PREDICTIONS ON AND ANALYSIS OF VIRAL PROTEINS ENCODED BY OVERLAPPING GENES

Mahvash Khosravi

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 2007

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

Master's Thesis
Committee

_____
A. Keith Dunker, PhD, Chair

_____
Pedro Romero, PhD

_____
Vladimir N. Uversky, PhD

Dedicated to the memory of my mother

Acknowledgements

I wish to express my gratitude to my thesis advisor Dr. A. Keith Dunker for his guidance, support and providing me the opportunity to work on this project. I would like to thank Dr. Pedro Romero for his advice and valuable input through the course of this research, as well as serving on my thesis committee. His guidance and encouragement is greatly appreciated. Also, I would like to thank Dr. Vladimir Uversky for serving on my thesis committee.

Finally I would like to express my gratitude and love for my family. I am indebted to my husband Dr. Hossein Hariri for his endless love, dedication, patience and encouragement. Thank you Hossein.

My deepest appreciation and love go to our beautiful daughters, our sweet engineer Ghazal and dear Homa who makes us so proud.

Abstract

Overlapping genes are adjacent genes that share a portion of their coding sequence. Such genes are often observed in the compact genomes of viruses, prokaryotes, and mitochondria. Overlapping genes are also seen in human and other mammalian genomes. Gene overlapping is a phenomenon to minimize genomic size and maximize encoding capacity. Overlapping genes produce different proteins. A major task in the post genomic era is the large-scale study of the structures and functions of proteins. Proteins play crucial roles in virtually all biological processes. In general it is assumed that 3-D structure determines the function of proteins, but many proteins or region of proteins may function in the absence of 3-D structure. The term "disordered" is used to describe these proteins. A large number of studies has shown that biological functions depend on both ordered and disordered proteins. Natively disordered regions are common and play essential roles in many proteins, especially, with regard to activities involved in signaling and regulation.

The goal of this research was the analysis of the ordered and disordered tendencies of viral proteins encoded by overlapping genes. Our hypothesis is that, in a pair of proteins or protein regions encoded by overlapping genes, at least one of the pair is disordered (or unstructured). Our hypothesis is based on the observation that structural proteins require highly specific amino acid sequences, while unstructured (disordered) sequences are essentially unconstrained. Thus, given a structural protein and its associated mRNA sequence, any sequence derived from an overlapping reading frame seems highly unlikely to have a sequence pattern commensurate with a structural protein; on the other hand, a sequence pattern consistent with a disordered protein seems much

more likely. We performed studies on the protein products of overlapping gene sequences, tested the hypothesis and addressed the following two questions: First do the proteins encoded by overlapping genes have opposite order-disorder content, that is, does the ordered part of one of the overlapping proteins correspond to a disordered part in the other overlapping protein? Second, does the encoded protein in the overlapping regions have more disordered amino acids than the non-overlapping regions?

Using our database of overlapping viral genes and the protein predictor PONDR VL3, we predicted the order-disorder of amino acids in the sequence of 97 viral protein samples. An analysis of the results supported our hypothesis and indicated that the ordered amino acids are mostly associated with non-overlapping regions while disordered amino acids are more prevalent in overlapping regions. In the overlapping regions for 52 protein pairs, we showed that most of the amino acid pairs facing each other on the protein sequences had at least one disorder for most cases. Out of 52 pairs, there were 3 protein pairs where there were no disordered amino acids and 22 protein pairs where there were no ordered amino acids on either sequence. The fraction of ordered pairs in the pool of overlapping regions of 52 protein pairs was 0.28. The non-overlapping region of 97 proteins had predominantly ordered proteins. The fraction of ordered amino acids in the pool of non-overlapping regions was determined to be 0.77.

Table of Contents

List of Tables

List of Figures

I. Introduction

Overlapping genes are adjacent genes that share a portion of their coding sequence. They are often observed in compact genome of viruses, prokaryotes, and mitochondria. Overlapping genes also occur in mammalian genomes.

Overlapping genes encode different protein products using the same nucleic acid sequence. Proteins play crucial roles in virtually all biological processes. For many years it was thought that 3-D structure of proteins determine their functions [1,2,3]. But there are proteins or region of proteins which lack 3-D structure, yet such proteins and regions function in the absence of any specific fixed structure [4,5]. These proteins, called natively disordered proteins, have many important roles in biological processes, specifically in cell cycle control, signaling and regulation [6]. Thus, detailed study and understanding of structure and function of natively disordered proteins is important and may eventually lead to finding cure for human diseases and novel medical products.

To the best of our knowledge, what is presented in this thesis is the first attempt to study disordered proteins expressed by overlapping genes. In this work the focus is on the protein products of viral overlapping gene sequences. First do the proteins encoded by overlapping genes have opposite order-disorder content, that is, does the ordered part of one of the overlapping proteins correspond to a disordered part in the other overlapping protein? Second do the overlapping regions of these proteins have a higher percentage of disordered amino acids than the non-overlapping regions?

In this introduction, we will first discuss the overlapping genes, proteins expressed by these genes, and the protein structure function paradigm. We then discuss

natively disordered proteins, how they are predicted, their frequency and function, their amino acid composition and their significance.

II. Background

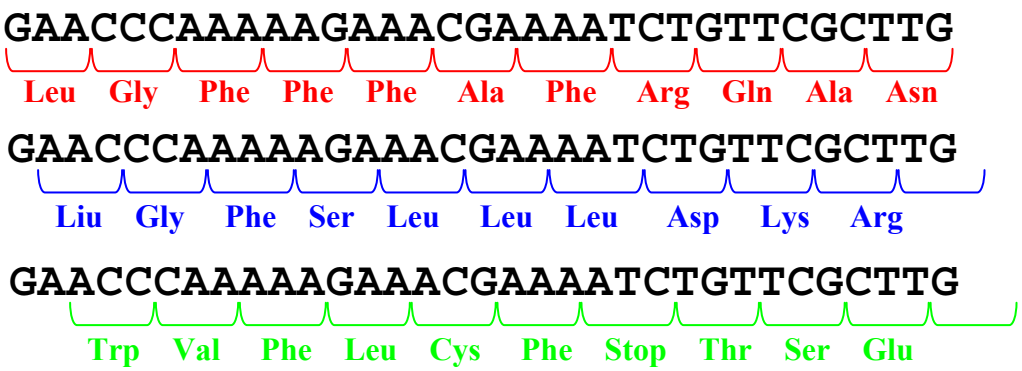
Overlapping genes

A gene is any given segment along the DNA that encodes instructions that allow a cell to produce a specific product, typically, a protein such as an enzyme that initiates a specific action. There are between 50,000 and 100,000 genes, and every gene is made up of thousands, sometimes even hundreds of thousands, of chemical bases [7].

Individual genes can overlap and share portion of their nucleotide sequence. Overlapping genes are defined as adjacent genes which share a portion of their nucleotide sequence [8,9]. Overlapping genes might have occurred by an overprinting mechanism or by rearrangements [10,11]. The overprinting mechanism is the process of generating new genes from pre-existing nucleotide sequences utilizing a frame shift phenomenon. This phenomenon allows overlapping genes to encode different protein products using the same nucleic acid sequence. Rearrangement is a process where the loss of a stop codon in a specific gene causes the gene to elongate to the stop codon of the next gene. Comparative analysis of the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae* has revealed a rearrangement process in which the overlapping genes are generated by mutations at the ends of coding regions [12]. The rearrangement process may occur in two ways; first is when the 3'-untranslated region and polyadenylation signal of a gene is lost, but somehow it may utilize the 3'-untranslated region and polyadenylation signal on the opposite strand of a neighboring gene. Second is when two genes somehow become neighbor and initially do not overlap, later one of the genes loses

its polyadenylation signal but ends up utilizing the polyadenylation signal that is present on the non-coding strand of the other gene [14].

An example of gene overlapping is described below. This example shows the frame shift phenomenon where changes in the reading frame of a nucleotide sequence leads to the production of different amino acid sequences.



The overlapping gene phenomenon has been suggested to be beneficial for the viruses and other organisms for several possible reasons. Overlapping genes can reduce the size of the genome without affecting the number of genes encoded. Overlapping genes can produce new proteins without increasing the size of genome. Overlapping genes can coordinate the expression level of functionally related genes. Overlapping genes can coordinate the expression of genes where the expression of one gene requires the deactivation of the other [11,12]. Gene overlapping is normally observed in compact genomes that have high rates of mutation such as viruses, bacteria and organelles like mitochondria [15]. Overlapping genes are also relatively frequent in human and other mammalian genome [16].

The origin and evolution of overlapping genes have been the subject of a number of studies which suggest that these genes are produced by evolutionary mechanisms to

decrease the genome size while increasing the number of genes [12,13,17,18,19,20]. It is speculated that the rates of evolution are slower in overlapping genes [21]. Recently, it has been suggested that evolution of overlapping genes occurs at a universal mutation rate across bacterial genomes [13]. More studies are needed to learn about the origin, evolution and cross-species conservation, and frequency and genome-wide distribution of overlapping genes in different genomes.

Overlapping genes may offer information about how coding and control sequences have evolved. It can also provide information about evolution patterns among classes of organisms [22]. Studies and comparison of overlapping genes in related species may help us understand how and under what conditions overlap evolved [13]. Gene overlap has been associated with a number of human disease genes since genomic rearrangements are likely to occur within overlapping regions possibly because of inconsistent sequence features common in these regions [23].

Protein structure-function paradigm

Proteins play crucial roles in virtually all biological processes. Proteins can act as enzymatic catalysts as well as assist in storage, coordination, transportation, and motion; provide mechanical support and immune protection; and aid in growth control, differentiation, nerve impulse transmission, and nerve impulse generation. Proteins, as a distinctive characteristic, have well-defined 3-D structures. According to the structure-function paradigm, the amino acid sequence specifies a protein's 3-D structure and the 3-D structure must be present for the protein to function. Fischer's "lock and key" proposal in 1894 [24] was an early concept that eventually led to the structure-function paradigm. Fischer used his studies on enzymes which hydrolyzed different bonds to draw

conclusions that led to his proposal. Studies by Wu in 1930's showed that the addition of heat or solutes to globular proteins causes their denaturation and loss of biological activity [25]. Pauling and Mirsky also reached the same conclusion as Wu [1]. Many years later, Anfinsen made the critical observation that ribonuclease denaturation is reversible and showed that the information needed to specify the 3-D structure of ribonuclease is contained in its amino acid sequence [26]. Thus, amino acid sequence is important because it specifies the conformation of protein. X-ray crystallography studies of structures of myoglobin and hemoglobin have also confirmed the structure-function paradigm [27]. In contrast to the focus on protein structure, Williams in his nuclear magnetic resonance spectroscopy revealed that some proteins lack defined and folded structures in solution [28].

Amino acids are the basic structural units of proteins which are linked by peptide bond and form a polypeptide chain. Each protein consists of one or more unique polypeptide chains. The amino acid sequence of a polypeptide chain forms the primary structure. Different regions of the amino acid sequence form local regular secondary structure, such as alpha helices or beta strands. Association of alpha helix and beta strands leads to folding of the protein.

Structural domains, which are compact globular units, are formed by interaction between elements of secondary structures and their side chains. The tertiary structure is the totally folded polypeptide chain and it may include one or more domains. Fully folded polypeptide chains may interact with other polypeptide chains and form a larger structure called subunit. The overall assembly is referred to as quaternary structure. By formation of such tertiary and quaternary structures, amino acids far apart in the sequence

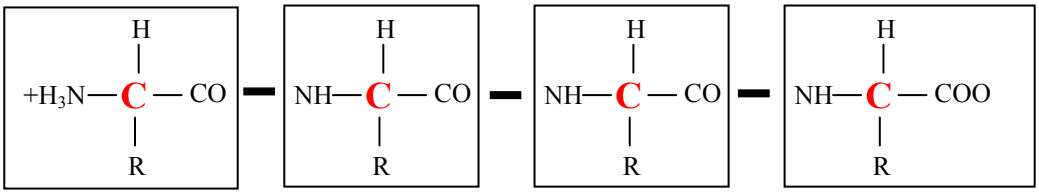
are brought close together in three dimensions and form a functional region, the active site. Proteins must recognize thousands of different molecules in the cell by detailed three-dimensional interactions, which require diverse and irregular structures of the protein molecules, the most important of which is their secondary structure [29].

Protein folding

As mentioned earlier, the amino acid sequence of a polypeptide chain forms the primary structure of a protein. In general, the information contained in the primary structure determines the manner by which protein folds. There are other forces that play roles in protein folding as will be discussed in this section.

There are twenty amino acids. A combination of amino acids makes a polypeptide chain from which proteins are formed. Each amino acid consists of an amino group, a carboxyl group, a hydrogen atom and a variable side chain (R group), which are bonded to a carbon atom called α-carbon. Each side chain differs in shape, size, charge, hydrogen bonding capability and chemical reactivity. Amino acids are linked by peptide bonds to form a polypeptide chain as shown in Figure 1. An amino acid unit in a polypeptide is called a residue.

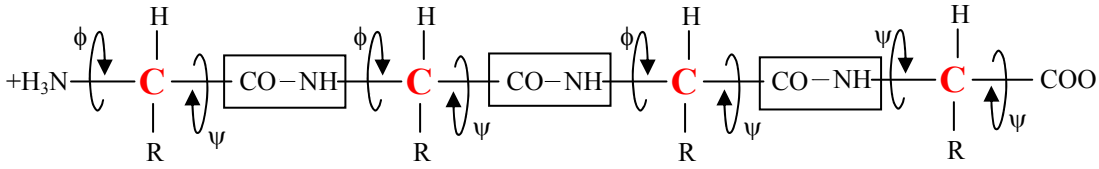
Figure 1. A Polypeptide chain with four amino acid residues



In the late 1930s, Linus Pauling and Robert Corey [30] carried out x-ray crystallographic studies on peptides and found that in a polypeptide chain, the CO-NH

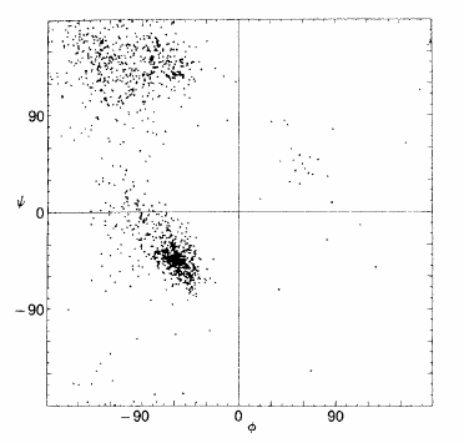
peptide unit is rigid and planar. In contrast, the bonds between α-carbon and NH and CO groups are single bonds which give them a large degree of rotational freedom. Rotation about the C_α-N bond is labeled ϕ and the one about C_α-C bond is labeled ψ as shown in Figure 2.

Figure 2. Rigid CO-NH bond and rotation of C_α-C and C_α-N bonds



Positive variation in ϕ corresponds to a clockwise rotation when viewed from C_α toward N. For ψ positive variation corresponds to a clockwise rotation when viewed from C_α toward C. The conformation corresponding to ϕ = ψ = 0 is when two CO-NH planes connected to a common C_α lie in the same plane. In principle ϕ and ψ can have any value between -180 to +180 degrees, however, many ϕ, ψ angular combinations are impossible because of steric collisions between atoms along the backbone or between backbone atoms and the side chain R group [31]. The polypeptide conformation has been represented by points on a ψ versus ϕ plot called Ramachandran plot. Figure 3 shows a Ramachandran plot for 1000 nonglycine residues in eight proteins [31].

Figure 3. Ramachandran plot for 1000 nonglycine residues in eight proteins



Since protein folding takes place in an aqueous environment, the interaction between polypeptide and water plays an important role in protein folding. The folding of water-soluble globular protein is due to minimizing the extent of exposure of hydrophobic group to the solvent. As a result the side chains are packed into the interior of the molecule which leads to a hydrophobic interior and a hydrophilic surface. The main chain folds into the interior with the side chains. The main polypeptide chain is hydrophilic because it is highly polar. In each peptide unit the NH group is hydrogen bond donor and the CO group is hydrogen bond acceptor. In a hydrophobic environment these polar groups are neutralized by hydrogen bond formation. This is facilitated by the formation of regular secondary structure within the interior of the protein molecule.

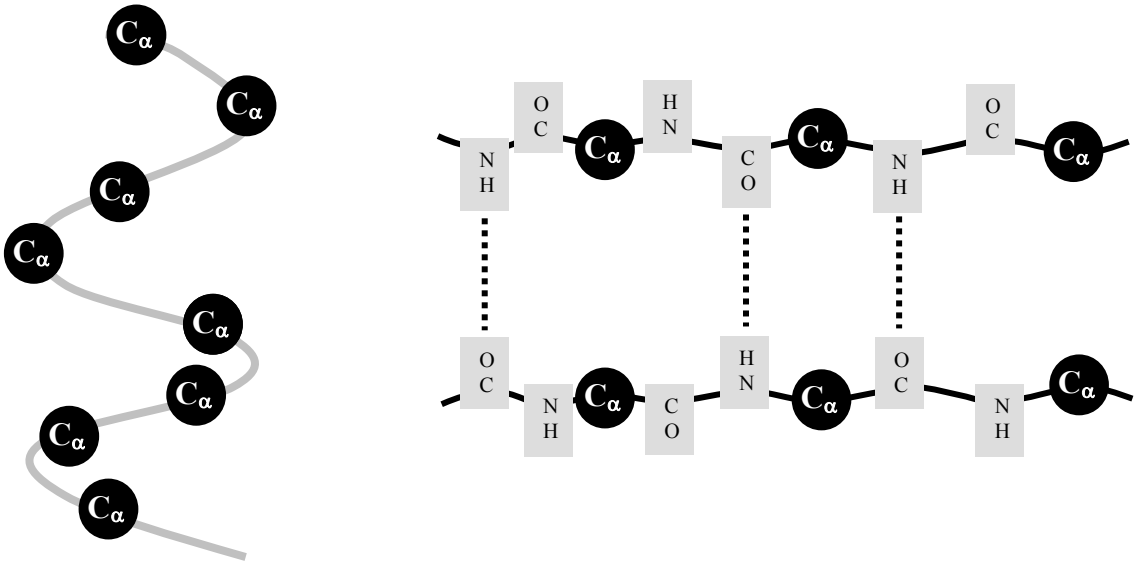
Finally, hydrogen bonding between elements of the peptide backbone leads to the formation of secondary structure.

Secondary structure of protein

In 1951 Pauling and Corey [30] proposed two models for the secondary structure of protein, alpha helix and beta pleated sheet. The alpha helix is a rodlike structure, as

shown in Figure 4, with a tightly coiled polypeptide forming the inner part of the rod and the side chains extend outward in a helical array.

