

# Comparison of Protein Structures by Transformation into Dihedral Angle Sequences

by

**Doug L. Hoffman**

A Dissertation submitted to the faculty of The University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computer Science.

Chapel Hill

1996

Approved by:

---

Raj K. Singh, Advisor

---

Bruce W. Erickson, Reader

---

Jan F. Prins, Reader

Copyright © 1996  
Doug L. Hoffman  
All rights reserved

**DOUG L. HOFFMAN. Comparison of Protein Structures by  
Transformation into Dihedral Angle Sequences.  
(Under the direction of Raj K. Singh.)**

**ABSTRACT**

Proteins are large complex organic molecules that are essential to the existence of life. Decades of study have revealed that proteins having different sequences of amino acids can possess very similar three-dimensional structures. To date, protein structure comparison methods have been accurate but costly in terms of computer time. This dissertation presents a new method for comparing protein structures using dihedral transformations. Atomic XYZ coordinates are transformed into a sequence of dihedral angles, which is then transformed into a sequence of dihedral sectors. Alignment of two sequences of dihedral sectors reveals similarities between the original protein structures. Experiments have shown that this method detects structural similarities between sequences with less than 20% amino acid sequence identity, finding structural similarities that would not have been detected using amino acid alignment techniques. Comparisons can be performed in seconds that had previously taken minutes or hours.

# Contents

<b>List of Tables</b>	<b>vii</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Abbreviations</b>	<b>xi</b>
<b>1 Introduction and thesis</b>	<b>1</b>
<b>2 Background</b>	<b>3</b>
2.1 What are proteins? . . . . .	3
2.2 Comparison of protein structures . . . . .	5
2.3 Previous comparison methods . . . . .	5
<b>3 Representing protein structure using dihedral angle descriptors</b>	<b>9</b>
<b>4 Dihedral sequence comparison</b>	<b>12</b>
4.1 A characterization of the problem . . . . .	12
4.2 Simplification of the structural representation . . . . .	14
4.3 The dihedral transformation . . . . .	15
4.3.1 Calculation of the dihedral angle . . . . .	15
4.3.2 Calculation of the $C_\beta$ position for glycine . . . . .	17
4.4 Dihedral sequence alignment . . . . .	18
4.4.1 The choice of sequence alignment algorithm . . . . .	21
4.4.2 Implementation in custom hardware . . . . .	23
4.5 Computational complexity . . . . .	23

<b>5</b>	<b>Analysis of binification error</b>	<b>25</b>
5.1	Direct measurement of positional uncertainty . . . . .	27
5.2	Propagated error using partial derivatives . . . . .	27
5.3	Propagated error using interval arithmetic . . . . .	28
5.4	Impact of propagated error on bin size . . . . .	29
<b>6</b>	<b>Classes of dihedral angle descriptors</b>	<b>31</b>
6.1	Main-chain dihedral angles . . . . .	31
6.2	Pendant dihedral angles . . . . .	32
6.3	Statistics of descriptor angle distributions . . . . .	33
<b>7</b>	<b>Construction of the score table</b>	<b>46</b>
7.1	Relative information content scaling . . . . .	46
7.2	Statistical diffusion of score values . . . . .	48
7.3	Impact of the mismatch score . . . . .	67
<b>8</b>	<b>Experimental results</b>	<b>68</b>
8.1	Comparison of protein structure . . . . .	69
8.2	Dihedral sequence alignment vs. 3D structure alignment . . . . .	69
8.2.1	Alignment of 1mbd and 1bab . . . . .	69
8.2.2	Alignment of 1cd8 and 2rhe . . . . .	85
8.2.3	Alignment of 1rcf and 4fxn . . . . .	91
8.3	Discussion . . . . .	99
<b>9</b>	<b>Conclusions</b>	<b>100</b>
9.1	Future work . . . . .	101
<b>A</b>	<b>Relationship between the dihedral angles <math>\psi</math>, <math>\phi</math>, and oo1</b>	<b>102</b>
A.1	Generalized rotational transformation matrix. . . . .	102
A.2	Positions of the carbonyl carbon and oxygen atoms. . . . .	103
A.3	$C_i$ and $O_i$ postitions in local coordinates. . . . .	104
A.4	$C_{i-1}$ and $O_{i-1}$ postitions in local coordinates. . . . .	104
A.5	Translated positions of the carbonyl carbon and oxygen atoms. . . . .	105
A.6	Translated $C_i$ and $O_i$ postitions. . . . .	105
A.7	Translated $C_{i-1}$ and $O_{i-1}$ postitions. . . . .	106
A.8	Computation of the oo1 dihedral angle. . . . .	107

<b>B List of protein structures used</b>	<b>108</b>
<b>Bibliography</b>	<b>122</b>

# List of Tables

2.1	Existing protein structure comparison methods. . . . .	7
5.1	Bond length statistics for cdaz. . . . .	27
5.2	Probability of correct bin membership. . . . .	30
6.1	Pendant dihedral angle descriptors . . . . .	34
6.2	Bin frequencies for the bb descriptors. . . . .	36
6.3	Bin frequencies for the bo descriptors. . . . .	37
6.4	Bin frequencies for the ob descriptors. . . . .	38
6.5	Bin frequencies for the oo descriptors. . . . .	39
7.1	Diffusion coefficients based on probability of correct attribution ( $a$ ). .	48
7.2	Numeric diffusion coefficients give probability $a = 0.86$ . . . . .	49
7.3	Table of alignment scores based on the bb1 descriptor. . . . .	51
7.4	Table of alignment scores based on the bb2 descriptor. . . . .	52
7.5	Table of alignment scores based on the bb3 descriptor. . . . .	53
7.6	Table of alignment scores based on the bb4 descriptor. . . . .	54
7.7	Table of alignment scores based on the bo1 descriptor. . . . .	55
7.8	Table of alignment scores based on the bo2 descriptor. . . . .	56
7.9	Table of alignment scores based on the bo3 descriptor. . . . .	57
7.10	Table of alignment scores based on the bo4 descriptor. . . . .	58
7.11	Table of alignment scores based on the ob1 descriptor. . . . .	59
7.12	Table of alignment scores based on the ob2 descriptor. . . . .	60
7.13	Table of alignment scores based on the ob3 descriptor. . . . .	61
7.14	Table of alignment scores based on the ob4 descriptor. . . . .	62
7.15	Table of alignment scores based on the oo1 descriptor. . . . .	63
7.16	Table of alignment scores based on the oo2 descriptor. . . . .	64

7.17	Table of alignment scores based on the oo3 descriptor. . . . .	65
7.18	Table of alignment scores based on the oo4 descriptor. . . . .	66
8.1	Possible experimental outcomes. . . . .	68
8.2	Table of bb descriptor alignment scores for 1mbd vs. 1bab. . . . .	70
8.3	Table of bo descriptor alignment scores for 1mbd vs. 1bab. . . . .	71
8.4	Table of ob descriptor alignment scores for 1mbd vs. 1bab. . . . .	72
8.5	Table of oo descriptor alignment scores for 1mbd vs. 1bab. . . . .	73
8.6	Alignment statistics for 1cd8 vs. 2rhe. . . . .	85
8.7	Alignment statistics for 1rcf vs. 4fxn. . . . .	91
B.1	List of protein structures used . . . . .	109



# List of Figures

2.1	Definition of the $\phi$ , $\psi$ , and $\omega$ dihedral angles. . . . .	6
3.1	Definition of the $\tau$ dihedral angle defined by four consecutive $C_\alpha$ atoms.	10
3.2	Definition of the oo1 dihedral angle. . . . .	11
4.1	The process of structure comparison using dihedral sequences. . . . .	13
4.2	Diagonal path graph . . . . .	20
4.3	The BioSCAN algorithm in C. . . . .	22
6.1	The protein backbone showing the $O$ and $C_\beta$ pendant atoms. . . . .	33
6.2	The atoms defining the sixteen pendant dihedral angles. . . . .	35
6.3	Bin frequency for bb1 $\alpha$ -helix and $\beta$ -sheet regions. . . . .	40
6.4	Bin frequency for bo2 $\alpha$ -helix and $\beta$ -sheet regions. . . . .	41
6.5	Frequency plots for the bb descriptors. . . . .	42
6.6	Frequency plots for the bo descriptors. . . . .	43
6.7	Frequency plots for the ob descriptors. . . . .	44
6.8	Frequency plots for the oo descriptors. . . . .	45
8.1	Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the bb descriptors. . . . .	76
8.2	Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the bb descriptors (continued).	77
8.3	Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the bo descriptors. . . . .	78
8.4	Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the bo descriptors (continued).	79
8.5	Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the ob descriptors. . . . .	80

8.6	Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the ob descriptors (continued).	81
8.7	Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the oo descriptors. . . . .	82
8.8	Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the oo descriptors (continued).	83
8.9	3D ribbon diagrams of the $\alpha$ -helical proteins myoglobin (a) and hemoglobin beta chain (b). . . . .	84
8.10	Dihedral sequence alignment of the human T-cell co-receptor CD8 (1cd8) with the immunoglobulin lambda chain of human Bence-Jones protein RHE (2rhe) using the bb1, bo1, bo3, and bo4 descriptors. . .	87
8.11	Dihedral sequence alignment of the human T-cell co-receptor CD8 (1cd8) with the immunoglobulin lambda chain of human Bence-Jones protein RHE (2rhe) using the oo1, oo3, and oo4 descriptors. . . . .	88
8.12	Alignment of the human T-cell co-receptor CD8 (1cd8) with the immunoglobulin lambda chain of human Bence-Jones protein RHE (2rhe).	89
8.13	3D ribbon diagrams of the human T-cell co-receptor(a) and the immunoglobulin lambda chain of human Bence-Jones protein RHE (b).	90
8.14	Six alignments of dihedral sequences for the flavodoxins 1rcf and 4fxn.	92
8.15	Alignment of sequences for the flavodoxins from <i>Anabaena 7120</i> (1rcf) and <i>Clostridium MP</i> (4fxn). . . . .	94
8.16	Alignment of segments of the oo2 dihedral sequences of the flavodoxins from <i>Anabaena 7120</i> (1rcf) and <i>Clostridium MP</i> (4fxn). . . . .	95
8.17	Three dihedral alignments for the flavodoxins 1rcf and 4fxn using the oo2 descriptor. . . . .	96
8.18	3D ribbon diagrams of the flavodoxins 1rcf (a) and 4fxn (b). . . . .	97
8.19	3D ribbon diagrams of the flavodoxins 1rcf (a) and 4fxn (b). . . . .	98

# List of Abbreviations

<b>1D</b>	One-dimensional
<b>3D</b>	Three-dimensional
<b>BioSCAN</b>	Biological Sequence Comparative Analysis Node
<b>BLOSUM</b>	Block Substitution Matrix
<b>BNL</b>	Brookhaven National Laboratory
<b>DNA</b>	deoxyribonucleic acid
<b>FMN</b>	flavin mononucleotide
<b>NMR</b>	Nuclear Magnetic Resonance
<b>PDB</b>	Protein Data Bank
<b>RMSD</b>	Root Mean Square Deviation
<b>RNA</b>	ribonucleic acid
<b>VLSIC</b>	Very Large Scale Integrated Circuit
<b>WWW</b>	World Wide Web

# Chapter 1

## Introduction and thesis

Many methods have been used to compare and classify the three-dimensional (3D) structure of proteins. The time-honored method is visual comparison of actual model proteins by a biochemist or a molecular biologist, drawing on the scientist's own knowledge of other existing structures. Initially such models were made of brass or plastic, now models are displayed on a computer screen using graphics programs designed specifically for this purpose. Using computers in this way still depends on a scientist's visual interpretation of the model, introducing an element of subjectivity.

Not surprisingly, automated comparison methods utilizing digital computers have been created in an attempt to reduce the subjectivity and time of comparisons. Several of these methods have yielded accurate and interesting results, but require a great deal of computation time. A computationally efficient algorithm that can accurately compare 3D protein structures, filtering the data available from the database of known structures, would be a useful tool for protein scientists.

The thesis of this dissertation is that,

two protein structures can be effectively compared by aligning their sequences of pendant dihedral angles.
--

This dissertation presents a method for comparing 3D protein structures that reduces the problem from three dimensions to one dimension based on a technique called dihedral transformation. It compares the transformed dihedral sequences using a one-dimensional sequence alignment algorithm. It will be shown that finding alignments between dihedral angle sequences correspond to finding alignments between the respective protein structures. Experimental results support this thesis and show

that this algorithm is an accurate and effective filter capable of generating useful, significant, and even unexpected results.

# Chapter 2

## Background

### 2.1 What are proteins?

Proteins are large complex organic molecules that are the fundamental material of life. Cell walls and internal structures are built from proteins. Muscles, tendons, hair, and finger nails are made of protein, as are fish scales and the exoskeletons of insects. Aside from forming the structure of living creatures proteins are Nature's own nanomachines, functioning as the assemblers and disassemblers of other organic molecules, essential to the energy production and storage mechanisms of cells. DNA (deoxyribonucleic acid) is called the code of life; what the code contains are instructions for making proteins.

The study of proteins started with the investigation of so called "albuminous" substances in the eighteenth century. These were substances, like egg whites and globulin from blood, that coagulate when heated rather than turn from a solid to a liquid like most other substances. By the mid nineteenth century attempts had been made to work out the basic chemical formula of such substances; the Dutch chemist Gerardus Johannes Mulder worked out a basic formula and named it "protein", from the Greek word for "of first importance" (Asimov, 1960, page 68). Mulder's formula was much simpler than those of actual proteins because methods for determining the chemical composition of such large molecules did not exist at the time. Never the less, the name "protein" came into common use for this entire class of substances.

What Mulder did not know, though by the end of the nineteenth century others were starting to suspect, was that proteins are composed of a set of standard building

blocks. Some of these building blocks had been isolated as early as the 1820's but the connection to proteins was not made till the end of the century. Each of these building blocks contains an amine group ( $NH_2$ ) and a carboxyl group ( $COOH$ ), joined by a central carbon atom, the alpha carbon ( $C_\alpha$ ). Because of the carboxyl group these substances have the characteristics of an acid, causing them to be christened "amino acids".

Also bound to the  $C_\alpha$  atom is a group of atoms, called the side-chain, that gives an amino acid its chemical identity. The simplest amino acid, glycine, has a side-chain consisting of only a hydrogen atom. All other amino acids have a more complex, carbon containing side-chain. In all, there are twenty different types of amino acids that are found in natural proteins.

Toward the end of the nineteenth century the connection between amino acids and proteins was being uncovered. It was believed that proteins were somehow formed from amino acid building blocks, though how the amino acids were joined was not yet understood. After the turn of the century Emil Fischer managed to form chains of amino acids. He discovered that these chains formed by always linking the carboxyl group of one amino acid to the amine group of another. Though the substances he synthesized were too small to exhibit protein like properties, each having fewer than twenty linked amino acids, Fischer believed that proteins were formed in this way and that proteins, when digested, broke down into similar amino acid chains. He named these amino acid chains "peptides" from the Greek word for "digestion" and the link between the carboxyl carbon and the amine group a "peptide bond" (Asimov, 1960, page 72).

Today we know that proteins are polypeptide chains that are long enough to fold up into a characteristic shape. This characteristic shape helps to determine the chemical and biological functions of the protein. It is thought that the folded shape of a protein molecule is determined by the types and order of its amino acids. Unfortunately, prediction of a protein's folded shape (fold) from its amino acid sequence is not currently possible.

Proteins with similar amino acid sequences are thought to be closely related in evolutionary terms and are expected to exhibit similar folds. Proteins that are closely related can be identified by comparing their amino acid sequences but, as proteins evolve their sequences may diverge, eventually leading to proteins with dissimilar sequences that have similar structure and function. It has been estimated that approximately 50% of globular protein folds are known at this time so it is expected

that many new folds still await discovery (Johnson et al., 1994). For these reasons the study of proteins requires comparison not only of amino acid sequence similarity but of 3D structural similarity as well.

## 2.2 Comparison of protein structures

Advances in X-ray crystallography and Nuclear Magnetic Resonance (NMR) techniques have significantly increased the rate of new protein structure determination. The repository of known protein structures, the Brookhaven National Laboratory (BNL) Protein Data Bank (PDB) (Bernstein et al., 1977), contains over 1000 distinct folds and more than 4000 individual file entries. The growth of this database parallels the earlier experiences of protein (SWISS-PROT) and nucleic acid (GenBank) sequence databases. As discovered some time ago with protein amino acid sequence data, computational methods for automated comparison and classification of protein structures are essential for the efficient utilization of the growing structure database.

The complexity of protein structure results from the flexibility of the polypeptide backbone. The folded shape of the amino acid backbone in a polypeptide chain is determined by three types of dihedral (torsional) angles: phi ( $\phi$ ), psi ( $\psi$ ), and omega ( $\omega$ ). The  $\phi$  angle is the dihedral angle defined by the positions of the  $C_{i-1}$ ,  $N_i$ ,  $C_i^\alpha$ , and  $C_i$  atoms; the  $\psi$  angle is defined by  $N_i$ ,  $C_i^\alpha$ ,  $C_i$ , and  $N_{i+1}$ ; and the  $\omega$  angle by  $C_i^\alpha$ ,  $C_i$ ,  $N_{i+1}$ , and  $C_{i+1}^\alpha$ , where  $i$  is the residue number (Fig. 2.1). The  $\omega$  angle describes two possible forms of the peptide bond: the *trans* form ( $\omega = 180$ ) and the *cis* form ( $\omega = 0$ ). In both cases small deviations from planarity can occur. The *cis* form is observed in only 0.5% of all non-proline residues, whereas 5.6% of prolines are found in the *cis* form (Thornton, 1992). Thus, the flexibility of the protein backbone is almost entirely defined by rotations (by angles  $\phi$  and  $\psi$ ) around single bonds adjacent to the  $C^\alpha$  atoms.

## 2.3 Previous comparison methods

Several methods have been developed to perform 3D alignments between protein structures. They include generalized forms of sequence alignment algorithms using dynamic programming (Orengo et al., 1992; Russell and Barten, 1993), the use of inter-atom geometric relationships (Holm and Sander, 1993), 3D clustering (Vriend and Sander, 1991), and statistically based methods utilizing Monte Carlo techniques



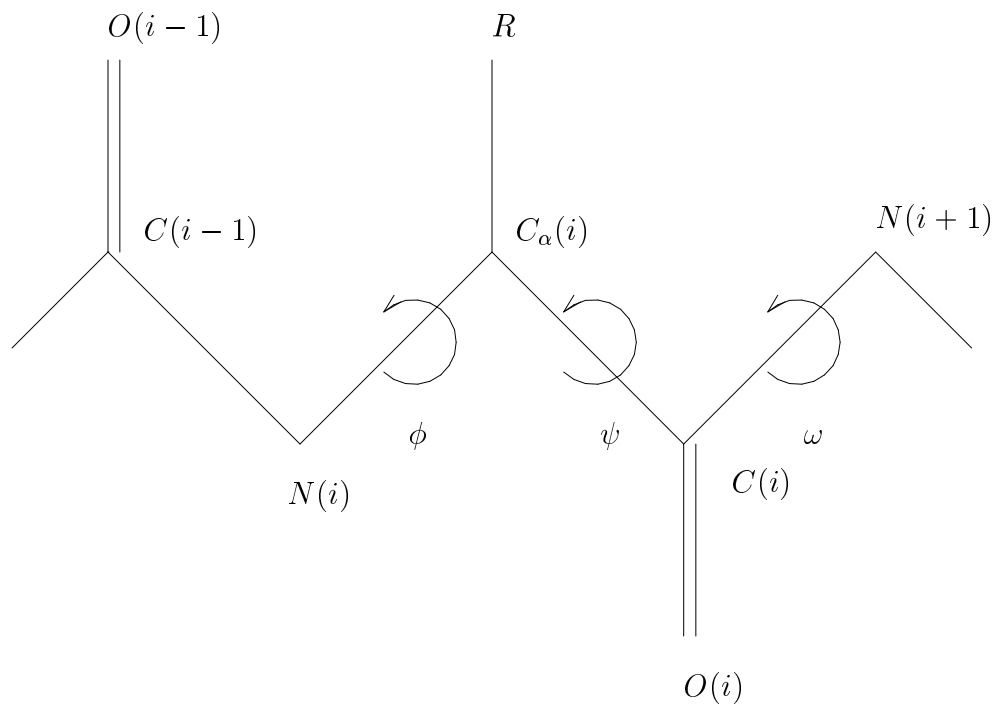


Figure 2.1: Definition of the  $\phi$ ,  $\psi$ , and  $\omega$  dihedral angles.

(Sali and Blundell, 1990). These methods have various drawbacks, including the need for an initial alignment, the inability to handle large insertions or deletions, failure to detect repeated patterns, and the requirement that the protein be broken into a limited number of secondary structures which are then aligned. Use of probabilistic methods to limit the size of the search space introduces the possibility of missing significant alignments. These problems can adversely impact the sensitivity and effectiveness of a comparison method. The most common limiting factor, however, is computational speed. The time required ranges from a few hours to tens of hours for comparing a single structure against any significant portion of the PDB.

Shown in Table 2.1 is a summary of recently developed and currently used structure comparison methods drawn from a survey by Orengo (Orengo, 1994). Most of these methods have been derived from earlier works and represent many years of algorithm and program development effort. When execution times have been reported they have been included in the table.

Method	Structural Descriptor	Algorithm used	CPU Time (minutes)	Reference
COMPARER	Multiple	Dynamic programing, simulated annealing	not published	(Sali and Blundell, 1990)
CONGENEAL	Intramolecular $C_\alpha$ distances	Overlap of distance matrices	not published	(Yee and Dill, 1993)
DALI	Intramolecular $C_\alpha$ distances	Monte Carlo optimization	5-10 min SPARC-1	(Holm and Sander, 1993)
Godzik method	Intramolecular $C_\alpha$ distances	Monte Carlo, simulated annealing	10-20 min (pair)	(Godzik et al., 1993)
Nussinov Wolfson method	Intramolecular vectors	Geometric hashing	1 min (pair) SUN4	(Nussinov and Wolfson, 1989) (Bachar et al., 1993)
PROTEP	Secondary structure, distances, and angles	Graph theory clique algorithm	15 min (search)	(Grindley et al., 1993) (Artymiuk et al., 1989)
SSAP	Intramolecular $C_\beta$ vectors	Double dynamic programing	25 min (search)	(Taylor and Orengo, 1989)
STAMP	Intermolecular $C_\alpha$ distances	Dynamic programing and rigid-body superposition	6 min (pair) SPARC-1	(Russell and Barten, 1993)
Subbiah method	Intermolecular $C_\alpha$ distances	Rigid-body superposition	not published	(Subbiah et al., 1993)
WHATIF	Intermolecular $C_\alpha$ distances (fragments)	Rigid-body superposition	not published	(Vriend and Sander, 1991)

Table 2.1: Existing protein structure comparison methods.

In summary, the faster methods are not sensitive to low scoring alignments or can omit significant alignments because of algorithmic optimizations that limit the search space. The more rigorous methods do not have these problems but are very time consuming.

## Chapter 3

# Representing protein structure using dihedral angle descriptors

The efficiency and effectiveness of a protein structure alignment method is rooted in the choice of structural representation. It has been shown that working directly from the atomic coordinates in three-space results in accurate but computationally intensive algorithms (Orengo et al., 1993). Efforts using a simplified structural representation, predominantly inter-atom distance, have resulted in faster algorithms but still require a great deal of time for a complete database search (on the order of hours). Our starting point in the search for a more efficient comparison method is with the structural representation.

An effective description of the 3D geometry of a protein can be achieved by considering simpler structural descriptors than atomic spatial coordinates (Levitt and Warshel, 1975; Warshel and Levitt, 1976; Laiter et al., 1995). Oldfield and Hubbard (Oldfield and Hubbard, 1994) have shown that protein structural analysis can be performed using only the  $C_\alpha$  atoms of the polypeptide backbone. For each ordered set of four successive  $C_\alpha$  atoms in the polypeptide backbone, they defined the dihedral angle  $\tau$  (Fig. 3.1). Their work indicated that the  $C_\alpha$  geometry of the protein chain constrains the  $\tau$  angle to fall into well defined regions, much like the  $\phi$  and  $\psi$  angles are constrained into well defined regions in a Ramachandran plot.

My colleagues and I have shown (Laiter et al., 1995) that the oo1 dihedral angle (called OCCO in the referenced paper), defined by the  $O_{i-1}$ ,  $C_{i-1}$ ,  $C_i$ , and  $O_i$  atoms (where  $i$  is the residue number), is a useful descriptor of protein secondary structures.

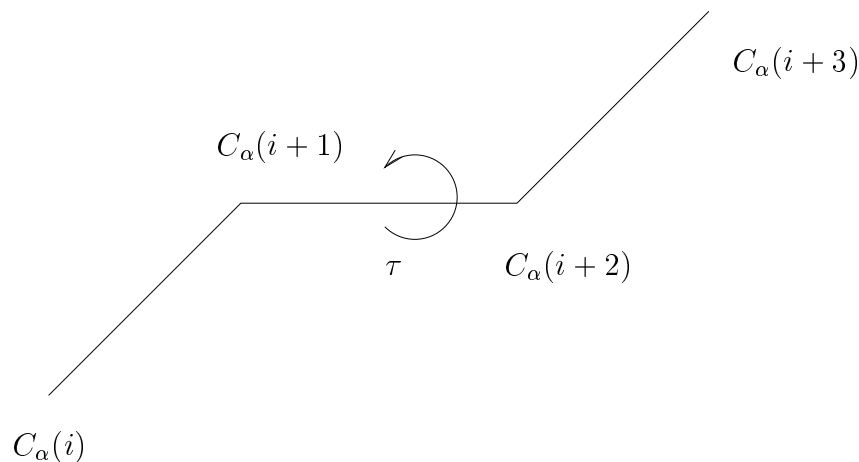


Figure 3.1: Definition of the  $\tau$  dihedral angle defined by four consecutive  $C_\alpha$  atoms.

We have found that it is capable of discriminating between secondary structures as effectively as the  $\phi$  and  $\psi$  angles. The relationship between the  $\phi$  and  $\psi$  angles and the oo1 dihedral angle is derived in Appendix A. As will be shown in Chapter 6 this particular dihedral angle is only one of a larger family of pendant dihedral angles.

The decision to use dihedral angles to represent protein structure leads to the question of how to compare the values of these angles. Applying a dihedral angle transformation to a protein structure results in a sequence of angle values that must be aligned in some way. The individual vector components could be compared as floating-point values, with values in the range of  $-180^\circ$  to  $180^\circ$ , but how a match should be scored is not clear. A further computational simplification is to binify the sequence elements, transforming the dihedral angle sequence into a dihedral sector sequence. Each dihedral sector is represented by a character from the English alphabet. Thus a dihedral angle sequence is effectively transformed into a string of characters, the dihedral sequence.

Dihedral angle descriptors such as  $\tau$  and oo1 provide a single measure of local structure for each residue, as opposed to the pair of angles when using  $\phi$  and  $\psi$ . It is straightforward to convert these dihedral angle values that describe the polypeptide backbone into a 1D structure sequence that can be analyzed by linear sequence

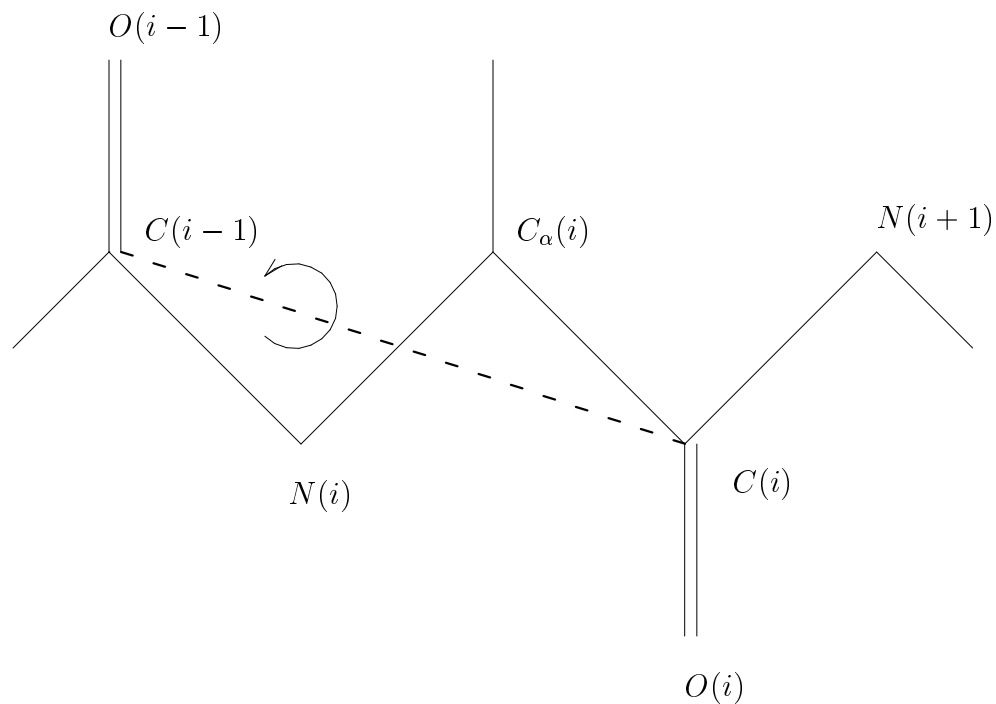


Figure 3.2: Definition of the oo1 dihedral angle. The backbone atoms of two consecutive amino acid residues are shown. The dotted line represents the  $C_{i-1}C_i$  line segment. The oo1 angle is defined by rotation (shown as solid arc) about this line segment.

comparison algorithms. The idea that a protein's 3D structure can be effectively represented by a character string is the basis for the dihedral sequence comparison algorithm.

# Chapter 4

## Dihedral sequence comparison

This chapter presents the algorithm used in comparing 3D structures using dihedral angle sequence alignment. An overview of the comparison process is shown in Figure 4.1. Starting with the XYZ atomic coordinates of a protein molecules constituent atoms a sequence of dihedral angles is generated, one angle associated with each amino acid of the protein.

First, a statement of the problem will be given in non-chemical terms. Second, the assumptions, based on structural constraints, that allow simplification of the structural representation will be presented. Third, the method of dihedral angle transformation for converting a 3D structure into a 1D character string is outlined. Fourth, the 1D character string, called a dihedral sequence, is used in an algorithmic step involving dihedral sequence alignment. The method used for aligning the dihedral sequences is described in the fifth part of this chapter. Finally, a discussion of the computational complexity of the algorithms is presented.

