

G protein-coupled receptors show unusual patterns of intrinsic unfolding

Veli-Pekka Jaakola^{1,2}, Jaime Prilusky³, Joel L. Sussman⁴
and Adrian Goldman^{1,5}

¹Institute of Biotechnology (Biocenter 3), University of Helsinki, PO Box 65, Viikinkaari 1, FIN-00014 Helsinki, Finland, ²Viikki Graduate School in BioSciences, PO Box 56, Viikinkaari 9, FIN-00014 Helsinki, Finland, ³Department of Biological Services and ⁴Department of Structural Biology, Weizmann Institute of Science, 76100 Rehovot, Israel

⁵To whom correspondence should be addressed.
E-mail: adrian.goldman@helsinki.fi

Intrinsically unstructured proteins (IUPs) or IUP-like regions often play key roles in controlling processes ranging from transcription to the cell cycle. *In silico* such proteins can be identified by their sequence properties; they have low hydrophobicity and high net charge. In this study, we applied the FoldIndex (<http://bioportal.weizmann.ac.il/fldbin/findex>) program to analyze human G protein-coupled receptors and compared them with membrane proteins of known structure and with IUPs. We show that human G protein-coupled receptor (GPCR) extramembranous domains include long (>50 residues) disordered segments, unlike membrane proteins of known structure. The predicted disorder occurred primarily in the N-terminal, C-terminal and third intracellular domain regions: 55, 69 and 56% of the human GPCRs were disordered in these regions, respectively. This increased flexibility may therefore be critical for GPCR function. Surprisingly, however, the kinds of residues used in GPCR unstructured regions were different than in hitherto-identified IUPs. The GPCR third intracellular loop domains contain very high percentages of Arg, Lys and His residues, especially Arg, but the percentage of Glu, Asp and Pro is no higher than in folded proteins. We propose that this has structural and functional consequences.

Keywords: G protein-coupled receptors/intrinsically unstructured proteins/membrane proteins/sequence prediction

Introduction

G protein-coupled receptors (GPCRs) are the largest family of mammalian cell surface receptors, responsible for primary communication between cells and their environment. Consequently, GPCR agonists and antagonists are important therapeutically in the treatment of pain, depression, hypertension, cardiac dysfunction, anxiety and inflammation (Spiegel and Weinstein, 2004): as many as 50% of currently used drugs affect GPCRs (Klabunde and Hessler, 2002). Understanding GPCR structure and function is therefore essential. Unfortunately, bovine rhodopsin is the only member of the GPCRs whose structure is known (Palczewski *et al.*, 2000). All GPCRs have seven transmembrane α -helices (TMs) connected by intracellular (1i–3i) and extracellular (1e–3e) domains/loops, with an N-terminal (N) and a C-terminal (C) domain (Baldwin

et al., 1997) (Figure 1). By amino acid sequence alignment, GPCRs can be classified into six main categories: rhodopsin-like (class A), secretin-like, metabotropic glutamate/pheromone, fungal pheromone, cAMP receptors and frizzled/smoothed family (Horn *et al.*, 2001), of which class A is the largest and most extensively studied. Within class A, most of the natural ligands and drugs bind in the middle of the bundle of seven α -helices, while the intracellular domains including the C-terminus bind G proteins and other signaling proteins and are therefore responsible for transmitting the ligand-induced signal (Gether and Kobilka, 1998). Consequently, it is interesting that some of these domains are not well defined in the rhodopsin crystal structures (Palczewski *et al.*, 2000; Okada *et al.*, 2002).

Some proteins and protein domains appear to have little or no ordered structure under physiological conditions when studied by nuclear magnetic resonance (NMR), small-angle X-ray scattering, circular dichroism (CD) spectroscopy, light scattering spectra and analytical ultracentrifugation (Uversky *et al.*, 2000; Dunker *et al.*, 2001, 2002; Uversky, 2002a; Dafforn and Rodger, 2004; Uversky and Fink, 2004). Such proteins, now called native unfolded or intrinsically unstructured proteins (IUPs) (Wright and Dyson, 1999), differ from the structures found in the protein database [PDB; <http://www.rcsb.org/pdb/> (Berman *et al.*, 2000)] in amino acid composition, sequence complexity, hydrophobicity, charge and flexibility (Tompa, 2002). Surprisingly, IUPs are evolutionarily stable, although they have their own substitution nature and speed (Iakoucheva *et al.*, 2001). Based on the sequence features mentioned above, tools have been developed for predicting whether a given primary protein sequence is an IUP or not (Uversky *et al.*, 2000; Linding *et al.*, 2003). Most IUPs have regulatory roles in basic cellular processes such as transcription, translation, signal transduction and the cell cycle, suggesting that the structural disorder is essential in these cellular processes (Wright and Dyson, 1999; Tompa, 2002; Ward *et al.*, 2004). One example of such a protein is the cyclin-dependent kinase inhibitor p21, which has an important role in the cell cycle (Johnson and Walker, 1999). Based on NMR and CD studies, it is unfolded *in vitro* but becomes folded when it interacts with cyclin-dependent kinase Cdk2 (Kriwacki *et al.*, 1996; Pavletich, 1999).

The N, C, intracellular (1i–3i) and extracellular (1e–3e) domains of GPCRs are usually fairly large (>30 residues), especially in the rhodopsin-like class A GPCRs (Baldwin *et al.*, 1997; Horn *et al.*, 2001). Using the intrinsic protein disorder prediction program FoldIndex, we show here that, for human GPCRs, 55% of the N-termini, 69% of the C-termini and 56% of the third intracellular loop appear to contain intrinsically unstructured regions (IURs). We therefore compared in detail the folding analysis of 147 human class A GPCRs with integral membrane proteins of known structure, with soluble proteins and with previously identified IUPs. Our

