CONSERVATION OF INTRINSIC DISORDER IN PROTEIN DOMAINS AND FAMILIES

Jessica Walton Chen

Submitted to the faculty of the Bioinformatics Graduate Program
in partial fulfillment of the requirements
for the degree
Master of Science
in the School of Informatics,
Indiana University
August 2005

Accepted by the Bioinformatics Program Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics.

Master's Committee

____
A. Keith Dunker, PhD

____
Pedro Romero, PhD

____
Vladimir N. Uversky, PhD

Dedication

I dedicate this thesis to the memory of my mother, Ellen Stone Walton.

Acknowledgements

I am grateful to my thesis advisor, Dr. A. Keith Dunker, for his advice, support, and encouragement. Thank you!

I would like to thank Dr. Pedro Romero for his help with many parts of this project, as well as for serving on my committee. Also, I would like to thank Dr. Vladimir Uversky for serving on my committee.

Thank you to the Indiana University School of Informatics for awarding me the MDL Excellence in Informatics fellowship in 2002.

Finally, thank you to Ya Yue Van, President of Molecular Kinetics, for providing me with a license for the PONDR® VL-XT program. (Access to PONDR® was provided by Molecular Kinetics [6201 La Pas Trail, Ste 160, Indianapolis, IN 46268; www.molecularkinetics.com; main@molecularkinetics.com] under license from the WSU Research Foundation.)

Abstract

Protein regions which lack a fixed structure are called ‘disordered’. These intrinsically disordered regions are not only very common in many proteins, they are also crucial to the function of many proteins, especially proteins involved in signaling and regulation. The goal of this work was to identify the prevalence, characteristics, and functions of conserved disordered regions within protein domains and families.

A database was created to store the amino acid sequences of nearly one million proteins and their domain matches from the InterPro database, a resource integrating eight different protein family and domain databases. Disorder prediction was performed on these protein sequences. Regions of sequence corresponding to domains were aligned using a multiple sequence alignment tool. From this initial information, regions of conserved predicted disorder were found within the domains. The methodology for this search consisted of finding regions of consecutive positions in the multiple sequence alignments in which a 90% or more of the sequences were predicted to be disordered. This procedure was constrained to find such regions of conserved disorder prediction that were at least 20 amino acids in length.

The results of this work were 3,653 regions of conserved disorder prediction, found within 2,898 distinct InterPro entries. Most regions of conserved predicted disorder detected were short, with less than 10% of those found exceeding 30 residues in length. Regions of conserved disorder prediction were found in protein domains from all available InterPro member databases, although with varying frequency. Regions of conserved disorder prediction were found in proteins from all kingdoms of life, including viruses. However, domains found in eukaryotes and viruses contained a

higher proportion of long regions of conserved disorder than did domains found in bacteria and archaea. In both this work and previous work, eukaryotes had on the order of ten times more proteins containing long disordered regions than did archaea and bacteria. Sequence conservation in regions of conserved disorder varied, but was on average slightly lower than in regions of conserved order. Both this work and previous work indicate that in some cases, disordered regions evolve faster, 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 protein families 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.

This work has shown 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.

Table of Contents

List of Tables

List of Figures

I. Introduction

Although the function of a protein is generally thought to arise from its 3-dimensional structure, in some cases, its lack of structure gives it function [1]. Protein regions which lack a fixed structure are called ‘disordered’. These intrinsically disordered regions are not only very common in many proteins, they are also crucial to the function of many proteins, especially proteins involved in signaling and regulation [2].

The study of intrinsic disorder is important for many reasons, but mainly because disorder is thought to be crucially important for many types of protein-protein interactions such as signal transduction, regulation, and cell cycle control [3]. In order to fully understand these processes, be it for general scientific knowledge or for medical research or drug development, intrinsic disorder must be fully understood.

There exist many methodologies for classifying proteins and protein regions into domains and families based on amino acid sequence [4-8]. The protein members of these domains and families are generally assumed to share common functionality. Like proteins in general, these domains are largely thought to derive their functions from their structure. However, it is likely that there are many examples of protein domains or families which have a common disordered region or which are entirely disordered, and that this is the basis for the shared functionality among the members of the family. It is important to know which families and domains have functional disordered regions, because when novel proteins are identified, membership in a protein family is often used to give the protein a potential function. Thus, if regions of conserved disorder were identified in protein families, it would enable the identification of new proteins that are likely to contain the same disordered region.

The goal of this work is to identify the prevalence, characteristics, and functions of conserved disordered regions within protein domains and families.

II. Background

A. The Protein Structure-Function Paradigm

Historically, a protein's function has been ascribed to its 3-dimensional structure. Emil Fischer's work, published in 1894, led him to the "lock and key" concept for enzymes and substrates, which postulated that it was the shape of these proteins that conferred their ability to bind with their substrates, and hence carry out their functions [9, 10]. Within the next 40 years, scientists showed that it was possible for a protein to lose its native activity and to gain this activity back again. By the 1930's, papers had been published on protein denaturation, stating in essence that proteins have a structure which confers a specific function, and that this structure, and hence the function, is lost by denaturation [11, 12].

Linderstrøm-Lang further refined the idea of protein structure, describing how proteins had a primary structure (the amino acid sequence), a secondary structure (such as helices, coils, and sheets), and a tertiary structure, formed by folding of the secondary structure [13, 14]. This tertiary structure was presumed to be what determined the protein's function. The use of X-ray crystallography to determine the structure of proteins starting in 1960 [15, 16] further cemented the position of the "sequence-structure-function" paradigm as accepted truth. However, there were signs indicating that not all parts of proteins were ordered within these structures determined by x-ray crystallography. Regions of some proteins had missing electron density, even in regions that were known to have function [17].

The idea that a protein's function may not require a rigid structure was first seen in the literature in 1950. Fred Karush was studying the protein serum albumin, which will bind to almost any hydrophobic molecule. He concluded that the protein was able to "exist in many molecular configurations of approximately equal energy" [18]. The protein serum albumin does not have a specific 3-dimensional shape, but instead can assume different shapes depending on the molecules around it.

Since then, more and more evidence has been found indicating that not only are some proteins intrinsically disordered, but also that this disorder is crucial to these proteins' functions. This evidence will be described in detail in the next section.

B. Intrinsically Disordered Protein: the New Paradigm

The new view of protein function is that although some proteins derive their functions from ordered regions, many others derive their function from regions that are disordered in the native state. When a protein or protein region is intrinsically disordered, the molecules that make up the protein do not occupy a fixed position in space relative to each other, but instead occupy different positions relative to each other over time and across different proteins with the same sequence. That is, two proteins with identical sequences that are ordered will have essentially the same structure, but two such proteins that are disordered will not only each have a different conformation, this conformation will vary for each protein over time [17].

Within the disordered protein world, there are thought to be two different types of disorder: extended and collapsed. Regions of extended disorder, sometimes called random coils, are defined by their lack of globularity. These extended disorder segments may have transient secondary structures which are sampled as part of the ensemble. In contrast, proteins exhibiting collapsed disorder, also called 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 trio of possibilities has been labeled the protein trinity [19]. Proteins can interconvert between these states due to various events such as binding to a partner.

1. Experimental Detection of Intrinsic Disorder

There are several ways that intrinsic disorder can be detected in experimental settings. One is by identifying missing residues in X-ray crystallography experiments. Although completely disordered proteins cannot generally be crystallized, proteins with disordered regions can sometimes be. Because the disordered regions will be in a different position within each protein, they will not scatter x-rays the same way, and so the region will lack electron density in the final structure [17]. Other reasons for missing electron density can be "wobbly domains" (i.e. those that are ordered within themselves but vary in location relative to the whole protein) and technical problems with the crystallization process [1].

Circular dichroism (CD) spectroscopy can also be used to identify intrinsic disorder in proteins. The combination of near- and far-UV CD is capable of distinguishing between the three protein states (order, extended disorder, and collapsed disorder), but is not able to provide the exact location of the disorder, only that it is present [1].

A more specific method of identifying disordered regions is NMR spectroscopy. Although the technical details will not be discussed, essentially NMR is capable of locating residues which are moving rapidly. That is, it can identify specific residues or regions that are disordered. However, there are technical difficulties in performing NMR spectroscopy on molten globule proteins, so it can only be used to locate extended disorder [1].

Finally, regions of disorder can also be identified by protease digestion. Proteases require a certain amount of unfolded sequence before they can cut, and so disordered regions are digested much more rapidly (on the order of 10^5 to 10^7 times faster) than ordered regions. So, combining proteolysis with mass spectrometry can identify proteolytically stable ordered regions, indicating that the remaining regions are likely to be disordered [20].

2. Prediction of Disorder from Amino Acid Sequence

Predictors of various types have been built to try to predict disorder tendencies both in individual residues of a protein and in proteins as a whole. The earliest such predictor used neural networks to predict disorder on a residue-by-residue basis and achieved an accuracy of 73% against testing data [21]. This early work was refined into the PONDR® VL-XT predictor, which was trained against long regions of disorder identified from regions missing in x-ray structures [22, 23]. Additional predictors were developed using different neural networks as well as logistic regression, some for predicting different “flavors” of disorder [24] and some for different length classes of disordered regions [25]. All of these predictors function by calculating values for different attributes of each residue, and feeding these into either a neural network or a linear predictor. Some of these attributes used for prediction of disorder include the frequency of certain amino acids or types of amino acids, hydropathy, and coordination number. Each attribute is calculated as the normalized value of the feature over a sliding window [23]. A numeric value between 0 and 1 is outputted for each residue, with 0.5 being the threshold between ordered (less than threshold) and disordered (more than threshold).

Other types of disorder predictors were built for predicting entirely disordered proteins, that is, they classify whole proteins as ordered or disordered. For example, the use of CDF (Cumulative Distribution Function) analysis for this purpose has been developed from the VL-XT predictor. It accumulates the frequency of disorder scores from VL-XT and, based on a boundary determined computationally, classifies the protein as ordered or disordered [26]. Another "binary" disorder prediction method uses charge-hydropathy plots, which can be used to linearly separate ordered and disordered proteins [27].

The accuracy of these methods varies based on the dataset they are tested against. Recently, the VL-XT predictor had an accuracy of about 73% when tested against a dataset from the fifth Critical Assessment of Structure Prediction [28]. The sensitivity of VL-XT was 58% for disordered regions of length 30 or less, and 79% for those longer than 30 residues.

3. Prevalence of Intrinsic Disorder

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 [21]. Later, this figure was updated to equal roughly 30% of the proteins in the SwissProt database [23].

When the VL-XT disorder predictor was applied to whole genomes, it was found that the amount of disorder within different species varied widely [26]. 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 [29]. 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.

4. Functions of Intrinsic Disorder

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 [29]. This work also found that disordered regions tended to be associated with molecular recognition functions.

C. Protein Family Databases

With the advent of bioinformatics techniques, many databases of protein families have been developed. These databases consist of groupings of protein sequences or parts of protein sequences based on some criteria. Most protein family databases attempt to group together proteins that have a common function. Because of the large number of protein sequences now known, this process must be automated in order to cover any substantial fraction of the known proteins. Therefore most protein family databases group their members based on primary protein sequence. Some databases have human-curated patterns, models, or profiles which they then apply on a large scale while others use various algorithms to develop the patterns without the aid of humans. Curated databases tend to be more accurate but will miss classifying protein families or domains that exist in nature but are unknown to humans. Some databases combine the two approaches, having some curated entries and some computer-generated.

The European Bioinformatics Institute's InterPro database combines several protein family databases together. Each database within InterPro identifies its family members based on a different technique. These techniques are summarized in the following paragraphs, and listed in Table 1.

The Pfam database [4] is a large, curated collection of protein families and domains. It uses hidden Markov models and multiple sequence alignments to identify members of its families. The Pfam database is the largest of the members of InterPro.

PIR SuperFamily [8] uses a hierarchical clustering method for identifying "homeomorphic" families. They define homeomorphic as "sharing full-length sequence similarity and a common domain architecture". The PIR SuperFamily database is also curated.

The PRINTS database [30] is a database of protein fingerprints. Their fingerprints are built by an iterative process involving manual sequence alignment and identification of conserved motifs (fingerprints). These motifs are then used to search the source database for more matches, which are used to adjust the frequency matrices of the motifs.

The ProDom database [31] is based partially on the Pfam database. In addition to the Pfam domains, which are curated, ProDom also contains domains built automatically using a position-specific iterative BLAST search. Thus ProDom consists of curated and non-curated domains.

PROSITE [5] contains two kinds of family and domain signatures. Most of its signatures are in the form of patterns, which are short regular expressions used to match similar regions of sequences. The rest of the signatures are called profiles, which are position-specific scoring matrices built using hidden Markov models.

The SMART database [7] claims to represent genetically mobile domains. It contains mainly many extracellular domains. These domain signatures are built, like Pfam and ProDom, using hidden Markov models and multiple sequence alignments.

Unlike the other InterPro member databases, the SUPERFAMILY database [32] is built from a source database of proteins of known structure. It uses hidden Markov models to build its families of structured proteins. Proteins within the same superfamily have "structural, functional and sequence evidence for a common evolutionary ancestor."

The final member database is TIGRFams [6]. This database groups related proteins into orthologous groups, termed "equivalogs". These groups are sets of "homologous proteins that are conserved with respect to function since their last common ancestor." TIGRFams also uses hidden Markov models and curated multiple sequence alignments to define their families.

Table 1. Member databases of InterPro v. 7.2

Name	Description
Pfam	Large curated collection of protein families and domains, built using hidden Markov models
PIR Superfamily	Clustering of PIR proteins based on evolutionary relationships. Clusters contain proteins that are homologous and "homeomorphic" (full-length sequence similarity and domain architecture).
PRINTS	Protein fingerprint database.
ProDom	Protein domain database, based partly on Pfam, partly on automated domain detection using PSI-BLAST.
PROSITE	Protein families and domains database. Mostly consists of "patterns" (regular expressions).
SMART	Database of genetically mobile domains.
SUPERFAMILY	Database of superfamilies from proteins on known structure. Based on the SCOP database.
TIGRFAMs	A database of protein families based on "equivalogs".

III. Methods

A. Materials

1. Hardware

All work was performed on a personal computer with a Pentium 4, 2.08e GHz processor and 512 MB of RAM, running Windows XP, Service Pack 2.

2. Software

Third party software used for this project is listed in Table 2. MySQL was used for a relational database to store and query data. XEmacs was used to write perl scripts to perform required tasks and calculations for the project. ActivePerl's implementation of perl was used to run these perl scripts. PONDR® VL-XT software was used to predict order and disorder of protein sequences. For generating multiple sequence alignments, CLUSTAL W was used. For some file parsing and manipulation of multiple sequence alignments, BioPerl, an open-source perl module for bioinformatics, was used. Finally, BLAST [33] queries were performed using the Blastall program.

Table 2. Third-party software used

Name	Source	Usage	License Terms
MySQL v4.0.20a-nt	http://www.mysql.com	Relational database	GNU General Public License
XEmacs v21.4	http://www.xemacs.org	Text editor	GNU General Public License
ActivePerl v5.8.3	http://www.activestate.com	Perl interpreter	GNU General Public License
PONDR® VL-XT v1.9	http://www.pondr.com	Prediction of order/disorder for protein sequences	Individually licensed from Molecular Kinetics
CLUSTAL W v1.83	http://www.ebi.ac.uk/clustalw/	Multiple sequence alignments	Free
BioPerl	http://www.bioperl.org	Bioinformatics package for perl programming	Perl Artistic License
Blastall v2.2.10	http://www.ncbi.nlm.nih.gov/BLAST/	Stand-alone BLAST queries	Free

3. *Other*

In building the initial database for this work, data from a two public databases, UniProt and InterPro, were downloaded and imported into a relational database. Information about these databases is listed in Table 3.

Table 3. Public databases used

Name	URL	Version	Description of Data
UniProt	http://www.uniprot.org	Release 1.9	Protein accession numbers and sequences
InterPro	http://www.ebi.ac.uk/interpro/	Release 7.2	InterPro entry names and accessions, InterPro to UniProt mappings

UniProt is a protein resource containing all of SwissProt, TrEMBL, and PIR protein information. InterPro is a domain database that integrates eight individual domain databases. Each InterPro entry consists of one or more signatures from one or more of the member databases representing a single domain concept. Because of different methods of detecting domains used in the different member databases, each match within an InterPro entry may span a different region of a protein.

InterPro also divides its entries by "type", which is the type of entity it represents. There are six types defined in the InterPro database: active site, binding site, domain, family, post-translational modification site, and repeat. Most InterPro entries are domains or families, as shown in Table 4. For the purposes of this work, the word "domain" will be used to generically represent all of these types.

Table 4. Types of InterPro Entries

Type	(Abbreviation)	Number of Entries
Active site	(AS)	26
Binding site	(BS)	20
Domain		2411
Family		8035
Post-translational modification site	(PTM)	20
Repeat		197

B. Procedures

The overall goal of this work was to characterize the function and extent of disordered regions in conserved domains. The procedures outlined in the following sections were designed to accomplish this goal.