### 4.1 A characterization of the problem

A protein molecule can be thought of as a collection of points in Euclidean space, each point representing the three-dimensional position of one atom. The problem at hand is to find all superpositions of two such sets of points in three space such that some local subregion in one set is equivalent to a subregion in the other set. A subregion is a subset of points belonging to a set where the points are adjacent to one another in three-space. To be equivalent the two subregions must contain the same number

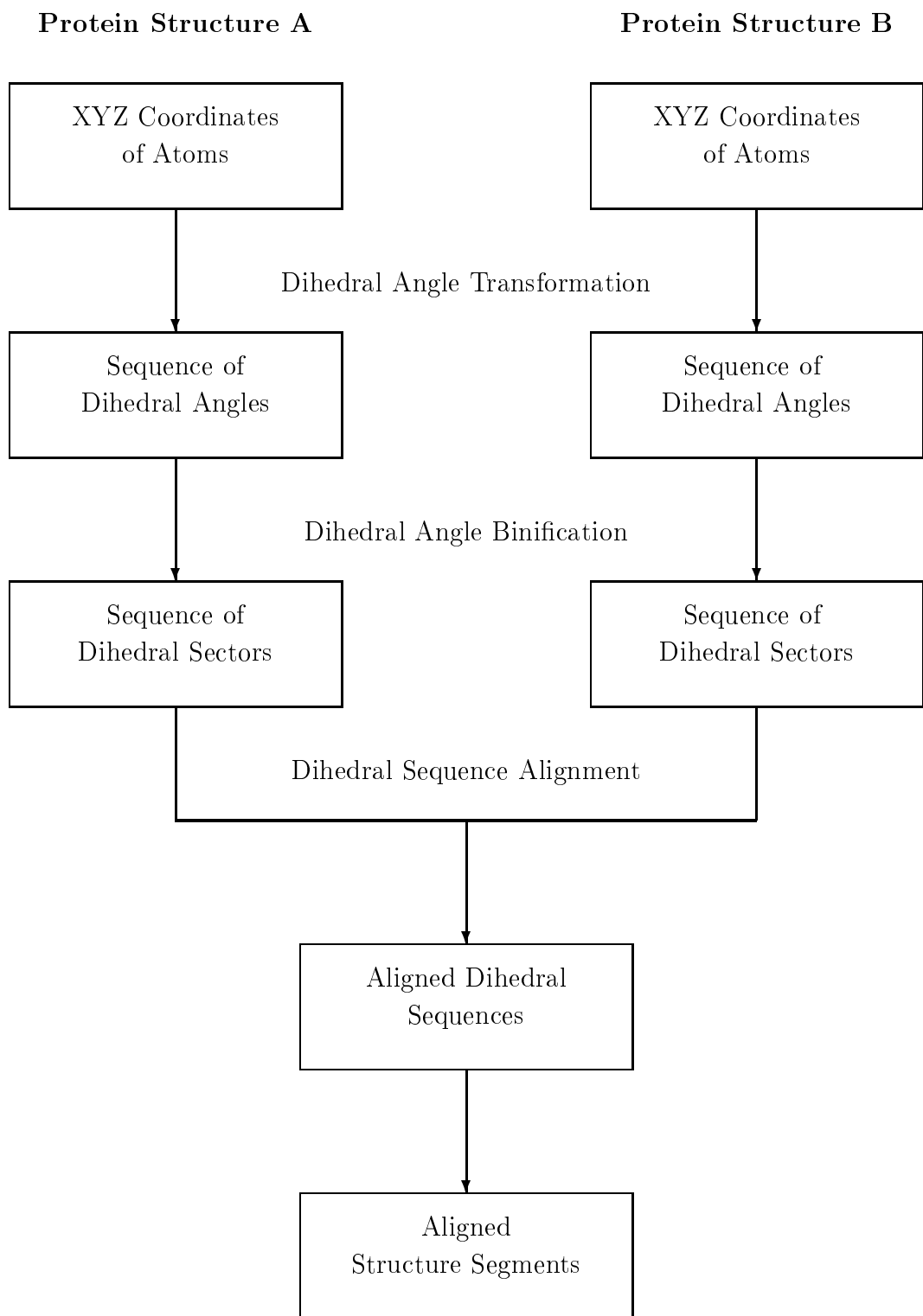


Figure 4.1: The process of structure comparison using dihedral sequences.



of points, the two subsets must be placed in a one-to-one correspondence, and the Root Mean Square Deviation (RMSD) between the positions of corresponding pairs of points must not exceed some given amount (the magnitude of this amount will be discussed later). The subregions can contain three or more points, up to the total size of the smaller collection. To obtain a superposition one set is subjected to rotations and translations in all three dimensions.

## 4.2 Simplification of the structural representation

As described above the problem is combinatorially prohibitive. However, a number of restrictions apply to the geometry of the points in a set allowing the problem to be simplified. First, the points are not independent of one another as they are connected together by rigid links that are fixed in length. The angle of incidence between any two adjacent links is also fixed.

These links connect each point to no more than two neighboring points. If the set is thought of as a graph, with the points as vertices and the links as edges, there are no cycles in the graph. The linked points appear as a chain or ribbon in space, tangled and compacted in on itself. There may be several disconnected ribbons in a set but each ribbon contains no branches and does not loop back on itself.

Second, the points along a ribbon are always grouped in sets of three. Each triplet has an N terminus and a C terminus, the third point forming the connection to the other two. In linking triplets together N termini will only link with C termini; there are no N to N or C to C links joining triplets. This restriction gives the ribbon a direction. This combination of ribbon topology and direction defines a natural ordering for the points in any collection.

For the three points in a triplet, if the N point is assigned position  $n$ , then the center point of the triplet is  $n + 1$  and the C point is  $n + 2$ . For any N point in a collection one of the following must be true.

1. If N has no link to a C point then N is the first point in a strand of ribbon and is labeled 0.
2. If N has a link to a C point and the the C point is labeled  $i$ , N is labeled  $i + 1$ .

These two rules for labeling points in a set define a unique ordering for the points in each individual ribbon. This allows the search space for subregion definitions to be

reduced to those subregions defined by segments of ribbon, a segment being defined as a subset of points belonging to a ribbon where each point is adjacent to one or more other points in the segment.

There is another simplification that can be made to the structural representation based on the observation that the points on a ribbon always appear as triplets in a fixed geometric relationship. By treating each triplet as a unit, with a single value from the dihedral transformation associated with each unit, the number of points to be compared can be reduced by two thirds.

### 4.3 The dihedral transformation

The process of transforming the 3D protein structure to a 1D character string (the dihedral sequence) entails calculating a dihedral angle for each unit in the ribbon. The angle value is assigned an alphabetic character by dividing the angle range of  $-180^\circ$  through  $180^\circ$  into bins of equal width and assigning one character to each bin. The width of each bin, and hence the size of the alphabet, is determined based on the expected uncertainty in the dihedral angle values. The uncertainty analysis is presented in Section 5.4. The bin width used in obtaining the results presented in this document was  $15^\circ$  such that, during binification,  $[-180^\circ, -165^\circ)$  is assigned an A,  $[-165^\circ, -150^\circ)$  is assigned a B, and so on.

There are a large number of dihedral angles that can be used for structure comparison, though some prove to be more effective than other. Regardless of the angle being used, the computation of a dihedral angle given four points in three-dimensional space is the same. For definitions of possible dihedral angles please refer to Chapter 6. The method for computing a dihedral angle is presented in the next section.

#### 4.3.1 Calculation of the dihedral angle

Given four points in three-dimensional space, represented by the vectors  $\vec{\mathbf{p}}_1$ ,  $\vec{\mathbf{p}}_2$ ,  $\vec{\mathbf{p}}_3$ , and  $\vec{\mathbf{p}}_4$ , where the points belonging to the triplets  $(\vec{\mathbf{p}}_1, \vec{\mathbf{p}}_2, \vec{\mathbf{p}}_3)$  and  $(\vec{\mathbf{p}}_2, \vec{\mathbf{p}}_3, \vec{\mathbf{p}}_4)$  are not co-linear, we can define the dihedral angle determined by these points,  $\tau$ , as a function of the twelve individual Cartesian coordinates as follows.

$$\begin{aligned}
\vec{\mathbf{p}}_1 &= \{x_1, y_1, z_1\} \\
\vec{\mathbf{p}}_2 &= \{x_2, y_2, z_2\} \\
\vec{\mathbf{p}}_3 &= \{x_3, y_3, z_3\} \\
\vec{\mathbf{p}}_4 &= \{x_4, y_4, z_4\}
\end{aligned} \tag{4.1}$$

We define three relative difference vectors  $\vec{\mathbf{v}}_{21}$ ,  $\vec{\mathbf{v}}_{23}$ , and  $\vec{\mathbf{v}}_{43}$  using the positional vectors defined in Equation 4.1.

$$\begin{aligned}
\vec{\mathbf{v}}_{21} &= \vec{\mathbf{p}}_2 - \vec{\mathbf{p}}_1 \\
\vec{\mathbf{v}}_{23} &= \vec{\mathbf{p}}_2 - \vec{\mathbf{p}}_3 \\
\vec{\mathbf{v}}_{43} &= \vec{\mathbf{p}}_4 - \vec{\mathbf{p}}_3
\end{aligned} \tag{4.2}$$

We then define the vectors  $\vec{\mathbf{n}}_1$  and  $\vec{\mathbf{n}}_2$  as vectors normal to the plains described by the triplets  $(\vec{\mathbf{p}}_1, \vec{\mathbf{p}}_2, \vec{\mathbf{p}}_3)$  and  $(\vec{\mathbf{p}}_2, \vec{\mathbf{p}}_3, \vec{\mathbf{p}}_4)$ , respectively. This is determined by taking the cross-products of the appropriate relative difference vectors.

$$\begin{aligned}
\vec{\mathbf{n}}_1 &= \vec{\mathbf{v}}_{21} \times \vec{\mathbf{v}}_{23} \\
\vec{\mathbf{n}}_2 &= \vec{\mathbf{v}}_{23} \times \vec{\mathbf{v}}_{43}
\end{aligned} \tag{4.3}$$

Finally, we compute the angle of incidence between the two normal vectors  $\vec{\mathbf{n}}_1$  and  $\vec{\mathbf{n}}_2$ .

$$\tau = \arccos\left(\frac{\vec{\mathbf{n}}_1 \cdot \vec{\mathbf{n}}_2}{\|\vec{\mathbf{n}}_1\| \|\vec{\mathbf{n}}_2\|}\right) \tag{4.4}$$

The scalar value  $\tau$  is the dihedral angle. Since the arccos function returns a value between  $0^\circ$  and  $180^\circ$  one final step is required to determine the sign of the rotation.

$$\begin{aligned}
\vec{\mathbf{r}} &= \vec{\mathbf{n}}_1 \times \vec{\mathbf{n}}_2 \\
\text{if } \vec{\mathbf{n}}_2 \cdot \vec{\mathbf{r}} < 0 &\text{ then } \tau = -\tau
\end{aligned} \tag{4.5}$$

### 4.3.2 Calculation of the $C_\beta$ position for glycine

Glycine, the simplest of all amino acids, has a side chain that consists of a single hydrogen atom in the position other amino acid have their  $C_\beta$  atom. Since the Cartesian coordinates of hydrogen atoms are normally not included in the PDB structure files, calculating dihedral angles involving  $C_\beta$  atoms for glycine requires determining the position where the associated  $C_\beta$  would have been in a non-glycine amino acid.

To compute the position of the missing  $C_\beta$  atom we start by defining the following three vectors, each relative to the position of the  $C_\alpha$  atom.

$$\begin{aligned}\vec{\mathbf{r}}_1 &= \{x_N - x_{C_\alpha}, y_N - y_{C_\alpha}, z_N - z_{C_\alpha}\} \\ \vec{\mathbf{r}}_2 &= \{x_C - x_{C_\alpha}, y_C - y_{C_\alpha}, z_C - z_{C_\alpha}\} \\ \vec{\mathbf{r}}_3 &= \{x_{C_\beta} - x_{C_\alpha}, y_{C_\beta} - y_{C_\alpha}, z_{C_\beta} - z_{C_\alpha}\}\end{aligned}\tag{4.6}$$

Normalizing these vectors yields three new vectors.

$$\begin{aligned}\vec{\mathbf{n}}_1 &= \frac{\vec{\mathbf{r}}_1}{\|\vec{\mathbf{r}}_1\|} \\ \vec{\mathbf{n}}_2 &= \frac{\vec{\mathbf{r}}_2}{\|\vec{\mathbf{r}}_2\|} \\ \vec{\mathbf{n}}_3 &= \frac{\vec{\mathbf{r}}_3}{\|\vec{\mathbf{r}}_3\|}\end{aligned}\tag{4.7}$$

We also define a fourth vector,  $\vec{\mathbf{n}}_4$ , as the cross product of  $\vec{\mathbf{r}}_1$  and  $\vec{\mathbf{r}}_2$ .

$$\vec{\mathbf{n}}_4 = \vec{\mathbf{r}}_1 \times \vec{\mathbf{r}}_2\tag{4.8}$$

Because the bonds of the  $C_\alpha$  atom form a tetrahedron, the following relationships hold.

$$\begin{aligned}\vec{\mathbf{n}}_1 \cdot \vec{\mathbf{n}}_3 &= \|\vec{\mathbf{n}}_1\| \|\vec{\mathbf{n}}_3\| \cos(109^\circ) \\ \vec{\mathbf{n}}_2 \cdot \vec{\mathbf{n}}_3 &= \|\vec{\mathbf{n}}_2\| \|\vec{\mathbf{n}}_3\| \cos(109^\circ) \\ \vec{\mathbf{n}}_4 \cdot \vec{\mathbf{n}}_3 &= \|\vec{\mathbf{n}}_4\| \|\vec{\mathbf{n}}_3\| \cos(54.5^\circ)\end{aligned}\tag{4.9}$$

Since the lengths of all the  $\vec{n}$  vectors is unity, these equations can be simplified as follows.

$$\begin{aligned}\cos(109^\circ) &= \vec{n}_1 \cdot \vec{n}_3 \\ \cos(109^\circ) &= \vec{n}_2 \cdot \vec{n}_3 \\ \cos(54.5^\circ) &= \vec{n}_4 \cdot \vec{n}_3\end{aligned}\tag{4.10}$$

This is a system of three equations in three unknowns that can be solved to yield values for the  $x$ ,  $y$ , and  $z$  components of the  $\vec{n}_3$  vector. Scaling this vector by the average length of the  $C_\alpha$  to  $C_\beta$  bond and adding it to the coordinates of the  $C_\alpha$  atom gives the desired values of  $x_{C_\beta}$ ,  $y_{C_\beta}$ , and  $z_{C_\beta}$ .

## 4.4 Dihedral sequence alignment

Comparison of 1D dihedral sequences is performed using sequence alignment techniques originally developed for comparing amino acid sequences. The need for automated methods for searching protein and nucleic acid sequence databases was recognized several a decades ago. Algorithms for detecting and measuring similarities between biological sequences have evolved over the last three decades, starting with the work of Devereux *et al.* on the WORDSEARCH program for the GCG package (Devereux et al., 1984), FASTP (Lipman and Pearson, 1985), FASTA (Pearson and Lipman, 1988), and including more recent algorithms, such as BLAST (Altschul et al., 1990). These algorithms have come to be accepted and well understood as an extensive body of research has been developed over the years.

One of the earliest exhaustive sequence comparison methods is by Needleman and Wunsch (Needleman and Wunsch, 1970). Subsequently, many algorithmic refinements have been offered, most notably by Smith and Waterman (Smith and Waterman, 1981) and Altschul *et al.* (Altschul et al., 1990). For computational purposes a DNA or RNA molecule is represented as a long string of characters drawn from a four-character alphabet. Similarly, protein molecules are represented as strings of characters from a twenty-character alphabet. Generally, given two sequences of lengths  $m$  and  $n$ , each of  $m^2$  contiguous segments of one sequence are compared with all  $n$  equal-length segments of the other sequence. The score of each segment pair is the sum of the scores of the aligned characters. A two-dimensional lookup table provides a score

for each possible pair of aligned characters. Several high-scoring segment alignments may be combined to estimate the overall similarity of the two sequences. Generally, a global similarity score can be computed between two sequences by these algorithms in  $O(nm)$  time.

Though a number of different alignment algorithms could be used for dihedral sequence alignment I have chosen to use the BioSCAN (Biological Sequence Comparative Analysis Node) linear similarity algorithm (White et al., 1991; Singh et al., 1993; Singh et al., 1996). This algorithm generates non-gapped alignments, very similar to BLAST (Altschul et al., 1990) but with greater sensitivity to long, low scoring alignments (Karlin and Altschul, 1993). A more detailed rationale for this choice is give in Section 4.4.1.

Each segment of the query sequence, defined as a contiguous substring of any length, is compared with each equal-length segment of every database sequence. The algorithm, in effect, traverses a search space that is proportional to the product of the length of the query sequence and the total length of the scanned database. For a sequence comparison between two sequences,  $A$  and  $B$ , with lengths  $m$  and  $n$ , respectively, the search space can be visualized as the interior of the rectangle in Figure 4.2.

Pairs of segments (one from each sequence) deemed similar are represented as diagonally-oriented line segments of various lengths. Segment pairs represented on the same diagonal correspond to the same alignment of sequences  $A$  and  $B$ . For any particular alignment between  $A$  and  $B$ , there is a constant difference in the indices of each pair of aligned characters. This provides a convenient way to number the alignments (and corresponding diagonals). Hence, diagonal  $(i - j)$  in Figure 4.2 contains the aligned letters  $A_i$  and  $B_j$ . The linear similarity score is mathematically expressed as:

$$S_{i,j}^L = \sum_{k=0}^{L-1} \mathcal{T}(A_{i-k}, B_{j-k}) \quad (4.11)$$

$S_{i,j}^L$  is the score of the segment pair of length  $L$  with rightmost elements  $A_i$  and  $B_j$  in alignment.  $\mathcal{T}$  is the score table. The general recurrence relation is given as:

$$S_{i,j} = S_{i-1,j-1} + \mathcal{T}(A_{i-k}, B_{j-k}) \quad (4.12)$$

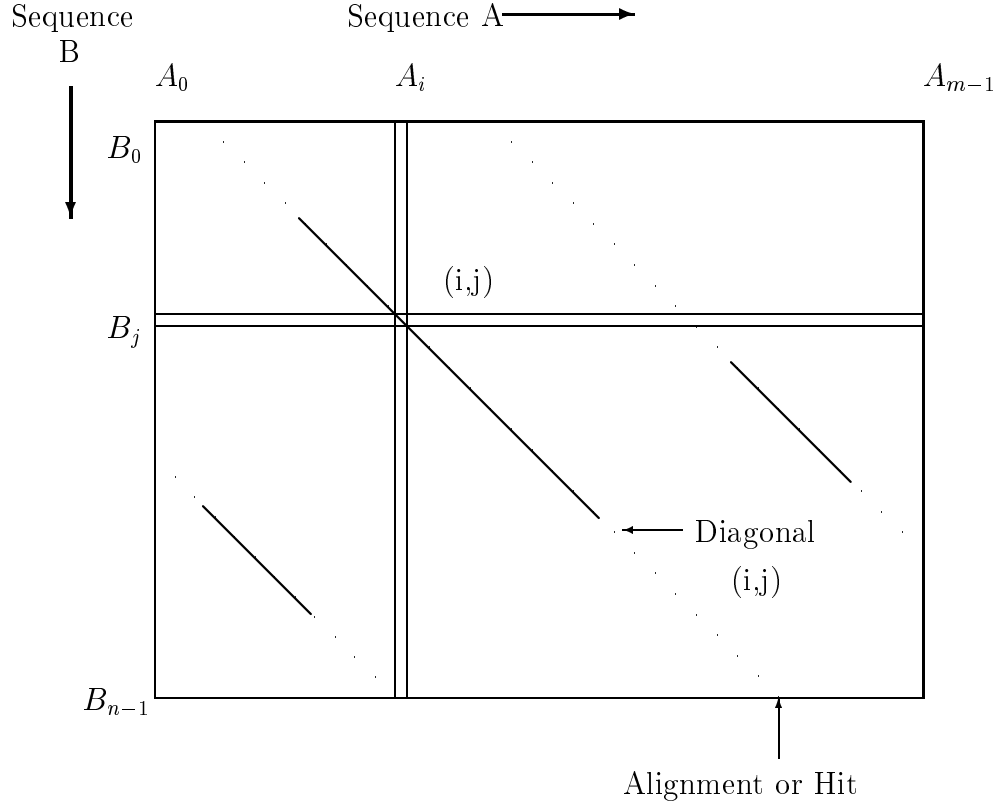


Figure 4.2: Diagonal path graph

For each possible alignment of A with B the following conditional logic is applied.

$$\begin{aligned}
 & \text{if } (S_{i-1,j-1} < 0) \quad S_{i,j} = \mathcal{T}(A_i, B_j) \\
 & \text{else if } (S_{i-1,j-1} \geq S^*) \quad S_{i,j} = S_{i-1,j-1} \\
 & \text{else} \quad S_{i,j} = S_{i-1,j-1} + \mathcal{T}(A_i, B_j)
 \end{aligned} \tag{4.13}$$

where,  $S_{x,y}$  is  $\max_L(S_{x,y}^L)$  while  $S_{x,y}$  is less than  $S^*$  (a user-specified score threshold).

Figure 4.3 describes the algorithm as a "C" program fragment. Note that the score sum is not allowed to accumulate a negative value; if the sum becomes negative, it is set to the alignment score for the current character pair. The threshold test identifies high-scoring alignments, insuring that a threshold-exceeding segment score is not lost

in trying to extend a high-scoring segment to include additional aligned characters.

#### 4.4.1 The choice of sequence alignment algorithm

There exist a number of sequence alignment algorithms that could be used to compare dihedral sequences. Before selecting the BioSCAN algorithm, I also considered the BLAST and the Smith-Waterman algorithms, two of the more popular algorithms used for protein sequence data comparison.

The BLAST algorithm utilizes a hashing technique that compares local blocks of characters. Its computational complexity is  $O(m)$ , where  $m$  is the length of the database that the query sequence is being compared with. If a local segment of six to eight characters has significant similarity (where significance is defined by user set-able parameters), then the location is marked for further investigation by more sophisticated software. Though fast this algorithm can miss marginally aligned subsequences that maintain a low average score over a fairly long subsequence length. The tendency has been observed by S. Altschul, the author of both the BLAST and BioSCAN algorithms.

The Smith-Waterman sequence alignment algorithm is based on dynamic programming techniques. Like the BioSCAN algorithm, its running time order of growth is  $O(mn)$ , where  $m$  is the length of the database and  $n$  is the length of the query sequence. This algorithm has many desirable features for the comparison of biosequences; insertions and deletions can be explicitly dealt with and an optimal set of mutation events transforming one sequence into another is produced. Though it can yield more data, these data are of little use in structure comparison. This nullifies any advantage the Smith-Waterman algorithm might have over the BioSCAN algorithm. Moreover, since the Smith-Waterman algorithm performs significantly more expensive computations in its inner loop the BioSCAN algorithm has a significant advantage in execution time.

In summary, the algorithm chosen is more sensitive than the BLAST algorithm and is more efficient than the dynamic programming algorithm while not sacrificing any useful capabilities. The next section will discuss the availability of this algorithm in a custom, parallel hardware implementation. The availability of the BioSCAN algorithm in hardware was not a major factor in selecting it for use in this application. However, my work on the implementation of the BioSCAN system (Hoffman, 1993b) familiarized me with the features and capabilities of its underlying algorithm.



```

#include <stdio.h>

#define M 28
#define N 28

/* BioSCAN declaration */
    unsigned int    La, Lb;    /* length A, B */
    char            A[La],    /* sequence A - padded */
                  B[Lb];    /* sequence B */
    signed int      T[M][N], /* similarity table */
                  Th;        /* threshold */
/* Outputs */
    signed int      S[Lb+1]; /* scores S[1..Lb] */

/* Temporaries */
    int            i, j;      /* index A and B */

extern report();

/* Algorithm */
main()
{
    for (i=0; i<Lb; ++i)    S[i] = 0;        /* initialize */
    for (i=0; i<La; ++i) {
        for (j=Lb-1; j>=0; --j)
            if (S[j] < 0) S[j+1] = T[A[i]][B[j]];
            else if (S[j] >= Th) S[j+1] = S[j];
            else S[j+1] = S[j] + T[A[i]][B[j]]; /* accumulate */
        if (S[Lb] >= Th) report(i);
    }
}

```

Figure 4.3: The BioSCAN algorithm in C.

### 4.4.2 Implementation in custom hardware

Some alignment algorithms lend themselves to efficient Very Large Scale Integrated Circuit (VLSIC) implementation. This has led to the development of a class of parallel processors designed expressly to operate on biosequence data. These efforts have resulted in PNAC (Lopresti, 1987), BISP (Chow et al., 1991), SPLASH (Gokhale et al., 1990), BSYS (Hughey and Lopresti, 1991), Bioccelerator (Compugen, 1995), and BioSCAN (White et al., 1991; Singh et al., 1993).

The BioSCAN system is a network-accessible, computational resource (Singh et al., 1996) that performs rapid exhaustive searches of DNA, RNA, and protein sequence databases. Originally intended for sequence comparison, I have adapted the BioSCAN computer to perform 3D protein structure comparison using 1D dihedral sequences (Hoffman et al., 1995). BioSCAN is comprised of hardware-based pre-filtering and software-based post-processing stages. The custom VLSIC implements the algorithm given in Equation 4.13. It identifies significant alignments which contain one or more high-scoring segment pairs. The system scans two million database characters per second and its performance is independent of the query sequence length. Each identified alignment is re-scanned and analyzed in software to determine the scores, lengths, and the end-points of all locally-optimal (non-overlapping) segment pairs.

## 4.5 Computational complexity

The dihedral sequence alignment method consists of two computationally distinct steps, the transformation of 3D structures into dihedral sequences and the search for aligned segments. The first step requires the calculation of a dihedral angle value (see Section 4.3.1) for each amino acid in a protein. These values form a 1D vector that is converted into a dihedral sequence by binifying the individual angle values. This process has computational complexity of  $O(n)$  though the constant factors are fairly high. Note that both the query structure and the structure database must be transformed into dihedral sequences. Of course the database need only be converted once, and is reused many times with new query structures.

The second step, aligning the dihedral sequences, has computational complexity of  $O(n^2)$  or more accurately  $O(mn)$ , where  $n$  is the length of the query dihedral sequence and  $m$  is the length of the database of known structures, with the expectation that  $m \gg n$ . The simplicity of the inner loop makes the BioSCAN algorithm computationally expedient, requiring a table lookup, an integer addition, and an in-

teger comparison to determine the result of the computation. Other methods that employ dynamic programming techniques are also  $O(n^2)$  but have much more complex computations in their inner loops. It is the austerity of the inner loop that gives the BioSCAN alignment algorithm its great computational advantage. The costly and complex floating-point calculations have been moved out of the inner loop and into the  $O(n)$  transformation step.

Several competing methods actually perform nested dynamic programming with a complexity of  $O(m^2n^2)$ . Other algorithms that directly compute RMSD in their inner loop usually rely on the quaternion method (Kearkey, 1990), since it provides a closed form solution for the superposition. Though the complexity of the quaternion method is  $O(n)$  its constant factors are large, involving a great deal of floating-point calculation. This algorithm forms an inner loop that must be run for all substructures of the query structure positioned with all possible alignments of the database. Such superposition algorithms are of order  $O(n^3)$ .

In practice the sequence alignment algorithm used by the dihedral sequence alignment method executes many times faster than competing algorithms. When used in conjunction with the BioSCAN custom hardware, this method is several orders of magnitude faster than the published times for the fastest competing methods (see Table 2.1).

# Chapter 5

## Analysis of binification error

An essential step in the algorithm being presented is the translation of the dihedral angle values into a sequence of characters suitable for use by sequence alignment algorithms. This is analogous to discretizing a continuous function. The dihedral angles are assigned to bins with discrete angle ranges (binified). Each of the bins is assigned an alphabetic character. The size of the alphabet is directly related to the size of the bins. For example, with bin size of  $20^\circ$  the alphabet size would be  $360^\circ/20^\circ = 18$  characters. It is the ordered list of these characters that form the structure's dihedral sequence. The process of binification is straightforward, but determining the size of the alphabet to use is not so simple.

Choosing the bin size and hence the size of the alphabet is a matter of tradeoffs. A small number of characters will reduce the sensitivity of the comparison by lumping together large ranges of angles. As shown in Section 6.3, the preferred range of angles associated with alpha helical secondary structures for some dihedral angle descriptors is a sharp peak that is only  $15^\circ$  to  $20^\circ$  wide. An alphabet containing only six characters would correspond to a bin size of  $60^\circ$ , making a match between alpha helical regions very fuzzy. To maintain the best possible selectivity during sequence alignment, the largest possible alphabet should be used.

The lower limiting factor on the bin size is uncertainty in the angle data being converted. Uncertainty in the atomic coordinates of a protein model are propagated through calculation to the dihedral angles themselves. Too large an alphabet will cause the angle data to be over sampled. To establish a reasonable minimum bin size, bounds must be placed on the error in the dihedral angles.

There is uncertainty associated with the atomic coordinates of any model of a molecular structure. This positional uncertainty derives partly from the modeling process and partly due to thermal vibration in the molecular structure itself. Whether displayed physically as a stick and ball model or electronically as an image on a computer screen, molecular structures appear to be rigid and un-moving. In reality, protein molecules are in constant motion; the individual atoms vibrate with thermal energy, bonds stretch and bend, and large pieces of the molecule wiggle about. Data collected from such active subjects, even in crystalline form, tends to be a bit blurred.

Accepting that the positions of the atoms in a protein molecule are dynamic, the process of determining the positions of the constituent atoms is itself not exact. Protein structures are resolved by two methods, X-ray crystallography and Nuclear Magnetic Resonance (NMR). Of the methods two X-ray crystallography is arguably more accurate, though to use this method a protein must be available in crystalline form.

Using either method is an iterative process, working from experimentally determined raw information that does not completely or unambiguously define the structure being studied. Using computer modeling tools (Xplor, O, etc.) models are constructed containing atom types and coordinates. These models are then checked for consistency against the experimental data. It is not uncommon for errors in published structures to remain uncorrected for several years. The same set of experimental data will often be resolved into significantly different structures by different researchers. PDB files based on NMR data may contain several dozen possible structure model, all derived from the same experimental data.

How accurate are the atomic coordinates in a PDB file? A good X-ray derived protein structure, like 4fxn (Smith et al., 1977), have a stated resolution of 1.8 Å for the entire structure. This translates to uncertainty of 0.1 Å for individual atoms (Cantor and Schimmel, 1980). Using this value as an initial error estimate it is possible to compute the propagated error in a typical dihedral angle computation. Calculating the propagated error using an initial coordinate uncertainty of 0.1 Å yields dihedral angle errors of greater than 10°. This level of error seemed unreasonable when compared with observed deviation in the bond angles. Bond angles vary around an average value and so variation in the angle data can be directly observed. Unfortunately, dihedral angles range over the full 360° of rotation with no meaningful average value around which variation can be measured. Because of this an alternative approach to directly measuring the positional uncertainty is needed.

## 5.1 Direct measurement of positional uncertainty

The heavy atoms of the protein backbone tend to be more precisely resolved than the average atom in a model structure. For the purposes of the dihedral transformation, we are most interested in the uncertainty in the relative positions of atoms, not absolute positional uncertainty with respect to the coordinated frame. This uncertainty is in the distances between the points used to calculate the dihedral angles. To find a bound on this relative positional error, direct measurement of the inter-atomic distances was performed, using several highly refined structures. Statistics for a typical protein structure are shown in Table 5.1.

measure	$N - C_\alpha$	$C_\alpha - C$	$C - N$
min	1.434 Å	1.489 Å	1.307 Å
mean	1.451 Å	1.519 Å	1.331 Å
max	1.477 Å	1.539 Å	1.363 Å
stddev	0.007 Å	0.007 Å	0.008 Å

Table 5.1: Bond length statistics for cdaz.