Figure 4. Alpha helix



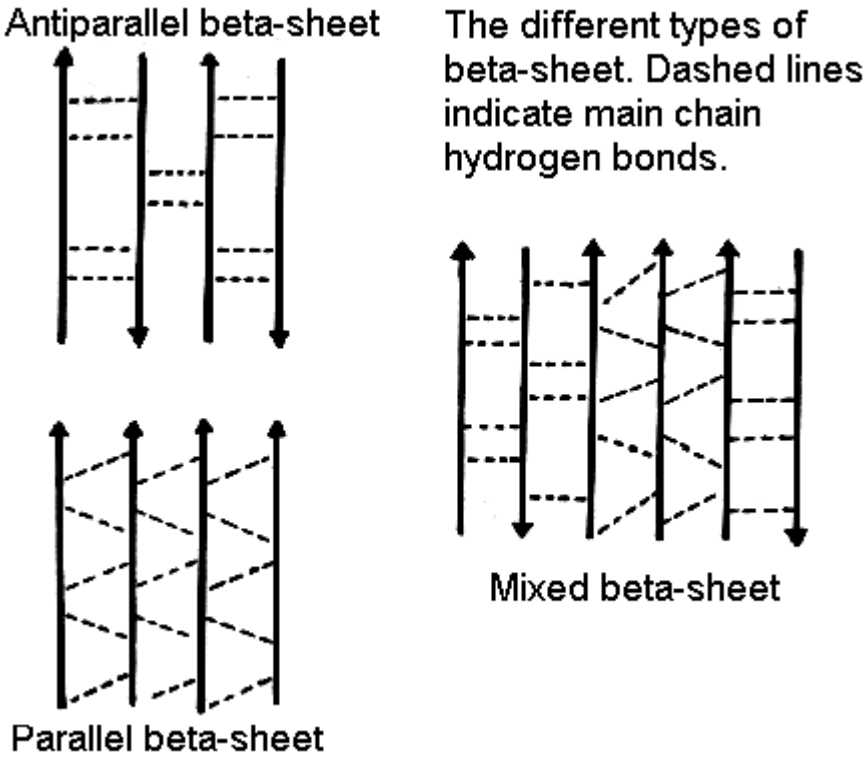
The most common secondary structure in currently known globular proteins is the alpha helix. The Alpha helix is stabilized by hydrogen bonds between the NH and CO groups of the main chain. All the hydrogen bonds in an alpha helix point in the same direction so the peptide units are aligned in the same orientation along a helical axis. There is a partial positive charge at the amino end and a partial negative charge at the carboxyl end of an alpha helix. This produces a significant net dipole for the alpha helix. These charges attract ligands of opposite charge. Alpha helix in a protein is basically the outside of the protein with one side of the helix facing the solution and the other side toward the hydrophobic interior of the protein.

Beta sheet is the second most common type of structural element found in currently known globular proteins. The beta sheet differs from alpha helix, in that it looks like a sheet and it is almost fully extended rather than being tightly coiled as in alpha helix. Beta pleated sheets are formed when two or more polypeptide chains are brought

together side by side. In this case the NH group of an amino acid residue on one chain forms a hydrogen bond with the CO group of the adjacent chain.

The strands of beta-sheets can run in one direction in a parallel arrangement. In an anti-parallel arrangement, sheets run in opposite directions. In a mixed-sheet arrangement some strands are parallel and others are anti-parallel as shown in Figure 5.

Figure 5. Beta sheet



Beta bends

In the previous section we discussed alpha helix and beta sheets as two forms of secondary structure of polypeptide chains. In order to fold this chain to a compact globular form, the polypeptide chains must be able to change direction. A commonly observed way to facilitate this change in direction is a beta bend. As shown in Figure 6, beta bend is a tight loop in which a CO group forms a hydrogen bond with the NH group of the residue three positions farther along in the polypeptide chain.

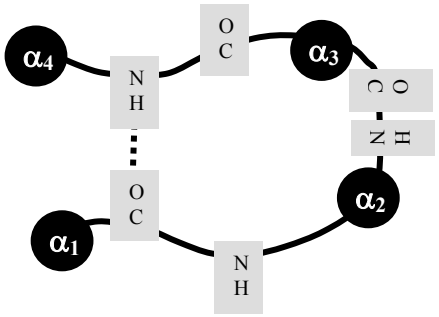


Figure 6. Beta bend

Tertiary structure of protein

Efficient packing of secondary structural elements is another important feature that leads to the tertiary structure of a protein. Alpha helices and beta sheets are packed together to form subunits or domains that are functional units of tertiary structure of a protein. In globular proteins, it has been frequently observed that adjacent alpha helices and beta sheets pack together and connected by loop regions to form a three dimensional protein structural motif as shown in Figure 7. The structural motifs are normally formed such that

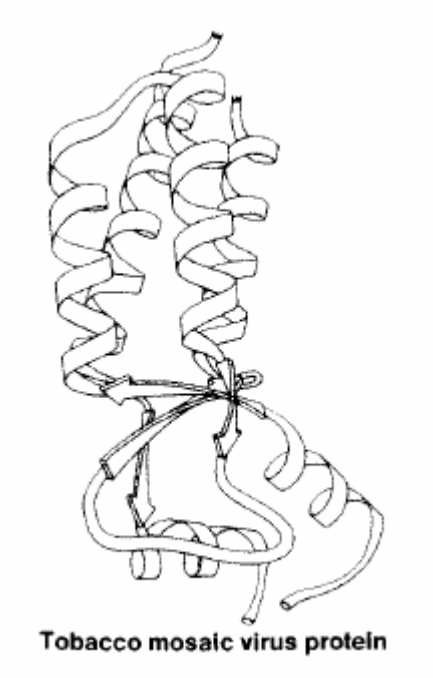


Figure 7. Three dimensional protein structural motif

they have a minimum accessible surface area. In some cases alpha helices may form complex and irregular geometries than structural motifs mentioned earlier. However, even in these cases it seems that the geometric restrictions that lead to close packing are still present.

Domains are classified into different main structural groups including alpha, beta and alpha/beta structures. In alpha structure, the core is built up exclusively from alpha helices. Beta structures comprise antiparallel beta sheets. Alpha/beta structures, consists of a combination of beta-alpha-beta motifs. The combination of domains in a single protein determines its overall function [29]. A protein may contain one or more structural domains. The domains of large proteins are usually connected by relatively flexible regions of polypeptide chains [31].

Natively disordered proteins - the new paradigm

Evidence is growing that dominant view of structure-function paradigm does not hold universally for all proteins. In contrast many proteins or region of proteins may function in the absence of 3-D structure. These proteins may lack specific 3-D structure and may be partially or completely unfolded in their native state. The terms natively denatured, intrinsically unstructured, or disordered proteins are used to describe these proteins. Natively disordered proteins or regions of protein usually have dynamic Φ and Ψ angles. Natively disordered proteins are characterized by X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, circular dichroism (CD) spectroscopy and protease sensitivity among many others.

Two ordered proteins with identical sequences would basically have the same structure. But two disordered proteins with the same sequence may each have a different conformation which may vary for each protein over time [32]. This is because molecules that make up natively disordered proteins do not reside in 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.

Natively disordered proteins are found in majority of species. Based on structure and function of these proteins, Tompa [33] has proposed to classify them as a separate entity. In general natively disordered proteins can be divided into two major groups: extended and collapsed [34]. The extended disorder refers to unfolded protein and regions that exist as random coil. While extended disordered proteins may have secondary structure, but due to fluctuation of Ramachandaran angles in the backbone, this structure is transient. The collapsed disorder refers to proteins and domains that resemble

molten globules which may have partially folded secondary structure with a dynamic tertiary structure [35]. Therefore, each protein may have three possible states: order, extended disorder, and collapsed disorder. Protein trinity refers to these three possibilities [36]. Proteins may change shape and take a form appropriate to any of the three states mentioned above depending on their environment.

Characterization of natively disordered proteins

Methods that are used to identify and characterize natively disordered proteins include NMR spectroscopy, X-ray crystallography, circular dichroism spectroscopy and protease sensitivity.

NMR spectroscopy

Nuclear magnetic resonance spectroscopy is a specific method that is used to identify structure and dynamic of natively disordered proteins. NMR can detect molecules which are moving rapidly. Natively disordered proteins have dynamic structures, i.e., they can convert between different states depending on the events through which they undergo. Thus, NMR can identify regions that are disordered as well as any transient secondary or tertiary structure that is present. NMR spectroscopy is associated with technical difficulties when it performs on molten globular proteins. Therefore, this technique is mostly used to detect extended disorder [4].

X-ray crystallography

Proteins that are ordered form crystals, but disordered proteins do not. When both ordered and disordered regions of a protein are subject to X-ray crystallography, the disordered proteins do not scatter x-rays the same way as ordered region do. This leads to missing electron density in the final structure [32]. As a result, completely disordered

proteins can not be studied by X-ray crystallography method. Also, disordered proteins are unlikely to form crystals in the first place. However, proteins consisting of both ordered and disordered regions can be studied because the ordered regions scatter x-rays, and the disordered regions occupy spaces between the ordered parts.

Circular dichroism (CD) spectroscopy

Proteins with tertiary structure can be detected by intense near-UV CD spectra while natively disordered proteins are characterized by low intensity near-UV CD spectra of low complexity. Far-UV CD spectrum is able to provide information about secondary structure of proteins. In circular dichroism spectroscopy, a combination of near- and far-UV CD is used to differentiate the ordered and disordered proteins (that is extended disorder and collapsed disorder). Circular dichroism spectroscopy can only provide information about presence of natively disordered regions but not about their locations within the sequence [4].

Protease sensitivity

Protease enzymes can be used to study natively disordered proteins. These enzymes digest specific sites of the unfolded protein sequence, thus, disordered proteins are digested rapidly. The rate of digestion of fully unfolded proteins can be on the order of 10^3 times faster than ordered proteins.

Due to the limitations associated with each of the method mentioned above, study of natively disordered proteins may require application of more than one method.

In biological systems disordered proteins do not degrade by proteases because they form a complex with binding partner and they are at least partially folded or in some cases are located in protease deficient regions of the cells.

A disordered protein example

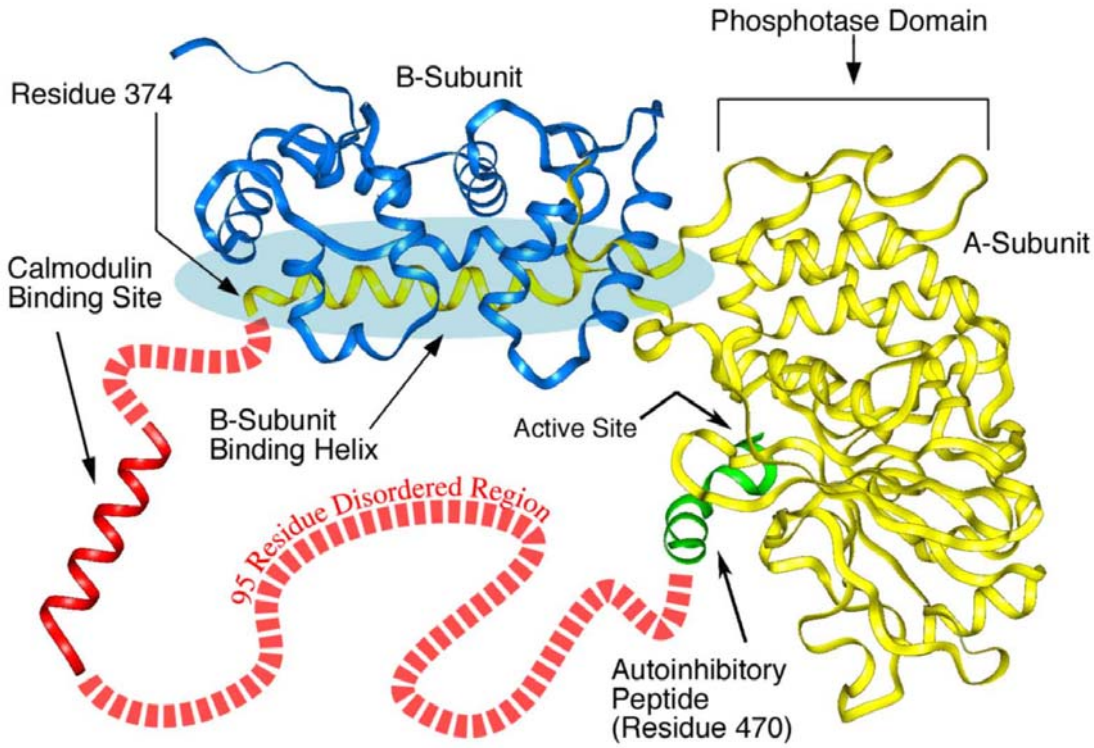
Calcineurin is involved in many biological responses including lymphocyte activation, neuronal and muscle development. Calcineurin is a major protein of the brain. It is a calcium calmodulin-dependent serine/threonine protein phosphatase. Calcineurin is composed of two catalytic A and B subunits. The A subunit contains the catalytic elements [37], a calmodulin binding domain [38], and autoinhibitory elements [39]. The B subunit is Ca^{2+}-binding protein which remains tightly associated with the A subunit. Calcium-calmodulin complex binds to the target helix within calcineurin, then autoinhibitory peptide disassociates from the active site and causes the phosphatase activity to turn on. Studies show that calcium-calmodulin complex wrap around the target helix, therefore, this region must lack tertiary structure and lie within the disordered regions.[40,41]

Figure 8 shows a 3-D structure of Calcineurin where A subunit is shown in yellow, B subunit is shown in blue, the autoinhibitory peptide in green, the location of 95-residue disorder region in red and camodulin binding site as red helix.

Frequency of natively disordered proteins

Studies show that on the average more than 30% of eukaryotic proteins and 4.2% of bacterial proteins are either completely or partially unfolded. Analysis of genomic sequence of different organisms shows that the proportion of sequence that code for natively disordered proteins depends on the complexity of the organism. Thus, disordered proteins are common in eukaryotes and not very common in bacteria [42].

Figure 8. 3-D structure of Calcineurin



Functions of disordered proteins

Although natively disordered proteins or regions of proteins lack specific order but they may have local and limited residual structure that allows them to interact and bind with different proteins, nucleic acids and membranes [43,44]. Studies show that disordered region might be of advantage to a protein because it allows efficient interaction with different regions of a single or a multiple target [45].

In a study of 115 disordered regions [5], twenty eight functions associated with 98 of these regions were identified. These functions were classified into four main groups as follows: molecular recognition, assembly/disassembly, protein modification and entropic chain activities

Molecular recognition

Many cellular activities such as gene expression and signal processing are dependent on dynamic and efficient macromolecular interactions which are facilitated by disordered regions. Protein binding, nucleic acid binding, and receptor-ligand binding are examples of these molecular interactions.

Assembly/disassembly is basically the same as molecular recognition and may be considered as its special case.

Protein modification

Protein modification, such as, phosphorylation, glycosylation, and methylation, occurs in the disordered regions. A recent study shows that chemical modification is also frequent in both RNA and protein chaperones [46].

Entropic chain activities

A collection of protein functions that depend on disordered regions without any induction of order is called “entropic chain activities”. This group of functions depends on the flexibility or rigidity of the disordered region.

Amino acid compositions of natively disordered proteins

Although functions of many proteins are determined by their 3-D structure, disordered proteins or regions possess biological functions, too. The sequences of natively disordered regions are evolutionary conserved and mainly consist of amino acids of low hydrophobicity with large net charge. The disordered regions may also have sequences of low complexity and high flexibility. [47]. The amino acid compositions of disordered proteins have a higher level of specific amino acids such as E,K,R,G,Q,S and P. They also have a lower level of the amino acids I,L,V,W,F,Y,C and N [47,48]. Based

on these specific amino acid composition disordered regions of proteins can be predicted. There are several programs that can identify these disordered regions.

Prediction of natively disordered proteins

Predictors of Natural Disordered Regions (PONDRs) is a neural network predictor that uses amino acid sequence data to predict disorder in a given region [48,49]. Basically PONDRs use sequence attributes taken over windows of 9 to 21 amino acids. The values used to train the neural network are average of attributes such as fractional compositions of the specific amino acids and hydropathy taken over these sequence windows around the residues of interest.

The earliest predictors of disordered proteins were the VL1 predictor [48], the N- and C- terminal predictors (XT) [50]. These predictors used feed forward neural networks to predict natively disorder proteins and had an relatively good accuracy (about 73%) against testing data [51]. Later on the PONDR VL-XT, which is a combination of the earliest versions, was developed. PONDR VL-XT uses neural networks to predict order–disorder class for every amino acid residue in a protein. The extensions added to PONDR describes the training data of each specific predictor and explained elsewhere [52].

The PONDR VL-XT was trained against long regions (40 or more residues) of disorder identified from regions missing in x-ray structures [48,50]. Additional predictors were developed using different neural networks as well as logistic regression. All of these predictors calculate values for different attributes of each amino acid residue and feed them into either a neural network or a linear predictor. The attributes used to predict the disorder amino acid residues are 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 [48]. PONDR VL-XT outputs for each residue are numeric value between 0 and 1. One is the ideal prediction for disorder and 0 is the ideal prediction for order. Usually PONDR VL-XT does not output these numbers, thus, a threshold of 0.5 is applied. Amino acid residues with values of greater or equal to 0.5 are assigned. Later on CDF (Cumulative Distribution Function) analysis from VL-XT predictor was developed to predict proteins which are completely disordered (wholly ordered). CDF summarizes the frequency of disorder scores from PONDR VL-XT. Based on a distribution of prediction scores it then classifies the protein as ordered or disordered [53].

The VL-XT predictor shows a higher accuracy to study short regions of either order or disorder. PONDR VL3 predictor that was used in this study was designed using longer sequence windows and showed a better prediction accuracy of order and disorder regions. PONDR VL3 has a high rate of accuracy of 85% and is based on averaging the outputs of an ensemble of 50 predictors. Therefore, it tends to predict disorder with less granularity [34].

Objective

As mentioned earlier, we would like to focus on the protein products of viral overlapping gene sequences in order to test our hypothesis that a pair of proteins encoded by overlapping genes have opposite order-disorder content. This means that an ordered amino acid on one sequence corresponds to a disordered amino acid on the other sequence. Moreover, the overlapping region of these proteins have a higher percentage of disordered amino acids than the non-overlapping region. In this study we will use 97

proteins (52 pairs) that are encoded by overlapping genes. We would like to predict the proportion of disordered proteins using PONDR VL3 predictor.

III. Materials and Methods

The available data for this research were in the form of a Microsoft Excel table. This table was provided to us by our collaborators from Architecture et Fonction des Macromolecules Biologiques (AFMB) at Ecole de l'AND in Marseille, France, who are studying viral proteins encoded by overlapping genes. The rows of the table were attributed to different viruses of interest. The numerous columns of the table provided information on different characteristics of each virus. We used the available data to investigate our hypothesis mentioned earlier. We accessed the computer hardware facilities as well as personal computers at the Center for Computational Biology and Bioinformatics at IUPUI in Indianapolis and a number of software to carry out this research.