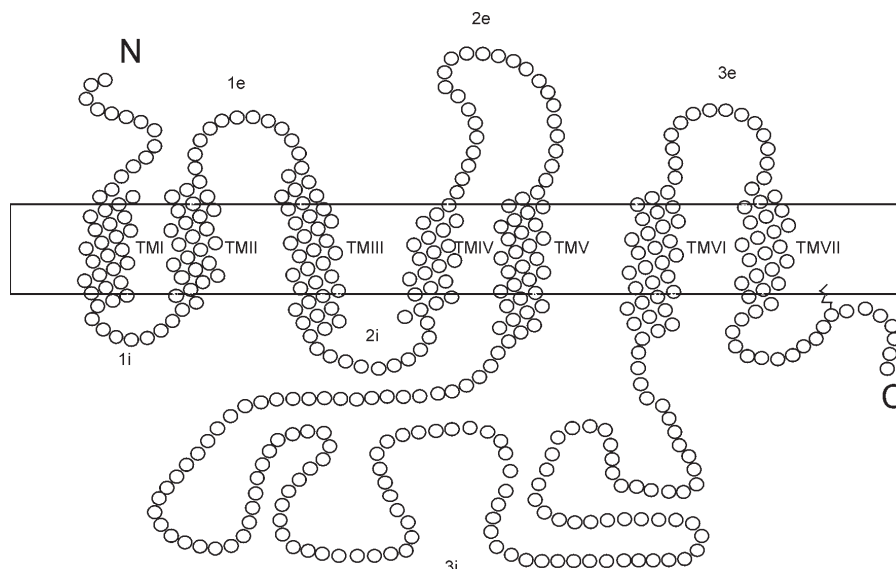


Fig. 1. Schematic representation of a GPCR. The seven α -helical transmembrane segments (TMI–VII) are connected by three extracellular (1–3e) and three intracellular (1–3i) domains/loops. The N-terminus is present on the extracellular side, whereas the C-terminus is present on the intracellular side.

analysis demonstrates that class A 3i domains usually contain long disordered stretches, unlike most membrane proteins of known structure. This suggests that intrinsic unfolding may play a role in GPCR functionality *in vivo*. Consistent with this, the 3i domains have an unusual amino acid distribution, suggesting that they form a hitherto-unidentified class of IUPs.

Materials and methods

Sequence analysis

Data set I. We chose a set of 343 human GPCRs sequences from the SWISS-PROT protein sequence data bank such that the entries were unique and cross-referenced according to the GPCRDB cross-reference database [http://www.gpcr.org/ (Horn *et al.*, 2001)]; a list of selected GPCRs is given in the Supplementary data available at *PEDS Online*. The boundaries for the helices were taken from the SWISS-PROT entries, which are based on the bovine rhodopsin crystal structure [http://www.gpcr.org/ (Horn *et al.*, 2001)]. We recorded the ID number, accession number and overall sequence length for each protein. We also calculated and collated the overall average FoldIndex [http://bioportal.weizmann.ac.il/fldbin/index (Zeev-Ben-Mordehai *et al.*, 2003)] of each extramembranous domain and the residue number of the start and end of each extramembranous domain (EMD). The FoldIndex (I_F) was calculated as in Zeev-Ben-Mordehai *et al.* (2003) using Equation 1:

$$I_F = 2.785H - C - 1.151 \quad (1)$$

where H is the mean hydrophobicity and C is the mean net charge. We initially calculated, using a minimal window length of 10 residues, for each EMD, the average most negative FoldIndex value and its length and the average second most negative FoldIndex value in each and its length in order to detect short IURs in the GPCRs. Having identified that long IURs often occurred, we subsequently used a window length of 25 residues. We recorded for each protein mean hydrophobicity, mean net charge, number of disordered regions, length

of the longest disordered region and number of disordered residues (Table I).

Data set II and reference sets. We further selected the 147 human class A GPCRs for hand classification. As a control group, we chose a representative set of solved X-ray structures of integral membrane proteins from the PDB: from the 143 solved membrane protein structures (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html), we produced a non-redundant set, by eliminating different photocycle states of bacteriorhodopsin, for example. Our final set contains 91 structures. The sequences were extracted from the PDB, because the sequence in the PDB can not be reconciled with the sequence databases in an automated fashion. Again using the FoldIndex server, we calculated folding plot, mean FoldIndex, mean net charge and mean net hydrophobicity using a window of 25 residues. We also used data from Uversky *et al.* (2000) for folded soluble proteins and proteins experimentally shown to be IUP.

For comparison, full sequences from data set I and the control set of membrane proteins (MPs) were analyzed using the neural network program DisEMBL [http://dis.embl.de/ (Linding *et al.*, 2003)], run with default parameters. We recorded, using the ‘remark465’ prediction, maximum length of disorder for each sequence from data set I and MPs. The neural network is trained to recognize disorder in sequences based on coordinates missing in the PDB files used in the training set (Linding *et al.*, 2003).

In the charge–hydrophobicity plots (Figure 2), the borderline between IUPs and native proteins was calculated as in Uversky *et al.* (2000) using Equation 2:

$$H_{\text{boundary}} = (C + 1.151)/2.785 \quad (2)$$

where H is the mean hydrophobicity and C is the mean net charge. Using FoldIndex, we considered a region to be unfolded if its FoldIndex value is ≤ -0.3 ; a value of -0.3 reduces the possibility of false positives (sequences ascribed as unfolded) for short loops. Amino acid sequence distributions were calculated using the VectorNTISuite 5.5 (InforMax, USA).

Table I. Comparison of mean net folding indexes, mean net charge and mean net hydrophobicity (more detailed comparison of membrane proteins of known structure is given in the Supplementary data)

	No. of proteins	Mean length ^b	Mean FoldIndex	Mean net charge	Mean net hydrophobicity
Reference proteins					
<i>Soluble proteins</i> ^a					
Folded	366 (100%)	–	–	0.04 ± 0.04	0.48 ± 0.03
IUP/natively unfolded	275 (75%)	–	–	0.12 ± 0.09	0.39 ± 0.05
MPs	91 (25%)	–	–	0.016 ± 0.002	0.511 ± 0.006
Transmembrane	91 (100%)	355	0.254 ± 0.016	0.016 ± 0.002	0.511 ± 0.006
7-TM	86 (95%)	343	0.258 ± 0.017	0.017 ± 0.002	0.512 ± 0.006
β-Barrel	4	266	0.499 ± 0.036	0.008 ± 0.003	0.603 ± 0.005
Monotopic membrane	25	360	0.078 ± 0.011	0.031 ± 0.005	0.452 ± 0.004
	5 (5%)	566	0.177 ± 0.021	0.007 ± 0.004	0.479 ± 0.007
Sample set					
Human GPCRs	343 (100%)	447	0.366 ± 0.005	0.025 ± 0.001	0.553 ± 0.002
Rhodopsin-like	147 (43%)	403	0.353 ± 0.007	0.026 ± 0.001	0.549 ± 0.002
3i domain	147	52	−0.145 ± 0.018	0.167 ± 0.008	0.419 ± 0.005
Largest i domain	147	77	−0.037 ± 0.156	0.110 ± 0.007	0.430 ± 0.005
Largest i domain (ex 3i)	82	69	0.014 ± 0.115	0.063 ± 0.005	0.441 ± 0.004
N-Terminal domain	147	50	0.070 ± 0.025	0.049 ± 0.009	0.456 ± 0.005
C-Terminal domain	147	57	−0.003 ± 0.022	0.055 ± 0.005	0.0432 ± 0.004

^aTaken from Uversky *et al.* (2000); the mean length and mean folding index were not analyzed.

