INTRINSICALLY UNSTRUCTURED PROTEINS AND THEIR FUNCTIONS

H. Jane Dyson and Peter E. Wright

Abstract | Many gene sequences in eukaryotic genomes encode entire proteins or large segments of proteins that lack a well-structured three-dimensional fold. Disordered regions can be highly conserved between species in both composition and sequence and, contrary to the traditional view that protein function equates with a stable three-dimensional structure, disordered regions are often functional, in ways that we are only beginning to discover. Many disordered segments fold on binding to their biological targets (coupled folding and binding), whereas others constitute flexible linkers that have a role in the assembly of macromolecular arrays.

NMR
Nuclear magnetic resonance (NMR) spectroscopy provides information on the three-dimensional structure and dynamics of biological molecules in solution.

Department of Molecular Biology and Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA.
e-mails: dyson@scripps.edu; wright@scripps.edu
doi:10.1038/nrm1589

The occurrence of unstructured regions of significant size (>50 residues) is surprisingly common in functional proteins[1,2]. In addition, the existence of functional unstructured proteins — for example, polypeptide hormones[3] — has been recognized for many years, and unstructured proteins were observed in intact cells in early proton NMR experiments[4]. However, the functional role of intrinsically disordered proteins in crucial areas such as transcriptional regulation, translation and cellular signal transduction[5–8] has only recently been recognized, as a consequence of the use of new paradigms in biochemical methodology. In particular, the availability of large amounts of sequence data coupled with gene-based functional analysis (BOX 1) has led to the extensive use of sequence analysis for the identification of intrinsically unstructured sequences. Another reason for the attention being paid to disordered regions of proteins is that techniques have recently been developed to analyse their structural propensities in solution. Crystal-structure analysis cannot provide information on unstructured states — it can only indicate their presence through the absence of electron density in local regions. However, spectroscopic methods such as NMR have now advanced in sensitivity and resolution, to the point at which the structural propensities and dynamics of sizeable disordered proteins in solution can be thoroughly characterized.

This review focuses on the identification of intrinsically unstructured proteins and describes their general characteristics. The functional roles that unstructured regions can have are summarized in the context of published examples, including several from the transcriptional activator cyclic-AMP-response-element-binding protein (CREB)-binding protein (CBP). Furthermore, the article also discusses the advantages and thermodynamic and biological consequences of the presence, in certain cellular machines, of regions that fold on binding to their physiological target or that function as flexible linkers.

Will a domain be folded or unfolded?

Predicting the three-dimensional (3D) structures of globular proteins from sequence data alone remains a key challenge, except in situations in which the protein has a high sequence homology to domains of known structure. On the other hand, identifying sequences that are likely to be intrinsically disordered — that is, that do not fold spontaneously into well-organized globular structures in the absence of stabilizing interactions — is comparatively straightforward.

Sequence signatures of intrinsic disorder. A signature of probable intrinsic disorder is the presence of low sequence complexity and amino-acid compositional bias, with a low content of bulky hydrophobic amino

Box 1 | Why have we only just discovered unfolded proteins?

Classic biochemical methods are strongly biased towards the production and characterization of folded, active proteins. The standard preparation methods run approximately as follows:

- Plant or animal tissues, or bacterial cells, are isolated and homogenized.
- The homogenate is assayed for the activity of interest.
- The homogenate is subjected to ammonium-sulphate fractionation, chromatography and/or gel filtration.
- The fractions are assayed for activity and the active protein is purified.
- The pure, active protein is sequenced.
- The three-dimensional structure is determined.

This methodology automatically selects folded proteins, because the formation of a homogenate invariably releases proteases and unfolded proteins are much more sensitive than folded proteins to degradation under these conditions. Also, if the unfolded domains are part of regulatory proteins, there might be only a few copies per cell, and they might not have a convenient activity to assay.

So, why are we discovering unfolded proteins now? We have only begun to understand the presence and roles of unfolded proteins since the advent of new paradigms in biochemical methodology. Instead of discovering a detectable activity and isolating it by purifying the protein, we now have access to a vast library of gene sequences. The use of genetic methods to isolate function, using mutants and knockouts, has been the way that most unfolded proteins have been identified so far. This identification process involves:

- Formulating a function (for example, control of transcription).
- Mapping the function to a particular gene and to a particular area in the gene.
- Transcribing the gene, producing the protein and purifying the protein.
- Examining its structure in solution by circular dichroism and NMR spectroscopy.
- If the protein is unfolded, trying to co-express it with a binding partner (also identified through genetic mapping).

CIRCULAR DICHROISM
The ultraviolet circular-dichroism spectrum uses the chirality or 'handedness' of biological molecules to provide information on secondary structure in solution.

FLUORESCENCE SPECTROSCOPY
Fluorescence spectroscopy of proteins gives information on the environment of aromatic rings, and can be used in conjunction with external probes to determine the distances between atoms in a molecule.

VIBRATIONAL CD SPECTROSCOPY
Vibrational circular dichroism (CD) is the chiroptical version of infra-red spectroscopy, and it gives information on the vibrations of individual bonds in a molecule.

RAMAN SPECTROSCOPY
Raman spectroscopy provides information on bond vibrations that is complementary to that provided by infra-red spectroscopy.

acids (Val, Leu, Ile, Met, Phe, Trp and Tyr), which would normally form the core of a folded globular protein, and a high proportion of particular polar and charged amino acids (Gln, Ser, Pro, Glu, Lys and, on occasion, Gly and Ala)[9,10]. The presence of such regions in transcriptional regulatory proteins was recognized more than 25 years ago[11]. Many of these regions function in transcriptional activation and they are often classified according to their amino-acid composition — for example, there are glutamine-rich, proline-rich and acidic activation domains. NMR studies of the acidic activation domain from herpes simplex virus confirmed that it is intrinsically unstructured[12]. Structural-disorder prediction was recently used to drive the design of laboratory experiments for the identification of binding sites within disordered regions, and the positions of these sites were later confirmed by crystal-structure analysis[13].

A number of computer programs are now available for the prediction of unstructured regions from amino-acid sequences. These include PONDR[14], FoldIndex[15], DisEMBL[16], GLOBPLOT 2 (REF. 17) and DISOPRED2 (REF. 18; see the online links box), as well as a 'support vector machine'[19]. An exhaustive summary of computational methods for the recognition of unstructured regions is beyond the scope of this review, but other review articles have recently addressed this subject in detail (see, for example, REFS 2,20). Analysis of sequence data for complete genomes indicates that intrinsically disordered proteins are highly prevalent, and that the proportion of proteins that contain such segments increases with the increasing complexity of an organism[1,18]. Database analysis indicates that proteins that are involved in eukaryotic signal transduction or that are associated with cancer have an increased propensity for intrinsic disorder[6]. More than 100 proteins have been shown experimentally to be either completely or partially disordered[1,2], and a database of protein disorder has recently been established[21] (DISPROT; see the online links box). A recent survey has shown that intrinsically unstructured proteins often contain repeat regions, and it was proposed that these regions might have evolved by repeat expansion[22].

Experimental characterization of disordered proteins. The key experimental method for obtaining systematic site-specific information on unstructured proteins is NMR spectroscopy. The application of NMR methodology to the study of unstructured proteins has recently been extensively reviewed elsewhere (see REFS 23,24). Other techniques, including CIRCULAR DICHROISM (CD), hydrodynamic measurements, FLUORESCENCE SPECTROSCOPY, VIBRATIONAL CD SPECTROSCOPY and RAMAN SPECTROSCOPY, as well as traditional biochemical studies of proteolytic susceptibility, can give important information (see REF. 25 for a comprehensive series of reviews).

General characteristics of disordered proteins. Proteins fall onto a structural continuum, from tightly folded single domains, to multidomain proteins that might have flexible or disordered regions, to compact but disordered MOLTEN GLOBULES and, finally, to highly extended, heterogeneous unstructured states (FIG. 1). This continuum has been interpreted in terms of a 'PROTEIN TRINITY' (ordered, molten globule and RANDOM COIL[20]) or 'PROTEIN QUARTET'[2], although it is clear from FIG. 1 that there are a number of different structural types within each of these subdivisions. In general, proteins with intrinsically disordered sequences cannot bury sufficient hydrophobic core to fold spontaneously into the highly organized 3D structures that characterize the proteins that are represented in the Protein Data Bank (see the online links box). In some cases, compact but disordered molten-globule-like states can be formed, or local regions of the sequence can have a propensity to adopt isolated and fluctuating elements of secondary structure (which is equivalent to the 'pre-molten globule' proposed by Uversky[2]). It is probable that proteins rarely, if ever, behave as true random coils, especially in non-denaturing media: even in their most highly unfolded states, proteins show a propensity to form local elements of secondary structure or hydrophobic clusters[25,26].

Many eukaryotic proteins are modular — that is, they contain independently folded globular domains that are separated by flexible linker regions. Linker sequences vary greatly in length and composition, but many are rich in polar, uncharged amino acids (such as Ser, Thr, Gln and Asn), in the small residues Ala and Gly, and in Pro residues. Many of these residues tend to bias the polypeptide chain towards the polyproline-II region of the RAMACHANDRAN PLOT[27,28]. This means that such

MOLTEN GLOBULE
The molten globule was originally defined with reference to the folding pathway of an ordered protein as a compact state of a protein, with native-like secondary structure but disordered tertiary structure.

PROTEIN TRINITY
In terms of their structure, proteins can be defined as being in one of three states — unfolded, molten globule or folded.

RANDOM COIL
This term refers to a 'statistical coil' with a random distribution of dihedral angles. In practice, no protein is ever a completely random coil, but the term is a convenient shorthand for the ensemble of conformations that occur for an unfolded protein.

linkers, although flexible, have a propensity to be highly extended. Compositionally biased linker sequences of significant length are found mainly in eukaryotic proteins[1,29], but short linker sequences of similar composition, known as Q-linkers, are also found in a number of bacterial regulatory proteins[30].

