Published in final edited form as:

J Proteome Res. 2006 April ; 5(4): 888–898. doi:10.1021/pr060049p.

Conservation of Intrinsic Disorder in Protein Domains and Families: II. Functions of Conserved Disorder†

Jessica Walton Chen^{‡,§}, Pedro Romero^{‡,¶}, Vladimir N. Uversky^{*,‡,§,□}, and A. Keith Dunker^{*,‡,§}

[‡]Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202, USA

[§]Molecular Kinetics, Inc., Indianapolis, IN 46268, USA

[¶]School of Informatics, Indiana University – Purdue University Indianapolis, IN, USA

[□]Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia

Abstract

Regions of conserved disorder prediction (CDP) were found in protein domains from all available InterPro member databases, although with varying frequency. These CDP regions were found in proteins from all kingdoms of life, including viruses. However, eukaryotes had one order of magnitude more proteins containing long disordered regions than did archaea and bacteria. Sequence conservation in CDP regions varied, but was on average slightly lower than in regions of conserved order. In some cases disordered regions evolve faster than ordered regions, in others they evolve slower, and in the rest they evolve at roughly the same rate.

A variety of functions were found to be associated with domains containing conserved disorder. The most common were DNA/RNA binding, and protein binding. Many ribosomal proteins also were found to contain conserved disordered regions. Other functions identified included membrane translocation and amino acid storage for germination. Due to limitations of current knowledge as well as the methodology used for this work, it was not determined whether or not these functions were directly associated with the predicted disordered region. However, the functions associated with conserved disorder in this work are in agreement with the functions found in other studies to correlate to disordered regions.

We have established that intrinsic disorder may be more common in bacterial and archaeal proteins than previously thought, but this disorder is likely to be used for different purposes than in eukaryotic proteins, as well as occurring in shorter stretches of protein. Regions of predicted disorder were found to be conserved within a large number of protein families and domains. Although many think of such conserved domains as being ordered, in fact a significant number of them contain regions of disorder that are likely to be crucial to their function.

Keywords

intrinsic disorder; protein structure-function; disorder prediction; PONDR

[†]This work was supported in part by the NIH grant R01 LM007688-01A1 (to A. K. D.) and the Programs of the Russian Academy of Sciences for the “Molecular and cellular biology” and “Fundamental science for medicine” (to V. N. U.)

Corresponding Author Footnote: To whom correspondence should be addressed at Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, 714 N. Senate St., Suite 250, Indianapolis, IN 46202. Phone: 317-278-9650; fax: 317-278-9217; E-mail: vversky@iupui.edu (V.N.U.) or kedunker@iupui.edu (A.K.D.).

Introduction

Many proteins have been shown to exist as dynamic ensembles of mutually transmuting conformations in which the atom positions and backbone Ramachandran angles are not fixed as in ordered proteins, but vary significantly over time with no specific equilibrium values and typically undergo non-cooperative conformational changes. In other words, such proteins or protein regions (known as intrinsically disordered (ID) or natively unfolded proteins) do not have rigid 3-D structure under physiological conditions *in vitro*.^{1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20} Importantly, ID proteins are known to carry out numerous vital biological functions,^{2; 4; 6; 8; 21} being intensively involved in cell signaling,^{4; 9; 10} recognition,^{9; 10} nucleic acid and protein-protein interactions.^{1; 2; 11; 12; 13; 14; 15}

Intrinsic protein disorder is a very abundant phenomenon. Although the actual percent of proteins that contain disordered regions in nature is unknown, it can be estimated using the disorder predictors described previously. Early efforts identified in excess of 15,000 proteins with disordered regions.²² Later, this figure was updated to equal roughly 30% of the proteins in the SwissProt database.²³ When the PONDR® VL-XT disorder predictor was applied to whole genomes, it was found that the amount of disorder within different species varied widely.²⁴ From 9% to 57% of archaeal proteins, from 13% to 52% of bacterial proteins, and from 48% to 63% of eukaryotic proteins contained predicted regions of disorder of length 30 or greater. However, when the length was restricted to 50 or greater, most archaeal and bacterial species had less than 10% of their proteins predicted to contain disorder, while eukaryotic species showed 25% or more proteins contained such long disordered regions.

A recent study using a different disorder predictor, DISOPRED2, confirmed this disparity in proportion of long disordered regions between kingdoms.²⁵ On average, 2% of archaeal proteins, 4.2% of bacterial proteins, and 33% of eukaryotic proteins contained regions of disorder of length 30 or greater. This difference was even more pronounced for disordered regions of length 50 or greater. This large jump in the percentage of proteins with long predictions of disorder in nucleated, rather than non-nucleated, organisms was both remarkable and unexpected. It has been pointed out that many of the disordered regions and most if not all of the completely disordered proteins are involved in cell signaling or regulation. To explain these and similar observations it has been hypothesized that the higher abundance of intrinsic disorder in eukaryotes could be a consequence of the increased need for cell signaling and regulation in higher organisms.^{1; 4; 24; 26; 27; 28}

Studies of the functions of disordered regions in proteins have been done using experimentally verified disordered regions as well as predicted regions. One survey of disordered regions' functions was carried out via a thorough literature search on 115 known disordered regions.² This work identified twenty-eight functions of disorder in 98 of the regions. Several of the identified functions dealt with molecular recognition, such as protein binding, nucleic acid binding, and receptor-ligand binding. Disorder is thought to be important for molecular recognition because it allows for binding with high specificity and low affinity as well as binding to different partners of different shapes. Some regulatory domains were also identified as disordered. This functionality benefits from intrinsic disorder much in the same way that molecular recognition does. Additional disordered regions were found to be sites of chemical modification, such as phosphorylation, glycosylation, and methylation. It is hypothesized that this is beneficial because the disordered region can fold directly onto the modifying enzyme, whereas an ordered region would need to align exactly with the enzyme.

For the previously described functions, it is theorized that the disordered region undergoes a transition to order upon binding its partner. However, some functions were identified for

disordered regions that did not require a transition to order. These functions include flexible linkers or spacers, and entropic springs, bristles, and clocks. It is the flexibility of the disordered regions that gives these regions their function.