^bThe mean length of membrane proteins of known structure is from the PDB files (total sequence length divided by number of chains). Mean hydrophobicity (Kyte and Doolittle, 1982) is $\Sigma(\text{hydrophobicity})/\text{sequence length}$; Mean net charge is $\Sigma(\text{net charge})/\text{sequence length}$.

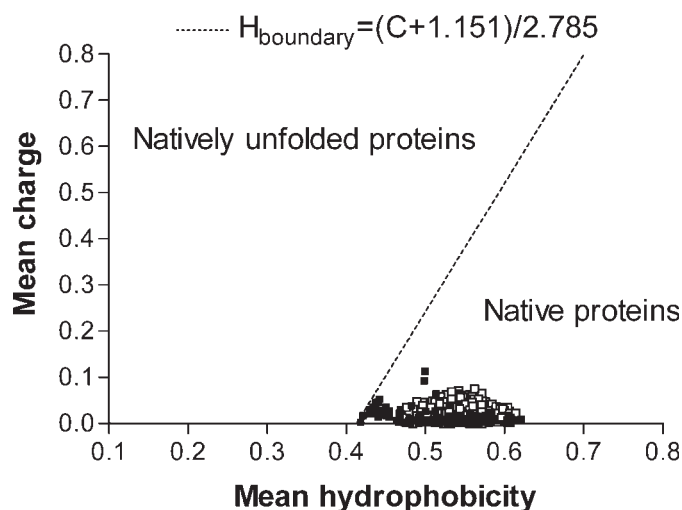


Fig. 2. Mean net charge and the mean hydrophobicity for 91 solved membrane proteins (black boxes) and 343 GPCRs (white boxes). The dashed line represents the border between intrinsically unstructured and native proteins calculated using Equation 2 (see text).

Results

Predicted disorder in human GPCRs and membrane proteins of known structure

The relationship of net mean hydrophobicity to net mean charge of both data set I [all human GPCRs ($n = 343$)] and the control set of solved membrane proteins ($n = 91$) was essentially the same (Figure 2); the overall FoldIndex value (\pm standard error) for full-length sequences was 0.366 (± 0.005) for GPCRs and 0.254 (± 0.016) for membrane proteins of known structure (MP). Both sets of proteins should be folded; if anything, the GPCRs are ‘more folded’ than the control set. We thought that this was probably due to the transmembrane helices in GPCRs and to study this closer, we classified the MPs as listed in the web site (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html; Table I). It is

clear that the FoldIndex server does not predict some of the secondary elements in β -barrel structures, such as porins, very well (overall score of 0.078 ± 0.011 , $n = 25$; Table I). Although porins as a whole were folded by FoldIndex, some areas of regular secondary structure were predicted to be disordered using default values. This problem was avoided by applying cut-off values (FoldIndex ≤ -0.3) to the data set. With these parameters, both FoldIndex and DisEMBL had similar length distributions of predicted unfolded regions (Figure 3).

Our work on α_2 -adrenergic receptors indicated that their third intracellular loop might be disordered (Jaakola *et al.*, 2005), which is consistent with the structures of bovine rhodopsin (Palczewski *et al.*, 2000; Okada *et al.*, 2002). We therefore examined the distribution of the longest disordered region of GPCRs and MPs. The distribution appeared to be somewhat different (Figure 3); GPCRs contain long regions (>30 residues) with predicted disorder (Table II, Figure 3), whereas most MPs, particularly seven α -helical transmembrane proteins (7-TMs), appear only to have shorter such regions. As anticipated, all the areas of GPCRs predicted to be unfolded were extramembranous.

We therefore analyzed the folding patterns of GPCR extramembranous domains using FoldIndex. For the N region, in 55% of the GPCRs there was at least one clear IUR (FoldIndex ≤ -0.3) region (Table II). The percentages for the C region and 3i region were 69 and 56%, respectively. Conversely, the 1i, 1e and 3e regions appeared to be less unfolded (Table II and Figure 4).

Detailed analysis of rhodopsin-like GPCRs

We also compared rhodopsin-like class A GPCRs (dataset II; $n = 147$) with the solved integral membrane proteins (above) and with data from Uversky *et al.* (2000) (Table I). We chose class A GPCRs because the only solved GPCR is bovine rhodopsin (Palczewski *et al.*, 2000) and because our previous studies have focused on the rhodopsin-like α_2 -adrenergic receptor (Liitti *et al.*, 1997; Bartus *et al.*, 2003; Sen *et al.*, 2003; Jaakola *et al.*, 2005).

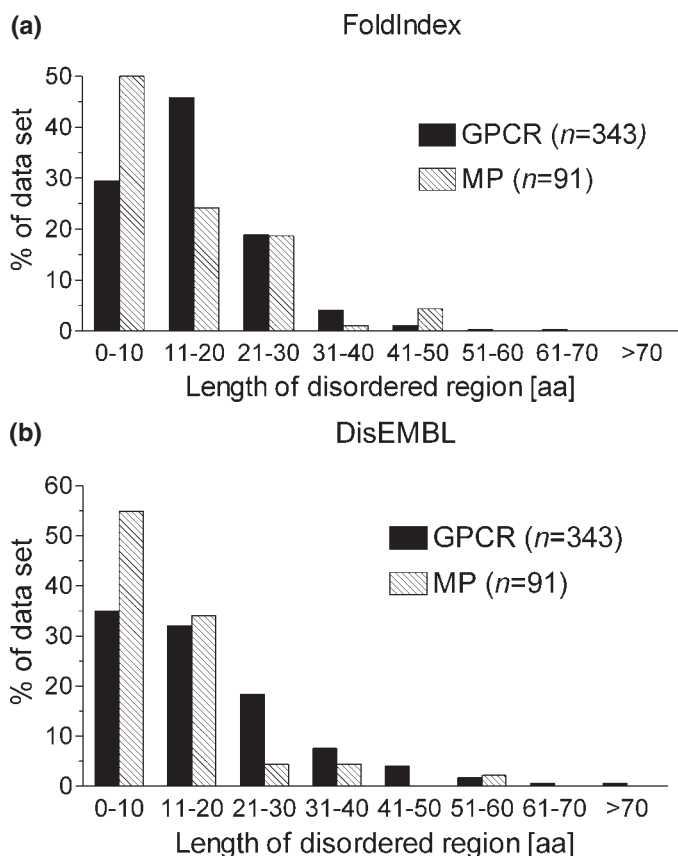


Fig. 3. Histogram showing the frequency distribution of the lengths of the predicted unfolded regions for membrane proteins of known structure and human GPCRs. The prediction was done using the FoldIndex (a) and DisEMBL (b) servers. FoldIndex cutoffs were score ≤ -0.3 (see Materials and methods). The four outliers (>40 residues predicted as unfolded) in MPs are PDB entries 1kmo, 1bcc, 1byg and 1zzv.

