J Mol Evol (2002) 55:104–110
DOI: 10.1007/s00239-001-2309-6

JOURNAL OF MOLECULAR EVOLUTION

Evolutionary Rate Heterogeneity in Proteins with Long Disordered Regions

Celeste J. Brown,[1] Sachiko Takayama,[1] Andrew M. Campen,[1] Pam Vise,[1] Thomas W. Marshall,[1] Christopher J. Oldfield,[1] Christopher J. Williams,[2] A. Keith Dunker[1]

[1] School of Molecular Biosciences, Washington State University, Pullman, WA 99164, USA
[2] Division of Statistics, University of Idaho, Moscow, ID 83844, USA

Received: 10 September 2001 / Accepted: 7 January 2002

Abstract. The dominant view in protein science is that a three-dimensional (3-D) structure is a prerequisite for protein function. In contrast to this dominant view, there are many counterexample proteins that fail to fold into a 3-D structure, or that have local regions that fail to fold, and yet carry out function. Protein without fixed 3-D structure is called intrinsically disordered. Motivated by anecdotal accounts of higher rates of sequence evolution in disordered protein than in ordered protein we are exploring the molecular evolution of disordered proteins. To test whether disordered protein evolves more rapidly than ordered protein, pairwise genetic distances were compared between the ordered and the disordered regions of 26 protein families having at least one member with a structurally characterized region of disorder of 30 or more consecutive residues. For five families, there were no significant differences in pairwise genetic distances between ordered and disordered sequences. The disordered region evolved significantly more rapidly than the ordered region for 19 of the 26 families. The functions of these disordered regions are diverse, including binding sites for protein, DNA, or RNA and also including flexible linkers. The functions of some of these regions are unknown. The disordered regions evolved significantly more slowly than the ordered regions for the two remaining families. The functions of these more slowly evolving disordered regions include sites for DNA binding. More work is needed to understand the underlying causes of the variability in the evolutionary rates of intrinsically ordered and disordered protein.

Key words: Disordered protein — Protein evolution — Rate heterogeneity

Introduction

The expanding field of structural genomics endeavors to ascertain the structural basis of protein function from amino acid sequence based upon sequence similarity to proteins with known structures (Burley 2000). An assumption behind this work is that structure is a prerequisite for function. However, from reports dating back to 1950, many unstructured or incompletely folded protein domains have been implicated in protein function (for a recent review see Dunker et al. 2001), thus calling into question the very basis of structural genomics efforts (Dunker and Obradovic 2001).

Proteins that fail to fold into a fixed three-dimensional (3-D) structure and yet carry out function in the incompletely folded state have been called "natively unfolded" (Weinreb et al. 1996), "intrinsically unstructured" (Wright and Dyson 1999), and "intrinsically disordered" (Dunker et al. 2001). Disordered protein is identified by missing electron density in X-ray crystallographic studies, by various results from nuclear magnetic resonance studies, and by a lack of signal in far-UV circular dichroism

Correspondence to: Celeste J. Brown; *email*: celesteb@disorder.chem.wsu.edu

studies. We have developed neural network predictors of protein disorder based upon these physically characterized regions of disordered proteins (Romero et al. 1997a, 1998, 1999, 2000, 2001; Garner et al. 1998; Li et al. 1999, 2000). Application of our predictors to various databases and genomes suggest that disorder is common (Romero et al. 1998; Dunker et al. 2000; Romero et al. 2001). An unexpected finding is that disordered protein apparently increases in commonness across the three kingdoms in the order eubacteria < archaea < eukarya. According to our predictions, more than 25% of the proteins in several eukaryotic genomes contain disordered regions 50 amino acids long or longer (Dunker et al. 2000). The commonness of disordered protein indicated by these findings strongly reinforce the earlier point that structural genomics efforts will be incomplete if protein disorder is not included on a systematic basis.

Further understanding of intrinsic disorder should come from the study of its evolution. It has been noted anecdotally that disordered amino acid sequences evolve more rapidly than ordered sequences. We showed that the disordered regions of eight calcineurins on average had a lower sequence similarity than their ordered regions (Dunker et al. 1998). Shaiu et al. (1999) noted that the disordered regions of topoisomerase II have more amino acid substitutions and insertions or deletions than the ordered regions of the same protein. From the assumption that sequence conservation derives from a 3-D structure, they suggested a similar evolutionary behavior for all disordered sequences. Ribosomal protein S4 (Sayers et al. 2000) and potassium channel subunits (Wissmann et al. 1999) are other proteins for which faster rates of evolution in the disordered regions have been noted. A counter example is flagellin, in which the ordered, central region of the protein has greater sequence diversity than the disordered termini (Vonderviszt et al. 1989). To our knowledge, a systematic investigation of rate heterogeneity in proteins with disordered sequences has not been done.