Another different function found for disordered regions is described as “structural mortar.” This was most commonly found in ribosomal proteins. For these disordered regions, although they become ordered in a sense on binding to the ribosomal RNA, they do not take on a particular ordered form, but rather whatever form they need in order to fill in the gaps in the ribosome structure. In this way, this function is distinct from those that require a disorder-to-order transition.

A more automated analysis of functions associated with disorder was carried out by comparing the frequency of certain functional annotations between predicted ordered and disordered regions.²⁵ This work also found that disordered regions tended to be associated with molecular recognition functions.

Based on the detailed analysis of literature data, a new view of protein function was postulated, which states that although some proteins derive their functions from ordered regions, many others derive their function from regions that are disordered in the native state.¹ Within the disordered protein world, there are thought to be two different types of disorder: extended and collapsed. Regions of extended disorder are defined by their lack of globularity. These extended disorder segments may be either completely unfolded (i.e., behave as random coils) or have transient secondary structures which are sampled as part of the ensemble (i.e., resemble the pre-molten globule state of globular proteins). In contrast, proteins exhibiting collapsed disorder, also called native molten globules, have persistent secondary structure, but no fixed tertiary structure. The molten globule is, as its name implies, globular in shape, but that shape is not fixed. Thus there are three possible states for a protein: order, extended disorder, and collapsed disorder. This set of possibilities has been labeled the protein trinity,²¹ or protein quartet.^{6; 7} Proteins can interconvert between these states due to various events such as binding to a partner and function can originate from any of these states or transitions between them.

It has been pointed out that intrinsically disordered proteins/regions are different from ordered ones by a number of features in their amino acid sequences and thus intrinsic disorder can be predicted.^{1; 22; 23; 29; 30; 31}

In our preceding paper, 3,653 regions of conserved disorder prediction were uncovered within 2,898 distinct InterPro entries.³² It has been shown that the majority of CPD regions were short, with less than 10% of those found exceeding 30 residues in length. The goal of this work was to identify the prevalence and functions of conserved disordered regions within protein domains and families.

Materials and Methods

Protein Data Bank BLAST Search

Once the conserved disorder regions had been found, the first issue to address was whether any of these putative disordered regions had ever been shown to be ordered; i.e., had their 3D structure determined by X-ray crystallography or NMR. A procedure was established to search for this information. Up to 100 protein regions matching each domain were used as query sequences in an all-against-all BLAST search. These sequences were the same ones that were used in the multiple sequence alignments. This BLAST search was done with default BLAST parameters, except the e-value was set to 0.001. For each domain containing a CPD, a record was kept of each hit against a PDB entry that had at least 70% or more sequence identity.

Next, each of these PDB hits was checked for overlap with CPD regions. Because the previous step involved a search on the entire protein domain, those BLAST hits that were in a different part of the domain than the CPD had to be excluded. To do this, a simple database search was performed for each CPD region to locate PDB BLAST hits whose start or end points, once converted to alignment-based positions, overlapped with the start or end points of the CPD region. The number of PDB hits within each CPD region was then tabulated. Additionally, the number of PDB hits within each CPD region that were most likely representative of matches to 3-D structures of complexes were counted. A PDB hit was labeled a “complex” if the chain label (the fifth character in the PDB ID) for the hit was not ‘A’ or ‘_’, both of which are used in the PDB database to represent structures of single chains. Although this methodology is not completely accurate, because some hits could be for chain ‘A’ in a complex, and some hits with multiple chains are not true complexes, it gives a general idea of the nature of the PDB hits.

Top CPDs by Kingdom

Each domain signature was classified according to the kingdom of the majority of its protein matches. Those domains for which 90% or more of its matching proteins belonged to the same kingdom were counted as “single kingdom” domains. Those where the most common kingdom was still less than 90% of the proteins were counted as “mixed kingdom” domains. The top 20 CPD regions for each kingdom were extracted from the database, ranked by score. The function of the domains containing each top CPD was researched using the InterPro database and literature searches. Additionally, for those CPDs with PDB matches, the matching PDB entries were examined to determine 1) if the structure was of a representative of the domain, 2) if the structure was the result of a complex, 3) if the CPD region of the domain was visible in the structure or whether it was missing and 4) which part of the CPD region was represented in the 3D structure.

Literature Search

Literature searches were conducted on the top five CPD regions for each kingdom. The purpose of these searches was to look for experimental evidence of disorder in the region identified as a CPD in the domain. This was done by searching PubMed (<http://www.pubmed.gov>) for the name of the domain plus one of the following words: disorder, disordered, unstructured, structure, NMR, crystal. Alternative names for the domain were used when available, from general literature about one or more proteins in the domain. Scientific journal articles found were then reviewed for evidence of intrinsic disorder.

Results

Top CPD Regions

Tables S1 through S5 in Supplementary Materials contain the top CPD regions for each kingdom. Each table lists the accession number(s) relevant to the domain, the name of the domain from the InterPro entry, the location of the CPD region within the domain, the description of the domain, summarized from the InterPro abstract, and a notation indicating if any 3-D matches were found for this part of the domain in PDB. The location of the region within the domain is the average start and end positions within each domain match, so is only an approximation.

Each of the domains in the previous tables was classified according to their known functions, based on the InterPro abstract. Table 1 shows the number of domains with CPDs for each kingdom having each function. The most common functions, besides unknown, were ‘DNA binding’, ‘ribosome structure’, ‘RNA binding’, and both kinds of protein binding (signaling/regulation and complex formation). Only eukaryotes and viruses had domains with function

‘protein binding (signaling or regulation)’ while only archaea and bacteria had domains with function ‘protein binding (complex formation)’.