Software

The software used for this research project is listed in Table 1. We also used Excel spreadsheet capabilities as needed. MySQL was used to create a relational database to store and query data.

MySQL is an open source relational database management system. It uses Structured Query Language (SQL), which is the most popular language for adding, accessing, searching and processing data in a database. Data definitions in SQL were used to create and alter the descriptions of the tables (or relations) of the database. SQL systems were used to specify primary keys and referential integrity constraints. Basic SQL queries like the select, delete, insert or alter statements were used for inserting information into and retrieving information from the database.

Table 1. List of software used for this research project

Name	Suppliers	Usage	License Terms
MySQL v4.1.14	http://www.mysql.com	Relational database	GNU General Public License
PONDR® VL3	http://www.pondr.com	Prediction of order/disorder for protein sequences	Individually licensed from Molecular Kinetics
XEmacs v21.3.1	http://www.xemacs.org	Text editor	GNU General Public License
Perl v5.8.0	http://www.activestate.com	Perl interpreter	GNU General Public License
EditPlus v2.12	http://www.editplus.com	Text editor	Individually licensed from Dawn Roberts

PONDR® VL3 was the software that we used to calculate the residual values for each amino acid in the protein sequence. We could use these residual values to predict order and disorder of the protein sequence.

We used XEmacs to prepare perl scripts that were needed to carry out some of the tasks and calculations for the project.

Perl v5.8.0 was used as interpreter to run the perl scripts.

Data analysis

Analysis of data required the following steps:

1. Data management
2. Database design, construction and implementation
3. Populating and query the database
4. Prediction of order/disorder of amino acid sequence encoded by overlapping genes using PONDRs

5. Extraction of the information related to the order/disorder of amino acid sequence.

Data management

As mentioned earlier data on viral proteins encoded by overlapping genes were provided to us our collaborators in France. The data included single and double stranded RNA viruses as well as a number of circular and linear DNA viruses as shown in Table 2.

Table 2. Dataset of Viral Proteins

Overlap	Acc number	Taxonomy	Organism	Genome type	Overlapped CDS 1 / Product 1	Overlapped CDS 2 / Product 2
1	**NC_001915**	Viruses; dsRNA viruses; Birnaviridae; Aquabirnavirus	Infectious pancreatic necrosis virus	ds-RNA & linear	**1**/ VP5 (modified, second ATG is used)	**2**/ polyprotein
2	**NC_004178**	Viruses; dsRNA viruses; Birnaviridae; Avibirnavirus	Infectious bursal disease virus	ds-RNA & linear	**1**/ VP5 protein	**2**/ VP2-4-3 polyprotein
3	**NC_004267**	Viruses; dsRNA viruses; Reoviridae; Orthoreovirus; Mammalian orthoreoviruses	Mammalian orthoreovirus 1	ds-RNA & linear	**1**/ minor capsid cell attachment protein sigma-1a	**2**/ nonstructural protein sigma-ibNS
4	**NC_003771**	Viruses; dsRNA viruses; Reoviridae; Oryzavirus	Rice ragged stunt virus	ds-RNA & linear	**1**/ RNA-dependent RNA polymerase	**2**/ P4b
5	**NC_003768**	Viruses; dsRNA viruses; Reoviridae; Phytoreovirus	Rice dwarf virus	ds-RNA & linear	**1**/ nonstructural protein	**2**/ OP-ORF (new)
6	**NC_001641**	Viruses; dsRNA viruses; Totiviridae; Totivirus	Saccharomyces cerevisiae virus L-BC (La)	ds-RNA & linear	**1**/ capsid	**2**/ RNA polymerase
7	**NC_001927**	Viruses; ssRNA negative-strand viruses; Bunyaviridae; Orthobunyavirus	Bunyamwera virus	ss-RNA & linear	**1**/ N protein	**2**/ NSs protein

Overlap	Acc number	Taxonomy	Organism	Genome type	Overlapped CDS 1 / Product 1	Overlapped CDS 2 / Product 2
8	**NC_001498**	Viruses; ssRNA negative-strand viruses; Mononegavirales; Paramyxoviridae; Paramyxovirinae; Morbillivirus	Measles virus	ss-RNA & linear	**1**/ phosphoprotein	**2**/ nonstructural C protein
9	**NC_002199**	Viruses; ssRNA negative-strand viruses; Mononegavirales; Paramyxoviridae; Paramyxovirinae	Tupaia paramyxovirus	ss-RNA & linear	**2**/ phosphoprotein	**3**/ nonstructural protein V
10	”	”	”	”	**2**/ phosphoprotein	**4**/ nonstructural protein C
11	**NC_005339**	Viruses; ssRNA negative-strand viruses; Mononegavirales; Paramyxoviridae; Paramyxovirinae	Mossman virus	ss-RNA & linear	**2**/ phosphoprotein	**3**/ V protein
12	”	”	”	”	**2**/ phosphoprotein	**4**/ C protein
13	**NC_001552**	Viruses; ssRNA negative-strand viruses; Mononegavirales; Paramyxoviridae; Paramyxovirinae; Respirovirus	Sendai virus	ss-RNA & linear	**2**/ C' protein	**3**/ P protein (co-factor of RNA polymerase)
14	”	”	”	”	**3**/ P protein (co-factor of RNA polymerase) [3/4]	**4**/ V protein (new)
15	**NC_002200**	Viruses; ssRNA negative-strand viruses; Mononegavirales; Paramyxoviridae; Paramyxovirinae; Rubulavirus	Mumps virus	ss-RNA & linear	**2**/ phoshoprotein	**3**/ V protein
16	**NC_001560**	Viruses; ssRNA negative-strand viruses; Mononegavirales; Rhabdoviridae; Vesiculovirus	Vesicular stomatitis Indiana virus	ss-RNA & linear	**2**/ NS protein	**3**/ Cprim protein (new)

Overlap	Acc number	Taxonomy	Organism	Genome type	Overlapped CDS 1 / Product 1	Overlapped CDS 2 / Product 2
17	**NC_002534**	Viruses; ssRNA positive-strand viruses, no DNA stage; Nidovirales; Arteriviridae; Arterivirus	Lactate dehydrogenase-elevating virus	ss-RNA & linear	**3**/ structural glycoprotein	**4**/structural glycoprotein
18	”	”	”	”	**4**/ structural glycoprotein	**5**/ structural glycoprotein
19	**NC_001633**	Viruses; ssRNA positive-strand viruses, no DNA stage; Barnaviridae; Barnavirus	Mushroom bacilliform virus	ss-RNA & linear	**1**/ nd	**2**/ nd
20	”	”	”	”	**2**/ nd	**3**/ nd
21	**NC_002035**	Viruses; ssRNA positive-strand viruses, no DNA stage; Bromoviridae; Cucumovirus	Cucumber mosaic virus	ss-RNA & linear	**1**/ RNA-dependent RNA polymerase	**2**/2b protein
22	**NC_003809**	Viruses; ssRNA positive-strand viruses, no DNA stage; Bromoviridae; Ilarvirus	Spinach latent virus	ss-RNA & linear	**2**/ putative polymerase	**2**/ putative 2b protein
23	**NC_001749**	Viruses; ssRNA positive-strand viruses, no DNA stage; Flexiviridae; Capillovirus	Apple stem grooving virus	ss-RNA & linear	**1**/241k polyprotein	**2**/36K protein
24	**NC_003499**	Viruses; ssRNA positive-strand viruses, no DNA stage; Flexiviridae; Carlavirus	Blueberry scorch virus	ss-RNA & linear	**5**/ Coat protein	**6**/16 kDa protein (putative nucleic acid-binding protein)
25	**NC_003093**	Viruses; ssRNA positive-strand viruses, no DNA stage; Flexiviridae; Mandarivirus	Indian citrus ringspot virus	ss-RNA & linear	**5**/ capsid protein CP	**6**/ putative 23 kDa nucleic acid binding protein
26	**NC_001642**	Viruses; ssRNA positive-strand viruses, no DNA stage; Flexiviridae; Potexvirus	Bamboo mosaic virus	ss-RNA & linear	**1**/ replicase	**2**/ hypothetical 14k protein
27	**NC_001658**	Viruses; ssRNA positive-strand viruses, no DNA stage; Flexiviridae; Potexvirus	Cassava common mosaic virus	ss-RNA & linear	**3**/ triple gene block protein 2	**4**/ triple gene block protein 3

Overlap	Acc number	Taxonomy	Organism	Genome type	Overlapped CDS 1 / Product 1	Overlapped CDS 2 / Product 2
28	**NC_001409**	Viruses; ssRNA positive-strand viruses, no DNA stage; Flexiviridae; Trichovirus	Apple chlorotic leaf spot virus	ss-RNA & linear	**2**/ movement protein	**3**/ coat protein
29	**NC_001434**	Viruses; ssRNA positive-strand viruses, no DNA stage; Hepatitis E-like viruses	Hepatitis E virus	ss-RNA & linear	**9**/ nd	**10**/ nd
30	**NC_003481**	Viruses; ssRNA positive-strand viruses, no DNA stage; Hordeivirus	Barley stripe mosaic virus	ss-RNA & linear	**3**/ beta C protein	**4**/ beta D protein
31	**NC_004730**	Viruses; ssRNA positive-strand viruses, no DNA stage; Pecluvirus	Indian peanut clump virus	ss-RNA & linear	**4**/ P14 protein	**5**/ P17 protein
32	**NC_003725**	Viruses; ssRNA positive-strand viruses, no DNA stage; Pomovirus	Potato mop-top virus	ss-RNA & linear	**2**/ triple-gene-block protein 2	**3**/ triple-gene-block protein 3
33	**NC_002568**	Viruses; ssRNA positive-strand viruses, no DNA stage; Sobemovirus	Sesbania mosaic virus	ss-RNA & linear	**2**/ polyprotein	**4**/ coat protein
34	**NC_004366**	Viruses; ssRNA positive-strand viruses, no DNA stage; Umbravirus	Tobacco bushy top virus	ss-RNA & linear	**3**/ unknown	**4**/ unknown
35	**NC_004146**	Viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; Alphanodavirus	Flock house virus	ss-RNA & linear	**1**/ protein A	**3**/ protein B2
36	**NC_003448**	Viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; Betanodavirus	Striped Jack nervous necrosis virus	ss-RNA & linear	**1**/ protein A	**2**/ protein B
37	**NC_005094**	Viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; unclassified Nodaviridae	Macrobrachium rosenbergii nodavirus	ss-RNA & linear	**1**/ RNA-dependent RNA polymerase	**2**/ B2 protein
38	**NC_001366**	Viruses; ssRNA positive-strand viruses, no DNA stage; Picornaviridae; Cardiovirus	Theilovirus	ss-RNA & linear	**1**/ viral polyprotein	**2**/ viral protein L (new)

Overlap	Acc number	Taxonomy	Organism	Genome type	Overlapped CDS 1 / Product 1	Overlapped CDS 2 / Product 2
39	**NC_001990**	Viruses; ssRNA positive-strand viruses, no DNA stage; Tetraviridae; Betatetravirus	Nudaurelia capensis beta virus	RNA & linear	**1**/ RNA-dependent RNA polymerase	**2**/ capsid protein
40	**NC_005899**	Viruses; ssRNA positive-strand viruses, no DNA stage; Tetraviridae; unclassified Tetraviridae	Dendrolimus punctatus tetravirus	ss-RNA & linear	**1**/ p17	**2**/ capsid protein p71
41	**NC_000939**	Viruses; ssRNA positive-strand viruses, no DNA stage; Tombusviridae; Aureusvirus	Pothos latent virus	ss-RNA & linear	**4**/ hypothetical protein, 27K	**5**/ hypothetical protein, 14K
42	**NC_003608**	Viruses; ssRNA positive-strand viruses, no DNA stage; Tombusviridae; Carmovirus	Hibiscus chlorotic ringspot virus	ss-RNA & linear	**1**/ RNA-dependent RNA polymerase	**3**/ P28 protein
43	”	”	”	”	**6**/ coat protein	**7**/ hypothetical protein
44	**NC_003627**	Viruses; ssRNA positive-strand viruses, no DNA stage; Tombusviridae; Machlomovirus	Maize chlorotic mottle virus	ss-RNA & linear	**4**/ p31 protein	**6**/ coat protein
45	**NC_003487**	Viruses; ssRNA positive-strand viruses, no DNA stage; Tombusviridae; Necrovirus	Tobacco necrosis virus D	ss-RNA & linear	**3**/7 kDa protein	**4**/7 kDa protein
46	**NC_003532**	Viruses; ssRNA positive-strand viruses, no DNA stage; Tombusviridae; Tombusvirus	Cymbidium ringspot virus	ss-RNA & linear	**4**/ putative movement protein	**5**/ core protein p19
47	**NC_004063**	Viruses; ssRNA positive-strand viruses, no DNA stage; Tymoviridae; Tymovirus	Turnip yellow mosaic virus	ss-RNA & linear	**1**/ overlapping protein/movement protein	**2**/ replicase/papain-like protease
48	**NC_001574**	Viruses; Retroid viruses; Caulimoviridae; Badnavirus	Cacao swollen shoot virus	DNA & circular	**3**/ polyprotein	**5**/ hypothetical protein
49	**NC_001719**	Viruses; Retroid viruses; Hepadnaviridae; Orthohepadnavirus	Arctic ground squirrel hepatitis B virus	DNA & circular	**1**/ e antigen precursor	**3**/ polymerase

Overlap	Acc number	Taxonomy	Organism	Genome type	Overlapped CDS 1 / Product 1	Overlapped CDS 2 / Product 2
50	”	”	”	”	**3**/ polymerase	**4**/ large envelope protein
51	”	”	”	”	**3**/ polymerase	**7**/ X protein
52	**NC_004324**	Viruses; Retroid viruses; Caulimoviridae; Caulimovirus	Cestrum yellow leaf curling virus	DNA & linear	**2**/ putative virion associated protein	**3**/ putative capsid protein

Table 2 shows some information on 52 pairs. The data provided to us by our collaborators include other information in addition to what is provided in Table 2. The descriptions of the information related to viruses are given in Table 3. A number of viruses in Table 2 share the same name, identification number, taxonomy, genome and overlap identification but different protein identification. In the database provided, in the column “protein number” we see a(b) and b(a) in the rows related to the same virus identification number. These rows refer to overlapping proteins.

Table 3. Description of Information on Viruses in Our Dataset

Virus Attribute	Description
Acc number	Virus identification number (included in Table 2)
Taxonomy	Virus Taxonomy
Organism	Virus name (included in Table 2)
Genome	Virus genome (included in Table 2)
Strain	Virus strain
GI	Gene information identifier
Tax_id	NCBI taxonomy identifier
genome length (bp)	Size of virus genome
Overlapping CDS	Overlap identification (included in Table 2)
begin(bp) overlap	The base pair number where overlap begins
end(bp) overlap	The base pair number where overlap ends
protein number	Refers to overlap identification (included in Table 2)
product	Protein product (included in Table 2)
sense of protein	Refers to sense strand that expresses the protein
prot_id of protein	Protein identification
GI of protein	Gene information identifier
lgth(bp) of protein	Length of DNA that produces the protein
lgth(aa) of protein	Protein length
seq aa of protein	Amino acid sequence of protein
begin(aa) of overlapping	Beginning, or start position, of an overlapping protein(amino acid) sequence
seq aa overlapping	Overlapping protein (amino acid) sequence
end(aa) of overlapping	End position of an overlapping protein (amino acid) sequence
length(aa) of overlapping	Length of an overlapping protein (amino acids)

Database design, construction and implementation

A relational database is basically a collection of organized tables. The process of designing a database involves a number of steps. Based on the nature of our research, we needed to build a relational database to store and retrieve information about viral

proteins. To access and retrieve the information we had to organize the data in relevant tables. First we identified the entities of our data model and decided that we needed four tables to organize the attributes of the entities as will be described later. Finally, we identified the relationship between the entities and the unique identifiers that facilitated the implementation of our data model.

VIRUS database schema:

Database schema describes the structure of tables and the relationship among them. The schema for VIRUS database consists of four tables as shown in Figure 9. This schema includes four tables, VIRUS_DES, CDS, OVERLAP and PONDRS. The table VIRUS_DES (virus description) includes general information on the virus as follows:

- Class: Refers to virus genome
- virusid: Refers to virus identification number (Acc number)
- name: Refers to virus name
- taxo: Refers to virus taxonomy
- gi : Refers to gene information identifier
- taxid: Refers to the NCBI taxonomy identifier
- lgthgenome: Refers to the size of virus genome

Classification and taxonomy of the viruses have been established by the International Committee on Taxonomy of Viruses (ICTV).

Table CDS includes information on genome sequence of the virus and the encoded proteins as follows:

- cdsprotid: Refers to protein id
- cdsgi : Refers to coding sequence

- overlapid: Refers to overlap id
- cdssense: Refers to gene sense strand
- cdscomplete: Refers to complete coding sequence
- cdsproduct: Refers to protein product
- cdsbegAN: Refers to the beginning or start position of gene (nucleotide) sequence
- cdsendAN: Refers to the end position of gene (nucleotide) sequence
- cdslgthAN: Refers to the length of gene (nucleotide) sequence
- cdslgthAA: Refers to the length of protein (amino acids)
- cdsfullseqAA: Refers to the sequence of protein.
- cdsfullseqAN: Refers to the sequence of the viral gene that expresses the above protein .