In the absence of their targets, modular proteins often behave as 'beads on a flexible string', where the function of the linker is, primarily, to enable a relatively unhindered spatial search by the attached domains[31]. However, binding can induce structure formation in linkers, which can have significant functional consequences. For example, the sequence-specific binding of CYS_2HIS_2 ZINC-FINGER PROTEINS to DNA causes the linker to fold, cap and thereby stabilize the preceding helix in the protein, and to orientate the next zinc finger correctly for binding in the major groove of DNA[32] (FIG. 1). This process has been likened to the action of an inducible 'snap-lock', which is activated by the binding of the flexible zinc-finger protein to its cognate DNA sequence. A similar role has been proposed for the flexible linker that connects the SRC-HOMOLOGY-2 (SH2) DOMAIN and SH3 DOMAIN of the kinases Hck (haematopoietic-cell kinase) and Src, which becomes 'locked' on phosphorylation[33]. The functional role of the linkers in zinc-finger proteins is emphasized by the Wilms' tumour protein WT1. Alternative splicing that lengthens a linker in WT1, and thereby increases its flexibility, abrogates DNA binding and alters both the function and subcellular localization of this protein[34].

Functions of intrinsic disorder in proteins

New examples of functional intrinsically disordered protein domains are constantly emerging, and the reader is referred to several recent reviews for a detailed survey (see REFS 1,2,5,8,35). Functions include the regulation of transcription and translation, cellular signal transduction, protein phosphorylation, the storage of small molecules, and the regulation of the self-assembly of large multiprotein complexes such as the bacterial flagellum and the ribosome. A recent review[36] highlights the occurrence of unfolded regions in proteins that function as chaperones for other proteins and for RNA molecules, and proposes that the unfolded regions work

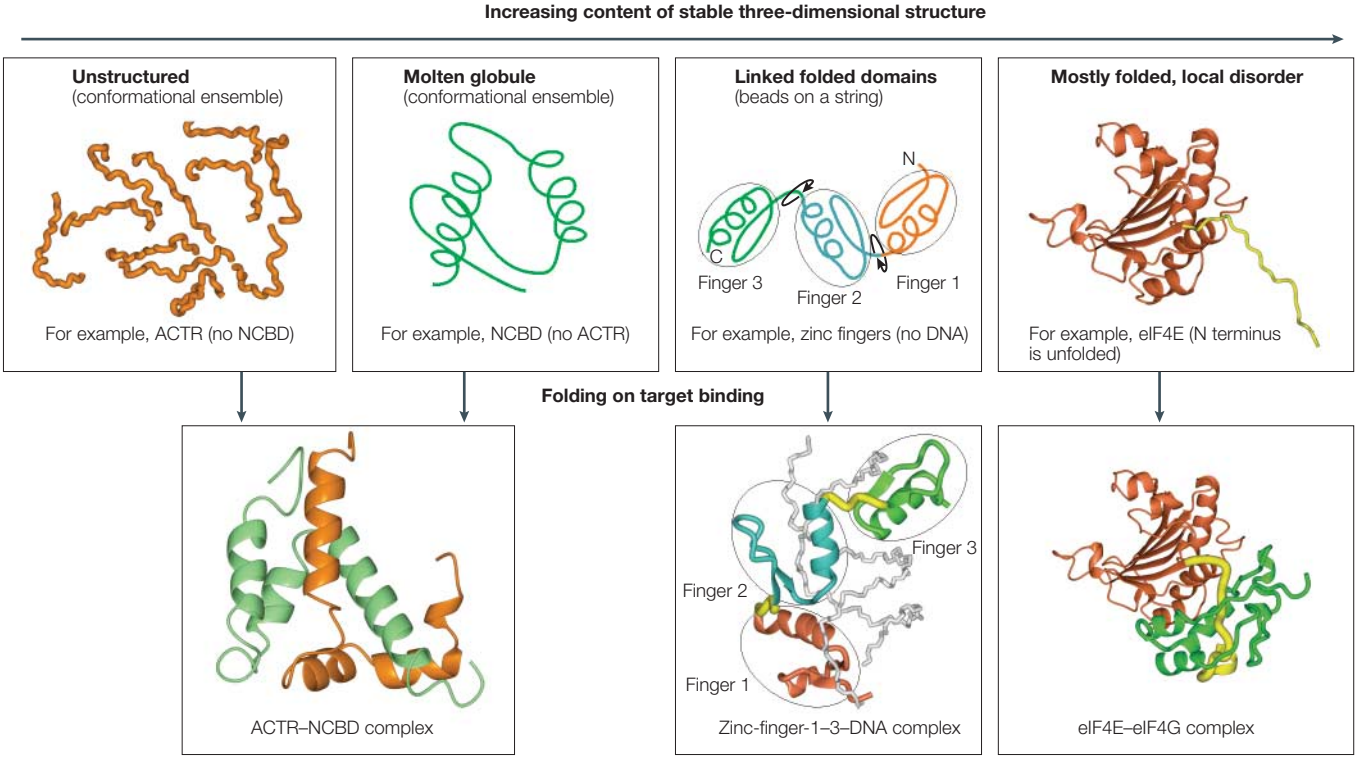


Figure 1 | **The continuum of protein structure.** The upper panels show examples on the continuum of protein structure: that is, an unstructured conformational ensemble, which is represented by the interaction domain of activator for thyroid hormone and retinoid receptors (ACTR)[91]; a molten globule-like domain such as the nuclear-receptor co-activator-binding domain (NCBD) of cyclic-AMP-response-element-binding protein (CREB)-binding protein (CBP)[89,91]; linked folded domains such as a construct that contains the first three zinc fingers of transcription factor-IIIA (TFIIIA)[122]; and free eukaryotic translation-initiation factor (eIF)4E[123], which is mostly folded with only local disorder. The lower panels show the structures of the domains in the upper panels when they are folded onto their biological target domains or sequences. The left-hand lower panel shows the mutually folded structure of a complex[91] (Protein Data Bank (PDB) file 1KBH) between the ACTR domain of the p160 co-activator (orange) and the NCBD domain of CBP (green). The second panel shows the well-ordered structure of the first three zinc-fingers of TFIIIA bound to an oligonucleotide that contains its cognate DNA sequence[124] (PDB file 1TF3). The third panel shows the complex between eIF4E (brown) and eIF4G, which highlights the mutual folding of the N-terminal tail of eIF4E (yellow) and eIF4G (green)[123] (PDB file 1RF8). All of the three-dimensional structure figures were drawn using MOLMOL[125].

PROTEIN QUARTET
A similar division for protein structure as the protein trinity, but the quartet includes a pre-molten globule state as well as the unfolded, molten-globule and folded states.

RAMACHANDRAN PLOT
A plot of the backbone dihedral angles ϕ and ψ for a polypeptide chain. Areas of low energy (greater probability) encompass angles that are observed in α-helical and β-sheet structures, and a part of the broad β-minimum is defined as the 'polyproline II' region.

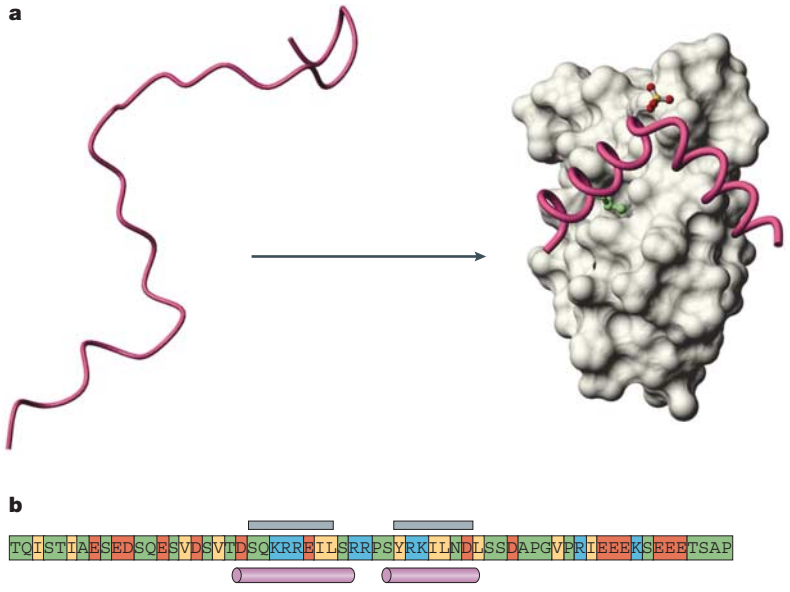
by binding to misfolded proteins and RNA molecules, such that they function as recognition elements and/or help in the loosening and unfolding of kinetically-trapped folding intermediates.

Many intrinsically disordered proteins undergo transitions to more ordered states or fold into stable secondary or tertiary structures on binding to their targets — that is, they undergo coupled folding and binding processes[5,37,38] (BOX 2). One of the most well-characterized examples is the kinase-inducible transcriptional-activation domain (KID) of CREB. The KID polypeptide is intrinsically disordered, both as an isolated peptide and in full-length CREB[39,40], but it folds to form a pair of orthogonal helices on binding to its target domain in CBP[41] (BOX 2). Interestingly, the intrinsic disorder of the KID can be reliably predicted from its amino-acid sequence, as can an inherent helical propensity in the region that undergoes the coupled folding and binding transition. Indeed, the identification of amphipathic elements embedded within regions of a protein that are predicted to be disordered might provide clues as to the location of potential functional sites. Coupled folding and binding might involve just a few residues — as is the case for the KID of CREB — or an entire protein domain. For example, the 116-residue N-terminal domain of DNA-fragmentation factor 45-kDa subunit (DFF45) is unstructured in solution, but folds into an ordered globular structure on forming a heterodimeric complex with DFF40 (REF. 42).

Box 2 | Coupled folding and binding

Coupled folding and binding is the process in which an intrinsically disordered protein, or region of a protein, folds into an ordered structure concomitant with binding to its target. There is an entropic cost to fold a disordered protein, which is paid for using the binding enthalpy (BOX 3). For example, the phosphorylated kinase-inducible domain (pKID) of the transcription factor cyclic-AMP-response-element-binding protein (CREB) is unstructured when it is free in solution[39,40], but it folds on forming a complex with the KID-binding (KIX) domain of CREB-binding protein (CBP)[41] (see figure, part a).

The amino-acid sequence of the KIX-binding region of CREB is shown in part b of the figure. The colour coding for the amino acids is: green for small residues, uncharged hydrophilic residues and Pro; yellow for hydrophobic residues; red for acidic residues; and blue for basic residues. Helices that were predicted using The PredictProtein server[117] (see the online links box), but which are not observed in the free peptide, are shown in grey above the sequence. The locations of the helices that are observed in pKID on its binding to KIX are shown in pink below the sequence.