PDB Matches to CPD Regions

There were 1,338 CPD regions (out of 3,653) that overlapped with one or more sequences in PDB. Of these, 774 had at least one PDB match that was not labeled as part of a complex. When individual alignment columns from CPD regions were checked for overlap with PDB regions, the percent of overlap decreased slightly to 33% overall and dropped to 13% for CPDs of length greater than 40. To illustrate this observation, Figure 1 displays a histogram of the percentage of CPD positions in various effective length ranges that matched PDB sequences. For CPDs with a length 30 or greater, the percentage of positions overlapping a PDB entry that is not a complex dropped to below 8%, and below 5% for CPDs with length 40 or greater. Figure 2 shows a histogram of PDB matches for CPD positions by kingdom. CPD regions in domains from multiple kingdoms had the highest percentage of columns matching a sequence position from PDB, and viruses had the lowest percentage. The different member database had widely varying percentages of CPD positions matched to PDB sequences. Pfam had the fewest CPD positions with overlapping known 3-D structure with about 22% overlapping in total. PIR Superfamily had the fewest positions overlapping with a PDB position from a non-complexed structure, with less than 9%. SUPERFAMILY had almost 95% of its CPD positions overlap with a 3D structure (Figure 3).

Literature Search

Results from literature searches on the domains containing the top five CPD regions for each kingdom, excluding those of unknown function, are summarized below. Graphs of the disorder prediction are displayed for each domain. Because of gaps in the alignments, the position of the CPD regions on the graphs will not correspond exactly with the position of the CPD region.

Dentin Matrix 1 (Eukaryota)—Related proteins bone sialoprotein and osteopontin were found to be completely disordered by NMR spectroscopy.³³ Based on sequence similarities, dentin matrix 1 is expected to be mostly or entirely disordered as well.³³ In this work, regions of conserved disorder prediction were found in the central portion of the domain, with the N and C terminals mostly predicted to be ordered (Figure 4A).

Fruit fly transformer (Eukaryota)—The member proteins of the transformer family are more highly diverged than other fruit fly proteins, with variable length repeats and an abundance of basic amino acids.³⁴ While there is no experimental evidence that the protein is disordered, there is also no evidence that it is ordered. The CPD region found extends through nearly all of the protein (Figure 4B).

Aspartyl beta-hydroxylase, N-terminal (Eukaryota)—The N-terminal end of this protein projects into the cytoplasm, followed by a transmembrane region, and then the C-terminal contains the catalytic domain. There is no evidence of order or disorder for the N-terminal region. The same gene can be alternatively spliced to form junctin, junctate (also known as humbug), and aspartyl beta-hydroxylase (BAH), which is the full protein. Only the latter contains the catalytic domain.³⁵ Both BAH and humbug have an apparent molecular weight roughly two times larger their actual molecular weight.³⁵ Figure 4C shows the location of the CPD region in the alignment.

Prion (Eukaryota)—Residues 1-124 of the human prion protein were found to be disordered by NMR.³⁶ A study of the hamster prion protein (residues 29-231) described the “random-coil nature of chemical shifts for residues 30-124” discovered by the heteronuclear [¹H]-¹⁵N nuclear Overhauser effect.³⁷ NMR studies of still other prion proteins also showed the N-

terminal tail (roughly 100 residues) was disordered.^{38; 39} The CPD region found in the prion family extends on average from residues 36 to 122 of the protein. Figure 4D shows the location of the CPD region in the alignment.

E-MAP-115 (Eukaryota)—This protein has an apparent molecular weight of 115,000, and a calculated molecular weight 84,051. The N-terminal region contains a microtubule binding region. There is a proline-rich area in the middle of the protein which possibly functions as a hinge.⁴⁰ The protein is regulated by phosphorylation.⁴¹ Most of E-MAP-115 is predicted to contain a conserved disordered region (Figure 4E).

Bacterial surface layer protein (Bacteria)—The N-terminal region of this protein, which contains the CPD region (Figure 5A), binds to a secondary cell wall polymer.⁴²

Translocated intimin receptor (Bacteria)—The translocated intimin receptor (Tir) is translocated out of bacterial cells and into the plasma membrane of host cells. The N-terminal and C-terminal portions of the protein protrude into the host cytoplasm, while the central portion is extracellular and binds intimin, which is secreted by the bacterial cells. The central portion of Tir protein has a known 3D structure when bound to intimin.⁴³ However, this segment of Tir is not within the N-terminal domain that contains CPD regions (Figure 5B). The structure of the N-terminal domain is unknown, but the N-terminal 100 residues are known to bind to a chaperone protein inside the bacterial cell, which facilitates the translocation of Tir.⁴⁴ This chaperone protein is required for stabilization and accumulation of Tir, suggesting that the binding of the chaperone to the N-terminal protects Tir from degradation.⁴⁴ Additionally, the Tir protein has a higher apparent molecular weight than expected.⁴⁵

Ribosomal protein L15, bacterial form (Bacteria)—In ribosomes, it is thought that the proteins act to stabilize the structure, that they function as a sort of mortar to fill in the gaps between the “RNA bricks”.⁴⁶ L15, and other ribosomal proteins, were shown to have extended segments when the entire large ribosome subunit was visualized by x-ray. These extended regions are “likely to be disordered outside the context provided by rRNA”.⁴⁶ The extended region of L15 covers residues 1 through 60 according to one model.⁴⁷ Figure 5C shows the location of the CPD region in the alignment.

HrpZ (Bacteria)—The HrpZ protein is secreted by bacteria. It binds to host cell membranes, probably forming an ion-pore. Experiments have shown that both the N-terminal residues 1-80 and C-terminal residues 201-345 bind to lipid bilayers. These terminal portions are also highly hydrophobic.⁴⁸ The CPD regions found in this work correspond approximately with these N- and C-terminal regions (Figure 5D).

Fertility inhibition (Bacteria)—The fertility inhibition (FinO) protein has been crystallized and its 3D structure determined. However, residues 1-25 had to be removed in order to crystallize the protein. Additionally, residues 26-32 were missing in the determined structure.⁴⁹ When FinO was exposed to trypsin through limited proteolysis, the fragment 62-170 showed to be protease resistant.⁵⁰ The predicted disordered region extends from residues 1 to 49 (Figure 5E).

