Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes†

Amrita Mohan,[a] William J. Sullivan Jr,[b] Predrag Radivojac,[a] A. Keith Dunker[c] and Vladimir N. Uversky*[cd]

Received 4th January 2008, Accepted 30th January 2008
First published as an Advance Article on the web 21st February 2008
DOI: 10.1039/b719168e

Parasitic protozoal infections have long been known to cause profound degrees of sickness and death in humans as well as animal populations. Despite the increase in the number of annotated genomes available for a large variety of protozoa, a great deal more has yet to be learned about them, from their fundamental physiology to mechanisms invoked during host–pathogen interactions. Most of these genomes share a common feature, namely a high prevalence of low complexity regions in their predicted proteins, which is believed to contribute to the uniqueness of the individual species within this diverse group of early-branching eukaryotes. In the case of *Plasmodium* species, which cause malaria, such regions have also been reported to hamper the identification of homologues, thus making functional genomics exceptionally challenging. One of the better accepted theories accounting for the high number of low complexity regions is the presence of intrinsic disorder in these microbes. In this study we compare the degree of disordered proteins that are predicted to be expressed in many such ancient eukaryotic cells. Our findings indicate an unusual bias in the amino acids comprising protozoal proteomes, and show that intrinsic disorder is remarkably abundant among their predicted proteins. Additionally, the intrinsically disordered regions tend to be considerably longer in the early-branching eukaryotes. An analysis of a *Plasmodium falciparum* interactome indicates that protein–protein interactions may be at least one function of the intrinsic disorder. This study provides a bioinfomatics basis for the discovery and analysis of unfoldomes (the complement of intrinsically disordered proteins in a given proteome) of early-branching eukaryotes. It also provides new insights into the evolution of intrinsic disorder in the context of adapting to a parasitic lifestyle and lays the foundation for further work on the subject.

Introduction

The study of early-branching eukaryotic cells yields great insights into the evolutionary landscape of cell developmental biology. Some of the oldest eukaryotic species are single-celled protozoa, a diverse array of organisms that live freely or have evolved into parasitic entities. Investigation of the parasitic varieties not only offers the benefit of studying ''living fossils'', but also may have rewards of therapeutic relevance. Among the myriad of parasitic protozoa are notorious pathogens that exert significant morbidity and mortality towards humans and livestock. Consequently, parasitic protozoal infections also have profound economic and socioeconomic ramifications.

Parasitic protozoa that have a substantial impact on humans include anaerobic organisms such as *Giardia lamblia* and *Entamoeba histolytica*. *G. lamblia*, a diplomonad, is one of the most common intestinal protozoal infections that cause diarrhoea.[1] *E. histolytica*, causing colitis and liver abscesses, is the second leading cause of death from parasitic diseases in the world, killing up to 100 000 people a year.[2]

The kinetoplastids include *Trypanosoma brucei* ssp. and *T. cruzi*, the causative agents of African sleeping sickness and Chagas' disease, respectively. An estimated 18 million persons are infected with *T. cruzi* in Latin America[3] and 300 000–500 000 cases of African sleeping sickness occur per year.[4] There are not many treatment regimes available for the *Trypanosoma* species, and some of these regimes are highly toxic.

Parasitic protozoa within the phylum Apicomplexa include *Plasmodium* spp., the causative agent of malaria. Malaria is one of the most catastrophic infectious diseases of our time, infecting ~500 million people (in 2002) and resulting in 1–3 million deaths, many of these occurring in children.[5,6] *Cryptosporidium* spp. is an intestinal infection that causes

[a] *School of Informatics, Indiana University, Bloomington, IN 47401, USA. E-mail: ammohan@indiana.edu. E-mail: predrag@indiana.edu*
[b] *Department of Pharmacology & Toxicology, Indiana University Center For AIDS Research, Indiana University School of Medicine, Indianapolis, IN 46202, USA. E-mail: wjsulliv@iupui.edu; Tel: +1-317 274-1573*
[c] *Institute of Intrinsically Disordered Proteins, Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202, USA. E-mail: kedunker@iupui.edu. E-mail: vuversky@iupu.edu; Fax: +1-317-278-9217; Tel: +1-317-278-6448*
[d] *Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, Moscow Region, 142290, Russia*
† Electronic supplementary information (ESI) available: Supplementary figures. See DOI: 10.1039/b719168e

substantial water-borne outbreaks that can result in serious strains on agricultural and medical resources.[7] *Toxoplasma gondii* causes congenital birth defects and has been linked to neurological disorders and behavior anomalies.[8–10] *Cryptosporidium* and *Toxoplasma* are also serious opportunistic infections of AIDS and other immunocompromised individuals,[11] and are classified by NIAID as category B pathogens that are relevant to biodefense research.[12]

Paradoxically, the devastating effects of protozoal pathogens have not translated into a wealth of research being done in the area. A great deal more has yet to be learned about these important organisms, from fundamental physiology to their host–pathogen interactions. With genome sequences now available for a wide variety of these protozoa, we now have the opportunity to explore phenomena that may simultaneously reveal valuable insights into basic cell biology and may ultimately translate into novel and effective therapies in these species. However, there are several complications that hamper functional genomic studies in protozoan parasites.[13] For example, most protozoal genomes consist of a high percentage of genes that lack apparent orthologues in other organisms.[14] Additionally, many of the genes in *Plasmodium* species have been found to encode relatively large proteins that contain a large number of low complexity regions.[15–17] One study has reported that >90% of all proteins in chromosomes 2 and 3 of *P. falciparum* contain such low complexity regions.[18] The uniqueness of genomes and the high abundance of the low complexity regions cause difficulties in identifying homologues of *Plasmodium* proteins.[18] Many *Plasmodium* proteins have been shown to be difficult to express in heterologous systems.[19] One of the reasons for low expression yields is the presence of intrinsically disordered regions.[18]

"Intrinsic disorder" refers to the lack of a relatively fixed structure in proteins. Some terms used to describe proteins exhibiting such a behavior include "rheomorphic",[20] "natively unfolded",[21,22] "intrinsically unstructured",[23,24] "intrinsically disordered",[25] and "natively disordered".[26] Many biologically active proteins remain unstructured, or incompletely structured, under physiological conditions (reviewed in[22–38]). Intrinsic disorder has been reported both at a regional as well as protein level and is characterized by dynamic ensembles of structures instead of one single structure. Some well-known functions of disordered proteins include molecular recognition and assembly (as encountered in signaling pathways), protein modification (*e.g.*, phosphorylation, acetylation, methylation, *etc.*), and entropic chain activities (*e.g.*, linkers, springs, and spacers).[32] The molecular recognition and assembly functions of disordered proteins are typically a direct consequence of disorder-to-order transitions that can readily occur in flexible, disordered regions of proteins. In contrast, the various entropic chain activities depend directly on the unstructured state without the need to form structure. The functional diversity provided by disordered regions has been conjectured to largely complement the functions of ordered proteins.

This complementarity hypothesis has been recently supported by a comprehensive analysis of the correlation of nearly all keywords in the SwissProt database with the presence or absence of predicted intrinsic disorder.[39–41] Out of the 710 Swiss-Prot functional keywords that were associated with at least 20 proteins, 238 were found to be strongly positively correlated with predictions of long intrinsically disordered regions, whereas 302 were strongly negatively correlated with such regions. Literature searches showed that the negatively correlated proteins are typically structured, and the top-ranking keywords for these proteins all end in "ase", indicative of keywords that describe enzymes, whose functions depend on active sites resulting from their ability to fold into globular 3D structures. The disorder-associated function list is rich in keywords describing signaling, regulation, and control. These functions often rely on highly-specific yet low affinity interactions of proteins with multiple partners and also on reversible posttranslational modification, both of which are features of disordered proteins.[39–41]

There are several computational studies evaluating the abundance of intrinsic disorder in proteins of *P. falciparum*. For example, 35% of proteins encoded by genes on chromosomes 2 and 3 are predicted to contain long (>40 consecutive residues) disordered regions.[42] A later study revealed that this number was an underestimate, suggesting that 52–67% of *P. falciparum* proteins harbor long disordered regions.[43] A more recent study evaluated the abundance of disordered proteins in several apicomplexan parasites, revealing that the proteomes of human malaria parasites (*P. falciparum* and *P. vivax*) and primate malaria parasites (*P. knowlesi*) are more enriched in intrinsically disordered proteins than those of rodent malaria parasites (*P. yoelii*, *P. chabaudi* and *P. berghei*).[18] Furthermore, proteins expressed in the sporozoites of *P. falciparum* were shown to be more intrinsically disordered compared to those expressed during other life cycle stages.[18]

In this study, we take advantage of the abundance of sequenced genomes for numerous eukaryotic microbes, pathogenic and nonpathogenic, to compare the degree of disordered proteins predicted to be expressed in these ancient eukaryotic cells. Our work is different from previous bioinformatics studies on early-branching eukaryotes as a different set of disorder predictors was employed. Our findings lend further support to the idea that abundant intrinsically disordered proteins in these organisms are likely to play significant, diverse roles in parasite biology and pathogenesis.

