A Framework for Comparing Flexible Objects Applied to Protein Structure Analysis

A Thesis Submitted to the Committee

In Partial Fulfillment of the Requirements

For the Degree of Master's of Science

In the Department of Computer Science

University of Saskatchewan

Saskatoon, Saskatchewan, CANADA.

by

Stephen D'Arcy O'Hearn

Spring, 2000

Permission to Use this Document

In presenting this thesis in partial fulfillment of the requirements for a post-graduate degree from the University of Saskatchewan, I agree that the libraries of this university may make it freely available for inspection. I further agree that permission for copying this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science,
University of Saskatchewan,
Saskatoon, Saskatchewan, Canada.
S7N 5A9

Abstract

This thesis presents a new framework which specifies rules for spatially decomposing objects for comparison and similarity detection. The framework is designed to optimize the trade-off between complexity and detail at all levels of analysis and promote accurate, efficient similarity detection between objects that flex, stretch, or otherwise moderately distort in shape.

A three-dimensional method specialized for structural similarity detection in protein molecules is derived from the framework. The method subjects each molecule to a spatial decomposition based on a recursive application of the octtree data structure (a special case of the binary space partition tree). Chemical properties are collated from neighbouring (and overlapping) regions of the molecules, scaled according to calculated weighting factors and decay functions, and mapped to special "aggregation" points within the cubic lattice. This enables rapid, efficient comparison of molecules.

The method is implemented as a computer program, ***MolCom3D****, which creates an octtree file that governs spatial analysis and constitutes a permanent multiple structure alignment of proteins. Separate experiments were conducted to calibrate the program and to verify that the spatial analysis and the resulting structure alignments are accurate. The accuracy of MolCom3D was found to be over 98 percent. Additionally, several avenues of future work resulting from the success of this research are identified.*

Acknowledgements

I would like to express my sincere gratitude to everyone who has demonstrated an interest in *the content* of this thesis.

I wish to thank my supervisor, Dr. Tony Kusalik, and Dr. Joe Angel for their guidance in both the computer science and chemical aspects of this project. I would also like to thank Dr. Kusalik for his thoroughness in editing this document.

I would also like to thank those who provided professional commentary and support for the ideas contained in this thesis, including Dr. Wilson Quail, Dr. Lata Presad, Dr. Mark Keil, Dr. Louis Delbaere, and Dr. Paul Mezey.

Finally, I wish to invite future researchers to explore this thesis and develop its ideas further.

Dedication

To my wife,

Michelle

and children,

Jenna and Emily

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Significant advances in X-ray crystallography, NMR spectroscopy, and other recent technologies have provided researchers with an enormous repository of protein sequence and structural data, and additional data is being added at a phenomenal rate. A natural way to use this data is to classify proteins and look for similarities. However, a fully automated method that *reliably* classifies proteins based on similarity of amino acid sequence or three-dimensional structure still remains elusive. Available methods for similarity detection, including the method developed here, endeavour to **align**, or literally put side-by-side, molecules under investigation in order to ascertain whether structural equivalences can be found in corresponding regions. Simply stated, if enough equivalent regions exist according to some reasonable threshold, the molecules are similar; if too few regions of equivalence can be found, the molecules are dissimilar.

Three major issues must be resolved before precise protein classification can be achieved through methods that rely on alignments. The first stems from a compromise that must be made between the level of detail chosen for the representation of molecules and the size of the search space required to perform comparisons of aligned regions. Detailed representations provide highly selective comparisons but preclude computational tractability in large molecules, such as most proteins, due to the enormous search space required for the comparisons. Highly detailed representations, therefore, are useful exclusively for small molecules. In contrast, highly aggregated representations feature substantially reduced search spaces amenable to practical computation but are prone to omission of information essential for reliable comparisons. Nevertheless, steady progress has been made in the development of approaches that detect similarity in molecules. Methods have emerged that are better suited for either small or large molecules, but are subject to the aforementioned compromise involving search space limitations. As a result, comparative methods that require highly detailed molecular representations are

typically limited either to small molecules or to slightly larger molecules already suspected of being similar.

The second difficulty, somewhat related to the first, involves the tendency of many optimization algorithms to converge to a local minimum value rather than to the global minimum. The propensity for this problem is directly attributable to two parameters, namely the level of detail in the representations, and the size of the molecular species being compared. Large values of these parameters necessarily result in massive search spaces for which only a relatively tiny window of consideration is possible at any given stage of the optimization, and as a result, such algorithms are predisposed to a greater likelihood of inaccuracy.

A final area of difficulty, largely unrelated to the other two, concerns the capacity of similarity detection algorithms to identify and concentrate on those portions of the molecules most likely to be biologically relevant. Biological relevance applies to portions containing either functional groups (typically located in the solvent-accessible regions), or structural elements that are highly conserved through evolution. Matching regions of biological relevance can be instrumental in detecting similarity. Unfortunately, many algorithms feature a somewhat arbitrary discretion mechanism for alignments. They are based on the simplifying assumption that *any* indication of excessive overall difference, even where differences are accumulated from nonessential regions, implies a lower likelihood of structural similarity, homology (that is, common evolutionary origin), or common biological function. Such algorithms are typically forced to introduce special place-holders called **alignment gaps** with penalties to offset their presence. This allows the algorithms to continue satisfying ordinal or geometric properties needed for the minimization of difference. However, gap usage tends to preferentially bias the sensitivity of homology or similarity detection for certain molecular compositions. For instance, gap usages causes sensitivity to be weaker in protein molecules wherein the majority of residues lie in looped regions while the highly conserved subunits that reveal homology are in the vast minority. Furthermore, purely geometric comparisons treat proteins as unconditionally rigid objects, and fail to make allowances for the subtle distortions in shape that occur when real proteins are crystallized for x-ray

crystallographic data collection. Such distortions can contribute collectively to an inaccurate detection of overall difference between proteins that are actually similar. In addition, proteins undergo subtle shape changes as they carry out normal biological functions. For example, in a process called **induced fit**, a substrate (a small molecule or ion) binds to a protein only after the protein adopts a slightly different conformation that allows the binding to occur. Similarity detection algorithms, then, should account for the fact that proteins flex and stretch, and that some alterations in the data might be necessitated in order to reveal similarity between proteins.

A fundamental challenge to automated protein classification, then, entails the formulation of a method that models and quantifies the degree of structural similarity between proteins in a way that is both biologically accurate *and* computationally efficient. In general, most current methods for modeling protein similarity involve the construction of an entity, such as a multiple sequence or multiple structure alignment, that readily identifies which regions of the proteins are similar, and which regions are dissimilar. This entity typically serves as a measure of **consensus** (or commonality) in either sequence or three-dimensional structure for every region of the protein. The collection of such consensus entities leads to a **prototype** (a model of a typical member of the collection). Prototypes are used, in turn, for the development of a classification system for proteins; each equivalence class is represented by one of the prototypes. Each prototype possesses all of the characteristic features (and possibly all of the evolutionarily conserved attributes) of its respective class, and allows newly discovered proteins to be readily categorized and the prototype to be updated. Additionally, a scoring scheme is provided by most methods for calculating the degree of similarity between class members and the prototype, and between the class members themselves. The scoring scheme allows quantitative (and sometimes statistical) decisions to be made as to whether a putative protein belongs within a given class.

Although contemporary alignment methods have enjoyed moderate success in detecting sequential or structural similarity between proteins, they cannot guarantee that high similarity scores imply similar biological function. Moreover, contemporary methods fail to overcome the difficulties identified above for several reasons. Firstly, a

fixed level of detail that typically bases comparisons exclusively on amino acid residue positions is applied throughout the algorithms. Secondly, the degree of similarity is based on an optimization of positional alignment of residues that requires the insertion of gaps to allow for optimizations that are reasonable [45, 44]. But the gap penalties necessitate parametric values that are difficult to realistically characterize. Finally, all portions encompassing the volume of the compared molecules are regarded as being equally likely to contribute to the biological function of the molecules.

This research introduces and explores the efficacy of a new method for protein similarity detection that should prove to be efficient, flexible, extensible, and most importantly, less sensitive to the problems discussed above. The method is an adaptation of a general framework also developed in this research for detecting structural similarity of moderately flexible objects. The mainstay of methods developed within this framework is a systematic decomposition and comparison of the spatial regions comprising each object. The method for protein similarity detection developed here is a three-dimensional derivation of the framework wherein octtrees[1] recursively divide the cubic volume around each object into eight equal sub-cubes. Roughly speaking, the properties within and around these sub-cubes can be compared efficiently. The application of this method to a given set of target proteins yields a multiple structure alignment in the form of a binary tree and an indication of the goodness of fit for each protein to the tree.

A description of the method for protein similarity detection and the general framework for flexible objects is deferred until appropriate background information has been presented. This background information is the substance of the next chapter.

[1] The spelling varies in the literature: 'octtree', 'octree', and 'oct-tree' are common. This research has adopted the spelling 'octtree' given that the Latin prefix for eight is *octo*. Consistent examples include 'octane' and 'octagon'. Note: 'Octotree' might also be reasonable. 'Octree' could be taken to refer to a tree relating to the eye (e.g., 'ocular').

Chapter 2

Background

The presentation of some background information is necessary before the rules of the general framework, and the protein similarity detection method derived from it can be discussed. This background information is organized into an introduction to proteins in §2.1 and protein similarity detection in §2.2, which includes a discussion of both sequence and structure comparison. Following this background, a short description of octtrees is given in §2.3.

2.1 Introduction to Proteins

The purpose of this section is to provide a basic introduction to protein chemistry that facilitates the reader's understanding of the rest of this thesis. Protein structure is hierarchical and therefore requires examination from a variety of perspectives that incorporate different levels of analysis. The following subsection formalizes some notions about levels of analysis.

2.1.1 Levels of Analysis

This section develops concepts important to the examination of complex systems using a "levels of analysis" approach. This approach is directly applicable to the investigation of protein structure and is critical to the design of the protein similarity detection algorithm developed in this research.

For the purposes of this thesis, the **level of analysis** can be defined as the degree of physical detail or conceptual abstraction chosen for the investigation or description of a system. Further, a level of analysis can be considered **appropriate** if the degree of physical detail (or conceptual abstraction) succinctly embodies the *essential* characteristics and functionality of some portion of the system that is under investigation.

In other words, the appropriate level of analysis incorporates enough specificity to allow characteristics and functionality of interest to be revealed in terms similar in detail to the original problem specification. Each level of analysis defines a set of tools applicable to the investigation in question, possibly including various branches of mathematics, particular algorithms, and so forth.

Suppose, for example, a system consists of a collection of barnyard fowl. The flock is to be segregated into chicks and ducklings (which are supposedly difficult to distinguish on the basis of appearance). The level of analysis, "birds, possessing feathers, wings, and legs" is not sufficiently detailed to manifest any differences. But adopting a slightly more detailed perspective, "the ability of birds, possessing feathers, wings, and legs to swim", reveals an obvious distinction: *the chicks cannot swim,* and are in want of rescuing. This analytical level is appropriate to the investigation of the system insofar as questions regarding segregation can be answered. However, a microscopic analysis of the nerves innervating the legs is excessively detailed. It is unduly complicated and will likely fail to reveal any of the sought-after information about the system.

Comprehensive descriptions or investigations of complex hierarchical systems—biological tissues, computer network topologies, and indeed, protein molecules—almost invariably require explanations focussed at more than one level of analysis. Moreover, each level delineates a category of concepts and components that have just the right amount of detail to effectively treat issues typically addressed at that level. The inspection of a hierarchical system can thus be guided by one or more frameworks (corresponding to overall perspectives on the system) prescribing levels of analysis that parallel the hierarchy. Thus investigations of the system can be conducted systematically rather than haphazardly.

2.1.2 What Are Protein Molecules?

Proteins are one of the three classes of biological polymers (the other two groups are nucleic acids and polysaccharides). Proteins serve diverse functions: as enzymes, they catalyze biological reactions; as transport molecules, they bind and convey biomolecules (like O_2) to various organs; as antibodies (immunoglobulins), they defend

organisms against invasion by viruses and bacteria; as structural entities, they form the major backbones of many tissues. In addition to these functions, numerous other examples exist. Figure 2.1 shows an example protein molecule, the myoglobin molecule, and its **heme** functional group responsible for binding O_2 [12]. The possible presence of certain **prosthetic** groups, like the heme group, along with other structural features of the protein molecule (especially its general shape) govern its function.

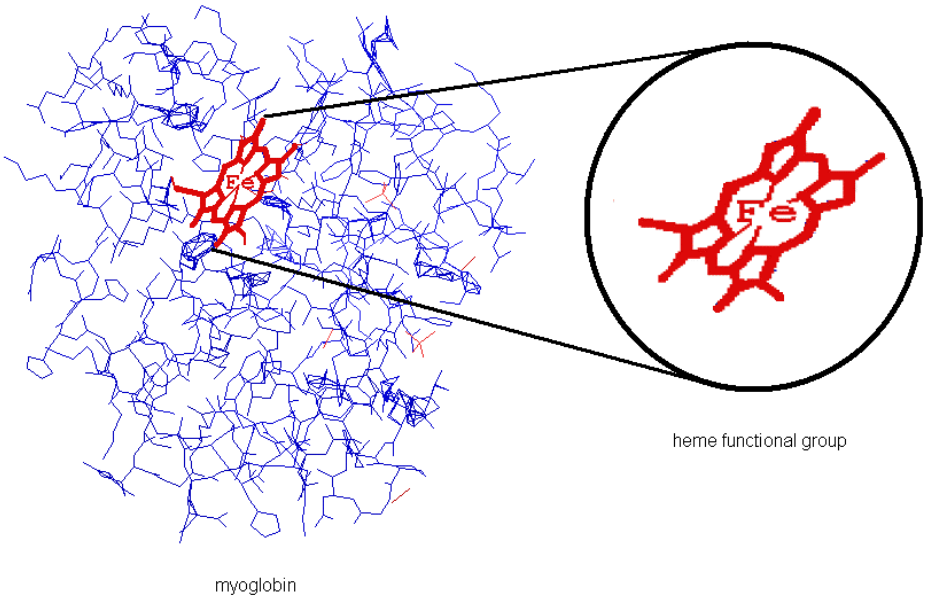


Figure 2.1: The Myoglobin Molecule

All naturally occurring proteins are synthesized by sequentially joining together amino acids from a set of twenty **standard amino acids**. They are called standard to differentiate them from other amino acids, like hydroxyproline, that result from modifications to the standard amino acids after protein synthesis. Single letter designators have been defined for naming the amino acids as listed in Table 2.1.

Amino acids are covalently bonded together through peptide bonds to form long amino acid sequences, aptly called **polypeptides**; long polypeptides, in turn, generally

fold in particular ways to form **proteins**. An amino acid that is part of a protein sequence is called an amino acid **residue** (because some of the atoms of the amino acid molecule are lost in the process of forming the peptide bond). All amino acids have a hydrogen atom, a carboxyl group, and an amino group attached to a central carbon atom called the **alpha carbon**, denoted **a-carbon**. A hydrocarbon **side chain** (or **R group**) is also attached to the α-carbon.

Amino Acid	*Designator*
alanine	A
asparagine or aspartate	B
cysteine	C
aspartate	D
glutamate	E
phenylalanine	F
glycine	G
histidine	H
isoleucine	I
lysine	K
leucine	L
methionine	M
asparagine	N
proline	P
glutamine	Q
arginine	R
serine	S
threonine	T
valine	V
tryptophan	W
unknown or nonstandard amino acid	X
tyrosine	Y
glutamine or glutamate	Z

Table 2.1: Amino Acid Single Letter Designators

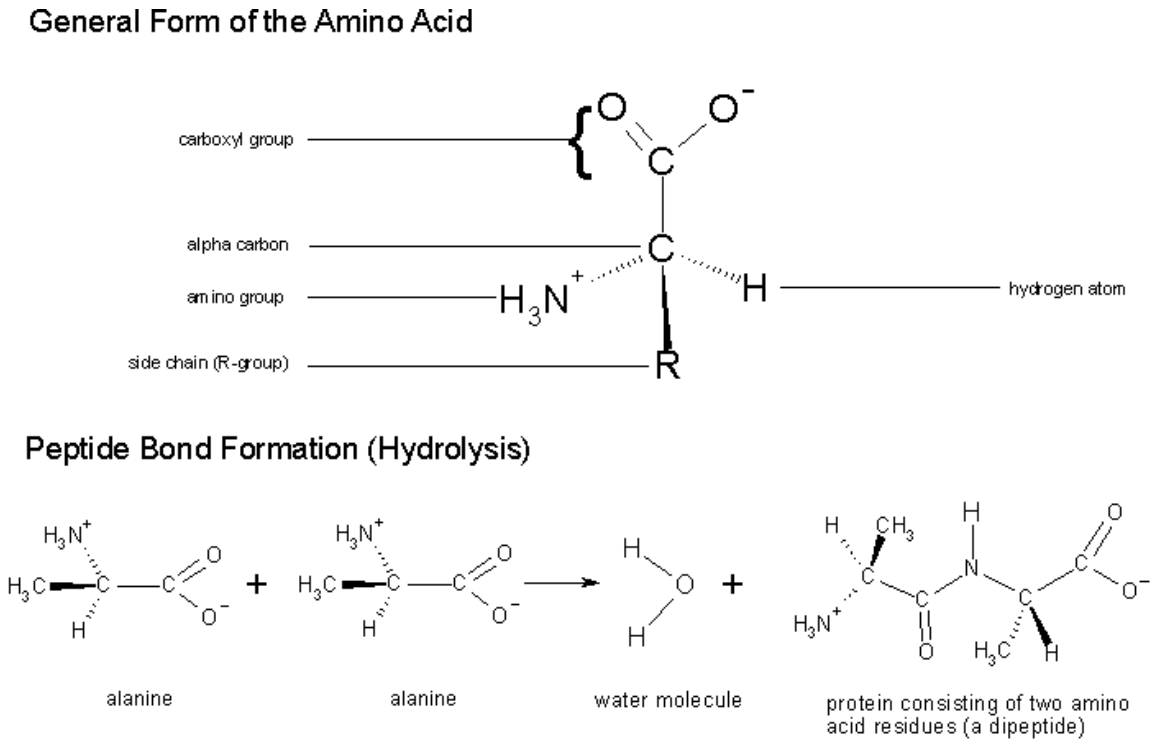


Figure 2.2: The Amino Acid General Form and Peptide Bond Formation

Figure 2.2 shows the general form of an amino acid and how peptide bonds are formed. The side chains vary in structure, size, hydrogen bonding affinity, ability to form disulfide bridges (cysteine only), and polarity.

Figure 2.3 shows a few examples of the twenty standard amino acids [26, 42]. These variations have a potentially profound effect on the overall shape and function of the protein. The following subsection discusses how these variations govern shape.

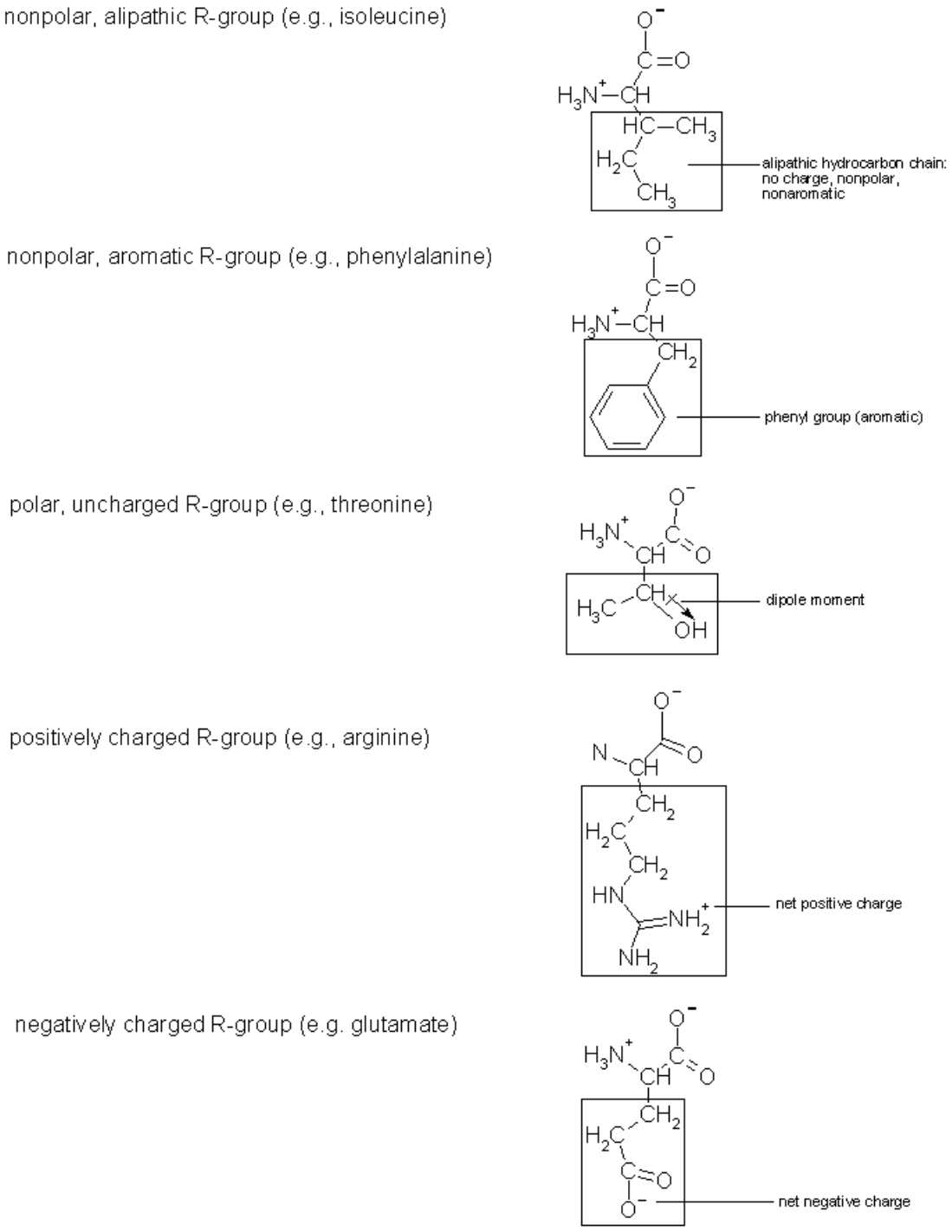


Figure 2.3: Some Representative Amino Acids

2.1.3 Side Chain Properties Govern Shape

The differences in the twenty amino acid side group structures give rise to the diversity of protein shape, and consequently, protein function. The physical properties giving rise to these differences are discussed in detail in this subsection. Taken together, these properties form the collection of comparative properties that can be examined by the similarity detection algorithm presented in this thesis.

Side chain structure and size determine the possible collection of **steric** interactions a given amino acid will have with the rest of the protein molecule. A steric interaction is a contact event that occurs between various parts of a molecule resulting from the space-filling properties of its parts, especially the amino acid side chains. Some side chains (like that of tryptophan) are rather bulky, and simply expressed, do not readily fit into small spaces.

The affinity for hydrogen and disulfide bond formation affects the capacity of amino acids to stabilize intermediate structural units (such as the α-helix discussed later) within the molecule. Disulfide (**—S—S—**) bonds are strong covalent bonds that form between two cysteine amino acid residues. As will be discussed in the following section, disulfide bonds constitute part of the primary structure of proteins.

Side chain polarity governs the polar attraction of an amino acid to water as indicated by so called **hydropathy** indices [25]. Negative indices indicate that a given amino acid is **hydrophilic** (that is, it is thermodynamically favourable when the amino acid is located on the solvent-accessible exterior surface of a protein). A positive index indicates that the amino acid is **hydrophobic** (the amino acid is best situated in a nonaqueous environment, such as within the interior regions of proteins where hydrophobic amino acids tend to aggregate in order to repel water).

As a general rule, nonpolar species tend to be hydrophobic. In contrast, polar species—either molecules with a net charge (ions), or uncharged molecules having a dipole moment—are generally hydrophilic. Thus the amino acids can be categorized readily on the basis of polarity. Nonpolar, alipathic (nonaromatic) amino acid residues—glycine, alanine, valine, leucine, isoleucine, and proline—promote hydrophobic

interactions, as well as methionine, which is nonpolar but not alipathic. Nonpolar, aromatic residues—phenylalanine and tryptophan—also promote hydrophobic interactions but not as strongly as the alipathic nonpolar counterparts. Polar, uncharged residues—serine, threonine, cysteine, asparagine, tyrosine, and glutamine—contain functional groups, such as the hydroxyl or thiol group, that form hydrogen bonds with water, making them more hydrophilic. Charged groups have a net charge and are the most hydrophilic. These include the rest of the amino acids: lysine, arginine, and histidine (net positively charged), and aspartate and glutamate (net negatively charged).

Collectively, these properties have a substantial impact on the shape of the protein molecule, both during protein synthesis (as the partially formed protein emerges from the ribosome) and after the final, thermodynamically stable shape has been adopted by the completely-formed protein. The precise set of events and conditions leading to this final shape is extremely complicated, and not surprisingly, poorly understood. For this thesis, however, it is sufficient to understand that the influences already discussed arise from the constitution of the side chains of the amino acid sequence. At every level of structure, the presence of a particular set of amino acids within certain regions of space *causes* all aspects of shape and all aspects of biological function.

The complicated interactions that cause and stabilize the final shape of protein molecules occur in concert. However, it does not follow that the examination of protein structure is necessarily a chaotic process. Instead, the overall structure can be justifiably regarded as a hierarchy of simpler structures (discussed in detail in the next section) for two reasons. First, studying and comparing protein structure is simplified considerably because superfluous detail can be suppressed and attention can be devoted to the particular structural elements under investigation. The second reason is considerably more important, and is a consequence of the natural process that yields shape: At different stages of protein synthesis, the relative contributions of the forces that influence shape fluctuate [26]. These changes result in the appearance of a hierarchy of structural elements or “building blocks” in the molecules. In particular, a successive aggregation of primitive building blocks into more complex building blocks occurs at every stage. Then, as the process continues, the forces themselves aggregate as a result of the newly formed

amalgam of building blocks; more complex building blocks are, in turn, aggregated into even more complex structural elements.

Observing proteins at different levels of structural hierarchy, then, is justified and necessary for analysis and similarity detection. What are the levels of structural analysis?

2.1.4 Protein Levels of Analysis

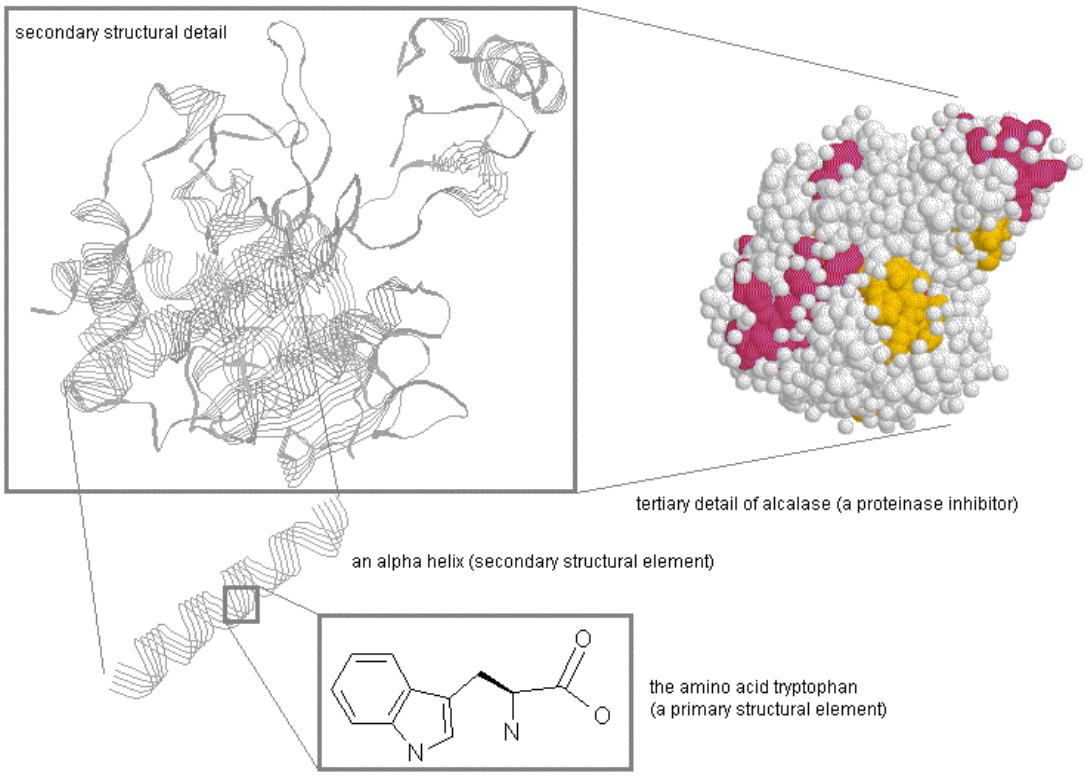
Protein molecules can be examined at several levels of structure [26]. The simplest level is called the **primary structure** and is defined by the linear sequence of amino acid residues along with the location of all disulfide bonds [42]. In essence, the primary structure captures all of the structural details brought about by covalent bonding. It is now a relatively routine process to determine the primary structure of a protein using the Edman degradation process (successive reactions of the protein with phenylisothiocyanate (C_6H_5—N=C=S) one amino acid at a time until the sequence has been determined) [5].

The **secondary structure** refers to conformations of periodic (or repetitive) structures of adjacent amino acid residues within localized regions of the protein [26, 5]. Secondary structure concerns three-dimensional arrangements of residues, but only over a short range. Only a limited number of energetically stable forms of secondary structural units are commonly observed because of chemical and spatial constraints that act on adjacent sections of the molecule. These constraints include **steric interactions** (generally observed in the nonpolar, alipathic residues that have bulky side chains), **weak interactions** (forces resulting from the thermodynamic tendency to have hydrophobic residues located in the interior of the molecule away from the aqueous environment, along with “van der Waals” forces resulting from the dynamic formation of intermediary dipoles), and **ionic interactions** (hydrogen bonding and the tendency of oppositely charged ionic residues to attract each other and be contained in an aqueous environment). Three common secondary structural units are the **a-helix**, the **b-conformation** (or equivalently, **pleated sheet**), and the **b-turn** (or equivalently, **b-bend**).