Table II. Number of negative FoldIndex scores, average, minimum and maximum length of human GPCR extramembranous domains (scores were calculated using a window of 10 amino acids and the numbers of <0 and ≤ -0.3 FoldIndex values are shown)

	Average length	Minimum length	Maximum length	Number (%) FoldIndex <0	Number (%) FoldIndex ≤ -0.3
N	102	6	2507	279 (81%)	190 (55%)
1i	10	3	37	130 (38%)	95 (28%)
1e	17	4	62	77 (22%)	42 (12%)
2i	19	7	40	158 (46%)	111 (32%)
2e	27	4	179	162 (47%)	103 (30%)
3i	34	4	239	244 (71%)	192 (56%)
3e	14	4	63	111 (32%)	75 (22%)
C	58	4	537	326 (95%)	237 (69%)

Class A GPCRs ($n = 147$), folded soluble proteins ($n = 275$) and membrane proteins of known structure ($n = 91$) all have similar overall mean net charge and mean net hydrophobicity (Table I and Figure 2). The class A intracellular domains nonetheless have IUR-like mean net charge properties (0.12 for natively undolded/IURs and 0.11 for the largest intracellular domain of class A GPCRs), unlike folded proteins, whose mean net charge is nearly neutral (0.04) (Table I). Of the GPCR loops, the 3i was by far the most unfolded and had the highest

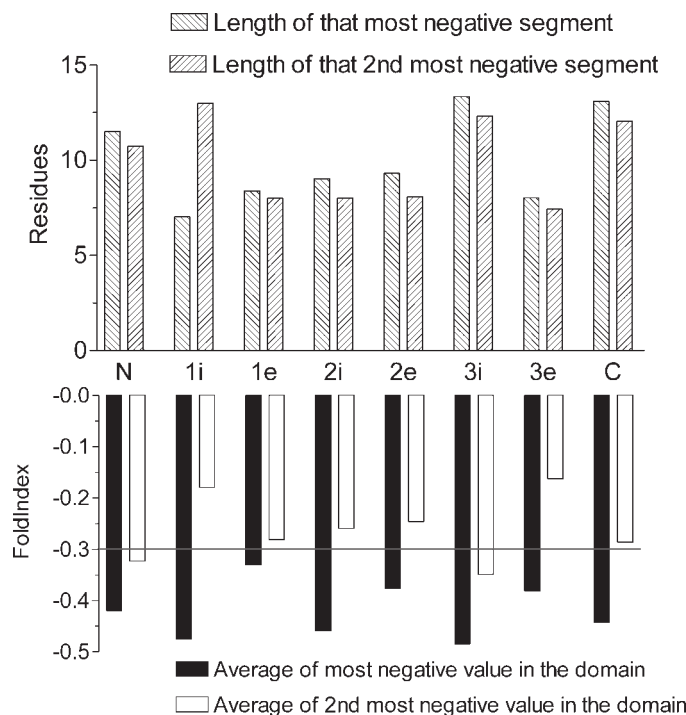


Fig. 4. Comparison of unfolding and sequence length for the set of 343 human GPCRs. The average most negative value and second most negative values are reported for each extramembranous domain (N, 1i, 1e, 2i, 2e, 3i, 3e and C) and the length of these values calculated using FoldIndex. The cut-off values of ≤ -0.3 is marked by a dotted line.

net charge (0.17) (Table I). Conversely, the N- and C-termini are marginally folded (FoldIndex score of 0.07 and -0.003 , respectively) with net charges similar to folded soluble proteins (Table I). There was no correlation between domain length and IUP-like nature; domains as long as 100 residues all appeared to be unfolded (Figure 5). All membrane proteins have higher net hydrophobicity than the soluble IUPs, even in the apparently disordered loops, (scores of 0.43–0.55 versus 0.39) (Table I). This suggested that there might be significant sequence differences between soluble IUPs and IURs of GPCRs.

We therefore compared the amino acid distribution of human rhodopsin-like GPCRs, their 3i domain and the N- and C-termini with the control data sets (Table III). The residue distribution in 3i loops appears to be significantly different from all other classes (Table III; Figure 6). In comparison with IUPs, 3i domains are very positively charged because the percentage of Arg residues is very high (Table III; Figure 6). This is also true for the C-termini, but not for the N-termini of GPCRs. In comparison with IUPs, 3i loops also have much more Ala and Leu, somewhat higher percentages of His, Cys and Trp, a similar percentage of Lys, but much less Asp and Glu (Table III; Figure 6). The C-termini are somewhat similar to the 3i domains, with raised Arg, Lys and His and lowered Asp and Glu. The N-termini are, however, different: Asp and Glu are not significantly low, whereas Lys, His and Arg are. This presumably reflects the different environments; the extracellular side of GPCRs is probably required to be ordered.

The 3i amino acid composition is, surprisingly, in some respects similar to that of folded proteins (Table III). The percentages of Ala, Cys, Glu, Ile, Leu, Met, Pro, Thr and Trp are similar. Conversely, 3i loops have very increased Arg and Lys

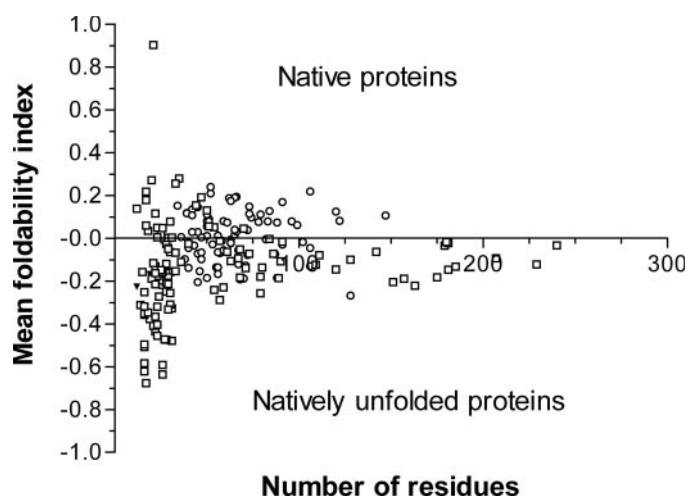


Fig. 5. Mean FoldIndex versus length, 3i domain of rhodopsin-like GPCR (squares); largest extramembranous domain of rhodopsin-like GPCR, if not 3i (circles); ordered extramembranous domain from a seven α -helical transmembrane protein (7-TM) (upright triangles); disordered (no coordinate in the PDB file) extramembranous domain from a 7-TM (inverted triangles).