A case study: transcriptional co-activators

CBP and its PARALOGUE p300 are modular transcriptional co-activators. They modify both chromatin and transcription factors through their intrinsic acetyltransferase activity, and also function as scaffolds for the recruitment and assembly of the transcriptional machinery[43–45]. Of the 2,442 amino acids that comprise the sequence of human CBP, more than 50% are in regions of the protein that are intrinsically disordered, according to many of the available prediction programs. Compositional bias in intrinsically disordered regions is exemplified by CBP, the sequence of which is shown in FIG. 2a. There are long amino-acid sequences between the folded domains that contain predominantly Gln, Pro, Ser, Gly, Thr and Asn residues. By contrast, most of the interaction domains that have been identified have an amino-acid composition that is more typical of globular proteins. Interestingly, there are a number of segments within the 'disordered' regions of CBP that contain relatively high proportions of hydrophobic and charged residues; these regions perhaps represent as-yet-unidentified interaction motifs. There are only seven domains in CBP/p300 that are capable of folding independently (FIG. 2b), and four of them require zinc binding to stabilize their tertiary structures (the transcriptional-adaptor zinc-finger-1 (TAZ1) domain, the plant homeodomain (PHD), a zinc-binding domain near the dystrophin WW domain (ZZ), and the transcriptional-adaptor zinc-finger-2 (TAZ2) domain). The 3D structures of the folded domains of CBP/p300 — except the PHD and histone acetyltransferase (HAT) domains — have been determined by NMR methods, and their functions as templates for coupled folding and binding processes are discussed below.

TAZ domains: scaffolds for assembly. The TAZ domains[46] of CBP/p300 are zinc-binding domains with a distinctive helical fold[47]. The TAZ1 and TAZ2 domains share significant sequence homology and adopt similar 3D structures, and the only significant differences are in the location of the third zinc-binding site and the C-terminal helix[47,48]. However, these subtle structural changes allow them to discriminate between different subsets of transcription factors. The TAZ2 domain is the site of interaction both with transactivation domains of viral oncoproteins such as E1A and with the tumour suppressor p53 (REFS 49–52), whereas TAZ1 mediates key interactions with hypoxia-inducible factor-1α (HIF1α) and thereby regulates the hypoxic response[53]. These ligands have all been found to be disordered in the free state. The C-terminal transcription-activation domain of HIF1α is unstructured in solution, but undergoes local folding transitions to form three short helices on binding

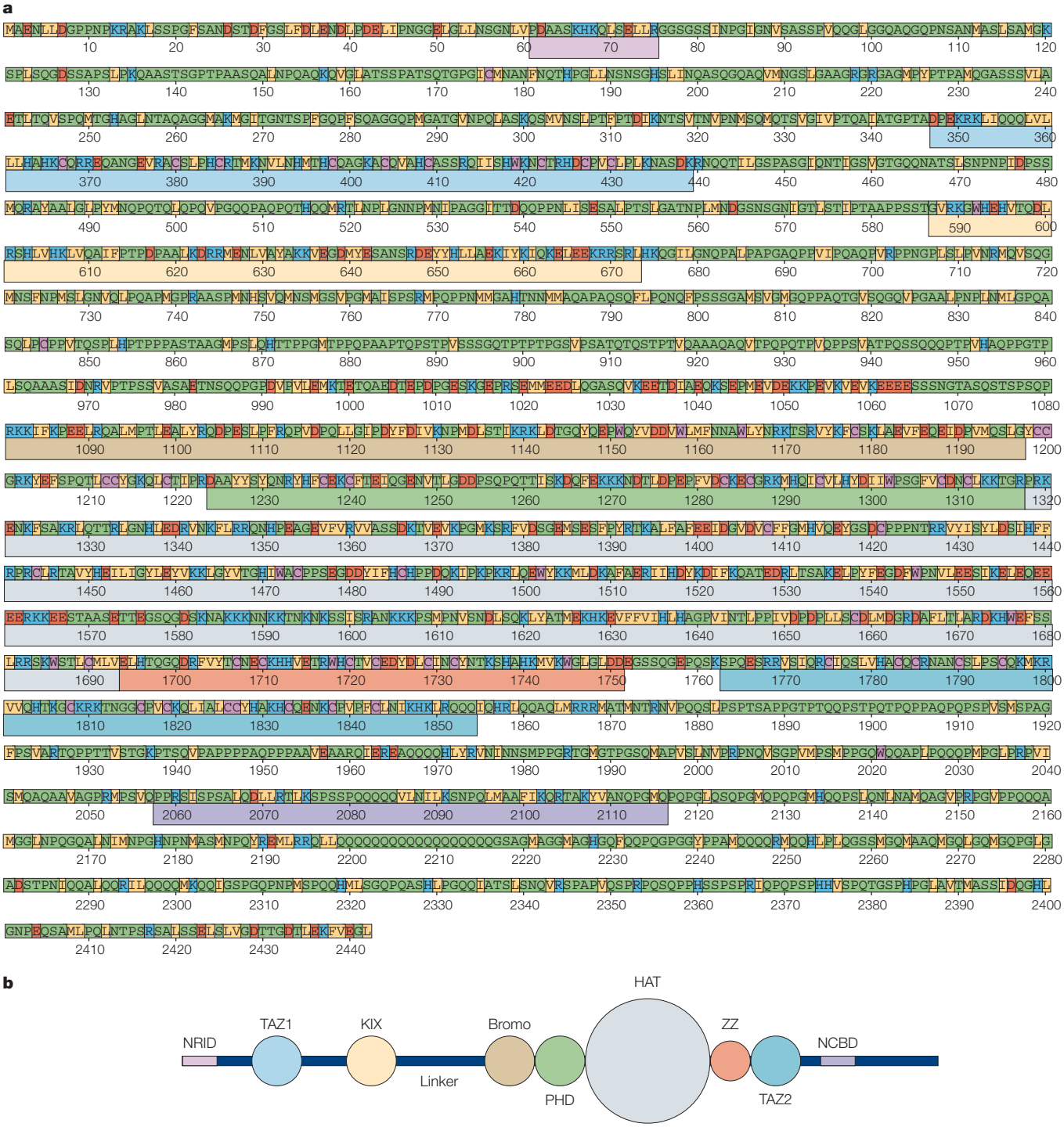


Figure 2 | **Structured and unstructured regions of CBP. a** | The amino-acid sequence of human cyclic-AMP-response-element-binding protein (CREB)-binding protein (CBP), which is colour coded to highlight the amino-acid-residue preferences in structured domains and intrinsically unstructured regions. The colour coding for amino acids is: green, small residues, uncharged hydrophilic residues and proline residues (that is, Asn, Gln, Ser, Thr, Gly, Ala and Pro); yellow, hydrophobic residues (that is, Val, Leu, Ile, Phe and Tyr); red, acidic residues (that is, Asp and Glu); blue, basic residues (that is, Lys, Arg and His); and purple, low-frequency residues (that is, Cys and Trp). The positions of the structured domains are highlighted by coloured boxes below the sequence, and the colours of these boxes correspond to the domain colours that are used in part **b**. **b** | A schematic representation of the domain structure of CBP. Structured domains are represented by circles. The unstructured nuclear-receptor co-activator-binding domain (NCBD) is shaded purple, and the pink shading at the N terminus represents the disordered nuclear-receptor-interaction domain (NRID). The cell-cycle regulatory domain-1 (CRD1; which is known to be between residues 1019–1082 of CBP) is not highlighted, because the boundaries of this domain have not been adequately defined. Bromo, bromodomain; HAT, histone acetyltransferase domain; KIX, KID-binding domain; PHD, plant homeodomain; TAZ1/2, transcriptional-adaptor zinc-finger-1/2; ZZ, zinc-binding domain near the dystrophin WW domain.

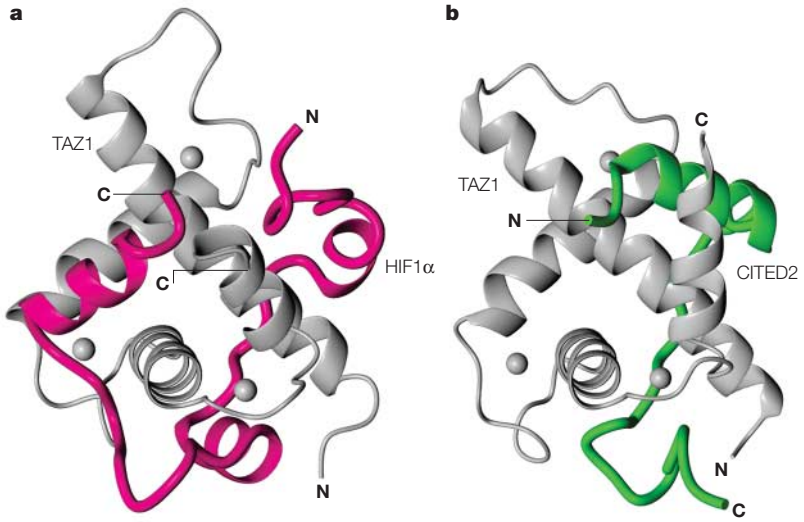


Figure 3 | **The structure of the TAZ1 domain in complex with two interaction domains.** **a** | The TAZ1-domain–HIF1α complex[48] (Protein Data Bank (PDB) accession code 1L8C). **b** | The TAZ1-domain–CITED2 complex[61] (PDB accession code 1R8U). Spheres show the location of zinc atoms. For further details, please refer to the main text. This figure was prepared using MOLMOL[125]. CITED2, second CBP/p300 interacting transactivator with glutamate (E)- and aspartate (D)-rich tail-2; HIF1α, hypoxia-inducible factor-1α; TAZ1, transcriptional-adapter zinc-finger-1.

to the TAZ1 domain of CBP/p300. The structure of the complex[48,54] (FIG. 3a) shows the HIF1α polypeptide wrapped almost entirely around the TAZ1 scaffold, thereby forming an extensive intermolecular interface. As a consequence, the binding affinity is very high (that is, the dissociation constant (K_d) = ~7 nM; REF. 48), much higher than could be achieved by interactions between two stable, folded proteins of comparable sizes. This reaction provides an example of enthalpy–entropy compensation (BOX 3), and the coupling of folding and binding allows the burial of an extremely large surface area, even when the interacting domains are quite small. Indeed, it has recently been pointed out that, in the absence of the coupled folding and binding of intrinsically unstructured protein domains, the proteins would have to be 2–3 times larger in order for them to form such extensive interfaces, which would either increase cellular crowding or the size of the cell itself[55]. There are a number of examples of extensive intermolecular interfaces that are formed by the folding of ligand molecules on binding, including those in the complexes $p27^{Kip1}$–cyclin-A–cyclin-dependent-kinase-2 (in which $p27^{Kip1}$ is an inhibitor of the complex; REF. 56), CADHERIN–β-catenin (REF. 57) and σ^{28}–FlgM (REF. 58), which indicates that this might be a general and important functional device that is endowed by disorder in binding sequences.

