doi:10.1016/S0022-2836(02)00736-2 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 322, 53–64

JMB

Loopy Proteins Appear Conserved in Evolution

Jinfeng Liu[1,2,3], Hepan Tan[2] and Burkhard Rost[2,3,4]*

[1]*Department of Pharmacology Columbia University, 630 West 168th Street, New York, NY 10032, USA*

[2]*Department of Biochemistry and Molecular Biophysics Columbia University, 650 West 168th Street BB217, New York NY 10032, USA*

[3]*North East Structural Genomics Consortium (NESG) Department of Biochemistry and Molecular Biophysics Columbia University, 650 West 168th Street BB217, New York NY 10032, USA*

[4]*Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St Nicholas Avenue, New York NY 10032, USA*

**Corresponding author*

Over the last decade, structural biologists have unravelled many proteins that appear natively disordered. Common assumptions are that many of these proteins adopt structure through binding and that the structural flexibility enables them to adopt different functions. Here, we investigated regions of more than 70 sequence-consecutive residues that have no regular secondary structure (NORS). Analysing 31 entirely sequenced organisms, we predicted five times as many proteins with NORS regions (loopy proteins) in eukaryotes (20%) than in prokaryotes and archaeas (4%). Thousands of these NORS regions were over 150 residues long. The amino acid composition of NORS regions differed from that of loops in PDB. Although NORS proteins had significantly more residues in low-complexity regions than other proteins, simple cut-off thresholds for sequence bias missed most NORS regions. On average, NORS regions were evolutionarily at least as conserved as their flanking regions. Furthermore, yeast proteins with NORS regions had more protein–protein interaction partners than other proteins. Regulatory and transcription-related functions were over-represented in loopy proteins, biosynthesis and energy metabolism were under-represented. Overall, our analysis confirmed that proteins with non-regular structures appear to play important functional roles, and they may adopt as yet unknown types of protein structures.

Keywords: disordered regions; protein function; protein–protein interactions; natively unstructured proteins; regular secondary structure

Introduction

Protein function may require flexible structures

The sequence of a protein largely determines its three-dimensional (3D) structure,[1–3] and structure often determines function.[4–13] Nevertheless, many proteins undergo changes in conformation upon binding to substrates or other ligands.[6,14–16] Some biological functions may require more flexible structures than others: for the catalytic activity of an enzyme the precise interaction between enzyme and substrate may be critical.[17,18] On the other hand, a structure that is intrinsically more flexible, more "loopy", may adapt more readily to different environments. Consequently, loopy structures may recognise many different biological targets.[19–28] For example, the serine/threonine phosphatase calcineurin becomes activated by binding a Ca^{2+}-calmodulin complex by a region that exists

Abbreviations used: 3D structure, three-dimensional structure, i.e. co-ordinates of all residues/atoms in a protein; COILS, prediction of coiled-coil regions from sequence on the basis of statistics and expert rules; DIP, database of interacting proteins; DSSP, automatic assignment of secondary structure and solvent accessibility from 3D co-ordinates; NORS, segment of more than 70 consecutive residues of NO regular secondary structure, i.e. without helix or strand (more precisely, we required that less than 12% of the residues in the respective region were in helix or strand and that at least one region of more than ten residues was exposed to solvent); NORS proteins, proteins with at least one NORS region; ORF, open reading frame (protein predicted by genome sequencing project); PDB, protein data bank of protein structures; PDBsub, sequence-unique subset of PDB with 1947 chains; PHDacc, profile-based neural network prediction of solvent accessibility; PHDsec, profile-based neural network prediction of secondary structure; SignalP, neural network-based prediction of signal peptides; SWISS-PROT, curated database with protein sequences and functional annotations; TrEMBL, automatic translation of EMBL nucleotide database of protein sequences.

E-mail address of the corresponding author: rost@columbia.edu

0022-2836/02/$ - see front matter

as a disordered ensemble.[29–31] Loopy structures appear to be important for macromolecular assembly, as exemplified by the assembly of the tobacco mosaic virus or the bacterial flagellum.[32]

Identifying disordered regions *in silico*

One class of "natively disordered" regions was identified initially through the observation that such regions are invisible in electron density maps, since the disorder prevented them from crystallising into well-ordered structures that scatter X-rays coherently. These regions appear to be characterised frequently by a particular bias in the use of amino acids, usually referred to as compositional bias or regions of low sequence complexity.[33–36] Romero and colleagues developed a method that predicts disordered regions by training a neural network to identify low-complexity regions longer than 40 residues.[37,38] Applying their method to the SWISS-PROT database,[39] they have found more than 15,000 protein regions that are putatively disordered.[37]

Here, we studied the problem of disordered proteins from a more structure-oriented perspective. We investigated regions of more than 70 residues that have very low content in regular secondary structure (helix or strand). These extended regions of no regular secondary structure (NORS) may still be sufficiently ordered to diffract X-rays and yield electron density maps. However, their lack of regular secondary structure is certainly intriguing. We found NORS regions to be particularly abundant in eukaryotic proteomes, to be evolutionarily conserved, and to be enriched in regulatory functions and in protein–protein interactions.

Results

Analysing loopy proteins in PDB

Visual classification into four types

We defined NORS regions as having at least one sequence-continuous fragment of over 70 residues with fewer than 12% of residues in regular secondary structure (helix or strand). We found less than 20 such proteins in a sequence-unique subset of PDB,[40] and then visually sorted these NORS proteins into types according to the structural context. Note that these types were by no means objective, i.e. were not based on a definition enabling automatic classification. We distinguished the following four types (Figure 1). (1) Connecting loops (Figure 1(a)) are long loops that connect structural domains or linked subunits (e.g. 1AA6,[41] 1BF2,[42] and 4DPV[43]). (2) Loopy ends (Figure 1(b)), i.e. long N-terminal or C-terminal loops (e.g. 1DHX,[44] 1B35 C chain,[45] and 1B0P[46]). (3) Wrapping loops (Figure 1(c)) are long loops wrapping around otherwise'normal globular domains (e.g. 2BAA,[47] 7CAT,[48,49] and 1CPO[50]). (4) Loopy domains (Figure 1(d)) are entire proteins or domains lacking regular secondary structure (e.g. 1TBI[51] and 1TAC[52]).

Functional reasons for NORS regions

Most NORS regions in PDB have enzymatic activities and/or are involved in substrate/ligand processes. For example, the turnover number of pyruvate ferredoxin oxidoredisoamylase (1B0P, A chain) increases by a factor of at least 5 upon reduction and complete removal of the loop.[46] A molecular dynamics simulation has indicated that the removal of the C-terminal domain VII may result in the formation of a new hydrophobic channel of only 7 Å. This channel could bridge the active site close to the molecular surface and could serve to evacuate reaction products. The observed increase in activity of the reduced enzyme compared to the oxidised form may be due to an easier flow of substrates and products toward and from the active site. For chloroperoxidase (1CPO)[50] and for the class II chitinases (2BAA),[47] the long loops are involved in substrate binding. The RGD loop is the key site for adhesive recognition and receptor interaction in HIVZ2 Tat protein (1TAC).[52] The loop of the sea raven type II antifreeze protein (SRAFP) (2AFP)[53] is part of the ice-binding site. The loops share inhibitor-binding and DNA-binding capabilities in carboxypeptidase (6CPA).[54]

Errors in predicting NORS regions

We optimised our definition of predicted NORS regions to yield a low false-positive rate when applying the method to PDB (Supplementary Material, Table S1): the predicted content in regular secondary structure (helix or strand) is below 12% over at least 70 consecutive residues, and at least ten consecutive residues are predicted to be exposed. On the basis of these criteria, we predicted NORS regions for 23 proteins from our sequence-unique subset of PDB. Five of these were identified when using the DSSP[55] assignments of the actual rather than the predicted secondary structure; five others were false positives. The remaining 13 proteins contained unusually long loopy regions although they were not detected when applying our criterion on the DSSP assignment from the 3D coordinates.