Table III. Amino acid frequencies of data sets of ordered and disordered proteins

Amino acid ^a	Globular ^b	MPs ^c	GPCRs ^d	3i ^d	N ^d	C ^d	IUP ^b
Ala	8.15	8.56	8.45	9.17	8.31	6.96	7.15
Cys	1.64	1.14	3.24	1.67	1.85	3.76	0.61
Asp	5.78	4.47	2.92	2.21	4.92	4.22	5.05
Glu	5.98	4.90	3.12	4.52	6.28	5.31	14.26
<i>Phe</i>	3.95	4.96	5.71	1.94	3.17	4.52	1.66
Gly	7.99	8.48	5.36	5.53	7.98	5.60	4.31
His	2.33	1.95	2.04	2.61	1.88	2.70	1.15
<i>Ile</i>	5.43	5.78	6.26	4.65	3.23	3.00	3.67
Lys	5.43	4.16	3.53	10.26	2.70	6.14	10.43
<i>Leu</i>	8.37	10.13	12.47	8.05	9.35	8.42	5.44
Met	2.03	2.66	2.62	2.21	2.73	1.50	1.30
<i>Asn</i>	4.66	3.85	3.82	2.98	6.50	3.74	2.06
Pro	4.61	4.63	5.18	4.20	8.25	6.26	12.07
Gln	3.69	3.36	2.58	4.80	3.53	4.28	4.46
Arg	4.61	4.56	5.27	13.89	3.58	8.94	4.21
Ser	6.31	6.23	7.96	7.78	10.79	11.71	6.91
Thr	6.15	5.76	6.05	5.61	6.48	5.55	5.14
Val	7.00	7.15	8.06	5.53	4.30	4.48	8.02
<i>Trp</i>	1.55	2.03	1.92	1.07	1.58	0.77	0.32
<i>Tyr</i>	3.64	3.53	3.51	1.76	2.59	0.83	1.42
X ^e		1.72					

^aAmino acid frequencies are in %.

^bData from Tompa (2002).

^cData from reference set of membrane proteins ($n = 91$).

^dData from human rhodopsin-like GPCRs ($n = 147$).

^eSide chains were not defined in the PDB entries. **Bold** are significantly enriched and *italic* depleted when IUPs are compared with folded proteins in the PDB. Dark and light shading, with boxes, are for increases $>3\%/> \times 3$ and $>1.5\%/> 2$, respectively, when compared with globular proteins. Dark and light shadings, without boxes, are for decreases for $<3\%/< \times 3$ and $<1.5\%/< 2$, respectively, when compared with globular proteins.

and very decreased percentages of Tyr, Phe, Gly, Asp and Val (Table III; Figure 6). The N-terminal domains are more like folded proteins than the 3i domain because they are not as highly charged, but they do have significantly raised levels of Ser and Pro, explaining why they appear to be only marginally folded. The C-terminal domains have very raised Ser and Arg, somewhat raised Pro and lowered Trp and Tyr. However,

it is intriguing that both the N- and C-terminal domains show raised Pro and Ser, whereas the 3i domains do not.

Discussion

Our analysis suggests that many human GPCRs contain structural flexibility not found in the membrane proteins whose structures have been solved, including bovine rhodopsin. Below we discuss the implications of this for GPCR cellular signaling, intrinsic disorder and structural studies.

Intrinsic disorder

Among the hundreds of IUPs and IURs found so far (Uversky and Fink, 2004), many act in important biological processes, including regulatory roles in transcription, translation, signal transduction and cell cycle control (Wright and Dyson, 1999; Dunker *et al.*, 2001; Tompa, 2002; Ward *et al.*, 2004). IUPs occur mainly in eukaryotic (33%) rather than eubacterial (4.2%) or archaean (2%) proteins (Ward *et al.*, 2004). *In vitro*, IUPs and IURs lack tertiary structure, do not have a tightly packed protein core and have a high degree of flexibility (Uversky, 2002b); lack of protein structure, just like protein structure, is encoded in the sequence. IURs identified so far have relatively little Trp, Tyr, Phe, Cys, Ile, Leu and Met but are significantly enriched in Pro, Glu, Lys, Ser and Gln (Romero *et al.*, 2001; Tompa, 2002). They also have many uncompensated charged groups, chiefly Glu and Lys, leading to a large net charge at neutral pH and low mean net hydrophobicity (Uversky *et al.*, 2000). For α_{2b} -AR, which we have studied extensively biochemically (Liitti *et al.*, 1997; Bartus *et al.*, 2003; Sen *et al.*, 2003; Jaakola *et al.*, 2005), it is clear that the long 3i is disordered (Figure 7). It is the site of extensive proteolysis and production of 3i truncations leads to more stable protein (V.-P.Jaakola *et al.*, unpublished work). In addition, the long 3i sequences are frequently so unusual that the disorder is clear even without sequence analysis. We show here, however, that this predicted disorder is GPCR-wide and has unusual sequence properties.

Regions of human rhodopsin-like GPCRs 3i domains have very high net mean charge (Table I), higher, even, than many other IURs. This is due to the very high percentage of Arg (13.9%), three times the proportion in IURs, membrane proteins or globular proteins (Table III; Figure 6). Intriguingly, Gln was increased, as in IURs, but not Glu, which was lower than in folded proteins. We could also see enhancement of Lys residues but, unlike IURs, not of Pro, Glu and Ser. Residues favoring folded structure, such as Ala, Trp, Ile and Leu, are not particularly infrequent in 3i domains and the proportion of Pro is the same as in folded proteins. GPCR 3i domains are thus predicted to form a unique class of IUR, which is supported by our results showing that the 3i of α_{2b} -adrenergic receptor is the site of extensive proteolysis (V.-P.Jaakola *et al.*, unpublished work). This is further confirmed by the properties of N- and C-terminal domains. The extracellular N-termini are marginally folded (Table I); they have low net charge, but large percentages of Ser and Pro. The marginally folded intracellular C-termini (Table I) have a net charge balance like that of the 3i domains but have large percentages of Ser and Pro.