Ebola nucleoprotein (Viruses)—No evidence of order or disorder for this protein was found. However, the C-terminal region of a nucleoprotein of another virus in the same order has been shown to be disordered.⁵¹ The graph of the disorder prediction for the family's alignment is shown in Figure 6A.

Minor capsid protein VI (Viruses)—This protein functions as a “cement” protein in holding together the virus structure. It also mediates uncoating of the virus during lytic infection.⁵² In a mature virus, protein VI is thought to form a trimer of dimers.⁵³ There are two predicted CPD regions in this protein (Figure 6B).

Ribosomal protein L19e (Eukaryota and Archaea)—As with ribosomal protein 11, discussed earlier, L19e contains a region of extended structure, which is thought to do disordered when not bound to rRNA. This extended region was seen to cover residues 52-90 in one study.⁴⁷

Ribosomal protein S8E (Eukaryota and Archaea)—Although the structure of ribosomal protein S8 has been determined, the structure of the proteins in family S8E, which is named based on sequence similarity to S8, has not.

Ribosomal protein L34e, C-terminal (Eukaryota and Archaea)—No evidence of order or disorder for this protein family was found, however, it has been described in preceding paragraphs how many ribosomal proteins are thought to contain regions of disorder that only take an ordered structure on binding to rRNA.⁴⁷

T-cell surface antigen CD2 (Eukaryota with viral homologues)—The N-terminal domain of this protein (roughly, residues 25 to 190) has a known structure.^{54; 55} The small peptide from residues 294 to 303, which is proline rich, was visualized in complex with a binding partner.⁵⁶ This small region is within the long CPD region from 285-335 (Figure 6C).

DNA topoisomerase, type II (Archaea)—This protein has a known structure from residues 58 to 97. This is theorized to be the DNA binding domain based on sequence homology.⁵⁷ This does not overlap with the CPD region (Figure 7).

Shikimate kinase (Archaea)—No evidence of order or disorder for this protein was found. Note that this protein is non-homologous to bacterial and eukaryotic shikimate kinase, which has a known structure.^{58; 59}

Finally, we did not find structural information related to the following proteins containing domains with top-five CPD regions: rubella capsid (Viruses), rubella membrane glycoprotein E2 (Viruses), circovirus coat protein (Viruses), gas vesicle synthesis (Archaea and Bacteria), DNA polymerase II large subunit DP2 (Archaea), ATP synthase A-type, A subunit (Archaea), and methyl-coenzyme M reductase operon protein C (Archaea).

Discussion

Functions of Conserved Disorder

The functions of domains containing conserved disordered regions may be used to speculate on the functions of conserved disordered regions. Because in most cases the CPD region only covered a part of the domain, it is possible that the disordered region is not required for the known function of the domain. However, given that this disorder is conserved through nearly all members of the domain, it seems likely that the disorder plays a role in at least one of the functions of the domain, whether that function is known or unknown. Since only the functions of domains containing the 20 longest CPD regions per kingdom were used for studying the function of conserved disorder, it is very possible that additional functions were present among the entire group of domains containing CPDs.

Most of the functions observed for domains containing CPD regions were shared across disordered regions from several kingdoms of life. All kingdoms had at least two domains whose

function was to bind DNA or RNA. There were numerous families of ribosomal proteins containing regions of conserved predicted disorder. Although none of the top 20 from eukaryota was a ribosomal protein family, several ribosomal proteins in the 'multiple kingdom' list were from both archaea and eukaryota. Both bacteria and eukaryota had a CPD region within a domain whose function was to bind cytoskeletal components. However, the bacterial protein with this function binds to (eukaryotic) host cytoskeletal components after the protein has been embedded in the host's cell membrane.

It is important to note that only eukaryota and viruses are predicted to use disordered regions for signaling and regulation via protein-protein binding. While there were conserved disordered regions predicted in bacterial and archaeal proteins that interact with other proteins, these interactions are part of complex formation, so are more permanent than the transient signaling and regulation interactions. Other work has also found that intrinsic disorder is especially prevalent in signaling and regulatory proteins.^{1; 4; 9; 10} This lends support to the theory that the use of disorder for ensuring interactions will have high specificity and low affinity (that is, for transient interactions) arose later evolutionarily than other uses for disorder, such as the "structural mortar" function of ribosomes.

There was also a difference among the different kingdoms' specific functions of DNA binding regions containing conserved disorder. In eukaryotic domains, most DNA-binding functions were in transcription regulation, with one functioning as a chromatin component. Among viruses, the DNA-binding functions were mainly for containing the viral genome. In bacteria and archaea, the functions were largely specific to DNA polymerase, DNA topoisomerase, or exonuclease activity.

There were two additional interesting functions associated with domains containing conserved disorder. These were 'membrane pore forming or crossing' and 'amino acid storage'. The former function label was assigned to protein families that were known to enter into a plasma membrane and either form a pore or cross through the membrane entirely. This function was mostly assigned to bacterial proteins. The translocated intimin receptor protein first crosses the bacterial cell wall to exit the cell, and then embeds itself into the target cell, where it facilitates bacterial attachment. The N-terminal of the intimin receptor protrudes into the host's cytoplasm to interact with various host proteins. This portion of the protein, which contains two conserved regions of predicted disorder, likely needs to be disordered in order to slip through the host's cell membrane. Similarly, the HrpZ protein penetrates host cell membranes and forms a pore through which virulence factors may pass. There was one protein family found in multiple kingdoms with this function: the colicin E3 translocation domain. This domain is found in antibiotic proteins that are encoded on plasmids, which are found mostly in bacteria as well, with one example in the unicellular eukaryotic parasite *E. cuniculi* and another in a Japanese rice hypothetical protein. This domain's function is to translocate the entire protein across the cell membrane.

The second interesting function of conserved disorder is in a domain whose function is amino acid storage in spores of some bacteria. The acid-soluble spore protein, gamma-type, has no known function other than to provide amino acids for bacterial spores after they germinate. This protein family is predicted to be mostly disordered, and is also highly gapped in its alignment. The use of disorder makes sense in this context, because the disordered protein would be quickly accessible to proteases for digestion into single amino acids for use in translation of new proteins.