In the α-helix, the backbone of the protein becomes tightly wound around the axis of the helix and the side chains of the amino acid residues radiate outwards. It forms as a result of the thermodynamic tendency to reduce the overall potential energy of the molecule through the maximization of hydrogen bonding. That is, the large number of hydrogen bonds made possible through the formation of helical structures makes the overall molecule more stable. Furthermore, because the side chains face out radially, steric repulsion is minimized. The β-conformation forms as a result of the protein backbone chain forming large zigzags such that different parts of the chain come to lie parallel to each other. Hydrogen bonds form between the parallel portions to maintain the β-conformation and to lower the potential energy of the molecule. Additionally, the smaller side chains are crowded into the interior of the sheet in order to avoid repulsive van der Waals forces [26, 42, 5]. This van der Waals avoidance is probably why the β-conformation is energetically preferable to the α-helix in some positions. The β-turn is commonly found where the backbone of the protein chain reverses direction in a tight 180° turn involving four amino acid residues (commonly found at the junctions of β-conformations). Glycine and Proline residues are often found in the β-turn because glycine is small and flexible (only a single hydrogen atom constitutes its side chain) and because proline (roughly speaking) contains a ring that orients the peptide bonds in a tight turn.

Tertiary structure simply refers to the three-dimensional arrangement of *all* atoms in the protein. Strictly speaking, the tertiary structure refers to the shape that results from the process of folding the protein backbone chain along with secondary structures already formed. Thus foldings are superimposed on the secondary structures during the synthesis of the protein. For this reason, a tertiary structure is often referred to simply as a **fold**.

Other levels of structure are commonly encountered in the literature, including supersecondary structure (conglomerations of secondary structural units) and quarternary structure (connected elements of tertiary structure). However, these structural abstractions are not explored in this thesis, and are thus not further discussed.



Generated using Rasmol (version 2.6) and ISIS Draw (version 2.1.3d) by Stephen O'Hearn.

Figure 2.4: Analytical Levels of Alcalase: A Typical Protein Molecule

Investigation of protein similarity can be conducted at one or more of the levels of structure already discussed. Figure 2.4 shows typical structural units examined at the various levels of structure. The following section discusses a number of methods that already exist for detecting protein similarity.

2.2 Protein Similarity Detection

Protein structure has been modeled and compared on the basis of primary structure (amino acid *sequence*), and to a lesser extent, on the basis of secondary and tertiary *structure* (helix and sheet locations, and three-dimensional atomic coordinates of

amino acid residues, respectively). A hybrid method, called **threading**, incorporates both sequence and structure.

The type of comparative method (sequence-based or structure-based) most effective for extracting biologically accurate information about proteins depends on several important associations among the concepts of **classification**, **homology**, **similarity**, **sequence**, **structure**, and **biological function**. Before describing current methods for protein similarity detection, it is important to clarify some of the associations and the implications surrounding these concepts:

- **Homology and similarity**: Homologous proteins *have a common evolutionary origin* [35]. That is, they are synthesized by organisms that have descended from a common ancestor. Homology is a quality that is inferred from a level of similarity (which is quantifiable). However, homology is not equivalent to similarity. Proteins are either homologous or they are not, whereas a range of quantities describes the degree of similarity.

- **Classification**: In the Structural Classification of Proteins (SCOP) database, proteins have been classified on the basis of homology (as inferred from similar protein sequences) into "families" [32]. However, homology is not the only basis for classification. Whenever sequence similarity is not apparent, proteins have been grouped into "superfamilies" based on structural and functional similarity, and "folds" based on similar topological arrangements of major secondary structural units.

- **Sequence, structure, and function**: Sequence *determines* structure. Structure *influences* biological function. However, sequence is not always a good predictor of biological function [40]. Furthermore, the reverse implication, that sequence is predictable from structure, is not sound because in some cases several sequences give rise to rather similar structures. By the same reasoning, biological function is not a good predictor of sequence.

2.2.1 Protein Sequence Comparison

A brief discussion of sequence comparison algorithms is appropriate insofar as it pertains to structural comparison. Early work in automated protein similarity detection began in earnest around 1970. At that time, effort focussed on amino acid *sequence* comparisons since primary structural information was much more readily available than the highly resolved tertiary information that is available today from crystallography and NMR spectroscopy. Automated algorithms for sequence similarity detection became widely recognized as a result of the Needleman and Wunsch algorithm [33]. The mainstay of this algorithm, and nearly all subsequently proposed sequence comparison algorithms, is the concept of the **pair-wise** (and later on, the **multiple**) **sequence alignment**.

Definition 2.1: Multiple Sequence Alignment. Given k sequences S_1, S_2, ..., S_k from an alphabet A of letters that includes a gap character "—", a *multiple alignment* of k sequences is a rectangular array of characters from A that satisfies the following conditions:

1. There are exactly k rows.
2. Row i is exactly the sequence S_i when the gaps are disregarded. Equivalently stated, only gap characters may be inserted into S_i in forming an alignment.
3. All k rows contain at least one character that is not a gap character.
4. Columns are identified wherein all characters are non-gap characters and all characters are considered to be equivalent (that is, "aligned") [18, 40].

Where $k = 2$, the sequence alignment is pair-wise; where $k \geq 3$, the alignment is multiple. The multiple alignment is the entire array of letters, not just the aligned columns. Figure 2.5 shows part of a multiple alignment for portions of several immunoglobulin protein fragments. The alphabet for aligning protein primary sequences typically consists of the

single letter amino acid designations indicated in Table 2.1. Aligned columns are indicated with rectangles where amino acid residues are identical (consensus positions).

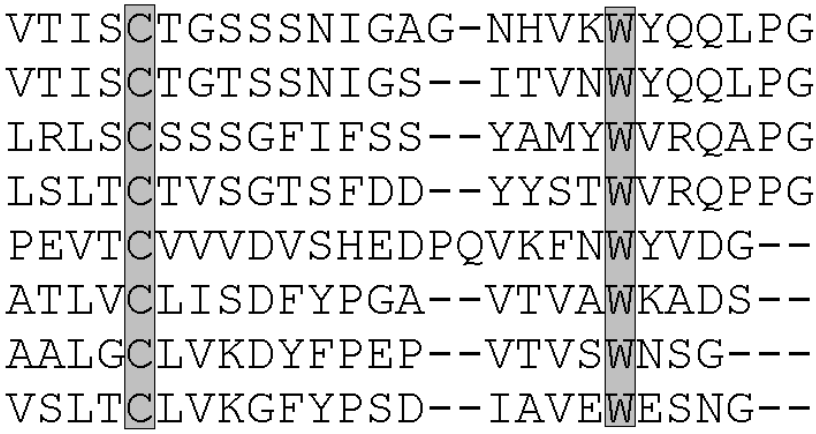


Figure 2.5: Example of a Multiple Sequence Alignment of Immunoglobulin Sequences

Needleman and Wunsch devised a scoring scheme for pair-wise alignments wherein a similarity matrix provides a score for every possible pair of aligned residues. A given alignment, then, receives a score based on the sum of the scores for all residues aligned, less a penalty for each gap introduced. The optimal alignment is found through a dynamic programming algorithm. In fact, most current sequence alignment optimization paradigms still use dynamic programming. Dynamic programming is a technique for ensuring that identical calculations are not repeated over and over again in recursive algorithms. This allows the algorithmic complexity to be deterministic and polynomial rather than exponential. In 1981, Smith and Waterman extended the Needleman and Wunsch algorithm, which handles only global alignments of entire sequences, to *local* alignments between protein subsequences [41]. These algorithms are generally considered to have solved the problem of aligning two sequences.

Since 1981, research in sequence alignments has focused on multiple alignments and on improving the biological accuracy of the results (especially by attempting to improve gap penalty choices) [45, 44, 48]. Several algorithms have been proposed for extending the original pair-wise sequence alignment algorithms [14, 3, 19, 24]. Multiple

sequence alignment is considerably more adept at extracting biologically important residues within aligned sequences, even if they are widely dispersed. It can be stated: Whereas pair-wise alignments are given to disagreement, multiple alignments are given to consensus. A multiply aligned sequence can readily reveal similarities, if any exist. This is due to the finding that residues tend to be biologically important at alignment positions where residues are of similar type in most of the sequences. Conversely, residues at positions exhibiting considerable variation are more likely to be replaceable without dramatically altering the activity of the protein. The details of multiple sequence alignment are beyond the scope of this thesis. However, it is sufficient to understand the notions of sequence alignment to the extent discussed in this section.

2.2.2 Protein Structure Comparison

Protein structure comparison is achieving greater importance as progressively more three-dimensional protein structures are resolved. Comparative techniques that employ structural information are useful in their own right and can complement the traditional sequence comparative techniques already discussed. The current empirical techniques for determining macromolecular structure are X-ray crystallography and NMR spectroscopy. Their widespread use has marked the advent of readily available atomic structural detail. The manual classification of structures through visual inspection has become insufficient to keep pace with newly resolved structural data. To wit, novel protein structures are currently resolved at a rate exceeding one per day, and this rate is escalating [21]. In 1992, approximately 300 protein structures were known [2]. By 1996, 2000 protein structures were known [10]. Currently, more than 8000 protein structures have been submitted to the Protein Data Bank [20]. Notwithstanding, newly published structures are found to be structurally similar to *previously determined* structures with increasing frequency [4]. This trend suggests that eventually an upper limit will be reached on the number of protein structural families observed. That is, protein structural diversity is apparently limited, and newly found structures will be categorized into *existing* families with increased likelihood as the number of structures grows. In fact, the

current 6500 structures are classified into about 450 families (or fold classifications) [21], and the upper limit on the number of topologically distinct fold classifications has been estimated to lie between 500 and 700 [4].

Since structure is more highly conserved than sequence through evolution [40], and the relationship between *structure* and biological function is stronger than the relationship between *sequence* and biological function [36, 29], much of the current research in comparative techniques has shifted towards structure. In fact, several biologically accurate classifications have been made through structure comparisons despite the absence of statistically detectable sequence similarity [21].

The problem of comparing three-dimensional shapes is a complex algorithmic problem. Structural comparison algorithms require (i) a representation of the chemical entities under comparison, (ii) an optimizable "objective function", (iii) a comparison algorithm, and (iv) a set of decision rules [37]. An *ab initio* (first principles) representation that incorporates the full complement of chemical properties of protein molecules would necessitate "heavy number crunching" methods that are not computationally tractable. It is thus incumbent on the algorithm to use a simplified representation that contains only the information deemed necessary to carry on biologically accurate comparisons. For example, the complexities of α-helices and β-conformations have been simplified as topological cartoon representations that facilitate efficient topology comparisons through constraint-based pattern matching [16, 15]. An objective function or, in general, an **objective relation** is a formulation of a quantitative mapping that is optimizable. A common objective function involves the intermolecular distances between superposed molecular structures. A comparison algorithm examines these distances and attempts to optimize the distances to within some threshold value. A set of decision rules, often statistically based from the results of many known examples, are then applied to accept or reject the measurement of similarity.

Several approaches have been explored for the comparison of three-dimensional structure. Holm and Sander have categorized these into dynamic programming algorithms (which will be referred to as **interstructural distance deviation** algorithms in

this thesis), distance matrix algorithms, three-dimensional clustering algorithms [21], and topology algorithms [30]. The method developed in this thesis is most closely akin to topology algorithms but is not, strictly speaking, formally topological.

The varied nature of these approaches, along with the approach presented herein, motivates the following new, *general* definition of **structural comparison** (or equivalently, **structural similarity detection**).

Definition 2.2: Structural comparison by alignment involves the process of forming a **multiple structure alignment**.

Definition 2.3: A **multiple structure alignment** of *k* objects is a computational entity representing the collection of spatial locations (or other referencing identifiers) mapped to regions deemed to coincide in all *k* objects, and deemed equivalent with respect to the consideration of (one or more) properties within and around those regions in all *k* objects.

The definition states that the spatial localities from which properties are considered (measured, predicted, conceptualized) need not geometrically coincide with the spatial locations actually attributed with the measurements. This abstraction between measurement locality and equivalenced location allows for physically reasonable departures from purely geometrical comparisons. Such geometric departures can increase sensitivity to similarity between nonrigid objects and are characteristic of topological algorithms and the algorithm developed herein. Of course, the definition does not preclude characterization of traditional algorithms that yield alignments based on purely geometric comparisons. In such algorithms, the measurement region (for example, the α-carbon position) simply coincides precisely with the geometric location where the property (α-carbon atom) is considered to exist. The coordinates of this position are used in the optimization of the objective function, and the corresponding portion of the alignment.

Informally then, a multiple structure alignment, according to the definition, shows where things are more or less alike according to properties measured in regions

considered to correspond in each structure. Region correspondence is a function of the property measured. For instance, α-carbon position is just that, the position of a particular carbon atom in a molecular group; this simple property involves matching a point and does not associate a definitive region with it. However, a property (say "45° α-helixness") might be defined at associated reference points for each structure compared. Then, the property in each structure might have a "scope" associated with it. The scope might, in this case, be realized as a step function that gives an incremental degradation in the influence of the property with distance of the observation from the reference point. Thus the types of properties dictate which regions coincide rather than simply the spatial coordinates at which measurements are taken. Coinciding regions are considered to represent alignments of structure.

The above definition is expected to be sufficiently robust to characterize not only the more conventional structural alignment methods that match, for example, α-carbon atom positions, but also algorithms based on topologic properties, and the algorithm developed in this research (which is reminiscent of topology). The definition allows alignments to be based on any physically or conceptually useful set of properties.

The current algorithms for structural comparison satisfy Definition 2.2 and can now be discussed.

2.2.2.1 Interstructural Distance Deviation Algorithms

The dynamic programming algorithms used for the optimization of *sequence* alignments have been adapted to *structure* alignments. For structure comparisons, the root-mean-square (RMS) deviation between equivalenced (i.e., matched) amino acid residues is optimized (minimized) by dynamic programming similar to the minimization of "edit distances" (or basically, number of consecutive gaps) in sequence alignments. If the structures are very similar, and if a *sequence* alignment is available to identify corresponding amino acid positions, RMS deviation is an effective and fast quantitative measure of similarity [10].

The decision about whether a given RMS deviation indicates similarity is usually based on the raw RMS score; if the RMS deviation measures below a chosen threshold, the structures are deemed similar. However, this measure of significance may not indicate true biological significance [1]. To increase accuracy, statistically-based decision rules have been introduced which use *p*-values and probability density functions derived from known structural information [27]. The probability density functions are an adaptation of earlier such functions used for sequence similarity detection.

A measurement resembling RMS, called the **Area Functional with Fit Comparison** (AFFC) has been proposed [10] as an alternative to RMS. The AFFC distance is found by a dynamic programming algorithm that minimizes the area of triangles formed between the α-carbons of two proteins. The minimum AFFC distance is related to this area. This method outperforms RMS for structures that have modest topological similarity, and neither requires an initial sequence alignment nor the introduction of gaps. Nonetheless, it is not intuitive that low AFFC distances indicate *biological* similarity in structure.

The multiple alignment of structures using the interstructural distance deviation approach can be thought of as the superposition of all structures that minimizes the distance measurement between corresponding parts of each structure. Furthermore, the superposition described satisfies Definition 2.3 because all regions in and around each structure are deemed to coincide as a result of the RMS (or whichever) deviation being below the threshold value (and as a result of introducing gaps into the structure where needed).

2.2.2.2 Distance Matrix Algorithms

Distance matrix algorithms construct, for each structure under comparison, one or more matrices containing quantitative values about structural relationships internal to the structure. Typically these values are distances between each amino acid residue and all the other residues within the structure. Thus a "structural environment" exists in the

matrix for each residue consisting of similar patterns of contact with neighbouring residues [38, 43]. Comparison of three-dimensional structures is then accomplished through the comparison of the constructed two-dimensional matrices and the discovery of optimal scores through dynamic programming [38] or both dynamic programming and Monte Carlo methods [43].

The information contained within the matrices has the advantage of being independent of the "coordinate frame" [38, 43]. That is, the information remains invariant with rotation and translation of the structures. Furthermore, distance matrices are less sensitive to moderate insertions and deletions of subsequences that result in subdomain (protein functional unit) displacements in relation to the two structures; topological equivalence is preserved in the matrices and remains detectable because relative trends are compared rather than absolute geometrical coordinates. In the method of Holm and Sander, patterns (in the form of small submatrices) are first matched in the matrices corresponding to each structure [43]. From these submatrices, hierarchically more complex submatrices are matched using a Monte Carlo optimization (which is simply a random selection of submatrices) to iteratively improve the submatrix matching overall score.

Distance matrix methods satisfy Definition 2.3 in that a one-dimensional alignment of amino acid residues is generated from an inspection of the matrices. This alignment, in effect, identifies structural environments for residues considered to be equivalent in the protein molecules. The equivalences are mapped to the amino acid residue locations at the center of the structural environment.

The major disadvantage of distance matrix algorithms is that the matrix comparisons themselves pose a computationally intensive problem. The matrix comparisons still require optimization strategies such as dynamic programming and Monte Carlo methods.

2.2.2.3 Three-dimensional Clustering Algorithms

Three-dimensional clustering algorithms endeavour to find the **common structural core**, which is defined as "a common set of structural elements similarly arranged in space" [37]. More simply stated, the common structural core is a collection of secondary structural unit clusters found in all structures in question.

Clustering has been done on the basis of common Spatial ARrangements of backbone Fragments (SARFs) of protein molecules [2, 1]. Initially, small sets of SARFs (fragments containing no gaps) are located. These are then unified into successively larger SARFs until the RMS deviation-based score stops improving and the SARFs with a similarity score above a certain threshold are reported for the collection of proteins. An alternative method for clustering uses a subgraph isomorphism algorithm to locate common subgraphs contained within complete graphs that represent the structural units in each protein [31]. The isomorphic subgraphs (subgraphs having identical node adjacency relationships) represent clusters within each full graph that are associated with secondary structural units common to each protein.

Clustering is effective for locating common **structural motifs** (that is, arrangements of secondary structure, or equivalently, **supersecondary structure**) present in a collection of proteins. The clusters serve as a basis for classification, and taken collectively, constitute an *alignment* of regions exhibiting similar secondary (or higher level) structure.

The clusters comprising the common core map *nongeometrically* to regions in all structures that coincide with respect to composition of similar secondary structure. Notwithstanding, Definition 2.3 is satisfied inasmuch as clusters are either implicitly or explicitly assigned identifying tags that allow reference to them and to the corresponding regions of common structure. Since the tagged regions are deemed to coincide, a nongeometric alignment is achieved.

2.2.2.4 Topology Algorithms

It has been found that evolutionary divergence in most protein families involves changes in the hydrophobic interior portions of the molecules [7, 6]. Amino acid replacements are accompanied by subtle shifts and rotations in secondary structural elements that allow for a largely invariant volume of the interior. More radical amino acid replacements are observed in the exterior looped regions. However, these alterations do not undermine biological function as evidenced by their continued phenotypic expression. Therefore, similarity detection based on topologic rather than geometric equivalence of secondary structural units is justified.

Topology is a relaxation of the geometric requirement that distances between all portions of an object must remain invariant for the object to maintain its identity. The term "rubber geometry" [30] characterizes topology. In general, topology is based on the **homeomorphism**, a "reversible continuous transformation that converts each point of the original object to a unique point of the new object", as when temporarily bending a rubber object [30]. Two objects are **topologically equivalent** if a homeomorphism exists between them. Of course, if two objects consist of identical arrangements of points in space (implying that the distances between their respective points are identical), the two objects are geometrically equivalent as well as topologically equivalent.

Attempts have been made to define topological equivalence in proteins. To this end, topological equivalence has been defined as "a sequential series of structurally equivalent residues" [29]. In this context, "structurally equivalent" means that a structural element of one molecule coincides with a similar element in the other molecule, within defined limits, and the elements are oriented in the same direction. An algorithm for topology prediction from secondary structure and a set of folding rules has been implemented in Prolog [34, 9]. In this algorithm, protein topology is based on the sequence, adjacency, and orientation of the residues comprising the units of secondary structure. Complete accuracy of secondary structure prediction is unnecessary for achieving over 70% accuracy in predicting topology. Topology prediction is useful for comparing new proteins to existing proteins with known topologies. The Prolog

algorithm has been enhanced through constraint logic programming, making it 60 times faster [8]. Attempts have also been made to simplify the representation of topological structure for efficient searching and topology matching by Gilbert et al. [16, 15]. In their method, complex three-dimensional topological structure is represented in two-dimensional diagrams. These diagrams represent secondary structural elements (SSEs) which can be aligned and scored using constraint logic programming.

Definition 2.3 is satisfied by topological algorithms since computational reference identifiers (like the SSE diagrams) are mapped to regions that are deemed equivalent through topology.

2.2.2.5 Specialties of the Current Structure Comparison Algorithms

Each algorithm class for structure comparison has proven to be adequate for its intended purpose. Interstructural distance deviation algorithms (typically RMS deviation algorithms) have been developed for efficiently aligning protein structures so that the *overall* RMS deviation of the amino acid residue positions is minimized. However, all residues are considered to be equally important during the alignment process and in the final score for the match.

Distance matrix algorithms find similar contact patterns of amino acid positions and compactly represent these in the form of matrices. These algorithms are well-suited for matching and comparing substructures within the proteins. However, the matrices still require analysis for similar amino acid residue contact patterns. This is a relatively complex process.

Three-dimensional clustering algorithms are adept at comparing common structural motifs and secondary structural elements based on backbone fragment orientations. However, small clusters found initially require assembly into successively larger clusters, posing a computationally difficult problem. Several threshold values must be defined to allow for small deviations in the backbone fragment orientations within the clusters.

Topological algorithms are aptly suited for molecules because of their tolerance for flexibility, a prominent characteristic of molecules. However, topological algorithms must, so to speak, "lock onto" the various structural portions, and "bend and twist" each portion into equivalent positions in order to verify a match—a formidable task that is often tackled using matrices [30] or graph theoretical methods [31]. Such methods tend to be rather computationally complex.

In addition to these disadvantages, these methods apply a fixed level of analysis. RMS deviation and distance matrices consider amino acid residue positions, clustering is based on backbone fragments, and topology algorithms have concentrated on a particular level of structure, such as the secondary structural elements discussed in §2.2.2.4.

Furthermore, except for topology, the other methods are not readily adaptable to other structures in general. Rather, they are highly specified for biological macromolecules.

2.2.2.6 Advantages of the New Structure Comparison Framework

The new structure comparison framework is fully described in Chapter 4. Like the current algorithms already discussed, it is based on the concept of the multiple structure alignment as described in Definition 2.3. However, it has been developed under a different philosophy. Its central tenet declares that the properties *of the space* surrounding the structural elements should be subject to analysis rather than the structural elements themselves. The *contents* of the space give rise to a myriad of observable and comparable properties.

The rationale behind this philosophy stems from the idea that structure manifests properties extending beyond simple identity. That is, structure is not the only attribute of space that can be matched. Structural elements may exert influences on the surrounding space that can be measured and compared. Apart from structure, electrical charge, statistical properties, radiosity, and many other properties can be cited as examples of observable entities that can arise from the content of the space. Consequently, the framework provides for the consideration of a dynamic collection of properties rather

than a single fixed property. Each property has a scope of analysis defined for it that matches its expected degree of influence in the physical system, and the range of spatial dimensions for which it should be considered. Existing methods, especially those for protein structure comparison, have not accommodated dynamic collections of properties as is designed here.

This treatment of properties, then, leads to a dynamic level of analysis. If the properties to be measured at any given dimension (that is, their scopes of analysis) are properly chosen, an appropriate level of analysis of the space (as defined previously) should be achieved. This benefit of dynamic properties will be investigated in this research.

A method is derived from the framework and verified for *three-dimensional* comparisons of *proteins*. However, the framework is applicable to objects in general, *not just to proteins*. Moreover, methods can be derived for comparing structure in spaces other than three dimensional space (see Future Work section). For instance, one-dimensional derivations lead to methods applicable to gap-free sequence comparison. Four-dimensional derivations of the framework could be used for time-based structure representation and comparison, which would be useful for investigating protein folding and modeling of cell content within living cells over time.

One of the most important features of the framework discussed in this thesis is the idea of overlapping spatial localities (described in detail in §4.2). The spatial localities concept specifies a scaled overlap in the observation scheme that redundantly maps observations to points in space that, in turn, are directly compared. As a result of measurement overlap, trends and tendencies are smoothed in the measured properties, and these smoothed measurements are examined and compared (see Future Work section for suggestions on enhancements to this measurement smoothing concept). This research will verify that the trend analysis brought about by the overlap accommodates moderate flexibility in structure.

2.2.2.7 Other Comparison Frameworks

This research is one of many projects that attempt to unify a collection of principles into a framework. Frameworks endeavour to generalize salient concepts of related processes and put them together into a basis that can be built upon.

For proteins, comparison on the basis of both sequence and structure has been unified in a statistical framework [27]. Probability density functions were derived for both sequence comparison and structure comparison raw values. Cumulative distribution functions were then determined for estimating statistical significance for either raw sequence scores or raw structure scores. The resulting framework provides a convenient statistical technique for comparing proteins on the basis of sequence, structure, or both.

Another example of a framework, called the Structured Adaptive Mesh Refinement Applications Infrastructure (SAMRAI), has been created to simplify development of **adaptive mesh refinement** (**AMR**) applications [22]. AMR applications localize important features of physical systems and processes and direct computing resources to these features. SAMRAI has been used in the development of applications for computing measurements on physical systems where the measurements vary considerably over the spatial domains occupied by the system and over the time measurements are taken. SAMRAI provides a collection of abstract classes and a set of operations from which particular applications can be derived. Applications for computational fluid dynamics and measurement of granular flow are under development within the SAMRAI framework. SAMRAI is expected to reduce code duplication, learning difficulty, and application development time since the underlying framework is part of every application developed.

2.3 Octtrees

The method to be presented for protein similarity detection is based on a three-dimensional derivation of the newly developed general framework. In three dimensions,

the recursive cubic subdivision specified by the framework is equivalent to the cubic mesh structure represented by the **octtree**. Based on operational definitions from the literature [11], the following general definition of the octtree is presented:

> **Definition 2.4:** An **octtree** is a hierarchical *data structure* which represents and locates aggregated feature information within a cubic region of space. It is constructed in accordance with a spatial decomposition of the cubic region by *recursively* subdividing this cubic region, and all cubic sub-regions, into (eight uniform) cubic sub-regions that become bounded by the three (orthogonal) dividing planes and the original (planar) faces of the cubic region being subdivided.

From Definition 2.4, a few concepts need to be made explicit. An octtree is neither the actual subdivided cubic region, nor the resulting mesh structure. Rather, it is a *data structure* in the form of a tree that represents and pinpoints features of interest that are found whenever the same cubic region is similarly broken down. The data structure is not restricted to any particular tree form: it can be octary (an m-ary tree with potentially eight child nodes per parent node), binary (with potentially two child nodes per parent node, where three levels are required to locate each sub-cube), or any other convenient form.

The octary tree simply considers each of the eight sub-cubes to be represented by a corresponding subtree. The binary tree representation is used in this research and considers the octtree to be a special case of a **binary space partition** (**BSP**) **tree** [11]. A BSP tree is a binary tree whose root node represents an n-dimensional spatial domain. Each node in the tree represents a hyperplane which partitions (divides) this space into two subspaces (note: partitioning in two does not imply division into halves). The BSP tree is equivalent to the octtree if the spatial region to be subdivided is a cube, and if the dividing planes cause the formation of eight sub-cubes of equal size.

Figure 2.6 shows a two-dimensional example of a BSP tree spatial decomposition. On the left-hand side of Figure 2.6 is "Area 1". On the right-hand side is the corresponding BSP tree. The leaf nodes of the BSP tree contain the summary information for the corresponding area represented in the tree. The interior nodes contain the information about the "hyperplane" (in the present case, the line segment in two

dimensions) position. In the example, Area 1 might be subdivided into sub-areas 1.1 and 1.2 by "hyperplane" A. These areas, in turn, might be partitioned by "hyperplanes" B and C, yielding respectively, the four sub-areas: 1.1.1 and 1.1.2, and, 1.2.1 and 1.2.2. The subdivision process continues until the area is sufficiently partitioned to yield information useful for the investigation. Homogeneity in the measurement of some property within the sub-region is a common subdivision criterion. If the property of interest measures approximately the same throughout the region, subdivision terminates, otherwise subdivision continues. As discussed in Chapter 4, the properties considered by the framework can be significantly more complex than homogeneity.

Figure 2.7 shows a spatial decomposition of a cubic region on its left-hand side, and the associated binary tree representation of the octtree on its right-hand side. Traversing the binary tree is equivalent to locating cubes, sub-cubes, sub-subcubes, and so forth (referred to simply as cubes hereafter), that are represented in the tree. For this research, the precise geometric mapping from the origin to the center of a given cube represented in the tree, along with its dimensions, are precisely defined in terms of a recurrence relation specified in the structure comparison framework presented in §4.2. The recurrence relation is associated with a tree traversal that locates the cube and assigns a number to it systematically. Binary numbers are assigned to the cubes in accordance with the traversal path taken to locate a particular cube. For each cube undergoing subdivision, this number comprises three bit positions (for basis vectors **i**, **j**, and **k**) that constitute surface normals to the three subdividing planes. Whenever a left branch is traversed, a zero is assigned to the corresponding bit position, and the sub-cube is located on the side of the dividing plane *against* the direction of the corresponding basis vector. Conversely, a right branch traversal results in the assignment of a one to the corresponding bit position, and the sub-cube to be located on the side of the dividing plane *along* the direction of the corresponding basis vector.