As a result of these measurements I have assumed a positional uncertainty of 0.01 Å for a typical backbone atom in a well resolved structure.

## 5.2 Propagated error using partial derivatives

To compute a first estimate of the propagated error in the dihedral angle the following formula can be used (Scarborough, 1966). Because this method is based on truncation of a Taylor series expansion, values computed in this way must be considered an estimate of the propagated error, not an exact value.

Let

$$Q = f(x_1, x_2, \dots, x_n) \quad (5.1)$$

where  $f$  is a function of directly measured quantities  $x_1, x_2, \dots, x_n$ . The estimated expected error in  $f$  can be expressed as

$$R = \sqrt{\left(\sigma_{x_1} \left| \frac{\partial Q}{\partial x_1} \right| \right)^2 + \left(\sigma_{x_2} \left| \frac{\partial Q}{\partial x_2} \right| \right)^2 + \cdots + \left(\sigma_{x_n} \left| \frac{\partial Q}{\partial x_n} \right| \right)^2} \quad (5.2)$$

where  $\sigma_{x_n}$  represents the probable error in  $x_n$ .

Using the initial estimate of atomic positional error of  $0.1\text{\AA}$  yields a probable error of  $\pm 12.7^\circ$  for the  $\psi$  dihedral angle. Using the more restrictive relative positional error of  $0.01\text{\AA}$  gives a value of  $\pm 1.27^\circ$ .

### 5.3 Propagated error using interval arithmetic

A more precise method of computing the propagated error in dihedral angle computation is to use interval arithmetic (Moore, 1975). Interval arithmetic is a simple technique for keeping track of the effects of starting a computation with approximate initial data. Because the error bounds are calculated at each step of the computation and propagated through all subsequent steps, the resulting error bounds are exact.

An *interval* of real numbers between  $a$  and  $b$  is represented by the symbol  $[a, b]$ . When performing arithmetic operations on such intervals the following rules are applied:

$$[a, b] + [c, d] = [a + c, b + d] \quad (5.3)$$

$$[a, b] - [c, d] = [a - d, b - c] \quad (5.4)$$

$$[a, b] \cdot [c, d] = [\min(ac, ad, bc, bd), \max(ac, ad, bc, bd)] \quad (5.5)$$

$$[a, b]/[c, d] = [a, b] \cdot [1/d, 1/c] \quad (5.6)$$

$$\text{(only defined if } 0 \notin [c, d]) \quad (5.7)$$

$$-[a, b] = [-b, -a] \quad (5.8)$$

$$c[a, b] = [ca, cb] \text{ (if } c > 0) \quad (5.9)$$

$$f([a, b]) = [\max f([a, b]), \min f([a, b])] \quad (5.10)$$

Using a positional uncertainty of  $\pm 0.01\text{\AA}$  results in an error interval of  $(87.2^\circ \dots 92.8^\circ)$  for an angle reading of  $90^\circ$ . This is an error of approximately  $\pm 2.8^\circ$ , a value some what larger than the error estimate given by the derivative method given presented in Section 5.2. Because the interval arithmetic method is more accurate

than the derivative method a value of  $2.8^\circ$  will be used as the deviation of dihedral angle measurement error in subsequent calculations.

## 5.4 Impact of propagated error on bin size

Using the value of  $2.8^\circ$  for the standard deviation of dihedral angle measurements and assuming a Gaussian distribution of the measurements, the probability of an angle being placed in the proper bin can be calculated using the following formula.

$$\frac{1}{w} \int_{-\frac{w}{2}}^{\frac{w}{2}} \int_{-\frac{w}{2}}^{\frac{w}{2}} g(x, m, \sigma) dx dm \quad (5.11)$$

Where  $w$  is the width of the bin and  $g(x, m, \sigma)$  is the Gaussian probability density function at point  $x$  with mean  $m$  and standard deviation  $\sigma$ .

$$g(x, m, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} \quad (5.12)$$

The probability that a dihedral angle value will be attributed to its correct bin, given an uncertainty with a standard deviation of  $2.8^\circ$ , is given in Table 5.2 for various bin widths.

As can be seen from inspecting the values in Table 5.2, a bin width of  $5^\circ$  will result in correct bin attribution only 57% of the time. These are not much better odds than tossing a coin and considering that the probability of a correct match then becomes  $p(a)p(b) = 0.3246$ , it becomes obvious that a bin width of  $5^\circ$  is too small for practical use. The odds improve to 78% for a bin width of  $10^\circ$  and to 85% for a bin width of  $15^\circ$ .



width (degrees)	probability of correct attribution
5	0.5698
6	0.6330
7	0.6824
8	0.7212
9	0.7519
10	0.7766
11	0.7969
12	0.8138
13	0.8281
14	0.8404
15	0.8511
20	0.8883

Table 5.2: Probability of correct bin membership.

# Chapter 6

## Classes of dihedral angle descriptors

The use of dihedral angles as descriptors of protein structure can be traced to the work of G. Ramachandran and V. Sasisekharen in the 1960's (Ramachandran and Sasisekharan, 1968). This work was based on observations of three dihedral angles that are described by the main-chain atoms of the peptide backbone,  $\phi$ ,  $\psi$ , and  $\omega$  (see Figure 2.1). Since then other dihedral angles have been employed as structural descriptors (Laiter et al., 1995; Levitt and Warshel, 1975; Oldfield and Hubbard, 1994; Warshel and Levitt, 1976). This chapter introduces a systematic way of classifying and describing dihedral angles to be used as protein structure descriptors. Furthermore, a family of angle classes, the pendant dihedral angles, will be described and arguments for their use in structural comparisons will be presented.

### 6.1 Main-chain dihedral angles

The traditional  $\phi$ ,  $\psi$ , and  $\omega$  dihedral angles are member of a class of descriptors that can be thought of as main-chain dihedral angles. They are defined by the positions of atoms that lie on the backbone, or main-chain, of a protein. The  $\phi$  angle is defined by the  $C_{i-1}$ ,  $N_i$ ,  $C_i^\alpha$ ,  $C_i$  atoms, the  $\psi$  angle is the dihedral angle  $N_i$ ,  $C_i^\alpha$ ,  $C_i$ ,  $N_{i+1}$ ; and the  $\omega$  angle is the dihedral angle  $C_i^\alpha$ ,  $C_i$ ,  $N_{i+1}$ ,  $C_{i+1}^\alpha$ , where  $i$  is the residue number (Fig. 2.1). In general these angles can be thought of as being defined by atoms  $j, j+1, j+2$ , and  $j+3$ , numbering the atoms sequentially along the protein backbone.

In the case of  $\phi$ ,  $\psi$ , and  $\omega$  the spacing between the atoms used in defining the angles is fixed at one. Other dihedral angles may be defined by using different spacing. One commonly used angle is defined by four successive  $C_\alpha$  atoms (Fig. 3.1). Designated the  $\tau$  angle by Oldfield and Hubbard (Oldfield and Hubbard, 1994) it has been used, in combination with other angles, as a descriptor of protein structure. It can be thought of as the angle defined by atoms  $j, j+3, j+6$ , and  $j+9$  with atom  $j$  always being a  $C_\alpha$ . There is no reason that the dihedral angles defined by every third backbone atom, where the first atom is not a  $C_\alpha$ , could not be used to measure structure. The use of  $C_\alpha$  atoms has a basis in the structure of the individual amino acids that comprise protein molecules in that the  $C_\alpha$  is the central atom of the three atoms an amino acid contributes to the protein backbone. Also, the side chain atoms, which give each amino acid its chemical identity, are connected to the main-chain by a covalent bond with the  $C_\alpha$ .

Other dihedral angles, defined using atoms from the protein backbone, can be stipulated. The general case uses atoms  $j, j+n, j+2n$ , and  $j+3n$  with  $n$  some fixed distance along the backbone. Observations of such descriptors show that the greater the value of  $n$  the more uniform the distribution of the angle values becomes. This is because of the larger number of included  $\phi$ ,  $\psi$ , and  $\omega$  angles, adding to the number of rotational degrees of freedom available to conform the measured segment.

## 6.2 Pendant dihedral angles

An alternative class of dihedral angle descriptors can be defined using atoms attached to the protein backbone as well as atoms along the backbone itself. Of particular interest are the carbonyl oxygen,  $O$ , and the first atom of the amino acid side chain, the beta carbon,  $C_\beta$ . I will refer to these atoms as pendant atoms (on the suggestion of Bruce Erickson) because they hang from the main-chain much like the pendant of a necklace (Fig. 6.1).

There are also pendant hydrogen atoms that are bonded to the nitrogen and carbon atoms of an amino acid but, since these hydrogen atoms are usually not present in the PDB files, the use of the pendant hydrogen makes little sense. There is an exception to the use of the pendant hydrogen, however.

In the case of the amino acid glycine the side chain consists only of a single hydrogen atom. Since these hydrogen atoms are most often not included in PDB

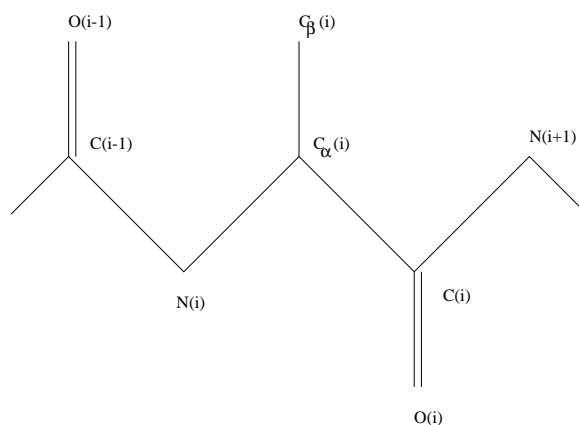


Figure 6.1: The protein backbone showing the  $O$  and  $C_{\beta}$  pendant atoms.

protein structure files an idealized position for the  $C_{\beta}$  must be computed for glycines. Details of how these computations are performed are detailed in Section 4.3.2.

Figure 6.2 graphically depicts the quadruplets of atoms that define the dihedral angle descriptors given in Table 6.1. Having defined these classes of dihedral angle descriptors, the next section will present frequency distributions and some statistical observations that will be used in the creation of the score tables in Chapter 7.

### 6.3 Statistics of descriptor angle distributions

The current PDB contains 4013 entries classified as proteins, peptides, and viruses (PDB, 1996). Many of these entries are variants of the same or similar structures. To get a more accurate statistical view of the frequency distributions for each of the dihedral angle descriptors I have elected to use the widely used 25% list of Hobohm and Sander (Hobohm et al., 1992; Hobohm and Sander, 1994), in which no two proteins share more than 25% sequence identity (for alignments of length 80 or more residues). Several of the structures listed in this dataset contain only  $C_{\alpha}$  positional data and cannot be used with the pendant dihedral angle descriptors. The files from the 25% list that have been included in compiling these statistics are listed in Appendix B.

It has long been known that the  $\phi$  and  $\psi$  angles fall mainly into distinct ranges (Ramachandran and Sasisekharan, 1968) so it is not surprising that other descriptor angles also possess preferred value ranges. Half of the descriptors (bb1, bb2, bo1,

name	dihedral atoms				backbone atoms spanned
	1	2	3	4	
bb1	$C_i^\beta$	$C_i^\alpha$	$C_{i+1}^\alpha$	$C_{i+1}^\beta$	4
bb2	$C_{i-1}^\beta$	$C_{i-1}^\alpha$	$C_{i+1}^\alpha$	$C_{i+1}^\beta$	7
bb3	$C_{i-1}^\beta$	$C_{i-1}^\alpha$	$C_{i+2}^\alpha$	$C_{i+2}^\beta$	10
bb4	$C_{i-2}^\beta$	$C_{i-2}^\alpha$	$C_{i+2}^\alpha$	$C_{i+2}^\beta$	13
bo1	$C_i^\beta$	$C_i^\alpha$	$C_i$	$O_i$	2
bo2	$C_{i-1}^\beta$	$C_{i-1}^\alpha$	$C_i$	$O_i$	5
bo3	$C_{i-1}^\beta$	$C_{i-1}^\alpha$	$C_{i+1}$	$O_{i+1}$	8
bo4	$C_{i-2}^\beta$	$C_{i-2}^\alpha$	$C_{i+1}$	$O_{i+1}$	11
ob1	$C_{i-1}$	$O_{i-1}$	$C_i^\alpha$	$C_i^\beta$	3
ob2	$C_{i-1}$	$O_{i-1}$	$C_{i+1}^\alpha$	$C_{i+1}^\beta$	6
ob3	$C_{i-2}$	$O_{i-2}$	$C_{i+1}^\alpha$	$C_{i+1}^\beta$	9
ob4	$C_{i-2}$	$O_{i-2}$	$C_{i+2}^\alpha$	$C_{i+2}^\beta$	12
oo1	$O_{i-1}$	$C_{i-1}$	$C_i$	$O_i$	4
oo2	$O_{i-1}$	$C_{i-1}$	$C_{i+1}$	$O_{i+1}$	7
oo3	$O_{i-2}$	$C_{i-2}$	$C_{i+1}$	$O_{i+1}$	10
oo4	$O_{i-2}$	$C_{i-2}$	$C_{i+2}$	$O_{i+2}$	13

Table 6.1: Pendant dihedral angle descriptors

bo3, ob2, ob3, oo1, and oo3) exhibit two distinct strong peaks that correspond to  $\alpha$ -helical and  $\beta$ -sheet regions. As an example of this strong separation between helical and sheet regions consider the bb1 frequency data presented in Figure 6.3.

It is evident from the plot that there are two distinct, well defined regions, one representing  $\alpha$ -helix centered around bin **Q** and one representing  $\beta$ -sheet centered around bin **Q**. There is practically no overlap between the two regions and both are narrowly defined, the  $\beta$ -sheet region less so than the  $\alpha$ -helix region. Compare this plot with the plot of bo2 frequencies shown in Figure 6.4.

In the bo2 plot it is obvious that the  $\alpha$ -helix and  $\beta$ -sheet regions overlap. Both regions are centered around the **G** sector bin. Because of this overlap it is expected that bo2 will be less able to differentiate  $\alpha$ -helix from  $\beta$ -sheet. When comparing structures that are dominated by one of these two secondary structure types, this lack of distinction between secondary structure types will probably not matter, but it could cause problems when comparing  $\alpha + \beta$  or  $\alpha/\beta$  structures.

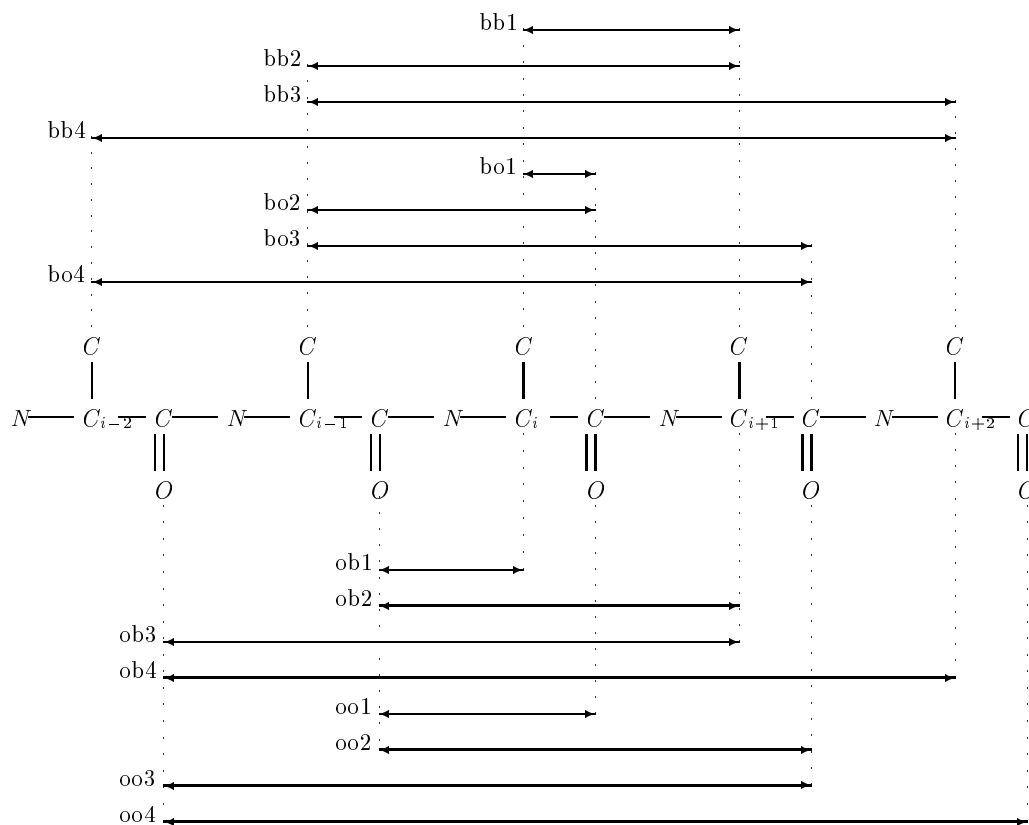


Figure 6.2: The atoms defining the sixteen pendant dihedral angles.

Plots of the frequency distributions for all sixteen descriptors are shown in Figures 6.5 through 6.8. In each distribution plot there appear three horizontal gray lines that represent, starting with the lowest line, the mean frequency, the mean plus one standard deviation, and the mean plus two standard deviations. A general observation can be made regarding all four descriptor families is that, as the residue span is increased from one to four, the frequency distributions tend to grow more uniform. The  $\alpha$ -helix and  $\beta$ -sheet peaks are still present in the descriptors that span four residues but their peaks are attenuated and their width increased. The effect of this increased uniformity will be discussed in Chapter 8.

bin	bb1	bb2	bb3	bb4
A	0.059809	0.023839	0.027445	0.030716
B	0.067068	0.020989	0.037153	0.033582
C	0.058147	0.019037	0.047459	0.035005
D	0.050971	0.019246	0.051699	0.035332
E	0.045768	0.020749	0.051783	0.034942
F	0.039132	0.023390	0.063086	0.033119
G	0.031634	0.024715	0.100838	0.029610
H	0.023242	0.025968	0.119435	0.029652
I	0.013387	0.027429	0.067862	0.030421
J	0.011382	0.027241	0.039630	0.031928
K	0.012701	0.029109	0.034561	0.033477
L	0.016918	0.036029	0.033585	0.038630
M	0.020864	0.045955	0.034697	0.049631
N	0.027054	0.054868	0.033952	0.097881
O	0.042777	0.057248	0.032608	0.123940
P	0.093831	0.051362	0.029827	0.065405
Q	0.178876	0.044222	0.028389	0.045658
R	0.071679	0.037428	0.029491	0.039831
S	0.029193	0.038618	0.027445	0.035015
T	0.019732	0.071484	0.023257	0.031454
U	0.015744	0.136634	0.021284	0.028588
V	0.014924	0.087944	0.021200	0.027945
W	0.019265	0.045037	0.020959	0.028124
X	0.035902	0.031458	0.022355	0.030116

Table 6.2: Bin frequencies for the bb descriptors.

bin	bo1	bo2	bo3	bo4
A	0.002921	0.017094	0.028275	0.025104
B	0.003354	0.014747	0.027972	0.025556
C	0.005748	0.013677	0.029194	0.025283
D	0.014757	0.014996	0.033489	0.023562
E	0.084323	0.025101	0.037460	0.025724
F	0.178646	0.114020	0.042862	0.031832
G	0.093652	0.223980	0.054450	0.052266
H	0.058338	0.117323	0.079632	0.108572
I	0.038844	0.076394	0.091220	0.134947
J	0.023137	0.053557	0.068650	0.080393
K	0.015521	0.041988	0.041932	0.055792
L	0.014128	0.034178	0.033876	0.043796
M	0.013251	0.023803	0.031023	0.034498
N	0.013044	0.016565	0.028411	0.032084
O	0.018658	0.016170	0.026446	0.030258
P	0.044355	0.016201	0.024169	0.030887
Q	0.100092	0.019296	0.025402	0.031255
R	0.114715	0.023990	0.029278	0.032902
S	0.086790	0.025963	0.042047	0.034466
T	0.043900	0.026420	0.052715	0.033385
U	0.015893	0.024509	0.053854	0.028599
V	0.006956	0.022495	0.045923	0.027917
W	0.004933	0.019182	0.038828	0.025524
X	0.004045	0.018351	0.032893	0.025398

Table 6.3: Bin frequencies for the bo descriptors.



bin	ob1	ob2	ob3	ob4
A	0.087349	0.027376	0.033551	0.036181
B	0.017326	0.051116	0.033708	0.031051
C	0.005800	0.177236	0.033645	0.026006
D	0.003544	0.126537	0.034697	0.023779
E	0.003368	0.035777	0.057775	0.023779
F	0.003648	0.027157	0.163358	0.024444
G	0.006434	0.037219	0.095405	0.024887
H	0.014447	0.054272	0.039638	0.030112
I	0.013127	0.064220	0.032341	0.035473
J	0.007442	0.058504	0.028104	0.040297
K	0.006652	0.051576	0.024698	0.043073
L	0.006236	0.043802	0.023730	0.043083
M	0.005789	0.037658	0.022448	0.040402
N	0.006091	0.031671	0.023615	0.037257
O	0.006797	0.026446	0.023741	0.035927
P	0.009853	0.021535	0.023026	0.039136
Q	0.029134	0.016196	0.025055	0.045067
R	0.057384	0.015977	0.026317	0.056646
S	0.089407	0.015632	0.032226	0.069564
T	0.094032	0.014033	0.041972	0.072435
U	0.088326	0.014096	0.051719	0.068013
V	0.094042	0.014963	0.049795	0.061057
W	0.141479	0.016969	0.042456	0.050798
X	0.202293	0.020031	0.036978	0.041532

Table 6.4: Bin frequencies for the ob descriptors.

bin	oo1	oo2	oo3	oo4
A	0.082381	0.018422	0.032468	0.028708
B	0.086923	0.023312	0.041573	0.031938
C	0.062497	0.026436	0.049006	0.036001
D	0.055419	0.027324	0.053054	0.038080
E	0.058818	0.030490	0.049448	0.039801
F	0.027450	0.035067	0.043676	0.034946
G	0.007671	0.035715	0.038135	0.030967
H	0.005488	0.028076	0.041026	0.027135
I	0.008117	0.024430	0.052276	0.026787
J	0.016827	0.025506	0.064231	0.025911
K	0.021421	0.031462	0.071727	0.029584
L	0.022222	0.046414	0.068321	0.040719
M	0.134224	0.105629	0.057817	0.059738
N	0.203748	0.197695	0.049490	0.077269
O	0.075209	0.098576	0.041047	0.079243
P	0.023074	0.054606	0.036042	0.070155
Q	0.010217	0.040155	0.032037	0.061532
R	0.006974	0.033698	0.028819	0.054419
S	0.007837	0.028265	0.025455	0.048075
T	0.007691	0.023521	0.025108	0.039072
U	0.007442	0.020344	0.024119	0.034661
V	0.009001	0.016927	0.023257	0.030703
W	0.016682	0.013500	0.025055	0.027410
X	0.042666	0.014430	0.026811	0.027146

Table 6.5: Bin frequencies for the oo descriptors.

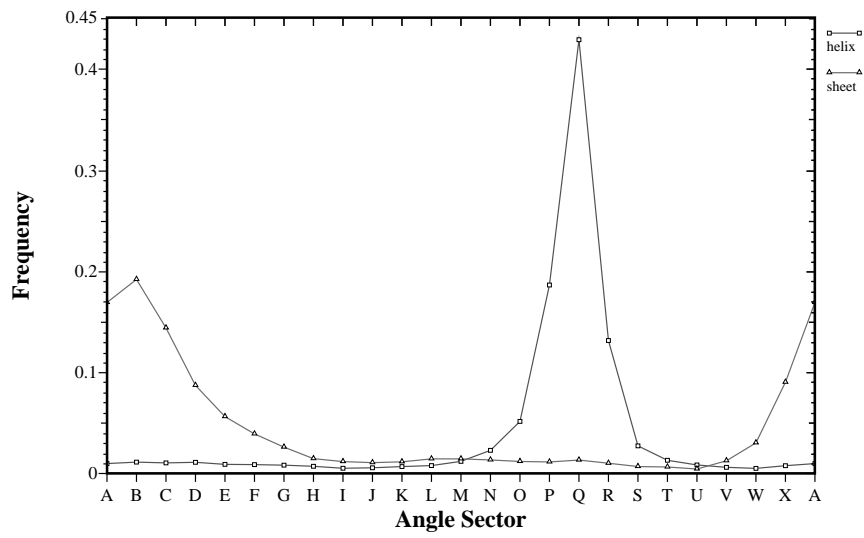


Figure 6.3: Bin frequency for bb1  $\alpha$ -helix and  $\beta$ -sheet regions.

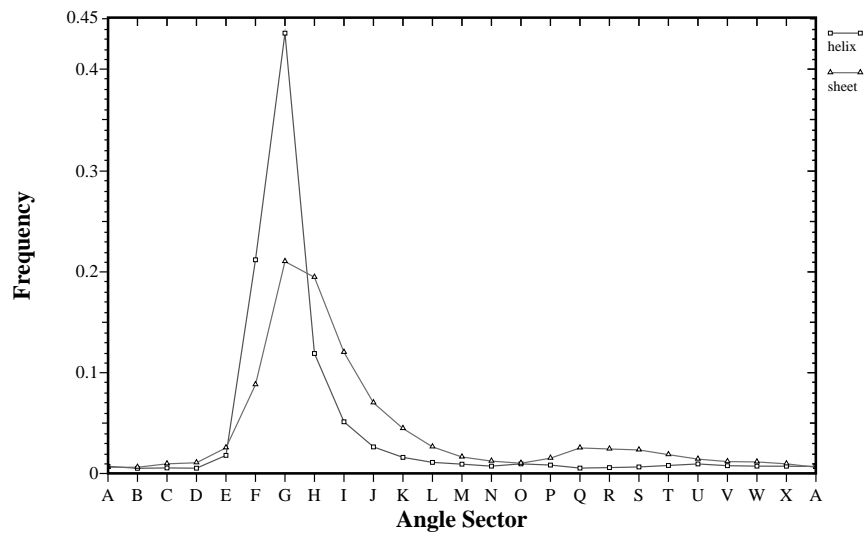
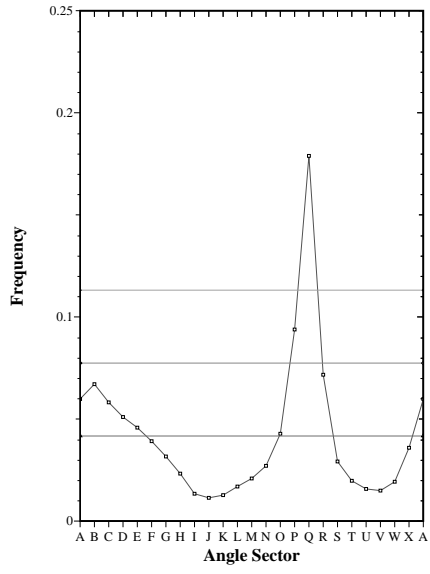
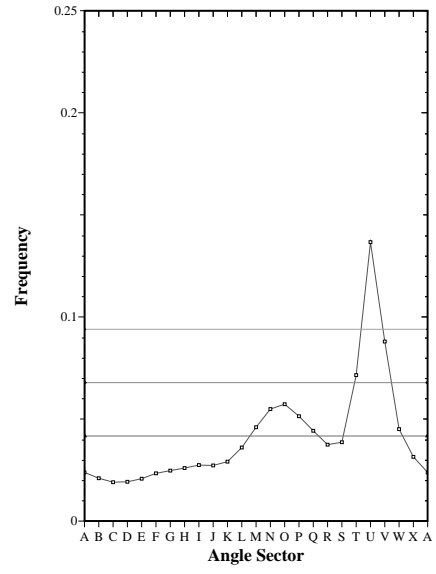


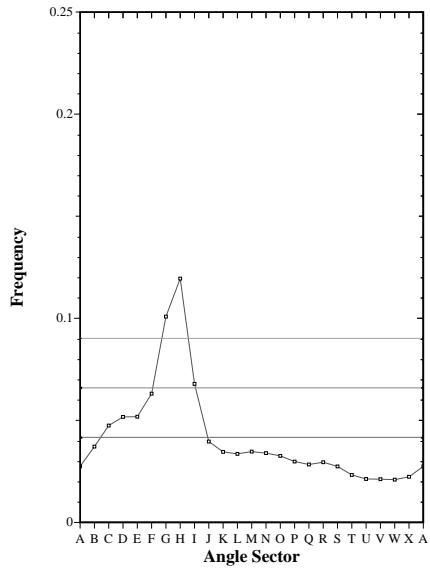
Figure 6.4: Bin frequency for bo2  $\alpha$ -helix and  $\beta$ -sheet regions.



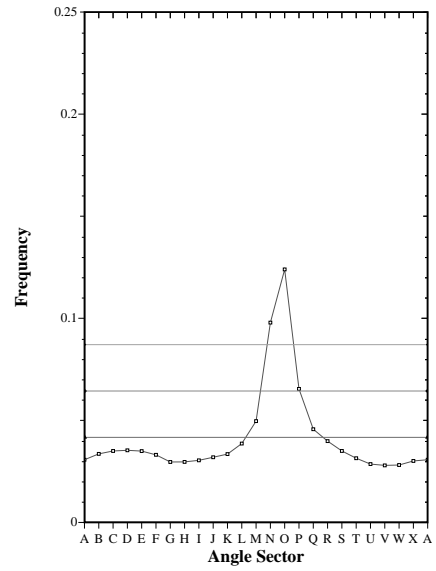
(a) bb1



(b) bb2

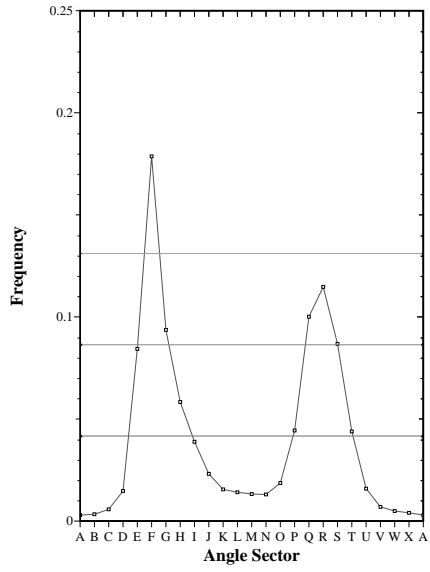


(c) bb3

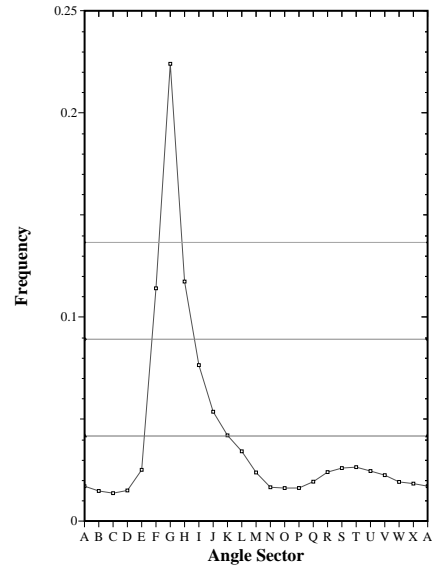


(d) bb4

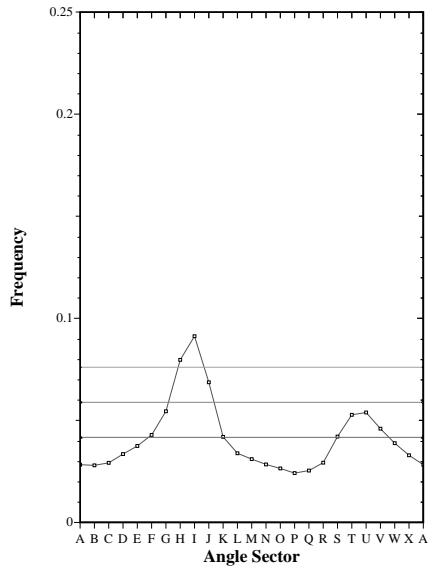
Figure 6.5: Frequency plots for the bb descriptors.



