Predicting the Triple Beta-Spiral Fold from Primary Sequence Data

by

Eben Louis Scanlon

Submitted to the Department of
Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degrees of

Master of Science in Electrical Engineering/Computer Science

and

Master of Business Administration

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2004

Author .....
Department of
Electrical Engineering and Computer Science
January 16, 2004

Certified by .....
Bonnie A. Berger
Professor of Applied Mathematics
Thesis Supervisor

Certified by .....
Roy E. Welsch
Professor of Statistics and Management
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Predicting the Triple Beta-Spiral Fold from Primary Sequence Data

by

Eben Louis Scanlon

Submitted to the Department of
Electrical Engineering and Computer Science
on January 16, 2004, in partial fulfillment of the
requirements for the degrees of
Master of Science in Electrical Engineering/Computer Science
and
Master of Business Administration

Abstract

The Triple β-Spiral is a novel protein structure that plays a role in viral attachment and pathogenesis. At present, there are two Triple β-Spiral structures with solved crystallographic coordinates – one from Adenovirus and the other from Reovirus. There is evidence that the fold also occurs in Bacteriophage SF6. In this thesis, we present a computational analysis of the Triple β-Spiral fold. Our goal is to discover new instances of the fold in protein sequence databases.

In Chapter 2, we present a series of sequence-based methods for the discovery of the fold. The final method in this Chapter is an iterative profile-based search that outperforms existing sequence-based algorithms. In Chapter 3, we introduce specific knowledge of the protein's structure into our prediction algorithms. Although this additional information does not improve the profile-based methods in Chapter 2, it does provide insight into the important forces that drive the Triple β-Spiral folding process. In Chapter 4, we employ logistic regression to integrate the score information from the previous Chapter into a single unified framework. This framework outperforms all previous methods in cross-validation tests.

We do not discover a great number of additional instances of the Triple β-Spiral fold outside of the Adenovirus and Reovirus families. The results of our profile based templates and score integration tools, however, suggest that these methods might well succeed for other protein structures.

Thesis Supervisor: Bonnie A. Berger
Title: Professor of Applied Mathematics

Thesis Supervisor: Roy E. Welsch
Title: Professor of Statistics and Management

Acknowledgments

The author wishes to acknowledge the *Leaders for Manufacturing Program* for its support of this work. He would also like to thank Professor Bonnie Berger, Professor Jonathan King, Professor Roy Welsch, and Peter Weigele for their guidance and support.

Contents

List of Figures

List of Tables

Chapter 1

Computational Background

1.1 Introduction

Proteins are fundamentally important biological molecules. Proteins perform many of the key functions of life including, but not limited to, enzymatic catalysis, transport and storage, control of growth, and transmission of nerve impulses. It is through proteins that DNA expresses the tremendous complexity and diversity of cellular processes. In addition, many human diseases can be traced to the malfunction of human proteins or the pathogenic properties of proteins in other organisms.

In this thesis, we present several novel computational methods for the analysis of proteins and protein structure. We focus particular attention on a protein fold that plays a key role in viral attachment and infection: the Triple β-Spiral. Although we focus particular attention on this fold, our goal is develop methods that are more broadly applicable.

Our analysis of the Triple β-Spiral fold is split into two main parts. In the first part, we analyze the amino acid sequence of the Triple β-Spiral fold. In the second part, we analyze the structure of this fold. In the remainder of this chapter, we give a basic introduction to protein structure and we present several common methods for the computational analysis of proteins. Readers familiar with these topics may wish to skip directly to Chapter 2.

1.2 Protein Sequence Determination

In its simplest form, a protein is composed of a linear chain of amino acid subunits. These subunits are joined together by a series of peptide bonds to form a long unbranched molecule with a uniform and repetitive backbone structure. There are twenty different amino acid subunits, each with a unique functional side chain. Differences in these functional side chains give each amino acid distinct chemical and physical properties. Usually, the twenty amino acids are designated either by a three-letter abbreviation or a one letter symbol. (See Figure 1-1) Each protein has a precisely defined sequence of amino acids. It is an organism's DNA that codes for each of these proteins through the degenerate genetic code in which three DNA nucleotides specify a single amino acid in a protein.

In the cell, proteins fold into complex three-dimensional structures. There are two common experimental methods for determining a protein's folded structure: X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR). Both of these methods are time-consuming and expensive to apply, and neither works in all cases. Researchers continue to dedicate resources to these studies, however, because a protein's structure is one key determinant of its function in the cell. In addition, discoveries that lead to greater understanding of folding properties of proteins promise deeper biological insight and elevated understanding of related problems.

Although a protein's three-dimensional structure is difficult to determine, its linear sequence is easier to find. For the past 50 years, experimental methods like *Edman Degradation* have allowed researchers to determine protein sequence [24]. More recently, the explosion of biological sequence data from the Human Genome Project [43] and related sequencing initiatives has led to a concomitant explosion in the number of proposed protein sequences. It is important to note, however, that some of the protein sequences derived from genome projects are not sequenced directly from cells, but rather computationally predicted from the core genomic data that these efforts provide. There is, therefore, a degree of uncertainty about the existence of these proposed proteins.

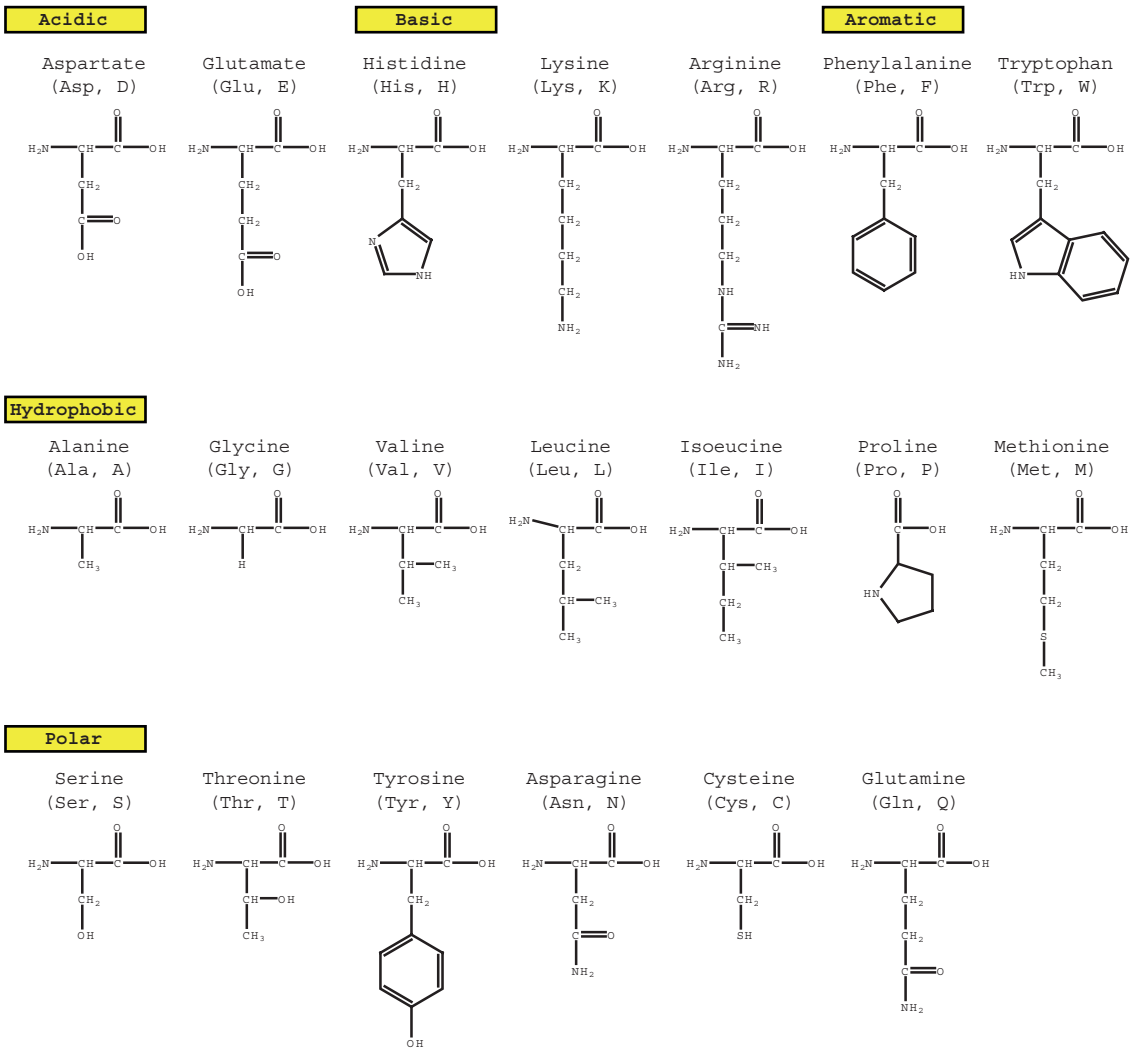

Figure 1-1: **The amino acids.** A list of all of the amino acids and their abbreviations and chemical formulae. The list is organized by chemical properties of the side-chains.

As the quantity of biological sequence information has grown, public databases have been created to hold and organize this data. Some of these databases are simply huge lists of biological sequences; others contain supplemental annotations and references. We will describe several of these databases in more detail later in this chapter.

Although the number of protein sequences has become immense over the past decade, the number of protein structures with known molecular coordinates has not kept pace. In part, this is attributable to the large expenditures of time and money

that are required to experimentally determine a protein's three-dimensional structure, and in part this is attributable to the fact that many proteins are not amenable to standard techniques for structure determination. For example, proteins that span cell membranes are notoriously difficult to solve through conventional techniques [12].

1.3 The Protein Folding Problem

As the number of protein sequences has far outstripped the number of known protein structures, computational biologists have worked to develop methods to deduce a protein's three-dimensional structure from its one-dimensional sequence. In part, these efforts are a continuation of earlier theoretical work in thermodynamics and quantum mechanics, and in part these efforts form a wholly novel statistics-based approach to protein structure.

At the most fundamental level, researchers are attempting to discover the forces and factors that drive proteins to fold into elegant three-dimensional structures and to elucidate the complex solvent interactions that proteins experience in cellular environments. At a more practical level, however, researchers have developed methods that give quite reliable protein structure predictions based upon a protein's one-dimensional sequence. Underlying most of these methods is the basic assumption that proteins with similar sequence (homologs) share a common evolutionary antecedent, and will therefore have the same three-dimensional structure. For the most part, these methods do not utilize thermodynamic or quantum mechanical principles.

The problem of predicting a protein's structure from its sequence is known as the "protein folding problem." A related but slightly different formulation of this problem is the "inverse folding problem." In this formulation, researchers begin with a protein fold, and then try to predict which protein sequences are compatible with this fold. Practically speaking, there is little difference between these two formulations, and the distinction between them is often blurred.

When two protein sequences are very similar to one another, there is little doubt that they share a common structure. When the degree of sequence similarity is

lower, however, the two protein sequences enter what is colloquially known as the "twilight zone" of protein sequences [55]. In this region of sequence similarity, it is not clear whether two proteins should be considered evolutionarily related. Advanced statistical techniques have been introduced to compare proteins in this region, and to make probabilistic statements about their relationships to one another. Later in this chapter we will describe several of these methods in more detail.

1.4 Protein Structure Basics

The first X-ray crystallographic image of a three-dimensional protein structure was attained for the protein myoglobin by John Kendrew in 1958 [42]. As soon as this structure was released, it became clear that proteins are much more structurally complex than DNA. Rather than folding into simple, regular, and symmetric structures, proteins fold into complex, irregular, and often asymmetric structures.

In spite of the complexity of protein structures, it is possible to break them down into smaller structural motifs. In 1951, seven years before the structure of myoglobin was first reported, Linus Pauling proposed two structural motifs that he expected to occur in three-dimensional proteins [53, 54]. He based these proposals on molecular models that he had constructed, and the properties of the atoms along the protein backbone. Pauling called these two structural elements the α-helix and the β-pleated sheet (or simply β-sheet.) He termed these two motifs "secondary-structural" elements to emphasize that he expected these to be small subunits of three-dimensional protein structures. Subsequent crystallization of many protein structures validated his theoretical models: the α-helix and the β-sheet occur frequently. Viewing a protein as a set of α-helices and β-sheets joined together by unstructured protein coils has proven the simplest way to visualize protein structures[1].

The α-helix is a simple right-handed helix, in which the protein backbone loops around and forms hydrogen bonds to itself with a periodicity of 3.6 amino acids per

[1]There are a number of more specialized secondary structural elements, but the α-helix and the β-sheet are by far the most common and the most useful.

turn. The β-sheet is composed of a set of β-strands that assemble together to form a continuous protein sheet. The signature of a single β-strand is that its backbone is stretched out into a fully extended conformation. In this fully extended conformation, two adjacent β-strands can engage in hydrogen bonding, with the backbone amide hydrogens bonding to the backbone carbonyl oxygen atoms on adjacent strands.

There are two main types of β-sheets: parallel and antiparallel. In parallel β-sheets, the strands that compose the sheet are arranged so that each strand has the same orientation from N-terminus to C-terminus. In antiparallel β-sheets the orientation is reversed. Because the peptide backbone of each of the strands in a β-sheet is fully extended, residue side chains are oriented out of the plane of the β-sheet. Consecutive side chains point out of opposite faces of the sheet, and every other side chain along a strand has the same orientation. One interesting aspect of β-sheets is that residue side chains display marked correlations in their hydrogen bonding pairs in the sheet formation [46].

Three-dimensional protein structures can be represented as a compilation of secondary structural elements. Figure 1-3 shows a TIM-Barrel, which consists of eight parallel β-strands forming a barrel shape, and eight α-helices that form a covering for the barrel [1]. The β-strands and α-helices are joined together by unstructured "random coil" regions.

In addition to forming complex three-dimensional structures, some proteins also join together with other proteins to form long-lived multi-protein complexes. This is called the "quaternary" structure of a protein. For example, in Figure 1-4, three α-helical proteins join together as permanent partners in a three-stranded coiled-coil domain. The protein fold that we will be studying in this thesis – the Triple β-Spiral – is similar to the three-stranded coiled-coil. It consists of three identical and permanently interacting protein chains.