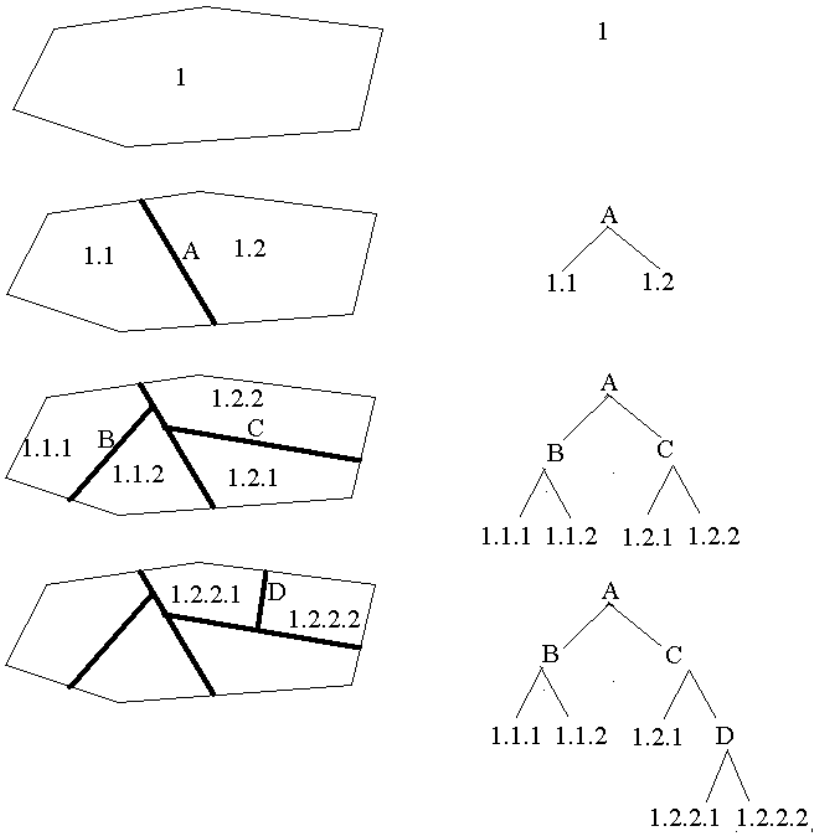


Figure 2.6: A BSP Tree Example in Two Dimensions

For this discussion it is sufficient to understand that three levels of the tree are necessary to locate any given cube in the mesh since a binary tree representation has been adopted. For any cube under subdivision, the root node of the subtree corresponds geometrically to a point at the center of the cube (called the aggregation point in the framework). Relative to this center point, the first tree level corresponds to the left or right half of the cube. Within this half, the next tree level corresponds to the top or bottom quarter of the original cube. Within this quarter, the next tree level corresponds to the back or front eighth of the original cube. The eight sub-cubes are called **octants** (analogous to quadrants in two dimensions).

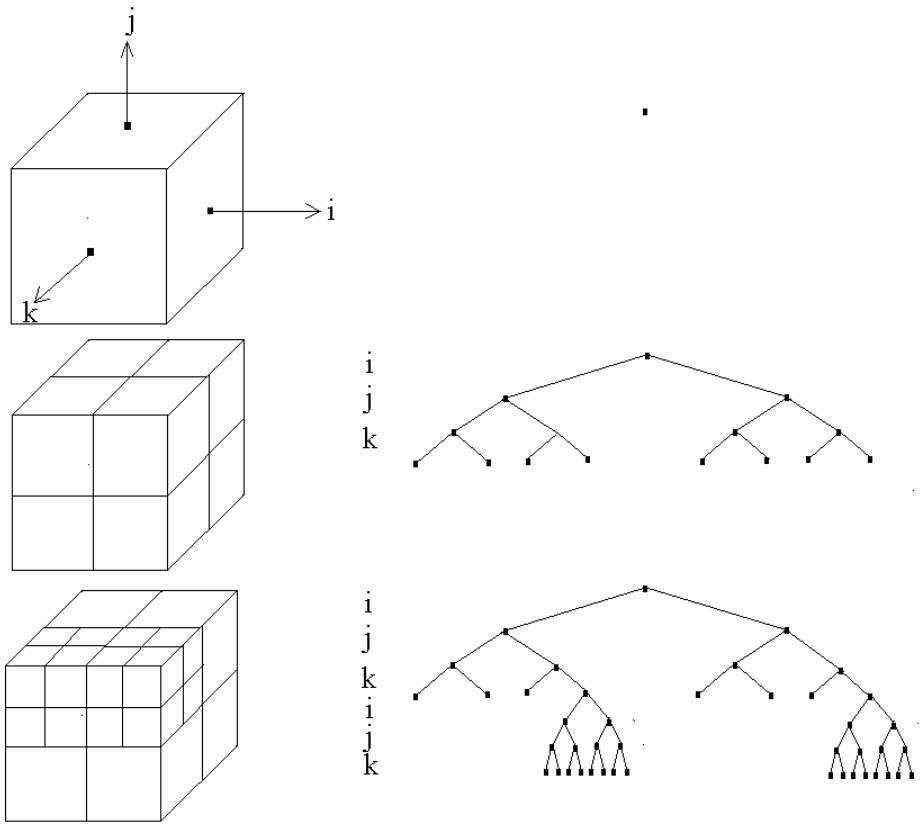


Figure 2.7: Spatial Decomposition Example and the Corresponding Octtree (Binary Tree Representation)

The remainder of this section discusses some specific applications of the octtree. Their application in this research is an extension of the more conventional octtree-based applications. The extension to octtree applications stems from the concept of **property scope** as discussed in the structure comparison framework presented in Chapter 4. A further extension involves the concept of the overlapping **spatial localities** described in this framework. A detailed explanation of octtree usage in the scope of this research is deferred until Chapter 4.

In general, octtrees are used to efficiently summarize information essential for the characterization and localization of objects within a system of objects in three-dimensional space. The summarized information has many potential uses. For instance,

the information can be used to compare the system to other systems, to store and reproduce salient features of the system, or to navigate robots within the system. These disparate examples have a common requirement for knowledge about the contents of a cubic region of space.

Octtrees have been used in graphical simulation systems for robotic visual systems to estimate the distance between objects under rigid motion [28]. In addition to storing object information, the octtrees maintained transformation matrices to limit recalculation of octtree spatial approximations and to reduce the accumulation of error.

Octtrees have also been used for storing summary information that directs a selective exploration of space (represented as volume data in a file) to regions of current interest for improving interactive rendering [46]. An adaptation of octtrees is used, called branch-on-need octtrees (BONO), that minimizes unnecessary exploration of regions that are not of current interest. Octtree branching occurs in conjunction with **isosurface** extraction from volume data. An isosurface is a surface of an object that has a constant value (to within a given threshold) for some quantity of interest.

Octtrees are an important feature of a spatial database program called SIERRA (Spatial Interface for Efficient Relational Retrieval and Analysis) [17]. SIERRA is a relational database that stores spatial relationships between the various components of a nervous system. It has been developed to support a tool, called NeuroSys, that explores connectivity patterns in nervous systems. The physical features of neurons, along with their positions, are efficiently represented using octtrees. Octtrees are initially built for the nervous system, and the relationships are then stored in the form of database tables for later querying.

Octtree usage is application-specific. The exact nature of the information represented, the actual data structure used for storing the information, the coordinate system, the labeling convention for the octants, and the conditions for terminating the spatial decomposition, all depend on the application. Enhancements to the *core* octtree paradigm can be added to traditional octtree-based algorithms. The core octtree paradigm is then responsible for directing spatial exploration to *formal spatial regions* (the actual

cubes of the associated mesh). Specific enhancements can then be applied by the application that enable measurements from outside the formal spatial region to get factored into the summary information for each formal region. This feature is explored in this research and is explained in chapter 4.

Chapter 3

Remainder of This Thesis

The previous chapter has presented some important background information, concepts, and definitions. Concepts in protein chemistry, notions of protein similarity detection through alignment, levels of analysis, properties, spatial decomposition, and octtree data structures have been addressed. In addition, several new general definitions have been offered for the *level of analysis*, the *appropriate* level of analysis, *structural comparison by alignment*, *multiple structure alignment*, and the *octtree* (in terms of the binary space partition tree).

The chapters that follow discuss the new paradigm for spatial comparison and its supporting computer algorithm, the testing procedure, the observations and results, and the conclusions and future ideas generated by this research. Chapter 4 elaborates on the new structure comparison framework presented in Chapter 2 and introduces the new protein comparison method derived from it. Chapter 5 discusses details specific to the development of the protein comparison method described in Chapter 4. This includes a discussion of the chemical properties compared and a description of all parametric values corresponding to these properties. In addition, Chapter 5 presents the actual algorithm and the octary tree file structure developed in this research. Chapter 6 discusses the testing procedure used to verify the protein comparison method. Chapter 7 gives the observations and results of the research, and Chapter 8 gives the conclusions and several suggestions for future research.

Chapter 4

Structure Comparison

The structure comparison framework mentioned in Chapter 2 is fully described in this chapter in two parts: an intuitive description is presented in §4.1 that is tailored to protein comparison, and a formal description for the comparison of objects in general follows in §4.2. The intuitive description is meant to serve as an introduction to the formal definitions making up the framework. The formal description has been provided for generality and completeness of detail and may be glossed over by readers interested solely in the comparison of proteins.

The framework consists of a set of rules for comparing objects through a recursive hypercubic spatial decomposition of the objects. The framework leads to a "comparison space" in the form of a binary tree that indicates regions of similarity. This research provides evidence that the framework is useful in general for comparing objects that flex, stretch, or otherwise undergo moderate distortions in shape.

The protein comparison method is designed to model essential structural features and detect whether molecules are similar by comparing these features. This method is a specific three-dimensional application derived from the structure comparison framework mentioned above, where the hypercubes form a recursive *cubic* lattice described by the octtree data structure. Properties examined within the various regions of the proteins are mapped to particular points in space by virtue of the octtrees (as discussed in §2.3). The resulting octtrees are shown by this research to constitute efficient, readily comparable representations of aligned objects in compliance with Definition 2.3.

4.1 The Protein Comparison Method

The structure comparison framework specifies rules for spatially decomposing and comparing *n*-dimensional objects. A binary space partition tree is constructed to

keep track of spatial decomposition of the objects being compared and to constitute a multiple structure alignment in accordance with Definition 2.3. This section informally introduces this structure comparison framework in the context of the derived protein comparison method in three dimensions.

The ultimate goal of the protein comparison method is to build a multiple structure alignment for the proteins under investigation in accordance with Definition 2.3. A special binary space partition tree (octtree in three dimensions) is built by enclosing each protein, in turn, by a cube (large enough to encompass every protein under examination), and then recursively subdividing this cube over successive iterations of the method while collecting spatial information about the proteins. As a necessary precondition, the proteins must be in the same approximate spatial orientation (currently another algorithm is used to structurally pre-align the proteins). Certain properties within this first cube are measured and compared. These properties might include, for example, the number of α-helices, the number of amino acid residues, the number of alipathic, aromatic, or charged residues. Alternatively, statistical properties, such as *p*-values, or RMS deviations of particular features might be examined, and so on. Any protein molecules that are sufficiently different, as specified by the list of properties, are discarded from further consideration by the algorithm. For the proteins that remain, the point at the center of this cube, represented by the root node of the octtree (a so-called **aggregation node**) stores the comparative summary information. For example, the average number of α-helices might be 10, and on average there might be 1200 amino acid residues. The actual properties examined in this research are discussed in Chapter 5.

The cube surrounding each protein remaining in the collection, if any, is then subdivided into eight (equal) sub-cubes; these cubic subdivision process continues recursively over successive iterations of the algorithm. In accordance with the size of the sub-cubes, a list of properties is measured within and around each sub-cube formed by the mesh. It is sufficient to understand, at this point, that the set of properties examined within and around any cube is contingent upon the dimensions of the mesh at each iteration of the method. The rationale behind this stems from the expectation that properties change in relative influence with variation in mesh size. For instance, as the

mesh size decreases, it might become important to measure amino acid residue identity and charge rather than the number or orientation of α-helices (after these have been equivalenced). It behooves the user of the method to choose appropriate properties at every mesh resolution to achieve biologically accurate results.

Figure 4.1 depicts part of a subdivision process for a protein molecule. The cube on the top right-hand side is called the level-0 cube and contains the entire protein. It gets subdivided into eight sub-cubes (or octants). The vectors **i**, **j**, and **k** denote the orientation of the cube in space. Properties, like α-helix count, might be examined at this level. This count is attributed to the aggregation point for comparison with other molecules similarly broken down.

The sub-cube on the top left-hand side of the diagram is one of the octants of the level-0 cube; it is one of eight level-1 cubes that get recursively subdivided. The cube at the bottom of the diagram is a magnified representation of this octant. At this level, the properties of interest might no longer consist of an α-helix count, but rather, charge and amino acid type, for example. The measured properties are mapped to the aggregation point at the center of the smaller level-1 sub-cube (and may get mapped to nearby aggregation points if overlapping measurements are used). This sub-cube (along with the other seven level-1 sub-cubes) is eventually subdivided, in turn, into eight level-2 sub-cubes, and the process continues. Subdivision terminates when the cube dimensions become small enough that no further properties require examination, when the maximum depth of subdivision is reached, or when a threshold of similarity is exceeded and the structures can be deemed equivalent.

So far, properties measured from within a given cube or sub-cube are mapped to the aggregation point at the center of the cube. However, the cubic region can be optionally extended into a larger cubic or spherical spatial locality having dimensions that are a multiple of the size of the cube. In this case, the measurements mapped to the aggregation point are taken from this larger area and are still mapped to the aggregation point (a decay function is applied to the measurement that varies with distance between the measurement and the aggregation point). These extended measurement localities

cause measurements to overlap and be less sensitive to shape distortions due to flexing of the protein and a more subtle trend analysis results.

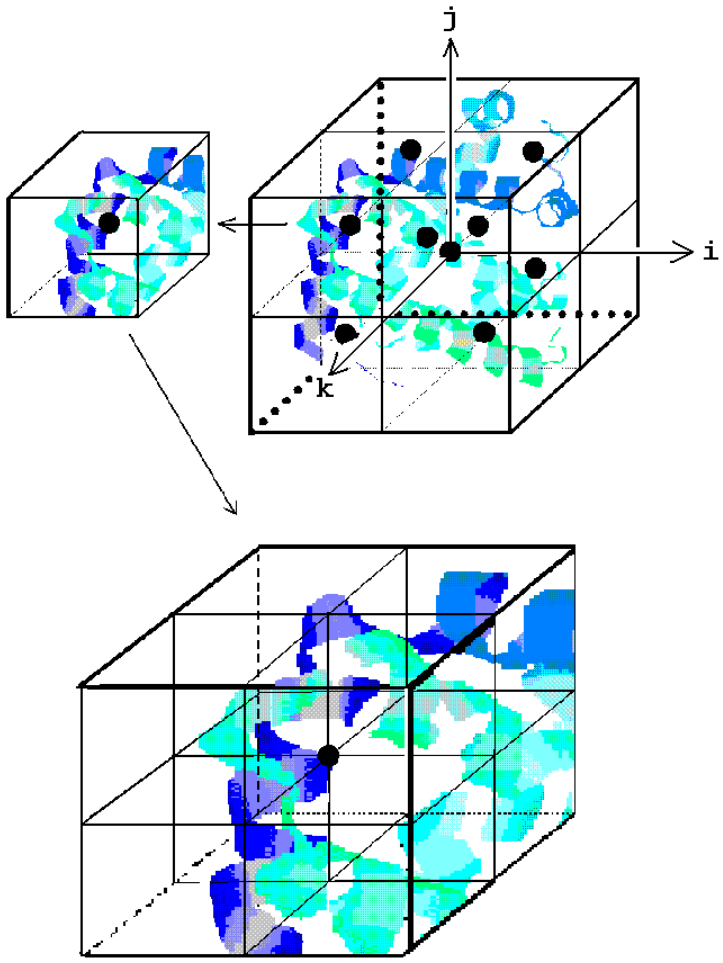


Figure 4.1: Octtree Spatial Decomposition Example

During the spatial decomposition, a binary tree is created that stores the aggregated measurements for all portions of the cubic space that are equivalent with respect to the properties considered in the collection. A subsequent traversal of the tree will locate the coordinates of the aggregation points corresponding to regions of similarity, and optionally, the values mapped to these points (if measured values are

stored in the tree). Since these points correspond to cubic regions, the regions with similar properties can be approximately reproduced in space. Thus the tree is a compact representation of the structural portions of each protein that are similar and constitutes a multiple structure alignment. As discussed in §5.2.2, the tree also serves as a basis for calculating an overall similarity score for the collection of proteins compared.

In summary, during the generation of the octtree, trends and tendencies in chemical attributes contained within corresponding cubic subdivisions of the volumes encompassing each protein in the target set are compared. An alignment results that consists of a set of points for each protein that summarizes the chemical attributes contained within neighbouring, interpenetrating cubic regions whose dimensions are greater than or equal to the dimensions of the associated subdivided cube. The measured chemical attributes culminate in directly comparable quantitative values. A series of quantitative values are associated with each point and each point is considered to be "imbued" with these values.

The next section gives a formal presentation of the details of the structure comparison framework. These details can be omitted by readers interested only in three-dimensional protein structure comparison. However, readers in need of a more explicit description for formulating methods that compare *n*-dimensional objects in general are encouraged to read the formal definition.

4.2 The Framework: Formal Definition

The following framework is hereby formally defined for conducting all-versus-all comparisons between members of a collection of (two or more) *n*-dimensional structures ($n \geq 1$) in terms of the following rules. The rules require that the structures are already approximately aligned. The Future Work section suggests a possible simple method for obtaining approximate initial alignments.

1. **Vector Space**:

   Each structure in the collection, and all regions within and around each structure, must be spanned by a finite set, S, of n orthonormal basis vectors $\{\boldsymbol{e}_0, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_{n-1}\}$ that exist in Euclidean space. The set C is hereby defined as the set of vectors describing the collection of structures under comparison.

2. **Mesh**:

   A hypercube, H_0, is defined in terms of S such that H_0 at least encloses the largest structure in the collection; that is, H_0 encloses C. H_L is referred to as a hypercube **at level *L***, or equivalently, a **level-*L*** hypercube.

   Spatial decomposition of the structures in the collection is accomplished by superimposing each structure within its own instance of H_0 and recursively subdividing H_0 into successively smaller hypercubes.

   In general, at every iteration, $L = \{0, 1, 2, ...\}$, of the spatial decomposition, the $(L+1)^n$ level-L hypercubes are potentially subdivided to form a total of $2^{n(L+1)}$ smaller level-$(L+1)$ hypercubes. That is, every level-L hypercube that requires further exploration gets subdivided into 2^n level-$(L+1)$ hypercubes.

   During each subdivision iteration, L, properties are measured "within and around" each level-L hypercube (as discussed in Rule 8) for the current structure, and the measurements are aggregated and attributed to the point at the center of each hypercube, called the **aggregation point** (described in Rule 7).

   The measurements at corresponding aggregation points for each structure are compared before the next level of hypercubic subdivision continues. The subdivision of a particular level-L hypercube terminates when one or more of the following conditions are met:

   I. the aggregated measurements at the aggregation point contained within a particular level-L hypercube are deemed sufficiently dissimilar to measurements aggregated to corresponding points for the other structures,
   II. the **critical similarity threshold** for the property is reached (see Rule 9),
   III. *all* properties in the property list would become out of **analytical scope** (see Rule 9) with increased analytical level L, or when
   IV. L becomes equal to L_{max} defined for the algorithm derived from this framework.

3. **Comparison**:

(a) **Structures with Structures**:

Comparing the collection of $k \geq 2$ structures is equivalent to generating exactly one binary tree in accordance with Rule 6. The tree systematically locates the aggregation points already alluded to in Rule 2 and fully described in Rule 7. The aggregation points map to **spatial localities**, as described in Rule 8, that are deemed equivalent in all k structures with respect to the measurement of a collection, P, of properties as defined in accordance with Rule 9.

(b) **Structures with a Tree**:

Corollary: In accordance with Rule 3(a), the tree described in Rule 6 can be compared to a collection of $k \geq 1$ structures. A tree traversal (as described in Rule 4) identifies spatial localities deemed equivalent in a collection of k' previously analyzed structures. The information aggregated from the corresponding localities "within and around" the k new structures (as discussed in Rule 8) is comparable to aggregated information contained in the tree such that the tree becomes updated with aggregated information for the $k+k'$ structures.

(c) **Tree with a Tree**:

Two trees of the form described in Rule 6 can be merged into a single new tree by traversing both original trees (as described in Rule 4) and adding to the new tree all "aggregation nodes" (described in Rule 6) that correspond to spatial localities which:

I. are represented in both original trees, and
II. are deemed equivalent in terms of aggregated information contained therein.

4. **Tree Traversal**:

Traversing any tree mentioned in Rules 3(a), 3(b), and 3(c) is equivalent to locating spatial localities, described in Rule 8, "within and around" all structures associated with the tree that are deemed equivalent with respect to the properties in P.

5. **Multiple Structure Alignment:**

Corollary: In respect of Rule 4, the tree constitutes a multiple structure alignment in accordance with Definition 2.3.

6. **Binary Tree Properties**:

A single binary tree is created (or updated) for the collection of structures that systematically locates aggregation points (described in Rule 7) and maintains aggregated information. Those aggregation points included in the tree map to spatial localities (described in Rule 8) *deemed equivalent* with respect to the properties defined (in accordance with Rule 9) in P. Conversely, all aggregation points mapped to localities deemed *nonequivalent* are excluded from the tree.

Thus the tree provides a simplified structural representation of areas of similarity in the collection of n-dimensional structures.

The tree has the following properties:

(a) The height of the tree, h_{max}, must be a multiple of n; that is, $h_{max} \bmod n = 0$.

(b) Each node contains, but is not limited to the containment of, two subtrees: a "left" subtree, T_L, and a "right" subtree, T_R.

(c) The root node, or equivalently, the node at level 0, must contain all information required for the generation and regeneration of the particular recursive subdivision of H_0, and the corresponding tree. In particular, this information includes, but is not limited to:

I. the edge length of the level-0 hypercube, denoted $|H_0|$, on which the dimensions of all higher level hypercubes are based according to $|H_{i+1}| = ½ |H_i|$,
II. an identification of each structure represented by the tree, and
III. the translation and rotation vectors for each structure relative to the original coordinates specified in terms of the basis vectors in the original specification of the structures.

The information contained here is considered to *specify a particular hypercubic decomposition* for the collection of objects.

(d) Nodes at any level, h, where $h \bmod n = 0$, contain a list of aggregated measurements for the properties defined. Such nodes are referred to as **aggregation nodes**.

(e) The tree is constructed in conjunction with the successive comparative iterations that yield the hypercubic spatial decomposition of the structures in the collection as described in Rule 2.

(f) The following recurrence relation maps a tree traversal path between two aggregation nodes in the tree (which will have length equal to the number of dimensions, n) to a vector. This vector is specified in terms of S and spans

between an aggregation point inside a level-L hypercube to an aggregation point inside a level-(L+1) hypercube.

Let $\xi(x) = \begin{cases} -1 & \text{if } x = 0 \\ 1 & \text{otherwise,} \end{cases}$ where $x \in \mathbf{Z}$.

Let $\boldsymbol{v}_{L,\, b_0 b_1 \ldots b_{n-1}}$ be defined as a vector from the aggregation point in a level-L hypercube to an aggregation point in a level-(L+1) hypercube according to:

$$\boldsymbol{v}_{L,\, b_0 b_1 \ldots b_{n-1}} = \begin{cases} \mathbf{0} \text{ (the zero vector)} & \text{if } L = 0 \\ \frac{1}{4} \left| H_{L-1} \right| \sum_{i=0}^{n-1} \xi(b_i)\, \boldsymbol{e}_i & \text{if } L > 0 \end{cases}$$

where:

- $b_i = 0$ if branch i along the traversal path leads to a left subtree, and
- $b_i = 1$ if branch i along the traversal path leads to a right subtree.

Thus traversals of the tree forge paths from one aggregation node to the next and have a geometric mapping to aggregation points within the hypercubes.

7. **Aggregation Points**:

An **aggregation point** is an n-dimensional vector (or equivalently, a point in n dimensional space) that identifies the center of a level-L hypercube. The coordinates of all aggregation points can be located by traversing the binary tree structure defined in Rule 6. Traversals cause the associated recurrence relation to be expanded to yield the sought-after vectors which lead to the corresponding point from the origin (the aggregation point of the level-0 hypercube).

The aggregation point is also the point to which all summarized property measurements taken from "within and around" the level-L hypercube (as discussed in Rule 8) are attributed.

8. **Spatial Localities**:

Spatial localities are regions of n-dimensional space from which property measurements are taken.

Specifically, spatial localities have the following properties:

(a) A spatial locality exists for every aggregation point defined in Rule 7.

(b) Spatial localities are centered at the associated aggregation points.

(c) Spatial localities can be defined either as hypercubes having edge length denoted $|G_L|$, or hyperspheres having radii denoted r_L.

(d) Spatial localities have dimensions scaled according to the edge length of the associated level-L hypercube, $|H_L|$, such that either:

$$|G_L| = E(L)|H_L|, \quad \text{for hypercubic spatial localities, or,}$$

$$r_L = E(L)^{\sqrt{n}/2}|H_L|, \quad \text{for hyperspherical spatial localities (based on the radii being at least half of the diagonal length of the level-}L\text{ hypercube),}$$

where $E(L) \geq 1$ is a user-defined function that controls the degree of overlap of the spatial localities as a function of hypercube level (this spatial locality overlap is what is meant by the general phrase "within and around").

(e) Spatial localities may overlap other spatial localities.

9. **Properties**:

A property is defined as any observable structural feature capable of being partitioned in conjunction with spatial decomposition.

A **property list** must be defined such that each property satisfies the following conditions:

(a) A property must be assigned an **observation point** for establishing distances between the point of observation and the aggregation points to which portions of the measurement are attributed.

(b) Measurements of the properties yield numeric quantities that can be aggregated and assigned to appropriate aggregation points after being subjected to scaling by the distance decay function (see part (e)). Each property must have a fractional scale factor associated with it to scale its relative influence at every scope of analysis (see part (d)). The scale factors must sum to unity.

(c) A **critical similarity threshold** is defined that constitutes a measurement discrepancy value below (or above) which properties are deemed dissimilar on comparison.

(d) Properties have a finite **scope of analysis** based on the edge length of the level-L hypercube. Maximum and minimum edge lengths must be specified between which measurement of the property is conducted.

(e) A property has a defined **distance decay function**, D(r), that is used to scale the measurements in accordance with distance to the aggregation points. D(r) must become 0 as $r \rightarrow \infty$.

Note: r is the distance between an aggregation point and the observation point of the property.

Note: The property list must include an overall indication of similarity. That is, the relative contribution of each property in effect at each scope of analysis must be determined for the aggregation of measurements for Rule 9(b) above. These contributions scale each property measurement during assignment to aggregation points.

10. **Sub-structural Collocation:**

Properties may be selected for which portions of structure can be shifted (through alterations of appropriate structural data files) to bring regions of strong similarity into closer proximity. Such data file alterations, in effect, result in collocating those sub-structures that match into a "consensus" region.

Collocation is accomplished through the partial construction of the tree, searching for measurements in common in *neighbouring* regions, shifting the structures in the data files, and then rebuilding the affected portion of the tree.

The scope of this research, however, precludes complete development of this concept. It will be discussed further in the Future Work section.

Chapter 5

The Protein Comparison Method

The structure comparison framework has been presented in the previous chapter. In addition, a derivation of a protein comparison method in three dimensions has been discussed. This chapter discusses details specific to the protein comparison method developed from the general framework. In particular, the settings of critical framework parameters, the details of the computer algorithm, and the chemical properties analyzed are described here.

5.1 Settings for the Derived Protein Comparison Method

The development of a protein comparison method involves the determination of several constants, functions, and definitions required by the rules of the structure comparison framework in §4.2. These settings are determined in part by intuition based on chemical literature (where indicated below) and through preliminary calibration testing conducted on a set of training proteins, as discussed in §6.1. The following items require specification for the protein comparison method:

- The edge length, $|H_0|$ of the initial level-0 cube surrounding each protein must be determined for Rule 2 so that all proteins examined during testing readily fit within it. An acceptable value of $|H_0|$ is 256 Å (1 Å = 10^{-10} m). This was the smallest value that is a convenient power of two that allowed all proteins tested in this research to fit inside the level-0 cube. This value would have to be increased if larger proteins were being examined.
- A maximum tree height must be determined for Rule 6 so that effective comparisons can be made without excessively large trees being constructed. Tree levels become excessive when additional levels do not significantly increase comparative accuracy. A reasonable value is 9 since the similarity scores tend to have stabilized just before

this level and do not change appreciably beyond this point. This is reasonable given that $|H_0|$ is 256 Å, and 9 subdivisions results in a level of analysis close to the feature size of individual atoms.

- The overlap functions, E(L), for Rule 8 are defined as constant functions of the form E(L) = C, where C ranges from 100% to 250%. The functions are constant because there is no benefit in varying the overlap characteristics with level of analysis given the properties examined in this research.

- Tables 5.1 to 5.4 on page 68 give the properties examined for Rule 9 (a detailed discussion the chemical properties is given in §5.3). In keeping with the scope of this research, the properties examined are kept as straightforward as possible and limited to structural elements. The properties have been broadly categorized into two groups. The first group contains four properties associated with secondary structure: the proportions of residues in, respectively, α-helices, β-sheets, turns, and looped regions. The second group contains five properties associated with primary structure: the proportions of residues having radical groups that are, respectively, alipathic, aromatic, polar, positively charged at biological pH, and negatively charged at biological pH.

- Each property requires the specification of observation points in accordance with Rule 9(a). For all properties measured, a natural choice is to have the observation points coincide with the positions of the atoms giving rise to the property (usually the α-carbon atom). For example, an α-helix is considered to have observation points coinciding with all α-carbon atoms within the helix. Several observation points, then, are considered to impart an "α-helixness" that emanates from the point and decays with radial distance.

- The relative contribution of each property in effect at each level of analysis must be specified for Rule 9(b). These contributions affect the degree to which each property measurement is scaled and summarized into the scores assigned to the various aggregation points. As the cubes become smaller with increasing cubic level, the contribution to the overall similarity score gradually shifts from the Group 1

properties to the Group 2 properties. This is accomplished by scaling the measurement of each property by weighting functions that depend on the cubic level variable, L. The weighting functions were derived through testing on calibration proteins and are discussed in §5.3.2.