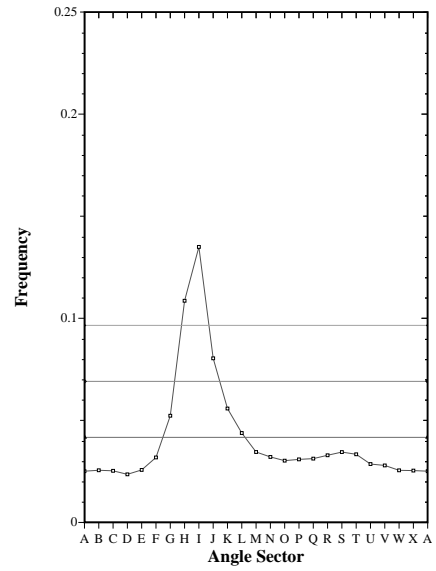
(a) bo1



(b) bo2

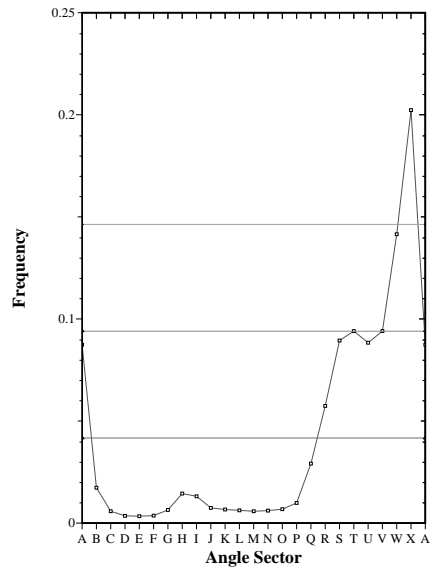


(c) bo3

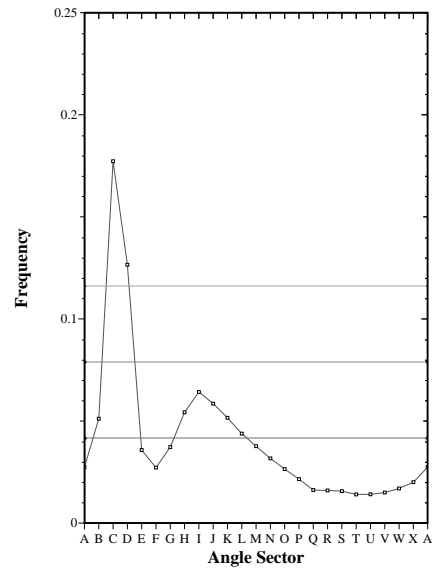


(d) bo4

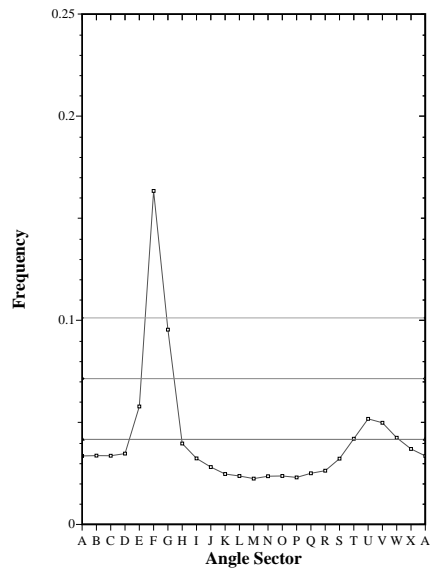
Figure 6.6: Frequency plots for the bo descriptors.



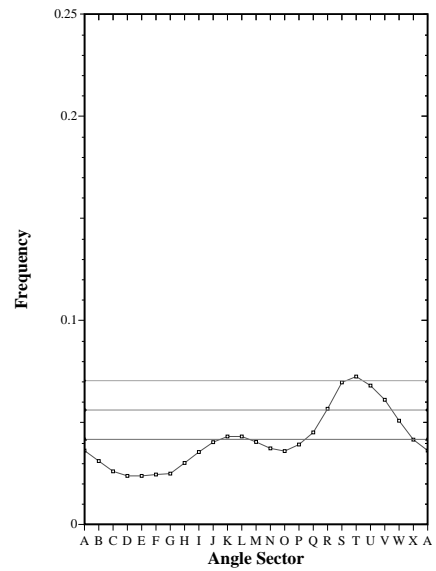
(a) ob1



(b) ob2

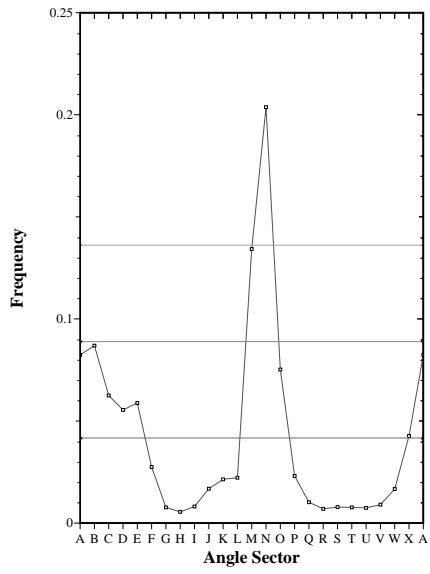


(c) ob3

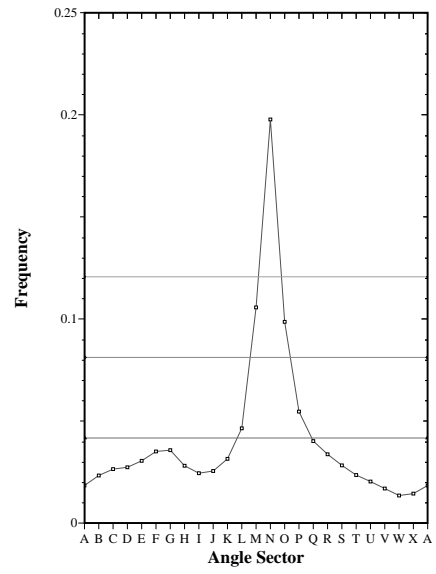


(d) ob4

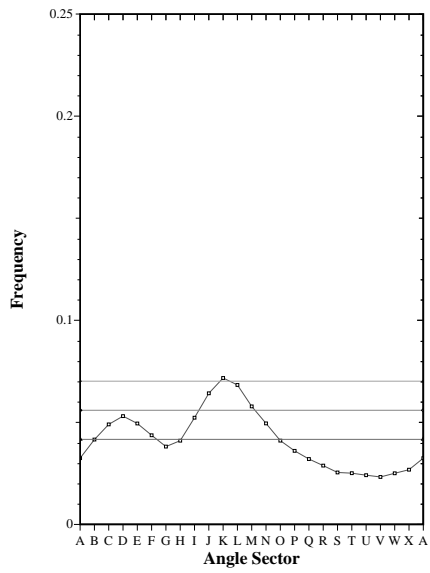
Figure 6.7: Frequency plots for the ob descriptors.



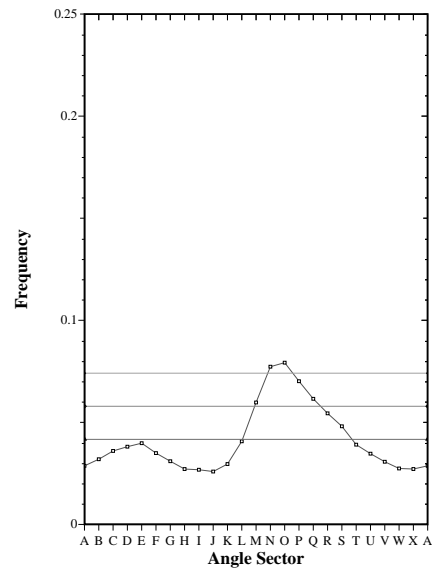
(a) oo1



(b) oo2



(c) oo3



(d) oo4

Figure 6.8: Frequency plots for the oo descriptors.



# Chapter 7

## Construction of the score table

It is well known that the sensitivity and effectiveness of sequence alignment algorithms is determined by the quality of the score table (Altschul et al., 1994). The construction of a score table for a particular dihedral descriptor starts with a two-dimensional 24-by-24 array with the entries on the main diagonal being 100 and all off diagonal entries being 0. The diagonal entries represent the score for matching an angle sector with a like angle sector. In a sense this score table represents very strict, identical sequence matching. To compensate for the variation in the frequency distributions of the various dihedral descriptors the score tables entries will be scaled in two ways: once to compensate for the variation in the frequency distributions and once to diffuse or blur the matching criteria to allow for inexact matches.

### 7.1 Relative information content scaling

To compensate for the relative frequencies of occurrence and the range of angle sector frequencies, the scaling factor,  $S_i$ , is applied. Note that  $h(f_x) = -\log_2(f_x)$  is the information content of the symbol  $x$  with frequency of occurrence  $f_x$  (Shannon, 1948). First we define  $h_{high}$ ,  $h_{low}$ ,  $h_i$ , and  $\bar{h}$  as follows:

$$\begin{aligned}
h_{high} &= -\log_2(f_{high}) \\
h_{low} &= -\log_2(f_{low}) \\
h_i &= -\log_2(f_i) \\
\bar{h} &= -\log_2(\bar{f})
\end{aligned} \tag{7.1}$$

where  $f_{high}$ ,  $f_{low}$ , and  $f_i$  are the frequencies observed for a given descriptor for its highest bin, lowest bin, and the  $i$ th bin, respectively, and  $\bar{f}$  is the mean frequency of the set of 24 bins. The scale factor is given as:

$$S_i = (1 - k) + kR_i \tag{7.2}$$

where

$$R_i = \frac{h_{high} - h_i}{h_{high} - h_{low}} \tag{7.3}$$

and

$$k = 1 - \frac{h_{high}}{\bar{h}} \tag{7.4}$$

Expanding equation 7.2 by substituting for  $k$

$$\begin{aligned}
S_i &= 1 - \left(1 - \frac{h_{high}}{\bar{h}}\right)(1 - R_i) \\
&= R_i + \frac{h_{high}}{\bar{h}}(1 - R_i) \\
&= \frac{h_{high}}{\bar{h}} + \left(1 - \frac{h_{high}}{\bar{h}}\right)R_i
\end{aligned} \tag{7.5}$$

For example, for the oo1 descriptor (Table 7.15) we have the following frequency values.

$$f_{high} = f_N = 0.2037, f_{low} = f_H = 0.0055, \bar{f} = 0.0417 \tag{7.6}$$

Therefore

$$h_{high} = 2.2955, h_{low} = 7.5064, \bar{h} = 4.5838 \quad (7.7)$$

$$k = 1 - \frac{2.2955}{4.5838} = 0.4992 \quad (7.8)$$

$$R_i = \frac{2.2955 - h_i}{2.2955 - 7.5064} \quad (7.9)$$

$$S_i = 0.5008 - 0.4992 \left( \frac{2.2955 - h_i}{5.2109} \right) \quad (7.10)$$

$$S_N = 0.5008, S_H = 1.0 \quad (7.11)$$

## 7.2 Statistical diffusion of score values

In order to compensate for binification error and to add some tolerance to the sequence alignments for inexact matches, the main diagonal score values are statistically diffused among the  $(i, i + 2)$  through  $(i + 2, i)$  diagonal entries. The calculation of the diffusion values is depicted in Table 7.1.

$b^2$	$ab$	$b^2$
$ab$	$a^2$	$ab$
$b^2$	$ab$	$b^2$

Table 7.1: Diffusion coefficients based on probability of correct attribution ( $a$ ).

The probability,  $a$ , of an individual angle sector occurring is taken to be the correct attribution probability as calculated in Section 5.4. This give a value of  $a \cong 0.86$ . In diffusing the score value of a table entry we do not wish to increase or decrease the sum of the positive scores. This implies that

$$a^2 + 4ab + 4b^2 = 1 \quad (7.12)$$

or

$$(a + 2b)^2 = 1 \quad (7.13)$$

Solving for  $b$  yields a value of 0.07. Using these values for  $a$  and  $b$  Table 7.1 can be rewritten by replacing the symbolic values with calculated values (Table 7.2). One diffusion pass was performed on entry  $T_{i,j}$  of a score table in the following way:

0.01	0.06	0.01
0.06	0.72	0.06
0.01	0.06	0.01

Table 7.2: Numeric diffusion coefficients give probability  $a = 0.86$

$$T'_{i,j} = 0.72 \times T_{i,j} \quad (7.14)$$

$$T'_{i-1,j-1} = T_{i-1,j-1} + 0.01 \times T_{i,j} \quad (7.15)$$

$$T'_{i-1,j} = T_{i-1,j-1} + 0.06 \times T_{i,j} \quad (7.16)$$

$$T'_{i-1,j+1} = T_{i-1,j+1} + 0.01 \times T_{i,j} \quad (7.17)$$

$$T'_{i,j-1} = T_{i,j-1} + 0.06 \times T_{i,j} \quad (7.18)$$

$$T'_{i,j+1} = T_{i,j+1} + 0.06 \times T_{i,j} \quad (7.19)$$

$$T'_{i+1,j-1} = T_{i+1,j-1} + 0.01 \times T_{i,j} \quad (7.20)$$

$$T'_{i+1,j} = T_{i+1,j-1} + 0.06 \times T_{i,j} \quad (7.21)$$

$$T'_{i+1,j+1} = T_{i+1,j+1} + 0.01 \times T_{i,j} \quad (7.22)$$

In practice a table form with only a single diffusion pass is very strict in terms of structure match, aligning structure segments that are almost identical and rejecting all others. This type of table might have application under some conditions. A second diffusion pass was performed on the resulting table to give the final table. The resulting positive scores find not only nearly identical dihedral alignments but also lower scoring alignments.

Applying relative information content scaling and two statistical diffusions to the frequency data for the bb1 dihedral angle descriptor generates the bb1 score table shown in Table 7.3. The values given have all been converted to integers using an initial scale of 0 to 100. The mismatch score was set to  $-30$ . Tables 7.4 through 7.18 contain the corresponding score tables for the remaining fifteen descriptors.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	35	10	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	12
b	10	34	10	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1
c	1	10	36	11	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	1	11	38	11	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	1	11	39	12	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	1	12	41	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	1	13	44	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	1	14	49	16	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	1	16	56	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	2	17	58	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	2	17	57	17	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	53	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	50	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	46	13	1	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	40	10	1	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	10	29	7	1	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	7	21	8	1	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	8	33	12	1	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	12	45	14	1	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	51	16	2	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	54	16	2	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	16	55	16	1
w	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	16	51	14
x	12	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	41

Table 7.3: Table of alignment scores based on the bb1 descriptor.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	55	17	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16
b	17	57	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2
c	1	17	59	18	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	2	18	58	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	2	17	57	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	2	17	55	16	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	2	16	54	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	2	16	53	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	1	16	52	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	1	16	52	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	1	16	51	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	47	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	43	12	1	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	12	40	12	1	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	12	39	12	1	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	12	41	13	1	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	44	14	1	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	47	14	1	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	46	12	1	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	12	35	9	1	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	9	25	8	1	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	8	32	11	1
w	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	11	43	14
x	16	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	49

Table 7.4: Table of alignment scores based on the bb2 descriptor.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	53	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17
b	15	48	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1
c	1	14	43	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	1	13	42	12	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	1	12	41	12	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	1	12	38	10	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	1	10	29	8	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	1	8	26	9	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	1	9	36	12	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	1	12	46	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	1	14	49	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	50	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	49	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	50	15	1	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	50	15	1	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	52	16	1	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	53	16	1	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	52	16	2	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	54	17	2	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	57	17	2	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	58	18	2	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	18	58	18	2
w	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	18	59	17
x	17	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	56

Table 7.5: Table of alignment scores based on the bb3 descriptor.



	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	56	17	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17
b	17	55	16	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2
c	1	16	54	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	2	16	53	16	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	1	16	54	16	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	2	16	55	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	2	17	57	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	2	17	57	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	2	17	57	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	2	17	56	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	2	17	55	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	16	51	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	46	11	1	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	11	31	8	1	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	8	26	10	1	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	10	40	13	1	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	48	15	1	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	51	16	1	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	54	16	2	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	56	17	2	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	58	18	2	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	18	59	18	2
w	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	18	58	17
x	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	56

Table 7.6: Table of alignment scores based on the bb4 descriptor.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	58	17	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17
b	17	57	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2
c	1	16	52	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	1	14	42	10	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	1	10	26	6	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	1	6	18	6	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	1	6	24	8	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	1	8	29	9	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	1	9	33	11	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	1	11	38	12	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	1	12	42	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	43	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	44	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	44	13	1	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	40	11	1	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	11	32	8	1	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	8	24	7	1	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	7	22	7	1	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	7	25	8	1	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	8	32	11	1	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	11	42	14	1	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	50	16	1
w	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	53	16
x	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	54

Table 7.7: Table of alignment scores based on the bo1 descriptor.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	55	17	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	16
b	17	57	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2
c	1	17	59	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	2	17	57	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	2	16	49	11	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	1	11	27	6	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	1	6	17	6	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	1	6	26	9	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	1	9	33	11	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	1	11	38	12	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	1	12	42	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	45	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	50	16	2	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	56	17	2	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	56	17	2	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	56	16	1	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	16	53	16	1	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	50	15	1	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	49	15	1	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	49	15	1	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	50	15	1	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	51	16	1
w	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	53	16
x	16	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	53

Table 7.8: Table of alignment scores based on the bo2 descriptor.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	55	17	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16
b	17	55	16	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2
c	1	16	54	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	2	16	51	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	1	15	49	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	1	14	46	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	1	13	41	11	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	1	11	33	9	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	1	9	30	10	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	1	10	36	12	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	1	12	46	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	51	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	53	16	2	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	55	17	2	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	57	17	2	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	58	17	2	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	57	17	2	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	54	15	1	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	15	47	13	1	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	42	12	1	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	12	41	13	1	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	45	14	1
w	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	48	15
x	16	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	51

Table 7.9: Table of alignment scores based on the bo3 descriptor.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	57	17	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17
b	17	57	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2
c	1	17	57	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	2	17	58	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	2	17	57	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	2	16	53	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	1	14	43	11	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	1	11	29	8	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	1	8	25	9	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	1	9	34	11	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	1	11	42	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	46	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	51	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	52	16	1	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	54	16	1	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	53	16	1	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	53	16	1	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	52	16	1	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	51	15	1	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	52	16	2	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	55	17	2	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	55	17	2
w	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	57	17
x	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	56

Table 7.10: Table of alignment scores based on the bo4 descriptor.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	25	10	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	6
b	10	41	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
c	1	14	53	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	1	17	58	18	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	2	18	59	18	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	2	18	58	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	2	16	52	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	1	14	44	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	1	13	44	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	1	14	50	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	1	15	51	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	52	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	53	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	52	16	1	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	51	15	1	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	47	12	1	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	12	36	10	1	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	10	29	8	1	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	8	24	7	1	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	7	24	7	1	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	7	24	7	1	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	7	24	6	1
w	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	6	20	5
x	6	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	5	16

Table 7.11: Table of alignment scores based on the ob1 descriptor.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	48	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15
b	13	39	9	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1
c	1	9	21	7	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	1	7	26	10	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	1	10	44	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	1	14	48	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	1	14	44	12	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	1	12	38	11	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	1	11	36	11	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	1	11	37	11	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	1	11	39	12	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	12	41	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	44	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	46	14	1	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	49	15	1	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	52	16	2	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	56	17	2	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	57	17	2	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	57	17	2	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	59	18	2	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	18	58	18	2	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	18	58	17	2
w	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	56	16
x	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	16	52

Table 7.12: Table of alignment scores based on the ob2 descriptor.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	51	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15
b	15	51	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1
c	1	15	51	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	1	15	50	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	1	14	41	9	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	1	9	23	8	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	1	8	32	12	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	1	12	48	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	1	15	52	16	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	1	16	54	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	2	17	57	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	58	18	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	18	59	18	2	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	18	58	17	2	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	58	17	2	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	58	17	2	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	57	17	2	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	56	16	1	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	16	52	15	1	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	47	14	1	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	43	13	1	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	44	14	1
w	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	47	14
x	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	48

Table 7.13: Table of alignment scores based on the ob3 descriptor.



	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	49	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14
b	15	52	16	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1
c	1	16	56	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	2	17	59	18	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	2	18	59	18	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	2	18	58	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	2	17	57	17	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	2	17	53	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	1	15	49	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	1	14	46	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	1	14	45	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	45	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	46	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	48	15	1	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	49	14	1	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	47	14	1	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	44	12	1	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	12	38	11	1	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	11	34	10	1	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	10	33	10	1	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	10	34	11	1	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	11	37	12	1
w	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	12	41	13
x	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	45

Table 7.14: Table of alignment scores based on the ob4 descriptor.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	28	8	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	9
b	8	27	8	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1
c	1	8	30	9	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	1	9	32	9	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	1	9	31	11	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	1	11	40	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	1	14	54	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	1	17	58	17	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	2	17	54	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	1	15	46	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	1	13	43	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	42	10	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	10	22	6	1	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	6	17	7	1	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	7	28	10	1	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	10	42	14	1	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	51	16	2	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	56	17	2	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	54	16	2	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	16	55	17	2	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	55	16	1	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	16	53	15	1
w	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	46	12
x	9	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	12	34

Table 7.15: Table of alignment scores based on the oo1 descriptor.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	54	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17
b	16	51	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2
c	1	15	49	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	1	15	48	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	1	14	47	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	1	14	44	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	1	13	44	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	1	14	48	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	1	15	50	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	1	15	49	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	1	14	46	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	40	10	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	10	28	7	1	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	7	19	7	1	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	7	29	10	1	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	10	38	12	1	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	12	42	13	1	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	45	14	1	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	48	15	1	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	50	15	1	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	53	16	2	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	55	17	2
w	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	58	17
x	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	56

Table 7.16: Table of alignment scores based on the oo2 descriptor.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	51	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	16
b	14	45	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1
c	1	13	42	12	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	1	12	40	12	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	1	12	41	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	1	13	44	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	1	14	47	14	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	1	14	45	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	1	13	40	11	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	1	11	35	10	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	1	10	33	10	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	10	34	11	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	11	38	12	1	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	12	41	13	1	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	46	14	1	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	49	15	1	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	51	16	2	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	54	17	2	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	56	17	2	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	57	17	2	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	58	18	2	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	18	59	17	2
w	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	57	17
x	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	54

Table 7.17: Table of alignment scores based on the oo3 descriptor.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	56	17	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17
b	17	53	16	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2
c	1	16	51	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
d	-30	1	15	49	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
e	-30	-30	1	15	48	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
f	-30	-30	-30	1	15	51	16	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
g	-30	-30	-30	-30	1	16	54	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
h	-30	-30	-30	-30	-30	2	17	57	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
i	-30	-30	-30	-30	-30	-30	2	17	58	18	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
j	-30	-30	-30	-30	-30	-30	-30	2	18	58	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
k	-30	-30	-30	-30	-30	-30	-30	-30	2	17	55	15	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
l	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	15	48	13	1	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30
m	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	38	10	1	-30	-30	-30	-30	-30	-30	-30	-30	-30
n	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	10	32	9	1	-30	-30	-30	-30	-30	-30	-30	-30
o	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	9	32	10	1	-30	-30	-30	-30	-30	-30	-30
p	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	10	34	11	1	-30	-30	-30	-30	-30	-30
q	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	11	38	12	1	-30	-30	-30	-30	-30
r	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	12	41	13	1	-30	-30	-30	-30
s	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	13	44	14	1	-30	-30	-30
t	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	14	49	15	1	-30	-30
u	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	15	52	16	2	-30
v	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	1	16	54	17	2
w	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	57	17
x	17	2	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	-30	2	17	56

Table 7.18: Table of alignment scores based on the oo4 descriptor.

## 7.3 Impact of the mismatch score

After applying relative information content scaling and two statistical diffusions the only unspecified score table values are those representing a mismatch. The negative values given to these table entries represent a penalty for misalignment. The value of a mismatch can be varied, providing control over sensitivity to long or low scoring alignments. A small negative mismatch score finds alignments with greater RMSD. A small negative value finds much longer dihedral alignments with more mismatches. Under some conditions this may be what is desired from a database search, finding alignments that are remotely similar. Tables with mismatch penalty values of  $-10$  and  $-20$  have these characteristics.

The BioSCAN algorithm depends on the negative mismatches scores to terminate alignment sequences so it is important not to set it too low. Too small a negative value will result in long meaningless alignments. At the extreme, use of a mismatch score of 0 permits a sufficiently long alignment to always exceed the detection threshold and produce meaningless output. Randomly generated sequences should tend to accumulate a slightly negative value to avoid finding meaningless alignments. In practice a mismatch score of  $-10$  is close to the smallest usable value, which is determined by the frequency distribution of the descriptor being used.

Using a mismatch score of  $-40$  or  $-50$ , finds tighter alignments that tend to be shorter in length with lower RMSD between the segments. Experience has shown that a mismatch score of  $-30$  is a good general purpose compromise.

# Chapter 8

## Experimental results

Existing dogma in the field of protein science says that sequence similarity implies structural similarity and that structural similarity implies functional similarity. Sequence comparison methods can provide an indication of sequence similarity. Possible outcomes combining amino acid sequence similarity, structural similarity, and functional similarity are tabulated in Table 8.1.

type	sequence similarity	structural similarity	functional similarity	Likelihood of occurrence
1	similar	similar	similar	expected
2	similar	similar	dissimilar	scarce
3	similar	dissimilar	similar	very scarce
4	similar	dissimilar	dissimilar	scarce
5	dissimilar	similar	similar	scarce
6	dissimilar	similar	dissimilar	very scarce
7	dissimilar	dissimilar	similar	scarce
8	dissimilar	dissimilar	dissimilar	expected

Table 8.1: Possible experimental outcomes.

The dogmatic, expected outcome (type 1) is not a very exciting result. Outcomes where the protein structures are dissimilar (types 3, 4, 6, and 7) are not found by the dihedral alignment method since proving a negative result is much harder than proving a positive one. The most exciting results are type 3, where expected structural similarity is not found, and type 6, where unexpected structural similarity is found.

## 8.1 Comparison of protein structure

The sensitivity of the score tables developed in Chapter 7 were tested by verifying that this method can correctly register the structural dissimilarity between proteins in Jones' list (Jones et al., 1992). This list contains proteins which do not have an apparent structural similarity. Our pairwise comparisons of all proteins in this list showed that each of these unique proteins aligned strongly only with itself. Performing these more than 10,000 comparisons required a total of 7.5 minutes for each of the two descriptors on the BioSCAN system.

We also evaluated the ability of this method in determining structural similarity relationships between protein structures in the collection of 167 representative similar protein structures given by Orengo et al. (Orengo et al., 1992). This paper clusters proteins into fold families and ranks the proteins within each family based on a complex 3D structural similarity score. The dihedral alignment method found the same proteins as similar.

## 8.2 Dihedral sequence alignment vs. 3D structure alignment

The dihedral alignment method proposed here identifies similar segments of two proteins. The ultimate goal of this research is to identify true 3D protein homology. The latter can be judged by the RMSD between the  $C_\alpha$  coordinates of corresponding segments of two 3D protein structures. The following three sections present comparisons between proteins with low amino acid sequence homology. These examples were specifically chosen to show alignments between proteins with secondary structure types of  $\alpha$ ,  $\beta$ , and  $\alpha/\beta$ .

### 8.2.1 Alignment of 1mbd and 1bab

The first alignment is between sperm-whale myoglobin (1mbd) (Phillips, 1980) and the beta chain of human hemoglobin Thionville (1bab) (Vasseur et al., 1992). Tables 8.2 through 8.5 summarizes the dihedral sequence alignments for all sixteen descriptors. The column labeled *desc* gives the descriptor type, *mismatch* the mismatch score, *score* the sum score for the alignment, *len* the length of the alignment, *avg* the average score per residue, *RMSD* the alignments Root Mean Square Devia-



tion in Angstroms, *start* the starting residue position of the alignment in the 1mbd protein, and *end* the ending residue position of the alignment in the 1mbd protein,

Inspection of the results show that increasing the mismatch penalty score for all descriptors produces shorter aligned sequences with smaller RMSD values. In many cases the same alignment is found using more than one mismatch score value. In these cases the larger negative value produces a lower total score but the RMSD remains the same.

desc	mismatch	score	len	avg	RMSD	start	end
bb1	-10	1281	137	9.35	2.28	6	142
	-20	982	117	8.39	1.57	26	142
	-30	742	117	6.34	1.57	26	142
	-40	528	95	5.56	1.56	26	120
	-50	298	17	17.53	0.63	26	42
		444	34	13.06	0.79	87	120
bb2	-10	1147	125	9.18	1.60	20	144
	-20	817	125	6.45	1.60	20	144
	-30	524	98	5.35	1.55	20	117
	-40	245	24	10.21	0.79	50	73
		451	35	12.89	1.01	83	117
	-50	439	31	14.16	0.71	87	117
bb3	-10	1123	138	8.14	2.25	7	144
	-20	811	118	6.87	1.59	27	144
	-30	554	111	4.99	1.52	27	132
	-40	385	36	10.69	1.06	82	117
	-50	355	36	9.86	1.06	82	117
bb4	-10	1239	137	9.04	2.24	7	143
	-20	889	137	6.49	2.24	7	143
	-30	579	117	5.10	1.58	27	143
	-40						
	-50						

Table 8.2: Table of bb descriptor alignment scores for 1mbd vs. 1bab.