Figure 9. Schema for VIRUS database

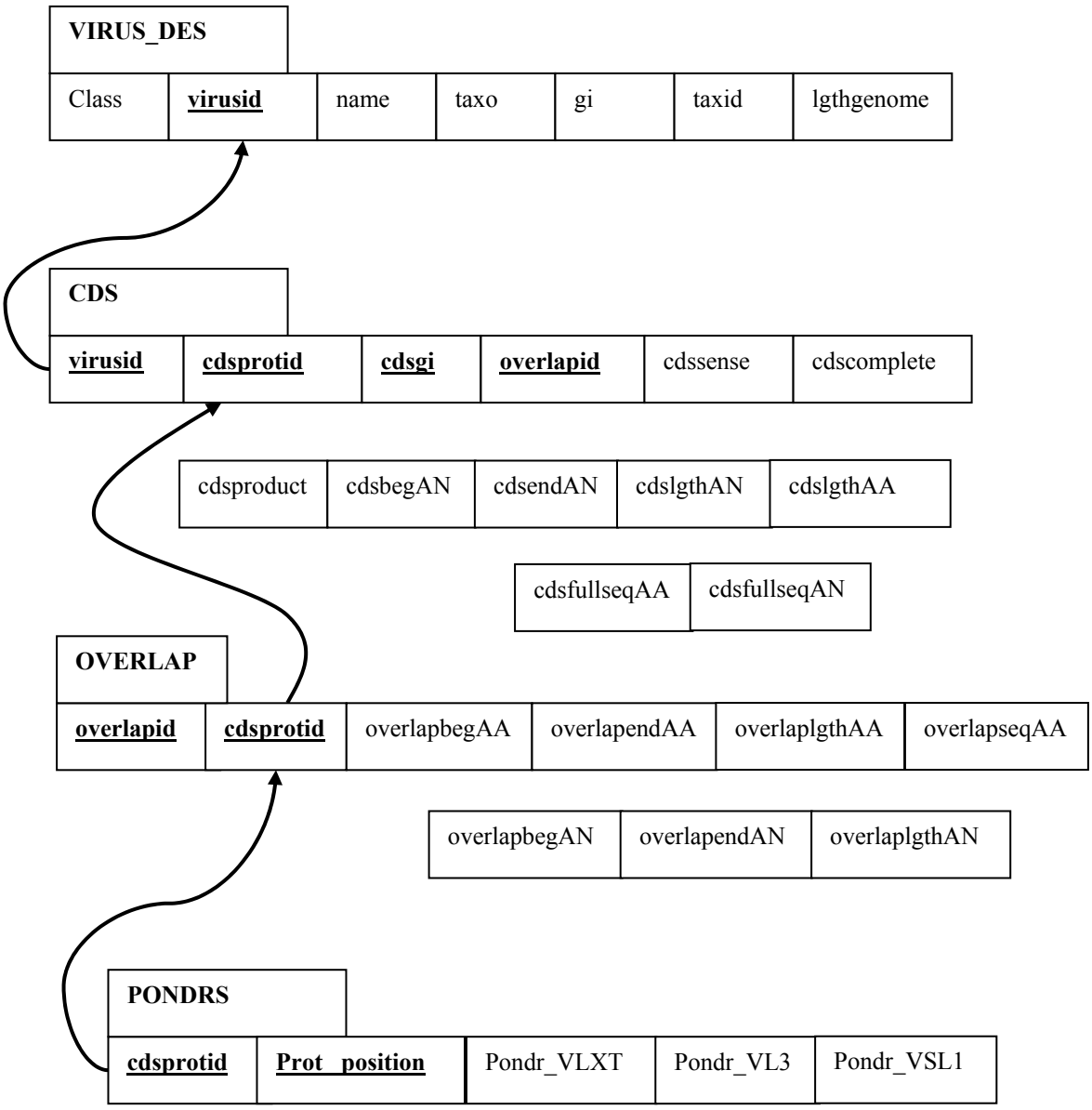


Table OVERLAP includes information on overlapping genes and their encoded proteins as follows:

- overlapbegAA: Refers to the beginning, or start position, of an overlapping protein (amino acid) sequence

- overlaplgthAA : Refers to the length of an overlapping protein (amino acids)
- overlapseqAA : Refers to an overlapping protein (amino acid) sequence
- overlapbegAN : Refers to the beginning or start position of an overlapping gene (nucleotide) sequence
- overlapendAN : Refers to the end position of an overlapping gene (nucleotide) sequence
- overlaplgthAN :Refers to the length of an overlapping gene (nucleotide) sequence

Table PONDRS includes the results predicted by the predictor as follows. Pondr_VL3 was used in this study but Pondr_VLXT and Pondr_VSL1 were added for future research.

- Prot_ position : Refers to position of amino acid sequence
- Pondr_VL3 : Refers to PONDR VL3
- Pondr_VLXT : Refers to PONDR VLXT
- Pondr_VSL1: Refers to PONDR VSL1

Each of the four tables shown in Figure 9 includes a primary key, identified by bold font. A primary key is used to uniquely identify tuples (rows) in a table and can be one or more attributes in the table. We have linked the tables by foreign keys as shown in Figure 9.

VIRUS database construction:

SQL commands were used to create tables, rows and columns as follows:

```
create table VIRUS_DES (
class text not null,
id varchar (20)  not null,
name text not null,
taxo text not null,
gi text not null,
taxid text not null,
lgthgenome text not null,
Primary key (id)
);
```

```
create table CDS
(
virusid varchar(30) not null,
cdsprotid varchar(30) not null,
cdsgi varchar (30) not null,
cdssense text not null,
cdscomplete text not null,
cdsproduct text not null,
cdsbegAN text not null,
cdsendAN text not null,
cdslgthAN text not null,
cdsfullseqAA text not null,
cdsfullseqAN text not null,
Primary key (virusid, cdsgi, cdsprotid),
foreign key (virusid) references VIRUS (id)
);
```

```
create table OVERLAP
(
overlapid varchar (30) not null,
cdsprotid varchar (30) not null,
overlap_begAA text,
overlap_endAA text,
overlap_lgthAA text,
overlap_seqAA text,
overlap_begAN text,
overlap_endAN text,
overlap_lgthAN text,
Primary key (cdsprotid, overlapid),
foreign key (cdsprotid) references CDS (virusid,
cdsgi, cdsprotid)
);
```

```
create table PONDRS
(
cdsprotid varchar (30) not null,
prot_position int (11),
pondr_VLXT float,
pondr_VL3 float,
pondr_VSL1 float,
Primary key (cdsprotid, prot_position),
foreign key (cdsprotid) references CDS (virusid,
cdsprotid, cdsgi)
);
```

Database implementation

Our data model was implemented in the open source relational database management system MySQL on a linux platform as shown in Figure 10.

Figure 10. Physical Schema Of DNA_VIRUS Database

Tables in DNA Virus database
CDS ()
OVERLAP ()
PONDRS()
VIRUS_DES()

CDS table

Field	Type	Null	Key	Default	Extra
virusid	varchar(30)		Primary		
cdsprotid	varchar(30)		Primary		
cdsgi	varchar(30)		Primary		
overlapid	varchar(30)		Primary		
cdssense	text				
cdscomplete	text				
cdsproduct	text				
cdsbegAN	text				
cdsendAN	text				
cdslgthAN	text				
cdsfullseqAA	text				
cdsfullseqAN	text				

OVERLAP Table

Field	Type	Null	Key	Default	Extra
overlapid	varchar(30)		Primary		
cdsprotid	varchar(30)		Primary		
overlap_begAA	text	yes		null	
overlap_endAA	text	yes		null	
overlap_lgthAA	text	yes		null	
overlap_seqAA	text	yes		null	
overlap-begAN	text	yes		null	
overlap_endAN	text	yes		null	
overlap_lgthAN	text	yes		null	

PONDRS Table					
Field	Type	Null	Key	Default	Extra
Cdsprotid	varchar(30)				
Prot_position	int(11)			0	
Pondr_VLXT	float	Yes		null	
Pondr_VL3	float	Yes		null	
Pondr_VSLI	float	yes		null	

VIRUS_DES Table					
Field	Type	Null	Key	Default	Extra
class	text				
virusid	varchar(20)		primary		
name	text				
taxo	text				
gi	text				
taxid	text				
lgthgenome	text				

Populating and query the database

Data were first filtered to remove nucleotide redundancy and then used to populate all tables of the VIRUS database except PONDRS table. PONDRS table is populated after protein prediction. Structured query language (SQL) was used to communicate with the database, do queries and extract the data that are stored in the database.

Database queries

Before connecting the database to PONDRs we performed a number of queries to test the database and extract information for protein sequence analysis. Examples of queries are shown in Figure 11.

Figure 11. Examples of queries

```
mysql> use VIRUS;
Database changed
mysql> show tables;
+--------------------+
| Tables_in_VIRUS |
+--------------------+
| CDS                |
| OVERLAP            |
| PONDRS             |
| VIRUS_DES          |
+--------------------+
4 rows in set (0.00 sec)

mysql> select overlapid from OVERLAP;

+--------------------------+
| overlapid                |
+--------------------------+
| OVERLAP; NC_001366-1(2)  |
| OVERLAP; NC_001409-2(3)  |
| OVERLAP; NC_001409-3(2)  |
| OVERLAP; NC_001560-2(3)  |
| OVERLAP; NC_001574-3(5)  |
| OVERLAP; NC_001574-5(3)  |
| OVERLAP; NC_001633-1(2)  |
| OVERLAP; NC_001633-2(1)  |
| OVERLAP; NC_001633-3(2)  |
| OVERLAP; NC_001641-1(2)  |
| OVERLAP; NC_001641-2(1)  |
| OVERLAP; NC_001642-1(2)  |
| OVERLAP; NC_001642-2(1)  |
| OVERLAP; NC_001658-3(4)  |
| OVERLAP; NC_001658-4(3)  |
| OVERLAP; NC_001719-1(3)  |
| OVERLAP; NC_001719-4(3)  |
| OVERLAP; NC_001719-7(3)  |
| OVERLAP; NC_001749-1(2)  |
| OVERLAP; NC_001749-2(1)  |
| OVERLAP; NC_001915-2(1)  |
| OVERLAP; NC_001927-1(2)  |
| OVERLAP; NC_001927-2(1)  |
| OVERLAP; NC_001990-1(2)  |
| OVERLAP; NC_001990-2(1)  |
| OVERLAP; NC_002035-1(2)  |
| OVERLAP; NC_000939-4(5)  |
| OVERLAP; NC_000939-5(4)  |
| OVERLAP; NC_002199-2(3)  |
| OVERLAP; NC_002199-3(2)  |
| OVERLAP; NC_002199-4(2)  |
| OVERLAP; NC_002200-2(3)  |
| OVERLAP; NC_002200-3(2)  |
| OVERLAP; NC_001434-9(10) |
| OVERLAP; NC_001434-10(9) |
```

| OVERLAP; NC_001552-2(3) |
| OVERLAP; NC_001552-3(2) |
| OVERLAP; NC_001498-2(3) |
| OVERLAP; NC_001498-3(2) |
| OVERLAP; NC_002534-3(4) |
| OVERLAP; NC_002534-4(3) |
| OVERLAP; NC_002534-5(4) |
| OVERLAP; NC_002568-2(4) |
| OVERLAP; NC_002568-4(2) |
| OVERLAP; NC_003093-5(6) |
| OVERLAP; NC_003093-6(5) |
| OVERLAP; NC_003448-1(2) |
| OVERLAP; NC_003448-2(1) |
| OVERLAP; NC_003481-3(4) |
| OVERLAP; NC_003481-4(3) |
| OVERLAP; NC_003487-3(4) |
| OVERLAP; NC_003487-4(3) |
| OVERLAP; NC_003499-5(6) |
| OVERLAP; NC_003499-6(5) |
| OVERLAP; NC_003532-4(5) |
| OVERLAP; NC_003532-5(4) |
| OVERLAP; NC_002035-2(1) |
| OVERLAP; NC_003608-1(3) |
| OVERLAP; NC_003608-6(7) |
| OVERLAP; NC_003608-7(6) |
| OVERLAP; NC_003627-4(6) |
| OVERLAP; NC_003627-6(4) |
| OVERLAP; NC_003725-2(3) |
| OVERLAP; NC_003725-3(2) |
| OVERLAP; NC_003768-1(2) |
| OVERLAP; NC_003771-1(2) |
| OVERLAP; NC_003771-2(1) |
| OVERLAP; NC_003809-1(2) |
| OVERLAP; NC_003809-2(1) |
| OVERLAP; NC_004063-1(2) |
| OVERLAP; NC_004063-2(1) |
| OVERLAP; NC_004146-1(3) |
| OVERLAP; NC_004146-3(1) |
| OVERLAP; NC_004178-1(2) |
| OVERLAP; NC_004178-2(1) |
| OVERLAP; NC_004267-1(2) |
| OVERLAP; NC_004267-2(1) |
| OVERLAP; NC_004366-3(4) |
| OVERLAP; NC_004366-4(3) |
| OVERLAP; NC_004730-4(5) |
| OVERLAP; NC_004730-5(4) |
| OVERLAP; NC_004324-2(3) |
| OVERLAP; NC_004324-3(2) |
| OVERLAP; NC_005094-1(2) |
| OVERLAP; NC_005094-2(1) |
| OVERLAP; NC_005339-2(3) |
| OVERLAP; NC_005339-3(2) |
| OVERLAP; NC_005339-4(2) |
| OVERLAP; NC_001915-1(2) |
| OVERLAP; NC_003768-2(1) |
| OVERLAP; NC_005339-2(4) |
| OVERLAP; NC_001552-3(4) |

```
| OVERLAP; NC_001552-4(3)  |
| OVERLAP; NC_001560-3(2)  |
| OVERLAP; NC_002534-4(5)  |
| OVERLAP; NC_001633-2(3)  |
| OVERLAP; NC_001366-2(1)  |
| OVERLAP; NC_003608-3(1)  |
| OVERLAP; NC_001719-3(1)  |
| OVERLAP; NC_001719-3(4)  |
| OVERLAP; NC_001719-3(7)  |
| OVERLAP; NC_005899-1(2)  |
| OVERLAP; NC_005899-2(1)  |
+--------------------------+
97 rows in set (0.02 sec)
```

Prediction of proteins encoded by overlapping genes

Amino acid sequences encoded by overlapping genes were converted into fasta format. Fasta format start with a title line which starts with a ">" symbol followed by lines of amino acid sequence data. The length of fasta formatted amino acid sequence data is 60 amino acid. An example of fasta format is shown in Figure 12.

Figure 12. Example of fasta format of protein sequences

```
>NP_690838\
MTNLQDQTQQIVPFIRSLLMPTTGPASIPDDTLEKHTLRSETSTYNLTVGDTGSGLIVFF
PGFPGSIVGAHYTLQSNGNYKFDQMLLTAQNLPASYNYCRLVSRSLTVRSSTLPGGVYAL
NGTINAVTFQGSLSELTDVSYNGLMSATANINDKIGNVLVGEGVTVLSLPTSYDLGYVRL
GDPIPAIGLDPKMVATCDSSDRPRVYTITAADDYQFSSQYQAGGVTITLFSANIDAITSL
SIGGELVFQTSVQGLILGATIYLIGFDGTAVITRAVAADNGLTAGTDNLMPFNIVIPTSE
ITQPITSIKLEIVTSKSGGQAGDQMSWSASGSLAVTIHGGNYPGALRPVTLVAYERVATG
SVVTVAGVSNFELIPNPELAKNLVTEYGRFDPGAMNYTKLILSERDRLGIKTVWPTREYT
DFREYFMEVADLNSPLKIAGAFGFKDIIRALRRIAVPVVSTLFPPAAPLAHAIGEGVDYL
LGDEAQAASGTARAASGKARAASGRIRQLTLAADKGYEVVANLFQVPQNPVVDGILASPG
ILRGAHNLDCVLREGATLFPVVITTVEDAMTPKALNSKMFAVIEGVREDLQPPSQRGSFI
RTLSGHRVYGYAPDGVLPLETGRVYTVVPIDGVWDDSIMLSKDPIPPIVGSSGNLAIAYM
DVFRPKVPIHVAMTGALNAYGEIENVSFRSTKLATAHRLGLKLAGPGAFDVNTGSNWATF
IKRFPHNPRDWDRLPYLNLPYLPPNAGRQYDLAMAASEFKETPELESAVRAMEAAANVDP
LFQSALSVFMWLEENGIVTDMANFALSDPNAHRMRNFLANAPQAGSKSQRAKYGTAGYGV
EARGPTPEGAQREKDTRISKKMETMGIYFATPEWVALNGHRGPSPGQLKYWQNTREIPDP
NEDYLDYVHAEKSRLASEGQILRAATSIYGAPGQAEPPQAFIDEVAKVYEVNHGRGPNQE
QMKDLLLTAMEMKHRNPRRAPPKPKPKPNVPTQRPPGRLGRWIRAVSDEDLE
```

To study and analyze protein sequences that are encoded by overlapping genes we ran PONDRs to predict the entire amino acid sequence of each sample and generate output data of PONDR_VL3. Using the output data we populate the table PONDR in database DNA_VIRUS using the Perl script given in Appendix 1.

Extraction of proteins encoded by overlapping genes

Next a Perl script was written to extract the predicted proteins of the overlapping genes. A sample output is shown in Figure 13. At the top of the output in Figure 13, the virus identification numbers NC_000939-4(5) and NC_000939-5(4) are shown followed by the select statements referring to the overlap locations. Next the amino acid sequence on the overlapping protein pair is given followed by the position of the amino acid on sequence 1, the residual values of the amino acids in both sequences, and the position of the amino acid on sequence 2. The residual value equal or greater than 0.5 indicates a predicted disorder and less than 0.5 indicates order.

Figure 13. A sample output

```
OVERLAP; NC_000939-4(5)
OVERLAP; NC_000939-5(4)
select * from OVERLAP where overlapid = "OVERLAP; NC_000939-5(4)"
select * from PONDRS where cdsprotid = "NP_051033" AND prot_position >=
44 AND prot_position <= 173
130 rows returned
select * from PONDRS where cdsprotid = "NP_051034" AND prot_position >=
1 AND prot_position <= 130
130 rows returned
130
```

	seq1pos	O/d	O/d	seq2pos
Amino acids: K M	44	O (0.168255)	d (0.628742)	1
Amino acids: W E	45	O (0.163555)	d (0.626855)	2
Amino acids: K N	46	O (0.15771)	d (0.624686)	3
Amino acids: I S	47	O (0.151991)	d (0.621748)	4
Amino acids: P Q	48	O (0.147354)	d (0.61707)	5
Amino acids: K T	49	O (0.143412)	d (0.610165)	6
Amino acids: Q G	50	O (0.140759)	d (0.604253)	7
Amino acids: G V	51	O (0.138932)	d (0.601007)	8
Amino acids: F L	52	O (0.139971)	d (0.59991)	9
Amino acids: Y C	53	O (0.142626)	d (0.599018)	10
Amino acids: A P	54	O (0.145522)	d (0.596876)	11
Amino acids: P N	55	O (0.145944)	d (0.593222)	12
Amino acids: I R	56	O (0.14448)	d (0.590186)	13
Amino acids: D C	57	O (0.144306)	d (0.589928)	14
Amino acids: V Q	58	O (0.148619)	d (0.592494)	15
Amino acids: K V	59	O (0.158146)	d (0.593509)	16
Amino acids: F C	60	O (0.168943)	d (0.594164)	17
Amino acids: V S	61	O (0.179586)	d (0.593942)	18
Amino acids: L H	62	O (0.187421)	d (0.594516)	19

Our study included protein prediction for overlapping and non-overlapping regions. The focus of this study was extracting the predicted proteins of the overlapping regions. We also performed some protein prediction in non-overlapping regions.

IV. Results

The ordered or disordered amino acids in the proteins encoded by overlapping genes were predicted by PONDR VL3 and the results were obtained. The above predictions include both overlapping and non-overlapping regions of proteins. As mentioned earlier we were given 52 protein pairs mentioned in Table 2 to analyze.