Herein, we begin characterization of the molecular evolution of disordered protein. Our purpose is to test the hypothesis that intrinsically disordered protein sequences evolve more rapidly than ordered sequences.

Methods

Proteins with both ordered regions and disordered regions 30 residues long or longer were chosen so that the order and disorder being compared had apparently had the same evolutionary histories. Furthermore, we chose proteins whose disorder was characterized by either X-ray crystallography or NMR so that our results could be interpreted more generally. We also included two proteins whose disordered regions were characterized by circular dichroism and limited proteolysis. Our method for constructing families of proteins with disordered regions and our method of analysis are outlined in Fig. 1. Disordered protein was identified either by missing electron density in X-ray crystal structure entries in PDB (Berman et al. 2000) or by word searches for "NMR" or "circular dichroism" and "disordered" or "unstructured" or "unfolded" in PubMed. The entire protein sequence was used to identify homologous members of the protein family by BLASTP searches (Altschul et al. 1990; 1997) of the nonredundant protein database at NCBI (www.ncbi.nlm.nih.gov). Distant homologues with only short regions of sequence similarity were eliminated to ensure that only homologous comparisons were made. All homologues (both paralogous and orthologous) were used except that identical sequences of the same length were reduced to a single representative per species. Homologues were aligned using the default settings of CLUSTALW (Thompson et al. 1994) at the Baylor College of Medicine web site (Smith et al. 1996). Alignments were not corrected by hand, to reduce the possibility of bias. Aligned sequences were then partitioned into ordered sequences and disordered sequences based on alignment with the structurally characterized sequence. Genetic distances between each pair of sequences in the ordered subset and between each pair in the disordered subset were calculated using Protdist from the PHYLIP computer package (Felsenstein 1993). Distance estimates by Protdist were based upon the Dayhoff PAM model (Dayhoff et al. 1978).

The hypothesis being tested is whether the average genetic distance between pairs of disordered sequences is significantly different from the average distance between pairs of ordered sequences within each family. The statistic used to test this hypothesis is the average of the difference between the ordered and the disordered genetic distance estimates for all pairwise comparisons within a family, $\Delta = \Sigma\,(O_{ij} - D_{ij}) / \{[n(n+1)/2] - n\}$, where O_{ij} is the pairwise genetic distance estimate between the ordered part of sequence i and the ordered part of sequence j, D_{ij} is the pairwise genetic distance estimate between the disordered part of sequence i and the disordered part of sequence j. The summation is over the ij pairs, where $i < j$, $i = 1,2, \ldots n{-}1$, $j = 2,3, \ldots n$, and n is the total number of sequences in the family.

To test the statistical significance of Δ, a sampling distribution for the test statistic under the null hypothesis of no difference between evolutionary rates of order and disorder, Δ_0, is required for each family. The sampling distribution of Δ_0 was estimated by randomly assigning amino acid positions to the disorder or order categories in proportion to the frequency of these categories in the original sequence. Genetic distances for the new data sets are then estimated, and the average difference is found as for the original set of sequences. This was done 1000 times, and the distribution of the 1000 Δ_0 was used as the sampling distribution. The sampling distribution indicates the frequency at which simulated Δ_0 values fall into a given range (bin) of values. The sampling distribution can be used to estimate the probability that we get a value that is equal to or more extreme than Δ simply by chance when the null hypothesis of no difference is true. When Δ falls completely outside of the range of simulated values the probability or p value is less than 1 in 1000 ($p < 0.001$), and the null hypothesis is rejected.

Results

Figure 2 illustrates the sampling distributions of the difference statistic Δ_0 for two protein families, replication protein A (RPA) (Jacobs et al. 1999) and tomato bushy stunt virus (TBSV) coat protein (Hopper et al. 1984). The y-axis is the number of times a simulated Δ_0 value fell into the bins whose midpoints are along the x-axis. Both families have seven members. The genetic distances for the viral