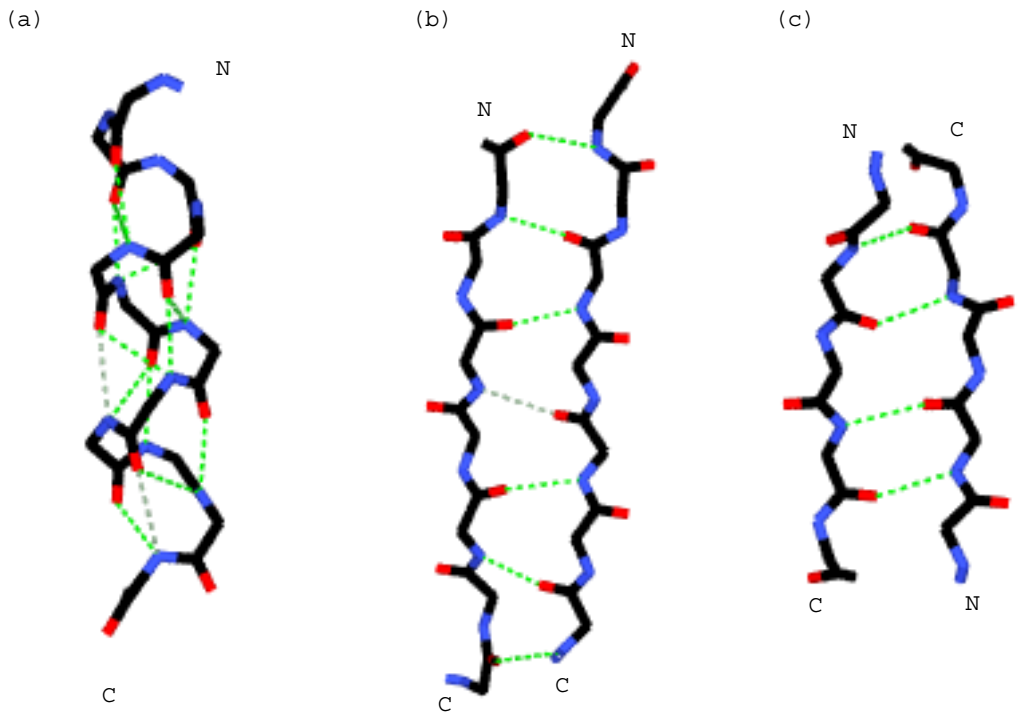

Figure 1-2: **Protein Secondary Structure.** (a) An α-helix with the protein backbone. (b) A parallel β-sheet with two β-strands. (c) An antiparallel β-sheet with two β-strands. All three figures show only the protein backbone (no side chains) and display backbone hydrogen bonds with dotted green lines. Atoms in the figures follow the normal conventions: black for carbon, red for oxygen, and blue for nitrogen. Hydrogen atoms are not shown. The orientation of each protein chain is denoted by the letters `N` and `C`. Note the tight turns in the α-helical backbone chain and the extended backbone conformation in the β-sheets.

1.5 Protein Databases

There are a number of public databases that contain information about proteins. Table 1.1 provides a brief summary of several of these, although the list is certainly not all-encompassing. Broadly speaking, the available protein databases can be split into two main categories: (1) those that provide primary protein information, and (2) those that provide annotations or classifications of primary protein information. An example of the former kind is the Research Collaboratory for Structural Bioinformatics Protein Data Bank (`PDB`), which contains the molecular coordinates for all proteins of known three-dimensional structure [8]. An example of the latter is the Structural Classification of Proteins (`SCOP`) which organizes proteins of known structure into a

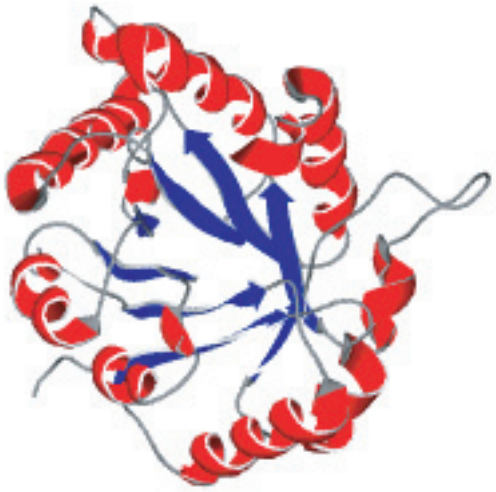

Figure 1-3: **A TIM Barrel.** This is an image of a TIM Barrel structure, which consists of eight parallel β-strands in a barrel covered by eight α-helices. The convention is to show β-strands as arrows, α-helices as cylinders, and random coil as either tubes or strings. There is no universal coloring scheme. Colors are chosen to help illustrate particular aspects of a three-dimensional structure. This is a view of the TIM-Barrel looking down the axis of the barrel.

Figure 1-4: **Coiled Coils.** A three-stranded coiled coil, consisting of three distinct interacting α-helices.

six-level hierarchy of structurally related proteins [51]. There are, of course, databases that break this simple categorization scheme. For example, the `Swiss-Prot` database provides both primary protein sequence data and extensive annotations for each of these sequences [9].

For the purposes of this thesis, several databases are of particular interest. The aforementioned PDB is the primary source for all solved protein structures, and is therefore of critical importance to protein folding research. Used in conjunction with a three-dimensional protein viewer[2], the PDB provides a view of the complex and beautiful secondary, tertiary and quaternary structures of proteins. There are two important supplements to the PDB. The first, the Astral database, provides extensively annotated and cross-referenced information about each protein domain within the PDB [13, 16]. The second, the SCOP database, provides a common language for discussing structurally similar proteins. Using some automation and a panel of experts, SCOP assigns every protein in the PDB to a distinct protein domain. The protein structure that we will discuss in this thesis – the Triple β-Spiral – is classified within SCOP, and the corresponding three-dimensional coordinates are contained in the PDB and Astral databases.

The Swiss-Prot database from the Swiss Institute of Bioinformatics is another database of key importance. The Swiss-Prot database contains a collection of extensively annotated protein sequences. Some of these sequences are also contained in the PDB, but there are also many sequences without known structure in the Swiss-Prot database. Each sequence in the Swiss-Prot database is assigned a unique identifier and is also given a short entry name. For example, the fiber protein from the Human Adenovirus Serotype 2 has the identifier P03275 and the entry name FIBP_ADE02.

Many proteins with known sequence are not annotated in Swiss-Prot. These sequences are compiled in the TrEMBL database [9]. Taken together, the Swiss-Prot and TrEMBL databases (the Swiss-Prot/TrEMBL database) contain a reasonably comprehensive list of known protein sequences. Swiss-Prot/TrEMBL is comparable in size and scope to the Non-Redundant Protein Database (NR) database from the National Center Biotechnology Information (NCBI).

Finally, we use several more specialized databases periodically throughout this thesis. We discuss these databases in more detail as they are used.

[2]Swiss-PDB at http://us.expasy.org/spdbv is one such tool.

Name	Brief Description	Number of Sequences
PDB	Contains 3-D coordinates of solved proteins	22,333
SCOP	PDB organized into hierarchical domains (1.63)	NA
Astral	PDB structures split into individual domains (1.63)	46,981
Swiss-Prot	Annotated protein information (41.20)	132,675
TrEMBL	Protein sequences not yet in Swiss-Prot (24.8)	940,641
SP/TrEMBL	Combined Swiss-Prot and TrEMBL protein sequences	1,073,316
NR	NCBI non-redundant protein sequence database	1,487,336

Table 1.1: **Protein Sequence Databases.** A list of some common protein databases and the number of elements that they contain as of August 15, 2003. Where appropriate, the version of the database is indicated in parentheses. In other cases, the databases are updated daily, weekly, or monthly and are not assigned a version number.

1.6 Protein Sequence Alignments

Up to this point, we have spoken of the similarity between protein sequences without precisely defining this term. Intuitively, two proteins are similar to one another if they share regions of similar amino acid residues. To formalize this intuition, we introduce the notion of an "alignment." In an alignment, two proteins are compared to attain the greatest degree of overlap between them. The degree of similarity between two protein sequences is then determined based upon this optimal alignment.

Although this methodology might seem somewhat arbitrary, there is a good biological justification for this approach. Underlying the comparison of two proteins is the assumption that they share a common evolutionary precursor. We would, therefore, like to ignore regions that are the result of insertions and mutations that do not affect basic protein structure.

1.6.1 Scoring Matrices

Figure 1-5 shows two alternative alignments between two protein fragments. Intuitively, the second alignment is preferable to the first because it has a greater degree of overlap between the two proteins. To formalize this intuition, we need to develop an alignment scoring function. The simplest such function is to assign a score of

+1 to every pair of identical residues and a score of −1 to every pair of mismatched residues. Under this scoring scheme, alignment (a) in Figure 1-5 would attain a score of −9 and alignment (b) would attain a score of −1.

```
(a) PGLSLDSNNALQVHTG
    AGLQISNNALAVKVG

(b) PGLSLDSNNALQVHTG
     AGLQISNNALAVKVG

(c) PGLSLDSNNALQVHTG
    AGLQI-SNNALAVKVG
```

Figure 1-5: **Protein Sequence Alignments.** Three possible alignments for fragments from an Adenovirus protein sequence. Intuitively, alignment (b) is preferable to (a) because it has a greater number (7 vs. 3) of aligned identical residues. Alignment (c) is preferable to both (a) and (b) because it has 9 aligned residues.

Although this scoring scheme works remarkably well considering its simplicity, it does not take into account the relationships between different types of amino acids. For example, Leucine and Isoleucine have quite similar chemical properties, and are more similar to one another than either is to Tryptophan. To capture the relationships between different amino acid residues, Henikoff and Henikoff created a 20×20 symmetric matrix called the `BLOSUM62` matrix (Table 1.2) that gives a score to every pair of amino acids[3]. The `BLOSUM62` matrix was constructed by hand-aligning a large set of related proteins and then counting the frequencies of residue pairings [30]. The entries in the `BLOSUM62` matrix give log-odds values for each of these frequencies. Using this matrix, alignment (a) in Figure 1-5 would attain a score of 29 and alignment (b) would attain a score of 36.

1.6.2 Alignment Gaps

Alignment (c) in Figure 1-5 shows the same two proteins from the previous example, but with a "gap" in the alignment between the two proteins. Intuitively, this alignment is preferable to either of the previous two because it has the greatest degree of

[3]In fact, there are several alternative matrices from which to choose. Each is useful under certain conditions. The interested reader is referred to the excellent introduction to bioinformatics by Durbin et al [22].

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Table 1.2: **The BLOSUM62 scoring matrix.** The BLOSUM62 matrix is used to score residue pairs in protein sequence alignments. It holds log-odds scores from hand-aligned sets of proteins. The true BLOSUM62 matrix also holds scores for a number of non-standard residue identifiers. We omit these in this figure.

overlap between the two proteins.

There is no formal probabilistic framework for incorporating gaps into the scoring scheme [3]. A simple and effective way to incorporate gaps, however, is to assign some arbitrary penalty (say -4) to aligning a residue with a gap. If we assign gap penalties in this way, then alignment (c) in Figure 1-5 has a score of 39.

From a biological perspective, a "gap" in an alignment represents a region between two proteins where either a deletion or an insertion has occurred in the genome. Since these insertions and deletions usually occur in fragments of several residues, ideally we would like our gap scoring function to penalize opening a gap more than extending this gap. One way to achieve this is to introduce an "affine" gap-penalty. In this scheme, opening a gap is penalized some large value and incrementally extending this gap is penalized some smaller fixed value. In practice, values of -11 for opening a gap and -1 for extending a gap have been found to work well [2].

1.6.3 Alignment Algorithms

The simplest way to find the optimal alignment ot two proteins (P_1 and P_2 of length n_1 and n_2 respectively) is to consider the score of every possible alignment between them. Unfortunately, the number of possible alignments between two proteins is equal to

$$\begin{pmatrix} n_1 + n_2 \\ n_1 \end{pmatrix} = \frac{(n_1 + n_2)!}{(n_1!)(n_2!)}$$

This is an impossibly large number of alignments to consider for two proteins of reasonable length.

Fortunately, if we want to find only the optimal alignment between two proteins we can employ a dynamic programming algorithm developed by Needleman and Wunsch [52]. In this algorithm, we construct a matrix of optimal "sub-scores" for each sub-alignment of two the proteins. It proceeds as follows:

1. Start directly before the first residue of each protein.

2. Consider each of three possibilities:

   (a) The next residue from P_1 is aligned with the next residue from P_2

   (b) The next residue from P_1 is aligned with a gap from P_2

   (c) The next residue from P_2 is aligned with a gap from P_1

3. Score each cell in the matrix by calculating the maximum of all paths that lead into this cell.

4. Return to step 2.

The key to this algorithm is that we only need to store the optimal score for each of the sub-alignments. That is, the optimal score for an alignment of length $i+1$ will depend only upon the optimal scores for each of the sub-alignments of length i – we do not need to consider all of the sub-optimal alignment scores. This algorithm reduces the complexity of the problem to $O(n_1 \times n_2)$.

Formally, if we let $A(i,j)$ be the optimal score for aligning P_1 up to residue i and P_2 up to residue j, then the score $A(i,j)$ can be calculated recursively as

$$A(i,j) = \max \begin{cases} A(i-1,j-1) + s(i,j) \\ A(i-1,j) - g \\ A(i,j-1) - g \end{cases}$$

where g is the gap penalty and $s(i,j)$ is the `BLOSUM62` score for aligning the residue at position i from P_1 with the residue at position j from P_2. Figure 1-6 shows a simple example of this alignment technique. To simplify our algorithm, we use a non-affine penalty of -4 for opening and extending gaps. Using affine gap penalties slightly increases the complexity of the algorithm.

There are two basic types of alignments: global and local. Figure 1-6 shows an example of a global alignment. In a global alignment, two sequences are compared throughout their entire lengths to determine the best-scoring overlap. In a local

(a)

		P	G	L	S	L	D	S	N	N	A	L	Q	V	H	T	G
	0	-4	-8	-12	-16	-20	-24	-28	-32	-36	-40	-44	-48	-52	-56	-60	-64
A	-4	-1	-4	-8	-11	-15	-19	-23	-27	-31	-32	-36	-40	-44	-48	-52	-56
G	-8	-5	5	1	-3	-7	-11	-15	-19	-23	-27	-31	-35	-39	-43	-47	-46
L	-12	-9	1	9	5	1	-3	-7	-11	-15	-19	-23	-27	-31	-35	-39	-43
Q	-16	-13	-3	5	9	5	1	-3	-7	-11	-15	-19	-18	-22	-26	-30	-34
I	-20	-17	-7	1	5	11	7	3	-1	-5	-9	-13	-17	-15	-19	-23	-27
S	-24	-21	-11	-3	5	7	11	11	7	3	-1	-5	-9	-13	-16	-18	-22
N	-28	-25	-15	-7	1	3	8	12	17	13	9	5	1	-3	-7	-11	-15
N	-32	-29	-19	-11	-3	-1	4	9	18	23	19	15	11	7	3	-1	-5
A	-36	-33	-23	-15	-7	-4	0	5	14	19	27	23	19	15	11	7	3
L	-40	-37	-27	-19	-11	-3	-4	1	10	15	23	31	27	23	19	15	11
A	-44	-41	-31	-23	-15	-7	-5	-3	6	11	19	27	30	27	23	19	15
V	-48	-45	-35	-27	-19	-11	-9	-7	2	7	15	23	26	34	30	26	22
K	-52	-49	-39	-31	-23	-15	-12	-9	-2	3	11	19	24	30	33	29	25
V	-56	-53	-43	-35	-27	-19	-16	-13	-6	-1	7	15	20	28	29	33	29
G	-60	-57	-47	-39	-31	-23	-20	-16	-10	-5	3	11	16	24	26	29	39

(b)

```
PGLSLDSNNALQVHTG
AGLQI-SNNALAVKVG
```

Figure 1-6: **Example DP Matrix.** (a) The dynamic programming matrix for aligning two Adenovirus sequence fragments. For simplicity, we have used a non-affine gap penalty of -4 for all gaps. The score for the final alignment is 39, which can be found in the bottom right-hand corner of the matrix. The optimal alignment can be reconstructed by tracing back through the matrix elements and choosing the maximum precursor at each step. The red arrows demonstrate this traceback. (b) The optimal alignment based upon this traceback.

alignment, the two sequences are searched for the highest-scoring region of localized overlap. In general, local alignments are more useful than global alignments because they do not impose the assumption of a complete shared structure. In practice, the algorithms for local and global alignment are almost identical. In a local alignment, however, the alignment score is not allowed to become negative, and the best local alignment is recovered from the maximal score at any point in the dynamic programming matrix.