- Critical similarity ("cut-off") scores for the aggregated measurements must be determined for Rule 9(c) that indicate when the property measurements are dissimilar. If property measurements are dissimilar, no further exploration of the associated subdivision cube (octant) is conducted. As a result of testing on the set of training proteins, a reasonable threshold value was found to be 55% similarity. The similarity score to be compared with this threshold is found by taking the ratio of the extreme values measured in the collection of proteins compared as discussed in §5.2.2.
- In accordance with Rule 9(d), maximum and minimum edge lengths of subdivision cubes must be determined for each property that define the scope of analysis over which the property is measured. These edge lengths correspond to the cubic level variable, L, where the property is in effect as governed by the weighting functions. That is, for this research, properties are defined to be "in scope" at all levels of analysis, but their contributions are governed by weighting functions discussed in §5.3.
- A decay function, $D(r)$, must be defined for each property for Rule 9(e) that is applied to properties aggregated from outside the current cubic region but within the cubic locality region (referred to as the "fringe region"). $D(r)$ has been derived as follows:

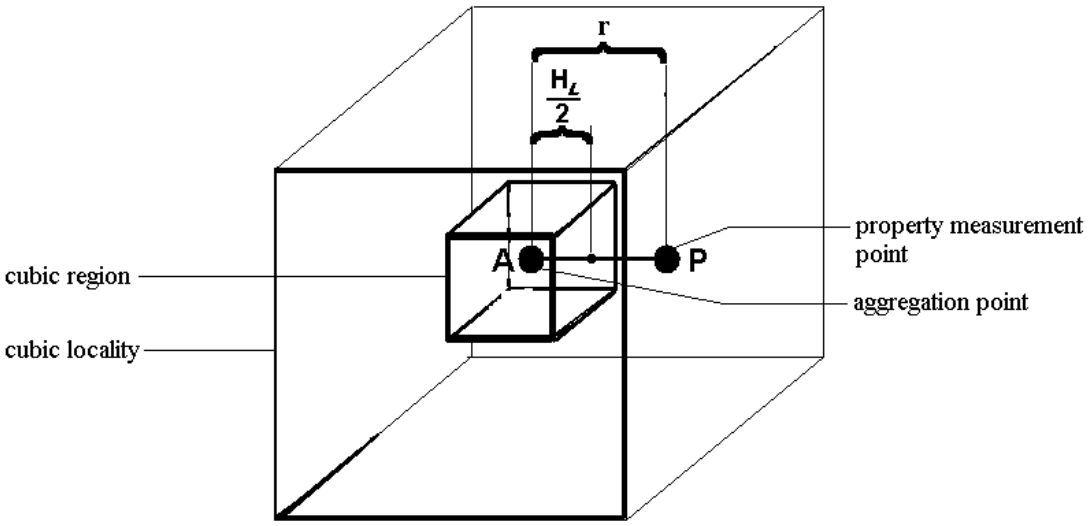


Figure 5.1: Applying the Decay Function to a Property in the Fringe Region

The domain of D(r) is $r \geq \frac{H_L}{2}$

The desired range of D(r) is $0 < D(r) \leq 1$

This range is achieved by defining the decay function as,

$$D(r) \equiv \frac{H_L^3}{8r^3}$$

Since,

$$r = \sqrt{(A_x - P_x)^2 + (A_y - P_y)^2 + (A_z - P_z)^2} \geq \frac{H_L}{2}$$

Multiplying equation by $\frac{2}{H_L}$ gives,

$$\frac{2r}{H_L} = \frac{2\sqrt{(A_x - P_x)^2 + (A_y - P_y)^2 + (A_z - P_z)^2}}{H_L} \geq 1$$

Given that $H_L > 0$ and $r > 0$, taking the reciprocal of the equation gives,

$$\frac{H_L}{2r} = \frac{H_L}{2\sqrt{(A_x - P_x)^2 + (A_y - P_y)^2 + (A_z - P_z)^2}} \leq 1$$

Cubing the above equation,

$$\frac{H_L^3}{8r^3} \leq 1$$

Furthermore, as $r \rightarrow \infty$,

$$\frac{H_L^3}{8r^3} \rightarrow 0$$

Thus, $0 < D(r) \leq 1$.

With the above definition, the value of $D(r)$ degrades with the cube of the radius. Moreover, since the volume increases by the cube of the radius, property contribution around the aggregation point is adjusted for volume.

5.2 Details of the Computer Algorithm

The protein comparison method developed here should be considered as a "proof-of-concept", demonstrating that a structure comparison method adequate for the classification of proteins in reasonable time can be constructed using the structure comparison framework as a template. The details of the computer algorithm, the similarity scoring scheme, the output of the computer program, and the format of the octtree file are discussed in this section.

5.2.1 Property Detection Algorithm

The processing involved with a cubic spatial decomposition has the potential to increase exponentially with increasing level of analysis because every cube is recursively divisible into eight sub-cubes. This section discusses how such an exponential increase is avoided by systematically locating property observation points and processing only the affected aggregation points. Cubic octants and the associated aggregation points are

examined and subdivided only if they are affected by properties within the associated spatial localities. Thus it is the presence of properties at a given level of analysis that governs the processing involved with spatial decomposition and not simply the entire accumulation of cubes at a given depth of the cubic lattice.

Before the algorithm can be presented, the notion of the fringe locality and fringe aggregation point must be discussed to simplify the description of the algorithm. The **fringe locality** of an aggregation point, A, includes simply the volume of the spatial locality corresponding to A, less the cubic region corresponding to A. A **fringe aggregation** point is an aggregation point with which only property observation points in the fringe locality are associated. That is, property observation points exist in the spatial locality of the aggregation point, but not within the cubic region of the aggregation point. Point A in Figure 5.1 is a fringe aggregation point because it has no property points in the cubic region corresponding to it, but has property point P corresponding to it in the fringe region.

The following efficient algorithm has been used to compare proteins on the basis of the property list already discussed.

ProteinsCompare (PDBList */*list of two or more PDB files for proteins being compared*/*)

1 From the protein data bank files, build three property observation point lists for the collection of proteins sorted by x-coordinate, y-coordinate, and z-coordinate. These coordinate lists are called, respectively, the x-list, y-list, and z-list. Mark all points in the x-list UNPROCESSED */*Note: It is not necessary to mark the y-list and z-list*/* Retain with every point added to the x-list the molecule identifier to which each property observation point belongs.

2 Initialize the level of analysis variable, L: $L \leftarrow 0$.

3 Initialize the octtree for processing at $L = 0$. */*The structure and operation of the octtree is discussed in the next section.*/*

4 **Loop** L from 0 to the chosen maximum value of L, L_{max}, or until all property observation points in the x-list have been marked COMPLETE */*meaning that the proteins in the collection have been found dissimilar and no octants remain to be subdivided*/* ...

a **While** property observation points exist in the x-list that are marked UNPROCESSED ...

1 The first property observation point, *P*, is obtained from the x-list that is marked UNPROCESSED.

2 Find the aggregation point, *A*, closest to *P*.

3 Remove *A* from the fringe-list (by marking it PROCESSED) if *A* has been added to this list from step 4-a-5-b.
*/*It is recommended that the reader ignore this step when reading this algorithm for the first time. Understanding this step is straightforward after reading step 4-a-5-b.*/*

4 Ascertain whether the cubic region corresponding to *A* is to be subdivided by referencing the octtree. A cubic region is to be subdivided if $L = 0$, or if a path exists in the octtree to the parent octant at level L-1 currently under subdivision.

5 **If** the region corresponding to *A* is to be subdivided,

a Use the coordinate lists to find all property observation points within the spatial locality of *A* and process each point.
*/*Processing a point involves marking the point PROCESSED in the x-list, examining the property, and adding its scaled value to the aggregated measurement information for the associated molecule. Scaling the measured value entails multiplying the measurement by 1 if the point is within the cubic region of the aggregation point, and multiplying the measurement by the decay function, D(r), if the point is in the fringe region.*/*

b Find the collection of aggregation points corresponding to the collection of property observation points within the cubic region (not including the current aggregation point).
*/*These are the aggregation points that are affected in some way by the property points in the cubic region under current examination. These aggregation points are potential fringe aggregation points, and are added to a list of points called the fringe-list. Most of the points in the fringe-list will be removed in step 4-a-3 if they are not, in fact, fringe aggregation points.*/*

6 Otherwise, **if** *A* is not to be subdivided,

a use the coordinate lists to find all points within the cubic region of *A* and mark each point COMPLETE. */*Points marked COMPLETE will be ignored by subsequent iterations of the algorithm.*/*

b **While** aggregation points exist in the fringe-list that are marked UNPROCESSED ...

1 Obtain first aggregation point from the fringe-list that is marked UNPROCESSED.

2 Mark aggregation point as PROCESSED.

3 Ascertain whether the aggregation point is a fringe aggregation point by ensuring that no points exist in the cubic region of the aggregation point.
*/*If points exist in the cubic region, the aggregation point entered the list by way of processing neighbouring aggregation points in step 4(a) after the current aggregation point was processed. Understanding the details of this normal situation is probably not worth the effort by the reader. It is sufficient to understand that this test is a necessary part of the algorithm.*/*

4 **If** the aggregation point is a fringe aggregation point,
- a Ascertain whether the cubic region corresponding to *A* is to be subdivided by referencing the octtree.
- b **If** the region corresponding to *A* is to be subdivided,
  - use the coordinate lists to find all property observation points within the fringe region of *A* and process each point */*of course, there will be no points in the cubic region of A*/*

b **Delete** the fringe-list (and re-initialize a new fringe-list for the next iteration).

5 Display the similarity scores of the comparison as discussed in §5.2.2.

A far simpler paradigm could be proposed as follows:

1 For each level of analysis, find all the aggregation points.
2 Process them.
3 Display the similarity scores of the comparison as discussed in §5.2.2.

However, this simpler paradigm disallows processing aggregation points one at a time and requires the storage of aggregated measurements for a potentially huge number of aggregation points at higher levels of analysis. The former algorithm is more complicated but is nevertheless efficient, and handles aggregation points one at a time.

5.2.2 Similarity Scoring Scheme

Recall from Chapter 4 that a cube enclosing each protein is subdivided recursively over successive iterations of the algorithm, and this subdivision process is governed by the formation of an octtree (the representation of the octtree is discussed in §5.2.1.2). In

addition to governing subdivision, the octtree also provides the structure from which raw similarity scores are generated at each level of analysis. During octtree formation, branches at a given level identify cubes whose aggregated measurements of properties in all molecules are found similar. A given cubic region of space encloses *similar* portions of the molecules if a path exists in the octtree to that cubic region; the region encloses *dissimilar* portions otherwise. This section discusses how aggregated measurements are determined to be similar or dissimilar, the **aggregate raw similarity score** (provided at each level of analysis), and the **cumulative volume-adjusted** (**CVA**) **similarity score** (provided at every level of analysis). In addition, two overall scores indicating similarity: the **minimum CVA similarity score**, and an **extrapolated** minimum CVA similarity score (provided whenever the algorithm is forced to terminate before reaching the highest level of analysis specified).

The principal governor of property measurement and aggregation throughout the course of the algorithm is the level of analysis, *L*. At each level of the algorithm, properties are measured around every aggregation point affected by at least one property and separate property measurements are maintained for each molecule. After an aggregation point has been processed (and before moving on to the next aggregation point), the measurements for each property are combined into an aggregate score for each molecule and these scores are compared (the actual properties measured and the weighting functions are discussed in §5.3). On a property-by-property basis, the lowest and highest measurements of each property (one measurement comes from each molecule) are taken and a **simple property score**, **M**, is calculated for each property,

$$M(P_i) = \frac{\text{lowest measurement of property } P_i}{\text{highest measurement of property } P_i} \times 100\%.$$

These simple property scores are then combined into an **octant decision score** using property weighting functions, $W_i(L)$, which depend on the level of analysis,

$$\text{octant decision score} = \frac{M(P_1)\times W_1(L) + M(P_2)\times W_2(L) + \ldots + M(P_n)\times W_n(L)}{W_1(L) + W_2(L) + \ldots + W_n(L)}$$

The octant decision score is used to determine whether the current octant should be further subdivided at the next level by comparing it with the exploration threshold value. The octtree is updated to indicate whether the octant corresponding to the currently processed aggregation point should be subdivided at the next level. Further subdivision will occur only if the octant decision score is greater than or equal to the exploration threshold value. As a simplifying condition for this research, the octtree indicates only whether octants are similar or dissimilar. No other property summary information is maintained in the octtree.

After all required aggregation points have been processed at a given level of analysis, the octtree will have been completely updated to that level. After completing the aggregation of measurements at a given level of analysis, an overall raw similarity score is provided based on the number of octants determined to be similar and the total number of octants examined at that level. This score is called the **aggregate raw similarity score**, $\mathbf{R}_L$, and is defined as,

$$\mathbf{R}_L = \frac{\text{number of octants deemed similar at level } L}{\text{total number of octants examined at level } L} \quad .$$

From the aggregate raw similarity scores over successive levels of analysis, a score called the **cumulative volume-adjusted** (**CVA**) **similarity score** is calculated at each level (except for level 0) which compensates for octant volume and for the fact that octants are only examined if they are affected by properties. It is defined as,

$$CVA_1 = R_0 ,$$

$$CVA_L = \frac{T_L}{8\,T_{L-1}} \times R_L + \left(1 - \frac{T_L}{8\,T_{L-1}}\right) \times CVA_{L-1} .$$

where:

- R_L is the aggregate raw similarity score for level of analysis L, and
- T_L is the total number of octants examined at level of analysis L.

Note: CVA_0 does not exist.

The CVA similarity score is composed of two terms. The first term scales the raw similarity score, R_L , by the proportion, $T_L / (8\ T_{L-1})$, that indicates how many of the original sub-cubes were subdivided between level *L*-1 and level *L*. The maximum number of sub-cubes that can be generated between levels *L*-1 and *L* is $8\,T_{L-1}$. The actual number of cubes subdivided at level *L*, T_L , is less than or equal to maximum number of subcubes $8\,T_{L-1}$ that can result. This ratio depends on the distribution of properties in the cubic regions at level *L*-1. The second term scales the CVA score from level *L*-1 by the ratio that indicates how many of the original subcubes were *not* subdivided between level *L*-1 and level *L*. In effect, this assigns the "prevailing" similarity score to the property-devoid regions interspersed amongst the sub-cubic regions housing properties. During the calibration testing, this scheme has been found to lead to smooth scoring with level of analysis progression and a fair overall similarity score. Exclusive counting of regions with properties was found (with unrecorded testing during program) to lead to harsher scores that failed to reflect the more subtle changes in molecular properties actually observed.

5.2.3 The MolCom3D Program

Figure 5.2 shows the output of the program, **MolCom3D**, with two protein molecules being compared using 7 levels of analysis and a cubic overlap of 250 percent. With the "-v" flag provided, the output shows the progress of the program, a table of summary results for each level of analysis, and finally an average CVA score for the proteins (the progress of the program is not shown without the "-v" flag). In the table of summary results, the first column gives the level of analysis. The next three columns give, respectively, the total number of octants examined at each level, the number of similar octants, and the number of dissimilar octants for each level. The last two columns give the aggregate raw similarity score and the CVA score for the level. After the table, an indication of whether the molecules are predicted to be "similar", "somewhat similar", or "different" is given, based on minimum CVA scores of 80% or more, under 80% but greater or equal to 70%, or under 70%, respectively.

```
\MolCom3D>MolCom3D -v -O 250 -L 7 -f 1KBS.pdb ref_1KBS_2CRT.pdb -allcarbons
.     .          .      .       .   .        .      .         .      .
.  M. .  . M     O O .   L .      c C    .O O     M. .  .M    3 3 .   D D .   .
    m  .  MM.       O       .   C    .     .  O     m  .    .     3    D   d
 .   M   m M    O        l  .      .      O        .   m  M     33  .    .  D  .
   M    m  .     .  o .         C     .       O.   M  M           3    D   d
       M         O O     L L L    c c     0 0     M .    M .  3 3  .  d  D .
.  M    .  M          .         .          .     .   .         .      .      .
           .          MolCom3D - Molecular Comparison Software
.         .           Proof-of-Concept Version 0.5a by Stephen O'Hearn, 1999   .

Building lists of atomic coordinates and property lists ...
The "-allcarbons" flag has been given.
Building lists will require much time.


Analyzing the following protein molecules:
.\1KBS.pdb
.\ref_1KBS_2CRT.pdb

Cubic Overlap:  250.0 %.  Maximum level of analysis:  7.

Measuring properties at all CARBON atom positions.
Much time will be required to perform calculations.

Processing level 0 of 7 (cube edge length = 256.00 angstroms)
Processing level 1 of 7 (cube edge length = 128.00 angstroms)
Processing level 2 of 7 (cube edge length = 64.00 angstroms)
Processing level 3 of 7 (cube edge length = 32.00 angstroms)
Processing level 4 of 7 (cube edge length = 16.00 angstroms)
Processing level 5 of 7 (cube edge length = 8.00 angstroms)
Processing level 6 of 7 (cube edge length = 4.00 angstroms)
Processing level 7 of 7 (cube edge length = 2.00 angstroms)
Molecular analysis complete.

                                                          cumulative volume-
level    total      similar     dissimilar    similarity   adjusted (CVA)
         octants    octants     octants       score (%)    similarity score (%)
-------------------------------------------------------------------------------
    0          1          1              0       100.00                 N/A
    1          8          8              0       100.00              100.00
    2         16         16              0       100.00              100.00
    3         49         47              2        95.92               98.44
    4         64         62              2        96.88               98.18
    5        159        136             23        85.53               94.25
    6        441        368             73        83.45               90.51
    7       1374       1285             89        93.52               91.68

Minimum cumulative volume-adjusted (CVA) similarity score ..... 90.5 %
*******************************************************************************
* Molecules have been judged                                          SIMILAR *
*******************************************************************************

\MolCom3D>
```

Figure 5.2: Output from the MolCom3D Program

The reader is reminded that the total number of octants examined between any two successive levels of analysis does not necessarily increase by a factor of eight because one or more of the octants at the higher level may be devoid of properties. Empty octants, however, contribute to the score by assigning the "prevailing" similarity score to the property-devoid regions interspersed amongst the sub-cubic regions housing properties. Under this convention, the empty space does not dominate the score (see §5.2.2 for a review on scoring).

In addition to the textual output, MolCom3D builds an octtree file called **octtree.binary** in the subdirectory containing MolCom3D (or in any other desired directory as configured). Currently, this file is in a binary format and is used exclusively by the program. The format of this file is described in the next section.

After the spatial comparison of the molecules has completed, MolCom3D uses the resulting octtree file to generate a PDB file for each protein containing only those portions of the molecules that are within cubic regions that have been deemed similar during the analysis. The top of Figure 5.3 shows the two original proteins compared during the program execution shown in Figure 5.2. The bottom of Figure 5.3 shows the corresponding Rasmol images from the PDB files regenerated by MolCom3D. Enlarged versions of these figures appear in Appendix A. These PDB files are named with the following convention:

"sim_overlap_<*overlap*>_lmax_<*maximum level*>_<original molecule name>.ent"

For the molecules examined, for example, two PDB files are generated called:

"sim_overlap_250_lmax_7_pdb2crt_.ent" and "sim_overlap_250_lmax_7_pdb1kbs_.ent"

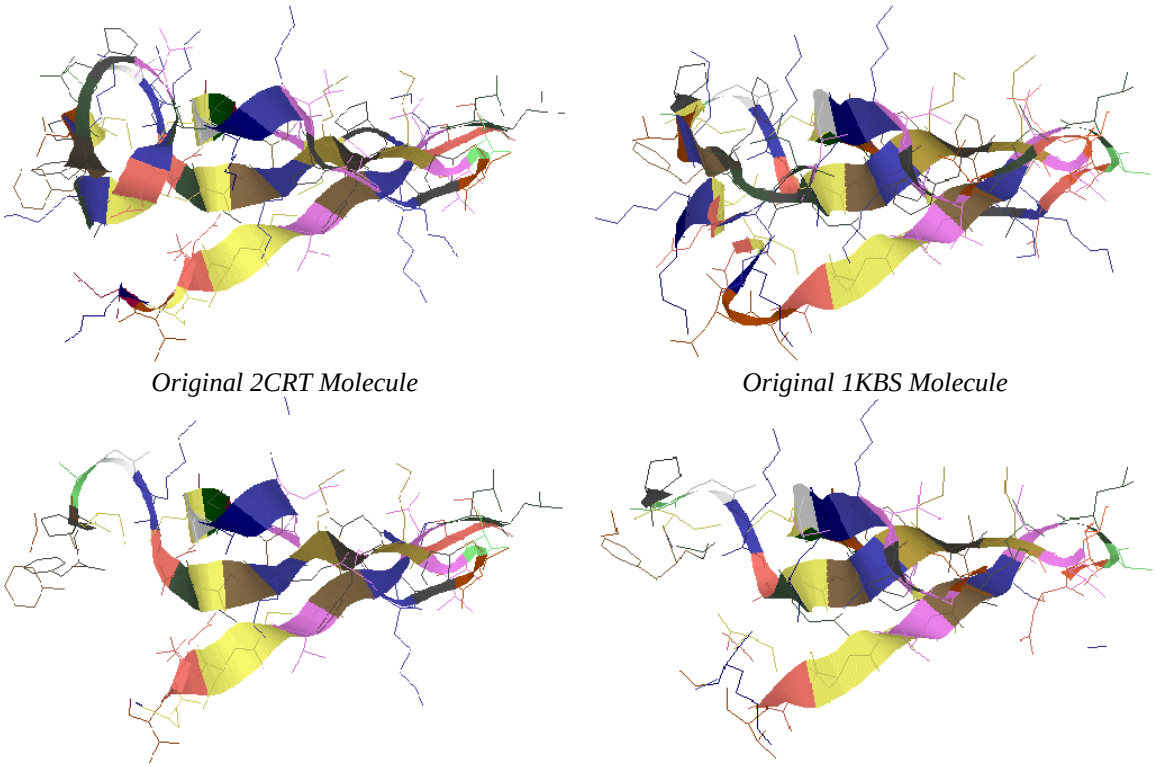


Figure 5.3: Similarity Indication in 2CRT and 1KBS

5.2.4 Details of the Octtree File Structure

Recall that the octtree constructed for this research only contains information about whether octants are similar or dissimilar. This is accomplished by providing paths to the similar octants. No other property summary information is maintained in the octtree. This simplification discounts Rule 3(b) of the structure comparison framework (the ability to compare structures with a tree), but the smaller octtrees were desirable for developing and testing the above "proof-of-concept" algorithm. Furthermore, extra information can be economically added using separate but corresponding files if required for future implementations. This is discussed in the Future Work section.

The octtree is represented as an octary tree (a tree with eight branches per node) in a file called **octtree.binary** which is created with every execution of the program. The octary tree structure is functionally equivalent to the binary tree structure described in Rule 6 of the structure comparison framework, but is more economical to implement using the raw bytes of a binary file. The general form of the octary tree file is a collection of five-byte records as shown in Figure 5.4.

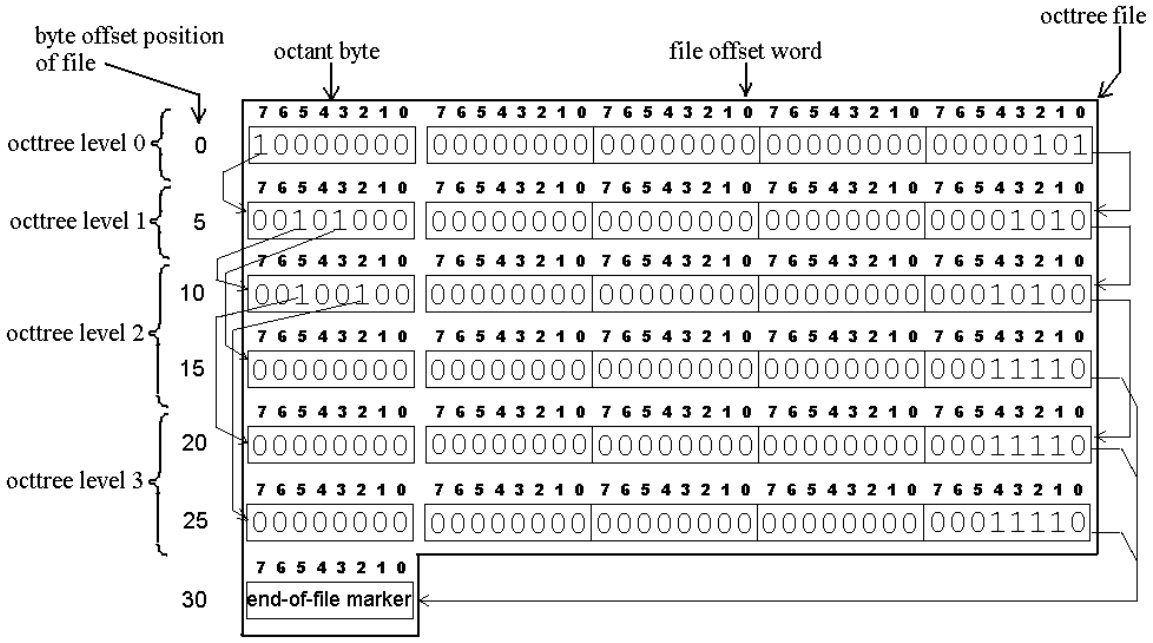


Figure 5.4: Structure of the"octtree.binary" File

Each five-byte record represents exactly one cube (or equivalently, one octant) and consists of an **octant byte** and a **file offset word** (four bytes). Since this implementation was straightforward and was sufficient for accomplishing the goals of this research, other implementations in the octtree literature were not considered. The *usage* rather than the implementation of the octree was the focus of this research, and other research using octtrees has been treated in §2.3.

It is helpful to remember that this octtree structure *represents paths to similar octants*. To ascertain whether an octant at a given level of analysis exists in the tree, the aggregation point at the center of the octant is sought using the following "OcttreeOctantExists" algorithm.

OcttreeOctantExists (A_x, A_y, A_z /* aggregation point */, L/* level of analysis */)

1 /* initialize local variables */
currentFileOffset ← 0
(x, y, z) ← (0, 0, 0) /* (x, y, z) is the current position – searching for the desired octant is equivalent to searching for the aggregation point (Ax, Ay, Az) at level L starting at the origin */
pathExists ← FALSE
curLevel ← 0
RESET octtree.binary file

2 WHILE (curLevel ≤ L) ∧ ¬pathExists ∧ ¬end-of-file

a /*
find the bit position of the current octant that leads to the next aggregation point by comparing each coordinate of the current location (x, y, z) to the sought-after location, the aggregation point at level L (Ax, Ay, Az)
*/
vector $_i$ ← 000_2 /* subscript 2 implies binary number representation */
vector $_j$ ← 000_2
vector $_k$ ← 000_2

IF ($A_x \geq x$)
vector $_i$ ← 001_2
IF ($A_y \geq y$)
vector $_j$ ← 010_2
IF ($A_z \geq z$)
vector $_k$ ← 100_2

/*
bit-wise OR of vectors gives the octtree branch number to traverse or the sought-after octant when curLevel becomes equal to L
*/
octantByteBitPos := vector $_i$ | vector $_j$ | vector $_k$

b /*
find the vectors to the next aggregation point
the vector sense function, x(x), is defined in framework Rule 6(f)
H_0 is defined 256Å
*/
$x \leftarrow x + \xi(\text{vector}_i) \times H_0 \times (1/2)^{curLevel+1}$
$y \leftarrow y + \xi(\text{vector}_j) \times H_0 \times (1/2)^{curLevel+1}$
$z \leftarrow z + \xi(\text{vector}_k) \times H_0 \times (1/2)^{curLevel+1}$

c octantByte ← readOctantByteFromFile (octtree.binary file)

d IF (octantByteSet [octantByteBitPos]) ∧ (A_x = x) ∧ (A_y = y) ∧ (A_z = z)
pathExists ← TRUE /* octant has been found */

ELSE
currentLevel ← currentLevel + 1
preceedingOctants ← the number of bit positions set to 1 to the left of octantByteBitPos
firstSubCubePos ← readFileOffsetWordFromFile (octtree.binary file)
goToByteOffset (firstSubCubePos + 5 * preceedingOctants)

3 RETURN pathExists

If, for example, the octtree in Figure 5.4 (page 63) is searched for the presence of similarity in the right-top-front sub-cube of the right-top-back sub-cube of the level-0 cubic region, the following events occur. The algorithm begins by examining the first byte of the octtree file. Bit position 7 (the only bit that is examined at level 0) has the setting '1'. This means that the level-0 cubic region has been analyzed and all molecules are similar with respect to the properties examined within and around this cubic region. The file offset pointer is moved to offset 5. The currently sought-after vector location changes from (0, 0, 0) to (64, 64, -64) given that the value of H_0 is 256 Å. From this, the bit position of the desired path is determined to be 2x011 (from disjunction of the z-branch, y-branch, and x-branch bit positions), or equivalently, branch 3 in decimal notation. Bit position 3 has the setting '1' in the octtree. Thus the right-top-back sub-cube of the level-0 cubic region exists. The offset of this sub-cube is calculated by referencing the offset contained in the file offset word and adding the number of higher-order bit positions having the setting '1', multiplied by 5 bytes apiece. This gives: 20 bytes + 1 higher-order 1-bits × 5 bytes per higher-order 1-bits = 25 bytes from start of file. The file position is moved to 25 and an examination of the octant byte shows that all bits are zero; although the sought-after octant exists, no similarity exists anywhere in this octant.