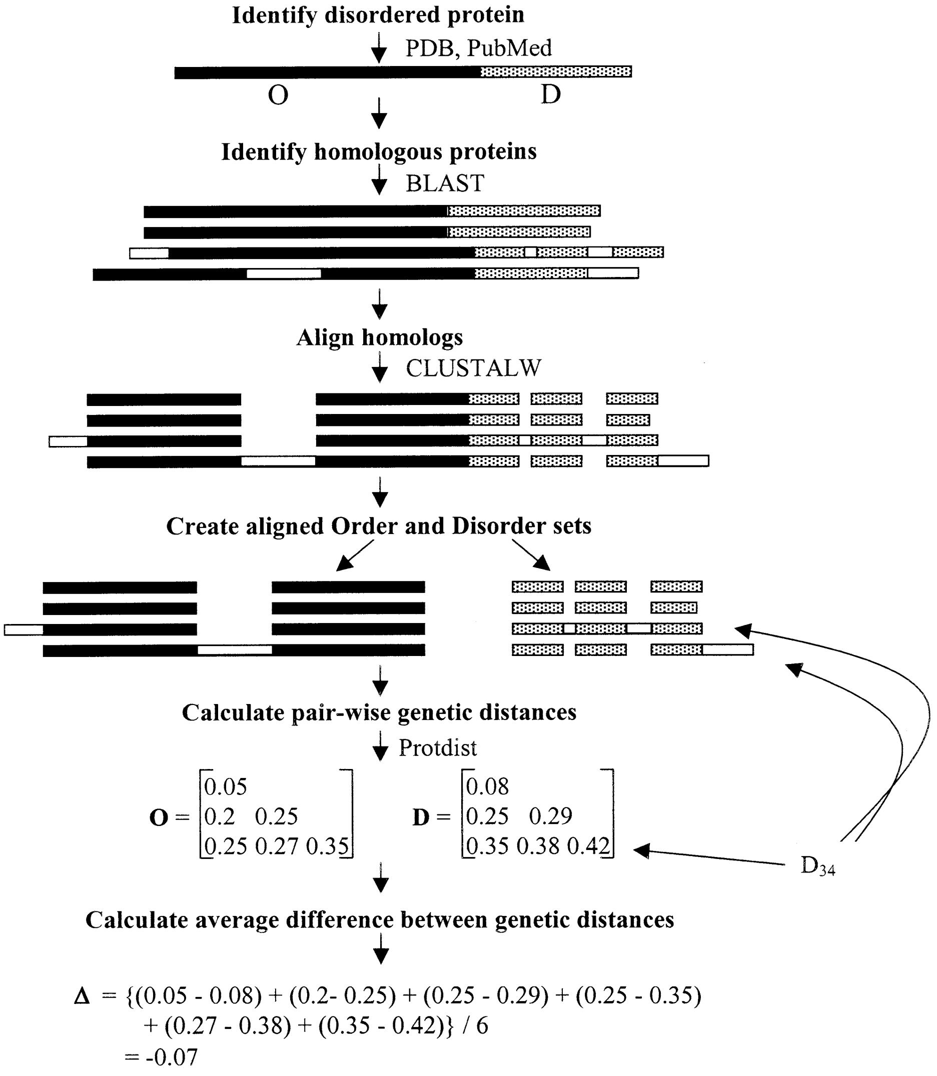
Identify disordered protein
PDB, PubMed
O
D
Identify homologous proteins
BLAST
Align homologs
CLUSTALW
Create aligned Order and Disorder sets
Calculate pair-wise genetic distances
Protdist
O =
0.05
0.2 0.25
0.25 0.27 0.35
D =
0.08
0.25 0.29
0.35 0.38 0.42
D34
Calculate average difference between genetic distances
Δ = {(0.05 - 0.08) + (0.2- 0.25) + (0.25 - 0.29) + (0.25 - 0.35)
+ (0.27 - 0.38) + (0.35 - 0.42)} / 6
= -0.07

Fig. 1. Procedures for identifying order and disorder in protein families and calculating average genetic distances between ordered and disordered protein. This procedure is followed for each protein family. *Black boxes* indicate ordered sequences, *gray boxes* indicate disordered sequences, and *white boxes* indicate insertions relative to the starting sequence.

coat protein family are much smaller than for RPA because only three viruses are represented among the seven sequences; the other sequences are strain variants of two of these viruses. The RPA family, on the other hand, represents the entire breadth of the eukaryotic kingdom, from yeast to rice to humans. Figure 2 clearly indicates the importance of determining the sampling distribution for each family. The Δ for the coat protein, −0.63, is significant at $p < 0.001$; that is, no simulated values were less than −0.63. This same Δ has a p value of 0.005 for the RPA distribution; that is, 5 of 1000 simulated Δ_0 had values lower than −0.63 and no values higher than 0.63.