1. Initial Database Creation

A relational database was created using standard SQL syntax. Into this database, protein and domain information was imported by parsing downloaded data files (see

Section A.3) using perl scripts. Protein information, from UniProt, added to the database included the protein's accession number, name, amino acid sequence, and the kingdom and species the protein is from. Domain information, from InterPro, added to the database included the domain's accession number, name, and type. In addition to this information, UniProt-to-InterPro mappings were imported from InterPro to the local database. Each mapping lists an InterPro entry, a member database accession number, a protein accession number, and start and end positions, which indicate the location of the domain match within the protein.

Additional tables and attributes were added to the database for containing data which was created in later stages of the project. The final form of the relational database contained all the necessary information for location and analyzing regions of conserved predicted disorder. The tables of the database are described in Table 5. Specific attributes for each relational table are described in Tables 6 through 18.

Table 5. Description of tables in relational database

Table Name	Description	Refers To

`align_char`	Represents characters of sequences within multiple sequence alignments	`align_column,`
		`protein_domain`

`align_column`	Represents columns of multiple sequence alignments of domains	`domain_signature`
`cpd`	Represents regions of conserved disorder within domain signatures	`domain_signature`
`cpd_pdb_match`	Represents overlaps in position between a CPD region and a region of sequence found in the PDB database	`cpd, pdb_hit`
`domain_signature`	Represents domains found in InterPro member databases	`ipr_entry, ipr_db`
`gap`	Represents horizontal gaps in the alignments of protein domains	`protein_domain`
`ipr_db`	Represents member databases of InterPro	
`ipr_entry`	Represents InterPro entries (which may consist of multiple domain signatures from member databases)	
`pdb_hit`	Represents BLAST hits on the PDB	`protein_domain`

	database for protein matches to a domain	
protein	Represents proteins from UniProt	
protein_domain	Represents domain signature matches in proteins	domain_signature, protein
residue	Represents residues of proteins	protein
species	Represents species whose proteins are in the database	

Table 6. Description of align_char table

Attribute	Description	Type
DbAccession†	The accession number of the domain signature this alignment is for	VARCHAR(15)
AlignPos†	The position in the alignment of this character (from 1 to length of alignment)	INTEGER
ProteinAccession†	The accession number of the protein whose sequence this character is in	VARCHAR(6)
ProteinStart†	The position in the protein that the domain match starts in	INTEGER
ProteinPos	The domain-based position of this character in the protein (non-gapped)	DOUBLE
AlignChar	The character as this position (amino acid letter or gap character)	CHAR(1)

†Primary key

Table 7. Description of align_column table

Attribute	Description	Type
DbAccession†	The accession number of the domain signature this alignment is for	VARCHAR(15)
AlignPos†	The position in the alignment of this column (from 1 to length of alignment)	INTEGER
PctDisorder	The percent of sequences in this column that are predicted to be disordered	DOUBLE
PctGapped	The percent of sequences that have gaps at this column position	DOUBLE
ShannonEntropy	The raw Shannon's entropy for this column	DOUBLE
NormEntropy	The normalized Shannon's entropy for this column	DOUBLE
AvgDomainPos	The average of all of non-gapped positions in the domain this column represents	DOUBLE

†Primary key

Table 8. Description of cpd table

Attribute	Description	Type
DbAccession†	The accession number of the domain signature this conserved region is in	VARCHAR(15)
Start†	The starting position in the alignment for this	INTEGER

	conserved region	
End†	The ending position in the alignment for this conserved region	INTEGER
PctDisorder	The actual percent of predicted disorder in the entire region	DOUBLE
GapArea	The percent of alignment characters in this region that are gaps	DOUBLE
EffLength	The "effective" length of the region, calculated as (End-Start)* GapArea	DOUBLE
DomainStart	The average domain-based position this region starts at (will be less than or equal to Start)	DOUBLE
DomainEnd	The average domain-based position this region ends at (will be less than or equal to End)	DOUBLE
PdbMatches	The number of pdb_hit entries that overlap the region between DomainStart and DomainEnd	INTEGER
PdbComplexes	The number of PdbMatches that are from complexes	INTEGER
Score	The priority score for this region, calculated as EffLength*PctDisorder	DOUBLE
AvgEntroopy	The average normalized Shannon's entropy for the alignment columns that make up this CPD	DOUBLE

†Primary key

Table 9. Description of cpd_pdb_match table

Attribute	Description	Type
DbAccession†	The accession number of the domain signature the conserved region is in	VARCHAR(15)
CPDStart†	The starting position in the alignment for the CPD region	INTEGER
CPDEnd†	The ending position in the alignment for the CPD region	INTEGER
PDBId†	The ID of the PDB sequence that overlaps the CPD region	VARCHAR(5)
ProteinAccession†	The accession number of the protein that matched this PDB entry	VARCHAR(6)
PDBStart	The alignment column the PDB match starts in	INTEGER
PDBEnd	The alignment column the PDB match ends in	INTEGER

†Primary key

Table 10. Description of domain_signature table

Attribute	Description	Type
DbAccession†	The accession number of the domain signature this alignment is for	VARCHAR(15)
IprAccession	The accession number of the InterPro entry this domain signature is a part of	VARCHAR(9)
DbCode	The code representing which member database this domain is part of (from the ipr_db table).	INTEGER
AlignNumProts	The number of proteins that were kept in the alignment for this domain	INTEGER
Kingdom	The kingdom to which the highest percentage of	VARCHAR(45)

	protein matches to this signature belong	
KingdomPct	The percent of proteins matches to this signature that belong to the kingdom specified in the Kingdom attribute	DOUBLE
AvgLength	The average length of the protein sequences matching this signature	DOUBLE

†Primary key

Table 11. Description of gap table

Attribute	Description	Type
DbAccession†	The accession number of the domain signature to identify the alignment the gap is in	VARCHAR(15)
ProteinAccession†	The accession number of the protein that the gap is in	VARCHAR(6)
ProteinStart†	The starting position in the protein of the domain match	INTEGER
InsertPos†	The “insertion point” of the gap in the protein sequence, protein-based position (non gapped)	INTEGER
StartPos	The starting column position of the gap in the alignment	INTEGER
EndPos	Then ending column position of the gap in the alignment	INTEGER
FlankingVLXT	Will be one of ‘D’, ‘O’, or ‘X’ to represent the disorder prediction of the residues to the immediate left and right of the gap (disordered, ordered, or mixed)	CHAR(1)
PctDomainLength	The length of the gap as a percent of the length of the domain	DOUBLE

†Primary key

Table 12. Description of ipr_db table

Attribute	Description	Type
DbCode†	The integer code assigned to this member database	INTEGER
DbName	The name of the member database	VARCHAR(24)
DbAbbrev	The abbreviation for the database which serves as a prefix for accession numbers	VARCHAR(5)

†Primary key

Table 13. Description of ipr_entry table

Attribute	Description	Type
IprAccession†	The accession number of the InterPro entry	VARCHAR(9)
IprName	The name of the InterPro entry	INTEGER
IprType	The type of the InterPro entry (see Table 4)	DOUBLE

†Primary key

Table 14. Description of pdb_hit table

Attribute	Description	Type
DbAccession†	The accession number of the domain this hit	VARCHAR(15)

	was in	
`PDBId`†	The PDB ID for the hit	`VARCHAR(5)`
`ProteinAccession`	The accession number of the protein matched	`VARCHAR(6)`
`EVal`	The E-value of the match	`DOUBLE`
`QueryStart`	The starting position of the hit within the protein's domain	`DOUBLE`
`QueryEnd`	The ending position of the hit within the protein's domain	`DOUBLE`
`HitStart`	The starting position of the hit within the PDB sequence	`DOUBLE`
`HitEnd`	The ending position of the hit within the PDB sequence	`DOUBLE`
`FracIdentical`	The % identity for the hit to this PDB ID	`DOUBLE`

†Primary key

Table 15. Description of `protein` table

Attribute	Description	Type
`ProteinAccession`†	The accession number of the protein (from UniProt)	`VARCHAR(6)`
`ProteinName`	The name of the protein	`VARCHAR(45)`
`Kingdom`	The kingdom of the species this protein is from	`VARCHAR(45)`
`SpeciesId`	This species id of the species this protein is from	`INTEGER`

†Primary key

Table 16. Description of `protein_domain` table

Attribute	Description	Type
`DbAccession`†	The accession number of the domain signature	`VARCHAR(15)`
`ProteinAccession`†	The accession number of the protein that this domain match is in	`VARCHAR(6)`
`Start`†	The amino acid position at which this domain match starts	`INTEGER`
`End`	The amino acid position at which this domain match ends	`INTEGER`

†Primary key

Table 17. Description of `residue` table

Attribute	Description	Type
`ProteinAccession`†	The accession number of the protein this residue is part of	`VARCHAR(6)`
`Position`†	The position number of this residue in the protein	`INTEGER`
`AminoAcid`	The one-letter amino acid code	`CHAR(1)`
`VLXT`	The PONDR® VL-XT score for this residue	`DOUBLE`

†Primary key

Table 18. Description of `species` table

Attribute	Description	Type
`SpeciesId`[†]	The unique number used to identify the species	`INTEGER`
`Species`	The name of the species	`VARCHAR(45)`
`Kingdom`	The kingdom the species belongs to	`VARCHAR(45)`
`NumCPDs`	The number of CPD regions that are in proteins from this species	`INTEGER`
`NumSeqs`	The total number of protein sequences in the database from this species	`INTEGER`

[†] Primary key

2. *Disorder Prediction*

PONDR® VL-XT, software for prediction of disorder tendency of protein sequences, was run against all of the amino acid sequences of the proteins in the database. The resulting disorder score was saved in the database for each amino acid position. This raw disorder prediction information was used in later steps to find consecutive regions of conserved disorder prediction.

3. *Conserved Predicted Disorder Discovery*

A methodology was developed to search for regions of conserved predicted disorder (CPD) in domains. Briefly, this methodology consists of finding regions of consecutive positions in a multiple sequence alignment of all domain matches in which a high percent of sequences are predicted to be disordered. This procedure, which is spelled out in detail in the following paragraphs, was followed for each individual domain signature in the database.

First, all of the protein regions matching the domain signature were extracted from the database. If there were more than 100 such matches, then 100 were randomly selected. The amino acid sequences of these protein regions were then aligned using default settings with CLUSTAL W. The resulting text files, containing the multiple sequence alignment (MSA), were stored in a designated directory.

Second, the character of the gaps in the MSA was analyzed. This was done to weed out any domains whose alignments were potentially incorrect. For this analysis, each gap's length (a gap is a sequential row of positions within a single sequence in an alignment containing gap characters) was compared to the length of the domain for which the MSA was done. Any protein sequence within the alignment which contained a gap whose length was greater than a certain percentage of the domain length was removed from consideration in the alignment. This cutoff percentage was decided on by looking at the mode and standard deviation of the value for all gaps. Proteins containing gaps that were more than two standard deviations from the mode in size were eliminated.

Third, the flanking residues for each gap were checked for their disorder prediction score. Each gap was characterized based on whether the amino acid before and the amino acid after the gap were both predicted to be ordered, both predicted to be disordered, or one of each ('mixed'). This information was saved to be used in the next step.

Fourth, the percent of included positions in each alignment column which were predicted to be disordered was calculated and stored in the database. For sequences in a column that were gapped at that position, the disorder/order prediction for the flanking amino acids was used, as described in the previous paragraph. If both were ordered or disordered, then the gap character was counted as ordered or disordered. If they were mixed, then the gap character order/disorder was decided by a weighted coin flip. The coin flip was weighted based on the proportion of ordered and disordered residues in the database.

Lastly, the alignment columns were searched for regions where 90% or more of the sequences were predicted to be disordered. The smallest consecutive region that was considered was 20 columns in a row. Information about each of these regions (the domain accession number, and the start and end positions in the alignment) was stored in the database.

Once these initial CPD regions had been found, several statistics were calculated for each one. The “gap area” was calculated as the percentage of characters in the alignment between the start and end points of the CPD that were gap characters. The “effective length” of the CPD was calculated by finding the average length of the protein sequences within the region, omitting gaps. The standard deviation for the effective length was also calculated. The two statistics, gap area and effective length, are related, as the higher the gap area, the smaller the effective length will be, relative to the actual length. Additionally, the overall percent of positions predicted to be disordered in the entire CPD was calculated. Although the lower limit was 90%, this was done to determine exactly how conserved the disorder prediction was.

A final set of statistics were calculated for each CPD: the average position of the CPD within the domain. While the start and end positions of the region found previously were alignment-based, these domain positions were domain-based. For example, if a CPD region were found by the methodology described thus far in a certain alignment of a domain from positions 1 to 40, it is likely that the actual ending position of this region, were it mapped onto an actual protein sequence, would be less than 40. The domain-based start and end positions were calculated by finding the average length of the protein sequences, not counting gaps, up to the start or end position.

From the initial set of CPD regions found as described in the preceding paragraphs, only those with 10 or more protein matches used in the alignment and an effective length of 20 or more were kept as true CPD regions. This eliminated those regions found from too small a set of protein matches, and those that, although having a raw length of at least 20, had too many gaps, shrinking the effective length below the acceptable threshold.

4. Ranking of CPD Regions

In order to focus on a smaller number of CPD regions, a scoring system was devised to indicate the priority level for each conserved disorder region. This score for each CPD was calculated as the effective length of the conserved region multiplied by the actual percent predicted disorder for that region of the alignment. The effective length, as explained previously, was calculated finding the average length of non-gapped sequences in the region. All CPD regions were ranked using this measure, from highest to lowest.

5. 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 3D 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.

6. *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.

7. 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.

8. Sequence Conservation

Sequence conservation for each alignment column was calculated by applying Shannon’s entropy formula:

$$H(X) = -\sum\nolimits_{1}^{i} p^{i} \ln(p^{i})$$

where H is the entropy value, X is the alignment column number, and p^{i} is the frequency of the i^{th} letter of the alphabet. For this project, the alphabet was all amino acids plus the gap character. The formula results in a number which represents the degree of variability in the amino acids represented in that column. This number was normalized by dividing by the maximum possible Shannon’s entropy score. For each alignment column, the maximum possible entropy is calculated based on the number of sequences in that column. These normalized entropy values can be directly compared, with a range from 0 (all sequences the same) to 1 (all sequences different, as much as possible).

The average entropy for each CPD region was calculated by averaging the values for each column involved in the CPD region. For comparison purposes, Conserved Predicted Order (CPO) regions were found in the same way that CPDs were found, with the difference being that 90% or more of sequences had to be predicted to be ordered rather than disordered. These CPOs were subject to the same effective length and minimum protein sequence restrictions as CPDs. The average entropy was then calculated for all CPOs.

Additionally, the average entropy for all "disordered" alignment columns (those with 90% or more sequences predicted to be disordered) as well as the average entropy for all "ordered" alignment columns (those with 90% or more sequences predicted to be ordered) was calculated.

C. Analysis

1. Method of Project Evaluation

The success of the project was evaluated based on whether or not a significant number of regions with conserved disorder prediction were found. Additionally, a majority of the regions found should have no known three-dimensional structure in unbound form in order to conclude that the regions found are actually likely to be disordered in real life.

2. Method of Results Analysis

Generally, results were analyzed by first extracting the relevant data from the database, sometimes using the grouping functionality of SQL to count or to find averages and standard deviations, and then either formatting the data into a table or graphing it in histogram form.

IV. Results

A. Database

The database constructed contained nearly one million proteins (961,216) from 62,305 different species. The most commonly represented kingdom was eukaryota, followed by bacteria. There were about 800 proteins whose classification was unknown. Table 19 shows some basic statistics on the proteins by Kingdom.

Table 19. Proteins in the database by Kingdom

Kingdom	# Proteins	# Species	Average Length
Archaea	28,888	336	318.5
Bacteria	342,300	7,111	344.7
Eukaryota	405,146	47,660	398.1
Viruses	184,101	7,425	247.5
Unclassified	463	23	163.2
Unknown	316	1	837.6
plasmids	1	1	260.0
transposons	1	1	287.0
Total	961,216	62,305	347.8

The database included 15,498 distinct domain signatures from the eight member databases, representing 10,709 InterPro entries. Pfam accounted for nearly half of the domain signatures. There were over 4.5 million domain matches to proteins. Over 90% of the proteins in the database contained a match to a Pfam domain. A breakdown of the domains and protein matches by member database is shown in Table 20.

Table 20. Domains and domain matches by database

Member Database	Domains	Domain Matches	Proteins with Domain Matches	Average Matches per Domain	Average Domain Match Length
Pfam	7,316	1,413,574	896,537	193	144
PIR Superfamily	406	8289	8,289	20	356
PRINTS	1,849	1,235,460	228,163	668	17
ProDom	993	185,352	165,298	186	136
PROSITE	1,752	857,410	441,789	489	57
SMART	659	337,000	168,072	511	91
SUPERFAMILY	602	488,936	345,782	812	125
TIGRFams	1,921	171,877	140,690	89	295
Total	15,498	4,697,898	963,428	303	95

B. Conserved Disorder Prediction

1. Multiple Sequence Alignments

Of the 15,498 domains in the database, 13,824 were successfully used to generate multiple sequence alignments. The rest had too few protein matches or resulted in CLUSTAL W errors for various reasons and were discarded. These alignments contained 3,129,498 columns in total.