Results

Amino acid compositional profiles

Amino acid compositions of fifteen early-branching eukaryotic organisms (the nine Apicomplexa *P. falciparum*, *P. berghei*, *P. chabaudi*, *P. vivax*, *P. yoelii*, *T. gondii*, *T. parva*, *C. hominis*, and *C. parvum*, two yeasts, *C. albicans* and *C. glabrata*, the kinetoplastid *T. brucei*, the Amoebozoa *D. discoideum* and *E. histolytica*, and the Metamonada *G. lamblia*) were compared with compositions of proteins from a representative disordered (Fig. 1A) and ordered (Fig. 1B) dataset, the free-living non-pathogenic protozoan *T. thermophila* (Fig. S1, see ESI†), the model multicellular eukaryote *C. elegans* (Fig. S2, see ESI†) and the prokaryote *V. cholerae* (Fig. S3, see ESI†). These figures depict fractional relative compositions, with the amino acids arranged from left to right in increasing order of

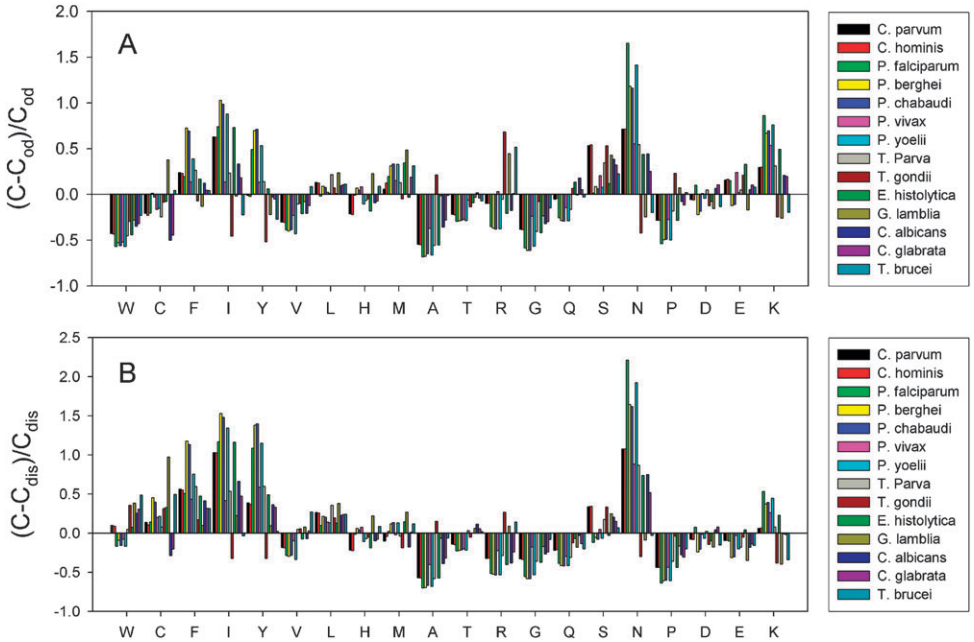

Fig. 1 Compositional profiling of early-branching eukaryotes in comparison with a set of ordered (A) and experimentally characterized disordered proteins (B). The bar for a given amino acid represents the fractional difference in composition between a given protein (or set of proteins) and a set of ordered proteins. The fractional difference is calculated as $(C_X - C_{\text{reference}})/C_{\text{reference}}$, where C_X is the content of a given amino acid in a given protein (or protein set), and $C_{\text{reference}}$ is the corresponding content in a set of reference proteins and plotted for each amino acid. The amino acid residues are arranged from the most rigid to the most flexible according to the Vihinen's flexibility scale, which is based on the averaged B-factor values for the backbone atoms of each residue type as estimated from 92 proteins.[44] Negative values indicate residues that the given protein/set has less than the reference set, positive values correspond to residues that are more abundant in a given dataset in comparison with the reference set.

surface accessibility in globular proteins (Vihinen flexibility scale).

One can observe various general patterns in each of these figures. For instance, parasitic protozoan datasets are significantly depleted in tryptophan (W) and enriched in lysine (K), in comparison to ordered sequences (Fig. 1A). However, in comparison to the disordered dataset, most of the protein sets are depleted in tryptophan (W) (Fig. 1B). It is interesting to note that in comparison to ordered as well as disordered sequences of the majority of these thirteen datasets are enriched in phenylalanine (F), isoleucine (I) and tyrosine (Y). Comparison of Fig. 1A and B suggests that early-branching eukaryotes represent a unique group whose proteomes are compositionally different when compared to typical ordered and intrinsically disordered proteins.

The high abundance of phenylalanine and tyrosine residues might be related to peculiarities of protein folding and/or functionality. Several intrinsically disordered proteins have been shown to be highly enriched in these residues. For example, multiple tyrosine residues were shown to be essential for the function of the Ewings sarcoma (EWS) fusion proteins (EFPs). EFPs are potent transcriptional activators interacting with other proteins required for mRNA biogenesis. A characteristic functionality of EFPs is associated with the EWS activation domain (EAD), containing multiple degenerate hexapeptide repeats (consensus SYGQQS) with a conserved tyrosine residue. This intrinsically disordered domain was shown to be responsible for transcriptional activation and cellular transformation.[45] Furthermore, these multiply conserved tyrosines were shown to be essential for the EAD function. Intriguingly, they can be effectively substituted by phenylalanine, showing that an aromatic ring can confer EAD function in the absence of tyrosine phosphorylation.[45] Other examples include a set of phenylalanine–glycine repeat-containing nucleoporins (FG-Nups), specific proteins from nuclear pore complexes (NPCs) that are embedded in the nuclear envelope of eukaryotic cells. There are 13 such proteins in the *Saccharomyces cerevisiae* NPC. They bind karyopherins and facilitate the transport of karyopherin–cargo complexes. All these proteins were shown to be intrinsically disordered and the FG repeat regions of Nups were shown to form a meshwork of random coils at the NPC through which nuclear transport proceeds.[46,47] Another example is immunoreceptor tyrosine-based activation motif (ITAM)-containing cytoplasmic domains of many immune receptors, which have been recently shown to represent a novel class of intrinsically disordered proteins.[48,49]

In comparison to their free-living, non-pathogenic counterpart (*i.e.*, *Tetrahymena thermophila*), pathogenic early-branching eukaryotes are observed to be significantly enriched in aspartic acid (D), proline (P) and valine (V), and polar residues such as tryptophan (W) and histidine (H). However, depletion of the polar residue glutamine (Q) appears to be common across all species in comparison to *T. thermophila* (Fig. S1, see ESI†). Compared to *C. elegans* or *V. cholerae*, depletion of tryptophan (W) and valine (V), both order-promoting residues, is apparent in the microbes analyzed (Fig. S2 and S3†). Although many other amino acids are also depleted in various

proteomes, W and V are the only two residues with consistent behavior across all species in comparison to *C. elgans* and *V. cholerae*. These figures also evidence for a pronounced lysine (K) content amongst most parasites.

Prevalence of long intrinsically disordered regions

To further evaluate the presence of intrinsic disorder in different proteomes we used a predictor of natural disordered regions (PONDR® VSL2B).[50] The abundance of intrinsic disorder in various organisms has been evaluated *via* the presentation of the fractions of proteins with predicted disordered regions of a given length.[31,51] Earlier, this approach was used to show the prevalence of intrinsic disorder in signaling and cancer-associated proteins[31] and in proteins involved in the cardiovascular disease.[51] Fig. 2A and S4A† show that intrinsic disorder is predicted to be relatively abundant in early-branching eukaryotes. The percentages of proteins with 30 or more consecutive residues predicted to be disordered by PONDR® VSL2 and PONDR® VLXT (corresponding numbers are shown in brackets) were 87.8 (89.8) for *T. gondii*, 80.3 (82.5) for *P. vivax*, 79.0 (81.0) for *P. falciparum*, 75.3 (76.8) for *D. discoideum*, 73.8 (75.1) for *C. parvum*, 72.4 (74.1) for *C. albicans*, 71.9 (73.1) for *C. glabrata*, 71.4 (72.4) for *T. thermophila*, 70.4 (72.0) for *T. brucei*, 69.7 (70.9) for *C. hominis*, 67.5 (68.9) for *T. parva*, 63.0 (64.5) for *P. yoelii*, 62.6 (64.1) for *S. cerevisiae*, 63.0 (64.3) for *C. elegans*, 58.2 (59.5) for *E. histolytica*, 52.1 (53.2) for *G. lamblia*, 42.5 (43.4) for *P. berghei*, 40.3 (41.3) for *P. chabaudi* and 24.9 (25.1) for *V. cholerae*. Earlier, using PONDR® VLXT it has been shown that a set of eukaryotic proteins from Swiss-Prot, EU_SW, and a set of ordered proteins from PDB, O_PDB_S25, contained 47(±4)% and 13(±4)% proteins with 30 or more consecutive residues predicted to be disordered.[31,51] Therefore, being compared to a set of ordered proteins, all the microbial proteomes were significantly enriched in proteins with long disordered regions. Furthermore, the vast majority of the early-branching eukaryotic organisms (except for *G. lamblia*, *P. berghei*, and *P. chabaudi*) had more proteins with long disordered regions than a set of typical eukaryotic proteins in Swiss-Prot.