The sampling distribution for RPA also illustrates the nonnormal nature of this distribution. A normal distribution would be symmetric around the 0 bin. Both distributions have modes at 0 but are skewed to the left with a few, highly negative values.

Table 1 lists the results for each of 26 protein families. The disordered regions of 6 proteins were determined by NMR, 17 by X-ray crystallography, and 1, the apoptosis regulator Bcl-x_L, by both NMR and crystallography. The disordered regions of two protein families were determined by a combination of far-UV circular dichroism and limited proteolysis. Family sizes ranged from 4 to 80. There was no significant difference ($p \leq 0.05$) in the average genetic distances of the ordered versus the disordered regions of five families. Nineteen proteins had significantly faster rates of evolution in their disordered regions, and two proteins had significantly slower rates in their disordered regions. There does not appear to be any relationship be-

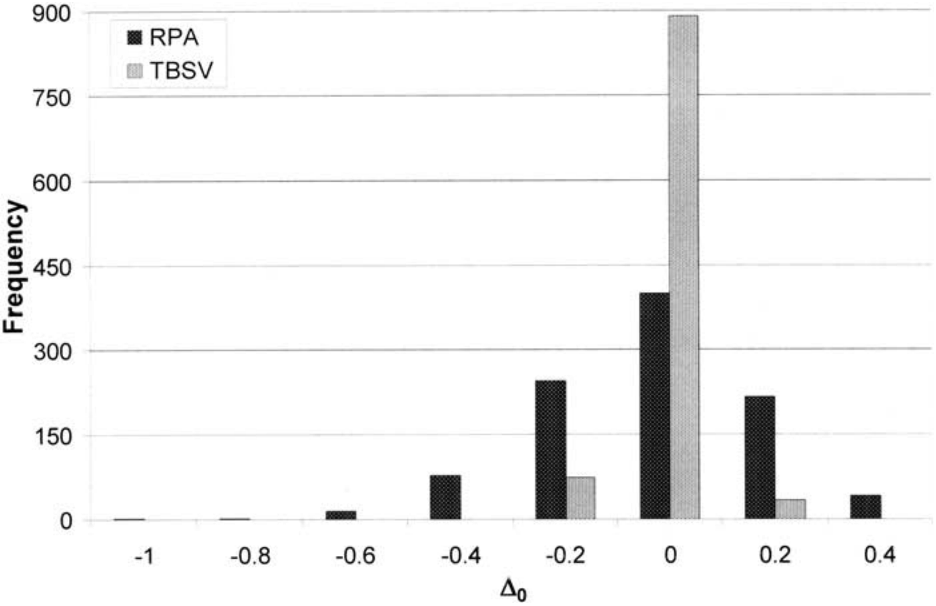

Fig. 2. Sampling distribution of the average difference in genetic distances of ordered and disordered proteins, Δ_0, for two protein families, replication protein A (RPA) and tomato bushy stunt virus coat protein (TBSV). The *y*-axis indicates the number of times simulated Δ_0 values fall into the bins whose midpoints are indicated on the *x*-axis.

Table 1. Average difference in genetic distance, Δ, between ordered and disordered regions of 26 protein families

Protein family	Reference	Detection method[a]	No. sequences	Δ[b]	*p* value[c]
Replication protein A	Jacobs et al. (1999)	NMR	7	−1.92	0.001
NF-KB p65	Schmitz et al. (1994)	NMR	4	−1.18	0.001
Glycyl-tRNA synthetase	Logan et acl. (1995)	X-Ray	24	−1.69	0.002
Regulator of G-protein signaling 4	Tesmer et al. (1997)	X-Ray	17	−0.96	0.001
Topoisomerase II	Berger et al. (1996)	X-Ray	28	−0.87	0.001
Calcineurin	Kissinger et al. (1995)	X-Ray	23	−0.84	0.001
c-Fos	Campbell et al (2000)	NMR	23	−0.82	0.001
Thyroid transcription factor	Tell et al. (1998)	CD, LP	12	−0.76	0.001
Sulfotransferase	Bidwell et al. (1999)	X-Ray	12	−0.74	0.013
Phenylalanine-tRNA synthetase	Mosyak et al. (1995)	X-Ray	14	−0.69	0.001
Coat protein, tomato bushy stunt virus	Hopper et al. (1984)	X-Ray	7	−0.63	0.001
Gonadotropin	Lapthorn et al. (1994)	X-Ray	9	−0.61	0.001
Coat protein, Sindbis virus	Choi et al. (1991)	X-Ray	6	−0.60	0.025
Histone H5	Aviles et al.(1978)	NMR	9	−0.41	0.001
Small heat shock protein	Kim et al. (1998)	X-Ray	6	−0.36	0.457
Telomere binding protein	Horvath et al. (1998)	X-Ray	8	−0.29	0.001
Cytochrome BC1	Iwata et al. (1998)	X-Ray	7	−0.27	0.034
DNA-lyase	Gorman et al. (1997)	X-Ray[d]	8	−0.18	0.001
Bcl-xL	Muchmore et al. (1996)	X-Ray, NMR	7	−0.13	0.001
Coat protein, southern bean mosaic virus	Silva and Rossmann (1985)	X-Ray	6	−0.09	0.100
α-Tubulin	Jimenez et al. (1999)	NMR	80	−0.06	0.034
Epidermal growth factor	Louie et al. (1997)	X-Ray	10	−0.03	0.736
Prion	Riek et al. (1997)	NMR	72	0.03	0.636
Glycine N-methyltransferase	Huang et al. (2000)	X-Ray	11	0.09	0.095
ssDNA binding protein	Tucker et al. (1994)	X-Ray	20	0.37	0.010
Flagellin	Vonderviszt et al. (1989)	LP	34	0.66	0.023

