San Jose State University
SJSU ScholarWorks

Master's Theses | Master's Theses and Graduate Research

2007

Exploring the viability of protein structure prediction using sequence entropy

Shalini Potluri
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_theses

Recommended Citation
Potluri, Shalini, "Exploring the viability of protein structure prediction using sequence entropy" (2007). *Master's Theses*. 3469.
DOI: https://doi.org/10.31979/etd.bp4h-2eme
https://scholarworks.sjsu.edu/etd_theses/3469

This Thesis is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Theses by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

EXPLORING THE VIABILITY OF PROTEIN STRUCTURE PREDICTION USING SEQUENCE ENTROPY

A Thesis

Presented to

The Faculty of the Department of Chemical and Materials Engineering

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Shalini Potluri

December 2007

UMI Number: 1452054

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI®

UMI Microform 1452054

Copyright 2008 by ProQuest LLC.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 E. Eisenhower Parkway
PO Box 1346
Ann Arbor, MI 48106-1346

© 2007

Shalini Potluri

ALL RIGHTS RESERVED

APPROVED FOR THE DEPARTMENT OF CHEMICAL

AND MATERIALS ENGINEERING

Dr. Brooke Lustig, Department of Chemistry

5/31/07

Dr. Melanie McNeil, Department of Chemical Engineering

Dr. Daryl Eggers, Department of Chemistry

APPROVED FOR THE UNIVERSITY

01/28/08

ABSTRACT

EXPLORING THE VIABILITY OF PROTEIN STRUCTURE PREDICTION USING SEQUENCE ENTROPY

by Shalini Potluri

Determination of the structure of a protein from the sequence of amino acids has been a major goal in computational biology and bioinformatics. A strong correlation between sequence entropy, and inverse packing density has been shown in recent studies indicated by the occurrence of two major regions, but with a lot of noise in the relationship data. One hundred and thirty query proteins and their sequence alignments are used to test modifications to sequence entropy calculations that significantly reduce the noise in the data. Gapped entropy, Gerstein-Altman entropy, and window average entropy offer improvement in terms of linear correlation but no significant improvement in the data noise is observed due to the introduction of 21st gap term, or Gerstein-Altman random entropy term. Averaging the sequence entropy that includes the 21st gap term within three neighbors resulted into smoothening of the entropy curve with no significant reduction in the data noise.

ACKNOWLEDGEMENTS

Dr. Brooke Lustig (Graduate Advisor)

Dr. Melanie McNeil (Department Advisor)

Dr. Daryl Eggers (Reading Committee Member)

William Yeh

Haihong Liao

My family and friends

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

Proteins are linear unbranched polymers of amino acids. Proteins form the very basis of life. As enzymes, they are the critical catalysts behind all of the biochemical reactions, which make biology work. As structural elements, they are the main constituents of our bones, muscles, hair, skin and blood vessels. As antibodies, they recognize invading elements and allow the immune system to get rid of the unwanted invaders. Proteins are composed of varying amounts of the 20 common amino acids, which in the intact protein are united through covalent chemical linkages called peptide bonds. The amino acids, linked together, form linear unbranched polymeric structures called polypeptide chains; such chains may contain hundreds of amino-acid residues, and the amino acids are arranged in a specific order for a given species of protein [1].

The physiological activity of most proteins is closely linked to their three-dimensional architecture. A protein has four distinct levels of structure as shown in Figure 1. Primary structure denotes the precise linear sequence of amino acids that constitutes the polypeptide chain of the protein molecule. The physical interaction of sequential amino-acid subunits results in a so-called secondary structure, which often can either be a twisting of the polypeptide chain approximating a linear helix (α-configuration), or a zigzag pattern (β-configuration). The secondary structures are held together by hydrogen bonds. Most globular proteins also undergo extensive folding of the chain into a complex three-dimensional geometry designated as tertiary structure. The tertiary structure is held together primarily by hydrophobic interactions but hydrogen

bonds, ionic interactions, and disulfide bonds are usually involved too. All protein molecules are simple unbranched chains of amino acids, but it is by coiling into a specific three-dimensional shape that they are able to perform their biological function. The tertiary structure that a protein assumes to carry out its physiological role inside a cell is known as the native state or sometimes the native conformation. A protein assumes tertiary structure by "folding". Two or more polypeptide chains that behave in many ways as a single structural and functional entity are said to exhibit quaternary structure. The separate chains are not linked through covalent chemical bonds but by weak forces of association [2].

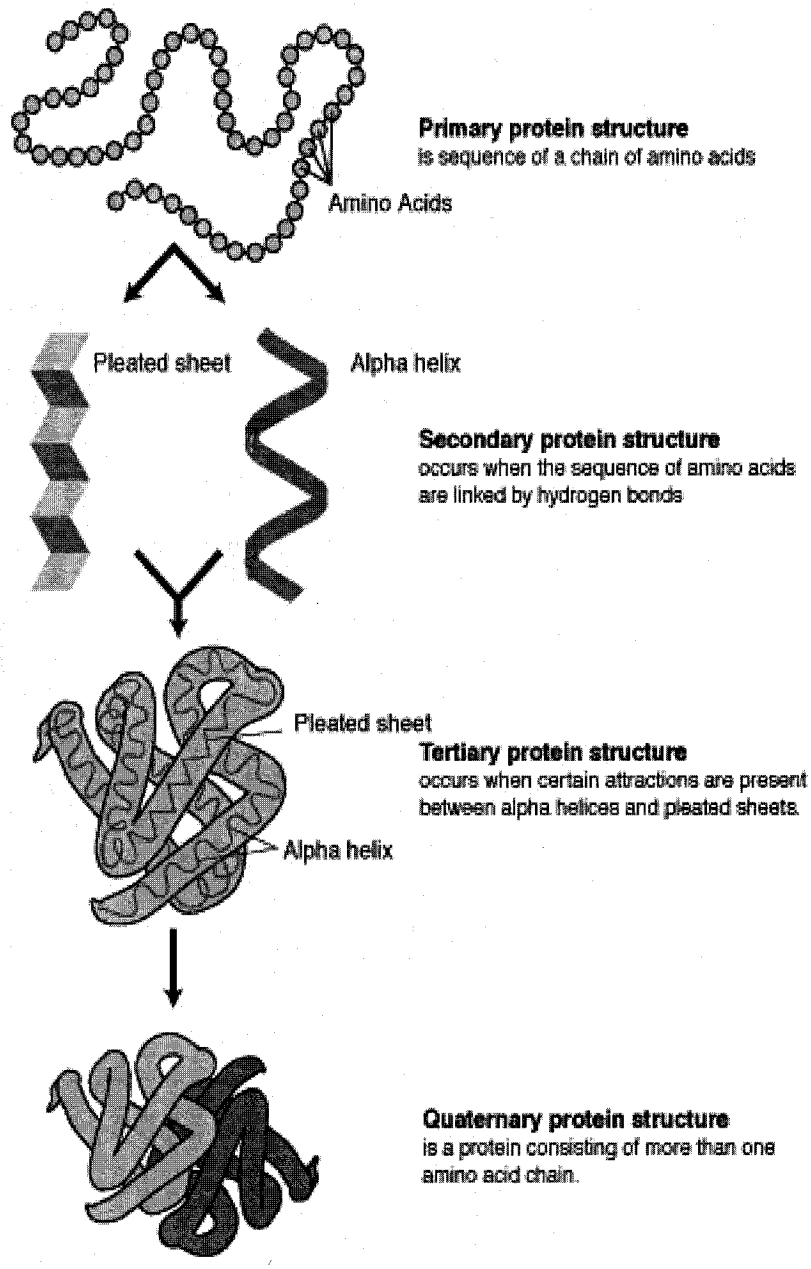

Figure 1. Schematic of protein structures [3].

The Protein Data Bank contains about 42,752 proteins of known structure, as of April 2007 [4]. Experimental methods to determine detailed protein structure, such as X-ray diffraction studies and nuclear magnetic resonance (NMR) analyses, are highly labor intensive. For some protein classes such as transmembrane proteins, the experimental methods to determine three-dimensional structure are not feasible. This creates interest in computational methods for protein structure prediction. Protein structure prediction is one of the most significant tasks tackled in computational structural biology. It has the aim of determining the three-dimensional structure of proteins from their amino acid sequences. In more formal terms, this is the prediction of protein tertiary structure from primary structure. Theoretical understanding of how proteins fold will allow scientists to predict the structure of a protein from its amino acid sequence. The structure of a protein determines its function. This in turn enables them to probe the function of the protein, understand substrate and ligand binding, devise intelligent mutagenesis and biochemical protein engineering experiments that improve specificity and stability, perform rational drug design, and design novel proteins. Given the usefulness of known protein structures in such valuable tasks as rational drug design, this is a highly active field of research.

The necessary information for predicting the structure of a protein is coded in the amino acid sequence of the protein, plus its native solution environment. Although sequence must determine structure, it is not yet possible to predict accurately the entire structure from sequence alone. There are many factors that come into play in the structure formation from sequence such as hydrophobic property of amino acids, packing

of the amino acids and sequence entropy of amino acid sequences. Prediction of a protein structure from the amino acid sequence must take in to account all the above mentioned factors. Hydrophobicity is the measure of miscibility of an amino acid with respect to water solvent. The hydrophobicity of the amino acids determines where the amino acid will be located in the final structure of the protein as shown in Figure 2. This is similar to the formation of the micelle from amphiphilic species.

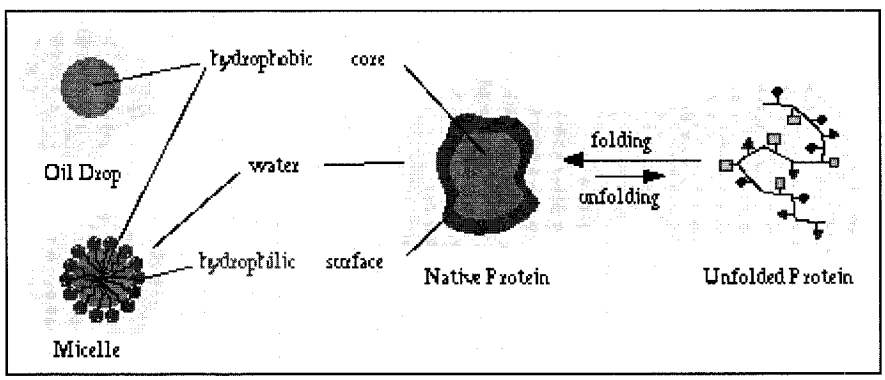

Figure 2. Hydrophobic nature of amino acids in a protein [4].

Unique protein secondary structures can be identified from variability patterns in amino acid sequence using information theory. Applying the Shannon entropy expression to nucleic acid sequence variability has been proven useful in identifying DNA control regions. This method was further extended to measure amino acid conservation in proteins. To understand protein stability and function, we have to understand the interplay between entropy, structure and sequence.

CHAPTER TWO

BACKGROUND

Some terms and concepts that are essential to this research work are introduced in the following paragraphs of this chapter.

2.1 Bioinformatics

Bioinformatics is the use of techniques from mathematics, informatics, statistics and computer science to solve biological problems. The terms bioinformatics and computational biology are often used interchangeably, although the former is, strictly speaking, a subset of the latter. A common thread in projects in bioinformatics and computational biology is the use of mathematical tools to extract useful information from noisy data produced by high-throughput biological techniques [6].

Identity is the extent to which two (nucleotide or amino acid) sequences are invariant. Conservation is the preservation of the physico-chemical properties of the original residue due to the changes at a specific position of an amino acid or (less commonly, DNA) sequence. Homology is the similarity attributed to descent from a common ancestor. Sequence alignment is the process of lining up two or more sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) in order to assess the degree of similarity and homology.

The computational biology tool best known among biologists is probably BLAST, an algorithm for searching large sequence (protein, DNA) databases [7]. BLAST is an acronym for basic local alignment search tool and NCBI (National Center for

Biotechnology Information) provides a popular implementation that searches their massive sequence databases. BLAST is an example of one of sequence alignment methods and enables a user to look for sequences that resemble a given sequence of interest. The BLAST program, which was first originated by Altschul and coworkers, pair-wise aligns a user-selected number of subject sequences in the chosen databases that are most similar to an input query sequence [7]. Protein databases presently contain large number of peptide sequences. The BLAST program is written in such a manner that it minimizes the time it spends on a sequence region whose similarity with the query sequence has little chance of exceeding some minimum alignment score. Computer-scripting languages such as PERL and Python are often used to interface with biological databases and parse output from bioinformatics programs.

Bioinformatics helps to bridge the gap between genome and proteome projects, for example in the use of a DNA sequence for protein identification. The complete proteome for an organism can be conceptualized as the complete set of proteins from all of the various cellular proteomes. This is very roughly the protein equivalent of the genome. Moreover the proteome has at least two levels of complexity lacking in the genome. The genome is defined by the sequence of nucleotides, whereas the proteome entails more than just the sum of the sequences of the proteins present. Knowledge of the proteome requires knowledge of (1) the structure of the proteins in the proteome and (2) the functional interaction between the proteins.

2.2 Sequence Homology

Protein structure prediction is an important application of bioinformatics. The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it. But, the protein can only function correctly if it is folded in a very special and individual way. The prediction of this folding just by looking at the amino acid sequence is quite difficult. One of the key principles in bioinformatics is homology. In the genomic branch of bioinformatics, homology is used to predict the function of a gene. Homologous genes are genes with similar nucleotide sequences. Homology theory predicts that, If gene A is homologous to gene B of which the function is already known, then gene B is likely to have a similar function [R]. In the structural branch of bioinformatics, homology is used to determine which parts of the protein are important in structure formation and interaction with other proteins [8].

Sequence regions that are homologous may be called conserved, concensus or canonical sequences and represent the most common choice of base or amino acid at each position. In a technique called homology modeling, this information is used to predict the structure of a protein once the structure of a homologous protein is known. This currently remains the only way to predict protein structures reliably. Sequence homology makes use of these similarities, comparing the query protein sequence with other sequences in the databases [9]. For protein analysis, the BLASTP is a BLAST program for comparing the query protein sequence with protein databases [7].

Insertion is the addition of a few base pairs into genetic sequence. Deletion is the removal of few base pairs from genetic sequence. Genetic mutations are the reason for the insertions and deletions [1].

2.3 Shannon Entropy

Entropy has important physical implications as the amount of disorder of a system. In microscopic system, entropy measures the degree of disorder in the system. In the protein study, the entropy can be defined as [10]

$$S = -\sum_{\text{protein state}} W\,(\text{protein state})\ \ln W\,(\text{protein state}) \qquad \text{Equation 1}$$

Here, "protein states" refers to all the conformational degrees of freedom necessary to specify the state of the folded protein. The probability of being in a particular state of the protein is W.

Shannon's entropy has a central role in information theory and is sometimes is referred to as measure of uncertainty, thus a measure of amount of information. Sequence entropy is the relevant Shannon entropy expression for proteins and nucleic acids. It is defined as

$$S_k = -\sum p_{jk} \ln p_{jk} \qquad (j = 1,20) \qquad \text{Equation 2}$$

Where p_{jk} is probability of observing a particular amino acid j at sequence position k [10]. The p ln p term is defined as zero if p=0. The more partitioning of a set of events [i.e., P_{jk} distribute more evenly], the larger is the Shannon entropy.

S_k ranges from 0 (only one residue in present at that position) to 3 (all 20 residues are equally represented in that position). Typically, positions with S_k >2.0 are considered variable, whereas those with S_k < 2 are consider conserved. Highly conserved positions are those with S_k <1.0. A minimum number of sequences is however required (~100) for S_k to describe the diversity of a protein family.

2.4 Packing Density

The folding of the polypeptide chain allows the van der Waals surfaces of atoms to fit together snugly, filling up most of the space in the interior [11]. The packing density of a residue or molecule is the ratio of the volume enclosed by the van der Waals surface to the volume occupied in the state in question (such as protein interior, crystal, and liquid).

2.5 Hydrophobicity

Proteins consist of several amino acids held together with peptide bonds. Each amino acid has a different R group. The R group determines whether the protein is hydrophobic or hydrophilic. Hydrophilic groups are typically polar, interacting with water by hydrogen bonding. For this reason, they are called "water loving". Hydrophobic groups, on the other hand, are nonpolar, unable to interact with water, and thus are referred to as "water fearing". The hydrophobicity of the amino acid determines where the amino acid will be located in the final structure of the protein [11].

In globular proteins, the hydrophobic R groups will be located on the inside of the protein, away from the water in the cytosol as shown in Figure 2. The hydrophilic R groups tend to be located on the outside of the protein, interacting with the water in the cytosol. An integral membrane protein, on the other hand, must have a stretch of 18-20 hydrophobic amino acids to cross the very hydrophobic interior of the bilipid membrane [12]. The hydrophobicity of the inside of the membrane is due to the long hydrocarbon chains of the lipid molecule. Hydrophilic amino acids are often restricted to the outside of the membrane.

CHAPTER THREE

LITERATURE REVIEW

3.1 Sequence Determines Structure Determines Function

The structure of a protein determines its function. But determination of the structure of a protein is dependent on many factors. The hypothesis that structure is uniquely determined by the specificity of the sequence has been verified for many, but not all, proteins of known structure [13]. It has been observed that particular proteins known as chaperones often play a role in the folding pathway. These proteins also aid in correcting misfolds. It is still generally assumed that the final structure is at the free-energy minimum [14,15]. Thus, all information about the native structure of a protein may be coded in the amino acid sequence, plus its particular solvent conditions. So far, the exact relationship between protein sequence and structure is not thoroughly understood [14,15]. The gap between the number of known sequences (>170,000 [16]) and the number of known structures (about 42,752 [17]) is widening rapidly. One of the most successful computational methodologies for bridging this gap is using sequence alignment to assist the molecular modeling [18].

Scoring is an important parameter for quantification of alignments when using search algorithms. The score reflects the degree of similarity between the two sequences being compared. The higher the score, the greater the degree of similarity [19]. The space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another is called the gap as shown in Figure 3. Introduction of a gap

causes the deduction of a fixed amount (the gap score) from the alignment score to prevent the accumulation of too many gaps in an alignment. This is called gap-open penalty. Extension of the gap to include additional amino acids is also penalized while scoring the alignment. This is called gap-extension penalty. The raw score S for an alignment is calculated by summing the scores for each aligned position and the scores for gaps. In the following figure, a sample DNA alignment is shown. In amino acid alignments, the score for an identity or a substitution is given by the specified substitution matrix (e.g. BLOSUM62) [20].

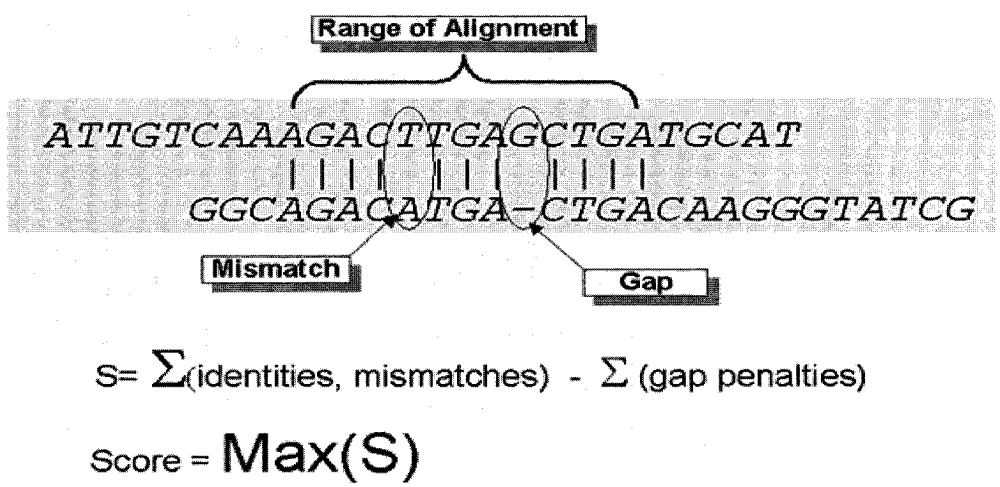

Figure 3. Sample of DNA alignment [21].

There are many protein comparison methods that implement fold identification. This is dependent on the theory that proteins with similar sequences tend to fold into similar structures. One method compares the target sequence with each of the database sequences independently in searching for related protein sequences and structures in a

pair-wise fashion. Another method known as homology structure building, first finds the template sequences with known structure that are similar to original sequence, and then uses them as a template to build the model of a final structure [22].

Altschul and coworkers developed another protein comparison approach that relies on multiple sequence comparisons to improve the sensitivity of the search [23]. For a given sequence, an initial set of homologous sequences from a sequence database was collected, a weighted multiple alignment was made from the query sequence database, and its homologues were scored. A position-scoring matrix was constructed from this alignment, and then this matrix was used to search the database for new homologs. The resulting Position-Specific Iterated BLAST (PSI-BLAST) program runs at approximately the same speed per iteration as BLAST, but in many cases is much more sensitive to weak but biologically relevant sequence similarities. So, several new biologically relevant sequences can be related using PSI-BLAST. Another approach uses multiple sequence alignments in combination with structural information predicted from the sequence of the target [24].

Gross and coworkers found that multiple alignments of secondary structure regions are useful in the identification of key hydrophobic residues when utilizing hydrophobic cluster analysis [25].

3.2 Hydrophobicity

Hydrophobicity plays a key role in protein structure and is often conserved during evolution [26]. Hydrophobic effects arise because water molecules take on new, thermodynamically unfavorable networks of hydrogen bonds when placed at a hydrophobic surface. The most obvious result of this is that the most hydrophobic residues tend to be buried in a protein core to avoid the unfavorable water structure. Therefore, interiors of proteins tend to contain fewer charged and polar residues and more nonpolar residues than the surfaces in contact with water [27]. Hydrophobic effects are not true bonds but they are the determinants of protein three-dimensional structure, and the burial of hydrophobic groups is a significant determinant of stabilization energy for proteins [28].

A large number of different hydrophobicity scales have been developed for amino acids. Significant differences exist among these hydrophobicity scales [27]. Among these, the Hopp-Woods scale utilized predictions of potential antigenic sites in globular proteins, assuming they are likely to be rich in charged and polar residues [29]. The scale is essentially a hydrophilic index resulting in nonpolar residues typically being assigned negative values. Hopp-Woods eventually optimized the Levitt scales determined from the measured free energy of transfer of individual amino acid from water to ethanol [30]. When the available experimental information was insufficient, the scales were estimated from the relationship between accessible surface area and hydrophobicity. The Hopp-Woods parameters are optimized from a number of predicted antigenic determinants.

Many groups of membrane proteins are observed to have alpha-helical secondary structure. Helices are believed to be the most common motif in membrane proteins. It is observed that this motif is relatively easy to spot in a sequence, and most of the transmembrane alpha helices have been successfully predicted using only molecular sequence data. Steitz and coworkers developed a hydrophobicity scale from the hydrophobic and hydrophilic components of transfer of amino acid side chains from water to a nonaqueous environment [27]. The scale has been specifically developed for amino acids in alpha helical structures. After choosing a residue hydrophobicity value, depending on the scale chosen, the total hydrophobicity of a sequence segment can be determined, and depending on the total polarity, its helical propensity can be predicted. Transmembrane helices are usually identified using this method. Transmembrane alpha-helical sequences are characterized by a largely, if not completely, hydrophobic stretch of around 20 amino acids. However, predictions of which sequences will fold into helices may differ slightly depending on which polarity scale is chosen.

Steitz and coworkers stress that many scales are based on side chains partitioning between an aqueous environment and a protein interior, which is a very different case than partitioning between aqueous and lipid environments [27]. The dielectric constant in the lipid environment is low and constant, whereas the prediction of dielectric constant in a protein interior is difficult and is highly variable. The Goldman, Engelman, Steitz (GES) hydrophobicity scale predicts the structure of known membrane helices better than many other known hydrophobic scales, especially those that are known to contain some polar residues [27]. Their method also takes into account the possibility of polar groups

interacting in the bilayer, which would increase the chance of recognizing a transmembrane sequence containing a few polar groups. In the GES scale, the free energy of transfers for both the hydrophobic and hydrophilic components of each amino acid were assigned an individual value. The hydrophobic component of the free energy of transfer for water-oil can be calculated from the surface area of an amino acid side chain in an alpha helix. The hydrophilic term essentially involves polar contributions arising from hydrogen bonding interaction. The hydrophilic term in the GES scale includes the energy required to convert the charged side chains to neutral species at pH 7. For carboxyl groups, the energy cost must be considered in two stages. In the first stage, there is the energy cost of removing the protonated group from contact with the aqueous environment, which involves approximately 4.3 kcal/mole. In the second stage, there is the energy required to protonate the carboxyl group, which is given by

$$\Delta G = -1.36(pK-7) \qquad \text{Equation 3}$$

Sharp, Honig and coworkers [31] have refined the existing experimentally determined residue hydrophobicity values [32]. These experiments determined changes in free energy only for individual amino acids from alcohol to water. Sharp and coworkers made corrections for the free energy changes by including the size of side chains too. All three hydrophobicity scales, the Goldman, Engelman, Steitz (GES) hydrophobicity scale, the Hopp-Woods hydrophobicity scale, and the Sharp–Honig hydrophobicity scale, are different in origin and represent a reasonable sample of the many such scales.

Lustig and coworkers observed strong correlations between sequence entropy and aggregate hydrophobicity [33]. Aggregate hydrophobicity was calculated by averaging Sharp-Honig, Hopp and Woods, and GES hydrophobicity values for each of the proteins. When sequence entropy and aggregate hydrophobicity were overlayed with respect to inverse of packing density, they observed a strong similarity between sequence entropy and aggregate hydrophobicity, even in the anomalous regions of high and low packing density. They concluded that hydrophobicity values correspond to sequence entropy because sequence entropy measures the ability of an amino acid to accommodate mutation and because hydrophobicity measures the degree of burial of an amino acid. This was an interesting correlation because hydrophobicity plays a major role in the correct folding of model protein chains.

3.3 Shannon Entropy

Information theory can be used to identify unique protein secondary structures from patterns of variability in amino acid sequences [34]. Lustig and coworkers have observed conventional generalized chain statistics alone may not be entirely useful in calculating the entropic penalty associated with loop closure in proteins and RNA [35,36]. By exploring large-scale sequence space, Larson et al. [37] have found that sequence entropy values cluster around a specific fold. Applying an expression of Shannon entropy to nucleic acid sequence variability has been proven useful in identifying DNA control regions. Valdar and coworkers further extended this method, so as to measure amino acid conservation in proteins [38,39].

Koehl and Levitt introduced an approach to explore and quantify sequence entropy with respect to protein structure. In their work, they have shown that Shannon-derived entropies for a protein sequence correlate with the entropies calculated from local physical parameters, including backbone geometry [40]. They suggested a twenty-first term for gaps in the calculation of sequence entropy. In their work, they found that the geometry and stability of a given structure defines the compatibility of the sequence space to a protein structure. They did their study for a small set of ten proteins. The sequence information contained in a sequence alignment was converted into a profile matrix with an array of vectors, one for each position in the sequence. Each of the vectors contained twenty-one values representing the frequencies of occurrence of all twenty types of amino acids plus the gap at the considered sequence position. Sequence entropy was calculated using the Shannon entropy expression for amino acids. Sequences of all protein structures that were homologous to the protein of interest were extracted from the fold classification (FSSP) database and then optimized by averaging the entropy per residue. The structural alignments of these homologous proteins were used to measure structural entropy, Sstr. For the small set of ten proteins, the entropy derived from the sequence information (Sseq) correlated well with the entropy derived from structure information (Sstr). This was an interesting observation, because Sseq and Sstr are two independent measures of the size of the same sequence space, derived from two different databases. So, if bias existed in these databases, it must be small.

When sequence entropy per residue was plotted as a function of the position in the protein sequence, they observed that the sequence spaces that are compatible with two

proteins of similar length possess different sizes. This was due to the dependency of their calculation on the geometry of the protein backbone, on the amino acid composition of the protein sequence, and on the stability of the fold. The calculated design entropy correlated well with the observed structural entropy. This explained the diversity in sequence space observed among known proteins sharing the similar fold. Their results suggest that sequence space exploration using sequence entropy values may be useful for identifying highly designable folds of a given protein.

Lustig and coworkers have also found applications using Shannon entropy exclusively for a large set of protein alignments [33]. They investigated the Shannon information entropy with respect to the local flexibility of globular proteins. Strong correlations were observed between sequence entropy (at each residue), residue flexibility and hydrophobicity. Strong linear correlation was observed between sequence entropy and the inverse packing density, when average sequence entropy was plotted against the inverse of packing density, except at the highest and low ranges of densities. This provided a quantitative relationship between sequence entropy and packing density and thus, an important structural measure for determining likely sites for mutation. They stressed that the packing at the residue level for coarse-grained protein structures exhibited a strong relation to sequence conservation, when it was averaged over large number of residues. This averaging was necessary to obtain a single representation of all the combinations in which a residue's atoms may be packed.

When sequence entropy was averaged over the set of all proteins, the entropy values correlated well with the regression line, whereas there was a lot of variation in the

entropy values calculated for each of the proteins. Sequence entropy values in Liao's data were calculated using the Shannon entropy expression for all twenty amino acids, and in this case gapped regions were ignored. Variability that is higher or lower than average is typically rationalized as resulting from noise in the data, and the introduction of a gap value at the twenty-first term for the calculation of sequence entropy might reduce the noise in the data.

3.4 Packing Density

The packing density of a given protein is defined as the ratio of the volume occupied by its van der Waals (VDW) envelope to the volume it actually occupies. The recent studies involving the measurement of the packing in proteins have shown that the packing inside proteins is somewhat tighter than observed initially. It was also observed that the overall packing efficiency of atoms in the protein core is greater than in crystals of organic molecules. When molecules are packed this tightly, small changes in packing efficiency are quite significant. In this scenario, the limitation on close packing is hard-core repulsion, so even a small change in the packing conformation results into a quite substantial change in the free energy. In addition to this, Richards and Lim [41] pointed out that the number of allowable configurations that a collection of atoms can adopt without hard-core overlap drops off very quickly as these atoms approach the close-packed limit.

Researchers studying protein structure using highly simplified two-dimensional lattice models have pointed out that tight packing in the protein core may drive or force

the formation of secondary structures. This theory has been tested on somewhat more realistic off-lattice models of protein structure [41]. The results of this calculation have been mixed in the sense that these models do observe high packing density driving the formation of secondary structure but to a much lesser degree than in the lattice models. The exceptionally tight packing in the protein core seems to require a precise jigsaw puzzle–like fitting together of the residues inside proteins. This theory holds well for the majority of atoms inside proteins. However, there are exceptions, and some studies have focused on these, showing how the packing inside proteins is punctuated by defects or cavities. These defects can accommodate buried water molecules, if they are large enough. The packing efficiency can also be studied by comparing the protein core to the protein surface [42]. This comparison is especially interesting from the packing density perspective because the protein surface is covered by water. The protein surface is known to be packed usually much less tightly than the protein core and in a distinctly different fashion. When comparative studies for packing were done at internal interfaces inside of proteins, particularly at domain-domain interfaces, packing was found to be closely coupled with protein flexibility.

Local packing density is measured as the alpha carbon packing density, which is calculated from the associated atomic coordinates of the alpha carbons. Bahar and coworkers observed a correlation between the local flexibility and the inverse of local packing density [43]. They observed that, the higher the inverse of local packing density, the higher is the local flexibility. Lustig and coworkers have found that there is a strong correlation between sequence variability and local amino acid flexibility [33].

3.5 Summary

The structure of a protein is determined by the linear sequence of amino acids. Past research shows the importance of the sequence in predicting the protein structure. Though protein structure prediction based on its sequence is far from completion, the growing databases of sequences may facilitate the daunting task. Many factors such as sequence entropy, hydrophobicity, and packing density play a major role in this sequence to structure prediction. To understand protein stability and function, we have to understand the interplay between entropy, structure and sequence. Sequence entropy calculations are currently useful in exploring aggregate behavior of proteins in general.

Koehl and Levitt [40] have shown that Shannon-derived entropies for a protein sequence for a set of ten proteins correlate well with the entropies calculated from local physical parameters, including backbone geometry. Lustig and coworkers [33] have observed that mutability as measured by sequence entropy was inversely correlated to packing density for a set of one hundred and thirty query proteins. They ignored gapped regions in the calculation of sequence entropy values of the proteins using the Shannon entropy expression.

The introduction of gap information at the twenty-first term in the calculation of protein sequence entropy values may significantly increase the correlation between the sequence entropy and the inverse of packing density.

CHAPTER FOUR

RESEARCH OBJECTIVE

4.1 Objective

The objective of my research is the following:

To implement and test modifications to sequence entropy calculations that significantly reduce the noise in the data when sequence entropy is plotted against the inverse of packing density.

Earlier studies suggest that, for most residue positions, sequence entropy is inversely correlated to their packing density. Calculation of sequence entropy values with an introduction of a gap term may reduce the variability in the data that is higher or lower than the average trend (which is typically rationalized away as resulting from noise in the data). Thus, this study will explore the connections between protein structure and sequence entropy, by the following:

- Recalculating alignment sets for additional query proteins for the set of 130 original query proteins [44]. Changing the default setting from a maximum of 100 alignments to 500 for a subset of 20 proteins while doing the BLASTP search.
    - This allows one to check how variable the sequence entropy results are when larger sets of alignments are allowed.
- Calculating the sequence entropy at each residue position for an additional term that involves gap information.

- Set filters to see whether any one of the following correlations offers improvement in the data noise of aggregate plots of sequence entropy versus inverse of packing density for 130 proteins and for each individual protein.
    - Gap included average sequence entropy versus inverse of packing density, Gerstein-Altman average entropy versus inverse of packing density, window averaged sequence entropy versus inverse of packing density, and normalized B-factor are utilized as optimum measure of inverse C_α packing density in correlation plots.
- Create a subset of 20 proteins from 130 proteins and check the consistencies in the above mentioned correlations.

CHAPTER FIVE
MATERIALS AND METHODS

5.1 Materials

- 130 query proteins and sequence alignment set from Liao [44].
- Sequence alignment and entropy calculation program by Yeh [45] with modifications to the script to incorporate the gap term.
- PERL script to calculate the packing densities of the query proteins.
- Computer with large bandwidth capabilities to access BLAST and PDB.
- Microsoft Excel for analyzing and merging data generated by PERL programs.

5.2 Method

The raw data for 130 query proteins and their sequence alignment sets were obtained from the Lustig group [33]. The raw data were then merged and used as input data for the modified PERL programs. Output files generated by the PERL programs were loaded into Microsoft Excel for analysis. Figure 4 gives an overview of the method that was followed to calculate the packing densities and sequence entropy values for the query proteins from the sequence alignments.

The below mentioned steps will be followed:

5.21 Packing Density and B-factor Values

- A set of 130 query proteins with known 3D structure data from the Protein Data Bank (PDB) were compiled.
- For each query protein residue, a 9 Å radius C_α packing density was calculated from its atomic coordinates from the PDB database.

- The inverse of packing density for each query residue was calculated by finding the reciprocal of C_α packing density of the residue.
- Atomic distance involving any two residues i and j of the protein was calculated using the following formula:

  $$D(i,j) = \sqrt{((x(i) - x(j))^2 + ((y(i) - y(j))^2 + ((z(i) - z(j))^2} \qquad \text{Equation 4}$$

- The temperature factor or B-factor of each query residue was taken from the PDB data file of each protein from the PDB website.
- Yeh's PERL program was modified to read and print the temperature factor value along with the packing density value for each query residue.

5.22 Sequence Alignment:

- Homologous protein sequences for the query proteins were searched using the BLASTP.
- On the BLASTP query page, the number of alignments was set to 10-100 by Liao for the 130 proteins. To see the effect of more alignments, the number of alignments was set to 10-500 for a subset of 20 proteins
- The resulting number of alignments for a particular query should be more than 10 to be of any statistical significance and hence, results with fewer alignments are discarded.

5.23 Sequence Entropy and Gerstein-Altman Entropy

- A gap term was introduced for the sequence entropy calculations.

- The Gerstein-Altman entropy was calculated using equation 5, where the first term is the most basic Shannon (sequence) entropy expression, the second term is the entropy for random substitution S^R for each protein, and P_j is the probability of observing amino acid type j in the whole sequence alignment, and j ranges from 1 to 20 [39].

  $$G.E. = -\sum_{j=1,20} P_{jk} \log_2 P_{jk} + \Sigma_{j=1,20} P_j \log_2 P_j \qquad \text{Equation 5}$$

  Single averaging for the Gerstein-Altman sequence entropy can be implemented where each residue position can be averaged within an interval of inverse of packing density. Similar single averaging for the sequence entropy, described by the first term in equation 5, can also be implemented [33].

- Gap-included sequence entropy for each residue position was calculated using the Equation 6 that was embedded in the PERL program.

  $$S.E. = -\sum_{j=1,21} P_{jk} \log_2 P_{jk} \qquad \text{Equation 6}$$

  The PERL program was modified to include the gaps in sequence entropy calculations, where the BLASTP query results file was used as an input to the PERL program. The gap-included sequence entropy values are then averaged within an interval of inverse of packing density.

- Window averaged entropy (W.A.E) for each residue position was calculated using Equation 7 that was embedded in the PERL program. S.E. in the Equation 7 was calculated using the Equation 6, where N is the length of the window used for averaging (additional averaging within an interval of inverse packing density can also be implemented).

  $$W.A.E. = (1/N) (\sum_{j=1,N} S.E.), \text{ where } N = 2, 3, 4, 5 \qquad \text{Equation 7}$$

5.3 Analysis

Various Correlation plots will be plotted to understand the relationships. The different plots that will be analyzed are as follows:

- Gap-included sequence entropy versus inverse of alpha carbon packing density for each of the query proteins.
- Aggregate gap-included single average sequence entropy (sequence entropy for all query proteins) versus inverse alpha carbon packing density.
- Gerstein-Altman entropy versus inverse of alpha carbon packing density for each of the query proteins.
- Gerstein-Altman single average entropy (G.E. for all query proteins) versus inverse alpha carbon packing density.
- Window averaged sequence entropy versus inverse alpha carbon packing density.
- B-factor values for each query residue versus gap included sequence entropy for all the 130 proteins.

5.31 Sequence Entropy versus Inverse Packing Density

- Sequence entropy is plotted on the ordinate and the inverse C_α packing density on the abcissa.
- Linear regression is applied to the correlation plot. R square values, slope, and intercept for each protein is calculated from the linear fit and regression [46].

5.32 Aggregate Sequence Entropy versus Inverse Packing Density

- Aggregate sequence entropy is plotted on ordinate and inverse C_α packing density on the abscissa. This plot will show the relationship between entropy and packing density.
- Single averaging is done by adding individual residue entropies for a particular C_α packing density interval from all query protein sets of alignments.

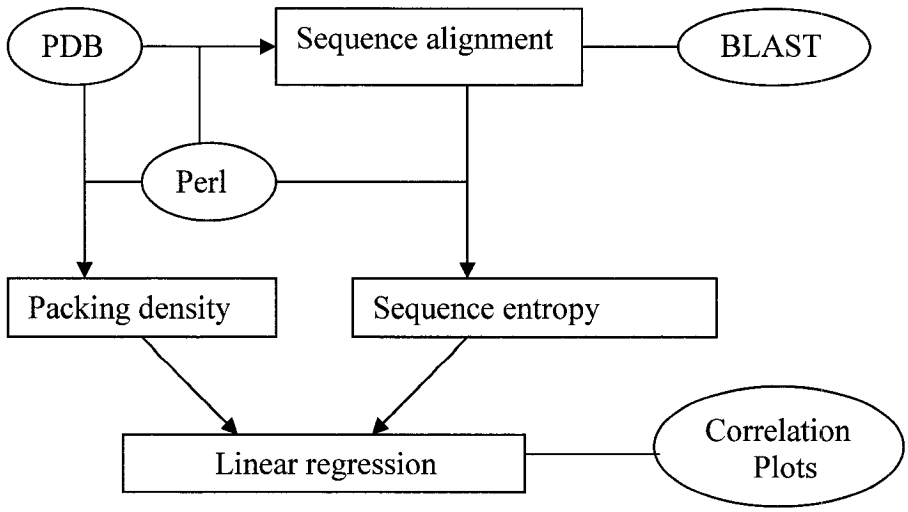

Figure 4. Flow diagram of the method.

CHAPTER SIX

RESULTS

6.1 Gap-included Single Average Entropy

The plot of aggregate average sequence entropy versus inverse packing density for all the 130 query proteins is shown in Figure 5. The average sequence entropy values were calculated from the alignment sequences generated by protein BLAST by including the gap term in the residue entropy calculation. The gaps were treated as an additional 21st term along with 20 amino acids in the residue entropy calculation. In the previous study done by the Lustig group [33], gaps were ignored in the residue entropy calculations for the average sequence entropy versus inverse density plots as shown in Figure 7. A gap term was introduced to see whether it offers any improvement in the noise of the plots. As shown in Figures 5 and 6 there are two main regions associated with, high and low density [33]. In the high-density region, there is a strong correlation between entropy and inverse packing density for inverse packing density values 0.040 to 0.083. There is no big improvement in the data noise due to gap introduction in the entropy calculations but a slight improvement was observed in the linear correlation. The linear correlation for the overall correlation plot in Figure 5 was found to be $y = 2.236x + 0.6507$, with a correlation coefficient of 0.2995 and $P<0.001$, whereas the linear correlation for the overall correlation plot in Figure 6 was found to be $y = 0.7996x + 0.6843$, with a correlation coefficient of 0.040 and $P<0.001$. Also the linear correlation

for major region I in the gap-included entropy plot shown in Figure 5 was found to be y = 17.862x - 0.2768, with a correlation coefficient of 0.9945 and P<0.001.

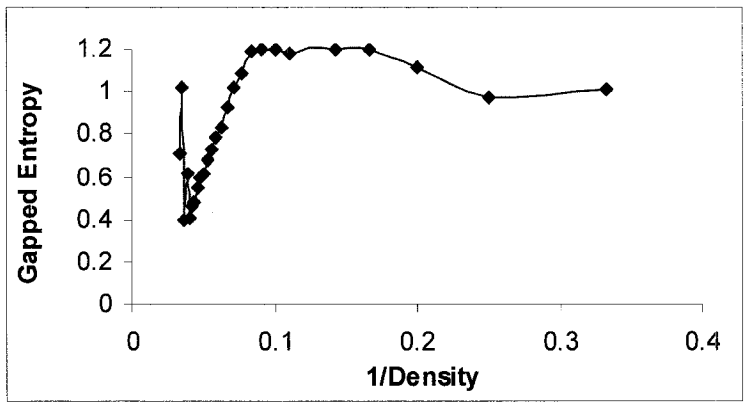

Figure 5. Aggregate graph of average sequence entropy plotted against inverse packing density, for the packing radius 9 Å for 130 query proteins. The average sequence entropy (ordinate) is calculated for each inverse packing density, by averaging the entropy values of all the residues within a density interval of one. The entropy values were calculated by taking into account of the gap term for each protein. The inverse packing density (abscissa) of each residue is determined from the C_α coordinates supplied by the PDB protein files, using the 9 Å packing radius.

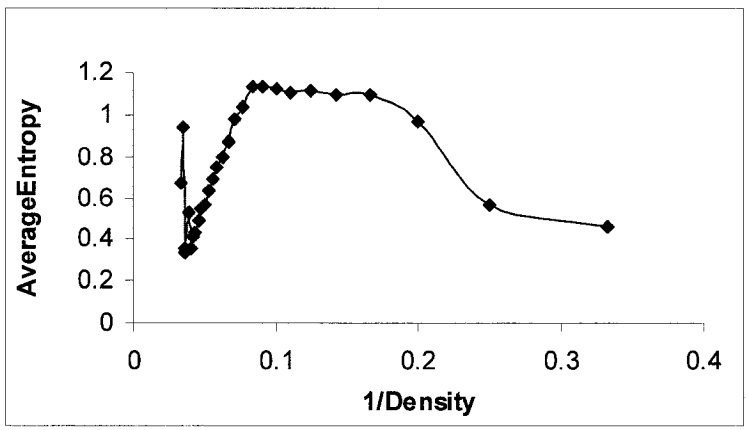

Figure 6. Aggregate graph of average sequence entropy plotted against inverse packing density, for the packing radius 9 Å for 130 query proteins, where entropy values were calculated by excluding the gap term for each query protein residue.

The correlation plot in Figure 5 consists of four regions, flanking region I with decreasing sequence entropy with respect to inverse packing density from inverse packing density 0.03 to 0.038 (35to 26 C_α), major region I with increasing sequence entropy with respect to inverse packing density from inverse packing density 0.04 to 0.083 (25 to 12 C_α), major region II with flattening of sequence entropy with respect to inverse packing density from inverse packing density 0.09 to 0.17 (11 to 6 C_α), and flanking region 2 with decreasing sequence entropy with respect to inverse packing density from inverse packing density 0.2 to 0.33 (5to 3 C_α).

There is a total of 42,253 residues for the 130 proteins, and flanking region 1 has 102 residues (0.24 % of total residues), major region I has 24, 392 residues (57.7 % of total residues), major region II has 16,977 residues (40.1 % of total residues), and flanking region 2 has 782 residues (1.8 % of total residues).

6.2 Gerstein-Altman Single Average Entropy

The plot of aggregate average Gerstein-Altman entropy versus inverse of packing density for all of the 130 query proteins is shown in Figure 7. The same trend is observed in the four regions for the average gerstein entropy plot that was calculated by taking the effect of random entropy for each protein into account. When compared to the average entropy plot shown in Figure 6 there is no significant improvement in the data noise for the average Gerstein-Altman plot shown in Figure 7. There is a slight improvement in the linear correlation of the average entropy within the packing density interval; the linear correlation for the plot in Figure 7 was found to be $y = -0.2065x + 0.2618$, with a

correlation coefficient of 0.055 and $P<0.001$, whereas the linear correlation for plot in Figure 6 was found to be $y = 0.7996x + 0.6843$, with a correlation coefficient of 0.040 and $P<0.001$.

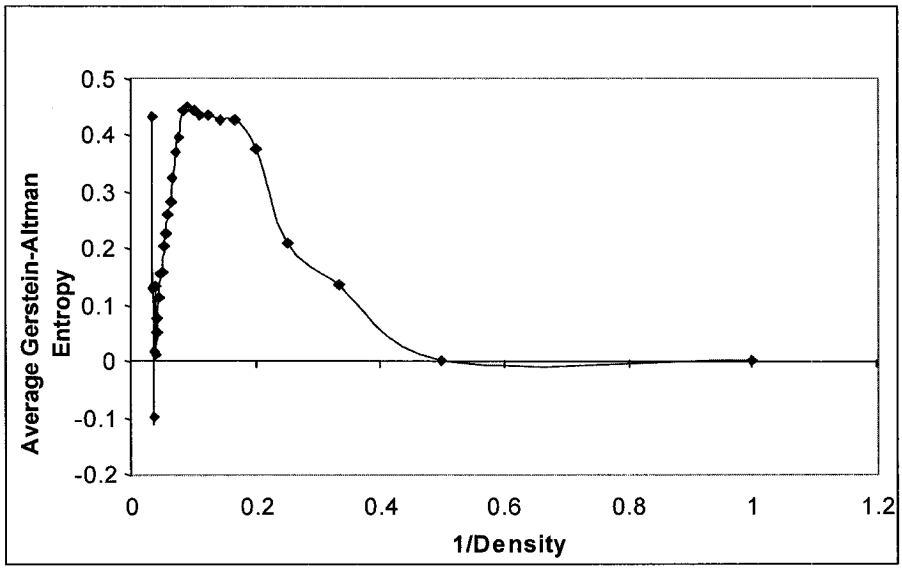

Figure 7. Aggregate graph of average Gerstein-Altman sequence entropy plotted against inverse packing density, for the packing radius 9 Å for 130 query proteins. The average Gerstein-Altman sequence entropy (ordinate) is calculated for each inverse packing density, by averaging the entropy values of all the residues within a density interval of one. Gerstein entropy values were calculated by taking into account the Gerstein-Altman term. The Gerstein-Altman sequence entropy was calculated by subtracting sequence entropy S_k with the random entropy S^R for each protein. Inverse packing density (abscissa) of each residue is determined from the C_α coordinates supplied by the PDB protein files, using the 9 Å packing radius.