2. Gap Analysis

The gaps in the alignments were analyzed in order to eliminate protein sequences that were a poor match to the domain alignment, as described in section B.3. There were 5,018,083 gaps in all sequences of all alignments, with an average of 363 gaps per alignment. The average raw gap length was 10.3 positions with a standard deviation of 38.6. The length of the gaps when calculated as a percentage of the domain length was on average 5.9% with a standard deviation of 17.6%. Figure 1 shows a histogram of the gap length as a function of domain length. The first bucket, which represents gaps whose lengths are between 0% and 2% of their domain length, accounts for 59% of the gaps.

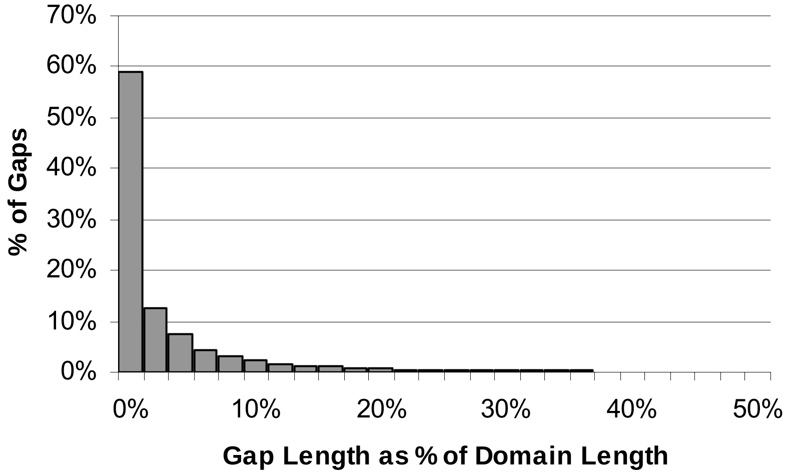

Figure 1. Histogram of gap length as a function of domain length

The cutoff for gap length was set at 35%, which is approximately two standard deviations from the mode. Any protein sequence containing a gap larger than 35% of the domain length was excluded from further analysis. A histogram of the percent of proteins containing gaps of a certain length, based on the domain length, is shown in Figure 2. There were 85,142 protein sequences that did not fit this criterion and were eliminated. Based on these eliminations, 10,802 domain signatures had 10 or more protein matches. Table 21 shows the number and percent of domains for each member database that had 10 or more protein matches.

Table 21. Domains with 10+ Protein Matches by Database

Member Database	Domains with 10+ Protein Matches	% of Initial Domains
Pfam	5,265	72.0%
PIR Superfamily	231	56.9%
PRINTS	1012	54.7%
ProDom	680	68.5%
PROSITE	1,289	73.6%
SMART	409	62.1%
SUPERFAMILY	369	61.3%
TIGRFams	1,547	80.5%
Total	10,802	69.7%

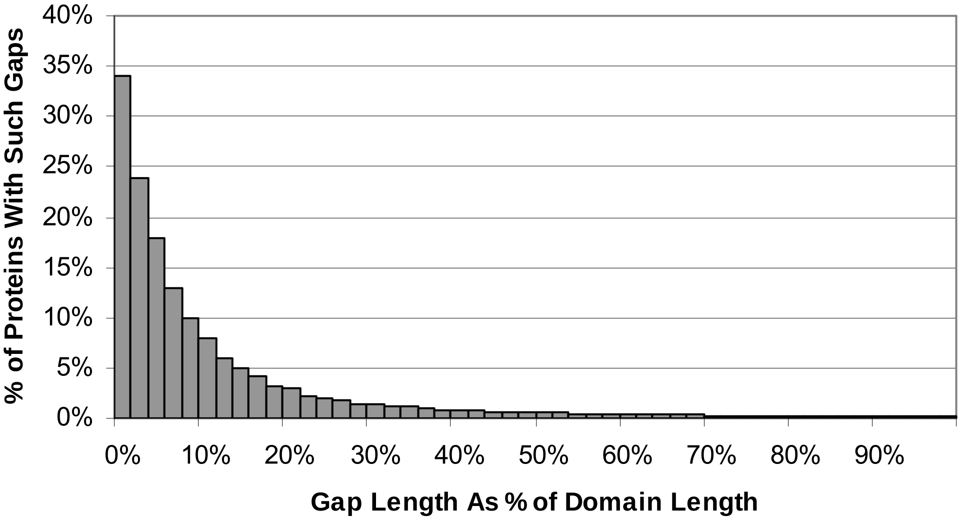

Figure 2. Histogram of gap length as a percentage of domain length, for proteins

Next, the disorder predictions for the residues before and after each gap were analyzed. Seventy-two percent of the gaps had ordered residues to each side, while almost twenty-four percent had disordered resides to each side. Only 4.4% of the gaps had mixed (one ordered and one disordered) flanking residues. Table 22 shows the count, percent, and average length of the different categories of gaps. ‘Mixed’ gaps were about three times longer, on average, than ordered or disordered-flanked gaps.

Table 22. Statistics on gaps, by flanking order/disorder category

Category	Count	% of Total	Average Length
Ordered	3,613,676	72.0%	9.4
Disordered	1,184,549	23.6%	9.5
Mixed	219,858	4.4%	29.2

3. Disorder Prediction for Alignments

Part of the procedure for finding the conserved disordered regions included finding the percent of disorder prediction for each column of each alignment, as explained in section B.3. The histogram for percent of sequences with disorder predicted is shown in Figure 3. This histogram does not include columns with no disorder predicted, of which there were 542,355 (18.8% of columns). The distribution is centered at around 10% with a long tail extending all the way to 100%.

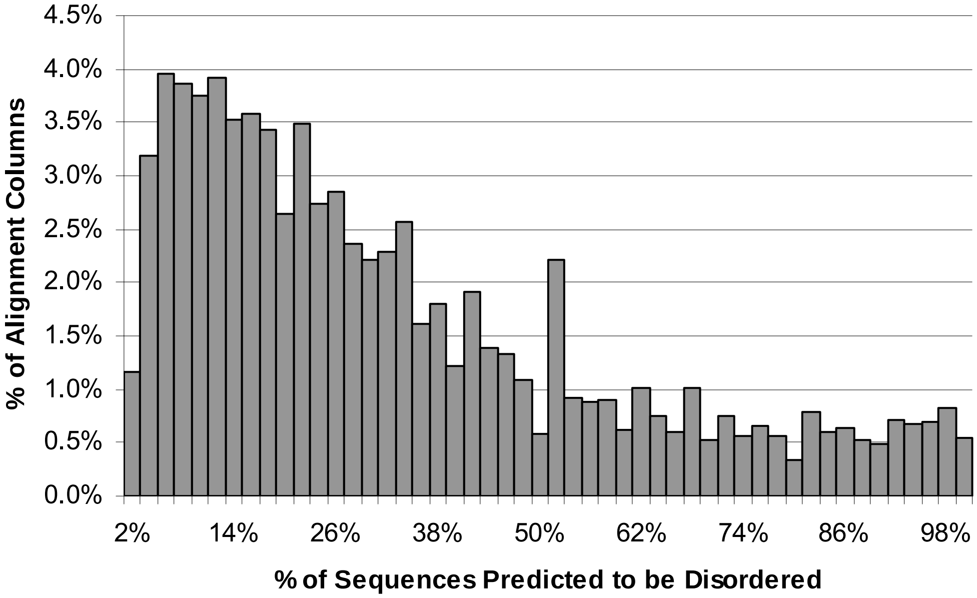

Figure 3. Histogram of percent predicted disorder for alignment columns. Columns with exactly 0% disorder, accounting for 18.8% of columns, were not included in the histogram.

4. Conserved Predicted Disorder Regions

A total of 3,653 Conserved Predicted Disorder (CPD) regions in 3,392 domains, representing 2,898 distinct InterPro entries, were discovered in the database. As explained in section B.3, only those regions with an effective length of 20 or greater and which were based on alignments of 10 or more sequences were included in the final set of CPD regions.

a. CPD Regions by InterPro Member Database

These CPDs were found in all eight member databases, although very few were found in the PRINTS database. The percent of domains containing CPDs for each database was calculated based on the number of domains that, after gap analysis and elimination of protein sequences, had at least 10 protein matches (see Table 21). Nearly half of the TIGRFams domains and almost 40% of Pfam domains contained at least one CPD region. Figure 4 compares the percent of domains containing CPDs for each database. Table 23 lists the number of regions found per database as well as the number of domains for each database containing one or more CPD region.

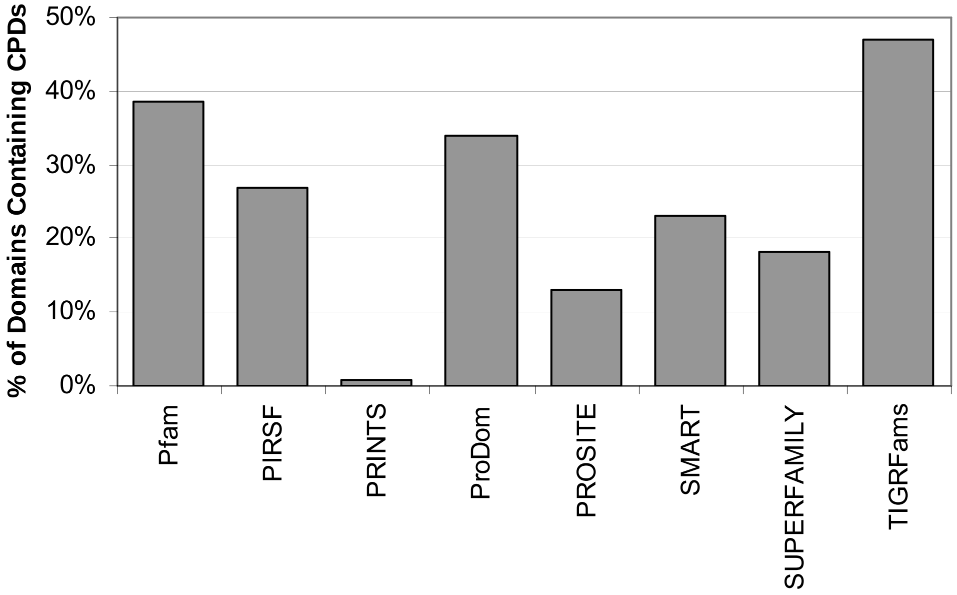

Figure 4. Percent of domains for each member database containing one or more CPD.

Table 23. CPD Regions per member database

Member Database	Number of CPD Regions	Number of Domains with CPD Regions	(% of qualified domains)
Pfam	2252	2036	(38.7%)
PIR Superfamily	74	62	(26.8%)
PRINTS	8	8	(0.8%)
ProDom	243	232	(34.0%)
PROSITE	167	172	(13.0%)
SMART	95	95	(23.2%)
SUPERFAMILY	69	68	(18.2%)
TIGRFams	745	726	(46.9%)
Total	3653	3403	(31.4%)

b. CPDs by Kingdom

Each domain was assigned to a kingdom (archaea, bacteria, eukaryota, viruses) based on the proteins that matched it. Those domains for which over 90% of the proteins with matching regions belonged to the same kingdom were assigned to that kingdom. Those domains for which no kingdom's proteins made up more than 90% of the matches were assigned to the kingdom 'Multiple', meaning the domain was present in proteins from more than one kingdom of life. The most common kingdom assignment was eukaryota, followed closely by Multiple. Archaea had the fewest domains assigned to it. Table 24 shows the number of domains assigned to each kingdom, as well as the number of domains assigned to each kingdom that had 10 or more protein matches.

Table 24. Domain Kingdom Assignments

Kingdom	Domains Assigned	Domains with 10+ Protein Matches
Archaea	270	125
Bacteria	3757	2930
Eukaryota	5553	3329
Viruses	1062	800
Multiple	4856	3618

There were CPD regions found in domains assigned each kingdom in the database. Table 25 shows the number of distinct domains in each kingdom that contained CPDs. The percent of domains is calculated out of the total number of distinct domains containing at least 10 protein matches, since those with fewer matches were not considered when searching for CPD regions.

Table 25. CPD Regions by Kingdom

Kingdom	Number of CPD Regions	Domains in Kingdom Containing CPDs (%)	Average CPD Effective Length
Archaea	53	50 (40.0%)	22.4
Bacteria	1174	1136 (38.8%)	22.6
Eukaryota	910	795 (23.9%)	25.9
Viruses	540	446 (55.9%)	27.2
Multiple	976	965 (26.7%)	22.6

More than half of the viral domains contained CPD regions. The CPD regions in viruses and eukaryotes were longer on average than other kingdoms.

Figure 5 shows the percent of CPDs assigned to each kingdom by different length classes. Only domains assigned to eukaryota and viruses had a significant proportion of long CPD regions.

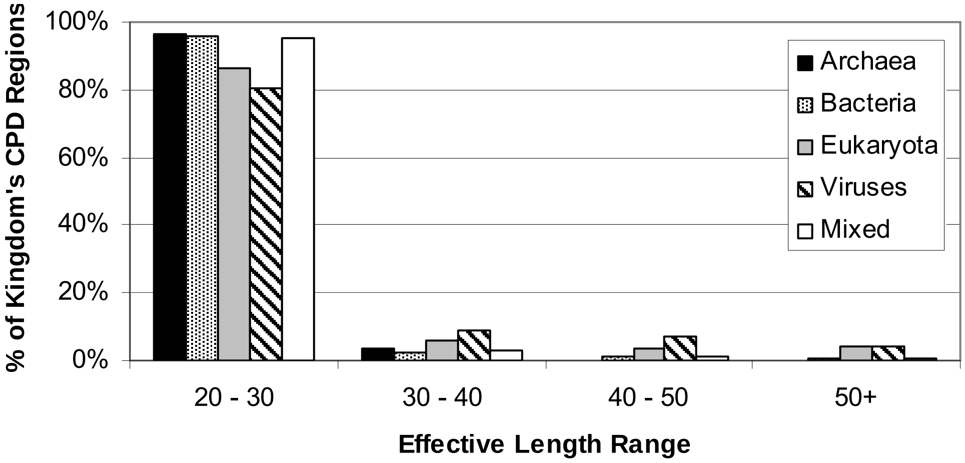

Figure 5. Histogram of CPD effective length classes by kingdom

Another way of looking at the prevalence of CPD regions in different kingdoms is by calculating the percent of proteins in each kingdom which contains at least one CPD region. As shown in Table 26, the percent of sequences containing a CPD of any length (noting that the minimum CPD length is 20) varies from about 19% for archaea to 37% for viruses, with eukaryota and bacteria falling in between. Similarly to the previous results, viruses and eukaryota had ten times more CPDs of length 50 or more than the other two kingdoms.

Table 26. Sequences with CPD Regions, by Kingdom

Kingdom	Percent of sequences with CPD regions			
	Any	Length ≥ 30	Length ≥ 40	Length ≥ 50
Archaea	18.8%	0.3%	0.1%	0.1%
Bacteria	30.0%	2.8%	2.1%	0.03%
Eukaryota	21.4%	3.7%	1.4%	1.3%
Viruses	36.9%	10.1%	6.7%	1.0%

c. CPD Length

The minimum required effective length of CPD regions was 20; the vast majority of CPDs (91%) had an effective length between 20 and 30. The largest effective length was 171, within the Dentin matrix 1 domain. Figure 6 shows a histogram of the effective length of CPD regions, for those CPD regions with an effective length between 20 and 40.

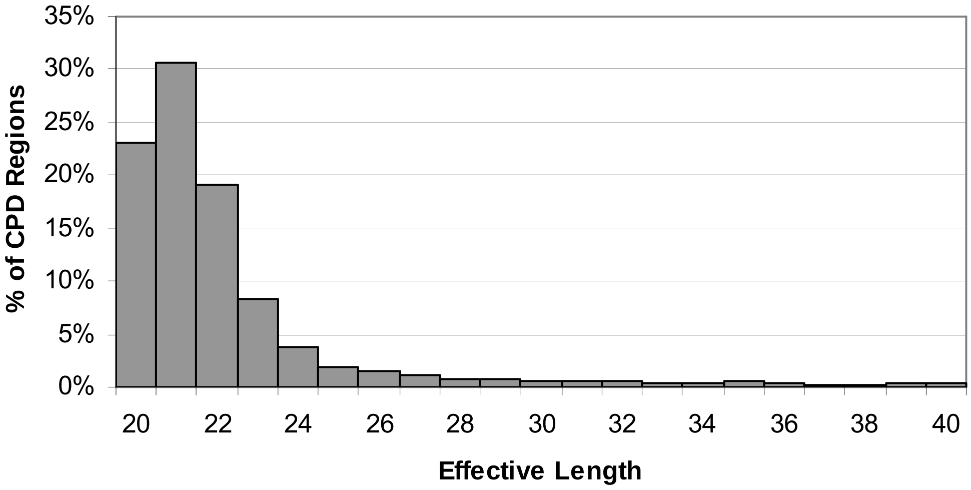

Figure 6. Length histogram of CPD regions.
There were 165 CPD regions with length greater than 40, not shown on histogram.

A relatively small number of CPD regions (less than 9%) exceeded 30 in length. Table

27 lists the number of regions at or above certain effective length thresholds.

Table 27. Long CPD Regions

Effective Length	Number of CPD regions	% of CPD regions
≥30	326	8.9%
≥40	168	4.6%
≥50	77	2.1%
≥60	39	1.1%

When the CPD effective length was taken as a fraction of the domain length, it showed that most CPD lengths were less than 15% of the domain. However, 316 (8.7%) of the CPDs covered more than half of the domain length, and 16 CPD regions covered the entire domain. Figure 7 shows a histogram of CPD effective length as a fraction of domain length.