Several alignment methods are in common use. Most of these provide some optimizations that make the alignment algorithm more rapid at the expense of complete determinism. For example, the Basic Local Alignment Search Tool (`BLAST`) searches

for short regions of protein identity and then tries to extend these regions through dynamic programming around these points [2].

1.6.4 Sequence Similarity

Once two sequences are aligned, the sequence identity and similarity are calculated as:

$$\% \text{ Identity} = \frac{\text{Number of identical aligned residues}}{\text{Length of shorter sequence}} \times 100$$

$$\% \text{ Similarity} = \frac{\text{Number of similar aligned residues}}{\text{Length of shorter sequence}} \times 100$$

where residues are considered similar if they have a positive alignment score in the `BLOSUM62` matrix.

As a rule of thumb, proteins that have sequence identity of greater than 25% or sequence similarity of greater than 40% have the same three-dimensional structure [55]. Unfortunately, the converse is not true: there are many instances of proteins with low sequence similarity that have the same three-dimensional structure.

Sequence alignment methods are often used to test all of the sequences in a protein sequence database against a desired target sequence. For example, a researcher who has just discovered a novel protein might align this protein with all of the sequences in the `Swiss-Prot` database to find other similar proteins. This technique can be useful, as sequence similarity is a rapid test that implies functional and structural relationships between proteins.

1.7 Profile Methods

Although sequence alignment methods provide a reasonable starting point for comparison of two sequences, ideally one would like to be able to compare a sequence to an entire family of related proteins rather than just a single representative of this family.

For example, Figure 1-7 shows an example of an alignment of a number of Adenovirus sequence fragments. There is more information contained in this alignment than in any single sequence from this alignment.

To take advantage of family relationships, we can use an alignment of related proteins to build a "profile" that represents the entire alignment [32]. This profile can then be used to search a sequence database for other members of the protein family. Profile search methods are more sensitive than search methods utilizing a single sequence because they incorporate position-specific information about conservation of residues at each position in the profile.

```
FIBP_ADEB3_114  PGLSLDSN-----NALQVHTG
FIBP_ADEB3_176  AGLQISN------NALAVKVG
FIBP_ADECC_34   KGLTESSP-----GTLAVNIS
FIBP_ADECC_98   DGLTFTSPLHKIENTVSLSIG
FIBP_ADE1A_52   TPLTTTG------GSLQLKVG
FIBP_ADE1A_86   TPLVKTG------HSIGLSLG
FIBP_ADEM1_118  APLQIND------GVLQLSFG
FIBP_ADEP3_99   SPITLTA------EGISLSLG
FIBP_ADEP3_129  APLQFQG------NALTLPLA
FIBP_ADEP3_144  AGLQNTD------GGMGVKLG
FIBP_ADE02_53   EPLDTSH------GMLALKMG
FIBP_ADE02_88   QPLKKTK------SNISLDTS
FIBP_ADE02_103  APLTITS------GALTVATT
FIBP_ADE02_118  APLIVTS------GALSVQSQ
```

Figure 1-7: **Aligned Adenovirus Sequences.** A set of Adenovirus protein sequence fragments aligned by `ClustalW`. Note the residue bias to `P` and `G` at column 2 in the profile.

1.7.1 Profile Construction

There are several ways to construct a profile from an alignment. The simplest way is to assign position specific probabilities based upon the maximum likelihood estimates for each residue. So for example, in Figure 1-7 we would assign a probability of 9/14 to Proline (`P`) at position 2. There are obvious problems with assigning probabilities in this way. These problems include:

1. Residues that do not occur will be assigned a probability of 0, indicating that we will *never* match this residue in any alignment.

2. This method does not take into account the background distribution of residues. This background distribution is not uniform.

To address these shortcomings, most profile construction algorithms use a probabilistic approach to assign position specific residue probabilities. The simplest such approach is to add a background-dependent "pseudocount" to each of the residue frequencies. In this approach, residue probabilities at each position in the profile are calculated according to

$$p_i(r) = \frac{c_r + A \times b_r}{n + A} \tag{1.1}$$

where n is the total number of observed residues at that position, c_r is the number of residues of type r and b_r is the background probability of this residue in all protein sequences (see Table 1.3.) The constant A determines what weight is given to the background probability vis-a-vis the residue counts. Usually A is given a value of 20 [22]. This formula has the rather appealing property that when there are few sequences in a profile the residue probabilities approximate the background distribution, and when there are many sequences in the profile, the residue probabilities approximate the maximum likelihood estimates. In practice, these probabilities are usually rescaled by

$$s_i(r) = \log_2 \frac{p_i(r)}{b_r}$$

to give a position-specific residue bit-score. Figure 1-8 shows the scores for the Adenovirus alignment calculated in this way. Although the background pseudocount method works well in most cases, it implicitly assumes that each of the sequences in the alignment is independent. If this assumption is incorrect – i.e. if the alignment contains many sequences that are not sufficiently evolutionarily diverged – then the profile will be biased in favor of the dependent sequences. To address this problem we can pre-weight each of the sequences in the alignment. Henikoff and Henikoff discuss several such sequence weighting schemes [31].

Residue	Prob	Score	Residue	Prob	Score
A	0.0776	0.634	M	0.0238	-1.073
C	0.0158	-1.664	N	0.0427	-0.226
D	0.0529	0.082	P	0.0487	-0.039
E	0.0656	0.391	Q	0.0392	-0.350
F	0.0406	-0.301	R	0.0525	0.070
G	0.0691	0.466	S	0.0695	0.475
H	0.0227	-1.139	T	0.0550	0.139
I	0.0591	0.242	V	0.0667	0.416
K	0.0596	0.253	W	0.0118	-2.080
L	0.0959	0.940	Y	0.0312	-0.681

Table 1.3: **Background Residue Probabilities.** Residue probabilities in the `Swiss-Prot` database. The scores are calculated as $\log_2(b_r/(1/20))$ where we divide b_r by 1/20 only to give an idea of whether the residue is more or less likely than we would expect based upon the uniform distribution.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	1.52	-0.77	0.19	0.05	-0.77	-0.77	-0.77	-0.77	0.11	-0.77	-0.77	-0.77	0.25	0.42	-0.77	0.02	0.73	-0.77	-0.77	-0.77
2	-0.77	-0.77	-0.77	-0.77	-0.77	1.44	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	2.59	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77
3	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	0.12	-0.77	2.19	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77
4	-0.77	-0.77	0.19	-0.77	-0.77	-0.77	-0.77	0.12	0.11	-0.77	-0.77	-0.77	-0.77	1.84	-0.77	0.02	1.70	0.04	-0.77	-0.77
5	-0.77	-0.77	-0.77	0.05	1.03	-0.77	-0.77	1.06	0.66	0.26	-0.77	0.35	-0.77	-0.77	-0.77	-0.77	0.73	0.04	-0.77	-0.77
6	-0.77	-0.77	0.19	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	0.35	-0.77	0.42	-0.77	0.89	2.28	-0.77	-0.77	-0.77
7	-0.05	-0.77	0.77	-0.77	-0.77	0.90	0.91	-0.77	0.11	-0.77	-0.77	0.35	-0.77	-0.77	-0.77	1.44	-0.77	-0.77	-0.77	-0.77
8	-0.77	-0.77	-0.77	0.05	-0.77	1.84	0.91	-0.77	-0.77	-0.77	-0.77	1.74	-0.77	-0.77	-0.77	0.02	-0.77	-0.77	-0.77	-0.77
9	1.31	-0.77	-0.77	-0.77	-0.77	0.53	-0.77	-0.77	-0.77	-0.77	0.87	0.35	-0.77	-0.77	-0.77	0.52	0.73	0.04	-0.77	-0.77
10	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	1.06	-0.77	1.74	0.87	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	0.04	-0.77	-0.77
11	0.79	-0.77	-0.77	-0.77	-0.77	0.53	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	1.50	-0.77	1.19	0.73	-0.77	-0.77	-0.77
12	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	1.60	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	1.69	-0.77	-0.77
13	-0.05	-0.77	0.19	-0.77	-0.77	-0.77	0.91	-0.77	1.36	-0.77	-0.77	0.35	0.25	0.42	-0.77	1.19	-0.77	-0.77	-0.77	-0.77
14	-0.77	-0.77	-0.77	-0.77	0.39	-0.77	-0.77	0.66	-0.77	0.86	0.87	-0.77	-0.77	-0.77	-0.77	0.02	1.13	0.56	-0.77	-0.77
15	-0.05	-0.77	-0.77	-0.77	-0.77	2.14	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	-0.77	0.42	-0.77	0.52	0.17	-0.77	-0.77	-0.77

Figure 1-8: **Profile for Aligned Adenovirus Sequences.** Profile scores for the Adenovirus alignment pictured in Figure 1-7. Note that we assume that positions 8 through 13 are gaps. The profile therefore has 15 positions.

1.7.2 PSI-BLAST

Profile methods have found wide use in computational biology. In particular, the `PSI-BLAST` program [3] (the successor to the `BLAST` program) uses profiles in the following way:

1. Given an initial target sequence, an alignment is generated by a database search via basic `BLAST`.

2. A profile is constructed from this alignment using a pseudo-count method based

on the BLOSUM62 matrix. This is slightly different from the background pseudocount method.

3. This profile is used to search the database, with each subsequent search producing new sequences to add to the profile.

PSI-BLAST has met with great success, and it is the most commonly used method for sequence alignments and homology search.

Although PSI-BLAST is effective at detecting weak homology among proteins, it must be used with care. This is because the incorporation of even one or two spurious sequences into a PSI-BLAST profile during early iterations can lead to the incorporation of many incorrect sequences in later iterations [57, 37].

1.7.3 Profile Hidden Markov Models

In addition to PSI-BLAST, several other profile methods are common in computational biology. One of these is the profile Hidden Markov Model (HMM) [23]. A profile HMM is constructed by first building a profile from a hand-selected and pre-aligned set of sequences. From this profile, an HMM is constructed with three states – Match, Insert, and Delete– for each position in the profile. In this model, a Match state emits residue symbols with probability according to the profile. The Insert state emits residue symbols with probability according to the background distribution[4], and the Delete state is a non-emitting state corresponding to a gap in the profile.

Profile HMMs differ from other profile methods because they associate a probability (a score) to the transition between each state. In this way, profile HMMs can incorporate position-specific gap and insertion penalties into a protein profile. This makes HMMs preferable to PSI-BLAST for matching entire protein domains.

To align a target sequence with a profile HMM, the sequence is passed through the HMM, and the most probable alignment of the sequence with the profile is determined. This calculation is carried out by a forward dynamic programming technique called the Viterbi algorithm. In this algorithm a frontier of available states is kept as each

[4]Actually, a slightly modified background distribution is used [22].

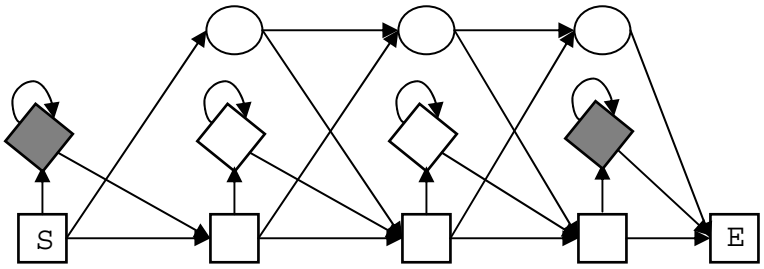

Figure 1-9: **A profile HMM.** In this model, `Match` states are represented by squares, `Insert` states are represented by diamonds, and `Delete` states are represented by circles. There are no transitions from `Delete` states to `Insert` states. The gray `Insert` states at the beginning and end of the model emit residues according to the background residue distribution, and is used to match parts of the sequence that are outside of the profile. The `Start` and `End` states (`S` and `E`) are special non-emitting states.

symbol in the target sequence is consumed by the model. At points where the same state can be entered from two different preceding states, only the state with the higher score (probability) is stored.

Formally, if we let

$$V_j^M(i) = \text{Best subsequence score of length } i \text{ leading to } \texttt{Match}\ j$$

$$V_j^I(i) = \text{Best subsequence score of length } i \text{ leading to } \texttt{Insert}\ j$$

$$V_j^D(i) = \text{Best subsequence score of length } i \text{ leading to } \texttt{Delete}\ j$$

then the Viterbi relations for the dynamic programming algorithm can be written as

$$V_j^M(i) = e_{M_j}(i) + \max \begin{cases} V_{j-1}^M(i-1) + a_{M_{j-1}M_j} \\ V_{j-1}^I(i-1) + a_{I_{j-1}M_j} \\ V_{j-1}^D(i-1) + a_{D_{j-1}M_j} \end{cases}$$

$$V_j^I(i) = e_{I_j}(i) + \max \begin{cases} V_j^M(i-1) + a_{M_j I_j} \\ V_j^I(i-1) + a_{I_j I_j} \\ V_j^D(i-1) + a_{D_j I_j} \end{cases}$$

$$V_j^D(i) = \max \begin{cases} V_{j-1}^M(i-1) + a_{M_{j-1} D_j} \\ V_{j-1}^I(i-1) + a_{I_{j-1} D_j} \\ V_{j-1}^D(i-1) + a_{D_{j-1} D_j} \end{cases}$$

where a_{XY} is the log-odds scores for the transition from state X to state Y, the $e_{M_j}(i)$ is the profile score for emitting the residue at position i in the protein from the j^{th} `Match` state, and $e_{I_j}(i)$ is the profile score for emitting the residue at position i in the protein from the j^{th} `Insert` state.

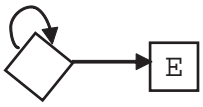

Figure 1-10: **HMM Null Model.** The null model against which a sequence alignment to a profile HMM is compared. Residues are emitted from the `Match` state in this model according to the background residue distribution. The transition probability control the length of the expected match. In practice, this is very close to 1 (350/351 for `HMMER`.)

Two points about the Viterbi algorithm deserve special attention. First, generally the Viterbi algorithm calculates the best path using log-odds scores (as above) rather than using the probabilities directly. Second, the calculation of the final bit-score for a protein is performed with respect to an underlying null model (see Figure 1-10). The final score of a sequence is then equal to:

$$\text{score} \quad = \quad \log_2 \frac{P(\text{seq}|\text{HMM})}{P(\text{seq}|\text{null})}$$

HMMs can model both global and local alignments, and can incorporate position specific gap and deletion penalties. These benefits make HMMs preferable to other profile methods for matching entire protein domains. For example, an HMM can perform a "glocal" match to a sequence in which a single sequence can match an entire model multiple times. This is useful if we expect that a protein structural motif will recur in a single protein chain. This greater flexibility, however, comes at a cost. Profile HMMs are more difficult to iterate than `PSI-BLAST`. They also require an externally-created initial alignment of hand-selected proteins for the construction of their profile.