The hypoxic response is subject to tight regulatory control. One of the control mechanisms involves the protein CITED2 (the second CBP/p300 interacting transactivator with glutamate (E)- and aspartate (D)-rich tail), which functions as a negative-feedback regulator by competing with HIF1α for binding to CBP/p300 (REF. 59). CITED2 is also intrinsically disordered in solution, and it folds on binding to the TAZ1 domain at a site that partially overlaps the binding site for HIF1α (REFS 60,61) (FIG. 3b). Therefore, we can envisage the competition of the two ligands for TAZ1 as being controlled by mass action — in the presence of an excess of one of the ligands, the other ligand could be 'peeled off' without the necessity for prior complete dissociation.

Regulating affinity by post-translational modification. The binding of intrinsically disordered proteins to their targets is often regulated by covalent modifications, which leads to simple biological switches, and several examples of this phenomenon are found in the CBP/p300 system. Such processes require binding to more than one target — that is, binding to a modifying enzyme and to the physiological receptor. Because the conformational requirements for binding to each target often differ, binding might be facilitated by the presence of structural disorder in the protein.

First, the hypoxic response will be considered, which is regulated at several levels, including the hydroxylation of HIF1α at specific Pro and Asn residues in normoxic cells[62–65]. The conserved Asn803 of HIF1α functions as a

CYS$_2$HIS$_2$ ZINC-FINGER PROTEIN
The Cys$_2$His$_2$ zinc finger is a common structural motif. It is a small sequence motif that contains two Cys and two His residues, which coordinate a single zinc ion. Tandem repeats of zinc fingers are common.

SH2 DOMAIN
The Src-homology-2 (SH2) domain, which is a peptide-binding domain of Src protein kinases, is a common structural motif. It binds peptides and proteins that contain phosphorylated Tyr residues.

SH3 DOMAIN
The Src-homology-3 domain, which is a peptide-binding domain of Src protein kinases, is a common structural motif. It binds polyproline sequences.

PARALOGUES
Sequences, or genes, that have originated from a common ancestral sequence, or gene, by a duplication event.

CADHERIN
A Ca^{2+}-binding membrane protein that mediates homophilic cell adhesion.

Box 3 | **Thermodynamic consequences of coupled folding and binding**

There is an entropic cost associated with the disorder-to-order transition that accompanies the binding of an intrinsically unstructured protein to its target. The key thermodynamic driving force for the binding reaction is generally a favourable enthalpic contribution, which gives an example of enthalpy–entropy compensation.

Coupled folding and binding frequently gives rise to a complex with high specificity and relatively low affinity, which is appropriate for signal-transduction proteins that must not only associate specifically to initiate the signalling process, but must also be capable of dissociation when signalling is complete. Another advantage of a system that uses components that fold on binding is that the conformational flexibility facilitates the post-translational modification of important transcription factors. Conformational flexibility allows a protein to bind to both its physiological target and to modifying enzymes. Several excellent examples of entropy–enthalpy compensation are provided by the transcriptional activator CBP (cyclic-AMP-response-element-binding protein (CREB)-binding protein) system. Specifically, by the phosphorylated kinase-inducible domain (pKID)–KID-binding-domain (KIX) interaction, the hypoxia-inducible factor-1α (HIF1α)–transcriptional-adaptor-zinc-finger-1 (TAZ1)-domain interaction and the activator for thyroid hormone and retinoid receptors (ACTR)–nuclear-receptor-co-activator-binding-domain (NCBD) interaction (please refer to the main text for further details). This mechanism for the reversible formation of complexes with extremely high specificity might prove to be a hallmark of the protein–nucleic-acid and protein–protein interactions that are involved in transcription and translation.

E3 UBIQUITIN LIGASE
E3 (enzyme-3) ubiquitin ligases are the enzymes that are responsible for attaching ubiquitin to proteins, which can target them for destruction by the 26S proteasome.

hypoxic switch: hydroxylation of Asn803 in normoxic cells impairs the binding of the HIF1α C-terminal transactivation domain to CBP/p300 (REF. 65). This Asn lies in a region of HIF1α that is intrinsically disordered in the free state, but that forms a helical structure when it is in complex with the TAZ1 domain of CBP/p300 (REFS 48,54) (FIG. 4a). However, binding to the enzyme that accomplishes the hydroxylation of Asn803 requires that the same region of HIF1α forms a highly extended structure, with the Asn side chain projecting into the enzyme active site[66,67] (FIG. 4b). The intrinsic plasticity of the disordered HIF1α activation domain means that its conformation can be adapted, at little energetic cost, to fit the requirements of its various targets. This idea was first raised in the context of the cyclin-dependent-kinase inhibitor p21 (REF. 68).

Hydroxylation of Pro564 in the oxygen-dependent degradation domain of HIF1α under normoxic conditions leads to the recruitment of the von-Hippel–Lindau tumour suppressor, a component of an E3 UBIQUITIN LIGASE, and to the subsequent proteolytic degradation of the HIF1α polypeptide[62–64]. As with the transactivation domain, the oxygen-dependent degradation domain of HIF1α seems to be intrinsically disordered[69], but it adopts an ordered β-strand-like structure at the exposed edge of a β-sheet on binding to the von-Hippel–Lindau tumour suppressor[69,70]. The interactions that are made by the hydroxyproline side chain are the primary determinant of specificity. The location of the Pro in a disordered region leaves it freely exposed for covalent modification and for signalling.

Protein interactions with the KID-binding (KIX) domain of CBP are also subject to regulation by post-translational modification. The KIX domain binds to a number of co-activators, including the KID of CREB[45]. The KID sequence must be phosphorylated at Ser133 (REF. 71) for high-affinity binding to the KIX domain: much of the binding energy for the pKID–KIX interaction is derived from specific electrostatic interactions of the covalently bound phosphate[72]. Both the phosphorylated and unphosphorylated forms of KID are unstructured in solution[39], but pKID folds into a pair of helices on binding to KIX[41] (BOX 2). In a similar manner as for the HIF1α transactivation domain, the intrinsic disorder in KID facilitates its interaction with a modifying enzyme, in this case the cAMP-dependent protein kinase A, which binds substrates that are in an extended conformation[73]. A systematic study of the sequence motifs that are associated with phosphorylation sites indicates that they occur predominantly within intrinsically disordered regions[7] (see also DISPHOS 1.3: Disorder-Enhanced Phosphorylation Sites Predictor in the online links box).

Other domains that interact with KIX are also intrinsically disordered in isolation: for example, the transcriptional-activation domain of Myb folds into a single helix on binding to the pKID-binding site of KIX[72], whereas the transactivation domain of the mixed lineage leukaemia (MLL) protein binds in an allosteric manner to a remote site on the KIX domain[74]. The N-terminal region of the HIV viral protein Tat, and the

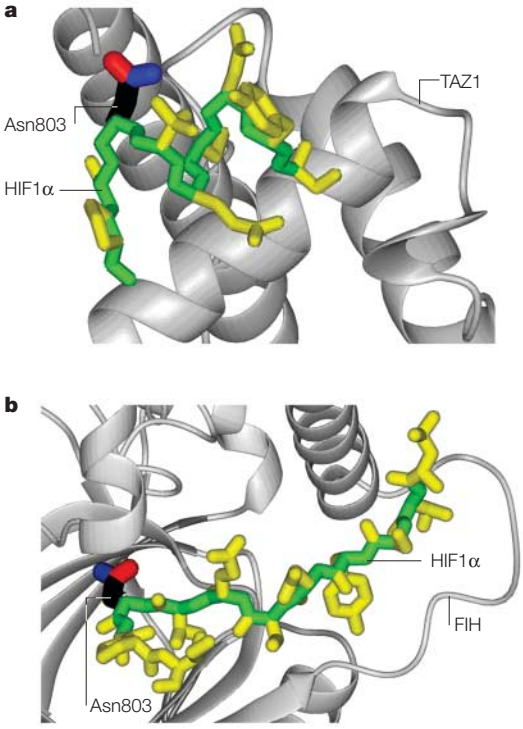


Figure 4 | **The structure of a HIF1α-interaction-domain fragment bound to two targets. a** | The structure of a fragment containing residues 798–805 of the HIF1α-interaction domain in complex with the TAZ1 domain of CBP[48] (Protein Data Bank (PDB) accession code 1L8C). In this structure, HIF1α has a regular helical backbone conformation (see the structural element that is highlighted in green). **b** | The identical fragment of HIF1α is shown bound to the asparagine hydroxylase FIH[66,126] (PDB accession code 1H2K). In this structure, HIF1α has an extended backbone conformation (again, see the structural element that is highlighted in green). The conserved Asn803 of HIF1α that functions as a hypoxic switch is highlighted in both parts of the figure (please refer to the main text for further details), whereas the other side chains of the HIF1α fragment are shown in yellow. This figure was prepared using MOLMOL[125]. CBP, cyclic-AMP-response-element-binding protein (CREB)-binding protein; FIH, factor inhibiting HIF; HIF1α, hypoxia-inducible factor-1α; TAZ1, transcriptional-adapter zinc-finger-1.

activation domains of the transcription factor Jun and the human T-cell leukaemia virus type-1 (HTLV-1) transcriptional activator Tax, also bind to KIX at the MLL site[75–77]. These domains are all disordered and fold into helical structures on binding to KIX. Post-translational modification of KIX itself — for example, methylation of Arg600 — modulates its affinity for pKID[78]. The KIX domain is of marginal stability[79] and becomes further destabilized on mutation of Arg600 (REF. 80). It is probable that methylation of Arg600 modulates the thermodynamic stability of KIX and therefore the affinity with which it binds its target transactivation domains. The KIX domain of CBP provides a scaffold for the integration of several signalling pathways, and this integration is achieved by numerous and various interactions with unstructured domains. Another well-characterized example of a folded domain that

SIGMOIDAL UNFOLDING CURVE
A folded protein with a stable three-dimensional structure unfolds cooperatively on addition of denaturant — that is, all of the molecules in the ensemble change from being fully folded to fully unfolded within a small range of denaturant concentration. This produces the sigmoidal unfolding curve that is typical for a folded protein.