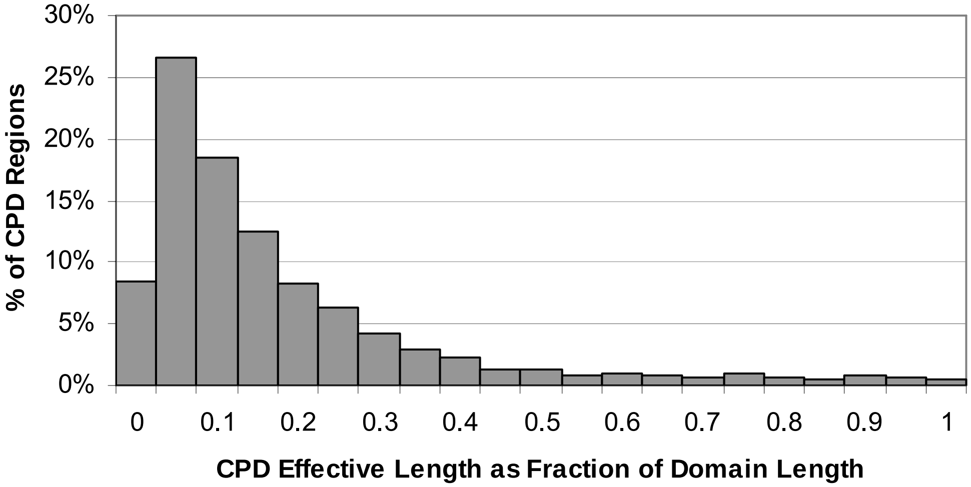

Figure 7. Histogram of CPD length as a fraction of domain length

d. Actual Percent Predicted Disorder

Although the minimum percent predicted disorder for a CPD region was 90%, most CPD regions had a much higher actual percent disorder. The percent of positions that were disordered within the region for all sequences was calculated as the actual percent disorder for the region. Almost 60% of the regions were 99% or more disordered, while only a few were actually 90% disordered. Figure 8 shows a histogram of the percent disorder for all CPD regions.

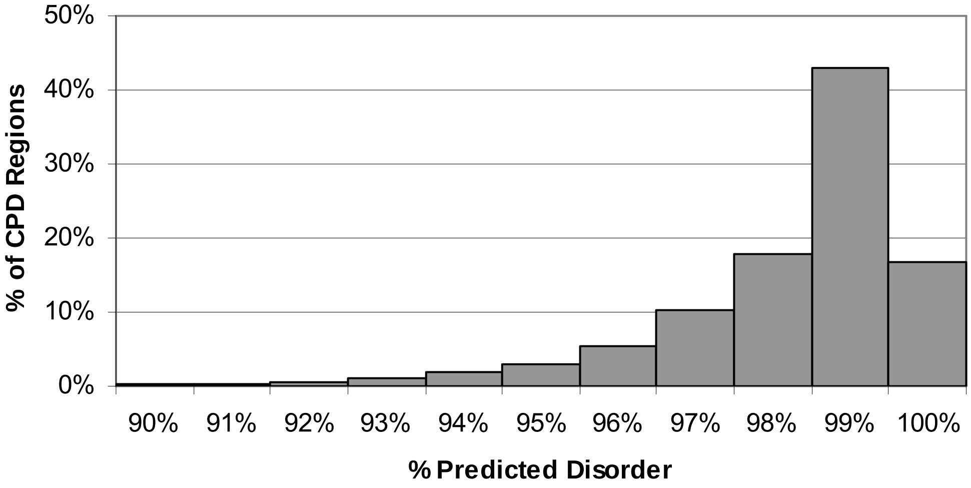

Figure 8. Histogram of percent predicted disorder of CPD regions.

e. Top CPD Regions

Tables 28 through 32 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 3D 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.

Table 28. Top 20 CPD Regions for Domains Predominantly in Eukaryota

	Accession	Domain Name	Region	Domain Description	3D[a]
1	PF07263	Dentin matrix 1	184-355	Transcription activation for osteoblast differentiation; regulation of dentin matrix formation; extracellular	
2	PF06495	Fruit fly transformer	1-137	RNA processing; alternative splicing of doublesex pre-mRNA; highly diverged [34]	
3	PF07263	Dentin matrix 1	20-127	Transcription activation for osteoblast differentiation; regulation of dentin matrix formation; extracellular	
4	PF05279	Aspartyl beta-hydroxylase, N-terminal	121-229	Domain at N-terminal of junctin, junctate and aspartyl beta-hydroxylase proteins, integral to endoplasmic reticulum membrane	
5	PF05334	Protein of unknown function DUF719	1-97	Domain within proteins of unknown function	
6	SM00157	Prion protein	36-122	Small glycoprotein; normal function unknown	M [35, 36]
7	PF04889	Cwf15/Cwc15 cell cycle control protein	97-182	Part of spliceosome, interacts with cdc15; function unknown	
8	PF05672	E-MAP-115	11-96	Microtubule-stabilizing protein; expressed mainly in epithelial cells	
9	TIGR01622	Splicing factor, CC1-like	40-120	RNA splicing factor; contains RNA recognition regions	
10	PF07169	Triadin	159-239	Domain within triadin protein; ryanodine receptor; calsequestrin binding protein	
11	PF02161	Progesterone receptor	182-257	N-terminal (modulatory) domain of progesterone receptor	
12	PIRSF002279	Notch	2103-2175	Transcription factor; membrane-bound; N-terminal half is extracellular, C-terminal half is intracellular	
13	PF01101, SM00527	High mobility group protein HMG14 and HMG17	1-69	Non-histone chromatin component; binds DNA	
14	PF04697	Pinin/SDK	1-69	N-terminal domain of proteins; may regulate protein-protein interactions	
15	PF02084	Bindin	3-70	Mediates species-specific adhesion of sperm to the egg surface during fertilization	
16	PF01034	Syndecan	243-306	Transmembrane heparan sulphate proteoglycans; may bind extracellular matrix components and growth factors	
17	PF06278	Protein of unknown function DUF1032	202-266	Unknown function	
18	PF05831	GAGE	35-99	Unknown function; implicated in human cancers	
19	PF05471	Podocalyxin	241-304	Membrane protein in glomerular epithelial cells; anti-adhesin function via charge repulsion	
20	PF05920	Coprinus cinereus mating-type protein	331-394	Regulates transcription of fungal mating genes	

[a] 'M' means region is missing in 3D structures.

Table 29. Top 20 CPD Regions for Domains Predominantly in Bacteria

	Accession	Domain Name	Region	Domain Description	3D[a]
1	PF04220	Protein of unknown function DUF414	1-88	Unknown function	
2	PF03217	Bacterial surface layer protein	1-56	S-layer precursor: forms layer that coats surface of bacteria	
3	PF07490	Translocated intimin receptor, N-terminal	1-51	Secreted by bacteria, embedded in target cell's membrane; facilitates bacterial attachment to host; binds host cytoskeletal proteins	
4	TIGR01071	Ribosomal protein L15, bacterial form	1-52	Part of large ribosomal subunit in bacteria	C [37]
5	PF04877	HrpZ	217-266	Binds to membranes, forms pore; may release nutrients or virulence factors; secreted	
6	PF05286	Fertility inhibition	1-49	Represses conjugated transfer of plasmids by binding RNA to block translation	x: 33-49 [38]
7	PF00700	Flagellin, C-terminal	1-49	C-terminal region of flagellin, which forms bacterial flagella	C, D [39-41]
8	PF06213	Cobalt chelatase, pCobT subunit	238-281	Subunit of aerobic cobalt cheltase; ATP-dependent formation of vitamin B12	
9	PF04156	IncA protein	1-44	Associated with homotypic fusion of inclusions of intracellular bacteria	
10	TIGR01105	UTP-glucose pyrophosphorylase, regulatory subunit	178-220	Non-catalytic subunit of UTP-glucose pyrophsphorylase; modulates enzyme activity	
11	PF04259, TIGR01442	Acid-soluble spore protein, gamma-type	1-40	Glutamine and asparagines-rich peptides used for storage of amino acids in spores of some bacteria	
12	PF07490	Translocated intimin receptor, N-terminal	190-231	Secreted by bacteria, embedded in target cell's membrane; facilitates bacterial attachment to host; binds host cytoskeletal proteins	
13	PIRSF-011502	Diol/glycerol dehydratase reactivating factor, large subunit	89-130	Alpha-subunit of complex that reactivates diol and glycerol dehydrogenase; functions like chaperone; interacts with ATP	C [42]
14	PIRSF-005259	Tripartite hybrid signal transduction histidine kinase, BarA type	266-307	Part of bacterial two-component signal transduction system	
15	PD004231, PF01649	Ribosomal protein S20	1-38	Part of small ribosomal subunit	C [43]
16	PD028235	Exonuclease VII, small subunit	1-38	Part of enzyme that catalyzes exonucleolytic cleavage	C (PDB: 1VP7)
17	PF03434	Protein of unknown function DUF276	1-40	Unknown function; encoded on extrachromosomal DNA	
18	TIGR01348	Dihydrolipoamide acetyltransferase	239-277	Part of pyruvate dehydrogenase complex, contains E3 binding region	
19	TIGR00060	Ribosomal protein L18	1-38	Part of large ribosomal subunit	m: 1-23, n: 24-38 [44]
20	PIRSF-018507	Propanediol/glycerol dehydratase, large subunit	439-476	Alpha subunit of holoenzyme that catalyzes diol/glycerol to deoxy aldehyde	C [45]

[a] 'C' means region has a known 3D structure while in a complex, 'D' means region shown to be disordered, 'm' means part of region missing in 3D structure, 'n' means part of region has a 3D structure observed via NMR spectroscopy, 'X' means region has a 3D structure observed via X-ray crystallography, 'x' means part of region has a 3D structure observed via X-ray crystallography.

Table 30. Top 20 CPD Regions for Domains Predominantly in Viruses

	Accession	Domain Name	Region	Domain Description	3D [a]
1	PF05750	Rubella capsid	1-144	Capsid protein of Rubella virus; interacts with RNA during virus assembly	
2	PF05749	Rubella membrane glycoprotein E2	1-144	Contained in lipid bilayer enveloping capsid; forms heterodimer with E1; directs membrane fusion during infection process	
3	PF06595	Borna disease virus P24	1-127	Unknown function	
4	PF05505	Ebola nucleoprotein	547-642	Encapsidates genomic RNA in Ebola and Marburg viruses	
5	PF02993	Minor capsid protein VI	138-229	Part of viral capsid, may link DNA-protein core to external capsid	
6	PF04162	Circovirus coat protein	1-76	Viral capsid component in circovirus	
7	PF06193	Orthopoxvirus A5L	163-244	Immunodominant late protein; found in virion core; required for virus maturation	
8	PF03276	Spumavirus gag protein	437-512	Core viral protein; cleaved to yield mature protein	
9	PD004155, PF03012	Phosphoprotein	1-69	Component of viral polymerases	
10	PD001068, PF00894	Luteovirus group 1 coat protein	1-65	Viral capsid structural protein, N-terminus possibly involved in protein-RNA interaction	
11	PF03012, PD004155	Phosphoprotein	155-214	Component of viral polymerases for Rhabdoviruses	x: 186-214 [46]
12	PF03014	Hepatitis E virus structural protein 2	79-137	May be involved in genomic RNA encapsidation; highly basic	
13	PF02443	Circovirus ORF-2 protein	1-59	Viral protein of unknown function	
14	PF00944	Peptidase S3, togavirin	65-121	Viral endopeptidase domain, serine-type; part of polyprotein; homo-dimer	X [47]
15	PF00513	Late protein L2	76-133	Viral capsid structural protein	
16	PF03910	Adenovirus minor core protein PV	156-211	Associates with nucleoli in cells infected by adenovirus	
17	PF04861	Circovirus VP2 protein	1-56	May be non-structural and required for virus assembly	
18	PF01516	VP6 blue-tongue virus inner capsid protein	171-227	Inner capsid protein that surrounds and possibly binds genomic viral RNA	
19	PF01516	VP6 blue-tongue virus inner capsid protein	76-131	Inner capsid protein that surrounds and possibly binds genomic viral RNA	
20	PF00761	Polyomavirus coat protein VP2	32-83	Internal coat protein	

[a] ‘x’ means part of region has a 3D structure observed via X-ray crystallography

Table 31. Top 20 CPD Regions for Domains from Multiple Kingdoms

	Accession	Domain Name	Region	Domain Description	3D [a]
1	PF01280, PD004823, SSF48140	Ribosomal protein L19e	53-124	Part of the large ribosomal subunit, in Eukaryota and Archaea	C [48]
2	TIGR00307, PD005658, PF01201	Ribosomal protein S8E	1-68	Part of the small ribosomal subunit, in Eukaryota and Archaea	
3	SSF46950	DNA-binding TFAR19-related protein	1-56	Ubiquitous, may be involved in apoptosis; in Eukaryota and Archaea	X [49]
4	PF05800	Gas vesicle synthesis	1-55	Required for gas vesicle synthesis; in Archaea and Bacteria	
5	PD005148	Ribosomal protein L34e, C-terminal	1-45	C-terminal end of part of large ribosomal subunit; in Eukaryota and Archaea	
6	PF05790	T-cell surface antigen CD2	285-335	Mediates T-cell adhesion; has cytoplasmic tail; family includes viral homologues	
7	PF05076	Suppressor of fused	307-350	Tumor suppressor; interacts with proteins to suppress; mainly in Eukaryota, homologues in several Bacteria	
8	PF01199	Ribosomal protein L34e	2-45	Part of large ribosomal subunit in Eukaryota and Archaea	
9	TIGR01305	Guanosine monophosphate reductase 1	1-41	Catalyzes reductive deamination of GMP→IMP, in Eukaryota and Bacteria	
10	PF00468, TIGR01030	Ribosomal protein L34	1-42	Part of the large ribosomal subunit, in Eukaryota and Bacteria	C [50]
11	PF01984	DNA-binding TFAR19-related protein	1-39	Ubiquitous, may be involved in apoptosis; in Eukaryota and Archaea	
12	TIGR01632	Ribosomal protein L11, bacterial	1-40	Part of the large ribosomal subunit, in Eukaryota and Bacteria	C [50] D [51]
13	PD003823, PF01655	Ribosomal protein L32e	1-39	Part of the large ribosomal subunit, in Eukaryota and Archaea	C [48]
14	TIGR00982	Ribosomal protein S23, eukaryotic and archaeal form	1-40	Part of the small ribosomal subunit, in Eukaryota and Archaea	C [52]
15	PF06992	Replication P	185-223	Promotes replication of phase chromosome; delivers DnaB helicase to replication origin	
16	PF05766	Bacteriophage Lambda NinG	53-91	Involved in recombination, in Viruses and Bacteria	
17	SSF69369	Colicin E3, translocation	1-38	N-terminal of plasma-encoded antibiotic protein; mediates translocation across membrane	C [53]
18	PD011777, PF01941	Archaeal S-adenosylmethionine synthetase (MAT)	1-37	Catalyzes synthesis of S-adenosylmethionine, binds ATP, in Archaea and Bacteria	
19	TIGR00523, PF01176	Eukaryotic initiation factor 1A (eIF-1A)	1-37	Helps ribosome dissociation, stabilizes binding of Met-tRNA to ribosome, in Eukaryota and Archaea	d: 1-28 [54] c: 15-37 [55]
20	PD001367	Ribosomal protein L11	1-39	Part of the large ribosomal subunit, in Eukaryota and Archaea	C [50] D [51]

[a] ‘C’ means region has a known 3D structure while in a complex, ‘c’ means part of region has a known 3D structure while in a complex, ‘d’ means part of region shown to be disordered, ‘m’ means part of region missing in 3D structure, ‘X’ means region has a 3D structure observed via X-ray crystallography.

Table 32. Top 20 CPD Regions for Domains Predominantly in Archaea

	Accession	Domain Name	Region	Domain Description	3D [a]
1	TIGR00354, PF03833	DNA polymerase II large subunit DP2	50-83	Part of DNA polymerase complex; binds DNA	
2	TIGR00264	Alpha-NAC-related protein	1-27	Hypothetical protein	
3	PD008669	DNA topoisomerase, type II	1-24	Archaeal topoisomerase	
4	PF05854	Non-histone chromosomal MC1	48-72	Protects DNA against thermal denaturation	X [56]
5	PF05942	Archaeal PaREP1	1-22	Function unknown	
6	PF05457	Sulfolobus transposase	3-26	Function unclear; no evidence for transposase activity	
7	PD012512	Protein of unknown function DUF211	1-23	Function unknown	
8	TIGR01043	ATP synthase A-type, A subunit	2-24	One of 10 subunits of Archaeal A-ATPase	
9	TIGR00489	Translation elongation factor aEF-1 beta	1-23	Involved in exchange of GDP for GTP on other subunit	N [57]
10	PIRSF003137	Methyl-coenzyme M reductase operon protein C	81-103	Part of complex for methanogenesis; exact function of protein C in complex unknown	
11	TIGR01546	Glyceraldehyde-3-phosphate dehydrogenase, type II	1-23	Catalyzes central step in glycolysis and glucogenesis	X [58]
12	TIGR00354, PF03833	DNA polymerase II large subunit DP2	1-22	Part of DNA polymerase complex; binds DNA	
13	TIGR01920	Shikimate kinase, archaea	1-22	Phosphorylates shikimate; part of aromatic amino acid biosynthesis	
14	PD016182	Protein of unknown function UPF0179	1-20	Unknown function	
15	TIGR00748	Putative condensing enzyme FabH-related	1-22	May be hydroymethylglutaryl-CoA synthase	
16	TIGR01506	Riboflavin synthase, archaeal	1-22	Catalyzes final step in riboflavin biosynthesis	
17	PF01908	Protein of unknown function DUF75	1-22	Unknown function, maybe transmembrane	
18	PD005228, PF02505	Methyl-coenzyme M reductase, protein D	1-22	Part of complex for methanogenesis; exact function of protein D in complex unknown	
19	TIGR00144	GHMP kinase family group 1	1-22	methanopterin biosynthesis; N-terminal may bind ATP	
20	PF05854	Non-histone chromosomal MC1	1-22	Protects DNA against thermal denaturation	X [56]

[a] 'X' means region has a 3D structure observed via X-ray crystallography, 'N' means region has a 3D structure observed via NMR spectroscopy

Each of the domains in the previous tables was classified according to their known functions, based on the InterPro abstract. Table 33 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)’.