Figure 5.5 shows an output listing from MolCom3D resulting from the comparison of three molecules at 110% overlap to 3 levels of analysis. Under the title "Final octtree" in the listing, the octtree file contents are displayed (this display is provided whenever MolCom3D is compiled with a macro, called FINAL_OCTTREE, defined as 1 rather than 0). The following number base indicator convention is used in the output listings: 0x *NUMBER* indicates that *NUMBER* is a hexadecimal quantity (base-16); 2x *NUMBER* indicates that NUMBER is a binary quantity (base-2); *NUMBER* by itself indicates that *NUMBER* is simply a decimal number (base-10).

```
\MolCom3D>MolCom3D -v -f pdb1era_.ent pdb1nxb_.ent pdb1kbs_.ent -O 110 -L 3
                                        .
                                        .
                                        .
                                                                  cumulative volume-
level    total      similar     dissimilar    similarity    adjusted (CVA)
         octants    octants     octants       score (%)     similarity score (%)
-------------------------------------------------------------------------------
    0          1           1             0       100.00                N/A
    1          8           7             1        87.50              87.50
    2          8           6             2        75.00              85.94
    3          7           3             4        42.86              81.23

Minimum cumulative volume-adjusted (CVA) similarity score ..... 81.2 %
*******************************************************************************
* Molecules have been judged                                          SIMILAR *
*******************************************************************************
Final octtree
0x00000000 2x10000000 0x00000005                    0 2x10000000           5
0x00000005 2x11101111 0x0000000A                    5 2x11101111          10
0x0000000A 2x00000001 0x0000002D                   10 2x00000001          45
0x0000000F 2x00000010 0x00000032                   15 2x00000010          50
0x00000014 2x00000100 0x00000037                   20 2x00000100          55
0x00000019 2x00010000 0x0000003C                   25 2x00010000          60
0x0000001E 2x00100000 0x00000041                   30 2x00100000          65
0x00000023 2x00000000 0x00000046                   35 2x00000000          70
0x00000028 2x10000000 0x00000046                   40 2x10000000          70
0x0000002D 2x00000001 0x0000004B                   45 2x00000001          75
0x00000032 2x00000010 0x0000004B                   50 2x00000010          75
0x00000037 2x00000000 0x0000004B                   55 2x00000000          75
0x0000003C 2x00010000 0x0000004B                   60 2x00010000          75
0x00000041 2x00000000 0x0000004B                   65 2x00000000          75
0x00000046 2x00000000 0x0000004B                   70 2x00000000          75
0x0000004B End Of File                             75 End Of File

\MolCom3D>_
```

Figure 5.5: Small"octtree.binary" File Example

The right-most three columns list, respectively, the file byte offset position, a binary display of the octant byte, and the file offset of the octant corresponding to the first

set bit position of the current octant. The left-most three columns give the same information, albeit with the offset positions represented in hexadecimal.

5.3 Chemical Properties and Weighted Comparisons

So far, this chapter has discussed the details of the protein comparison algorithm and the scoring scheme used to aggregate and compare measurements of chemical properties during the spatial decomposition of molecules. This section discusses the characteristics of these chemical properties and how each property's contribution to the octant decision score varies over successive levels of analysis of the algorithm.

5.3.1 Property Comparison in this Research

MolCom3D compares properties on the basis of the α-carbon atoms of the amino acid residues by default. However, it can also compare all carbon atoms, or all atoms irrespective of type. The desired unit of comparison is selected by specifying one of the command line arguments: "-alphacarbons", "-allcarbons", or "-all". Comparing all atoms in the molecules requires considerably more processing time than only comparing carbon atoms. In the tables, the "unit" can be either amino acid residues, carbon atoms, or atoms of any type. If the unit is the amino acid residue, the property observation point is considered to be located at the α–carbon position. Otherwise, the property observation point is considered to coincide with the atom position.

Two groups of properties involving secondary structure (Group 1(a) and Group 1(b)), and two groups of properties involving primary structure (Group 2(a) and Group 2(b)) are examined in this research and are listed in Tables 5.1 to 5.4. The properties included in Groups 1(a) and 2(a) are expected to exert a stronger influence on the overall shape and biological activity of the molecules than the Groups 1(b) and 2(b) counterparts [40, 26]. This is based on the expectation that properties more highly conserved through

evolution (those listed in groups 1(a) and 2(a)) are more influential [40] and serve as stronger indicators of similarity.

Secondary Structural Properties (More Influential)	
Property	Property Description
Helix Proportion	Proportion of units within helical structures
Sheet Proportion	Proportion of units within sheet structures

Table 5.1: Group 1(a) Properties

Secondary Structural Properties (Less Influential)	
Property	Property Description
Turn Proportion	Proportion of units within turns
Loop Proportion	Proportion of units within looped regions

Table 5.2: Group 1(b) Properties

Primary Structural Properties (More Influential - Hydrophilic)	
Property	Property Description
Negative Proportion	Proportion of units comprising negatively charged residues
Positive Proportion	Proportion of units comprising positively charged residues
Polar Proportion	Proportion of units comprising polar residues

Table 5.3: Group 2(a) Properties

Primary Structural Properties (Less Influential - Hydrophobic)	
Property	Property Description
Aromatic Proportion	Proportion of units comprising aromatic residues
Alipathic Proportion	Proportion of units comprising alipathic residues

Table 5.4: Group 2(b) Properties

5.3.2 Weighting Functions

As the level of analysis of the algorithm increases, the scoring influence gradually shifts from the Group 1 properties to the Group 2 properties. This is accomplished through level-dependent weighting functions whose values decline with increasing levels of analysis for the Group 1 properties, and increase with increasing levels of analysis for the Group 2 properties. The weighting functions are sigmoid curves of the form:

$$W(L) = \frac{ab + cL^d}{b + L^d}$$

The empirical weighting coefficients a, b, c, and d must be derived through mathematical software [23]. The sigmoid curves lead to a smoother transition of property mixture with level of analysis progression. Sigmoid curves cause transitions in property contributions that are expected to parallel the "S-shaped" property influences in nature [26]. The weighting coefficients used in this research for each property group are listed in Table 5.5. Sigmoid curves have been fit to sets of discrete values that provide approximate weighting factor trends established by chemical intuition and verified through calibration testing (discussed in Chapter 6). Table 5.6 lists the set of discrete values for each group used to generate the sigmoid curves that yielded the most accurate similarity scoring during calibration testing. The corresponding weighting curves are shown in Figure 5.6.

Coefficients of W(L) Used in the MolCom3D Program

Group	Sigmoid Function Coefficient			
	coefficient a	coefficient b	coefficient c	coefficient d
1(a)	0.22074	0.003587	6.0706	–3.5938
1(b)	0.11037	0.003587	3.0353	–3.5938
2(a)	1.98130	34981	6.0644	6.5523
2(b)	0.99065	34981	3.0322	6.5523

Table 5.5: Coefficients of Sigmoid Weighting Function

Discrete Sigmoid Curve Generator Weighting Values				
Level, *L*	Desired Approximate Weighting Factor At Level L, W(*L*)			
	Group 1(a)	Group 1(b)	Group 2(a)	Group 2(b)
1	6.00	3.00	2.00	1.00
2	6.00	3.00	2.00	1.00
3	5.00	2.50	2.00	1.00
4	4.00	2.00	3.00	1.50
5	3.00	1.50	4.00	2.00
6	2.00	1.00	5.00	2.50
7	2.00	1.00	6.00	3.00
8	0.50	0.25	6.00	3.00
9	0.50	0.25	6.00	3.00
10	0.50	0.25	6.00	3.00
11	0.50	0.25	6.00	3.00
12	0.50	0.25	6.00	3.00
13	0.50	0.25	6.00	3.00
14	0.50	0.25	6.00	3.00

Table 5.6: Discrete Weighting Factors

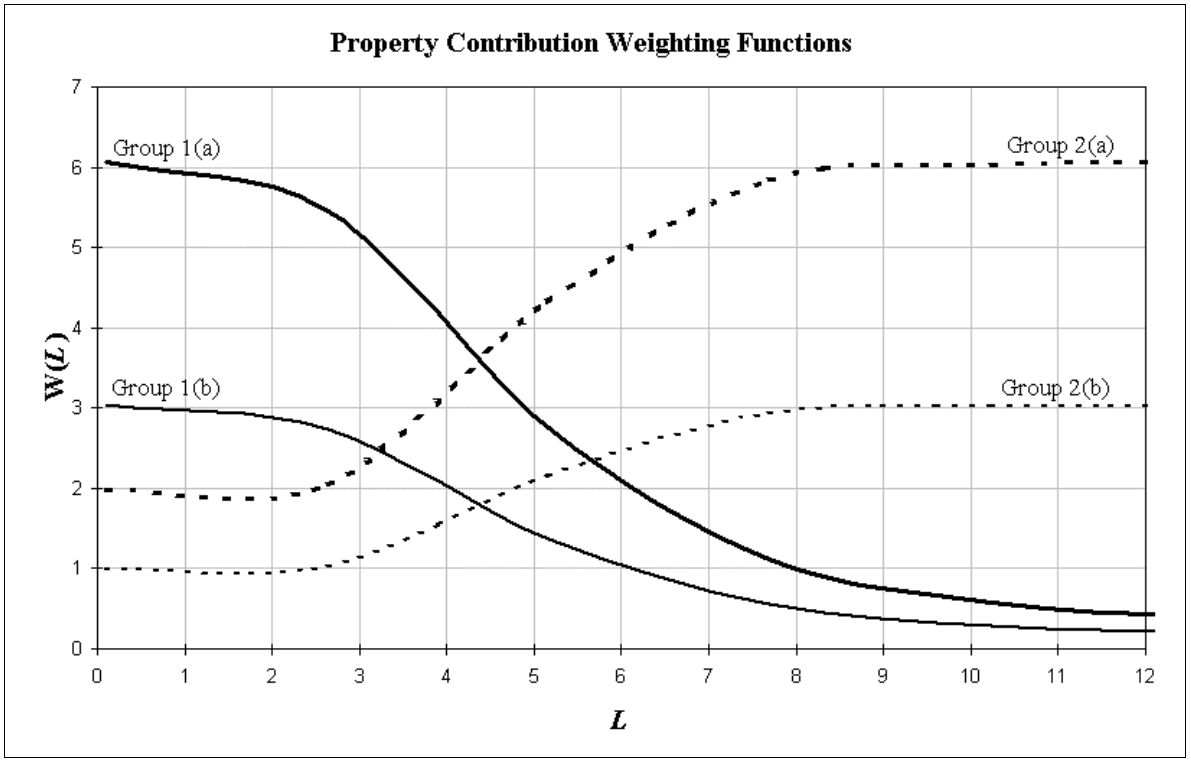


Figure 5.6: Weighting Function Sigmoid Curves

As the level of analysis, *L*, advances, primary structural similarity comes to influence the octant decision score more strongly than secondary structural similarity. In each group, the less influential sub-groups are half as influential at every level as the more influential sub-groups as indicated by their relative weighting functions.

Chapter 6

Testing of the Protein Comparison Method

Two phases of testing were conducted in developing the protein comparison method: **calibration testing** and **verification testing**. Distinct sets of proteins were used during each phase of testing. Calibration tests were conducted during the development of the method in order to establish reasonable property weighting functions and settings for some of the method parameters identified in Chapter 5. Verification tests were conducted after the development of the method with proteins not used during the calibration to confirm that the calibrated protein comparison method indeed functions as expected.

In keeping with the scope of this research, the framework and its derived method requires an approximate initial alignment for the objects under comparison. Consequently, proteins were rotated and translated into initial alignment orientations by a program called **lsqkab** (version 3.4) for subsequent comparison by the MolCom3D program. Lsqkab is part of a suite of protein crystallography programs created in accordance with the Collaborative Computational Project (CCP4), an initiative undertaken by the U.K. Biotechnology and Biological Sciences Research Council [47]. This software was made available for this research in the laboratory of Dr. Delbaere at the Department of Biochemistry, University of Saskatchewan.

The right-hand side of Figure 6.1 illustrates a protein being translated and rotated into a new orientation that approximately matches a reference protein on the left. The two proteins are shown at the bottom of Figure 6.1 oriented similarly. Orientation is based on minimizing the RMS distance between the two structures. (The actual proteins shown, PDB identifiers 1AHO (left) and 1NRA (right), are from the same SCOP classification, 1.7.3.6.1. As expected, they are quite similar in structure having an RMS deviation of 4.6 Å.) In addition to providing initial alignments, lsqkab also provided the RMS deviation scores used to verify the scores given by the protein comparison method. As a future alternative for providing approximate initial alignments, the Future Work

section discusses a simple method stemming from this research. This method is applicable to objects in general and is not limited to the alignment of proteins.

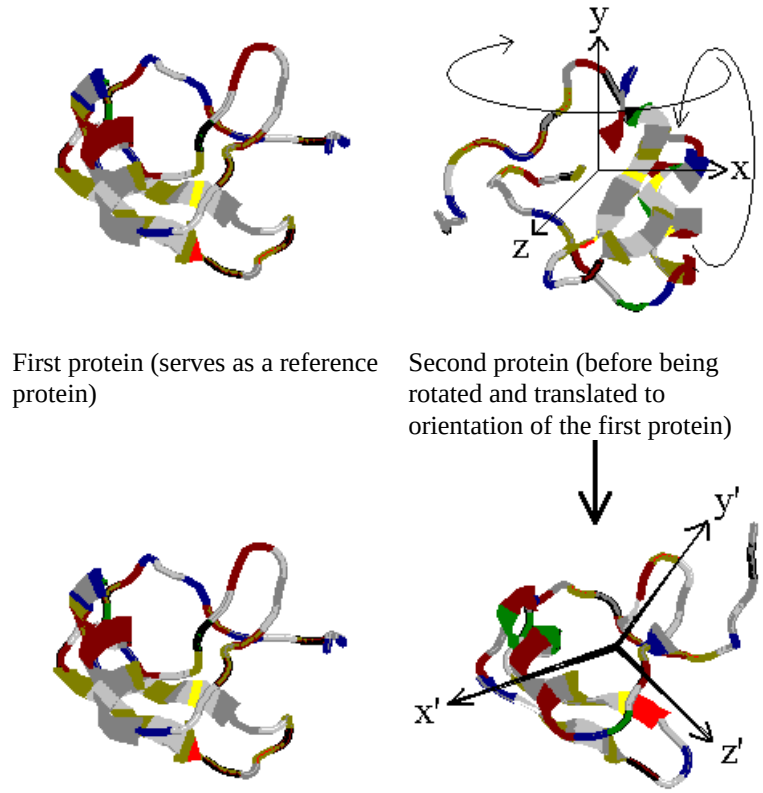


Figure 6.1: Approximate Initial Orientation Alignment of Proteins

In both phases of testing, pairs of structures with varying degrees of similarity were compared by the MolCom3D program. The RMS deviations from the lsqkab program were considered to be reasonable benchmark indicators of similarity for the pairs of structures. The RMS values ranged from roughly 0 Å to 20 Å. The minimum-CVA similarity scores produced by MolCom3D were validated against three sources of information:

- lsqkab-provided RMS deviations,
- **Structural Classification of Proteins** (**SCOP**) database [32] introduced in §2.2, and
- visual inspection.

The principal and quantitative verifier of similarity was the RMS deviation. The accuracy of the CVA similarity scores was verified by considering the proportion of protein pairs correctly classified under the "70-10 rule" by MolCom3D. The same classification was used in both the calibration and verification testing phases of this study. An RMS deviation at or below 10 Å was considered to indicate similarity and an RMS deviation above 10 Å was considered to indicate dissimilarity. This value was set for this research based on visually inspecting pairs of molecules at several RMS deviations. A mean minimum-CVA score was considered correct if its classification (based on the 70% threshold) matched the RMS-based classification (based on the 10 Å threshold) for a given pair of proteins. In terms of MolCom3D's mean minimum-CVA scores, pairs were considered to be similar by MolCom3D if the CVA score was 70% or higher, and were considered dissimilar if the score was below 70%. This classification rule will be referred to as the "70-10 rule" hereinafter.

The other two verifiers of similarity, SCOP classification and visual inspection, were used mainly for additional qualitative confirmation of similarity and dissimilarity during the selection of the protein pairs used for calibration. After the proteins were selected, the "70-10 rule" was used for validating the scores.

The ability of MolCom3D to accurately score and classify protein pairs in both the calibration and verification phases of testing was ultimately assessed by considering three quantitative indicators of accuracy. These indicators were all based on comparing MolCom3D's mean minimum-CVA similarity scores with the RMS deviations provided by lsqkab. The indicators, listed from highest to lowest precedence, included:

- the correlation of the mean minimum-CVA similarity score with the RMS deviation for the pairs,

- the error rate in classifying the proteins in each pair as similar or dissimilar as indicated by the "70-10 rule", and
- the separation of minimum-CVA similarity scores awarded to groups of protein pairs that are similar and dissimilar.

These indicators are discussed further in Chapter 7 where the results of the calibration and verification tests are presented.

6.1 Calibration Testing

Many of the parameters assigned during the development of the MolCom3D program were straightforward and somewhat arbitrary, and did not require calibration for the scope of this research. Instead, the calibration testing for the empirical weighting function parameters and the octant decision score threshold served as a "sanity check" for the non-calibrated parameter assignments. That is, a successful calibration of the these parameters indicated that the non-calibrated parameters were assigned acceptable values (A failed attempt to calibrate the software would have entailed revisiting some of the non-calibrated parameter assignments). The non-calibrated parameter assignments have been discussed in Chapter 5 and include: the edge length of the level-0 cube, $|H_0|$; the maximum tree height; the overlap functions, E(L); property observation point locations, property scopes of analysis (in terms of maximum and minimum octant edge lengths); and the decay function, D(r).

The collection of proteins chosen for calibrating the remaining parameters for the MolCom3D software adhered to the following criteria:

1. The proteins were chosen from two different **fold classifications** from the SCOP database (as indicated by the first three digits of the classification number). A fold classification consists of proteins known to have similar structural features. This allowed for the selection of protein pairs that are decidedly different.

2. The proteins within each fold classification were also selected from the same **superfamily** from the SCOP database (except for one protein as indicated below). A superfamily consists of proteins having similar structural and functional similarity.

Tables 6.1 and 6.2 list the proteins chosen for calibrating the protein comparison method. These tables contain two groups of calibration proteins that were carefully selected in order to guarantee that highly similar and highly dissimilar pairs of proteins could be formed to allow for an accurate calibration. Each table represents a collection of similar proteins chosen from a different fold classification; the proteins in these tables will be referred to as Group 1 and Group 2 proteins, respectively. One of the Group 1 proteins also differs in **superfamily classification**. Proteins in different superfamilies are less likely to have a common evolutionary origin. However, after visually inspecting many candidate proteins for calibration testing, it has been found that proteins from different superfamilies are often dissimilar in overall three-dimensional structure. Nevertheless, this protein has major structural similarity with the proteins of the other superfamily listed in the table.

Group 1: Fold "knottins (1.7.3)"				
SCOP Classification Number	PDB Identifier	Superfamily	Family	Protein Name
1.7.3.6.1	1AHO	scorpion toxin-like	long-chain scorpion toxins	toxin II
1.7.3.6.2	1LQH	scorpion toxin-like	short-chain scorpion toxins	insectotoxin
1.7.3.6.1	1NRA	scorpion toxin-like	long-chain scorpion toxins	neurotoxin V, CSE V
1.7.3.6.1	1PTX	scorpion toxin-like	long-chain scorpion toxins	scorpion toxin II
1.7.3.6.1	2SN3	scorpion toxin-like	long-chain scorpion toxins	scorpion neurotoxin (variant 3)

Table 6.1: Proteins Used for Calibration Testing (Group 1)

Group 2: Fold "snake toxin-like (1.7.5)"				
SCOP Classification Number	PDB Identifier	Superfamily	Family	Protein Name
1.7.5.1.1	1COD	snake toxin-like	snake venom toxins	cobrotoxin
1.7.5.1.1	1CRE	snake toxin-like	snake venom toxins	cardiotoxin II
1.7.5.1.1	1ERA	snake toxin-like	snake venom toxins	erabutoxin B
1.7.5.1.1	1FAS	snake toxin-like	snake venom toxins	fasciculin 1
1.7.5.1.1	1KBS	snake toxin-like	snake venom toxins	cytotoxin 4
1.7.5.1.1	1NXB	snake toxin-like	snake venom toxins	neurotoxin B
1.7.5.1.1	2CDX	snake toxin-like	snake venom toxins	cardiotoxin CTX I
1.7.5.1.1	2CRT	snake toxin-like	snake venom toxins	cardiotoxin III

Table 6.2: Proteins Used for Calibration Testing (Group 2)

Calibration proceeded as follows. Pairs of proteins were formed from the groups and compared using MolCom3D compiled with various octant decision score thresholds and sets of weighting factors. The octant decision score thresholds ranged from 50% to 65%. Figure 5.6 shows the base weighting factor curves and Table 7.2 lists various linear combinations of the base weighting factor curves tested. A limit of 9 cubic subdivisions was imposed for each execution of MolCom3D since the minimum CVA score was found to not change appreciably beyond this level during undocumented, informal tests conducted during the development of the software. The degree of cubic overlap ranged from 100% to 200%. The program was run over this cubic overlap range in increments of 25%, and the minimum-CVA similarity scores were averaged into a mean minimum-CVA similarity score for each pair.

Calibration was considered acceptable when at least 95% of the pairs of proteins formed from Groups 1 and 2 were correctly classified as similar or dissimilar according to the "70-10 rule". Furthermore, the (Pearson) correlation between CVA similarity score and RMS deviation value had to be –0.8 or lower (the reader is reminded that the correlation value is negative). It is possible that several different calibration settings would result in CVA similarity scores that meet these specifications. However, one such realization was considered sufficient for this research.

A total of 78 protein pairs were formed and tested. The pairs were generated from all possible pairings of proteins that could be formed from within each group and between each group. As indicated by the SCOP classification numbering, by visual inspection, and by RMS deviations, 38 pairs were similar and 40 pairs were dissimilar. The 38 similar pairs were formed from the 10 pairs that could be formed from Group 1 and the 28 pairs that could be formed from Group 2. The 40 dissimilar pairs were formed by having one member of the pair originating from each group. The results of calibration testing are presented in §7.1 and the empirical data is presented in Appendix B.1 under "Calibration Test Data".

6.2 Verification Testing

The proteins chosen for verifying the MolCom3D software are listed in Tables 6.3 and 6.4. These proteins adhered to the same criteria used for calibration testing.

Two groups of proteins from different folds were tested and are referred to as Group 3 and Group 4 proteins, respectively. A total of 153 protein pairs were formed; as indicated by the SCOP classification numbering, 76 pairs were formed from within a SCOP classification (55 pairs came from the 11 proteins in Group 3 and 21 pairs came from the 7 proteins of Group 4), and 77 pairs were formed between SCOP classifications (one member of the pair originated from each group). The pairs in Group 3 tended to have low RMS values (under 10 Å), but some pairs had high RMS values despite being from the same classification. All pairs in Group 4 had high RMS values. The results of verification testing are presented in §7.2 and the empirical data is presented in Appendix B.2 under "Verification Test Data".

The accuracy of the mean minimum-CVA similarity scores was verified by considering the proportion of protein pairs correctly classified under the "70-10 rule" by the calibrated version of MolCom3D. The mean minimum-CVA similarity scores were calculated by averaging the minimum-CVA scores over a range of cubic overlaps (100% to 250%). As a further verification of the calibrated MolCom3D program, the mean minimum-CVA similarity scores were correlated with the RMS deviations for the pairs of

proteins compared. A strong negative correlation would verify that high minimum-CVA scores, indicating high similarity, would tend to occur with low RMS deviations, and *vice versa*. The results of this testing are presented in §7.2.

Group 3: Fold "Microbial Ribonucleases (1.4.1)"				
SCOP Classification Number	PDB Identifier	Superfamily	Family	Protein Name
1.4.1.1.2	1FUS	Microbial Ribonucleases	Microbial Ribonucleases	hydrolase (endoribonuclease)
1.4.1.1.1	1GMP	Microbial Ribonucleases	Microbial Ribonucleases	hydrolase (guanyloribonuclease)
1.4.1.1.2	1RCL	Microbial Ribonucleases	Microbial Ribonucleases	hydrolase (endoribonuclease)
1.4.1.1.7	1RDS	Microbial Ribonucleases	Microbial Ribonucleases	hydrolase (endoribonuclease)
1.4.1.1.1	1RGE	Microbial Ribonucleases	Microbial Ribonucleases	hydrolase (endoribonuclease)
1.4.1.1.3	1RGK	Microbial Ribonucleases	Microbial Ribonucleases	hydrolase (endoribonuclease)
1.4.1.1.3	1RGL	Microbial Ribonucleases	Microbial Ribonucleases	hydrolase (endoribonuclease)
1.4.1.1.7	1RMS	Microbial Ribonucleases	Microbial Ribonucleases	hydrolase (endoribonuclease)
1.4.1.1.1	1SAR	Microbial Ribonucleases	Microbial Ribonucleases	hydrolase (endoribonuclease)
1.4.1.1.3	2AAE	Microbial Ribonucleases	Microbial Ribonucleases	hydrolase (endoribonuclease)
1.4.1.1.3	9RNT	Microbial Ribonucleases	Microbial Ribonucleases	hydrolase (endoribonuclease)

Table 6.3: Proteins Used for Verification Testing (Group 3)

Group 4: Fold "Lysozyme-like (1.4.1)"				
SCOP Classification Number	PDB Identifier	Superfamily	Family	Protein Name
1.4.2.1.2.1	193L	Lysozyme-like	C-type Lysozyme	hydrolase (O-glycosyl)
1.4.2.1.2.1	1HEW	Lysozyme-like	C-type Lysozyme	hydrolase (O-glycosyl)
1.4.2.1.2.1	1HWA	Lysozyme-like	C-type Lysozyme	hydrolase (O-glycosyl)
1.4.2.1.2.1	1LMA	Lysozyme-like	C-type Lysozyme	hydrolase (O-glycosyl)
1.4.2.1.2.1	1LZB	Lysozyme-like	C-type Lysozyme	hydrolase (O-glycosyl)
1.4.2.1.2.1	1RFP	Lysozyme-like	C-type Lysozyme	hydrolase
1.4.2.1.2.1	6LYT	Lysozyme-like	C-type Lysozyme	hydrolase (O-glycosyl)

Table 6.4: Proteins Used for Verification Testing (Group 4)

Chapter 7

Observations and Results

The results of this research indicate that the structure comparison framework, the protein comparison method, and the computer program, MolCom3D, indeed accurately indicate structural similarity.

The details of the calibration and verification phases of this research have been discussed in Chapter 6. This chapter presents the results of calibration and verification testing. Calibration and verification were conducted with different objectives, and consequently, their results are treated separately.

7.1 Calibration Test Results

The primary objective of the calibration testing was to determine, if possible, at least one set of parameter assignments for the protein comparison method that results in the output of reasonable similarity scores. Similarity scores were considered reasonable if they demonstrated a tendency to increase linearly from 0% to 100% as the degree of similarity rises from extreme dissimilarity to identity. This tendency was indicated by correlating the mean minimum-CVA scores with the RMS deviations. If multiple parameter assignments produced a successful calibration, a secondary objective was to select the set of assignments[2] that maximized the difference between the mean scores awarded to pairs of similar and dissimilar molecules and minimized the number of individual errors committed in awarding these scores. An error is committed whenever a score of under 70% is awarded to a pair of proteins with RMS deviations below 10 Å, or

[2] Finding the optimum set of parameter assignments would require an enormous (and unwarranted) amount of calibration testing on thousands of protein collections.

whenever a score of 70% or more is awarded to proteins with RMS deviations of 10 Å or more.

During the calibration phase, both objectives were achieved. As a result, a version of MolCom3D exists that has been calibrated to give accurate similarity scores in accordance with the properties described in §5.3.1 for pre-aligned, one-subunit proteins.

The first parameter to be calibrated was the critical similarity threshold for Rule 9(c). This threshold, called the **octant decision score** in the MolCom3D program, has been discussed in §5.2.2. The octant decision score was tested with values ranging from 50% to 65% in increments of 5% as shown in Table 7.1. The best threshold score was found to be **55%**. The value for this parameter was found to be easily determined.

Sets of Property Contribution Functions						
Octant Decision Score (%)	Description of Curve Set	Mean CVA Score of Similar Proteins (%)	Mean CVA Score of Dissimilar Proteins (%)	Separation of Mean CVA Scores Between Similar and Dis-similar Proteins (%)	Number of Errors in Scoring 78 Pairs of Calibration Proteins	Correlation of Mean CVA Scores with RMS Deviations
50	1×Group 1(a) 1×Group 1(b) 1×Group 2(a) 1×Group 2(b)	88.50	17.21	71.26	0	–0.81
55	1×Group 1(a) 1×Group 1(b) 1×Group 2(a) 1×Group 2(b)	83.11	5.67	77.44	0	–0.87
60	1×Group 1(a) 1×Group 1(b) 1×Group 2(a) 1×Group 2(b)	78.45	0.00	78.45	8	–0.90
65	1×Group 1(a) 1×Group 1(b) 1×Group 2(a) 1×Group 2(b)	69.09	0.00	69.09	20	–0.92

Table 7.1: Octant Decision Scores Tested During Calibration

Scores of 60% or more were found to be decidedly too restrictive, because they caused termination of cubic exploration prematurely. This resulted in a tendency to misclassify similar proteins as being dissimilar and this tendency became more prominent as the threshold increased above 55%. It is important to acknowledge that cubes deemed similar and explorable in subsequent iterations of the algorithm get subsequent opportunities to be judged dissimilar, if dissimilarity is missed the first time. Conversely, a threshold of 50% caused a tendency for dissimilar proteins to be considered similar.

Five sets of weighting function sigmoid curves were tested and were based on the curves for the four property groups shown in Figure 5.6. The results are given in Table 7.2. Set 1 was found to be the best set of property weighting function curves. This set resulted in the greatest separation of CVA scores awarded to similar and dissimilar proteins, resulted in no erroneous classifications, and had the most negative correlation. The strong negative correlation of almost –0.9, of course, indicates that to a large extent, minimum-CVA similarity score increases linearly with decreasing RMS deviation.

The Set 1 weighting functions are the most chemically intuitive. It seems reasonable to assign more weight to comparisons of helices and sheets than to loops and turns [40]; it also seems reasonable to emphasize charged portions of molecules more strongly than alipathic portions because ionic forces are considerably more influential than van der Waals forces [26]. The property weighting functions of Set 1 resulted in the best mean CVA score separation, the best correlation, and no errors. Thus Set 1 was chosen for use in the verification testing along with the octant decision score of 55%.

Set 5 produced the worst performance. This set was meant to demonstrate that reversing the order of property group examination, that is, examining primary properties at low levels of analysis and secondary properties at high levels of analysis, would give poor performance. The other sets of curves resulted in performances that were intermediate between Set 1 and Set 5.