Figures 8.1 through 8.7 show the dihedral alignments found using 16 dihedral descriptors with five mismatch scores. Also shown is an alignment of the corresponding amino acid sequences. Identical amino acid residues are echoed in the middle line. The distal His64 and proximal His93 of myoglobin are near the heme iron atom. Double-headed arrows mark the positions of the eight alpha helices (A-H). The dipep-

desc	mismatch	score	len	avg	RMSD	start	end
bo1	-10	1387	139	9.98	2.29	6	144
	-20	1207	139	8.68	2.29	6	144
	-30	1027	139	7.39	2.29	6	144
	-40	902	92	9.80	1.56	26	117
	-50	832	92	9.04	1.56	26	117
bo2	-10	1258	137	9.18	2.28	6	142
	-20	1008	137	7.36	2.28	6	142
	-30	793	117	6.78	1.57	26	142
	-40	593	117	5.07	1.57	26	142
	-50	279	14	19.93	0.29	26	39
		384	27	14.22	0.79	50	76
		471	31	15.19	0.71	89	117
bo3	-10	1054	138	7.64	2.31	4	141
	-20	711	117	6.08	1.55	25	141
	-30	319	23	13.87	0.65	51	73
		523	50	10.46	0.97	88	137
	-40	299	23	13.00	0.65	51	73
		463	50	9.26	0.97	88	137
	-50	403	50	8.06	0.97	88	137
bo4	-10	1260	137	9.20	2.24	7	143
	-20	880	137	6.42	2.24	7	143
	-30	339	23	14.74	0.65	51	73
		540	45	12.00	0.98	97	141
	-40	329	23	14.30	0.65	51	73
		480	45	10.67	0.98	97	141
	-50	319	23	13.87	0.65	51	73
		420	45	9.33	0.98	97	141

Table 8.3: Table of bo descriptor alignment scores for 1mbd vs. 1bab.

desc	mismatch	score	len	avg	RMSD	start	end
ob1	-10	898	140	6.41	2.30	5	144
	-20	578	140	4.13	2.30	5	144
	-30	311	92	3.38	1.56	26	117
	-40	261	21	12.43	4.30	51	71
		255	33	7.73	0.76	85	117
	-50	261	21	12.43	4.30	51	71
		225	23	9.78	0.78	85	107
ob2	-10	1717	140	12.26	2.29	6	145
	-20	1497	140	10.69	2.29	6	145
	-30	1277	140	9.12	2.29	6	145
	-40	1057	140	7.55	2.29	6	145
	-50	859	121	7.10	1.59	25	145
ob3	-10	1661	139	11.95	2.24	7	145
	-20	1341	139	9.65	2.24	7	145
	-30	1021	139	7.35	2.24	7	145
	-40	701	139	5.04	2.24	7	145
	-50	346	15	23.07	0.32	25	39
		608	35	17.37	0.87	84	118
ob4	-10	1098	129	8.51	1.97	13	141
	-20	710	118	6.02	1.55	24	141
	-30	413	22	18.77	0.71	94	115
	-40	393	22	17.86	0.71	94	115
	-50	373	22	16.95	0.71	94	115

Table 8.4: Table of ob descriptor alignment scores for 1mbd vs. 1bab.

desc	mismatch	score	len	avg	RMSD	start	end
oo1	-10	1355	139	9.75	2.29	6	144
	-20	1195	139	8.60	2.29	6	144
	-30	1035	139	7.45	2.29	6	144
	-40	875	139	6.29	2.29	6	144
	-50	715	139	5.14	2.29	6	144
oo2	-10	1295	136	9.52	2.27	6	141
	-20	1035	136	7.61	2.27	6	141
	-30	775	136	5.70	2.27	6	141
	-40	620	92	6.74	1.38	50	141
	-50	535	67	7.99	1.33	50	116
oo3	-10	994	120	8.28	1.58	23	142
	-20	770	95	8.11	1.55	23	117
	-30	560	95	5.89	1.55	23	117
	-40	398	51	7.80	1.60	23	115
	-50	327	19	17.21	0.44	23	41
		297	28	10.61	0.71	88	115
oo4	-10	1160	121	9.59	1.58	23	142
	-20	840	121	6.94	1.58	23	142
	-30	558	93	6.00	1.57	23	115
	-40	317	14	22.64	0.32	23	36
		431	21	20.52	0.72	95	115
	-50	317	14	22.64	0.32	23	36
		431	21	20.52	0.72	95	115

Table 8.5: Table of oo descriptor alignment scores for 1mbd vs. 1bab.

tide Glu18-Ala19 of myoglobin has no counterpart in the beta chain of hemoglobin.

By comparison of their X-ray crystallographic structures, myoglobin (Mb) segment 20-147 is structurally homologous with hemoglobin (Hb) segment 19-146. For example, the RMSD for the 125 pairs of  $C_\alpha$  atoms for aligning Mb-(20-144) with Hb-(19-143) is only 1.60 Å. This structure-based alignment covers helices B-H. Of the 80 dihedral analysis results reported here, 15 results found the full alignment of helices B-H. The longest such result was obtained using dihedral descriptor bb2 with a mismatch score of  $-10$  or  $-20$ , which aligned Mb-(20-144) with Hb-(19-143). Other excellent results were obtained using bb1 ( $-20$ ,  $-30$ ), bb3 ( $-20$ ,  $-30$ ), bb4 ( $-30$ ), bo2 ( $-30$ ,  $-40$ ), bo3 ( $-20$ ), ob2 ( $-50$ ), ob4 ( $-20$ ), oo3 ( $-10$ ), and oo4 ( $-10$ ,  $-20$ ). No descriptor, span, or mismatch score consistently produced this best alignment. Another 7 results found the shorter alignment of helices B-G. This very good result was obtained using bb1 ( $-40$ ), bo1 ( $-40$ ,  $-50$ ), ob1 ( $-30$ ), oo3 ( $-20$ ,  $-30$ ), and oo4 ( $-30$ ). Another very good result, the shorter alignment of helices D-H, was obtained using oo2 ( $-40$ ). A total of 13 of the 16 pendant dihedral descriptors (all except bo4, ob3, and o1) provided a correct dihedral alignment of Mb and Hb covering most of their structurally similar residues. Thus most of these pendant dihedral descriptors found most of the supersecondary structures shared by these two  $\alpha$ -helical proteins.

The hemoglobin chain is missing two residues (between Val18 and Asn19) that are present in myoglobin, so myoglobin segment 1-17 is structurally homologous with hemoglobin segment 2-18. This alignment of their A helices requires a different registration than that of the registration of the major alignments discussed above. If helix A and helices B-H are aligned in the same registration, the segments containing helix A must be misaligned. Indeed, 31 results aligned helices A-H of Mb with those of Hb in a single large alignment. For example, all five results using the oo1 descriptor had Mb-(6-144) aligned with Hb-(5-143). The RMSD for these 139 pairs of Ca atoms increased to 2.29 Å. This misalignment of helix A occurred more often for smaller negative mismatch scores (13 times for  $-10$ , 9 times for  $-20$ , 5 times for  $-30$ , 3 times for  $-40$ , and only once for  $-50$ ). In a sense, this type of result (overalignment) is an error of commission.

In contrast, 5 results properly aligned helices B-C and helices F-G in the same registration but failed to align the intervening helices D-E. Similarly, 8 results properly aligned helices D-E and helices F-G in the same registration but failed to align the intervening loop. These segmented results occurred more often for larger negative mismatch scores (only once for  $-30$ , 5 times for  $-40$ , and 6 times for  $-50$ ). Also, the

result for bo2 ( $-50$ ) properly aligned helices B-C, helices D-E, and helices F-G in the same registration but failed to align the intervening two loops. In a sense, this type of result (underalignment) is an error of omission. But in another sense, it suggests that the folding of the intervening unaligned segments is probably less conserved than the folding of the aligned helical segments.

Finally, of the 80 results examined only two results (bb4 with a mismatch score of  $-40$  or  $-50$ ) failed to give a significant dihedral alignment of myoglobin with the hemoglobin beta chain. This type of result (nonalignment) is an error of omission.

Taken together, these results indicate that the best mismatch score for general use with these 16 pendant dihedral descriptors is  $-30$ . A value of  $-10$  produced excessive overalignment of an adjacent segment (13 of 16 results), whereas a value of  $-50$  gave excessive underalignment of an intervening segment (7 of 16 results).

A visual comparison the ribbon diagrams for these two proteins is shown in Figure 8.9. The segments identified by dihedral sequence alignment are indicated in red. The residues colored green are the first residues of each segment. The last residues in each segment are colored orange.

Table (m) 1...5...10...15...20...25...30...35...40...45...50...55...60...65...70...77

```

bb1 (-10) :...:<-----
bb1 (-20) :...:<-----
bb1 (-30) :...:<-----
bb1 (-40) :...:<-----
bb1 (-50) :...:<----->.....

bb2 (-10) :...:<-----
bb2 (-20) :...:<-----
bb2 (-30) :...:<-----
bb2 (-40) :...:<----->...
bb2 (-50) :...:<-----

bb3 (-10) :...:<-----
bb3 (-20) :...:<-----
bb3 (-30) :...:<-----
bb3 (-40) :...:<-----
bb3 (-50) :...:<-----

bb4 (-10) :...:<-----
bb4 (-20) :...:<-----
bb4 (-30) :...:<-----
bb4 (-40) :...:<-----
bb4 (-50) :...:<-----

...<-helix A-----><-helix B-----><--C-->.....<--D--><-helix E----->
1...5...10...15...20...25...30...35...40...45...50...55...60...65...70...77
1mbd VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFD RFKHLKTEAEMKASEDLKKHGVTVLTALGAILK
.L...E...V...W.KV .V...G...L RL...P.T...F..F..L.T.....K.HG..VL.A...L.
1bab VHLTPEEKSAVTALWGKV NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSGDLA
1...5...10...15.18 19...25...30...35...40...45...50...55...60...65...70...76
...<-helix A-----> <-helix B-----><--C-->.....<--D--><-helix E----->

```

Figure 8.1: Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the bb descriptors.

```

Table (m) 78....85...90...95..100..105..110..115..120..125..130..135..140..145.....153
bb1 (-10) ----->:.....
bb1 (-20) ----->:.....
bb1 (-30) ----->:.....
bb1 (-40) ----->.....
bb1 (-50) .....<----->.....
bb2 (-10) ----->:.....
bb2 (-20) ----->:.....
bb2 (-30) ----->.....
bb2 (-40) .....<----->.....
bb2 (-50) .....<----->.....
bb3 (-10) ----->:.....
bb3 (-20) ----->:.....
bb3 (-30) ----->.....
bb3 (-40) .....<----->.....
bb3 (-50) .....<----->.....
bb4 (-10) ----->:.....
bb4 (-20) ----->:.....
bb4 (-30) ----->:.....
bb4 (-40) .....
bb4 (-50) .....
.....<---F--->.....<-helix G----->.....<-helix H----->.....
78....85...90...95..100..105..110..115..120..125..130..135..140..145.....153
1mbd KKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
.....L...H...K.....VL.....F...Q.A..K.....A.KY.
1bab HLDNLKGTfATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVAGVANALAHKYH
77.80...85...90...95..100..105..110..115..120..125..130..135..140..146
.....<---F--->.....<-helix G----->.....<-helix H----->...

```

Figure 8.2: Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the bb descriptors (continued).



```

Table (m) 1...5...10...15...20...25...30...35...40...45...50...55...60...65...70....77

bo1 (-10) :...:<-----
bo1 (-20) :...:<-----
bo1 (-30) :...:<-----
bo1 (-40) :...:.....:<-----
bo1 (-50) :...:.....:<-----

bo2 (-10) :...:<-----
bo2 (-20) :...:<-----
bo2 (-30) :...:.....:<-----
bo2 (-40) :...:.....:<-----
bo2 (-50) :...:.....:<----->:.....<----->.

bo3 (-10) :...<-----
bo3 (-20) :...:.....<-----
bo3 (-30) :...:.....:<----->:...
bo3 (-40) :...:.....:<----->:...
bo3 (-50) :...:.....:<----->:...

bo4 (-10) :...:<-----
bo4 (-20) :...:<-----
bo4 (-30) :...:.....:<----->:...
bo4 (-40) :...:.....:<----->:...
bo4 (-50) :...:.....:<----->:...

...<-helix A-----><-helix B-----><--C-->.....<--D--><-helix E----->
1...5...10...15...20...25...30...35...40...45...50...55...60...65...70....77
1mbd VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFD RFKHLKTEAEMKASEDLKKHGVTVLTALGAILK
.L...E...V...W.KV .V...G...L RL...P.T...F..F..L.T.....K.HG..VL.A...L.
1bab VHLTPEEKSAVTALWGKV NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSGDLA
1...5...10...15.18 19...25...30...35...40...45...50...55...60...65...70....76
...<-helix A-----> <-helix B-----><--C-->.....<--D--><-helix E----->

```

Figure 8.3: Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the bo descriptors.

```

Table (m) 78....85...90...95..100..105..110..115..120..125..130..135..140..145....153

bo1 (-10) ----->:.....
bo1 (-20) ----->:.....
bo1 (-30) ----->:.....
bo1 (-40) ----->:.....
bo1 (-50) ----->:.....

bo2 (-10) ----->:.....
bo2 (-20) ----->:.....
bo2 (-30) ----->:.....
bo2 (-40) ----->:.....
bo2 (-50) ..:.....<----->:.....

bo3 (-10) ----->:.....
bo3 (-20) ----->:.....
bo3 (-30) ..:.....<----->:.....
bo3 (-40) ..:.....<----->:.....
bo3 (-50) ..:.....<----->:.....

bo4 (-10) ----->:.....
bo4 (-20) ----->:.....
bo4 (-30) ..:.....<----->:.....
bo4 (-40) ..:.....<----->:.....
bo4 (-50) ..:.....<----->:.....

.....<---F--->....<-helix G----->....<-helix H----->....
78....85...90...95..100..105..110..115..120..125..130..135..140..145....153
1mbd KKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
.....L...H..K.....VL.....F....Q.A..K.....A.KY.
1bab HLDNLKGTfATLSELHCdKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVvAGVANALAHKYH
77.80...85...90...95..100..105..110..115..120..125..130..135..140...146
.....<---F--->....<-helix G----->....<-helix H----->...

```

Figure 8.4: Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the bo descriptors (continued).

```
Table (m) 1...5...10...15...20...25...30...35...40...45...50...55...60...65...70....77
ob1 (-10) :...<-----
ob1 (-20) :...<-----
ob1 (-30) :.....:<-----
ob1 (-40) :.....:<----->.....
ob1 (-50) :.....:<----->.....

ob2 (-10) :...:<-----
ob2 (-20) :...:<-----
ob2 (-30) :...:<-----
ob2 (-40) :...:<-----
ob2 (-50) :.....:<-----

ob3 (-10) :.....<-----
ob3 (-20) :.....<-----
ob3 (-30) :.....<-----
ob3 (-40) :.....<-----
ob3 (-50) :.....<----->:.....

ob4 (-10) :.....<-----
ob4 (-20) :.....<-----
ob4 (-30) :.....
ob4 (-40) :.....
ob4 (-50) :.....

..<-helix A-----><-helix B-----><--C-->.....<--D--><-helix E----->
1...5...10...15...20...25...30...35...40...45...50...55...60...65...70....77
1mbd VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDLKKHGVTVLTALGAILK
.L...E...V...W.KV .V...G...L RL...P.T...F..F..L.T.....K.HG..VL.A...L.
1bab VHLTPEEKSAVTALWGKV NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMGNPKVKAHGKKVLGA FSGDLA
1...5...10...15.18 19...25...30...35...40...45...50...55...60...65...70....76
...<-helix A----> <-helix B-----><--C-->.....<--D--><-helix E----->
```

Figure 8.5: Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the ob descriptors.

Table (m) 78....85...90...95..100..105..110..115..120..125..130..135..140..145....153

```

ob1 (-10) ----->:.....
ob1 (-20) ----->:.....
ob1 (-30) ----->:.....
ob1 (-40) ...:....<----->:.....
ob1 (-50) ...:....<----->:.....

ob2 (-10) ----->:.....
ob2 (-20) ----->:.....
ob2 (-30) ----->:.....
ob2 (-40) ----->:.....
ob2 (-50) ----->:.....

ob3 (-10) ----->:.....
ob3 (-20) ----->:.....
ob3 (-30) ----->:.....
ob3 (-40) ----->:.....
ob3 (-50) ...:....<----->:.....

ob4 (-10) ----->:.....
ob4 (-20) ----->:.....
ob4 (-30) ...:....<----->:.....
ob4 (-40) ...:....<----->:.....
ob4 (-50) ...:....<----->:.....

.....<---F--->.....<-helix G----->.....<-helix H----->.....
78....85...90...95..100..105..110..115..120..125..130..135..140..145....153
1mbd KKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
.....L...H..K.....VL.....F....Q.A..K.....A.KY.
1bab HLDNLKGTfATLSELHCDKLHVDPENFRLLGNVLCVLAHHFGKEFTPPVQAAYQKVAGVANALAHKYH
77.80...85...90...95..100..105..110..115..120..125..130..135..140...146
.....<---F--->.....<-helix G----->.....<-helix H----->...

```

Figure 8.6: Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the ob descriptors (continued).

```

Table (m) 1...5...10...15...20...25...30...35...40...45...50...55...60...65...70....77

oo1 (-10) :...:<-----
oo1 (-20) :...:<-----
oo1 (-30) :...:<-----
oo1 (-40) :...:<-----
oo1 (-50) :...:<-----

oo2 (-10) :...:<-----
oo2 (-20) :...:<-----
oo2 (-30) :...:<-----
oo2 (-40) :...:.....<-----
oo2 (-50) :...:.....<-----

oo3 (-10) :...:.....<-----
oo3 (-20) :...:.....<-----
oo3 (-30) :...:.....<-----
oo3 (-40) :...:.....<-----
oo3 (-50) :...:.....<----->.....

oo4 (-10) :...:.....<-----
oo4 (-20) :...:.....<-----
oo4 (-30) :...:.....<-----
oo4 (-40) :...:.....<----->.....
oo4 (-50) :...:.....<----->.....

...<-helix A-----><-helix B-----><--C-->.....<--D--><-helix E----->
1...5...10...15...20...25...30...35...40...45...50...55...60...65...70....77
1mbd VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDKHLKTEAEMKASEDLKKHGVTVLTALGAILK
.L...E...V...W.KV .V...G...L RL...P.T...F..F..L.T.....K.HG..VL.A...L.
1bab VHLTPEEKSAVTALWGKV NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSGDLA
1...5...10...15.18 19...25...30...35...40...45...50...55...60...65...70....76
...<-helix A-----> <-helix B-----><--C-->.....<--D--><-helix E----->

```

Figure 8.7: Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the oo descriptors.

Table (m) 78....85...90...95..100..105..110..115..120..125..130..135..140..145....153

```

oo1 (-10) ----->:.....
oo1 (-20) ----->:.....
oo1 (-30) ----->:.....
oo1 (-40) ----->:.....
oo1 (-50) ----->:.....

oo2 (-10) ----->.....
oo2 (-20) ----->.....
oo2 (-30) ----->.....
oo2 (-40) ----->.....
oo2 (-50) ----->.....

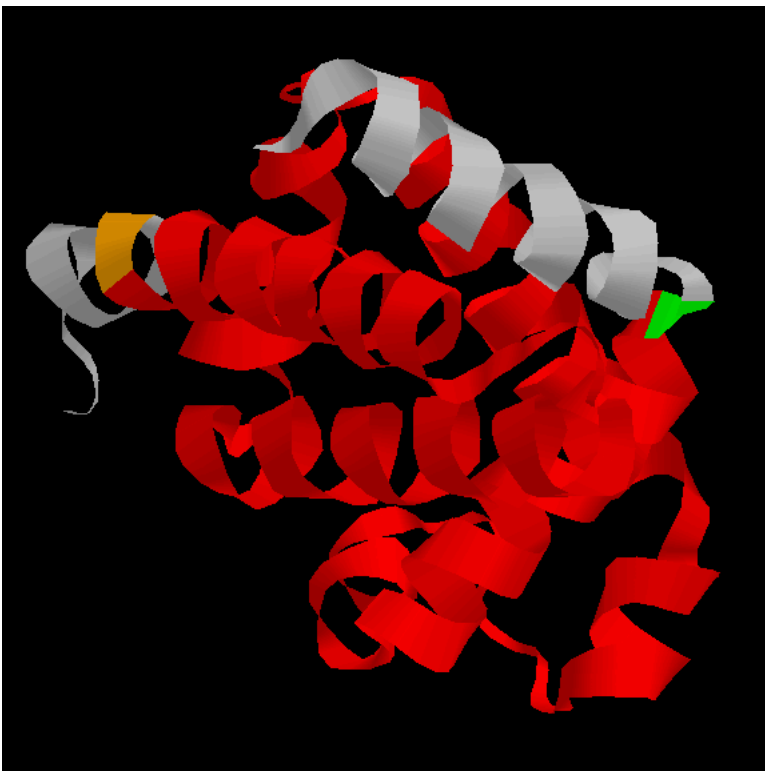
oo3 (-10) ----->.....
oo3 (-20) ----->.....
oo3 (-30) ----->.....
oo3 (-40) ----->.....
oo3 (-50) .....<----->.....

oo4 (-10) ----->.....
oo4 (-20) ----->.....
oo4 (-30) ----->.....
oo4 (-40) .....<----->.....
oo4 (-50) .....<----->.....

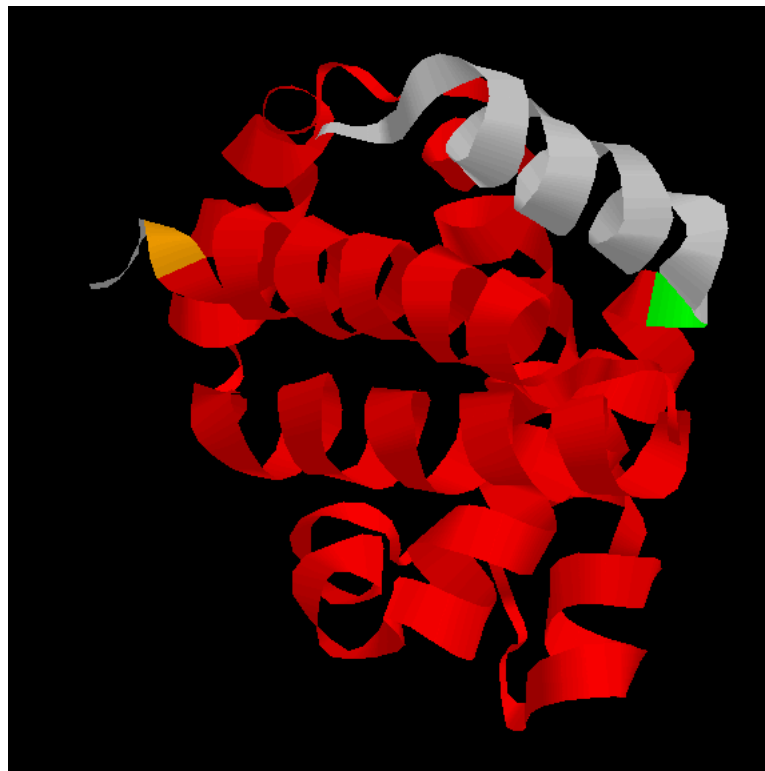
.....<---F--->.....<-helix G----->.....<-helix H----->.....
78....85...90...95..100..105..110..115..120..125..130..135..140..145....153
1mbd KKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
.....L...H..K.....VL.....F....Q.A..K.....A.KY.
1bab HLDNLKGTfATLSELHCDKLHVDPENFRLLGNVLCVLAHHFGKEFTPPVQAAYQKVAGVANALAHKYH
77.80...85...90...95..100..105..110..115..120..125..130..135..140...146
.....<---F--->.....<-helix G----->.....<-helix H----->...

```

Figure 8.8: Alignment of sperm-whale myoglobin (1mbd) and the beta chain of human hemoglobin Thionville (1bab) for the oo descriptors (continued).



(a) 1mbd



(b) 1bab

Figure 8.9: 3D ribbon diagrams of the  $\alpha$ -helical proteins myoglobin (a) and hemoglobin beta chain (b). The segments in red were found to be in alignment by the bb2 dihedral descriptor using a mismatch score of  $-20$ .

### 8.2.2 Alignment of 1cd8 and 2rhe

The second set of dihedral alignments that we will consider in detail is comparison of the human T-cell co-receptor CD8 (1cd8) (Leahy et al., 1992) with the immunoglobulin lambda chain of human Bence-Jones protein RHE (2rhe) (Furey et al., 1983). Each of the 16 pendant dihedral descriptors was used with the three useful mismatch scores ( $-20$ ,  $-30$ , and  $-40$ ). Table 8.6 summarizes the segment statistics for all meaningful alignments found in the same registration for the 16 positive results. Since the 3D structure of CD8 is not fully determined below residue 57, an alignment containing residues below that position would not be valid.

descriptor	score	length	average	RMSD
bb1 (-20)	517	31	16.68	0.82
bb1 (-30)	476	27	17.63	0.82
bb1 (-40)	451	25	18.04	0.75
bo1 (-20)	317	28	11.32	0.97
bo3 (-20)	557	26	21.42	0.70
bo3 (-30)	537	26	20.65	0.70
bo3 (-40)	517	26	19.88	0.70
bo4 (-20)	503	26	19.35	0.70
oo1 (-20)	472	32	14.75	0.94
oo1 (-30)	442	32	13.81	0.94
oo1 (-40)	412	32	12.88	0.94
oo3 (-20)	537	29	18.52	0.76
oo3 (-30)	442	32	13.81	0.94
oo3 (-40)	477	29	16.45	0.76
oo4 (-20)	523	31	16.87	0.95
oo4 (-30)	483	31	15.58	0.95

Table 8.6: Alignment statistics for 1cd8 vs. 2rhe.

As can be seen from Table 8.6, the length of the aligned segments vary between 27 and 31 amino acid residues and the RMSD for pairs of  $C_\alpha$  atoms varied between 0.70 Å and 0.97 Å. The ob type descriptors are conspicuously absent. Of the bb descriptors only bb1 is present. The bo family is represented by bo3 and bo4 and all the oo descriptors are present except oo2.

The alignments are show in Figure 8.10 for the bb1, bo3 and bo4 descriptors and in Figure 8.11 for the oo1, oo3, and oo4 descriptors. The dihedral sector letters



are aligned by the sequence alignment algorithm. Exact matches are echoed on the line between each pair of dihedral sequences. Aligned pairs of dihedral sectors that contribute positively to the alignment score but are not exact matches are shown with a “+” sign on the same line. All of these alignments have the same registration between the sequences. Shown at the bottom of each figure is an alignment of the amino acid sequences of the two proteins annotated to show the secondary structural features of 2rhe. All of these alignments cover three  $\beta$ -strands and two turns. Thus dihedral sequence comparison can produce alignments of supersecondary structures present in  $\beta$ -sheet proteins.

Figure 8.12 shows a visual summary of all of these alignments, which are represented as double-headed arrows. Each alignment has a common core starting at position 71 (identified by bb1 and bo1) and extending to position 93 (identified by bo3 and bo4). Some descriptors extend this alignment to positions 67 (bb1 and oo1) and 98 (bo1, oo1, and oo4).

As in the Mb/Hb example, the oo1 descriptor found the same segment regardless of mismatch score and also finds the longest segment, 32 amino acid residues long with an RMSD of 0.94 Å. The best alignment in terms of RMSD was found by bo3 starting at position 68 and extending for 26 residues having an RMSD of 0.70 Å.

A visual comparison of the ribbon diagrams for these two proteins is presented in Figure 8.13. The segments in red were found to be in alignment by the oo1 dihedral descriptor using a mismatch score of  $-30$ . The residues colored green are the first residues of each segment. The last residues in each segment are colored orange.

# A

Table (m) 60...65...70...75...80...85...90...95...100...105...110...114		
1cd8	67	OXUEBBOIQBBBAACMBAGQQTGAXEAXCA 97
bb1 (-20)		+X B+O+Q+BB+++ +AGQQT++A+ AX A
2rhe	62	PXABBXOHQABBBBBPAAGQQTUEABBAXGA 92
1cd8	71	BBOIQBBBAACMBAGQQTGAXEAXCA 97
bb1 (-30)		B+O+Q+BB+++ +AGQQT++A+ AX A
2rhe	66	BXOHQABBBBBPAAGQQTUEABBAXGA 92
1cd8	71	BBOIQBBBAACMBAGQQTGAXEAX 95
bb1 (-40)		B+O+Q+BB+++ +AGQQT++A+ AX
2rhe	66	BXOHQABBBBBPAAGQQTUEABBAX 90
1cd8	71	TROWIRQQQPEPPSFHIRWRRRTSSSSP 98
bo1 (-20)		+R+++++QQQ+ +S+HI+ +R+SSS++
2rhe	66	SRPXJSSQQQQLRSGHIQTQRRSSSRQ 93
1cd8	68	JTUXTMPWMUUTRGIGVJQEIoTBTv 93
bo3 (-20,-30,-40)		+T+ +M++M+UT+ +++JQ+IO++TV
2rhe	63	LTVTRMRVMTUTSVHITJQDIOSXTV 88
1cd8	68	WHIKBGUMBJIGTUVVWGTQUEJOKK 93
bo4 (-20)		++I B++ ++++ U+++G+QUE++K+
2rhe	63	XIIGBHVJAIJIKUWUVGSQUEIMKJ 88

# B

```

.....:.....:.....:.....:.....:.....:.....:.....:.....:.....
60...65...70...75...80...85...90...95...100...105...110...114
1cd8  AGELDTQRFSGKRLGDTFVLTLSDFRRENEGYYFCSALSNSIMYFSHFVPVFLPA
.....RFS....G....L..S....E.E..Y.C.A...S.....L..
2rhe  LPSGVSDRFSASKSGTSASLAISGLESEDEADYYCAAWNDSLDEPGFGGGTKLTV
55...60...65...70...75...80...85...90...95...100...105...109
.....<strA3>.<-strA2-->.....<strB3>.....<strB4->..
.....<t4>....<t5>.....<t6>.....<t7>.....

```

Figure 8.10: Dihedral sequence alignment of the human T-cell co-receptor CD8 (1cd8) with the immunoglobulin lambda chain of human Bence-Jones protein RHE (2rhe) using the bb1, bo1, bo3, and bo4 descriptors. **A.** Alignment of their dihedral sequences. The descriptor and the mismatch score (m) are indicated. **B.** Alignment of their amino acid sequences.

# A

Table (m)	60...65...70...75...80...85...90...95...100...105...110...114
1cd8	67 MBBAAXXUQBBBAXNACCNOODABAACAXCW 98
oo1 (-20,-30,-40)	+++AAX+U+B++B++ +NOODAB++C+X++
2rhe	62 NCAAAAXWUPBAXBBADKGBNOODABBBBCXXEA 93
1cd8	68 RDCBXSCXVEDCBNOPFSALQWECDEEDC 96
oo3 (-20,-30,-40)	+++B+SCX+++C+ 0+ ++++ECD++++
2rhe	63 TEBBWSCXUCBCCEOQQVXKRXCDFDBD 91
1cd8	68 GQOMHNKMJRPOCDDGGPDDEMQPSRQRNXQ 98
oo4 (-20,-30)	+Q++H+++++PO +++ +D++QPSR+R+ Q
2rhe	63 IQNLH MJLIPPOSEEFSTBDFNQPSRORPFQ 93

# B

```

:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....
60...65...70...75...80...85...90...95...100...105...110...114
1cd8 AGELDTQRFSGKRLGDTFVLTLSDFRRENEGYYFCSALSNSIMYFSHFVPVFLPA
.....RFS....G....L..S....E.E..Y.C.A...S.....L..
2rhe LPSGVSDRFSASKSGTSASLAISGLESEDEADYYCAAWNDSLDEPGFGGGTKLTV
55...60...65...70...75...80...85...90...95...100...105...109
.....<strA3>.<-strA2-->.....<strB3>.....<strB4->..
.....<t4>....<t5>.....<t6>.....<t7>.....