NORS regions, on average, are depleted of hydrogen bonds

We investigated in the following way whether NORS regions found in PDB were indeed flexible in structure: we calculated the number of hydrogen bonds within and between NORS regions, as well as between NORS regions and non-NORS regions. Between residues within NORS regions, we counted, on average, 0.66 hydrogen bond per residue. Between residues in non-NORS regions of similar length, we counted 1.209 hydrogen bonds. These two values differed significantly ($t = -7.8$,

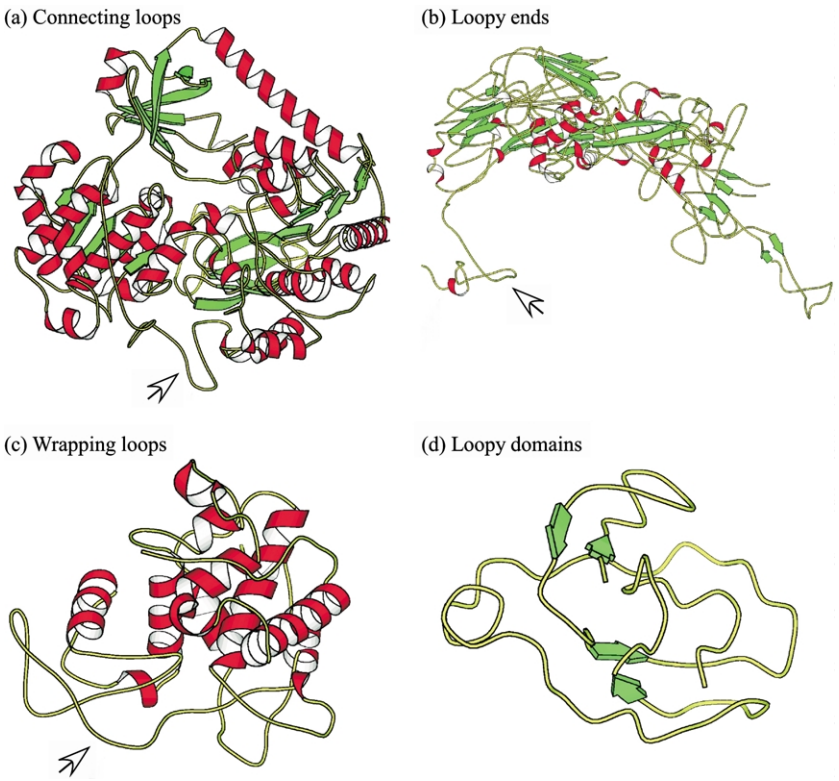


Figure 1. Four types of PDB proteins with NORS regions. NORS regions are defined as having at least 70 consecutive residues with less than 12% regular secondary structure (helix or strand). We found four types of proteins. (a) Connecting loops: long loops that connect two domains or chains (shown, formate dehydrogenase H, 1AA6,[41]). In the isoamylase (1BF2), the 96 residue loop connects domains N and A, and forms an inter-domain bridge across the barrel. The corresponding loop is absent from the alpha-amylase family enzymes lacking domain N, but a domain analysis based on the distance map showed the loop to be included in domain A.[42] The loop EF (residues 279–334) in the DNA-containing capsid of canine parvovirus (4DPV) contacts several neighbouring beta-strands, thus apparently accounting for the specificity in the assembly of interactions.[43] (b) Loopy ends: long N or C-terminal regions that lack regular secondary structure (shown, hexon from adenovirus type 2, 1DHX[44]). The overall shape of the trimeric hexon molecule of the adenovirus type 2 hexon (1DHX) is unusual and may be divided into a pseudo-hexagonal base rich in secondary structure, and a triangular top formed from three long loops.[44] The hexon top consists of intimately interacting loops (l1, l2 and l4) emerging from P1 and P2 in the base; it has a triangular shape not exhibiting the pseudo-symmetry of the base. In general, temperature factors are good indicators of atomic flexibility. The N-terminal arm at the base and the loop l1 at the top of 1DHX have the highest average temperature values in that structure. Similarly, the loop insertions between β-strands E and F, and G and H of CPV capsid protein (1B35, C chain) add specificity to the assembly interaction by forming inter-bridging sheets.[45] (c) Loopy wraps: long loopy regions wrapping around globular domains (shown, class II chitinase, 2BAA[47]). The third domain of beef liver catalase (7CAT, residues 321–436) is referred to as the wrapping domain; it forms an outer layer to each subunit. It lacks discernible secondary structure in a long stretch between residues 366 and 420. However, this domain contains the essential helix with the proximal ligand Tyr357. The wrapping domain forms a short secondary structure with an identical region of the P-axis-related-subunit.[49] (d) Loopy domains: entire structures that have almost no regular secondary structure (shown extra-cellular domain of T beta RI, 1TBI[51]). The Arg78-Gly79-Asp80 (RGD loop) of the HIV-1 *trans*-activating regulatory protein TAT (1TAC) is a key site for adhesive recognition and receptor interaction. This region is solvent-exposed at the tip of a hairpin structure that is experimentally well defined by several NOESY cross-peaks, as reflected in the low variation between the NMR ensemble for this loop. This low variation is unusual, given that RGD loops seem to be very flexible in most proteins studied so far, e.g. the structure of decorsin (1DEC) has a rigid RGD loop similar to that of the HIVZ2 Tat protein.[104] The rigidity of the HVIZ2 Tat protein RGD loop structure may be due to two close proline residues, Pro77 and Pro81, flanking the loop.

$p < 0.001$). Similarly, we found about 0.13 hydrogen bond per residue between NORS residues and the rest of the proteins (non-local), while non-NORS regions had 0.27 non-local hydrogen bond ($t = -3.3$, $p = 0.001$). Thus, NORS regions appeared significantly less stabilised by hydrogen bonds than non-NORS regions.

Predicting NORS regions in entire proteomes

Many proteins with NORS regions in proteomes

We predicted a high fraction of proteins with NORS regions in each of the 31 entirely sequence proteomes that we tested (Figure 2; Supplementary Material, Table S2). The numbers differed considerably between the three kingdoms: for most in archaebacteria and prokaryotes we predicted NORS regions in less than 5% of all proteins (exception, *Aeropyrum pernix*, 13%), while the values were 17–30% for eukaryotes (Figure 2(a)). For eukaryotes, 7–15% of the entire residue mass was predicted in eukaryotic NORS regions (Figure 2(b)). Most NORS regions were between 70 and 130 residues long (Figure 2(c)). Almost all extremely long NORS regions (>500 residues) were found in eukaryotes.