Table 33. Functions of domains containing CPDs, by kingdom

Function	Number of Domains Containing CPDs with Function [a]					
	Archaea	Bacteria	Eukaryota	Viruses	Multiple	Total
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				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 27-31) were counted

f. 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. Table 34 summarizes the results of PDB matches to CPD regions.

Table 34. CPD Regions with PDB Matches

CPDs overlapping with PDB entries	1,338	(36.6%)
in complex	564	(15.4%)
alone	774	(21.2%)

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. Table 35 shows the percent of positions in CPDs that overlapped with a known structure in the PDB.

Table 35. CPD Positions Overlapping PDB Sequence Positions

CPD Length	Positions overlapping with PDB structures		
	Any	In Complex	Alone
≥20	33.0%	14.3%	18.7%
≥30	19.0%	13.7%	7.7%
≥40	13.0%	8.7%	4.3%
≥50	13.4%	8.8%	4.6%

Figure 9 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.

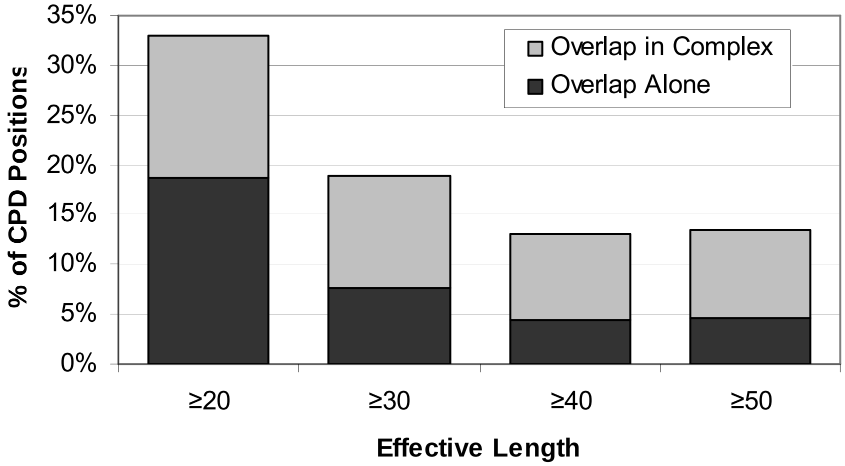

Figure 9. Histogram of PDB matches to CPD positions, by effective length

Table 36 breaks down the PDB matches 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. Figure 10 shows a histogram of PDB matches for CPD positions by kingdom.

Table 36. CPD Positions Overlapping PDB Sequence Positions, by Kingdom

Kingdom	Positions overlapping with PDB structures		
	Any	**In Complex**	**Alone**
Archaea	18.5%	10.8%	7.7%
Bacteria	34.6%	15.3%	19.3%
Eukaryota	23.7%	8.9%	14.8%
Viruses	11.9%	4.4%	7.5%
Multiple	56.3%	25.9%	30.4%

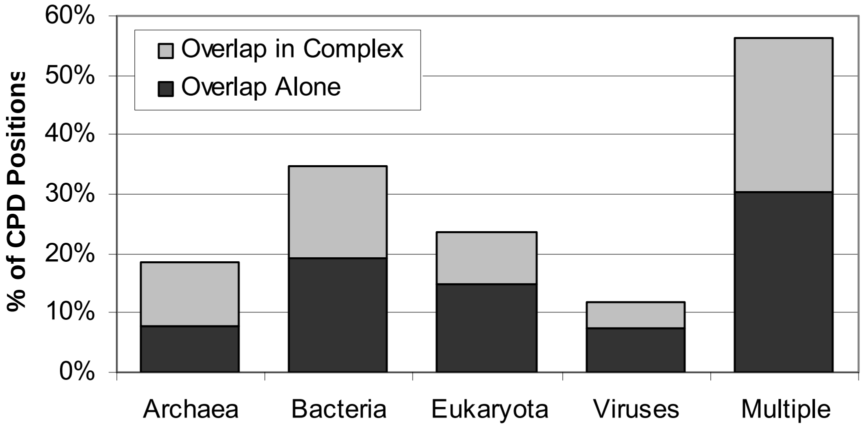

Figure 10. Histogram of PDB matches to CPD positions, by kingdom

The different member database had widely varying percentages of CPD positions matched to PDB sequences. Pfam had the fewest CPD positions with overlapping known 3D 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 11).

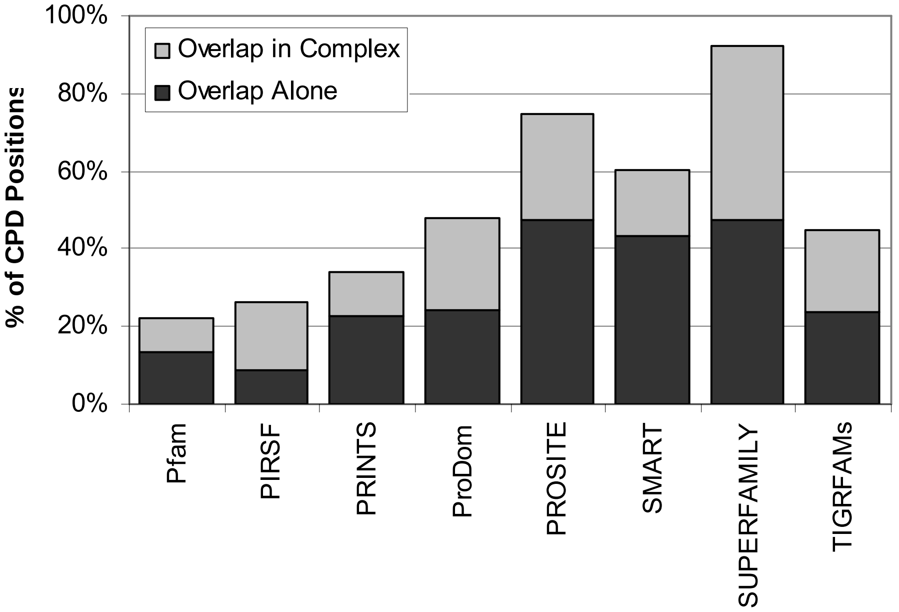

Figure 11. Histogram of PDB matches to CPD positions, by member database

5. *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 as stated in Tables 20-24.

Dentin Matrix 1 (Eukaryota): Related proteins bone sialoprotein and osteopontin were found to be completely disordered by NMR spectroscopy [59]. Based on sequence similarities, dentin matrix 1 is expected to be mostly or entirely disordered as well [59]. 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 12).

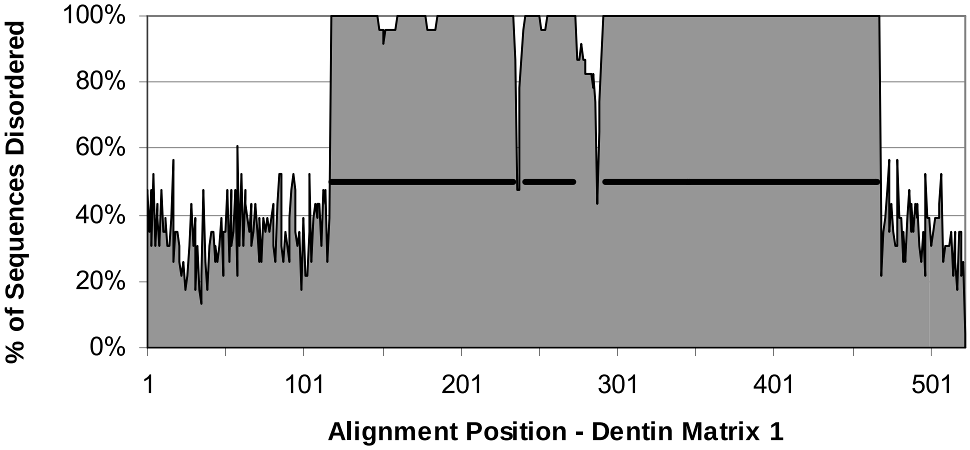

Figure 12. Graph of disorder prediction for dentin matrix 1 protein family. The CPD regions are shown as horizontal black lines.

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 13).

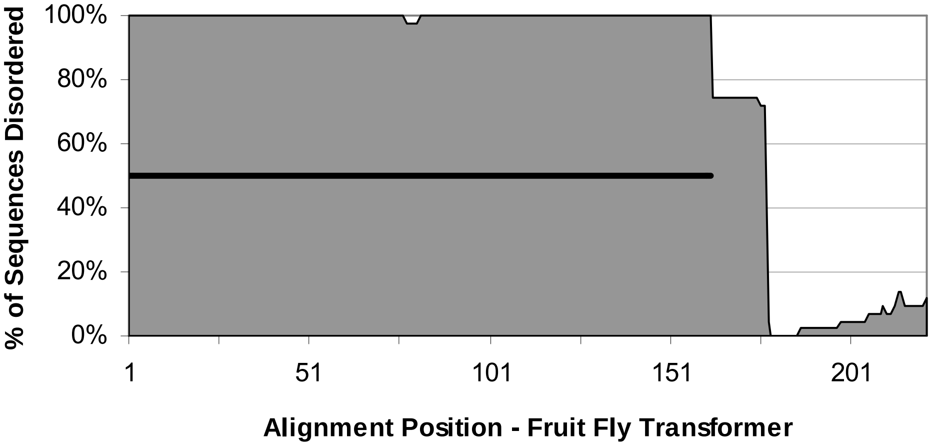

Figure 13. Graph of disorder prediction for fruit fly transformer family. The CPD region is shown as a black horizontal line.

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 [60]. Both BAH and humbug have an apparent molecular weight roughly two times larger their actual molecular weight [60]. Figure 14 shows the location of the CPD region in the alignment.

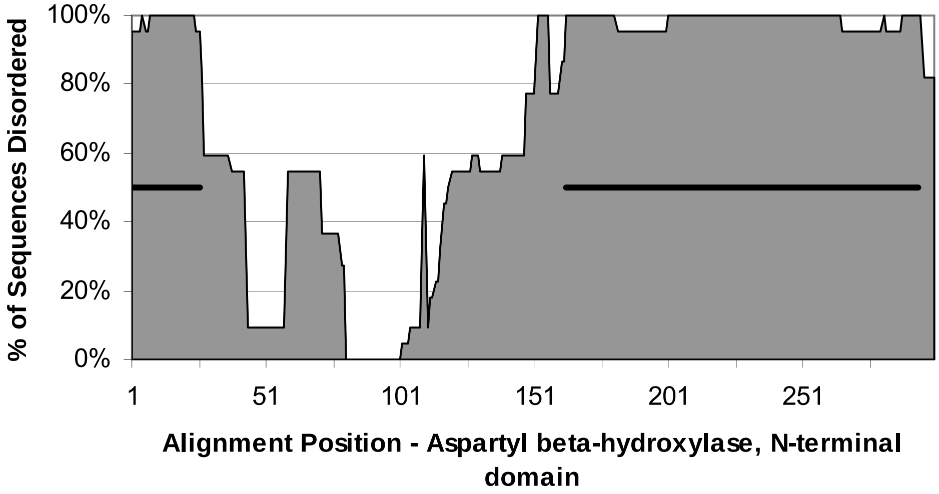

Figure 14. Graph of disorder prediction for N-terminal domain of aspartyl β-hydroxylase. The CPD regions are shown as black horizontal lines.

Prion (Eukaryota): Residues 1-124 of the human prion protein were found to be disordered by NMR [35]. 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 [^{1}H]-^{15}N nuclear Overhauser effect [61]. NMR studies of still other prion proteins also showed the N-terminal tail (roughly 100 residues) was disordered [36]. The CPD region found in the prion family extends on average from residues 36 to 122 of the protein. Figure 15 shows the location of the CPD region in the alignment.

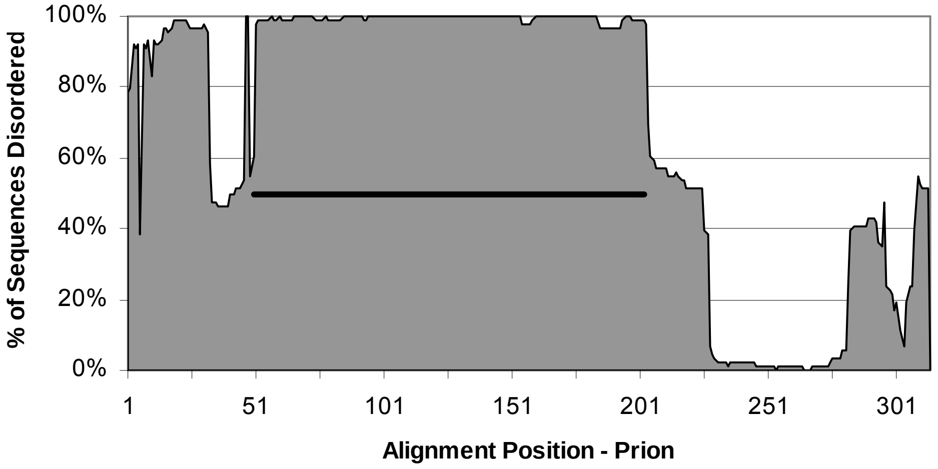

**Figure 15. Graph of disorder prediction for prion family.
The CPD region is shown as black horizontal line.**

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 [62]. The protein is regulated by phosphorylation [63]. Most of E-MAP-115 is predicted to contain a conserved disordered region (Figure 16).

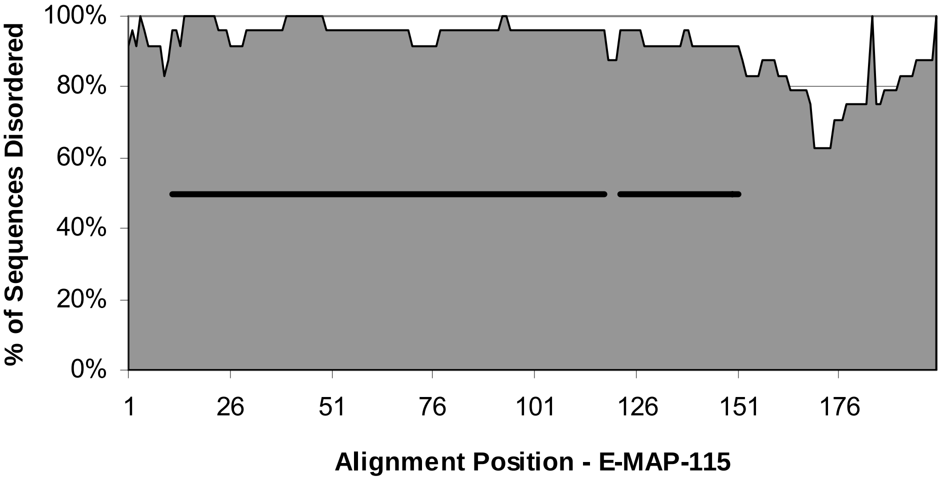

**Figure 16. Graph of disorder prediction for E-MAP-115 family.
The CPD regions are shown as black horizontal lines.**

Bacterial surface layer protein (Bacteria): The N-terminal region of this protein, which contains the CPD region (Figure 17), binds to a secondary cell wall polymer [64].

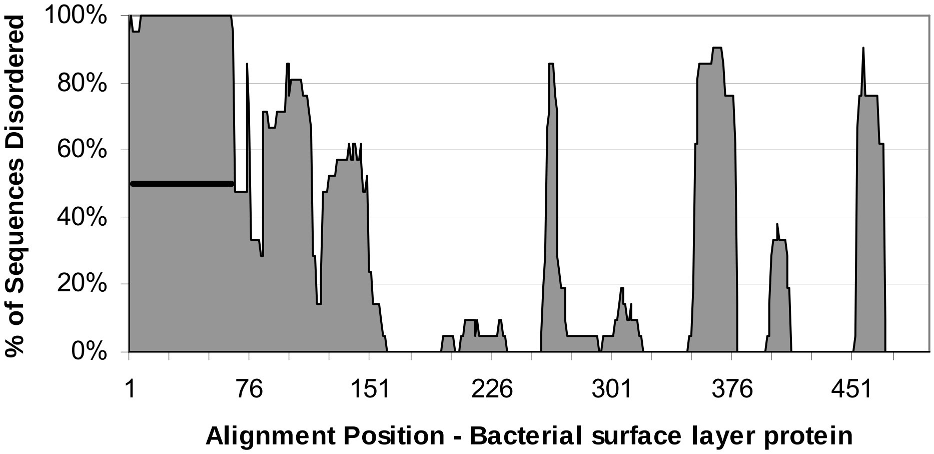

Figure 17. Graph of disorder prediction for bacterial surface layer protein family. The CPD region is shown as a black horizontal line.