Overlapping regions of protein pairs

First, the results for the overlapping regions of protein pairs were considered. The output of PONDR VL3 for one of these protein pairs, identified by NC_000939-4(5) and NC_000939-5(4), in the overlapping region is presented in Appendix 2. In this output, the first section includes SQL statements to extract the encoded proteins. These statements first refer to the protein pair identifier followed by other identifications. At the end of this section we can read the number of amino acid pairs, which is 130 in this case. In the second section of the output that follows the dotted lines, amino acid names, sequence positions and PONDR VL3 predictions for both sequences in the overlapping region are shown. For example, the first amino acid pair shown in Appendix 2 is K and M. The K amino acid is at position 44 on sequence 1 and the M amino acid is at position 1 on sequence 2. The prediction is that sequence 1 is ordered, identified by (O) with the amino acid residual value of 0.168255 and sequence 2 is disordered, identified by (d) with the amino acid residual value of 0.628742. The results in section 2 include 130 amino acid pairs.

At the end of the rows for 130 pairs of amino acid the total number of predicted disordered and ordered amino acids for sequence 1 and sequence 2 are calculated and shown followed by their respective percentages. The total number of disordered (d)

amino acids on sequence 1 is given as19. This is followed by the total number of amino acids on sequence 1 which are not disordered (O) and is 111. Same data for sequence 2 follow. After that the percents of (d) and (O) on each sequence are given. Finally, the last five lines in Appendix 2 include numbers related to pair analysis where both sequences are considered.

The results indicate that out of the total overlap length of 130, there are 63 amino acid pairs where at least one amino acid on either sequence is (d). The rest 67 amino acid pairs are (O-O), i.e., ordered on both sequence. As shown in Appendix 2, from the 63 amino acid pairs with at least one disorder, 0 is (d-d), 19 are (d-O) and 44 are (O-d). To clarify the pair analysis, the following example is useful. Consider two sequences shown below. Here amino acid sequence 1 is OOOddOdO and amino acid sequence 2 is ddOddOdd. If we consider these two sequences in pair opposite to each other like:

```
amino acid sequence 1 is    OOOddOdO
amino acid sequence 2 is    ddOddOdd
```

Then the pair analysis would be as follows:

Number of amino acid pairs where both are ordered (O-O)	2
Number of amino acid pairs where there is at least one disorder (d-O, O-d, d-d)	6
Number of amino acid pairs where both are disordered (d-d)	3

The results of order and disorder predictions for the protein pair in Appendix 2 were imported to Excel spreadsheet together with the respective protein identifiers. In our spreadsheet, the row that belongs to the protein pair in Appendix 2 would look as shown in Figure 14.

Figure 14. Results of protein prediction in Excel Spreadsheet

ID		P-ND	P-BSD	P-SQ1D	P-SQ2D	P-Total
NC_000939-4(5)	NC_000939-5(4)	67	0	19	44	130

In our Excel spreadsheet we adopted the following abbreviations:

Abbreviation	Description
P-ND	Amino acid pairs where both ordered (O-O)
P-BSD	Amino acid pairs where both disordered (d-d)
P-SQ1D	Amino acid pairs where amino acid on sequence 1 is disordered and amino acid on sequence 2 is ordered (d-O)
P-SQ2D	Amino acid pairs where amino acid on sequence 2 is disordered and amino acid on sequence 1 is ordered (O-d)
P-Total	Total number of amino acid pairs in overlap

The generated output for 52 protein pairs by PONDR VL3 were imported to Excel spreadsheet. The results are reported in Appendix 3. The row related to the sample in Appendix 2 is shaded.

The percentages of O-O, d-d, d-O and O-d for each overlapping protein pair were calculated in Excel as shown in Figure 15. In this figure, P-TD refers to the percent of amino acid pairs in the overlap where there is at least one disorder (d-d, d-O and O-d). P-TD is equal to the sum of P-BSD, P-SQ1D and P-SQ2D. The sum of P-ND and P-TD should be 100%.

Figure 15. Percentages of order and disorder calculated by Excel Spreadsheet

							%					
ID		P-ND	P-BSD	P-SQ1D	P-SQ2D	P-Total	P-ND	P-BSD	P-SQ1D	P-SQ2D	P-TD	Total
NC_000939-4(5)	NC_000939-5(4)	67	0	19	44	130	51.54	0.00	14.62	33.85	48.46	100.00

These percentages were plotted as a bar chart as shown in Figure 16. The x-axis in Figure 16 is the number of overlapping protein pairs (52). In the chart in Figure 16 we start with 22 protein pairs that have no O-O, i.e., the amino acids pairs are either d-d, d-O or O-d. The 23rd protein pair has 5.5% O-O and the remaining 94.5% has at least one disorder (d-d, d-O or O-d). As is observed in Figure 16 the percentage of O-O increases until it becomes 100% in 50th protein pair. In Figure 17 the breakdown of percentages of O-d, d-O, d-d and O-O for all overlapping protein pairs are shown.

Figure 16. Analysis of Order-Disorder for Overlap Proteins

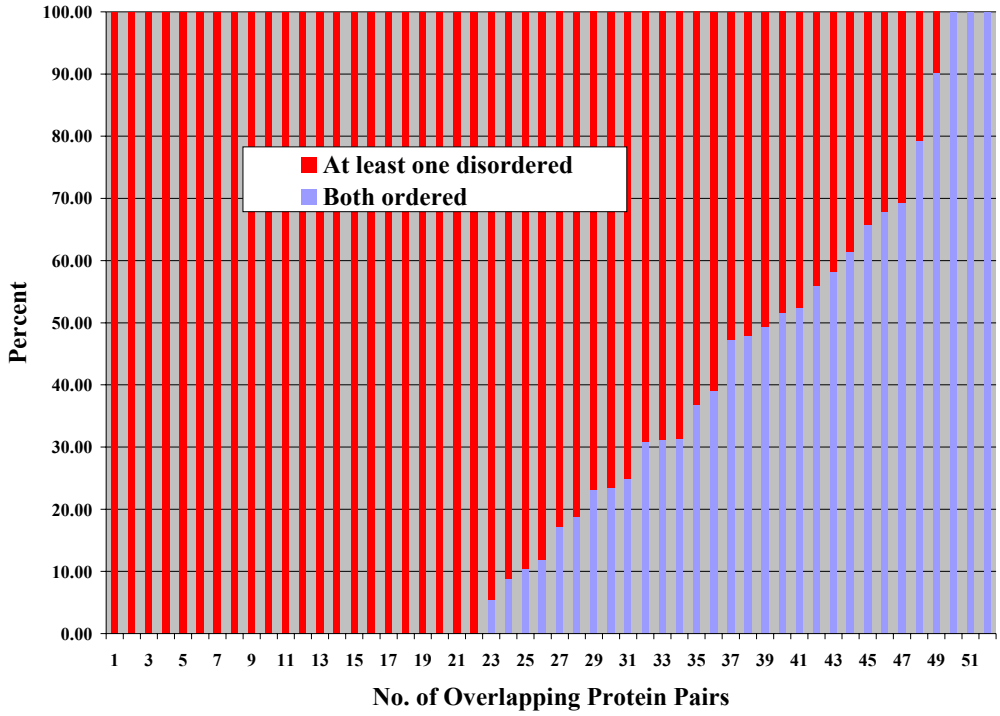
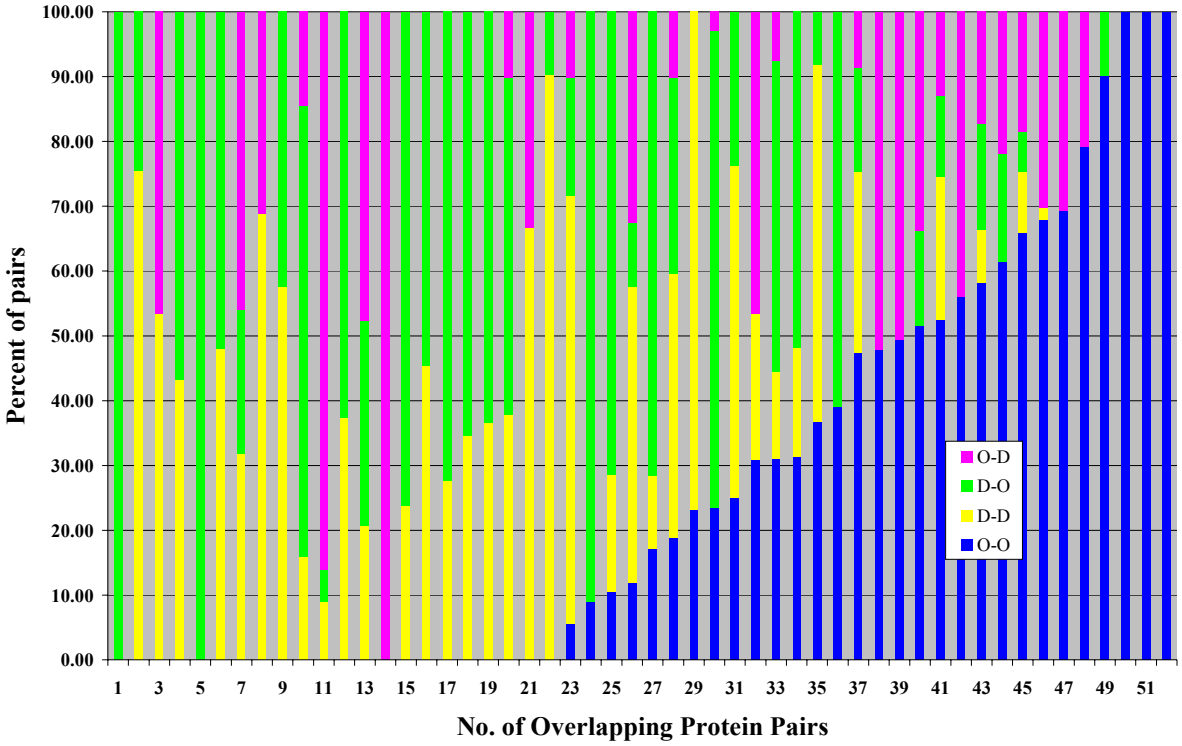


Figure 17. Analysis of Overlap Proteins with Disorder Breakdown



Non-overlapping regions of proteins

We then considered the output of PONDR VL3 for non-overlapping regions of 52 protein pairs. A sample of PONDR VL3 output for two proteins with protein numbers NP_051033 and NP_051034 is presented in Appendix 4. These two proteins belong to the protein pair to which we referred in Appendix 2. Data in Appendix 4 shows the residual values for each amino acids in the entire protein sequence, i.e., overlapping and non-overlapping regions for the two proteins. The overlapping region is shaded in Appendix 4. Out of 52 protein pairs, or 104 samples, 7 samples generated repeating order and disorder information which was redundant. Therefore, those 7 samples were removed from the pool for the analysis of non-overlapping region. The results generated by

PONDR VL3 for 97 proteins were imported to Excel spreadsheet and tabulated. The tabulated results in our spreadsheet for two protein pairs would look as shown in Figure 18. This figure shows the length, orders and disorders for the proteins in the overlap region (which was earlier discussed), followed by the similar information for the entire sequence and non-overlap region.

Figure 18. Results of protein prediction for non-overlap region

cdsprotid	Length of overlap section	O in overlap section	D in overlap section	Length of entire sequence	D in entire sequence	O in entire sequence	Length of nonoverlap section	D in nonoverlap section	O in nonoverlap section
NP_051033	130	111	19	242	88	154	112	69	43
NP_051034	130	86	44	130	44	86	0	0	0

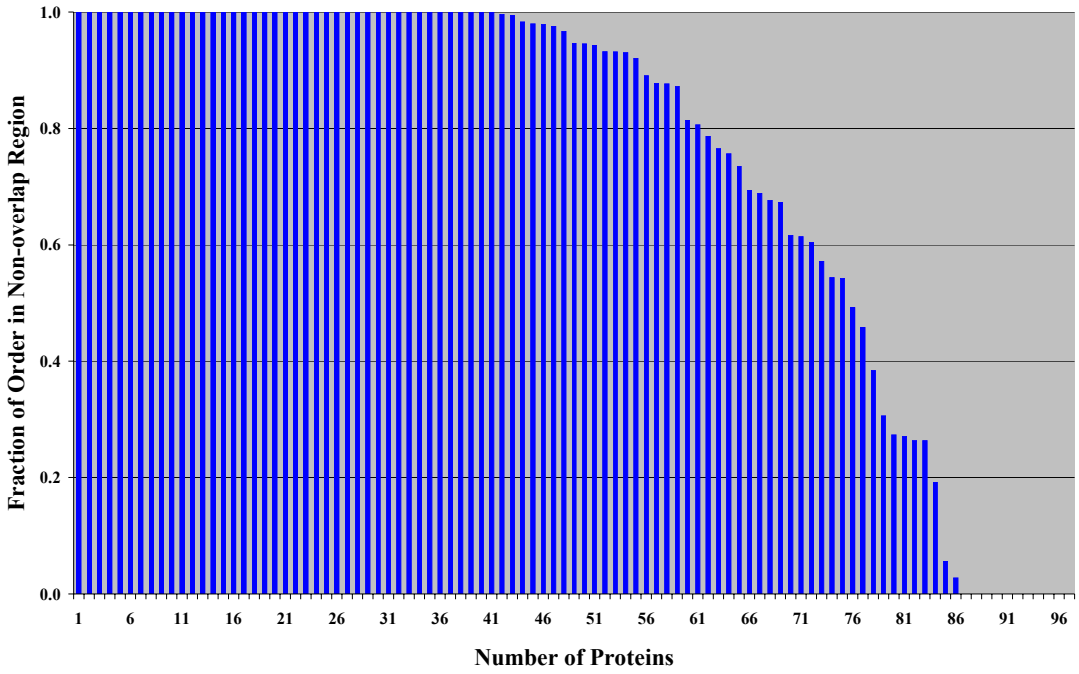
The generated results for non-overlap regions for 97 proteins are reported in Appendix 5. The row related to the sample in Appendix 4 is shaded. Using the data in Appendix 5 we calculated the fraction of ordered amino acids (O) in each protein sequence. We sorted data to the proteins with the higher fraction of (O) in a descending order. A plot of the data is shown in Figure 19. A global fraction of (O) in the entire pool of 97 proteins was also calculated as 0.77.

Bootstrapping

The results generated above were based on a dataset with 97 viral proteins or 52 protein pairs. To test the reliability of our results we applied bootstrapping. This is a statistical method used to examine if a particular dataset is biased. We used boostrapping in two cases, first for the fraction of O-O pairs in the overlapping region and second the fraction of disorder in the entire sequence. In our boostrapping, we used a repeated

random sampling with replacement from our original sample of 52 protein pairs or 97 proteins to provide 10,000 new pseudoreplicate samples, from which sampling variance and margin of error could be estimated at a given confidence level.

Figure 19. Fraction of Ordered Amino Acids in the Non-overlap Region of Proteins



V. Discussion

Overlapping genes are adjacent genes which share a portion of their nucleotide sequence. They are often observed in compact genome of viruses, prokaryotic genome, and organelles like mitochondria. They may also be present in human and other mammalian genome. These organisms take advantage of overlapping genes to produce new proteins without increasing the size of genome. Overlapping genes produce different proteins. A major work in post genomic era is large-scale study of structures and functions of proteins. Although, in general, it is assumed that 3-D structure of a protein determines the function of proteins, but many proteins or regions of proteins may function in the absence of 3-D structure. The term disordered is used to describe these proteins. Based on a large number of studies, biological functions depend on both ordered and disordered proteins.

Disordered regions of proteins can be predicted using specific amino acid composition of these regions. There are several programs that can identify these disordered regions. PONDR VL3, a neural network predictor that uses amino acid sequence data to predict disorder in a given region, was used in this study.

In the results section we performed studies on 97 proteins (52 protein pairs) encoded by overlapping gene to decide the order or disorder of amino acids in the sequence of each protein. The length of amino acid sequence in overlapping regions for the above proteins were at least 31 and at most 626. Also we analyzed each protein sequence for the percentage of disordered amino acids in its overlapping and non-overlapping regions. The entire length of amino acid sequence for the above proteins,

including the overlapping and non-overlapping regions, were at least 62 and at most 2303.

As mentioned earlier, based on our hypothesis, most often, in a pair of proteins encoded by overlapping genes at least one is disordered (unstructured). This is believed to be attributed to the creation of these proteins by overprinting the sequence of a pre-existing gene. The overprinting mechanism may impose a constraint where the genetic code would not allow encoding of two structured proteins in different reading frames.

Figure 20 shows that there are only 3 protein pairs out of 52, indicated by 100% blue bars, where there are no disorder on either sequence of the pair. Moreover, as highlighted in Figure 20, for about 39 protein pairs out of 52, the length of blue bar (percent of O-O) is less than 50%.

Figure 20 shows that there are 22 protein pairs where there is no O-O (indicated by 100% red bar). There are 40 pairs where there are some O-O amino acid pairs (indicated by blue bar). The total number of amino acid pairs in the overlapping region for all the proteins under study (52 pairs) is 7219 and the total number of amino acid pairs which are ordered (O-O) in the overlapping regions for all proteins is 2014. Therefore, the global fraction of O-O pairs in the overlapping region would be obtained by dividing 2014 by 7219 which is 0.28. Bootstrapping of 10,000 random samples using the data on our 52 protein pairs shows that this fraction, i.e., fraction of O-O pairs, would also be 0.28 with 95% confidence.

The above results indicate that according to our hypothesis, for 52 pairs of proteins encoded by overlapping genes that were studied, most often, at least one is disordered and the O-O pairs are less than 30%.