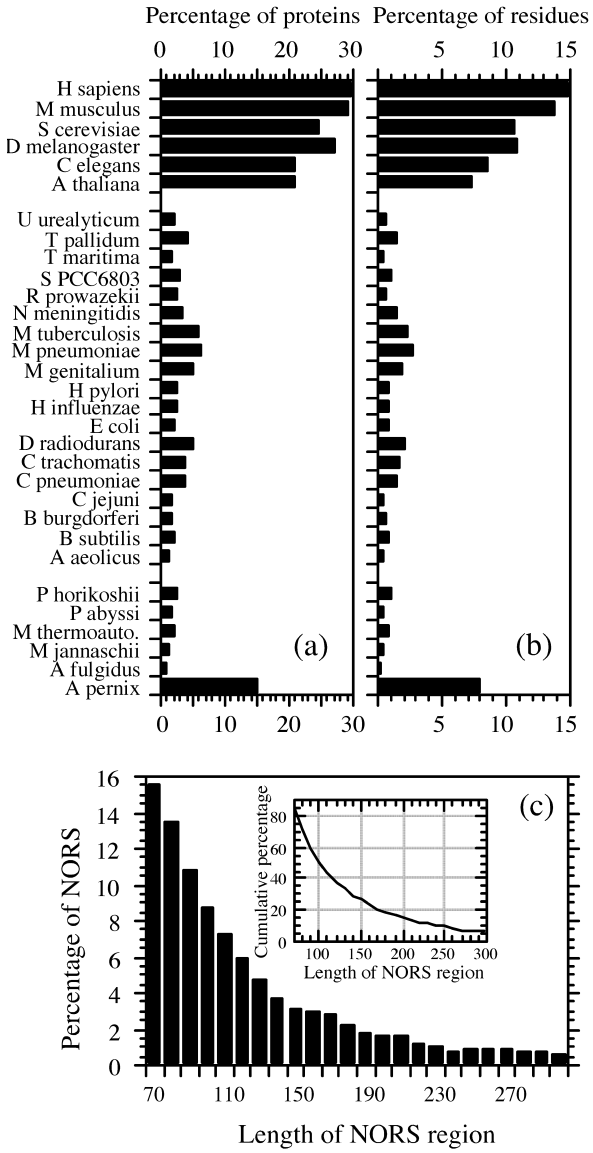


Figure 2. Many NORS proteins were predicted in proteomes. We predicted many NORS regions in 31 entirely sequenced organisms. NORS proteins appeared particularly abundant in eukaryotes. (a) The upper left graph gives the percentage of proteins in respective proteome for which we predicted at least one NORS region. (b) The upper right graph illustrates the percentage of all the residues of the respective proteome for which we predicted in a NORS region (note the difference in scales between (a) and (b)). (c) The lower graph gives the percentage of all predicted NORS regions that are between N and $N+10$ residues long (note that, by definition, NORS regions are longer than 70 residues). Surprisingly, almost 15% of all the predicted NORS regions extend over more than 200 residues (inset in (c)).

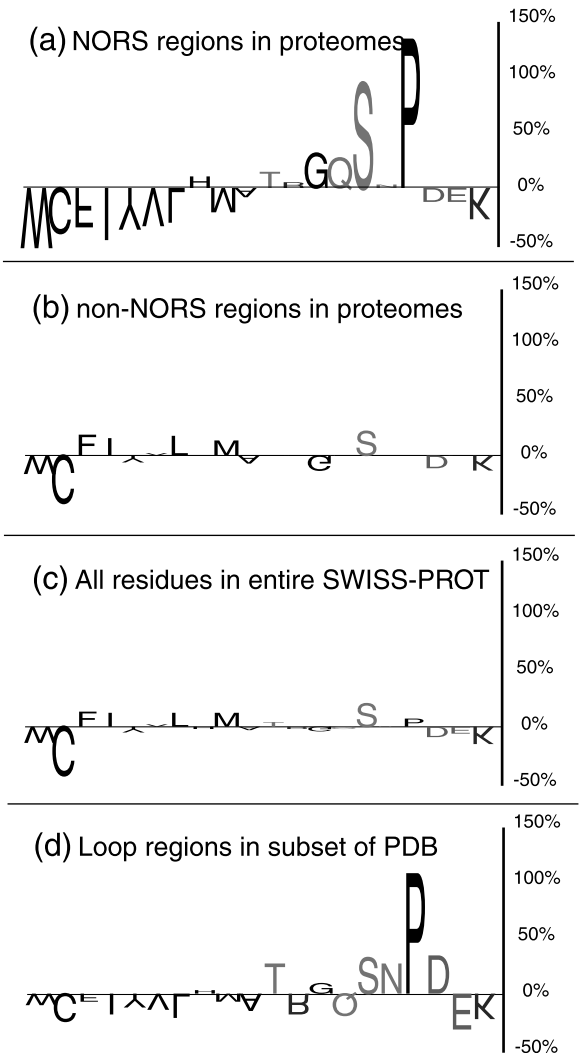
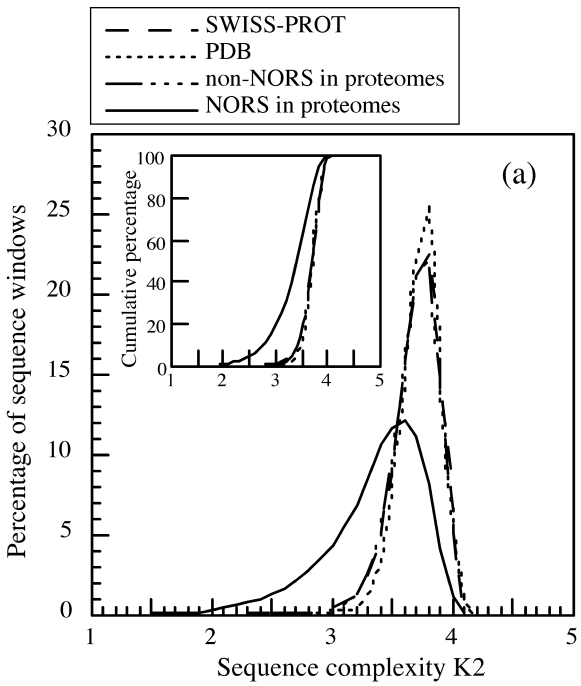


Figure 3. NORS regions use particular amino acids. The height of the one-letter amino acid code is proportional to the abundance of the respective acid in each data set. The actual value is the difference in occurrence with respect to the frequency observed in a sequence-unique subset of PDB: $(f_i^{a} - f_i^{\text{PDB\_unique}})/f_i^{\text{PDB\_unique}}$. Inverted letters indicate acids that are less frequent than "expected". The amino acids are sorted by flexibility,[56] with the more rigid on the left. Overall, NORS regions are as abundant in more flexible residues as loop regions in PDB. However, we found considerably more serine (S), glutamine (Q), and glycine (G) and considerably fewer arginine (R), aspartic acid (D), glutamic acid (E), tryptophan (W), and phenylalanine (F) residues in NORS regions than in loop regions, in general.

NORS regions have specific amino acid composition

We compared the NORS regions predicted in the proteomes (Figure 3(a)) to non-NORS regions in the same set of proteins (Figure 3(b)), as well as to all proteins in SWISS-PROT (Figure 3) and to all residues without regular secondary structure in a sequence-unique subset of PDB (Figure 3(d)). The most rigid amino acid residues (WCFIYVLM), as measured by the Vihinen scale,[56] which reflects the side-chain motion, were severely underrepresented, while only some of flexible amino acid residues (GQSP) were over-represented (Figure 3(a)). Although loop residues (Figure 3(d)) exhibited a similar trend, the amino acid



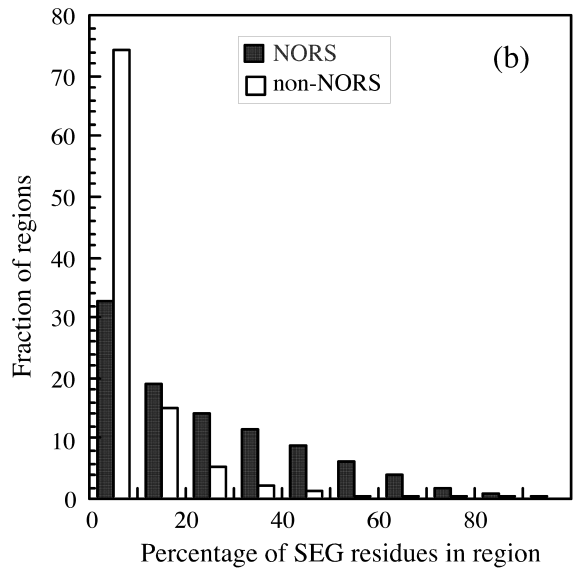


Figure 4. Most NORS regions have compositional bias similar to that of PDB proteins. We measured sequence composition in two slightly different ways: by the Shannon entropy averaged over segments of 45 consecutive residues (K_2, equation (1)) and by the percentage of low-complexity residues assigned by the program SEG.[35] (a) The distribution of the Shannon entropy was shifted towards values lower than that for non-NORS regions. (b) Similarly, NORS regions had significantly more residues of low complexity than non-NORS regions. However, if we choose a threshold in complexity that considers only 1% of the PDB proteins to have low-complexity segments longer than 45 residues, we detect only 16% of the NORS regions predicted (cumulative percentage given in the inset in (a)).

composition of NORS regions differed significantly from that of loop residues. More specifically, NORS regions were more depleted in WFVD, and more