[a] Disordered state detected by NMR (nuclear magnetic resonance), X-Ray (X-ray crystallography), CD (circular dichroism), and LP (limited proteolysis).

[b] Negative values of Δ indicate that disordered regions are evolving more rapidly than ordered regions.

[c] For a two-sided test of the null hypothesis.

[d] Useful crystallization only in the absence of most of the disordered region.

tween family size or method of detecting disorder and whether disordered regions evolve more rapidly or more slowly.

Discussion

Our survey of proteins with ordered and disordered sequences indicates that, generally, disordered protein does evolve more rapidly than ordered. There are a few exceptions, which are discussed below.

Simulated sampling distributions allow reliable tests for each protein family. The null hypothesis for our sampling distribution is that the residues in any aligned amino acid position are as likely to be in a region of disorder as a region of order, and hence the pairwise genetic distances are equal. From previous work, however, we know that this is not the case. The amino acid composition of disordered regions of proteins is very different from the composition of ordered protein. For example, disordered protein has fewer aromatic amino acids, and more charged amino acids, than ordered protein (Xie et al. 1998; Romero et al. 2001; Williams et al. 2001). The aromatic amino acids in general have a lower substitution rate than the charged amino acids, and the difference in rates of evolution between ordered and disordered protein may be due to this difference in amino acid composition. The development of evolutionary models for disordered protein will clarify the importance of the amino acid composition and the physicochemical properties of disordered protein to their evolutionary rate.

Another possible explanation for the generally faster rates of evolution in disordered protein is that the disordered protein does not perform any particular function, and therefore its evolution is unconstrained. Indeed, the disordered regions of chorionic gonadotropin can be deleted without apparently affecting the known activity of this protein. Also, the functions for the disordered regions of glycyl-tRNA synthetase, the signal transduction inhibitor RGS4, small heat shock protein, DNA-lyase, and epidermal growth factor are unknown. However, the absence of effects on a given activity and absence of known function do not rule out the possibility of another function for the region of disorder. Furthermore, the other, rapidly evolving, disordered regions studied do have known functions, including binding to other molecules such as DNA, RNA, protein, and substrate. When involved in molecular interactions, these proteins typically undergo disorder-to-order transitions upon binding to their ligands (phenylalanine-tRNA synthetase and transcription factor c-Fos). For these proteins, a faster rate of evolution for the disordered region cannot be explained by a lack of function.

Another possible explanation is that not having a fixed structure is the function of the disordered region. Since there are many potential amino acid sequences that can lead to being unstructured, such a function could lead to very rapid rates of evolution. Two functions that fit this category are flexible linkers, as found in RPA and topoisomerase II, and target display, as found in calcineurin and Bcl-x_L. The function of the flexible linker in RPA is to tether two structured domains together so that they can separately attach to their respective targets (Jacobs et al. 1999). The disordered loop in the breast cancer gene, Bcl-x_L, contains protease digestion and phosphorylation sites that play critical roles in programmed cell death (Chang et al. 1997; Clem et al. 1998). We have proposed that having these sites in a disordered region increases their availability for binding by proteases or kinases (Dunker et al. 2001).