There are several good publicly available HMM's for use by the research community [23, 41]. In this thesis we focus on `HMMER` by Sean Eddy, which is the most widely used and referenced HMM. Figure 1-11 shows a `HMMER` model file constructed from the alignment in Figure 1-7. The details of this file can be found in the `HMMER` documentation and in Durbin et al.'s book on HMM's [22].

The `Pfam` database

One novel use of Profile HMMs is the `Pfam` database [5]. In the `Pfam` database, expert curators create protein profiles by hand-selecting related families of proteins.[5] These profiles are then compiled into HMMs using the `HMMER` tool. These HMM's are used to search the `Swiss-Prot/TrEMBL` database for other potential family members. In this way, sequences in the `Swiss-Prot/TrEMBL` database can be automatically assigned to membership in protein families.

[5] Actually there are two `Pfam` databases. One is curated manually and the other by automated clustering algorithms.

```
HMMER2.0  [2.3.1]
NAME  adeno-example
LENG  18
ALPH  Amino
RF    no
CS    no
MAP   yes
COM   hmmbuild adeno-example.hmm adeno-example.aln
COM   hmmcalibrate --seed 0 adeno-example.hmm
NSEQ  14
DATE  Sat Sep  6 21:49:36 2003
CKSUM 2479
XT      -8455     -4  -1000  -1000  -8455     -4  -8455     -4
NULT      -4  -8455
NULE     595  -1558     85    338   -294    453  -1158...
EVD   -10.951290   0.377361
        A     C      D     E      F      G..    P     Q      R     S      T      V     W     Y
     m->m  m->i   m->d  i->m   i->i   d->m   d->d  b->m   m->e
      -93      * -4000
 1   2072 -2966    463   366  -3286  -2460..  234   972  -1216    87    948  -2588  -3150 -2466
 -   -149  -500    233    43   -381    399..  394    45     96   359    117   -369   -294  -249
 -     -4 -9060 -10102  -894  -1115   -701  -1378   -93      *
 2  -4178 -4440  -5600 -5973  -6579   2413.. 3642 -5961  -6018 -4436  -4618  -5865  -5878 -6592
 -   -149  -500    233    43   -381    399..  394    45     96   359    117   -369   -294  -249
 -     -4 -9060 -10102  -894  -1115   -701  -1378     *      *
 3  -4728 -4095  -7080 -6533  -2158  -6877..-5879 -5048  -5788 -6345  -4571  -2308  -4005 -4212
 -   -149  -500    233    43   -381    399..  394    45     96   359    117   -369   -294  -249
 -     -4 -9060 -10102  -894  -1115   -701  -1378     *      *
 4  -1500 -2803    292  -866  -3060  -2498..-2592  2494  -1269    90   2201   -184  -3029 -2389
 -   -149  -500    233    43   -381    399..  394    45     96   359    117   -369   -294  -249
 -     -4 -9060 -10102  -894  -1115   -701  -1378     *      *
                                                 .
                                                 .
                                                 .
17  -1763 -1582  -4109 -3476    735  -3322..-3364 -2695  -2877  -207   1645   1084  -2052 -1715
 -   -149  -500    233    43   -381    399..  394    45     96   359    117   -369   -294  -249
 -     -4 -9060 -10102  -894  -1115   -701  -1378     *      *
18    240 -2742  -4147 -4119  -5094   3096..-3706   631  -4224  1103    276  -3778  -5258 -5022
 -      *     *      *     *      *      *.     *     *      *     *      *      *      *     *
 -      *     *      *     *      *      *      *     *      0
//
```

Figure 1-11: **A Sample HMMER File.** This HMMER model file was constructed from the Adenovirus alignments in this chapter. Profile positions 5 through 16 are not shown. The most important part of this model are the numbered rows. For each profile position, the first row gives the emission scores from the Match state, the second row gives the emission scores from the Insert state, and the next row gives the transition scores. Note the high emission scores for Proline and Glycine in the second profile position. These scores differ slightly from those in our hand-computed profile because HMMER uses Dirichlet mixtures rather than the background pseudo-count method to derive its scores [22]. Scores in this model are multiplied by 1000.

1.7.4 Calcualtion of E-values

When PSI-BLAST and HMMER are used to search a sequence database, the output is a set of sequences that match the profile and a "bit-score" for each sequence. This bit-score measures how well a sequence aligns to a profile. Karlin and Altschul showed that these bit-scores follow an extreme-value distribution for profile-based

search methods [40]. The significance of a bit-score can therefore be estimated by comparing the bit-score to the expected number of sequences that would attain this bit-score or higher for a database of a given size.

From this estimation, both `PSI-BLAST` and `HMMER` calculate an "E-value" which is a measure of the number of sequences that we would expect to attain this bit-score or higher by random chance. By default, both `PSI-BLAST` and `HMMER` output all sequences in a database that match a profile with E-value 10 or lower.

1.8 Other Protein Structure Tools

Many more specialized protein structure prediction methods are in widespread use. In this section we attempt to give a very brief overview of the field.

1.8.1 PROSITE

The `PROSITE` database contains a large list of regular expression patterns corresponding to protein sequence and structure motifs [25]. These regular expressions are similar to protein profiles, but they are considerably simpler, as they do not take into account the relative frequency of different residues at each position in the motif, nor do they incorporate background residue frequencies.

The greatest shortcoming of the `PROSITE` model is that novel sequences cannot match a regular expression unless they match it at every position. Although this is an acceptable restriction for simple protein screens, it is too great a restriction for a true search method. In order to discover weakly homologous protein partners, a search method must allow for a close but imperfect match between a target and a sequence motif. For this reason the `PROSITE` database is of limited utility, and has been largely supplanted by `Pfam`.

1.8.2 Threading Methods

To this point, all of the computational methods that we have discussed take only the one dimensional sequence of a protein into account. Although these methods have proven quite effective, it is sometimes desirable to also incorporate the proposed three-dimensional structure of a protein. The most common way to accomplish this is to start with a three-dimensional structural template, and then to "thread" a proposed set of amino acids onto this template [48]. The quality of the threading is calculated by an energy function, usually one that takes the specific environment (solvation energy, hydrogen bonds with neighbors, etc.) of each residue into account.

Unfortunately, the calculation of the best threading for a given structural template is NP-complete [44], so in most cases the initial threading is determined by a sequence alignment. Indeed, the greatest bottleneck to threading is getting this initial alignment of target to template correct. The most common threading method is `GenThreader` by David T. Jones [38, 35].

1.8.3 Molecular Dynamics and Ab Initio Methods

Molecular dynamics and *ab initio* methods find their roots in thermodynamics and quantum mechanics [62, 18]. In these methods, the entire protein and all of its atoms are taken into account. In theory, using these methods one could perform a complex energy optimization to find the folded configuration of a protein from first physical principles. In practice, this problem is far beyond the reach of modern computation and substantial simplifications are required to render these approaches tractable.

One molecular dynamics method that has demonstrated great success in recent years is `Rosetta` [15, 14]. `Rosetta` uses a database of three and four amino acid fragments and their known patterns of association to predict protein structure from primary amino acid sequence.

1.8.4 BetaWrap

`BetaWrap` is a specialized variant of a threading method developed by Phil Bradley, Bonnie Berger, Lenore Cowen, Jonathan King, and Matthew Menke [10]. The goal of `BetaWrap` is to find sequences that are compatible with the Right-Handed Parallel β-Helix fold.

`BetaWrap` differs from other structure prediction methods because it bases its prediction primarily upon the observed correlations of hydrogen bonded residues in β-sheets. As we mentioned previously, residue side-chains in β-sheets display marked statistical preferences for paired packing interactions [46, 47]. `BetaWrap` uses these preferences to determine the most probable overlap of prospective β-strands with one another. The `BetaWrap` algorithm integrates this technique with specialized expert knowledge to successfully predict the occurrence of the Right-Handed Parallel β-Helix fold. This is the first instance of residue correlation in β-sheets being used for protein structure prediction purposes. We will return to `BetaWrap`'s methods in Chapter 3.

Figure 1-12: **A Right-Handed Parallel β-Helix.** This is a portion of a Right-Handed Parallel β-Helix. The three faces are shown in red, blue, and yellow. The protein is a single chain.

1.8.5 Sequence-Structure Methods

In the past several years, several methods have emerged that analyze protein folds by simultaneously considering both their sequence and their structure. In these methods, protein families are first screened to identify sequence positions that are well-conserved (in essence, creating a profile for the family). Proteins with known structure in the

family are then compared to identify structural positions with conserved sequence-structure patterns or correlated mutations. Two automated methods that have met with considerable success in this domain are `Trilogy` [11] and `Conservatism-of--Conservatism` [49].

Chapter 2

Sequence Analysis

2.1 Introduction

In the first chapter of this thesis we presented an overview of computational methods for the analysis of proteins. In this chapter, we turn our attention to a particular protein fold: the Triple β-Spiral. We apply several of the methods that we discussed in the first chapter to this fold, and we discuss the results of these experiments. We also develop several novel modifications to existing computational methods and apply them to this fold.

Our ultimate goal in this chapter and the next one is to find as yet uncharacterized instances of the Triple β-Spiral in protein sequence databases – i.e. address the "inverse-folding problem." Along the way, however, we will provide a detailed description of the morphology and function of the Triple β-Spiral and its role in viral pathogenesis. We will also discuss how its recent crystallization has changed our understanding of the fold sequence, and we will critically reexamine studies of the Triple β-Spiral that took place before the structure was known.

In this chapter, we focus our attention mainly on an analysis of the Triple β-Spiral protein sequence. That is, we apply homology-based methods like `PSI-BLAST` and `HMMER` that rely solely on linear amino acid sequence for their results. Although our analysis is informed (and in some cases enhanced) by a prior knowledge of the fold structure, we do not explicitly take this structure into account. The next chapter

incorporates structural information into these homology-based methods for a more complete analysis of the Triple β-Spiral.

2.2 Fold Morphology and Function

The Triple β-Spiral is a protein fold that was first identified and classified by Mark J. van Raaij and coworkers in 1999 [63]. The Triple β-Spiral is a processive homotrimer consisting of three identical interacting protein chains.

At present, there are two instances of the Triple β-Spiral fold with solved structure in the PDB. The first is the penton fiber from Human Adenovirus Serotype 2 (Ad2) and the second is the σ1 attachment protein from the Dearing strain of Reovirus (Rσ1) [17]. In both instances, the three protein chains that compose the Triple β-Spiral are the product of a single gene.

In both Ad2 and Rσ1, the Triple β-Spiral fold serves as a fibrous connector from the main virus capsid to a C-terminal knob that binds to host cell-surface receptor proteins (see Figure 2-1). In Ad2, the sole fibrous fold in this fiber is the Triple β-Spiral. In Rσ1, the fiber consists of both the Triple β-Spiral fold and other fibrous domains, probably a three-stranded α-helical coiled-coil [65, 6].

The Triple β-Spiral is a remarkably stable fold. It is resistant to heat, protease, and detergent under most conditions [63]. This stability is a common theme among viral capsid and attachment proteins, and has undoubtedly evolved in response to host immune response and the demands of harsh extracellular conditions.

2.2.1 Triple β-Spiral Structure

Each of the three identical chains of the Triple β-Spiral is composed of a series of repeated structural elements (see Figure 2-2). Each of these structural elements is composed of:

1. A β-strand that runs parallel to the fiber axis

2. A long solvent-exposed loop, and

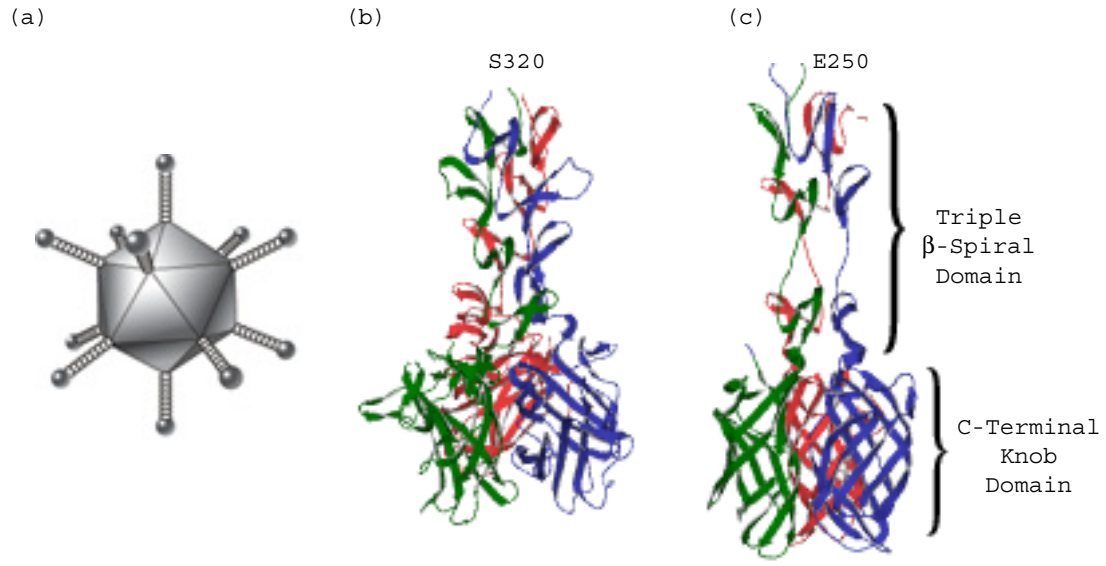

Figure 2-1: **Triple β-Spiral Fibers.** (a) A representation of an icosahedral virus capsid with 12 fibers extending from the 12 vertices. (b) A portion of the Human Adenovirus serotype 2 attachment fiber. Both the Triple β-Spiral shaft and the C-terminal knob are shown. This is only a fragment of the shaft (beginning at residue 319) as the rest of the fiber was not crystallized. (c) A portion of the Reovirus $\sigma 1$ attachment protein showing the Triple β-Spiral domain and the C-terminal knob. Note the flexible spacer inserted into the Triple β-Spiral domain.

3. A second β-strand antiparallel to the first, and slightly skewed to the fiber axis

Successive structural elements along the same chain are connected together by a tight β-turn. The three chains of the Triple β-Spiral wrap tightly around one another with a slight right-handed twist. Approximately seven repeated structural elements constitute one full turn about the fiber axis.

The repeated structural element on each chain engage in hydrogen bonding and hydrophobic interactions with the identical structural element on its two sister chains, and also with the previous, and next sequential repeats on the same chain. The result is a long fiber (approximately 15 Å in diameter) with a buried hydrophobic core and a covering of solvent exposed loops. Figure 2-3 shows the hydrogen bonding pattern within two successive structural elements on the same chain and between identical structural elements on sister chains.