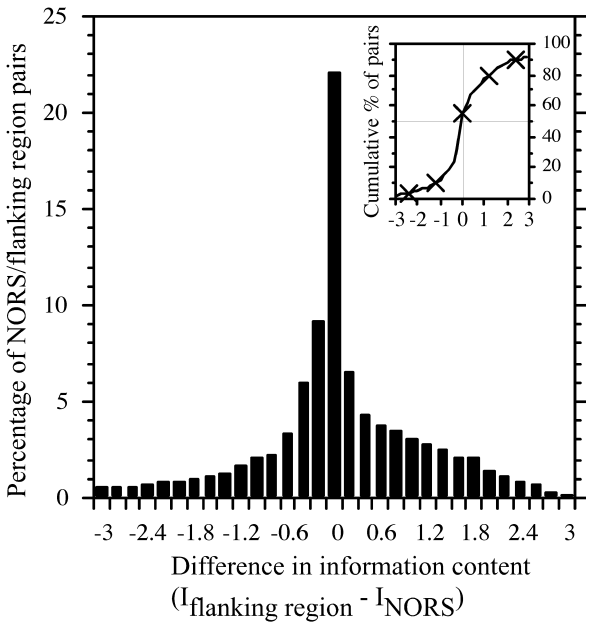


Figure 5. NORS regions are as conserved as flanking regions. In order to investigate whether NORS regions were evolutionarily conserved, we measured the difference in the information contents between alignments in NORS and their flanking regions ($I_{\text{flanking region}} - I_{\text{NORS}}$, equation (3)). The percentage values were compiled over all pairs of NORS/flanking regions, i.e. the total number of pairs was twice the number of NORS regions found. The inset gives the cumulative percentages. The difference in conservation between NORS and flanking regions was not significant. In other words, NORS regions appeared, on average, to be as conserved as non-NORS regions.

enriched in QSP. NORS regions and loops shared some similarities (high P, low E) that distinguished them from SWISS-PROT proteins and non-NORS proteins in proteomes.

Low-complexity and NORS regions differed

We compared the complexity (K_2, equation (1)) between NORS and non-NORS regions (Figure 4). NORS regions were clearly shifted toward lower complexity values (Figure 4(a)): about 16% of the NORS regions had K_2 values below 2.9, while only 1% of the fragments in non-NORS, PDB proteins, or SWISS-PROT proteins were below this value (Figure 4(a), inset). We monitored low complexity using a slightly different definition; namely, the percentage of residues considered to be of low complexity according to the widely used method SEG.[35] Consistent with the findings for the K_2 distribution, NORS regions had a higher fraction of SEG residues (Figure 4(b)). However, more than 80% of the NORS regions predicted could not have been identified only by applying some threshold in low complexity.

NORS regions are as conserved as flanking regions

In multiple alignments of evolutionarily diverged protein families, we typically observe two kinds of consecutive regions:[57–60] (1) regions that can be aligned over the entire length; and (2) regions for which some of the family members have insertions. The usual assumption is that regions with long deletions/insertions are functionally less important. To determine whether NORS regions were evolutionarily conserved, we compared the information content (equation (3)) in NORS regions to that of the N-terminal and C-terminal non-NORS segments. For more than 20% of all NORS regions, we could not distinguish between the conservation of the NORS and of its flanking regions (no difference in information content, Figure 5). For about 56% of all pairs of NORS/flanking region, the NORS had a similar or higher information content, i.e. was evolutionarily equally or more conserved (negative values, Figure 5, inset). A detailed analysis revealed that the differences in evolutionary conservation were not statistically significant. This suggested that NORS regions evolved according to evolutionary constraints similar to those applied to the flanking regions.

NORS proteins had slightly more interaction partners than non-NORS proteins

We analysed the Database of Interacting Proteins (DIP),[61] which lists all protein–protein interactions unravelled by the first two large-scale yeast-two-hybrid experiments.[62] We found that 3464 (72%) of all predicted non-NORS yeast proteins had one or more binding partners. In contrast, about 1126 (79%) of all 1556 predicted NORS proteins in yeast had at least one interaction partner (Figure 6). This difference was statistically significant ($z = 5.18$, $p < 0.001$, equation (2)). The comprehensive experimental analysis of protein–protein interactions in *Helicobacter pylori*[63] yielded similar results: 64% of the NORS and 44% of the non-NORS proteins had interaction partners ($z = 2.41$, $p = 0.008$, equation (2)). Most data from yeast two-hybrid experiments do not reveal the precise regions involved in protein–protein interfaces. However, we found 37 examples of protein–protein interactions in DIP for which the regions of interaction overlapped with NORS regions for at least 70 residues (Supplementary Material, Table 3S). For example, the yeast protein YKR025w, a potential subunit of RNA polymerase III, interacts with RPC4_YEAST (YDL150W, RNA polymerase III chain C53) *via* region 62–200,[64] which coincides with predicted NORS region at 82–155 (Supplementary Material, Table S3). Another family of examples were revealed by a genome-wide two-hybrid screening showing that Lsm (like sm) proteins interact with some splicing factors and proteins involved in mRNA turnover.[65]

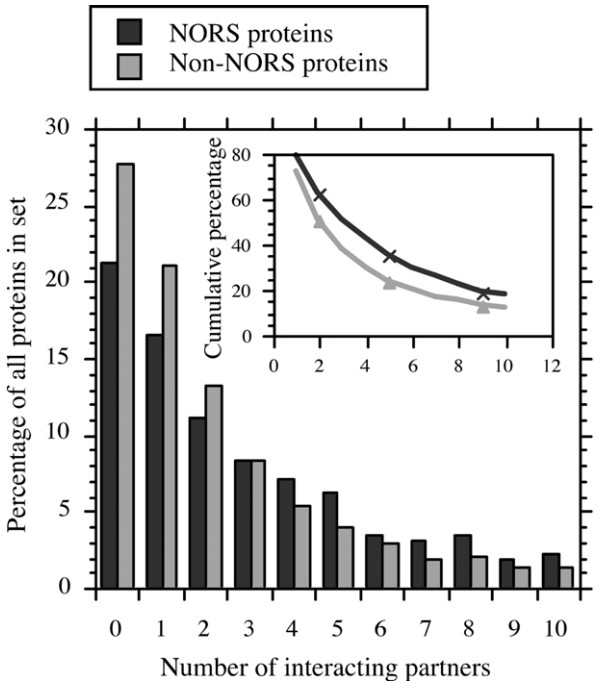


Figure 6. NORS proteins interacted more than non-NORS proteins. We compared the number of interacting partners annotated in DIP[61] between predicted NORS and non-NORS proteins. We found that considerably more NORS than non-NORS proteins had one or more interaction partners (the inset gives the cumulative percentages). The difference between the distributions for NORS and non-NORS proteins was statistically significant ($z = 5.18$, $p < 0.001$, equation (2)).

Many of these protein–protein interactions of the Lsm proteins seemed to be mediated by predicted NORS regions (Supplementary Material, Table S3). Finally, NORS regions appeared to be involved in interactions with actin-related proteins (Supplementary Material, Table S3).[66]

NORS proteins are often related to regulation and transcription

For all predicted NORS proteins, we searched for functional annotations in SWISS-PROT. We found a variety of descriptions, including numerous carbohydrate modification sites, phosphorylation sites, disulphide bridges, and catalytic active sites. NORS regions occurred in many transcription factors, and frequently were found spanning half of the zinc-finger motifs and the residues preceding these. Residues immediately upstream of homeodomains were often in NORS regions. Monika Riley introduced classes of cellular function to characterise the functional content of proteomes.[67,68] We assigned such classes automatically through the program EUCLID.[69,70] We classified about 45–65% of all proteins into one of 14 functional classes at a level reported to yield 70% correct classifications; namely, above 30% pairwise sequence identity.[71] Notably, NORS proteins were significantly under-represented in most