6.3 Window Averaged Sequence Entropy

Neighbors are the amino acid residues that are adjacent to each other within a given packing radius. To see the effect of different window lengths for averaging, Figure 8 shows the average sequence entropy plots for 9 Å packing radius, using 2, 3, 4, and 5 neighbors. The window averaging serves as a tool for smoothening the curve [46]. The significance of different window sizes for averaging is to determine what window length greatly reduces standard deviation and reduces noise associated with data points away from linear correlation. Summing the single average entropy values corresponding to the window length and then dividing by the respective window size determines the window-averaged entropy. The window averaged sequence entropy of three neighbors is determined by summing the sequence entropy values of residue and above and below neighbors of a given residue and later dividing by three.

Table 1 shows the mean, median, and standard deviation of average entropy in each bin of C_α packing density with respect to each window length for averaging sequence entropy. It shows averaging with a window size of three neighbors greatly reduces the standard deviation compared to averaging with other window sizes, but causes only a slight reduction in the mean and median values. However, except for the reduction in standard deviation with a window size of three and a slight overall difference in magnitude for each window averaged sequence entropy plot, as shown in Figure 8 all the average entropy curves with different window sizes are effectively identical.

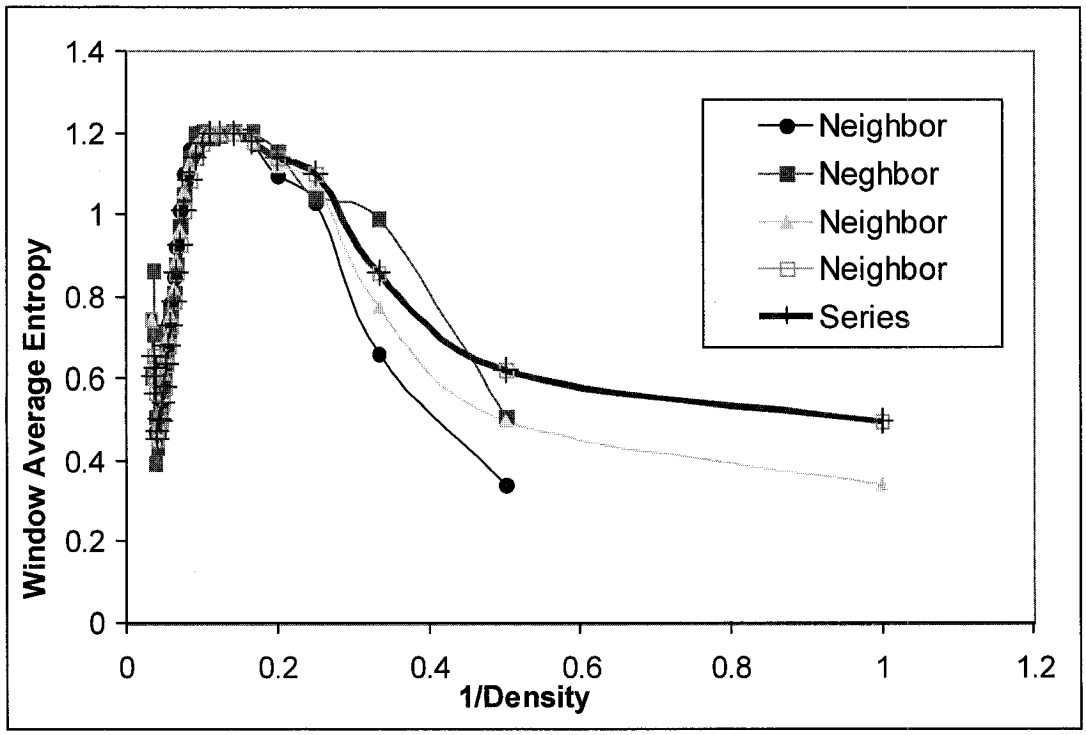

Figure 8. Window Averaged Correlation Plots (Window Averaging-3)
Aggregate graph of window averaged entropy plotted against inverse packing density, for the packing radius 9 Å for 130 query proteins. The average sequence entropy is calculated as explained in Figure 1. The length of window used for window averaging was 2, 3, 4, and 5 neighbors.

Table 1. The mean, median, and standard deviation of each window-averaged entropy, calculated using window length of 2, 3, 4, 5.

Window Averaging Length for Sequence Entropy	2	3	4	5	Gapped Entropy without Window Averaging
Mean	0.800	0.790	0.804	0.810	0.792
Median	0.782	0.740	0.750	0.735	0.806
Standard Deviation	0.500	0.314	0.604	0.605	0.500

Figure 9 shows the standard deviations of average entropy with respect to each packing density. It shows that the window average entropy with a window length of three neighbors greatly reduces the standard deviation compared to an average entropy calculated by taking into account the gap term. The average standard deviation for the window average entropy is 0.313, compared to 0.5 for the gapped entropy. There is a 38% reduction in standard deviation for averaging the entropy with a window length of three neighbors. However, except for the reduction in standard deviation with window averaging, the two types of averaged sequence entropy are effectively identical.

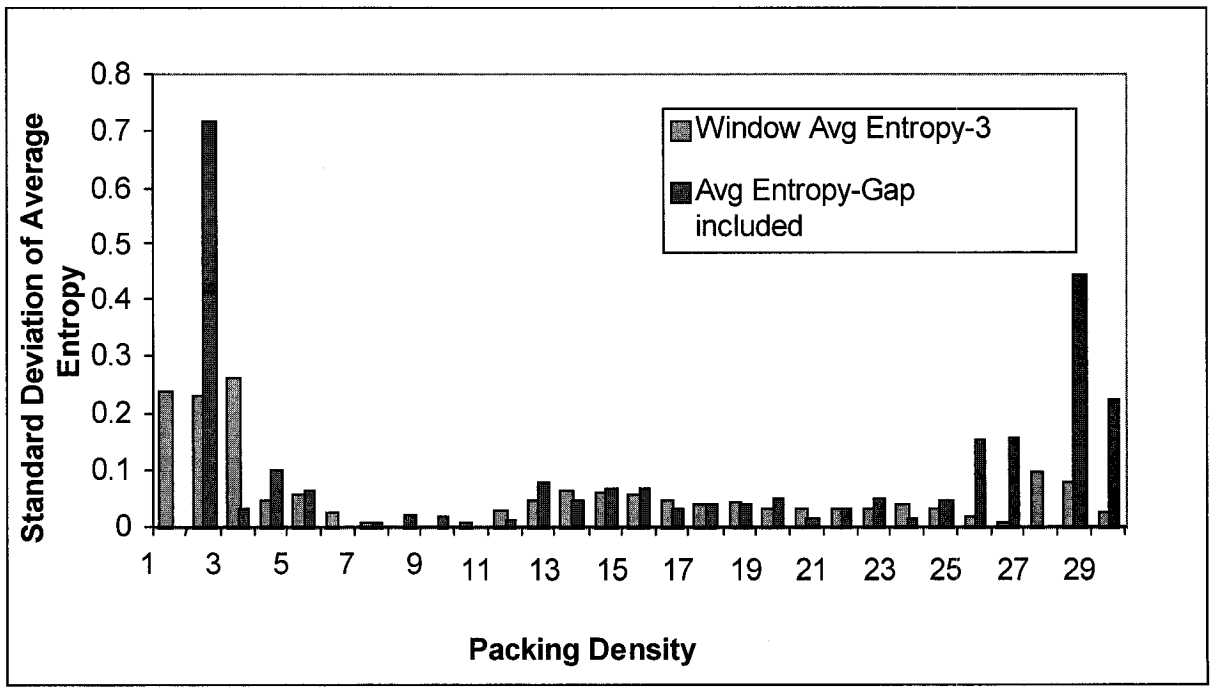

Figure 9. Standard deviations of average entropy with respect to each packing density. Estimated average standard deviation for window average entropy is 0.313, and 0.5 for average entropy.

6.4 Correlation Pattern for 20 Proteins

A subset of 20 proteins was selected from 130-protein list to check whether the same kind of correlation exists between sequence entropy and inverse density for these 20 proteins as shown in Figure 5. The list of 20 proteins is shown in Table 2.

Criteria for the selection of these 20 proteins were:

(1). 20% of them, i.e. at least four proteins, should have multifunctional sites and multiple chains.

(2). Proteins from different size ranges i.e. proteins with both a small and large number of residues.

There is a total of 7474 query residues for the 20 proteins, and flanking region 1 (35to 26 C_α) has 19 query residues (0.25 % of total residues), major region I (25 to 12 C_α) has 4068 query residues (54.4 % of total residues), major region II (11to 6 C_α) has 3260 query residues (43.6 % of total residues), and flanking region 2 (5 to 3 C_α) has 127 query residues (1.6 % of total residues).

Figure 10 shows that the correlation pattern between sequence entropy and inverse density for these 20 proteins is similar to that of the 130 query proteins. Aggregate correlation plots for the highly populated region (packing of 25 to 12 C_α within 9 Å radius) are shown in Figures 10 and 11. These plots show that the average entropy values for the 130 query proteins and for the 20 query proteins are strongly

linearly correlated with respect to inverse packing density. The straight-line fit for the aggregate average sequence entropy versus inverse packing density for the 130 proteins is $y = 17.862x - 0.2768$; the correlation coefficient is 0.9945; $P<0.001$. The straight-line fit for the aggregate entropy plot for the 20 proteins is effectively identical: $y = 17.544x - 0.2781$; correlation coefficient is 0.9553; $P<0.001$.

Table 2. List of 20 proteins with their residue count, where shaded proteins [47] can be classified as multimeric proteins.

1a1i - 85	1bf2 – 750	1e3q - 596	**4dfr - 159**
1a1s - 314	1bg3 - 918	1omd – 108	5acn – 754
1a3s - 160	1boh – 295	**1ton – 235**	**5rub – 465**
1agm - 470	1crc – 105	**2cts - 437**	**7cat – 506**
1aln - 294	**1dht - 327**	3psg - 370	**8atc - 310**

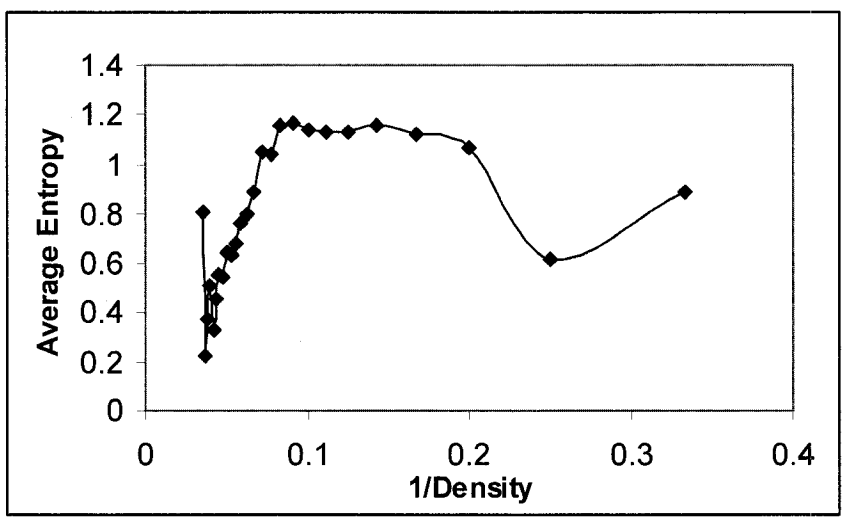

Figure 10. The aggregate graph of average sequence entropy plotted against inverse packing density, for the 9 Å packing radius of 20 query proteins. The straight-line fit for this plot is $y = 1.6395x + 0.6511$; correlation coefficient is 0.1632; $P<0.001$.

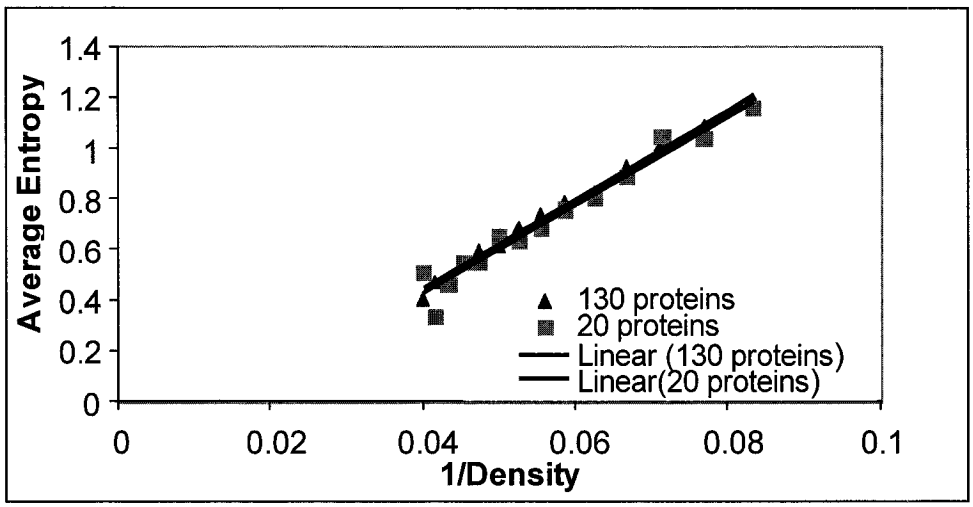

Figure 11. Linear regression of selected regions (packing of 25 to 12 alpha carbon atoms within 9 Å radius) for correlation plots involving 24,392 residues for 130 proteins and involving 4068 residues for 20 proteins. Aggregate average entropy and inverse density are calculated as mentioned in Figure 6. The straight-line fit for the aggregate average sequence entropy (triangles) versus inverse packing density for 130 proteins is $y = 17.862x - 0.2768$; correlation coefficient is 0.9945; $P<0.001$. The straight-line fit for the aggregate entropy plot for 20 proteins (squares) is effectively identical: $y = 17.544x - 0.2781$; correlation coefficient is 0.9553; $P<0.001$.

6.5 Frequency Distributions

Various frequency distributions of the 20 query proteins {listed in Table 2} residues are shown in Figure 12. Frequency distribution plot of total of 7474 residues of all 20 proteins with respect to each packing density interval shown in Figure 12A is slightly right- skewed. Packing density values range from 0 to 30 and average, standard deviation, standard error values for the frequency per density interval are 0.804, 0.296

and 0.82 respectively. Packing density values of 1, 2, 29, and 30 C_α atoms within 9 Å radii have no residues, while packing density value of 10 C_α atoms has the largest number of 769 residues.

Frequency distribution data for the length of query proteins is shown in Figure 12B. The average, median, mode number of amino-acid length per query protein for all 20 proteins are 2, 2, 1 respectively. The range of query residue values is from 85 to 918 with 1a1i having 85 residues and 1bg3 having 918 residues.

The query sequence alignment search by NCBI BLAST generates a result of 3285 aligned protein sequences for all 20 proteins. Alignments were excluded if bit score values fall below 100 and alignments should be greater than or equal to 40% of the best bit score to be included in the result set. Calculations for a set of 130 query proteins showed 40% of the best BLASTP bit score as a reasonable cutoff with respect to sequence entropy calculations within each C_α packing density interval [45]. Earlier sequence entropy calculations done by Lustig's group allowed only a maximum number of 100 alignments. This maximum default value was initially designed to limit any size bias, however an alternative of effectively unlimited size was also explored (this was met by 500 alignments, maximum alignment value allowed by BLASTP). This was done because, for proteins with a high degree of homology, number of alignments may need to be larger than 100 to generate reasonable sequence entropy values. The frequency distribution of number of alignments is shown in Figure 12C. The alignment values range from 7 to 500. 1bf2 has the least number of alignments with a value of 7 and 1a1ihas highest number of alignments with a value of 500.

The bit score is the most important criterion for selecting alignments from an alignment set generated by BLASTP. The alignment bit score is calculated by normalizing the raw score with respect to the scoring system. The bit score values range from 197 to 1829 for 20 query proteins. Protein 1a1i has the highest bit score value of 197, and protein 1bg3 has a value of 1829. The frequency distribution of BLAST bit scores shown in Figure 12D is consistent with the right-skewed distribution for 130 protein set BLAST scores [33].

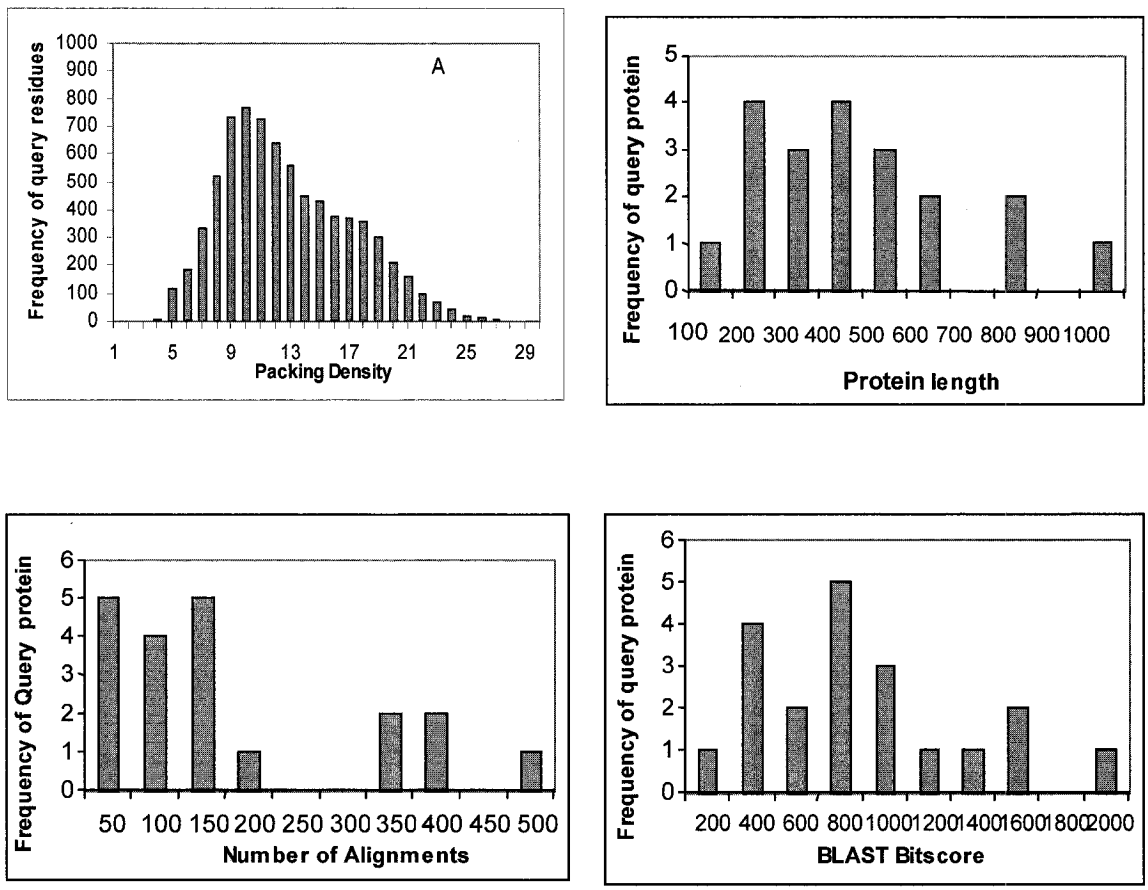

Figure 12. Various frequency distributions of 20 query proteins and BLASTP alignments.

6.6 Individual Protein Correlation Plots

Examples of protein correlation plots (sequence entropy for each query residues plotted against inverse of C_α packing density) showing the best, medium, and worst fit among the 130 query protein set are given in Figures 13A thru 13C. Representatives of the range of plots with corresponding P values are lactate dehydrogenase (5ldh, 333 residues), hexokinase (1bg3, 901 residues), and troponin (4tnc, 160 residues). The respective straight-line fits for the raw data are $y = 11.707x - 0.297$ ($P<0.001$), $y = 9.0421x + 0.2307$ ($P<0.001$), and $y = 3.9208x + 0.4608$ ($P<0.001$), and the respective correlation coefficients are 0.1792, 0.1039, and 0.0197. The correlation plots between sequence entropy and inverse of C_α packing density show a large scattering of data points for all 130 proteins. However, a similar correlation trend is observed for the straight-line fits of all proteins. The respective straight-line fits of average sequence entropies with respect to inverse of C_α packing density are $y = 13.309x - 0.3915$, $y = 7.1014x + 0.277$, $y = 0.8465x + 0.6427$, and the respective correlation coefficients are 0.9162, 0.7937, and 0.0061. Correlation coefficients between average sequence entropies and the inverse of C_α packing density are larger than the ones between the raw sequence entropy and inverse of the C_α packing density for most proteins. This trend remains valid for the three proteins shown in Figure 13. For an individual protein; two major regions and two flanking regions can be indicated. Interestingly, the best linear correlation fit may correspond to the least representative example of such four-region behavior.

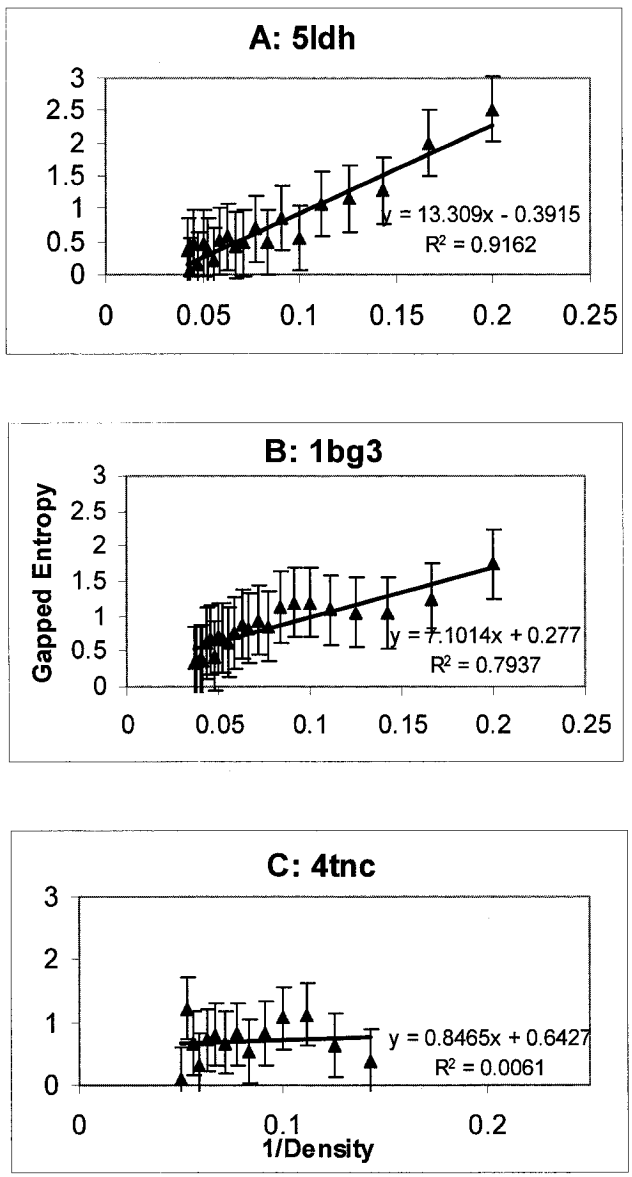

Figure 13. Correlation plots of sequence entropy and inverse of packing density for a range of sample proteins to show best, medium, and worst fit. A) For dehydrogenase (5ldh, 333 residues), the straight-line fit for raw data is y = 11.707x - 0.297; correlation coefficient is 0.1792. B) For hexokinase (1bg3, 901 residues), the straight-line fit for raw data is y = 9.0421x + 0.2307; correlation coefficient is 0.1039. C) For troponin (4tnc, 160 residues), the straight-line fit for raw data is y = 3.9208x + 0.4608; correlation coefficient is 0.0197.

6.7 Normalized B-factor Values

B-factors or temperature factors of amino acid residues provide experimental information about the flexibility and dynamics of a protein. Different scales are used to measure B-factors in different protein structures due to differences in their refinement procedures. So, raw data may be normalized to compare B-factors of different protein structures [48]. The B-factor values of 130 proteins were normalized using equation 10, where B' is the normalized B-factor, B is the B-factor value, <B> is the average value of all C_α atoms, and $\sigma(B)$ is the standard deviation of the B-value for the chosen protein.

$$B' = \{B - \langle B \rangle\}/\sigma(B) \qquad \text{Equation 10}$$

The total number of residues and total entropy were then calculated within each bin of normalized B-factor values. Average entropy with respect to B-factor values was then calculated by dividing the total entropy by the total residues within each normalized B-factor bin. Each B-factor bin is 0.1 times the temperature value, and temperature values that were calculated for 130 proteins from the PDB ranged from –2 to +6. So, there are 80 bins of averaging for the plot shown in Figure 14. The plot of aggregate average sequence versus normalized B-factor for all of the 130 proteins is also shown in Figure 14. There is a linear correlation between temperature factor or B-factor and the mean square displacement of an atom, and thus B-factor values provide an indication of protein flexibility [49].

A correlation of similar pattern to that of average sequence entropy versus inverse density plot as shown in Figure 5 was expected for the average sequence entropy

versus normalized B-factor plot as shown in Figure 14, as inverse density is a measure of flexibility of a protein. But, Figure 14 shows a different correlation and, hence, the normalized B –factor plot was discarded as a proper tool for checking the noise reduction in the data.

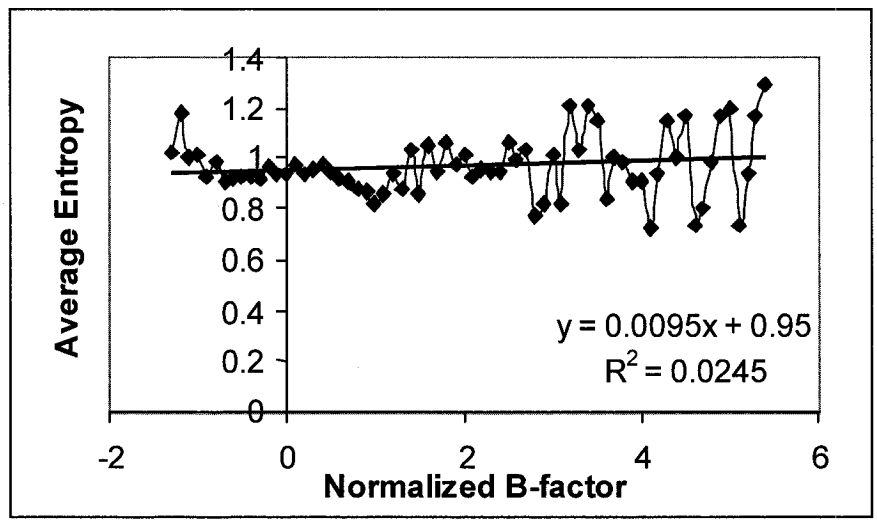

Figure 14. The aggregate graph of average sequence entropy plotted against normalized B-factor values for 130 query proteins.

Four different correlation plots – average entropy versus inverse density, gapped (i.e. gap included) entropy versus inverse density, window average entropy versus inverse density, and Gerstein-Altman entropy versus inverse density and linear correlation data tables for these plots are shown for each of 130 proteins in appendices A, and B respectively.

CHAPTER SEVEN

DISCUSSION

The connections between sequence entropy, protein flexibility, and structure are continuously being explored. Shannon entropy or sequence entropy is essentially a measure of sequence variability [39]. The sequence entropy of a protein was observed to be closely related to the packing density and hydrophobicity of a protein [33]. Investigation of the correlation between sequence entropy and flexibility (measured by inverse C_α packing density) within a 9 Å packing radius for 130 query proteins yielded a four- region behavior. The size of the 130 query proteins ranged from 85 to 901 residues. Major region I was shown by the Lustig group to span the range of C_α packing densities from 12 to 25 amino acids indicating a strong correlation between sequence entropy and inverse packing density, and this region contains a majority of the total residues for the 130 query proteins. At higher inverse densities, nearly all of the remaining query residues were associated with relatively constant sequence entropy, most likely corresponding to the surface of a protein. Again, the 9 Å C_α packing radius was found to be the optimum in the previous studies involving sequence entropy and inverse packing density [45].

The major objectives of this research were to reduce noise in the correlation plots obtained for sequence entropy and inverse of C_α packing density, and to develop and evaluate methods to reduce noise associated with sequence entropy. Three aggregate plots for 130 query proteins – gap-included average entropy, Gerstein-Altman entropy,

and window-averaged entropy, all plotted against inverse of C_α packing density for a 9 Å packing radius - show some improvement in data noise in terms of linear correlation compared to gap-excluded average entropy plotted against inverse of packing density.

When the data shape for all the aggregate plots shown in Figures 5, 6, 7 and 8 is compared, the overall trend of the data shape looks similar with the occurrence of two major regions and two flanking regions. Data shape comparison between gap-included average entropy plot shown in Figure 5 and gap-excluded average entropy plot shown in Figure 6 shows that the introduction of a gap-term in the sequence entropy calculations moderates a downward shift in the correlation in the flanking region 2. Flanking region 2 is the most flexible region of all the four regions (with packing density of 5 to 3 C_α for a 9 Å packing radius), and the correlation shift might have occurred due to the increase in flexibility because of the introduction of the gap term. There are 42,253 residues for 130 proteins, and only 782 residues out of them lie in the flanking region 2, so there is not enough representative data for this region to make a conclusive observation.

The individual graphs of sequence entropy versus inverse of packing density are plotted for each protein to see if there is an improvement in the noise. These individual plots are shown in Appendix A. The majority of these individual plots exhibit a similar pattern of occurrence of four regions as observed in the aggregate plots. The slope, intercept, and R^2 values obtained from the linear fit and regression of each of the plots are tabulated into tables. These tables are shown in Appendix A. Few of the proteins were observed to contain a negative slope in the linear fit. There are only 3 proteins (1ebv, 1lz1, and 3pgk) out of 130 proteins with negative slopes in the linear fit for the major

region I. The most sensitive and quantitative signature of anomalous behavior is the presence of a negative slope, but when all 130 plots were visually inspected, additional proteins were observed to contain a flat major region I along with a lack of indication of any other behavior.

The major region I in the 130-protein aggregate gap set entropy plot with respect to inverse of C_α packing density, shown in Figures 5 and 12 (from 25 to 12 C_α), was found to contain 57.7 % of the total residues that exhibit a strong linear dependence between sequence entropy and inverse packing density. This indicates a strong correlation between sequence variability and local amino acid flexibility, because ability of a local structure to accommodate mutation increases with the increased flexibility at a residue position [36]. The containment of a large fraction of residues in major region I is consistent with the pattern observed by Lustig and coworkers where 74.9 % of the total residues were observed in the major region I [33]. The major region II (from 11 to 6 C_α) was found to contain 40.1 % of the total residues that exhibit flattening of sequence entropy with respect to inverse packing density, indicating the presence of strongly hydrophobic residues.

Anomalous behavior is observed in the two flanking regions, the regions above packing densities of 26 C_α and below densities of 5 C_α. The region above 26 C_α (ranging from 26 to 30 C_α) contains only 102 residues and the region below 5 C_α (ranging from 5 to 3 C_α) contains 782 residues out of a total of 42,253 residues.

A strong correlation between sequence variability (mutability) and local amino acid flexibility is consistent with a similar pattern noted by Lustig and coworkers with

respect to peptide binding to RNA [35, 36]. The ability of a protein structure to accommodate mutation was indicated by an enhanced flexibility at a particular residue position. These residue positions are found primarily in flexible 3-D features such as loops and possibly in alternative residue contacts.

Disorder in a protein has been measured by many means, such as protease digestion where sites of hypersensitivity indicate disorder, X-ray diffraction where residues missing from electron density maps indicate disorder, and NMR spectroscopy where sharp peaks indicate disorder [50]. Shannon entropy, often referred to as sequence entropy, is also a measure of disorder of a system and has been applied to amino acid sequences to measure the disorder in a protein computationally. Koehl and Levitt quantified sequence entropy with respect to conformational entropy [40].

Sequence entropy measures the sequence variability and residue conservation of amino acids. Gerstein and Altman [39] compared sequence conservation to structure conservation by measuring the entropy of a position relative to the random alignment of sequences. Gerstein-Altman entropy is calculated by subtracting the random entropy that takes into account of the random occurrence of each amino acid from the Shannon entropy for each protein. The aggregate Gerstein-Altman entropy curve, which is calculated by considering the possible effect of random entropy, is found to be comparable to the gap included average entropy and window-averaged entropy curves.

Earlier studies by Koehl and Levitt [40] indicate that entropy derived from sequence information was similar to the entropy derived from structure information for a small set of ten proteins, when a gap was included as a twenty-first term at the considered

sequence position. So, inclusion of a gap in sequence entropy calculations is an appropriate tool in studying correlations between sequence entropy and flexibility of a protein.

The original sequence alignments of the 130 query proteins obtained from the BLASTP search utilized in this research were 3 years old, so a representative set of 20 proteins was selected, and an updated set of sequence alignments was obtained using the current version of BLASTP. Interestingly, the aggregate plot of average sequence entropy versus inverse of the C_α packing density for 20 proteins is observed to contain the same four regions (two major regions and two flanking regions), similar to the aggregate sequence entropy curve observed for 130 proteins. Initial concern about using a complete set of BLASTP alignments attempts to minimize any concerns of bias. A limited number of 130 protein query set and the 20- protein query set exceeded the original 100-alignment limit. The frequency distribution for 20 proteins between frequency of query protein and number of alignments after the correction, reveals eight proteins with a number of alignments greater than 100.

Window average entropy is calculated by averaging the entropy values of closest neighbors for each residue [51]. Earlier studies, done by Galzitskaya and Melnik [52] in predicting protein domain boundaries using sequence information alone, indicate that averaging entropy within a small window size provides additional entropy information. Increasing the window size (i.e. neighbors included in the window for average entropy calculation) smoothes the entropy profile; lowering it increases the resolution of the plot but results in the introduction of more noise. Defining an optimum window size is the

best compromise between a good resolution of the plot and low noise. The objective of this research was to study the effects of different window sizes (2, 3, 4 and 5 neighbors within a window) on the noise level. While doing the single average entropy calculations, it was observed that the window size of three neighbors significantly reduces the standard deviation values. Window average entropy, calculated by using a window size of three was found to be the optimum in terms of resolution and tolerable noise level for the individual protein plots. This approach significantly improved the linear correlation for the individual protein plots as shown in the appendices. This improvement might be because, the variations in the packing of the secondary structure are minimal for three neighbors as compared to 4, 5, 6 neighbors. Variations in the packing of the secondary structure are minimum for two neighbors as compared to three neighbors, but averaging the entropy for two neighbors biases data in one direction.

CHAPTER EIGHT

CONCLUSIONS

Various filters were introduced to observe the affect on the noise level of the aggregate average sequence entropy plots; gap included sequence entropy curve, Gerstein-Altman entropy curve, window average entropy curve, and normalized B-factor curve. Gap-included average entropy (gapped entropy) offers some improvement in terms of linear correlation but not a significant reduction in noise level. The linear correlation coefficient R was determined to be 0.1352 and the corresponding P value was calculated to be less than 0.001 for the gapped entropy plot, whereas the linear correlation coefficient R was determined to be 0.031 and the corresponding P value was calculated to be less than 0.001 for the average entropy plot. Two major regions and two flanking regions are observed in the sequence entropy versus inverse of Cα packing density correlation plot.

Subsequently, the Gerstein-Altman entropy calculated considering the random entropy effect also offers some improvement in linear correlation between sequence entropy and inverse of Cα packing density with no significant reduction in the noise level. The linear correlation coefficient R was determined to be 0.055 and the corresponding P value was calculated to be less than 0.001 for the Gerstein-Altman entropy plot. Window averaging with a window size three significantly improves the linear coefficient values of correlation plots between average entropy and inverse density for most of the 130 query proteins. The average standard deviation for the window average entropy is 0.313, compared to 0.5 for the gapped entropy.

Correlation pattern between sequence entropy and inverse density for a subset of 20 query proteins is similar to that of 130 query proteins. The straight-line fit for the aggregate average sequence entropy versus inverse packing density for the 130 proteins is $y = 17.862x - 0.2768$; the correlation coefficient is 0.9945; $P<0.001$. The straight-line fit for the aggregate entropy plot for the 20 proteins is effectively identical: $y = 17.544x - 0.2781$; correlation coefficient is 0.9553; $P<0.001$. Correlation pattern for normalized B-factor curve plotted between aggregate average sequence entropy and normalized B-factor is different to the correlation pattern observed between average sequence entropy and inverse density, hence is not utilized to further explore the relationship between sequence entropy and B-factor at individual protein level.

SUGGESTED FUTURE STUDIES

Further understanding the protein structure relations with sequence entropy is of great interest and should include:

1. Exploring the reasons for the fluctuations in sequence entropy values in anomalous regions (high density and low density regions).
2. Exploring the usage of PSI BLAST or other means in the sequence alignment search to take care of the redundancies in the sequence alignment sets.
3. Exploring the usage of B-factor information to characterize the flexible portions of a protein.

REFERENCES

1. C. Branden and J. Tooze, Introduction to Protein Structure, (Garland Pub. New York, 1991).

2. C. Sander and R. Schneider, "*Database of homology-derived protein structures and the structural meaning of sequence alignment*," Proteins **9**, 56-68 (1991).

3. Nation Human genome research institute Website (2005) Illustration [Online] Available at www.genome.gov/Pages/Hyperion/DIR/VIP/Glossary/Illustration/protein.shtml (accessed in Feb. 2005).

4. RCSB Protein Data Bank (2005) Information portal [Online] Available at http://www.rcsb.org/pdb/ (accessed in March 2005).

5. Connolly, M (1996) Molecular volume and protein packing [Online] Available at http://www.netsci.org/Science/Compchem/feature14g.html (accessed in March 2005).

6. NCBI website (2004) Glossary [Online] Available at http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html (accessed in May 2007).

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

8. M.O. Dayhoff, W.C. Barker and L.T. Hunt, "*Establishing homologies in protein sequences,"* Meth. Enzymol. **91,** 524-545 (1983).

9. Bevan, D (1997) Molecular modeling of proteins and nucleic acids {Online] Available at http://www.biochem.vt.edu/modeling/homology.html#Introduction (accessed in Feb. 2005).

10. H. Kono and J. G. Saven, "*Statistical theory for protein combinatorial libraries*," J. Mol. Biol. **306**, 607-628 (2001).

11. J. Tsai, R. Taylor, C. Chothia and M. Gerstein, "*The packing density in proteins: standard radii and volumes,*" J. Mol. Biol. **290**, 253-266 (1999).

12. J. Kyte and R. Doolittle, "*A simple method for displaying the hydropathic character of a protein,*" J. Mol. Biol. **157,** 105-132 (1982).

13. B. Rost, C. Sander and R. Schneider, "*Protein fold recognition by prediction based threading*," J. Mol. Biol. **270**, 471-480 (1997).

14. D. Jones, "*Protein structure prediction in the post genomic era*," Curr. Opinion Struct. Biol. **10**, 371-379 (2000).

15. D. Baker and A. Salij, "*Protein structure prediction and structural genomics*," Science **294**, 93-96 (2001).

16. Swiss-Prot protein knowledge base (2005) Release notes [Online] Available at http://us.expasy.org/sprot/relnotes/spwrnew.html (accessed in March 2005).

17. RCSB Protein Data Bank (2005) An information portal [Online] Available at http://www.rcsb.org/pdb/ (accessed in March 2007).

18. M.O. Dayhoff, W.C. Barker and L.T. Hunt, "*Nucleic acid sequence database v: completely sequences genomes,*" DNA. **4,** 275-280 (1983).

19. S. B. Needlman and C. D. Wunsch, "*A general method applicable to the search for similarities in the amino acid sequence of two proteins*," J. Mol. Biol. **48**, 443-53 (1970).

20. S. Henikoff and J.G. Henikoff, "*Position based sequence weights*," J. Mol. Biol. **243**, 574-578 (1994).

21. NCBI website (2004) Alignment scoring [Online] Available at http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Alignment_Scores2.html accessed in March 2005.

22. M. A. Marti-Renom, A. C. Fiser and A. Sali, "*Comparative protein structure modeling of genes and genomes*," Ann. Rev. Biophy. 29, 291-325, (2000).

23. S.F. Altschul, T.L. Madden, A.A. Scaffer and D.J. Lipman, *"Gapped BLAST and PSI-BLAST : A new generation of protein database search programs,"* Nucl. Acids Res. **25,** 3389-3402 (1997).

24. D. Fischer and D. Eisenberg, "*Assigning folds to the proteins encoded by the genome of mycoplasma genitalium*," Proct. Natl. Acad. Sci. **94**, 11929-11934 (1997).

25. E.A. Gross and G.R. Li, "*Prediction of structural and functional relation ships of repeat 1 of human interphotoreceptor retinoid binding protein*," Mol. Vision. **6**, 30-39 (2000).

26. S. Miyazawa and R.L. Jernigan, "*A new substitution matrix for protein sequence searches based on contact frequencies in protein structures*," Proct. Engr., **6**, 267-278 (1993).

27. D. M. Engelman, T.A. Steitz and Goldman, "*Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins*," Annu. Rev. Biophys. Chem. **115**, 321-353 (1986).

28. W. Kauzmann, "*Some factors in the interpretation of protein denaturation*," Adv. Prot. Chem. **14**, 1-63 (1959).

29. T. P. Hopp and K. R. Woods, "*Prediction of protein antigenic determinants from amino acid sequences*," Proct. Nat. Acad. Sci., **78**, 3824-3828 (1981).

30. M. Levitt, "*A simplified representation of protein conformations for rapid simulation protein folding*," J. Mol. Biol., **104**, 59-107 (1976).

31. K. A. Sharp, A.Nicholls, R. Friedman and B. Honig, "*Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models*," Biochemistry, **30**, 9686-9697 (1991).

32. J. L. Fauchere and V. Pliska, "*Hydrophobic parameters of aminoacid chains from the partitioning of N-acetyl amino acid amides*," Eur. J. Med. Chem., **18**, 369-375 (1983).

33. H. Liao, W.Yeh, D. Chiang, R.L. Jernigan and B. Lustig, " *Protein sequence entropy is closely related to packing density and hydrophobicity*," Protein Engineering Design and Selection, **18**(2), 59-64 (2005).

34. H. Kono and J. G. Saven, "*Statistical theory for protein combinatorial libraries*," J. Mol. Biol. **306**, 607-628 (2001).

35. B. Lustig, I. Bahar and R. Jernigan, "*RNA bulge entropies in the unbound stae correlate with peptide binding strengths for HIV-1 and BIV TAR RNA because of improved conformational access*," Nucleic Acids Res. **26**, 5212-5217 (1998).

36. M. Hsieh, E. D. Collins, T. Blomquist and B. Lustig, "*Flexibility of BIV TAR –Tat: models of peptide binding*," J. Biomol. Struct. & Dyn. **20**, 243-251 (2002).

37. S. M. Larson, J. L. England and J. R. Pande, "*Thoroughly sampling sequence space: Large scale protein design of structural ensembles*," Protein Sci. **11**, 2804-2813 (2002).

38. W. S. J. Valdar and J. M. Thornton, "*Protein-protein interfaces: analysis of amino acid conservation in homodimers*," Proteins: Structure, Function, and Genetics. **42**, 108-124 (2001).

39. M Gerstein and R. B. Altman, "*Average core structures and variability measures for protein families:* ," J. Mol. Biol. **48**, 227-241, (1995).

40. P. Koehl and M. Levitt, "*Protein topology and stability define the space of allowed sequences*," Proct. Natl. Acad. Sci., **99**, 1280-1285 (2002).

41. F. Richards and W. Lim, "*An analysis of packing in the protein folding problem*," *Q. Rev. Biophys.* **26**, 423–498 (1994).

42. G. Makhatadze and P. Privalov, "*Energetics of protein structure*," Adv. Protein Chem. **47**, 307–425 (1995).

43. Bahar, R.A. Ali and B. Erman, "*Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential*," Folding and Design, **2**, 173-181, (1997).

44. H. Liao, "*Flexibility and sequence variability in proteins*," Thesis (2005).

45. W.Yeh, "*Detailed analysis of protein sequence entropy, hydrophobicity, and flexibility*," Thesis (2005).

46. P. Bevington and D. K. Robinson, Data Reduction and Error Analysis for the Physical Sciences, 3rd ed. (McGraw-Hill Pub. New York, 2003), pp. 102-108.

47. Protein models website (2007) Model of 1crc [Online] Available at http://www.pqs.ebi.as.uk/ (accessed in October 2007).

48. Z. Yuan, T.L. Bailey and R.D. Teasdale, "*Prediction of protein B-factor profiles*," Proteins, **58**, 905-912, (2005).

49. Z. Yuan, J. Zhao and Z.X. Wang, "*Flexibility analysis of enzyme active sites by crystallographic temperature factors*," Protein Engineering, **16**, 109-114, (2003).

50. P. Romero, X. Li and C.J. Brown, *Sequence complexity of a disordered protein*," Proteins, **42**, 38-48, (2001).

51. Average entropy (2006) Window averaging [Online] Available at http://www.doe-mbi.ucla.edu/~luki/serbeta/help.php?f=entropy_win (accessed in December, 2006).

52. O.V. Galzitskaya and B.S. Melnvik, "*Prediction of protein domain boundaries from sequence alone*," Protein Science, **12**, 696-701, (2003).

APPENDIX A- Individual Correlation Plots for 130 Proteins.

Four different correlation plots are shown for each of the130 proteins

Top-left: Entropy versus inverse density

Top-right: Gapped (i.e. gap included) entropy versus inverse density

Bottom-left: Gerstein-Altman entropy versus inverse density

Bottom-right: Window average entropy versus inverse density

Error bars corresponding to the standard deviation as calculated from the data are shown for each scatter plot. These are derived from the averaging of all sequence entropies within a density bin. The overall region for the plots corresponds to the packing density values of 1 to 30 C_α atoms, and major region I corresponds to higher packing densities of 12 to 25 C_α atoms.

■ – denotes major region 1 data points, and ▲- denotes overall region data points for the correlation plots of all the proteins.