There was one example of a conserved disordered region within a domain that has an entropic function. The podocalyxin family contains membrane proteins whose function is to keep parts of certain epithelial cells separated by charge repulsion. This sort of entropic function can best

be carried out by a disordered region, as it is free to sample various configurations, and can thus cover a larger volume of space than could an ordered protein. Therefore it is not surprising to see proteins with this function predicted to contain a conserved disordered region.

Finally, three conserved disordered regions were predicted to fall within protein families with catalytic function. Two of these were in archaeal families, and one was in a family that is mostly archaeal with a few bacterial members. Catalytic enzymes are thought to function via an ordered structure, so this seems at first like an error. However, no 3D structure was found for the part of these families which contained the CPD region. Therefore it is conceivable that these regions do not fall within the catalytic domain of these proteins. There are several other examples of disordered regions in enzymes, most of which do not have a function assigned to the disordered region.²

Assessment of Methodology

There are several ways of assessing the accuracy and usefulness of the conserved disorder prediction methodology. One is to check for experimental evidence of disorder within the regions identified as CPD regions. Although laboratory work to verify disorder is not common, there are indirect signs that can indicate, although not prove, the existence of intrinsic disorder. This paper detailed the results of literature searches for 25 of the top domains in which CPD regions were found that did not have any overlap with a known 3-D structure. For one of these domains, direct experimental evidence was found to support the predicted region of disorder. For ten domains, indirect evidence that the protein contained a disordered region was found. The other fourteen domains had no evidence of order or disorder.

The direct evidence of disorder came from the prion family of proteins. The region of conserved predicted disorder extends from residues 36 to 122 of the family, while residues 1 to 124 for human prion and 30 to 124 for hamster prion were shown to be disordered by NMR spectroscopy. This overlaps almost exactly with the CPD region. The graph of the disorder prediction for the alignment of the prion family shows that many of the first 30 or so residues also have highly conserved disorder, but there are segments where the percent of sequences disordered drops to below 50%, which is why the conserved disorder regions starts at 36. It may be that in some species (such as humans), the first 30 residues are entirely disordered, and in some species, there are regions of order in the first 30 residues. The CPD discovery methodology only finds regions of disorder that are conserved across nearly all members of the family, so many regions that are in fact disordered in some proteins will not be within a CPD region.

The fertility inhibition protein has a known structure through most of its length. In order to crystallize the protein, the first 25 residues had to be removed. This is an indirect sign of disorder, as disordered regions can inhibit crystal formation. Additionally, once the remainder of the protein was crystallized, residues 26 to 32 were missing in the crystal. This is also a strong indicator of disorder. The final evidence is that limited proteolysis shows that the protein is protease sensitive up to around residue 60. Disordered regions are more accessible to proteases and so digest more quickly than ordered proteins. The CPD region for this protein family extends from residues 1 to 49. Between the missing residues in the x-ray structure and the protease sensitivity, there is strong evidence that most of this region is in fact disordered.

Two of the ribosomal protein families researched, L15 and L19e, had very strong indirect evidence of disorder. It is thought that the ribosomal proteins act to stabilize the ribosome, in that they function as a sort of mortar to fill in the gaps between the RNA molecules. The large ribosomal subunit's structure has been visualized, and both L15 and L19e had regions of extended conformation. Likely, these regions are disordered, and when they bind to the ribosome structure, they take on whatever shape is necessary to bind the various parts together.

These extended regions in the ribosomal proteins coincided very closely with the regions of conserved predicted disorder. In L15, the observed extended region was from positions 1 to 60 and the CPD region was from positions 1 to 52. In L19e, the observed extended region was from positions 52 to 90, and the CPD region was from positions 53 to 124.

Seven additional domains researched had indirect evidence for disorder. Both aspartyl beta-hydroxylase and E-MAP-115 protein family members have been observed to have high apparent molecular weights. Although there can be other reasons for this phenomenon, such as post-translational modifications, there are examples where the reason for aberrant mobility turns out to be disorder⁶⁰. The indirect evidence for dentin matrix 1 is simply its similarity to related proteins which have been shown experimentally disordered. The fruit fly transformer family is highly diverged and full of variable length repeats. This points to the protein being disordered because regions of repeats have low complexity, and low complexity is associated with a lack of order.^{23; 61}

The translocated intimin receptor (Tir) is likely to have a disordered region at its N-terminus, which coincides with one of its predicted disordered regions. Its first 100 residues are known to bind to a chaperone protein, and without the chaperone protein, Tir is unstable and doesn't accumulate in the cell. Disordered regions are more sensitive to degradation than ordered regions, but can be protected from digestion by binding to a partner and undergoing a disorder-to-order transition. The fact that the N-terminal region of Tir requires binding to a partner to avoid degradation is a strong indication that it is disordered. Additionally, Tir has a high apparent molecular weight, which, as just explained previously, is also evidence for disorder.

Finally, both the T-cell surface antigen CD2 family and the DNA topoisomerase type II family have regions of known structure over some of their length, but no known structure over the region predicted to be disordered. It is possible that these regions are ordered, but it seems likely that if that were the case, then the whole protein would have been used to make the crystals for determining the 3D structure. This is therefore indirect evidence for disorder in these protein families.

Overall, almost half of the protein domains researched had at least indirect evidence to support the prediction of disorder within the domain. Most of the domains without any evidence were from viral and archaeal proteins.

Another method for checking the accuracy of the CPD discovery method is to check for experimentally verified regions of order within the supposed disordered regions. This will give some sense of the error rate for predicting this kind of disorder. Because the disorder predictor used is not perfect, it was expected that some regions identified as disordered would actually be ordered in real life. Because the accuracy of VL-XT increases with the increasing size of a region of consecutive disorder, it was also expected that long CPD regions would have fewer errors than shorter ones. This in fact turned out to be true.

The percent of positions within CPD regions found that overlap with a position in a sequence with known 3D structure can be used to estimate the error for this conserved disorder prediction methodology. Because those proteins which have a structure while in a complex may still be disordered when unbound, only those CPD regions that have a 3D structure alone were considered as true errors. Using this estimate, the error rate for prediction of positions of conserved disorder is around 19% for regions of length 20 or more, approximately 8% for regions of length 30 or greater, and less than 5% for regions of length 40 and over.