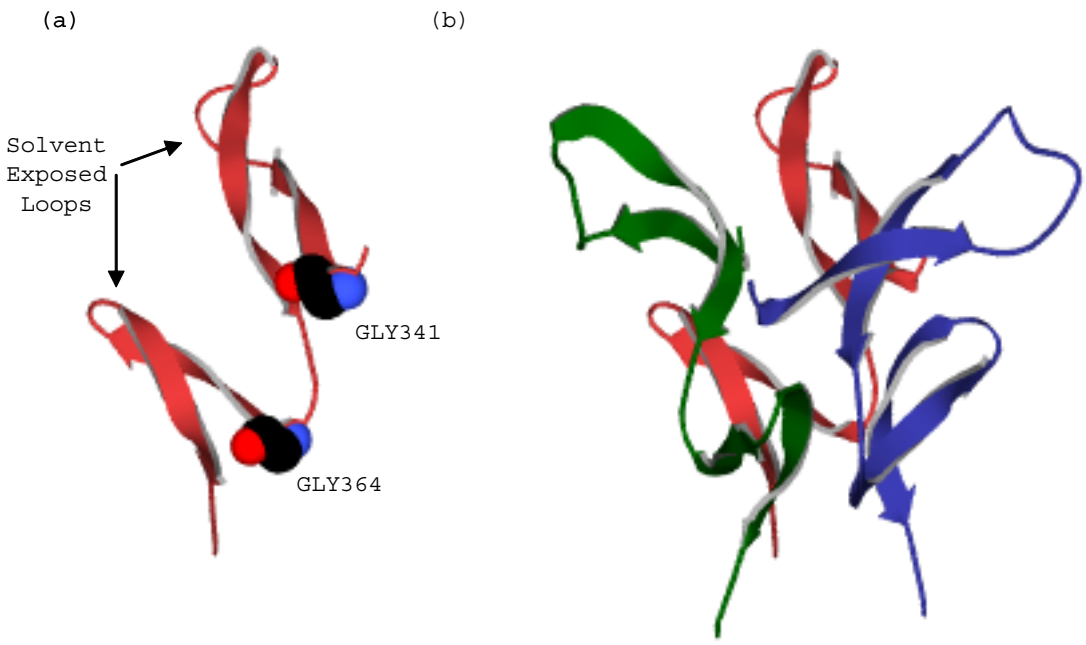

Figure 2-2: **Triple β-Spiral Structural Repeats.** (a) Two structural repeats from the Ad2 Triple β-Spiral fold. The Glycine's at the β-turn positions are highlighted. (b) Three identical double repeats from the Ad2 Triple β-Spiral showing the packed conformation of the fiber.

2.2.2 Adenovirus Family

Adenoviruses form a diverse group of human and animal viruses. In humans, Adenoviruses are responsible for infections such as pneumonia, cystitis, conjunctivitis, and one form of the common cold [50]. Human Adenovirus serotype 2 (Ad2) has been implicated as an agent of myocarditis [33].

The Adenovirus carries its genetic material as double-stranded DNA (dsDNA) inside of an icosahedral capsid (see Figure 2-1). Protruding from each of the twelve vertices of this capsid is the long fiber that contains the Triple β-Spiral domain and the C-terminal knob. This fiber is attached to the capsid by a domain at the N-terminal end of the fiber protein.

There are twenty-five Adenovirus fiber sequences in the `Swiss-Prot` database. Of these, only `FIBP_ADE02` (Ad2) has solved structure, but it is *very* likely that the other Adenovirus fiber proteins also contain the Triple β-Spiral fold. In almost every Adenovirus, the sequences of the twelve attachment fibers are identical to one another.

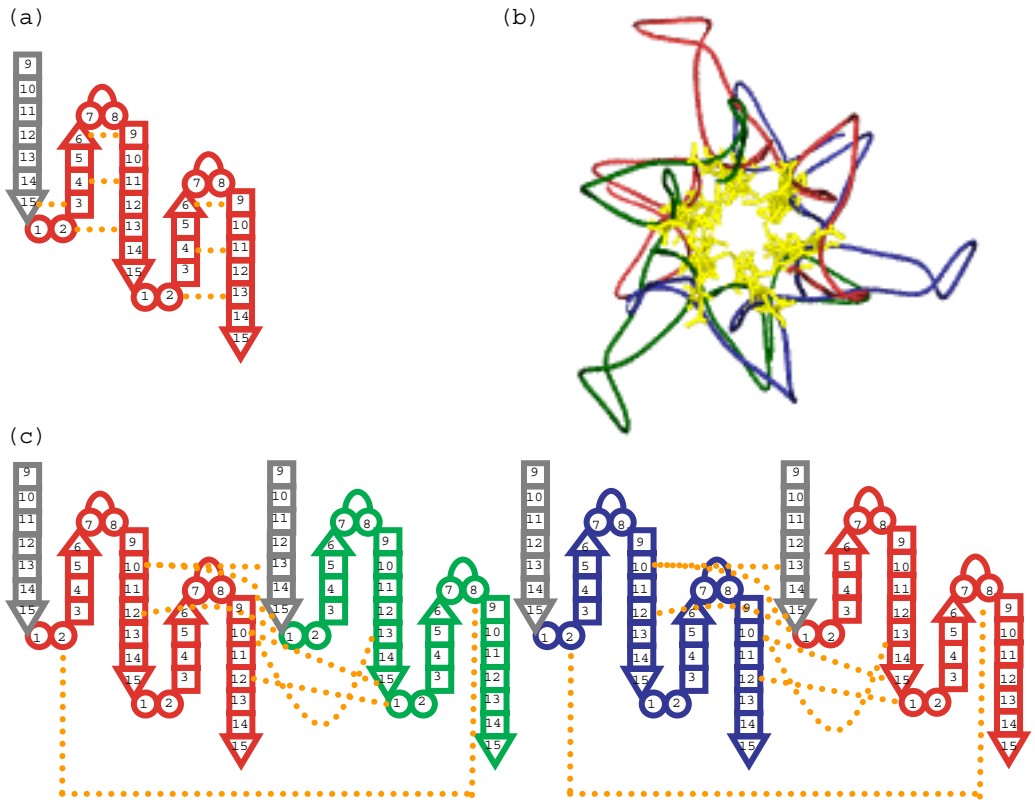

Figure 2-3: **Triple β-Spiral Hydrogen Bonding Pattern.** (a) A schematic representation of two full repeats of the Triple β-Spiral fold (in red) showing the intra-chain hydrogen bonding pattern of the Triple β-Spiral repeats. (b) A view of the Adenovirus Triple β-Spiral fiber axis. Hydrophobic residue side chains forming the core of the molecule are shown in yellow. (c) A schematic representation of the inter-chain hydrogen bonding pattern among identical repeat elements from the three chains. Chains are shown in red, green, and blue. The second red element is the same as the first, and is shown to demonstrate the spatial arrangement of the three structural units. For clarity, hydrogen bonds between the second and third chains have been omitted, but they are identical to the patterns shown.

In three cases, however, there are two variants of the fiber protein in the Adenovirus genome. These two variants still form homotrimeric fibers, but the fibers are evenly distributed between the two types (six of each for the twelve total icosahedral vertices). It is not known what advantage, if any, the dual fibers confer.

Name	ID	Type	Length
FIB1_ADE40	P18047	Human adenovirus type 40	547
FIB1_ADE41	P14267	Human adenovirus type 41	562
FIB1_ADEG1	Q64761	Avian adenovirus gal1 (strain Phelps) (Fowl adenovirus 1)	710
FIB2_ADE40	P18048	Human adenovirus type 40	387
FIB2_ADE41	P16883	Human adenovirus type 41	387
FIB2_ADEG1	Q64762	Avian adenovirus gal1 (strain Phelps) (Fowl adenovirus 1)	410
FIBP_ADE02	P03275	Human adenovirus type 2	582
FIBP_ADE03	P04501	Human adenovirus type 3	319
FIBP_ADE04	P36844	Human adenovirus type 4	426
FIBP_ADE05	P11818	Human adenovirus type 5	581
FIBP_ADE07	P15141	Human adenovirus type 7	343
FIBP_ADE08	P36845	Human adenovirus type 8	362
FIBP_ADE09	P36846	Human adenovirus type 9	362
FIBP_ADE12	P36711	Human adenovirus type 12	587
FIBP_ADE15	P36847	Human adenovirus type 15	367
FIBP_ADE1A	P35773	Human adenovirus type 11 (Ad11A) (strain BC34)	325
FIBP_ADE1P	P35774	Human adenovirus type 11 (Ad11P) (strain Slobiski)	325
FIBP_ADE31	P36848	Human adenovirus type 31	556
FIBP_ADEB3	Q03553	Bovine adenovirus type 3 (Mastadenovirus bos3)	976
FIBP_ADECC	Q65961	Canine adenovirus type 1 (strain CLL)	543
FIBP_ADECG	P22230	Canine adenovirus type 1 (strain Glaxo)	543
FIBP_ADECR	Q96689	Canine adenovirus type 1 (strain RI261)	543
FIBP_ADECT	Q65914	Canine adenovirus type 2 (strain Toronto A 26-61)	542
FIBP_ADEM1	P19721	Mouse adenovirus type 1 (MAV-1)	613
FIBP_ADEP3	Q83457	Porcine adenovirus type 3 (PAV-3)	448
VSI1_REOVD	P03528	Reovirus (type 3 / strain Dearing)	455
VSI1_REOVJ	P04507	Reovirus (type 2 / strain D5/Jones)	462
VSI1_REOVL	P04506	Reovirus (type 1 / strain Lang)	470

Table 2.1: **Triple β-Spiral Fiber Sequences in `Swiss-Prot`.** This is a list of all of the Adenovirus and Reovirus sequences in the `Swiss-Prot` database and the organisms they infect. The length is the number of amino acids in the fiber.

```
>sp|P03275|1qiu|FIBP_ADE02 Fiber protein - Human adenovirus type 2.
MKRARPSEDTFNPVYPYDTETGPPTVPFLTPPFVSPNGFQESPPGVLSLRVSEPLDTSHG
MLALKMGSGLTLDKAGNLTSQNVTTVTQPLKKTKSNISLDTSAPLTITSGALTVATTAPL
IVTSGALSVQSQAPLTVQDSKLSIATKGPITVSDGKLALQTSAPLSGSDSDTLTVTASPP
LTTATGSLGINMEDPIYVNNGKIGIKISGPLQVAQNSDTLTVVTGPGVTVEQNSLRTKVA
GAIGYDSSNNMEIKTGGGMRINNNLLILDVDYPFDAQTKLRLKLGQGPLYINASHNLDIN
YNRGLYLFNASNNTKKLEVSIKKSSGLNFDNTAIAINAGKGLEFDTNTSESPDINPIKTK
IGSGIDYNENGAMITKLGAGLSFDNSGAITIGNKNDDKLTLWTTPDPSPNCRIHSDNDCK
FTLVLTKCGSQVLATVAALAVSGDLSSMTGTVASVSIFLRFDQNGVLMENSSLKKHYWNF
RNGNSTNANPYTNAVGFMPNLLAYPKTQSQTAKNNIVSQVYLHGDKTKPMILTITLNGTS
ESTETSEVSTYSMSFTWSWESGKYTTETFATNSYTFSYIAQE
```

Figure 2-4: **Ad2 Sequence.** The full sequences in FASTA format for the Ad2 attachment fiber. Portions of the sequences with solved crystal structure (319-582) are shown with a yellow background. The fiber portion (53-394) of the molecule is shown in red.

2.2.3 Reovirus Family

Like Adenoviruses, Reoviruses infect both humans and animals. In humans, Reoviruses primarily infect children and cause respiratory and gastrointestinal illnesses [17]. In contrast to Adenovirus, however, Reovirus carries its genetic material as double-stranded RNA (dsRNA).

```
>sp|P03528|1kke|VSI1_REOVD Sigma 1 protein precursor - Reovirus type 3.
MDPRLREEVVRLIIALTSDNGASLSKGLESRVSALEKTSQIHSDTILRITQGLDDANKRI
IALEQSRDDLVASVSDAQLAISRLESSIGALQTVVNGLDSSVTQLGARVGQLETGLADVR
VDHDNLVARVDTAERNIGSLTTELSTLTLRVTSIQADFESRISTLERTAVTSAGAPLSIR
NNRMTMGLNDGLTLSGNNLAIRLPGNTGLNIQNGGLQFRFNTDQFQIVNNNLTLKTTVFD
SINSRIGATEQSYVASAVTPLRLNSSTKVLDMLIDSSTLEINSSGQLTVRSTSPNLRYPI
ADVSGGIGMSPNYRFRQSMWIGIVSYSGSGLNWRVQVNSDIFIVDDYIHICLPAFDGFSI
ADGGDLSLNFVTGLLPPLLTGDTEPAFHNDVVTYGAQTVAIGLSSGGAPQYMSKNLWVEQ
WQDGVLRLRVEGGGSITHSNSKWPAMTVSYPRSFT
```

Figure 2-5: **Rσ1 Sequence.** The full sequences in FASTA format for the Rσ1 attachment fiber. Portions of the sequences with solved crystal structure (247-455) are shown with a yellow background. The Triple β-Spiral portion (175-294) of the molecule is shown in red.

Like the Adenovirus, the Reovirus capsid is an icosahedron, with fibers extending from the twelve capsid vertices. These fibers attach to the capsid at their N-terminal end, and exhibit a knob that has been implicated in host cell-receptor binding. As we mentioned previously, the Rσ1 fibers are thought to contain a trimeric α-helical coiled-coil in addition to the Triple β-Spiral fold. Only the C-terminal end of the fiber

has been crystallized. This is the part of the fiber that contains the Triple β-Spiral and C-terminal knob shown in Figure 2-1.

There are three Reovirus serotypes with sequences in the `Swiss-Prot` database. These three serotypes are Dearing, Lang, and Jones (with `Swiss-Prot` id's respectively of `VSI1_REOVD`, `VSI1_REOVL`, and `VSI1_REOVJ`.) It is the Dearing strain of the Reovirus whose structure has been determined [17].

2.2.4 Sequence Alignments

Following the methods outlined in Chapter 1, the natural first step after determining the similar structure of the Ad2 and Rσ1 fibers is to determine their degree of sequence similarity. As previously mentioned, greater sequence similarity indicates that sequences are less evolutionarily diverged. Sequences with similarity greater than 40% are presumed to fold into the same tertiary formation, and are by extension, considered not to be sufficiently distinct for rigorous cross-validation analysis.

Figure 2-6 presents the results of a sequence alignment generated by `ClustalW` for the two solved Triple β-Spiral structures [61]. Table 2.2 provides the same information in summary form for all of the Adenovirus and Reovirus Fiber sequences in `Swiss-Prot`. Note that the Ad2 and Rσ1 structures exhibit a weak but significant sequence similarity ($> 40\%$). This indicates that sensitive homology modeling tools should be able to detect this similarity in sequence databases.

In spite of the sequence similarity between the Ad2 and Rσ1 fibers, we will treat these two sequences separately. We continue in this way because most existing homology-modeling tools fail to detect the relationship between Ad2 and Rσ1 in sequence database searches. We will explore these methods and why they fail more fully later in this chapter. Although strict computational validation is not possible, we hope that if we provide reasonable candidates for membership in this fold, then wet-lab biologists may further test the structural properties of our candidates.