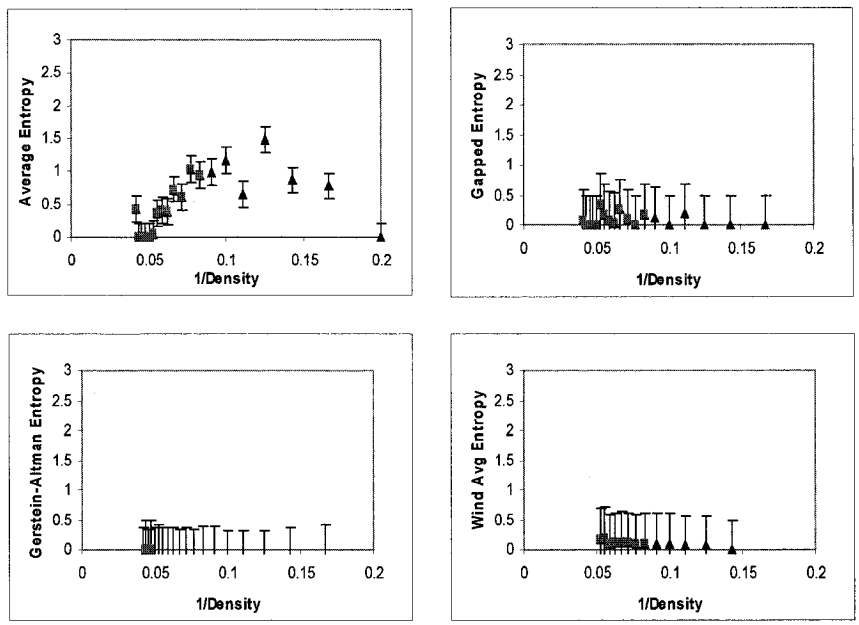

Figure A1. Various correlation plots for protein1A1I

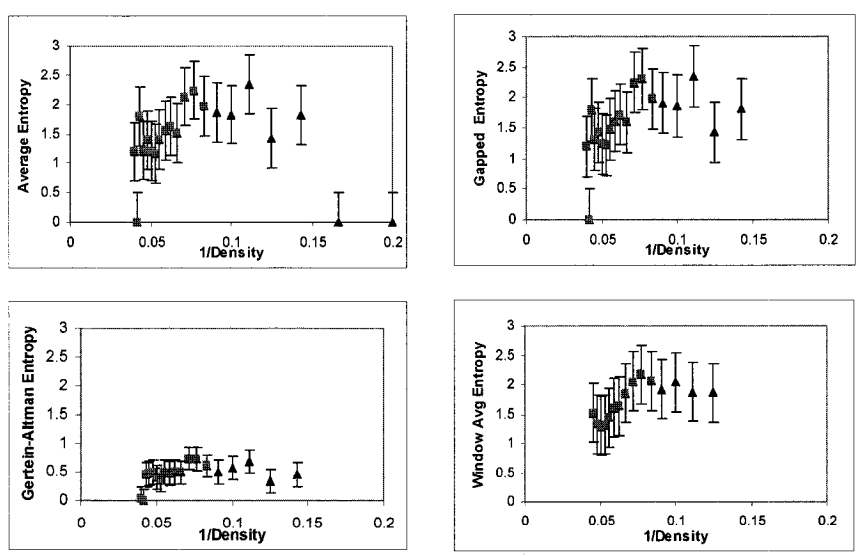

Figure A2. Various correlation plots for protein 1A1S

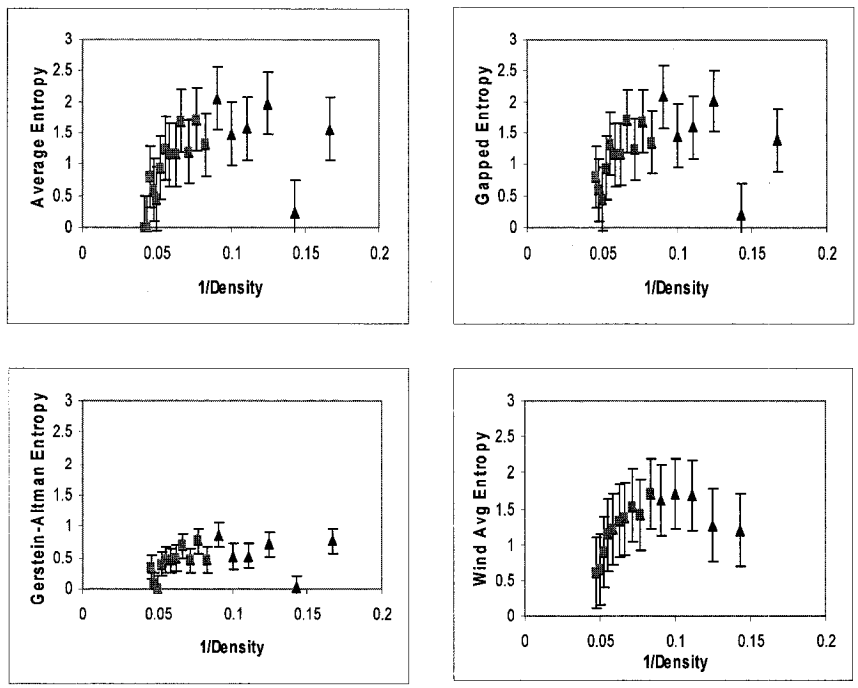

Figure A3. Various correlation plots for protein 1A3C

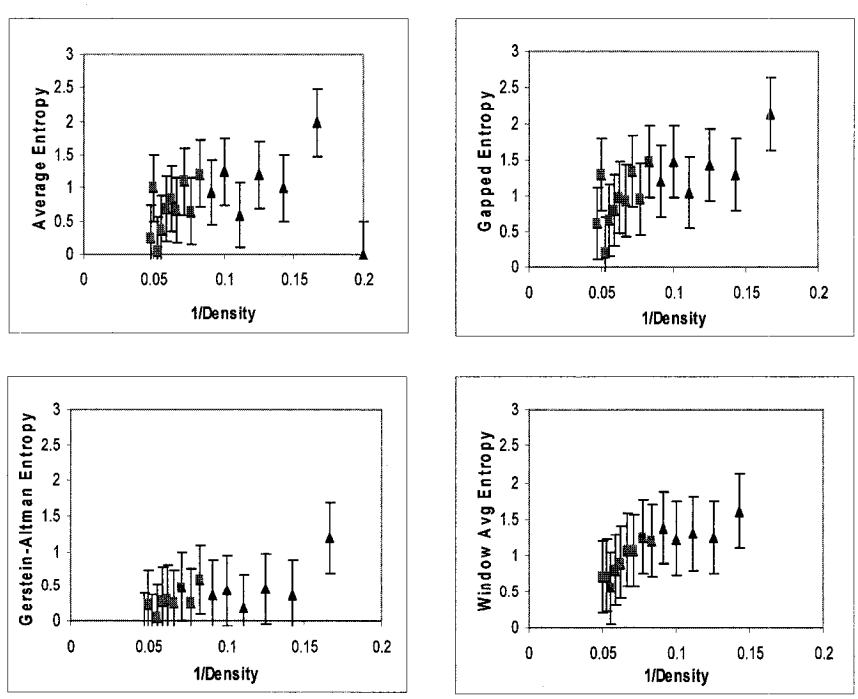

Figure A4. Various correlation plots for protein 1A3S

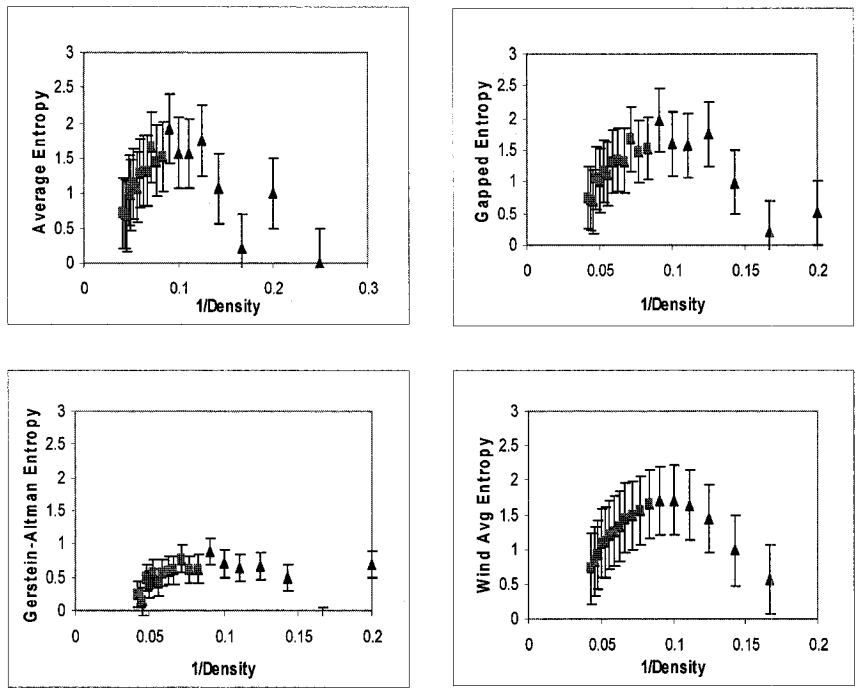

Figure A5. Various correlation plots for protein 1A5Z

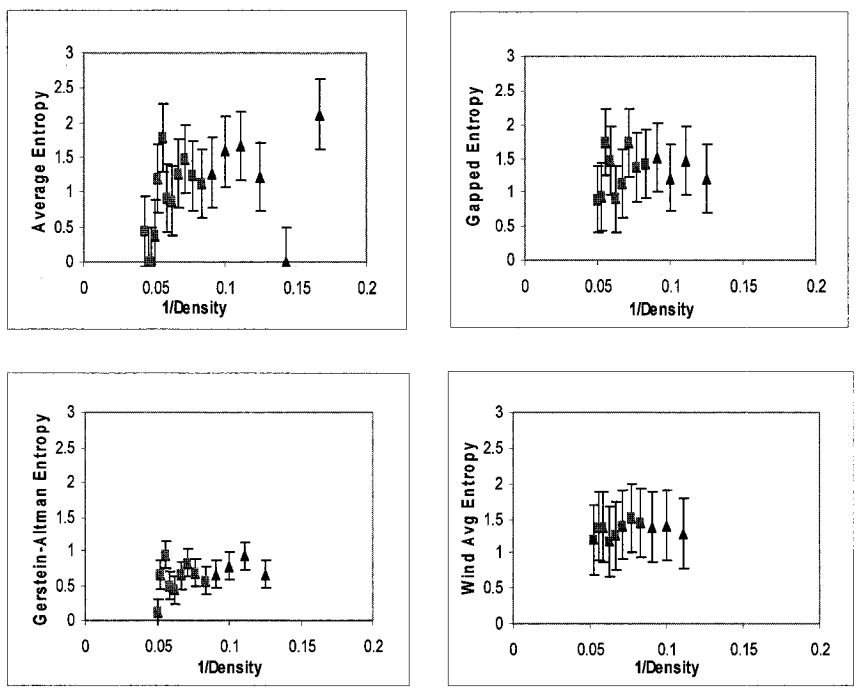

Figure A6. Various correlation plots for protein 1A6F

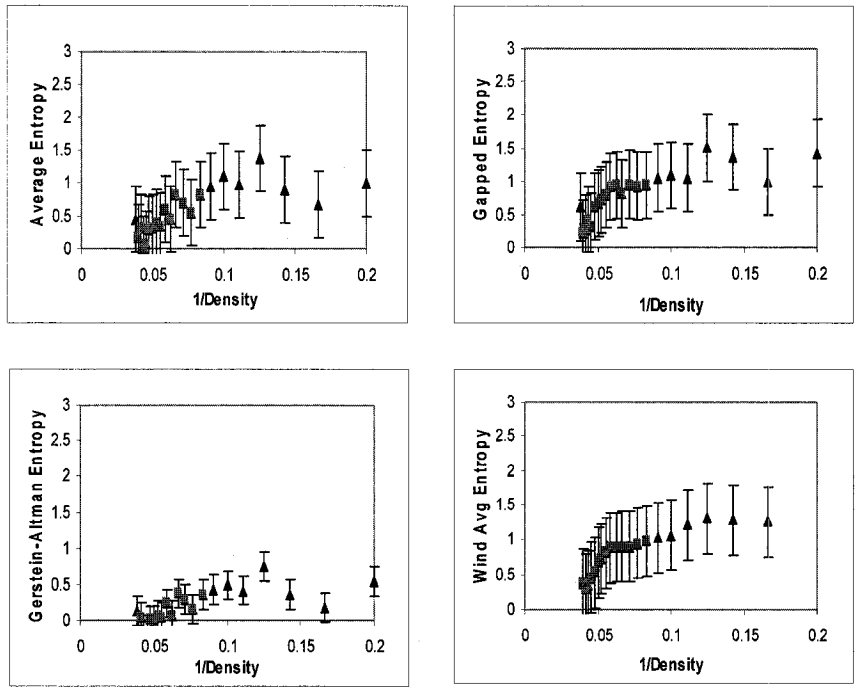

Figure A7. Various correlation plots for protein 1A6Q

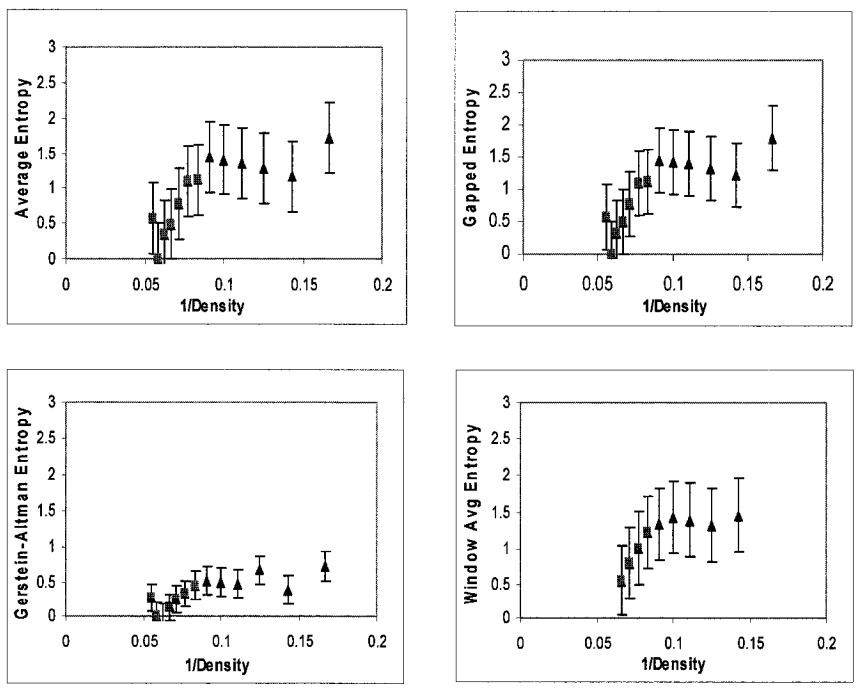

Figure A8. Various correlation plots for protein 1A32

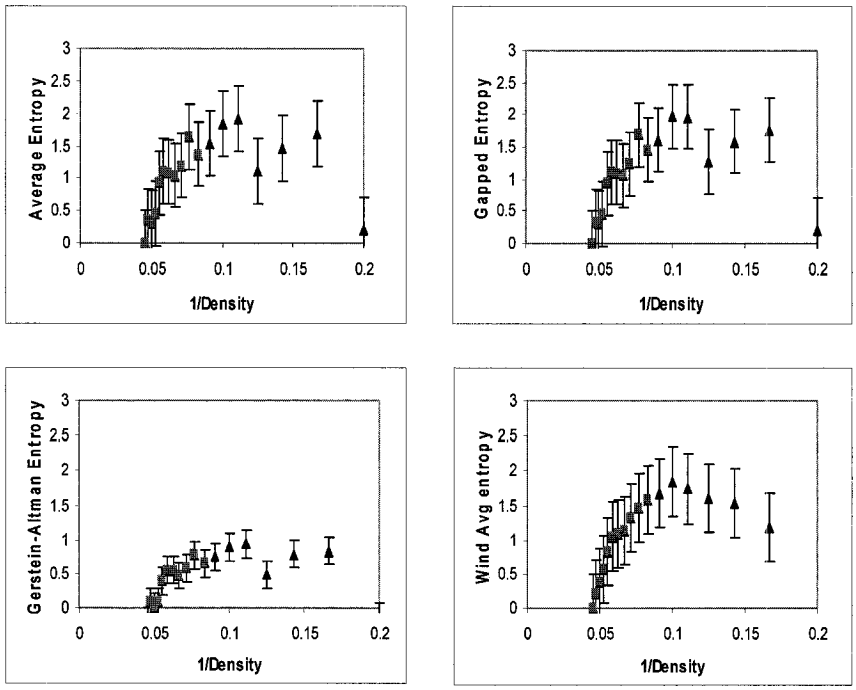

Figure A9. Various correlation plots for protein 1A48

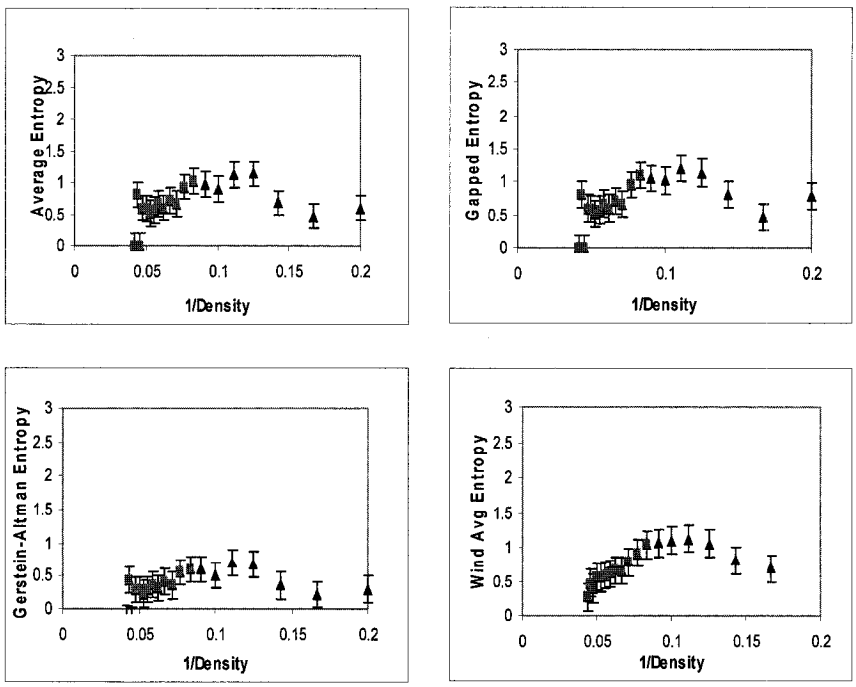

Figure A10. Various correlation plots for protein 1A59

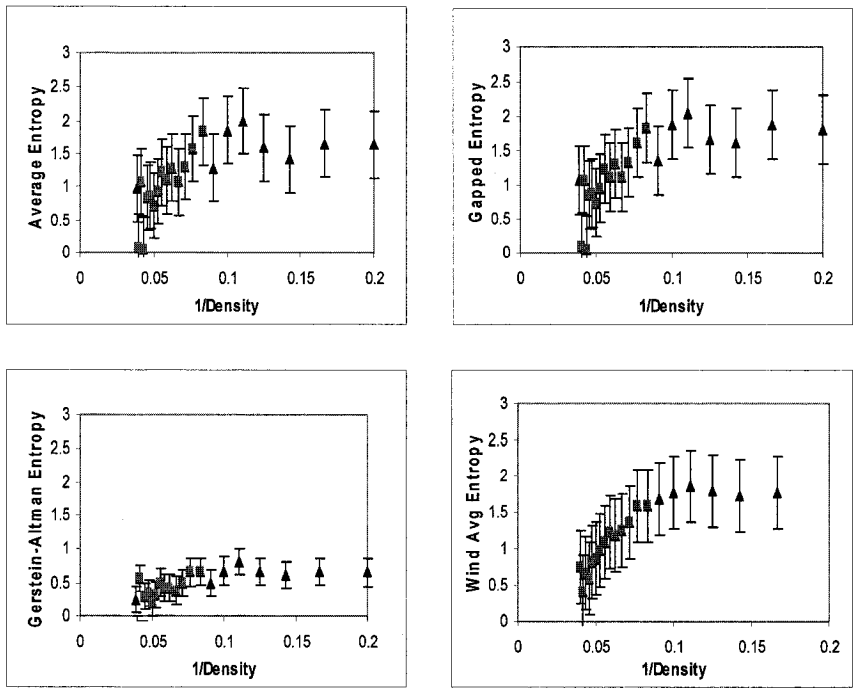

Figure A11. Various correlation plots for protein 1AAT

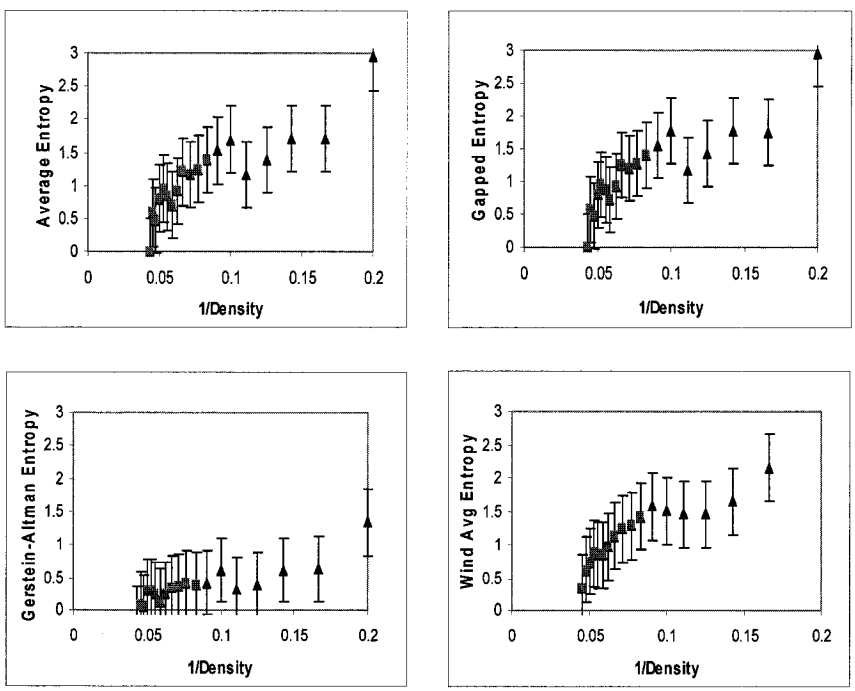

Figure A12. Various correlation plots for protein 1AB4

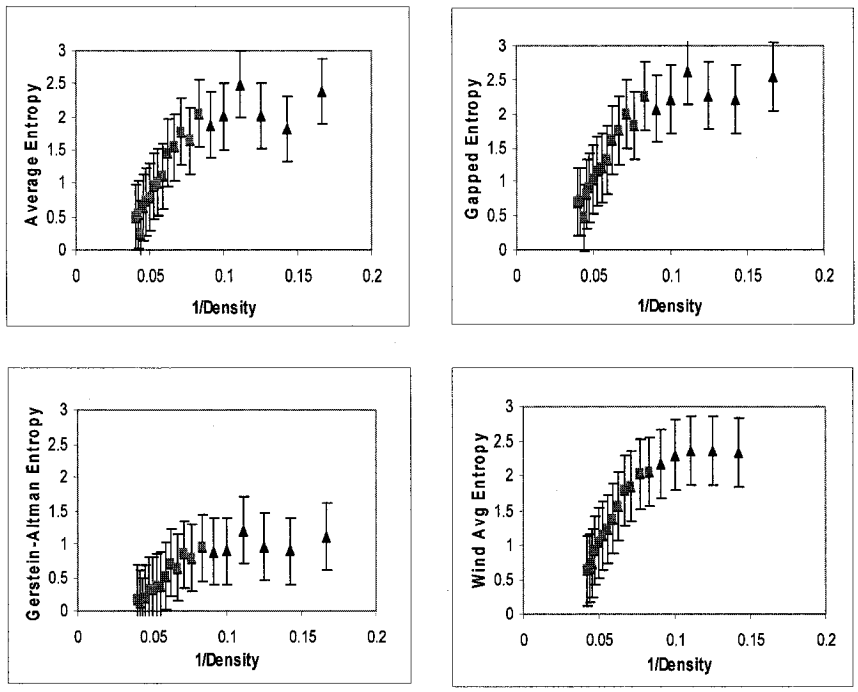

Figure A13. Various correlation plots for protein 1ACB

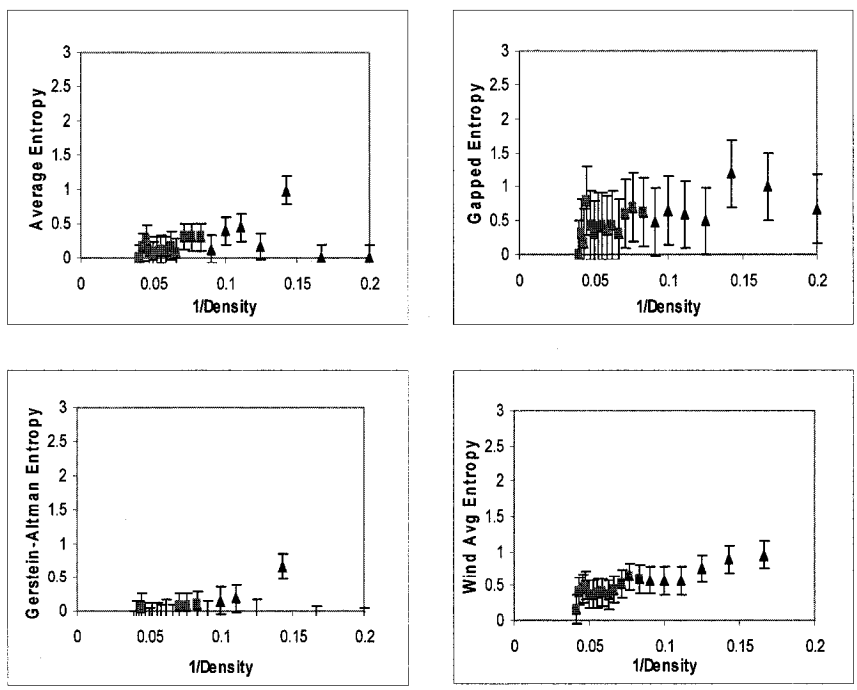

Figure A14. Various correlation plots for protein 1ADD

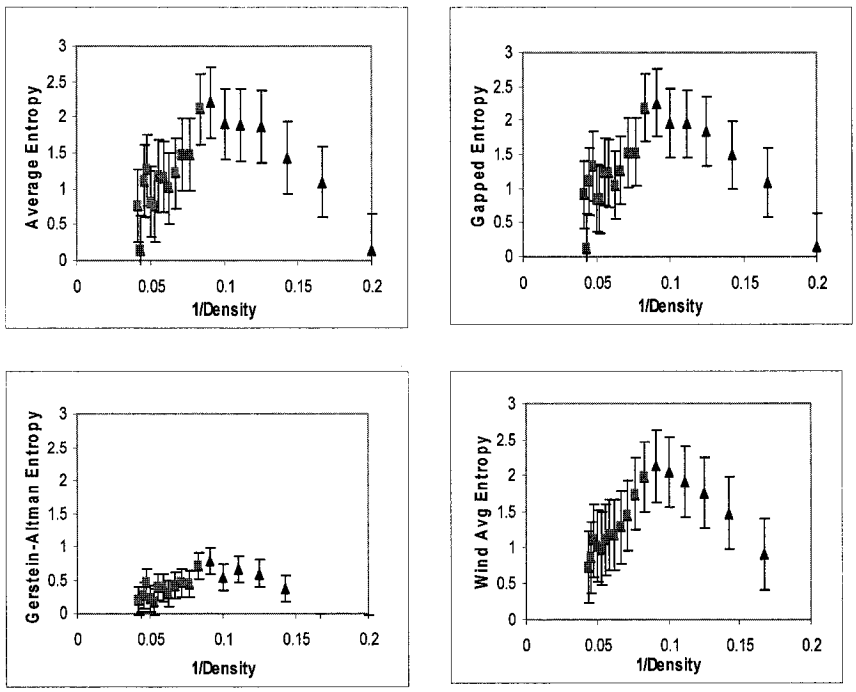

Figure A15. Various correlation plots for protein 1ADI

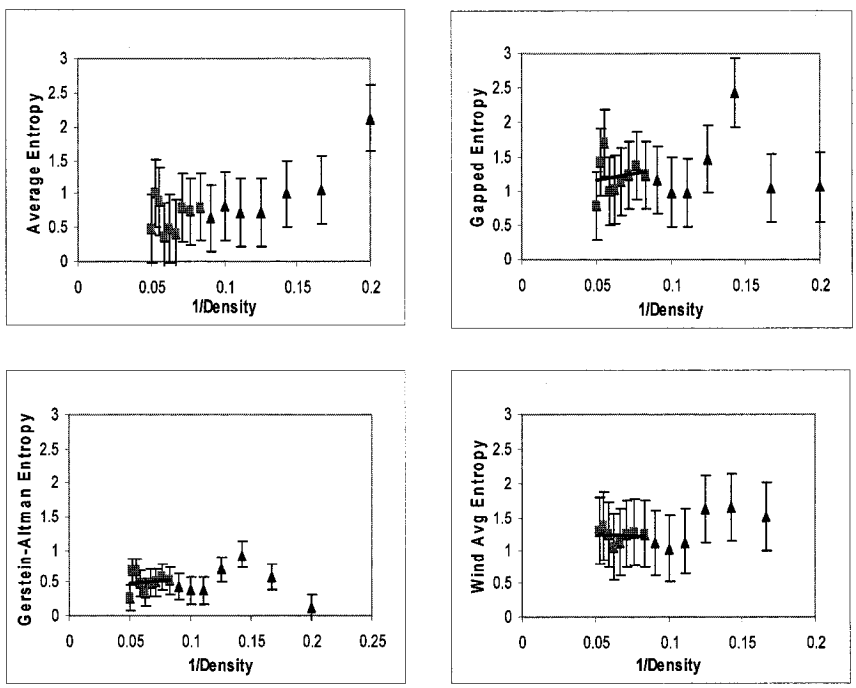

Figure A16. Various correlation plots for protein 1AE4

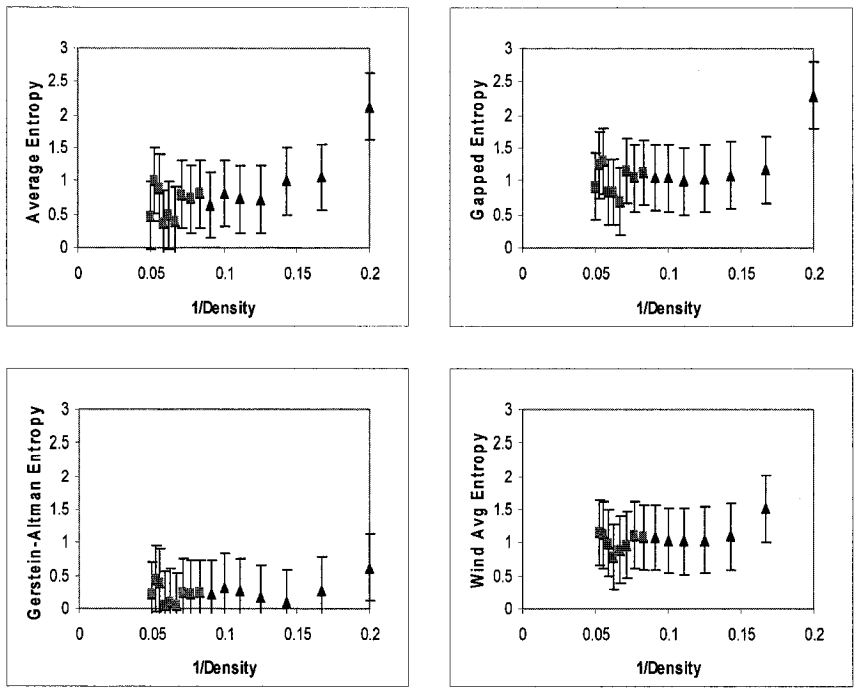

Figure A17. Various correlation plots for protein 1AF3

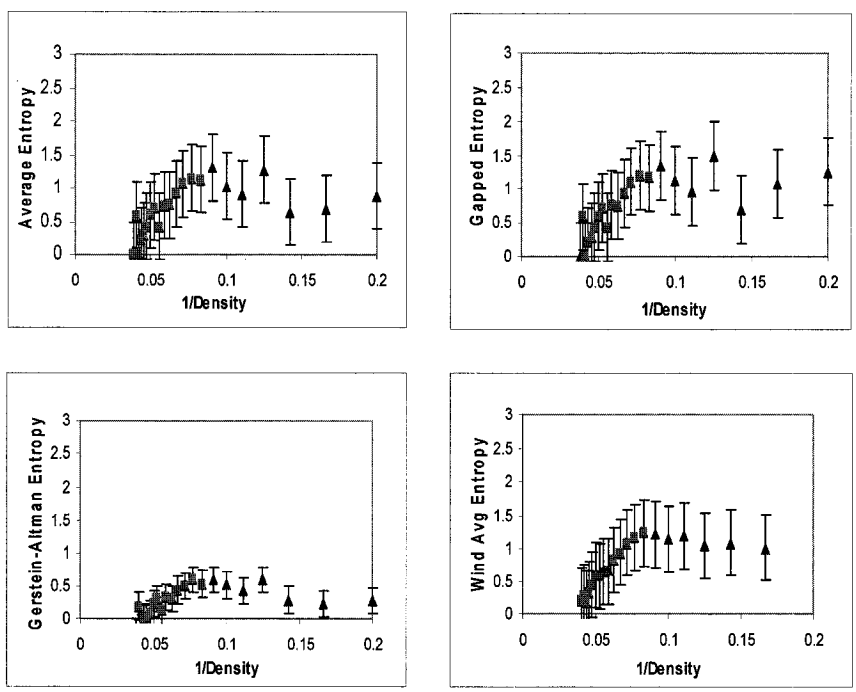

Figure A18. Various correlation plots for protein 1AGM

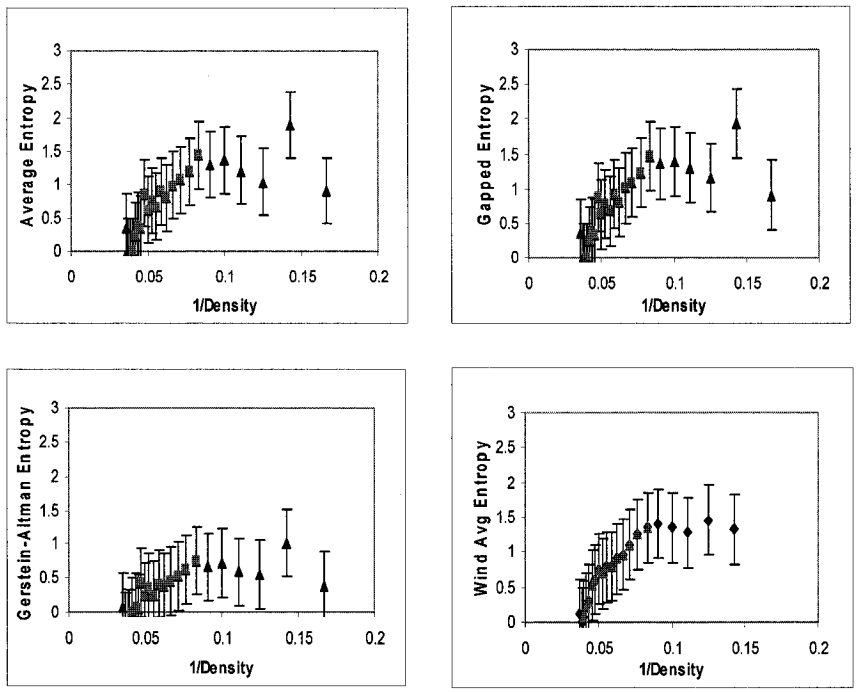

Figure A19. Various correlation plots for protein 1AGX

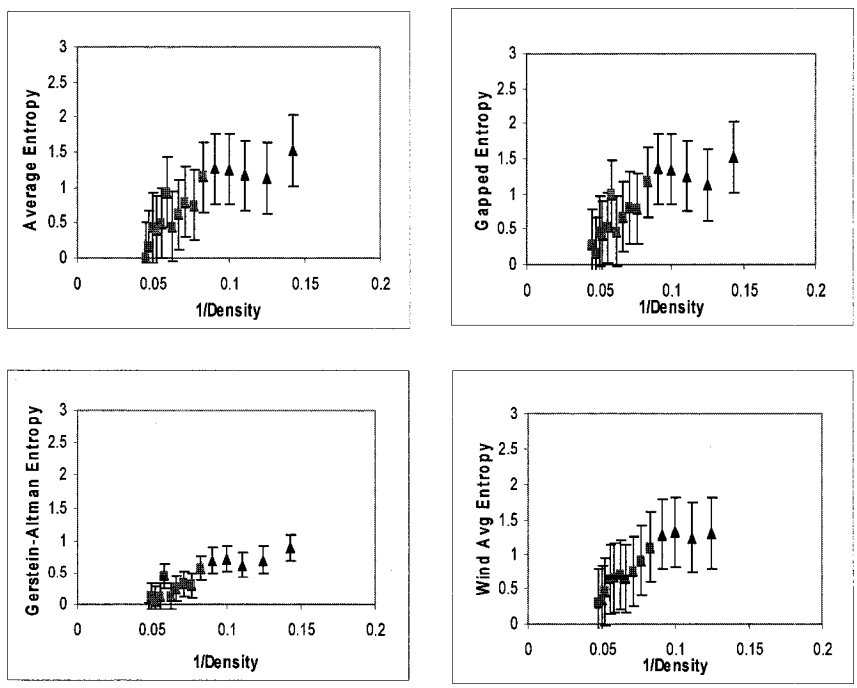

Figure A20. Various correlation plots for protein 1AHA

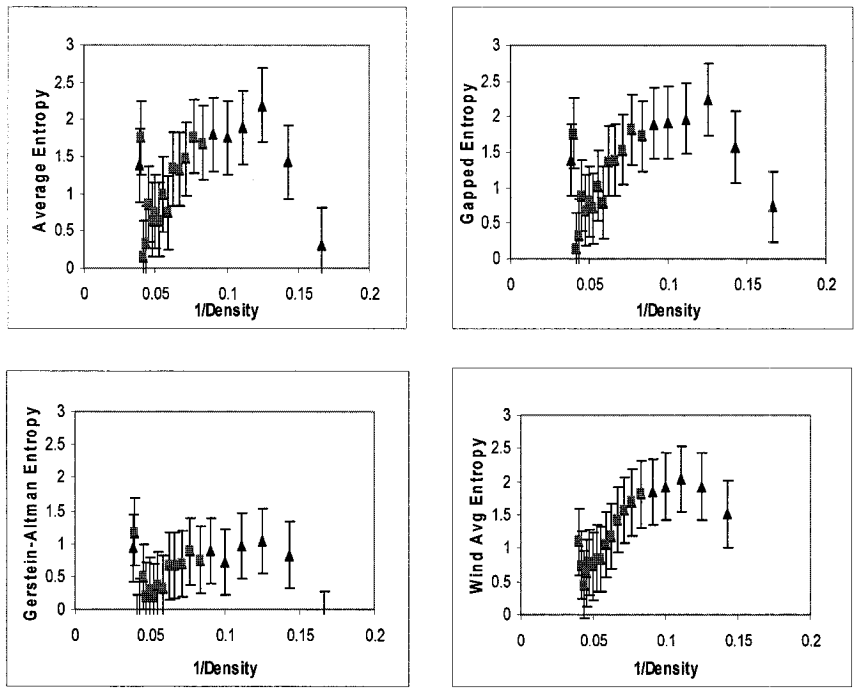

Figure A21. Various correlation plots for protein 1AHN

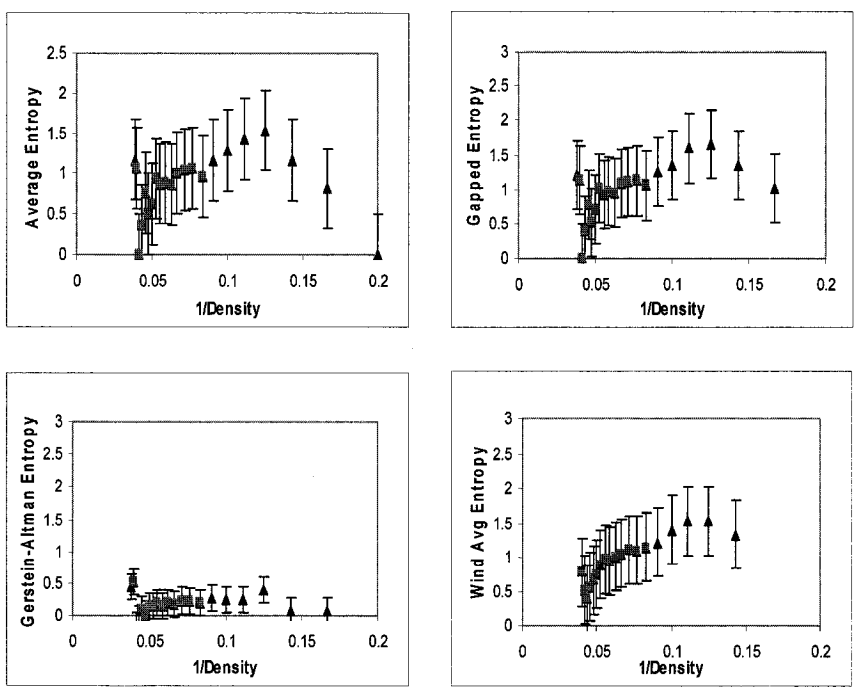

Figure A22. Various correlation plots for protein 1AI2

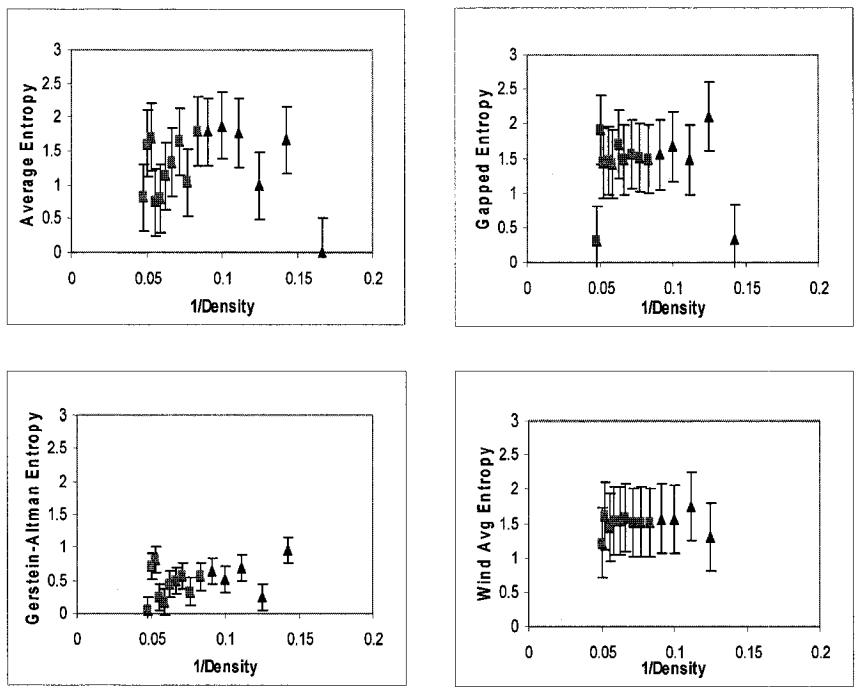

Figure A23. Various correlation plots for protein 1AK2

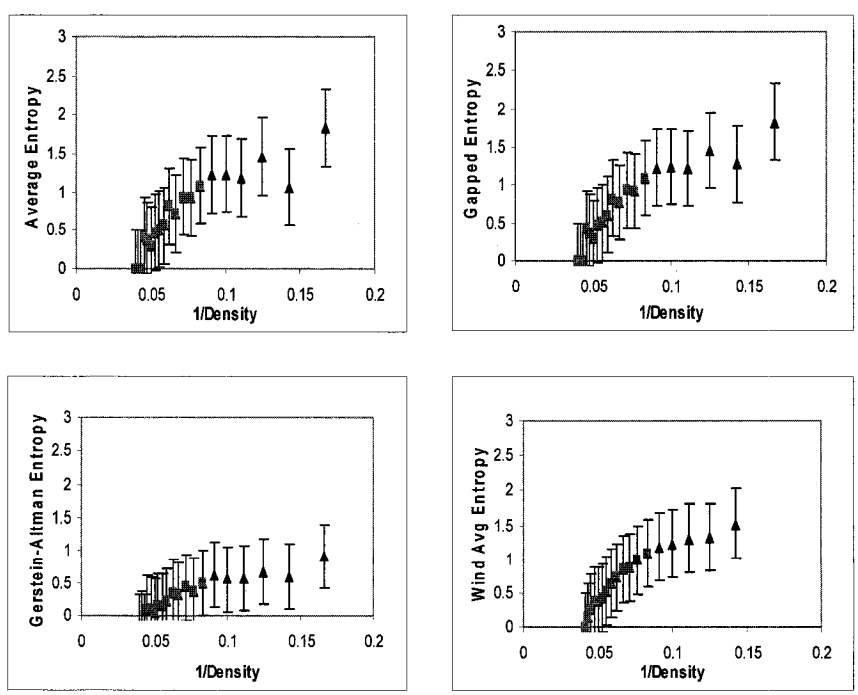

Figure A24. Various correlation plots for protein 1AKO

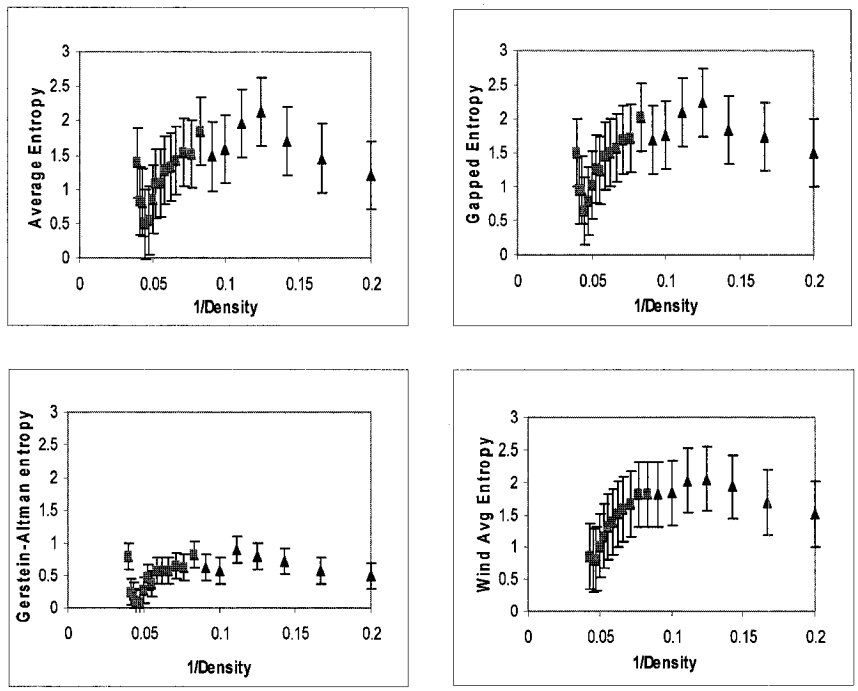

Figure A25. Various correlation plots for protein 1AL8

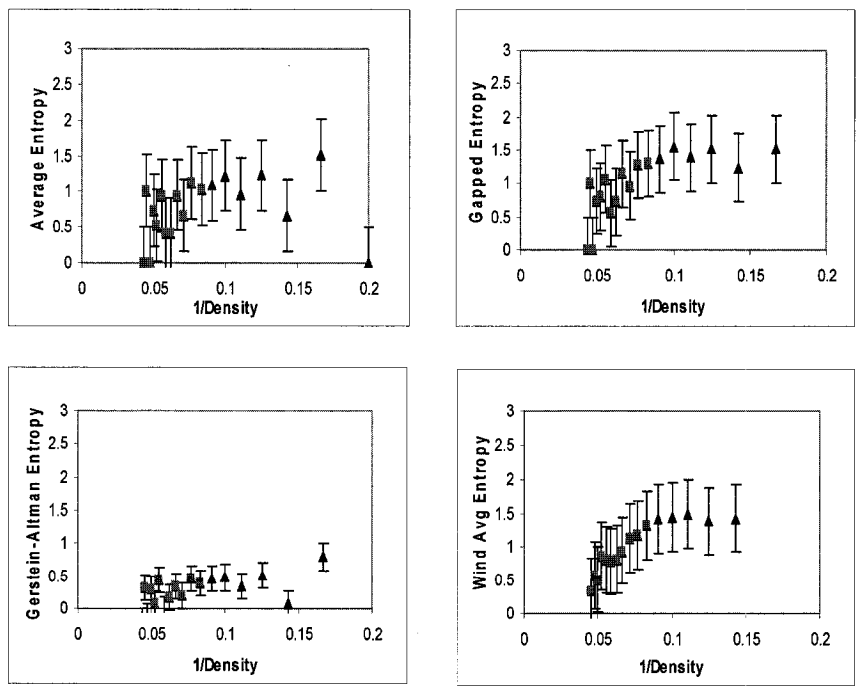

Figure A26. Various correlation plots for protein 1ALC

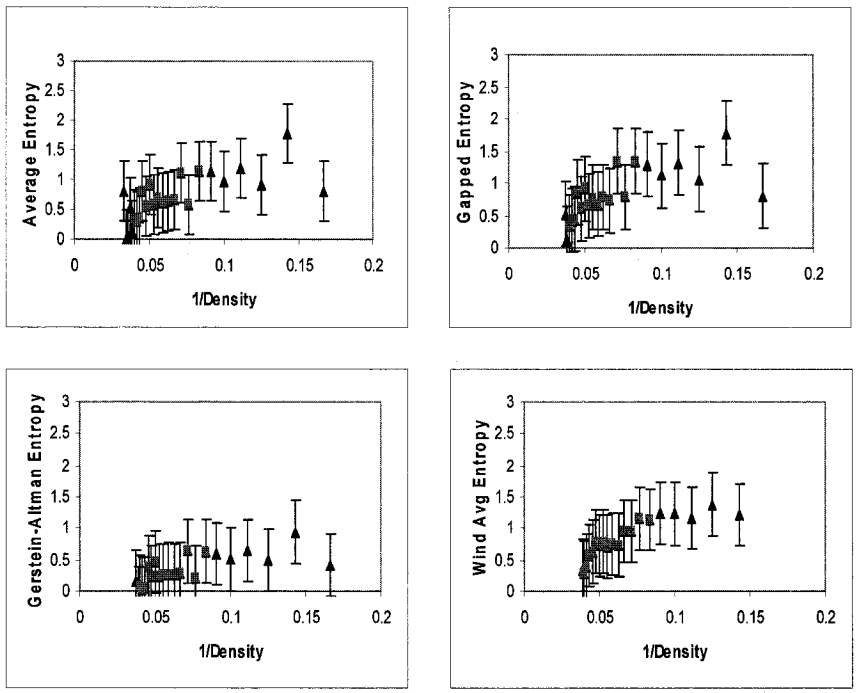