This error rate, based on PDB matches, varied based on kingdom. Although viruses and archaea had the lowest error rate, it may be simply because viral and archaeal proteins have not been as extensively studied, so there are fewer protein structures known. The structural genomics

initiatives are largely focused on eukaryotic and bacterial proteins. Because of this, the error rates for conserved disorder prediction in eukaryotic and bacterial domains are likely to be closer to the true error rate, at 15% and 18% for CPD regions of any length 20 or greater, respectively. Domains with member proteins in multiple kingdoms had a very high error rate, at 30%. It is unclear why this would be the case.

There was also a difference in error rates for prediction of conserved disorder in domains from different InterPro member databases. Pfam and PIRSF had the lowest percent of CPD positions overlapping with positions of known structure, with 13% and 9% respectively. PROSITE, SMART, and SUPERFAMILY domains all had an error rate in excess of 40%. As mentioned earlier, because SUPERFAMILY is built from proteins of known structure, it is not surprising that most of the CPD regions had a known structure. The fact that 18% of SUPERFAMILY's domains contained a CPD region is significant, because that nearly matches the observed error rate for overlap with PDB regions across all regions of conserved disorder. It is likely that those member databases with a high error rate just have more domains and families of known structure than the others, leading to more than the average number of PDB hits.

In summary, the inaccuracies in predicting regions of conserved disorder are the same as those in predicting regions of disorder in a single sequence, because the same disorder predictor is used. The VL-XT program is more accurate the longer the disordered region gets, and so long conserved disordered regions are also more accurately predicted. Although short CPD regions (length less than 30) are more likely to be wrong, the observed rate of error, on average 19%, is not high enough warrant an exclusion of short CPDs from consideration.

Conclusions

We have seen that protein domains of various functions appear to contain regions of disorder conserved across nearly all family members, including protein binding, nucleic acid binding, ribosome structure, and some more unusual functions such as membrane translocation. A difference was seen in the type of functions associated with conserved disorder between eukaryotes and prokaryotes. Eukaryotic proteins seem to use disorder for transient binding purposes (signaling and regulation) while prokaryotic proteins seem to use disorder for longer lasting interactions, such as complex formation.

There are several extensions to this work that might improve the accuracy or provide more information about regions of conserved disorder in protein domains and families. Firstly, using different kinds of disorder predictors could be used to improve the accuracy or the sensitivity of this search. Combining disorder predictors and looking for regions predicted to be disordered by the majority of them would likely improve the accuracy by reducing the false positive rate. Combining disorder predictors and looking for regions predicted to be disordered by any of them would likely improve the sensitivity by identifying regions of conserved disorder that were not identified using just one predictor. Another option would be to use a combination of a short disorder predictor and a long disorder predictor, to be able to more accurately identify short regions of conserved disorder. This work was limited to use of a single disorder predictor because of time constraints; this methodology required running the predictor on nearly a million protein sequences, which is a time-consuming process.

Extending the functional classification of domains containing conserved disordered regions to all such domains found would result in a more complete picture of the functions of conserved disorder. For this work, only a subset of domains was studied for function, due to the time-intensive nature of this kind of research. It is possible there are many other as-yet unknown functions of conserved disorder that did not occur in the subset selected.

Based on the results of this work, intrinsic disorder may be more common in bacterial and archaeal proteins than previously thought, but this disorder is likely to be used for different purposes than in eukaryotic proteins, as well as occurring in shorter stretches of protein.

In conclusion, some predicted regions of intrinsic disorder were found to be conserved within protein families and domains. Although many think of such conserved domains as being ordered, in fact a significant number of them contain regions of disorder that are likely to be crucial to their function.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hips KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. Intrinsically disordered protein. *J Mol Graph Model* 2001;19:26–59. [PubMed: 11381529]
2. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. *Biochemistry* 2002;41:6573–6582. [PubMed: 12022860]
3. Dunker AK, Brown CJ, Obradovic Z. Identification and functions of usefully disordered proteins. *Adv Protein Chem* 2002;62:25–49. [PubMed: 12418100]
4. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 2002;323:573–584. [PubMed: 12381310]
5. Uversky VN. Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding crossroads: which way to go? *Cell Mol Life Sci* 2003;60:1852–1871. [PubMed: 14523548]
6. Uversky VN. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* 2002;11:739–756. [PubMed: 11910019]
7. Uversky VN. What does it mean to be natively unfolded? *Eur J Biochem* 2002;269:2–12. [PubMed: 11784292]
8. Uversky VN, Gillespie JR, Fink AL. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* 2000;41:415–427. [PubMed: 11025552]
9. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *Febs J* 2005;272:5129–5148. [PubMed: 16218947]
10. Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J Mol Recognit* 2005;18:343–384. [PubMed: 16094605]
11. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol* 2002;12:54–60. [PubMed: 11839490]
12. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 2005;6:197–208. [PubMed: 15738986]
13. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 1999;293:321–331. [PubMed: 10550212]
14. Namba K. Roles of partly unfolded conformations in macromolecular self-assembly. *Genes Cells* 2001;6:1–12. [PubMed: 11168592]
15. Demchenko AP. Recognition between flexible protein molecules: induced and assisted folding. *J Mol Recognit* 2001;14:42–61. [PubMed: 11180561]
16. Fink AL. Natively unfolded proteins. *Curr Opin Struct Biol* 2005;15:35–41. [PubMed: 15718131]
17. Tompa P. Intrinsically unstructured proteins. *Trends Biochem Sci* 2002;27:527–533. [PubMed: 12368089]
18. Gunasekaran K, Tsai CJ, Kumar S, Zanuy D, Nussinov R. Extended disordered proteins: targeting function with less scaffold. *Trends Biochem Sci* 2003;28:81–85. [PubMed: 12575995]