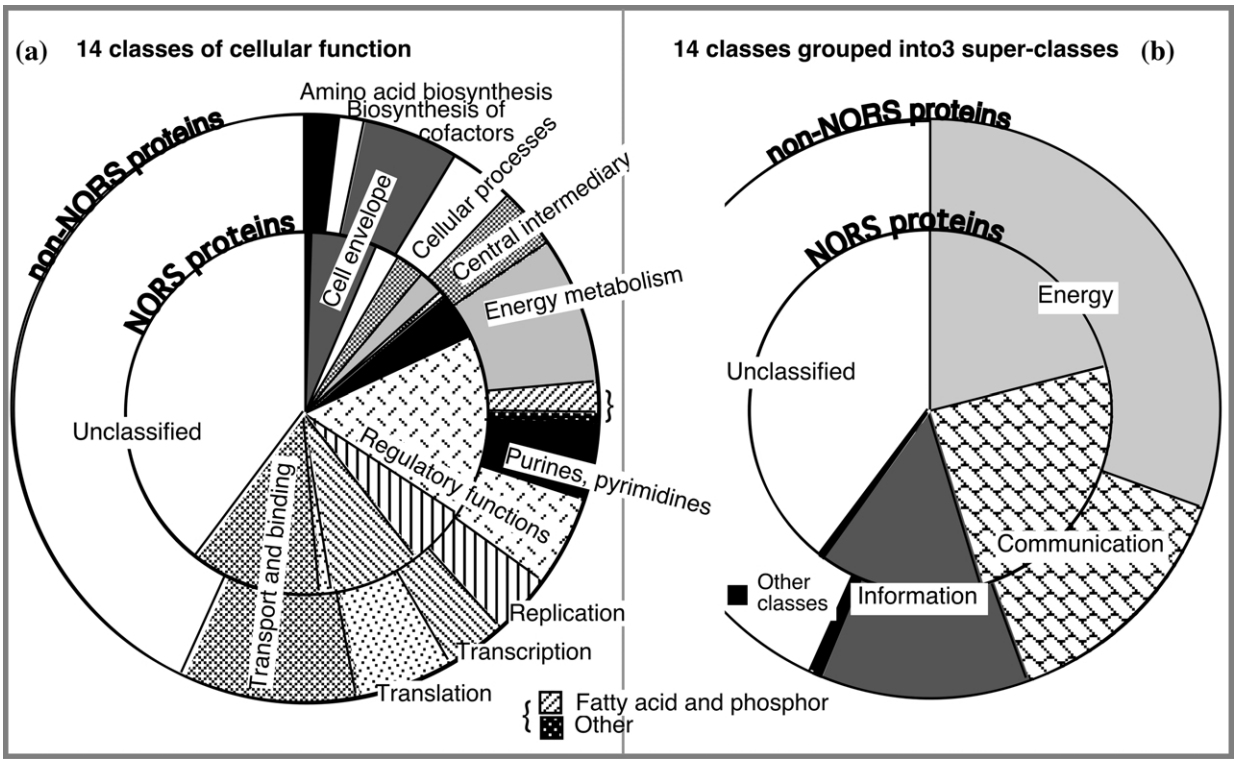


Figure 7. NORS proteins unique in their spectrum of cellular functional classes. The program EUCLID[69] sorts proteins of experimentally known function into classes of cellular function. For each proteome, we compared the fraction of NORS proteins in each of these classes to that of the non-NORS proteins. Here, we show the averages over all 31 proteomes (inner circle: NORS proteins, outer circle: non-NORS proteins). (a) The upper graph separates all 14 classes assigned by EUCLID, (b) the lower graph groups the 14 classes into three "super-classes".

biosynthesis classes (Amino acid biosynthesis, Biosynthesis of cofactors, prosthetic groups, and carriers, Fatty acid and phospholipid metabolism, Purines, pyrimidines, nucleosides, and nucleotides), in Energy metabolism, and in Translation compared to non-NORS proteins (Figure 7(a)). In contrast, Regulatory Functions and Transcription classes were more abundant in NORS proteins (Figure 7(a)). This was consistent with our observation that NORS appeared in many transcription factors. When grouping the 14 classes into three super-classes, energy, information and communication,[69] we found that NORS proteins were more often associated with Communication (24(±1)% *versus* 14(±1)%; Figure 7(b) and less often with Energy (21(±1)% *versus* 31(±2)%; Figure 7(b)) than were non-NORS proteins.

Discussion and Conclusion

Do NORS, disordered, natively unfolded regions and structural switches differ?

It is commonly assumed that regions of non-regular secondary structure (turns, loops, or bends) are more flexible than are the networks of backbone hydrogen bonds stabilising helices and strands. In fact, two-thirds of all globular protein structures fall into a rather narrow window of 50–65% regular secondary structure, and only 1% of the proteins longer than 70 residues have less than 20% regular secondary structure.[72] Experimentalists are beginning to find increasing evidence of proteins that appear to be unfolded in their native, unbound conformation[19,21,22,24,25,73] or that can undergo considerable conformational changes upon binding;[74–81] and that structural switches can be predicted from sequence.[16,82] Dunker and colleagues developed a method predicting natively disordered regions (labelled Dunker regions).[33,34,37] Uversky *et al.* claimed that natively unfolded proteins could be identified through their net charge and hydrophobicity;[73] Zetina claimed that many natively unfolded proteins contain a particular helix-unfolding sequence-motif.[83] Here, we focus on regions longer than 70 residues that have an unusually low content of regular secondary structure (NORS). We expect that NORS regions overlap with regions identified by the Dunker group: both Dunker regions and our NORS regions have more segments of low complexity than typical globular proteins (K_2, Figure 4).[38] In contrast, the Dunker regions and NORS differed in their amino acid composition: while R and E were more abundant in Dunker regions, the usage of R was similar between NORS and non-NORS (Figure 3) and E was even less frequent in NORS than in non-NORS (Figure 3). The under-representation of C in Dunker regions also did not correspond to the

observation in NORS regions. Finally, Dunker *et al.* predicted almost twice as many disordered regions in eukaryotes than we predicted NORS regions (Figure 2 *versus* Dunker *et al.*[28]). Dunker and colleagues published their 20 strongest predictions: for all, we also predicted NORS regions, although for two predictions, Dunker regions and NORS did not overlap (Supplement Material, Table S4). We could not verify that all NORS regions were strictly confined to a particular region in hydrophobicity *versus* net charge plot as has been postulated for natively unfolded proteins.[73] Clearly, the NORS regions we predicted did not overlap with regions typically labelled as structural switches. In summary, the various attempts at characterising non-regular regions in proteins identified sets of regions that overlapped to only some extent.

Are NORS regions important for function?

Although NORS regions were abundant in low-complexity residues, they were evolutionarily as conserved as flanking regions (Figure 5). This suggested that NORS regions are important for function. One way in which NORS regions could play important functional roles is through protein–protein interactions. Calcineurin is one particular example for protein–protein interactions that are mediated by disordered regions: the flexibility of a 95 residue segment in subunit A is important for calmodulin-binding.[84,85] The analysis of the yeast two-hybrid results (Figure 6) confirmed that proteins with NORS regions have, on average, more interaction partners than other proteins. The seemingly more active role of NORS proteins in protein–protein interaction might be explained by the hypothesis that NORS regions might be stabilised by protein–protein interactions (induced fit). We also found NORS proteins to be more often related to regulatory functions and transcription than non-NORS proteins (Figure 7). We do not know the precise role that NORS regions play in transcription factors. An obvious hypothesis is that the conformational adaptability of NORS regions enables different regulation. This might explain why NORS proteins appeared particularly abundant in eukaryotes (Figure 2).

New types of protein structures?

We predicted over 20,000 proteins with NORS regions over 130 residues in eukaryotes alone. Currently, we have no example for any of these in PDB. Large-scale efforts at determining all protein structures (structural genomics initiatives)[6,86–92] may help to decide whether loopy proteins constitute a new class of protein structures. However, while we have no data supporting speculations of how these proteins look or what they do, we could refute the assumption that NORS regions constitute some ancient carry-over that is functionally unimportant.

Methods

Data sets

Source of proteome sequences

We obtained the sequences for all 31 organisms that we analysed from the public domain. We downloaded most ORFs from NCBI†. The exceptions were *Homo sapiens* (from SWISS-PROT release 39 and TrEMBL[39] database release 15), *Caenorhabditis elegans*‡, *Drosophila melanogaster*§, and *Mus musculus*‖.