HEAT-CAPACITY CHANGE
During thermal denaturation, a folded protein shows a typical bell-shaped transition in the plot of heat capacity versus temperature (measured, for example, by differential scanning calorimetry). Unfolded proteins do not show this behaviour.

SUMO
(small ubiquitin-like modifier). SUMO proteins are ubiquitin-like proteins that post-translationally modify proteins to control their localization and activity.

interacts with several different unstructured partners is β-catenin[81,82], which coordinates distinct signal-transduction pathways through interactions with: cadherins[57,83]; the transcription factor T-cell factor (TCF), which is the downstream effector of the Wnt signalling pathway [84,85]; the adenomatous polyposis coli (APC) tumour-suppressor protein[86]; an inhibitor of TCF, which is known as ICAT (inhibitor of β-catenin and TCF4)[87]; and p300 (REF. 88).

Synergistic folding and binding: NCBD and ACTR. In contrast to the well-ordered TAZ and KIX domains discussed above, the nuclear-receptor co-activator-binding domain (NCBD; also called IBiD to reflect its function in binding interferon regulatory factors[89]), which comprises residues 2,058–2,116 of human CBP (FIG. 2), is intrinsically disordered. Although the NCBD is highly helical, it has the characteristics of a molten globule rather than of a stably folded protein with an ordered tertiary structure[90] (FIG. 1). The ACTR (activator for thyroid hormone and retinoid receptors) domain of the p160 nuclear-receptor co-activator is also disordered in the free state (FIG. 1), and the NCBD and the ACTR domain undergo mutual synergistic folding on forming a complex[91]. The NCBD has a biased amino-acid composition. It is highly enriched in Gln, Pro, Ser and Leu residues, and cannot bury enough of its hydrophobic surface to form a stable globular structure in the absence of its binding partner (FIG. 1). This is unequivocally shown by the denaturation behaviour of isolated NCBD, which shows neither the SIGMOIDAL UNFOLDING CURVE nor the characteristic HEAT-CAPACITY CHANGE that accompany the unfolding of a globular protein[90,91].

Interactions of CBP/p300 with p53. CBP and p300 interact with p53, and have an important role in regulating its stability and transcriptional response[92]. In common with the CBP and p300 co-activators, p53 is modular and contains both structured and disordered domains. The N-terminal and C-terminal regions of p53, which contain the transcriptional-activation domain and a regulatory domain, respectively, are intrinsically disordered, both in the context of the intact protein[93,94] and in isolation[95]. The activity of these regions is tightly regulated by post-translational modification (phosphorylation, acetylation and ubiquitylation) at numerous sites[96,97] (BOX 4). The transcriptional-activation domain of p53 interacts with CBP and p300 through the folded TAZ2 domain[50–52], presumably by way of a coupled folding and binding process. The interaction of p53 with CBP is also mediated by the disordered C-terminal regulatory domain, which binds specifically to the bromodomain of CBP following the acetylation of Lys382 of p53 in response to DNA damage[98]. On binding to CBP, the p53 peptide folds into a β-turn conformation that allows side chains that are adjacent to the acetyllysine residue to make specific interactions with the CBP bromodomain (FIG. 5).

The linker regions of CBP. Inspection of FIG. 2 immediately highlights the highly biased amino-acid composition of the linker sequences that connect the structured domains of CBP. A characteristic feature of the linkers is that they are highly conserved between species in amino-acid composition, but not in sequence. There is a high proportion of polar residues (~70%) — a percentage that is retained from *Caenorhabditis elegans* to *Homo sapiens* — despite significant differences in linker length. Examination of individual linkers shows that they sometimes contain embedded regions that have a high sequence conservation and an increased content of charged or hydrophobic residues; some of these segments map to known interaction or regulatory motifs and the others are probable candidates for such functions. For example, the nuclear-receptor-interaction domain at the N terminus of vertebrate CBP/p300 is predicted to be disordered, but it contains a characteristic KHKXLXXLL motif (residues 66–74 of human CBP) that mediates binding to nuclear receptors[99–101]. A transcriptional-repression domain — the cell-cycle regulatory domain-1 (CRD1) — has been mapped to the C-terminal end of the disordered linker between the KIX domain and the bromodomain (between residues 1,019–1,082 in CBP, although the exact boundaries of this domain have not been defined)[102]. The CRD1 motif is rich in charged residues and contains two SUMO-modification sites that are required for transcriptional repression[103]. Finally, embedded within the HAT domain is a 60-residue disordered loop, as predicted by DISPROT[21] (see the online links box), in which more than 90% of the amino acids are charged or polar. This loop has recently been described as an autocatalytic switch[104], which regulates acetyltransferase activity using a mechanism that involves the auto-acetylation of several lysine residues that are contained within this loop.

Apart from these specific interaction motifs, the key function of the linkers in CBP/p300 seems to be to provide flexibility for the assembly of the transcriptional apparatus on promoters that contain numerous regulatory proteins. CBP and p300 have evolved as transcriptional scaffolds that provide binding sites for numerous transcriptional regulatory proteins, both within the structured domains and at specific recognition motifs in the disordered linker regions. Linkers of

Box 4 | Regulatory regions and post-translational modification

An important regulatory mechanism that has emerged in recent years involves the post-translational 'marking' of regulatory regions at numerous sites to form a 'code' that determines the biological response. The most well-known example of this is the histone code, in which histone tails are subject to modification by acetylation, phosphorylation, methylation and ubiquitylation, and which has a fundamental role in regulating access to DNA[118,119]. The N-terminal tails of histones are disordered in isolated histone proteins[120] and in the crystal structure of the nucleosome core particle[121]. Other regulatory regions that are subject to numerous covalent modifications, such as the N- and C-terminal regulatory domains of p53 (REFS 93,94), are also known to be intrinsically disordered. The intrinsic disorder and inherent flexibility in these regions is probably essential for their function: it leaves the side chains exposed for modification by several different enzymes — for example, kinases and phosphatases, acetyltransferases and deacetylases, methylases, and ubiquitin ligases — and provides the flexibility that is needed to allow adaptation to the varying conformational requirements of the active sites of these enzymes.

UBIQUITIN–PROTEASOME SYSTEM
Refers to the targeting of proteins for destruction by the 26S proteasome through the attachment of ubiquitin.

PEST-SEQUENCE MOTIF
This term refers to an amino-acid sequence that is enriched in Pro (P), Glu (E), Ser (S) and Thr (T) residues. PEST domains are frequently found in signalling, regulatory and adhesion molecules.

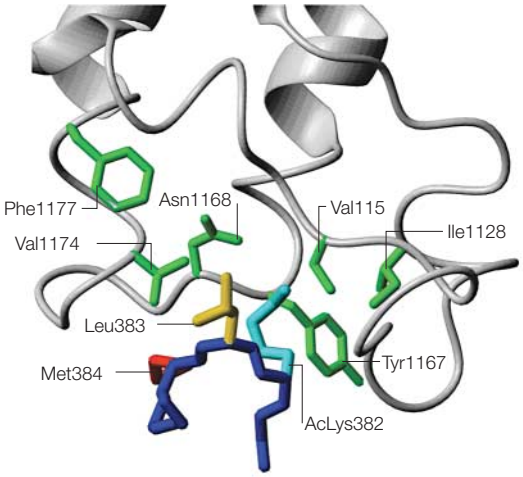


Figure 5 | **The structure of a CBP-bromodomain–acetyl-lysine complex.** A close-up view of the peptide-binding site in a complex of the bromodomain of CBP (grey) with a peptide (dark blue) that contains an acetyl-lysine (AcLys) residue (light blue) and was derived from the sequence of p53 (REF. 98; Protein Data Bank accession code 1JSP). On binding to CBP, the p53 peptide (only the portion of the peptide that interacts with the bromodomain is shown) folds into a β-turn conformation that allows the side chains that are adjacent to the AcLys residue (which are shown in red and yellow) to make specific interactions with the CBP bromodomain. Interacting residues on the CBP bromodomain are shown in green. This figure was prepared using MOLMOL[125]. CBP, cyclic-AMP-response-element-binding protein (CREB)-binding protein.

the appropriate length and composition seem to be essential for the correct assembly of large complexes[31]. Changes in the length or composition of the Pro- and Gln-repeat regions are thought to be a way of modulating transcriptional activation[105]. The presence of long linker regions in transcriptional-activator proteins such as CBP/p300 (FIG. 2) facilitates the assembly of the active transcriptional complex through interactions that occur over a distance scale of hundreds to thousands of angstroms. For example, a linker that is 400 residues in length — such as the one between the KIX domain and the bromodomain of CBP/p300 — has the ability to span distances that range from less than 100 Å to more than 1,500 Å. This gives the transcriptional activator the flexibility to interact with numerous transcriptional-regulatory factors that are bound with varying geometries over large promoter/enhancer regions (hundreds to thousands of base pairs). In addition, it is easy to envisage that flexible linker regions would facilitate the recruitment of other protein factors to the complex, through a mechanism known as 'fly-casting'[106].

CBP and p300 were recently described by Smith *et al.*[107] as "…'molecular interpreters' that can parse and/or conjugate the regulatory 'words', 'phrases', and 'sentences' of the genome." This function is undoubtedly facilitated by the intrinsically disordered linkers, which allow these scaffolding proteins to 'read' the genomic language that is encoded in the combinatorial arrangements of promoter-bound transcription factors.

More roles for intrinsically disordered proteins

An in-depth examination of all of the examples of functional unfolded proteins in processes such as transcription, translation, signal transduction and cell-cycle control will have to form the subject of a different review. However, to conclude this review, we survey a few areas in which intrinsically unstructured proteins have been found to have important roles.