Figure 20. Overlapping protein pairs with more than 50% disorder

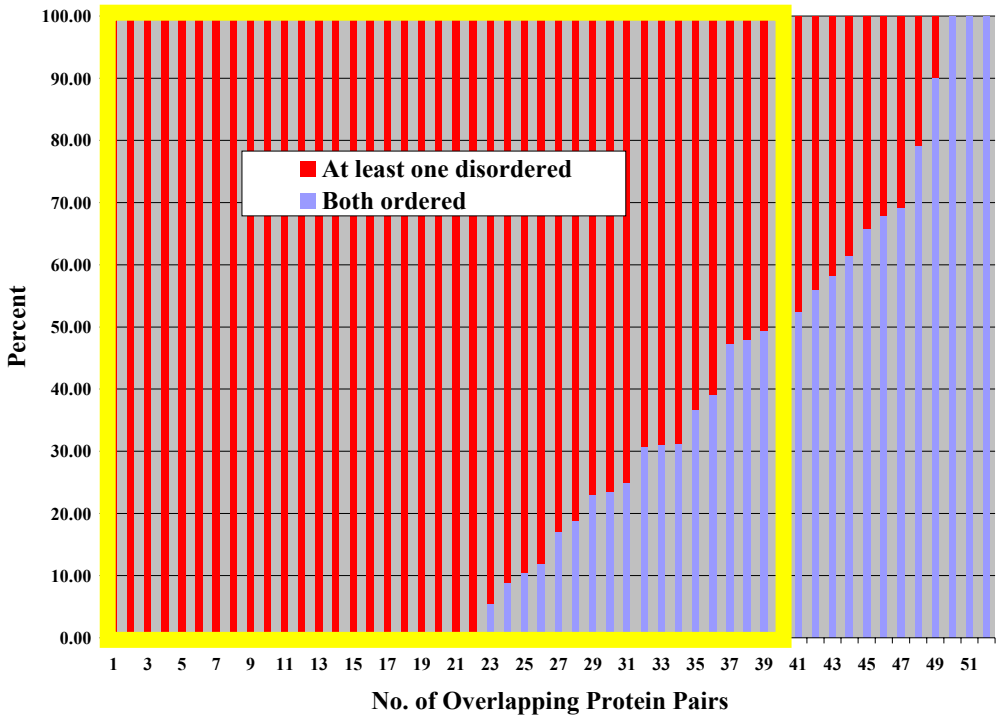
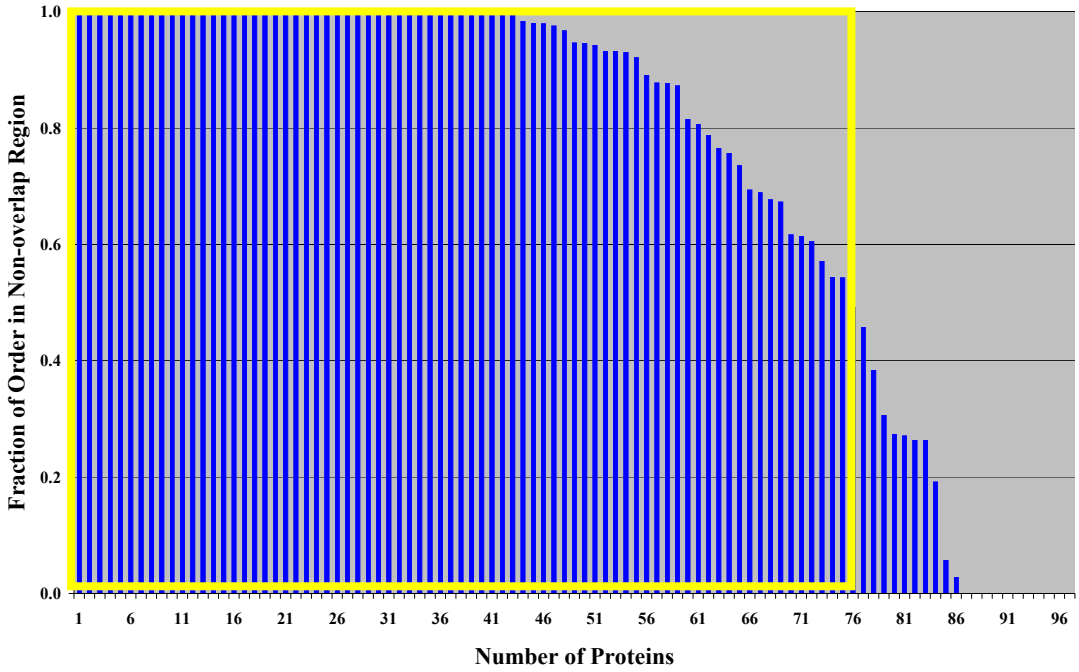


Figure 21 shows that out of 97 proteins, there are 80 proteins for which the percent of ordered amino acids (O) in the non-overlapping region, indicated by red bars, is greater than 50% (as highlighted). Data in Appendix 5 indicate that the total number of ordered amino acids in the non-overlapping region is 25428 and the total length of the non-overlapping region is 32946. This will result in a fraction of ordered amino acids in the non-overlapping region equal to 0.77. This is another indication that the ordered amino acids are mostly associated with the non-overlapping region while the disorderd amino acids are prevalent in overlapping region. This is another support for our hypothsis.

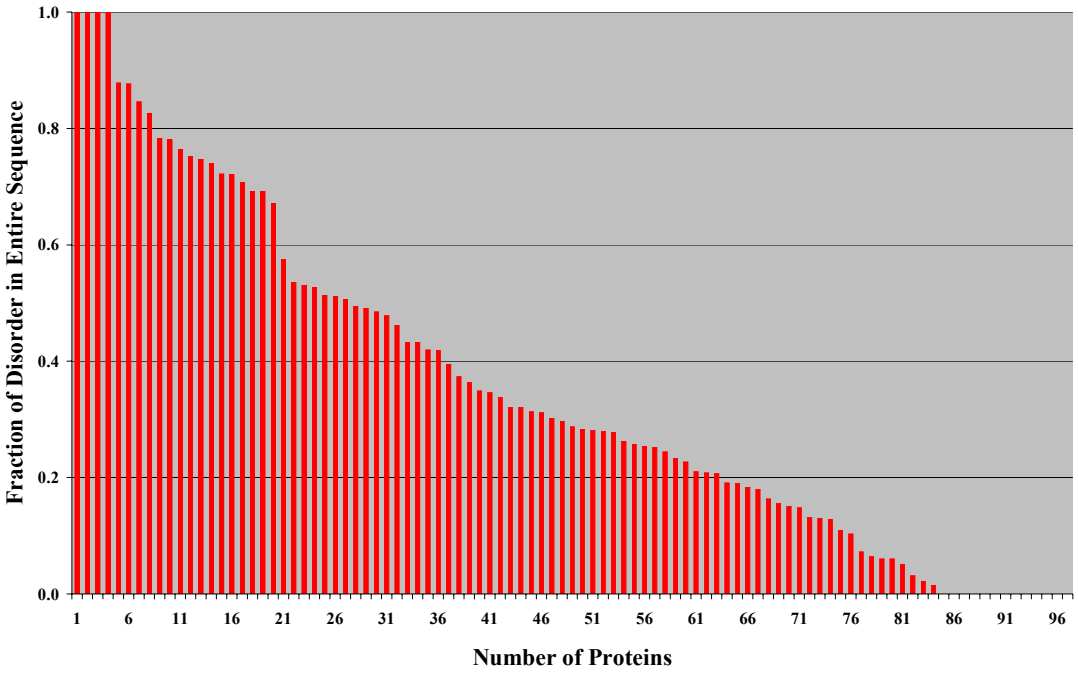
Further analysis was performed on the entire sequence of proteins which included both overlapping and non-overlapping regions. The fraction of disorder in the entire

Figure 21. Fraction of Ordered Amino Acids in the Non-overlap Region of Proteins



sequence of proteins is shown in Figure 22. As shown in Appendix 5 the total length of the entire sequences of 97 proteins under study is 47543 and the number of disordered amino acids in these sequences is 14476. This will result in a fraction of disordered amino acids in the entire sequence equal to 0.30. Bootstrapping of 10,000 random samples using the data for the entire sequence of 97 proteins shows that the fraction of disorder in the entire sequence would be 0.31 with 95% confidence.

Figure 22. Fraction of Disordered Amino Acids in the Entire Sequence of Proteins



VI. Conclusions

In our study we focused on the proteins encoded by overlapping genes. These proteins are produced due to frame shift phenomenon where changes in the reading frame of a nucleotide sequence leads to the production of different amino acid sequences.

Unlike most proteins that have a 3-D structure, the majority of proteins that are encoded by overlapping genes, are unstructured and thus, called disordered. Although, in general, it is assumed that 3-D structure of a protein determines the function of proteins, but based on a large number of studies, biological functions depend on both ordered and disordered proteins.

We developed a method to predict and analyze the proteins encoded by overlapping genes. Our method included design, construction and implementation of a database, populating and query of the database and finally prediction of proteins encoded by overlapping genes using the database and the protein predictor PONDR VL3. Using our method we could predict the order-disorder of amino acids in the sequence of 97 viral protein samples that were provided to us. The results we generated in this study were tabulated and analyzed to provide the number and fraction of ordered and disordered amino acids in the overlapping, non-overlapping regions and the entire sequence of 97 protein samples under study.

The objective of our study was to investigate our hypothesis that most often, in a pair of proteins encoded by overlapping genes at least one is disordered (unstructured). In another word, in the sequence of proteins produced by overlapping genes, an ordered amino acid on one sequence corresponds to a disordered amino acid on the other sequence in most of the cases. This is believed to be attributed to constraints where the

genetic code would not allow encoding of two structured proteins in different reading frames.

We considered 97 samples given to us as 52 pairs in one set of analysis and as 97 protein sequences in a different analysis. When each of the 52 protein pairs were considered, we showed that most of the amino acid pairs facing each other on the protein sequences had at least one disorder for most cases. There were only 3 protein pairs out of 52 where there were no disorder on either sequence of the protein pair. On the other hand, there were 22 protein pairs out of 52 where there were no O-O amino acid pair and in 39 protein pairs, the percent of O-O amino acid pairs was less than 50%. The global fraction of O-O pairs in the pool of overlapping regions of 52 protein pairs was 0.28. Bootstrapping of 10,000 random samples with 95% confidence also resulted in the same fraction. The pair analysis of proteins encoded by overlapping genes, supported our hypothesis

When 97 proteins were considered one sequence at a time, there were 80 proteins for which the percent of ordered amino acids in the non-overlapping region, was greater than 50%. The fraction of ordered amino acids in the pool of 97 proteins in their non-overlapping regions was calculated to be 0.77. This is another indication that the ordered amino acids are mostly associated with the non-overlapping region while the disorderd amino acids are prevalent in overlapping region. This is another support for our hypothsis.

VII. Recommendations for future work

We recommend to expand this study by applying the methodology developed in this work to new datasets of different organisms. Moreover, amino acid composition of the overlapping genes could be studied to see which amino acids promote order or disorder.

It has been shown that overlapping genes may play an important role as transcriptional and translational regulators of gene expression, which in turn, determine the function of proteins. In the future protein product of overlapping genes could be studied to see what kind of function they might be involved.

In this study we observed that an average 28% of the overlaps were O-O. The relationship between the function of protein and percent O-O in the overlap can be studied. We also saw a limited number that had 100% O-O in their overlap. The location of this overlap can be studied to see if it occurred in the middle or at either ends of the entire protein sequence. The work on 100% O-O in overlap can be further extended by a homologue study using blast search.

Our hypothesis can be studied using protein products of overlapping genes in related species and comparisons can be made using the results of our viral protein study.

Many human diseases are associated with overlapping genes because of anomalous sequence features in this region. Moreover, it has been reported that 80% of cancer-associated proteins predicted to have large regions of disorder. Study of protein products of this region could shed further light on the possible role of disordered proteins produced by overlapping genes in human diseases

VIII. References

1. Mirsky, A.E., and Pauling, L., On the structure of native, denatured and coagulated proteins. *Proc Natl Acad Sci USA* 1936;22:439-47.

2. Hagerman, P.J.I., From sequence to structure to function, *Curr. Opin. Struct. Biol.*, 1996; 6(3):277-280.

3. Orengo, C.A., and Todd, A.E., From protein structure to function, structure to function. *Curr. Opin. Struct. Biol.*, 1999; 9:374-382.

4. 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., Interinsically disordered protein. *J.Mol. Graph. Model.* 2001; 19: 26-59.

5. Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M., and Obradovic, Z., Intrinsic disorder and protein function. *Biochemistry*, 2002; 41(21): 6573-6582.

6. Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z., and Dunker, A.K., Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, 2002;323(3):573-584.

7. Thomson, M.W., McInnes, R.R., and Willard, H.F., Genetics in Medicine, W.B. Saunders Company, 5th Edition 1991.

8. Barrell, B.G., Air, G.M., and Hutchison, C.A., Overlapping genes in bacteriophage-Psix174. *Nature,* 1976; 264: 34-41.

9. Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M., and Smith, M., Nucleotide-sequence of bacteriophage Phichi174 DNA. *Nature* 1977; 265: 687-695.

10. Shintani, S., OhUigin, C., Toyosawa, S., Michalova, V., and Klein, J., Origin of overlap gene, The case of TCP1 and ACAT2. *Genetics* 1999; 152:743-754.

11. Keese, P.K., and Gibbs, A., Origine of genes: "Big bang" or continuous creation? *Proc. Natl. Acad. Sci.* 1992; 89:9489-9493.

12. Fukuda, Y., Washio, T., and Tomita, M., Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae. Nucleic Acids Res.* 1999; 27: 1847-1853.

13. Fukuda, Y., Nakayama, Y., and Tomita, M., On dynamics of overlapping genes in bacterial genomes. *Gene* 2003; 323:181-187.

14. Krakauer, D.C., Evolutionary principles of genomic compression. Comments on Theor. Biol. 2002.

15. Normark, S., Bergstrom, S., Edlund, T., Grundstrom, T., Jaurin, B., Lindberg, F.P., and Olsson, O., Overlapping genes. *Annu. Rev. Genet.* 1983; 17:499-525.

16. Kiyosawa, H. and Abe, K., Speculations on the role of natural antisense transcripts in mammalian X chromosome evolution. *Cytogenet. Genome Res.* 2002; 99:151-156.

17. Fukuda, Y., Washio, T., and Tomita, M., Evolution of overlapping genes: Comparative genomic of *Mycoplasma genitalium* and *Mycoplasma pneumoniae.* The ninth workshop in Genome Informatics 1998.

18. Spencer, C.A., Gietz, R.D., and Hodgetts, R.B., Overlapping transcription units in the Dopa decarboxylase region of *Drosophila. Nature* 1986; 322: 279-281.

19. Williams, T. and Fried, M., A mouse locus at which transcription from both DNA strands produces messenger-RNAs complementary at their 3' ends. *Nature* 1986; 322: 275-279.

20. Wellington, C.L., Bauer, C.E., and Beatty, J.T., Photosynthesis gene superoperons in purple nonsulfur bacteria: The tip of the iceberg. *Can. J. Microbiol.* 1992; 38: 20-27.

21. Miyata, T., and Yasunaga, T., Evolution of overlapping genes. Nature. 1978;272:532-535

22. Rogozin, I., Spiridonov, A., Sorokin, A., Wolf, Y., King, J., Tatusov, R., and Koonin, E., Purrifying and directional selection in overlapping prokaryotic genes. *Trends Genet.* 2002; 18(5):228-232.

23. Karlin, S., Chen, C., Gentles, A., and Cleary, M., Associations between human disease genes and overlapping gene groups and multiple amino acid runs. *PNAS* 2002; 26:17008-17013.

24. Fischer, E., Einfluss der configuration auf die wirkung der enzyme. *Berichte Deutsch chemische Gesellschaft* 1894; 27:2985-93.

25. Wu, H., Studies on denaturation of proteins XIII: a theory of denaturation. *Chinj physiol* 1931; 1:219-34.

26. Anfinsen, C.B., Principles that govern the folding of protein chains. *Science,* 1973; 181:223-230.

27. Kendrew, J.C., Dickerson, R.E., et al. Structure of myoglobin: a three-dimensional Fourier synthesis at 2 A resolution. *Nature* 1960; 206:757-763.

28. Williams, R.J., The conformational mobility of proteins and its functional significance. *Biochem. Soc. Trans.* 1978; 6:1123-1126.

29. Carl Branden, C., and Tooze, J., Introduction to Protein Structure, Garland Publishing Inc, 1991.

30. Stryer, L., Biochemistry, 2nd Edition, 1975, W.H. Freeman and Company

31. Zubay, G., Biochemistry, 2nd Edition, 1988, Macmillan Publishing Company.

32. 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.

33. Tompa, P., Intrinsically unstructured proteins. *Trends Biochem Sci* 2002; 27:527-33.

34. Romero, P., Obradovic, Z., and Dunker, A.K., Natively disordered proteins. *Appl. Bioinformatics* 2004; 3(2-3):105-113.

35. Kuwajima, K., The molten globule state as a clue for understanding the folding and cooperativity of globular protein structure. Proteins 1989; 6:87-103.

36. Dunker, A.K., Obradovic, Z., The protein trinity -- linking function and disorder. *Nat. Biotechnol*. 2001; 19(9):805-806.

37. Winkler, M.A., Merat, D.L., Tallant, E.A., Hawkins, S., and Cheung, W.Y., *Proc. Natl. Acad. Sci. USA* 1984; 81:3054-3058.

38. Kincaid, R.L., Nightingale, M.S., and Martin, B.M., *Proc. Natl. Acad. Sci. USA* 1988; 85:8983-8987.

39. Hashimoto, Y., Perrino, B.A., and Soderling, T.R., J. Biol. Chem. 1990; 265:1924-1927.

40. Klee, C.B., Crouch, T.H., and Krinks, M.H., Calcineurin: acalcium- and calmodulin-binding protein of the nervous system. *Proc. Natl. Acad. Sci USA* 1979; 76:6270-6273.

41. Klee, C.B., Draetta, G.F., and Hubbard, M.J., *Enzymol.Relat. Areas Mol. Biol.* Calcineurin. 1988; 61:149-200.

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

43. Gunasekaran, K, Tsai, C.J., Kumar, S., Zanuy, D., and Nussinov, R., Extended disordered proteins: targeting function with less scaffold. *Trends Biochem Sci* 2003; 28:81-85.

44. Fuxreiter, M., Simon, I., Friedrich, P., and Tompa, P., Preformed structuralelements feature in partner recognition by intrinsically unstructured proteins. *J.Mol.Biol* 2004; 338:1015-1026.

45. Liu, J., Lu, M., An aniline-zipper structure determined by long range intermolecular interactions. J. Biol. Chem. 2002; 277:48708-48713.

46. Tampa, P., and Csermely, P., The role of structural disorder in the function of RNA and protein chaperones. FASEB J. 2004; 18:1169-1175.

47. Uversky, V.N., Gillespie, J.R., and Fink, A.L., Why are ‘‘natively unfolded’’proteins unstructured under physiologic conditions? *Proteins* 2000; 41:415-427.

48. Romero, P., Obradovic, Z., Li, X., Garner, E., Brown, C., and Dunker, A.K., Sequence complexity of disordered protein. *Proteins: Structure Funct Gen* 2001; 42:38-48.

49. Romero, P., Obradovic, Z., and Dunker, A.K., Sequence data analysis for long disordered regions prediction in calcineurin family. *Genome Informatics* 1997; 8:110-124.

50. 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.

51. Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Garner, E., Guilliot, S., and Dunker, A.K., Thousands of proteins likely to have long disordered regions. *Pac. Symp. Biocomput.,* 1998: 437-448.

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

53. 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: 161-171.