When analyzed by percentages of residues, the proteins with 30 or more consecutive residues predicted to be disordered by PONDR® VSL2 (PONDR® VLXT) were distributed as follows (see Fig. 2B and S4B†): 58.3 (36.1) for *T. gondii*, 42.7 (7.3) for *P. falciparum*, 41.5 (21.2) for *P. vivax*, 37.9 (6.5) for *T. thermophila*, 37.1 (6.9) for *P. yoelii*, 34.4 (16.4) for *D. discoideum*, 29.6 (6.1) for *P. chabaudi*, 29.5 (13.7) for *C. albicans*, 29.2 (4.2) for *P. berghei*, 28.5 (14.5) for *C. glabrata*, 27.7 (15.6) for *C. elegans*, 27.6 (13.5) for *S. cerevisiae*, 27.5 (17.8) for *T. brucei*, 27.2 (10.4) for *C. hominis*, 26.2 (9.4) for *C. parvum*, 24.6 (12.5) for *G. lamblia*, 23.5 (9.4) for *T. parva*, 19.6 (6.6) for *E. histolytica*, and 6.2 (6.7) for *V. cholerae*. According to earlier PONDR® VLXT estimations, there were (6.5 ± 0.5)% and (1.48 ± 0.45)% residues in long disordered regions of eukaryotic proteins from Swiss-Prot and of non-homologous ordered proteins from PDB, respectively.[31,51]

The data suggest that sequences from early-branching eukaryotes contained more disordered residues than eukaryotic Swiss-Prot proteins and ordered PDB proteins. Finally, we have established that many protozoan sequences (between 20 and 60% depending on the organism) contained very long

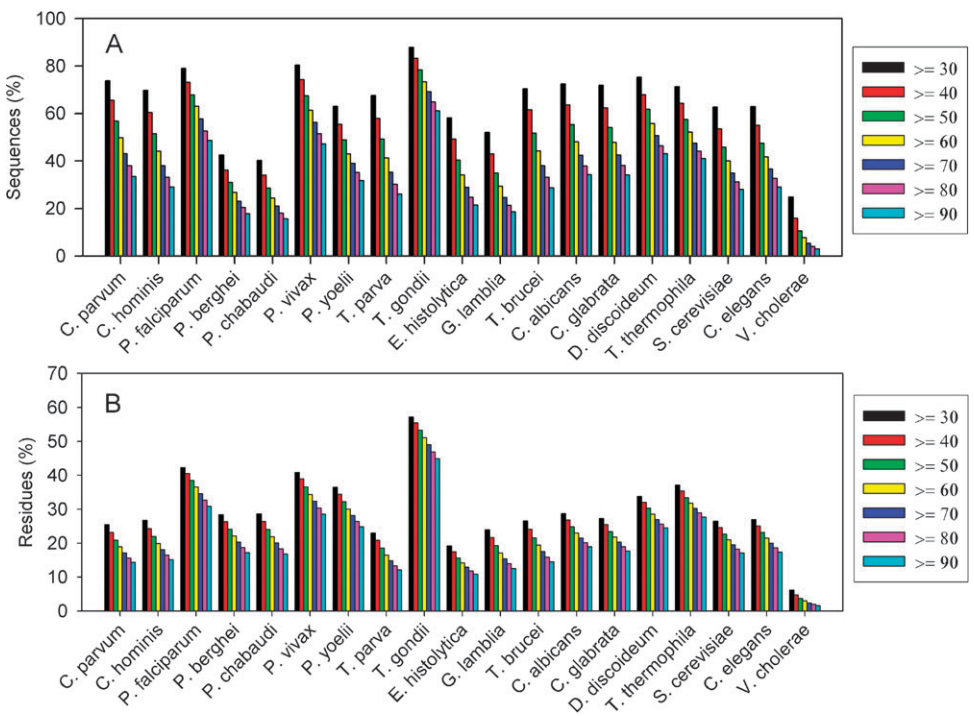

Fig. 2 PONDR® VSL2 disorder prediction results on 19 proteomes: *C. parvum*, *C. hominis*, *P. falciparum*, *P. berghei*, *P. chabaudi*, *P. vivax*, *P. yoelii*, *T. parva*, *T. gondii*, *E. histolytica*, *G. lamblia*, *T. brucei*, *C. albicans*, *C. glabrata*, *D. discoideum*, *T. thermophila*, *S. cerevisiae*, *C. elegans*, and *V. cholerae*. (A) Percentages of proteins in the 19 proteomes with ≥ 30 to ≥ 90 consecutive residues predicted to be disordered. (B) Percentages of residues in these 19 proteomes predicted to be disordered within segments of length $\geq$ the value on the *x*-axis.

Table 1 CH and CDF prediction results for all 19 organisms with number of predicted α-MoRFs

Organism	Total proteins used for predictions	Number of proteins with one or more α-MoRFs	Number of α-MoRFs	Number of proteins predicted disordered by both CH and CDF	Number of proteins predicted disordered only by CH	Number of proteins predicted disordered only by CDF	Number of proteins predicted ordered by both CH and CDF
C. parvum	3801	810	1375	367	238	357	2839
C. hominis	3884	720	1165	438	258	365	2823
P. falciparum	5400	1348	2757	509	1854	69	2968
P. berghei	10 459	787	1204	745	2262	188	7264
P. chabaudi	12 968	1133	1614	1235	2748	245	8740
P. vivax	5330	1878	4832	1287	440	897	2706
P. yoelii	7238	1155	2083	660	1729	159	4690
T. parva	4070	595	866	463	232	305	3070
T. gondii	7793	3761	11 889	1300	73	3969	2451
E. histolytica	9766	991	1389	1113	705	487	7461
G. lamblia	9646	1242	1993	792	239	2025	6590
T. brucei	8758	2016	3530	944	146	2221	5447
C. albicans	6068	1472	2738	949	255	809	4055
C. glabrata	5271	1292	2390	763	210	849	3449
D. discoideum	4031	1153	2305	666	189	552	2624
T. thermophila	26 212	5873	12 121	2562	5703	726	17 221
S. cerevisiae	10 868	2302	4247	1448	433	1552	7435
C. elegans	38 336	8379	15 002	4339	751	6890	26 356
V. cholerae	3829	72	76	92	60	327	3350

disordered regions (with lengths ≥ 90 consecutive residues). Thus, early-branching eukaryotic proteins are heavily enriched in predicted disorder in comparison with typical eukaryotic proteins from Swiss-Prot and especially with ordered proteins from PDB.

CDF and CH-plot analyses

The sequences of protozoan proteins were also used to predict whether these proteins are likely to be mostly disordered using two binary predictors of intrinsic disorder: the charge-hydropathy plot (CH-plot)[22,52] and the cumulative distribution function analysis (CDF analysis).[52] Both these methods perform binary classification of whole proteins as either mostly disordered or mostly ordered, where mostly ordered indicates proteins that contain more ordered residues than disordered residues and mostly disordered indicates proteins that contain more disordered residues than ordered residues.[52] A simultaneous observation of low mean hydropathy and relatively high net charge is typical for the "natively unfolded" proteins, which are characterized by the lack of compact, collapsed structure.[22] Therefore, ordered and disordered proteins plotted in CH-space can be separated to a significant degree by a linear boundary, with proteins located above this boundary line being natively unfolded and with proteins below the boundary line being ordered.[22,52] CDF analysis summarizes the per-residue disorder predictions by plotting PONDR scores against their cumulative frequency, which allows ordered and disordered proteins to be distinguished based on the distribution of prediction scores. In this case, order–disorder classification is based on whether a CDF curve is above or below a majority of boundary points: proteins with high PONDR scores will have CDF curves that have low cumulative values over most of the CDF curve, and proteins with low PONDR scores will have CDF curves that have high cumulative values over most of the CDF curve.[52] The results of CH-plot and CDF analyses are shown in Fig. S5, S6 and S7.†

Data presented in Fig. S5A, S6† and Table 1 show that early-branching eukaryotes appear to contain many compact proteins. In fact, for the majority of proteomes studied the average content of proteins predicted to be natively unfolded is $\sim$4.5%. However, *P. falciparum*, *P. berghei*, *P. chabaudi*, *P. yoelii* and *T. thermophila* are predicted to contain a very significant number of proteins that are potentially natively unfolded (34%, 22%, 21%, 24% and 22%, respectively). The curves presented in Fig. S5B and S7† clearly indicate that most pathogenic protozoa as well as *C. albicans* have high cumulative values over most of the CDF curve and hence are largely ordered. Fig. S5† also shows that *V. cholerae* contains the most ordered proteome amongst all organisms studied. This is in accordance with earlier reports suggesting that eukaryotes have higher disorder content in comparison to prokaryotes.[42,52] *C. elegans* and *S. cerevisiae* appear to have an apparent balance between the number of ordered and disordered proteins. The results of CDF and CH-plot analyses are further summarized in Table 1, which shows that most of the pathogenic proteomes are predicted to be ordered by both these methods. *T. gondii* however, appears to serve as the single exception to this rule with only 31% of proteins unanimously predicted as ordered. *V. cholerae* appears to be the most ordered of all organisms with as much as 88% agreement between CH-plot and CDF analyses.

Table 1 also shows that there is a sizeable discrepancy between these two methods and the level of disorder predicted by CDF is on average 1.25-fold higher than that predicted by CH-plots. The difference between these two methods in the magnitude of predicted disorder is generally similar to previously published data.[51,52] This difference was explained by the fact that the CH-plot is a linear classifier that takes into account only two parameters of the particular sequence—charge and hydropathy,[22] whereas CDF analysis is dependent upon the output of the PONDR® VLXT predictor, a nonlinear neural network classifier, which was trained to distinguish order and disorder based on a significantly larger feature

space that explicitly includes net charge and hydropathy.[53,54] According to these methodological differences, CH-plot analysis is predisposed to discriminate proteins with substantial amounts of extended disorder (random coils and pre-molten globules) from proteins with globular conformations (molten globule-like and rigid well-structured proteins). On the other hand, PONDR-based CDF analysis may discriminate all disordered conformations including molten globules from rigid well-folded proteins. Therefore, this discrepancy in the disorder prediction by CDF and CH-plot might provide a computational tool to discriminate "natively unfolded" proteins from native molten globules, which might predicted to be disordered by CDF, but compact by CH-plot. This model is consistent with the behavior of several intrinsically disordered proteins. Work is currently in progress to analyze the generality of this hypothesis. In application to protozoan proteins this means that some of them are predicted to be extended, whereas others can possess molten globule-like properties.