Protein–nucleic-acid recognition. DNA-binding proteins seem to have evolved an ingenious array of techniques for dealing with the thermodynamic and kinetic challenges of binding to specific DNA sequences, and many of them involve the use of unfolded or partly folded proteins. A role for induced protein folding in sequence-specific DNA binding was proposed more than a decade ago by Spolar and Record[108], on the basis of the large heat-capacity changes that result from DNA–protein complex formation. Some of these processes involve the large-scale folding of entire domains — for example, the basic region of the basic leucine-zipper (bZip) DNA-binding domain[109] — whereas others involve the folding of local disordered loops or linkers between folded domains[32].

Many RNA-binding proteins also contain unstructured regions[2]. For example, the ribosomal protein L5 seems to associate with 5S ribosomal RNA by a mutual induced-fit mechanism: both RNA and protein are significantly more structured in the complex than in the free state[110].

Regulation through degradation. The relative instability of the intrinsically disordered proteins that are involved in transcription and signalling might provide a further level of cellular control through proteolytic degradation. The targeted degradation of transcriptional-activation domains by the UBIQUITIN–PROTEASOME SYSTEM might even have a direct role in transcriptional activation[111]. Targeted degradation also seems to have a role in the regulation of cadherins and therefore in the stabilization of β-catenin. The entire cytoplasmic domain of E-cadherin is intrinsically disordered and exposes PEST-SEQUENCE MOTIFS. These motifs function as signals for ubiquitin–proteasome-mediated degradation, but they become inaccessible on E-cadherin binding to β-catenin[57,83].

The functions of linker sequences demand that they are flexible and extended, and have a relatively high long-term stability. The low proportion of hydrophobic residues in linkers might be significant for their function. Misfolded proteins are targeted for destruction in the cell, and it is thought that the key to this targeting is the identification of solvated hydrophobic amino acids by cellular-destruction mechanisms such as the ubiquitin–proteasome pathway. Linker sequences must necessarily be unfolded, but need to be resistant to proteolysis. It is probable that the lack of hydrophobic residues in linker sequences is related to the requirement of an intrinsically unfolded protein segment for long-term stability. Furthermore, experimentally, it has been found that polyglutamine repeats are resistant to degradation by eukaryotic proteasomes[112], which is

POLYGLUTAMINE-REPEAT DISORDER
This term refers to the group of, frequently genetic, diseases that arise due to the expansion of regions of repeated glutamine sequences — for example, Huntington's disease.

consistent with the observed accumulation of such sequences in POLYGLUTAMINE-REPEAT DISORDERS. The particular amino-acid composition of linker sequences also stops them folding to form local structures that might interact nonspecifically with other proteins.

The biological 'cost' of disordered proteins

Disordered sequences are essential for the function of transcriptional activators and numerous other signalling and regulatory proteins. However, this function does not come without 'cost': these intrinsically disordered regions are the sites of many chromosomal translocations that are associated with disease. For example, translocations that fuse regions of CBP or p300 to segments of MOZ (monocytic zinc-finger leukaemia protein) or MLL are associated with human leukaemias[45,113]. The breakpoints in CBP/p300 for all of these translocations are located in the N-terminal disordered region or in the linker between the KIX domain and the bromodomain (FIG. 2). This phenomenon probably reflects the structural organization of the protein. Translocations in disordered regions leave the folded domains intact and thereby lead to a fusion protein with aberrant functions. By contrast, truncations or translocations in the regions of a gene that encode fully structured domains would almost certainly lead to the production of a misfolded protein, which would be rapidly degraded by the cellular machinery and would therefore not produce a disease phenotype.

Disordered regions can also have a biological cost in terms of the promotion and proliferation of protein-folding diseases. Neurodegenerative diseases such as prion diseases or Parkinson's disease are associated with intrinsically disordered proteins. A number of transcription factors, many of which contain glutamine-rich motifs, have been implicated in polyglutamine diseases[114]. For example, human CBP, which contains a stretch of 18 Gln residues, is sequestered in polyglutamine aggregates in the brains of patients with Huntington's disease, and it has been proposed that the disease might result from interference with the normal transcriptional functions of CBP[115]. A recent survey found that human proteins that have numerous tracts of extremely low complexity, most commonly consisting of Gln, Ser or acidic residues, are often associated with neurological diseases and cancer[116].

Implications and future directions

Clearly, much work still needs to be done to characterize functional unfolded proteins. The advent of powerful computational methods to screen protein sequences, and even the sequences of entire genomes, for intrinsic disorder will undoubtedly reveal many more proteins that belong to this class. Parallel progress in functional genomics will advance our understanding of the functions of these disordered regions of proteins. The role of sequences of low complexity that are between structured domains is just beginning to be addressed. In addition, the degree of polypeptide mobility that is consistent with (or required for) protein function is still unclear. It is probable that the present concepts of proteins and their functions will evolve to encompass a continuum; from fully unstructured proteins that fold on binding their target, through strings of protein domains that assemble on binding their target, to relatively rigid proteins with mobile functional regions. Our concept of a functional protein must therefore evolve from a static picture to a highly dynamic one, in which several conformations that are consistent with various aspects of function are represented.

1. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradovic, Z. Intrinsic disorder and protein function. *Biochemistry* **41**, 6573–6582 (2002).
2. Uversky, V. N. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **11**, 739–756 (2002).
Provides a comprehensive review of the field of unfolded proteins with much valuable analysis. It contains the most up-to-date table of unfolded proteins.
3. Boesch, C., Bundi, A., Oppliger, M. & Wüthrich, K. ^{1}H nuclear-magnetic-resonance studies of the molecular conformation of monomeric glucagon in aqueous solution. *Eur. J. Biochem.* **91**, 209–214 (1978).
4. Daniels, A. J., Williams, R. J. P. & Wright, P. E. The character of the stored molecules in chromaffin granules of the adrenal medulla: a nuclear magnetic resonance study. *Neuroscience* **3**, 573–585 (1978).
5. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999).
6. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z. & Dunker, A. K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **323**, 573–584 (2002).
7. Iakoucheva, L. M. *et al*. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **32**, 1037–1049 (2004).
8. Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* **27**, 527–533 (2002).
9. Romero, P. *et al*. Sequence complexity of disordered protein. *Proteins* **42**, 38–48 (2001).
10. Vucetic, S., Brown, C. J., Dunker, A. K. & Obradovic, Z. Flavors of protein disorder. *Proteins* **52**, 573–584 (2003).
11. Mitchell, P. J. & Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245**, 371–378 (1989).
12. O'Hare, P. & Williams, G. Structural studies of the acidic transactivation domain of the Vmw65 protein of herpes simplex virus using ^{1}H NMR. *Biochemistry* **31**, 4150–4156 (1992).
13. Longhi, S. *et al*. The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. *J. Biol. Chem.* **278**, 18638–18648 (2003).
14. Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E. & Dunker, A. K. Identifying disordered regions in proteins from amino acid sequences. *Proc. IEEE Int. Conf. Neural Netw.* **1**, 90–95 (1997).
15. Uversky, V. N., Gillespie, J. R. & Fink, A. L. Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins* **41**, 415–427 (2000).
16. Linding, R. *et al*. Protein disorder prediction: implications for structural proteomics. *Structure (Camb.)* **11**, 1453–1459 (2003).
17. Linding, R., Russell, R. B., Neduva, V. & Gibson,T. J. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* **31**, 3701–3708 (2003).
18. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645 (2004).
19. Weathers, E. A., Paulaitis, M. E., Woolf, T. B. & Hoh, J. H. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Lett.* **576**, 348–352 (2004).
20. Dunker, A. K. *et al*. Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59 (2001).
21. Vucetic, S. *et al*. DisProt: a database of protein disorder. *Bioinformatics* **21**, 137–140 (2005).
22. Tompa, P. Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays* **25**, 847–855 (2003).
23. Dyson, H. J. & Wright, P. E. Nuclear magnetic resonance methods for elucidation of structure and dynamics of disordered states. *Methods Enzymol.* **339**, 258–270 (2001).
24. Dyson, H. J. & Wright, P. E. Unfolded proteins and protein folding studied by NMR. *Chem. Rev.* **104**, 3607–3622 (2004).
25. Rose, G. D. in *Advances in Protein Chemistry* Vol. 62, (eds Richards, F. M., Eisenberg, D. S. & Kuriyan, J.) (Academic Press, San Diego, 2002).
This book provides an excellent introduction to the biophysical techniques that can be used to identify and characterize unfolded proteins and protein domains.
26. Dyson, H. J. & Wright, P. E. Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance. *Adv. Protein Chem.* **62**, 311–340 (2002).
27. Rucker, A. L., Pager, C. T., Campbell, M. N., Qualls, J. E. & Creamer, T. P. Host-guest scale of left-handed polyproline II helix formation. *Proteins* **53**, 68–75 (2003).
28. Shi, Z., Olson, C. A., Rose, G. D., Baldwin, R. L. & Kallenbach, N. R. Polyproline II structure in a sequence of seven alanine residues. *Proc. Natl Acad. Sci. USA* **99**, 9190–9195 (2002).
29. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C. & Brown, C. J. Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* **11**, 161–171 (2000).
30. Wootton, J. C. & Drummond, M. H. The Q-linker: a class of interdomain sequences found in bacterial multidomain regulatory proteins. *Protein Eng. Des. Sel.* **2**, 535–543 (1989).