Figure A27. Various correlation plots for protein 1ALN

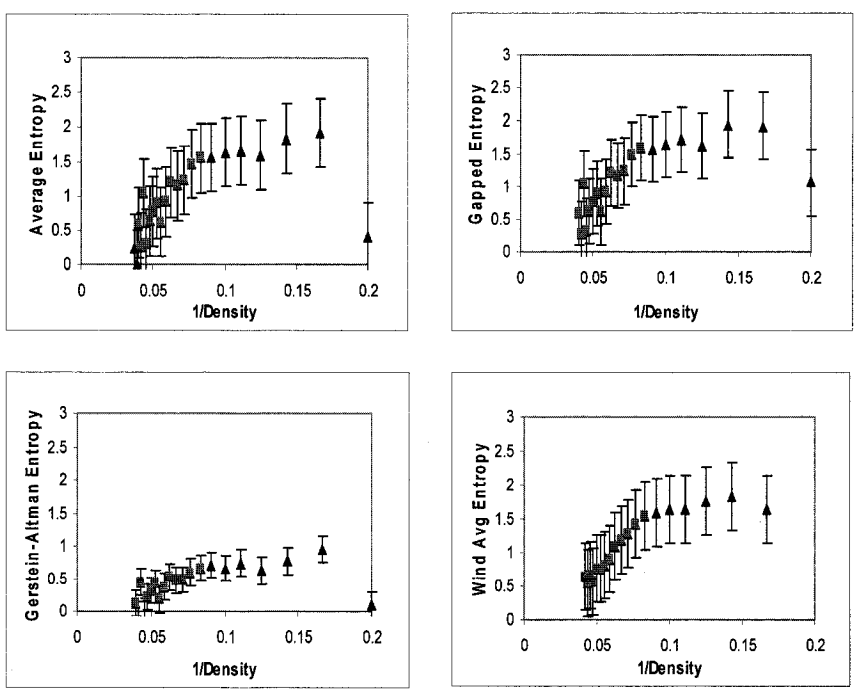

Figure A28. Various correlation plots for protein 1AMN

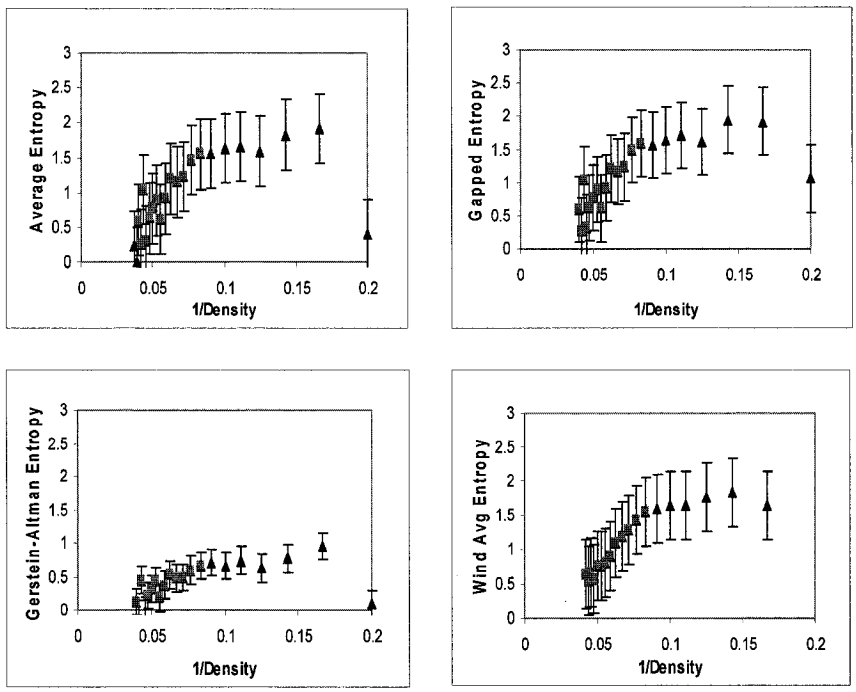

Figure A29. Various correlation plots for protein 1AMP

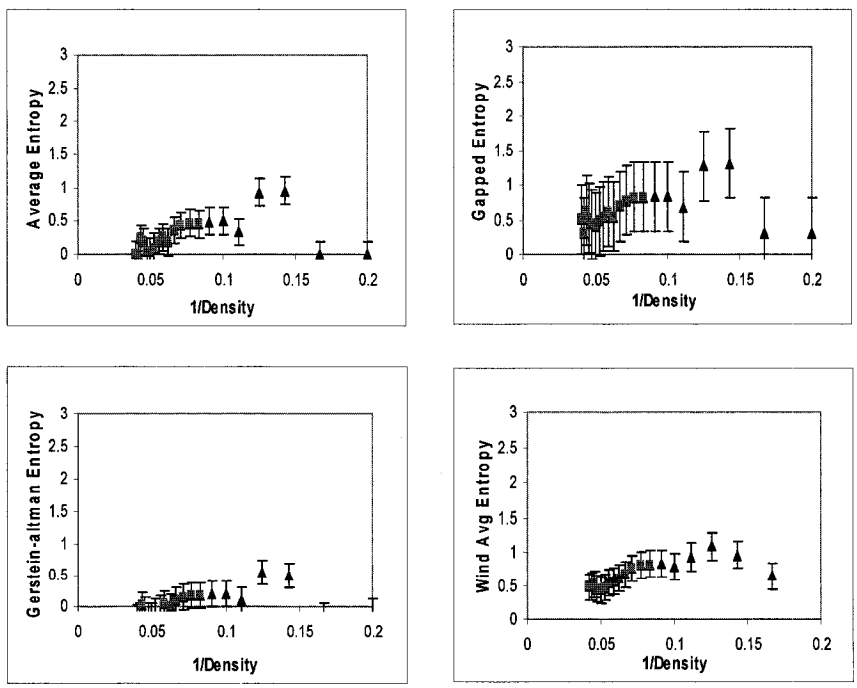

Figure A30. Various correlation plots for protein 1AN9

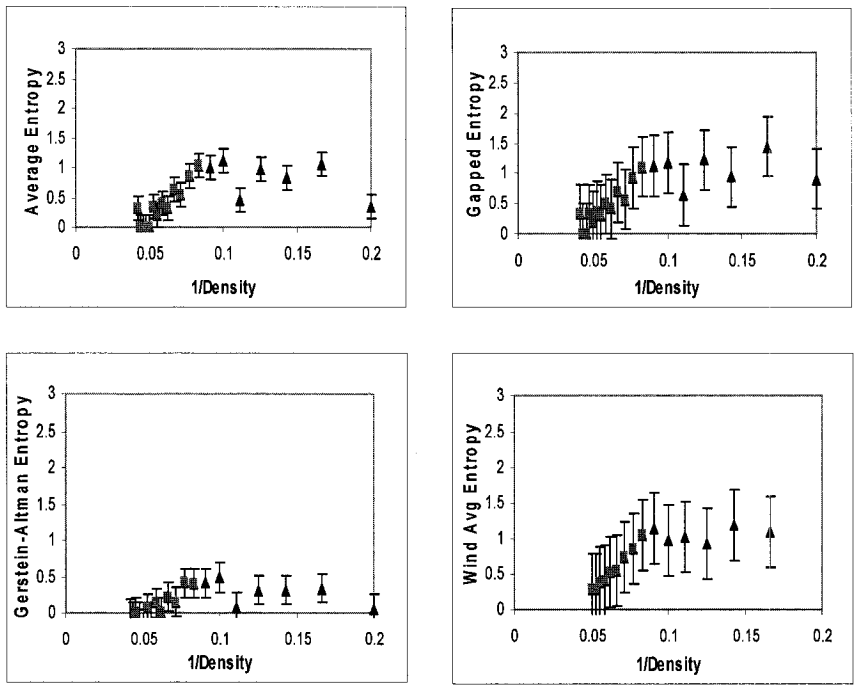

Figure A31. Various correlation plots for protein 1ANG

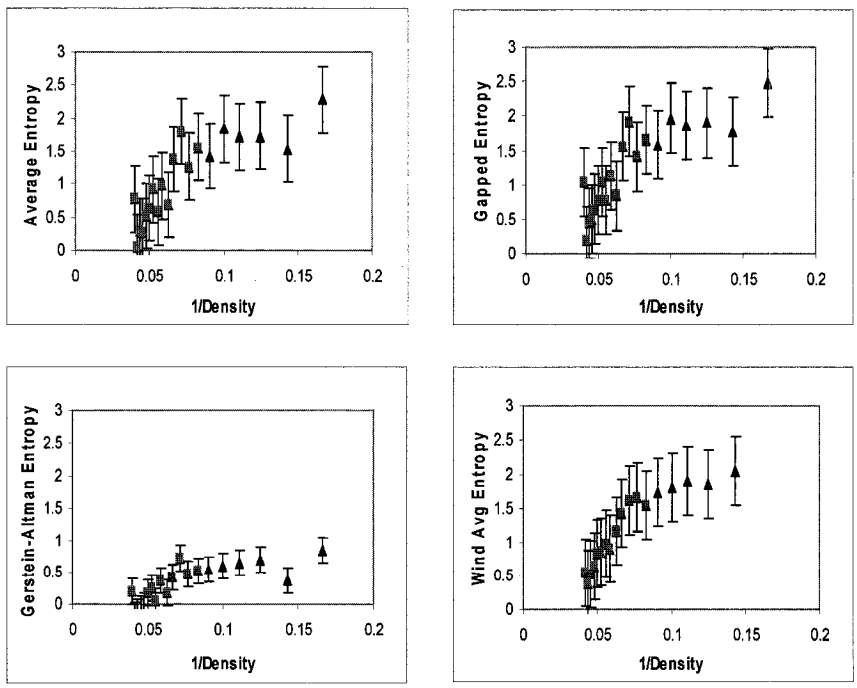

Figure A32. Various correlation plots for protein 1AO5

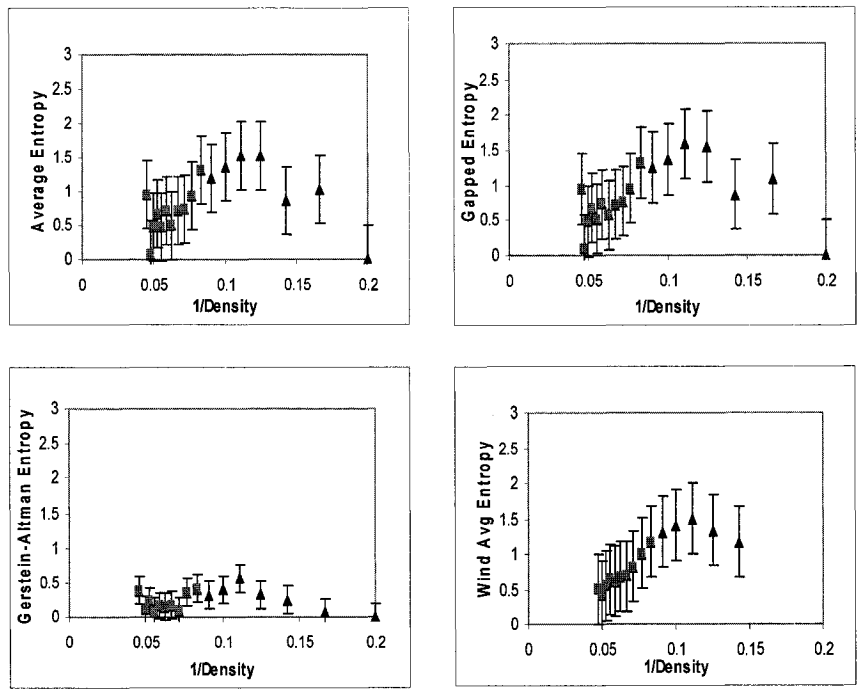

Figure A33. Various correlation plots for protein 1AOB

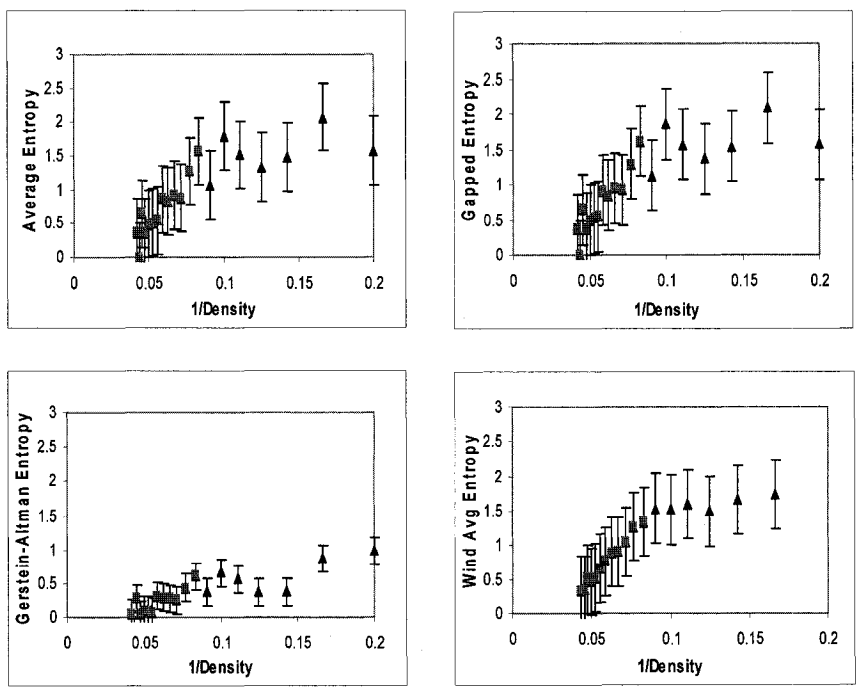

Figure A34. Various correlation plots for protein 1AQH

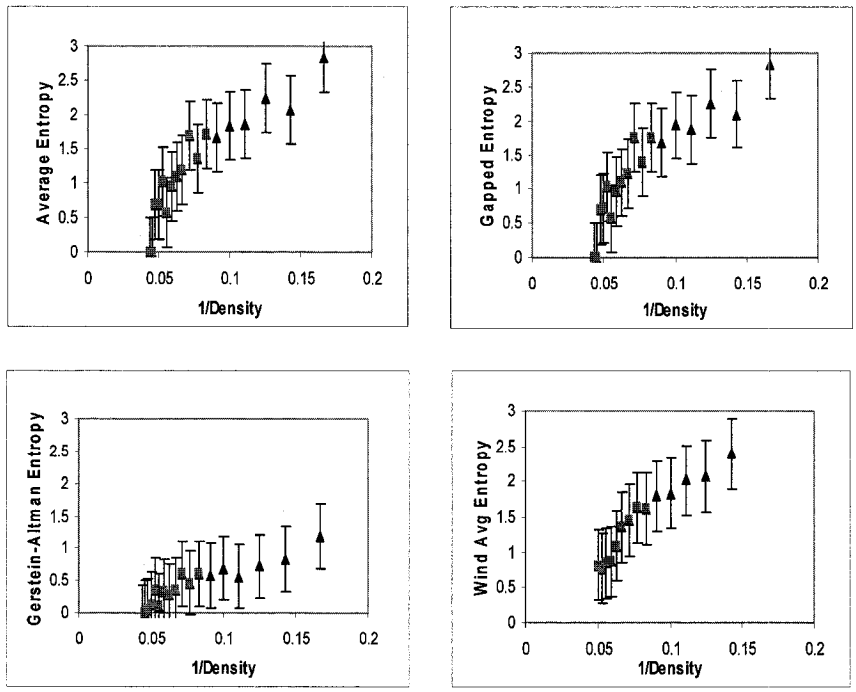

Figure A35. Various correlation plots for protein 1AQO

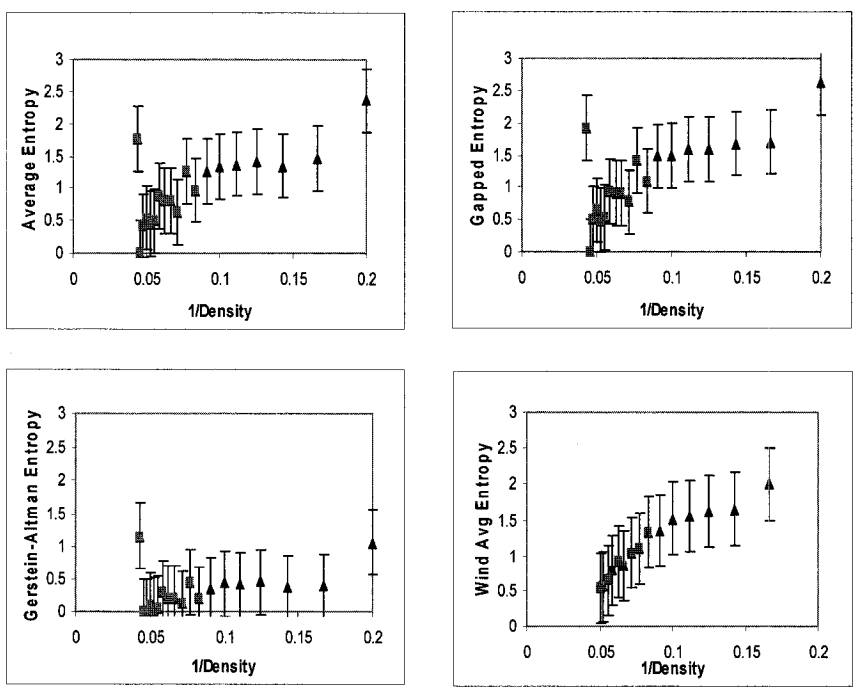

Figure A36. Various correlation plots for protein 1ATP

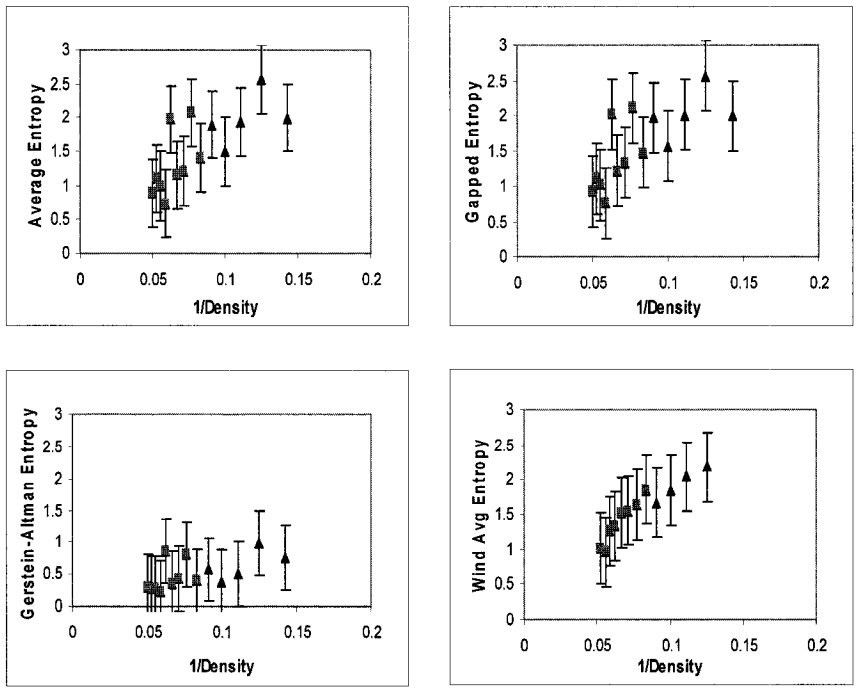

Figure A37. Various correlation plots for protein 1AV5

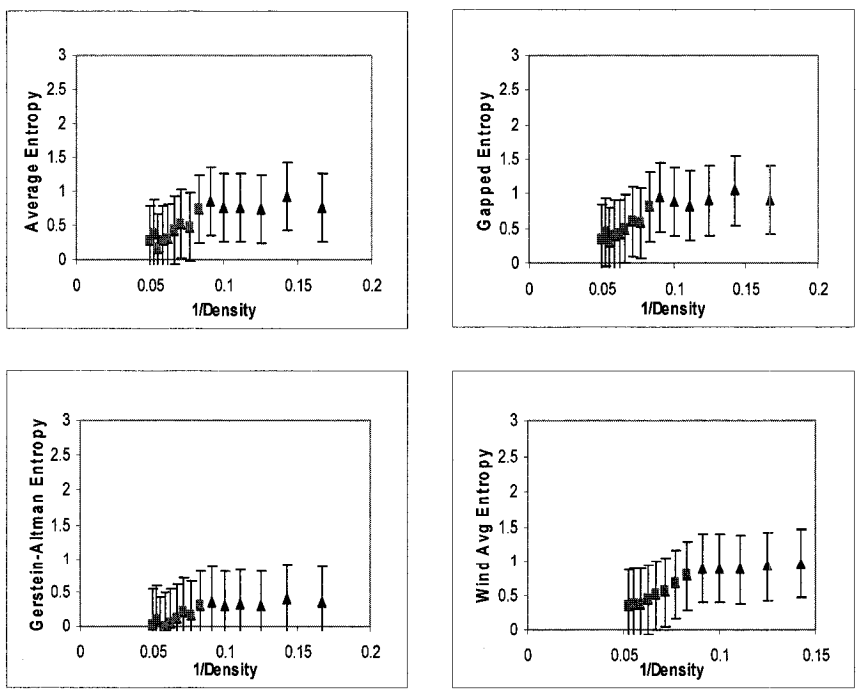

Figure A38. Various correlation plots for protein 1AV6

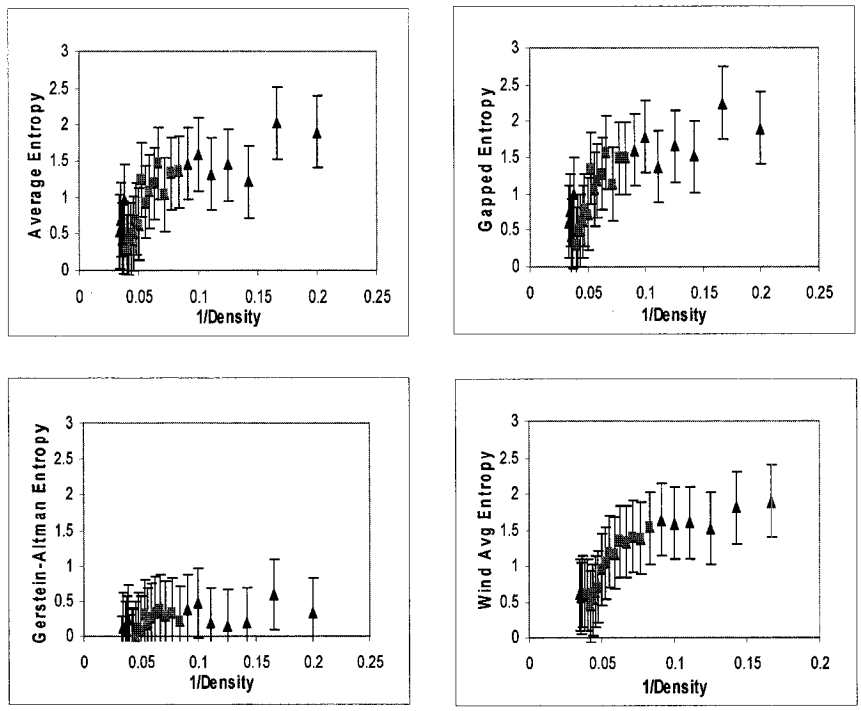

Figure A39. Various correlation plots for protein 1AV7

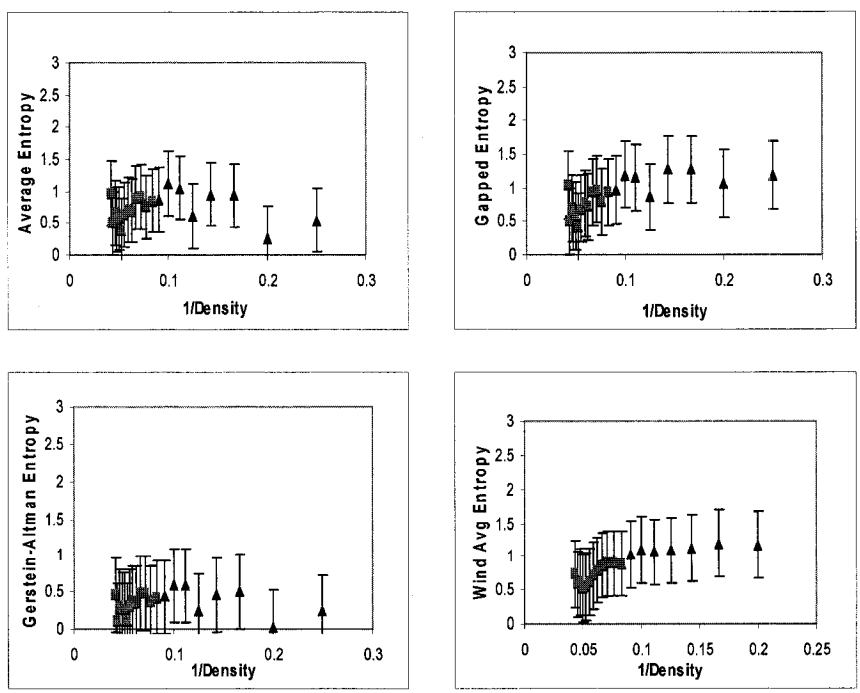

Figure A40. Various correlation plots for protein 1AW5

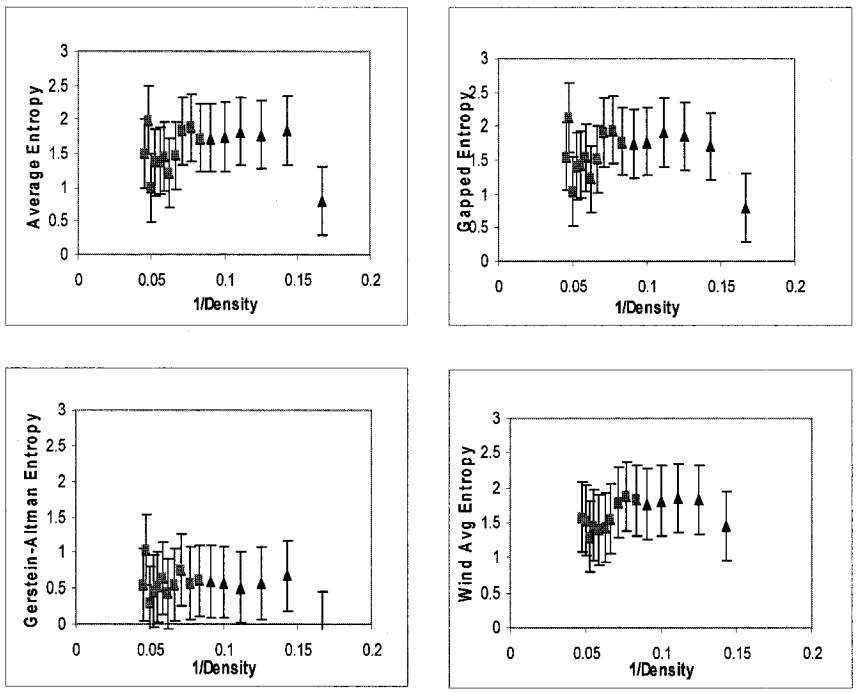

Figure A41. Various correlation plots for protein 1AW9

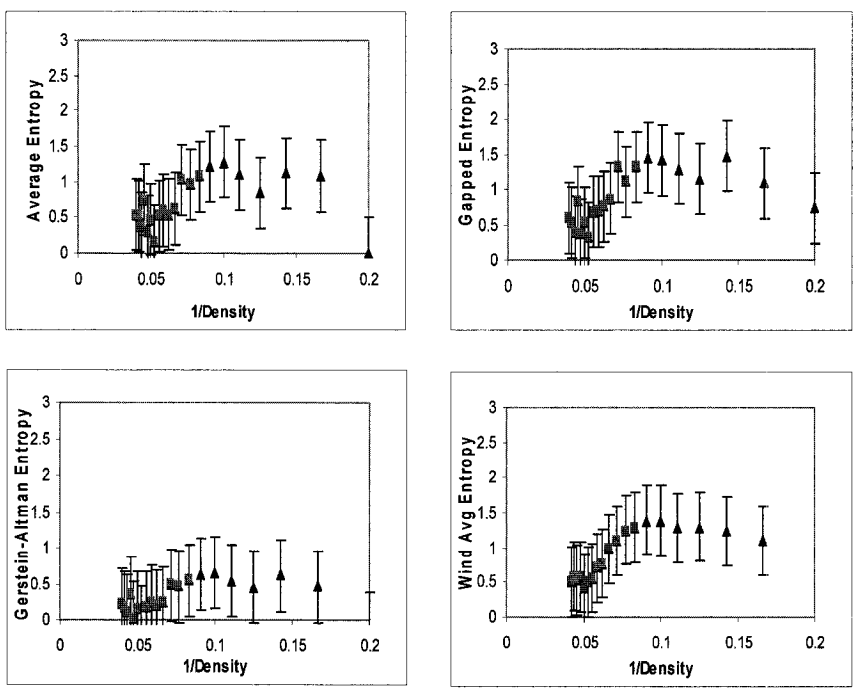

Figure A42. Various correlation plots for protein 1AYE

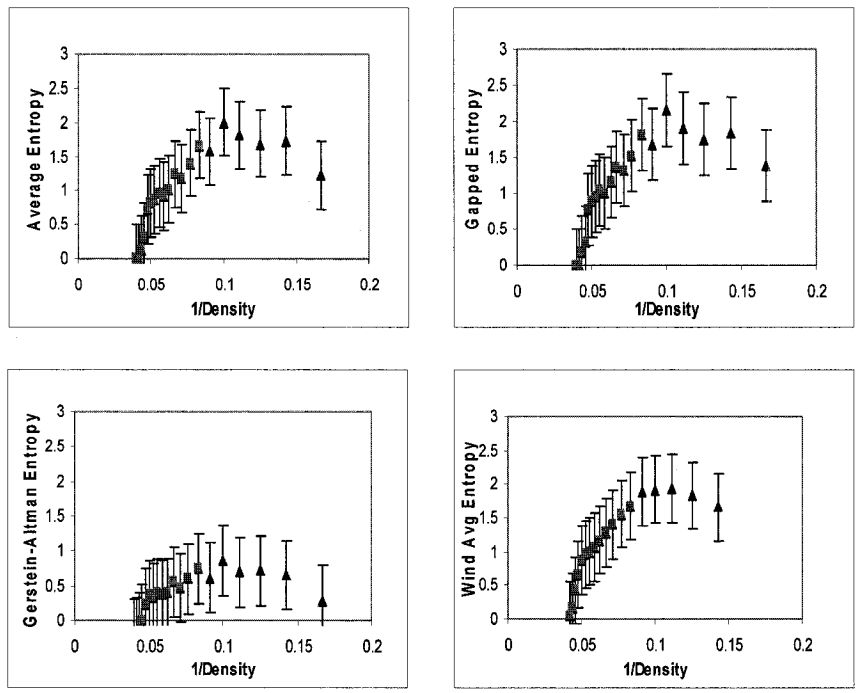

Figure A43. Various correlation plots for protein 1AYL

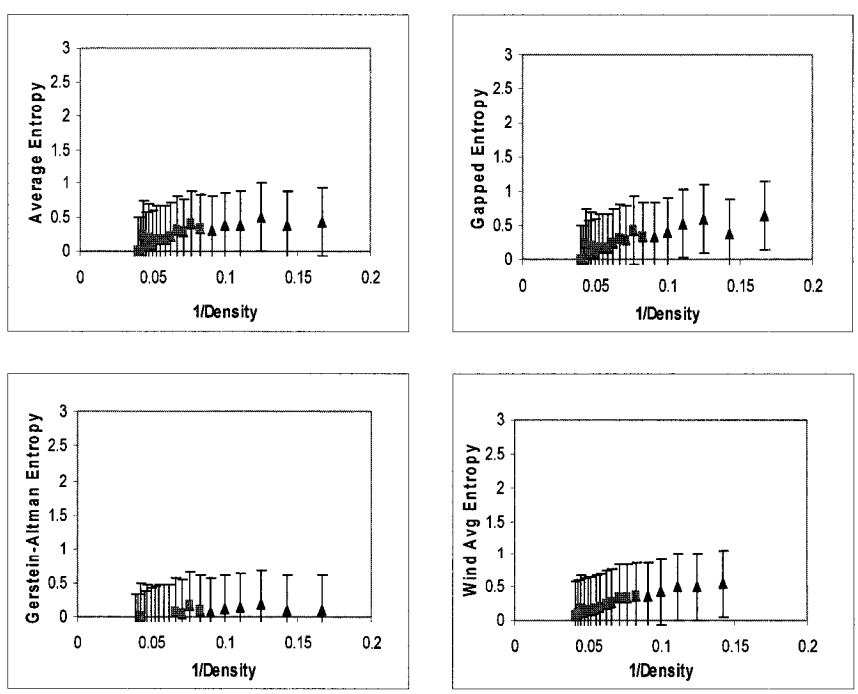

Figure A44. Various correlation plots for protein 1AYX

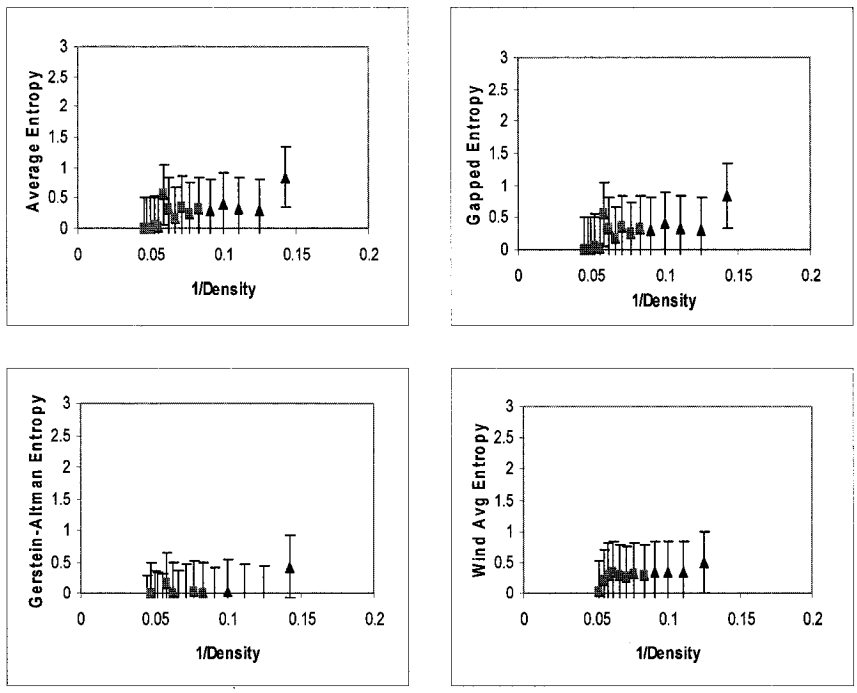

Figure A45. Various correlation plots for protein 1AZI

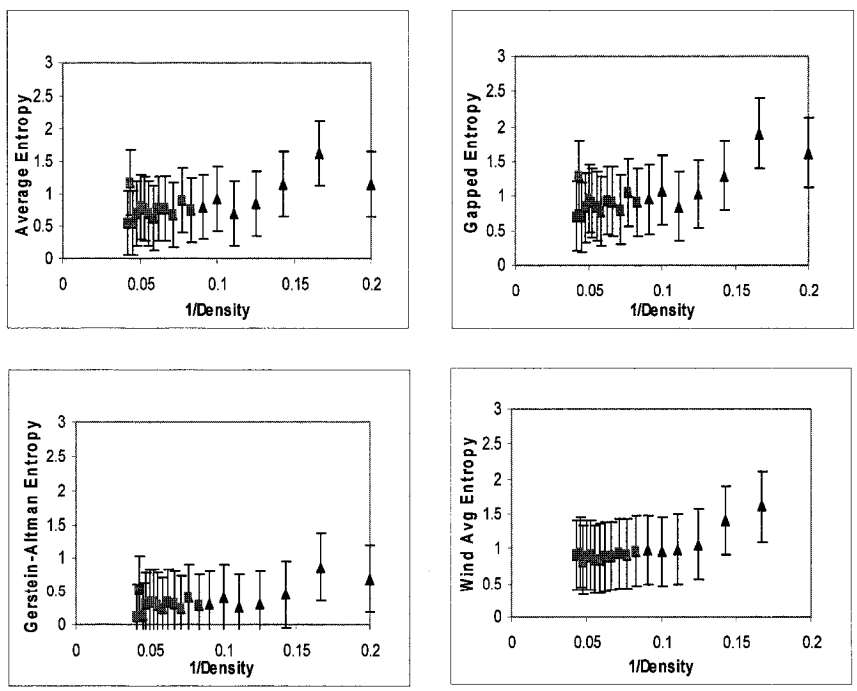

Figure A46. Various correlation plots for protein 1BA3

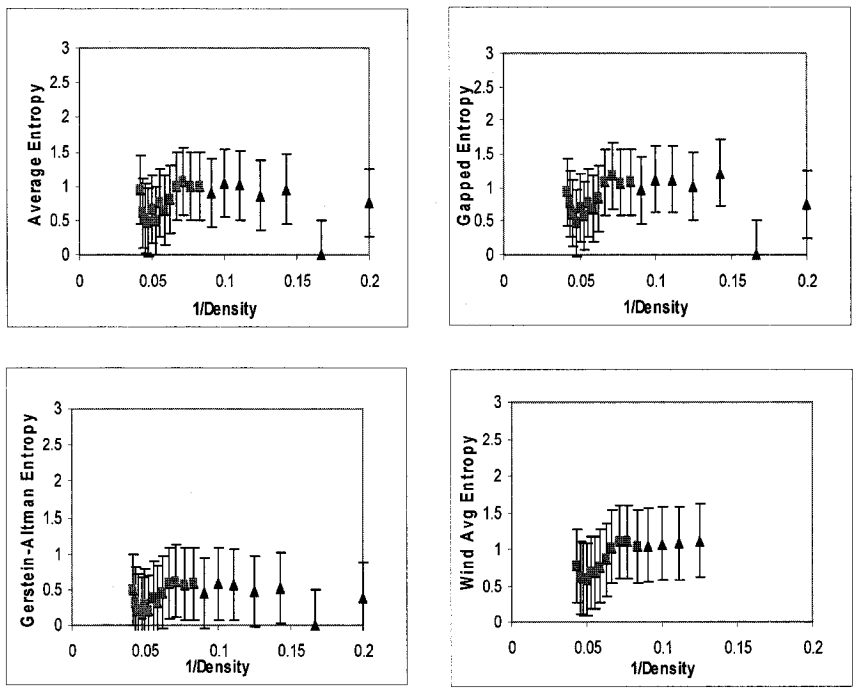

Figure A47. Various correlation plots for protein 1BC2

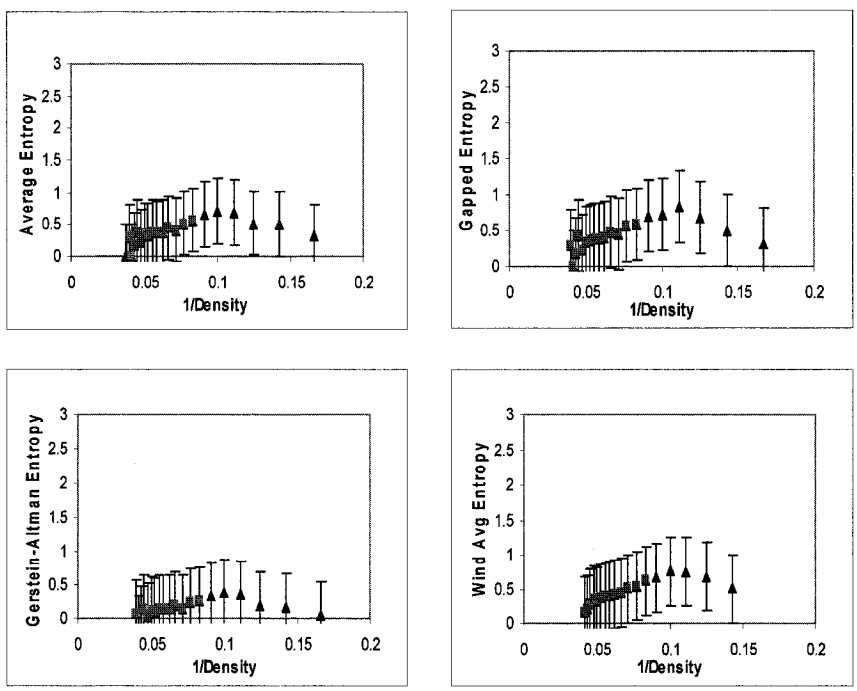

Figure A48. Various correlation plots for protein 1BF2

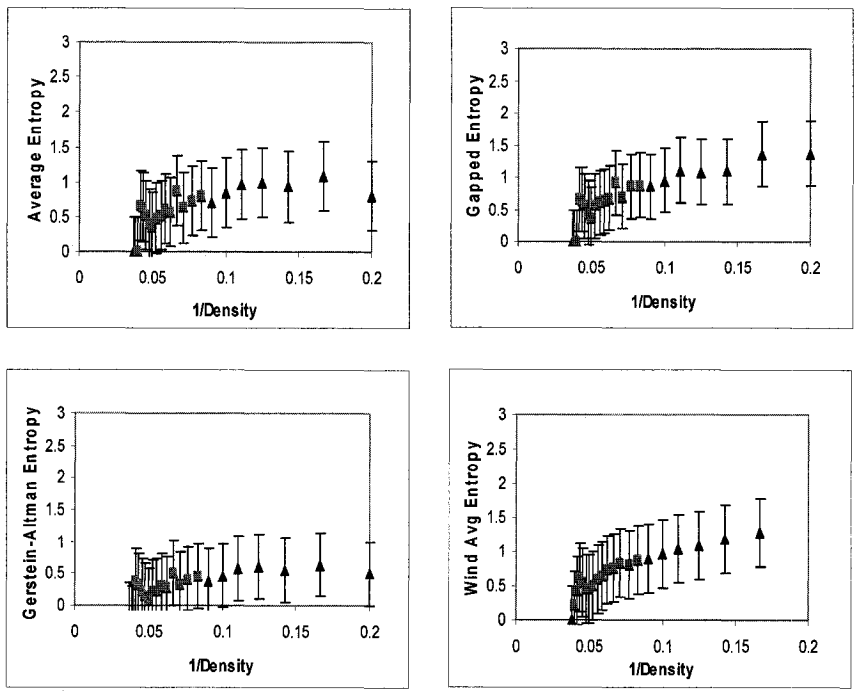

Figure A49. Various correlation plots for protein 1BFD

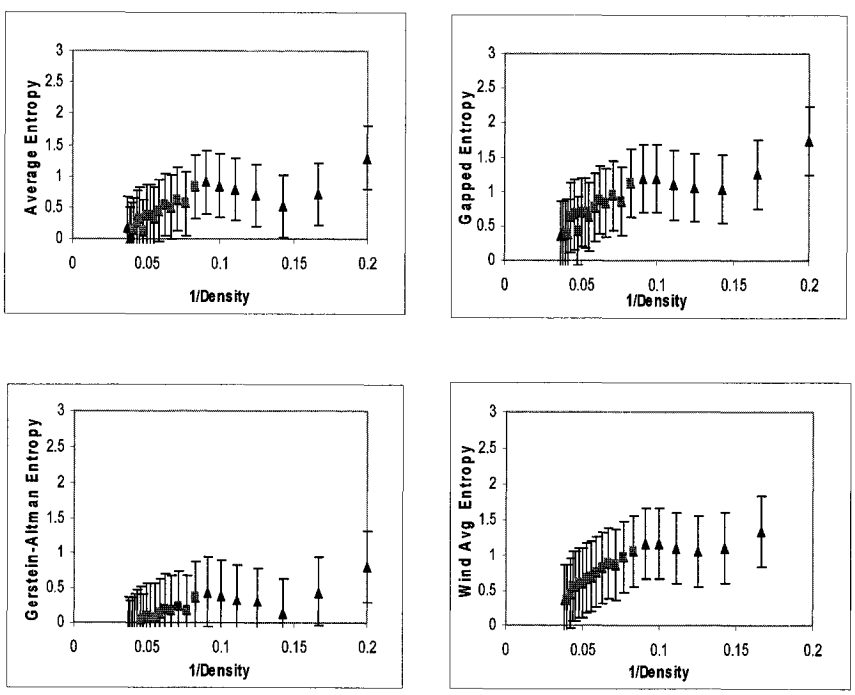

Figure A50. Various correlation plots for protein 1BG3

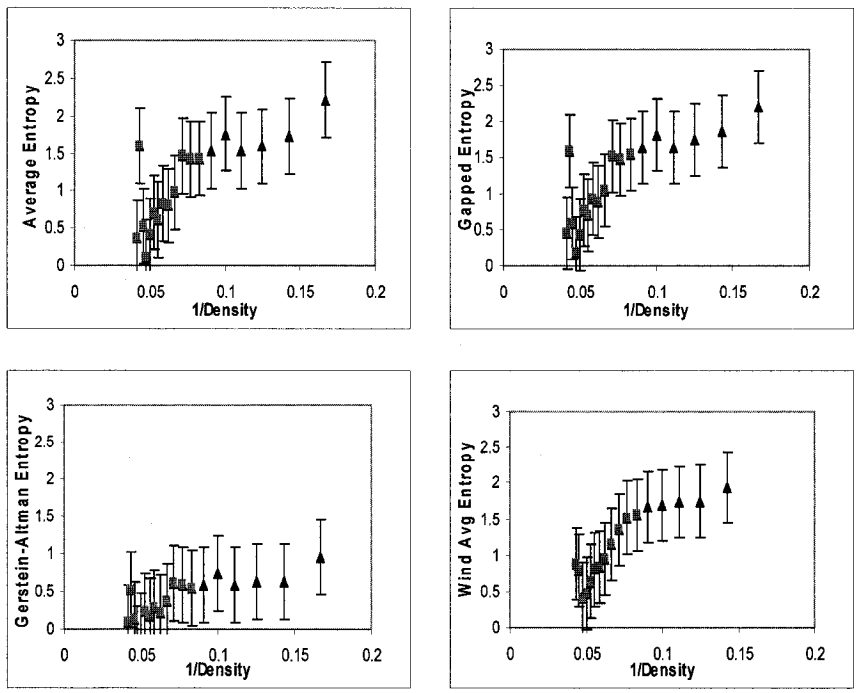

Figure A51. Various correlation plots for protein 1BGO

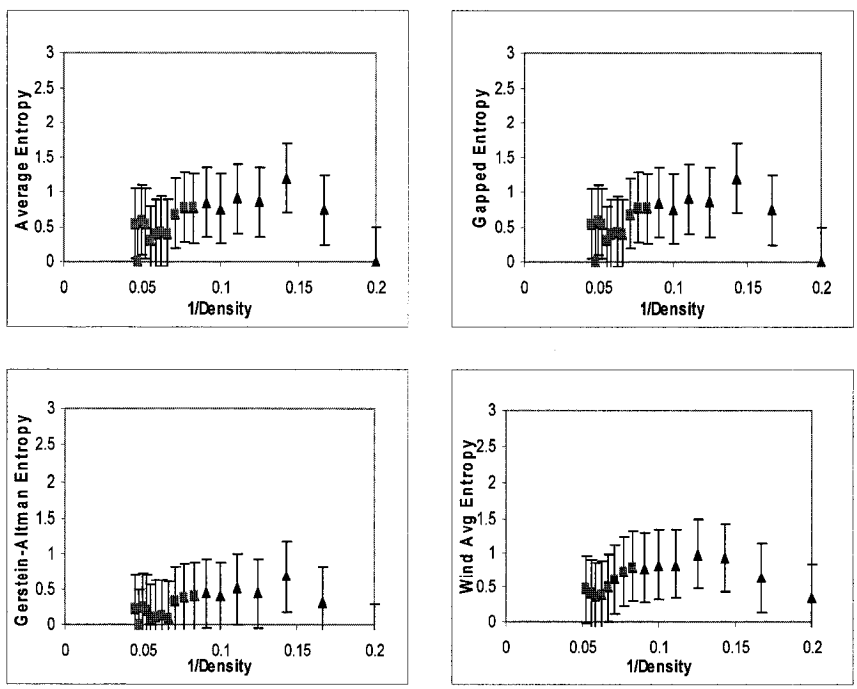

Figure A52. Various correlation plots for protein 1BIA

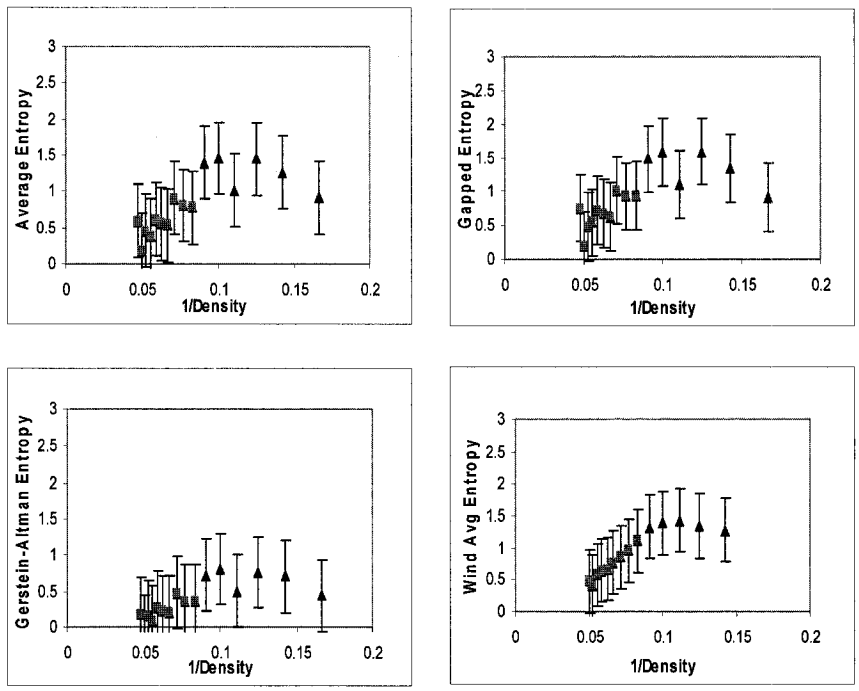

