sequence of globular proteins [15,16]. The remarkable accuracy of neural network predictors developed for recognizing disorder by sequence inputs only [9,10,14] also underlines this conclusion. Furthermore, the long-range distribution of certain amino acids (proline, glutamine, acidic, basic) in IUP sequences is far from being random, as for many IUPs organization into domains defined by the prevalence of certain amino acids is clearly discernible (Fig. 2). Remarkably, the TADs of transcription factors, now known to be unstructured, are traditionally classified on the same basis (e.g. proline- and glutamine-rich, acidic) and these distinct structural classes often appear to be functionally incongruous [7]. This correspondence probably infers modular assembly of IUPs, which is underscored by the remarkable correlation of their characteristic amino-acid enrichment pattern with domains identified by functional criteria (Fig. 2).

IUPs are not disordered, but pliable

IUPs resemble denatured globular proteins and are often considered to be random coil or random-coil-like, although a true random coil does not exist even under strongly denaturing conditions [50]. A systematic analysis of hydrodynamic and far-UV CD data by Uversky has revealed that IUPs do display residual structure and fall into two distinct structural classes; that is, they behave as either coil-like or premolten-globule-like proteins [11]. It is tempting to speculate that their residual structure, approached experimentally by CD [3,6,11,13,25], ROA [47] and nuclear magnetic resonance [4,51,52] (see also Box 1), is linked to their function in a way that the protein preferentially samples conformational states that prevail in its complex with its target. This has been explicitly suggested for FlgM [51] and CREB TAD [52], for example.

A further point with respect to structure subserving function is that IUPs sometimes recognize and bind each other in a process termed 'cofolding' or 'synergistic folding', such as in the interaction between Bob1 and Oct1 TAD [53], RNA Pol II and the TADs of transcription factors, or multiple vesicle-associated proteins [46] or CBP/p300 and p160 nuclear receptor coactivators [54]. These cases can hardly be reconciled with a fully featureless conformational state. More importantly, the above examples suggest that structure correlates with function in the case of IUPs; that is, the classical structure–function paradigm only needs to be extended to encompass these highly flexible conformational states. Attempts to accommodate this view into a unifying concept have already been made.

The 'protein trinity' proposal [9] starts from prior work [55] suggesting that proteins can exist in three primary forms – folded globular, molten globular and unfolded – and goes on to suggest that the functions of native proteins can arise from any one of the three forms and from transitions between them. Uversky has extended this concept to the 'protein quartet' model, which also includes the premolten globular state in a similar manner [11]. These approaches are based on the structural analogy of IUPs with the folding intermediates of globular proteins. In light of the foregoing discussions, however, IUPs have several distinctive features that justify defining them as a unique category. For example, their open conformation is largely preserved in their complex with their target (Fig. 1); that is, they undergo local disorder–order transition or folding but do not fold into a compact globular state (not even *in vivo* [10]). Further, their function is intimately linked with disorder, which shows evolutionary stability [44]. Finally, the conformation they adopt is largely defined by their interacting partner, and not so much by their amino acid sequence as with globular proteins. To emphasize these unique features, the generic term 'pliable' is suggested to denote these proteins.

Conclusion

Intrinsically unstructured, or pliable, proteins are distinct from folded proteins with a well-defined

Acknowledgements
This work was supported by grant T 32360 from the Hungarian Scientific Research Fund (OTKA) and by a Bolyai János Scholarship. I thank Dr István Simon for reading the manuscript and Dr Peter Friedrich for his suggestions, in particular the term 'pliable' for intrinsically unstructured proteins. The many useful comments of three anonymous referees and Dr Keith Dunker are also appreciated.

three-dimensional structure. Their commonness and clear structural, functional and evolutionary separation justify their recognition as a unique protein category. This, however, is only the first step towards acknowledging their importance – their true potential has yet to be fully appreciated. For those involved with structural biology, nevertheless, the journey towards this goal will predictably be a most exciting and rewarding undertaking in the years ahead.