Appendix 1

Perl script for populating the table PONDR in DNA_VIRUS database

```
#! /usr/bin/perl
use DBI;
use FileHandle;
my ($dbh, $sth);
my ($user_name)="mkhosrav";
my ($password)="4QqM37eF6s";
my $table = "PONDRS";
my $db = "DNA_VIRUS";
my $data_file="DNAss_linout.txt";
$dbh=DBI-
>connect("DBI:mysql:database=DNA_VIRUS;host=localhost",$mkhosrav,$4QqM37eF6
s,{RaiseError=>1});
open(DAT,$DNAss_linout.txt)|| die("Could not open file: $!");
@lines = <DAT>;
close DAT;
foreach $line(@line){
if ($line= ~ m/#/)
{
$line= ~ S/#//;
$cdsprotid = $line;}
elsif ($line!~ M/\S/)
```

```
{
@data=split (" ", $line);
print $data[0];
print $data[1];
print $data[2];
$query="insert into PONDRS (cdsprotid, prot_position, PONDR_VL3) values
($cdsprotid, $data[0], $data[2])";
$sth= dbh->prepare($query);
$sth->execute();
}
}
$sth->finish();
$dbh->disconnect();
```

Appendix 2

A sample of output for prediction of order and disorder of proteins in the overlapping region

```
OVERLAP; NC_000939-4(5)
OVERLAP; NC_000939-5(4)
select * from OVERLAP where overlapid = "OVERLAP; NC_000939-5(4)"
select * from PONDRS where cdsprotid = "NP_051033" AND prot_position >=
44 AND prot_position <= 173
130 rows returned
select * from PONDRS where cdsprotid = "NP_051034" AND prot_position >=
1 AND prot_position <= 130
130 rows returned
130
seq1pos     O/d   O/d   seq2pos
-----------------------
 Amino acids: K M44   O (0.168255)  d (0.628742)   1
 Amino acids: W E45   O (0.163555)  d (0.626855)   2
 Amino acids: K N46   O (0.15771)   d (0.624686)   3
 Amino acids: I S47   O (0.151991)  d (0.621748)   4
 Amino acids: P Q48   O (0.147354)  d (0.61707)    5
 Amino acids: K T49   O (0.143412)  d (0.610165)   6
 Amino acids: Q G50   O (0.140759)  d (0.604253)   7
 Amino acids: G V51   O (0.138932)  d (0.601007)   8
 Amino acids: F L52   O (0.139971)  d (0.59991)    9
 Amino acids: Y C53   O (0.142626)  d (0.599018)   10
 Amino acids: A P54   O (0.145522)  d (0.596876)   11
 Amino acids: P N55   O (0.145944)  d (0.593222)   12
 Amino acids: I R56   O (0.14448)   d (0.590186)   13
 Amino acids: D C57   O (0.144306)  d (0.589928)   14
 Amino acids: V Q58   O (0.148619)  d (0.592494)   15
 Amino acids: K V59   O (0.158146)  d (0.593509)   16
 Amino acids: F C60   O (0.168943)  d (0.594164)   17
 Amino acids: V S61   O (0.179586)  d (0.593942)   18
 Amino acids: L H62   O (0.187421)  d (0.594516)   19
 Amino acids: T T63   O (0.195137)  d (0.593076)   20
 20seq2 disordered
 Amino acids: P T64   O (0.201413)  d (0.589753)   21
 Amino acids: H Y65   O (0.209206)  d (0.58643)    22
 Amino acids: I I66   O (0.218029)  d (0.583107)   23
 23seq2 disordered
 Amino acids: S R67   O (0.228422)  d (0.579783)   24
 Amino acids: E E68   O (0.238379)  d (0.57646)    25
 25seq2 disordered
 Amino acids: R S69   O (0.248252)  d (0.573137)   26
 Amino acids: A S70   O (0.258021)  d (0.569814)   27
 Amino acids: Q G71   O (0.2704)    d (0.566491)   28
 Amino acids: V Q72   O (0.284319)  d (0.563168)   29
 Amino acids: R G73   O (0.298677)  d (0.559845)   30
 Amino acids: G G74   O (0.310698)  d (0.556522)   31
 31seq2 disordered
 Amino acids: V R75   O (0.320918)  d (0.553199)   32
 Amino acids: V Q76   O (0.329199)  d (0.549876)   33
 Amino acids: K A77   O (0.335734)  d (0.546553)   34
 Amino acids: L C78   O (0.34063)   d (0.54323)    35
```

```
Amino acids: V R79   O (0.34507)   d (0.539907)   36
Amino acids: D F80   O (0.349925)  d (0.536584)   37
Amino acids: S T81   O (0.354849)  d (0.533261)   38
Amino acids: R R82   O (0.358642)  d (0.529938)   39
39seq2 disordered
Amino acids: D F83   O (0.360413)  d (0.526615)   40
Amino acids: L V84   O (0.361245)  d (0.523292)   41
Amino acids: S T85   O (0.363247)  d (0.519969)   42
Amino acids: P Q86   O (0.366261)  d (0.516646)   43
Amino acids: S P87   O (0.370165)  d (0.513323)   44
Amino acids: R R88   O (0.371909)  O (0.4523)     45
Amino acids: E V89   O (0.369587)  O (0.38784)    46
Amino acids: L V90   O (0.361249)  O (0.320718)   47
Amino acids: Y S91   O (0.350751)  O (0.306974)   48
Amino acids: R E92   O (0.339616)  O (0.294188)   49
Amino acids: S Q93   O (0.331223)  O (0.279732)   50
Amino acids: K G94   O (0.3229)    O (0.265332)   51
Amino acids: E I95   O (0.316261)  O (0.251962)   52
Amino acids: F Q96   O (0.307946)  O (0.241968)   53
Amino acids: N Y97   O (0.300035)  O (0.235196)   54
Amino acids: I R98   O (0.291681)  O (0.229073)   55
Amino acids: G S99   O (0.284926)  O (0.222146)   56
Amino acids: H W100  O (0.277606)  O (0.215086)   57
Amino acids: G L101  O (0.269677)  O (0.207027)   58
Amino acids: L S102  O (0.258508)  O (0.197978)   59
Amino acids: V D103  O (0.24556)   O (0.188037)   60
Amino acids: I R104  O (0.23223)   O (0.177213)   61
Amino acids: E G105  O (0.221738)  O (0.168888)   62
Amino acids: G F106  O (0.214016)  O (0.164301)   63
Amino acids: S P107  O (0.209983)  O (0.163215)   64
Amino acids: Q A108  O (0.207681)  O (0.163626)   65
Amino acids: L T109  O (0.206839)  O (0.162108)   66
Amino acids: P L110  O (0.204695)  O (0.159318)   67
Amino acids: F L111  O (0.201722)  O (0.155536)   68
Amino acids: C S112  O (0.197002)  O (0.152579)   69
Amino acids: L T113  O (0.192496)  O (0.150301)   70
Amino acids: P S114  O (0.187574)  O (0.14749)    71
Amino acids: V G115  O (0.182377)  O (0.144328)   72
Amino acids: G G116  O (0.176845)  O (0.141105)   73
Amino acids: D L117  O (0.173314)  O (0.137687)   74
Amino acids: Y S118  O (0.172612)  O (0.134025)   75
Amino acids: P T119  O (0.173061)  O (0.129457)   76
Amino acids: L T120  O (0.173404)  O (0.123551)   77
Amino acids: Q I121  O (0.174493)  O (0.116544)   78
Amino acids: F R122  O (0.178005)  O (0.110201)   79
Amino acids: E G123  O (0.184458)  O (0.106322)   80
Amino acids: V H124  O (0.191875)  O (0.104729)   81
Amino acids: T G125  O (0.19946)   O (0.10448)    82
Amino acids: V V126  O (0.205579)  O (0.104506)   83
Amino acids: L A127  O (0.212177)  O (0.104315)   84
Amino acids: Q V128  O (0.217475)  O (0.103776)   85
Amino acids: S T129  O (0.22331)   O (0.102923)   86
Amino acids: Q I130  O (0.229232)  O (0.101915)   87
Amino acids: F Q131  O (0.237641)  O (0.0999731)  88
Amino acids: R G132  O (0.247765)  O (0.0985853)  89
Amino acids: E D133  O (0.259101)  O (0.0965168)  90
Amino acids: T S134  O (0.270505)  O (0.0954874)  91
```

```
 Amino acids: A K135  O (0.28191)   O (0.0956794)  92
 Amino acids: N S136  O (0.293314)  O (0.0969503)  93
 Amino acids: L L137  O (0.304719)  O (0.0970088)  94
 Amino acids: Y L138  O (0.316123)  O (0.0951305)  95
 Amino acids: S N139  O (0.327528)  O (0.0923114)  96
 Amino acids: T F140  O (0.338932)  O (0.0914421)  97
 Amino acids: S C141  O (0.350337)  O (0.0920859)  98
 Amino acids: V R142  O (0.361741)  O (0.0944435)  99
 Amino acids: E V143  O (0.373146)  O (0.0976489)  100
 Amino acids: W A144  O (0.38455)   O (0.101804)   101
 Amino acids: R Y145  O (0.395955)  O (0.106974)   102
 Amino acids: M D146  O (0.407359)  O (0.112257)   103
 Amino acids: M V147  O (0.418764)  O (0.117923)   104
 Amino acids: S F148  O (0.430168)  O (0.124381)   105
 Amino acids: S H149  O (0.441573)  O (0.132517)   106
 Amino acids: T H150  O (0.452977)  O (0.143772)   107
 Amino acids: T P151  O (0.464382)  O (0.157478)   108
 Amino acids: P V152  O (0.475786)  O (0.173772)   109
 Amino acids: L V153  O (0.487191)  O (0.192083)   110
 Amino acids: S Q154  O (0.498595)  O (0.212542)   111
 Amino acids: R S155  d (0.514991)  O (0.234928)   112
 Amino acids: V E156  d (0.536376)  O (0.256765)   113
 Amino acids: R V157  d (0.562753)  O (0.278909)   114
 Amino acids: S C158  d (0.589129)  O (0.295554)   115
 Amino acids: V H159  d (0.615506)  O (0.308007)   116
 Amino acids: M G160  d (0.641882)  O (0.315913)   117
 Amino acids: G S161  d (0.668258)  O (0.324561)   118
 Amino acids: A G162  d (0.694635)  O (0.333676)   119
 Amino acids: A P163  d (0.721011)  O (0.342237)   120
 Amino acids: Q A164  d (0.747388)  O (0.350608)   121
 Amino acids: Q T165  d (0.773764)  O (0.359037)   122
 Amino acids: P S166  d (0.800141)  O (0.367866)   123
 Amino acids: A D167  d (0.821531)  O (0.377165)   124
 Amino acids: M E168  d (0.838629)  O (0.387531)   125
 Amino acids: K I169  d (0.853569)  O (0.399986)   126
 Amino acids: L T170  d (0.870614)  O (0.414747)   127
 Amino acids: Q T171  d (0.888659)  O (0.431481)   128
 Amino acids: P K172  d (0.904826)  O (0.45017)    129
 Amino acids: N F173  d (0.919749)  O (0.459772)   130
19 seq1 disordered
111 seq1 not disordered
44 seq2 disordered
86 seq2 not disordered

14.6153846153846 percent seq1 disordered
85.3846153846154 percent seq1 not disordered
33.8461538461538 percent seq2 disordered
66.1538461538461 percent seq2 not disordered
63 disordered
67 not disordered
0 both disordered
19 seq1 disordered
44 seq2 disordered
```

Appendix 3

Result of order and disorder prediction of proteins in the overlapping regions

Acc number		Sequence 1					Sequence 2				No. of pairs			
		Protein identification	Overlap position		No. of predicted disorder	No. of predicted order	Protein identification	Overlap position		No. of predicted disorder	No. of predicted order	P-ND (O-O)	P-BSD (d-d)	P-SQ1D (d-O)	P-SQ2D (O-d)
NC_001915-1(2)	NC_001915-2(1)	NP_X10000	3	133	90	41	NP_047196	1	131	22	109	41	22	68	0
NC_004178-1(2)	NC_004178-2(1)	NP_690837	16	149	111	23	NP_690838	1	134	15	119	23	15	96	0
NC_004267-1(2)	NC_004267-2(1)	NP_694621	21	139	73	46	NP_694622	1	119	25	94	37	16	57	9
NC_003771-1(2)	NC_003771-2(1)	NP_620541	160	485	0	326	NP_620542	1	326	165	161	161	0	0	165
NC_003768-1(2)	NC_003768-2(1)	NP_620538	91	182	51	41	NP_X10001	1	92	72	20	11	42	9	30
NC_001641-1(2)	NC_001641-2(1)	NP_042580	649	697	31	18	NP_042581	1	49	27	22	18	27	4	0
NC_001927-1(2)	NC_001927-2(1)	NP_047213	7	107	0	101	NP_047214	1	101	21	80	80	0	0	21
NC_001498-2(3)	NC_001498-3(2)	NP_056919	8	193	132	54	NP_056920	1	186	95	91	35	76	56	19
NC_002199-2(3)	NC_002199-3(2)	NP_054691	230	282	1	52	NP_054692	230	282	17	36	36	1	0	16
NC_002199-2(4)	NC_002199-4(2)	NP_X10002	9	161	153	0	NP_054693	1	153	88	65	0	88	65	0
NC_005339-2(3)	NC_005339-3(2)	NP_958049	244	295	52	0	NP_958050	244	295	19	33	0	19	33	0
NC_005339-2(4)	NC_005339-4(2)	NP_X10003	11	162	114	38	NP_958051	1	152	78	74	38	78	36	0
NC_001552-2(3)	NC_001552-3(2)	NP_056872	8	215	111	97	NP_056873	1	208	208	0	0	111	0	97
NC_001552-3(4)	NC_001552-4(3)	NP_X10004	318	369	40	12	NP_X10005	318	369	40	12	12	40	0	0
NC_002200-2(3)	NC_002200-3(2)	NP_054708	156	224	59	10	NP_054709	1	224	21	48	0	11	48	10
NC_001560-2(3)	NC_001560-3(2)	NP_041713	25	91	67	0	NP_X10006	1	67	29	38	0	29	38	0
NC_002534-3(4)	NC_002534-4(3)	NP_065672	184	227	0	44	NP_065673	1	44	0	44	44	0	0	0
NC_002534-4(5)	NC_002534-5(4)	NP_X10007	156	191	0	36	NP_065674	1	36	0	36	36	0	0	0

			Sequence 1					Sequence 2					No. of pairs		
NC_001633-1(2)	NC_001633-2(1)	NP_042508	3	179	0	177	NP_042509	1	177	0	177	177	0	0	0
NC_001633-2(3)	NC_001633-3(2)	NP_X10008	605	657	53	0	NP_042510	1	53	0	53	0	0	53	0
NC_002035-1(2)	NC_002035-2(1)	NP_049324	778	857	55	25	NP_619631	1	80	80	0	0	55	0	25
NC_003809-1(2)	NC_003809-2(1)	NP_620678	696	797	75	27	NP_620679	1	102	3	99	24	0	75	3
NC_001749-1(2)	NC_001749-2(1)	NP_044335	1584	1903	50	269	NP_044336	1	320	89	230	210	30	20	59
NC_003499-5(6)	NC_003499-6(5)	NP_612812	268	312	41	4	NP_612813	1	45	0	45	4	0	41	0
NC_003093-5(6)	NC_003093-6(5)	NP_203557	226	325	14	86	NP_203558	1	100	95	5	0	9	5	86
NC_001642-1(2)	NC_001642-2(1)	NP_042582	421	547	107	20	NP_042583	1	127	97	30	7	84	23	13
NC_001658-3(4)	NC_001658-4(3)	NP_042697	63	112	0	50	NP_042698	1	50	22	28	28	0	0	22
NC_001409-2(3)	NC_001409-3(2)	NP_040552	356	460	105	0	NP_040553	1	105	0	105	0	0	105	0
NC_001434-10(9)	NC_001434-9(10)	NP_056788	14	123	110	0	NP_056787	1	110	83	27	0	83	27	0
NC_003481-3(4)	NC_003481-4(3)	NP_604488	69	131	33	30	NP_604489	1	63	43	20	0	13	20	30
NC_004730-4(5)	NC_004730-5(4)	NP_835266	71	122	0	52	NP_835267	1	52	16	36	36	0	0	16
NC_003725-2(3)	NC_003725-3(2)	NP_620439	72	119	0	48	NP_620440	1	48	25	23	23	0	0	25
NC_002568-2(4)	NC_002568-4(2)	NP_066392	900	962	42	21	NP_066394	1	63	63	0	0	42	0	21
NC_004366-3(4)	NC_004366-4(3)	NP_733849	6	237	232	0	NP_733850	1	232	64	168	0	64	168	0
NC_004146-1(3)	NC_004146-3(1)	NP_689444	900	998	99	0	NP_689446	1	99	45	54	0	45	54	0
NC_003448-1(2)	NC_003448-2(1)	NP_599247	893	967	75	0	NP_599248	1	75	28	47	0	28	47	0
NC_005094-1(2)	NC_005094-2(1)	NP_919036	901	1033	133	0	NP_919037	1	133	46	87	0	46	87	0
NC_001366-1(2)	NC_001366-2(1)	NP_040350	5	160	95	61	NP_X10009	1	156	0	156	61	0	95	0
NC_001990-1(2)	NC_001990-2(1)	NP_048059	1316	1925	546	64	NP_048060	1	610	110	500	64	110	436	0
NC_005899-1(2)	NC_005899-2(1)	YP_025095	32	158	114	13	YP_025096	1	127	61	66	0	48	66	13
NC_000939-4(5)	NC_000939-5(4)	NP_051033	44	173	19	111	NP_051034	1	130	44	86	67	0	19	44
NC_003608-1(3)	NC_003608-3(1)	NP_619671	4	212	51	157	NP_X10010	1	209	53	155	121	17	34	36
NC_003608-6(7)	NC_003608-7(6)	NP_619676	5	228	22	202	NP_619677	1	224	0	224	202	0	22	0
NC_003627-4(6)	NC_003627-6(4)	NP_619720	130	279	66	84	NP_619722	1	150	55	95	71	42	24	13
NC_003487-3(4)	NC_003487-4(3)	NP_608313	13	62	0	50	NP_608314	1	50	50	0	0	0	0	50
NC_003532-4(5)	NC_003532-5(4)	NP_613263	11	182	39	133	NP_613264	1	172	119	53	53	39	0	80
NC_004063-1(2)	NC_004063-2(1)	NP_663296	3	628	626	0	NP_663297	1	626	149	477	0	149	477	0
NC_001574-3(5)	NC_001574-5(3)	NP_041734	1721	1834	19	95	NP_041736	1	114	25	89	70	0	19	25
NC_001719-1(3)	NC_001719-3(1)	NP_043862	166	217	52	0	NP_X10011	1	52	25	27	0	25	27	0
NC_001719-3(4)	NC_001719-4(3)	NP_X10012	188	614	148	279	NP_043865	1	427	149	278	224	94	54	55

		Sequence 1					Sequence 2					No. of pairs			
NC_001719-3(7)	NC_001719-7(3)	NP_X10013	793	877	46	39	NP_043868	1	85	66	19	0	27	19	39
NC_004324-2(3)	NC_004324-3(2)	NP_861408	148	178	31	0	NP_861409	1	31	28	3	0	28	3	0
Total pool of 52 protein pairs												2014	1653	2530	1022

Appendix 4

A sample of output for order and disorder predictions in the entire sequence of a protein pair