The weak but significant sequence similarity between the Ad2 and Rσ1 folds has important implications for the computational analysis of these sequences. Because these folds evidence a sequence similarity greater than 40%, it is not possible for us

	FIB1_ADE40	FIB1_ADE41	FIB1_ADEG1	FIB2_ADE40	FIB2_ADE41	FIB2_ADEG1	FIBP_ADE02	FIBP_ADE03	FIBP_ADE04	FIBP_ADE05	FIBP_ADE07	FIBP_ADE08	FIBP_ADE09	FIBP_ADE12	FIBP_ADE15	FIBP_ADE1A	FIBP_ADE1P	FIBP_ADE31	FIBP_ADEB3	FIBP_ADECC	FIBP_ADECG	FIBP_ADECR	FIBP_ADECT	FIBP_ADEM1	FIBP_ADEP3	VSI1_REOVD	VSI1_REOVJ	VSI1_REOVL
FIB1_ADE40	1.00	0.98	0.45	0.65	0.66	0.54	0.56	0.71	0.60	0.56	0.57	0.70	0.64	0.61	0.63	0.61	0.61	0.57	0.57	0.47	0.47	0.47	0.49	0.49	0.48	0.55	0.57	0.45
FIB1_ADE41	0.98	1.00	0.48	0.70	0.70	0.57	0.55	0.71	0.56	0.54	0.48	0.64	0.61	0.61	0.66	0.58	0.62	0.55	0.56	0.44	0.44	0.44	0.49	0.51	0.53	0.55	0.52	0.52
FIB1_ADEG1	0.45	0.48	1.00	0.48	0.47	0.73	0.41	0.61	0.47	0.46	0.57	0.54	0.56	0.43	0.53	0.57	0.57	0.46	0.53	0.47	0.47	0.47	0.47	0.48	0.53	0.61	0.51	0.48
FIB2_ADE40	0.65	0.70	0.48	1.00	0.98	0.49	0.74	0.63	0.46	0.75	0.53	0.70	0.62	0.67	0.63	0.52	0.54	0.61	0.59	0.55	0.55	0.55	0.55	0.52	0.47	0.43	0.47	0.47
FIB2_ADE41	0.66	0.70	0.47	0.98	1.00	0.49	0.72	0.64	0.49	0.73	0.52	0.62	0.61	0.71	0.64	0.52	0.53	0.66	0.56	0.54	0.54	0.54	0.55	0.52	0.47	0.42	0.48	0.45
FIB2_ADEG1	0.54	0.57	0.73	0.49	0.49	1.00	0.54	0.58	0.44	0.47	0.59	0.43	0.43	0.49	0.43	0.58	0.57	0.52	0.63	0.49	0.49	0.49	0.51	0.54	0.49	0.49	0.43	0.41
FIBP_ADE02	0.56	0.55	0.42	0.74	0.72	0.54	1.00	0.68	0.51	0.87	0.66	0.71	0.64	0.54	0.70	0.67	0.56	0.57	0.53	0.47	0.47	0.47	0.51	0.45	0.52	0.58	0.52	0.48
FIBP_ADE03	0.71	0.71	0.61	0.63	0.64	0.58	0.68	1.00	0.51	0.74	0.77	0.55	0.60	0.61	0.58	0.78	0.77	0.69	0.68	0.64	0.64	0.64	0.61	0.69	0.59	0.53	0.58	0.53
FIBP_ADE04	0.60	0.56	0.47	0.46	0.49	0.44	0.51	0.51	1.00	0.57	0.54	0.44	0.43	0.54	0.44	0.54	0.57	0.54	0.55	0.54	0.54	0.54	0.50	0.55	0.49	0.42	0.40	0.43
FIBP_ADE05	0.56	0.54	0.46	0.75	0.73	0.47	0.87	0.74	0.57	1.00	0.71	0.63	0.67	0.52	0.69	0.65	0.68	0.59	0.50	0.49	0.49	0.49	0.50	0.46	0.49	0.57	0.49	0.46
FIBP_ADE07	0.57	0.48	0.57	0.53	0.52	0.59	0.66	0.77	0.54	0.71	1.00	0.53	0.57	0.65	0.57	0.96	0.96	0.59	0.75	0.62	0.62	0.62	0.61	0.54	0.56	0.46	0.49	0.57
FIBP_ADE08	0.70	0.64	0.54	0.70	0.62	0.43	0.71	0.55	0.44	0.63	0.53	1.00	0.95	0.71	0.93	0.56	0.56	0.58	0.61	0.65	0.65	0.65	0.65	0.57	0.52	0.47	0.39	0.36
FIBP_ADE09	0.64	0.61	0.56	0.62	0.61	0.43	0.64	0.60	0.43	0.67	0.57	0.95	1.00	0.61	0.95	0.58	0.59	0.58	0.60	0.61	0.61	0.61	0.63	0.58	0.48	0.46	0.37	0.35
FIBP_ADE12	0.61	0.61	0.43	0.67	0.71	0.49	0.54	0.61	0.54	0.52	0.65	0.70	0.61	1.00	0.67	0.62	0.65	0.88	0.58	0.44	0.44	0.44	0.48	0.47	0.52	0.48	0.57	0.52
FIBP_ADE15	0.63	0.66	0.53	0.63	0.64	0.43	0.74	0.58	0.44	0.69	0.57	0.93	0.95	0.68	1.00	0.56	0.58	0.61	0.58	0.59	0.59	0.59	0.63	0.61	0.51	0.46	0.37	0.35
FIBP_ADE1A	0.61	0.58	0.57	0.52	0.52	0.58	0.67	0.78	0.54	0.65	0.96	0.56	0.58	0.62	0.56	1.00	0.95	0.62	0.72	0.63	0.63	0.63	0.58	0.57	0.57	0.49	0.54	0.52
FIBP_ADE1P	0.61	0.62	0.57	0.54	0.53	0.57	0.48	0.77	0.57	0.68	0.96	0.56	0.59	0.65	0.58	0.95	1.00	0.59	0.73	0.57	0.57	0.57	0.54	0.56	0.62	0.46	0.51	0.47
FIBP_ADE31	0.57	0.55	0.46	0.61	0.66	0.52	0.57	0.69	0.54	0.59	0.59	0.58	0.58	0.88	0.61	0.57	0.59	1.00	0.55	0.42	0.42	0.42	0.39	0.45	0.47	0.55	0.53	0.44
FIBP_ADEB3	0.57	0.56	0.53	0.59	0.56	0.63	0.53	0.68	0.55	0.50	0.73	0.61	0.60	0.58	0.58	0.72	0.73	0.55	1.00	0.56	0.56	0.56	0.56	0.57	0.58	0.55	0.52	0.48
FIBP_ADECC	0.47	0.44	0.47	0.55	0.54	0.49	0.47	0.64	0.54	0.50	0.62	0.65	0.61	0.44	0.59	0.63	0.56	0.42	0.56	1.00	1.00	1.00	0.91	0.45	0.55	0.48	0.48	0.45
FIBP_ADECG	0.47	0.44	0.47	0.55	0.54	0.49	0.47	0.64	0.54	0.50	0.62	0.65	0.61	0.44	0.59	0.63	0.56	0.42	0.56	1.00	1.00	1.00	0.91	0.45	0.55	0.48	0.48	0.45
FIBP_ADECR	0.47	0.44	0.47	0.55	0.54	0.49	0.47	0.64	0.54	0.50	0.62	0.65	0.61	0.44	0.59	0.63	0.56	0.42	0.56	1.00	1.00	1.00	0.91	0.45	0.55	0.48	0.48	0.45
FIBP_ADECT	0.49	0.49	0.47	0.55	0.55	0.51	0.51	0.61	0.50	0.50	0.61	0.65	0.63	0.48	0.63	0.58	0.54	0.39	0.56	0.91	0.91	0.91	1.00	0.47	0.53	0.50	0.45	0.43
FIBP_ADEM1	0.49	0.51	0.49	0.52	0.52	0.52	0.45	0.69	0.55	0.46	0.54	0.57	0.58	0.47	0.61	0.57	0.56	0.45	0.57	0.45	0.45	0.45	0.47	1.00	0.50	0.50	0.49	0.48
FIBP_ADEP3	0.48	0.54	0.53	0.47	0.47	0.49	0.52	0.59	0.49	0.49	0.56	0.52	0.48	0.52	0.51	0.57	0.62	0.47	0.58	0.55	0.55	0.55	0.53	0.50	1.00	0.51	0.46	0.45
VSI1_REOVD	0.55	0.55	0.61	0.43	0.42	0.49	0.58	0.53	0.42	0.57	0.48	0.47	0.46	0.48	0.46	0.49	0.46	0.57	0.55	0.48	0.48	0.48	0.50	0.50	0.51	1.00	0.52	0.49
VSI1_REOVJ	0.57	0.52	0.51	0.47	0.48	0.43	0.52	0.58	0.40	0.49	0.49	0.39	0.37	0.57	0.37	0.54	0.51	0.53	0.52	0.48	0.48	0.48	0.45	0.49	0.46	0.52	1.00	0.74
VSI1_REOVL	0.45	0.52	0.48	0.47	0.45	0.41	0.48	0.51	0.43	0.46	0.57	0.36	0.35	0.52	0.35	0.52	0.47	0.44	0.48	0.46	0.46	0.46	0.43	0.48	0.45	0.49	0.74	1.00

Table 2.2: **Adenovirus and Reovirus Sequence Similarity.** This table summarizes the sequence similarity between each of the Adenovirus and Reovirus Triple β-Spiral sequences in `Swiss-Prot`. The alignments were generated with `ClustalW` and the degree of sequence similarity was calculated by dividing the number of similar residues by the length of the shorter sequence. Note that three of the Adenovirus sequence (CC, CR, CT) are identical in their Triple β-Spiral regions. They differ in other parts of the fiber chains, so we include them as separate entries.

```
CLUSTAL W (1.81) multiple sequence alignment


FIBP_ADE02      EPLDTSHGMLALKMGSGLTLDKAGNLTSQNVTTVTQPLKKTKSNISLDTSAPLTITSGAL
VSI1_REOVD      ------------------------------------------------------------


FIBP_ADE02      TVATTAPLIVTSGALSVQSQAPLTVQDSKLSIATKGPITVSDGKLALQTSAPLSGSDSDT
VSI1_REOVD      --------------------APLSIRNNRMTMGLNDGLTLSGNNLAIR------------
                                    ***::::.::::. :. :*:*..:**::

FIBP_ADE02      LTVTASPPLTTATGSLGINMEDPIYVNNGKIGIKISGPLQVAQNSDTLTVVTGPGVTVEQ
VSI1_REOVD      -----------LPGNTGLNIQN--------------GGLQFRFNTDQFQIVN--------
                         .*. *:*:::              * **.  *:* : :*.

FIBP_ADE02      NSLRTKVAGAIGYDSSNNMEIKTGGGMRINNNLLILDVDYPFDAQTKLRLKLGQGPLYIN
VSI1_REOVD      ----------------NNLTLKTTVFDSINSRIGATEQSYVASAVTPLRLNS--------
                              **: :**     **..:   : .*  .* * ***:

FIBP_ADE02      ASHNLDINYNRGLYLFNASNNTKKLEVSIKKSSGLNFDNTAIAINAGKGLEFDTNTSESP
VSI1_REOVD      --------------------STKVLDMLIDSST---------------------------
                                    .** *:: *..*:

FIBP_ADE02      DINPIKTKIGSGIDYNENGAMITKLGSGLSFDNSGAITIGNK
VSI1_REOVD      ---------------------------LEINSSGQLTVRST
                                           *.::.** :*: ..
```

Figure 2-6: **Triple β-Spiral sequence alignments.** An alignment of the Triple β-Spiral portions of the Ad2 and Rσ1 attachment proteins. The two proteins are 31% identical and 58% similar to one another. The alignment was constructed with `ClustalW`.

to present a rigorous leave-one-out cross-validation of methods that we develop to predict the occurrence of this fold. This is a very important point. In contrast with `BetaWrap`, which at the time of its implementation had seven distinct representative structures with less than 40% sequence similarity, the Triple β-Spiral fold has only one such representative.

The weak but significant similarity between the Ad2 and Rσ1 sequences does present one interesting question about their evolutionary relationship to one another. As we noted earlier, Adenovirus carries its genetic material as double-stranded DNA, whereas Reovirus carries its genetic material as double-stranded RNA. Usually, viruses that carry their genetic material in different forms are considered to be unrelated to one another, or if related, only through a very ancient evolutionary precursor. This suggests that the Triple β-Spiral fold has either spread laterally through widely divergent virus types, or that in fact the Triple β-Spiral is an ancient protein

fold [65].

2.2.5 Triple β-Spiral Sequence Repeats

Both the Ad2 and Rσ1 fibers contain a repeated sequence motif that corresponds to their structural repeat. Green and coworkers recognized and characterized this sequence repeat in several different Adenoviruses in 1983, almost 20 years before the structure of the Ad2 fiber was definitively determined [26]. Green's motif contained 15 residue positions, labeled *a* through *o*.

Now that the structure of the Ad2 fiber is known, we observe that the repeated sequence motif that Green characterized contained:

1. The second β-strand from one structural unit of the Triple β-Spiral fold,
2. The residues constituting the β-turn, and
3. The first β-strand from the following structural unit of the Triple β-Spiral fold.

Figure 2-7 gives the break-down of the Ad2 fiber sequence into repeated elements based upon Green's scheme. Note that the solvent exposed loop is not included in the sequence repeat motif. It was natural that Green would exclude the solvent exposed loop from the Triple β-Spiral sequence repeat because this loop region is variable in both length and residue composition. Before the structure of the Ad2 fiber was known, the loop region was an intuitive breaking point for the repeated motif.

Green's characterization of the Adenovirus was influential. Subsequent papers that discussed the Adenovirus sequence adopted his labeling scheme, and often referred to this repeat motif.[1] There were even several papers published suggesting (incorrect) structures based upon this repeated motif [60].

[1] In previous characterizations of the Adenovirus fiber sequences, researchers categorized the repeat pattern that Green had identified into two subtly different sub-groups: the G-repeat and the P-repeat. These repeat patterns are almost identical. The G-repeat is characterized by having a Glycine at positions `h` and `j`. The P-repeat has a Proline at `h` and a hydrophobic residue at `j`. These classifications are not well conserved in all Adenovirus and Reovirus fibers, though, and we do not use them here.