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 [65]. However, this segment of Tir is not within the N-terminal domain that contains CPD regions (Figure 18). 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 [66]. 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 [66]. Additionally, the Tir protein has a higher apparent molecular weight than expected [67].

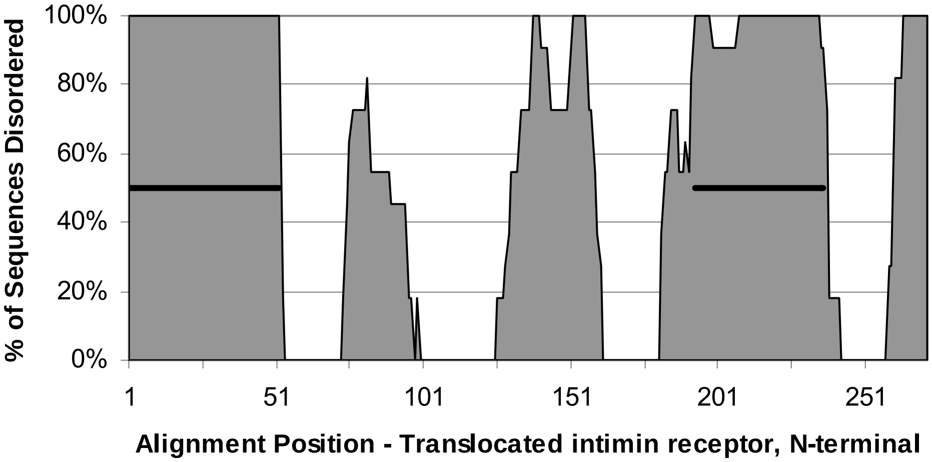

Figure 18. Graph of disorder prediction for the N-terminal of the translocated intimin receptor family.
The CPD regions are shown as black horizontal lines.

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” [68]. 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” [68]. The extended region of L15 covers residues 1 through 60 according to one model [48]. Figure 19 shows the location of the CPD region in the alignment.

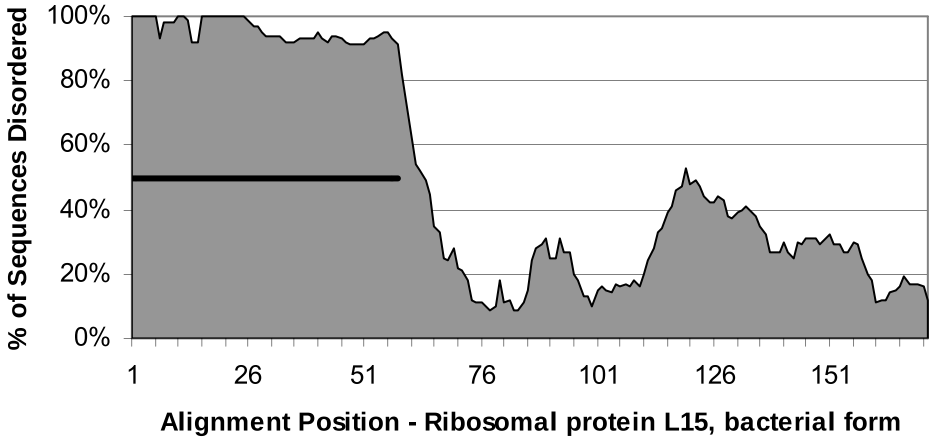

Figure 19. Graph of disorder prediction for Ribosomal protein L15 bacterial form. The CPD regions are shown as black horizontal lines.

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 [69]. The CPD regions found in this work correspond approximately with these N- and C-terminal regions (Figure 20).

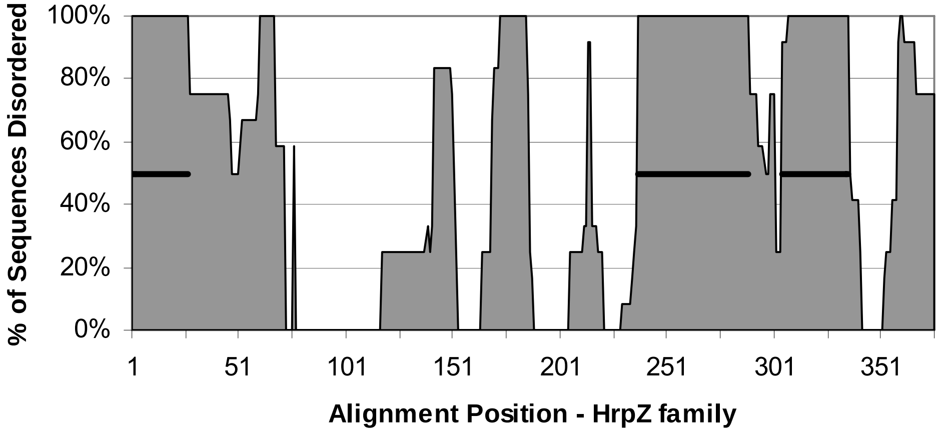

Figure 20. Graph of disorder prediction for the HrpZ family. The CPD regions are shown as black horizontal lines.

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 [38]. When FinO was exposed to trypsin through limited proteolysis, the fragment 62-170 showed to be protease resistant [70]. The predicted disordered region extends from residues 1 to 49 (Figure 21).

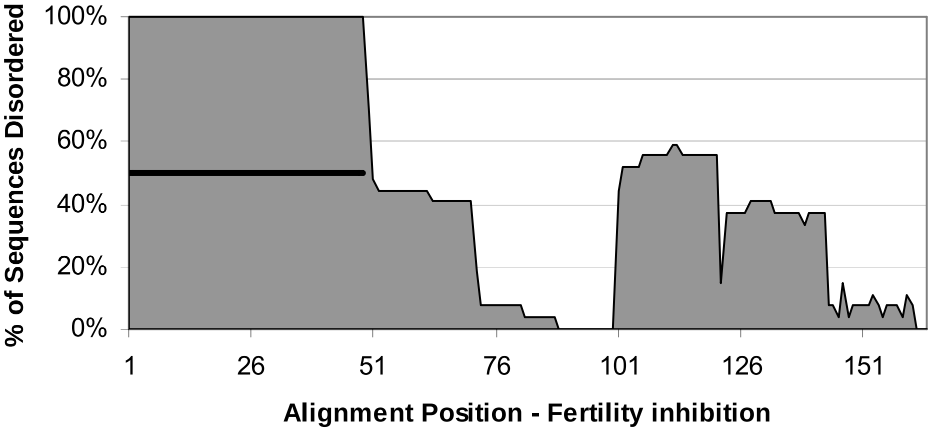

Figure 21. Graph of disorder prediction for the fertility inhibition protein family. The CPD region is shown as a black horizontal line.

Rubella capsid (Viruses): No evidence of order or disorder for this protein was found.

Rubella membrane glycoprotein E2 (Viruses): No evidence of order or disorder for this protein was found.

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 [71]. The graph of the disorder prediction for the family's alignment is shown in Figure 22.

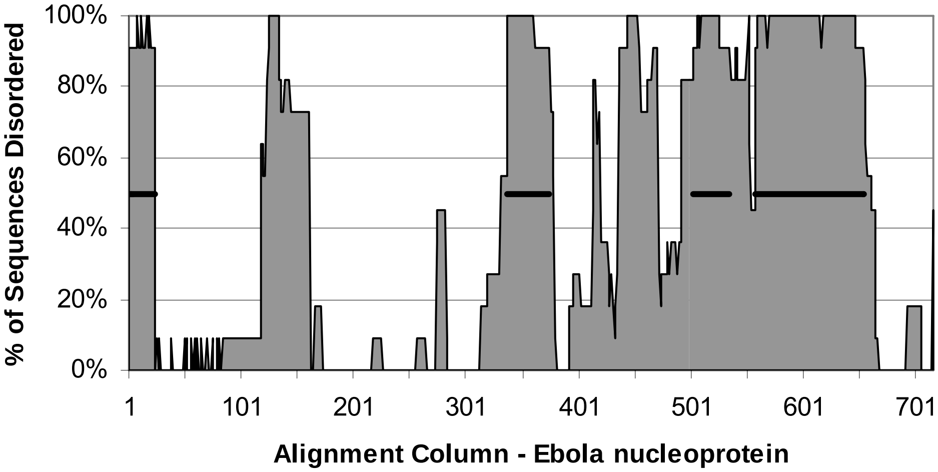

Figure 22. Graph of disorder prediction for the Ebola nucleoprotein. The CPD regions are shown as black horizontal lines.

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 [72]. In a mature virus, protein VI is thought to form a trimer of dimers [73]. There are two predicted CPD regions in this protein (Figure 23).

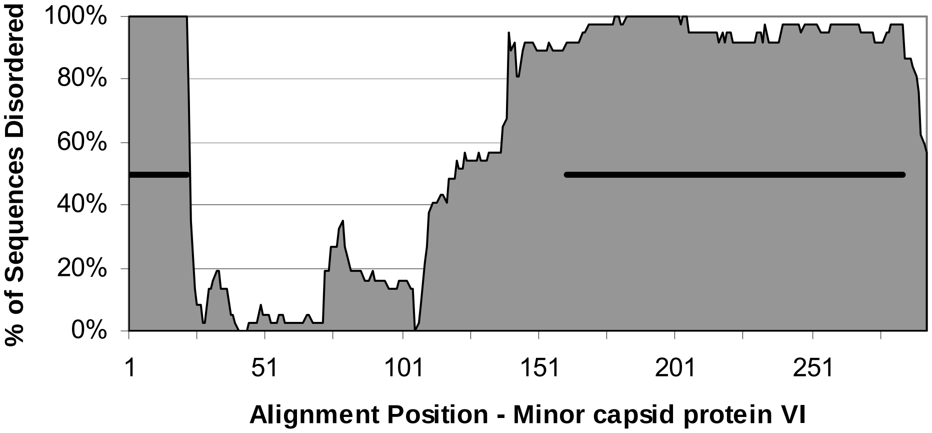

Figure 23. Graph of disorder prediction for minor capsid protein VI. The CPD regions are shown as black horizontal lines.

Circovirus coat protein (Viruses): No evidence of order or disorder for this protein was found.

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 [48].

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.

Gas vesicle synthesis (Archaea and Bacteria): No evidence of order or disorder for this protein was found.

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 [48].

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 [74, 75]. The small peptide from residues 294 to 303, which is proline rich, was visualized in complex with a binding partner [76]. This small region is within the long CPD region from 285-335 (Figure 24).

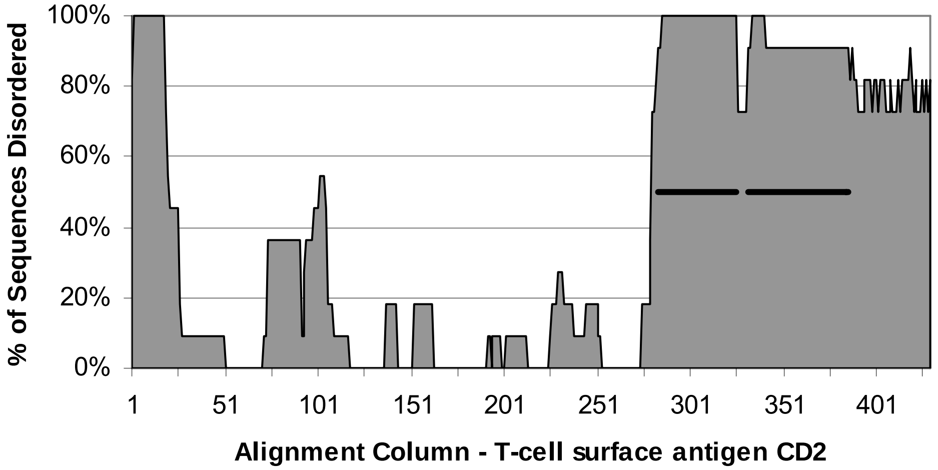

Figure 24. Graph of disorder prediction for T-cell surface antigen CD2. The CPD regions are shown as black horizontal lines.

DNA polymerase II large subunit DP2 (Archaea): No evidence of order or disorder for this protein was found.

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 [77]. This does not overlap with the CPD region (Figure 25).

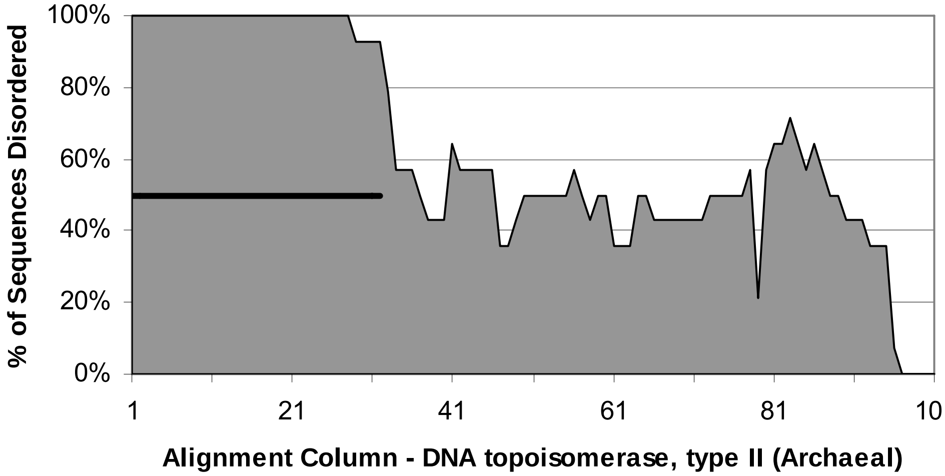

Figure 25. Graph of disorder prediction for DNA topoisomerase, type II. The CPD region is shown as a black horizontal line.

ATP synthase A-type, A subunit (Archaea): No evidence of order or disorder for this protein was found.

Methyl-coenzyme M reductase operon protein C (Archaea): No evidence of order or disorder for this protein was found.

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 [78, 79].

C. Sequence Conservation

Shannon's entropy was calculated for all alignment columns as a measure of sequence conservation. On average, disordered alignment columns (those with 90% or more sequences disordered at that position) had a higher entropy value than ordered alignment columns (those with 90% or more sequences predicted to be ordered at that position). However, when just the alignment columns that were part of either a CPD or CPO region were taken, the entropy values were roughly equal. Table 37 shows the average entropy values. Figure 26 shows a histogram of the entropy values for disordered and ordered alignment columns, and Figure 27 shows the histogram for alignment columns in CPD and CPO regions.

Table 37. Average Shannon's entropy for alignment columns

	90%+ Disordered	**90%+ Ordered**
All alignment columns	0.38	0.33
Alignment columns in CPD or CPO regions	0.34	0.33

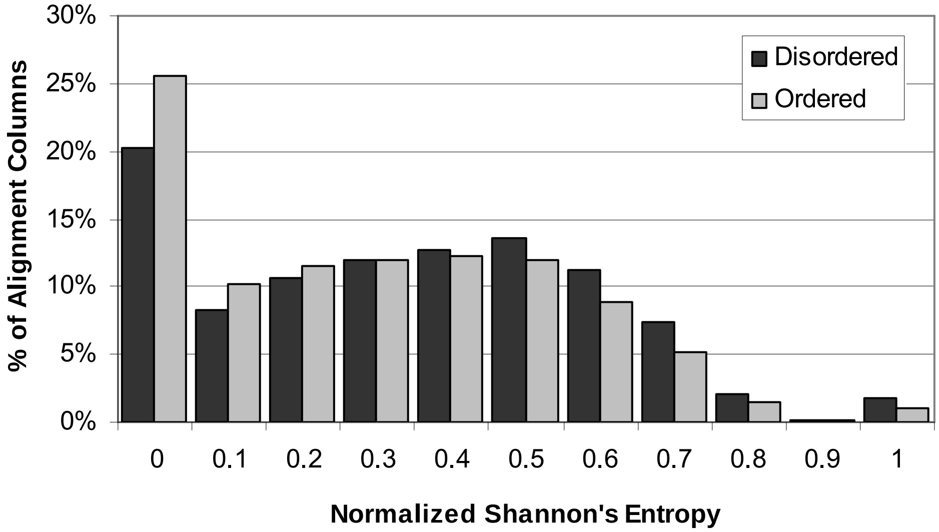

Figure 26. Histogram of Shannon's Entropy for disordered and ordered alignment columns

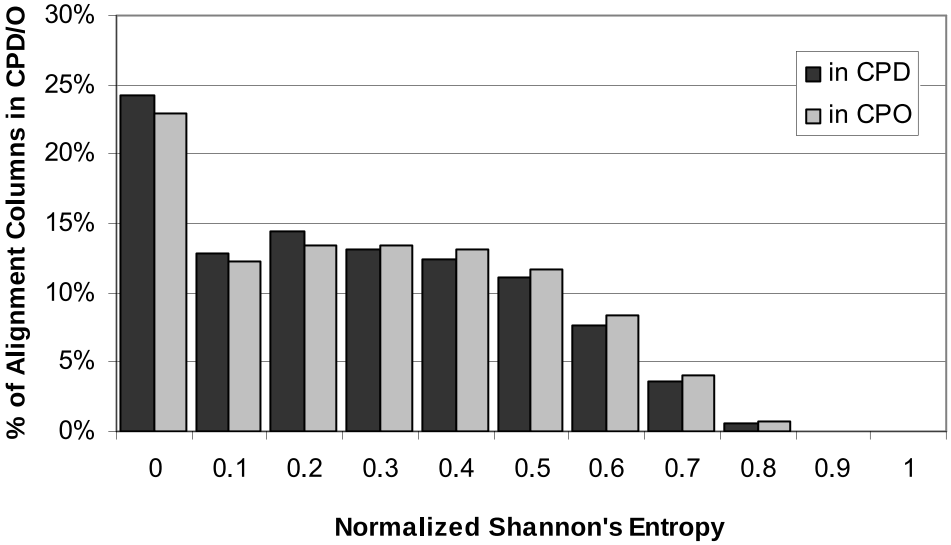

Figure 27. Histogram of Shannon's entropy for alignment columns that were part of CPD or CPO regions

The average Shannon's entropy value was calculated for all CPD and CPO regions. The average of these values for CPD regions was 0.35 with a standard deviation of 0.15, and for CPO regions was 0.34 with a standard deviation of 0.14. A histogram of the average Shannon's entropy values for CPD and CPO regions is shown in Figure 28.