Overall, the decrease in negative charge and increase in positive charge in the 3i and C-terminal domains presumably reflects the fact that they are inside the cell and close to the membrane. However, the positive to negative ratio in the

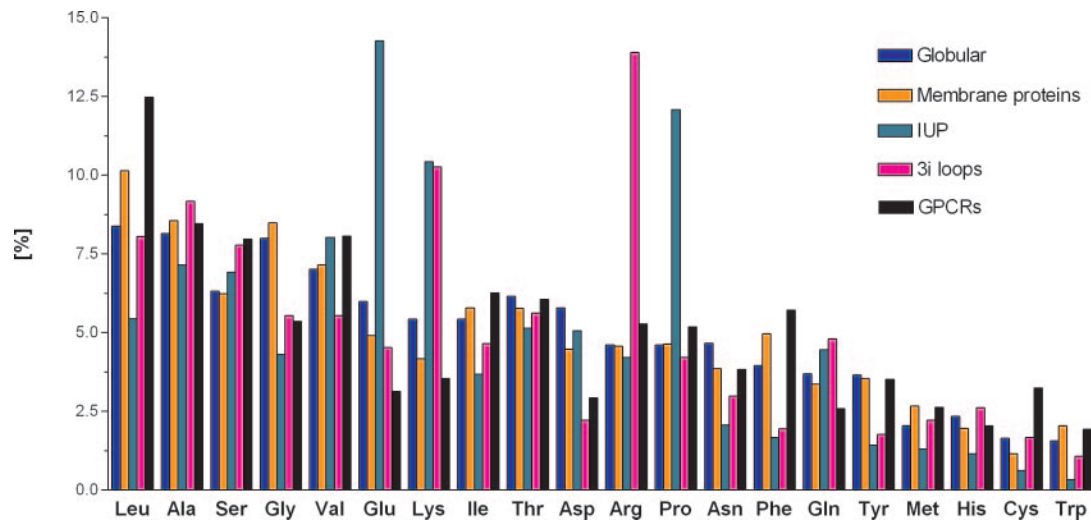


Fig. 6. Comparison of amino acid composition of globular proteins, IUPs, sample set II of human rhodopsin-like GPCRs, 3i domains of human rhodopsin-like GPCRs and reference set of membrane proteins of known structure.

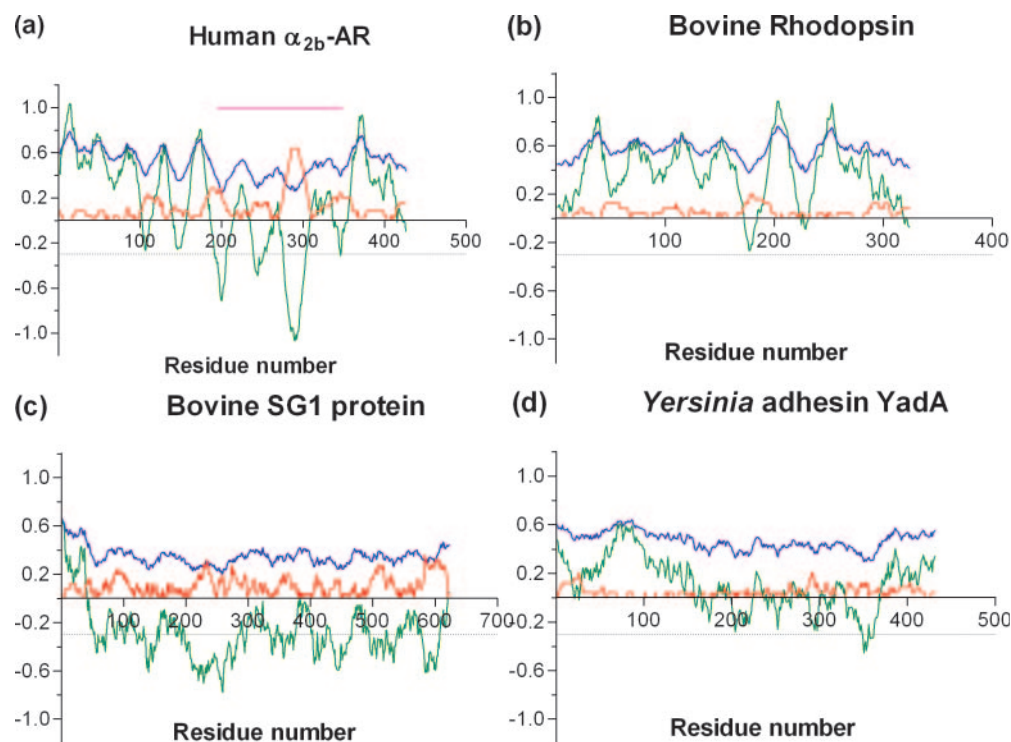


Fig. 7. Folding prediction plots. The figures were produced using FoldIndex (a) human α_{2b} -AR, (b) bovine opsin [1F88 (3)], (c) bovine SG1 protein (unfolded) and (d) *Yersinia* adhesin YadA [1P9H (47)]. A window of 25 residues was used. The FoldIndex value is marked as a green line, the mean net hydrophobicity (48) is marked as a blue line and the mean net charge is marked as a red line. The cut-off value of ≤ -0.3 is indicated as a dotted line. The α_{2b} -AR 3i is marked as a solid purple line. The seven α -helical motifs are seen from the folding plots. The analysis shows that the 3i loops of α_{2b} -AR and bovine rhodopsin are clearly predicted to be unfolded.

3i domains is very large, almost 4:1, whereas the C-terminal domain ratio is <2 . The intracellular 3i is predicted to be unfolded in the absence of interacting partners. Membrane proteins of solved structure do not show an increase in Arg and Lys, but this may be because these proteins are integral to the membrane; the 3i loop is on the surface of the membrane. Conversely, although charged residues such as Arg and Lys are frequently used for positioning the TM helices, as set-screws and stop-transfer signals, the very high percentage of Arg and Lys in 3i indicates that they must also play other roles (von Heijne and Gavel, 1988; von Heijne, 1989).

Cellular signaling of GPCRs

Each G protein-coupled receptor has several partners in the cellular signaling cascade, such as G proteins (Preinerger and Hamm, 2004), arrestins (Lefkowitz and Whalen, 2004) and GPCR kinases (Lefkowitz *et al.*, 2002). Furthermore, the receptor environment changes during various re-localization processes in the living cell. Interactions with specific co-proteins can cause structural changes in IUPs and IURs by binding to them. Moreover, the mean net charge and hydrophobicity of the complex will presumably be more similar to those seen in typical folded proteins. There are many examples

of such regulation in soluble proteins (Kriwacki *et al.*, 1996; Wright and Dyson, 1999). Many of the GPCR partners that have recently been identified interact via the C terminus (Brakeman *et al.*, 1997; Klein *et al.*, 1997; Hall *et al.*, 1998; Lezcano *et al.*, 2000) or 3i domains (Wu *et al.*, 1997; Prezeau *et al.*, 1999; Heuss and Gerber, 2000), which we predict to be the most unfolded regions in GPCRs (Table II).