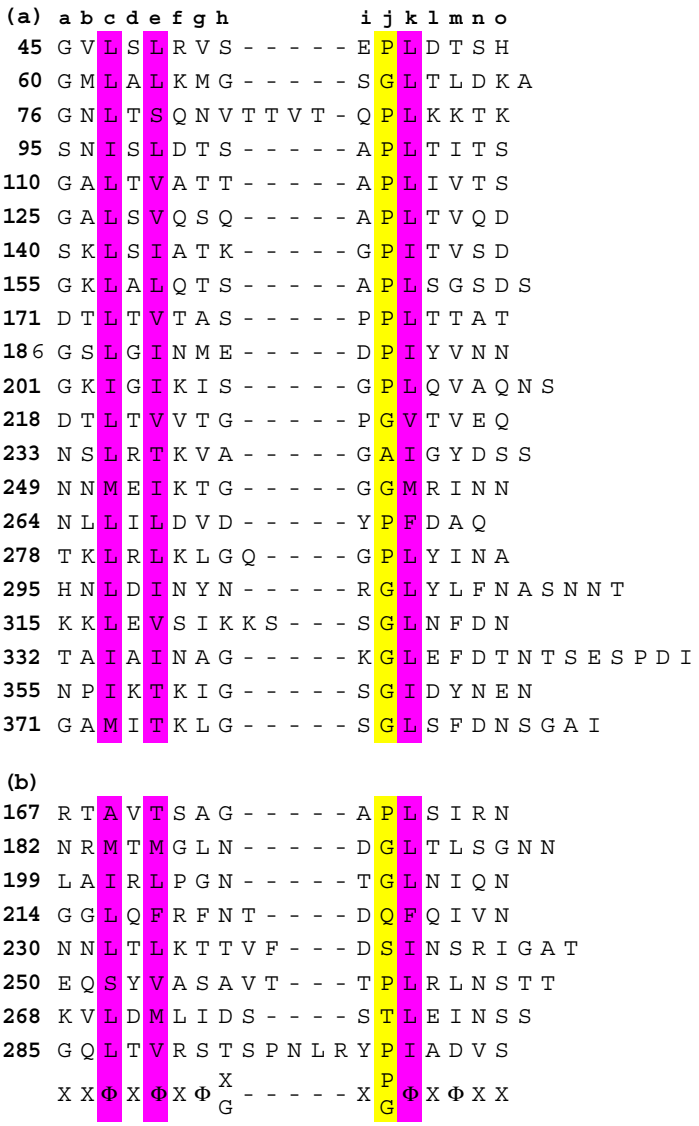

Figure 2-7: **Triple β-Spiral Repeats.** The sequence repeat pattern for the (a) Ad2 and (b) Rσ1 fibers. The bottom of the figure shows a canonical repeat pattern: `X` for any residue, Φ for hydrophobic residues and `PG` in key structural positions. The residues contributing to the hydrophobic core are shown in pink and the residues at the β-turn position are shown in yellow.

Of course, Green's characterization of the Adenovirus sequence repeat was not incorrect. It is the nature of a repeated sequence motif that one can choose any arbitrary starting point and find subsequent repeated elements based upon this starting point. The only flaw with this approach is that a poorly-chosen starting point will

lead to mis-characterization of the first and last elements of the chain. In fact, this problem did occur with the Adenovirus sequences, and for many years researchers referred to the "half-repeat" at the C-terminal end of the fiber shaft. With the benefit of hindsight, we now know that the C-terminal end of the Adenovirus contains a full structural repeat. Green had chosen a starting point that was one-half repeat too early in the sequence.

After solving the crystal structure of the Triple β-Spiral fold, van Raaij and coworkers noted the discrepancy between the previous sequence repeat motif and the natural structural repeat. Based upon the inter and intra-chain hydrogen bonding pattern in the Triple β-Spiral fold (see Figure 2-3), van Raaij and coworkers suggested that a sequence repeat of residues g through f would be a natural sequence repeat for the Triple β-Spiral fold. In this sequence repeat scheme, the solvent exposed loop is included in the middle of each repeat, and the two paired beta-strands flank this loop region. The insertion between positions h and i is incorporated into a single repeat element.

In this thesis, we have chosen a sequence repeat that differs slightly from the repeat suggested by van Raiij and coworkers (see Figure 2-8.) Rather than beginning our sequence repeat at g, we begin at position i. (We also relabeled the positions of the repeat motif to 1 through 15 instead of a through o). Like the motif suggested by van Raaij and coworkers, this scheme incorporates the solvent-exposed loop between the two paired beta-strands. Unlike their suggested motif, however, our motif treats the insertion between positions 15 and 1 (h and i) as being outside the repeat element. We think that this is a more intuitive way to view the fiber shaft and its sequence repeat. Pragmatically speaking, there is almost no difference between our characterization and the one suggested by van Raaij. We find that our repeats are more natural to work with and visualize.

This discussion points to the somewhat arbitrary nature of characterizing proteins that contain repeated structural or sequence elements. Often, we can analyze portions of the fold without taking into account the precise start and end points of the structure. For example, the `BetaWrap` program successfully characterized Right-Handed

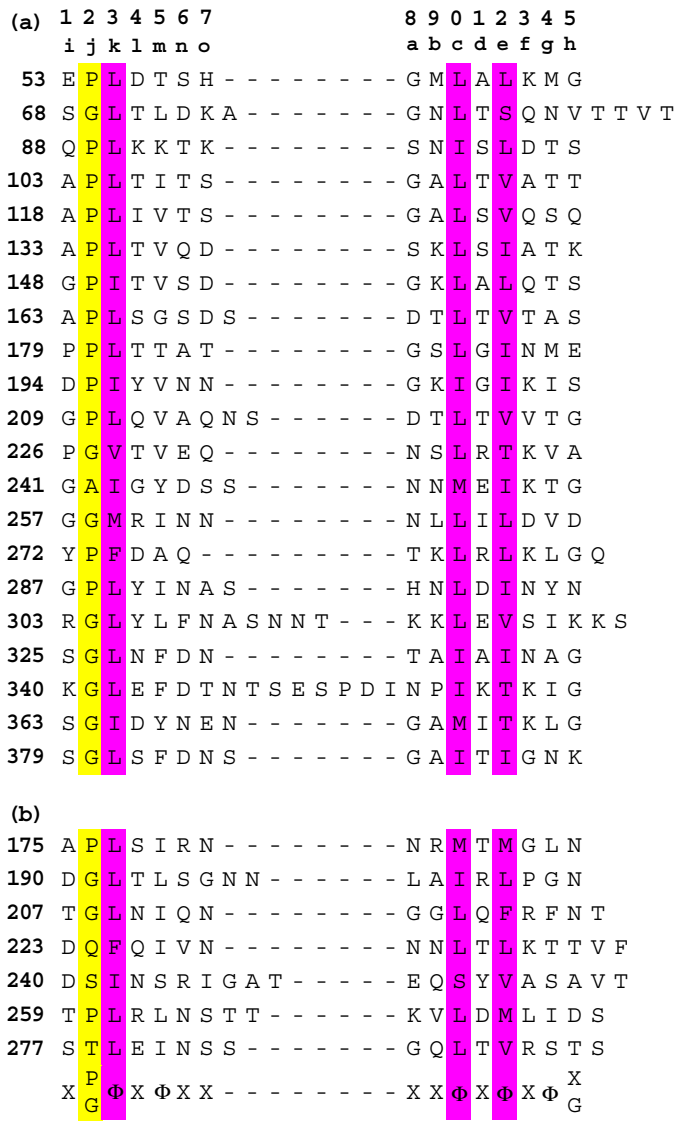

Figure 2-8: **Triple β-Spiral Repeats.** The sequence repeat pattern for the (a) Ad2 and (b) Rσ1 fibers. Note that we have relabeled the positions in the repeat from a-o to 1-15 (with 10-15 as 0-5 at the end). The bottom of the figure shows a canonical repeat pattern: `X` for any residue, Φ for hydrophobic residues and `PG` at key structural positions. The residues contributing to the hydrophobic core are shown in pink and the residues at the β-turn position are shown in yellow.

Parallel β-Helix proteins by analyzing only several components of a long, repetitive structure [10].

In spite of the somewhat arbitrary nature of selecting a repeating unit, we think

that it is important to define the sequence and structural repeat of the Triple β-Spiral fold to include the paired β-strands and the solvent exposed loop. Because the Triple β-Spiral fiber almost certainly arose through multiple duplication of a repeated structural element, the stabilizing hydrogen bonds and solvent interactions of this configuration suggest that it is this element that is duplicated. Any other duplication would leave a less stable, fragmentary structural element. In addition, we will see as we progress through our analysis that characterizing the repeat correctly lends additional insight into the Triple β-Spiral fold. For example, we will observe a biased residue composition in the solvent-exposed loop region that previous analyses neglected to observe. We suspect that previous analyses did not report this residue bias because the solvent-exposed loop residues were not included in Green's original repeat motif.

2.2.6 Automated Discovery of Sequence Repeats

Several existing computational tools detect sequence repeats in protein sequence databases [20, 4, 28, 45]. Because the Triple β-Spiral evidences a repeated sequence motif, we tested two of these tools to determine whether they could detect the repeated regions in the Triple β-Spiral fold.

The first tool that we tested was Coward and Drabløs' Periodicity Tester Program. Figure 2-9 shows a sample of the output from this program. In this figure, fragment lengths with a low p-value indicate a potential internal sequence repeat. If this pattern is repeated at integral multiples of the the length, then there is a strong indication of a Unfortunately, this program detects no clear repeat pattern in the Ad2 or Rσ1 sequences.

The second tool that we tested was Heger and Holm's Rapid Automatic Detection and Alignment of Repeats (`RADAR`) program. This program detected 6 double-repeat regions (30 residues each) for the Ad2 fiber and 44 for the Rσ1 fiber. Figure 2-10 shows the results.

Overall, these programs did not effectively identify the Triple β-Spiral sequence repeat, and we are doubtful that existing automated methods are the equal of Green's

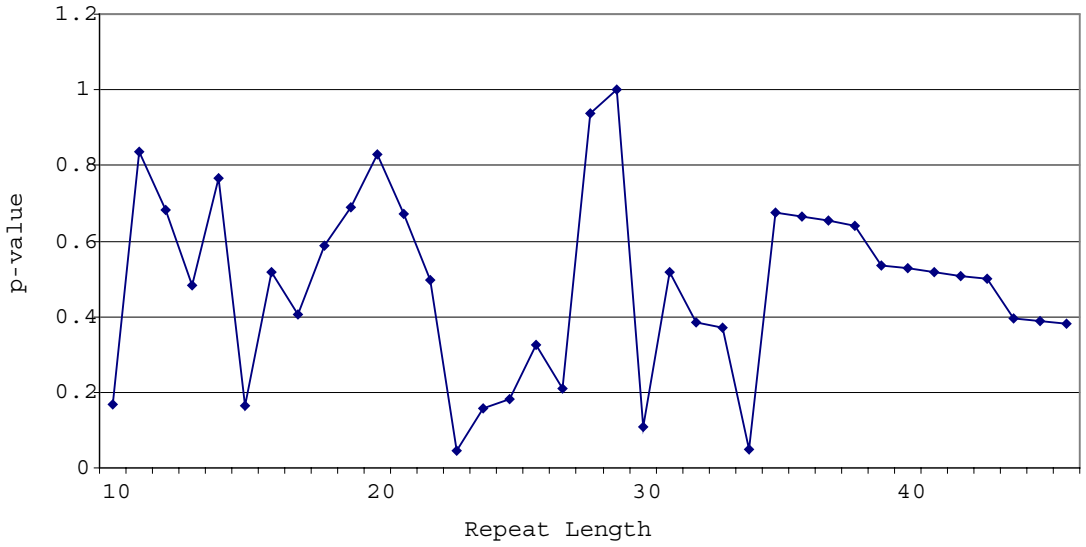

Figure 2-9: **Periodicty Finder Results.** The output from running the Periodicity Finder program on the Ad2 fiber repeat. Note that the p-value should attain a minimum for expected repeat lengths. Thus, for the Adenovirus fibers we would expect clear minima at 15, 30, and 45 residues.

```
(a)
   45-62  ----------GVLSLRVSEPLDT-SHGML
   65-82  K------MG-SG-LTLDKAGNLTS-QN---
  87-112  --TQPLKKT-KSNISLDTSAPLTI-TSGAL
 115-142  ATTAPLIVT-SGALSVQSQAPLTV-QDSKL
 160-188  QTSAPLSGSdSDTLTVTASPPLTT-ATGSL
 191-217  NMEDPIYVN-NGKIGIKISGPLQVaQNS--

(b)
   20-63  NGASLS-K-GLESRVSALEKT-SQI---HSDTIL-----RITQGLDDANKRIIAL
  96-140  NGLDSSvT-QLGARVGQLETGlADVrvdHDNLVA-----RV----DTAERNIGSL
 159-192  ----------ESRISTLERT-AVT---SAGAPLsirnnRMTMGLND------GL
 207-248  TGLNIQ-NgGLQFRFNT-DQF-QIV---NNNLTL-----KTT-VFDSINSRIGA-
```

Figure 2-10: **Results of RADAR.** The results of running the RADAR program on the (a) Ad2 and (b) Rσ1 fiber sequences. RADAR detects 6 total repeat regions in Ad2 (out of 21 total). The repeat regions do not coincide well with Green's repeats. RADAR detects 4 total repeats in Rσ1, though these repeats do not coincide with the Rσ1 repeats. We show only the top scoring repeat group for RADAR.

original observations. We suspect that both of these methods perform poorly with the repeat pattern in the Triple β-Spiral fold because of the variable length loop regions and insertions between the β-turn. These inserted regions break the integral periodicty of the Triple β-Spiral repeat pattern.

2.3 Computational Analysis of the Triple β-Spiral

In this section we will present a computational analysis of the sequence of the Triple β-Spiral fold. We begin our analysis by presenting a relatively simple regular expression based search tool. We then proceed to more advanced homology-modeling methods.

Throughout this chapter, we will develop a series of progressively more advanced models for predicting the Triple β-Spiral fold from primary sequence data. To organize our discussion, we will number each model as it is introduced. Because every model that we present will require a set of training sequences, we split each model into two sub-models. The first sub-model is trained using Adenovirus fiber sequences. The second is trained using Reovirus fiber sequences. In cases where it is relevant, we specify which Adenovirus and Reovirus fiber sequences we used to create the model.

We will test each of the models that we develop by using it to search both the `Swiss-Prot` and `PDB` databases. For the `Swiss-Prot` database, we will record the following information:

1. Which of the 25 Adenovirus Fiber proteins the model finds,
2. Which of the 3 Reovirus Fiber proteins the model finds, and
3. How many other proteins the model finds and what they are.

For the `PDB` database we will record the the rank of both the Ad2 (`1QIU`) and Rσ1 (`1KKE`) fiber structures. Table 2.3 provides an example of the summary results for each method. At the end of this chapter, we will also present a table (Table 2.14) that summarizes and compares all of the models that we present.