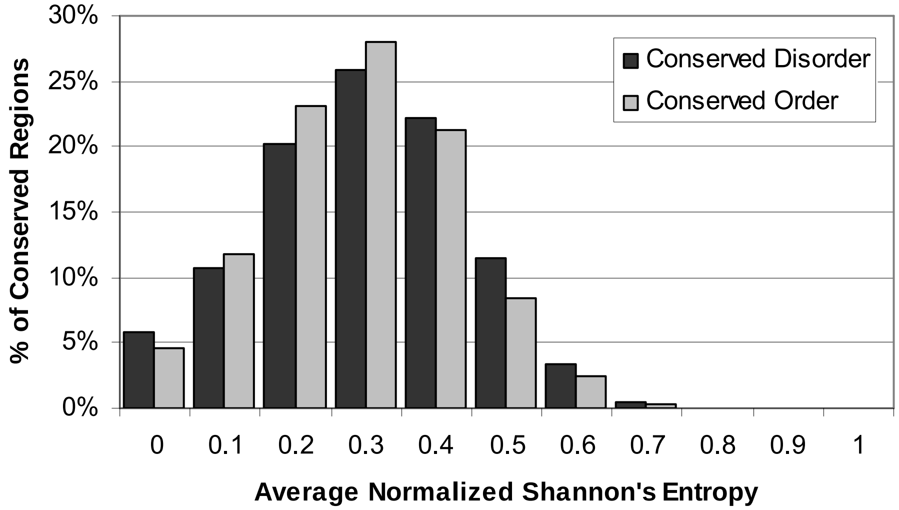

Figure 28. Histogram of average Shannon's Entropy for all CPD and CPO regions

There were 1,511 different domains containing both CPD regions and CPO regions. The average CPD entropy and the average CPO entropy were calculated for all of these domains. On average, the CPD entropy was slightly higher than the CPO entropy (with the average difference between CPD entropy and CPO entropy for domains being 0.015 with a standard deviation of 0.10).

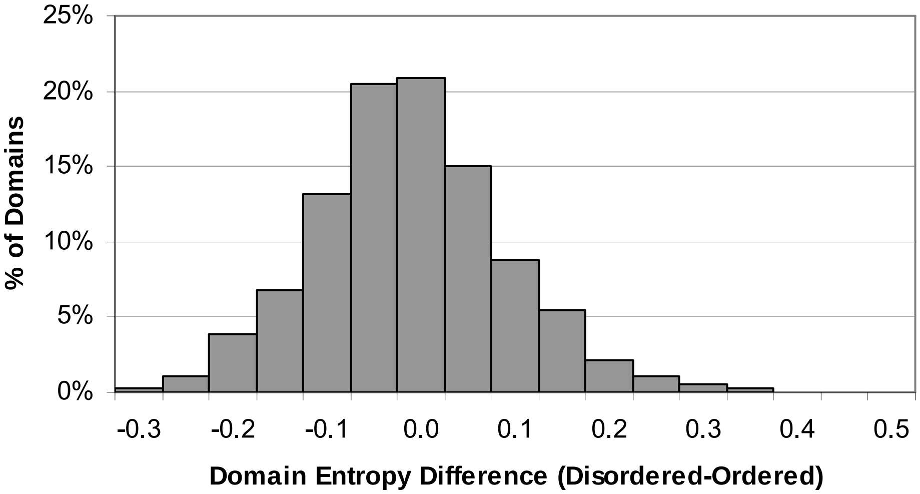

Figure 29. Histogram of difference between average CPD and CPO entropy for domains containing both types of conserved regions

However, the histogram of the difference between the two entropy values (Figure 29) shows that for some domains, the disordered regions are more conserved in terms of amino acid sequence than the ordered regions.

V. Discussion

A. Prevalence and Characteristics of Conserved Disorder

Regions of conserved disorder prediction were found in protein domains from all available InterPro member databases. The percent of domains from each member database containing one or more CPDs varied from 0.8% to 47%. The PRINTS database had CPD regions in only 0.8% of its domains. This is due to the short length of most PRINTS region matches. With the average PRINTS domain match length of 17, it would be impossible for many regions to contain a CPD region of 20 or longer.

In the TIGRFams database, 47% of its domains contained CPD regions. The TIGRFams database is built by grouping proteins into clusters of orthologous groups, which implies a similar function across the members. This may be why such a high percentage of domains contained CPD regions: if the function is conserved across protein members, and if the disorder is necessary for the function, then the disorder will be conserved. In contrast, databases which group more distant family members together, which may have diverged in function, may be more likely to have domains in which disorder tendencies are not conserved across all members of the family.

The Pfam and ProDom databases also had fairly high percentages of domains containing CPDs, at 39% and 34%, respectively. These databases build families in different ways than TIGRFams, in that they do not cluster orthologues. Although these databases supposedly classify domains and families of similar function, it may be that they are including more distant relatives, leading to slightly less conservation of disorder. The fact that TIGRFAMs domains have on average less than half as many protein members as Pfam and ProDom, indicating a more exclusive family membership, lends support to this theory.

It is surprising that 18% of SUPERFAMILY's domains contained CPD regions, given that the database is derived from proteins of known structure. As expected, nearly all of its CPD regions had a known structure, and 50% had a known structure alone. However, only two CPDs (0.5%) derived from SUPERFAMILY had length 30 or greater and had a known 3D structure not in a complex. The implications of this for the accuracy of this work with respect to shorter regions of conserved disorder will be discussed in section V.C.

Regions of conserved disorder prediction were found in all kingdoms of life, including viruses. When considering CPD regions of all lengths, viruses have the greatest proportion of proteins containing conserved disorder, and archaea have the least. However, when only long CPD regions are considered, viruses and eukaryota have far more conserved disorder (roughly 1% of proteins) than archaea and bacteria (0.1% of proteins). This finding is in line with previous work [1, 29], showing that eukaryotic proteomes have a higher long disorder (>50) content than bacterial and archaeal proteomes. In both this work and previous work, eukaryotes had on the order of ten times more proteins containing long disordered regions than did archaea and bacteria. The fact that viruses also have higher disorder content has not previously been reported, so cannot be verified by comparison to earlier work.

The difference between this and earlier work is that the percent of domains containing conserved disordered regions per kingdom is roughly a factor of ten less than the percent of proteins containing long disordered regions. One interpretation for this is that many disordered regions are not within regions of sequence conservation. A certain level of sequence conservation is required for membership in a protein domain or family, since the sequence is what is used to identify the domain signature. It is very likely that there are other regions of conserved disorder which are not in regions of sequence conservation, which would therefore not be detected by the methodology used for this project.

Most regions of conserved predicted disorder detected were short. The previous work on disorder prediction done using PONDR® VL-XT has focused on long (>30 residues) disordered regions. This was done because the VL-XT predictor is more accurate over longer stretches of continuous disorder. Although short disordered regions are known to exist in nature, they have a different amino acid composition than long disordered regions, and so are difficult to predict with current long disorder predictors [25]. This work focused on regions of disorder that were are least 20 residues long. Although the error rate for PONDR® VL-XT predictions at this length is higher than for longer stretches of disorder, it was thought that the methodology used to find conserved disorder, using multiple sequence alignments, might improve the accuracy compared to predicting short disorder regions in a single sequence. That is, a prediction of disorder might be more likely to be correct if that prediction was found in the same region in nearly all family members.

Although most regions of conserved disorder prediction were short, there were a significant number than exceeded 30 in length. Most of these were in domains from eukaryotic or viral proteins. As explained above, this is in line with previous work, which found that long regions of intrinsic disorder were much more prevalent in eukaryotes than in prokaryotes.

Sequence conservation in regions of conserved disorder varied, but was on average slightly lower than in regions of conserved order. Positions within CPD regions were just slightly less conserved, based on Shannon's entropy, than positions within regions of conserved order. When all alignment columns were considered, not just those in conserved order/disorder regions, regions of disorder were noticeably less conserved than regions of order. That is, among positions in the alignments that were 90% or more disordered, those within CPD regions were more conserved than those that weren't. This indicates that these CPD regions are important for some function of the protein, since both the sequences and the disorder tendencies are conserved.

When the average Shannon's entropy was taken for regions of conserved order and disorder within a single domain, it was seen that for most domains, regions of conserved order and disorder were about equally conserved in sequence. However, 25% of domains had a higher degree of sequence conservation within conserved disorder as compared to conserved order. Roughly 18% of domains had a lower degree of sequence conservation within conserved disorder than within conserved order. In previous work studying rates of evolution in disordered regions, it was found that nearly 75% of disordered regions evolved faster, and therefore would have lower sequence conservation among family members, than ordered regions within the same protein [80]. Only 9% of the disordered regions were more conserved than the ordered regions, and the remaining

18% were equally conserved. However, the sample size for the previous work was much smaller (26 families), and the families used were built using a BLAST search, rather than taken from protein family databases. Additionally, the previous work used only families with a region of experimentally characterized disorder, whereas this work used predicted disorder. Because of these differences in methodology, the results cannot be viewed as contradictory. In fact, both indicate that in some cases, disordered regions evolve faster, in others they evolve slower, and in the rest they evolve at roughly the same rate.

B. 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]. 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]

C. 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. Section IV.B.6 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 3D 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 disorder [81]. 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, 82].

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.

VI. Conclusion

In this work, regions of conserved predicted disorder were identified in domains from all member databases of InterPro and in domains occurring in all kingdoms of life. Although most of these conserved disordered regions were relatively short, between 20 and 30 residues, some were long. These long regions of conserved disorder were much more common in protein families and domains occurring in eukaryotic organisms and viruses. However, conserved predicted disorder was much less common than predicted disorder in general.

This work has also shown that protein domains and families have regions of conserved disorder as well as conserved sequence. Most conserved disordered regions had sequence conservation greater than or equal to that in conserved ordered regions within the same protein. This indicates that disorder tendencies are kept in these proteins, indicating that their function depends on disorder. 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.

VII. References

1. Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, W., Garner, E.C., Obradovic, Z., *Intrinsically disordered protein.* J. Mol. Graph. Model., 2001. **19**(1): p. 26-59.
2. Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M., Obradovic, Z., *Intrinsic disorder and protein function.* Biochemistry, 2002. **41**(21): p. 6573-6582.
3. Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z., Dunker, A.K., *Intrinsic disorder in cell-signaling and cancer-associated proteins.* J. Mol. Biol., 2002. **323**(3): p. 573-584.
4. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna,

A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., Eddy, S.R., *The Pfam protein families database.* Nucleic Acids Res., 2004. **32**(Database issue): p. D138-141.

5. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K., Bairoch, A., *The PROSITE database, its status in 2002.* Nucleic Acids Res., 2002. **30**(1): p. 235-238.
6. Haft, D.H., Selengut, J.D., White, O., *The TIGRFAMs database of protein families.* Nucleic Acids Res., 2003. **31**(1): p. 371-373.
7. Ponting, C.P., Schultz, J., Milpetz, F., Bork, P., *SMART: identification and annotation of domains from signalling and extracellular protein sequences.* Nucleic Acids Res., 1999. **27**(1): p. 229-232.
8. Wu, C.H., Nikolskaya, A., Huang, H., Yeh, L.S., Natale, D.A., Vinayaka, C.R., Hu, Z.Z., Mazumder, R., Kumar, S., Kourtesis, P., Ledley, R.S., Suzek, B.E., Arminski, L., Chen, Y., Zhang, J., Cardenas, J.L., Chung, S., Castro-Alvear, J., Dinkov, G., Barker, W.C., *PIRSF: family classification system at the Protein Information Resource.* Nucleic Acids Res., 2004. **32**(Database issue): p. D112-114.
9. Fischer, E., *Einfluss der Configuration auf die Wirkung der Enzyme.* Ber. Dt. Chem. Ges., 1894. **27**: p. 2985-2993.
10. Lemieux, R.U., Spohr, U., *How Emil Fischer was led to the lock and key concept for enzyme specificity.* Adv. Carbohydr. Chem. Biochem., 1994. **50**: p. 1-20.
11. Mirsky, A.E., Pauling, L., *On the structure of native, denatured, and coagulated proteins.* Proc Natl Acad Sci U S A, 1936. **22**(7): p. 439-447.
12. Edsall, J.T., *Hsien Wu and the first theory of protein denaturation (1931).* Adv. Protein Chem., 1995. **46**: p. 1-5.
13. Linderstrøm-Lang, K.U., The Lane Medical Lectures, 1952. **6**: p. 1-115.
14. Linderstrøm-Lang, K.U., Shellman, J.A., *Protein structure and enzyme activity*, in *The Enzymes*, P.D. Boyer, Editor. 1959, Academic Press: New York. p. 443-510.
15. Kendrew, J.C., Dickerson, R.E., Strandberg, B.E., Hart, R.G., Davies, R.D., Phillips, D.C. and Shore, V.C., *Structure of myoglobin: a three-dimensional Fourier synthesis at 2.0Å resolution.* Nature, 1960. **185**: p. 422-427.
16. Perutz, M.F., Rossman, M.G., Cullis, A.F., Muirhead, H., Will, G. and North, A.C.T., *Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5Å resolution, obtained by X-ray analysis.* Nature, 1960. **185**: p. 416-422.
17. Daughdrill, G.W., Pielak, G.J., Uversky, V.N., Cortese, M.S., and Dunker, A.K., *Natively Disordered Proteins*, in *Protein Folding Handbook. Part II.*, J. Buchner, Kiefhaber, T., Editor. 2005, WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim. p. 275-357.
18. Karush, F., *Heterogeneity of the binding sites of bovine serum albumin.* J. Am. Chem. Soc., 1950. **72**(6): p. 2705-2713.
19. Dunker, A.K., Obradovic, Z., *The protein trinity -- linking function and disorder.* Nat. Biotechnol., 2001. **19**(9): p. 805-806.
20. Iakoucheva, L.M., Kimzey, A.L., Masselon, C.D., Bruce, J.E., Garner, E.C., Brown, C.J., Dunker, A.K., Smith, R.D., Ackerman, E.J., *Identification of intrinsic order and disorder in the DNA repair protein XPA.* Protein Sci., 2001. **10**(3): p. 560-571.
21. Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Garner, E., Guilliot, S., Dunker, A.K., *Thousands of proteins likely to have long disordered regions.* Pac. Symp. Biocomput., 1998: p. 437-448.
22. Li, X., Romero, P., Rani, M., Dunker, A.K., Obradovic, Z., *Predicting Protein Disorder for N-, C-, and Internal Regions.* Genome Inform. Ser. Workshop Genome Inform., 1999. **10**: p. 30-40.
23. Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., Dunker, A.K., *Sequence complexity of disordered protein.* Proteins, 2001. **42**(1): p. 38-48.
24. Vucetic, S., Brown, C.J., Dunker, A.K., and Obradovic, Z., *Flavors of protein disorder.*

Proteins, 2003. **52**(4): p. 573-584.

25. Radivojac, P., Obradovic, Z., Smith, D.K., Zhu, G., Vucetic, S., Brown, C.J., Lawson, J.D., Dunker, A.K., *Protein flexibility and intrinsic disorder.* Protein Sci., 2004. **13**(1): p. 71-80.

26. Dunker, A.K., Obradovic, Z., Romero, P., Garner, E.C., Brown, C.J., *Intrinsic protein disorder in complete genomes.* Genome Inform. Ser. Workshop Genome Inform., 2000. **11**: p. 161-171.

27. Uversky, V.N., Gillespie, J.R., Fink, A.L., *Why are "natively unfolded" proteins unstructured under physiologic conditions?* Proteins, 2000. **41**(3): p. 415-427.

28. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C.J., Dunker, A.K., *Predicting intrinsic disorder from amino acid sequence.* Proteins, 2003. **53**: p. 566-572.

29. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., Jones, D.T., *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.* J. Mol. Biol., 2004. **337**(3): p. 635-645.

30. Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A., Zygouri, C., *PRINTS and its automatic supplement, prePRINTS.* Nucleic Acids Res., 2003. **31**(1): p. 400-402.

31. Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D., Kahn, D., *ProDom: automated clustering of homologous domains.* Brief. Bioinform., 2002. **3**(3): p. 246-251.

32. Gough, J., Karplus, K., Hughey, R., Chothia, C., *Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structur.* J. Mol. Biol., 2001. **323**(4): p. 903-919.

33. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman D.J., *Basic local alignment search tool.* J. Mol. Biol., 1990. **215**(3): p. 403-410.

34. Kulathinal, R.J., Skwarek, L., Morton, R.A., Singh, R.S., *Rapid evolution of the sex-determining gene, transformer: structural diversity and rate heterogeneity among sibling species of Drosophila.* Mol.Biol. Evol., 2003. **20**(3): p. 441-452.

35. Zahn, R., Liu, A., Luhrs, T., Riek, R., von Schroetter, C., Lopez Garcia, F., Billeter, M., Calzolai, L., Wider, G., Wuthrich, K., *NMR solution structure of the human prion protein.* Proc Natl Acad Sci U S A, 2000. **97**(1): p. 145-150.

36. Calzolai, L., Lysek, D.A., Perez, D.R., Guntert, P., Wuthrich, K., *Prion protein NMR structures of chickens, turtles, and frogs.* Proc Natl Acad Sci U S A, 2005. **102**(3): p. 651-655.

37. Vila-Sanjurjo, A., Ridgeway, W.K., Seymaner, V., Zhang, W., Santoso, S., Yu, K., Cate, J.H., *X-ray crystal structures of the WT and a hyper-accurate ribosome from Escherichia coli.* Proc Natl Acad Sci U S A, 2003. **100**(15): p. 8682-8687.

38. Ghetu, A.F., Gubbins, M.J., Frost, L.S., Glover, J.N., *Crystal structure of the bacterial conjugation repressor finO.* Nat. Struct. Biol., 2000. **7**(7): p. 565-569.

39. Aizawa, S.I., Vonderviszt, F., Ishima, R., Akasaka, K., *Termini of Salmonella flagellin are disordered and become organized upon polymerization into flagellar filament.* J. Mol. Biol., 1990. **211**(4): p. 673-677.

40. Samatey, F.A., Imada, K., Nagashima, S., Vonderviszt, F., Kumasaka, T., Yamamoto, M., Namba, K., *Structure of the bacterial flagellar protofilament and implications for a switch for supercoiling.* Nature, 2001. **410**(6826): p. 331-337.

41. Vonderviszt, F., Kanto, S., Aizawa, S., Namba, K., *Terminal regions of flagellin are disordered in solution.* J. Mol. Biol., 1989. **209**(1): p. 127-133.

42. Liao, D.I., Reiss, L., Turner, I., Dotson, G., *Structure of glycerol dehydratase reactivase: a new type of molecular chaperone.* Structure, 2003. **11**(1): p. 109-119.

43. Gao, H., Sengupta, J., Valle, M., Korostelev, A., Eswar, N., Stagg, S.M., Van Roey, P., Agrawal, R.K., Harvey, S.C., Sali, A., Chapman, M.S., Frank, J., *Study of the structural dynamics of the E coli 70S ribosome using real-space refinement.* Cell, 2003. **113**(6): p.

789-801.

44. Turner, C.F., Moore, P.B., *The solution structure of ribosomal protein L18 from Bacillus stearothermophilus.* J. Mol. Biol., 2004. **335**(3): p. 679-684.

45. Liao, D.I., Dotson, G., Turner, I. Jr, Reiss, L., Emptage, M., *Crystal structure of substrate free form of glycerol dehydratase.* J Inorg. Biochem., 2003. **93**(1-2): p. 84-91.

46. Mavrakis, M., McCarthy, A.A., Roche, S., Blondel, D., Ruigrok, R.W., *Structure and function of the C-terminal domain of the polymerase cofactor of rabies virus.* J. Mol. Biol., 2004. **343**(4): p. 819-831.

47. Tong, L., Wengler, G., Rossmann, M.G., *Refined structure of Sindbis virus core protein and comparison with other chymotrypsin-like serine proteinase structures.* J. Mol. Biol., 1993. **230**(1): p. 228-247.

48. Klein, D.J., Moore, P.B., Steitz, T.A., *The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit.* J. Mol. Biol., 2004. **340**(1): p. 141-177.

49. Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K.L., Wu, N., McIntosh, L.P., Gehring, K., Kennedy, M.A., Davidson, A.R., Pai, E.F., Gerstein, M., Edwards, A.M., Arrowsmith, C.H., *Structural proteomics of an archaeon.* Nat. Struct. Biol., 2000. **7**(10): p. 903-909.

50. Harms, J., Schluenzen, F., Zarivach, R., Bashan, A., Gat, S., Agmon, I., Bartels, H., Franceschi, F., Yonath, A., *High resolution structure of the large ribosomal subunit from a mesophilic eubacterium.* Cell, 2001. **107**(5): p. 679-688.

51. Ban, N., Nissen, P., Hansen, J., Capel, M., Moore, P.B., Steitz, T.A., *Placement of protein and RNA structures into a 5 A-resolution map of the 50S ribosomal subunit.* Nature, 1999. **400**(6747): p. 841-847.

52. Spahn, C.M., Gomez-Lorenzo, M.G., Grassucci, R.A., Jorgensen, R., Andersen, G.R., Beckmann, R., Penczek, P.A., Ballesta, J.P., Frank, J., *Domain movements of elongation factor eEF2 and the eukaryotic 80S ribosome facilitate tRNA translocation.* EMBO J., 2004. **23**(5): p. 1008-1019.

53. Soelaiman, S., Jakes, K., Wu, N., Li, C., Shoham, M., *Crystal structure of colicin E3: implications for cell entry and ribosome inactivation.* Mol. Cell., 2001. **8**(5): p. 1053-1062.

54. Fletcher, C.M., Pestova, T.V., Hellen, C.U., Wagner, G., *Structure and interactions of the translation initiation factor eIF1.* EMBO J., 1999. **18**(9): p. 2631-2637.

55. Battiste, J.L., Pestova, T.V., Hellen, C.U., Wagner, G., *The eIF1A solution structure reveals a large RNA-binding surface important for scanning function.* Mol. Cell., 2000. **5**(1): p. 109-119.

56. Paquet, F., Culard, F., Barbault, F., Maurizot, J.C., Lancelot, G., *NMR solution structure of the archaebacterial chromosomal protein MC1 reveals a new protein fold.* Biochemistry, 2004. **43**(47): p. 14971-14978.

57. Kozlov, G., Ekiel, I., Beglova, N., Yee, A., Dharamsi, A., Engel, A., Siddiqui, N., Nong, A., Gehring, K., *Rapid fold and structure determination of the archaeal translation elongation factor 1beta from Methanobacterium thermoautotrophicum.* J. Biomol. NMR, 2000. **17**(3): p. 187-194.

58. Isupov, M.N., Fleming, T.M., Dalby, A.R., Crowhurst, G.S., Bourne, P.C., Littlechild, J.A., *Crystal structure of the glyceraldehyde-3-phosphate dehydrogenase from the hyperthermophilic archaeon Sulfolobus solfataricus.* J. Mol. Biol., 1999. **291**(3): p. 651-660.

59. Fisher, L.W., Torchia, D.A., Fohr, B., Young, M.F., Fedarko, N.S., *Flexible structures of SIBLING proteins, bone sialoprotein, and osteopontin.* Biochem. Biophy. Res. Commun., 2001. **280**(2): p. 460-465.

60. Dinchuk, J.E., Henderson, N.L., Burn, T.C., Huber, R., Ho, S.P., Link, J., O'Neil, K.T., Focht, R.J., Scully, M.S., Hollis, J.M., Hollis, G.F., Friedman, P.A., *Aspartyl beta -hydroxylase (Asph) and an evolutionarily conserved isoform of Asph missing the catalytic domain share exons with junctin.* J. Biol. Chem., 2000. **275**(50): p. 39543-39554.

61. Donne, D.G., Viles, J.H., Groth, D., Mehlhorn, I., James, T.L., Cohen, F.E., Prusiner, S.B., Wright, P.E., Dyson, H.J., *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**(25): p. 13452-13457.

62. Masson, D., Kreis, T.E., *Identification and molecular characterization of E-MAP-115, a novel microtubule-associated protein predominantly expressed in epithelial cells.* J. Cell. Biol., 1993. **123**(2): p. 357-371.

63. Masson, D., Kreis, T.E., *Binding of E-MAP-115 to microtubules is regulated by cell cycle-dependent phosphorylation.* J. Cell. Biol., 1995. **131**(4): p. 1015-1024.

64. Ilk, N., Kosma, P., Puchberger, M., Egelseer, E.M., Mayer, H.F., Sleytr, U.B., 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**(24): p. 7643-7646.

65. Luo, Y., Frey, E.A., Pfuetzner, R.A., Creagh, A.L., Knoechel, D.G., Haynes, C.A., Finlay, B.B., Strynadka, N.C., *Crystal structure of enteropathogenic Escherichia coli intimin-receptor complex.* Nature, 2000. **405**(6790): p. 1073-1077.

66. Abe, A., de Grado, M., Pfuetzner, R.A., Sanchez-Sanmartin, C., Devinney, R., Puente, J.L., Strynadka, N.C., Finlay, B.B., *Enteropathogenic Escherichia coli translocated intimin receptor, Tir, requires a specific chaperone for stable secretion.* Mol. Microbiol., 1999. **33**(6): p. 1162-1175.

67. Kenny, B., Finlay, B.B., *Intimin-dependent binding of enteropathogenic Escherichia coli to host cells triggers novel signaling events, including tyrosine phosphorylation of phospholipase C-gamma1.* Infect. Immun., 1997. **65**(7): p. 2528-2536.

68. Ban, N., Nissen, P., Hansen, J., Moore, P.B., Steitz, T.A., *The complete atomic structure of the large ribosomal subunit at 2.4 A resolution.* Science, 2000. **289**(5481): p. 905-920.

69. Lee, J., Klusener, B., Tsiamis, G., Stevens, C., Neyt, C., Tampakaki, A.P., Panopoulos, N.J., Noller, J., Weiler, E.W., Cornelis, G.R., Mansfield, J.W., 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**(1): p. 289-294.

70. Ghetu, A.F., Gubbins, M.J., Oikawa, K., Kay, C.M., Frost, L.S., Glover, J.N., *The FinO repressor of bacterial conjugation contains two RNA binding regions.* Biochemistry, 1999. **38**(42): p. 14036-14044.

71. Bourhis, J.M., Johansson, K., Receveur-Brechot, V., Oldfield, C.J., Dunker, K.A., 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**(2): p. 157-167.

72. Wiethoff, C.M., Wodrich, H., Gerace, L., Nemerow, G.R., *Adenovirus protein VI mediates membrane disruption following capsid disassembly.* J. Virol., 2005. **79**(4): p. 1992-2000.

73. Stewart, P.L., Fuller, S.D., Burnett, R.M., *Difference imaging of adenovirus: bridging the resolution gap between X-ray crystallography and electron microscopy.* EMBO J., 1993. **12**(7): p. 2589-2599.

74. Bodian, D.L., Jones, E.Y., Harlos, K., Stuart, D.I., Davis, S.J., *Crystal structure of the extracellular region of the human cell adhesion molecule CD2 at 2.5 A resolution.* Structure, 1994. **2**(8): p. 755-766.

75. Jones, E.Y., Davis, S.J., Williams, A.F., Harlos, K., Stuart, D.I., *Crystal structure at 2.8*

A resolution of a soluble form of the cell adhesion molecule CD2. Nature, 1992. **360**(6401): p. 232-239.

76. Freund, C., Kuhne, R., Yang, H., Park, S., Reinherz, E.L., Wagner, G., *Dynamic interaction of CD2 with the GYF and the SH3 domain of compartmentalized effector molecules.* EMBO J., 2002. **21**(22): p. 5985-5995.
77. Nichols, M.D., DeAngelis, K., Keck, J.L., Berger, J.M., *Structure and function of an archaeal topoisomerase VI subunit with homology to the meiotic recombination factor Spo11.* EMBO J., 1999. **18**(21): p. 6177-6188.
78. Daugherty, M., Vonstein, V., Overbeek, R., Osterman, A., *Archaeal shikimate kinase, a new member of the GHMP-kinase family.* J. Bacteriol., 2001. **183**(1): p. 292-300.
79. Krell, T., Coggins, J.R., Lapthorn, A.J., *The three-dimensional structure of shikimate kinase.* J. Mol. Biol., 1998. **278**(5): p. 983-997.
80. Brown, C.J., Takayama, S., Campen, A.M., Vise, P., Marshall, T.W., Oldfield, C.J., Williams, C.J., Dunker, A.K., *Evolutionary rate heterogeneity in proteins with long disordered regions.* J. Mol. Evol., 2002. **55**(1): p. 104-110.
81. Iakoucheva, L.M., Kimzey, A.L., Masselon, C.D., Smith, R.D., Dunker, A.K., Ackerman, E.J., *Aberrant mobility phenomena of the DNA repair protein XPA.* Protein Sci., 2001. **10**(7): p. 1353-1362.
82. Romero, P., Obradovic, Z., and Dunker, A.K., *Folding minimal sequences: the lower bound for sequence complexity of globular proteins.* FEBS Letters, 1999. **462**(3): p. 363-367.

Appendix A: Curriculum Vitae

Jessica Walton Chen
1425 Everett Avenue, Apt 1
Louisville, KY 40204
(502) 456-3328
jessica.w.chen@gmail.com

Experience

Molecular Kinetics, Inc. Indianapolis, IN
Bioinformatics Programmer 8/2003 – 8/2004
Designed and implemented a database of cancer proteins for SBIR Phase I project. Built web-based interface for data entry using Perl and Apache. Developed Perl scripts for statistical and graphical analysis of data. Interviewed, hired, trained, and managed small team in annotating proteins for database.

Developed algorithm for analysis of protein sequence conservation using alignments and entropy calculations. Investigated usage of conservation in determining protein structure/domain boundaries.

Wrote one Small Business Innovation Research (SBIR) Phase I grant and one SBIR Phase II grant, as well as multiple progress reports.

IUPUI Department of Computer Engineering Indianapolis, IN
Research Assistant 1/2003 – 8/2003
Implemented a Java Swing application for building and executing workflows of bioinformatics services as part of the SIBIOS (Service Integration for Bioinformatics Services) project. Built this application as a standard menu-and-toolbar-driven desktop application, deployable via the web using Java WebStart technology. Functionality of the application included adding, deleting, and editing nodes in the workflow to represent services and data filters, as well as execution control and results viewing and exporting.

Implemented web services using commercial web services platform for communication between the workflow builder and the server components, which handled service discovery, service schema definition, and execution of workflow services.

Researched and analyzed web-based bioinformatics services to determine which should be integrated into SIBIOS. Refined an ontology of biological and bioinformatics terms used to define services. Constructed XML-based schema files representing inputs, outputs and execution parameters of bioinformatics services. Trained intern to construct schema files.

Informative, Inc (formerly Recipio) Brisbane, CA
Software Engineer 5/1998 – 7/2002

Used Java and JSP to build reporting client that communicated with a SQL server. Implemented changes to reporting engine to support new features including customizable charts and cross-tab segmentation reports. Wrote detailed design documents, participated in design reviews.

Translated non-technical idea for web-based customer interaction portal into interactive HTML mock-up which succeeded in selling product idea to clients. Collaborated with a small distributed group to design, build, test, debug, and release this product in a tight time frame. Used Java, JDBC, SQL and servlets and n-tier architecture. First implementation of the product went live after three months and grew to over 50,000 users in the first year.

Worked with a team to re-architect customer relationship software. The software consolidated and analyzed guided and free-form text input from users obtained via a simple web interface.

Developed prototype and alpha versions of Java application for product management. Designed and coded features for product management based on specifications. Collaborated with product managers and marketing to improve product and requirements. Redesigned user interface; interacted with user interface design consultant.

Caltech Department of Computer Science Pasadena, CA
Research Intern 6/1997 – 8/1997
Collaborated with another student to design and implement a distributed agent bartering system in Java. Exploited existing infrastructure for messaging and component management. (http://www.infospheres.caltech.edu/papers/vsm/paper.html)

Stanford Department of Molecular and Cellular Physiology Stanford, CA
Research Assistant 6/1996 – 9/1996
Assisted graduate student with experiments involving calcium signaling in T lymphocytes. Maintained cell lines, analyzed data, performed video microscopy.

Education

Indiana University-Purdue University Indianapolis Indianapolis, IN
Master of Bioinformatics August 2005
Recipient of 2002 MDL Excellence in Informatics Fellowship
Coursework includes bioinformatics, molecular biology, databases, and statistics.
GPA: 3.9

Stanford University Stanford, CA
Bachelor of Science in Biological Sciences April 1998
Minored in Computer Science. Coursework included programming methodology, data structures and algorithms, object-oriented systems design, multivariable mathematics, molecular biology, and genetics.
GPA: 3.7