Fig. 3 compares the results of the CH-plot and CDF analyses by showing the distributions of proteins in each proteome within the CH–CDF phase space. In these plots, each spot corresponds to a single protein and its coordinates are calculated as a distance of this protein from the boundary in the corresponding CH-plot (*y*-coordinate) and an averaged distance of the corresponding CDF curve from the boundary (*x*-coordinate). Positive and negative *y* values correspond to proteins which, according to CH-plot analysis, are predicted to be natively unfolded or compact, respectively. Whereas positive and negative *x* values are attributed to proteins that, by the CDF analysis, are predicted to be ordered or intrinsically disordered, respectively. Therefore, each plot contains four quadrants (see an explanatory panel in the low right corner of Fig. 3): (−, −) contains proteins predicted to be disordered by CDF, but compact by CH-plot (*i.e.*, proteins possibly with molten globule-like properties); (−, +) includes proteins predicted to be disordered by both methods (*i.e.*, random coils and pre-molten globules); (+, −) contains ordered proteins; (+, +) includes proteins predicted to be disordered by CH-plot, but ordered by the CDF analysis. The sharp cut-off at the right side of each plot is due to the upper limit of a difference between the CDF curve (which has a maximum value of 1.0) and the boundary separating IDPs and ordered proteins in CDF plots. Analysis of the (−, −) and (−, +) quadrants in Fig. 3 shows that the majority of the wholly disordered proteins from *C. elegans*, *S. cerevisiae*, and *V. cholerae* likely possess molten globule-like properties. In contrast, protozoan proteomes are generally characterized by a more balanced distribution between compact and extended disordered proteins. This balance is also observed in the case of *C. albicans* and *C. glabrata* proteomes demonstrating some prevalence for the extended disordered proteins.

α-MoRF predictions

Often, intrinsically disordered regions in proteins are involved in protein–protein interactions and molecular recognitions.[24,25,33–36] It has been pointed out that many flexible proteins or regions undergo disorder-to-order transitions upon binding, which is crucial for recognition, regulation, and signaling.[22,23,25,37,38,55–57] A correlation has been established between the specific pattern in the PONDR® VL-XT curve and the ability of a given short disordered region to undergo disorder-to-order transitions on binding.[58] Based on these specific features an α-MoRF predictor was recently

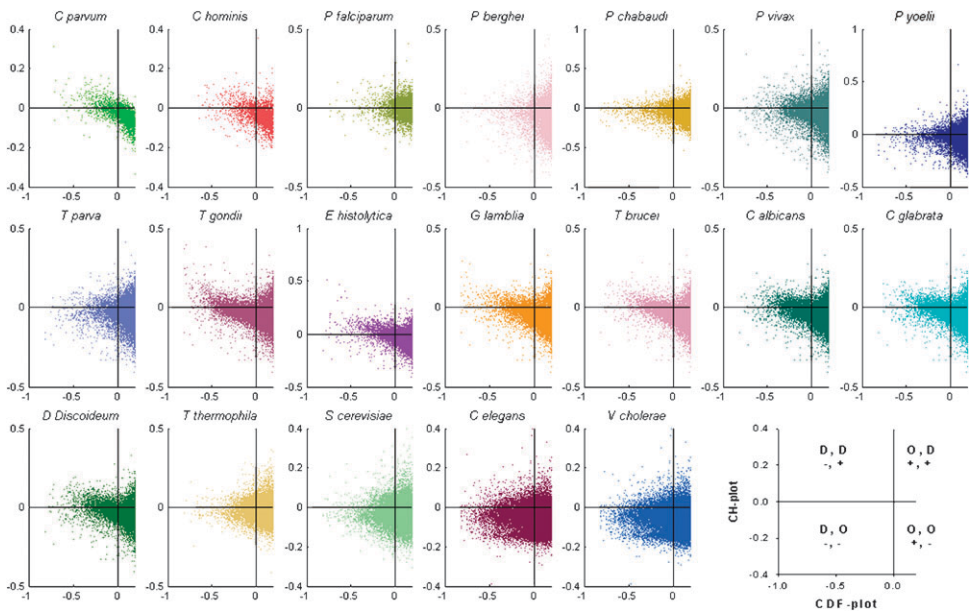

Fig. 3 Comparison of the PONDR® CDF and CH-plot analyses of whole protein order-disorder *via* distributions of proteins in each proteome within the CH-CDF phase space. Each spot represents a single protein whose coordinates were calculated as a distance of this protein from the boundary in the corresponding CH-plot (*y*-coordinate) and an averaged distance of the corresponding CDF curve from the boundary (*x*-coordinate). Four quadrants in each plot correspond to the following predictions: (−, −) proteins predicted to be disordered by CDF, but compact by CH-plot; (−, +) proteins predicted to be disordered by both methods; (+, −) contains ordered proteins; (+, +) includes proteins predicted to be disordered by CH-plot, but ordered by the CDF analysis. This is further illustrated by an explanatory plot at the bottom right corner.

developed.[55] Not all MoRF regions share these same features and some of them form β- or irregular structures rather than the α-helix.[56,57] Therefore, these predicted MoRFs are only fractions of the total numbers of MoRFs for each organism. Interestingly, recent studies showed that not all specific interactions between intrinsically disordered proteins are necessarily accompanied by the disorder-to-order transitions.[48,59,60] It has been hypothesized that such interactions of the immune receptor signaling subunits could play a crucial role in immune signaling.[61]

The application of the α-MoRF predictor to various datasets reveals that helix forming molecular recognition features are highly abundant in proteins from early-branching eukaryotes, suggesting the potential for extensive interaction networks in these proteomes. Table 1 shows that, on average, ~20% protozoan proteins contain α-MoRFs, ranging from 7.5% in *P. berghei* and ending with 48.3% in *T. gondii*. The number of α-MoRF-containing proteins in the prokaryotic representative *V. cholerae* is considerably smaller (1.9%). Importantly, in each proteome some long, highly disordered proteins have multiply predicted α-MoRF regions (Table S1†) that may potentially serve as binding sites for multiple proteins. For example, *C. elegans* protein CE25234 (4900 amino acid residues) has 49 predicted α-MoRFs. Similarly, *T. gondii* proteins 44.m02695 (putative protein phosphatase 2C, 3966 amino acids) and 42.m03467 (mediator complex subunit SOH1-related, 4253 amino acids) contain 24 and 22 predicted α-MoRFs, respectively.

Analysis of the *P. falciparum* protein–protein interaction map

One goal with this analysis was to study a published interaction map of *P. falciparum* with special emphasis on the role of intrinsic disorder in such a network. This map includes 2321 proteins involved in 19 979 protein–protein interactions.[62] A log–log plot of the number of proteins *versus* the number of interactions reveals that the network closely mimics the properties of a scale-free network (Fig. 4). In other words, there are few proteins participating in a high number of interactions (hubs) and a large number of proteins having few or no interactions. This finding is further supported by the regression analysis which shows that the data fits a linear equation with a negative slope. The fact that the R^2 value (0.9) is close to 1 is indicative of a reasonably good fit.

According to PONDR® VSL2B, the level of predicted intrinsic disorder in proteins form the interaction map of *P. falciparum* is about 45%. This number is marginally higher than the overall amount of disorder present in all annotated proteins from *P. falciparum* (41.6%) and is significantly higher than the level of intrinsic disorder in the *C. elegans* proteome (35.9%), the *V. cholerae* proteome (22.2%), as well as the mean disorder in all early-branching eukaryotes (39.0%, see Table 2). Furthermore, a correlation score of 0.13 (p-value = 0.0001) was obtained between the per protein PONDR® VLXT score and the number of interactions in which it participates. This indicates the presence of a weak association between intrinsic disorder and connectivity in the *P. falciparum* protein–protein interaction map.

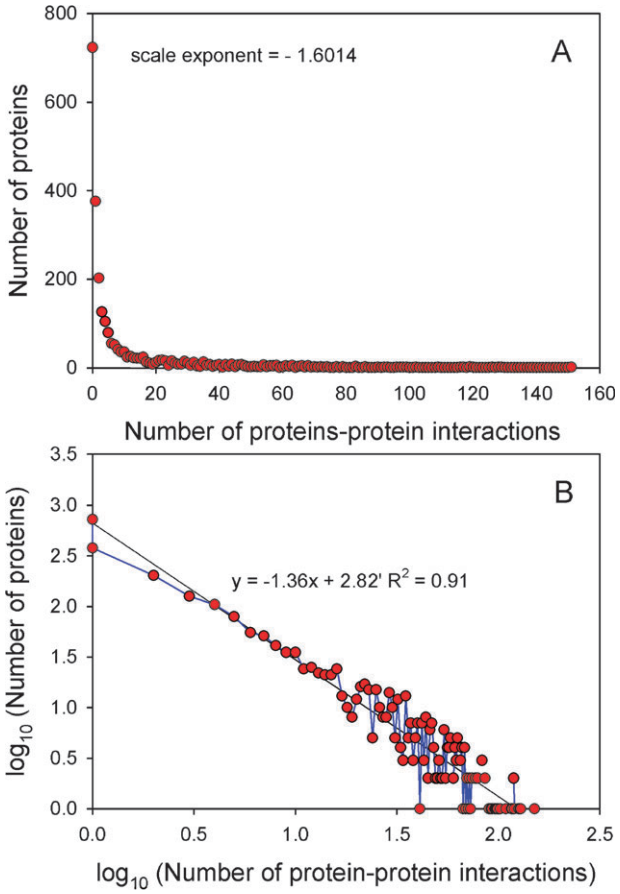

Fig. 4 Analysis of *P. falciparum* interaction map. (A) Number of protein–protein interactions (*x*-axis) *vs.* number of proteins (*y*-axis) based on *P. falciparum* interaction map published in Wuchty and Ipsaro.[62] (B) Log–log plot obtained using data from Fig. 4A. Data regressed onto linear equation obtained using least squares method.