Sequence-unique subset of PDB (PDBsub)

To reduce the bias from mutation studies, we restricted our analysis of PDB to a sequence-unique subset. This subset was defined as having no pair with more than 33 pairwise identical residues over more than 100 residues aligned. More precisely, the HSSP distance[93] was below 0 for any pair in the set. We maintain a weekly update of such a set through our EVA server.[94] The set used for this study contained 1947 protein chains.

Prediction methods

Secondary structure, membrane helices and solvent accessibility

We obtained multiple sequence alignments by searching with the dynamic programming method MaxHom[95] against SWISS-PROT.[39] The resulting alignments were subsequently filtere[93] and used as input for PHDsec,[96,97] PHDacc,[96,98] and PHDhtm.[99] For all methods, we used the default parameters. For proteins of known structure, we assigned accessibility and secondary structure with DSSP.[55] In particular, we used the following convention to convert the eight DSSP states into three classes: DSSP HGIhelix (H), DSSP'EBstrand (E), and all other to non-regular (L). Buried residues were defined as those with a relative accessibility to solvent of $<16\%$.

Secreted proteins and coiled-coil regions

We predicted signal peptides using the program SignalP,[100] considering a protein to contain a signal peptide if the mean *S* value was above the default threshold. We predicted coiled-coil regions with COILS,[101] using a window size of 28 and a probability threshold of 0.9.

Sequence complexity

Low sequence-complexity regions were determined by SEG,[35] using default parameters. Sequence compositional complexity K_2 of a sequence window was calculated as described:[35]

$$K_2 = -\sum_{i=1}^{N} \frac{n_i}{L}\left(\log_2 \frac{n_i}{L}\right) \quad (1)$$

where N represents the number of letters in the alphabet (20 for amino acid residues in protein) and n_i is the

† ftp://ncbi.nlm.nih.gov/genbank/genomes
‡ From www.sanger.ac.uk/Projects/C_elegans/wormpep/, wormpep65
§ From www.fruitfly.org/
‖ From www.ensembl.org/Mus_musculus/

occurrence of amino acid i in sequence window of length L. In particular, we chose segments of 45 consecutive residues to measure compositional bias.

Functional classification

We classified cellular function using the program EUCLID.[70] The SWISS-PROT homologues input to EUCLID were identified by MaxHom (pairwise sequence identity >30%). EUCLID assigned the following 14 categories of cellular function:[102] amino acid biosynthesis; biosynthesis of cofactors, prosthetic groups, and carriers; cell envelope; cellular processes; central intermediary metabolism; energy metabolism; fatty acid and phospholipid metabolism; other categories; purines, pyrimidines, nucleosides, and nucleotides; regulatory functions; replication; transcription; translation; and transport and binding proteins. Finally, we added the class Unclassified, listing all those proteins for which we either did not find homologues in SWISS-PROT or that could not be classified by EUCLID.

Definition of no regular secondary structure region (NORS)

We identified NORS regions (extended regions of NO regular secondary structure) in the following way. First, we applied all programs (PHDsec, PHDacc, PHDhtm, COILS, and SignalP). Then, we compiled the percentage of residues with regular structural signals (regular secondary structure, transmembrane helices, coiled-coil regions, signal peptides) over sliding windows of 70 consecutive residues. NORS were assigned if both following conditions applied. (1) The regular structural content was below 12%, i.e. there were less than 12% helix/strand/coiled-coil/membrane helix/signal peptide. (2) We found at least one continuous segment longer than ten residues within which all residues were exposed to solvent. NORS regions were extended in both directions as long as the above two criteria remained valid.

Calculation of inter-region and intra-region hydrogen bonds

We extracted hydrogen bond information for all residues in PDBsub (see above) through the DSSP program.[55] We calculated the number of inter-region and intra-region hydrogen bonds per residue for each NORS region in PDB, and averaged over all regions. For non-NORS regions, these numbers were obtained for all 70-residue sequence windows from the data set by randomly selecting 2000 such windows in order to avoid over-sampling of overlapping sequence windows.

Statistical analysis

We applied the standard Student's t-test to determine whether the difference between the number of hydrogen bonds of NORS regions and non-NORS regions was significant. We also tested the differences between the proportions of two populations p_1 and p_2 in the following way:

$$z = (\bar{p}_1 - \bar{p}_2)/\hat{\sigma}_{p_1-p_2} \tag{2}$$

where $\hat{\sigma}_{p_1-p_2} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}$, and

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

The one-tailed probability was then obtained through the normal distribution table.

Calculation of information content in a sequence segment

The information content of sequence segments was determined on the basis of the multiple sequence alignment generated by MaxHom, according to the method described by Gorodkin *et al.*[103] Briefly, the information content of position i of the alignment I_i was calculated as follows:

$$I_i = \sum_{k \in A} I_{ik} = \sum_{k \in A} q_{ik} \log_2 \frac{q_{ik}}{p_k} \tag{3}$$

where $A = \{A,C,D,\ldots,W,Y,-\}$ is the set of 20 amino acid residues including gaps (−), q_{ik} is the fraction of amino acid k at position i. When k is not −, p_k equals the *a priori* distribution of the amino acid for SWISS-PROT database. p_- is set to 1. The average information content of a sequence segment was then calculated by taking the average of individual positions within the segment.

Possible functional annotation of NORS

For proteins with NORS regions that were contained in SWISS-PROT, we extracted functional annotations from the FT entry. For other proteins, we aligned each NORS region against SWISS-PROT (using MaxHom), and extracted functional annotations for homologues that had more than 50% pairwise identical residues over more than 100 aligned residues, i.e. an HSSP distance above 15.[93] Wherever possible, we kept only annotations that were related explicitly to the NORS regions.

Acknowledgments

We thank Henrik Nielsen (CBS, Denmark) for providing the source code for SignalP and for his generous help in using this program, to Andrei Lupas (MPI Tübingen) for helpful suggestions about using the COILS program, and to Henry Bigelow (Columbia) for crucial comments on the manuscript. We thank Florencio Pazos, Damien Devos and Alfonso Valencia (all CNB Madrid) for supplying and helping with the program EUCLID.; and Alexei Murzin (MRC Cambridge) for useful discussions. Finally, we thank to the undisclosed reviewer who suggested analysing the hydrogen bonding networks of NORS regions. The work of J.L. and B.R. was supported by grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institute of Health. Last, not least, thanks to all those who deposit their experimental data into public databases, and to those who maintain these databases.

References

1. Anfinsen, C. B. & Scheraga, H. A. (1975). Experimental and theoretical aspects of protein folding. *Advan. Protein Chem.* **29**, 205–300.
2. Ellis, R. J., Dobson, C. & Hartl, U. (1998). Sequence does specify protein conformation. *Trends Biochem. Sci.* **23**, 468.

3. Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223–230.
4. Blomberg, N. & Nilges, M. (1997). Functional diversity of PH domains: an exhaustive modelling study. *Fold. Des.* **2**, 343–355.
5. Jones, S., van Heyningen, P., Berman, H. M. & Thornton, J. M. (1999). Protein–DNA interactions: a structural analysis. *J. Mol. Biol.* **287**, 877–896.
6. Lima, C. D., Klein, M. G. & Hendrickson, W. A. (1997). Structure-based analysis of catalysis and substrate definition in the HIT protein family. *Science*, **278**, 286–290.
7. Moult, J. & Melamud, E. (2000). From fold to function. *Curr. Opin. Struct. Biol.* **10**, 384–389.
8. Murzin, A. G. (1996). Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.* **6**, 386–394.
9. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143.
10. Jones, S., Daley, D. T., Luscombe, N. M., Berman, H. M. & Thornton, J. M. (2001). Protein–RNA interactions: a structural analysis. *Nucl. Acids Res.* **29**, 943–954.
11. Irving, J. A., Whisstock, J. C. & Lesk, A. M. (2001). Protein structural alignments and functional genomics. *Proteins: Struct. Funct. Genet.* **42**, 378–382.
12. Russell, R. B. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* **279**, 1211–1227.
13. Wallace, A. C., Borkakoti, N. & Thornton, J. M. (1997). TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **6**, 2308–2323.
14. Trowbridge, I. S. (1991). Endocytosis and signals for internalization. *Curr. Opin. Cell Biol.* **3**, 634–641.
15. Unwin, N. (1998). The nicotinic acetylcholine receptor of the Torpedo electric ray. *J. Struct. Biol.* **121**, 181–190.
16. Young, M., Kirshenbaum, K., Dill, K. A. & Highsmith, S. (1999). Predicting conformational switches in proteins. *Protein Sci.* **8**, 1752–1764.
17. Rozovsky, S. & McDermott, A. E. (2001). The time scale of the catalytic loop motion in triosephosphate isomerase. *J. Mol. Biol.* **310**, 259–270.
18. Rozovsky, S., Jogl, G., Tong, L. & McDermott, A. E. (2001). Solution-state NMR investigations of triosephosphate isomerase active site loop motion: ligand release in relation to active site loop dynamics. *J. Mol. Biol.* **310**, 271–280.
19. Wright, P. E. & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. *J. Mol. Biol.* **293**, 321–331.
20. Perutz, M. F. (1992). What are enzyme structures telling us? *Faraday Discuss.* 1–11.
21. Okada, K., Hirotsu, K., Hayashi, H. & Kagamiyama, H. (2001). Structures of *Escherichia coli* branched-chain amino acid aminotransferase and its complexes with 4-methylvalerate and 2-methylleucine: induced fit and substrate recognition of the enzyme. *Biochemistry*, **40**, 7453–7463.
22. Yaremchuk, A., Tukalo, M., Grotli, M. & Cusack, S. (2001). A succession of substrate induced conformational changes ensures the amino acid specificity of *Thermus thermophilus* prolyl-tRNA synthetase: comparison with histidyl-tRNA synthetase. *J. Mol. Biol.* **309**, 989–1002.
23. Claussen, H., Buning, C., Rarey, M. & Lengauer, T. (2001). FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **308**, 377–395.
24. Weiss, M. A. (2001). Floppy SOX: mutual induced fit in hmg (high-mobility group) box-DNA recognition. *Mol. Endocrinol.* **15**, 353–362.
25. Wyatt, R., Kwong, P. D., Desjardins, E., Sweet, R. W., Robinson, J., Hendrickson, W. A. & Sodroski, J. G. (1998). The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature*, **393**, 705–711.
26. Wang, Y., Sha, M., Ren, W. Y., van Heerden, A., Browning, K. S. & Goss, D. J. (1996). pH-dependent and ligand induced conformational changes of eucaryotic protein synthesis initiation factor eIF-(iso)4F: a circular dichroism study. *Biochim. Biophys. Acta*, **1297**, 207–213.
27. Zhou, G., Ellington, W. R. & Chapman, M. S. (2000). Induced fit in arginine kinase. *Biophys. J.* **78**, 1541–1550.
28. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S. *et al.* (2001). Intrinsically disordered protein. *J. Mol. Graph. Model*, **19**, 26–59.
29. Manalan, A. S., Krinks, M. H. & Klee, C. B. (1984). Calcineurin: a member of a family of calmodulin-stimulated protein phosphatases. *Proc. Soc. Expt. Biol. Med.* **177**, 12–16.
30. Manalan, A. S. & Klee, C. B. (1983). Activation of calcineurin by limited proteolysis. *Proc. Natl Acad. Sci. USA*, **80**, 4291–4295.
31. Kissinger, C. R., Parge, H. E., Knighton, D. R., Lewis, C. T., Pelletier, L. A., Tempczyk, A. *et al.* (1995). Crystal structures of human calcineurin and the human FKBP12–FK506–calcineurin complex. *Nature*, **378**, 641–644.
32. Namba, K. (2001). Roles of partly unfolded conformations in macromolecular self-assembly. *Genes Cells*, **6**, 1–12.
33. Dunker, A. K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J. *et al.* (1998). Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac. Symp. Biocomput.* **3**, 473–484.
34. Garner, E., Cannon, P., Romero, P., Obradovic, Z. & Dunker, A. K. (1998). Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization. Genome inform ser workshop. *Genome Inform.* **9**, 201–213.
35. Wootton, J. C. & Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571.
36. Dunker, A. K. & Obradovic, Z. (2001). The protein trinity-linking function and disorder. *Nature Biotechnol.* **19**, 805–806.
37. Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E., Garner, E., Guilliot, S. & Dunker, A. K. (1998). Thousands of proteins likely to have long disordered regions. *Pac. Symp. Biocomput.* **3**, 437–448.
38. Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J. & Dunker, A. K. (2001). Sequence complexity of disordered protein. *Proteins: Struct. Funct. Genet.* **42**, 38–48.
39. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT:protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
40. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.