References

1 Lo Conte, L. *et al.* (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* 30, 264–267
2 Weinreb, P.H. *et al.* (1996) NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry* 35, 13709–13715
3 Kriwacki, R.W. *et al.* (1996) Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl. Acad. Sci. U. S. A.* 93, 11504–11509
4 Daughdrill, G.W. *et al.* (1997) The C-terminal half of the anti-sigma factor, FlgM, becomes structured when bound to its target, sigma 28. *Nat. Struct. Biol.* 4, 285–291
5 Konno, T. *et al.* (1997) A circular dichroism study of preferential hydration and alcohol effects on a denatured protein, pig calpastatin domain I. *Biochim. Biophys. Acta* 1342, 73–82
6 Schweers, O. *et al.* (1994) Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. *J. Biol. Chem.* 269, 24290–24297
7 Triezenberg, S.J. (1995) Structure and function of transcriptional activation domains. *Curr. Opin. Genet. Dev.* 5, 190–196
8 Wright, P.E. and Dyson, H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. *J. Mol. Biol.* 293, 321–331
9 Dunker, A.K. *et al.* (2001) Intrinsically disordered protein. *J. Mol. Graph. Model.* 19, 26–59
10 Dunker, A.K. *et al.* (2002) Intrinsic disorder and protein function. *Biochemistry* 41, 6573–6582
11 Uversky, V.N. (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* 11, 739–756
12 Uversky, V.N. *et al.* (2000) Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins* 41, 415–427
13 Kim, T.D. *et al.* (2000) Thermal behavior of proteins: heat-resistant proteins and their heat-induced secondary structural changes. *Biochemistry* 39, 14839–14846
14 Dunker, A.K. *et al.* (2000) Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* 11, 161–171
15 Wootton, J.C. (1994) Sequences with 'unusual' amino acid compositions. *Curr. Opin. Struct. Biol.* 4, 413–421
16 Romero, P. *et al.* (1999) Folding minimal sequences: the lower bound for sequence complexity of globular proteins. *FEBS Lett.* 462, 363–367
17 Gerstein, M. (1998) How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold. Design* 3, 497–512
18 Dyson, H.J. and Wright, P.E. (2002) Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* 12, 54–60
19 Mukhopadhyay, R. and Hoh, J.H. (2001) AFM force measurements on microtubule-associated proteins: the projection domain exerts a long-range repulsive force. *FEBS Lett.* 505, 374–378
20 Spolar, R.S. and Record, M.T., Jr (1994) Coupling of local folding to site-specific binding of proteins to DNA. *Science* 263, 777–784
21 Demchenko, A.P. (2001) Recognition between flexible protein molecules: induced and assisted folding. *J. Mol. Recog.* 14, 42–61
22 Parker, D. *et al.* (1999) Role of secondary structure in discrimination between constitutive and inducible activators. *Mol. Cell. Biol.* 19, 5601–5607
23 Hauer, J.A. *et al.* (1999) Binding-dependent disorder-order transition in PKI alpha: a fluorescence anisotropy study. *Biochemistry* 38, 6774–6780
24 Fletcher, C.M. and Wagner, G. (1998) The interaction of eIF4E with 4E-BP1 is an induced fit to a completely disordered protein. *Protein Sci.* 7, 1639–1642
25 Wallon, G. *et al.* (2000) Model for stathmin/OP18 binding to tubulin. *EMBO J.* 19, 213–222
26 Radhakrishnan, I. *et al.* (1997) Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator:coactivator interactions. *Cell* 91, 741–752
27 Horwitz, S.B. *et al.* (1997) The microtubule-destabilizing activity of metablastin (p19) is controlled by phosphorylation. *J. Biol. Chem.* 272, 8129–8132
28 Pavletich, N.P. (1999) Mechanisms of cyclin-dependent kinase regulation: structures of Cdks, their cyclin activators, and Cip and INK4 inhibitors. *J. Mol. Biol.* 287, 821–828
29 Russo, A.A. *et al.* (1996) Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. *Nature* 382, 325–331