31. Zhou, H. X. Quantitative account of the enhanced affinity of two linked scFvs specific for different epitopes on the same antigen. *J. Mol. Biol.* **329**, 1–8 (2003).
32. Laity, J. H., Dyson, H. J. & Wright, P. E. DNA-induced α-helix capping in conserved linker sequences is a determinant of binding affinity in Cys_2-His_2 zinc fingers. *J. Mol. Biol.* **295**, 719–727 (2000).
33. Young, M. A., Gonfloni, S., Superti-Furga, G., Roux, B. & Kuriyan, J. Dynamic coupling between the SH2 and SH3 domains of c-Src and Hck underlies their inactivation by C-terminal tyrosine phosphorylation. *Cell* **105**, 115–126 (2001).
34. Laity, J. H., Dyson, H. J. & Wright, P. E. Molecular basis for modulation of biological function by alternate splicing of the Wilms' tumor suppressor protein. *Proc. Natl Acad. Sci. USA* **97**, 11932–11935 (2000).
35. Namba, K. Roles of partly unfolded conformations in macromolecular self-assembly. *Genes Cells* **6**, 1–12 (2001).
36. Tompa, P. & Csermely, P. The role of structural disorder in the function of RNA and protein chaperones. *FASEB J.* **18**, 1169–1175 (2004).
37. Dyson, H. J. & Wright, P. E. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* **12**, 54–60 (2002).
Provides a comprehensive survey of protein folding and binding events.
38. Demchenko, A. P. Recognition between flexible protein molecules: induced and assisted folding. *J. Mol. Recognit.* **14**, 42–61 (2001).
39. Radhakrishnan, I., Pérez-Alvarado, G. C., Dyson, H. J. & Wright, P. E. Conformational preferences in the Ser^{133}-phosphorylated and non-phosphorylated forms of the kinase inducible transactivation domain of CREB. *FEBS Lett.* **430**, 317–322 (1998).
40. Richards, J. P., Bächinger, H. P., Goodman, R. H. & Brennan, R. G. Analysis of the structural properties of cAMP-responsive element-binding protein (CREB) and phosphorylated CREB. *J. Biol. Chem.* **271**, 13716–13723 (1996).
41. Radhakrishnan, I. *et al*. 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 (1997).
42. Zhou, P., Lugovskoy, A. A., McCarty, J. S., Li, P. & Wagner, G. Solution structure of DFF40 and DFF45 N-terminal domain complex and mutual chaperone activity of DFF40 and DFF45. *Proc. Natl Acad. Sci. USA* **98**, 6051–6055 (2001).
43. Giordano, A. & Avantaggiati, M. L. p300 and CBP: partners for life and death. *J. Cell. Physiol.* **181**, 218–230 (1999).
44. Blobel, G. A. CREB-binding protein and p300: molecular integrators of hematopoietic transcription. *Blood* **95**, 745–755 (2000).
45. Goodman, R. H. & Smolik, S. CBP/p300 in cell growth, transformation, and development. *Genes Dev.* **14**, 1553–1577 (2000).
46. Ponting, C. P., Blake, D. J., Davies, K. E., Kendrick-Jones, J. & Winder, S. J. ZZ and TAZ: new putative zinc fingers in dystrophin and other proteins. *Trends Biochem. Sci.* **21**, 11–13 (1996).
47. De Guzman, R. N., Liu, H. Y., Martinez-Yamout, M., Dyson, H. J. & Wright, P. E. Solution structure of the TAZ2 (CH3) domain of the transcriptional adaptor protein CBP. *J. Mol. Biol.* **303**, 243–253 (2000).
48. Dames, S. A., Martinez-Yamout, M., De Guzman, R. N., Dyson, H. J. & Wright, P. E. Structural basis for Hif-1α/CBP recognition in the cellular hypoxic response. *Proc. Natl Acad. Sci. USA* **99**, 5271–5276 (2002).
49. Eckner, R. *et al*. Molecular cloning and functional analysis of the adenovirus E1A-associated 300-kD protein (p300) reveals a protein with properties of a transcriptional adaptor. *Genes Dev.* **8**, 869–884 (1994).
50. Avantaggiati, M. *et al*. Recruitment of p300/CBP in p53-dependent signal pathways. *Cell* **89**, 1175–1184 (1997).
51. Lill, N. L., Grossman, S. R., Ginsberg, D., DeCaprio, J. & Livingston, D. M. Binding and modulation of p53 by p300/CBP coactivators. *Nature* **387**, 823–827 (1997).
52. Gu, W., Shi, X. L. & Roeder, R. G. Synergistic activation of transcription by CBP and p53. *Nature* **387**, 819–823 (1997).
53. Arany, Z. *et al*. An essential role for p300/CBP in the cellular response to hypoxia. *Proc. Natl Acad. Sci. USA* **93**, 12969–12973 (1996).
54. Freedman, S. J. *et al*. Structural basis for recruitment of CBP/p300 by hypoxia-inducible factor-1α. *Proc. Natl Acad. Sci. USA* **99**, 5367–5372 (2002).
55. Gunasekaran, K., Tsai, C. J., Kumar, S., Zanuy, D. & Nussinov, R. Extended disordered proteins: targeting function with less scaffold. *Trends Biochem. Sci.* **28**, 81–85 (2003).
Points out that the extensive interfaces observed when unfolded proteins bind their targets would only be possible for fully structured proteins if they were much larger. However, increased protein size would result in greater macromolecular crowding or would require cells to be larger.
56. Russo, A. A., Jeffrey, P. D., Patten, A. K., Massagué, J. & Pavletich, N. P. Crystal structure of the $p27^{Kip1}$ cyclin-dependent-kinase inhibitor bound to the cyclin A–Cdk2 complex. *Nature* **382**, 325–331 (1996).
57. Huber, A. H. & Weis, W. I. The structure of the β-catenin/E-cadherin complex and the molecular basis of diverse ligand recognition by β-catenin. *Cell* **105**, 391–402 (2001).
58. Sorenson, M. K., Ray, S. S. & Darst, S. A. Crystal structure of the flagellar σ/anti-σ complex σ^{28}/FlgM reveals an intact σ factor in an inactive conformation. *Mol. Cell* **14**, 127–138 (2004).
59. Bhattacharya, S. *et al*. Functional role of p35srj, a novel p300/CBP binding protein, during transactivation by HIF-1. *Genes Dev.* **13**, 64–75 (1999).
60. Freedman, S. J. *et al*. Structural basis for negative regulation of hypoxia-inducible factor-1α by CITED2. *Nature Struct. Biol.* **10**, 504–512 (2003).
61. De Guzman, R. N., Martinez-Yamout, M., Dyson, H. J. & Wright, P. E. Interaction of the TAZ1 domain of CREB-binding protein with the activation domain of CITED2: regulation by competition between intrinsically unstructured ligands for non-identical binding sites. *J. Biol. Chem.* **279**, 3042–3049 (2004).
62. Jaakkola, P. *et al*. Targeting of HIF-α to the von Hippel–Lindau ubiquitylation complex by O_2-regulated prolyl hydroxylation. *Science* **292**, 468–472 (2001).
63. Ivan, M. *et al*. HIFα targeted for VHL-mediated destruction by proline hydroxylation: implications for O_2 sensing. *Science* **292**, 464–468 (2001).
64. Yu, F., White, S. B., Zhao, Q. & Lee, F. S. HIF-1α binding to VHL is regulated by stimulus-sensitive proline hydroxylation. *Proc. Natl Acad. Sci. USA* **98**, 9630–9635 (2001).
65. Lando, D., Peet, D. J., Whelan, D. A., Gorman, J. J. & Whitelaw, M. L. Asparagine hydroxylation of the HIF transactivation domain: a hypoxic switch. *Science* **295**, 858–861 (2002).
66. Elkins, J. M. *et al*. Structure of factor-inhibiting hypoxia-inducible factor (HIF) reveals mechanism of oxidative modification of HIF-1α. *J. Biol. Chem.* **278**, 1802–1806 (2003).
67. Dann, C. E. III, Bruick, R. K. & Deisenhofer, J. Structure of factor-inhibiting hypoxia-inducible factor 1: an asparaginyl hydroxylase involved in the hypoxic response pathway. *Proc. Natl Acad. Sci. USA* **99**, 15351–15356 (2002).
68. Kriwacki, R. W., Hengst, L., Tennant, L., Reed, S. I. & Wright, P. E. Structural studies of $p21^{Waf1/Cip1/Sdi1}$ in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl Acad. Sci. USA* **93**, 11504–11509 (1996).
69. Hon, W. *et al*. Structural basis for the recognition of hydroxyproline in HIF-1α by pVHL. *Nature* **417**, 975–978 (2002).
70. Min, J. H. *et al*. Structure of an HIF-1α–pVHL complex: hydroxyproline recognition in signaling. *Science* **296**, 1886–1889 (2002).
71. Chrivia, J. C. *et al*. Phosphorylated CREB binds specifically to nuclear protein CBP. *Nature* **365**, 855–859 (1993).
72. Zor, T., Mayr, B. M., Dyson, H. J., Montminy, M. R. & Wright, P. E. Roles of phosphorylation and helix propensity in the binding of the KIX domain of CREB-binding protein by constitutive (c-Myb) and inducible (CREB) activators. *J. Biol. Chem.* **277**, 42241–42248 (2002).
73. Knighton, D. R. *et al*. Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* **253**, 414–420 (1991).
74. Goto, N. K., Zor, T., Martinez-Yamout, M., Dyson, H. J. & Wright, P. E. Cooperativity in transcription factor binding to the coactivator CREB-binding protein (CBP). The mixed lineage leukemia protein (MLL) activation domain binds to an allosteric site on the Kix domain. *J. Biol. Chem.* **277**, 43168–43174 (2002).
75. Vendel, A. C. & Lumb, K. J. NMR mapping of the HIV-1 Tat interaction surface of the KIX domain of the human coactivator CBP. *Biochemistry* **43**, 904–908 (2004).
76. Campbell, K. M. & Lumb, K. J. Structurally distinct modes of recognition of the KIX domain of CBP by Jun and CREB. *Biochemistry* **41**, 13956–13964 (2002).
77. Vendel, A. C., McBryant, S. J. & Lumb, K. J. KIX-mediated assembly of the CBP–CREB–HTLV-1 tax coactivator-activator complex. *Biochemistry* **42**, 12481–12487 (2003).
78. Xu, W. *et al*. A transcriptional switch mediated by cofactor methylation. *Science* **294**, 2507–2511 (2001).
79. Radhakrishnan, I. *et al*. Structural analyses of CREB–CBP transcriptional activator–coactivator complexes by NMR spectroscopy: implications for mapping the boundaries of structural domains. *J. Mol. Biol.* **287**, 859–865 (1999).
80. Wei, Y., Horng, J. C., Vendel, A. C., Raleigh, D. P. & Lumb, K. J. Contribution to stability and folding of a buried polar residue at the CARM1 methylation site of the KIX domain of CBP. *Biochemistry* **42**, 7044–7049 (2003).
81. Shapiro, L. β-catenin and its multiple partners: promiscuity explained. *Nature Struct. Biol.* **8**, 484–487 (2001).
82. Daniels, D. L., Eklof, S. K. & Weis, W. I. β-catenin: molecular plasticity and drug design. *Trends Biochem. Sci.* **26**, 672–678 (2001).
83. Huber, A. H., Stewart, D. B., Laurents, D. V., Nelson, W. J. & Weis, W. I. The cadherin cytoplasmic domain is unstructured in the absence of β-catenin. A possible mechanism for regulating cadherin turnover. *J. Biol. Chem.* **276**, 12301–12309 (2001).
84. Graham, T. A., Weaver, C., Mao, F., Kimelman, D. & Xu, W. Crystal structure of a β-catenin/Tcf complex. *Cell* **103**, 885–896 (2000).
85. Graham, T. A., Ferkey, D. M., Mao, F., Kimelman, D. & Xu, W. Tcf4 can specifically recognize β-catenin using alternative conformations. *Nature Struct. Biol.* **8**, 1048–1052 (2001).
86. Eklof Spink, K., Fridman, S. G. & Weis, W. I. Molecular mechanisms of β-catenin recognition by adenomatous polyposis coli revealed by the structure of an APC–β-catenin complex. *EMBO J.* **20**, 6203–6212 (2001).
87. Graham, T. A., Clements, W. K., Kimelman, D. & Xu, W. The crystal structure of the β-catenin/ICAT complex reveals the inhibitory mechanism of ICAT. *Mol. Cell* **10**, 563–571 (2002).
88. Daniels, D. L. & Weis, W. I. ICAT inhibits β-catenin binding to Tcf/Lef-family transcription factors and the general coactivator p300 using independent structural modules. *Mol. Cell* **10**, 573–584 (2002).
89. Lin, C. H. *et al*. A small domain of cbp/p300 binds diverse proteins. Solution structure and functional studies. *Mol. Cell* **8**, 581–590 (2001).
90. Demarest, S. J., Deechongkit, S., Dyson, H. J., Evans, R. M. & Wright, P. E. Packing, specificity, and mutability at the binding interface between the p160 coactivator and CREB-binding protein. *Protein Sci.* **13**, 203–210 (2004).
91. Demarest, S. J. *et al*. Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature* **415**, 549–553 (2002).
Presents the first example of two domains that are largely, if not completely, unfolded when free in solution, but that fold together when they interact to produce a complex of exceptionally high specificity and with a large surface area of binding.
92. Grossman, S. R. p300/CBP/p53 interaction and regulation of the p53 response. *Eur. J. Biochem.* **268**, 2773–2778 (2001).
93. Bell, S., Klein, C., Muller, L., Hansen, S. & Buchner, J. p53 contains large unstructured regions in its native state. *J. Mol. Biol.* **322**, 917–927 (2002).
94. Ayed, A. *et al*. Latent and active p53 are identical in conformation. *Nature Struct. Biol.* **8**, 756–760 (2001).
Provides important information on the N- and C-terminal domains of p53, which are unstructured in the context of the full-length protein, as well as isolated in solution.
95. Dawson, R. *et al*. The N-terminal domain of p53 is natively unfolded. *J. Mol. Biol.* **332**, 1131–1141 (2003).
96. Prives, C. & Hall, P. A. The p53 pathway. *J. Pathol.* **187**, 112–126 (1999).
97. Alarcon-Vargas, D. & Ronai, Z. p53–Mdm2 — the affair that never ends. *Carcinogenesis* **23**, 541–547 (2002).
98. Mujtaba, S. *et al*. Structural mechanism of the bromodomain of the coactivator CBP in p53 transcriptional activation. *Mol. Cell* **13**, 251–263 (2004).
99. Kamei, Y. *et al*. A CBP integrator complex mediates transcriptional activation and AP-1 inhibition by nuclear receptors. *Cell* **85**, 403–414 (1996).
100. Heery, D. M., Kalkhoven, E., Hoare, S. & Parker, M. G. A signature motif in transcriptional co-activators mediates binding to nuclear receptors. *Nature* **387**, 733–736 (1997).
101. Darimont, B. D. *et al*. Structure and specificity of nuclear receptor–coactivator interactions. *Genes Dev.* **12**, 3343–3356 (1998).
102. Snowden, A. W., Anderson, L. A., Webster, G. A. & Perkins, N. D. A novel transcriptional repression domain mediates $p21^{WAF1/CIP1}$ induction of p300 transactivation. *Mol. Cell. Biol.* **20**, 2676–2686 (2000).
103. Girdwood, D. *et al*. p300 transcriptional repression is mediated by SUMO modification. *Mol. Cell* **11**, 1043–1054 (2003).
104. Thompson, P. R. *et al*. Regulation of the p300 HAT domain via a novel activation loop. *Nature Struct. Mol. Biol.* **11**, 308–315 (2004).
105. Gerber, H. P. *et al*. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263**, 808–811 (1994).