The main GPCR partner is G_{α} , which has been shown by many biochemical/pharmacological studies to interact with the 2i, 3i and C-terminal regions of GPCRs [see reviews (Hamm, 2001; Preinerger and Hamm, 2004)]. The inactive structure of bovine rhodopsin reveals only a little about these interactions, as large structural movements occur upon receptor activation and signaling. Presumably some of these interactions involve the interaction of pre-formed structures, whereas others are based on linear sequence recognition. For instance, soluble rhodopsin receptor-mimetic peptide studies with NMR reveal significant ordering both at the receptor C-terminus and in the flexible C-terminal regions of G_{α} (Brabazon *et al.*, 2003). However, ordering of the 3i of rhodopsin receptor-mimetic peptide was not seen during the interactions of the G protein peptides (Brabazon *et al.*, 2003). These findings suggest that the interaction between the C-terminus of rhodopsin and G_{α} might be based on linear sequence recognition, consistent with the analysis presented here.

Both zebra fish and human α_2 -adrenergic receptors have extremely long 3i (>100 residues) and even the most divergent regions of the zebra fish receptor show clear molecular fingerprints for each subtype (Ruuskanen *et al.*, 2004). Each of these 3i domains is predicted to be unfolded and consequently the disorder must be relevant for function, as it has been preserved during 400 million years of evolution. In particular, despite the lack of negatively charged residues in 3i domains in general (Table III; Figure 6), human α_{2a} - and α_{2b} -adrenergic receptor have long negatively charged regions: the former contains ³⁰¹DLEES₄DHAE and the latter ²⁹⁴EDEAE₁₂CE and ²⁴⁵EKEEGETPED, some or all of which are conserved among most α_{2a} - and α_{2b} -adrenergic receptors, including zebra fish. α_{2c} -Adrenergic receptor, however, does *not* have an equivalent negatively charged region. Such regions presumably have functional significance and interact with other proteins, but they are not required for coupling to G proteins or to other currently identified interacting partners. Other interactions and interacting partners may therefore remain to be discovered, such as the recently identified interactions of β -arrestins (Wu *et al.*, 1997) and spinophilin (Richman *et al.*, 2001) with 3i domain of GPCRs. Similarly, 14–3–3 ζ protein, which can cause conformational change in target proteins (Bridges and Moorhead, 2004), binds to the 3i of α_2 -ARs (Prezeau *et al.*, 1999).

Implications for structural studies

The loops and coils connecting the more regular parts of protein secondary structure are often more disordered than the core structure itself. However, the loops found in the PDB are usually fairly short, <10 residues length on average (Espadaler *et al.*, 2004), although there are exceptions (Abdel-Meguid *et al.*, 1984). Crystallizability and lack of disordered structure are correlated.

One approach to dealing with this problem, adopted by most if not all of the structural genomics projects, is to eliminate as crystallization targets all proteins with long disordered

regions: the so-called ‘low-hanging fruit’ approach (Linding *et al.*, 2003). However, this would mean that structural studies of GPCRs are impossible; also, if IURs are common in eukaryota and used in signaling (Dunker *et al.*, 2002; Pe’er *et al.*, 2004), it will not be possible to study key signaling molecules. Consequently, picking the ‘low-hanging fruit’ may lead to the proverbial ‘drunks under the lamp-post’ problem. The keys are not there, but the light is bright.

Co-crystallization of GPCRs with cognate protein ligands may help. An obvious alternative (Linding *et al.*, 2003) is to modify or remove the IUR sequences, but this requires identifying such regions as we have done above and then testing the modified protein to ensure that the functionality is unchanged. For instance, the structure of the *Yersinia enterocolitica* YadA head group could be solved (Nummelin *et al.*, 2004), but only once the leucine-triple helix stalk had been removed (H.Nummelin, personal communication). This region (residues 225–380) appears to be disordered by FoldIndex (Figure 7).

Conclusions

This and other recent work suggest it is probably incorrect to speak of a single type of IUP or IUR. All IUPs and IURs have low hydrophobicity and high net charge, but the distribution of amino acids can differ (Vucetic *et al.*, 2003). For instance, Lu and Hansen (2004) recently showed that the linker histone C-terminal domain (CTD) appears to form a sequence-dependent structure in the presence of DNA, rather than interacting in a completely unstructured charge-dependent fashion. This implies that it, too, is an IUR, but the percentage of lysine in CTD is 36–41%, even higher than in IUPs. There may therefore be as many different kinds of IURs and IUPs as there are IUR and IUP functions.

Supplementary data

The proteins studied are listed in the Supplementary data, available at *PEDS* Online. Also, more detailed comparison of membrane proteins of known structure is given in the Supplementary data.

Acknowledgements

The study was funded by grants to Adrian Goldman from the Finnish National Technology Agency (grant 40272/01), the Academy of Finland (grant 78766) and the Sigrid Juselius Foundation and by a grant to Veli-Pekka Jaakola from the Magnus Ehrnrooth Foundation (2004). It was also funded by grants to Joel L.Sussman from the European Commission Structural Proteomics Project (SPINE) (QLG2-CT-2002-00988), the Ministry of Science and Technology’s grant to the Israel Structural Proteomics Center and the Minerva Foundation. J.L.S. is the Morton and Gladys Pickman Professor of Structural Biology

References

- Abdel-Meguid,S.S., Grindley,N.D., Templeton,N.S. and Steitz,T.A. (1984) *Proc. Natl Acad. Sci. USA*, **81**, 2001–2005.
- Baldwin,J.M., Schertler,G.F. and Unger,V.M. (1997) *J. Mol. Biol.*, **272**, 144–164.
- Bartus,C.L., Jaakola,V.P., Reusch,R., Valentine,H.H., Heikinheimo,P., Levay,A., Potter,L.T., Heimo,H., Goldman,A. and Turner,G.J. (2003) *Biochim. Biophys. Acta*, **1610**, 109–123.
- Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.
- Brabazon,D.M., Abdulaev,N.G., Marino,J.P. and Ridge,K.D. (2003) *Biochemistry*, **42**, 302–311.
- Brakeman,P.R., Lanahan,A.A., O’Brien,R., Roche,K., Barnes,C.A., Haganir,R.L. and Worley,P.F. (1997) *Nature*, **386**, 284–288.
- Bridges,D. and Moorhead,G.B. (2004) *Sci. STKE*, re10.
- Dafforn,T.R. and Rodger,A. (2004) *Curr. Opin. Struct. Biol.* **14**, 541–546.
- Dunker,A.K. *et al.* (2001) *J. Mol. Graph. Model.*, **19**, 26–59.