Finally, faster rates of evolution in disordered protein may be due to positive selection for variability within regions of disorder or strong purifying selection within regions of order (Yang and Bielawski 2000). Future investigations of the DNA sequences of these protein families will test these possibilities.

There are five protein families that do not have significantly different rates of evolution in their ordered and disordered regions. One of the protein families that shows no significant difference between rates is prion. The normal function of the prion protein is not known, but various lesions are caused by the accumulation of prion whose structure has switched from α-helix to β-sheet (Pan et al. 1993). The disordered region may be important for binding copper ions at the cell surface and may become structured upon copper binding (Aronoff-Spencer et al. 2000). Much of the disordered region is composed of octamer repeats that are the site of copper binding; maintaining the octamer sequence may constrain the evolutionary rate in this region.

The two proteins with slower rates of evolution in their disordered regions have well-characterized functions. The flexible loop of the adenovirus ssDNA binding protein forms part of the ssDNA binding interface and is essential for high-affinity binding to ssDNA. This region acts to unwind the ssDNA, thereby enabling replication (Dekker et al. 1998).

The other protein with a slower-evolving region of disorder is flagellin. The disordered segment in flagellin becomes ordered upon polymerization to form the flagellar filament (Aizawa et al. 1990). The ordered regions form the outside of the flagellar filament, whereas the disordered regions form the interior. From their surface-exposed position, the ordered regions are available as targets for antibodies, thus perhaps leading to positive selection for increased sequence diversity. This last example illustrates that factors controlling the relative rates of evolution of ordered and disordered regions of protein can be complex and suggests that testing for molecular adaptation in these proteins may be a fruitful avenue for future research.

The analyses reported here are based on very simple evolutionary models, yet 19 of the 26 families were shown with high statistical confidence to have disordered regions that evolve more rapidly than their ordered regions. By exploring the issues raised in our discussion, we will be able to gain further insight into the molecular evolution of intrinsically disordered protein.

Acknowledgments. We thank Z. Obradovic, P. Radivojac, G. Daughdrill, H. Wichman, and two anonymous reviewers for their helpful comments. Also, we would like to thank an anonymous reviewer for pointing out that an earlier sampling distribution was too liberal. We gratefully acknowledge support from NSF Grant CSE-IIS-9711532 to Z. Obradovic and A.K.D. and NIH Grant 1R01-LM06916 to A.K.D. and Z. Obradovic.

References

Aizawa SI, Vonderviszt F, Ishima R, Akasaka K (1990) Termini of salmonella flagellin are disordered and become organized upon polymerization into flagellar filament. J Mol Biol 211: 673–677

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402

Aronoff-Spencer E, Bums CS, Avdievich NI, Gerfen GJ, Peisach J, Antholine WE, Ball HL, Cohen FE, Prusiner SB, Millhauser GL (2000) Identification of the Cu(2+) binding sites in the N-terminal domain of the prion protein by EPR and CD spectroscopy. Biochemistry 39: 13760–13771

Aviles FJ, Chapman GE, Kneale GG, Crane-Robinson C, Bradbury EM (1978) The conformation of histone H5. Isolation and characterisation of the globular segment. Eur J Biochem 88: 363–371

Berger JM, Gamblin SJ, Harrison SC, Wang JC (1996) Structure and mechanism of DNA topoisomerase II. Nature 379: 225–232

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28: 235–242

Bidwell LM, McManus ME, Gaedigk A, Kakuta Y, Negishi M, Pedersen L, Martin JL (1999) Crystal structure of human catecholamine sulfotransferase. J Mol Biol 293: 521–530

Burley S (2000) An overview of structural genomics. Nat Struct Biol 7: 932–934

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

Chang BS, Minn AJ, Muchmore SW, Fesik SW, Thompson CB (1997) Identification of a novel regulatory domain in Bcl-x_L and Bcl-2. EMBO J 16: 968–977

Choi HK, Tong L, Minor W, Dumas P, Boege U, Rossmann MG, Wengler G (1991) Structure of Sindbis virus core protein reveals a chymotrypsin-like serine proteinase and the organization of the virion. Nature 354: 37–43

Clem RJ, Cheng EH, Karp CL, Kirsch DG, Ueno K, Takahashi A, Kastan MB, Griffin DE, Earnshaw WC, Veliuona MA, Hardwick JM (1998) Modulation of cell death by Bcl-x_L through caspase interaction. Proc Natl Acad Sci USA 95: 554–559

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. Atlas Prot Seq Struct 5: 345–352

Dekker J, Kanellopoulos PN, van Oosterhout JA, Stier G, Tucker PA, van der Vliet PC (1998) ATP-independent DNA unwinding by the adenovirus single-stranded DNA binding protein requires a flexible DNA binding loop. J Mol Biol 277: 825–838.