In general, the measures used to indicate accurate calibration (correlation, classification error rate, CVA score separation) were in agreement. A good separation of

scores amongst similar and dissimilar proteins implied few errors, and implied a strong negative correlation. MolCom3D was indeed accurately calibrated.

Sets of Property Contribution Functions						
Curve Set	Description of Curve Set	Mean CVA Score of Similar Proteins (%)	Mean CVA Score of Dissimilar Proteins (%)	Separation of Mean CVA Scores Between Similar and Dis-similar Proteins (%)	Number of Errors in Scoring 78 Pairs of Calibration Proteins	Correlation of Mean CVA Scores with RMS Deviations
1	1×Group 1(a) 1×Group 1(b) 1×Group 2(a) 1×Group 2(b)	83.11	5.67	77.44	0	–0.87
2	1×Group 1(a) 2×Group 1(b) 1×Group 2(a) 1×Group 2(b)	81.77	20.51	61.26	3	–0.80
3	1×Group 1(a) 1×Group 1(b) 1×Group 2(a) 2×Group 2(b)	83.40	8.70	74.70	0	–0.86
4	1×Group 1(a) 2×Group 1(b) 1×Group 2(a) 2×Group 2(b)	81.77	20.12	61.65	3	–0.81
5	1×Group 1(a) 2×Group 1(b) 1×Group 2(a) 2×Group 2(b) Group 1 and Group 2 curves interchanged	83.04	45.41	37.64	5	–0.78

Table 7.2: Weighting Function Curve Sets Tested During Calibration

7.2 Verification Test Results

The primary objective of the verification tests was to demonstrate that the calibrated MolCom3D accurately differentiates similar and dissimilar proteins on a collection of proteins not used in the calibration.

One indicator of accuracy of the calibrated software was the number of errors made in judging similarity in the 153 pairs of molecules over the 7 tested degrees of cubic overlap (100% to 250% cubic overlap in increments of 25%). The number of errors in the $7 \times 153 = 1071$ protein comparisons was found to be 15 (98.6% accuracy). All 15 errors occurred at 100% overlap and the errors disappeared once some actual cubic overlap was introduced.

A second indicator of accuracy was the correlation of the mean minimum-CVA similarity scores to the RMS deviations of structure provided by the lsqkab software. The verification tests revealed strong negative correlations. In considering all protein pairs with nonzero CVA scores, irrespective of similarity and dissimilarity, the correlation was –0.89; in considering only the similar proteins, the correlation was –0.96. Several protein pairs were awarded CVA scores of zero because there was insufficient similarity between the molecules for MolCom3D to reach the maximum level of analysis. It was more reasonable to consider correlation without these zero scores included in the calculation because CVA scores are linear from roughly 35% to 100%. Below 35%, the algorithm usually terminates with a CVA score of 0% (indicating that no octants remain to be processed due to extreme dissimilarity). The Future Work section discusses making the 0% to 34% range more linear. Given that a good overall correlation of –0.89 existed, a linear regression line was determined to be,

$$\text{CVA score (RMS deviation)} = -3.166\ \%/\text{Å} \times \text{RMS deviation} + 99.64\%$$

This line predicts a CVA score for a protein pair given the RMS deviation for the structures. A scatter plot corresponding to the verification test data showing the above linear regression line is shown in Figure 7.1.

A final indicator of accuracy was to consider the tendencies of the mean minimum-CVA scores (that is, the mean, mean minimum-CVA scores) for the similar and dissimilar protein pairs. As was done for correlation, the calculation of the mean CVA scores excluded scores of zero. The mean CVA score awarded to the similar proteins was 93%, the mean CVA score awarded to the dissimilar proteins (not including zeros) was 48%, and the separation of these means was 45%. For the calibration testing, these three values including the zero scores were, 93%, 31%, and 62%, respectively.

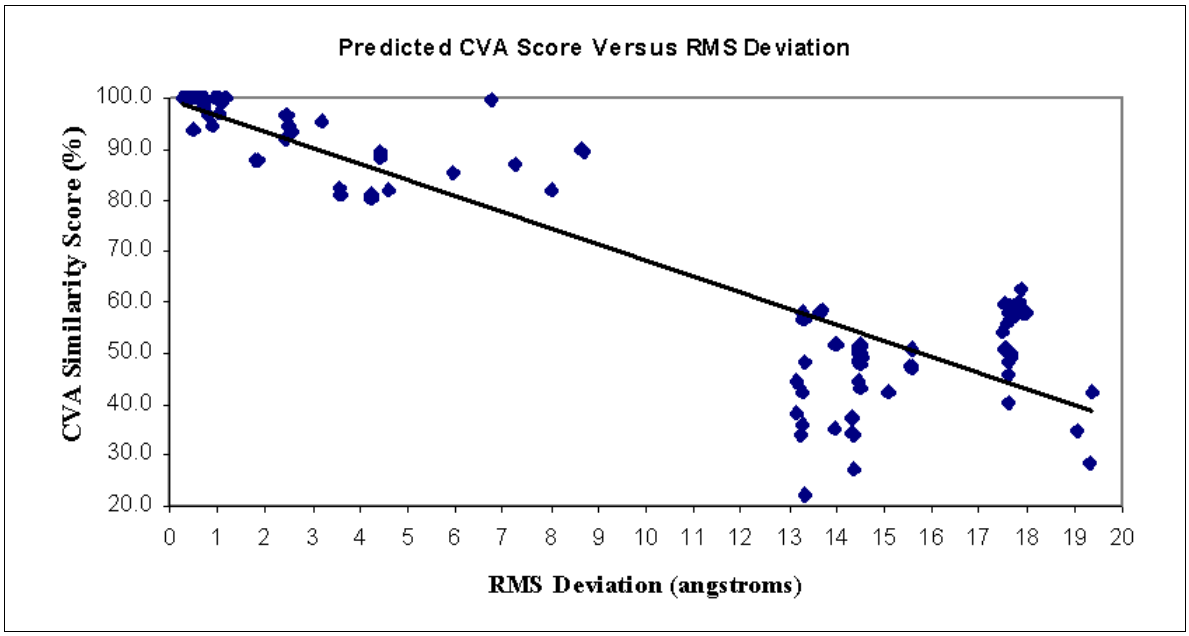


Figure 7.1: Regression Line for Predicted CVA Score Versus RMS Deviation

The empirical data for verification testing is listed in Table B.2.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

This research has introduced a new framework and a derived method for similarity detection that is efficient, flexible, and extensible. Assuming that the properties examined are appropriately chosen, an appropriate amount of detail is considered over each level of comparative analysis.

8.1.1 A Comparison of MolCom3D with Commonly Used RMS Algorithms

RMS deviation algorithms align proteins so that the overall RMS deviation of the amino acid residue positions is minimized. The resulting RMS value gives a good overall indication of structural similarity. Furthermore, programs such as lsqkab produce this value in well under a minute on an SGI workstation.

The minimum-CVA similarity score of MolCom3D also gives a good overall indication of structural similarity, as evidenced by its low error rate and correlation with RMS deviation values. Run times on individual protein pairs are in the order of 1 to 4 minutes (depending on the degree of cubic overlap for a particular run). Then what are the major advantages of MolCom3D?

The first advantage is that MolCom3D provides an octtree. In addition to the overall indication of similarity given by the CVA score, MolCom3D's octtree indicates *where* similarity exists in the collection of structures. The octtree constitutes a structural alignment of similar portions of the structures compared. A second advantage is the capacity of MolCom3D to dynamically alter the influence of the properties compared as the level of analysis increases. Properties are thus compared at a level of analysis that is *appropriate* for the expected relative influence at any given degree of spatial detail.

Although modest execution times were not an objective of this research and were not specifically tested, execution times allowed all of the calibration and verification tests to be completed in under 12 hours on a 100 MHz pentium processor running Windows NT 4.0 (involving approximately 1500 comparisons spanning a variety of cubic overlaps from 100% to 250%).

8.1.2 Contributions to Computer Science

It has been demonstrated that at least one algorithm and computer program can be derived from the structure comparison framework. Presumably, other objects besides protein molecules can be compared using algorithms derived from the framework.

Several important ideas have been presented, tested, and shown to be successful. The major ideas studied in this research were:

- the overlapping spatial localities concept,
- partitioning of space to limit combinatorial searches using an octtree,
- the creation of an algorithm that is directed by presence of properties and occupancy of space rather than by the total possible number of cubes in the cubic lattice.

8.1.3 Contributions to Chemistry

A “proof-of-concept” computer program, MolCom3D, has been constructed that effectively scores collections of proteins based on structural similarity. Moreover, a sound basis has been devised for developing programs for other types of molecules. Furthermore, this research has invented a new multiple structural alignment paradigm based on the octtree that indicates regions of similarity.

8.2 Future Work

Several avenues of future research result from this “proof-of-concept” research. These are discussed in order from the straightforward to the more difficult challenges.

The program requires improvement for practical use. Several issues of efficiency were not addressed in favour of expediency in meeting the goals of this research. For example, hash tables and trees could be used in place of lists. The choice of better data structures could result in an order of magnitude improvement in the execution time. More efficacious properties could be added to greatly enhance accuracy. For example, volume-limited statistical properties might be considered (based on the cubic lattice) [27]. Volume-limited RMS deviations of various structures could be considered [14]. The CVA scoring scheme would also be improved if the linear range were expanded to encompass scores in the 0% to 35% range (such that CVA scores under 35% did not have a tendency to fall off sharply to 0%).

The program could be altered to compare gap-less DNA nucleotide sequence tendencies. This might be accomplished by combining sets of nucleotides into overlapping "meta-sequences" using a one-dimensional derivation of the comparison framework (that would be called a sequence comparison framework).

The substructure collocation idea discussed in Rule 10 requires development. This rule defines alterations of structural data and causes shifting of structures towards regions of structural consensus before delving to deeper levels of analysis. This could be accomplished as the tree is built. Variably located, but related sub-structures (like α-helices) could be located by the tree and mapped to the aggregation points as usual. The tree could be used to subsequently "collocate" these structures through the alteration of the structural data files. Then analysis can delve into more highly refined properties within the newly formed structures. Unfortunately, the scope of this research precluded collocation, and it must wait until the future to be developed.

A pre-alignment method is required so that the algorithm is not restricted to minimum RMS pre-alignments. It might be possible to invent a set of non-alignment "rotatory properties" that give reasonable comparisons. The objects under comparison would be revolved around axes in two dimensions forming three-dimensional solids whose property densities around their centroids would be directly comparable and automatically aligned.

Exploration of complementarity in surfaces might prove to be successful using a modified version of the framework. This would be useful in studying, for example, rational drug design and enzyme-substrate binding.

References

1. Alexandrov, N. N., and Fischer, D. Analysis of Topological and Nontopological Structural Similarities in the PDB: New Examples With Old Structures. *Proteins* **25**:354-365 (1996).

2. Alexandrov, N. N., Takahashi, K., and Go, N. Common Spatial Arrangements of Backbone Fragments in Homologous and Non-homologous Proteins. *Journal of Molecular Biology* **225**:5-9 (1992).

3. Bailey, T. L., and Gribskov, M. Combining Evidence Using p-values: Application to Sequence Homology Searches. *Bioinformatics* **14**(1):48-54 (1998).

4. Blundell, T. L., and Johnson, M. S. Catching a Common Fold. *Protein Science* **2**(6):877-883 (1993).

5. Brown, W. H. (1982). Introduction to Organic Chemistry. Boston, Massachusetts: Willard Grant Press.

6. Chothia, C. and Lesk, A. M. Evolution of Proteins Formed by β-Sheets: II. The Core of the Immunoglobulin Domains. *Journal of Molecular Biology* **160**:325-342 (1982).

7. Chothia, C. and Lesk, A. M. Evolution of Proteins Formed by β-Sheets: I. Plastocyanin and Azurin. *Journal of Molecular Biology* **160**:309-323 (1982).

8. Clark, D. A., Rawlings, C. J., Shirazi, J., Verson, A., and Reeve, M. Protein Topology Prediction through Parallel Constraint Logic Programming. *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology* 1993.

9. Clark, D. A., Shirazi, J., and Rawlings, C. J. Protein Topology Prediction Through Constraint-based Search and the Evaluation of Topological Folding Rules. *Protein Engineering* **4**(7):751-760 (1991).

10. Falicov, A., Cohen, F. E. A Surface of Minimum Area Metric for the Structural Comparison of Proteins. *Journal of Molecular Biology* **258**:871-892 (1996).

11. Foley, J. D., van Dam, A., Feiner, S. K., Hughes, J. F. (1990). *Computer Graphics: Principles and Practice*. Reading, Massachusetts: Addison-Wesley Publishing Company.

12. Fox, S. I. (1992). *Human Physiology*. Oxford, England: Wm. C. Brown Publishers, Inc.

13. Fullen G. Research Group in Practical Computer Science (1996). http://www.techfak.uni-bielefeld.de/bcd/Curric/MulAli/node2.html

14. Gerstein, M. and Levitt, M. Comprehensive Assessment of Automatic Structural Alignment Against a Manual Standard, the SCOP Classification of Proteins. *Protein Science* **7**:445-456 (1998).

15. Gilbert, D. and Westhead, D. Formalising TOPS Cartoons, the Development of an Interactive Similarity Search Mechanism Over TOPS Databases and a TOPS-based Structure Comparison Method. http://www.ebi.ac.uk/~drg/epsrc-vf/epsrc-vf.html.

16. Gilbert, D., Westhead, D., and Thornton, J. A Computer System to Perform Topology-based Protein Structure Comparison (Personal communications - to appear in Bioinformatics).

17. Glassy, L., Starkey, J., and Jacobs G. SIERRA: A Spatial Data System for Neuronal Data (1997). http://nervana.montana.edu/publications/neurosys/SIERRA.html

18. Gusfield (1997). *Algorithms on Strings, Trees, and Sequences: Computer Sceince and Computational Biology*. Cambridge: Press Syndicate of the University of Cambridge.

19. Henikoff, S., Pietrokovski, S., Henikoff, J. G. Superior Performance in Protein Homology Detection with the BLOCKS Database Servers. *Nucleic Acids Research* **26**:309-312 (1998).

20. Holm L., Sander C. New Structure—Novel Fold? *Structure* **5**:165-171 (1997).

21. Holm, L. and Sander, C. Searching Protein Structure Databases Has Come of Age. *Proteins* **19**:165-173 (1994).

22. Hornung, R. D., and Kohn, S. R. The Use of Object-oriented Design Patterns in the SAMRAI Structured AMR Framework. http://www.llnl.gov/CASC/SAMRAI/pubs/samrai_oo98.pdf.

23. Hyams, D. (1997). Curve Expert Version 1.34. http://www.ebicom.net/~dhyams/cmain.htm.

24. Karlin, S., Bucher, P., Brendel, V., and Altschul, S. F. Statistical Methods and Insights for Protein and DNA Sequences. *Annual Review of Biophysics and Biophysical Chemistry* **20**:175-203 (1991).

25. Kyte, J. and Doolittle, R. F. A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology* **157**:105-132 (1982).

26. Lehninger, A. L., Nelson, D. L., and Cox, M. M. (1993). *Principles of Biochemistry*. New York, New York: Worth Publishers, Inc.

27. Levitt, M., and Gerstein, M. A Unified Statistical Framework for Sequence Comparison and Structure Comparison. *Proceedings of the National Academy of Sciences USA* **95**:5913-5920 (1998).

28. Major, F., Malenfant J., and Stewart, N. Distance Between Objects Represented by Octtrees Defined in Different Coordinate Systems. *Computers and Graphics* **13**(4):497-503 (1989).

29. Matthews, B. W., and Rossmann, M. G. Comparison of Protein Structures. Methods in Enzymology **115**:397-420 (1985).

30. Mezey, P. G. (1993). *Shape in Chemistry: An Introduction to Molecular Shape and Topology*. VCH Publishers, Inc.

31. Mitchell, E. M., Artymiuk, P. J., Rice, D. W., and Willett, P. Use of Techniques Derived from Graph Theory to Compare Secondary Structure Motifs in Proteins. *Journal of Molecular Biology* **212**:151-166 (1989).

32. Murzin, A. G., Brenner, S. E., Hubbard, T. J. P., and Chothia C. (1998). Introduction to Structural Classification of Proteins. http://scop.mrc-lmb.cam.ac.uk/scop/intro.html

33. Needleman, S. B., Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology* **48**:443-453 (1970).

34. Rawlings, C. J., Taylor, W. R., Nyakairu, J., Fox J., and Sternberg, M. J. E. Reasoning About Protein Topology Using the Logic Programming Language Prolog. *Journal of Molecular Graphics* **3**(4):151-157 (1985).

35. Reeck, G. R., de Haen, C., Teller, D. C., Doolittle, R. F., Fitch, W. M., Dickerson, R. E., Chanbon, P., McLachlan, A. D., Margoliash, E., Jukes, T. H., and Zuckerkandll, E. "Homology" in Proteins and Nucleic Acids: A Terminology Muddle and a Way out of It. *Cell* **50**:667-668 (1987).

36. Sali, A., and Blundell, T. L. Definition of General Topological Equivalence in Protein Structures. *Journal of Molecular Biology* **212**:403-428 (1990).

37. Sander C., Holm, L. Mapping the Protein Universe. *Science* **273**:595-602 (1996).

38. Sander, C., and Holm, L. Protein Structure Comparison by Alignment of Distance Matrices. *Journal of Molecular Biology* **233**:123-138 (1993).

39. Schmidt, J. P., Chen, C. C., Cooper, J., and Altman, R. B. A Surface Measure for Probabilistic Structural Computations. *ISMB-98* 148-156 (1998).

40. Shortle, D. Structure Prediction: Folding Proteins by Pattern Recognition. *Current Biology* **7**(3):151-154 (1998).

41. Smith, T. F., Waterman, M. S. Comparison of Biosequences. *Advances in Applied Mathematics* **2**: 482-289 (1981).

42. Solomons, T. W. G. (1988). *Organic Chemistry*. Toronto, Ontario: John Wiley & Sons, Inc.

43. Taylor, W. R., and Orengo, C. A. Protein Structure Alignment. *Journal of Molecular Biology* **208**:1-22 (1989).

44. Thompson, J. D., Higgins D. G., and Gibson, T. J. CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research* **22**(22):4673-4680 (1994).

45. Vingron, M. and Waterman M. Sequence Alignment and Penalty Choice. *Journal of Molecular Biology* **235**:1-12 (1994).

46. Welhelms, J., and van Gelder, A. Octrees for Faster Isosurface Generation. *ACM Transactions on Graphics* **11**(3):201-227 (1992).

47. CCP4: Collaborative Computational Project No. 4 (1994). The CCP4 Suite: Programs for Protein Crystallography. Acta Crystallographica, Sect. P, **50**, 760-763.

48. Yao, J., Levitt, M., and Lee C. A Structural View of Gaps in Sequence Analysis. Poster 86 in *Sixth International Conference on Intelligent Systems for Molecular Biology* (1998).

Appendix A

Enlarged Images

A.1 Comparison of Proteins 2CRT and 1KBS

Figures A.1 to A.4 are enlarged images corresponding to the proteins shown in Figure 5.3 on page 62.

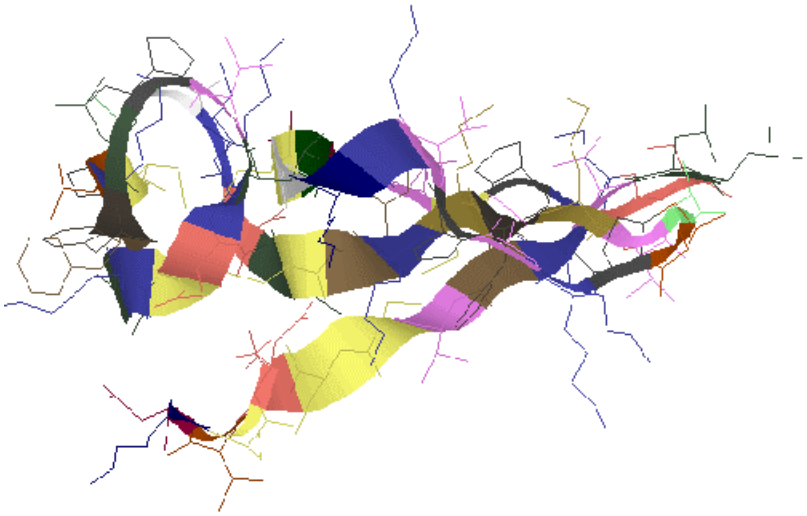


Figure A.1: Original 2CRT Molecule

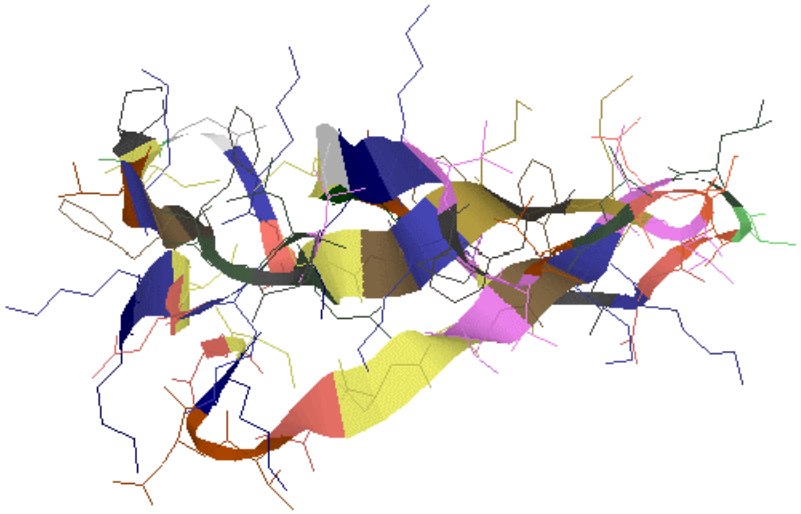


Figure A.2: Original 1KBS Molecule

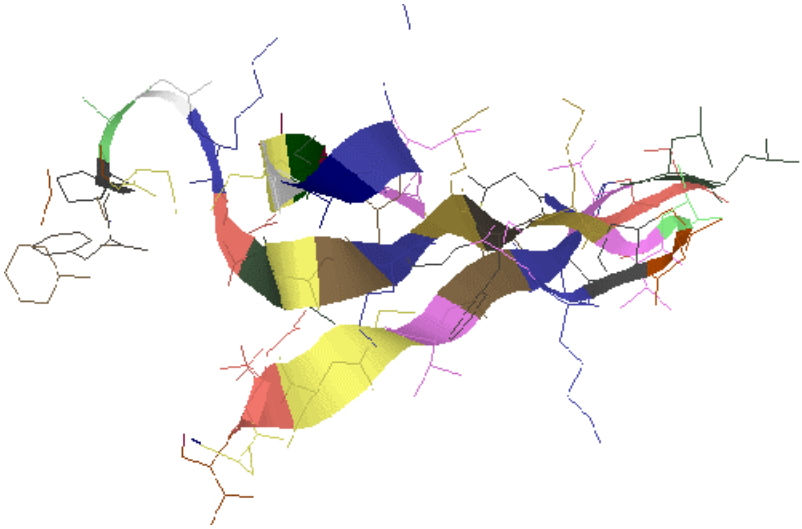


Figure A.3: Portions of 2CRT Bearing Similarity to Corresponding Portions of 1KBS

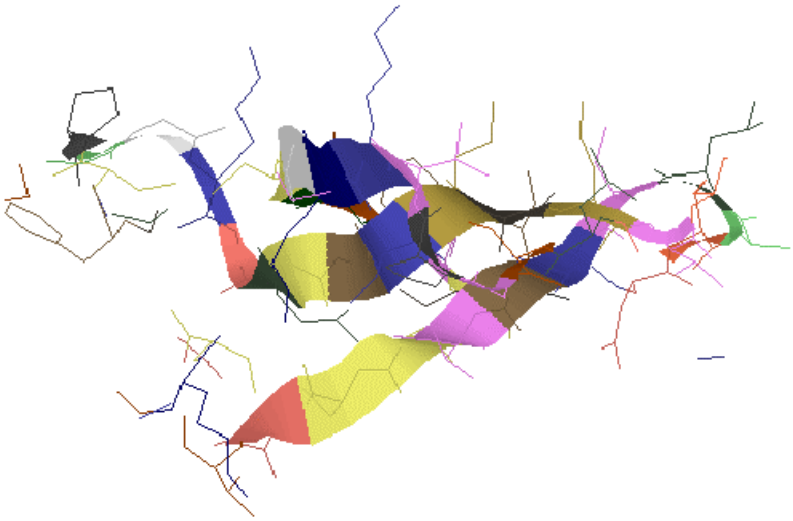


Figure A.4: Portions of 1KBS Bearing Similarity to Corresponding Portions of 2CRT

Appendix B

Empirical Data

B.1 Calibration Test Data

Calibration Data for Set 1 Weighting Function Curves (55% octant decision score)

Protein Pair PDB Identifiers		RMS Score (Å)	Mean Minimum-CVA Similarity Score From MolCom3D (%)	Standard Deviation in Minimum-CVA Similarity Score (%)	Error Where Marked *X*	Observed Minimum-CVA Similarity Score (%) At Degree of Cubic Overlap (%)				
						100	125	150	175	200
1AHO	1LQH	4.7	92.4	2.8		75	94	93	88	95
1AHO	1NRA	4.8	88.3	4.3		95	91	91	82	89
1AHO	1PTX	0.3	100.0	0.0		100	100	100	100	100
1AHO	2SN3	8.4	81.7	4.7		50	89	80	80	79
1LQH	1NRA	7.3	84.1	4.7		82	87	86	77	87
1LQH	1PTX	4.4	90.9	4.1		91	95	94	88	87
1LQH	2SN3	8.7	78.4	5.1		63	85	78	72	79
1NRA	1PTX	4.7	85.8	3.7		96	90	82	84	87
1NRA	2SN3	9.4	78.0	6.4		63	84	75	70	83
1PTX	2SN3	9.0	72.3	5.3		94	71	68	70	80
1COD	1CRE	7.9	79.8	3.4		63	83	80	75	82
1COD	1ERA	2.7	92.0	2.0		97	95	91	90	92
1COD	1FAS	6.5	81.7	2.4		88	83	81	79	84
1COD	1KBS	7.6	84.5	4.7		86	90	85	78	85
1COD	1NXB	6.2	82.5	2.1		86	84	83	79	84
1COD	2CDX	8.1	75.3	6.8		88	82	73	67	79
1COD	2CRT	7.7	85.5	3.1		88	89	83	83	87
1CRE	1ERA	7.5	82.2	5.2		75	88	83	75	83
1CRE	1FAS	4.8	76.8	3.9		75	82	76	73	76
1CRE	1KBS	4.6	97.1	0.7		99	98	97	97	96
1CRE	1NXB	9.6	72.1	6.3		50	78	70	64	76
1CRE	2CDX	4.8	88.7	2.7		88	92	87	86	90
1CRE	2CRT	3.7	96.2	0.4		100	97	97	96	96
1ERA	1FAS	5.8	78.9	2.8		94	81	81	75	79
1ERA	1KBS	7.2	82.5	6.3		98	89	81	75	85
1ERA	1NXB	5.7	92.8	1.6		97	93	94	93	91
1ERA	2CDX	7.6	77.1	4.1		92	82	77	73	77
1ERA	2CRT	7.4	85.0	5.0		97	90	87	78	85
1FAS	1KBS	4.9	74.1	3.5		78	75	72	71	79
1FAS	1NXB	8.4	76.1	6.3		87	73	70	78	84
1FAS	2CDX	4.9	75.4	3.0		91	78	72	74	78
1FAS	2CRT	4.6	74.4	4.1		83	75	73	70	80
1KBS	1NXB	9.3	78.9	7.4		87	88	81	70	77
1KBS	2CDX	5.1	87.2	1.1		85	88	87	86	89
1KBS	2CRT	4.6	96.4	2.1		100	100	96	95	95
1NXB	2CDX	7.5	74.6	5.1		84	78	72	69	80
1NXB	2CRT	7.3	70.1	7.6		86	75	63	64	78
2CDX	2CRT	4.4	89.0	2.4		92	92	89	86	89
1AHO	1COD	12.3	0.0	0.0		0	0	0	0	0
1AHO	1CRE	11.0	0.0	0.0		0	0	0	0	0
1AHO	1ERA	12.0	0.0	0.0		0	0	0	0	0
1AHO	1FAS	11.6	0.0	0.0		0	0	0	0	0
1AHO	1KBS	10.7	0.0	0.0		0	0	0	0	0
1AHO	1NXB	13.1	0.0	0.0		0	0	0	0	0
1AHO	2CDX	11.0	0.0	0.0		0	0	0	0	0

1AHO	2CRT	11.1	0.0	0.0		0	0	0	0	0
1LQH	1COD	12.1	0.0	0.0		0	0	0	0	0
1LQH	1CRE	11.5	0.0	0.0		0	0	0	0	0
1LQH	1ERA	12.4	0.0	0.0		0	0	0	0	0
1LQH	1FAS	11.7	0.0	0.0		0	0	0	0	0
1LQH	1KBS	11.1	0.0	0.0		0	0	0	0	0
1LQH	1NXB	13.2	0.0	0.0		0	0	0	0	0
1LQH	2CDX	11.4	0.0	0.0		0	0	0	0	0
1LQH	2CRT	11.6	0.0	0.0		0	0	0	0	0
1NRA	1COD	12.0	16.9	8.2		50	6	25	18	19
1NRA	1CRE	11.0	0.0	0.0		0	0	0	0	0
1NRA	1ERA	11.7	0.0	0.0		0	0	0	0	0
1NRA	1FAS	10.7	0.0	0.0		0	0	0	0	0
1NRA	1KBS	11.0	21.7	6.7		50	25	25	12	25
1NRA	1NXB	12.3	0.0	0.0		0	0	0	0	0
1NRA	2CDX	11.0	0.0	0.0		0	0	0	0	0
1NRA	2CRT	11.0	0.0	0.0		0	0	0	0	0
1PTX	1COD	12.2	40.2	5.4		63	38	38	37	48
1PTX	1CRE	10.8	0.0	0.0		0	0	0	0	0
1PTX	1ERA	12.0	43.5	11.1		56	42	36	36	60
1PTX	1FAS	11.6	0.0	0.0		0	0	0	0	0
1PTX	1KBS	10.6	51.4	5.0		63	55	47	48	57
1PTX	1NXB	13.0	0.0	0.0		0	0	0	0	0
1PTX	2CDX	10.8	0.0	0.0		0	0	0	0	0
1PTX	2CRT	11.0	0.0	0.0		0	0	0	0	0
2SN3	1COD	9.9	0.0	0.0		0	0	0	0	0
2SN3	1CRE	10.9	19.0	12.3		68	34	23	14	6
2SN3	1ERA	9.9	0.0	0.0		0	0	0	0	0
2SN3	1FAS	10.3	34.3	8.7		59	35	43	37	22
2SN3	1KBS	10.6	0.0	0.0		0	0	0	0	0
2SN3	1NXB	11.0	0.0	0.0		0	0	0	0	0
2SN3	2CDX	10.5	0.0	0.0		0	0	0	0	0
2SN3	2CRT	10.9	0.0	0.0		0	0	0	0	0