Table 2 PONDR® VLXT and VSL2 scores for all organisms

EARLY-BRANCHING EUKARYOTES: PATHOGENIC		
Organism	VLXT (%)	VSL2B (%)
C. parvum	23.33	36.86
C. hominis	24.14	38.18
P. falciparum	17.83	41.57
P. berghei	12.59	32.78
P. chabaudi	13.97	33.78
P. vivax	30.25	44.31
P. yoelii	15.90	37.94
T. parva	22.26	34.48
T. gondii	46.14	57.40
E. histolytica	21.05	32.28
G. lamblia	30.56	40.16
T. brucei	33.76	38.09
C. albicans	27.56	39.43
C. glabrata	29.35	38.20
NON-PATHOGENIC		
T. thermophila	17.85	42.69
D. discoideum	25.94	39.32
S. cerevisiae	26.84	37.05
MODEL EUKARYOTE		
C. elegans	27.49	35.93
MODEL PROKARYOTE		
V. cholerae	23.02	22.24

It is intriguing to note that 10 randomly selected *P. falciparum* proteins that are largely disordered (*i.e.*, possessed a predicted disorder score >97%) but participate in one or no interactions have almost no PROSITE functional or CDD structural annotations associated with them. Likewise, almost no functional information is available for 10 randomly picked proteins involved in one or no interactions and with a predicted disorder score less than 5% (Table 3B, C). In contrast to earlier cases, most functional as well as structural information was available for the top 10 proteins with the highest number of protein–protein interactions and the top 10 proteins with more than 15 interactions (*i.e.*, putative hub proteins) but with a predicted disorder score less than 5% (Table 3A, D).

Interestingly, α-MoRF predictions for these data reveal that, of the 529 putative hub proteins (*i.e.*, proteins involved in 10 or more protein–protein interactions), 134 contain one or more predicted α-MoRF regions (25.3%). In comparison to this, 600 of the 1792 likely non-hub proteins had a corresponding α-MoRF prediction (33.5%). Both these numbers are higher that the average number of eukaryotic proteins with predicted α-MoRFs (~23%) and are significantly higher than a number of MoRF-containing proteins in bacteria (~3%) and archaea (~2.5%). In other words, both protein sets are highly enriched in disordered segments that are potentially involved in molecular recognition and that undergo disorder-to-order transitions upon interaction with their binding partners. Interestingly, non-hub proteins on average contain more α-MoRFs than hubs. On the other hand, hub and non-hub proteins are characterized by the PONDR® VLXT scores of 24.5% and 19.7%, respectively. This apparent discrepancy can be explained by the fact that MoRFs are short ordered regions (around 20 residues) located within long disordered regions. Therefore, higher MoRF content should correspond to lower overall disorder score.

Discussion

Despite a significant increase in the number of annotated genomes available for many protozoan and other early-branching eukaryotic organisms, much still remains to be learned about them. A large part of their fundamental physiology remains to be better understood, and the nature of many aspects of host-pathogen interactions continues to evade us. Many of these genomes are predicted to encode an abundance of low complexity regions. Such low complexity regions are typically indicative of the presence of intrinsic disorder. This study analyzes sequenced genomes corresponding to both pathogenic and non-pathogenic protozoa as well as fungi in order to study and compare the degree of disordered proteins expressed in these ancient eukaryotic cells. Our results indicate that early-branching eukaryotic proteins are generally enriched in predicted disorder as compared to typical eukaryotic proteins found in Swiss-Prot as well as ordered proteins from PDB. On average, the number of regions from protozoa with ≥30 consecutive disordered residues is twice the number found in a representative set of proteins from Swiss-Prot. This number translates to a 7-fold increase in comparison to the number of similar regions from a PDB Select 25 set of proteins.

Results from fractional amino acid compositions derived using ordered and disordered datasets from parasitic protozoa are found to be significantly depleted in tryptophan while continuing to remain enriched in lysine, in comparison to ordered proteins. Interestingly, in comparison to disordered proteins, most of the pathogenic early branching eukaryotes are found to be depleted in tryptophan. It is worthy to note that in comparison to ordered as well as disordered sequences, a majority of the 13 apicomplexan pathogen datasets are enriched in phenylanaline, isoleucine and tyrosine. Collectively, these observations may indicate unique properties associated with protozoal proteomes that differentiates them from typical ordered and intrinsically disordered proteins. For example, the overall reduction of tryptophan in *T. gondii* proteins is likely to be an adaptation of its intracellular lifestyle; *T. gondii* has evolved tryptophan auxotrophy and thus may have minimized its reliance on this essential amino acid by excluding it from its proteins.[63]

Cumulative distribution fraction curves clearly indicate that most pathogens have high cumulative values over most of their respective CDF curves. This suggests the possibility that most such organisms may largely contain compact proteins. *Vibrio cholerae* is the most ordered organism amongst all organisms studied. However, exceptions to this observation exist in the cases of *P. falciparum*, *P. berghei*, *P. chabaudi*, *P. yoelii* and *T. thermophila* (Table 1). These organisms contain a significant number of proteins that are potentially natively unfolded. This observation supports earlier findings suggesting that eukaryotes have higher disorder content in comparison to prokaryotes.[42,52] *C. elegans* and *S. cerevisiae* appear to have an apparent balance between ordered and disordered proteins.

We also studied a published protein–protein interaction dataset from *P. falciparum* consisting of 2321 proteins and 19 979 protein–protein interactions.[62] We find this network closely mimicking the properties of a scale-free network. 45% of these proteins contained intrinsic disorder.

This number is slightly higher than the overall amount of disorder present in all annotated proteins from *P. falciparum* (41.6%) and significantly higher than the level of intrinsic disorder in the *C. elegans* proteome (35.9%) and the *V. cholerae* proteome (22.2%), as well as the mean disorder in all early-branching eukaryotes (39.0%). A correlation study between overall disorder scores for each of these proteins and the number of interactions partaken by it indicates a low correlation coefficient of 0.13 (p-value = 0.0001). This number is suggestive of a weak association between intrinsic disorder and the degree of connectivity of proteins from this interaction dataset.

These data are in a good agreement with the results of several recently published studies, where the roles of disorder in protein–protein interaction networks were tested.[34,64–68] First, a set of structurally characterized hub proteins was assembled and analyzed.[34] Several hub proteins were found to be entirely disordered from one end to the other, and yet capable of binding large numbers of partners. Other hubs contained both ordered and disordered regions. For these

Table 3 Functional and structural annotations corresponding to top 10 proteins with (A) Highest number of interactions; (B) One or no protein–protein interaction(s) while having predicted disorder greater than 97%; (C) One or no protein–protein interaction(s) while having predicted disorder less than 5%; (D) Functional and structural annotations corresponding to top 5 proteins with more than 15 protein–protein interactions while having a predicted disorder score less than 5%

		Prosite domains associated	Conserved structural domains
Top 10 proteins with highest number of protein–protein interactions A	PF11_0061	Histone H4 signature	Histone H4
	PF13_0061	ATP synthase gamma subunit signature	ATP synthase
	PFD0665c	AAA-protein family signature	AAA (involved in membrane fusion, proteolysis and DNA replication)
	PF14_0655	DEAD-box RNA helicase Q motif profile, Superfamilies 1 and 2 helicase ATP-binding type-1 domain profile, Superfamilies 1 and 2 helicase C-terminal domain profile	DEAD box helicases, helicase superfamily C-terminal domain
	PF14_0068	Fibrillarin signature	Fibrillarin
	PF11_0096	Serine/Threonine protein kinases active-site signature	S/T kinase domains
	PF07_0054	Histone H2B signature	Histone 2B
	PF10_0366	Solute carrier (Solcar) repeat profile	Mitochondrial carrier protein
	PFC0400w	No information available	Ribosomal protein (N)
	PFL0585w	Ubiquitin domain profile, Ubiquitin domain signature	Ubiquitin (all-through)
Top 10 proteins with 0/1 protein–protein interaction & >97% predicted disorder (VSL2) B	PF10_0252	No information available	Cyt. C. oxidase chaperone (C)
	PFE0990w	No information available	No information available
	PF10_0280	No information available	No information available
	PFB0490c	No information available	No information available
	PFA0475c	No information available	Domain with unknown function
	PFA0420w	No information available	No information available
	PF11_0035	No information available	No information available
	PF13_0342	No information available	No information available
	PFL0155c	No information available	No information available
	PFB0095c	No information available	SbcC, ATPase involved in DNA repair [DNA replication, recombination, and repair], PPE-repeat proteins [Cell motility and secretion], Spumavirus gag protein
Top 10 proteins with 0/1 protein–protein interaction & <5% predicted disorder (VSL2) C	PF11_0150	No information available	Rhomboid, Rhomboid family. This family contains integral membrane proteins that are related to Drosophila rhomboid protein. Members of this family are found in bacteria and eukaryotes.
	PF14_0415	Dephospho-CoA kinase (DPCK) domain profile	Dephospho-coenzyme A kinase (DPCK, EC 2.7.1.24) catalyzes the phosphorylation of dephosphocoenzyme A (dCoA) to yield CoA, which is the final step in CoA biosynthesis
	PFC0935c	No information available	Glycosyl transferase family 4
	MAL13P1.117	Zinc finger DHHC-type profile	Uncharacterized protein containing DHHC-type Zn finger
	PFA0455c	ELO family signature	ELO, GNS1/SUR4 family. Members of this family are involved in long chain fatty acid elongation systems that produce the 26-carbon precursors for ceramide and sphingolipid synthesis. Predicted to be integral membrane proteins, in eukaryotes they are probably located on the endoplasmic reticulum.
	PF14_0317	No information available	Microsomal signal peptidase 12 kDa subunit
	PFE0660c	No information available	Uridine phosphorylase
	PF14_0399	G-protein coupled receptors family 1 signature, ADP-ribosylation factors family signature	Arl2, Arl2 subfamily. Arl2 (Arf-like 2) GTPases are members of the Arf family that bind GDP and GTP with very low affinity
	PF11_0370	No information available	No information available
	PFE0730c	No information available	Ribose 5-phosphate isomerase type A (RPI_A) subfamily; RPI catalyzes the reversible conversion of ribose-5-phosphate to ribulose 5-phosphate