41. Boyington, J. C., Gladyshev, V. N., Khangulov, S. V., Stadtman, T. C. & Sun, P. D. (1997). Crystal structure of formate dehydrogenase H: catalysis involving Mo, molybdopterin, selenocysteine, and an Fe4S4 cluster. *Science*, **275**, 1305–1308.
42. Katsuya, Y., Mezaki, Y., Kubota, M. & Matsuura, Y. (1998). Three-dimensional structure of Pseudomonas isoamylase at 2.2 Å resolution. *J. Mol. Biol.* **281**, 885–897.
43. Xie, Q. & Chapman, M. S. (1996). Canine parvovirus capsid structure, analyzed at 2.9 Å resolution. *J. Mol. Biol.* **264**, 497–520.
44. Athappilly, F. K., Murali, R., Rux, J. J., Cai, Z. & Burnett, R. M. (1994). The refined crystal structure of hexon, the major coat protein of adenovirus type 2, at 2.9 Canine parvovirus capsid structure, analyzed at 2.9 Å resolution resolution. *J. Mol. Biol.* **242**, 430–455.
45. Tate, J., Liljas, L., Scotti, P., Christian, P., Lin, T. & Johnson, J. E. (1999). The crystal structure of cricket paralysis virus: the first view of a new virus family. *Nature Struct. Biol.* **6**, 765–774.
46. Chabriere, E., Charon, M. H., Volbeda, A., Pieulle, L., Hatchikian, E. C. & Fontecilla-Camps, J. C. (1999). Crystal structures of the key anaerobic enzyme pyruvate:ferredoxin oxidoreductase, free and in complex with pyruvate. *Nature Struct. Biol.* **6**, 182–190.
47. Hart, P. J., Pfluger, H. D., Monzingo, A. F., Hollis, T. & Robertus, J. D. (1995). The refined crystal structure of an endochitinase from *Hordeum vulgare* L. seeds at 1.8 Å resolution. *J. Mol. Biol.* **248**, 402–413.
48. Fita, I. & Rossmann, M. G. (1985). The active center of catalase. *J. Mol. Biol.* **185**, 21–37.
49. Fita, I. & Rossmann, M. G. (1985). The NADPH binding site on beef liver catalase. *Proc. Natl Acad. Sci. USA*, **82**, 1604–1608.
50. Sundaramoorthy, M., Terner, J. & Poulos, T. L. (1995). The crystal structure of chloroperoxidase: a heme peroxidase–cytochrome P450 functional hybrid. *Structure*, **3**, 1367–1377.
51. Jokiranta, T. S., Tissari, J., Teleman, O. & Meri, S. (1995). Extracellular domain of type I receptor for transforming growth factor-beta: molecular modelling using protectin (CD59) as a template. *FEBS Letters*, **376**, 31–36.
52. Bayer, P., Kraft, M., Ejchart, A., Westendorp, M., Frank, R. & Rosch, P. (1995). Structural studies of HIV-1 Tat protein. *J. Mol. Biol.* **247**, 529–535.
53. Gronwald, W., Loewen, M. C., Lix, B., Daugulis, A. J., Sonnichsen, F. D., Davies, P. L. & Sykes, B. D. (1998). The solution structure of type II antifreeze protein reveals a new member of the lectin family. *Biochemistry*, **37**, 4712–4721.
54. Kim, H. & Lipscomb, W. N. (1990). Crystal structure of the complex of carboxypeptidase A with a strongly bound phosphonate in a new crystalline form: comparison with structures of other complexes. *Biochemistry*, **29**, 5546–5555.
55. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
56. Vihinen, M., Torkkila, E. & Riikonen, P. (1994). Accuracy of protein flexibility predictions. *Proteins: Struct. Funct. Genet.* **19**, 141–149.
57. Taylor, W. R. (1997). Multiple sequence threading: an analysis of alignment quality and stability. *J. Mol. Biol.* **269**, 902–943.
58. Bork, P. & Gibson, T. J. (1996). Applying motif and profile searches. *Methods Enzymol.* **266**, 162–184.
59. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S. *et al.* (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucl. Acids Res.* **29**, 22–28.
60. Pascarella, S. & Argos, P. (1992). Analysis of insertions/deletions in protein structures. *J. Mol. Biol.* **224**, 461–471.
61. Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M. & Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucl. Acids Res.* **30**, 303–305.
62. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R. *et al.* (2000). A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
63. Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S. *et al.* (2001). The protein–protein interaction map of *Helicobacter pylori*. *Nature*, **409**, 211–215.
64. Flores, A., Briand, J. F., Gadal, O., Andrau, J. C., Rubbi, L., Van Mullem, V. *et al.* (1999). A protein–protein interaction map of yeast RNA polymerase III. *Proc. Natl Acad Sci. USA*, **96**, 7815–7820.
65. Fromont-Racine, M., Mayes, A. E., Brunet-Simon, A., Rain, J. C., Colley, A., Dix, I. *et al.* (2000). Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast*, **17**, 95–110.
66. Bon, E., Recordon-Navarro, P., Durrens, P., Iwase, M., Toh, E. A. & Aigle, M. (2000). A network of proteins around Rvs167p and Rvs161p, two proteins related to the yeast actin cytoskeleton. *Yeast*, **16**, 1229–1241.
67. Riley, M. (1993). Function of the gene products in *Escherichia coli*. *Microbiol. Rev.* **57**, 862–952.
68. Riley, M. & Labedan, B. (1997). Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module. *J. Mol. Biol.* **268**, 857–868.
69. Tamames, J., Ouzounis, C., Sander, C. & Valencia, A. (1996). Genomes with distinct function composition. *FEBS Letters*, **389**, 96–101.
70. Tamames, J., Ouzounis, C., Casari, G., Sander, C. & Valencia, A. (1998). EUCLID: automatic classification of proteins in functional classes by their database annotations. *Bioinformatics*, **14**, 542–543.
71. Devos, D. & Valencia, A. (2000). Practical limits of function prediction. *Proteins: Struct. Funct. Genet.* **41**, 98–107.
72. Andersen, C. A. F., Palmer, A. G., Brunak, S. & Rost, B. (2002). Continuous assignment of secondary structure correlates with protein flexibility. *Structure* **10**, 175–184.
73. Uversky, V. N., Gillespie, J. R. & Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: Struct. Funct. Genet.* **41**, 415–427.
74. Ashkenazi, G., Ripoll, D. R., Lotan, N. & Scheraga, H. A. (1997). A molecular switch for biochemical logic gates: conformational studies. *Biosens. Bioelectron.* **12**, 85–95.
75. Azuma, Y., Renault, L., Garcia-Ranea, J. A., Valencia, A., Nishimoto, T. & Wittinghofer, A. (1999). Model of the ran–RCC1 interaction using

biochemical and docking experiments. *J. Mol. Biol.* **289**, 1119–1130.
76. Falke, S., Fisher, M. T. & Gogol, E. P. (2001). Structural changes in GroEL effected by binding a denatured protein substrate. *J. Mol. Biol.* **308**, 569–577.
77. Noel, J. R. (1997). Turning off the Ras switch with the flick of a finger. *Nature Struct. Biol.* **4**, 677–681.
78. Simeonidis, S., Stauber, D., Chen, G., Hendrickson, W. A. & Thanos, D. (1999). Mechanisms by which IkappaB proteins control NF-kappaB activity. *Proc. Natl Acad. Sci. USA*, **96**, 49–54.
79. Sola, M., Lopez-Hernandez, E., Cronet, P., Lacroix, E., Serrano, L., Coll, M. & Parraga, A. (2000). Towards understanding a molecular switch mechanism: thermodynamic and crystallographic studies of the signal transduction protein cheY. *J. Mol. Biol.* **303**, 213–225.
80. Solano, R., Fuertes, A., Sanchez-Pulido, L., Valencia, A. & Paz-Ares, J. (1997). A single residue substitution causes a switch from the dual DNA binding specificity of plant transcription factor MYB.Ph3 to the animal c-MYB specificity. *J. Biol. Chem.* **272**, 2889–2895.
81. Stouten, P. F. W., Sander, C., Wittinghofer, A. & Valencia, A. (1993). How does the switch II region of G-domains work? *FEBS Letters*, **320**, 1–6.
82. Kirshenbaum, K., Young, M. & Highsmith, S. (1999). Predicting allosteric switches in myosins. *Protein Sci.* **8**, 1806–1815.
83. Zetina, C. R. (2001). A conserved helix-unfolding motif in the naturally unfolded proteins. *Proteins: Struct. Funct. Genet.* **44**, 479–483.
84. Kissinger, C. R., Parge, H. E., Knighton, D. R., Lewis, C. T., Pelletier, L. A., Tempczyk, A. *et al.* (1995). Crystal structures of human calcineurin and the human FKBP12–FK506–calcineurin complex. *Nature*, **378**, 641–644.
85. Meador, W. E., Means, A. R. & Quiocho, F. A. (1992). Target enzyme recognition by calmodulin: 2.4 Å structure of a calmodulin–peptide complex. *Science*, **257**, 1251–1255.
86. Shapiro, L. & Harris, T. (2000). Finding function through structural genomics. *Curr. Opin. Biotechnol.* **11**, 31–35.
87. Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T. *et al.* (1999). Structural genomics: beyond the human genome project. *Nature Genet.* **23**, 151–157.
88. Sali, A. (1998). 100,000 protein structures for the biologist. *Nature Struct. Biol.* **5**, 1029–1032. See comments.
89. Rost, B. (1998). Marrying structure and genomics. *Structure*, **6**, 259–263.
90. Blundell, T. L. & Mizuguchi, K. (2000). Structural genomics: an overview. *Prog. Biophys. Mol. Biol.* **73**, 289–295.
91. Gaasterland, T. (1998). Structural genomics taking shape. *Trends Genet. Sci.* **14**, 135.
92. Thornton, J. (2001). Structural genomics takes off. *Trends Biochem. Sci.* **26**, 88–89.
93. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94.
94. Eyrich, V., MartÍ-Renom, M. A., Przybylski, D., Fiser, A., Pazos, F., Valencia, A. *et al.* (2001). EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*, **17**, 1242–1243.
95. Sander, C. & Schneider, R. (1991). Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.
96. Rost, B. & Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Struct. Funct. Genet.* **19**, 55–72.
97. Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599.
98. Rost, B. & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins: Struct. Funct. Genet.* **20**, 216–226.
99. Rost, B., Casadio, R. & Fariselli, P. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **5**, 1704–1718.
100. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1–6.
101. Lupas, A. (1996). Prediction and analysis of coiled-coil structures. *Methods Enzymol.* **266**, 513–525.
102. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Kelley, J. M. *et al.* (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403. See comments..
103. Gorodkin, J., Heyer, L. J., Brunak, S. & Stormo, G. D. (1997). Displaying the information contents of structural RNA alignments: the structure logos. *CABIOS*, **13**, 583–586.
104. Krezel, A. M., Wagner, G., Seymour-Ulmer, J. & Lazarus, R. A. (1994). Structure of the RGD protein decorsin: conserved motif and distinct function in leech proteins that affect blood clotting. *Science*, **264**, 1944–1947.

Edited by M. Levitt

(Received 4 January 2002; received in revised form 18 June 2002; accepted 10 July 2002)

http://www.academicpress.com/jmb

Supplementary Material for this paper comprising four Tables is available on IDEAL