Calibration Data for Set 2 Weighting Function Curves (55% octant decision score)

Protein Pair PDB Identifiers		RMS Score (Å)	Mean Minimum-CVA Similarity Score From MolCom3D (%)	Standard Deviation in Minimum-CVA Similarity Score (%)	Error Where Marked *X*	Observed Minimum-CVA Similarity Score (%) At Degree of Cubic Overlap (%)				
						100	125	150	175	200
1AHO	1LQH	4.7	95.4	1.7		98	96	96	93	97
1AHO	1NRA	4.8	91.2	2.2		96	93	92	88	92
1AHO	1PTX	0.3	100.0	0.0		100	100	100	100	100
1AHO	2SN3	8.4	72.0	2.4		50	75	70	71	73
1LQH	1NRA	7.3	87.7	3.5		92	90	88	83	90
1LQH	1PTX	4.4	92.2	3.2		94	95	95	92	88
1LQH	2SN3	8.7	59.1	4.0	*X*	63	63	55	57	63
1NRA	1PTX	4.7	88.8	3.6		97	93	88	85	90
1NRA	2SN3	9.4	77.8	7.9		63	87	78	68	79
1PTX	2SN3	9.0	66.7	7.4	*X*	91	68	62	60	77
1COD	1CRE	7.9	81.1	2.1		63	81	84	79	81
1COD	1ERA	2.7	94.5	2.3		97	97	95	92	94
1COD	1FAS	6.5	79.4	3.7		75	84	81	75	79
1COD	1KBS	7.6	87.7	3.1		88	90	88	83	89
1COD	1NXB	6.2	77.9	3.9		50	80	82	77	73
1COD	2CDX	8.1	75.8	4.1		75	78	74	71	81
1COD	2CRT	7.7	88.7	2.4		88	90	88	86	91
1CRE	1ERA	7.5	79.4	4.1		63	81	81	73	82
1CRE	1FAS	4.8	75.3	3.5		63	80	74	72	76
1CRE	1KBS	4.6	96.4	1.3		88	98	96	97	95
1CRE	1NXB	9.6	71.1	9.9		38	84	66	62	72
1CRE	2CDX	4.8	86.8	2.5		75	90	85	85	87
1CRE	2CRT	3.7	94.7	1.0		100	96	94	95	93
1ERA	1FAS	5.8	75.5	2.6		74	76	77	72	77
1ERA	1KBS	7.2	86.6	4.6		99	92	87	81	88

1ERA	1NXB	5.7	87.7	1.9		95	85	87	89	89
1ERA	2CDX	7.6	75.4	3.8		84	81	76	73	72
1ERA	2CRT	7.4	88.2	3.8		97	92	89	83	89
1FAS	1KBS	4.9	72.2	4.4		77	73	67	71	78
1FAS	1NXB	8.4	73.7	7.0		68	69	68	74	83
1FAS	2CDX	4.9	77.4	3.5		92	79	75	74	82
1FAS	2CRT	4.6	73.7	4.4		87	73	74	69	79
1KBS	1NXB	9.3	73.2	2.9		75	77	71	71	73
1KBS	2CDX	5.1	84.5	1.1		63	83	85	85	85
1KBS	2CRT	4.6	97.6	1.5		100	99	99	96	96
1NXB	2CDX	7.5	72.5	5.3		82	77	72	65	77
1NXB	2CRT	7.3	62.4	8.8	*X*	79	67	54	56	72
2CDX	2CRT	4.4	87.4	3.1		75	91	85	84	89
1AHO	1COD	12.3	38.8	13.7		75	32	50	23	50
1AHO	1CRE	11.0	0.0	0.0		0	0	0	0	0
1AHO	1ERA	12.0	25.0	5.5		63	18	24	26	32
1AHO	1FAS	11.6	0.0	0.0		0	0	0	0	0
1AHO	1KBS	10.7	31.4	9.9		63	35	17	36	38
1AHO	1NXB	13.1	0.0	0.0		0	0	0	0	0
1AHO	2CDX	11.0	0.0	0.0		0	0	0	0	0
1AHO	2CRT	11.1	11.2	6.0		50	8	8	8	20
1LQH	1COD	12.1	35.6	12.3		63	23	44	27	48
1LQH	1CRE	11.5	0.0	0.0		0	0	0	0	0
1LQH	1ERA	12.4	34.0	15.6		50	21	45	20	50
1LQH	1FAS	11.7	0.0	0.0		0	0	0	0	0
1LQH	1KBS	11.1	45.1	4.5		63	46	45	39	50
1LQH	1NXB	13.2	0.0	0.0		0	0	0	0	0
1LQH	2CDX	11.4	0.0	0.0		0	0	0	0	0
1LQH	2CRT	11.6	31.5	10.2		63	38	35	16	38
1NRA	1COD	12.0	53.7	2.8		50	54	53	51	57
1NRA	1CRE	11.0	0.0	0.0		0	0	0	0	0
1NRA	1ERA	11.7	51.9	3.5		75	56	53	48	52
1NRA	1FAS	10.7	0.0	0.0		0	0	0	0	0
1NRA	1KBS	11.0	58.2	12.5		50	75	53	45	60
1NRA	1NXB	12.3	0.0	0.0		0	0	0	0	0
1NRA	2CDX	11.0	0.0	0.0		0	0	0	0	0
1NRA	2CRT	11.0	62.1	4.6		63	61	66	56	65
1PTX	1COD	12.2	51.4	15.4		56	44	38	50	73
1PTX	1CRE	10.8	0.0	0.0		0	0	0	0	0
1PTX	1ERA	12.0	54.5	11.8		50	48	47	51	72
1PTX	1FAS	11.6	0.0	0.0		0	0	0	0	0
1PTX	1KBS	10.6	60.7	5.9		69	63	54	58	67
1PTX	1NXB	13.0	0.0	0.0		0	0	0	0	0
1PTX	2CDX	10.8	0.0	0.0		0	0	0	0	0
1PTX	2CRT	11.0	53.9	12.9		63	58	42	46	70
2SN3	1COD	9.9	0.0	0.0		0	0	0	0	0
2SN3	1CRE	10.9	32.3	8.9		60	38	27	23	41
2SN3	1ERA	9.9	0.0	0.0		0	0	0	0	0
2SN3	1FAS	10.3	55.4	15.5		68	35	60	54	73
2SN3	1KBS	10.6	0.0	0.0		0	0	0	0	0
2SN3	1NXB	11.0	0.0	0.0		0	0	0	0	0
2SN3	2CDX	10.5	34.2	6.1		71	31	43	30	32
2SN3	2CRT	10.9	0.0	0.0		0	0	0	0	0

Calibration Data for Set 3 Weighting Function Curves (55% octant decision score)

Protein Pair PDB Identifiers		RMS Score (Å)	Mean Minimum-CVA Similarity Score From MolCom3D (%)	Standard Deviation in Minimum-CVA Similarity Score (%)	Error Where Marked *X*	Observed Minimum-CVA Similarity Score (%) At Degree of Cubic Overlap (%)				
						100	125	150	175	200
1AHO	1LQH	4.7	92.4	2.3		87	93	93	89	94
1AHO	1NRA	4.8	88.3	3.7		95	92	91	83	88
1AHO	1PTX	0.3	100.0	0.0		100	100	100	100	100
1AHO	2SN3	8.4	81.6	6.0		50	91	80	78	78
1LQH	1NRA	7.3	83.4	4.3		82	87	84	77	86
1LQH	1PTX	4.4	89.8	3.9		91	95	90	87	87
1LQH	2SN3	8.7	78.4	6.1		75	87	75	73	80
1NRA	1PTX	4.7	85.6	2.9		97	89	83	83	87
1NRA	2SN3	9.4	78.7	7.5		63	86	74	71	84
1PTX	2SN3	9.0	73.4	4.1		93	70	72	72	79
1COD	1CRE	7.9	81.1	2.1		63	81	84	79	81
1COD	1ERA	2.7	92.0	3.2		99	96	91	89	92
1COD	1FAS	6.5	80.5	1.6		88	79	81	79	83
1COD	1KBS	7.6	84.2	3.6		86	88	85	80	84
1COD	1NXB	6.2	84.3	2.0		87	87	84	82	84
1COD	2CDX	8.1	78.9	4.7		87	84	78	73	81
1COD	2CRT	7.7	83.9	3.0		88	87	83	80	86
1CRE	1ERA	7.5	82.1	4.7		75	85	85	75	83
1CRE	1FAS	4.8	76.3	4.0		75	81	74	72	78
1CRE	1KBS	4.6	96.9	1.8		99	100	96	96	96
1CRE	1NXB	9.6	73.3	6.7		50	81	73	65	75
1CRE	2CDX	4.8	88.4	2.6		88	92	88	86	89
1CRE	2CRT	3.7	97.2	1.2		100	96	97	98	98
1ERA	1FAS	5.8	79.4	2.8		94	81	82	75	80
1ERA	1KBS	7.2	82.6	5.0		97	88	82	76	84
1ERA	1NXB	5.7	93.8	1.6		98	94	95	95	92
1ERA	2CDX	7.6	76.6	4.5		93	81	80	71	75
1ERA	2CRT	7.4	84.8	4.7		95	90	86	79	84
1FAS	1KBS	4.9	74.8	2.3		80	75	75	72	77
1FAS	1NXB	8.4	78.4	3.9		87	76	76	77	84
1FAS	2CDX	4.9	75.2	3.5		91	78	73	71	78
1FAS	2CRT	4.6	73.6	3.8		81	71	75	70	78
1KBS	1NXB	9.3	79.6	7.6		87	89	81	71	78
1KBS	2CDX	5.1	89.4	0.4		87	89	89	89	90
1KBS	2CRT	4.6	96.5	2.1		100	99	98	95	95
1NXB	2CDX	7.5	73.5	5.2		83	74	72	68	81
1NXB	2CRT	7.3	70.9	7.9		89	73	65	64	81
2CDX	2CRT	4.4	89.9	2.7		93	93	89	87	91
1AHO	1COD	12.3	0.0	0.0		0	0	0	0	0
1AHO	1CRE	11.0	0.0	0.0		0	0	0	0	0
1AHO	1ERA	12.0	0.0	0.0		0	0	0	0	0
1AHO	1FAS	11.6	0.0	0.0		0	0	0	0	0
1AHO	1KBS	10.7	0.0	0.0		0	0	0	0	0
1AHO	1NXB	13.1	0.0	0.0		0	0	0	0	0
1AHO	2CDX	11.0	0.0	0.0		0	0	0	0	0
1AHO	2CRT	11.1	0.0	0.0		0	0	0	0	0
1LQH	1COD	12.1	0.0	0.0		0	0	0	0	0
1LQH	1CRE	11.5	0.0	0.0		0	0	0	0	0
1LQH	1ERA	12.4	0.0	0.0		0	0	0	0	0
1LQH	1FAS	11.7	0.0	0.0		0	0	0	0	0
1LQH	1KBS	11.1	0.0	0.0		0	0	0	0	0
1LQH	1NXB	13.2	0.0	0.0		0	0	0	0	0
1LQH	2CDX	11.4	0.0	0.0		0	0	0	0	0
1LQH	2CRT	11.6	0.0	0.0		0	0	0	0	0
1NRA	1COD	12.0	21.1	4.5		50	25	25	18	17
1NRA	1CRE	11.0	0.0	0.0		0	0	0	0	0
1NRA	1ERA	11.7	22.9	4.1		63	25	25	25	17
1NRA	1FAS	10.7	0.0	0.0		0	0	0	0	0
1NRA	1KBS	11.0	42.4	9.0		50	48	42	30	49
1NRA	1NXB	12.3	0.0	0.0		0	0	0	0	0

1NRA	2CDX	11.0	0.0	0.0		0	0	0	0	0
1NRA	2CRT	11.0	0.0	0.0		0	0	0	0	0
1PTX	1COD	12.2	39.5	5.5		50	38	37	36	48
1PTX	1CRE	10.8	0.0	0.0		0	0	0	0	0
1PTX	1ERA	12.0	36.4	4.1		50	42	36	32	36
1PTX	1FAS	11.6	0.0	0.0		0	0	0	0	0
1PTX	1KBS	10.6	52.4	7.0		75	53	48	47	62
1PTX	1NXB	13.0	0.0	0.0		0	0	0	0	0
1PTX	2CDX	10.8	0.0	0.0		0	0	0	0	0
1PTX	2CRT	11.0	35.8	7.6		69	44	34	26	39
2SN3	1COD	9.9	0.0	0.0		0	0	0	0	0
2SN3	1CRE	10.9	29.2	8.1		67	32	23	23	39
2SN3	1ERA	9.9	0.0	0.0		0	0	0	0	0
2SN3	1FAS	10.3	48.4	12.3		59	35	52	43	64
2SN3	1KBS	10.6	0.0	0.0		0	0	0	0	0
2SN3	1NXB	11.0	0.0	0.0		0	0	0	0	0
2SN3	2CDX	10.5	19.9	12.5		70	15	38	10	17
2SN3	2CRT	10.9	0.0	0.0		0	0	0	0	0

Calibration Data for Set 4 Weighting Function Curves (55% octant decision score)

Protein Pair PDB Identifiers		RMS Score (Å)	Mean Minimum-CVA Similarity Score From MolCom3D (%)	Standard Deviation in Minimum-CVA Similarity Score (%)	Error Where Marked ***X***	Observed Minimum-CVA Similarity Score (%) At Degree of Cubic Overlap (%)				
						100	125	150	175	200
1AHO	1LQH	4.7	94.6	1.7		98	96	95	92	96
1AHO	1NRA	4.8	90.8	3.4		97	94	93	87	90
1AHO	1PTX	0.3	100.0	0.0		100	100	100	100	100
1AHO	2SN3	8.4	71.9	2.3		50	75	70	71	72
1LQH	1NRA	7.3	86.8	4.4		90	92	87	81	88
1LQH	1PTX	4.4	91.4	4.4		95	96	94	90	86
1LQH	2SN3	8.7	60.8	2.0	***X***	63	63	60	59	63
1NRA	1PTX	4.7	88.3	3.9		97	93	87	84	89
1NRA	2SN3	9.4	77.4	6.5		63	83	77	69	81
1PTX	2SN3	9.0	66.0	8.5	***X***	93	66	59	60	78
1COD	1CRE	7.9	82.5	4.1		63	86	84	77	83
1COD	1ERA	2.7	93.6	2.6		99	97	92	92	93
1COD	1FAS	6.5	79.7	1.0		85	81	79	79	79
1COD	1KBS	7.6	86.7	3.0		88	90	88	83	87
1COD	1NXB	6.2	80.2	3.8		50	86	79	78	77
1COD	2CDX	8.1	77.3	4.8		75	84	76	72	77
1COD	2CRT	7.7	86.2	2.4		88	88	85	84	89
1CRE	1ERA	7.5	77.6	3.9		63	79	80	72	80
1CRE	1FAS	4.8	74.6	3.3		63	79	74	71	75
1CRE	1KBS	4.6	96.3	1.6		88	98	96	95	96
1CRE	1NXB	9.6	71.5	6.5		38	80	69	65	72
1CRE	2CDX	4.8	86.5	1.8		88	89	86	85	87
1CRE	2CRT	3.7	95.6	0.5		100	96	96	95	96
1ERA	1FAS	5.8	76.2	2.8		74	78	78	72	77
1ERA	1KBS	7.2	84.4	5.4		97	90	84	77	86
1ERA	1NXB	5.7	90.3	0.8		96	91	91	90	89
1ERA	2CDX	7.6	75.4	3.2		84	79	77	72	73
1ERA	2CRT	7.4	86.5	3.8		98	90	88	81	87
1FAS	1KBS	4.9	74.2	4.2		77	76	73	69	79
1FAS	1NXB	8.4	76.0	5.1		85	71	73	78	82
1FAS	2CDX	4.9	74.9	4.4		91	77	73	70	80
1FAS	2CRT	4.6	72.5	5.0		83	72	71	68	80
1KBS	1NXB	9.3	74.3	3.0		75	77	76	71	74
1KBS	2CDX	5.1	87.1	1.2		63	86	89	86	87
1KBS	2CRT	4.6	97.0	1.6		100	99	98	96	95
1NXB	2CDX	7.5	71.1	4.0		81	74	68	68	76
1NXB	2CRT	7.3	63.4	7.8	***X***	86	69	58	56	71
2CDX	2CRT	4.4	88.0	2.2		75	90	87	86	90
1AHO	1COD	12.3	25.8	5.6		75	32	23	20	29
1AHO	1CRE	11.0	0.0	0.0		0	0	0	0	0

1AHO	1ERA	12.0	35.2	15.8		63	18	26	45	52
1AHO	1FAS	11.6	0.0	0.0		0	0	0	0	0
1AHO	1KBS	10.7	36.0	9.7		63	47	28	28	41
1AHO	1NXB	13.1	0.0	0.0		0	0	0	0	0
1AHO	2CDX	11.0	0.0	0.0		0	0	0	0	0
1AHO	2CRT	11.1	28.2	9.9		50	16	35	38	25
1LQH	1COD	12.1	26.6	7.4		38	20	37	22	27
1LQH	1CRE	11.5	0.0	0.0		0	0	0	0	0
1LQH	1ERA	12.4	28.0	17.7		50	11	34	17	50
1LQH	1FAS	11.7	0.0	0.0		0	0	0	0	0
1LQH	1KBS	11.1	46.2	3.9		63	46	48	41	50
1LQH	1NXB	13.2	0.0	0.0		0	0	0	0	0
1LQH	2CDX	11.4	0.0	0.0		0	0	0	0	0
1LQH	2CRT	11.6	37.7	12.2		61	46	40	20	45
1NRA	1COD	12.0	52.7	1.5		50	52	53	51	55
1NRA	1CRE	11.0	0.0	0.0		0	0	0	0	0
1NRA	1ERA	11.7	45.1	9.8		75	51	54	44	32
1NRA	1FAS	10.7	0.0	0.0		0	0	0	0	0
1NRA	1KBS	11.0	55.1	12.0		50	71	49	43	57
1NRA	1NXB	12.3	0.0	0.0		0	0	0	0	0
1NRA	2CDX	11.0	0.0	0.0		0	0	0	0	0
1NRA	2CRT	11.0	59.4	4.6		63	61	63	53	61
1PTX	1COD	12.2	47.4	17.4		56	44	37	36	73
1PTX	1CRE	10.8	0.0	0.0		0	0	0	0	0
1PTX	1ERA	12.0	50.6	13.5		50	46	41	45	71
1PTX	1FAS	11.6	0.0	0.0		0	0	0	0	0
1PTX	1KBS	10.6	56.1	10.1		75	58	49	48	70
1PTX	1NXB	13.0	0.0	0.0		0	0	0	0	0
1PTX	2CDX	10.8	0.0	0.0		0	0	0	0	0
1PTX	2CRT	11.0	50.4	14.1		63	50	39	42	70
2SN3	1COD	9.9	0.0	0.0		0	0	0	0	0
2SN3	1CRE	10.9	33.2	6.5		60	37	27	29	41
2SN3	1ERA	9.9	0.0	0.0		0	0	0	0	0
2SN3	1FAS	10.3	57.9	11.3		61	45	60	56	72
2SN3	1KBS	10.6	0.0	0.0		0	0	0	0	0
2SN3	1NXB	11.0	0.0	0.0		0	0	0	0	0
2SN3	2CDX	10.5	33.1	3.0		68	31	35	30	36
2SN3	2CRT	10.9	0.0	0.0		0	0	0	0	0

Calibration Data for Set 5 Weighting Function Curves (55% octant decision score)

Protein Pair PDB Identifiers		RMS Score (Å)	Mean Minimum-CVA Similarity Score From MolCom3D (%)	Standard Deviation in Minimum-CVA Similarity Score (%)	Error Where Marked *X*	Observed Minimum-CVA Similarity Score (%) At Degree of Cubic Overlap (%)				
						100	125	150	175	200
1AHO	1LQH	4.7	98.6	0.7		63	98	100	98	99
1AHO	1NRA	4.8	95.2	1.9		98	96	97	93	95
1AHO	1PTX	0.3	100.0	0.0		100	100	100	100	100
1AHO	2SN3	8.4	82.8	4.2		75	89	83	79	80
1LQH	1NRA	7.3	90.7	1.3		63	91	90	89	92
1LQH	1PTX	4.4	94.2	2.6		88	97	95	94	91
1LQH	2SN3	8.7	76.8	4.4		75	81	73	73	81
1NRA	1PTX	4.7	93.4	2.0		99	96	92	92	94
1NRA	2SN3	9.4	76.2	6.9		50	83	74	67	80
1PTX	2SN3	9.0	75.9	4.5		85	78	76	69	80
1COD	1CRE	7.9	62.1	0.8	*X*	50	61	63	63	63
1COD	1ERA	2.7	97.6	1.9		88	99	95	99	97
1COD	1FAS	6.5	79.7	1.8		75	81	81	77	79
1COD	1KBS	7.6	90.2	1.6		38	92	92	89	89
1COD	1NXB	6.2	78.2	3.9		75	80	77	74	83
1COD	2CDX	8.1	78.0	2.5		63	75	78	79	81
1COD	2CRT	7.7	90.6	1.4		75	89	92	92	90
1CRE	1ERA	7.5	83.1	2.4		63	84	85	80	84
1CRE	1FAS	4.8	74.5	3.1		50	74	72	72	79
1CRE	1KBS	4.6	95.4	0.8		100	96	95	94	96

1CRE	1NXB	9.6	70.3	4.0		38	74	72	65	69
1CRE	2CDX	4.8	86.6	2.4		75	90	87	84	85
1CRE	2CRT	3.7	94.7	1.3		99	97	95	94	94
1ERA	1FAS	5.8	76.3	3.7		91	81	76	73	76
1ERA	1KBS	7.2	87.5	1.8		88	90	87	87	86
1ERA	1NXB	5.7	85.9	1.6		95	88	85	85	86
1ERA	2CDX	7.6	75.8	1.7		87	78	75	75	75
1ERA	2CRT	7.4	86.7	0.7		87	86	88	87	86
1FAS	1KBS	4.9	73.5	4.6		80	78	69	70	77
1FAS	1NXB	8.4	75.0	3.9		80	73	71	76	80
1FAS	2CDX	4.9	73.3	6.0		86	69	69	74	81
1FAS	2CRT	4.6	73.2	3.6		74	72	75	69	77
1KBS	1NXB	9.3	76.2	5.2		75	83	75	71	76
1KBS	2CDX	5.1	84.9	1.5		63	85	84	83	87
1KBS	2CRT	4.6	98.4	0.7		100	98	99	99	97
1NXB	2CDX	7.5	71.9	4.6		78	74	71	66	77
1NXB	2CRT	7.3	65.4	7.3	***X***	80	69	60	59	74
2CDX	2CRT	4.4	87.3	2.5		87	90	86	85	88
1AHO	1COD	12.3	56.4	17.1		25	69	63	31	62
1AHO	1CRE	11.0	41.6	21.1		4	18	62	30	57
1AHO	1ERA	12.0	55.6	11.1		13	62	64	57	40
1AHO	1FAS	11.6	44.2	19.3		13	73	32	34	38
1AHO	1KBS	10.7	58.5	19.1		25	78	71	37	49
1AHO	1NXB	13.1	19.0	4.2		25	13	19	20	23
1AHO	2CDX	11.0	12.0	2.7		25	8	12	13	15
1AHO	2CRT	11.1	62.6	1.7		13	64	63	60	63
1LQH	1COD	12.1	45.2	15.7		25	61	56	29	35
1LQH	1CRE	11.5	41.6	18.8		13	58	57	21	31
1LQH	1ERA	12.4	60.0	3.5		38	64	60	55	61
1LQH	1FAS	11.7	34.8	16.9		13	24	60	25	30
1LQH	1KBS	11.1	69.3	5.4		13	76	69	63	70
1LQH	1NXB	13.2	26.1	10.4		0	16	24	24	40
1LQH	2CDX	11.4	29.3	9.2		25	24	21	42	30
1LQH	2CRT	11.6	72.3	4.3	***X***	25	78	73	68	70
1NRA	1COD	12.0	59.0	3.5		38	54	63	59	59
1NRA	1CRE	11.0	58.9	5.2		50	63	56	53	63
1NRA	1ERA	11.7	68.8	5.8		38	69	77	65	64
1NRA	1FAS	10.7	65.4	2.4		25	66	68	65	62
1NRA	1KBS	11.0	73.8	5.7	***X***	38	81	74	67	74
1NRA	1NXB	12.3	29.2	14.0		13	46	17	36	18
1NRA	2CDX	11.0	19.1	8.7		0	12	13	31	21
1NRA	2CRT	11.0	76.6	3.8	***X***	25	79	78	71	78
1PTX	1COD	12.2	49.6	11.2		40	40	50	44	65
1PTX	1CRE	10.8	40.1	15.0		56	44	30	27	60
1PTX	1ERA	12.0	50.4	13.6		44	44	44	43	71
1PTX	1FAS	11.6	37.4	10.8		53	31	35	30	53
1PTX	1KBS	10.6	58.8	10.1		63	62	55	48	71
1PTX	1NXB	13.0	32.0	3.4		44	29	33	29	36
1PTX	2CDX	10.8	46.0	12.6		63	44	37	38	64
1PTX	2CRT	11.0	48.4	15.8		63	44	38	40	72
2SN3	1COD	9.9	19.3	11.2		58	23	8	13	33
2SN3	1CRE	10.9	32.4	21.9		65	10	26	31	63
2SN3	1ERA	9.9	36.5	19.0		70	23	32	26	64
2SN3	1FAS	10.3	58.6	10.0		74	49	51	66	69
2SN3	1KBS	10.6	29.4	27.2		72	9	17	22	69
2SN3	1NXB	11.0	36.1	6.9		67	34	40	27	43
2SN3	2CDX	10.5	30.3	18.0		78	13	17	49	42
2SN3	2CRT	10.9	32.4	22.2		46	10	21	39	61

Calibration Data for Octant Decision Score = 50% (Using Set 1 Weighting Function Curves)											
Protein Pair PDB Identifiers		RMS Score (Å)	Mean Minimum-CVA Similarity Score From MolCom3D (%)	Standard Deviation in Minimum-CVA Similarity Score (%)	Error Where Marked *X*	Observed Minimum-CVA Similarity Score (%) At Degree of Cubic Overlap (%)				
						100	125	150	175	200
1AHO	1LQH	4.7	95.7	1.9		97	97	96	93	97
1AHO	1NRA	4.8	93.5	2.4		97	97	94	91	93
1AHO	1PTX	0.3	100.0	0.0		100	100	100	100	100
1AHO	2SN3	8.4	87.4	3.3		63	92	87	86	85
1LQH	1NRA	7.3	90.4	3.3		91	93	91	86	91
1LQH	1PTX	4.4	93.5	3.2		95	97	95	93	89
1LQH	2SN3	8.7	83.6	5.8		75	90	78	80	88
1NRA	1PTX	4.7	91.5	3.0		97	94	92	87	93
1NRA	2SN3	9.4	84.0	5.5		63	90	82	78	87
1PTX	2SN3	9.0	83.0	3.3		96	87	81	80	85
1COD	1CRE	7.9	85.4	4.0		85	90	85	80	87
1COD	1ERA	2.7	95.6	2.2		99	99	95	94	95
1COD	1FAS	6.5	87.1	2.1		88	88	87	84	89
1COD	1KBS	7.6	89.3	2.4		96	91	91	86	90
1COD	1NXB	6.2	90.1	3.0		87	94	89	87	91
1COD	2CDX	8.1	82.9	4.4		88	83	82	78	89
1COD	2CRT	7.7	91.4	1.7		98	92	91	89	93
1CRE	1ERA	7.5	89.2	3.0		88	92	91	85	89
1CRE	1FAS	4.8	81.7	6.2		87	90	79	76	81
1CRE	1KBS	4.6	99.1	1.1		99	100	99	100	98
1CRE	1NXB	9.6	80.0	6.3		63	87	74	75	84
1CRE	2CDX	4.8	93.1	1.1		88	94	93	91	94
1CRE	2CRT	3.7	97.9	1.2		100	97	97	98	99
1ERA	1FAS	5.8	83.6	2.1		94	86	84	81	83
1ERA	1KBS	7.2	88.6	3.1		99	92	88	84	90
1ERA	1NXB	5.7	95.6	1.7		98	95	98	96	94
1ERA	2CDX	7.6	82.5	2.9		94	86	82	79	83
1ERA	2CRT	7.4	90.6	3.6		98	95	91	86	91
1FAS	1KBS	4.9	81.4	3.2		80	85	80	78	83
1FAS	1NXB	8.4	85.6	2.5		90	88	83	84	88
1FAS	2CDX	4.9	83.8	4.5		93	88	82	78	87
1FAS	2CRT	4.6	80.2	3.0		90	77	80	79	84
1KBS	1NXB	9.3	80.8	6.8		87	90	75	77	82
1KBS	2CDX	5.1	90.9	1.2		92	89	91	91	92
1KBS	2CRT	4.6	98.3	1.3		100	100	99	97	98
1NXB	2CDX	7.5	81.5	4.5		89	84	81	75	86
1NXB	2CRT	7.3	80.7	5.8		90	85	76	75	86
2CDX	2CRT	4.4	92.9	2.3		96	95	93	90	94
1AHO	1COD	12.3	21.8	6.4		75	25	25	12	25
1AHO	1CRE	11.0	0.0	0.0		0	0	0	0	0
1AHO	1ERA	12.0	9.6	4.9		63	5	8	8	17
1AHO	1FAS	11.6	0.0	0.0		0	0	0	0	0
1AHO	1KBS	10.7	18.2	6.4		63	23	23	10	16
1AHO	1NXB	13.1	0.0	0.0		0	0	0	0	0
1AHO	2CDX	11.0	0.0	0.0		0	0	0	0	0
1AHO	2CRT	11.1	0.0	0.0		0	0	0	0	0
1LQH	1COD	12.1	0.0	0.0		0	0	0	0	0
1LQH	1CRE	11.5	0.0	0.0		0	0	0	0	0
1LQH	1ERA	12.4	27.3	11.4		63	11	33	27	38
1LQH	1FAS	11.7	0.0	0.0		0	0	0	0	0
1LQH	1KBS	11.1	25.0	0.0		63	25	25	25	25
1LQH	1NXB	13.2	0.0	0.0		0	0	0	0	0
1LQH	2CDX	11.4	0.0	0.0		0	0	0	0	0
1LQH	2CRT	11.6	0.0	0.0		0	0	0	0	0
1NRA	1COD	12.0	46.6	3.8		63	48	41	47	50
1NRA	1CRE	11.0	0.0	0.0		0	0	0	0	0
1NRA	1ERA	11.7	36.8	1.4		73	38	38	35	38
1NRA	1FAS	10.7	0.0	0.0		0	0	0	0	0
1NRA	1KBS	11.0	58.1	5.5		50	63	58	50	62