Table 3 (*continued*)

		Prosite domains associated	Conserved structural domains
Top 5 proteins with ≧15 protein–protein interactions & <5% predicted disorder (VSL2)	PFB0200c	Aminotransferases class-I pyridoxal-phosphate attachment site	TyrB, Aspartate/tyrosine/aromatic aminotransferase
D	PF13_0156	Proteasome B-type subunits signature	Proteasome_beta_type_7, proteasome beta type-7 subunit
	PF14_0142	Serine/threonine specific protein phosphatases signature	Protein phosphatase 2A homologues, catalytic domain. Large family of serine/threonine phosphatases, including PP1, PP2A and PP2B (calcineurin) family members
	PF14_0676	Proteasome B-type subunits signature	Proteasome_beta_type_2, proteasome beta type-2 subunit

hubs, many, but not all, of the interactions mapped to the regions of disorder. Two highly structured hubs were found. For both of these structured hubs, the binding regions of the partners were entirely disordered.[34] Overall, this initial study suggested two primary mechanisms by which disorder is utilized in protein–protein interaction networks, either using disordered regions to interact with multiple structured partners or using structured proteins to interact with multiple disordered partners.

These ideas were further tested *via* bioinformatics studies on collections of hub proteins. Several of these studies support one of the two primary mechanisms, namely the common use of disordered regions by hub proteins to bind to multiple partners.[64–68] Further bioinformatics studies refine the analysis further with the suggestion that disorder is very commonly used for regions that bind sequentially to multiple partners (so called "date hubs"[68]). Finally, bioinformatics investigations of the binding partners of two mostly structured hubs, calmodulin and 14-3-3, suggest that the binding regions of their partners are very likely to be located in regions of disorder.[69,70]

Analysis of the possible roles of intrinsic disorder in protein–protein interactions suggested that one disordered region could bind to many partners (one-to-many signaling) and that flexibility would enable multiple disordered regions to bind to one site on one partner (many-to-one signaling).[71] The peculiarities of these two binding mechanisms were recently studied *via* the careful analysis of the illustrative structures of a one-to-many example (namely, the disordered regions in p53 binding to their many partners) and a many-to-one example (namely the binding site of 14-3-3 associating with many different disordered partners).[72]

For the one-to-many signaling example (using the structures currently in the PDB), a single disordered region of p53 was observed to form a helix when associating with one partner, a sheet with a second partner, an irregular structure with a third partner, and an irregular structure with a completely different trajectory with a fourth partner. The set of residues involved in these interactions exhibited a very high extent of overlap along the sequence.[72] Plotting the changes in the solvent accessible surface area for each amino acid *versus* its position in the sequence gives a binding profile. The binding profiles for the single region of p53 bound to four different partners were completely different. It is as if the same sequence is "read" by the different partners in entirely different ways.[72]

Table 4 Summary of number of sequences, average sequence length, and ambiguous residues corresponding to each dataset used in this study

Organism	Number of annotated sequences available and used	Average sequence length (rounded to the nearest integer)	Number ambiguous residues (replaced by alanine)
C. parvum	3806	597	355
C. hominis	3886	452	318
P. falciparum	5411	751	71
P. berghei	12 235	245	2029
P. chabaudi	15 007	194	2223
P. vivax	5352	682	368
P. yoelii	7861	433	8689
T. parva	4079	465	4
T. gondii	7793	720	10 287
E. histolytica	9766	389	0
G. lamblia	9646	351	0
T. brucei	8758	502	8779
C. albicans	6125	479	128
C. glabrata	5271	502	204
D. discoideum	4032	668	4124
T. thermophila	27 424	605	60 746
S. cerevisiae	11 081	435	1316
C. elegans	38 398	465	718
V. cholerae	3887	299	20

For a many-to-one signaling example (using structures currently in the PDB), five disordered sequences associated within a single binding groove in 14-3-3 were studied. The flexibility of the disordered regions enabled them to fit into a common binding site. Not only backbone flexibility, but also side-chain flexibility is implicated in the movements needed for the different sequences to be able to fit into the common binding site.[72] Importantly, this analysis also revealed that flexibility on the structured protein side of the complex (*i.e.*, within the 14-3-3 protein) also played a very important role in enabling the binding of many disordered segments to a single partner.[72] We believe that signaling and regulation interactions in protein–protein interactions networks of early-branching eukaryotes can also be described in terms of one-to-many and many-to-one models.

The abundance of intrinsic disorder we have identified using a set of bioinformatics tools which include compositional profiling, disorder prediction by PONDR® VLXT and VSL2, CH-plot and CDF analysis as well as α-MoRF prediction in the proteomes of early-branching eukaryotes, many of which are pathogenic, emphasizes the need for more research into their contribution to cellular physiology. While our analysis of a published *P. falciparum* interactome revealed a weak correlation between disorder and the proclivity to engage in protein–protein interactions, more such networks need to be evaluated to address this function. Additionally, it has been proposed that regions of intrinsic disorder in pathogenic protozoa provide a flexible means to facilitate host cell invasion and thwart immune mechanisms.[18] Given the high degree and unusual nature of the intrinsically disordered regions we have analyzed here, it is clear that further steps to elucidate their biological roles in the context of parasite physiology and pathogenesis would be effort well spent.

Conclusions

The abundance of intrinsically disordered proteins in several early-branching eukaryotic organisms has been evaluated. An unusual bias in the amino acids comprising protozoal proteomes, high abundance of intrinsic disorder among their proteins and prevalence of long disordered regions were indicated by bioinformatics methods. An analysis of a *Plasmodium falciparum* interactome indicated that protein–protein interactions may be at least one function of the intrinsic disorder. This study provides the results of the large-scale bioinformatics analysis of proteins expressed by several genomes of early-branching eukaryotes, involves the computational identification of all the intrinsically disordered proteins in these organsims and provides some clues of the role of these disordered proteins in physiological and pathological functions. Therefore, we provide here bioinformatics basis for the discovery and analysis of unfoldomes (the complement of intrinsically disordered proteins in a given proteome) of early-branching eukaryotes. Altogether, these analyses provide new insights into the evolution of intrinsic disorder in the context of adapting to a parasitic lifestyle and lay the foundation for further work on the subject.

Materials and methods

Sequences and datasets

Various online databases were used as sources for annotated genomes corresponding to the following species: *Plasmodium falciparum* (excluding mitochondrial and plastid proteins), *P. berghei*, *P. chabaudi*, *P. vivax*, *P. yoelii* (Release 3.4),[73] *Toxoplasma gondii* (Release 4.1),[74] *Theileria parva* (http://www.tigr.org),[75] *Cryptosporidium hominis* and *Cryptosporidium parvum* (http://cryptodb.org/cryptodb/),[76,77] *Candida albicans and C. glabrata*,[78] *Entamoeba histolytica* (http://www.tigr.org), *Giardia lamblia*[79] and *Trypanosoma brucei* (http://www.tigr.org). In addition to these, annotated data corresponding to the non-pathogenic free-living protozoan *Tetrahymena thermophila* (http://www.tigr.org), the slime mold *Dictyostelium discoideum* (http://dictybase.org/), and the yeast *Saccharomyces cerevisiae* (http://www.yeastgenome.org/), were obtained to serve as control organisms. *Caenorhabditis elegans* (http://www.wormbase.org/) and *Vibrio cholerae* (http://www.tigr.org) were used as models for a multicellular eukaryote and prokaryote, respectively. All occurrences of ambiguous residues such as B, X, or Z in the datasets were replaced by alanine, due to its neutrality to order as well as disorder. The total numbers of sequences, mean sequence lengths, and number of ambiguous residues for each working dataset have been summarized in Table 4.