19. Bracken C, Iakoucheva LM, Romero PR, Dunker AK. Combining prediction, computation and experiment for the characterization of protein disorder. *Curr Opin Struct Biol* 2004;14:570–576. [PubMed: 15465317]
20. Daughdrill, GW.; Pielak, GJ.; Uversky, VN.; Cortese, MS.; Dunker, AK. Natively disordered proteins. In: Buchner, J.; Kiefhaber, T., editors. *Handbook of Protein Folding*. Wiley-VCH, Verlag GmbH & Co KGaA; Weinheim, Germany: 2005. p. 271-353.
21. Dunker AK, Obradovic Z. The protein trinity--linking function and disorder. *Nat Biotechnol* 2001;19:805–806. [PubMed: 11533628]
22. Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Garner E, Guilliot S, Dunker AK. Thousands of proteins likely to have long disordered regions. *Pac Symp Biocomput* 1998:437–448. [PubMed: 9697202]
23. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. *Proteins* 2001;42:38–48. [PubMed: 11093259]
24. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* 2000;11:161–171.
25. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 2004;337:635–645. [PubMed: 15019783]
26. Vucetic S, Brown CJ, Dunker AK, Obradovic Z. Flavors of protein disorder. *Proteins* 2003;52:573–584. [PubMed: 12910457]
27. Liu J, Rost B. Comparing function and structure between entire proteomes. *Protein Sci* 2001;10:1970–1979. [PubMed: 11567088]
28. Sigler PB. Transcriptional activation. Acid blobs and negative noodles. *Nature* 1988;333:210–212. [PubMed: 3367995]
29. Romero P, Obradovic Z, Dunker AK. Sequence data analysis for long disordered regions prediction in the calcineurin family. *Genome Informatics* 1997;8:110–124. [PubMed: 11072311]
30. Romero, P.; Obradovic, Z.; Kissinger, C.; Villafranca, JE.; Dunker, AK. Identifying disordered regions in proteins from amino acid sequence. 1997 Proceedings of International Conference on Neural Networks; 1997. p. 90-95.
31. Dunker AK, Garner E, Guilliot S, Romero P, Albrecht K, Hart J, Obradovic Z, Kissinger C, Villafranca JE. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput* 1998:473–484. [PubMed: 9697205]
32. Chen CW, Romero P, Uversky VN, Dunker AK. Conservation of intrinsic disorder in protein domains and families: I. A Database of conserved predicted disordered regions. *J Proteome Res*. 2006
33. Fisher LW, Torchia DA, Fohr B, Young MF, Fedarko NS. Flexible structures of SIBLING proteins, bone sialoprotein, and osteopontin. *Biochem Biophys Res Commun* 2001;280:460–465. [PubMed: 11162539]
34. Kulathinal RJ, Skwarek L, Morton RA, Singh RS. Rapid evolution of the sex-determining gene, transformer: structural diversity and rate heterogeneity among sibling species of *Drosophila*. *Mol Biol Evol* 2003;20:441–452. [PubMed: 12644565]
35. Dinchuk JE, Henderson NL, Burn TC, Huber R, Ho SP, Link J, O'Neil KT, Focht RJ, Scully MS, Hollis JM, Hollis GF, Friedman PA. Aspartyl beta -hydroxylase (Asph) and an evolutionarily conserved isoform of Asph missing the catalytic domain share exons with junctin. *J Biol Chem* 2000;275:39543–39554. [PubMed: 10956665]
36. Zahn R, Liu A, Luhrs T, Riek R, von Schroetter C, Lopez Garcia F, Billeter M, Calzolari L, Wider G, Wuthrich K. NMR solution structure of the human prion protein. *Proc Natl Acad Sci U S A* 2000;97:145–150. [PubMed: 10618385]
37. Donne DG, Viles JH, Groth D, Mehlhorn I, James TL, Cohen FE, Prusiner SB, Wright PE, Dyson HJ. Structure of the recombinant full-length hamster prion protein PrP(29-231): the N terminus is highly flexible. *Proc Natl Acad Sci U S A* 1997;94:13452–13457. [PubMed: 9391046]
38. Lysek DA, Schorn C, Nivon LG, Esteve-Moya V, Christen B, Calzolari L, von Schroetter C, Fiorito F, Herrmann T, Guntert P, Wuthrich K. Prion protein NMR structures of cats, dogs, pigs, and sheep. *Proc Natl Acad Sci U S A* 2005;102:640–645. [PubMed: 15647367]