Figure A53. Various correlation plots for protein 1BLZ

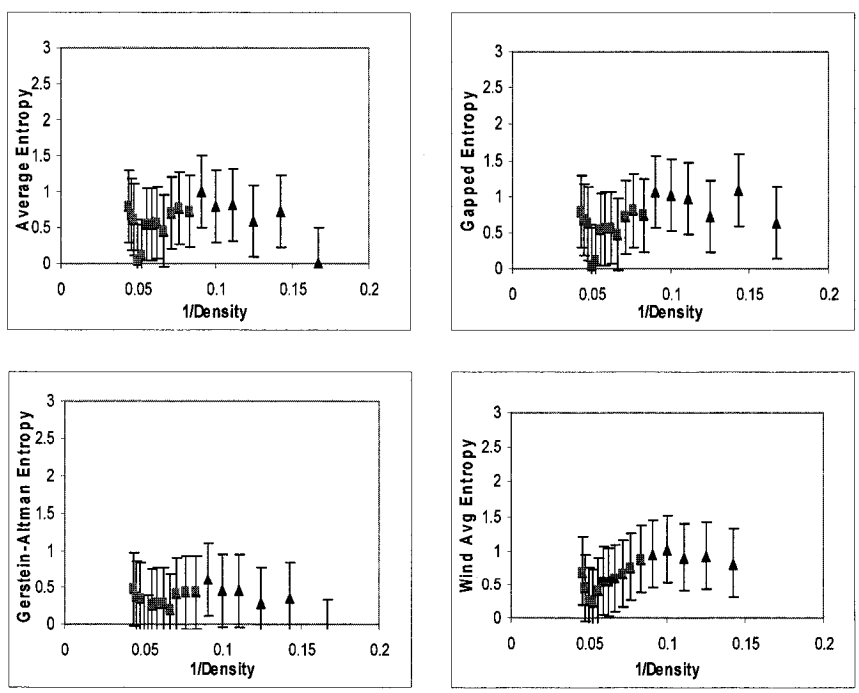

Figure A54. Various correlation plots for protein 1BN6

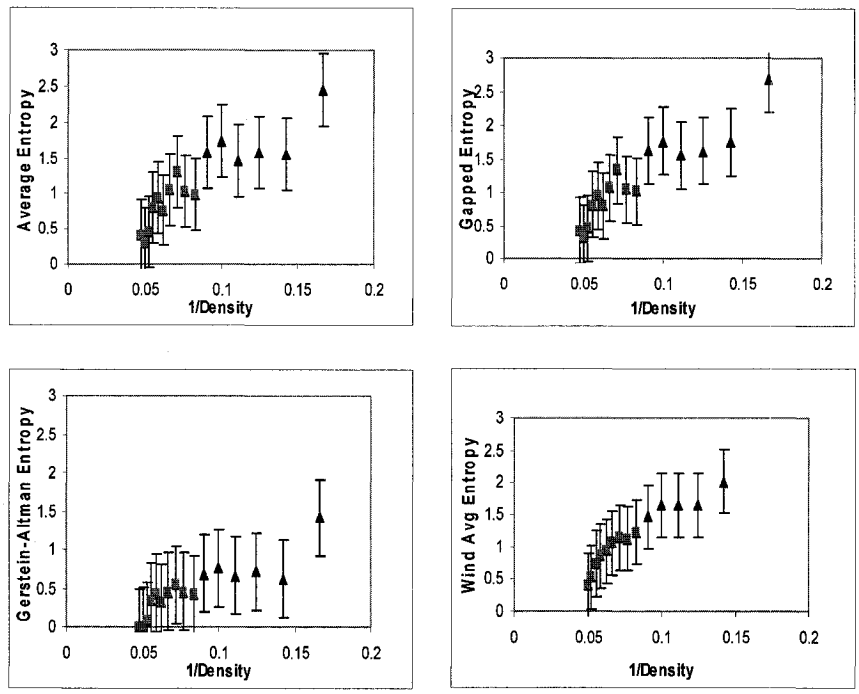

Figure A55. Various correlation plots for protein 1BO6

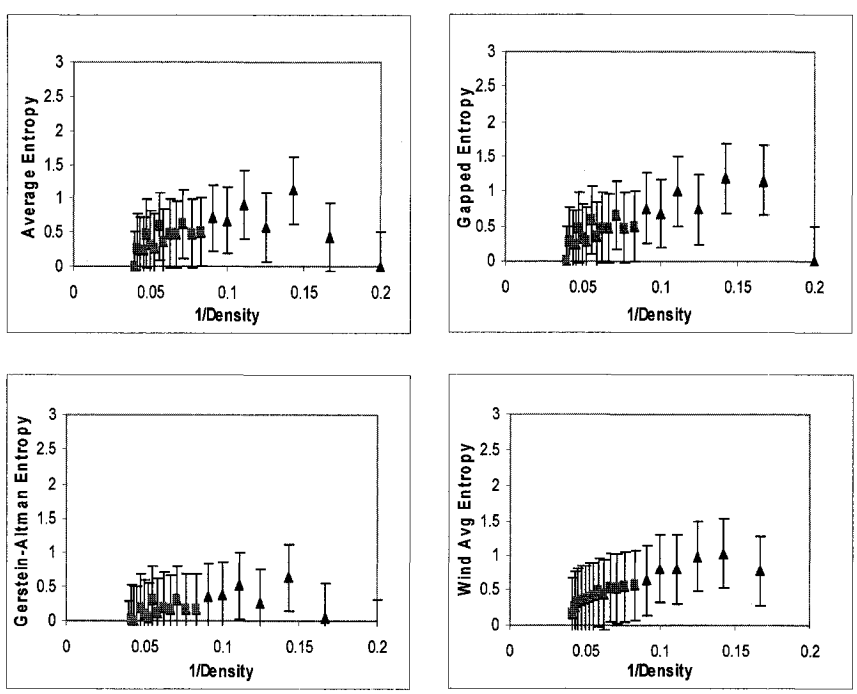

Figure A56. Various correlation plots for protein 1BOH

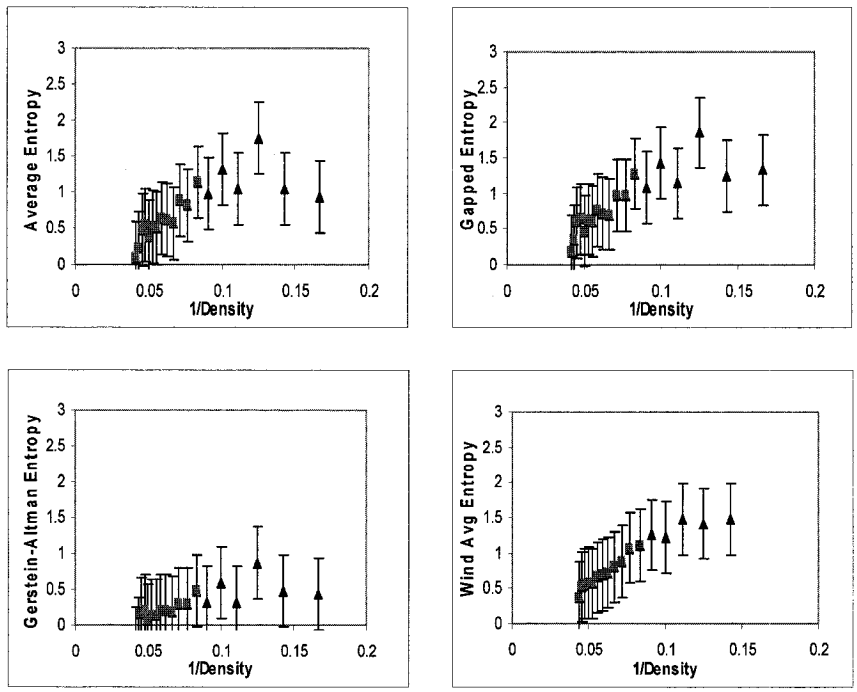

Figure A57. Various correlation plots for protein 1BSI

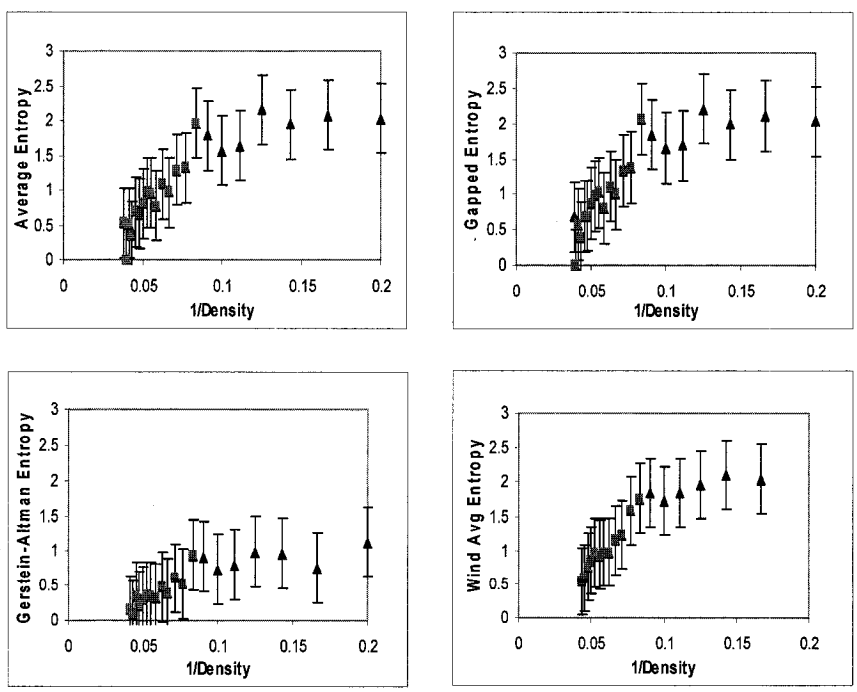

Figure A58. Various correlation plots for protein 1BT3

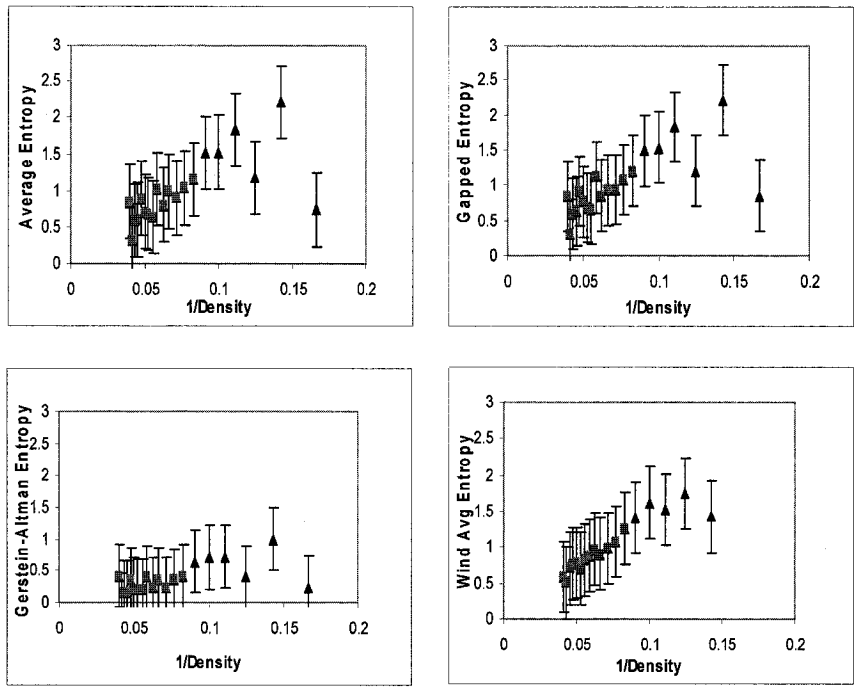

Figure A59. Various correlation plots for protein 1BUL

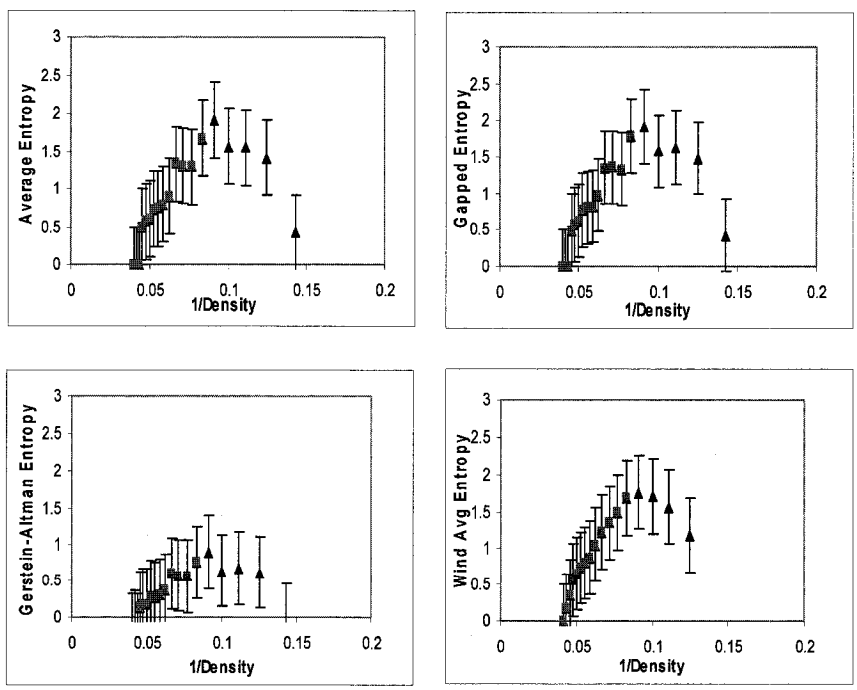

Figure A60. Various correlation plots for protein 1BXQ

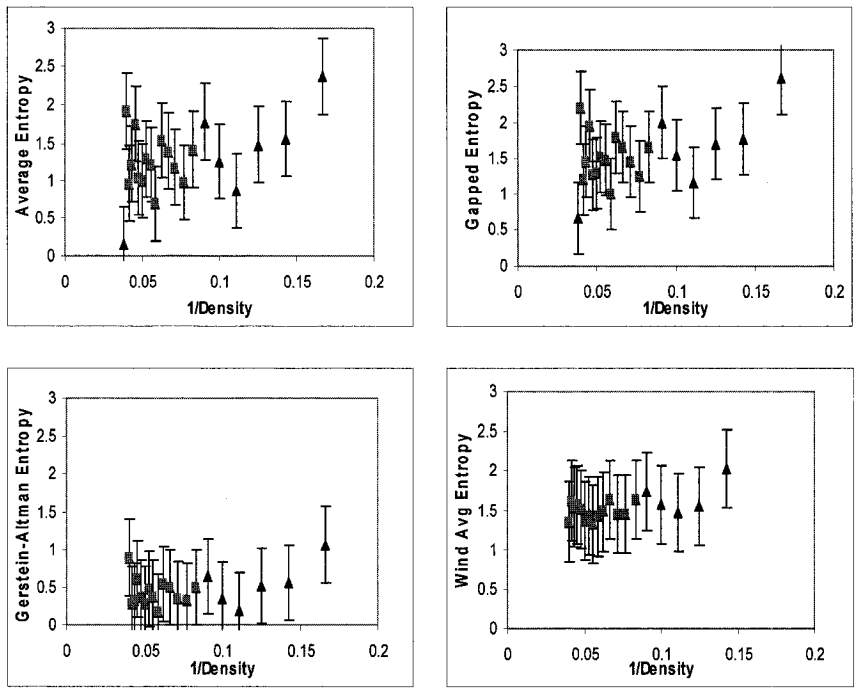

Figure A61. Various correlation plots for protein 1BYT

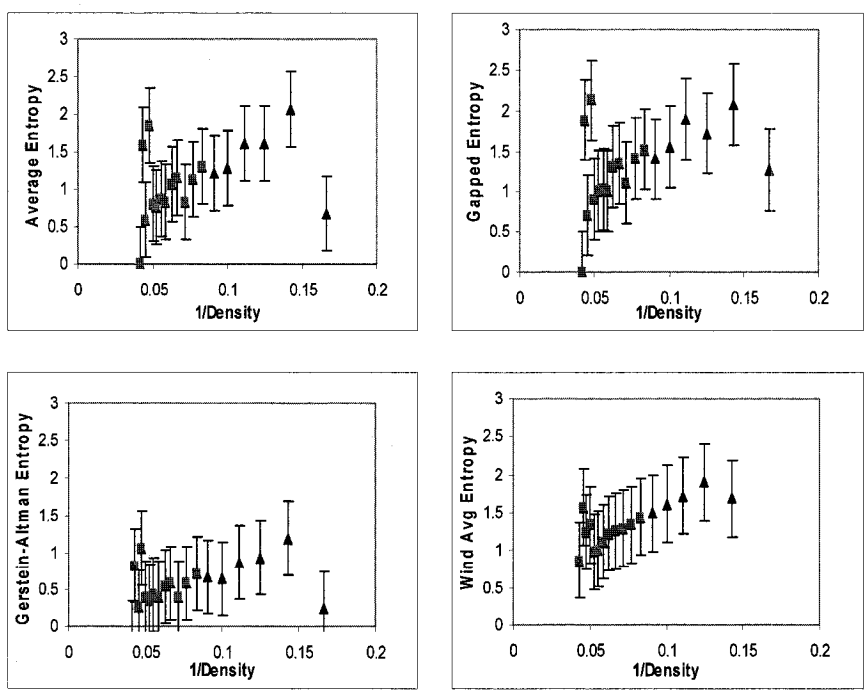

Figure A62. Various correlation plots for protein 1CBO

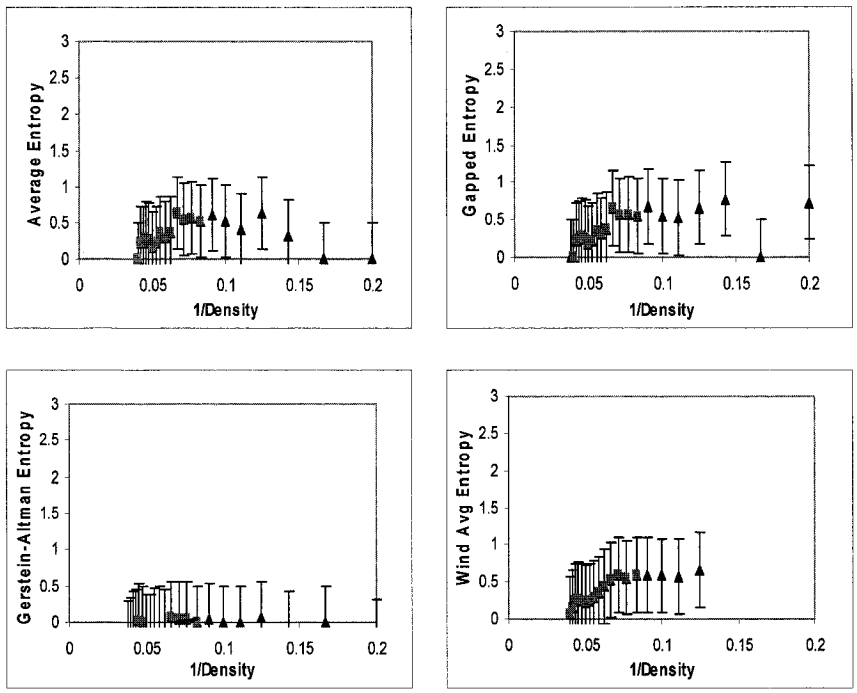

Figure A63. Various correlation plots for protein 1CEX

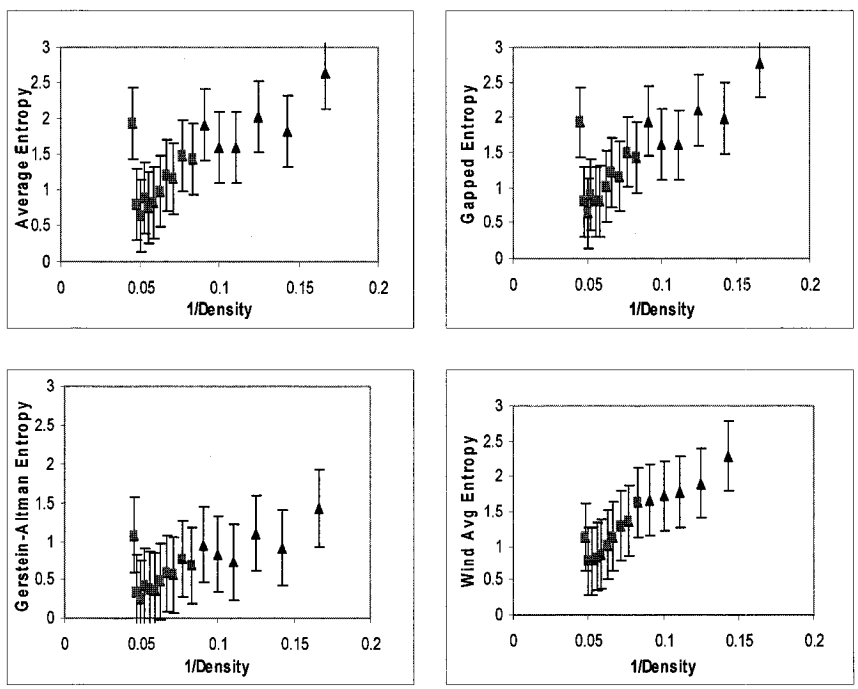

Figure A64. Various correlation plots for protein 1CJX

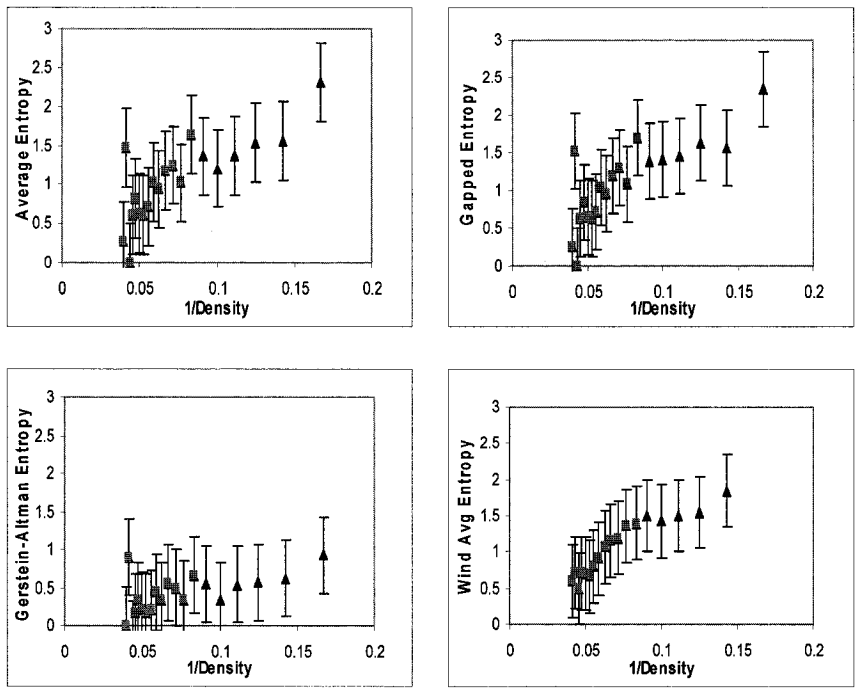

Figure A65. Various correlation plots for protein 1CK6

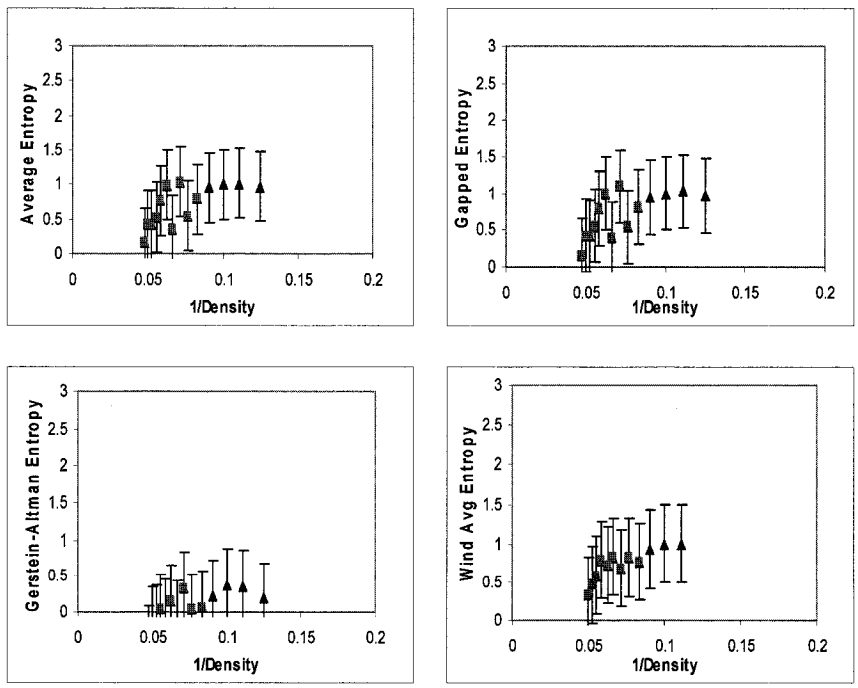

Figure A66. Various correlation plots for protein 1CRC

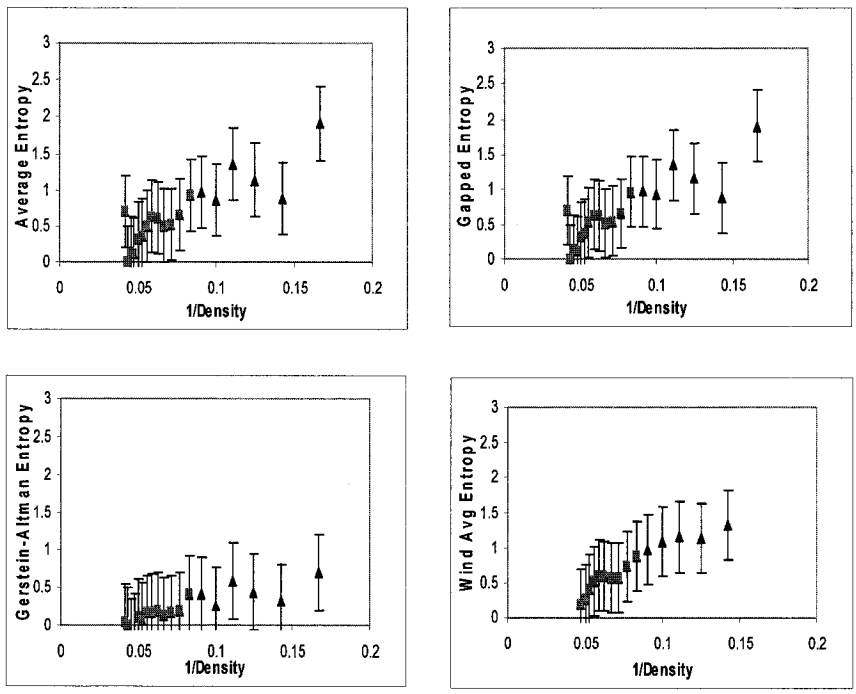

Figure A67. Various correlation plots for protein 1CRM

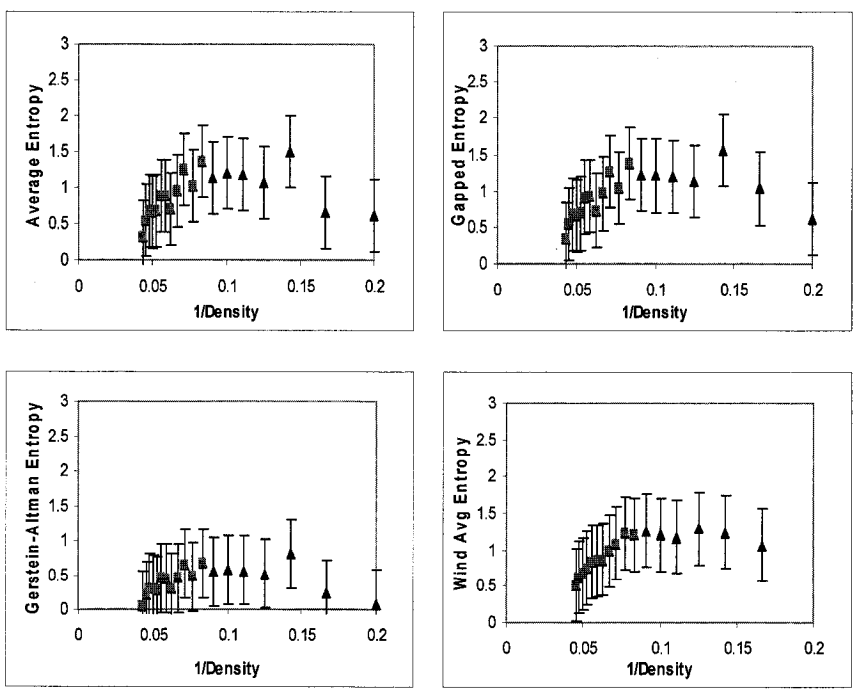

Figure A68. Various correlation plots for protein 1CRZ

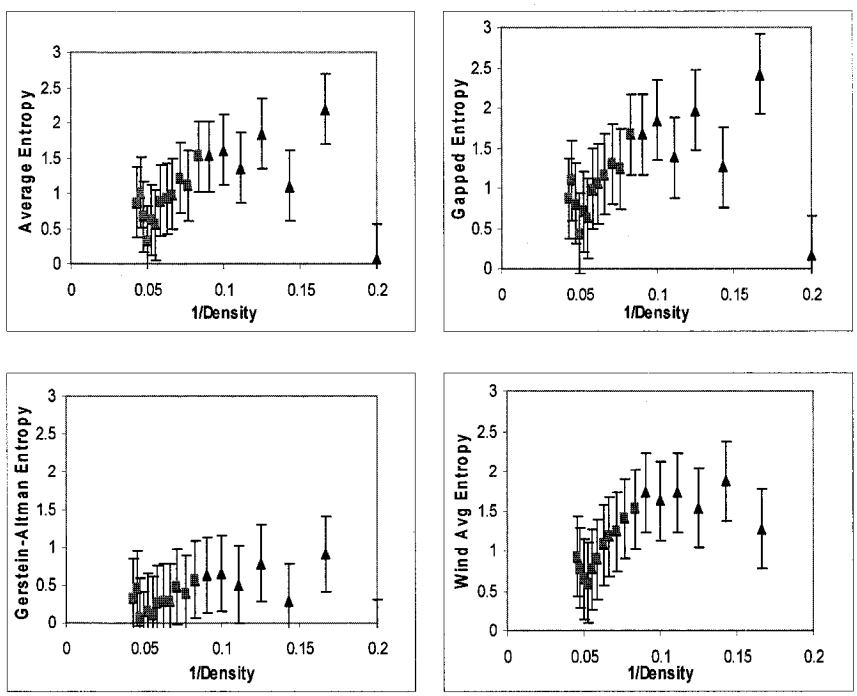

Figure A69. Various Correlation plots for protein 1CSR

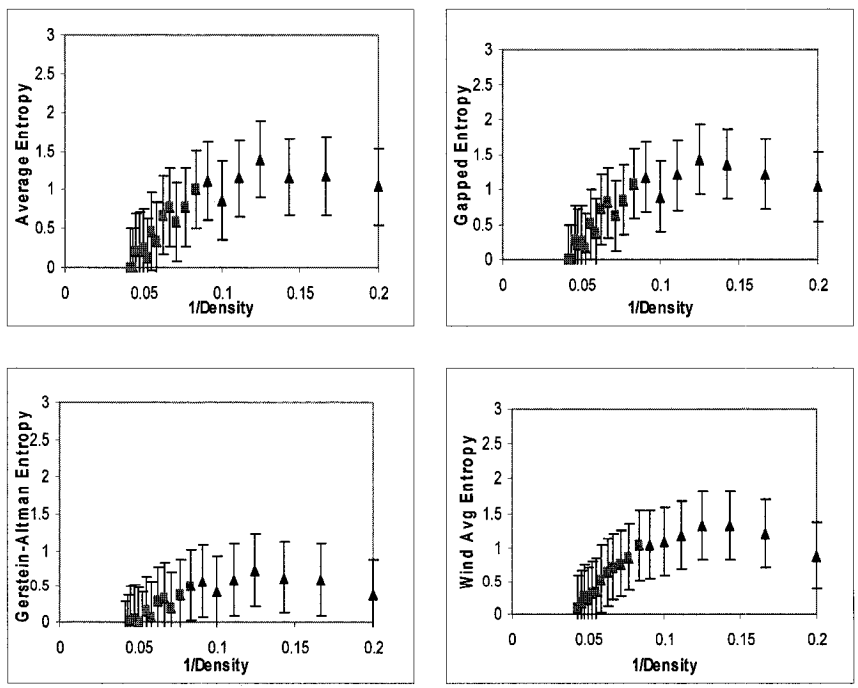

Figure A70. Various Correlation plots for protein 1D6M

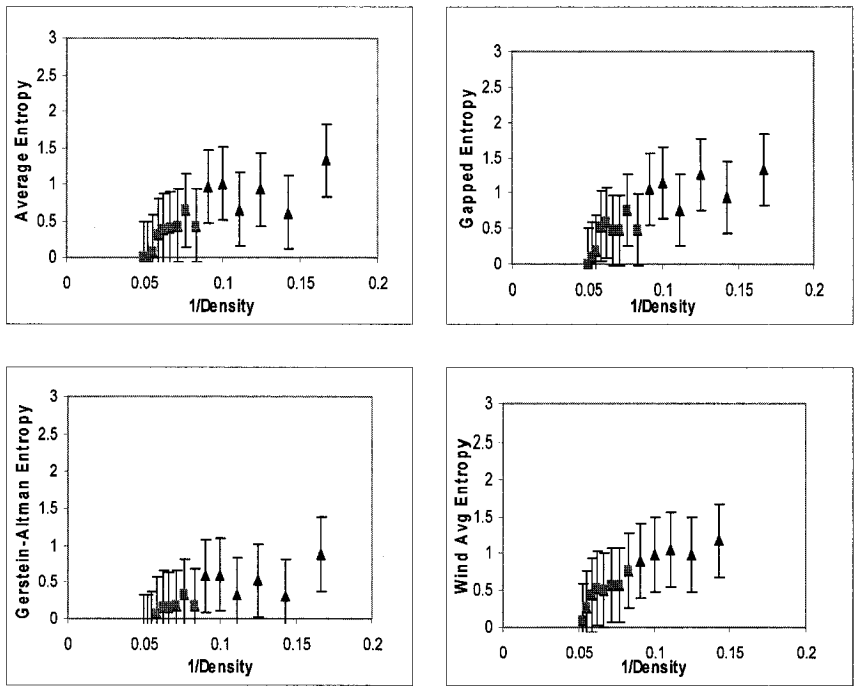

Figure A71. Various Correlation plots for protein 1DAJ

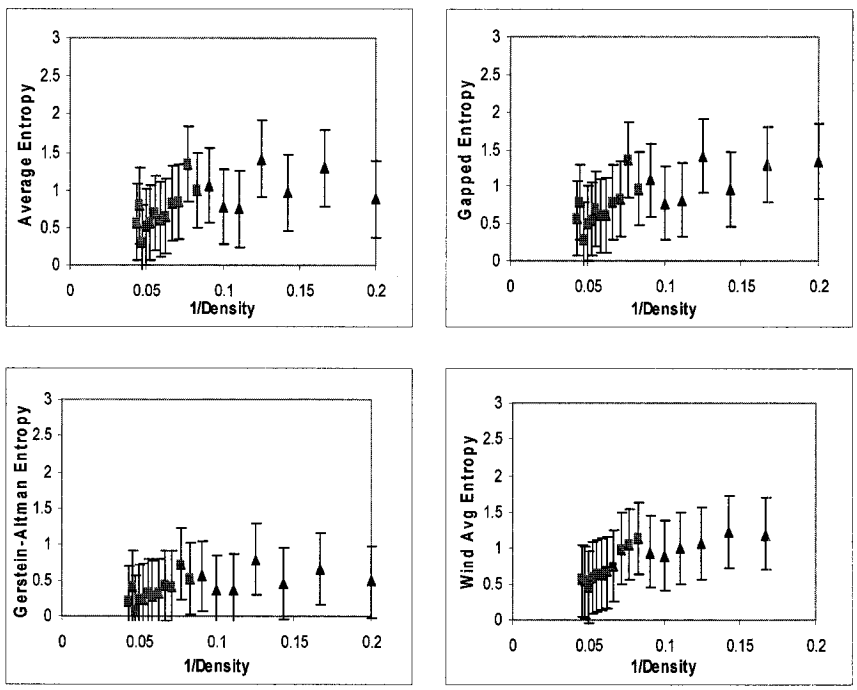

Figure A72. Various correlation plots for protein 1DCS

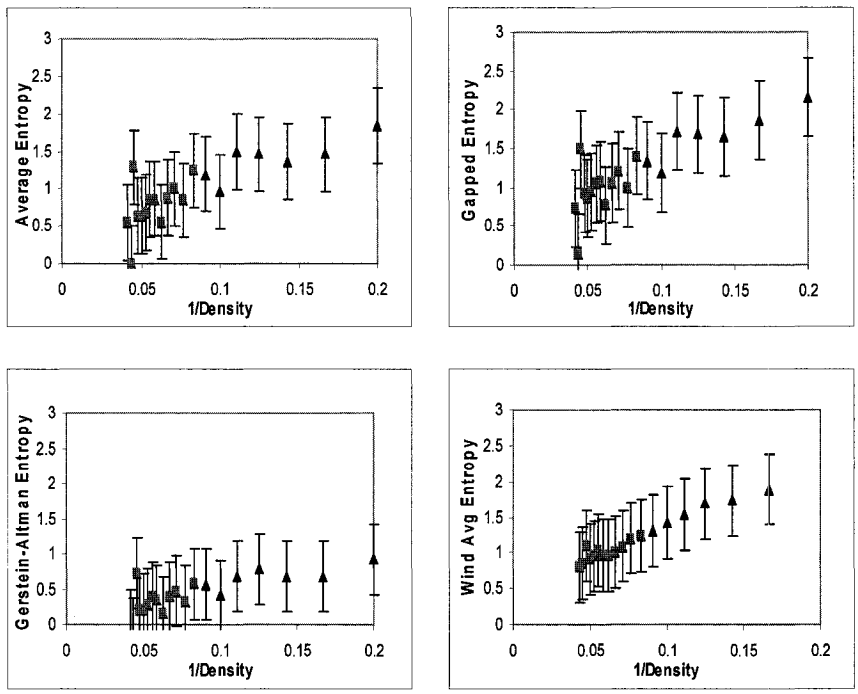

Figure A73. Various correlation plots for protein 1DHS

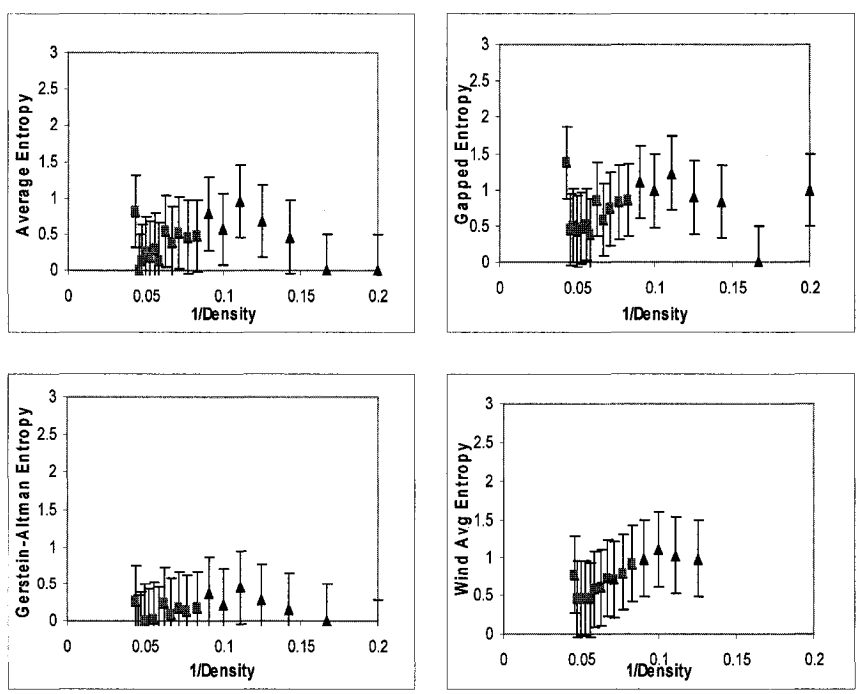

Figure A74. Various correlation plots for protein 1DHT

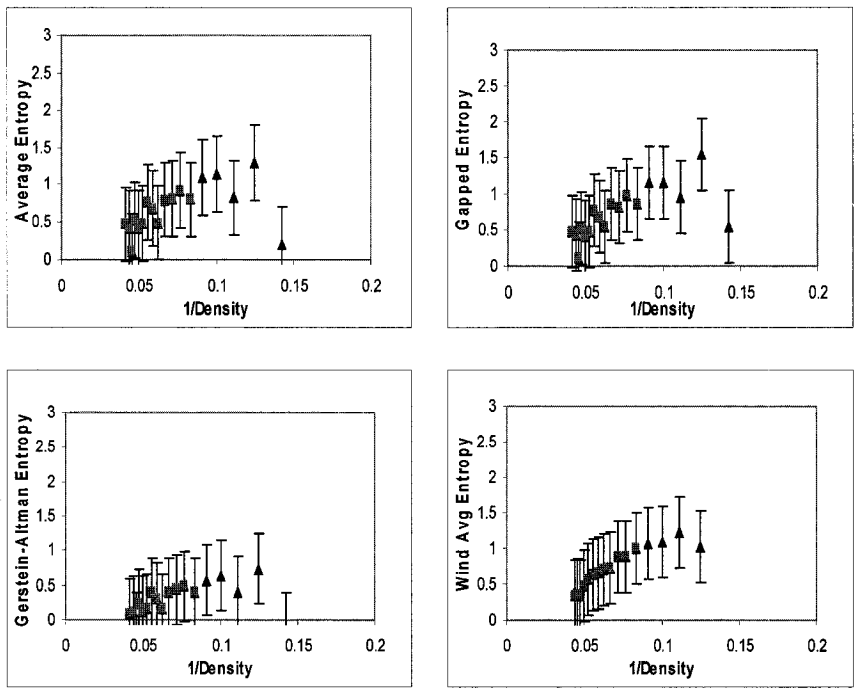

Figure A75. Various correlation plots for protein 1DIN

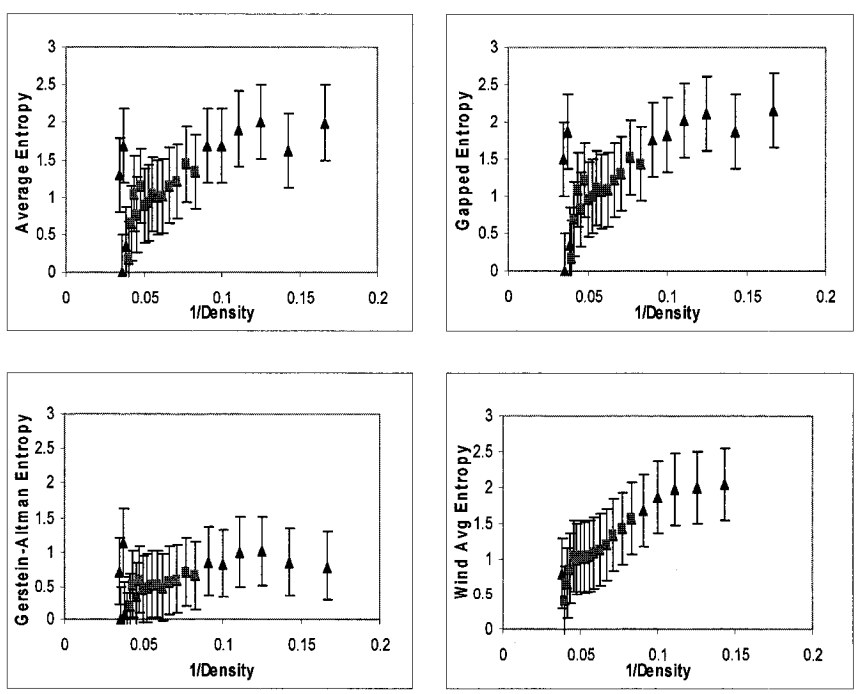

Figure A76. Various correlation plots for protein 1DMR

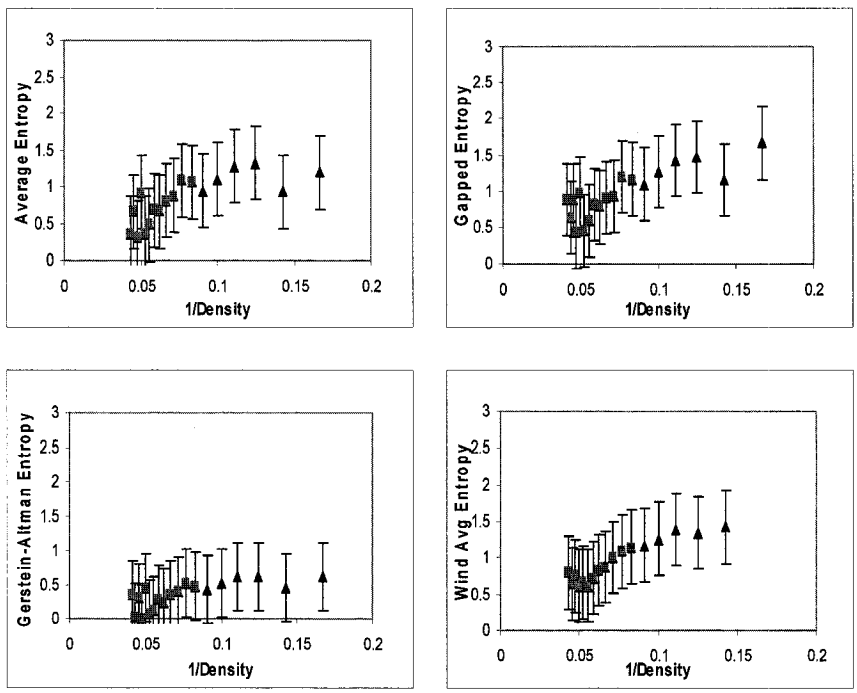

Figure A77. Various correlation plots for protein 1E1K

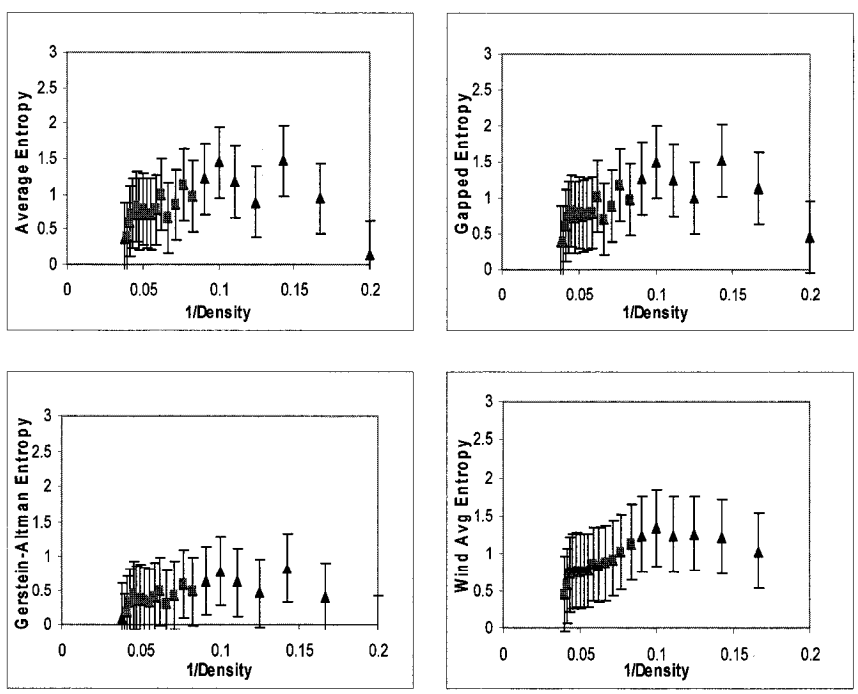

Figure A78. Various correlation plots for protein 1E3H

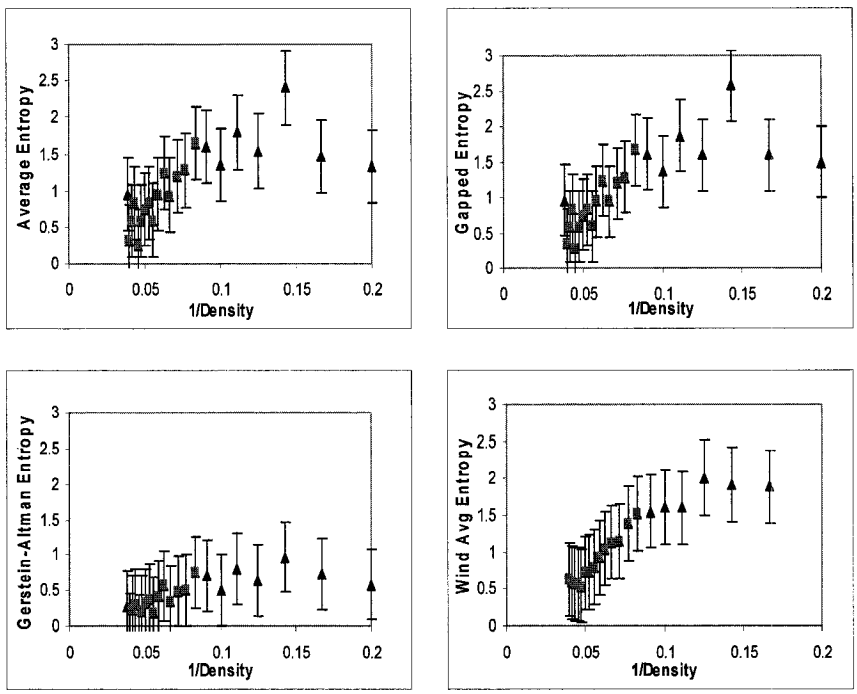