```

Figure 8.11: Dihedral sequence alignment of the human T-cell co-receptor CD8 (1cd8) with the immunoglobulin lambda chain of human Bence-Jones protein RHE (2rhe) using the oo1, oo3, and oo4 descriptors. **A.** Alignment of their dihedral sequences. The descriptor and the mismatch score (m) are indicated. **B.** Alignment of their amino acid sequences.

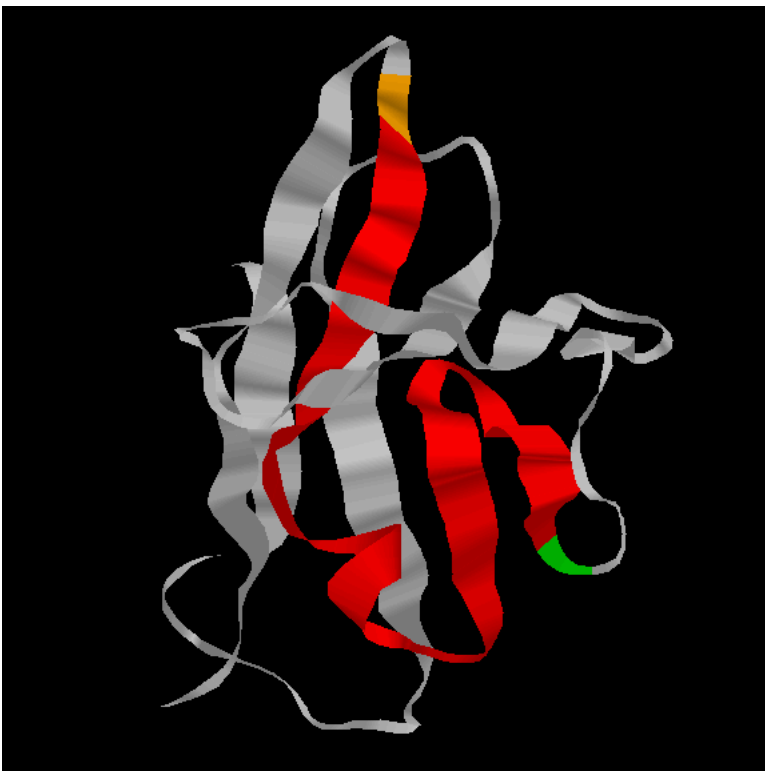
## A

Table (m)	60...65...70...75...80...85...90...95..100..105..110.114
bb1 (-20)	:.....<----->.....
bb1 (-30)	:.....<----->.....
bb1 (-40)	:.....<----->.....
bo1 (-20)	:.....<----->.....
bo3 (-20)	:.....<----->.....
bo3 (-30)	:.....<----->.....
bo3 (-40)	:.....<----->.....
bo4 (-20)	:.....<----->.....
oo1 (-20)	:.....<----->.....
oo1 (-30)	:.....<----->.....
oo1 (-40)	:.....<----->.....
oo3 (-20)	:.....<----->.....
oo3 (-30)	:.....<----->.....
oo3 (-40)	:.....<----->.....
oo4 (-20)	:.....<----->.....
oo4 (-30)	:.....<----->.....

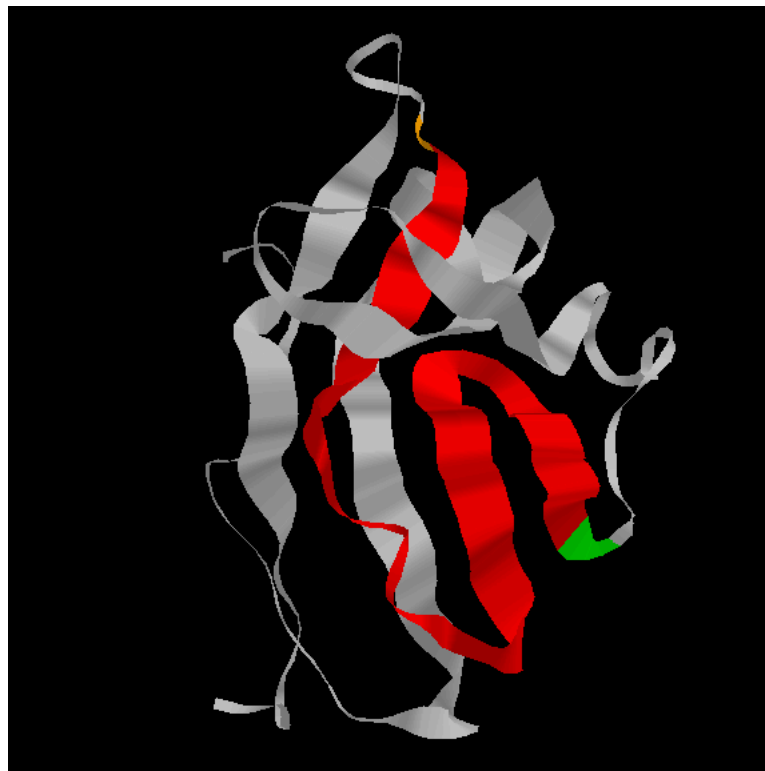
## B

	:.....
	60...65...70...75...80...85...90...95..100..105..110.114
1cd8	AGELDTQRFSGKRLGDTFVLTLSDFRRENEGYFCSALSNSIMYFSHFVPVFLPA
	.....RFS....G....L..S....E.E..Y.C.A...S.....L..
2rhe	LPSGVSDRFSASKSGTSASLAISGLESEDEADYYCAAWNDSLDEPGFGGGTKLTV
	55...60...65...70...75...80...85...90...95..100..105.109
	.....<strA3>.<-strA2-->.....<strB3>.....<strB4->..
	.....<t4>....<t5>.....<t6>.....<t7>.....

Figure 8.12: Alignment of the human T-cell co-receptor CD8 (1cd8) with the immunoglobulin lambda chain of human Bence-Jones protein RHE (2rhe). **A.** Alignment of their dihedral sequences. The score table and the mismatch score (m) are indicated. **B.** Alignment of their amino acid sequences. Identical residues are echoed in the middle line. Double-headed arrows mark the positions of four beta strands (strA3, strA2, strB3, strB4) and four beta turns (t4-t7) of RHE. The 3D structure of CD8 is not fully determined below residue 57.



(a) 1cd8



(b) 2rhe

Figure 8.13: 3D ribbon diagrams of the human T-cell co-receptor(a) and the immunoglobulin lambda chain of human Bence-Jones protein RHE (b). The segments in red were found to be in alignment by the oo1 dihedral descriptor using a mismatch score of  $-30$ .

### 8.2.3 Alignment of 1rcf and 4fxn

The third alignment considered in detail is between the oxidized form of flavodoxin from *Anabaena 7120* (1rcf) (Rao et al., 1992) and the semiquinone form of flavodoxin from *Clostridium MP* (4fxn) (Smith et al., 1977). Each of the 16 pendant dihedral descriptors was used with a mismatch score of  $-30$ . Table 8.7 summarizes the alignment statistics for all meaningful alignments found in the same registration for the 7 positive results.

descriptor	score	start	end	length	average	RMSD
bb1 (-30)	261	28	87	60	4.35	2.55
bo1 (-30)	375	13	88	76	4.93	3.15
bo4 (-30)	284	29	71	43	6.60	2.34
ob2 (-30)	406	48	87	40	10.15	2.71
oo1 (-30)	387	14	89	76	5.09	3.09
oo2 (-30)	303	94	115	22	13.77	2.87
	264	48	69	22	12.00	2.14
	232	5	23	19	12.21	0.47
oo3 (-30)	281	39	68	30	9.37	2.33

Table 8.7: Alignment statistics for 1rcf vs. 4fxn.

This set of alignments (Figure 8.14) shows a less exact but still significant structural similarity of these two proteins. Seven of the 16 descriptors (bb1, bo1, bo4, ob2, oo1, and oo2) found significant alignments using a miss penalty of  $-30$ . As in the other two examples, both the bb1 and oo1 descriptors provide an alignment. The oo1 descriptor found one of the two longest (76 residues) and least similar (RMSD of 3.09 Å) results. The 60-residue bb1 alignment is structurally more similar (RMSD of 2.55 Å). The oo3 alignment is most similar (RMSD of 2.33 Å), but at only 30 residues is only half the length of the alignment found by bb1. Because these two proteins are structurally less similar than in the previous two examples, their alignments vary much more in length and position.

Six alignments of dihedral sequences for 1rcf and 4fxn are shown in Figure 8.15. Part **A** shows alignment of segments of their amino acid sequences using the BioSCAN system and the indicated BLOSUM (Block Substitution Matrix) score tables. The probability that an alignment is due to chance is given in parentheses. Part **B** shows annotated alignment of their amino acid sequences. Identical amino-acid residues are echoed in the middle line. Alpha helices are marked <helix#>, the  $3_{10}$  helix as

Table (m)	10...	15...	20...	25...	30...	35...	40...	45...	50...	55...	60...	65...	70...	75...	80...	85...	89
1rcf																	
bb1 (-30)																	
4fxn																	
	28	REABCACDSORDGQSRROCOAABPGDDANGRRFGGPQROQRROTRRPDAUPQCAACXNM	87														
		R+AB+A+ +O++ +++ ++A+ + ++G R+G P ++Q+++ +P U+++A++X++															
	27	RDABBAEGROSCDTPPQPRFNBAXMJEGXPGJRGDPNPQQPQQOSPQQUNREABBXP	86														
1rcf	13	EFFFEGFFGEEFHHLHQQQQSPQHGISRGGGHRHPRQPAQRQSLNFSRRFFGFGFFGIGIHRPJRIRQQSQSDS	88														
bo1 (-30)		+FF+E+FF+E +++ + +Q+++++S+++ ++ R+P+++ ++Q+ N ++R+++F+F+++++++ +R+R+++++S															
4fxn	12	FFEEEEFFEEOEFGIITTSQQQPFFHSQFFSEHKRIPPPQDRTQUWNLTSRGHFFEFEGFHFPIIRHRSRRRTCS	87														
1rcf																	
bo4 (-30)																	
4fxn	29	EKJIJWHSQQPGRIJIOFKAVHHGIRNUFOUNIGMSHHJGJIG	71														
		+KJI+++S++P+ I I+ ++++++ +++N++++ ++++++															
	28	CKJIKAJSRPICIFIPINCTGGEKARIENVNKHNTNFHHHHH	70														
1rcf																	
ob2 (-30)																	
4fxn	48	XHHHAFILIEPRNOPDCDBCDDBDCAKIEBQJHHJGSH	87														
		XHHH FI ++++++++ +CD++D+++ E+Q +++G++															
	47	XHHHTFIOJWDRQMPNBXWCCDCDDBDCAREAQMJGIGUG	86														
1rcf	14	NNMMNNMMNNNTOCDAABCWCONODCNONNOOBMABXBAWDBCJLKDDENMMMMNNOONONBAOEJCBACAAABA	89														
oo1 (-30)		N+++NN+M+++NN +++++C+++N+ ++ O+B +++++AW++C K++EN+ M++++N+++ +E + A++A+++															
4fxn	13	NMNNNNMMNNMNOKDCXXCADPNNGXMNDAOQBPAAXAAWFDSCXKBEENOKMNNMMNNNOAKNEFDEAABAWDB	88														
1rcf																	
oo3 (-30)																	
4fxn	39	WTAJLJJJGQRBCBADEHNXQAQMWFKKLL	68														
		WTA +J + +++++BA+ X Q+++++ L															
	38	WTASJJWPJRQXBBAEHLAXVJQKXGMLOL	67														
Table (m)	10...	15...	20...	25...	30...	35...	40...	45...	50...	55...	60...	65...	70...	75...	80...	85...	89

Figure 8.14: Six alignments of dihedral sequences for the flavodoxins 1rcf and 4fxn.

<3:10>, beta strands as <strand#>, beta turns as <t#>, and sites in contact with the prosthetic group flavin mononucleotide as {FMN#}. Part **C** shows alignment of segments of their dihedral sequences using the BioSCAN system and the indicated dihedral score tables (mismatch score = -30). The aligned dihedral segments from 1rcf are noted in parentheses. The part of a dihedral alignment that is different from that shown in **B** is marked as <===.

Part **B** shows by two criteria (amino acid sequence alignment and conserved structural features) that segments 4-23, 24-29, and 31-89 of 1rcf are aligned in different registrations with 4fxn. As shown in part **C**, the parts of the four dihedral alignments that extend below position 31 of 1rcf are misaligned because they are in the wrong registration. This is most pronounced for the long bo1 and oo1 alignments. Like them, the more conservative bb1 alignment successfully recognized the pairing of strands 1, 3, and 4 as well as  $\alpha$ -helix 2 and the second cluster of residues that interact with FMN, the essential nonprotein component. This FMN-binding site and strand 3 were successfully aligned by all 7 dihedral descriptors. Thus, multiple dihedral sequence comparisons found a supersecondary structure containing mixture of helices and  $\beta$ -strands.

Of particular interest is the result found for the oo2 descriptor consisting of three short alignments each in a different registration. In Figure 8.16 identical amino-acid residues are echoed in the middle line. The aligned dihedral segments from 1rcf are noted. Alpha helices are marked <helix#>, the  $3_{10}$  helix as <3:10>, beta strands as <strand#>, beta turns as <t#>, and sites in contact with the prosthetic group flavin mononucleotide as {FMN#}.

Figure 8.17 shows the three dihedral alignments for the oo2 descriptor. The first alignment is 22 residues long in the +3 (5-2) registration. The second alignment is also 22 residues long but in the +1 (48-47) registration. The last alignment is 19 residues long and is in the +6 (84-88) registration. Each of the three alignments correctly pair a set of FMN-binding residues, which are indicated as {FMN#} in Figure 8.16.

A visual comparison of the ribbon diagrams for these two flavodoxins is shown in Figure 8.18. The segments in red were found to be in alignment by the bb1 dihedral descriptor using a mismatch score of -30. The residues colored green are the first residues of each segment. The last residues in each segment are colored orange.

Figure 8.19 shows ribbon diagrams with the three alignments found by the oo2 dihedral descriptor. Each of these alignment, colored red, green, and orange, are in a different registration.



**A**

```

.  blosum80  (p = 0.31) .      .      .      .      .  blosum40 (p = 0.042)      .      .
.  <-----> .      .  <----->
.      .      .      .      .      .      .      .
.  blosum100 (p = 0.45) .      .      .      .      .  blosum60 (p = 0.025)      .      .
.  <-----> .      .      <----->

```

**B**

```

.....{FMN1}.....          .....{FM2}.....          .....{
<str2>.....          .....<stra1>.....<stra3>.....<strand4->
.....          .....<t1>.....<t2>.....          .....<t3>.....<t4>.....
.....<---helix1  -->.....<3:10->.....          .....<heli2>.....
1...5...10...15...20... 24...30...35...40...45...50...55...60...65...70...75...80...85..89
1rcf SKKIGLFYGTQTGKTESVAEIIR  DEFGNDVVTHHDVSQAEVTDLNDYQYLIIGCPTWNIGELQSDWEGLYSELDDVDFNGKLVAYFGTG
      Y  TG TE  AE I    E G D V    VS    L    LI GC    L    E    GK VA FG
4fxn  MKIVYWSGTGNTKMAELIAKGIIESGKD VNTINVSDVNIDELLNEDILILGCSAMGDEVLEESEFEPFIEEISTKISGKKVALFGSY
1...5...10...15...20...25..29  ...35...40...45...50...55...60...65...70...75...80...85..88
.....<-----helix1----->..  ...<t1>.....<t2>.....<t3>.....<helix2->...<t4>.....
<st2>.....          .....< stra1>.....          .....<st3>.....          .....<strand4>
.....{FMN1}.....          .....{FMN2}.....          .....{F

```

**C**

```

1rcf 1...5...10...15...20... 24...30...35...40...45...50...55...60...65...70...75...80...85..89
.      .      .      .      .  <===-----bb1 (28-87)----->
.      .  <=====-----bo1 (13-88)----->
.      .      .      .      .  <=-----bo4 (29-71)----->      .      .
.      .      .      .      .      .      .  <--ob2 (48-87)----->
.      .  <=====-----oo1 (14-89)----->
.      .      .      .      .      .      .  <-----oo3 (39-68)----->      .      .

```

Figure 8.15: Alignment of sequences for the flavodoxins from *Anabaena 7120* (1rcf) and *Clostridium MP* (4fxn). **A.** Alignment of segments of their amino acid sequences using the indicated BLOSUM score tables. **B.** Annotated alignment of their amino acid sequences. **C.** Alignment of segments of their dihedral sequences using the indicated dihedral score tables (mismatch score =  $-30$ ).

```

.....{FMN1}.....{FM2}.....{
<str2>.....<stra1>.....<stra3>.....<strand4->
.....<t1>.....<t2>.....<t3>.....<t4>.....
.....<---helix1-->.....<3:10->.....<heli2>.....

1...5...10...15...20...24...30...35...40...45...50...55...60...65...70...75...80...85..89
1rcf SKKIGLFYGTQTGKTESVAEIIR DEFGNDVVTHHDVSQAEVTDLNDYQYLIIGCPTWNIGELQSDWEGLYSELDDVDFNGKLVAYFGTG
      Y   TG TE  AE I   E G D V   VS   L   LI GC   L   E   GK VA FG
4fxn  MKIVYWSGTGNTKMAELIAKGIIESGKD VNTINVSDVNIDELLNED ILILGCSAMGDEVLEESEFEPFIEEISTKISGKKVALFGSY

1...5...10...15...20...25..29 ...35...40...45...50...55...60...65...70...75...80...85.88
.....<---helix1----->.. ...<t1>.....<t2>.....<t3>.....<helix2->...<t4>.....
<str2>.....<stra1>.....<st3>.....<strand4>
.....{FMN1}.....{FMN2}.....{F

<---oo2 (5-23)---->                <-----oo2 (48-69)---->

.{FMN4}.....
.....<s5>
.....<t5>.....
.....<-helix3-->.....

93....100..105..110....118
1rcf AYADNFQDAIGILEEKISQRGGKTVG
      Y           EE   G   V
4fxn SYGWGDGKWMRDFEERMNGYGCVVVE
87.90...95..100..105....112

.....<--helix3-->.....
4>.....<st5>
{FM3}.....

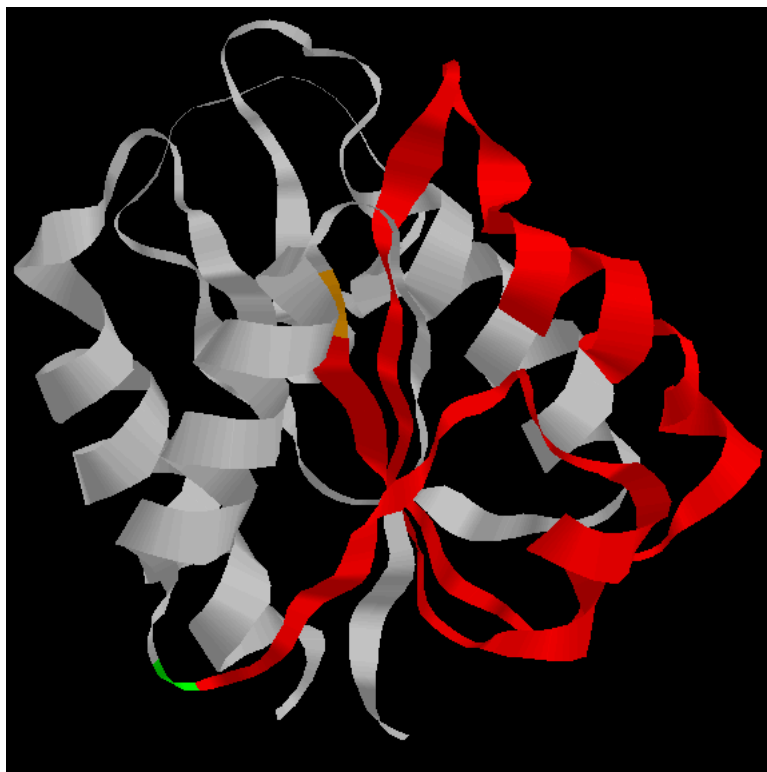
<-----oo2 (94-115)---->

```

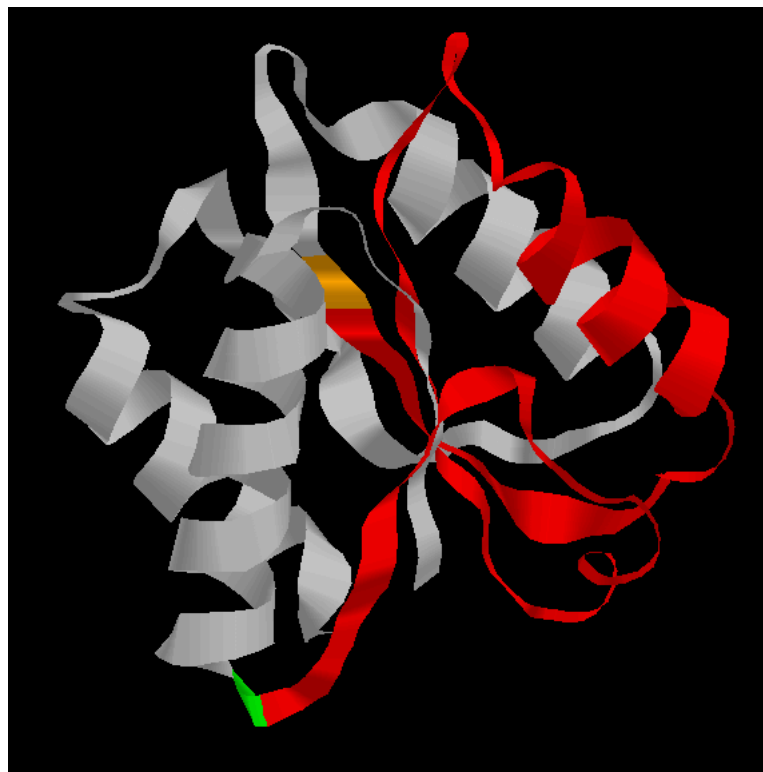
Figure 8.16: Alignment of segments of the oo2 (−30) dihedral sequences of the flavodoxins from *Anabaena 7120* (1rcf) and *Clostridium MP* (4fxn). The aligned dihedral segments from 1rcf are noted.

Table (m)	1...5...10...15...20...25...30...35...40...45...50...55...60...65...70	
1rcf	5 NNOPMWFIJMNNMNNNNNN 23	48 DNNMNLORPAJMATUGNMNNMO 69
oo2 (-30)	+NO+++ I+++N+++NNN+	D++MNL+ + + A++G+++N+O
4fxn	2 MNONLXIIKLONOOOONNO 20	47 DLMNMLQURJHUASVGOLMNNO 68
Table (m)	71..75...80...85...90...95..100..105..110..115..120	
1rcf	94 AICJDLPMNMNMOMNNMFIBRP 115	
oo2 (-30)	+I + L+MN+N+++NNM++BR+	
4fxn	88 WILKLLOMNNNONONNMHKBRN 109	

Figure 8.17: Three dihedral alignments for the flavodoxins 1rcf and 4fxn using the oo2 descriptor.

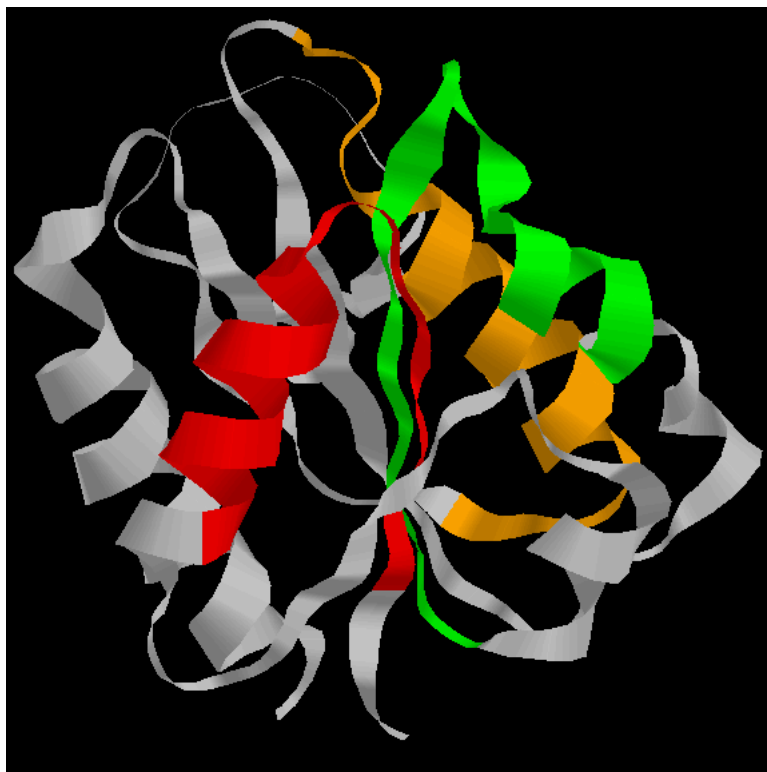


(a) 1rcf

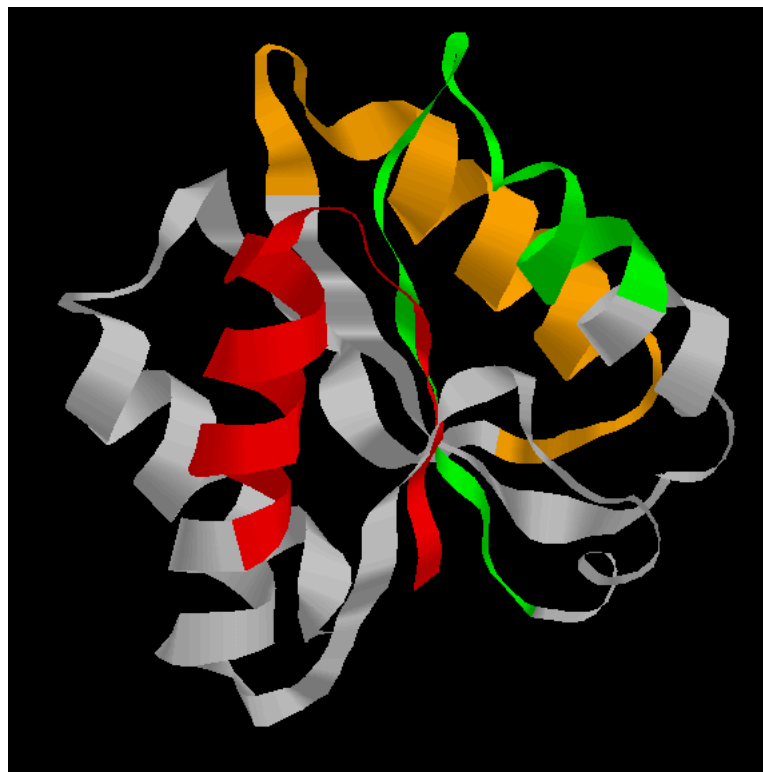


(b) 4fxn

Figure 8.18: 3D ribbon diagrams of the flavodoxins 1rcf (a) and 4fxn (b). The segments in red were found to be in alignment by the bb1 dihedral descriptor using a mismatch score of  $-30$ .



(a) 1rcf



(b) 4fxn

Figure 8.19: 3D ribbon diagrams of the flavodoxins 1rcf (a) and 4fxn (b). The segments in red, green, and orange were found to be in alignment by the oo2 dihedral descriptor using a mismatch score of  $-30$ .

## 8.3 Discussion

Our dihedral alignment method of protein structure comparison is simple and straightforward. It finds structurally similar local protein folds that agree with results of a much more computationally intensive, multi-parameter method of 3D comparison (Orengo et al., 1992). The speed of dihedral sequence comparison when implemented on the BioSCAN system is one of its most attractive features. Comparison of a folded protein with a database of 3754 folded proteins requires on the average just 12 seconds. In contrast, comparison of a folded protein with a database of 720 folded proteins by the method of Orengo et al on a Sun-4/280 workstation required 1.7 CPU hours (Orengo et al., 1992). A comparison of the speed of the BioSCAN algorithm on various architectures has been described (Hoffman, 1993a). This ability to compare a new protein structure against the entire PDB database of folded proteins in a matter of seconds encourages the everyday interactive use of such database searches.

The sensitivity of dihedral sequence comparison is determined by the score table used. By adjusting the mismatch score used in the score table, tolerance for intervening sequences of dissimilar structure can be increased or decreased to fit the problem. Searches for small, highly conserved structures, such as “foldons” (Panchenko et al., 1996), are best accomplished using a  $-40$  or  $-50$  mismatch score. When searching for less exact folding motifs, a more lenient mismatch score of  $-30$  or  $-20$  is indicated. Our experimental results indicate that a mismatch score of  $-10$  is not useful because it usually results in overalignment.

All 16 of the pendant dihedral angle descriptors are capable of detecting various secondary structure motifs, such as  $\alpha$ -helical,  $\beta$ -strand,  $\alpha + \beta$ , and  $\alpha/\beta$ . Dihedral sequence comparison can find significant alignments of supersecondary structures such as multiple  $\alpha$ -helices or  $\beta$ -strands or a combination of helices and strands. This method considers not only secondary structural elements but also turns, loops, and even disorganized regions when scoring a local alignment.

# Chapter 9

## Conclusions

I have demonstrated that 3D structure comparisons can be successfully performed using a 1D dihedral sequence alignment technique. The effectiveness of dihedral sequence alignment is demonstrated by its ability to detect significant local structural similarities in two folded proteins that have marginal amino acid sequence similarity. Tests using representative sets of protein structures have shown this method to be accurate and robust.

In algorithmic terms, dihedral transformation successfully reduces an alignment problem in three dimensions into a sequence alignment problem in one dimension. The computationally intensive task of transforming the structural representation of the folded proteins from atomic XYZ coordinates into dihedral sequences is performed prior to executing the main loop of the alignment algorithm. The transformation process has computational complexity of  $O(n)$ . Using dihedral sequences the main loop of the alignment algorithm is very simple, requiring only a table lookup, an integer addition, and a comparison test. No computationally intensive calculations are performed inside of the main loop. The alignment process has computational complexity of  $O(mn)$ .

A further practical advantage of this method is the parsimonious representation of the protein structure database. For a protein with  $N$  amino acid residues the dihedral sector sequence is  $N$  bytes long. If the dihedral angles are not binified, two to four times this storage space would be required to represent the dihedral sequence. If atomic XYZ coordinates are used the storage requirements are twelve times  $N$  even when using a single atom position per amino acid. If PDB files are used as the direct source of structural data, more than 200 MBytes would be required to represent the

structures stored in 1.7 MBytes after dihedral transformation and binification. It is not the amount of storage space required to hold the database on secondary storage that is important here, but the amount of data that must be read when performing a scan of the database. Even a small workstation can cache a 1.7 MByte data file in main memory but few systems have enough main memory to cache 200 MBytes. I/O bandwidth becomes a limiting factor with such large volumes of data.