39. Calzolari L, Lysek DA, Perez DR, Guntert P, Wuthrich K. Prion protein NMR structures of chickens, turtles, and frogs. *Proc Natl Acad Sci U S A* 2005;102:651–655. [PubMed: 15647366]
40. Masson D, Kreis TE. Identification and molecular characterization of E-MAP-115, a novel microtubule-associated protein predominantly expressed in epithelial cells. *J Cell Biol* 1993;123:357–371. [PubMed: 8408219]
41. Masson D, Kreis TE. Binding of E-MAP-115 to microtubules is regulated by cell cycle-dependent phosphorylation. *J Cell Biol* 1995;131:1015–1024. [PubMed: 7490279]
42. Ilk N, Kosma P, Puchberger M, Egelseer EM, Mayer HF, Sleytr UB, Sara M. Structural and functional analyses of the secondary cell wall polymer of *Bacillus sphaericus* CCM 2177 that serves as an S-layer-specific anchor. *J Bacteriol* 1999;181:7643–7646. [PubMed: 10601228]
43. Luo Y, Frey EA, Pfuetzner RA, Creagh AL, Knoechel DG, Haynes CA, Finlay BB, Strynadka NC. Crystal structure of enteropathogenic *Escherichia coli* intimin-receptor complex. *Nature* 2000;405:1073–1077. [PubMed: 10890451]
44. Abe A, de Grado M, Pfuetzner RA, Sanchez-Sanmartin C, Devinney R, Puente JL, Strynadka NC, Finlay BB. Enteropathogenic *Escherichia coli* translocated intimin receptor, Tir, requires a specific chaperone for stable secretion. *Mol Microbiol* 1999;33:1162–1175. [PubMed: 10510231]
45. Kenny B, Finlay BB. Intimin-dependent binding of enteropathogenic *Escherichia coli* to host cells triggers novel signaling events, including tyrosine phosphorylation of phospholipase C- γ 1. *Infect Immun* 1997;65:2528–2536. [PubMed: 9199415]
46. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* 2000;289:905–920. [PubMed: 10937989]
47. Klein DJ, Moore PB, Steitz TA. The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J Mol Biol* 2004;340:141–177. [PubMed: 15184028]
48. Lee J, Klusener B, Tsiamis G, Stevens C, Neyt C, Tampakaki AP, Panopoulos NJ, Noller J, Weiler EW, Cornelis GR, Mansfield JW, Nurnberger T. HrpZ(PspH) from the plant pathogen *Pseudomonas syringae* pv. phaseolicola binds to lipid bilayers and forms an ion-conducting pore in vitro. *Proc Natl Acad Sci U S A* 2001;98:289–294. [PubMed: 11134504]
49. Ghetu AF, Gubbins MJ, Frost LS, Glover JN. Crystal structure of the bacterial conjugation repressor finO. *Nat Struct Biol* 2000;7:565–569. [PubMed: 10876242]
50. Ghetu AF, Gubbins MJ, Oikawa K, Kay CM, Frost LS, Glover JN. The FinO repressor of bacterial conjugation contains two RNA binding regions. *Biochemistry* 1999;38:14036–14044. [PubMed: 10529250]
51. Bourhis JM, Johansson K, Receveur-Brechot V, Oldfield CJ, Dunker KA, Canard B, Longhi S. The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. *Virus Res* 2004;99:157–167. [PubMed: 14749181]
52. Wiethoff CM, Wodrich H, Gerace L, Nemerow GR. Adenovirus protein VI mediates membrane disruption following capsid disassembly. *J Virol* 2005;79:1992–2000. [PubMed: 15681401]
53. Stewart PL, Fuller SD, Burnett RM. Difference imaging of adenovirus: bridging the resolution gap between X-ray crystallography and electron microscopy. *Embo J* 1993;12:2589–2599. [PubMed: 8334984]
54. Bodian DL, Jones EY, Harlos K, Stuart DI, Davis SJ. Crystal structure of the extracellular region of the human cell adhesion molecule CD2 at 2.5 Å resolution. *Structure* 1994;2:755–766. [PubMed: 7994575]
55. Jones EY, Davis SJ, Williams AF, Harlos K, Stuart DI. Crystal structure at 2.8 Å resolution of a soluble form of the cell adhesion molecule CD2. *Nature* 1992;360:232–239. [PubMed: 1279440]
56. Freund C, Kuhne R, Yang H, Park S, Reinherz EL, Wagner G. Dynamic interaction of CD2 with the GYF and the SH3 domain of compartmentalized effector molecules. *Embo J* 2002;21:5985–5995. [PubMed: 12426371]
57. Nichols MD, DeAngelis K, Keck JL, Berger JM. Structure and function of an archaeal topoisomerase VI subunit with homology to the meiotic recombination factor Spo11. *Embo J* 1999;18:6177–6188. [PubMed: 10545127]
58. Daugherty M, Vonstein V, Overbeek R, Osterman A. Archaeal shikimate kinase, a new member of the GHMP-kinase family. *J Bacteriol* 2001;183:292–300. [PubMed: 11114929]

59. Krell T, Coggins JR, Laphorn AJ. The three-dimensional structure of shikimate kinase. *J Mol Biol* 1998;278:983–997. [PubMed: 9600856]
60. Iakoucheva LM, Kimzey AL, Masselon CD, Smith RD, Dunker AK, Ackerman EJ. Aberrant mobility phenomena of the DNA repair protein XPA. *Protein Sci* 2001;10:1353–1362. [PubMed: 11420437]
61. Romero P, Obradovic Z, Dunker AK. Folding minimal sequences: the lower bound for sequence complexity of globular proteins. *FEBS Lett* 1999;462:363–367. [PubMed: 10622726]

Figure 1.
Histogram of PDB matches to CPD positions, by effective length

Figure 2.
Histogram of PDB matches to CPD positions, by kingdom

Figure 3.
Histogram of PDB matches to CPD positions, by member database

Figure 4.

Graph of disorder prediction for eukaryotic proteins with domains containing the top five CPD regions: dentin matrix 1 protein family (A); fruit fly transformer family (B); N-terminal domain of aspartyl β -hydroxylase family (C); prion family (D); and E-MAP-115 family (E). The CPD regions are shown as horizontal black lines.

Figure 5. Graph of disorder prediction for bacterial proteins with domains containing the top five CPD regions: surface layer protein family (A); N-terminal of the translocated intimin receptor family (B); bacterial ribosomal protein L15 family (C); HrpZ family (D); the fertility inhibition protein family (E). The CPD region is shown as a black horizontal line.

Figure 6.

Graph of disorder prediction for the viral proteins with domains containing the top five CPD regions: Ebola nucleoprotein family (A); minor capsid protein VI family (B); T-cell surface antigen CD2 (C). The CPD regions are shown as black horizontal lines.

Figure 7.
Graph of disorder prediction for DNA topoisomerase, type II. The CPD region is shown as a black horizontal line.

Table 1

Functions of domains containing CPDs, by kingdom

Function	Archaea	Bacteria	Number of Domains Containing CPDs with Function ^a			Multiple	Total
			Eukaryota	Viruses			
Protein binding (signaling or regulation)			7	2		2	11
Protein binding (complex formation)	5	5					10
DNA binding	5	1	4	2		1	13
RNA binding		2	2	6		1	11
Small molecule binding	2						2
Cytoskeletal binding		1	2				3
Polymerization		2					2
Membrane pore forming or crossing		3				1	4
Ribosome structure		3				9	12
Enzymatic/catalytic function	2	3				1	3
Coat/capsid formation				7			7
Amino acid storage		1					1
Entropic function			1				1
Unknown	6	3	4	6			19

^a Only domains containing CPD regions from the Top 20 lists (Tables 1-4) were counted