A brief description of these organisms and their corresponding pathologies is presented below. *P. falciparum* causes the most dangerous form of malaria in humans. *P. vivax* is the most frequent and widely distributed cause of benign, but recurring (tertian) malaria in humans. *P. berghei*, *P. chabaudi*, and *P. yoelii* are three of the four malaria parasites of African murine rodents. *T. gondii* causes toxoplasmosis in warm-blooded vertebrates. *T. parva* is the causative agent of East Coast Fever (ECF), an acute, tick-borne disease causing high rates of morbidity and mortality in cattle. *Cryptosporidium* species cause diarrhoeal illness. *C. albicans* is a diploid fungus (a form of yeast) capable of causing opportunistic oral and genital infections in humans. *C. glabrata* is now recognized as a highly opportunistic pathogen of the urogenital tract as well as of the bloodstream in immunocompromised individuals. *E. histolytica* and *G. lamblia* are anaerobic protozoan parasites that infect the GI tract. *T. brucei* is parasitic protist that causes African trypanosomiasis (sleeping sickness) in humans and animals. *D. discoideum* (slime mold) is a soil-living amoeba that exists in uni- and multi-cellular forms. *T. thermophila* is a non-pathogenic free-living ciliated protozoan. *Saccharomyces cerevisiae* is a species of the budding yeast. *C. elegans* is a free-living nematode. *V. cholerae* is a gram negative bacterium that causes cholera in humans.

Compositional profiling

To gain an insight into the relationships between sequence and disorder, amino acid compositions from different datasets were compared using an approach recently developed for intrinsically disordered proteins.[25,80] To this end, the fractional difference in composition between a given set of proteins

and a set of reference proteins (either a set of ordered proteins[54] or disordered proteins form DisProt database,[81,82] or proteins from *Tetrahymena thermophila*, *Caenorhabditis elegans* or *Vibrio cholerae*) was calculated for each amino acid residue. The fractional difference was calculated as $(C_X - C_{\text{reference}})/C_{\text{reference}}$, where C_X is the content of a given amino acid in a given protein (or protein set), and $C_{\text{reference}}$ is the corresponding content in a set of reference proteins and plotted for each amino acid. In corresponding plots, the amino acids were arranged from the most rigid to the most flexible according to the Vihinen's flexibility scale, which is based on the averaged *B*-factor values for the backbone atoms of each residue type as estimated from 92 proteins.[44]

Predictions of intrinsic disorder

Disorder predictions for proteins corresponding to each of the above listed organisms were made using PONDR® VLXT[53,54] and VSL2B.[50] Cumulative distribution function curves or CDF curves[52] were generated for each dataset using PONDR® VLXT scores for each of the 19 organisms. CDF analysis discriminates between order and disorder by means of a boundary value. This value can be interpreted as a measure of proportion of residues with low and high disorder predictions. Additionally, charge-hydropathy distributions (CH-plots) were also analyzed for these organisms using methods as described in Uversky *et al.*[22]

α-MoRF predictions

The predictor of α-helix forming Molecular Recognition Features, α-MoRF, is based on observations that predictions of order in otherwise highly disordered proteins corresponds to protein regions that mediate interaction with other proteins or DNA. This predictor focuses on short binding regions within long regions of disorder that are likely to form helical structure upon binding.[55] It uses a stacked architecture, where PONDR® VLXT is used to identify short predictions of order within long predictions of disorder and then a second level predictor determines whether the order prediction is likely to be a binding site based on attributes of both the predicted ordered region and the predicted surrounding disordered region. An α-MoRF prediction indicates the presence of a relatively short (20 residues), loosely structured helical region within a largely disordered sequence.[55] Such regions gain functionality upon a disorder-to-order transition induced by binding to partner.[56,57]

Functional annotations

Additionally, we also made use of a protein–protein interaction map from *P. falciparum* published recently.[62] This map contains 19 979 interactions involving 2321 proteins. This map was generated by using logistic regression methods to interpret protein–protein interactions involved in conserved protein interactions, their underlying domain interactions and supplemental experimental data.[62] Our goal for working with this map was to compare the connectivity of *P. falciparum* proteins (*i.e.* how many interactions a given protein participates in) and their extent of intrinsic disorder.

Finally, functional and structural annotations were obtained using PROSITE[83] and Conserved Domain Database.[84]

Acknowledgements

This work was supported by grants from the National Institutes of Health LM007688-0A1 and GM071714-01A2 (A.K.D., and V.N.U.) and the Indiana Genomics Initiative (INGEN) (A.K.D.). INGEN is supported in part by Lilly Endowment Inc. Research in the W.J.S. laboratory is supported by grants from the National Institutes of Health (GM065051) and American Heart Association (0750201Z). The Programs of the Russian Academy of Sciences for the "Molecular and cellular biology" and "Fundamental science for medicine" provided partial support to V.N.U.

References

1 R. D. Adam, *Microbiol. Rev.*, 1991, **55**, 706.
2 S. L. Stanley, Jr, *Lancet*, 2003, **361**, 1025.
3 A. R. Teixeira, N. Nitz, M. C. Guimaro, C. Gomes and C. A. Santos-Buch, *Postgrad. Med. J.*, 2006, **82**, 788.
4 http://www.who.int/mediacentre/factsheets/fs259/en/.
5 R. W. Snow, M. H. Craig, U. Deichmann and D. le Sueur, *Parasitol. Today*, 1999, **15**, 99.
6 R. W. Snow, C. A. Guerra, A. M. Noor, H. Y. Myint and S. I. Hay, *Nature*, 2005, **434**, 214.
7 N. J. Hoxie, J. P. Davis, J. M. Vergeront, R. D. Nashold and K. A. Blair, *Am. J. Public Health*, 1997, **87**, 2032.
8 J. Flegr, *Schizophr. Bull.*, 2007, **33**, 757.
9 E. Rorman, C. S. Zamir, I. Rilkis and H. Ben-David, *Reprod. Toxicol.*, 2006, **21**, 458.
10 R. H. Yolken, S. Bachmann, I. Ruslanova, E. Lillehoj, G. Ford, E. F. Torrey and J. Schroeder, *Clin. Infect. Dis.*, 2001, **32**, 842.
11 S. Y. Wong and J. S. Remington, *AIDS*, 1993, **7**, 299.
12 T. R. Slifko, H. V. Smith and J. B. Rose, *Int. J. Parasitol.*, 2000, **30**, 1379.
13 M. Meissner, C. Agop-Nersesian and W. J. Sullivan, Jr, *Appl. Microbiol. Biotechnol.*, 2007, **75**, 963.
14 M. J. Gardner, N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M. S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser and B. Barrell, *Nature*, 2002, **419**, 498.
15 L. Aravind, L. M. Iyer, T. E. Wellems and L. H. Miller, *Cell*, 2003, **115**, 771.
16 E. Pizzi and C. Frontali, *Genome. Res.*, 2001, **11**, 218.
17 L. Brocchieri, *Genome. Res.*, 2001, **11**, 195.
18 Z. P. Feng, X. Zhang, P. Han, N. Arora, R. F. Anders and R. S. Norton, *Mol. Biochem. Parasitol.*, 2006, **150**, 256.
19 C. Mehlin, E. Boni, F. S. Buckner, L. Engel, T. Feist, M. H. Gelb, L. Haji, D. Kim, C. Liu, N. Mueller, P. J. Myler, J. T. Reddy, J. N. Sampson, E. Subramanian, W. C. Van Voorhis, E. Worthey, F. Zucker and W. G. Hol, *Mol. Biochem. Parasitol.*, 2006, **148**, 144.
20 C. Holt and L. Sawyer, *J. Chem. Soc., Faraday Trans.*, 1993, **89**, 2683.
21 P. H. Weinreb, W. Zhen, A. W. Poon, K. A. Conway and P. T. Lansbury, Jr, *Biochemistry*, 1996, **35**, 13709.
22 V. N. Uversky, J. R. Gillespie and A. L. Fink, *Proteins*, 2000, **41**, 415.
23 P. E. Wright and H. J. Dyson, *J. Mol. Biol.*, 1999, **293**, 321.
24 P. Tompa, *Trends Biochem. Sci.*, 2002, **27**, 527.
25 A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R.

Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner and Z. Obradovic, *J. Mol. Graph. Model*, 2001, **19**, 26.
26 G. W. Daughdrill, G. J. Pielak, V. N. Uversky, M. S. Cortese and A. K. Dunker, in *Protein Folding Handbook*, ed. J. Buchner and T. Kiefhaber, Wiley-VCH, Verlag GmbH & Co. KGaA, Weinheim, Germany, 2005, p. 271.
27 A. K. Dunker and Z. Obradovic, *Nat. Biotechnol.*, 2001, **19**, 805.
28 V. N. Uversky, *Protein Sci.*, 2002, **11**, 739.
29 V. N. Uversky, *Eur. J. Biochem.*, 2002, **269**, 2.
30 V. N. Uversky, *Cell Mol. Life Sci.*, 2003, **60**, 1852.
31 L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradovic and A. K. Dunker, *J. Mol. Biol.*, 2002, **323**, 573.
32 A. K. Dunker, C. J. Brown and Z. Obradovic, *Adv. Protein Chem.*, 2002, **62**, 25.
33 A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva and Z. Obradovic, *Biochemistry*, 2002, **41**, 6573.
34 A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva and V. N. Uversky, *FEBS J.*, 2005, **272**, 5129.
35 P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky and A. K. Dunker, *Biophys. J.*, 2007, **92**, 1439.
36 V. N. Uversky, C. J. Oldfield and A. K. Dunker, *J. Mol. Recognit.*, 2005, **18**, 343.
37 H. J. Dyson and P. E. Wright, *Curr. Opin. Struct. Biol.*, 2002, **12**, 54.
38 H. J. Dyson and P. E. Wright, *Nat. Rev. Mol. Cell Biol.*, 2005, **6**, 197.
39 S. Vucetic, H. Xie, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic and V. N. Uversky, *J. Proteome Res.*, 2007, **6**, 1899.
40 H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic and V. N. Uversky, *J. Proteome Res.*, 2007, **6**, 1917.
41 H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky and Z. Obradovic, *J. Proteome Res.*, 2007, **6**, 1882.
42 A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner and C. J. Brown, *Genome Inform. Ser. Workshop Genome Inform.*, 2000, **11**, 161.
43 S. Vucetic, C. J. Brown, A. K. Dunker and Z. Obradovic, *Proteins*, 2003, **52**, 573.
44 M. Vihinen, *Protein Eng.*, 1987, **1**, 477.
45 K. P. Ng, G. Potikyan, R. O. Savene, C. T. Denny, V. N. Uversky and K. A. Lee, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 479.
46 D. P. Denning, V. Uversky, S. S. Patel, A. L. Fink and M. Rexach, *J. Biol. Chem.*, 2002, **277**, 33447.
47 D. P. Denning, S. S. Patel, V. Uversky, A. L. Fink and M. Rexach, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 2450.
48 A. Sigalov, D. Aivazian and L. Stern, *Biochemistry*, 2004, **43**, 2049.
49 A. B. Sigalov, D. A. Aivazian, V. N. Uversky and L. J. Stern, *Biochemistry*, 2006, **45**, 15731.
50 K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker and Z. Obradovic, *BMC Bioinformatics*, 2006, **7**, 208.
51 Y. Cheng, T. LeGall, C. J. Oldfield, A. K. Dunker and V. N. Uversky, *Biochemistry*, 2006, **45**, 10448.
52 C. J. Oldfield, Y. Cheng, M. S. Cortese, C. J. Brown, V. N. Uversky and A. K. Dunker, *Biochemistry*, 2005, **44**, 1989.
53 X. Li, P. Romero, M. Rani, A. K. Dunker and Z. Obradovic, *Genome Inform. Ser. Workshop Genome Inform.*, 1999, **10**, 30.
54 P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown and A. K. Dunker, *Proteins*, 2001, **42**, 38.
55 C. J. Oldfield, Y. Cheng, M. S. Cortese, P. Romero, V. N. Uversky and A. K. Dunker, *Biochemistry*, 2005, **44**, 12454.
56 A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker and V. N. Uversky, *J. Mol. Biol.*, 2006, **362**, 1043.
57 V. Vacic, C. J. Oldfield, A. Mohan, P. Radivojac, M. S. Cortese, V. N. Uversky and A. K. Dunker, *J. Proteome Res.*, 2007, **6**, 2351.
58 E. Garner, P. Romero, A. K. Dunker, C. Brown and Z. Obradovic, *Genome Inform. Ser. Workshop Genome Inform.*, 1999, **10**, 41.
59 A. B. Sigalov, A. V. Zhuravleva and V. Y. Orekhov, *Biochimie*, 2007, **89**, 419.
60 P. Tompa and M. Fuxreiter, *Trends Biochem. Sci.*, 2008, **33**, 2.
61 A. B. Sigalov, *Trends Pharmacol. Sci.*, 2006, **27**, 518.
62 S. Wuchty and J. J. Ipsaro, *J. Proteome Res.*, 2007, **6**, 1461.
63 E. R. Pfefferkorn, M. Eckel and S. Rebhun, *Mol. Biochem Parasitol*, 1986, **20**, 215.
64 A. Patil and H. Nakamura, *FEBS Lett.*, 2006, **580**, 2041.
65 D. Ekman, S. Light, A. K. Bjorklund and A. Elofsson, *Genome Biol.*, 2006, **7**, R45.
66 C. Haynes, C. J. Oldfield, F. Ji, N. Klitgord, M. E. Cusick, P. Radivojac, V. N. Uversky, M. Vidal and L. M. Iakoucheva, *PLoS Comput. Biol.*, 2006, **2**, e100.
67 Z. Dosztanyi, J. Chen, A. K. Dunker, I. Simon and P. Tompa, *J. Proteome Res.*, 2006, **5**, 2985.
68 G. P. Singh and D. Dash, *Proteins*, 2007, **68**, 602.
69 P. Radivojac, S. Vucetic, T. R. O'Connor, V. N. Uversky, Z. Obradovic and A. K. Dunker, *Proteins*, 2006, **63**, 398.
70 D. M. Bustos and A. A. Iglesias, *Proteins*, 2006, **63**, 35.
71 A. K. Dunker, E. Garner, S. Guilliot, P. Romero, K. Albrecht, J. Hart, Z. Obradovic, C. Kissinger and J. E. Villafranca, *Pac. Symp. Biocomput.*, 1998, 473.
72 C. J. Oldfield, J. Meng, J. Y. Yang, V. N. Uversky and A. K. Dunker, in *The 2007 International Conference on Bioinformatics and Computational Biology*, ed. H. R. Arabnia, M. Q. Yang, and J. Y. Yang, CSREA Press, USA, Las Vegas and Nevada, USA, 2007, vol. 2, p. 553.
73 A. Bahl, B. Brunk, J. Crabtree, M. J. Fraunholz, B. Gajria, G. R. Grant, H. Ginsburg, D. Gupta, J. C. Kissinger, P. Labo, L. Li, M. D. Mailman, A. J. Milgram, D. S. Pearson, D. S. Roos, J. Schug, C. J. Stoeckert, Jr and P. Whetzel, *Nucleic Acids Res.*, 2003, **31**, 212.
74 J. C. Kissinger, B. Gajria, L. Li, I. T. Paulsen and D. S. Roos, *Nucleic Acids Res.*, 2003, **31**, 234.
75 M. J. Gardner, R. Bishop, T. Shah, E. P. de Villiers, J. M. Carlton, N. Hall, Q. Ren, I. T. Paulsen, A. Pain, M. Berriman, R. J. Wilson, S. Sato, S. A. Ralph, D. J. Mann, Z. Xiong, S. J. Shallom, J. Weidman, L. Jiang, J. Lynn, B. Weaver, A. Shoaibi, A. R. Domingo, D. Wasawo, J. Crabtree, J. R. Wortman, B. Haas, S. V. Angiuoli, T. H. Creasy, C. Lu, B. Suh, J. C. Silva, T. R. Utterback, T. V. Feldblyum, M. Pertea, J. Allen, W. C. Nierman, E. L. Taracha, S. L. Salzberg, O. R. White, H. A. Fitzhugh, S. Morzaria, J. C. Venter, C. M. Fraser and V. Nene, *Science*, 2005, **309**, 134.
76 P. Xu, G. Widmer, Y. Wang, L. S. Ozaki, J. M. Alves, M. G. Serrano, D. Puiu, P. Manque, D. Akiyoshi, A. J. Mackey, W. R. Pearson, P. H. Dear, A. T. Bankier, D. L. Peterson, M. S. Abrahamsen, V. Kapur, S. Tzipori and G. A. Buck, *Nature*, 2004, **431**, 1107.
77 M. S. Abrahamsen, T. J. Templeton, S. Enomoto, J. E. Abrahante, G. Zhu, C. A. Lancto, M. Deng, C. Liu, G. Widmer, S. Tzipori, G. A. Buck, P. Xu, A. T. Bankier, P. H. Dear, B. A. Konfortov, H. F. Spriggs, L. Iyer, V. Anantharaman, L. Aravind and V. Kapur, *Science*, 2004, **304**, 441.
78 R. C. Geer and E. W. Sayers, *Brief Bioinform.*, 2003, **4**, 179.
79 A. G. McArthur, H. G. Morrison, J. E. Nixon, N. Q. Passamaneck, U. Kim, G. Hinkle, M. K. Crocker, M. E. Holder, R. Farr, C. I. Reich, G. E. Olsen, S. B. Aley, R. D. Adam, F. D. Gillin and M. L. Sogin, *FEMS Microbiol. Lett.*, 2000, **189**, 271.
80 V. Vacic, V. N. Uversky, A. K. Dunker and S. Lonardi, *BMC Bioinformatics*, 2007, **8**, 211.
81 S. Vucetic, Z. Obradovic, V. Vacic, P. Radivojac, K. Peng, L. M. Iakoucheva, M. S. Cortese, J. D. Lawson, C. J. Brown, J. G. Sikes, C. D. Newton and A. K. Dunker, *Bioinformatics*, 2005, **21**, 137.
82 M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic and A. K. Dunker, *Nucleic Acids Res.*, 2007, **35**, D786.
83 E. de Castro, C. J. Sigrist, A. Gattiker, V. Bulliard, P. S. Langendijk-Genevaux, E. Gasteiger, A. Bairoch and N. Hulo, *Nucleic Acids Res.*, 2006, **34**, W362.
84 A. Marchler-Bauer, J. B. Anderson, P. F. Cherukuri, C. DeWeese-Scott, L. Y. Geer, M. Gwadz, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, G. H. Marchler, M. Mullokandov, B. A. Shoemaker, V. Simonyan, J. S. Song, P. A. Thiessen, R. A. Yamashita, J. J. Yin, D. Zhang and S. H. Bryant, *Nucleic Acids Res.*, 2005, **33**, D192.