Actual run times for this algorithm when comparing a single protein against the entire PDB containing 3754 protein structures requires about 12 seconds using the BioSCAN custom hardware. Run time on a workstation comparable to an HP 735/99 is under 80 seconds. This is in contrast to the fastest previous methods that required 15 to 25 minutes for a similar search.

## 9.1 Future work

Numerous avenues for future investigation exist. Finding new metrics for automatically classifying protein folds is an immediate area for investigation. A more exciting prospect is investigating the use of long-span dihedral descriptors for use in identifying non-local structural similarities. Descriptors with spans of twenty, fifty, or one hundred residues may be able to identify large repetitive motifs, such as those found in  $\alpha/\beta$  barrels, Greek keys, and jelly rolls (Richardson and Richardson, 1989). This application would require generation of new score tables and score evaluation methods.

After two decades of research, new methods of table generation for nucleic acid and amino acid sequence alignment continue to emerge. With this new application of sequence alignment techniques there is little doubt that the sensitivity of the score tables can be improved. It is possible that a block substitution method, similar to the approach used in generating the BLOSUM scoring tables for amino acid sequence alignments (Henikoff and Henikoff, 1992), can generate score tables with sensitivities aimed at particular local folds or protein families.

As things now stand, the software tools developed during the course of my research are effective but certainly not usable by a non-computer scientist. It would be good to develop these tools to a state similar to the BioSCAN tools which allow public access via the World Wide Web (WWW) (<http://genome.cs.unc.edu>). This would require generation of more easily interpreted alignment output, more efficient multi-descriptor report consolidation software, and a friendlier user interface.



# Appendix A

## Relationship between the dihedral angles $\psi$ , $\phi$ , and $\text{oo1}$

The following is a generalized derivation of the  $\text{oo1}$  dihedral angle from the standardly defined  $\phi$  and  $\psi$  dihedral angles.

### A.1 Generalized rotational transformation matrix.

In general the position of an atom in space can be represented by a vector  $\mathbf{v}$  with components  $v_x, v_y, v_z$ . This vector can be transformed from its own frame of reference to a new frame of reference, the result of rotation about an axis, by multiplying the vector by a *transformation matrix*.

$$\mathbf{v}' = \mathbf{R}\mathbf{v} \tag{A.1}$$

To transform Cartesian coordinates by rotation around the x-axis the following rotational transformation matrix is used:

$$\mathbf{R}_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{pmatrix} \tag{A.2}$$

Where  $\theta$  is the angle of rotation. Similarly, to perform rotational transformation about the z-axis the following matrix is used:

$$\mathbf{R}_z = \begin{pmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (\text{A.3})$$

## A.2 Positions of the carbonyl carbon and oxygen atoms.

To compute the positions of the  $C_{i-1}$ ,  $O_{i-1}$ ,  $C_i$ , and  $O_i$  from the associated  $\phi$  and  $\psi$  angles is done by defining two local coordinate systems; one for the  $C_i$  and  $O_i$ , and another for the  $C_{i-1}$  and  $O_{i-1}$ . These coordinate systems both have the  $C_i^\alpha$  as their origin but define the local x-axis differently.

The coordinate system used to compute the  $C_i$  and  $O_i$  positions defines the x-axis to be along the  $C_i^\alpha$ - $C_i$  bond while the coordinate system used to compute the  $C_{i-1}$  and  $O_{i-1}$  positions defines the x-axis to be along the  $C_{i-1}^\alpha$ - $N_{i-1}$  bond. Defining the atom positions in terms of  $\phi$  and  $\psi$  is simply a matter of trigonometry, requiring knowledge of the bond lengths and angles involved.

The following bond lengths are defined:

$$\begin{aligned} l_1 &= \|C^\alpha - C\| \\ l_2 &= \|C - O\| \\ l_3 &= \|N - C^\alpha\| \\ l_4 &= \|C - N\| \end{aligned} \quad (\text{A.4})$$

The following bond angles are defined:

$$\begin{aligned}
\theta_1 &= \angle C^\alpha C O \\
\theta_2 &= \angle N C^\alpha C \\
\theta_3 &= \angle N C O \\
\theta_4 &= \angle C^\alpha N C
\end{aligned} \tag{A.5}$$

### A.3 $\mathbf{C}_i$ and $\mathbf{O}_i$ postitions in local coordinates.

$$\mathbf{C}_i = \begin{pmatrix} l_1 \\ 0 \\ 0 \end{pmatrix} \tag{A.6}$$

$$\mathbf{O}_i = \begin{pmatrix} l_1 + l_2 \cos(180^\circ - \theta_1) \\ l_2 \sin(180^\circ - \theta_1) \\ 0 \end{pmatrix} \tag{A.7}$$

### A.4 $\mathbf{C}_{i-1}$ and $\mathbf{O}_{i-1}$ postitions in local coordinates.

$$\mathbf{C}_{i-1} = \begin{pmatrix} l_3 + l_4 \cos(180^\circ - \theta_4) \\ l_4 \sin(180^\circ - \theta_4) \\ 0 \end{pmatrix} \tag{A.8}$$

$$\mathbf{O}_{i-1} = \begin{pmatrix} l_3 + l_4 \cos(180^\circ - \theta_4) + l_2 \cos(\theta_3 + \theta_4) \\ l_4 \sin(180^\circ - \theta_4) - l_2 \sin(\theta_3 + \theta_4) \\ 0 \end{pmatrix} \quad (\text{A.9})$$

## A.5 Translated positions of the carbonyl carbon and oxygen atoms.

The atomic positions given in the previous section assume  $\phi$  and  $\psi$  angles of  $0^\circ$ . The next step in the calculation is applying rotational transformations for  $\phi$  and  $\psi$  to the  $\mathbf{C}_{i-1}$  and  $\mathbf{O}_{i-1}$  positions and the  $\mathbf{C}_i$  and  $\mathbf{O}_i$  positions, respectively.

## A.6 Translated $\mathbf{C}_i$ and $\mathbf{O}_i$ postitions.

Transformation of the  $\mathbf{C}_i$  and  $\mathbf{O}_i$  postitions with regard to  $\psi$  requires the application of the x-axis rotational transformation matrix given in equation A.2.

let

$$\mathbf{R}_\psi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \psi & -\sin \psi \\ 0 & \sin \psi & \cos \psi \end{pmatrix} \quad (\text{A.10})$$

then

$$\mathbf{C}'_i = \mathbf{R}_\psi \mathbf{C}_i \quad (\text{A.11})$$

and

$$\mathbf{O}'_i = \mathbf{R}_\psi \mathbf{O}_i \quad (\text{A.12})$$

## A.7 Translated $C_{i-1}$ and $O_{i-1}$ positions.

The  $C_{i-1}$  and  $O_{i-1}$  positions must be transformed not only by the  $\phi$  rotation about the local x-axis, but must also be transformed from their local coordinate system to the same frame of reference as the  $C_i$  and  $O_i$ . This second transformation is accomplished by a rotation about the z-axis by the angle of  $\angle NC^\alpha C$  ( $\theta_2$ ). To arrive at the final coordinates for  $C_i$  and  $O_i$  first the transformation for  $\phi$  is applied and then the transformation for  $\theta_2$  is applied.

let

$$\mathbf{R}_\phi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \phi & \sin \phi \\ 0 & -\sin \phi & \cos \phi \end{pmatrix} \quad (\text{A.13})$$

and

$$\mathbf{R}_{\theta_2} = \begin{pmatrix} \cos \theta_2 & -\sin \theta_2 & 0 \\ \sin \theta_2 & \cos \theta_2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (\text{A.14})$$

then

$$\mathbf{C}'_{i-1} = \mathbf{R}_{\theta_2} \mathbf{R}_\phi \mathbf{C}_{i-1} \quad (\text{A.15})$$

and

$$\mathbf{O}'_{i-1} = \mathbf{R}_{\theta_2} \mathbf{R}_\phi \mathbf{O}_{i-1} \quad (\text{A.16})$$

## A.8 Computation of the oo1 dihedral angle.

The last step in completing the relationship is computing the oo1 angle as the dihedral angle formed by the  $O'_{i-1}$ ,  $C'_{i-1}$ ,  $C'_i$ , and  $O'_i$  atoms.

$$\mathbf{oo1}_i = \text{dihedral}(\mathbf{O}'_{i-1}, \mathbf{C}'_{i-1}, \mathbf{C}'_i, \mathbf{O}'_i) \quad (\text{A.17})$$

The oo1 dihedral angle is the angle between the vector normal the the plane defined by  $O'_{i-1}$ ,  $C'_{i-1}$ , and  $C'_i$ , and the vector normal to the plane defined by  $C'_{i-1}$ ,  $C'_i$ , and  $O'_i$ .

Let three vectors be defined as follows:

$$\begin{aligned} \vec{\mathbf{v}}_1 &= \mathbf{C}'_{i-1} - \mathbf{O}'_{i-1} \\ \vec{\mathbf{v}}_2 &= \mathbf{C}'_{i-1} - \mathbf{C}'_i \\ \vec{\mathbf{v}}_3 &= \mathbf{O}'_i - \mathbf{C}'_i \end{aligned} \quad (\text{A.18})$$

then

$$\vec{\mathbf{n}}_1 = \vec{\mathbf{v}}_1 \times \vec{\mathbf{v}}_2 \quad (\text{A.19})$$

and

$$\vec{\mathbf{n}}_2 = \vec{\mathbf{v}}_2 \times \vec{\mathbf{v}}_3 \quad (\text{A.20})$$

$$\mathbf{oo1} = \arccos \left( \frac{\vec{\mathbf{n}}_1 \cdot \vec{\mathbf{n}}_2}{\|\vec{\mathbf{n}}_1\| \|\vec{\mathbf{n}}_2\|} \right) \quad (\text{A.21})$$

# Appendix B

## List of protein structures used

No pairs of chains in the following section have more than 25% identical residues for sequences longer than 80 residues (higher for shorter sequences according to the function derived in Schneider and Sander (Schneider and Sander, 1991)).

- Id — PDB identifier (fifth letter is chain identifier)
- len — number of amino acid residues
- $\alpha$  — number of residues in  $\alpha$ -helical conformation (DSSP-assignment)
- $\beta$  — number of residues in beta conformation (DSSP assignment)
- res — resolution in angstroms
- Rfac — R-factor
- M — Method (X for X-ray structure, N for NMR structure)
- compound — compound description

Table B.1: List of protein structures used

Id	len	$\alpha$	$\beta$	res	Rfac	M	compound
125D	43	8	0	-1.00	-1.00	N	CD2-gal4 65-residue DNA-binding
1AAF	55	6	0	-1.00	-1.00	N	HIV-1 Nucleocapsid Protein
1AAJ	105	3	47	1.80	0.15	X	Amicyanin (Apo Form)
1AAK	151	49	31	2.40	0.22	X	Ubiquitin Conjugating Enzyme
1AB2	109	17	17	-1.00	-1.00	N	Proto-oncogene Tyrosine Kinase
1ABK	211	129	0	2.00	0.18	X	Endonuclease III (Ace Reg 60184)
1ABRB	267	33	98	2.14	0.19	X	Abrin-a Complexed With Two Sugar
1ADD	349	179	51	2.40	0.18	X	Adenosine Deaminase Complexed
1ADS	315	121	47	1.60	0.20	X	Aldose Reductase Complex w Nadph
1AEP	153	122	0	2.70	0.21	X	Apolipoprotein III
1AFP	51	0	15	-1.00	-1.00	N	Antifungal Protein fr Aspergillus
1AGX	331	96	66	2.90	0.17	X	Glutaminase-asparaginase
1ALKA	449	142	94	2.00	0.18	X	Alkaline Phosphatase
1AMG	417	145	65	2.20	0.17	X	1,4-alpha-d-glucan Maltotetra
1AMP	291	115	50	1.80	0.16	X	Aminopeptidase Aeromonas Proteol
1AORA	605	272	86	2.30	0.15	X	Aldehyde Ferredoxin Oxidoreduct
1AOZA	552	55	210	1.90	0.20	X	Ascorbate Oxidase
1APLC	59	38	0	2.70	0.23	X	MAT Alpha2 Homeodomain Complexed
1APME	341	130	53	2.00	0.19	X	c-AMP-dependent Protein Kinase)
1ARB	263	41	85	1.20	0.15	X	Achromobacter Protease I
1ARS	396	181	57	1.80	0.21	X	Aspartate Aminotransferase
1ASH	147	114	0	2.15	0.18	X	Hemoglobin (Domain One)
1ASZA	490	145	118	3.00	0.20	X	Aspartyl TRNA Synthetase (Asprs)
1ATNA	372	153	77	2.80	0.21	X	Deoxyribonuclease I Complex
1ATY	46	32	0	-1.00	-1.00	N	F1F0 ATP Synthase Subunit C
1BAA	243	98	6	2.80	0.20	X	Endochitinase (26 Kd)
1BABB	146	118	0	1.50	0.16	X	Hemoglobin Thionville
1BAM	200	80	49	1.95	0.19	X	Restriction Endonuclease BamHI
1BARA	127	15	52	2.70	0.18	X	Acidic Fibroblast Growth Factor
1BBL	37	17	0	-1.00	-1.00	N	Dihydrolipoamide Succinyltransf
1BBPA	173	23	83	2.00	0.20	X	Bilin Binding Protein (BBP)
1BBT1	186	27	55	2.60	0.17	X	Foot-and-mouth Disease Virus
1BBT2	210	22	94	2.60	0.17	X	Foot-and-mouth Disease Virus
1BCFA	158	124	0	2.90	0.21	X	Bacterioferritin (Cytochrome B1)
1BET	107	0	65	2.30	0.22	X	Beta-nerve Growth Factor
1BGEB	159	113	2	2.20	0.19	X	Granulocyte Colony-stimulating
1BIP	122	40	6	-1.00	-1.00	N	Bifunctional Trypsin
1BMDA	327	149	62	1.90	0.15	X	Malate Dehydrogenase (Bacterial)
<i>continued on next page</i>							



<i>continued from previous page</i>							
Id	len	$\alpha$	$\beta$	res	Rfac	M	compound
1BPB	242	105	42	2.30	0.19	X	DNA Polymerase Beta (Polymerase)
1BSAA	107	24	25	2.00	0.17	X	Barnase (G Endonuclease)
1BVP1	349	119	64	2.60	0.18	X	Bluetongue Virus 10 (Usa) Vp7
1BW4	125	24	40	-1.00	-1.00	N	Barwin, Barley Seed Protein
1C5A	66	46	0	-1.00	-1.00	N	Des-ARG74-complement C5a
1CAUA	181	20	77	2.30	0.19	X	Canavalin (Jack Bean Vicilin)
1CAUB	184	25	71	2.30	0.19	X	Canavalin (Jack Bean Vicilin)
1CBN	46	16	5	0.83	0.11	X	Crambin
1CCF	42	0	10	-1.00	-1.00	N	Coagulation Factor X Calcium
1CCR	111	47	2	1.50	0.19	X	Cytochrome c
1CDTA	60	0	27	2.50	0.20	X	Cardiotoxin (Toxin III)
1CELA	434	57	159	1.81	0.18	X	1,4-beta-d-glucan Cellobiohyd
1CEWI	108	22	52	2.00	0.20	X	Cystatin
1CFB	205	3	91	2.00	0.20	X	Drosophila Neuroglian
1CGT	684	156	209	2.00	0.17	X	Cyclodextrin Glycosyltransf
1CHL	36	8	7	-1.00	-1.00	N	Chlorotoxin (NMR, 7 Structures)
1CHMA	401	150	95	1.90	0.18	X	Creatine Amidinohydrolase
1CHRA	370	135	55	3.00	0.20	X	Chloromuconate Cycloisomerase
1CID	177	3	100	2.80	0.23	X	CD4 (Domains 3 And 4)
1CKSA	74	12	9	2.10	0.17	X	Cyclin-dependent Kinase Subunit
1CLC	541	202	79	1.90	0.20	X	Molecule
1CMCA	105	43	15	1.80	0.20	X	Met Holo-repressor
1COLA	197	155	0	2.40	0.18	X	Colicin A (C-terminal Domain)
1CPCA	162	126	0	1.66	0.18	X	C-phycocyanin
1CPCB	172	128	0	1.66	0.18	X	C-phycocyanin
1CPT	412	215	48	2.30	0.19	X	Cytochrome P450-terp
1CRL	534	199	82	2.06	0.13	X	Lipase (Triacylglycerol Hydro)
1CSEI	63	14	21	1.20	0.18	X	Subtilisin Carlsberg
1CSKA	58	0	27	2.50	0.22	X	C-src Kinase (SH3 Domain)
1CSP	67	3	37	2.50	0.20	X	Major Cold Shock Protein (Cspb)
1CTAA	34	18	1	-1.00	-1.00	N	Troponin C Site III - Site III
1CTL	85	0	0	-1.00	-1.00	N	Molecule
1CTM	250	27	101	2.30	0.20	X	Cytochrome F (Reduced)
1CTN	538	147	148	2.30	0.16	X	Chitinase A (PH 5.5, 4 Deg C)
1CTT	294	107	57	2.20	0.19	X	Cytidine Deaminase Complex
1CUS	197	87	30	1.25	0.16	X	Cutinase
1DDT	523	190	140	2.00	0.20	X	Diphtheria Toxin (Dimeric)
<i>continued on next page</i>							

<i>continued from previous page</i>							
Id	len	$\alpha$	$\beta$	res	Rfac	M	compound
1DEC	39	0	9	-1.00	-1.00	N	Decorsin (Nmr, 25 Structures)
1DFNA	30	0	16	1.90	0.19	X	Defensin HNP-3
1DHR	236	95	58	2.30	0.15	X	Dihydropteridine Reductase
1DLC	584	197	176	2.50	0.18	X	Delta-endotoxin Cryiia (Bt13)
1DMC	31	7	0	-1.00	-1.00	N	CD-6 Metallothionein-1 (CD-6 Mt)
1DSBA	188	98	21	2.00	0.17	X	Dsba (Disulfide Bond Formation)
1DTS	220	87	45	1.65	0.17	X	Dethiobiotin Synthase
1DYNA	113	16	47	2.20	0.20	X	Dynamin (Pleckstrin Homology)
1ECA	136	104	0	1.40	0.18	X	Hemoglobin (Erythrocrurin)
1EDE	310	129	44	1.90	0.16	X	Haloalkane Dehalogenase
1EFT	405	90	138	2.50	0.20	X	Elongation Factor Tu (EF-TU)
1ENH	54	38	0	2.10	0.20	X	Engrailed Homeodomain
1EPAB	164	30	71	2.10	0.20	X	Epididymal Retinoic Acid-binding
1EPTA	43	3	21	1.80	0.18	X	Porcine E-Trypsin
1ERD	40	24	0	-1.00	-1.00	N	Pheromone ER-2
1ERIA	261	88	53	2.70	0.17	X	Eco Ri Endonuclease
1ERL	40	26	0	1.59	0.20	X	Mating Pheromone ER-1
1ERP	38	22	0	-1.00	-1.00	N	Pheromone ER-10
1ETB1	118	7	61	1.70	0.16	X	Transthyretin (Prealbumin)
1FBAA	360	160	54	1.90	0.18	X	Fructose-1,6-bisphosphate
1FCA	55	7	10	1.80	0.14	X	Ferredoxin
1FCDA	401	114	103	2.53	0.24	X	Flavocytochrome C Sulfide Dehydro
1FCDC	174	93	103	2.53	0.24	X	Flavocytochrome C Sulfide Dehydro
1FCT	32	13	0	-1.00	-1.00	N	Ferredoxin Chloroplastic Transit
1FHA	172	125	0	2.40	0.20	X	Ferritin (H-chain) Mutant
1FIAB	78	52	2	2.00	0.19	X	Fis Protein (Inversion Factor)
1FKF	107	13	43	1.70	0.17	X	FK506 Binding Protein (FKBP)
1FNC	296	74	92	1.70	0.15	X	Ferredoxin
1FOD4	47	12	1	2.60	0.21	X	Foot-and-mouth Disease Virus
1FRPA	321	115	82	2.00	0.19	X	Fructose-1,6-bisphosphatase
1FRUB	99	73	49	2.20	0.21	X	Fc (IGG) Receptor (Neonatal)
1FXRA	64	15	16	2.30	0.18	X	Ferredoxin I (4FE-4S)
1GAL	581	201	138	2.30	0.18	X	Glucose Oxidase
1GARA	205	68	48	1.96	0.17	X	Glycinamide Ribonucleotide
1GATA	60	13	8	-1.00	-1.00	N	Erythroid Transcription Factor
1GBS	185	107	7	1.80	0.18	X	Lysozyme
1GCA	309	139	60	1.70	0.19	X	GlucoseGalactose binding protein
<i>continued on next page</i>							

<i>continued from previous page</i>							
Id	len	$\alpha$	$\beta$	res	Rfac	M	compound
1GDHA	320	117	58	2.40	0.19	X	D-glycerate Dehydrogenase (Apo)
1GDM	153	118	0	1.70	0.15	X	Leghemoglobin (Oxy)
1GHSA	306	124	53	2.30	0.18	X	1,3-beta-glucanase
1GKY	186	83	42	2.00	0.17	X	Guanylate Kinase Complex
1GLCG	489	173	120	2.65	0.17	X	Glycerol Kinase Complexed
1GLT	296	105	88	2.00	0.19	X	Glutathione Synthase
1GMFA	119	64	10	2.40	0.20	X	Granulocyte-macrophage
1GOF	639	21	262	1.70	0.18	X	Galactose Oxidase (PH 4.5)
1GOX	350	155	48	2.00	0.19	X	Glycolate Oxidase
1GP1A	185	59	33	2.00	0.17	X	Glutathione Peroxidase
1GPB	823	415	130	1.90	0.19	X	Glycogen Phosphorylase b (T State)
1GPH1	465	139	114	3.00	0.18	X	Glutamine Phosphoribosylpyrophos
1GPR	158	17	61	1.90	0.16	X	Glucose Permease (Domain IIA)
1GPS	47	10	17	-1.00	-1.00	N	Gamma-1-P Thionin
1GRJ	151	70	37	2.20	0.21	X	Grea Transcript Cleavage Factor
1GSQ	202	120	16	2.40	0.18	X	Glutathione S-transferase (GST)
1GTRA	529	168	133	2.50	0.21	X	Glutaminyl-trna Synthetase Complex
1HAR	196	71	40	2.20	0.21	X	Hiv-1 Reverse Transcriptase
1HBQ	177	16	80	1.70	0.20	X	Retinol Binding Protein (Apo)
1HCE	118	3	62	-1.00	-1.00	N	Hisactophilin (NMR)
1HCGB	51	3	16	2.20	0.17	X	Blood Coagulation Factor Xa
1HCNA	85	4	47	2.60	0.20	X	Human Chorionic Gonadotropin
1HCNB	110	4	58	2.60	0.20	X	Human Chorionic Gonadotropin
1HCRA	52	25	1	1.80	0.23	X	Hin Recombinase (DNA-binding)
1HDCA	253	123	40	2.20	0.19	X	3-alpha, 20-beta-hydroxysteroid
1HDGO	332	103	91	2.50	0.17	X	Holo-D-glyceraldehyde-3-phosphate
1HEX	345	160	63	2.50	0.18	X	3-Isopropylmalate Dehydrogenase
1HFC	157	40	25	1.56	0.17	X	Fibroblast Collagenase
1HFI	62	0	22	-1.00	-1.00	N	Factor H, 15th C-module Pair
1HFT	197	16	97	2.40	0.22	X	Tissue Factor (Extracellular)
1HGJA	328	31	136	2.70	0.22	X	Hemagglutinin (Bromelain Digested)
1HJRA	158	66	38	2.50	0.16	X	Holliday Junction Resolvase
1HKS	106	19	6	-1.00	-1.00	N	Heat Shock Transcription Factor
1HLB	157	105	0	2.50	0.15	X	Hemoglobin (Sea Cucumber)
1HLEA	344	122	120	1.95	0.18	X	Horse Leukocyte Elastase Inhib
1HLEB	31	122	19	1.95	0.18	X	Horse Leukocyte Elastase Inhib
1HMCB	143	0	0	2.50	0.20	X	Human Macrophage Colony
<i>continued on next page</i>							

<i>continued from previous page</i>							
Id	len	$\alpha$	$\beta$	res	Rfac	M	compound
1HMPA	214	60	59	2.50	0.19	X	Hypoxanthine Guanine Phosphorib
1HMY	328	96	58	2.50	0.20	X	Hhai DNA (Cytosine-c5-)-methyl
1HNF	179	9	97	2.50	0.19	X	CD2 (Human)
1HNS	47	9	2	-1.00	-1.00	N	H-NS (DNA-binding Domain)
1HOE	74	0	36	2.00	0.20	X	Alpha-Amylase Inhibitor HOE-467A
1HPH	37	15	0	-1.00	-1.00	N	Human Parathyroid Hormone
1HRF	67	0	2	-1.00	-1.00	N	Heregulin-alpha (Epidermal GF)
1HRNB	334	54	154	1.80	0.21	X	Renin Complexed With Polyhydroxy
1HSLA	238	86	50	1.89	0.20	X	Histidine-binding Protein Complex
1HTMD	123	86	6	2.50	0.22	X	Hemagglutinin Ectodomain
1HTP	131	33	47	2.20	0.18	X	H-protein Complexed w Lipoic Acid
1HTRP	43	17	7	1.62	0.18	X	Progastricsin (Pepsinogen C)
1HUCB	205	43	52	2.10	0.16	X	Cathepsin B
1HUMA	69	15	14	-1.00	-1.00	N	Human Macrophage Inflammatory
1HUW	166	115	0	2.00	0.18	X	Human Growth Hormone Mutant
1HVD	313	227	0	2.00	0.19	X	Annexin V (Lipocortin V)
1IAE	200	65	38	1.83	0.14	X	Astacin w Zinc Replaced By Nickel
1IAG	201	74	40	2.00	0.17	X	Adamalysin II (Proteinase II)
1ICA	40	10	4	-1.00	-1.00	N	Insect Defensin A
1IFC	131	15	77	1.19	0.17	X	Intestinal Fatty Acid Binding
1IGP	175	31	55	2.20	0.22	X	Inorganic Pyrophosphatase
1ILK	151	112	0	1.80	0.16	X	Interleukin-10
1ILR1	145	15	74	2.10	0.20	X	Interleukin 1 Receptor Antag
1IRK	303	119	52	2.10	0.20	X	Insulin Receptr (Tyrosine Kinase)
1ISCA	192	98	23	1.80	0.18	X	Iron(III) Superoxide Dismutase
1ISUA	62	9	12	1.50	0.17	X	High-potential Iron-sulfur
1IVD	388	4	154	1.80	0.23	X	Influenza A Subtype N2
1KAB	136	36	42	1.80	0.19	X	Staphylococcal Nuclease Mutant
1KNB	186	9	76	1.70	0.16	X	Adenovirus Type 5 Fiber
1KNT	55	12	17	1.60	0.19	X	Collagen Type VI Kunitz-type C
1KTX	37	4	4	-1.00	-1.00	N	Kalioxin (Ktx)
1L92	162	107	15	1.70	0.15	X	Lysozyme Mutant w Cys 54 Repl
1LBA	146	40	30	2.20	0.19	X	Lysozyme Mutant w Ala 6 Repl
1LENA	181	3	86	1.80	0.17	X	Lectin (Lentil) (Monoclinic)
1LGAA	343	141	22	2.03	0.15	X	Lignin Peroxidase (LIP) (Ferric)
1LIS	131	94	0	1.90	0.19	X	Lysin
1LKI	172	116	4	2.00	0.19	X	Leukemia Inhibitory Factor (LIF)
<i>continued on next page</i>							

<i>continued from previous page</i>							
Id	len	$\alpha$	$\beta$	res	Rfac	M	compound
1LLDA	313	140	59	2.00	0.18	X	L-lactate Dehydrogenase
1LOBD	47	3	29	2.00	0.18	X	Lectin (Legume, Isolectin I)
1LPBB	449	104	128	2.46	0.18	X	Lipase Complexed With Colipase
1LPE	144	118	0	2.25	0.17	X	Apolipoprotein-E3 (LDL Receptor)
1LTSA	185	59	43	1.95	0.18	X	Heat-labile Enterotoxin (LT)
1LTSC	41	34	43	1.95	0.18	X	Heat-labile Enterotoxin (LT)
1LTSD	103	23	38	1.95	0.18	X	Heat-labile Enterotoxin (LT)
1LYP	32	30	0	-1.00	-1.00	N	Cap18 (Residues 106 - 137)
1MAT	263	70	73	2.40	0.18	X	Methionine Aminopeptidase
1MDC	132	13	71	1.75	0.17	X	Fatty Acid Binding Protein
1MINA	468	230	68	2.20	0.23	X	Nitrogenase Molybdenum-iron
1MINB	522	275	55	2.20	0.23	X	Nitrogenase Molybdenum-iron
1MMOB	384	241	2	2.20	0.17	X	Methane Monooxygenase Hydrolase
1MMOG	162	105	4	2.20	0.17	X	Methane Monooxygenase Hydrolase
1MRB	31	0	0	-1.00	-1.00	N	CD-7 Metallothionein-2a (Alpha)
1MRJ	247	99	62	1.60	0.17	X	Alpha-trichosanthin Complexed
1MSC	129	22	46	2.00	0.20	X	Bacteriophage Ms2 Unass Coat
1MSEC	105	62	0	-1.00	-1.00	N	C-myb DNA-binding Domain Complex
1MUP	157	20	73	2.40	0.19	X	Major Urinary Protein Complex
1MYLB	40	24	5	2.40	0.21	X	Arc Repressor Mutant
1MYPC	462	167	30	3.00	0.26	X	Myeloperoxidase
1NAR	289	108	67	1.80	0.16	X	Narbonin
1NBAA	253	108	38	2.00	0.19	X	N-Carbamoylsarcosine Amidohydrol
1NFP	228	104	45	1.60	0.17	X	LuxF Gene Product
1NHKL	143	64	28	1.90	0.17	X	Nucleoside Diphosphate Kinase
1NIPA	283	109	42	2.90	0.18	X	Nitrogenase Iron Protein
1NNT	328	99	69	2.30	0.16	X	Ovotransferrin (Monoferric)
1OLBA	517	175	116	1.80	0.17	X	Oligo-peptide Binding Protein
1OMA	48	0	10	-1.00	-1.00	N	Omega-aga-ivb
1OMP	370	164	74	1.80	0.21	X	D-maltodextrin-binding Protein
1OXY	573	206	111	2.40	0.17	X	Hemocyanin Subunit Ii Complex
1OYB	399	150	58	2.00	0.17	X	Old Yellow Enzyme (Oxidized)
1PAA	30	7	2	-1.00	-1.00	N	Yeast Transcription Factor Adr1
1PBE	391	160	112	1.90	0.16	X	P-hydroxybenzoate Hydroxylase
1PBP	321	127	74	1.90	0.15	X	Phosphate Transport
1PCRH	240	56	61	2.65	0.19	X	Photosynthetic Reaction Center
1PCRM	302	183	12	2.65	0.19	X	Photosynthetic Reaction Center
<i>continued on next page</i>							