Figure A79. Various correlation plots for protein 1E3Q

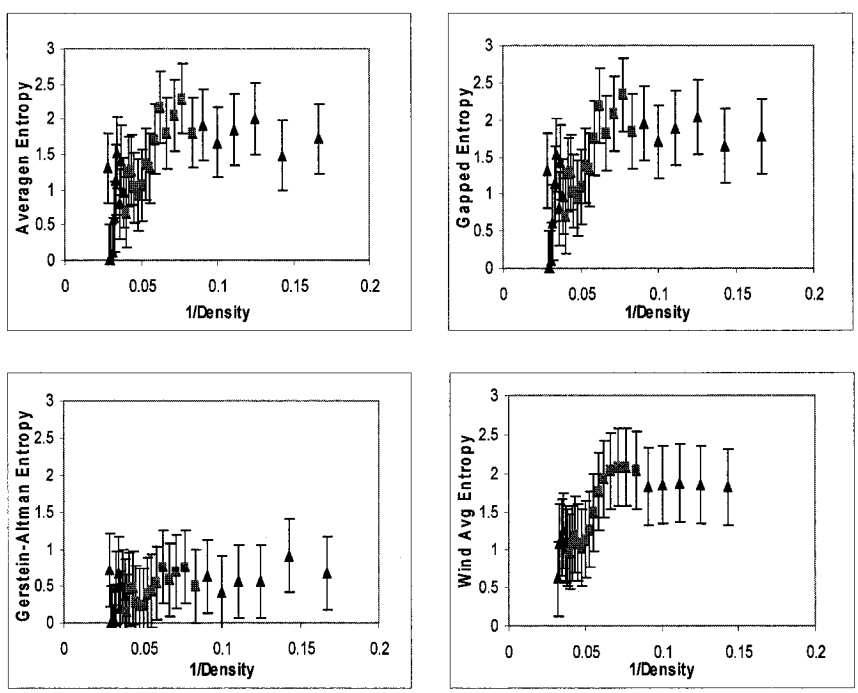

Figure A80. Various correlation plots for protein 1E5M

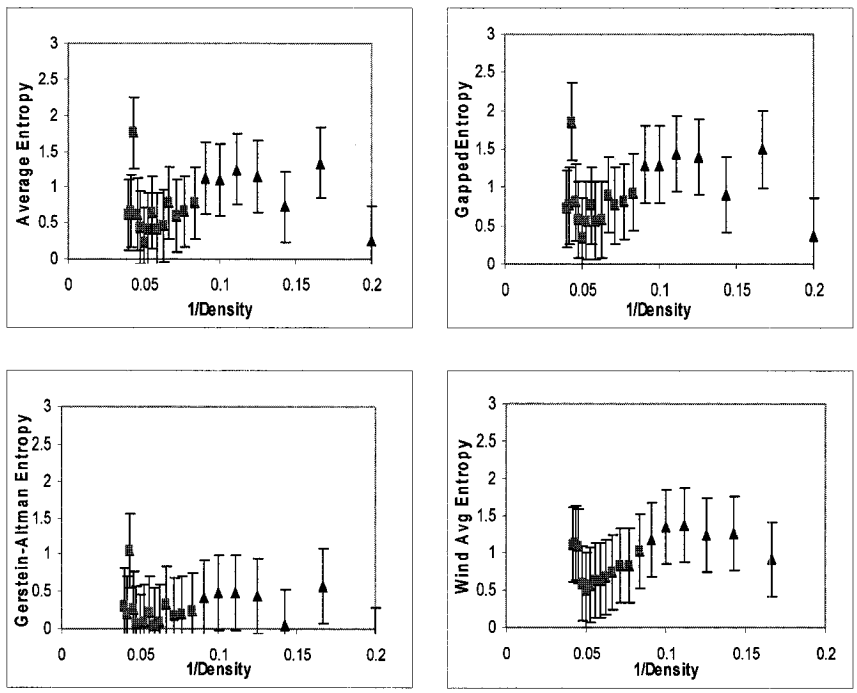

Figure A81. Various correlation plots for protein 1EBV

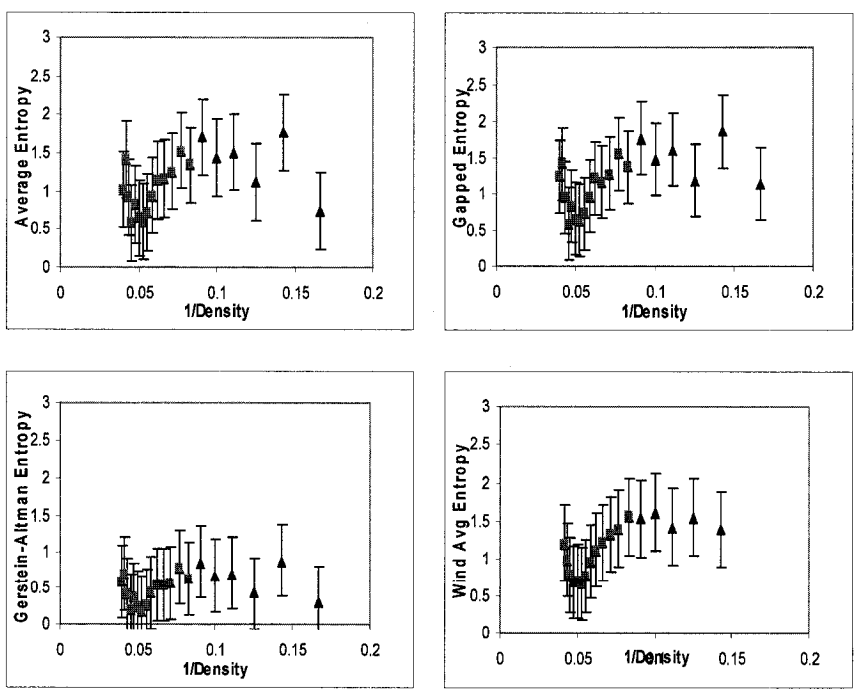

Figure A82. Various correlation plots for protein 1EEH

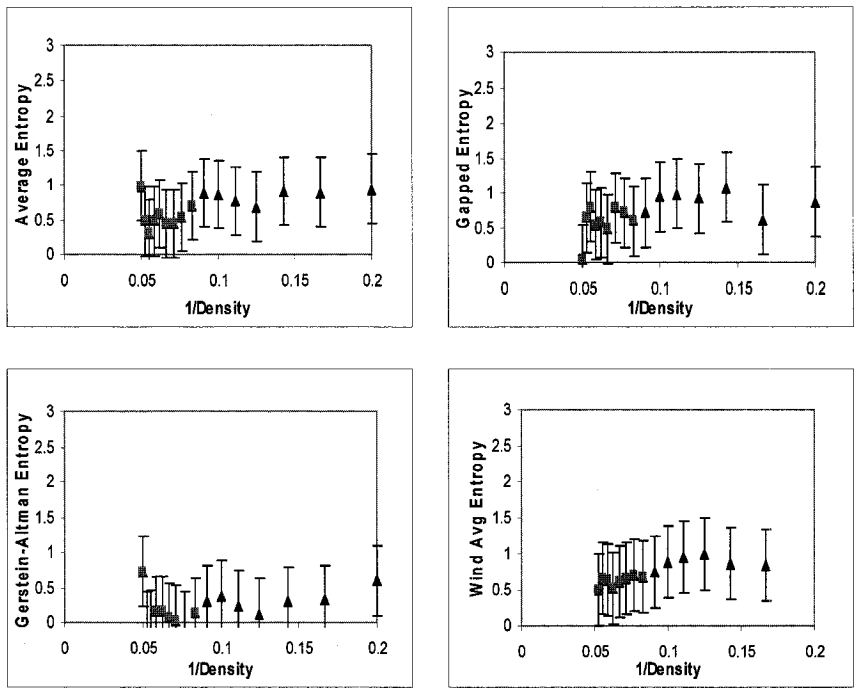

Figure A83. Various correlation plots for protein 1HGU

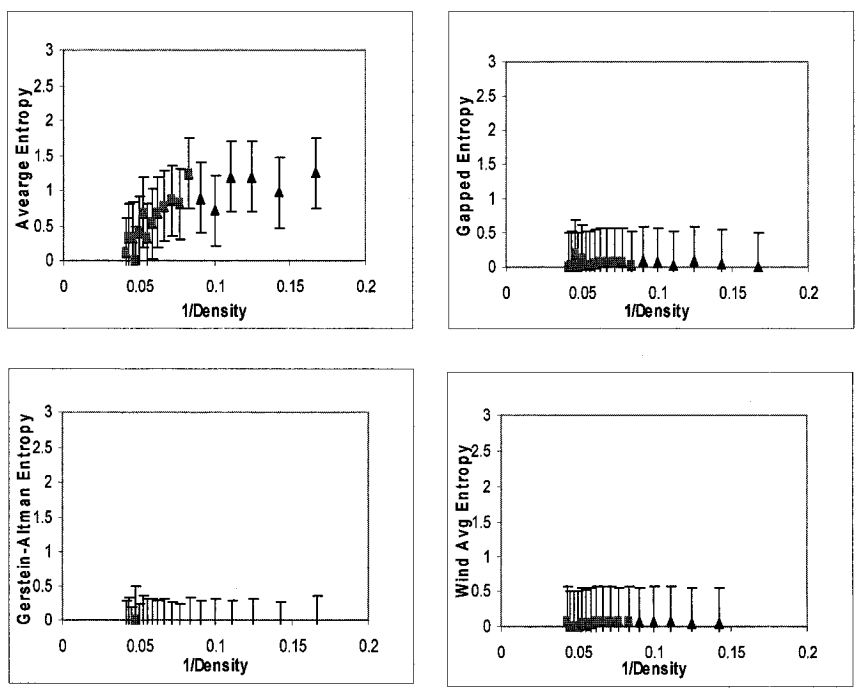

Figure A84. Various correlation plots for protein 1LZ1

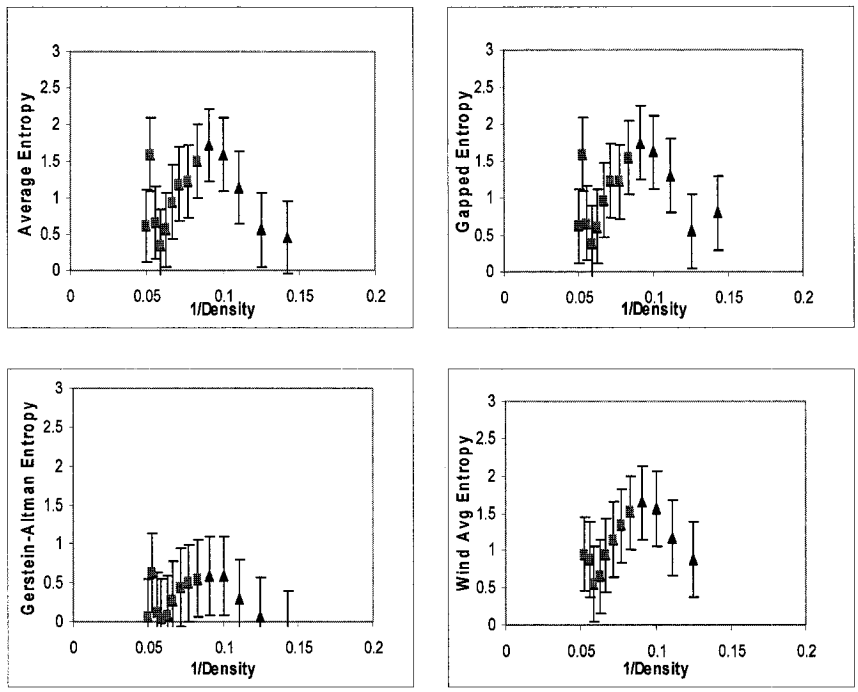

Figure A85. Various correlation plots for protein 1OMD

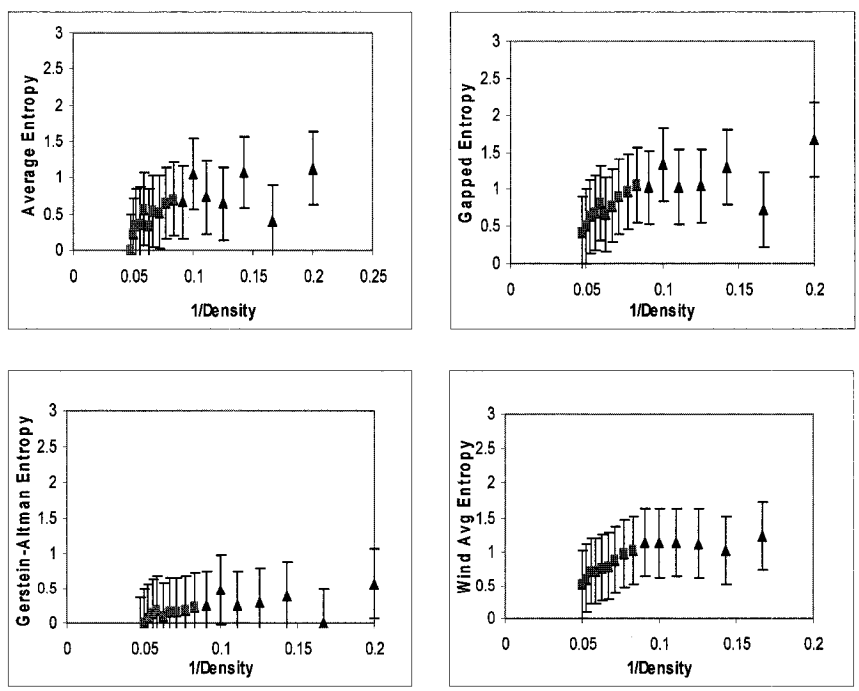

Figure A86. Various correlation plots for protein 1RBP

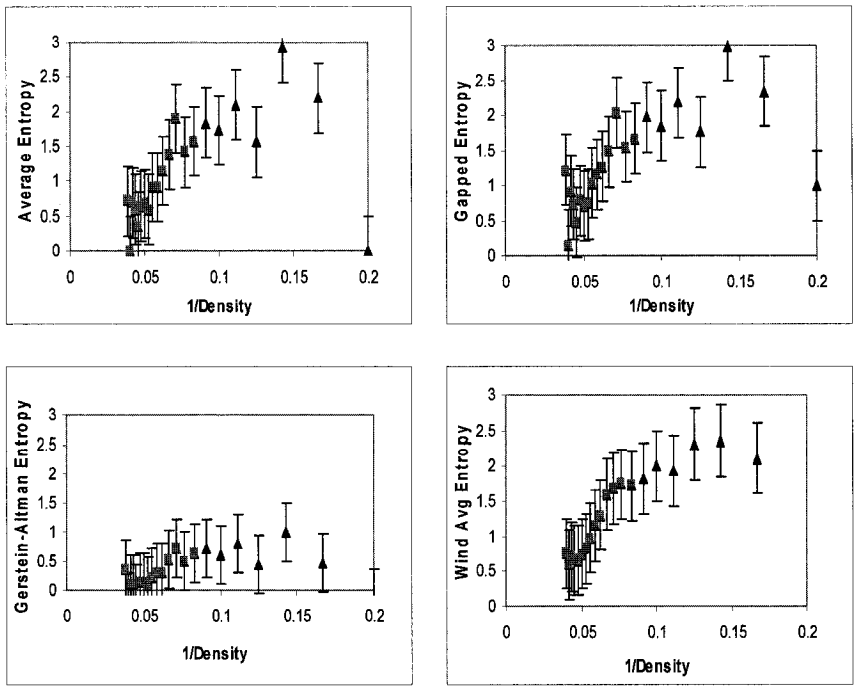

Figure A87. Various correlation plots for protein 1TON

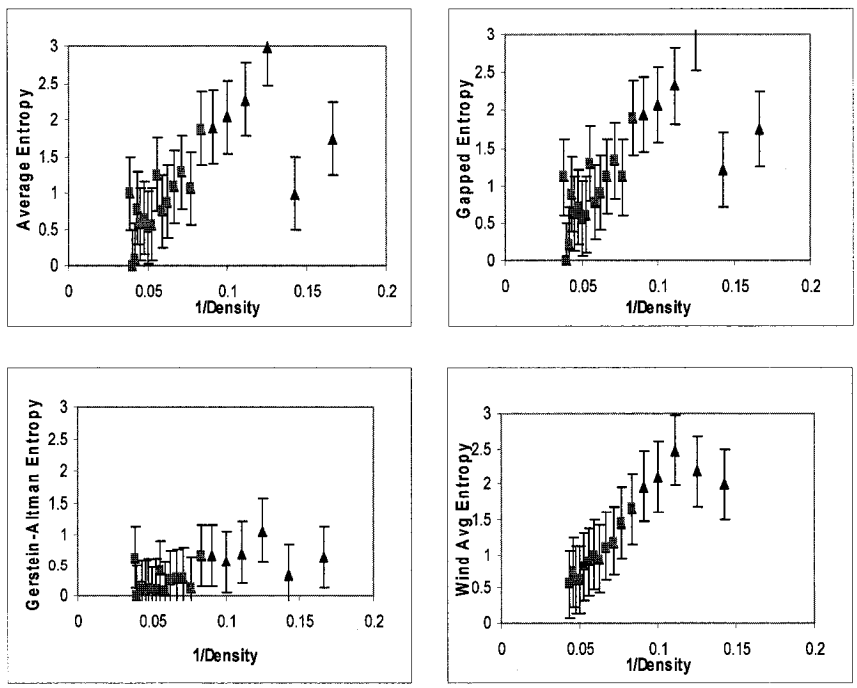

Figure A88. Various correlation plots for protein 2ACT

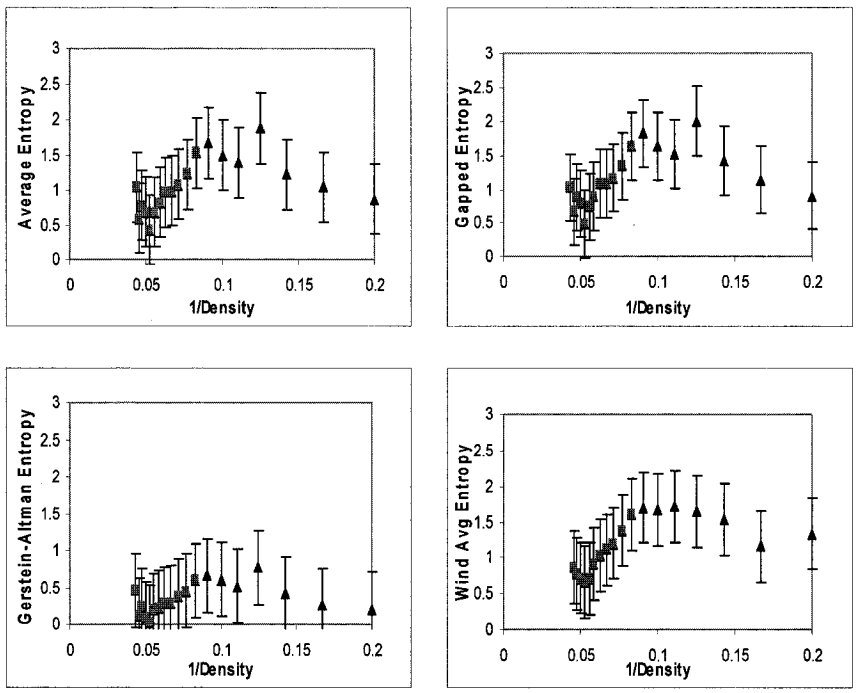

Figure A89. Various correlation plots for protein 2CTS

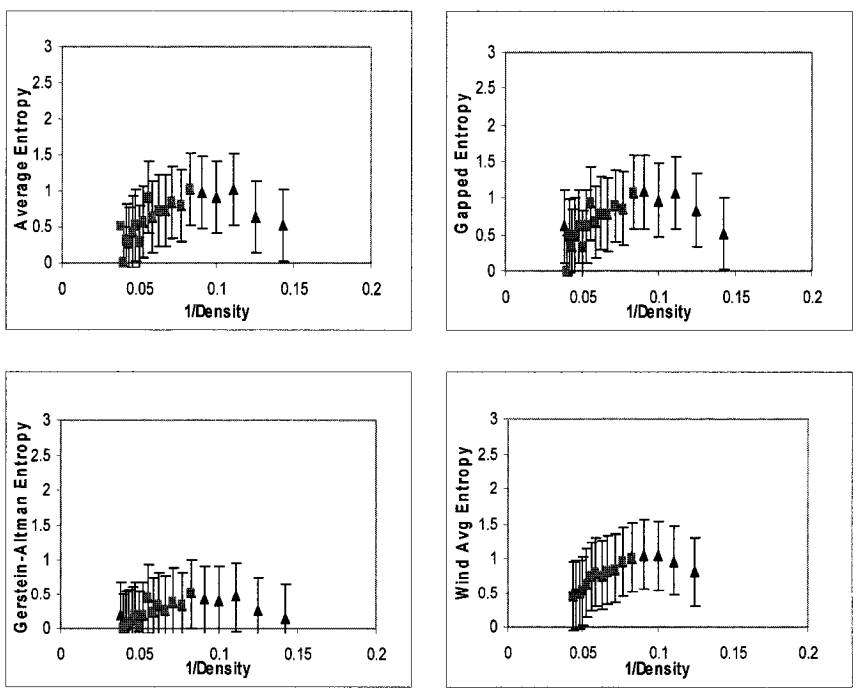

Figure A90. Various correlation plots for protein 2LBP

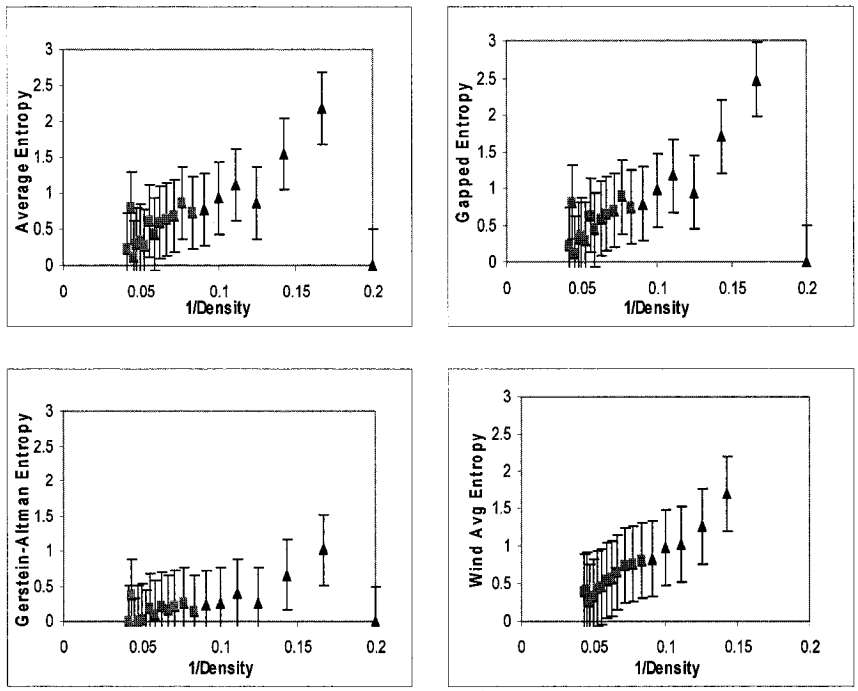

Figure A91. Various correlation plots for protein 2LDX

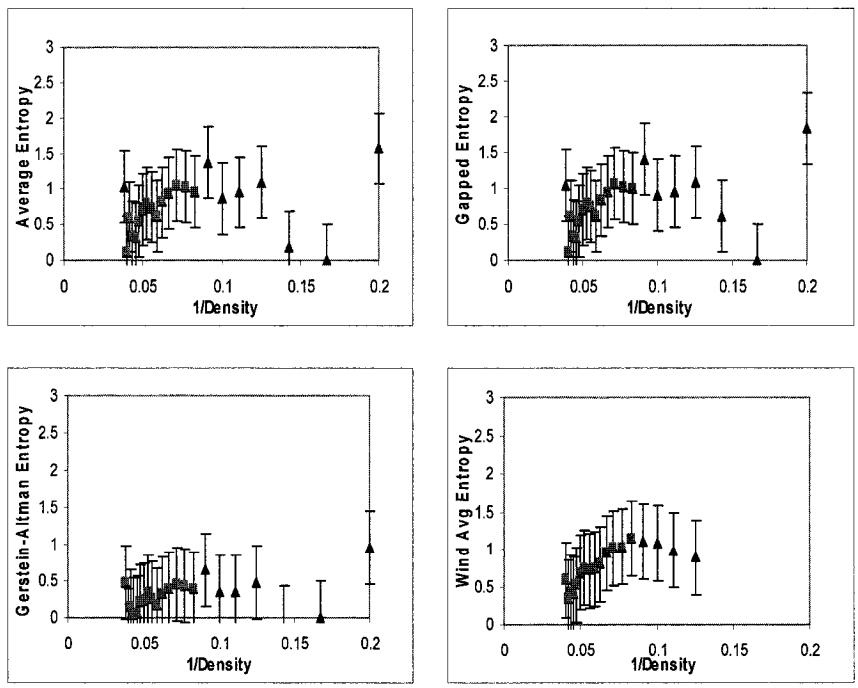

Figure A92. Various correlation plots for protein 2LIV

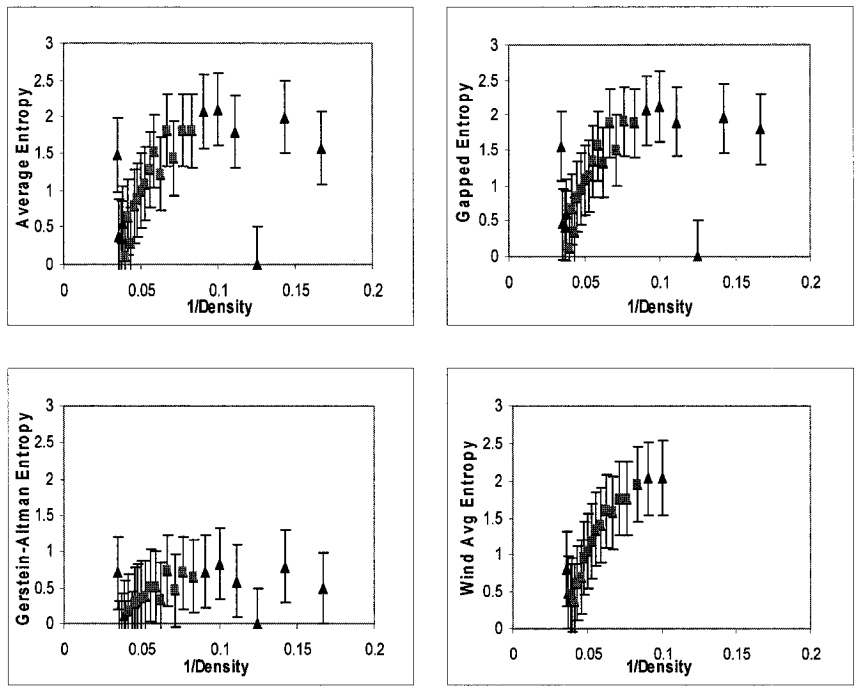

Figure A93. Various correlation plots for protein 2PRK

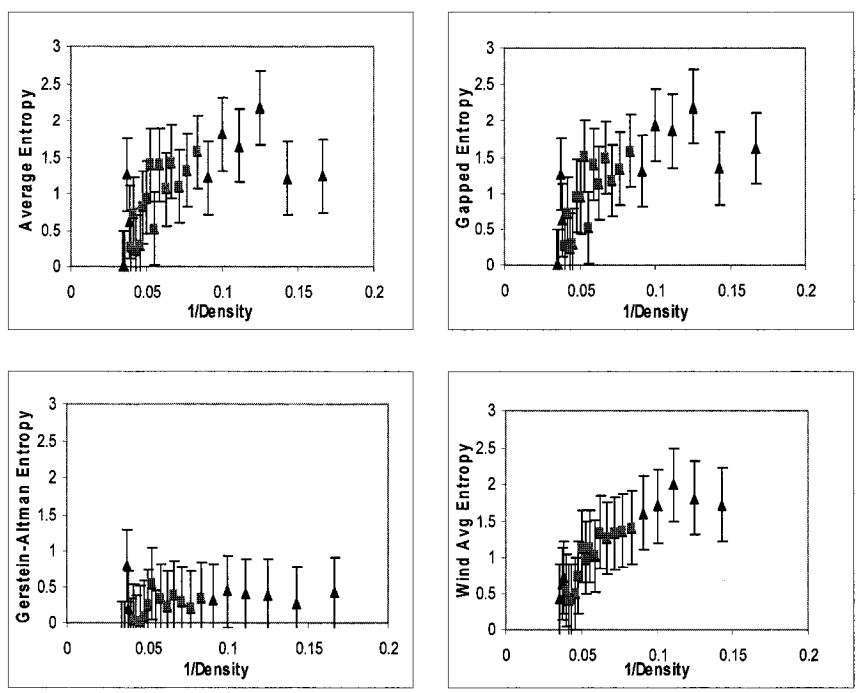

Figure A94. Various correlation plots for protein 2RN2

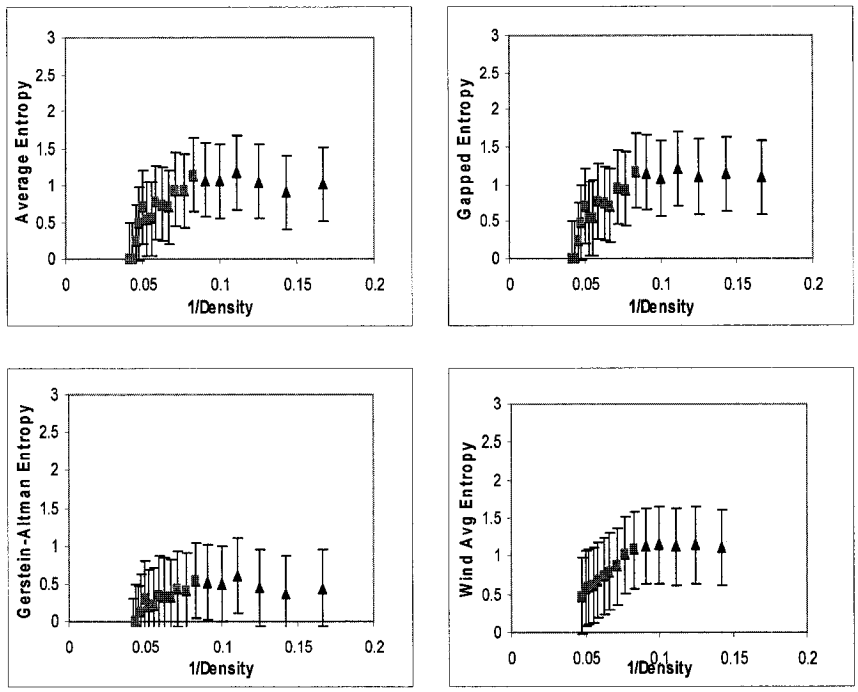

Figure A95. Various correlation plots for protein 2TAA

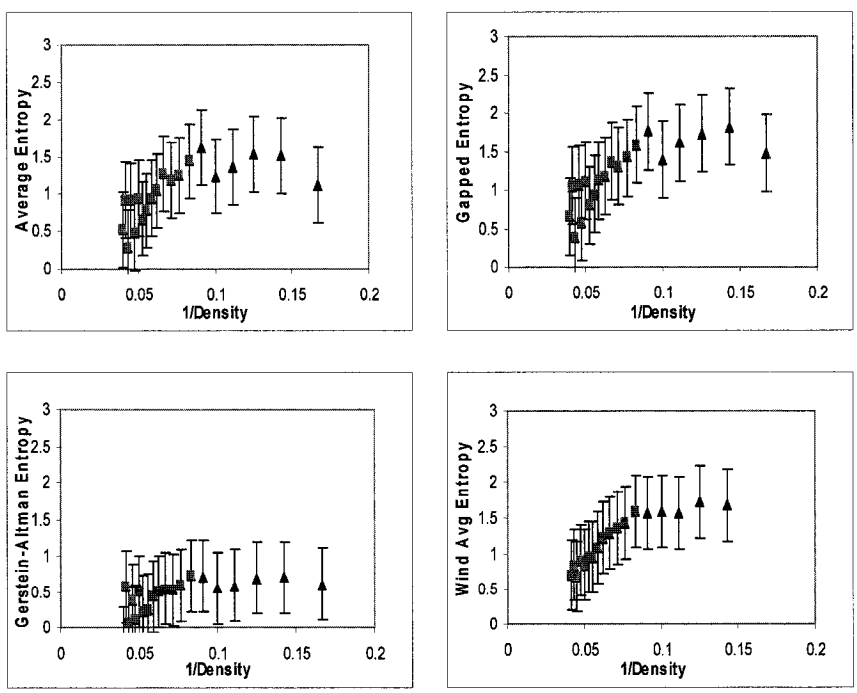

Figure A96. Various correlation plots for protein 3BLM

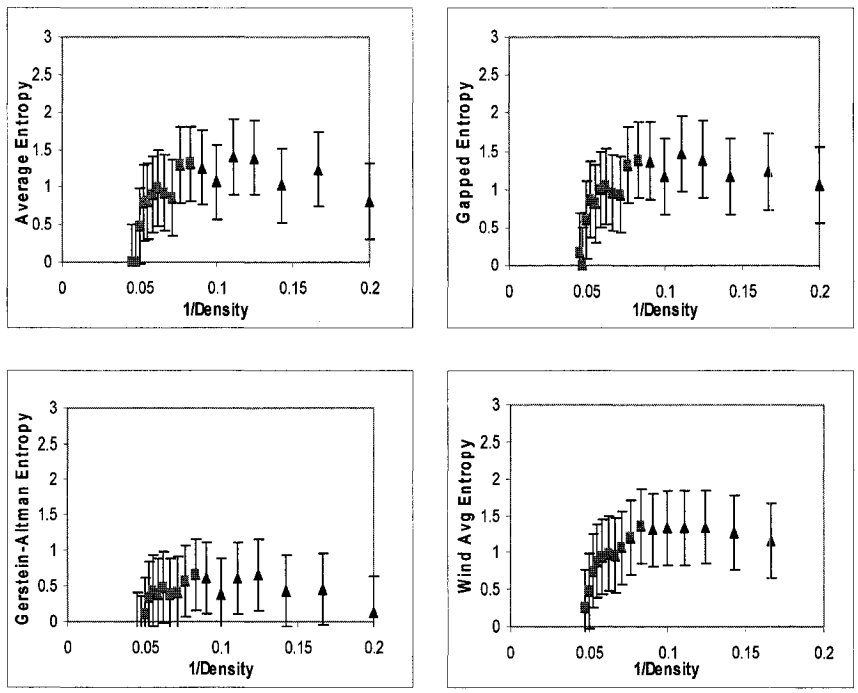

Figure A97. Various correlation plots for protein 3CLA

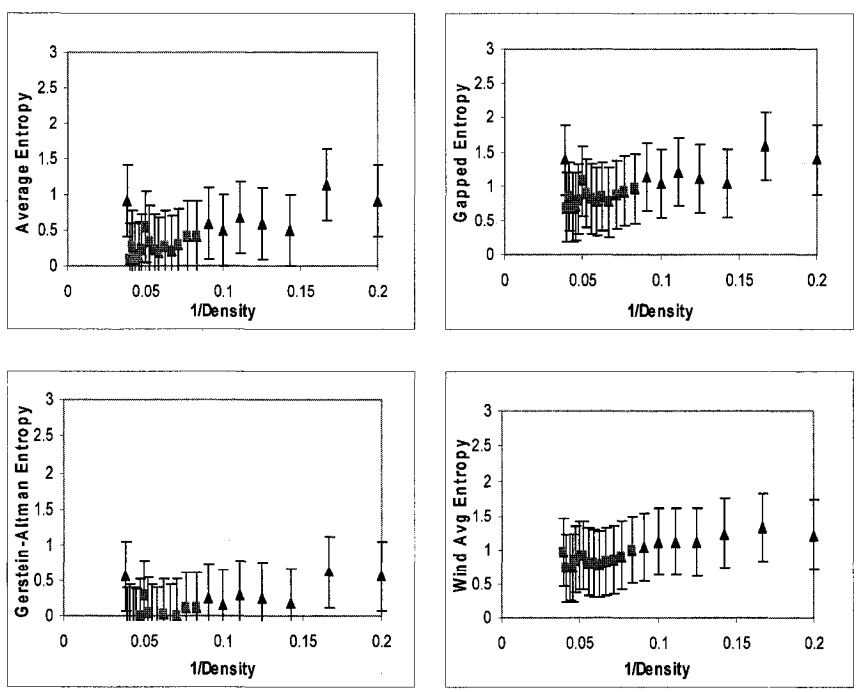

Figure A98. Various correlation plots for protein 3CNA

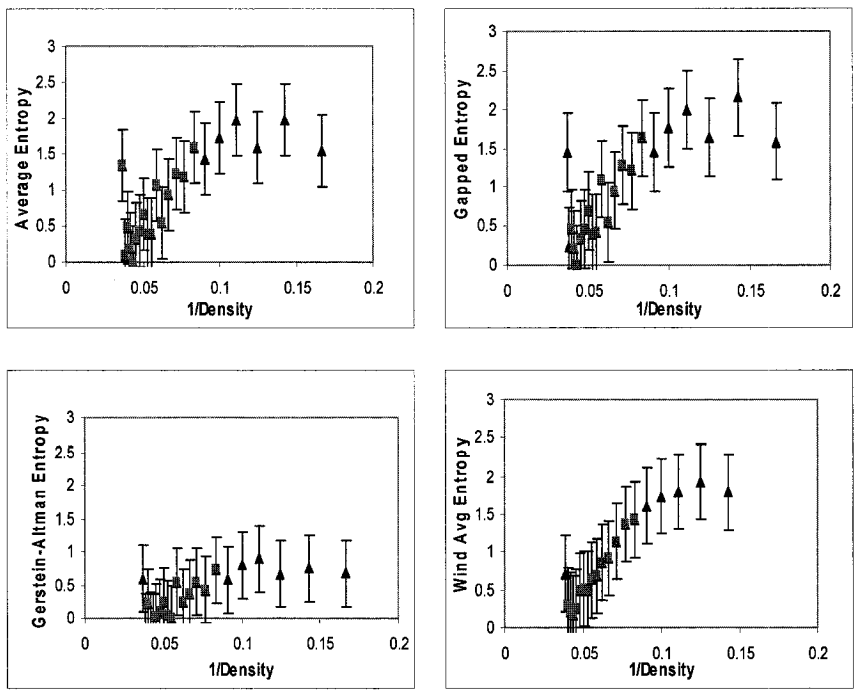

Figure A99. Various correlation plots for protein 3EST

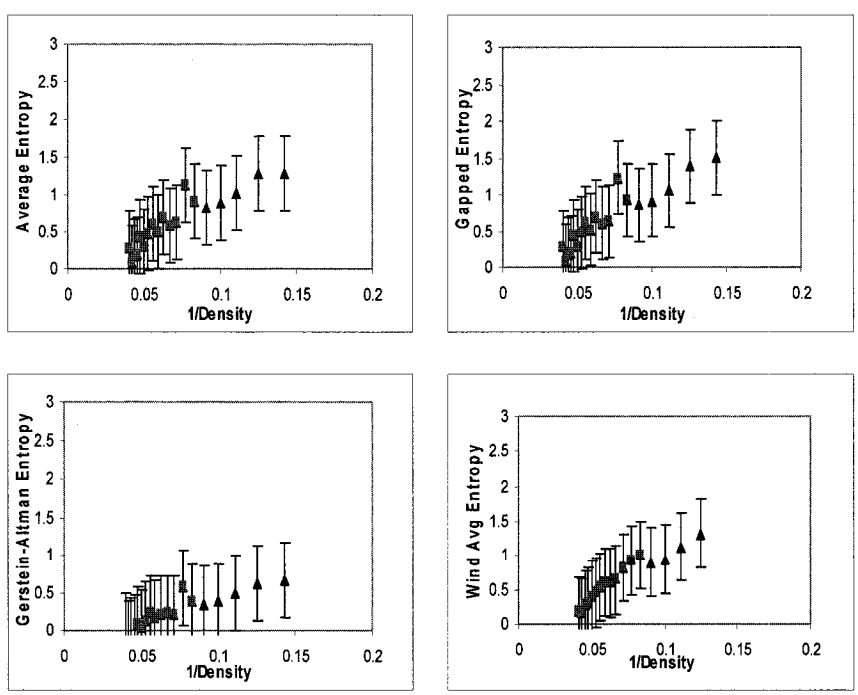

Figure A100. Various correlation plots for protein 3GBP

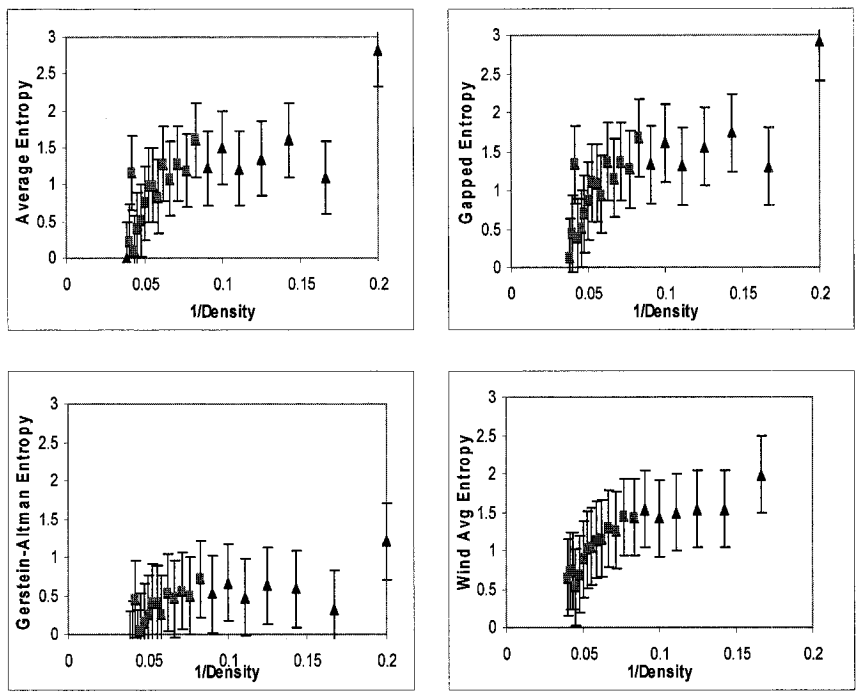

Figure A101. Various correlation plots for protein 3GRS

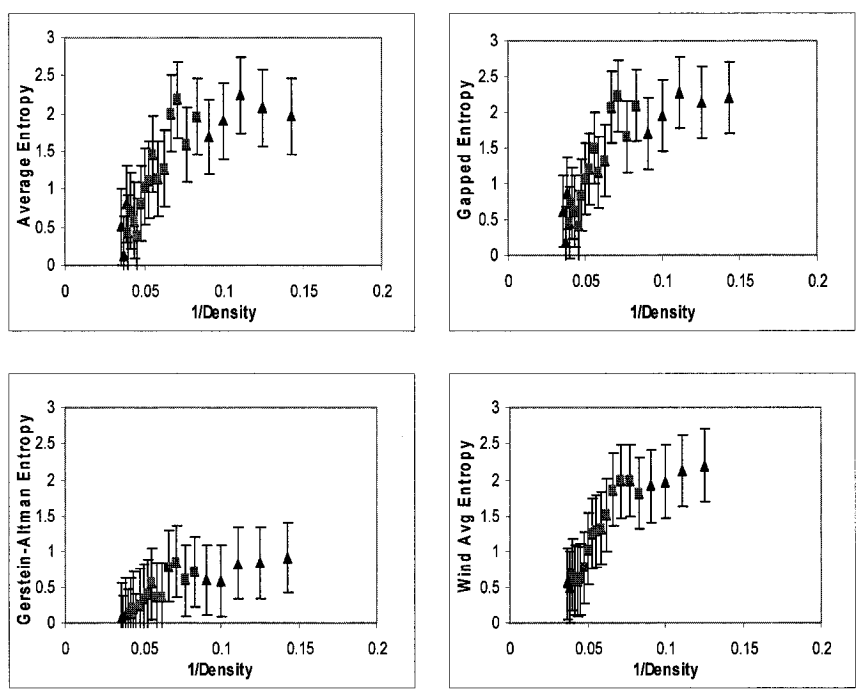

Figure A102. Various correlation plots for protein 3PFK

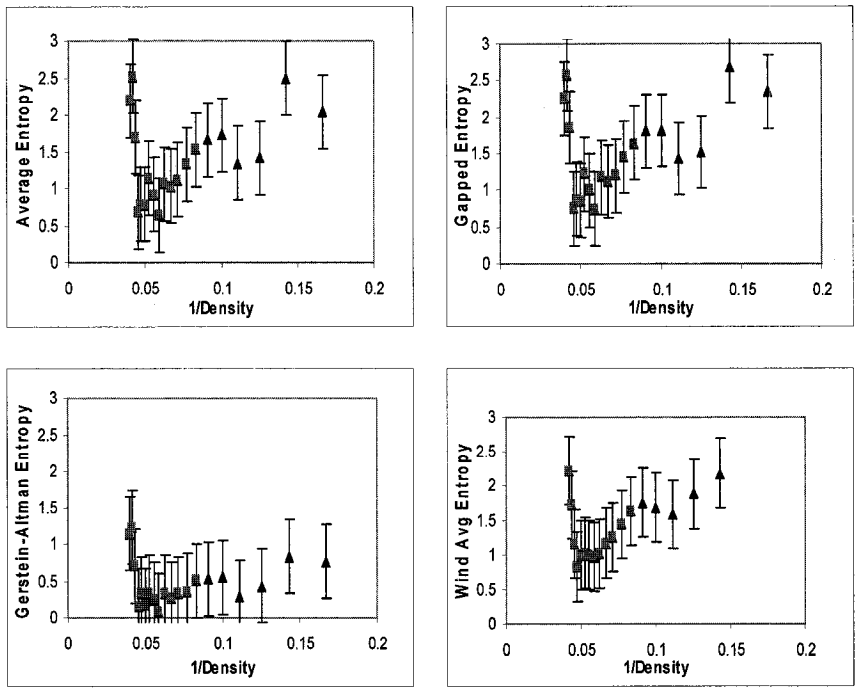

Figure A103. Various correlation plots for protein 3PGK

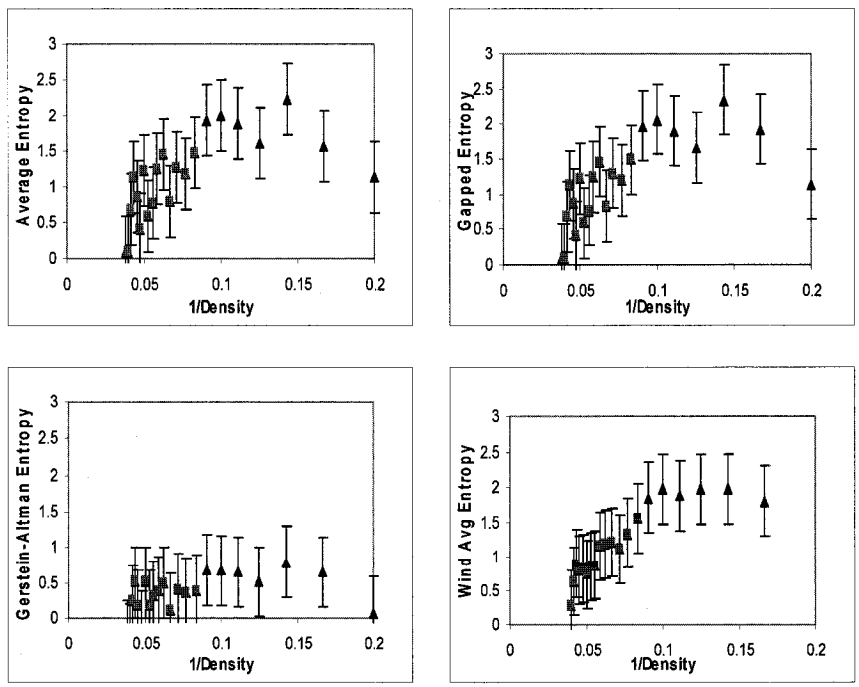

Figure A104. Various correlation plots for protein 3PGM

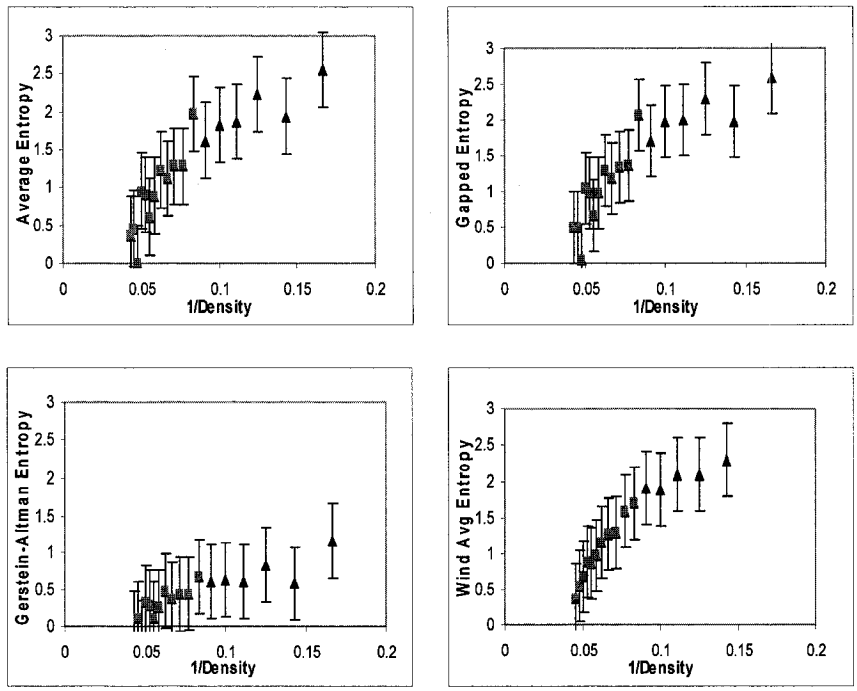