106. Shoemaker, B. A., Portman, J. J. & Wolynes, P. G. Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl Acad. Sci. USA* **97**, 8868–8873 (2000).
Describes a possible rationale for the presence of unstructured linker regions in multidomain proteins: a conformational ensemble for part of an interaction domain ensures that the volume of the surrounding solution is sampled extensively, which increases the likelihood of encountering the binding partner.
107. Smith, J. L. *et al*. Kinetic profiles of p300 occupancy *in vivo* predict common features of promoter structure and coactivator recruitment. *Proc. Natl Acad. Sci. USA* **101**, 11554–11559 (2004).
108. Spolar, R. S. & Record, M. T. Coupling of local folding to site-specific binding of proteins to DNA. *Science* **263**, 777–784 (1994).
109. Patel, L., Abate, C. & Curran, T. Altered protein conformation on DNA binding by Fos and Jun. *Nature* **347**, 572–574 (1990).
110. DiNitto, J. P. & Huber, P. W. Mutual induced fit binding of *Xenopus* ribosomal protein L5 to 5S rRNA. *J. Mol. Biol.* **330**, 979–992 (2003).
111. Salghetti, S. E., Caudy, A. A., Chenoweth, J. G. & Tansey, W. P. Regulation of transcriptional activation domain function by ubiquitin. *Science* **293**, 1651–1653 (2001).
112. Venkatraman, P., Wetzel, R., Tanaka, M., Nukina, N. & Goldberg, A. L. Eukaryotic proteasomes cannot digest polyglutamine sequences and release them during degradation of polyglutamine-containing proteins. *Mol. Cell* **14**, 95–104 (2004).
113. Yang, X. J. The diverse superfamily of lysine acetyltransferases and their roles in leukemia and other diseases. *Nucleic Acids Res.* **32**, 959–976 (2004).
114. McCampbell, A. & Fischbeck, K. H. Polyglutamine and CBP: fatal attraction? *Nature Med.* **7**, 528–530 (2001).
115. Nucifora, F. C. *et al*. Interference by Huntingtin and atrophin-1 with CBP-mediated transcription leading to cellular toxicity. *Science* **291**, 2423–2428 (2001).
116. Karlin, S., Brocchieri, L., Bergman, A., Mrazek, J. & Gentles, A. J. Amino acid runs in eukaryotic proteomes and disease associations. *Proc. Natl Acad. Sci. USA* **99**, 333–338 (2002).
117. Rost, B. & Liu, J. The PredictProtein server. *Nucleic Acids Res.* **31**, 3300–3304 (2003).
118. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
119. Berger, S. L. Histone modifications in transcriptional regulation. *Curr. Opin. Genet. Dev.* **12**, 142–148 (2002).
120. Hartman, P. G., Chapman, G. E., Moss, T. & Bradbury, E. M. Studies on the role and mode of operation of the very-lysine-rich histone H1 in eukaryote chromatin. The three structural regions of the histone H1 molecule. *Eur. J. Biochem.* **77**, 45–51 (1977).
121. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
122. Brüschweiler, R., Liao, X. & Wright, P. E. Long-range motional restrictions in a multidomain zinc-finger protein from anisotropic tumbling. *Science* **268**, 886–889 (1995).
123. Gross, J. D. *et al*. Ribosome loading onto the mRNA cap is driven by conformational coupling between eIF4G and eIF4E. *Cell* **115**, 739–750 (2003).
124. Wuttke, D. S., Foster, M. P., Case, D. A., Gottesfeld, J. M. & Wright, P. E. Solution structure of the first three zinc fingers of TFIIIA bound to the cognate DNA sequence: determinants of affinity and sequence specificity. *J. Mol. Biol.* **273**, 183–206 (1997).
125. Koradi, R., Billeter, M. & Wüthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graphics* **14**, 51–55 (1996).
126. Lee, C., Kim, S. J., Jeong, D. G., Lee, S. M. & Ryu, S. E. Structure of human FIH-1 reveals a unique active site pocket and interaction sites for HIF-1 and von Hippel–Lindau. *J. Biol. Chem.* **278**, 7558–7563 (2003).

Acknowledgments

We would like to thank past and present members of our laboratories for their contributions to the ideas that are expressed in this review. We are particularly grateful to M. Martinez-Yamout for continuing important contributions and for critically reading the manuscript. Our work is supported by grants from the National Institutes of Health.

Competing interests statement

The authors declare no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:
Interpro: http://www.ebi.ac.uk/interpro/
bromodomain | HAT | KID | KIX | NCBD | PHD | TAZ1 | TAZ2 | ZZ
Protein Data Bank: http://www.rcsb.org/pdb/
1H2K | 1JSP | 1KBH | 1L8C | 1RF8 | 1R8U | 1TF3
Swiss-Prot: http://us.expasy.org/sprot/
β-catenin | CBP | CITED2 | CREB | DFF40 | DFF45 | Hck | HIF1α | $p27^{Kip1}$ | p53 | p300 | Src | WT1

FURTHER INFORMATION

DisEMBL — Intrinsic Protein Disorder Prediction 1.4: http://dis.embl.de
DISPHOS 1.3 — Disorder-Enhanced Phosphorylation Sites Predictor: http://core.ist.temple.edu/pred/pred.html
DISPROT — Database of Protein Disorder: http://divac.ist.temple.edu/disprot/database.php
FoldIndex: http://bip.weizmann.ac.il/fldbin/findex
GLOBPROT 2 — Intrinsic Protein Disorder, Domain & Globularity Prediction: http://globplot.embl.de
PONDR (Predictors of Natural Disordered Regions): http://www.pondr.com/
The DISOPRED2 Disorder Prediction Server: http://bioinf.cs.ucl.ac.uk/disopred/
The PredictProtein server: http://www.embl-heidelberg.de/predictprotein/predictprotein.html
Access to this links box is available online.