NP_051033			NP_051034		
1	M	0.4263188541	1	M	0.6287424564
2	E	0.4210431874	2	E	0.6268553734
3	I	0.4103756845	3	N	0.6246856451
4	Q	0.4013533592	4	S	0.6217484474
5	S	0.3937372863	5	Q	0.6170695424
6	L	0.3872836828	6	T	0.6101647019
7	D	0.3812186420	7	G	0.6042528152
8	G	0.3759230673	8	V	0.6010074019
9	V	0.3698520362	9	L	0.5999103189
10	L	0.3629848063	10	C	0.5990181565
11	G	0.3546615839	11	P	0.5968763232
12	E	0.3473497331	12	N	0.5932222605
13	E	0.3415260613	13	R	0.5901864171
14	L	0.3376469910	14	C	0.5899282098
15	A	0.3352956772	15	Q	0.5924942493
16	I	0.3289301991	16	V	0.5935085416
17	Q	0.3169742525	17	C	0.5941638350
18	N	0.3013938367	18	S	0.5939415097
19	E	0.2880397737	19	H	0.5945159197
20	V	0.2782653272	20	T	0.5930755734
21	K	0.2688401639	21	T	0.5897526145
22	K	0.2598593235	22	Y	0.5864295959
23	I	0.2507073581	23	I	0.5831065178
24	L	0.2433348447	24	R	0.5797834992
25	L	0.2368623018	25	E	0.5764604211
26	S	0.2304195911	26	S	0.5731374621
27	H	0.2238274366	27	S	0.5698144436
28	K	0.2177044600	28	G	0.5664914250
29	T	0.2132986337	29	Q	0.5631683469
30	T	0.2100027949	30	G	0.5598453879
31	K	0.2072314024	31	G	0.5565223694
32	A	0.2048592418	32	R	0.5531993508
33	I	0.2024080008	33	Q	0.5498762727
34	L	0.1999487430	34	A	0.5465533137
35	P	0.1969751120	35	C	0.5432302356
36	L	0.1936348677	36	R	0.5399072170
37	A	0.1898535937	37	F	0.5365841985
38	P	0.1857631058	38	T	0.5332611203
39	I	0.1810358167	39	R	0.5299381614
40	S	0.1768899411	40	F	0.5266151428
41	Q	0.1739731282	41	V	0.5232921243
42	F	0.1726955622	42	T	0.5199690461
43	S	0.1711555868	43	Q	0.5166460872
44	K	0.1682546586	44	P	0.5133228302
45	W	0.1635548472	45	R	0.4523003101
46	K	0.1577104777	46	V	0.3878404200
47	I	0.1519905925	47	V	0.3207175434
48	P	0.1473541111	48	S	0.3069736660
49	K	0.1434118748	49	E	0.2941881418
50	Q	0.1407587975	50	Q	0.2797319591

51	G	0.1389324963	51	G	0.2653318942
52	F	0.1399712563	52	I	0.2519617081
53	Y	0.1426260024	53	Q	0.2419683188
54	A	0.1455216408	54	Y	0.2351964265
55	P	0.1459440589	55	R	0.2290729284
56	I	0.1444802135	56	S	0.2221457958
57	D	0.1443055123	57	W	0.2150861174
58	V	0.1486189216	58	L	0.2070267648
59	K	0.1581459790	59	S	0.1979777366
60	F	0.1689433306	60	D	0.1880372614
61	V	0.1795858145	61	R	0.1772130281
62	L	0.1874209046	62	G	0.1688884497
63	T	0.1951367855	63	F	0.1643005013
64	P	0.2014129162	64	P	0.1632147580
65	H	0.2092055678	65	A	0.1636258364
66	I	0.2180292755	66	T	0.1621076614
67	S	0.2284219116	67	L	0.1593181342
68	E	0.2383794338	68	L	0.1555356532
69	R	0.2482522130	69	S	0.1525788158
70	A	0.2580212057	70	T	0.1503012329
71	Q	0.2703996599	71	S	0.1474897563
72	V	0.2843189538	72	G	0.1443284601
73	R	0.2986767590	73	G	0.1411047876
74	G	0.3106977940	74	L	0.1376874596
75	V	0.3209180832	75	S	0.1340254992
76	V	0.3291994631	76	T	0.1294572800
77	K	0.3357338011	77	T	0.1235513091
78	L	0.3406301737	78	I	0.1165436134
79	V	0.3450701237	79	R	0.1102013364
80	D	0.3499245346	80	G	0.1063217893
81	S	0.3548491299	81	H	0.1047294140
82	R	0.3586422205	82	G	0.1044798866
83	D	0.3604131639	83	V	0.1045063809
84	L	0.3612449169	84	A	0.1043152884
85	S	0.3632472754	85	V	0.1037761867
86	P	0.3662614822	86	T	0.1029228568
87	S	0.3701654673	87	I	0.1019146219
88	R	0.3719085157	88	Q	0.0999731123
89	E	0.3695870638	89	G	0.0985852703
90	L	0.3612492979	90	D	0.0965167880
91	Y	0.3507513106	91	S	0.0954874381
92	R	0.3396156728	92	K	0.0956793651
93	S	0.3312228024	93	S	0.0969502926
94	K	0.3228998482	94	L	0.0970088318
95	E	0.3162614107	95	L	0.0951305330
96	F	0.3079462349	96	N	0.0923113525
97	N	0.3000348508	97	F	0.0914420709
98	I	0.2916814387	98	C	0.0920859054
99	G	0.2849255800	99	R	0.0944434628
100	H	0.2776061594	100	V	0.0976488590
101	G	0.2696771324	101	A	0.1018036008
102	L	0.2585083544	102	Y	0.1069739833
103	V	0.2455599755	103	D	0.1122568250
104	I	0.2322298139	104	V	0.1179232895
105	E	0.2217383534	105	F	0.1243807152
106	G	0.2140158415	106	H	0.1325165480
107	S	0.2099828124	107	H	0.1437723488

108	Q	0.2076812238
109	L	0.2068393975
110	P	0.2046948224
111	F	0.2017216533
112	C	0.1970018744
113	L	0.1924956292
114	P	0.1875739843
115	V	0.1823773831
116	G	0.1768452674
117	D	0.1733143777
118	Y	0.1726116687
119	P	0.1730613261
120	L	0.1734036952
121	Q	0.1744927913
122	F	0.1780048609
123	E	0.1844580024
124	V	0.1918754131
125	T	0.1994600743
126	V	0.2055794448
127	L	0.2121766210
128	Q	0.2174746245
129	S	0.2233098000
130	Q	0.2292315364
131	F	0.2376412749
132	R	0.2477651685
133	E	0.2591006458
134	T	0.2705051601
135	A	0.2819096744
136	N	0.2933141887
137	L	0.3047187030
138	Y	0.3161232173
139	S	0.3275277317
140	T	0.3389322758
141	S	0.3503367901
142	V	0.3617413044
143	E	0.3731458187
144	W	0.3845503330
145	R	0.3959548473
146	M	0.4073593616
147	M	0.4187638760
148	S	0.4301683903
149	S	0.4415729046
150	T	0.4529774189
151	T	0.4643819332
152	P	0.4757864475
153	L	0.4871909618
154	S	0.4985953867
155	R	0.5149905086
156	V	0.5363762379
157	R	0.5627526641
158	S	0.5891291499
159	V	0.6155056357
160	M	0.6418820620
161	G	0.6682584882
162	A	0.6946349144
163	A	0.7210113406
164	Q	0.7473878264

108	P	0.1574780196
109	V	0.1737724692
110	V	0.1920834035
111	Q	0.2125422508
112	S	0.2349284440
113	E	0.2567650974
114	V	0.2789087296
115	C	0.2955537736
116	H	0.3080069721
117	G	0.3159132302
118	S	0.3245606124
119	G	0.3336759508
120	P	0.3422371447
121	A	0.3506084383
122	T	0.3590371609
123	S	0.3678661287
124	D	0.3771648407
125	E	0.3875309527
126	I	0.3999863565
127	T	0.4147466719
128	T	0.4314810038
129	K	0.4501699507
130	F	0.4597721696

165	Q	0.7737643123
166	P	0.8001406789
167	A	0.8215308189
168	M	0.8386289477
169	K	0.8535690308
170	L	0.8706142306
171	Q	0.8886590004
172	P	0.9048259854
173	N	0.9197490811
174	F	0.9320861697
175	K	0.9425940514
176	M	0.9510643482
177	S	0.9587990642
178	L	0.9655229449
179	E	0.9714501500
180	S	0.9759386182
181	S	0.9794287682
182	K	0.9817369580
183	G	0.9837498069
184	G	0.9852137566
185	G	0.9864628911
186	M	0.9874262810
187	K	0.9877581596
188	P	0.9874918461
189	H	0.9865961075
190	Q	0.9852035046
191	K	0.9831839204
192	K	0.9805755615
193	S	0.9774577022
194	S	0.9726486802
195	K	0.9664878249
196	P	0.9587950110
197	N	0.9507102966
198	G	0.9420812130
199	H	0.9335042834
200	S	0.9251780510
201	R	0.9166570306
202	R	0.9078938961
203	G	0.8988140225
204	N	0.8898977637
205	L	0.8810328841
206	S	0.8722431064
207	G	0.8640000820
208	E	0.8563218117
209	V	0.8492720723
210	G	0.8425521255
211	G	0.8361129761
212	S	0.8298907876
213	S	0.8232998252
214	S	0.8165758252
215	S	0.8096819520
216	L	0.8029594421
217	P	0.7961145043
218	S	0.7891162038
219	G	0.7817986608
220	A	0.7739841938
221	Q	0.7666627765

222	T	0.7600855827	
223	G	0.7545640469	
224	E	0.7492833734	
225	W	0.7463501096	
226	I	0.7455788255	
227	D	0.7463886142	
228	N	0.7467793822	
229	D	0.7468461990	
230	Y	0.7469899058	
231	G	0.7469524741	
232	D	0.7474706173	
233	G	0.7483258247	
234	S	0.7497397065	
235	S	0.7512066960	
236	E	0.7530043721	
237	Y	0.7550315857	
238	S	0.7565171123	
239	G	0.7574564815	
240	V	0.7574818134	
241	S	0.7567691207	
242	T	0.7561950684	

Appendix 5

Result of order and disorder predictions in the entire sequence for 97 protein samples

Sequence	Protein Pair	cdsprotid	Length of overlap section	O in overlap section	D in overlap section	Length of entire sequence	D in entire sequence	O in entire sequence	Length of nonoverlap section	D in nonoverlap section	O in nonoverlap section
1	1	NP_X10000	131	41	90	133	92	41	2	2	0
2	1	NP_047196	131	109	22	972	201	771	841	179	662
3	2	NP_690837	134	23	111	149	126	23	15	15	0
4	2	NP_690838	134	119	15	1012	111	901	878	96	782
5	3	NP_694621	119	46	73	418	131	287	299	58	241
6	3	NP_694622	119	94	25	119	25	94	0	0	0
7	4	NP_620541	326	326	0	1255	0	1255	929	0	929
8	4	NP_620542	326	161	165	326	165	161	0	0	0
9	5	NP_620538	92	41	51	312	123	189	220	72	148
10	5	NP_X10001	92	20	72	92	72	20	0	0	0
11	6	NP_042580	49	18	31	697	42	655	648	11	637
12	6	NP_042581	49	22	27	863	44	819	814	17	797
13	7	NP_047213	101	101	0	233	0	233	132	0	132
14	7	NP_047214	101	80	21	101	21	80	0	0	0
15	8	NP_056919	186	54	132	507	366	141	321	234	87
16	8	NP_056920	186	91	95	186	95	91	0	0	0
17	9	NP_054691	53	52	1	527	462	65	474	461	13
18	9	NP_054692	53	36	17	282	233	49	229	216	13
19	10	NP_054693	153	65	88	153	88	65	0	0	0
20	11	NP_958049	52	0	52	542	408	134	490	356	134
21	11	NP_958050	52	33	19	295	198	97	243	179	64
22	12	NP_958051	152	74	78	152	78	74	0	0	0
23	13	NP_056872	208	97	111	215	114	101	7	3	4
24	13	NP_056873	208	0	208	568	499	69	360	291	69
25	14	NP_X10005	52	12	40	489	362	127	437	322	115
26	15	NP_054708	69	10	59	391	206	185	322	147	175
27	15	NP_054709	224	104	120	224	120	104	0	0	0
28	16	NP_041713	67	0	67	265	131	134	198	64	134
29	16	NP_X10006	67	38	29	67	29	38	0	0	0
30	17	NP_065672	44	44	0	227	0	227	183	0	183
31	17	NP_065673	44	44	0	191	0	191	147	0	147
32	18	NP_065674	36	36	0	175	0	175	139	0	139
33	19	NP_042508	177	177	0	179	0	179	2	0	2
34	19	NP_042509	177	177	0	657	184	473	480	184	296
35	20	NP_042510	53	53	0	420	9	411	367	9	358
36	21	NP_049324	80	25	55	857	154	703	777	99	678
37	21	NP_619631	80	0	80	110	110	0	30	30	0

Sequence	Protein Pair	cdsprotid	Length of overlap section	O in overlap section	D in overlap section	Length of entire sequence	D in entire sequence	O in entire sequence	Length of nonoverlap section	D in nonoverlap section	O in nonoverlap section
38	22	NP_620678	102	27	75	797	130	667	695	55	640
39	22	NP_620679	102	99	3	195	3	192	93	0	93
40	23	NP_044335	320	270	50	2105	269	1836	1785	219	1566
41	23	NP_044336	320	230	90	320	90	230	0	0	0
42	24	NP_612812	45	4	41	312	144	168	267	103	164
43	24	NP_612813	45	45	0	147	0	147	102	0	102
44	25	NP_203557	100	86	14	325	136	189	225	122	103
45	25	NP_203558	100	5	95	222	157	65	122	62	60
46	26	NP_042582	127	20	107	1365	178	1187	1238	71	1167
47	26	NP_042583	127	30	97	127	97	30	0	0	0
48	27	NP_042697	50	50	0	112	0	112	62	0	62
49	27	NP_042698	50	28	22	97	22	75	47	0	47
50	28	NP_040552	105	0	105	460	199	261	355	94	261
51	28	NP_040553	105	105	0	193	6	187	88	6	82
52	29	NP_056788	110	0	110	660	212	448	550	102	448
53	29	NP_056787	110	27	83	123	96	27	13	13	0
54	30	NP_604488	63	30	33	131	33	98	68	0	68
55	30	NP_604489	63	20	43	155	43	112	92	0	92
56	31	NP_835266	52	52	0	122	0	122	70	0	70
57	31	NP_835267	52	36	16	155	16	139	103	0	103
58	32	NP_620439	48	48	0	119	0	119	71	0	71
59	32	NP_620440	48	23	25	190	25	165	142	0	142
60	33	NP_066392	63	21	42	962	253	709	899	211	688
61	33	NP_066394	63	0	63	268	77	191	205	14	191
62	34	NP_733849	232	0	232	237	237	0	5	5	0
63	34	NP_733850	232	168	64	244	76	168	12	12	0
64	35	NP_689444	99	0	99	998	148	850	899	49	850
65	35	NP_689446	99	54	45	106	52	54	7	7	0
66	36	NP_599247	75	0	75	983	187	796	908	112	796
67	36	NP_599248	75	47	28	75	28	47	0	0	0
68	37	NP_919036	133	0	133	1045	163	882	912	30	882
69	37	NP_919037	133	87	46	133	46	87	0	0	0
70	38	NP_040350	156	61	95	2303	138	2165	2147	43	2104
71	38	NP_X10009	156	156	0	156	0	156	0	0	0
72	39	NP_048059	610	64	546	1925	617	1308	1315	71	1244
73	39	NP_048060	610	500	110	612	112	500	2	2	0
74	40	YP_025095	127	13	114	158	114	44	31	0	31
75	40	YP_025096	127	66	61	643	97	546	516	36	480
76	41	NP_051033	130	111	19	242	88	154	112	69	43
77	41	NP_051034	130	86	44	130	44	86	0	0	0
78	42	NP_619671	209	158	51	735	53	682	526	2	524
79	42	NP_X10010	209	156	53	209	53	156	0	0	0

Sequence	Protein Pair	cdsprotid	Length of overlap section	O in overlap section	D in overlap section	Length of entire sequence	D in entire sequence	O in entire sequence	Length of nonoverlap section	D in nonoverlap section	O in nonoverlap section
80	43	NP_619676	224	202	22	345	22	323	121	0	121
81	43	NP_619677	224	224	0	224	0	224	0	0	0
82	44	NP_619720	150	84	66	279	117	162	129	51	78
83	44	NP_619722	150	95	55	236	55	181	86	0	86
84	45	NP_608313	50	50	0	62	0	62	12	0	12
85	45	NP_608314	50	0	50	65	65	0	15	15	0
86	46	NP_613263	172	133	39	189	56	133	17	17	0
87	46	NP_613264	172	53	119	172	119	53	0	0	0
88	47	NP_663296	626	0	626	628	628	0	2	2	0
89	47	NP_663297	626	477	149	1844	522	1322	1218	373	845
90	48	NP_041734	114	95	19	1834	554	1280	1720	535	1185
91	48	NP_041736	114	89	25	131	25	106	17	0	17
92	49	NP_043862	52	0	52	217	53	164	165	1	164
93	49	NP_X10011	52	27	25	877	226	651	825	201	624
94	50	NP_043865	427	278	149	427	149	278	0	0	0
95	51	NP_043868	85	19	66	138	66	72	53	0	53
96	52	NP_861408	31	0	31	178	133	45	147	102	45
97	52	NP_861409	31	3	28	501	243	258	470	215	255
Total			13639	7235	6404	43304	12471	30833	29665	6067	23598

Appendix 6
Curriculum Vitae

Mahvash Khosravi
200 Scranton Ct.
Zionsville, IN 46077
(317) 873-5923
mkhosrav@iupui.edu

Education:

2007	Master of Science in Bioinformatics. Indiana University Purdue University, Indianapolis, IN
1991	Ph.D. degree in Molecular Biology Illinois Institute of Technology, Department of Biology, Chicago, IL Thesis title : "Use of Genetic Engineering to Optimize Protein Production in Recombinant Escherichia coli"
1987	M.S. in Molecular Biology Illinois Institute of Technology, Department of Biology, Chicago, IL Thesis title: "Effect of Plasmid Size on Growth and Physiology of Recombinant Escherichia coli"
1977	Jondi-Shapour University, Iran B.S. in Biology

Experience:

Teaching and undergraduate research at the Department of Chemistry and Life Sciences, Rose-Hulman Institute of Technology, Terre Haute, Indiana. Teaching areas have been molecular biology and genetics.

Research and clinical experience as visiting research faculty and postdoctoral fellow at Indiana University Medical Center. Research areas have been neurodegenerative diseases, hereditary diseases, eukaryotic genetics and study and characterization of novel cerebellar cDNA clones.