- Dunker,A.K., Brown,C.J. and Obradovic,Z. (2002) *Adv. Protein Chem.*, **62**, 25–49.
- Espadaler,J., Fernandez-Fuentes,N., Hermoso,A., Querol,E., Aviles,F.X., Sternberg,M.J. and Oliva,B. (2004) *Nucleic Acids Res.* **32**, Database issue, D185–D188.
- Gether,U. and Kobilka,B.K. (1998) *J. Biol. Chem.*, **273**, 17979–17982.
- Hall,R.A. et al. (1998) *Nature*, **392**, 626–630.
- Hamm,H.E. (2001) *Proc. Natl Acad. Sci. USA*, **98**, 4819–4821.
- Heuss,C. and Gerber,U. (2000) *Trends Neurosci.*, **23**, 469–475.
- Horn,F., Vriend,G. and Cohen,F.E. (2001) *Nucleic Acids Res.*, **29**, 346–349.
- Iakoucheva,L.M., Kimzey,A.L., Masselon,C.D., Bruce,J.E., Garner,E.C., Brown,C.J., Dunker,A.K., Smith,R.D. and Ackerman,E.J. (2001) *Protein Sci.*, **10**, 560–571.
- Jaakola,V.-P., Rehn,M., Moeller,M., Alexiev,U., Goldman,A. and Turner,G.J. (2005) *Proteins*, in press.
- Johnson,D.G. and Walker,C.L. (1999) *Annu. Rev. Pharmacol. Toxicol.*, **39**, 295–312.
- Klabunde,T. and Hessler,G. (2002) *Chembiochem*, **3**, 928–944.
- Klein,U., Ramirez,M.T., Kobilka,B.K. and von Zastrow,M. (1997) *J. Biol. Chem.*, **272**, 19099–19102.
- Kriwacki,R.W., Hengst,L., Tennant,L., Reed,S.I. and Wright,P.E. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 11504–11509.
- Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.
- Lefkowitz,R.J. and Whalen,E.J. (2004) *Curr. Opin. Cell Biol.*, **16**, 162–168.
- Lefkowitz,R.J., Pierce,K.L. and Luttrell,L.M. (2002) *Mol. Pharmacol.*, **62**, 971–974.
- Lezcano,N., Mrzljak,L., Eubanks,S., Levenson,R., Goldman-Rakic,P. and Bergson,C. (2000) *Science* **287**, 1660–4.
- Liitti,S., Narva,H., Marjamäki,A., Hellman,J., Kallio,J., Jalkanen,M. and Matikainen,M.T. (1997) *Biochem. Biophys. Res. Commun.*, **233**, 166–172.
- Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. and Russell,R.B. (2003) *Structure (Camb.)*, **11**, 1453–1459.
- Nummelin,H., Merckel,M.C., Leo,J.C., Lankinen,H., Skurnik,M. and Goldman,A. (2004) *EMBO J.*, **23**, 701–711.
- Okada,T., Fujiyoshi,Y., Silow,M., Navarro,J., Landau,E.M. and Shichida,Y. (2002) *Proc. Natl Acad. Sci. USA*, **99**, 5982–5987.
- Palczewski,K. et al. (2000) *Science*, **289**, 739–745.
- Pavletich,N.P. (1999) *J. Mol. Biol.*, **287**, 821–828.
- Pe'er,I., Felder,C.E., Man,O., Silman,I., Sussman,J.L. and Beckmann,J.S. (2004) *Proteins*, **54**, 20–40.
- Preininger,A.M. and Hamm,H.E. (2004) *Sci. STKE*, re3.
- Prezeau,L., Richman,J.G., Edwards,S.W. and Limbird,L.E. (1999) *J. Biol. Chem.*, **274**, 13462–13469.
- Richman,J.G., Brady,A.E., Wang,Q., Hensel,J.L., Colbran,R.J. and Limbird,L.E. (2001) *J. Biol. Chem.*, **276**, 15003–15008.
- Romero,P., Obradovic,Z., Li,X., Garner,E.C., Brown,C.J. and Dunker,A.K. (2001) *Proteins*, **42**, 38–48.
- Ruuskanen,J.O., Xhaard,H., Marjamäki,A., Salaneck,E., Salminen,T., Yan,Y.L., Postlethwait,J.H., Johnson,M.S., Larhammar,D. and Scheinin,M. (2004) *Mol. Biol. Evol.*, **21**, 14–28.
- Sen,S., Jaakola,V.P., Heimo,H., Engstrom,M., Larjomaa,P., Scheinin,M., Lundstrom,K. and Goldman,A. (2003) *Protein. Expr. Purif.*, **32**, 265–75.
- Spiegel,A.M. and Weinstein,L.S. (2004) *Annu. Rev. Med.*, **55**, 27–39.
- Tompa,P. (2002) *Trends Biochem. Sci.*, **27**, 527–533.
- Uversky,V.N. (2002a) *Protein Sci.*, **11**, 739–756.
- Uversky,V.N. (2002b) *Eur. J. Biochem.*, **269**, 2–12.
- Uversky,V.N. and Fink,A.L. (2004) *Biochim. Biophys. Acta*, **1698**, 131–153.
- Uversky,V.N., Gillespie,J.R. and Fink,A.L. (2000) *Proteins*, **41**, 415–427.
- Ward,J.J., Sodhi,J.S., McGuffin,L.J., Buxton,B.F. and Jones,D.T. (2004) *J. Mol. Biol.*, **337**, 635–645.
- von Heijne,G. (1989) *Nature*, **341**, 456–458.
- von Heijne,G. and Gavel,Y. (1988) *Eur. J. Biochem.*, **174**, 671–678.
- Wright,P.E. and Dyson,H.J. (1999) *J. Mol. Biol.*, **293**, 321–331.
- Wu,G., Krupnick,J.G., Benovic,J.L. and Lanier,S.M. (1997) *J. Biol. Chem.*, **272**, 17836–17842.
- Vucetic,S., Brown,C.J., Dunker,A.K. and Obradovic,Z. (2003) *Proteins*, **52**, 573–584.
- Zeev-Ben-Mordehai,T., Rydberg,E.H., Solomon,A., Toker,L., Auld,V.J., Silman,I., Botti,S. and Sussman,J.L. (2003) *Proteins: Struct. Funct. Bioinf.*, **53**, 758–767.

Received January 18, 2005; accepted January 28, 2005

Edited by Mirek Cygler