30 Holt, C. *et al.* (1996) Ability of a beta-casein phosphopeptide to modulate the precipitation of calcium phosphate by forming amorphous dicalcium phosphate nanoclusters. *Biochem. J.* 314, 1035–1039
31 Muenzer, J. *et al.* (1979) Properties of proline-rich proteins from parotid glands of isoproterenol-treated rats. *J. Biol. Chem.* 254, 5629–5634
32 Graham, T.A. *et al.* (2000) Crystal structure of a beta-catenin/Tcf complex. *Cell* 103, 885–896
33 Chang, J.F. *et al.* (1999) Oct-1 POU and octamer DNA co-operate to recognise the Bob-1 transcription co-activator via induced folding. *J. Mol. Biol.* 288, 941–952
34 Wolstein, O. *et al.* (2000) Specific interaction of TAFII105 with OCA-B is involved in activation of octamer-dependent transcription. *J. Biol. Chem.* 275, 16459–16465
35 Mogridge, J. *et al.* (1998) Independent ligand-induced folding of the RNA-binding domain and two functionally distinct antitermination regions in the phage lambda N protein. *Mol. Cell* 1, 265–275
36 Ban, N. *et al.* (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* 289, 905–920
37 Fisher, L.W. *et al.* (2001) Flexible structures of SIBLING proteins, bone sialoprotein, and osteopontin. *Biochem. Biophys. Res. Commun.* 280, 460–465
38 Campbell, K.M. *et al.* (2000) Intrinsic structural disorder of the C-terminal activation domain from the bZIP transcription factor Fos. *Biochemistry* 39, 2708–2713
39 Orphanides, G. and Reinberg, D. (2002) A unified theory of gene expression. *Cell* 108, 439–451
40 Pontius, B.W. (1993) Close encounters: why unstructured, polymeric domains can increase rates of specific macromolecular association. *Trends Biochem. Sci.* 18, 181–186
41 Namba, K. (2001) Roles of partly unfolded conformations in macromolecular self-assembly. *Genes Cells* 6, 1–12
42 Rechsteiner, M. and Rogers, S.W. (1996) PEST sequences and regulation by proteolysis. *Trends Biochem. Sci.* 21, 267–271
43 Williams, R.M. *et al.* (2001) The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac. Symp. Biocomput.* 6, 89–100
44 Iakoucheva, L.M. *et al.* (2001) Identification of intrinsic order and disorder in the DNA repair protein XPA. *Protein Sci.* 10, 560–571
45 Adzhubei, A.A. and Sternberg, M.J. (1993) Left-handed polyproline II helices commonly occur in globular proteins. *J. Mol. Biol.* 229, 472–493
46 Williamson, M.P. (1994) The structure and function of proline-rich regions in proteins. *Biochem. J.* 297, 249–260
47 Syme, C.D. *et al.* (2002) A Raman optical activity study of rheomorphism in caseins, synucleins and tau. *Eur. J. Biochem.* 269, 148–156
48 Conway, K.A. *et al.* (2000) Fibrils formed *in vitro* from alpha-synuclein and two mutant forms linked to Parkinson's disease are typical amyloid. *Biochemistry* 39, 2552–2563
49 von Bergen, M. *et al.* (2000) Assembly of tau protein into Alzheimer paired helical filaments depends on a local sequence motif ((306)VQIVYK(311)) forming beta structure. *Proc. Natl. Acad. Sci. U. S. A.* 97, 5129–5134
50 Shortle, D. (1996) The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB J.* 10, 27–34
51 Daughdrill, G.W. *et al.* (1998) The C-terminal half of the anti-sigma factor FlgM contains a dynamic equilibrium solution structure favoring helical conformations. *Biochemistry* 37, 1076–1082
52 Hua, Q.X. *et al.* (1998) Transcriptional activator–coactivator recognition: nascent folding of a kinase-inducible transactivation domain predicts its structure on coactivator binding. *Biochemistry* 37, 5858–5866
53 Lee, L. *et al.* (2001) Expression of the Oct-1 transcription factor and characterization of its interactions with the Bob1 coactivator. *Biochemistry* 40, 6580–6588
54 Demarest, S.J. *et al.* (2002) Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature* 415, 549–553
55 Ptitsyn, O.B. and Uversky, V.N. (1994) The molten globule is a third thermodynamical state of protein molecules. *FEBS Lett.* 341, 15–18