Figure A105. Various correlation plots for protein 3PSG

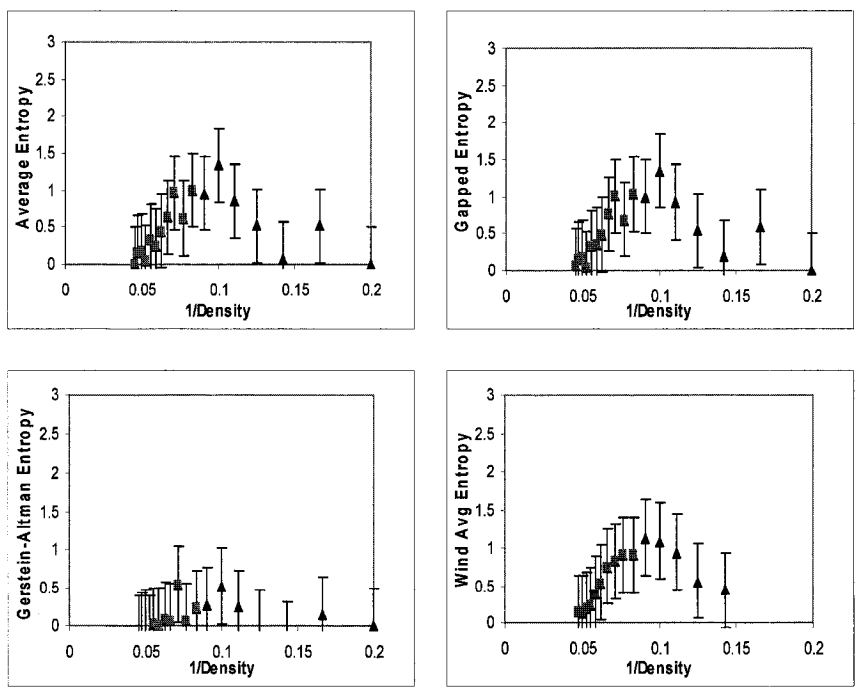

Figure A106. Various correlation plots for protein 3RN3

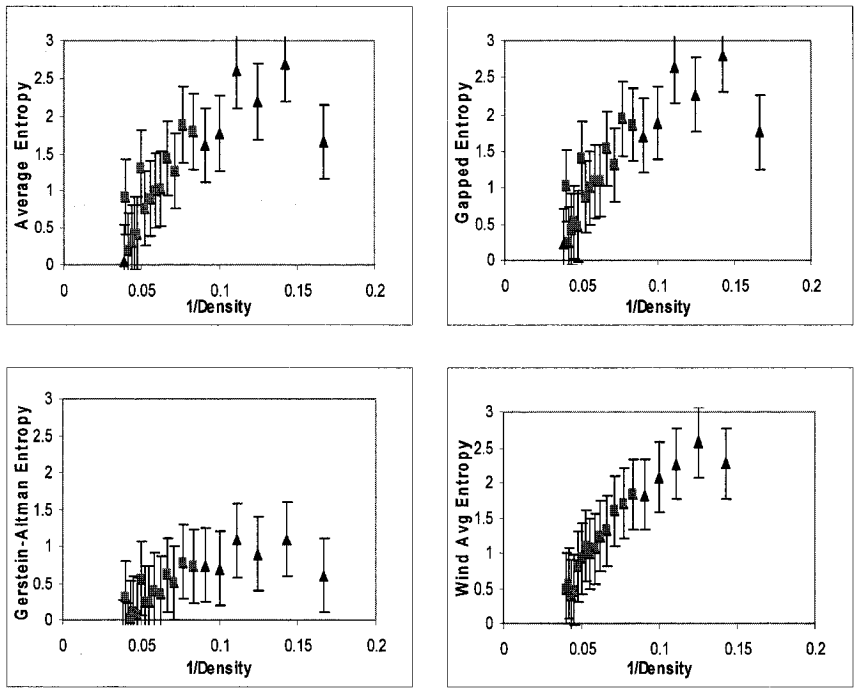

Figure A107. Various correlation plots for protein 3RP2

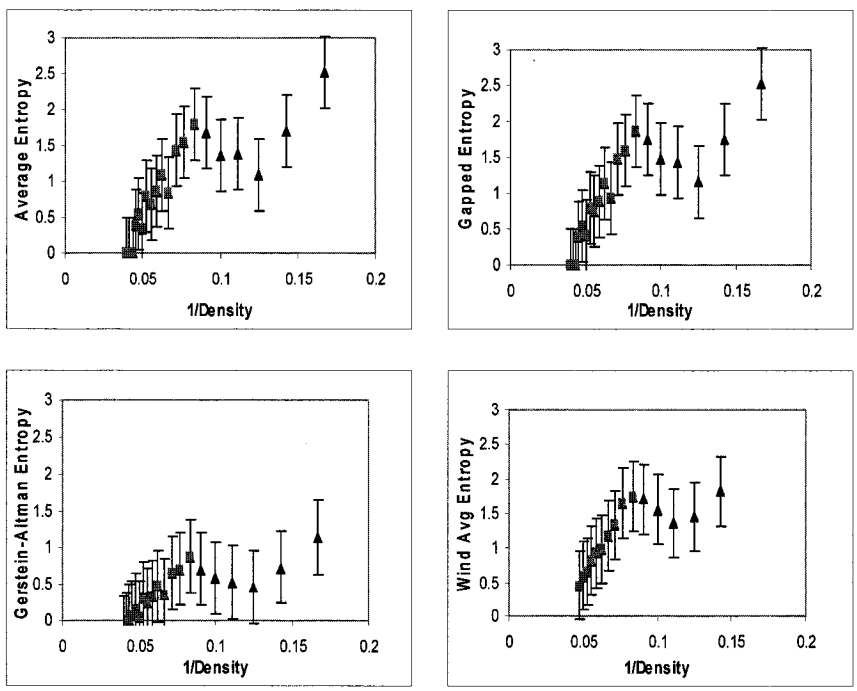

Figure A108. Various correlation plots for protein 4APE

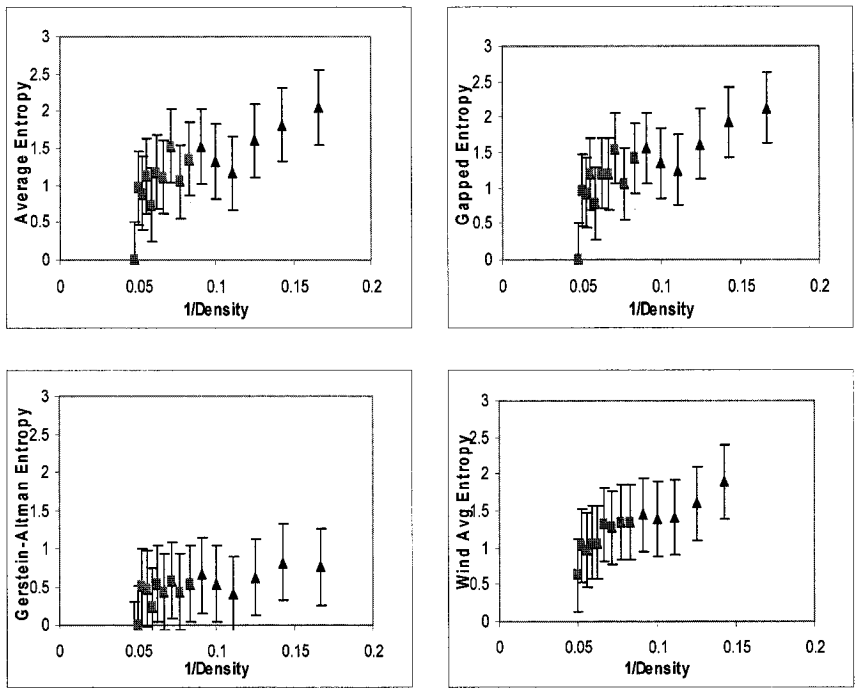

Figure A109. Various correlation plots for protein 4DFR

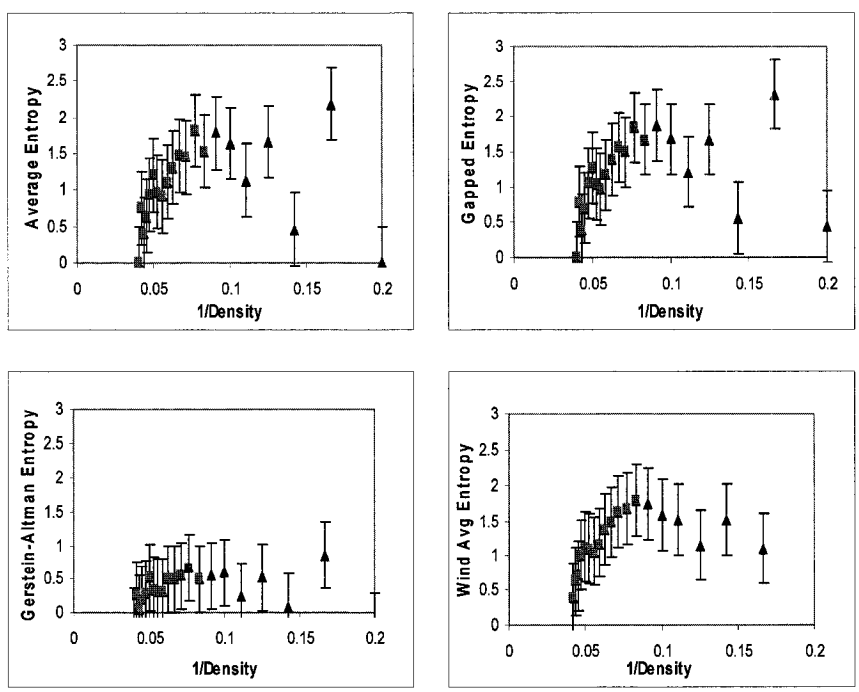

Figure A110. Various correlation plots for protein 4MDH

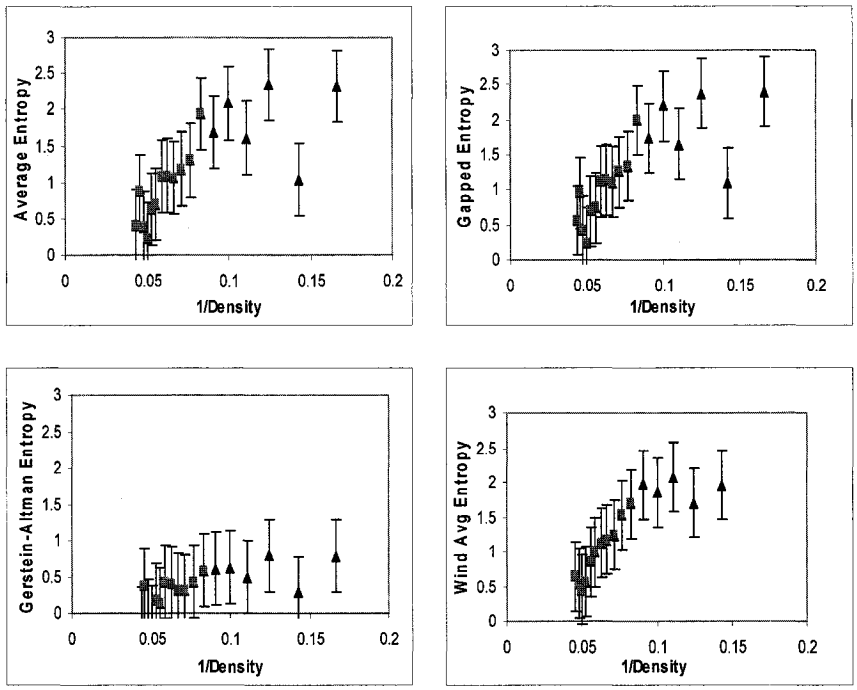

Figure A111. Various correlation plots for protein 4PEP

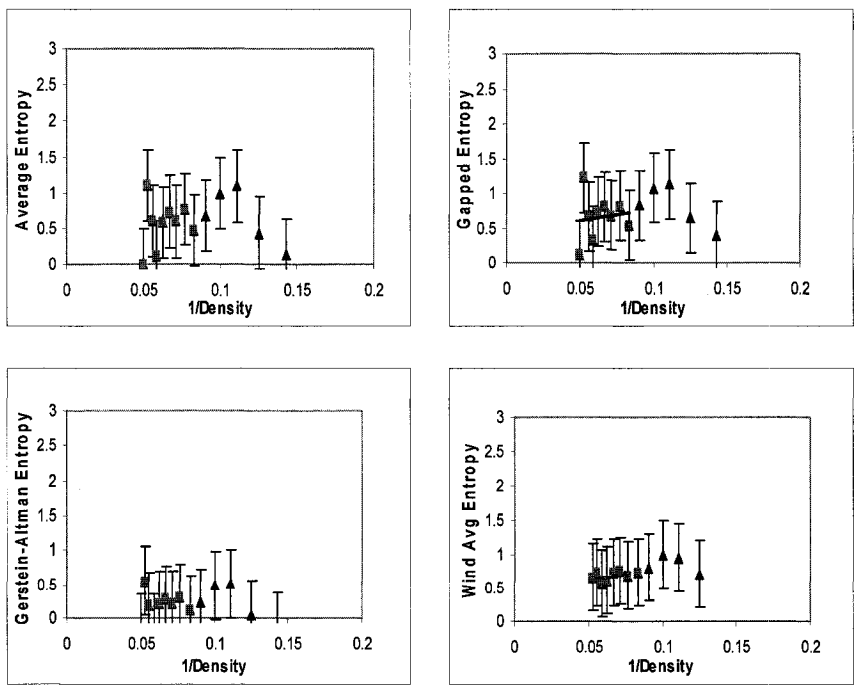

Figure A112. Various correlation plots for protein 4TNC

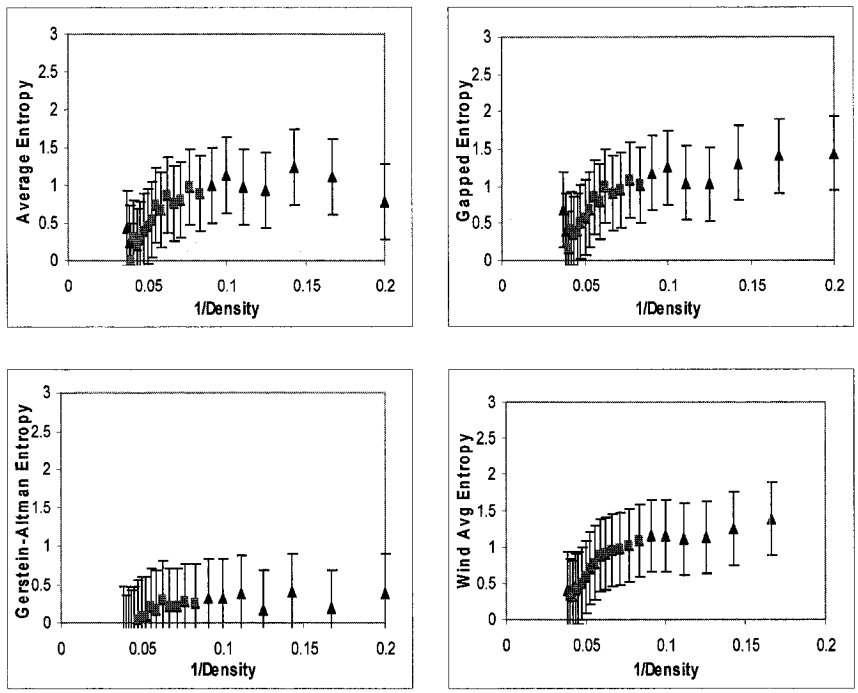

Figure A113. Various correlation plots for protein 5ACN

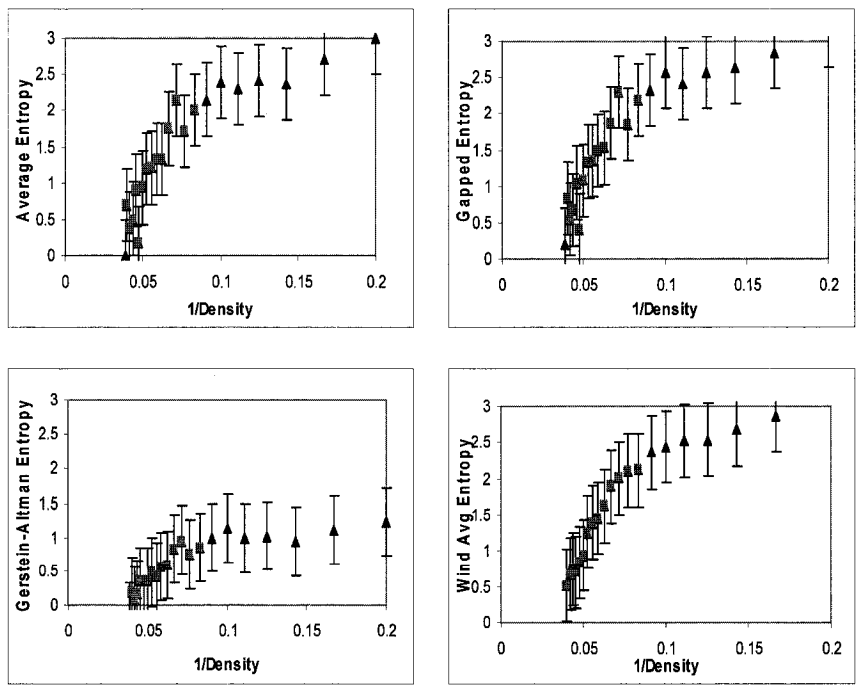

Figure A114. Various correlation plots for protein 5CHA

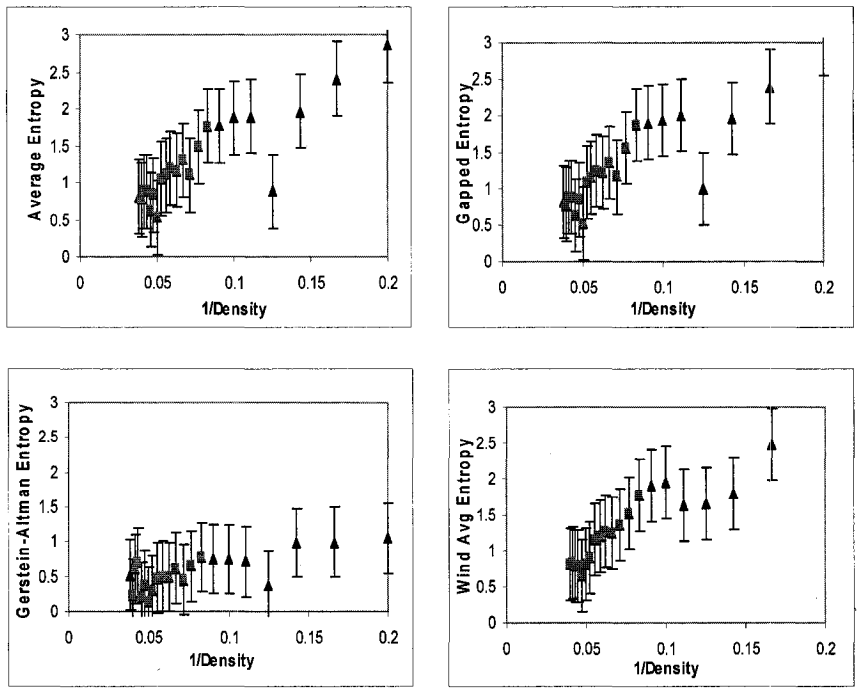

Figure A115. Various correlation plots for protein 5CPA

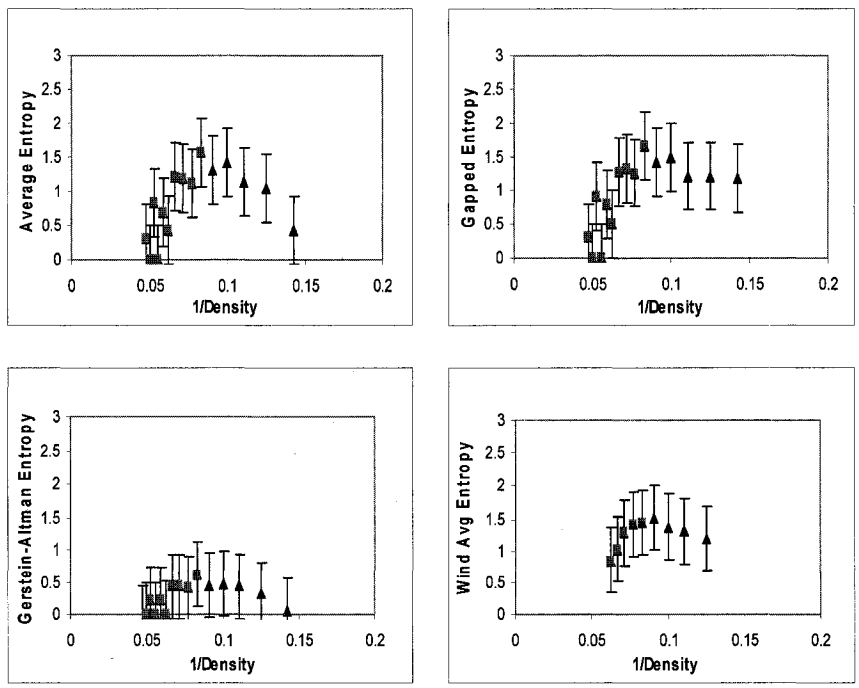

Figure A116. Various correlation plots for protein 5CPV

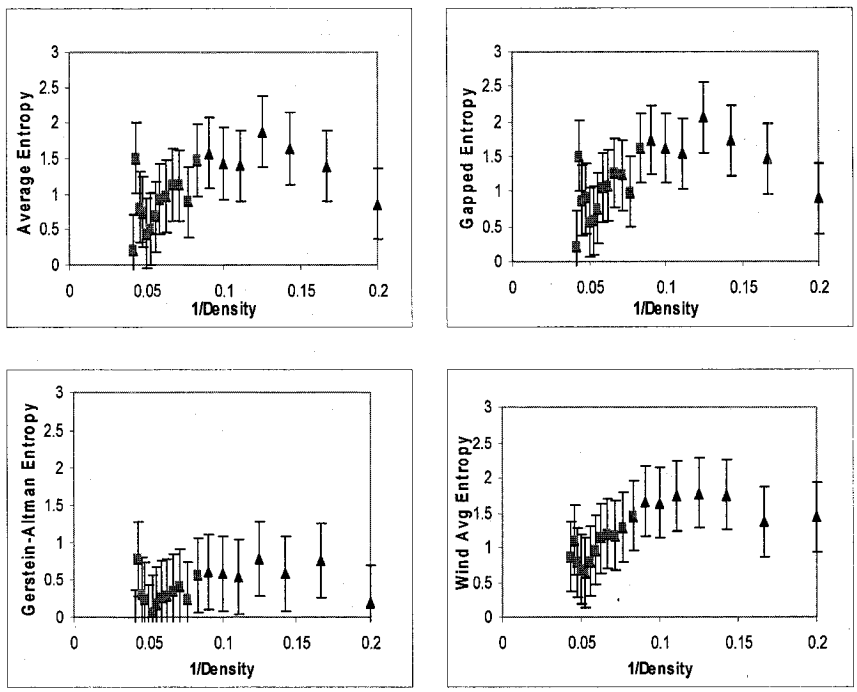

Figure A117. Various correlation plots for protein 5CTS

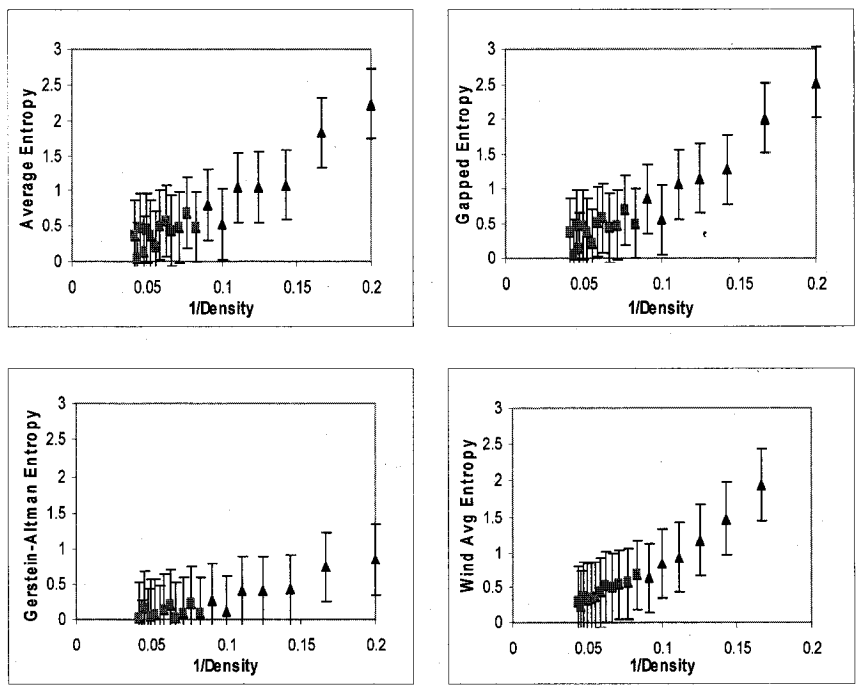

Figure A118. Various correlation plots for protein 5LDH

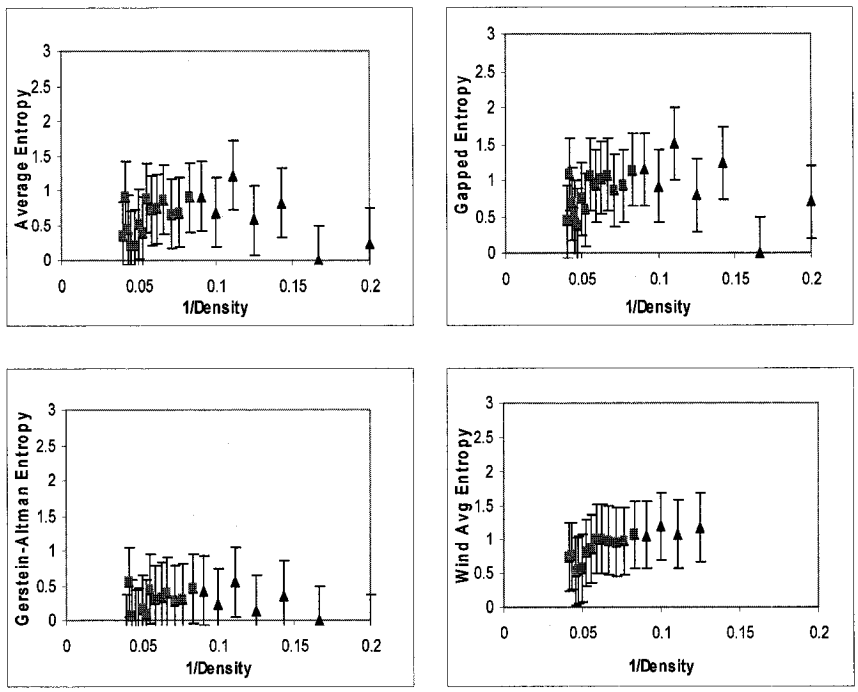

Figure A119. Various correlation plots for protein 5RUB

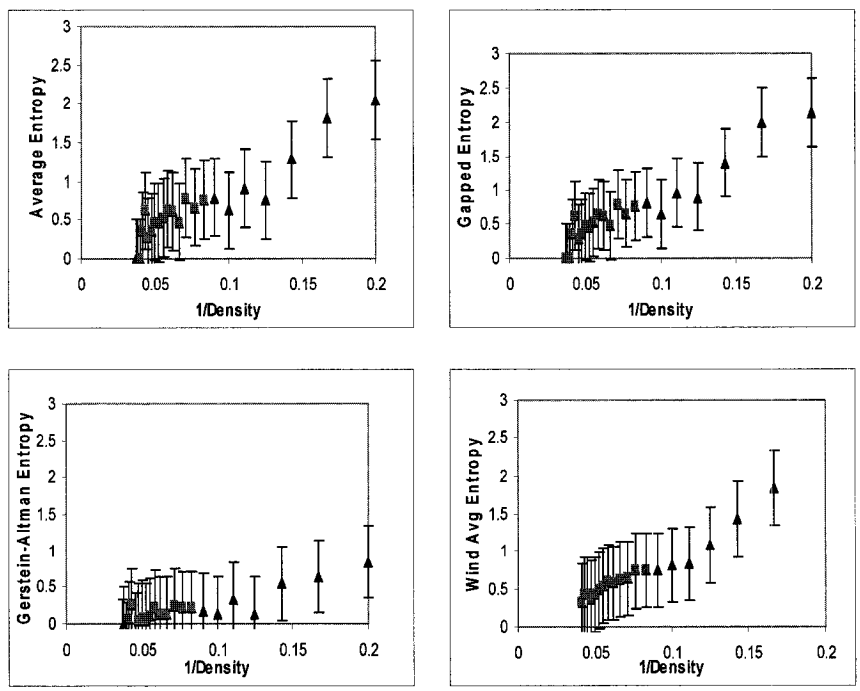

Figure A120. Various correlation plots for protein 6LDH

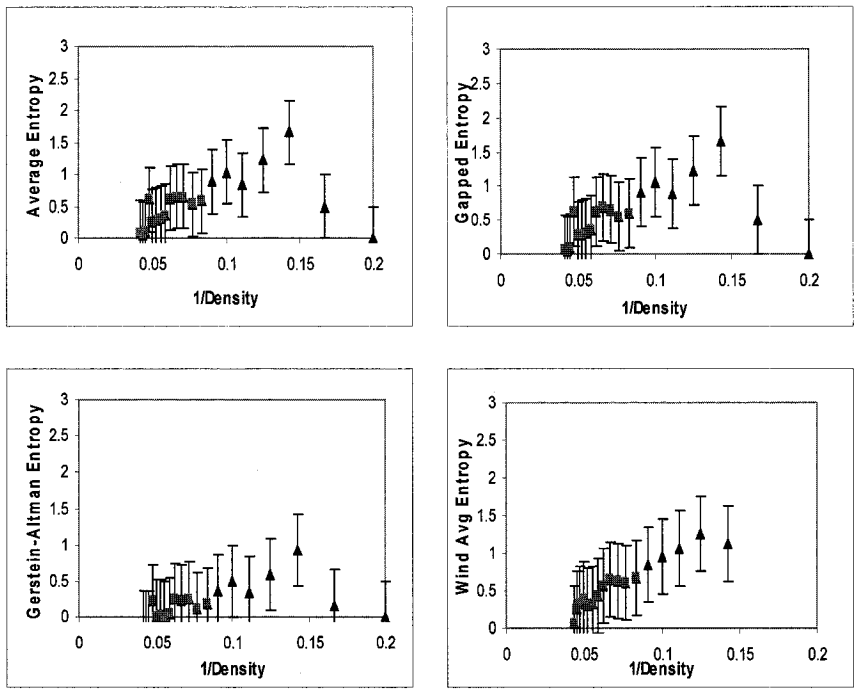

Figure A121. Various correlation plots for protein 6XIA

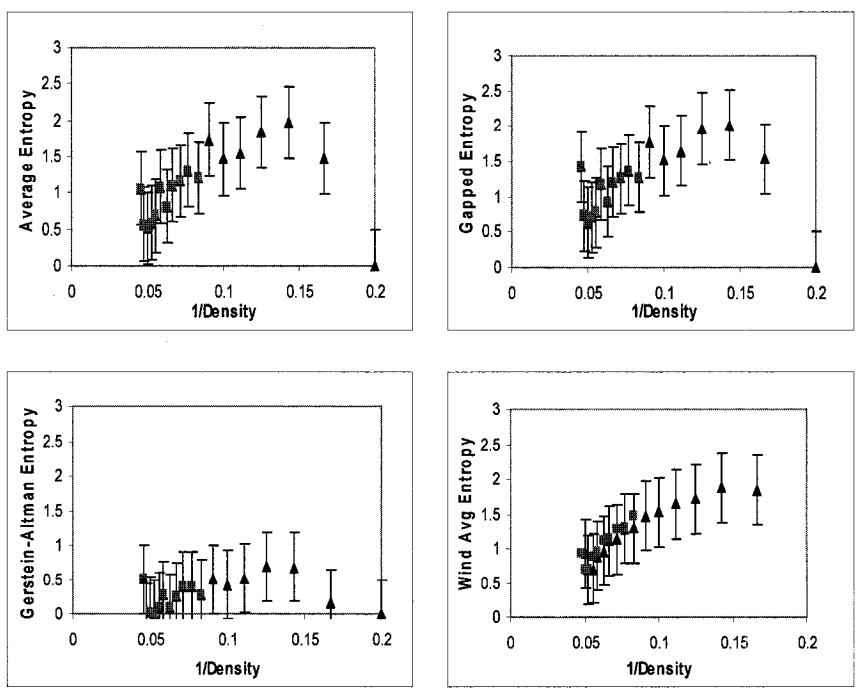

Figure A122. Various correlation plots for protein 7API

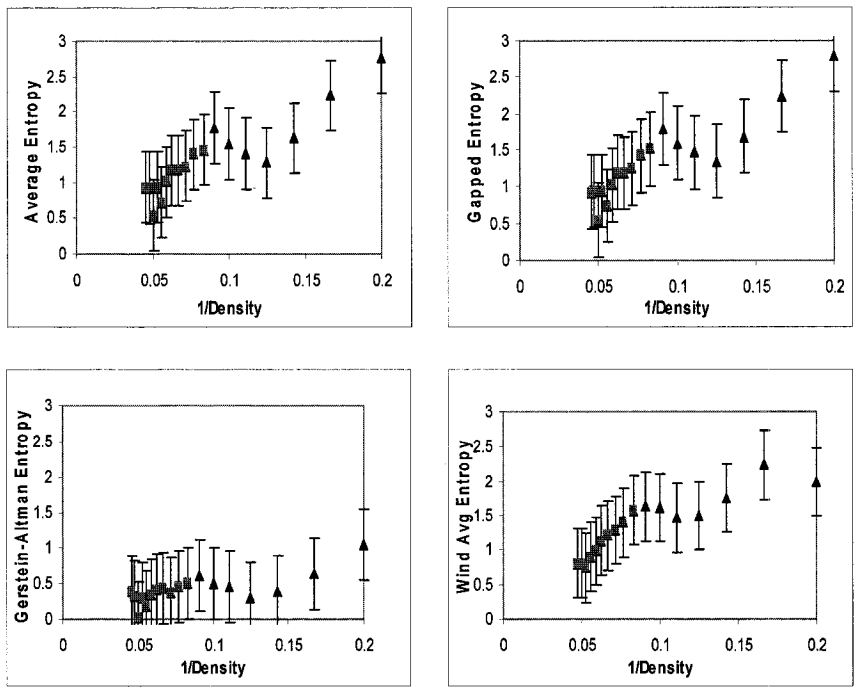

Figure A123. Various correlation plots for protein 7CAT

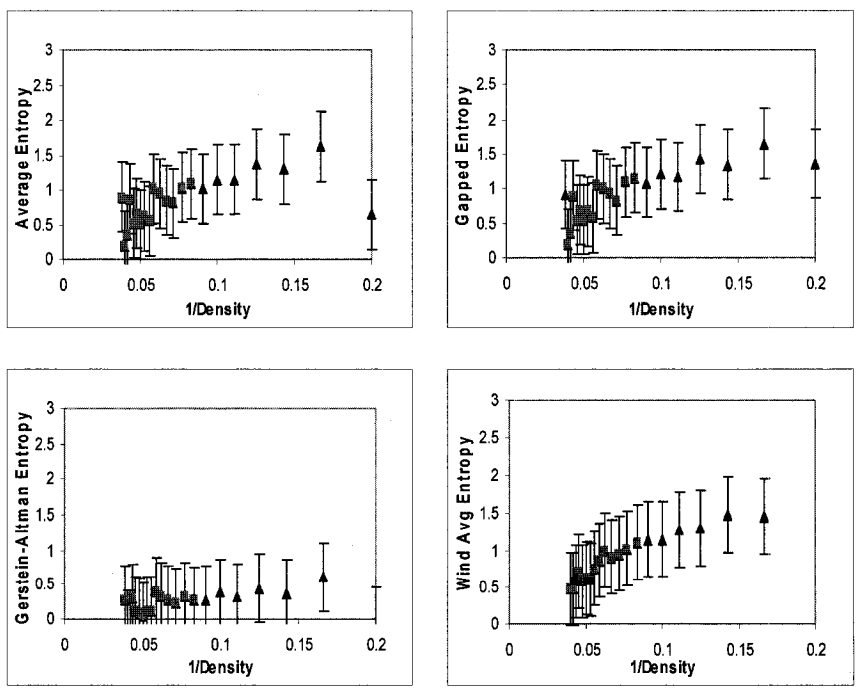

Figure A124. Various correlation plots for protein 8ADH

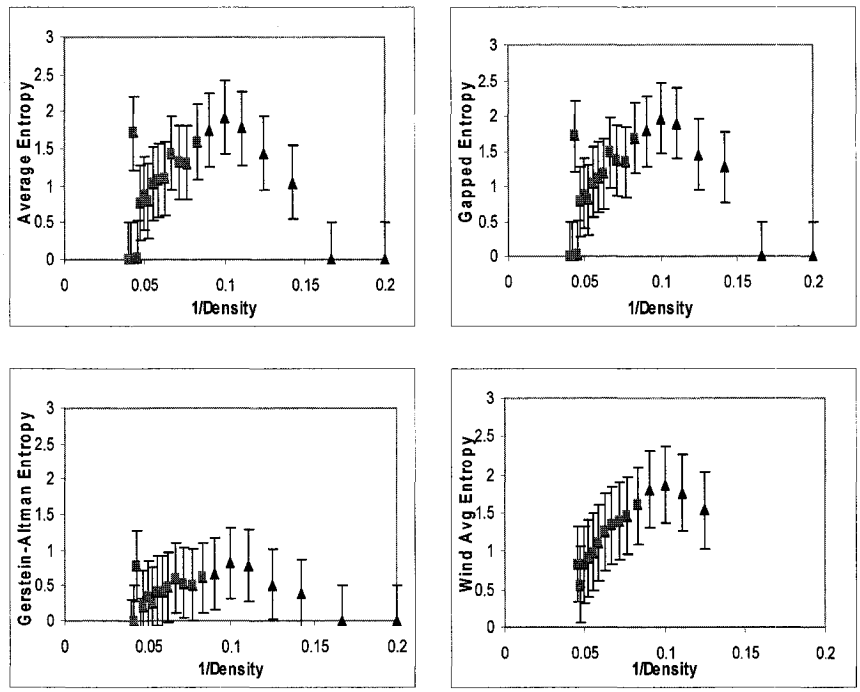

Figure A125. Various correlation plots for protein 8ATC

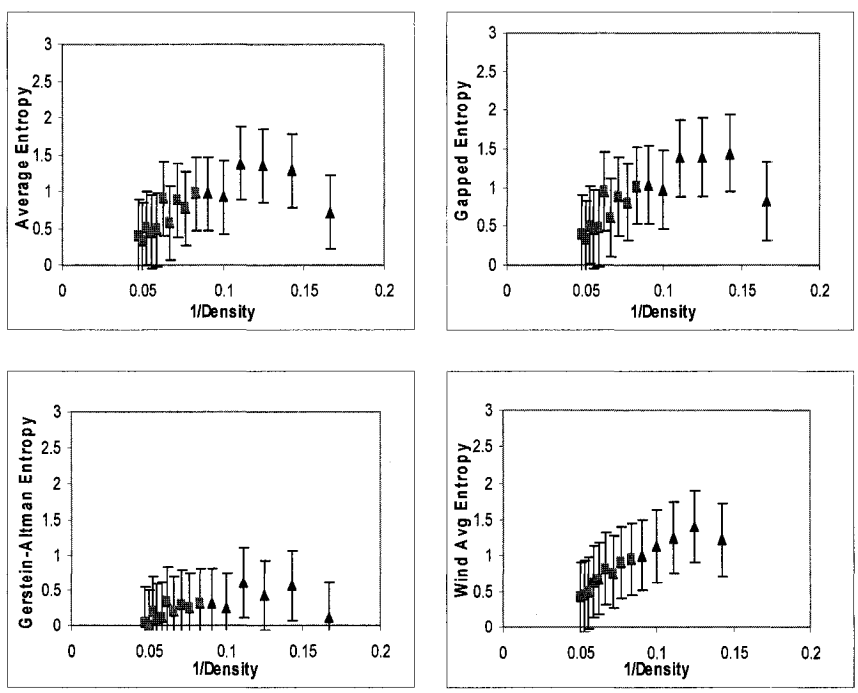

Figure A126. Various correlation plots for protein 8DFR

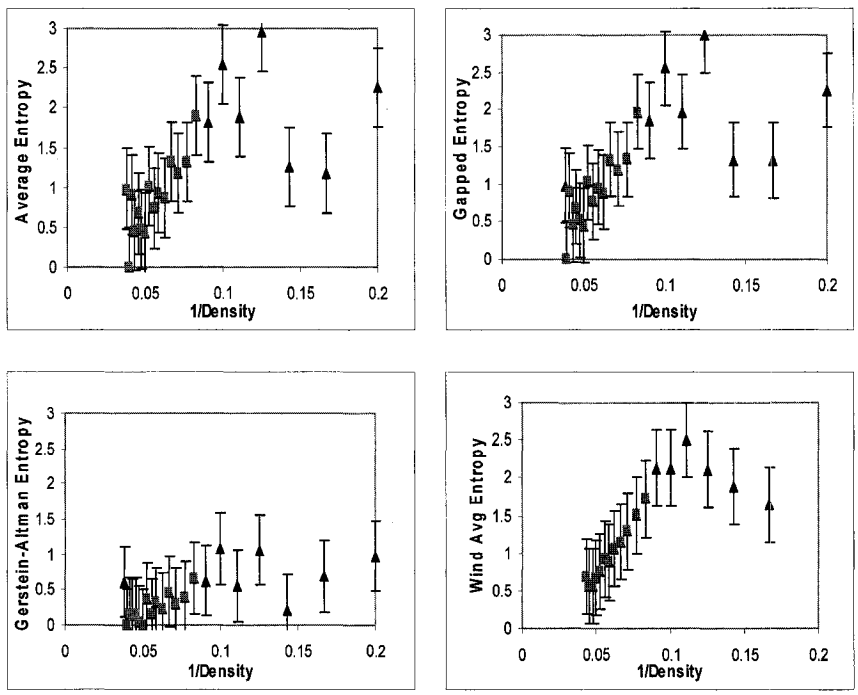

Figure A127. Various correlation plots for protein 9PAP

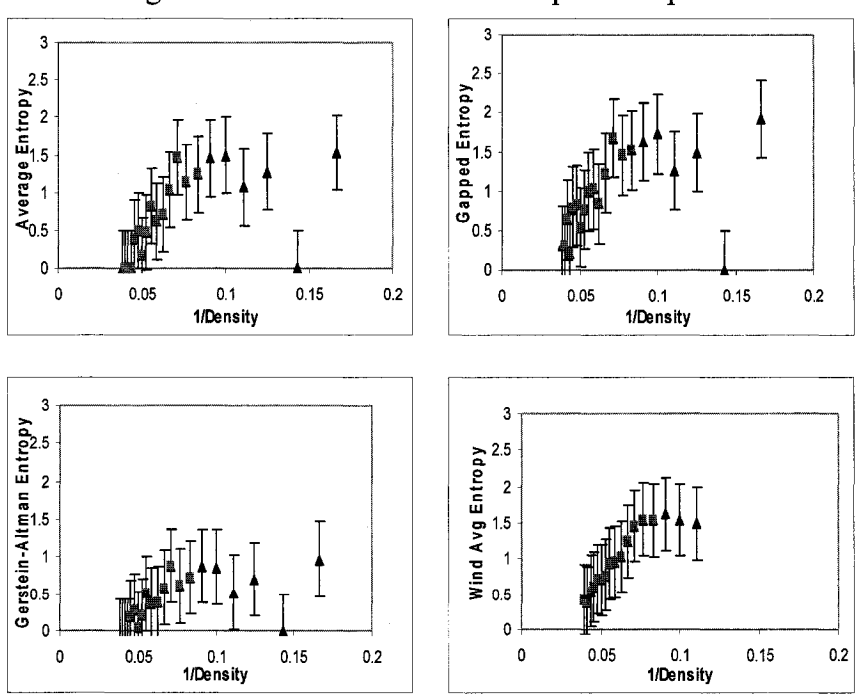

Figure A128. Various correlation plots for protein 9WGA

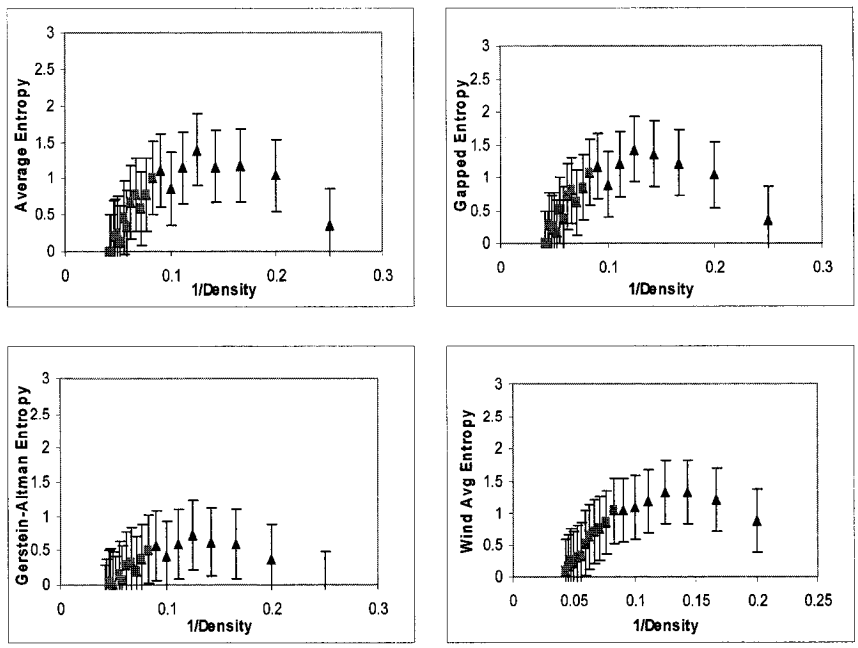

Figure A129. Various correlation plots for protein 1BIT

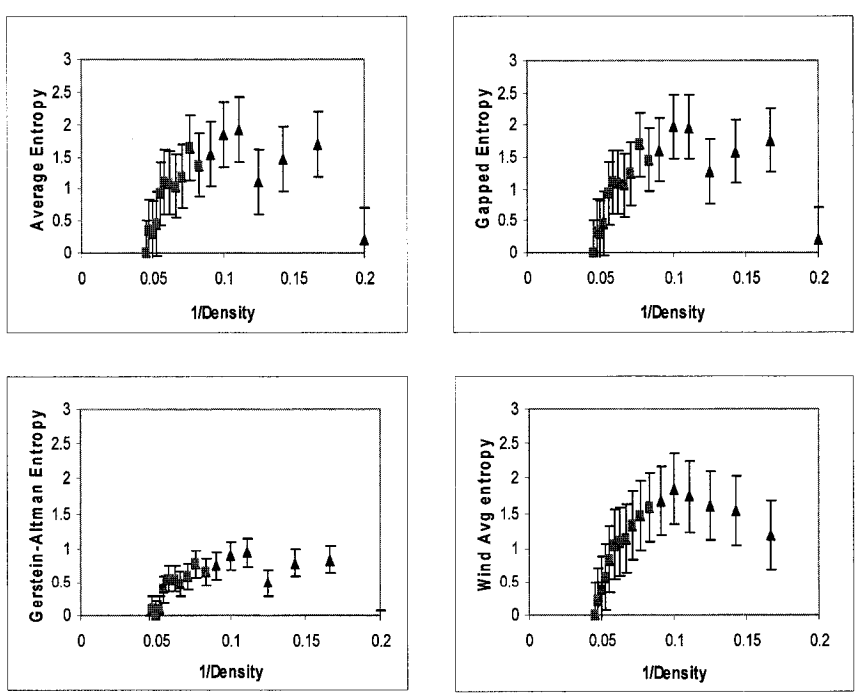

Figure A130. Various correlation plots for protein 1A48

APPENDIX B- Various Data tables for 130 Proteins.