Dunker AK, Obradovic Z (2001) The protein trinity—Linking function and disorder. Nat Biotech 19: 805–806

Dunker AK, Garner E, Guilliot S, Romero P, Albrecht K, Hart J, Obradovic Z, Kissinger C, Villafranca JE (1998) Protein disorder and the evolution of molecular recognition: Theory, predictions and observations. Pacific Symp Biocomput 3: 473–484

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

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

Felsenstein J (1993) PHYLIP (phylogen inference package). Distributed by the author, Department of Genetics, University of Washington, Seattle

Garner E, Cannon P, Romero P, Obradovic Z, Dunker AK (1998) Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization. Genome Inform 9: 201–213

German MA, Morera S, Rothwell DG, de La Fortelle E, Mol CD, Tainer JA, Hickson ID, Freemont PS (1997) The crystal structure of the human DNA repair endonuclease HAP1 suggests the recognition of extra-helical deoxyribose at DNA abasic sites. EMBO J 16: 6548–6558

Hopper P, Harrison SC, Sauer RT (1984) Structure of tomato bushy stunt virus. V. Coat protein sequence determination and its structural implications. J Mol Biol 177: 701–713

Horvath MP, Schweiker VL, Bevilacqua JM, Ruggles JA, Schultz SC (1998) Crystal structure of the *Oxytricha nova* telomere end binding protein complexed with single strand DNA. Cell 95: 963–974

Huang Y, Komoto J, Konishi K, Takata Y, Ogawa H, Gomi T, Fujioka M, Takusagawa F (2000) Mechanisms for auto-inhibition and forced product release in glycine N-methyltransferase: Crystal structures of wild-type, mutant R175K and S-adenosylhomocysteine-bound R175K enzymes. J Mol Biol 298: 149–162

Iwata S, Lee JW, Okada K, Lee JK, Iwata M, Rasmussen B, Link TA, Ramaswamy S, Jap BK (1998) Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex. Science 281: 64–71

Jacobs DM, Lipton AS, Isern NG, Daughdrill GW, Lowry DF, Gomes X, Wold MS (1999) Human replication protein A: global fold of the N-terminal RPA-70 domain reveals a basic cleft and flexible C-terminal linker. J Biomol NMR 14: 321–331

Jimenez MA, Evangelic JA, Aranda C, Lopez-Brauet A, Andreu D, Rico M, Lagos R, Andreu JM, Monasterio O (1999) Helicity of α(404-451) and β(394-445) tubulin C-terminal recombinant peptides. Protein Sci 8: 788–799

Kim KK, Kim R, Kim SH (1998) Crystal structure of a small heat-shock protein. Nature 394: 595–599

Kissinger CR, Parge HE, Knighton DR, Lewis CT, Pelletier LA, Tempczyk A, Kalish VJ, Tucker KD, Showalter RE, Moomaw

EW, Gastinel LN, Habuka N, Chen X, Maldanado F, Barker JE, Bacquet R, Villafranca JE (1995) Crystal structures of human calcineurin and the human FKBP12-FK506- calcineurin complex. Nature 378: 641–644

Lapthom AJ, Harris DC, Littlejohn A, Lustbader JW, Canfield RE, Machin KJ, Morgan FJ, Isaacs NW (1994) Crystal structure of human chorionic gonadotropin. Nature 369: 455–461

Li X, Romero P, Rani M, Dunker AK, Obradovic Z (1999) Predicting protein disorder for N-, C-, and internal regions. Genome Inform 10: 30–40

Li X, Obradovic Z, Brown CJ, Garner EC, Dunker AK (2000) Comparing predictors of disordered protein. Genome Inform 11: 172–184

Logan DT, Mazauric MH, Kern D, Moras D (1995) Crystal structure of glycyl-tRNA synthetase from *Thermus thermophilus*. EMBO J 14: 4156–4167

Louie GV, Yang W, Bowman ME, Choe S (1997) Crystal structure of the complex of diphtheria toxin with an extracellular fragment of its receptor. Mol Cell 1: 67–78

Mosyak L, Reshetnikova L, Goldgur Y, Delarue M, Safro MG (1995) Structure of phenylalanyl-tRNA synthetase from *Thermus thermophilus*. Nat Struct Biol 2: 537–547