<i>continued from previous page</i>							
Id	len	$\alpha$	$\beta$	res	Rfac	M	compound
1PDC	45	3	11	-1.00	-1.00	N	Seminal Fluid Protein Pdc-109
1PDGA	87	0	56	3.00	0.21	X	Platelet-derived Growth Factor Bb
1PETA	31	17	6	-1.00	-1.00	N	Tumor Antigen P53
1PFIA	46	42	0	3.00	-1.00	F	Major Coat Protein Of Pf1 Virus
1PGA	56	14	24	2.07	0.17	X	Protein G (B1 IGG-binding Domain)
1PHO	330	7	186	3.00	0.22	X	Phosphoporin (Phoe)
1PHP	394	167	67	1.65	0.16	X	3-Phosphoglycerate Kinase
1PII	452	173	91	2.00	0.17	X	N-(5'phosphoribosyl)anthranilate
1PKN	514	198	98	2.90	0.19	X	Pyruvate Kinase Complexed
1PLQ	258	48	114	2.30	0.19	X	Cell Nuclear Antigen
1PLS	113	18	39	-1.00	-1.00	N	Pleckstrin (N-term Pleckstrin)
1PMY	123	17	44	1.50	0.20	X	Pseudoazurin (Cupredoxin)
1PNT	157	69	23	2.20	0.17	X	Tyrosine Phosphatase
1POA	118	55	11	1.50	0.14	X	Phospholipase A2
1POC	134	36	28	2.00	0.19	X	Phospholipase A2 Complex
1POXA	585	250	96	2.10	0.16	X	Pyruvate Oxidase Mutant
1PPBL	36	11	0	1.92	0.16	X	Alpha-thrombin Complex
1PPI	496	140	118	2.20	0.15	X	Alpha Amylase (Ppa) Complex
1PPN	212	55	47	1.60	0.16	X	Papain Cys-25 With Bound Atom
1PPT	36	18	0	1.37	0.28	X	Avian Pancreatic Polypeptide
1PRCC	333	138	10	2.30	0.19	X	Photosynthetic Reaction Center
1PRS	173	18	56	-1.00	-1.00	N	Development-specific Protein
1PRTA	224	73	60	2.90	0.20	X	Pertussis Toxin
1PRTC	196	39	73	2.90	0.20	X	Pertussis Toxin
1PRTD	110	15	61	2.90	0.20	X	Pertussis Toxin
1PRTF	98	16	45	2.90	0.20	X	Pertussis Toxin
1PSM	38	9	0	-1.00	-1.00	N	Spam-h1 (Residues 90 - 127)
1PSPA	106	25	10	2.50	0.20	X	Pancreatic Spasmolytic
1PTX	64	10	25	1.30	0.15	X	Scorpion Toxin II
1PXTB	348	138	78	2.80	0.20	X	Peroxisomal 3-ketoacyl-COA
1PYAB	228	62	82	2.50	0.15	X	Pyruvoyl-dependent Histidine
1PYDA	537	221	97	2.40	0.20	X	Pyruvate Decarboxylase (Pdc)
1PYP	281	42	36	3.00	-1.00	X	Inorganic Pyrophosphatase
1QORA	326	113	82	2.20	0.14	X	Quinone Oxidoreductase Complex
1RBLM	109	28	26	2.20	0.20	X	Ribulose 1,5 Bisphosphate Carboxy
1RCB	129	81	6	2.25	0.22	X	Interleukin 4
1REC	185	121	10	1.90	0.19	X	Recoverin (Calcium Sensor In Vision)
<i>continued on next page</i>							

<i>continued from previous page</i>							
Id	len	$\alpha$	$\beta$	res	Rfac	M	compound
1RET	43	25	0	-1.00	-1.00	N	Gamma Delta Resolvase (DNA Binding)
1RGD	71	23	0	-1.00	-1.00	N	Glucocorticoid Receptor
1RIBA	340	235	14	2.20	0.17	X	Protein R2 Of Ribonucleotide
1ROPA	56	50	0	1.70	0.18	X	Rop
1RPA	342	155	41	3.00	0.21	X	Prostatic Acid Phosphatase Complex
1RSY	135	13	61	1.90	0.19	X	Synaptotagmin I (First C2 Domain)
1RTM1	149	53	38	1.80	0.22	X	Mannose-binding Protein A
1RTP1	109	65	4	2.00	0.18	X	Alpha-parvalbumin
1RVAA	244	82	70	2.00	0.16	X	Eco RV Endonuclease Complex
1S01	275	82	54	1.70	0.15	X	Subtilisin BPN(Prime) 8350
1SACA	204	11	94	2.00	0.18	X	Serum Amyloid P Component (SAP)
1SCMA	60	53	0	2.80	0.20	X	Myosin (Regulatory Domain)
1SCMC	149	70	10	2.80	0.20	X	Myosin (Regulatory Domain)
1SCUA	288	107	65	2.50	0.22	X	Succinyl-coa Synthetase
1SCUB	388	162	92	2.50	0.22	X	Succinyl-coa Synthetase
1SCY	31	10	8	-1.00	-1.00	N	Scyllatoxin (Leiurotoxin I)
1SLTA	133	0	82	1.90	0.17	X	S-Lectin (A Vertebrate 14 KDa)
1SPF	35	25	0	-1.00	-1.00	N	Pulmonary Surfactant-associated
1SRGA	116	6	69	1.80	0.17	X	Streptavidin Complexed
1SRYA	421	179	92	2.50	0.18	X	Seryl-tRNA Synthetase
1STO	208	88	43	2.60	0.18	X	Orotate Phosphoribosyltransfer
1TAB1	36	21	7	2.30	0.20	X	Trypsin Complex Bowman-Birk Inhib
1TADC	318	163	43	1.70	0.21	X	Transducin-alpha (Gt-alpha-gdp-alf)
1TAP	60	8	13	-1.00	-1.00	N	Factor Xa Inhibitor
1TCA	317	119	48	1.55	0.16	X	Lipase (Triacylglycerol Hydrolase)
1TFI	50	0	17	-1.00	-1.00	N	Transcriptional Elongation Factor
1TGSI	56	9	11	1.80	0.19	X	Trypsinogen Complex With Porcine
1THTA	294	120	59	2.10	0.23	X	Thioesterase
1THV	207	25	81	1.75	0.17	X	Thaumatococcus Isoform A (Orthorhombic)
1TIB	269	90	63	1.84	0.19	X	Lipase (Triacylglycerol Acylhydrol)
1TIE	166	7	69	2.50	0.21	X	Erythrina Trypsin Inhibitor
1TLCA	265	93	68	2.10	0.02	X	Thymidylate Synthase Complexed
1TLK	103	3	61	2.80	0.18	X	Telokin
1TML	286	110	37	1.80	0.18	X	Endo-1,4-beta-d-glucanase
1TNRA	144	7	82	2.85	0.16	X	Tumor Necrosis Factor Receptor
1TNRR	139	3	61	2.85	0.16	X	Tumor Necrosis Factor Receptor
1TPH1	245	110	41	1.80	0.18	X	Triosephosphate Isomerase Complex
<i>continued on next page</i>							

<i>continued from previous page</i>							
Id	len	$\alpha$	$\beta$	res	Rfac	M	compound
1TPLA	426	182	69	2.30	0.16	X	Tyrosine Phenol-lyase
1TPN	50	0	18	-1.00	-1.00	N	Tissue-type Plasminogen Activator
1TRB	316	91	89	2.00	0.18	X	Thioredoxin Reductase Mutant
1TRKA	678	303	88	2.00	0.16	X	Transketolase
1TRZB	30	14	3	1.60	0.17	X	Insulin (T3R3) Complex w Two Zinc
1TSSA	194	35	86	2.50	0.23	X	Toxic Shock Syndrome Toxin 1
1TVT	75	4	0	-1.00	-1.00	N	Transactivator Protein
1ULA	289	87	58	2.75	0.20	X	Purine Nucleoside Phosphorylase
1URK	130	9	15	-1.00	-1.00	N	Plasminogen Activator (Urokinase)
1VAAA	274	76	109	2.30	0.17	X	Mhc Class I H-2Kb Complexed
1VIL	126	24	31	-1.00	-1.00	N	Villin (Domain One Residues)
1WAS	146	104	0	2.70	0.19	X	Bacterial Aspartate Receptor
1WFBA	37	34	0	1.50	0.18	X	Antifreeze Protein Isoform Hplc6
1WHTA	256	96	53	2.00	0.18	X	Serine Carboxypeptidase II
1WHTB	153	56	34	2.00	0.18	X	Serine Carboxypeptidase II
1WSYA	248	126	33	2.50	0.25	X	Tryptophan Synthase
1WSYB	385	175	69	2.50	0.25	X	Tryptophan Synthase
1XNB	185	10	115	1.49	0.17	X	Xylanase (Endo-1,4-beta-xylanase)
1XSOA	150	9	63	1.49	0.10	X	Cu, Zn Superoxide Dismutase
1YPTB	280	105	58	2.50	0.17	X	Protein-tyrosine Phosphatase
1YTBA	180	49	61	1.80	0.20	X	Tata-box Binding Protein (Ytbp)
1ZAAC	85	35	12	2.10	0.18	X	Zif268 Immediate Early Gene
2ACG	125	43	41	2.50	0.18	X	Profilin II
2ACHB	35	112	14	2.70	0.18	X	Alpha1 Antichymotrypsin
2AK3B	221	103	38	1.85	0.19	X	Adenylate Kinase Isoenzyme-3
2ALP	198	14	109	1.70	0.13	X	Alpha-lytic Protease
2ATCB	152	9	7	3.00	0.27	X	Aspartate Carbamoyltransferase
2AYH	214	13	111	1.60	0.14	X	1,3-1,4-beta-D-Glucan 4 Glucanohyd
2AZAA	129	21	46	1.80	0.16	X	Azurin (Oxidized)
2BBKH	355	11	178	1.75	0.17	X	Methylamine Dehydrogenase (Madh)
2BBVC	321	47	119	2.80	0.22	X	Black Beetle Virus Capsid Protein
2BDS	43	0	19	-1.00	-1.00	N	BDS-I
2BOPA	85	27	30	1.70	0.20	X	Bovine Papillomavirus-1 E2
2BPA1	426	109	123	3.00	0.21	X	Bacteriophage PhiX174 Capsid
2BPA2	175	109	91	3.00	0.21	X	Bacteriophage PhiX174 Capsid
2BPA3	36	109	1	3.00	0.21	X	Bacteriophage PhiX174 Capsid
2BTFP	139	38	43	2.55	0.20	X	Beta-actin-profilin Complex
<i>continued on next page</i>							



<i>continued from previous page</i>							
Id	len	$\alpha$	$\beta$	res	Rfac	M	compound
2CAS	548	51	163	3.00	0.21	X	Canine Parvovirus Empty Capsid
2CBA	258	42	81	1.54	0.15	X	Carbonic Anhydrase II
2CBH	36	0	14	-1.00	-1.00	N	C-terminal Domain Of Cellobiohyd
2CDV	107	30	12	1.80	0.18	X	Cytochrome C3
2CHSA	114	35	33	1.90	0.19	X	Chorismate Mutase
2CND	260	61	92	2.50	0.19	X	Nadh-dependent Nitrate Reductase
2CP4	405	208	42	2.10	0.18	X	Cytochrome P450cam (Camphor)
2CPL	164	23	53	1.63	0.18	X	Cyclophilin A
2CRD	37	10	7	-1.00	-1.00	N	Charybdotoxin
2CRO	65	40	0	2.35	0.20	X	434 Cro Protein
2CTC	307	116	54	1.40	0.16	X	Carboxypeptidase A Complex
2DKB	431	174	80	2.10	0.18	X	2,2-Dialkylglycine Decarboxylase
2DNJA	253	76	75	2.00	0.17	X	Deoxyribonuclease I (DNase I)
2DRI	271	122	61	1.60	0.19	X	D-Ribose-binding Protein Complex
2DRPA	63	22	13	2.80	0.19	X	Tramtrack Protein (2 Zinc-fingers)
2EBN	285	87	70	2.00	0.16	X	Endo-beta-n-acetylglucosaminid
2ECH	49	0	6	-1.00	-1.00	N	Echistatin
2END	137	70	4	1.45	0.16	X	Endonuclease V
2ER7E	330	37	152	1.60	0.14	X	Endothia Aspartic Proteinase
2FCR	173	65	35	1.80	0.17	X	Flavodoxin
2GSTA	217	108	20	1.80	0.16	X	Glutathione S-transferase
2HBG	147	114	0	1.50	0.13	X	Hemoglobin (Deoxy)
2HHMA	272	90	60	2.10	0.17	X	Human Inositol Monophosphatase
2HIPA	71	10	18	2.50	0.18	X	High Potential Iron Sulfur
2HMZA	113	77	1	1.66	0.18	X	Hemerythrin (Adizomet)
2HNQ	278	98	58	2.80	0.20	X	Protein-tyrosine Phosphatase 1b
2HNTE	67	7	21	2.50	0.16	X	Gamma-thrombin
2HPDA	457	244	52	2.00	0.17	X	Cytochrome P450 (BM-3)
2HSP	71	0	11	-1.00	-1.00	N	Phospholipase C-gamma (SH3 Domain)
2IHL	129	52	14	1.40	0.17	X	Lysozyme (Japanese Quail)
2KAIB	152	18	43	2.50	0.22	X	Kallikrein A Complex With Bovine
2KAUB	101	3	38	2.00	0.18	X	Molecule
2KAUC	566	169	130	2.00	0.18	X	Molecule
2LGSA	445	166	103	2.80	0.23	X	Glutamine Synthetase Complexed
2LIV	344	153	68	2.40	0.18	X	Leucine/Isoleucine
2MADL	124	12	36	2.25	0.21	X	Methylamine Dehydrogenase (Madh)
2MEV1	268	19	68	3.00	0.22	X	Mengo Encephalomyocarditis Virus
<i>continued on next page</i>							

<i>continued from previous page</i>							
Id	len	$\alpha$	$\beta$	res	Rfac	M	compound
2MGE	154	122	0	1.70	0.16	X	Myoglobin (Met) Mutant
2MHU	30	3	0	-1.00	-1.00	N	CD-7 Metallothionein-2
2MNR	357	149	78	1.90	0.16	X	Mandelate Racemase
2MTAC	147	62	2	2.40	0.18	X	Methylamine Dehydrogenase Complex
2OHXA	374	107	96	1.80	0.17	X	Alcohol Dehydrogenase (Holo Form)
2PAC	82	26	0	-1.00	-1.00	N	Cytochrome C551
2PCDA	200	27	62	2.15	0.17	X	Protocatechuate 3,4-dioxygenase
2PDE	43	5	0	-1.00	-1.00	N	Dihydrolipoamide Acetyltransferase
2PF1	121	17	15	2.20	0.17	X	Prothrombin Fragment 1
2PFKD	305	135	61	2.40	0.17	X	Phosphofructokinase
2PGD	473	266	44	2.00	0.20	X	6-Phosphogluconate Dehydrogenase
2PIA	321	54	97	2.00	0.19	X	Phthalate Dioxygenase Reductase
2PMGA	561	179	129	2.70	0.22	X	Phosphoglucomutase
2POR	301	20	177	1.80	0.19	X	Porin (Crystal Form B)
2REB	303	133	76	2.30	0.21	X	Rec A Protein
2RN2	155	54	47	1.48	0.20	X	Ribonuclease H
2RSLB	120	55	22	2.30	0.20	X	Gamma Delta Resolvase
2RSPB	113	7	47	2.00	0.14	X	Rous Sarcoma Virus Protease
2SAS	185	117	4	2.40	0.20	X	Sarcoplasmic Calcium-binding
2SCPA	174	111	8	2.00	0.18	X	Sarcoplasmic Calcium-binding
2SH1	48	0	25	-1.00	-1.00	N	Neurotoxin I (SH I)
2SIL	381	22	175	1.60	0.17	X	Sialidase (Neuraminidase)
2SN3	65	8	16	1.20	0.19	X	Scorpion Neurotoxin (Variant 3)
2SNV	151	6	76	2.80	0.20	X	Sindbis Virus Capsid Protein
2STV	184	21	88	2.50	-1.00	X	Satellite Tobacco Necrosis Virus
2TBVA	287	7	111	2.90	0.20	X	Tomato Bushy Stunt Virus
2TGI	112	24	46	1.80	0.17	X	Transforming Growth Factor-beta
2TMDA	729	236	127	2.40	0.15	X	Trimethylamine Dehydrogenase
2TMVP	154	69	7	2.90	0.10	F	Intact Tobacco Mosaic Virus
2TPRA	482	162	132	2.40	0.18	X	Trypanothione Reductase
2TS1	317	172	32	2.30	0.23	X	Tyrosyl-transfer RNA Synthetase
2ZTAA	31	29	0	1.80	0.18	X	GCN4 Leucine Zipper
3AAHA	571	54	215	2.40	0.20	X	Methanol Dehydrogenase (MEDH)
3AAHB	57	22	2	2.40	0.20	X	Methanol Dehydrogenase (MEDH)
3CD4	178	12	86	2.20	0.20	X	CD4 (N-terminal Frag)
3CHY	128	58	22	1.66	0.15	X	CheY
3CLA	213	63	61	1.75	0.16	X	Type III Chloramphenicol Acetyl
<i>continued on next page</i>							

<i>continued from previous page</i>							
Id	len	$\alpha$	$\beta$	res	Rfac	M	compound
3COX	500	134	114	1.80	0.16	X	Cholesterol Oxidase
3DFR	163	40	53	1.70	0.15	X	Dihydrofolate Reductase Complex
3EBX	62	0	27	1.40	0.14	X	Erabutoxin b
3EGF	53	0	11	-1.00	-1.00	N	Epidermal Growth Factor (Egf)
3GAPB	205	71	38	2.50	0.21	X	Catabolite Gene Activator
3GLY	470	227	32	2.20	0.14	X	Glucoamylase-471
3HHRC	196	10	92	2.80	0.22	X	Human Growth Hormone Complex
3HSC	382	154	108	1.90	0.21	X	Heat-shock Cognate 70kd Protein
3IL8	68	18	19	2.00	0.19	X	Interleukin 8
3MDDA	385	216	63	2.40	0.17	X	Medium Chain Acyl-coa Dehydrogen
3MONA	44	0	27	2.80	0.19	X	Monellin
3SDHA	145	120	0	1.40	0.16	X	Hemoglobin I (Homodimer)
3SGBI	50	10	12	1.80	0.12	X	Proteinase B From Streptomyces
4BLMA	256	111	48	2.00	0.15	X	Beta-Lactamase (Penicillinase)
4CPAI	37	116	7	2.50	0.20	X	Carboxypeptidase A-alpha (Cox)
4ENL	436	197	74	1.90	0.15	X	Enolase (2-Phospho-D-glycerate
4FXN	138	50	31	1.80	0.20	X	Flavodoxin (Semiquinone Form)
4GCR	174	16	84	1.47	0.18	X	Gamma-B Crystallin (Previously
4RHV1	273	31	88	3.00	0.16	X	Rhinovirus 14 (HRV14)
4RHV3	236	21	75	3.00	0.16	X	Rhinovirus 14 (HRV14)
4RHV4	40	7	1	3.00	0.16	X	Rhinovirus 14 (HRV14)
4SBVA	199	30	73	2.80	0.25	X	Southern Bean Mosaic Virus Coat
4SGBI	51	12	15	2.10	0.14	X	Serine Proteinase B Complex
4TGF	50	0	14	-1.00	-1.00	N	Des-val1,Val2-transforming
4XIAA	393	188	41	2.30	0.16	X	D-Xylose Isomerase , D-Sorbitol
4ZNF	30	9	2	-1.00	-1.00	N	Zinc Finger (NMR)
5P21	166	60	44	1.35	0.20	X	c-H-ras p21 Protein
5RUBA	436	182	82	1.70	0.18	X	Rubisco (Ribulose-1,5-bisphos)
5ZNF	30	11	4	-1.00	-1.00	N	Zinc-finger (Zfy-6t)
6FABL	214	12	105	1.90	0.21	X	Antigen-binding Fragment
6TAA	476	157	99	2.10	0.20	X	Alpha Amylase (Taka Amylase)
7APIB	36	103	18	3.00	0.19	X	Modified Alpha1-Antitrypsin
7CCP	291	149	20	2.20	0.16	X	Cytochrome C Peroxidase Mutan
7PTI	58	12	15	1.60	0.17	X	Bovine Pancreatic Trypsin Inhib
7RSA	124	26	44	1.26	0.15	X	Ribonuclease A (Phosphate-free)
8ABP	305	145	67	1.49	0.17	X	L-Arabinose-binding Protein
8ACN	753	264	155	2.00	0.16	X	Aconitase Complex w Nitroisocit
<i>continued on next page</i>							

<i>continued from previous page</i>							
Id	len	$\alpha$	$\beta$	res	Rfac	M	compound
8ATCA	310	126	48	2.50	0.17	X	Aspartate Carbamoyltransferase
8CATA	498	162	84	2.50	0.19	X	Catalase
8FABB	215	19	106	1.80	0.17	X	Fab Fragment Human Immunoglob
8RXNA	52	9	12	1.00	0.15	X	Rubredoxin
8TLNE	316	131	54	1.60	0.17	X	Thermolysin Complexed w Val-Lys
9RNT	104	16	30	1.50	0.14	X	Ribonuclease T1 Complex
9WGAA	171	41	28	1.80	0.17	X	Wheat Germ Agglutinin

# Bibliography

- Altschul, S. F., Boguski, M. S., Gish, W., and Wootton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genetics*, 6:119–129.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410.
- Artymiuk, P., Mitchell, E., Rice, D., and Willett, P. (1989). Searching techniques for databases of protein structures. *Journal of Information Science*, 15:287–298.
- Asimov, I. (1960). *The intelligent man's guide to the biological sciences*. Basic Books, Inc.
- Bachar, O., Fischer, D., Nussinov, R., and Wolfson, H. (1993). A computer vision based technique for 3d sequence independent structural comparison of proteins. *Protein Engineering*, 6:279–288.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Jr., E. F. M., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112:535–542.
- Cantor, C. and Schimmel, P. (1980). *Techniques for the study of biological structure and function, Biophysical chemistry part 2*. W. H. Freeman and Company.
- Chow, E. T., Hunkapiller, T., Peterson, J. C., Zimmerman, B. A., and Waterman, M. S. (1991). A systolic array processor for biological information signal processing. *Proceedings of the International Conference on Supercomputing*, pages 216–223.
- Compugen (1995). *Biocellulator User Guide*. Compugen Ltd., Hamacabim St. 17, Petah-Tikva, 49220 Israel.
- Devereux, J., Haerberli, P., and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the vax. *Nucleic Acids Research*, 12:387–395.

- Furey, Jr, W., Wang, B. C., Yoo, C. S., and Sax, M. (1983). Structure of a novel bence-jones protein (rhe) fragment at 1.6 angstroms resolution. *Journal of Molecular Biology*, 167:661.
- Godzik, A., Skolnik, J., and Kolinski, A. (1993). Regularities in interaction patterns of globular proteins. *Protein Engineering*, pages 801–810.
- Gokhale, M., Holme, B., Akopser, Kunze, D., Lopresti, D., Lucas, S., and abd P. Olsen, R. (1990). Splash: A reconfigurable linear logic array. *Proceedings of the International Conference on Parallel Processing*, pages 526–532.
- Grindley, H., Artymiuk, P., Rice, D., and Willett, P. (1993). Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *Journal of Molecular Biology*, pages 707–721.
- Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89:10915–10919.
- Hobohm, U. and Sander, C. (1994). Enlarged representative set of protein structures. *Protein Science*, 3:533.
- Hobohm, U., Scharf, M., Schneider, R., and Sander, C. (1992). Selection of a representative set of structures from the brookhaven protein data bank. *Protein Science*, 1:409–417.
- Hoffman, D., Laiter, S., Singh, R. K., Vasiman, I. I., and Tropsha, A. (1995). Rapid protein structure classification using one-dimensional structure profiles on the BioSCAN parallel computer. *CABIOS*, 11(6):675–679.
- Hoffman, D. L. (1993a). A comparison of the BioSCAN algorithm on multiple architectures. Technical Report TR93-050, Department of Computer Science, University of North Carolina, Chapel Hill, NC.
- Hoffman, D. L. (1993b). Design of the BioSCAN server software. Technical Report TR93-049, Department of Computer Science, University of North Carolina, Chapel Hill, NC.
- Holm, L. and Sander, D. (1993). Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, 233:123–138.
- Hughey, R. P. and Lopresti, D. P. (1991). B-sys: A 470-processor programmable systolic array. *Proceedings of the International Conference on Parallel Processing*, pages 580–586.
- Johnson, M., Srinivasan, N., Sowdhamini, R., and Blundell, T. (1994). Knowledge-based protein modeling. *Critical Reviews in Biochemistry and Molecular Biology*, 29(1):1–64.

- Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, 358:86–89.
- Karlin, S. and Altschul, S. F. (1993). Application and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Sciences*, 90:5873–5877.
- Kearkey, S. K. (1990). An algorithm for the simultaneous superposition of a structural series. *Journal of Computational Chemistry*, 11:1187–1192.
- Laiter, S., Hoffman, D. L., Singh, R. K., Vasiman, I. I., and Tropsha, A. (1995). Pseudotorsional occo backbone angle as a single descriptor of protein secondary structure. *Protein Science*, 4(8):1633–1643.
- Leahy, D. J., Axel, R., and Hendrickson, W. A. (1992). Crystal structure of a soluble form of the human T cell co-receptor cd8 at 2.6 angstroms resolution. *Cell*, 68:1145.
- Levitt, M. and Warshel, A. (1975). Computer simulation of protein folding. *Nature*, 253:694–698.
- Lipman, D. and Pearson, W. (1985). Rapid and sensitive protein similarity searches. *Science*, 227:1435–1440.
- Lopresti, D. (1987). P-nac: A systolic array for comparing nucleic acid sequences. *Computer*, 20(7):98–99.
- Moore, R. E. (1975). *Mathematical elements of scientific computing*. Holt, Rinehart and Winston, Inc.
- Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453.
- Nussinov, R. and Wolfson, H. (1989). Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proceedings of the National Academy of Sciences*, 88:10495–10499.
- Oldfield, T. and Hubbard, R. (1994). Analysis of  $C_\alpha$  geometry in protein structures. *Proteins*, 18:324–337.
- Orengo, C. (1994). Classification of protein folds. *Current Opinion in Structural Biology*, 4:429–440.
- Orengo, C. A., Brown, N. P., and Taylor, W. R. (1992). Fast structure alignment for protein databank searching. *Proteins*, 14:139–167.

- Orengo, C. A., Flores, T. P., Jones, D. T., Taylor, W. R., and Thornton, J. M. (1993). Recurring structural motifs in proteins with different functions. *Current Biology*, 3:131–139.
- Panchenko, A., Luthey-Schulten, Z., and Wolynes, P. (1996). Foldons, protein structural modules, and exons. *Proceedings of the National Academy of Sciences*, 93:2008–2013.
- PDB (1996). *Protein Data Bank quarterly newsletter*. Brookhaven National Laboratory, P.O. Box 5000, Upton, NY 11973-5000 USA.
- Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85:2444–2448.
- Phillips, S. E. V. (1980). Structure and refinement of oxymyoglobin at 1.6 angstroms resolution. *Journal of Molecular Biology*, 142:531.
- Ramachandran, G. and Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Adv. Protein Chemistry*, pages 283–437.
- Rao, S., Shaffie, F., Yu, C., Satyshur, K., Stockman, B., Markley, J., and Sundaralingam, M. (1992). Structure of the oxidized long-chain flavodoxin from anabaena 7120 at 2 angstroms resolution. *Protein Science*, 1:1413.
- Richardson, J. S. and Richardson, D. C. (1989). Principles and patterns of protein conformation. In Fasman, G. D., editor, *Prediction of protein structure and the principles of protein conformation*, pages 1–98. Plenum Press.
- Russell, R. and Barten, G. J. (1993). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, 14:309–323.
- Sali, A. and Blundell, T. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *Journal of Molecular Biology*, 212:403–428.
- Scarborough, J. B. (1966). *Numerical Mathematical Analysis*. The John Hopkins Press.
- Schneider, R. and Sander, C. (1991). Selection of a representative set of structures from the brookhaven protein data bank. *Proteins*, 9(56).
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell Systems Technical Journal*, 27(379-423,623-656).
- Singh, R. K., Hoffman, D., Tell, S. G., and White, C. T. (1996). BioSCAN: A network sharable computational resource for searching biosequence databases. *CABIOS*, 12(3):191–196.



- Singh, R. K., Tell, S. G., White, C. T., Hoffman, D. L., Chi, V. L., and Erickson, B. W. (1993). A scalable systolic multiprocessor system for analysis of biological sequences. *Proceedings of the Symposium on Integrated Systems*, pages 167–182.
- Smith, T. F. and Waterman, M. S. (1981). Comparison of biosequences. *Adv. Appl. Math.*, 2:482–489.
- Smith, W. W., Burnett, R. M., Darling, G. D., and Ludwig, M. L. (1977). Structure of the semiquinone form of flavodoxin from clostridium mp. extension of 1.8 angstroms resolution and some comparisons with the oxidized state. *Journal of Molecular Biology*, 117:195.
- Subbiah, S., Laurents, D., and Levitt, M. (1993). Structural similarity of dna-binding domains of bacteriophage repressors and the globin core. *Current Biology*, 3:141–148.
- Taylor, W. and Orengo, C. (1989). Protein structure alignment. *Journal of Molecular Biology*, 208:1–22.
- Thornton, J. M. (1992). Protein structures: the end point of the folding pathway. In Creighton, T. E., editor, *Protein folding*, pages 59–81. Freeman and Co.
- Vasseur, C., Blouquit, Y., Kister, J., Prome, D., Kavanaugh, J. S., Rogers, P. H., Guillemin, C., Arnone, A., Galacteros, F., Poyart, C., Rosa, J., and Wajcman, H. (1992). Hemoglobin thionville: An alpha-chain variant with a substitution of a glutamate for valine at na-1 and having an acetylated methionine nh2 terminus. *Journal of Biological Chemistry*, 267:12682.
- Vriend, G. and Sander, C. (1991). Detection of common three-dimensional substructures in proteins. *Proteins*, 11:52–58.
- Warshel, A. and Levitt, M. (1976). Folding and stability of helical proteins: Carpmiogen. *Journal of Molecular Biology*, 106:421–437.
- White, C. T., Singh, R. K., Reintjes, P. B., Lampe, J., Erickson, B. W., Dettloff, W. D., Chi, V. L., and Altschul, S. F. (1991). BioSCAN: A VLSI-based system for biosequence analysis. *Proceedings of the International Conference on Computer Design: VLSI in Computers & Processors*, pages 504–509.
- Yee, D. P. and Dill, K. A. (1993). Families and the structural relatedness among globular proteins. *Protein Science*, 2:884–899.