Table B1 shows the residue count and estimate of standard deviation of gapped entropy for overall density region and major I region respectively for 130 proteins.

Tables B2 and B3 show linear correlation data for overall density region and major region I of 130 proteins respectively. In these tables, following notations are used. AE – average entropy, GE – gapped entropy, WE – window average entropy, and GAE – Gerstein-Altman entropy. R - r-square value, S – slope, I – intercept, S.D. – standard deviation from linear fit of the individual plots plotted between AE, GE, WE, GAE against density.

Table B1. Standard deviation data of gapped entropy for major region I and overall density regions.

Protein (PDB) ID	Total Query Residue count	Major Region I Residue Count	Estimated S.D of Gapped Entropy for all regions	Estimated S.D of Gapped Entropy for Major region1
1A1I	85	55	0.110	0.120
1A1S	313	243	0.357	0.364
1A32	85	31	0.368	0.298
1A3C	166	112	0.522	0.404
1A3S	158	112	0.449	0.385
1A48	298	208	0.609	0.532
1A59	377	268	0.330	0.314
1A5Z	312	245	0.454	0.315
1A6F	107	70	0.287	0.334
1A6Q	363	279	0.350	0.269
1AAT	411	317	0.540	0.496
1AB4	477	308	0.632	0.396
1ACB	241	179	0.662	0.543
1ADD	349	278	0.268	0.209
1ADI	431	322	0.581	0.475
1AE4	324	266	0.391	0.271
1AF3	145	88	0.351	0.208
1AGM	470	384	0.428	0.366
1AGX	331	260	0.521	0.403
1AHA	246	189	0.427	0.303
1AHN	169	130	0.590	0.536
1AI2	414	326	0.382	0.333
1AK2	219	143	0.490	0.421
1AKO	268	194	0.518	0.363
1AL8	344	259	0.422	0.397
1ALC	122	86	0.470	0.435
1ALN	294	216	0.394	0.298
1AMN	526	431	0.504	0.401
1AMP	291	242	0.336	0.270
1AN9	340	240	0.275	0.157
1ANG	123	78	0.379	0.289
1AO5	236	171	0.624	0.501
1AOB	264	179	0.408	0.314
1AQ0	306	237	0.637	0.418
1AQH	448	367	0.564	0.420
1ATP	336	238	0.568	0.294

Protein (PDB)ID	Total Query Residue count	Major Region I Residue Count	Estimated S.D of Gapped Entropy for all regions	Estimated S.D of Gapped Entropy for Major region1
1AV5	113	64	0.540	0.485
1AV6	289	204	0.255	0.158
1AV7	273	218	0.526	0.437
1AW5	321	228	0.257	0.193
1AW9	216	155	0.347	0.324
1AYE	401	310	0.382	0.326
1AYL	532	406	0.648	0.570
1AYX	492	386	0.183	0.124
1AZI	153	97	0.219	0.183
1BA3	539	405	0.303	0.156
1BC2	206	156	0.227	0.217
1BF2	750	632	0.201	0.154
1BFD	523	418	0.386	0.235
1BG0	356	267	0.583	0.476
1BG3	901	711	0.352	0.216
1BIA	292	198	0.291	0.175
1BIT	222	143	0.490	0.421
1BLZ	327	236	0.411	0.249
1BN6	291	216	0.281	0.242
1BO6	282	195	0.620	0.329
1BOH	292	229	0.328	0.172
1BSI	495	401	0.421	0.291
1BT3	336	257	0.584	0.427
1BUL	265	216	0.453	0.239
1BXQ	323	260	0.613	0.555
1BYT	836	173	0.431	0.322
1CB0	268	204	0.514	0.530
1CEX	197	146	0.226	0.183
1CJX	352	270	0.589	0.389
1CK6	336	272	0.548	0.460
1CRC	103	76	0.301	0.297
1CRM	256	199	0.455	0.248
1CRZ	397	296	0.320	0.304
1CSR	435	300	0.554	0.335
1D6M	603	434	0.457	0.342
1DAJ	206	127	0.413	0.249
1DCS	282	199	0.319	0.268
1DHS	344	243	0.462	0.333
1DHT	284	207	0.296	0.284

Protein (PDB)ID	Total Query Residue count	Major Region I Residue Count	Estimated S.D of Gapped Entropy for all regions	Estimated S.D of Gapped Entropy for Major region1
1DIN	233	176	0.343	0.241
1DMR	779	657	0.548	0.334
1E1K	455	351	0.339	0.240
1E3H	577	411	0.321	0.190
1E3Q	533	418	0.553	0.388
1E5M	411	316	0.538	0.500
1EBV	550	430	0.401	0.344
1EEH	431	333	0.377	0.322
1HGU	186	101	0.248	0.227
1LZ1	130	93	0.031	0.030
1OMD	107	77	0.462	0.440
1RBP	174	121	0.320	0.201
1RHD	293	214	0.264	0.197
1TON	227	171	0.688	0.510
2ACT	218	172	0.718	0.430
2CTS	437	301	0.436	0.313
2LBP	346	270	0.245	0.231
2LDX	331	244	0.554	0.241
2LIV	344	275	0.310	0.289
2PRK	279	217	0.628	0.563
2RN2	155	97	0.620	0.487
2TAA	478	375	0.290	0.256
3BLM	257	200	0.404	0.339
3CLA	213	140	0.403	0.424
3CNA	237	179	0.244	0.109
3EST	240	172	0.649	0.469
3GBP	305	237	0.411	0.307
3GRS	461	346	0.583	0.392
3PFK	319	250	0.681	0.605
3PGK	415	325	0.597	0.562
3PGM	230	160	0.625	0.414
3PSG	365	284	0.697	0.525
3RN3	124	80	0.394	0.361
3RP2	224	160	0.751	0.525
4APE	330	268	0.581	0.488
4DFR	158	97	0.490	0.436
4MDH	333	260	0.576	0.506
4PEP	326	254	0.656	0.473
4TNC	160	83	0.311	0.317

Protein (PDB) ID	Total Query Residue count	Major Region I Residue Count	Estimated S.D of Gapped Entropy for all regions	Estimated S.D of Gapped Entropy for Major region1
5ACN	753	609	0.361	0.290
5CHA	236	173	0.860	0.597
5CPA	307	241	0.639	0.364
5CPV	108	71	0.319	0.413
5CTS	429	293	0.482	0.389
5LDH	333	218	0.616	0.178
5RUB	436	347	0.290	0.258
6LDH	329	233	0.520	0.210
6XIA	387	267	0.415	0.235
7API	339	233	0.432	0.294
7CAT	498	331	0.539	0.288
8ADH	374	291	0.366	0.291
8ATC	310	232	0.488	0.467
8DFR	186	119	0.361	0.249
9PAP	212	168	0.717	0.426
9WGA	170	130	0.488	0.438

Table B2. Linear correlation data for overall density region of 130 proteins.

PDB ID	AE-R	AE-S	AE-I	GE-R	GE-S	GE-I	WE-R	WE-S	WE-I	GAE-R	GAE-S	GAE-I
1A1I	0.009	-0.572	0.551	0.086	0.379	0.007	0.038	0.150	0.032	0.080	-0.397	-0.092
1A1S	0.078	-4.318	1.770	0.239	8.623	0.984	0.468	8.444	1.112	0.125	2.228	0.304
1A32	0.619	11.497	-0.089	0.658	12.316	-0.146	0.618	9.749	0.222	0.619	5.331	-0.137
1A3C	0.190	7.565	0.518	0.054	3.396	0.963	0.249	6.094	0.810	0.094	2.206	0.287
1A3S	0.005	-0.658	0.831	0.561	9.477	0.308	0.758	9.235	0.324	0.545	6.405	-0.219
1A48	0.027	1.662	0.897	0.030	1.855	0.922	0.414	10.140	0.292	0.050	1.381	0.343
1A59	0.008	-0.524	0.696	0.000	-0.093	0.691	0.350	4.265	0.399	0.081	1.493	0.233
1A5Z	0.125	-3.076	1.428	0.039	-2.024	1.360	0.001	0.319	1.232	0.005	0.385	0.478
1A6F	0.162	7.151	0.454	0.031	2.148	1.139	0.087	1.618	1.220	0.173	3.848	0.344
1A6Q	0.093	2.028	0.401	0.647	6.378	0.332	0.773	7.712	0.252	0.483	3.705	-0.088
1AAT	0.430	7.559	0.589	0.520	8.814	0.544	0.718	10.895	0.415	0.391	3.875	0.124
1AB4	0.788	12.501	0.094	0.788	12.659	0.104	0.834	11.471	0.234	0.767	6.060	-0.157
1ACB	0.712	15.814	0.167	0.742	16.164	0.345	0.804	19.018	0.184	0.701	8.369	-0.030
1ADD	0.058	1.212	0.099	0.449	4.056	0.185	0.830	4.826	0.149	0.075	1.136	-0.095
1ADI	0.005	0.954	1.169	0.002	0.561	1.249	0.163	5.077	0.969	0.051	-1.449	0.460
1AE4	0.577	7.015	0.151	0.415	5.123	0.632	0.360	2.798	0.814	0.145	1.309	0.126
1AF3	0.577	7.015	0.151	0.020	1.241	1.132	0.331	3.225	0.990	0.007	-0.372	0.538
1AGM	0.225	4.084	0.387	0.394	7.492	0.577	0.478	9.344	0.456	0.442	2.097	0.109
1AGX	0.529	10.137	0.074	0.544	10.622	0.058	0.729	13.550	-0.106	0.508	6.385	-0.109
1AHA	0.771	13.242	-0.247	0.759	12.692	-0.150	0.869	13.973	-0.214	0.808	9.500	-0.373
1AHN	0.092	4.992	0.830	0.185	7.031	0.743	0.647	13.840	0.275	0.017	1.505	0.438
1AI2	0.002	0.405	0.858	0.338	6.143	0.551	0.784	9.753	0.294	0.001	0.126	0.174
1AK2	0.032	-2.647	1.514	0.007	-1.448	1.542	0.027	0.923	1.440	0.148	3.345	0.229
1AKO	0.807	12.659	-0.214	0.849	13.253	-0.241	0.882	14.202	-0.277	0.818	7.609	-0.272
1AL8	0.045	1.678	1.131	0.147	2.890	1.214	0.349	5.460	1.032	0.011	0.449	0.469
1ALC	0.011	-0.848	0.804	0.505	9.327	0.258	0.748	10.937	0.178	0.337	4.110	-0.054
1ALN	0.483	7.826	0.167	0.666	15.614	0.126	0.783	14.521	0.182	0.570	6.999	0.000
1AMN	0.281	6.858	0.478	0.448	7.622	0.521	0.765	11.004	0.310	0.214	2.875	0.207
1AMP	0.018	0.990	0.387	0.373	4.644	0.175	0.529	5.823	0.122	0.012	0.628	0.129
1AN9	0.078	1.757	0.150	0.048	1.357	0.536	0.526	3.918	0.369	0.127	1.625	-0.073
1ANG	0.285	4.707	0.122	0.552	7.135	0.055	0.674	7.585	0.101	0.201	1.972	-0.017
1AO5	0.706	14.504	0.000	0.729	14.790	0.140	0.815	16.604	0.025	0.592	6.472	-0.161
1AOB	0.014	0.896	0.750	0.025	1.215	0.748	0.725	10.750	0.061	0.078	1.014	0.129
1AQ0	0.832	19.332	-0.255	0.822	19.442	-0.242	0.915	17.640	0.023	0.861	8.491	-0.253
1AQH	0.672	10.205	0.141	0.661	10.362	0.152	0.814	12.348	0.037	0.735	5.437	-0.114
1ATP	0.571	9.543	0.201	0.616	11.133	0.205	0.908	12.125	0.104	0.245	3.545	0.017
1AV5	0.547	14.047	0.378	0.546	13.906	0.435	0.885	15.543	0.337	0.351	5.109	0.088
1AV6	0.675	5.720	0.059	0.733	6.208	0.119	0.838	7.605	0.020	0.690	3.607	-0.120
1AV7	0.684	9.215	0.346	0.686	9.942	0.396	0.779	11.358	0.315	0.343	2.662	-0.022
1AW5	0.009	-0.368	0.770	0.485	3.183	0.591	0.755	4.418	0.489	0.007	-0.233	0.373
1AW9	0.001	-0.198	1.576	0.011	-1.020	1.689	0.197	3.014	1.389	0.159	-2.433	0.750
1AYE	0.058	1.979	0.557	0.263	4.426	0.535	0.545	7.359	0.342	0.058	1.234	0.211
1AYL	0.530	12.315	0.118	0.527	13.067	0.152	0.686	16.572	-0.024	0.349	5.121	-0.001

PDB ID	AE-R	AE-S	AE-I	GE-R	GE-S	GE-I	WE-R	WE-S	WE-I	GAE-R	GAE-S	GAE-I
1AYX	0.683	3.278	0.002	0.784	4.488	-0.062	0.937	4.839	-0.075	0.571	2.232	-0.153
1AZI	0.555	5.801	-0.186	0.555	5.801	-0.186	0.575	3.616	0.007	0.379	3.141	-0.271
1BA3	0.483	4.000	0.509	0.610	5.336	0.569	0.779	4.930	0.591	0.504	2.791	0.122
1BC2	0.005	-0.401	0.806	0.001	-0.211	0.866	0.670	6.942	0.390	0.000	-0.002	0.410
1BF2	0.338	3.286	0.137	0.276	2.934	0.223	0.634	4.774	0.113	0.169	1.461	0.036
1BFD	0.507	4.996	0.224	0.730	7.477	0.123	0.815	7.944	0.119	0.511	3.931	0.000
1BG0	0.651	13.078	0.111	0.674	13.313	0.168	0.811	15.122	0.042	0.634	6.496	-0.106
1BG3	0.660	5.840	0.038	0.794	7.101	0.277	0.797	7.102	0.292	0.723	4.532	-0.193
1BIA	0.000	-0.045	0.579	0.000	-0.045	0.579	0.054	1.139	0.531	0.003	-0.190	0.274
1BIT	0.807	12.659	-0.214	0.849	13.253	-0.241	0.882	14.202	-0.277	0.818	7.609	-0.272
1BLZ	0.487	7.791	0.166	0.417	7.483	0.292	0.778	11.281	0.022	0.497	5.147	-0.048
1BN6	0.001	-0.272	0.605	0.249	3.907	0.359	0.563	6.219	0.179	0.008	-0.537	0.332
1BO6	0.813	14.350	-0.069	0.853	16.138	-0.174	0.905	15.561	-0.102	0.772	8.542	-0.234
1BOH	0.062	1.504	0.340	0.211	3.402	0.236	0.826	6.125	0.065	0.026	0.749	0.104
1BSI	0.561	8.369	0.102	0.692	9.733	0.128	0.918	11.851	0.011	0.547	5.083	-0.135
1BT3	0.732	12.215	0.202	0.709	12.085	0.257	0.827	13.438	0.240	0.724	6.046	0.055
1BUL	0.400	8.082	0.386	0.441	8.354	0.391	0.846	11.495	0.190	0.336	3.928	0.063
1BXQ	0.314	11.078	0.114	0.312	11.433	0.122	0.623	17.412	-0.202	0.226	5.220	-0.029
1BYT	0.222	6.146	0.826	0.223	5.638	1.132	0.364	3.243	1.300	0.152	2.400	0.280
1CB0	0.135	5.043	0.719	0.194	6.297	0.835	0.632	7.607	0.784	0.120	3.078	0.334
1CEX	0.014	-0.547	0.385	0.259	2.799	0.169	0.769	6.474	-0.040	0.004	0.118	-0.046
1CJX	0.650	12.579	0.350	0.690	13.721	0.289	0.913	15.337	0.122	0.589	7.025	0.118
1CK6	0.646	11.800	0.169	0.653	12.283	0.177	0.876	12.630	0.158	0.375	4.661	0.054
1CRC	0.491	8.762	0.049	0.462	8.523	0.085	0.728	8.781	0.099	0.536	6.831	-0.443
1CRM	0.764	11.142	-0.186	0.763	11.213	-0.178	0.915	11.231	-0.165	0.743	5.229	-0.190
1CRZ	0.061	1.741	0.761	0.136	2.671	0.717	0.503	4.986	0.581	0.005	0.327	0.389
1CSR	0.098	-2.466	1.234	0.019	-1.067	1.232	0.529	8.332	0.533	0.079	-1.127	0.437
1D6M	0.254	3.908	0.288	0.236	3.949	0.324	0.563	7.031	0.121	0.167	2.060	0.068
1DAJ	0.692	9.191	-0.264	0.715	9.912	-0.198	0.866	10.701	-0.225	0.688	7.106	-0.366
1DCS	0.094	-1.518	0.913	0.028	0.793	0.736	0.722	6.075	0.338	0.102	-0.996	0.467
1DHS	0.260	3.618	0.623	0.224	3.653	0.836	0.958	8.854	0.489	0.275	2.145	0.222
1DHT	0.007	-0.544	0.449	0.011	0.787	0.669	0.736	7.995	0.155	0.000	0.050	0.098
1DIN	0.180	4.531	0.346	0.423	7.499	0.187	0.835	10.562	-0.006	0.147	3.009	0.082
1DMR	0.561	11.193	0.383	0.557	12.163	0.406	0.902	14.761	0.225	0.350	5.145	0.212
1E1K	0.565	6.607	0.301	0.702	7.908	0.359	0.880	9.120	0.271	0.523	3.970	0.050
1E3H	0.033	1.333	0.733	0.162	2.923	0.659	0.611	5.412	0.516	0.024	0.790	0.332
1E3Q	0.494	8.276	0.446	0.555	9.321	0.394	0.870	13.061	0.144	0.511	4.302	0.097
1E5M	0.356	10.177	0.677	0.385	10.800	0.660	0.546	10.726	0.787	0.301	4.116	0.176
1EBV	0.031	1.560	0.630	0.041	1.836	0.758	0.264	4.093	0.607	0.011	-0.626	0.311
1EEH	0.155	4.032	0.810	0.291	5.634	0.747	0.557	8.419	0.541	0.107	1.964	0.368
1HGU	0.389	3.094	0.393	0.261	2.873	0.435	0.649	3.549	0.414	0.202	2.276	-0.005
1LZ1	0.645	8.476	0.032	0.025	-0.214	0.073	0.113	0.288	0.027	0.001	-0.059	-0.199
1OMD	0.001	-0.445	1.039	0.012	1.763	0.911	0.186	6.571	0.571	0.027	-1.434	0.415
1RBP	0.454	4.639	0.159	0.555	5.389	0.414	0.704	5.303	0.446	0.387	2.422	-0.028
1RHD	0.051	1.639	0.357	0.121	2.606	0.302	0.589	4.458	0.158	0.042	0.933	0.124

PDB ID	AE-R	AE-S	AE-I	GE-R	GE-S	GE-I	WE-R	WE-S	WE-I	GAE-R	GAE-S	GAE-I
1TON	0.195	7.495	0.573	0.374	9.519	0.595	0.802	15.138	0.223	0.098	2.254	0.176
2ACT	0.488	14.365	0.080	0.508	14.543	0.126	0.844	19.514	-0.159	0.432	5.806	-0.111
2CTS	0.102	1.650	0.920	0.246	2.854	0.926	0.279	3.475	0.883	0.226	1.404	0.242
2LBP	0.254	4.818	0.292	0.256	4.814	0.363	0.576	6.332	0.320	0.221	2.405	0.077
2LDX	0.051	2.023	0.490	0.190	4.287	0.351	0.961	12.434	-0.226	0.035	0.900	0.110
2LIV	0.049	1.993	0.597	0.133	3.391	0.522	0.572	7.647	0.267	0.121	1.930	0.145
2PRK	0.236	8.864	0.546	0.244	9.118	0.586	0.880	28.682	-0.476	0.208	3.990	0.105
2RN2	0.406	10.193	0.297	0.502	12.147	0.222	0.798	14.692	0.113	0.190	3.018	0.003
2TAA	0.489	6.988	0.189	0.580	8.156	0.128	0.749	7.344	0.292	0.409	3.770	0.013
3BLM	0.461	7.004	0.513	0.585	8.564	0.561	0.811	10.788	0.426	0.390	4.337	0.121
3CLA	0.212	4.313	0.532	0.280	4.818	0.564	0.433	6.002	0.530	0.058	1.274	0.267
3CNA	0.108	1.720	0.260	0.147	1.674	0.823	0.784	3.537	0.672	0.092	1.259	0.003
3EST	0.628	13.898	-0.059	0.627	14.228	-0.043	0.866	18.442	-0.339	0.548	6.785	-0.132
3GBP	0.841	11.274	-0.157	0.863	12.743	-0.235	0.915	13.100	-0.249	0.841	7.192	-0.280
3GRS	0.605	10.739	0.190	0.645	10.603	0.333	0.795	9.534	0.445	0.491	5.100	-0.014
3PFK	0.678	18.068	0.059	0.705	18.869	0.064	0.807	21.688	-0.077	0.695	9.038	-0.184
3PGK	0.165	6.472	0.908	0.211	7.620	0.934	0.324	8.297	0.807	0.015	1.058	0.399
3PGM	0.365	8.072	0.515	0.411	9.068	0.469	0.767	12.614	0.277	0.183	2.930	0.101
3PSG	0.807	17.414	-0.124	0.791	17.311	-0.033	0.887	19.460	-0.157	0.764	7.698	-0.189
3RN3	0.004	-0.311	0.509	0.005	0.374	0.498	0.210	5.529	0.171	0.001	0.075	0.085
3RP2	0.659	17.071	-0.028	0.666	16.982	0.068	0.893	21.259	-0.186	0.576	8.094	-0.151
4APE	0.703	15.915	-0.215	0.697	16.181	-0.193	0.614	12.176	0.244	0.650	7.813	-0.196
4DFR	0.633	10.604	0.311	0.643	11.075	0.318	0.818	9.650	0.479	0.466	4.993	0.032
4MDH	0.014	1.562	0.982	0.051	2.939	0.951	0.179	4.602	0.888	0.000	0.092	0.350
4PEP	0.588	14.008	0.091	0.585	14.004	0.149	0.769	16.879	-0.037	0.488	5.418	-0.080
4TNC	0.000	-0.091	0.601	0.006	0.847	0.643	0.371	3.319	0.487	0.002	-0.315	0.218
5ACN	0.492	5.434	0.255	0.717	6.923	0.302	0.768	8.098	0.225	0.482	2.514	-0.033
5CHA	0.756	17.062	0.157	0.762	16.990	0.322	0.817	19.214	0.214	0.645	7.552	0.017
5CPA	0.770	12.025	0.357	0.772	12.699	0.351	0.810	12.389	0.381	0.553	4.258	0.232
5CPV	0.141	6.594	0.323	0.353	10.891	0.099	0.683	0.549	0.577	0.101	2.389	0.076
5CTS	0.123	2.060	0.890	0.253	3.226	0.903	0.377	4.024	0.860	0.236	1.747	0.223
5LDH	0.902	11.594	-0.290	0.916	13.309	-0.392	0.965	12.403	-0.328	0.858	5.461	-0.269
5RUB	0.018	-0.917	0.691	0.000	-0.147	0.860	0.674	6.942	0.419	0.018	-0.647	0.280
6LDH	0.870	10.609	-0.172	0.895	11.614	-0.221	0.940	10.249	-0.084	0.777	4.829	-0.198
6XIA	0.163	3.023	0.302	0.280	4.393	0.212	0.916	10.921	-0.209	0.319	2.799	-0.048
7API	0.144	-2.985	1.334	0.057	-1.766	1.337	0.879	10.544	0.348	0.062	-0.810	0.347
7CAT	0.433	6.051	0.744	0.317	5.354	0.822	0.804	8.779	0.550	0.558	2.865	0.160
8ADH	0.354	4.667	0.494	0.659	6.727	0.399	0.879	8.181	0.301	0.126	1.434	0.117
8ATC	0.001	-0.341	1.036	0.000	0.019	1.044	0.725	13.973	0.264	0.004	0.360	0.354
8DFR	0.454	6.440	0.259	0.527	7.398	0.216	0.869	10.307	0.021	0.281	2.581	0.042
9PAP	0.439	11.018	0.346	0.463	11.419	0.344	0.602	13.612	0.242	0.443	4.957	0.022
9WGA	0.314	8.828	0.074	0.238	7.478	0.449	0.830	18.600	-0.170	0.335	5.741	-0.028

Table B3. Linear correlation data for major I density region of 130 proteins.

PDB ID	AE-R	AE-S	AE-I	GE-R	GE-S	GE-I	WE-R	WE-S	WE-I	GAE-R	GAE-S	GAE-I
1A1I	0.769	24.08	-1.02	0.034	0.604	-0.01	0.12	0.753	0.088	0.152	-1.591	-0.021
1A1S	0.507	28.58	-0.17	0.516	29.58	-0.17	0.857	24.18	0.186	0.565	11.428	-0.201
1A32	0.714	34.37	-1.71	0.714	34.37	-1.71	0.989	40.38	-2.13	0.515	12.74	-0.664
1A3C	0.674	34.28	-1.05	0.588	24.97	-0.4	0.863	28.49	-0.59	0.452	12.25	-0.33
1A3S	0.391	19.61	-0.55	0.372	19.76	-0.31	0.85	19.99	-0.36	0.64	15.785	-0.767
1A48	0.824	37.45	-1.42	0.85	39.53	-1.53	0.912	40.39	-1.59	0.774	22.204	-0.995
1A59	0.5	16.02	-0.34	0.532	17.25	-0.4	0.927	15.88	-0.32	0.519	12.502	-0.433
1A5Z	0.839	22	-0.13	0.826	21.58	-0.09	0.939	22.35	-0.11	0.645	11.172	-0.167
1A6F	0.393	27.85	-0.77	0.148	11.33	0.56	0.385	7.025	0.874	0.088	6.219	0.193
1A6Q	0.733	16.06	-0.49	0.736	16.93	-0.28	0.809	16.14	-0.23	0.702	10.541	-0.505
1AAT	0.692	30.06	-0.71	0.698	30.38	-0.71	0.912	25.54	-0.43	0.46	13.847	-0.449
1AB4	0.769	26.43	-0.72	0.776	27.12	-0.74	0.917	24.17	-0.54	0.636	10.179	-0.382
1ACB	0.945	38.76	-1.13	0.949	38.71	-0.93	0.967	37.78	-0.89	0.924	20.9	-0.739
1ADD	0.479	5.597	-0.17	0.314	8.565	-0.07	0.494	6.336	0.069	0.426	4.252	-0.289
1ADI	0.654	28.51	-0.55	0.63	28.37	-0.48	0.922	27	-0.39	0.552	11.548	-0.346
1AE4	0.014	2.438	0.505	0.017	3.133	1.009	0.016	-1.185	1.312	0.017	1.54	0.406
1AF3	0.014	2.438	0.505	0.001	0.486	0.987	0	-0.205	1.028	0.049	-2.717	0.395
1AGM	0.812	24.33	-0.75	0.836	32.73	-0.88	0.965	33.77	-0.93	0.805	16.209	-0.619
1AGX	0.853	26.78	-0.79	0.858	27.36	-0.81	0.916	26.31	-0.76	0.829	17.034	-0.657
1AHA	0.917	19.66	-0.57	0.698	20.41	-0.63	0.917	19.66	-0.57	0.624	13.726	-0.647
1AHN	0.441	25.76	-0.43	0.461	26.64	-0.45	0.814	27.87	-0.52	0.23	13.366	-0.303
1AI2	0.358	13.66	0.009	0.368	14.79	0.002	0.758	15.33	-0.02	0.081	3.134	-0.028
1AK2	0.113	11.48	0.538	0.1	11.19	0.728	0.155	3.963	1.241	0.016	2.551	0.279
1AKO	0.902	24.98	-0.91	0.905	25.33	-0.92	0.955	24.48	-0.86	0.872	15.114	-0.697
1AL8	0.637	23.39	-0.19	0.683	24	-0.06	0.927	28.28	-0.37	0.412	11.335	-0.191
1ALC	0.326	17.15	-0.37	0.502	23.96	-0.63	0.89	22.44	-0.53	0.327	9.903	-0.397
1ALN	0.449	12.71	-0.06	0.702	28.5	-0.58	0.923	28.31	-0.59	0.694	16.152	-0.507
1AMN	0.749	25.24	-0.54	0.76	25.6	-0.55	0.969	25	-0.52	0.645	12.988	-0.399
1AMP	0.718	16.1	-0.44	0.725	16.82	-0.47	0.775	16.55	-0.46	0.742	12.646	-0.535
1AN9	0.778	10.74	-0.4	0.701	9.64	0.028	0.891	9.681	0.023	0.777	8.113	-0.466
1ANG	0.822	23.16	-0.99	0.849	22.4	-0.86	0.978	23.01	-0.91	0.772	11.69	-0.601
1AO5	0.722	31.82	-0.97	0.688	30.44	-0.74	0.915	31.93	-0.85	0.692	16.246	-0.711
1AOB	0.43	16.63	-0.33	0.457	17.11	-0.34	0.925	18.78	-0.47	0.203	7.016	-0.259
1AQ0	0.812	39.31	-1.43	0.815	40.41	-1.47	0.929	30.22	-0.78	0.834	16.047	-0.69
1AQH	0.859	28.6	-0.95	0.865	29.39	-0.98	0.983	26.25	-0.81	0.716	12.015	-0.489
1ATP	0.064	8.931	0.221	0.083	10.93	0.193	0.962	22.53	-0.58	0.007	-2.117	0.354
1AV5	0.312	22.84	-0.18	0.355	24.52	-0.25	0.937	27.73	-0.44	0.181	8.802	-0.132
1AV6	0.738	12.78	-0.42	0.827	12.7	-0.33	0.967	14.04	-0.42	0.704	8.497	-0.447
1AV7	0.735	25.19	-0.53	0.757	26.99	-0.54	0.863	25.29	-0.42	0.586	10.532	-0.453
1AW5	0.213	5.939	0.356	0.229	6.96	0.335	0.685	9.295	0.177	0.271	4.787	0.053
1AW9	0.154	9.561	0.946	0.119	9.004	1.041	0.585	12.55	0.788	0	0.197	0.573
1AYE	0.545	14.58	-0.22	0.702	20.03	-0.4	0.887	21.47	-0.5	0.509	8.951	-0.269

PDB ID	AE-R	AE-S	AE-I	GE-R	GE-S	GE-I	WE-R	WE-S	WE-I	GAE-R	GAE-S	GAE-I
1AYL	0.88	35.77	-1.23	0.885	39.26	-1.35	0.892	35.85	-1.14	0.825	19.734	-0.838
1AYX	0.722	7.414	-0.23	0.724	7.741	-0.25	0.951	7.061	-0.2	0.768	6.168	-0.375
1AZI	0.398	9.528	-0.4	0.398	9.528	-0.4	0.315	5.297	-0.1	0.208	4.051	-0.311
1BA3	0.008	1.062	0.689	0.009	1.089	0.835	0.215	1.482	0.81	0.011	0.851	0.247
1BC2	0.499	11.2	0.114	0.524	12.15	0.122	0.782	13.93	-0.01	0.544	8.875	-0.114
1BF2	0.647	8.279	-0.12	0.667	9.188	-0.15	0.939	9.365	-0.16	0.593	6.112	-0.241
1BFD	0.416	10.18	-0.02	0.516	12.35	-0.1	0.79	12	-0.08	0.406	8.059	-0.193
1BG0	0.395	22.67	-0.45	0.44	23.76	-0.45	0.79	26.22	-0.62	0.466	12.763	-0.469
1BG3	0.88	14.6	-0.44	0.818	14.32	-0.11	0.962	14.26	-0.1	0.87	9.587	-0.459
1BIA	0.395	11.39	-0.21	0.395	11.39	-0.21	0.82	13.63	-0.36	0.427	6.888	-0.222
1BIT	0.902	24.98	-0.91	0.905	25.33	-0.92	0.955	24.48	-0.86	0.872	15.114	-0.697
1BLZ	0.569	13.48	-0.26	0.537	15.32	-0.27	0.972	20.04	-0.57	0.603	9.835	-0.385
1BN6	0.101	5.944	0.194	0.113	6.341	0.176	0.497	11.25	-0.14	0.1	4.777	-0.015
1BO6	0.636	21.53	-0.54	0.643	22.16	-0.57	0.866	23.67	-0.63	0.634	13.485	-0.54
1BOH	0.514	8.847	-0.13	0.514	9.047	-0.13	0.875	8.725	-0.1	0.427	6.479	-0.26
1BSI	0.84	18.76	-0.53	0.858	19.6	-0.46	0.964	16.85	-0.29	0.691	11.482	-0.511
1BT3	0.836	30.27	-0.83	0.84	32.77	-0.94	0.944	27.41	-0.61	0.817	14.861	-0.476
1BUL	0.566	12.42	0.088	0.567	13.17	0.073	0.905	14.52	-0	0.203	4.285	0.012
1BXQ	0.918	36.98	-1.36	0.925	39.11	-1.44	0.959	37.52	-1.34	0.901	20.503	-0.895
1BYT	0.023	-3.617	1.459	0.02	-3.322	1.702	0.04	1.526	1.398	0.031	-2.321	0.557
1CB0	0.054	8.026	0.515	0.072	10.68	0.56	0.134	5.733	0.883	0.059	5.333	0.182
1CEX	0.745	11.19	-0.3	0.754	11.66	-0.32	0.91	11.94	-0.34	0.386	3.269	-0.217
1CJX	0.096	9.762	0.502	0.099	9.836	0.509	0.748	20.55	-0.2	0.057	4.635	0.255
1CK6	0.425	21.22	-0.34	0.438	22.33	-0.37	0.933	21.71	-0.36	0.191	8.745	-0.163
1CRC	0.293	13.08	-0.22	0.297	13.59	-0.23	0.533	11.04	-0.04	0.431	11.134	-0.712
1CRM	0.479	13.74	-0.34	0.504	14.61	-0.38	0.843	15.5	-0.44	0.695	8.878	-0.403
1CRZ	0.829	20.6	-0.4	0.834	21.21	-0.42	0.96	18.36	-0.24	0.748	12.165	-0.34
1CSR	0.536	18.42	-0.21	0.611	20.32	-0.21	0.831	22.47	-0.37	0.328	8.421	-0.229
1D6M	0.892	23.02	-0.92	0.887	24.24	-0.96	0.977	23.35	-0.9	0.867	15.044	-0.757
1DAJ	0.756	16.94	-0.8	0.55	16.25	-0.65	0.817	17.21	-0.67	0.735	13.337	-0.79
1DCS	0.59	15.86	-0.22	0.562	15.59	-0.21	0.919	17.83	-0.36	0.61	10.38	-0.275
1DHS	0.277	13.05	0.01	0.224	11.87	0.284	0.709	8.679	0.492	0.23	8.229	-0.173
1DHT	0.103	5.687	0.011	0.023	3.363	0.465	0.54	9.751	0.042	0.233	5.223	-0.255
1DIN	0.626	13.44	-0.19	0.682	14.96	-0.26	0.954	16.9	-0.38	0.639	10.228	-0.348
1DMR	0.615	18	-0.04	0.627	19.36	-0.05	0.824	19.74	-0.06	0.51	10.752	-0.148
1E1K	0.606	16.25	-0.28	0.398	11.4	0.15	0.759	12.48	0.069	0.358	7.926	-0.183
1E3H	0.532	9.653	0.226	0.583	10.64	0.191	0.861	11.57	0.145	0.46	7.482	-0.07
1E3Q	0.795	25.36	-0.58	0.806	25.53	-0.59	0.957	23.15	-0.44	0.686	13.47	-0.441
1E5M	0.715	30.64	-0.26	0.72	31.07	-0.26	0.874	30.61	-0.23	0.502	10.139	-0.11
1EBV	0.008	-2.382	0.779	0.004	-1.485	0.866	0.003	-0.856	0.844	0.042	-3.869	0.452
1EEH	0.312	12.74	0.283	0.248	11.74	0.372	0.571	17.01	0.04	0.214	6.524	0.092
1HGU	0.006	-1.353	0.64	0.176	8.388	0.038	0.427	4.417	0.331	0.146	-8.1	0.644
1LZ1	0.817	23.13	-0.8	0	-0.032	0.06	0.305	1.258	-0.03	0.016	-0.709	-0.16
1OMD	0.237	18.98	-0.27	0.261	19.82	-0.3	0.646	24.27	-0.61	0.289	10.956	-0.402
1RBP	0.761	15.65	-0.55	0.894	15.97	-0.26	0.958	14.04	-0.13	0.641	7.318	-0.348

PDB ID	AE-R	AE-S	AE-I	GE-R	GE-S	GE-I	WE-R	WE-S	WE-I	GAE-R	GAE-S	GAE-I
1RHD	0.015	2.939	0.248	0.015	2.939	0.248	0.659	8.869	-0.14	0.02	1.965	0.039
1TON	0.772	31.42	-0.85	0.643	28.27	-0.51	0.922	31.23	-0.7	0.666	14.743	-0.557
2ACT	0.596	26.04	-0.63	0.555	24.98	-0.51	0.953	25.4	-0.54	0.191	7.523	-0.229
2CTS	0.612	18.4	-0.2	0.663	19.8	-0.19	0.87	22.68	-0.39	0.437	8.715	-0.237
2LBP	0.697	16.55	-0.35	0.701	17.64	-0.37	0.927	14.23	-0.14	0.73	9.951	-0.347
2LDX	0.464	12.23	-0.21	0.463	12.34	-0.2	0.902	13.32	-0.27	0.185	4.753	-0.166
2LIV	0.729	17.71	-0.32	0.746	18.31	-0.35	0.894	17.34	-0.26	0.369	7.416	-0.153
2PRK	0.81	36.15	-0.94	0.816	37.23	-0.94	0.911	36.09	-0.86	0.673	15.805	-0.524
2RN2	0.593	26.32	-0.56	0.571	26.93	-0.56	0.784	24.49	-0.42	0.361	9.029	-0.348
2TAA	0.81	23.57	-0.78	0.82	24.19	-0.81	0.988	16.48	-0.28	0.789	13.985	-0.582
3BLM	0.693	20.09	-0.24	0.681	20.51	-0.13	0.96	21.42	-0.18	0.49	13.146	-0.375
3CLA	0.748	30.62	-1.11	0.742	29.4	-0.97	0.835	24.92	-0.67	0.712	17.283	-0.732
3CNA	0.249	4.747	-0.01	0.22	3.737	0.625	0.176	2.642	0.707	0.141	3.072	-0.177
3EST	0.46	22.98	-0.58	0.802	30.79	-1.06	0.973	29.91	-1.02	0.674	15.851	-0.673
3GBP	0.835	19.45	-0.6	0.817	20.33	-0.64	0.975	20.11	-0.63	0.825	12.325	-0.558
3GRS	0.624	25.48	-0.56	0.632	25.04	-0.44	0.895	20.9	-0.19	0.616	14.904	-0.526
3PFK	0.798	38.36	-0.99	0.813	39.93	-1.03	0.889	36.87	-0.86	0.72	17.978	-0.634
3PGK	0.057	-9.9	1.806	0.047	-8.935	1.847	0.001	-0.994	1.322	0.163	-10.291	1.03
3PGM	0.421	19.45	-0.17	0.441	20.13	-0.2	0.824	21.25	-0.25	0.171	7.281	-0.155
3PSG	0.791	35.83	-1.21	0.779	35.97	-1.14	0.961	33.47	-1.01	0.726	15.274	-0.631
3RN3	0.851	25.69	-1.15	0.862	26.98	-1.18	0.936	26.43	-1.15	0.472	10.363	-0.565
3RP2	0.74	33.29	-0.92	0.717	32.56	-0.79	0.948	33.31	-0.86	0.672	17.693	-0.672
4APE	0.935	40.78	-1.58	0.946	42.67	-1.65	0.99	36.59	-1.26	0.942	22.535	-1.003
4DFR	0.448	23.3	-0.46	0.451	24.16	-0.48	0.752	17.93	-0.04	0.422	14.214	-0.537
4MDH	0.771	31.21	-0.74	0.768	32.49	-0.75	0.907	30.17	-0.59	0.617	12.4	-0.36
4PEP	0.798	32.88	-1.05	0.768	32.18	-0.95	0.938	32.18	-0.98	0.546	13.584	-0.572
4TNC	0.03	5.115	0.225	0.014	3.276	0.446	0.166	2.62	0.517	0.027	3.155	-0.032
5ACN	0.828	19.91	-0.57	0.839	19.42	-0.41	0.911	19.13	-0.38	0.763	9.018	-0.401
5CHA	0.814	39.99	-1.11	0.834	39.96	-0.95	0.951	41.14	-1.03	0.781	20.271	-0.686
5CPA	0.765	21.24	-0.16	0.789	23.68	-0.26	0.913	23	-0.22	0.287	7.644	0.031
5CPV	0.708	38.41	-1.67	0.713	41.29	-1.78	0.711	0.507	0.594	0.76	17.332	-0.855
5CTS	0.269	14.72	0.024	0.331	16.84	-0.01	0.652	15.92	0.057	0.108	5.942	-0.079
5LDH	0.381	8.222	-0.07	0.386	8.325	-0.08	0.927	9.786	-0.16	0.261	4.644	-0.217
5RUB	0.308	10.29	0.021	0.369	11.47	0.168	0.637	11.42	0.168	0.298	8.443	-0.253
6LDH	0.564	11.44	-0.15	0.653	14.73	-0.36	0.937	9.937	-0.04	0.391	6.026	-0.235
6XIA	0.574	12.96	-0.36	0.57	13.35	-0.38	0.82	13.38	-0.36	0.474	7.77	-0.382
7API	0.554	17.3	-0.14	0.306	13.1	0.247	0.881	20.45	-0.23	0.187	6.565	-0.195
7CAT	0.741	19.11	-0.12	0.753	20.14	-0.17	0.969	23.32	-0.38	0.375	6.503	-0.061
8ADH	0.443	12.65	0.022	0.636	16.97	-0.21	0.89	14.6	-0.06	0.297	5.826	-0.141
8ATC	0.458	28.35	-0.68	0.491	30.35	-0.77	0.914	24.73	-0.39	0.392	13.585	-0.43
8DFR	0.732	16.78	-0.42	0.744	18.06	-0.48	0.938	17.25	-0.43	0.61	7.702	-0.294
9PAP	0.64	26.22	-0.58	0.768	31.01	-0.87	0.965	28.68	-0.73	0.692	11.538	-0.411
9WGA	0.851	32.66	-1.24	0.815	28.92	-0.73	0.971	28.78	-0.72	0.815	20.374	-0.833