Muchmore SW, Sattler M, Liang H, Meadows RP, Harlan JE, Yoon HS, Nettesheim D, Chang BS, Thompson CB, Wong SL, Ng SL, Fesik SW (1996) X-ray and NMR structure of human Bcl-x_L, an inhibitor of programmed cell death. Nature 381: 335–341

Pan KM, Baldwin M, Nguyen J, et al. (1993) Conversion of α-helices into β-sheets features in the formation of the scrapie prion proteins. Proc Natl Acad Sci USA 90: 10962–10966

Riek R, Hornemann S, Wider G, Glockshuber R, Wuthrich K (1997) NMR characterization of the full-length recombinant murine prion protein, mPrP(23-231). FEBS Lett 413: 282–288

Romero P, Obradovic Z, Dunker AK (1997a) Sequence data analysis for long disordered regions prediction in the calcineurin family. Genome Inform 8: 110–124

Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Dunker AK (1997b) Identifying disordered regions in proteins from ammo acid sequences. Proc IEEE Int Conf Neural Networks 1: 90–95

Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Guilliot S, Garner E, Dunker AK (1998) Thousands of proteins likely to have long disordered regions. Pacific Symp Biocomput 3: 437–448

Romero P, Obradovic ZC, Dunker AK (2000) Intelligent data analysis for protein disorder prediction. Artificial Intelligence Rev 14: 447–484

Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK (2001) Sequence complexity of disordered protein. Proteins 42: 38–48

Sayers EW, Gerstner RB, Draper DE, Torchia DA (2000) Structural preordering in the N-terminal region of ribosomal protein S4 revealed by heteronuclear NMR spectroscopy. Biochemistry 39: 13602–13613

Schmitz ML, dos Santos Silva MA, Altmann H, Czisch M, Holak TA, Baeuerle PA (1994) Structural and functional analysis of the NF-KB p65 C terminus. An acidic and modular transactivation domain with the potential to adopt an alpha-helical conformation. J Biol Chem 269: 25613–25620

Shaiu WL, Hu T, Hsieh TS (1999) The hydrophilic, protease-sensitive terminal domains of eucaryotic DNA topoisomerases have essential intracellular functions. Pacific Symp Biocomput 4: 578–589

Silva AM, Rossmann MG (1985) The refinement of southern bean mosaic virus in reciprocal space. Acta Crystallogr B41 :147–157

Smith RF, Wiese BA, Wojzynski MK, Davison DB, Worley KC (1996) BCM Search Launcher — An integrated interface to molecular biology data base search and analysis services available on the World Wide Web. Genome Res 6: 454–462

Tell G, Perrone L, Fabbro D, Pellizzari L, Pucillo C, De Felice M, Acquaviva R, Formisano S, Damante G (1998) Structural and functional properties of the N transcriptional activation domain of thyroid transcription factor-1: Similarities with the acidic activation domains. Biochem J 329: 395–403

Tesmer JJ, Herman DM, Gilman AG, Sprang SR (1997) Structure of RGS4 bound to A1F4-activated $G_{i\alpha 1}$: Stabilization of the transition state for GTP hydrolysis. Cell 89: 251–261

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680

Tucker PA, Tsernoglou D, Tucker AD, Coenjaerts FE, Leenders H, van der Vliet PC (1994) Crystal structure of the adenovirus DNA binding protein reveals a hook-on model for cooperative DNA binding. EMBO J 13: 2994–3002

Vonderviszt F, Kanto S, Aizawa S, Namba K (1989) Terminal regions of flagellin are disordered in solution. J Mol Biol 209: 127–133

Weinreb PH, Zhen W, Poon AW, Conway KA, Lansbury Jr. PT, (1996) NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. Biochemistry 35: 13709–13715

Williams RM, Obradovic Z, Mathura V, Braun W, Garner EC, Young J, Takayama S, Brown CJ, Dunker AK (2001) The protein non-folding problem: Amino acid determinants of intrinsic order and disorder. Pacific Symp Biocomput 6: 89–100

Wissmann R, Baukrowitz T, Kalbacher H, Kalbitzer HR, Ruppersberg JP, Pongs O, Antz C, Fakler B (1999) NMR structure and functional characteristics of the hydrophilic N terminus of the potassium channel β-subunit Kvβ1.1. J Biol Chem274: 35521–35525

Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293: 321–331

Xie Q, Arnold GE, Romero P, Obradovic Z, Garner E, Dunker AK (1998) The sequence attribute method for determining relationships between sequence and protein disorder. Genome Inform 9: 193–200

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaption. Trends Ecol Evol 15: 496–503