1NRA	1NXB	12.3	0.0	0.0		0	0	0	0	0
1NRA	2CDX	11.0	0.0	0.0		0	0	0	0	0
1NRA	2CRT	11.0	37.5	0.0		63	38	38	38	38
1PTX	1COD	12.2	53.6	13.6		63	50	44	47	74
1PTX	1CRE	10.8	0.0	0.0		0	0	0	0	0
1PTX	1ERA	12.0	59.9	10.8		63	50	57	58	75
1PTX	1FAS	11.6	0.0	0.0		0	0	0	0	0
1PTX	1KBS	10.6	63.9	7.1		75	69	62	55	70
1PTX	1NXB	13.0	0.0	0.0		0	0	0	0	0
1PTX	2CDX	10.8	0.0	0.0		0	0	0	0	0
1PTX	2CRT	11.0	60.4	11.3		69	63	58	47	74
2SN3	1COD	9.9	0.0	0.0		0	0	0	0	0
2SN3	1CRE	10.9	55.3	10.5		82	45	52	54	70
2SN3	1ERA	9.9	0.0	0.0		0	0	0	0	0
2SN3	1FAS	10.3	63.7	5.6		71	67	61	57	70
2SN3	1KBS	10.6	0.0	0.0		0	0	0	0	0
2SN3	1NXB	11.0	0.0	0.0		0	0	0	0	0
2SN3	2CDX	10.5	51.0	9.9		80	63	55	43	42
2SN3	2CRT	10.9	0.0	0.0		0	0	0	0	0

For Calibration Data for Octant Decision Score = 55% (Using Set 1 Weighting Function Curves)
See Calibration Data for Set 1 Weighting Function Curves (55% octant decision score)

Calibration Data for Octant Decision Score = 60% (Using Set 1 Weighting Function Curves)

Protein Pair PDB Identifiers		RMS Score (Å)	Mean Minimum-CVA Similarity Score From MolCom3D (%)	Standard Deviation in Minimum-CVA Similarity Score (%)	Error Where Marked *X*	Observed Minimum-CVA Similarity Score (%) At Degree of Cubic Overlap (%)				
						100	125	150	175	200
1AHO	1LQH	4.7	91.5	2.5		75	95	92	89	91
1AHO	1NRA	4.8	85.2	3.2		96	88	86	81	85
1AHO	1PTX	0.3	100.0	0.0		100	100	100	100	100
1AHO	2SN3	8.4	67.6	5.1	*X*	38	75	64	64	67
1LQH	1NRA	7.3	77.6	4.1		75	79	78	72	81
1LQH	1PTX	4.4	88.0	5.4		87	94	90	86	82
1LQH	2SN3	8.7	59.5	3.8	*X*	63	62	60	54	63
1NRA	1PTX	4.7	84.9	4.2		75	90	82	82	86
1NRA	2SN3	9.4	74.2	5.8		50	80	72	67	77
1PTX	2SN3	9.0	63.1	7.9	*X*	87	64	59	55	74
1COD	1CRE	7.9	78.1	3.9		63	82	76	74	80
1COD	1ERA	2.7	90.9	3.6		97	96	89	87	92
1COD	1FAS	6.5	75.0	2.0		75	74	76	73	77
1COD	1KBS	7.6	81.6	4.6		75	87	83	76	81
1COD	1NXB	6.2	80.3	2.8		88	84	81	77	80
1COD	2CDX	8.1	75.3	7.4		75	85	72	68	76
1COD	2CRT	7.7	81.7	3.2		73	85	80	78	83
1CRE	1ERA	7.5	74.3	4.6		70	77	78	68	75
1CRE	1FAS	4.8	72.2	4.2		61	77	70	68	74
1CRE	1KBS	4.6	93.7	1.5		99	96	93	92	94
1CRE	1NXB	9.6	66.8	6.1	*X*	50	72	68	58	69
1CRE	2CDX	4.8	83.5	3.6		88	89	82	81	82
1CRE	2CRT	3.7	94.6	1.4		100	93	97	95	94
1ERA	1FAS	5.8	72.9	3.0		91	74	74	69	75
1ERA	1KBS	7.2	79.6	4.5		93	85	80	74	80
1ERA	1NXB	5.7	90.5	2.7		96	93	91	92	87
1ERA	2CDX	7.6	72.2	5.0		89	77	76	67	69
1ERA	2CRT	7.4	79.3	4.4		92	85	79	74	80
1FAS	1KBS	4.9	69.6	5.4	*X*	76	73	66	64	75
1FAS	1NXB	8.4	72.4	4.2		79	69	73	69	78
1FAS	2CDX	4.9	71.6	4.6		90	75	71	65	75
1FAS	2CRT	4.6	68.9	3.9	*X*	79	66	68	67	75
1KBS	1NXB	9.3	71.0	9.2		75	81	74	59	70
1KBS	2CDX	5.1	84.1	3.0		73	88	86	81	82
1KBS	2CRT	4.6	94.4	2.4		100	97	96	92	92

1NXB	2CDX	7.5	66.2	7.9	X	83	68	63	58	76
1NXB	2CRT	7.3	63.2	7.2	X	81	70	59	55	69
2CDX	2CRT	4.4	85.9	3.9		63	90	86	81	87
1AHO	1COD	12.3	0.0	0.0		0	0	0	0	0
1AHO	1CRE	11.0	0.0	0.0		0	0	0	0	0
1AHO	1ERA	12.0	0.0	0.0		0	0	0	0	0
1AHO	1FAS	11.6	0.0	0.0		0	0	0	0	0
1AHO	1KBS	10.7	0.0	0.0		0	0	0	0	0
1AHO	1NXB	13.1	0.0	0.0		0	0	0	0	0
1AHO	2CDX	11.0	0.0	0.0		0	0	0	0	0
1AHO	2CRT	11.1	0.0	0.0		0	0	0	0	0
1LQH	1COD	12.1	0.0	0.0		0	0	0	0	0
1LQH	1CRE	11.5	0.0	0.0		0	0	0	0	0
1LQH	1ERA	12.4	0.0	0.0		0	0	0	0	0
1LQH	1FAS	11.7	0.0	0.0		0	0	0	0	0
1LQH	1KBS	11.1	0.0	0.0		0	0	0	0	0
1LQH	1NXB	13.2	0.0	0.0		0	0	0	0	0
1LQH	2CDX	11.4	0.0	0.0		0	0	0	0	0
1LQH	2CRT	11.6	0.0	0.0		0	0	0	0	0
1NRA	1COD	12.0	0.0	0.0		0	0	0	0	0
1NRA	1CRE	11.0	0.0	0.0		0	0	0	0	0
1NRA	1ERA	11.7	0.0	0.0		0	0	0	0	0
1NRA	1FAS	10.7	0.0	0.0		0	0	0	0	0
1NRA	1KBS	11.0	0.0	0.0		0	0	0	0	0
1NRA	1NXB	12.3	0.0	0.0		0	0	0	0	0
1NRA	2CDX	11.0	0.0	0.0		0	0	0	0	0
1NRA	2CRT	11.0	0.0	0.0		0	0	0	0	0
1PTX	1COD	12.2	0.0	0.0		0	0	0	0	0
1PTX	1CRE	10.8	0.0	0.0		0	0	0	0	0
1PTX	1ERA	12.0	0.0	0.0		0	0	0	0	0
1PTX	1FAS	11.6	0.0	0.0		0	0	0	0	0
1PTX	1KBS	10.6	0.0	0.0		0	0	0	0	0
1PTX	1NXB	13.0	0.0	0.0		0	0	0	0	0
1PTX	2CDX	10.8	0.0	0.0		0	0	0	0	0
1PTX	2CRT	11.0	0.0	0.0		0	0	0	0	0
2SN3	1COD	9.9	0.0	0.0		0	0	0	0	0
2SN3	1CRE	10.9	0.0	0.0		0	0	0	0	0
2SN3	1ERA	9.9	0.0	0.0		0	0	0	0	0
2SN3	1FAS	10.3	0.0	0.0		0	0	0	0	0
2SN3	1KBS	10.6	0.0	0.0		0	0	0	0	0
2SN3	1NXB	11.0	0.0	0.0		0	0	0	0	0
2SN3	2CDX	10.5	0.0	0.0		0	0	0	0	0
2SN3	2CRT	10.9	0.0	0.0		0	0	0	0	0

Calibration Data for Octant Decision Score = 65% (Using Set 1 Weighting Function Curves)

Protein Pair PDB Identifiers		RMS Score (Å)	Mean Minimum-CVA Similarity Score From MolCom3D (%)	Standard Deviation in Minimum-CVA Similarity Score (%)	Error Where Marked X	Observed Minimum-CVA Similarity Score (%) At Degree of Cubic Overlap (%)				
						100	125	150	175	200
1AHO	1LQH	4.7	84.0	4.2		63	87	84	78	87
1AHO	1NRA	4.8	77.9	3.2		83	81	80	74	77
1AHO	1PTX	0.3	100.0	0.0		100	100	100	100	100
1AHO	2SN3	8.4	49.5	3.9	X	25	53	44	50	51
1LQH	1NRA	7.3	72.3	6.7		69	80	76	68	65
1LQH	1PTX	4.4	77.8	6.9		78	88	74	74	75
1LQH	2SN3	8.7	31.7	12.2	X	50	25	25	27	50
1NRA	1PTX	4.7	75.2	4.7		63	80	72	70	78
1NRA	2SN3	9.4	56.8	16.2	X	38	71	59	34	63
1PTX	2SN3	9.0	41.4	19.0	X	75	25	49	27	65
1COD	1CRE	7.9	66.2	2.4	X	38	69	65	64	67
1COD	1ERA	2.7	82.9	3.6		83	84	83	78	87
1COD	1FAS	6.5	66.6	7.2	X	50	74	67	57	69
1COD	1KBS	7.6	71.8	6.4		63	78	72	63	75
1COD	1NXB	6.2	71.9	5.8		50	81	71	68	68
1COD	2CDX	8.1	60.0	3.2	X	38	60	57	58	65

1COD	2CRT	7.7	73.4	6.5		38	80	71	65	77
1CRE	1ERA	7.5	66.5	5.7	X	36	71	69	58	67
1CRE	1FAS	4.8	66.6	5.6	X	38	73	66	59	69
1CRE	1KBS	4.6	90.5	2.9		88	94	89	87	91
1CRE	1NXB	9.6	59.9	6.2	X	25	67	60	52	61
1CRE	2CDX	4.8	77.2	3.0		75	81	74	77	77
1CRE	2CRT	3.7	89.0	1.3		97	90	90	88	88
1ERA	1FAS	5.8	66.6	3.7	X	83	71	68	62	66
1ERA	1KBS	7.2	69.7	4.8	X	90	75	69	64	72
1ERA	1NXB	5.7	84.3	1.3		95	83	86	84	84
1ERA	2CDX	7.6	62.6	3.5	X	80	66	64	58	62
1ERA	2CRT	7.4	70.4	5.2		88	75	73	63	70
1FAS	1KBS	4.9	58.4	7.0	X	73	59	54	53	68
1FAS	1NXB	8.4	62.0	6.0	X	68	64	56	59	69
1FAS	2CDX	4.9	64.2	4.3	X	77	70	60	62	66
1FAS	2CRT	4.6	61.0	5.4	X	79	61	60	55	68
1KBS	1NXB	9.3	65.1	6.7	X	49	73	64	57	66
1KBS	2CDX	5.1	75.3	4.4		50	82	75	72	73
1KBS	2CRT	4.6	90.1	2.5		98	94	89	88	89
1NXB	2CDX	7.5	57.8	8.8	X	70	62	56	47	67
1NXB	2CRT	7.3	50.1	5.1	X	75	50	46	47	57
2CDX	2CRT	4.4	79.1	2.2		50	80	78	76	82
1AHO	1COD	12.3	0.0	0.0		0	0	0	0	0
1AHO	1CRE	11.0	0.0	0.0		0	0	0	0	0
1AHO	1ERA	12.0	0.0	0.0		0	0	0	0	0
1AHO	1FAS	11.6	0.0	0.0		0	0	0	0	0
1AHO	1KBS	10.7	0.0	0.0		0	0	0	0	0
1AHO	1NXB	13.1	0.0	0.0		0	0	0	0	0
1AHO	2CDX	11.0	0.0	0.0		0	0	0	0	0
1AHO	2CRT	11.1	0.0	0.0		0	0	0	0	0
1LQH	1COD	12.1	0.0	0.0		0	0	0	0	0
1LQH	1CRE	11.5	0.0	0.0		0	0	0	0	0
1LQH	1ERA	12.4	0.0	0.0		0	0	0	0	0
1LQH	1FAS	11.7	0.0	0.0		0	0	0	0	0
1LQH	1KBS	11.1	0.0	0.0		0	0	0	0	0
1LQH	1NXB	13.2	0.0	0.0		0	0	0	0	0
1LQH	2CDX	11.4	0.0	0.0		0	0	0	0	0
1LQH	2CRT	11.6	0.0	0.0		0	0	0	0	0
1NRA	1COD	12.0	0.0	0.0		0	0	0	0	0
1NRA	1CRE	11.0	0.0	0.0		0	0	0	0	0
1NRA	1ERA	11.7	0.0	0.0		0	0	0	0	0
1NRA	1FAS	10.7	0.0	0.0		0	0	0	0	0
1NRA	1KBS	11.0	0.0	0.0		0	0	0	0	0
1NRA	1NXB	12.3	0.0	0.0		0	0	0	0	0
1NRA	2CDX	11.0	0.0	0.0		0	0	0	0	0
1NRA	2CRT	11.0	0.0	0.0		0	0	0	0	0
1PTX	1COD	12.2	0.0	0.0		0	0	0	0	0
1PTX	1CRE	10.8	0.0	0.0		0	0	0	0	0
1PTX	1ERA	12.0	0.0	0.0		0	0	0	0	0
1PTX	1FAS	11.6	0.0	0.0		0	0	0	0	0
1PTX	1KBS	10.6	0.0	0.0		0	0	0	0	0
1PTX	1NXB	13.0	0.0	0.0		0	0	0	0	0
1PTX	2CDX	10.8	0.0	0.0		0	0	0	0	0
1PTX	2CRT	11.0	0.0	0.0		0	0	0	0	0
2SN3	1COD	9.9	0.0	0.0		0	0	0	0	0
2SN3	1CRE	10.9	0.0	0.0		0	0	0	0	0
2SN3	1ERA	9.9	0.0	0.0		0	0	0	0	0
2SN3	1FAS	10.3	0.0	0.0		0	0	0	0	0
2SN3	1KBS	10.6	0.0	0.0		0	0	0	0	0
2SN3	1NXB	11.0	0.0	0.0		0	0	0	0	0
2SN3	2CDX	10.5	0.0	0.0		0	0	0	0	0
2SN3	2CRT	10.9	0.0	0.0		0	0	0	0	0

B.2 Verification Test Data

Verification Data Using the Set 1 Weighting Function Curves

Protein Pair PDB Identifiers		RMS Score (Å)	Mean Minimum-CVA Similarity Score From MolCom3D (%)	Standard Deviation in Minimum-CVA Similarity Score (%)	Error Where Marked *X*	Observed Minimum-CVA Similarity Score (%) At Degree of Cubic Overlap (%)						
						100	125	150	175	200	225	250
1RGK	1RGL	0.3	100.0	0.0		100	100	100	100	100	100	100
1GMP	1RGE	0.3	100.0	0.0		100	100	100	100	100	100	100
1RGL	9RNT	0.5	99.9	0.0		100	100	100	100	100	100	100
1RGK	9RNT	0.5	99.9	0.1		100	100	100	100	100	100	100
1GMP	1SAR	0.5	100.0	0.1		100	100	100	100	100	100	100
193L	1LZB	0.5	93.7	3.8		83	94	88	92	98	94	97
2AAE	9RNT	0.6	100.0	0.0		100	100	100	100	100	100	100
1RGE	1SAR	0.6	100.0	0.0		100	100	100	100	100	100	100
193L	1RFP	0.7	100.0	0.0		100	100	100	100	100	100	100
1HEW	1LZB	0.7	99.9	0.1		100	100	100	100	100	100	100
1RDS	1RMS	0.7	100.0	0.0		100	100	100	100	100	100	100
1RFP	6LYT	0.7	98.3	2.4		83	94	98	100	98	100	100
1LZB	1RFP	0.7	98.7	2.4		83	94	100	99	100	99	100
1HEW	1RFP	0.8	96.5	5.3		83	98	100	100	98	97	86
193L	6LYT	0.9	94.6	2.6		83	94	94	91	98	94	97
1HEW	6LYT	0.9	99.9	0.1		100	100	100	100	100	100	100
1LZB	6LYT	1.0	99.9	0.2		100	100	100	100	100	100	100
1HEW	1LMA	1.0	99.8	0.3		100	100	100	100	100	99	100
193L	1LMA	1.0	97.1	2.6		88	98	94	98	94	99	99
1LMA	1LZB	1.0	99.9	0.1		99	100	100	100	100	100	100
1LMA	1RFP	1.1	99.2	0.8		98	100	99	99	98	100	100
1LMA	6LYT	1.2	99.8	0.2		99	100	100	100	100	100	100
1FUS	1RGL	1.8	87.9	1.8		96	90	90	88	88	86	85
1FUS	9RNT	1.8	87.9	1.6		96	90	89	88	88	87	86
1FUS	1RGK	1.8	87.8	1.7		97	90	89	88	88	87	85
1HWA	1RFP	2.4	96.6	1.0		97	98	97	96	96	97	96
1HEW	1HWA	2.4	92.0	1.0		83	91	93	91	93	91	93
1HWA	1LZB	2.5	94.4	1.8		88	95	97	94	93	95	92
193L	1HWA	2.5	96.8	1.0		99	97	98	96	98	97	95
1HWA	6LYT	2.5	94.4	1.6		88	95	97	94	93	95	93
1HWA	1LMA	2.6	93.5	3.1		75	96	96	93	94	95	88
1FUS	1RCL	3.2	95.5	1.4		99	96	97	96	94	96	94
1RCL	1RGL	3.6	81.1	3.8		86	80	75	79	84	84	85
1RCL	9RNT	3.6	82.5	3.9		86	80	78	79	87	85	85
1RCL	1RGK	3.6	81.3	4.0		86	81	74	79	84	84	84
1RCL	1RDS	4.2	81.2	4.3		86	80	73	81	85	83	85
1RCL	1RMS	4.2	80.1	4.2		86	78	74	78	84	84	83
1RDS	9RNT	4.4	89.4	1.9		88	91	90	88	86	91	91
1RMS	9RNT	4.4	88.6	2.6		97	88	90	84	89	90	92
1RDS	1RGK	4.4	89.6	1.4		88	92	90	88	89	90	89
1RGK	1RMS	4.4	88.4	2.3		94	90	86	85	89	90	90
1RDS	1RGL	4.4	89.4	1.7		88	92	90	87	88	90	89
1RGL	1RMS	4.4	88.4	1.8		94	88	87	86	89	91	90
1FUS	1RMS	4.6	81.8	3.4		89	86	78	78	82	83	84
1FUS	1RDS	4.6	81.9	3.1		89	87	81	77	81	82	83
193L	1HEW	6.0	85.1	2.8		83	80	86	85	87	85	88
1RGL	2AAE	6.7	99.5	0.6		100	100	99	100	99	100	99
1RGK	2AAE	6.8	99.8	0.4		100	100	100	100	99	100	100
1FUS	2AAE	7.2	87.0	2.1		97	91	85	85	87	87	86
1RCL	2AAE	8.0	81.9	4.0		86	80	77	78	87	85	84
1RDS	2AAE	8.7	89.7	1.0		93	88	91	90	91	89	90
1RMS	2AAE	8.7	89.4	2.2		97	89	88	86	91	90	91
1RGE	1HWA	13.1	0.0	0.0		0	0	0	0	0	0	0
1GMP	1HWA	13.1	38.2	9.4		64	52	40	41	25	31	40

1SAR	1HWA	13.2	44.5	6.0	65	51	50	47	35	42	42
1GMP	1LMA	13.2	34.0	12.1	51	21	26	48	28	31	50
1GMP	193L	13.3	0.0	0.0	0	0	0	0	0	0	0
1SAR	1LMA	13.3	36.0	14.6	48	20	28	30	30	57	51
1SAR	193L	13.3	0.0	0.0	0	0	0	0	0	0	0
1RGE	1LMA	13.3	42.4	11.2	54	25	49	48	32	51	51
1RGE	193L	13.3	0.0	0.0	0	0	0	0	0	0	0
1GMP	1RFP	13.3	19.4	5.1	62	19	12	19	16	25	25
1GMP	6LYT	13.3	58.0	2.0	54	59	58	55	60	57	60
1GMP	1LZB	13.3	56.7	2.1	46	60	57	53	57	57	57
1SAR	1RFP	13.3	22.1	10.3	68	19	19	19	8	37	32
1SAR	6LYT	13.3	48.4	14.9	52	30	56	29	58	58	60
1SAR	1LZB	13.3	48.1	14.7	52	30	55	29	58	58	59
1RGE	6LYT	13.3	56.8	4.7	67	49	56	56	62	57	61
1RCL	1LMA	13.6	0.0	0.0	0	0	0	0	0	0	0
1RCL	193L	13.6	0.0	0.0	0	0	0	0	0	0	0
1RCL	1RFP	13.7	0.0	0.0	0	0	0	0	0	0	0
1RCL	6LYT	13.7	57.8	3.9	66	59	54	52	61	61	60
1RCL	1LZB	13.7	58.1	3.1	75	60	55	53	61	61	59
1RCL	1HWA	13.8	0.0	0.0	0	0	0	0	0	0	0
1FUS	1LMA	13.9	35.3	7.9	47	35	34	20	43	39	40
1FUS	193L	14.0	0.0	0.0	0	0	0	0	0	0	0
1FUS	1RFP	14.0	0.0	0.0	0	0	0	0	0	0	0
1FUS	6LYT	14.0	51.6	12.5	65	62	54	27	60	55	52
1FUS	1LZB	14.0	51.8	12.1	67	62	52	29	60	56	53
1RDS	193L	14.3	0.0	0.0	0	0	0	0	0	0	0
1RDS	1LMA	14.3	0.0	0.0	0	0	0	0	0	0	0
1RDS	1RFP	14.3	0.0	0.0	0	0	0	0	0	0	0
1RDS	6LYT	14.3	37.3	16.3	65	43	17	20	36	57	52
1RMS	193L	14.3	0.0	0.0	0	0	0	0	0	0	0
1RDS	1LZB	14.3	34.4	13.2	67	43	16	22	36	38	51
1RMS	1LMA	14.3	0.0	0.0	0	0	0	0	0	0	0
1FUS	1HWA	14.3	0.0	0.0	0	0	0	0	0	0	0
1RMS	1RFP	14.3	0.0	0.0	0	0	0	0	0	0	0
1RMS	6LYT	14.4	33.8	10.1	74	37	29	18	37	34	49
1RMS	1LZB	14.4	27.1	14.6	75	14	13	17	36	35	48
1RGK	193L	14.4	0.0	0.0	0	0	0	0	0	0	0
1RGK	1LMA	14.4	48.3	9.5	48	46	36	40	59	50	59
9RNT	193L	14.5	0.0	0.0	0	0	0	0	0	0	0
1RGK	1RFP	14.5	0.0	0.0	0	0	0	0	0	0	0
1RGL	193L	14.5	0.0	0.0	0	0	0	0	0	0	0
1RGK	6LYT	14.5	50.5	11.6	44	48	38	35	59	61	61
9RNT	1LMA	14.5	44.6	12.6	0	37	30	34	53	51	62
1RGL	1LMA	14.5	48.6	10.0	48	49	36	38	59	51	59
9RNT	1RFP	14.5	0.0	0.0	0	0	0	0	0	0	0
1RGL	1RFP	14.5	0.0	0.0	0	0	0	0	0	0	0
9RNT	6LYT	14.5	49.8	11.1	52	39	43	37	59	62	59
1RGK	1LZB	14.5	51.1	11.3	52	48	39	37	59	61	63
1RGL	6LYT	14.5	51.3	11.3	48	47	39	38	59	61	64
2AAE	193L	14.5	0.0	0.0	0	0	0	0	0	0	0
2AAE	1LMA	14.5	43.3	8.9	58	38	34	39	56	53	41
9RNT	1LZB	14.5	49.8	11.8	48	36	44	38	59	62	60
1RGL	1LZB	14.5	51.5	11.1	52	48	39	38	59	61	62
2AAE	1RFP	14.5	0.0	0.0	0	0	0	0	0	0	0
2AAE	6LYT	14.5	47.8	10.0	62	36	44	41	58	62	45
2AAE	1LZB	14.5	49.1	11.8	59	35	40	40	59	62	58
1RDS	1HWA	14.6	0.0	0.0	0	0	0	0	0	0	0
1RMS	1HWA	14.7	0.0	0.0	0	0	0	0	0	0	0
1RGK	1HWA	14.7	0.0	0.0	0	0	0	0	0	0	0
1RGL	1HWA	14.8	0.0	0.0	0	0	0	0	0	0	0
9RNT	1HWA	14.8	0.0	0.0	0	0	0	0	0	0	0
2AAE	1HWA	14.8	0.0	0.0	0	0	0	0	0	0	0
1RCL	1HEW	14.9	0.0	0.0	0	0	0	0	0	0	0
1FUS	1HEW	15.1	42.3	11.4	59	44	40	21	50	51	49
1RDS	1HEW	15.3	0.0	0.0	0	0	0	0	0	0	0
1RMS	1HEW	15.3	0.0	0.0	0	0	0	0	0	0	0
1RGK	1HEW	15.6	47.3	8.7	23	44	39	37	53	55	57
1RGL	1HEW	15.6	47.4	9.1	23	45	39	36	55	53	58
9RNT	1HEW	15.6	47.0	9.7	23	38	36	40	56	55	57

2AAE	1HEW	15.6	50.8	6.0	57	47	43	47	56	56	57
1SAR	9RNT	17.5	54.0	6.7	72	61	56	50	59	56	43
1GMP	9RNT	17.5	50.6	13.1	73	62	56	27	59	44	56
1FUS	1RGE	17.5	59.6	4.0	70	65	55	55	62	62	59
1RGE	1RGL	17.6	55.6	4.6	64	50	60	50	59	58	56
1FUS	1SAR	17.6	50.6	12.3	60	56	26	52	57	60	52
1RGE	1RGK	17.6	55.9	4.2	65	50	60	51	59	59	57
1FUS	1GMP	17.6	49.8	11.9	61	55	26	52	55	60	51
1RGL	1SAR	17.6	50.2	6.8	38	45	44	43	55	59	55
1RDS	1SAR	17.6	48.1	9.5	71	47	36	47	42	53	64
1GMP	1RDS	17.6	58.1	4.4	76	61	60	50	63	56	59
1RDS	1RGE	17.6	40.1	18.1	64	44	13	26	44	50	64
1RGK	1SAR	17.6	45.8	11.6	38	26	43	42	51	59	54
1GMP	1RGL	17.6	59.3	4.4	73	64	59	52	62	58	61
1GMP	1RGK	17.6	59.2	4.4	73	64	59	52	62	58	61
1RMS	1SAR	17.6	49.8	10.4	65	41	37	45	55	57	64
1GMP	1RMS	17.6	56.6	7.9	78	64	56	54	63	60	43
1RGE	1RMS	17.7	49.2	15.1	68	55	30	30	62	59	59
1RGE	9RNT	17.7	59.1	1.7	63	59	61	58	62	59	57
1RCL	1RGE	17.8	57.7	5.9	72	58	51	51	60	67	60
1RCL	1SAR	17.8	59.9	5.7	77	66	57	54	61	54	67
1GMP	1RCL	17.9	62.7	4.1	76	58	67	62	68	58	63
1RGE	2AAE	17.9	57.8	2.1	60	60	59	54	59	57	58
1SAR	2AAE	18.0	57.9	3.3	72	62	56	54	62	59	55
1GMP	2AAE	18.0	57.8	3.3	75	61	55	53	61	59	58
1RGE	1RFP	19.1	0.0	0.0	0	0	0	0	0	0	0
1RGE	1LZB	19.1	34.9	7.5	28	26	32	31	32	45	43
1RGE	1HEW	19.3	28.4	4.5	0	29	32	32	20	32	26
1SAR	1HEW	19.4	42.4	15.6	68	25	59	28	38	63	42
1GMP	1HEW	19.4	42.5	16.3	68	25	59	28	38	65	40

Note: Using the mean minimum-CVA similarity score, no classification errors were made, and consequently the “Error” column is unmarked.

The End