2.3.1 Model 1: Regular-Expression Search

As we discussed in Chapter 1, the `PROSITE` database contains a large set of regular expressions that occur in proteins. Because the Triple β-Spiral fold contains a repeated pattern that is easy to characterize with a regular expression, we might expect the `PROSITE` database to contain a consensus motif for this fold. Unfortunately,

Fiber	Adenovirus	Reovirus
FIBP_ADE02	X	
FIBP_ADE03	X	
FIBP_ADE04	X	
FIBP_ADE05	X	
FIBP_ADE07	X	
FIBP_ADE08	X	
FIBP_ADE09	X	
FIBP_ADE12	X	
FIBP_ADE15	X	
FIBP_ADE31	X	
FIBP_ADE1A	X	
FIBP_ADE1P	X	
FIBP_ADEB3	X	
FIBP_ADECC	X	
FIBP_ADECG	X	
FIBP_ADECR	X	
FIBP_ADECT	X	
FIBP_ADEP3	X	
FIBP_ADEM1	X	
FIB1_ADE40	X	
FIB2_ADE40	X	X
FIB1_ADE41	X	X
FIB2_ADE41	X	
FIB1_ADEG1	X	
FIB2_ADEG1		
VSI1_REOVD		X
VSI1_REOVL	X	
VSI1_REOVJ		X
1QIU (PDB Rank)	1	-
1KKE (PDB Rank)	345	1
Number of Repeats	56	7
Additional Hits	1233	89

Table 2.3: **Template Summary Output Table.** This table is an example of how we will summarize the output from a model. The first column summarizes the output from the model trained with Adenovirus sequences, and the second column summarizes the output from the model trained with Reovirus sequences. The first 28 rows contain an `X` if the model finds that sequence in `Swiss-Prot`. If that fiber was used to train the model then the `X` is red. The next two rows summarize the performance of the model on the `PDB` database. In this example, the Adenovirus-trained model found the Reovirus fiber sequence (`1KKE`) in the `PDB` with a rank of 345 and the Reovirus-trained model did not find `1QIU` at all. The next row summarizes the total number of sequence repeats found by the model in the `Swiss-Prot` database. Repeats are only counted for Adenovirus fibers and Reovirus fibers in their respective columns. The last row has the number of additional hits (outside of Adenovirus and Reovirus fibers) that this model matched in the `Swiss-Prot` database.

Ad2 Repeat	.[PG][LVIMF].[TLKIVGAFY]...{0,8}..[LIM].[LSIVT].[MNTSAIVLY].
Rσ1 Repeat	.[PGQ][LFI].[ILS]...{0,3}..[LIMS].[LFMV].[LGFTSI].

Table 2.4: **Triple β-Spiral Repeat Regular Expressions.** The regular expressions corresponding to the sequence repeat pattern in the Ad2 and Rσ1 Triple β-Spiral folds.

the PROSITE database does not contain regular expressions for either the Adenovirus or Reovirus sequence repeats. Although we are not certain why PROSITE has not characterized these patterns, it is probable that either the fold is too recent, or that it occurs too infrequently to be categorized.

Because PROSITE does not contain a pattern corresponding to the Triple β-Spiral, we implemented a flexible regular-expression search tool[2] for searching for regular expressions in sequence databases. Table 2.4 gives the regular expression patterns corresponding to the Ad2 and Rσ1 sequence repeats, and Table 2.5 presents the results of searching the Swiss-Prot and PDB databases for these regular expressions using our tool. We counted a sequence as a "hit" if it contained the regular expression pattern at least 4 times. We ranked our hits by dividing the number of patterns in a sequence by the sequence length. This quotient gives a rough indication of the density of the repeats in the sequence.

Although this method does discover a fair number of Adenovirus and Reovirus fibers, it provides far too many other hits to be of any reasonable utility. Furthermore, it suffers from the shortcomings that we discussed in Chapter 1. To wit, it is too restrictive in terms of β-strand residue composition and it is too permissive in regions of amino acid insertion. In addition, it does not take into account the background residue distribution or the number of amino acids between repeated occurrences of the sequence motif. For these reasons, we do not consider the results of this search method to be particularly reliable, and we will delay a discussion of the sequences that it finds until after we have presented more sophisticated and sensitive homology modeling methods.

[2]Available on the web at http://theory.lcs.mit.edu/~eben/mthesis.

Fiber	Adenovirus	Reovirus
FIBP_ADE02	X	
FIBP_ADE03		
FIBP_ADE04	X	
FIBP_ADE05	X	
FIBP_ADE07		
FIBP_ADE08	X	
FIBP_ADE09		
FIBP_ADE12	X	
FIBP_ADE15	X	
FIBP_ADE31	X	
FIBP_ADE1A		
FIBP_ADE1P		
FIBP_ADEB3	X	X
FIBP_ADECC	X	
FIBP_ADECG	X	
FIBP_ADECR	X	
FIBP_ADECT	X	
FIBP_ADEP3	X	
FIBP_ADEM1	X	
FIB1_ADE40	X	X
FIB2_ADE40		
FIB1_ADE41	X	X
FIB2_ADE41		
FIB1_ADEG1	X	X
FIB2_ADEG1		
VSI1_REOVD		X
VSI1_REOVL	X	
VSI1_REOVJ	X	
1QIU (PDB Rank)	208	-
1KKE (PDB Rank)	2845	35
Number of Repeats	114	6
Additional Hits	11965	213

Table 2.5: **Model 1: Regular Expression.** This table presents matches for the regular expression model. Note that we did not distinguish between P and G type repeats. A `Swiss-Prot` hit is anything with more than four matches to the regular expression in `Swiss-Prot`. Ranking is done by dividing by the length of the sequence.

2.3.2 Model 2: PSI-BLAST

We performed a PSI-BLAST search for additional instances of the Triple β-Spiral by using both the Ad2 and Rσ1 fiber sequences as search seeds. Table 2.6 presents the results of this search against the Swiss-Prot and PDB databases using the standard PSI-BLAST E-value cutoff of 10. In all four cases PSI-BLAST converged before 20 iterations had completed. Table 2.6 reports only those sequences that were present at the last iteration. Table 2.7 has the same results but incorporates all of the PSI-BLAST hits from every iteration. Note that although the Adenovirus-seeded Swiss-Prot search does match all three Reovirus sequences during intermediate iterations, these have disappeared by the time the model converges. Note also the very large number of hits in the Swiss-Prot database that this model produces. It is unlikely that even a small fraction of these hits are true Triple β-Spirals.

The disparity between the final converged results of PSI-BLAST and its intermediate iterations point to its primary shortcoming: it is quite sensitive to false-positives, and the incorporation of only a few spurious hits in early iterations can lead subsequent iterations to incorporate wholly unrelated sequences. These unrelated sequences can eventually drown out the signal from the original seed sequence, and lead to unstable or unreliable results. This is evident in both the types and number of hits that PSI-BLAST produces. In short, although PSI-BLAST is quite sensitive at detecting distant homology, this sensitivity has a concomitant negative effect on its specificity.

The incorporation of spurious hits into repeated iterations of the PSI-BLAST algorithm is a well-known problem and recurs frequently [37]. One remedy to this problem is to hand-tailor the results of successive iterations to eliminate suspected false positive hits. Although this approach has some merit, we did not pursue it in this thesis because we believe that incorporating this type of prior expectation into an automated method will lead to results that are at best biased, and at worst wholly unreliable.

Another failing of the PSI-BLAST search method is that a PSI-BLAST profile does

Fiber	Adenovirus	Reovirus
FIBP_ADE02	X	
FIBP_ADE03	X	
FIBP_ADE04	X	
FIBP_ADE05	X	
FIBP_ADE07	X	
FIBP_ADE08	X	
FIBP_ADE09	X	
FIBP_ADE12	X	
FIBP_ADE15	X	
FIBP_ADE31	X	X
FIBP_ADE1A	X	
FIBP_ADE1P	X	
FIBP_ADEB3	X	X
FIBP_ADECC	X	
FIBP_ADECG	X	
FIBP_ADECR	X	
FIBP_ADECT	X	
FIBP_ADEP3	X	
FIBP_ADEM1	X	X
FIB1_ADE40	X	X
FIB2_ADE40	X	X
FIB1_ADE41	X	X
FIB2_ADE41	X	X
FIB1_ADEG1	X	
FIB2_ADEG1	X	
VSI1_REOVD		X
VSI1_REOVL		
VSI1_REOVJ		
1QIU (PDB Rank)	1	-
1KKE (PDB Rank)	34	1
Number of Repeats	NA	NA
Additional Hits	52	9

Table 2.6: **Model 2:** `PSI-BLAST`**.** Hits from the `PSI-BLAST` algorithm from only the last (converged) iteration.

Fiber	Adenovirus	Reovirus
FIBP_ADE02	X	
FIBP_ADE03	X	
FIBP_ADE04	X	
FIBP_ADE05	X	
FIBP_ADE07	X	
FIBP_ADE08	X	
FIBP_ADE09	X	
FIBP_ADE12	X	
FIBP_ADE15	X	
FIBP_ADE31	X	X
FIBP_ADE1A	X	
FIBP_ADE1P	X	
FIBP_ADEB3	X	X
FIBP_ADECC	X	
FIBP_ADECG	X	
FIBP_ADECR	X	
FIBP_ADECT	X	
FIBP_ADEP3	X	
FIBP_ADEM1	X	X
FIB1_ADE40	X	X
FIB2_ADE40	X	X
FIB1_ADE41	X	X
FIB2_ADE41	X	X
FIB1_ADEG1	X	
FIB2_ADEG1	X	
VSI1_REOVD	X	X
VSI1_REOVL	X	
VSI1_REOVJ	X	
1QIU (PDB Rank)	1	-
1KKE (PDB Rank)	34	1
Number of Repeats	NA	NA
Additional Hits	114	9

Table 2.7: **Model 2: PSI-BLAST Complete.** Hits from the PSI-BLAST algorithm from every iteration. Note the larger number of Adenovirus and Reovirus fiber hits. Note also the large number of Swiss-Prot hits. Most of these are probably false positives.

not allow for position specific gap and insertion penalties. We can see from the sequence repeats in the Ad2 and Rσ1 fibers that there is a preference for insertions at the location of the solvent exposed loop and in the turn region between repeated structural elements. We would therefore like to search for local matches to these sequences that penalize gaps and insertions differentially at these positions.

One final shortcoming of `PSI-BLAST` that is specific to the Triple β-Spiral sequences is that it is difficult to reconstruct an entire set of repeated sequence elements for a single Adenovirus fiber from the results of a `PSI-BLAST` search. As we progressed through our investigation of the Triple β-Spiral fold, we realized that one of the most arduous aspects of working with the Adenovirus fiber shaft is the task of splitting individual fibers into repeated sequence elements. Although Chroboczek and coworkers did provide hand-constructed alignments of 20 Adenovirus sequences [19] in 1995, they were unable to determine the repeats in some parts of some Adenovirus fibers. Several new sequences have also been introduced since their work was published. We therefore recognized a need for an automated way to split a single Adenovirus fiber into repeated sequence elements – in essence, automating and updating Chroboczek's work. Because `PSI-BLAST` provides alignments only in high-scoring regions of local homology, it is unsuitable for splitting up Adenovirus fiber sequences into their constituent repeats. Several of the methods that we introduce later in this chapter are more effective at this task.

2.3.3 Model 3: `Pfam`

As we discussed in Chapter 1, the `Pfam` database contains an extensive set of Profile HMM's that are created by hand-selecting and aligning specific protein sequences [5]. The `Pfam` database contains entries for both the Adenovirus and Reovirus fiber folds. The `Pfam` identifier for the Adenovirus fiber is `PF00608`, and the `Pfam` identifier for the Reovirus fiber is `PF01664`. We will discuss each of these separately.

Adenovirus Fiber – PF00608

The curators of the Pfam database constructed the Profile HMM Pfam entry for the Adenovirus fiber fold by hand-selecting 63 regions from 14 different Adenovirus fibers. Each of these regions consists of two sequence repeats (see Figure 2-11). The repeats follow Green's scheme, and so begin at position 8 in our labeling scheme. These double-repeat regions were aligned using ClustalW to create a final aligned set that was 31 residues in length. The Pfam curators created a profile HMM from this

```
FIBP_ADEM1/283-312    KGSLGINWGEGIQVKE.QKITLKVTPANGLA
FIBP_ADE08/75-104     TGKLTVNTEPPLHLTN.NKLGIALDAPFDVI
FIBP_ADE05/185-214    TGSLGIDLKEPIYTQN.GKLGLKYGAPLHVT
FIBP_ADE02/185-214    TGSLGINMEDPIYVNN.GKIGIKISGPLQVA
FIBP_ADECG/304-333    GGSLTVATGPGLSHIN.GTIAAVIGAGLKFE
FIBP_ADECG/41-71      PGTLAVNISPPLTFSNLGAIKLSTGAGLILK
FIB1_ADE40/274-303    GSKLIINLGPGLQMSN.GAITLALDAALPLQ
FIBP_ADEM1/73-102     GNTLSLRLNKPLKRTA.KGLQLLLGSGLSVN
FIB1_ADE40/244-273    NNSLSLGVNPPFLITD.SGLAMDLGDGLALG
FIBP_ADE12/50-79      PGVLALNYKDPIVTEN.GTLTLKLGDGIKLN
FIB2_ADE40/42-72      PGVLALKYTDPITTNAKHELTLKLGSNITLQ
FIBP_ADE08/43-72      PGVLSLKLADPITINN.QNVSLKVGGGLTLQ
FIBP_ADE07/61-90      DGVLTLKCLTPLTTTG.GSLQLKVGGGLTID
                            •
                            •
                            •
```

Figure 2-11: Pfam **Adenovirus Fiber Training Sequences.** A sample of the alignment used to train the Pfam Adenovirus Fiber HMMER model. There were 63 total sequences from 14 different Adenovirus Fibers used. Note that because the final alignment is so compact, it seems clear to us that the Pfam curators chose these particular 63 double-repeats because they did not contain extensive insertions in the solvent-exposed loop region.

alignment with the HMMER tool. We used this HMM to search the Swiss-Prot and PDB databases.[3] Table 2.8 presents the results, and Figure 2-12 shows some of the output from this search. The total number of repeats discovered is an indication of how good this method is at picking out the individual repeats on each fiber shaft.

Though this search matches fewer sequences than PSI-BLAST, it does match every Adenovirus fiber. This complete coverage confirms the close evolutionary relationship that exists between all the Adenoviruses. This is a relationship that we already noted in Table 2.2.

[3]Note that we search only the Swiss-Prot database and not the combined Swiss-Prot/TrEMBL databases. Our results are therefore a subset of the hits represented on the Pfam web site.

Fiber	Adenovirus	Reovirus
FIBP_ADE02	X	
FIBP_ADE03	X	
FIBP_ADE04	X	
FIBP_ADE05	X	
FIBP_ADE07	X	
FIBP_ADE08	X	
FIBP_ADE09	X	
FIBP_ADE12	X	
FIBP_ADE15	X	
FIBP_ADE31	X	
FIBP_ADE1A	X	
FIBP_ADE1P	X	
FIBP_ADEB3	X	
FIBP_ADECC	X	
FIBP_ADECG	X	
FIBP_ADECR	X	
FIBP_ADECT	X	
FIBP_ADEP3	X	
FIBP_ADEM1	X	
FIB1_ADE40	X	
FIB2_ADE40	X	
FIB1_ADE41	X	
FIB2_ADE41	X	
FIB1_ADEG1	X	
FIB2_ADEG1	X	
VSI1_REOVD		X
VSI1_REOVL		X
VSI1_REOVJ		X
1QIU (PDB Rank)	1	-
1KKE (PDB Rank)	245	1
Number of Repeats	288	-
Additional Hits	0	4

Table 2.8: **Model 3: `Pfam`.** This table presents Hits from the HMM constructed from the `Pfam` training alignment.

```
FIBP_ADEM1_104        DGQLESSegiseA.DA.PL...QI...ND..GVLQL.........SF...
FIBP_ADEB3_37         EATLAML.....V.EK.PL...TF...DK.eGALTL.........GV...
FIBP_ADE05_318        KLEVNLS.....T.AK.GL...MF...DA..TAIAI.........NA...
FIB1_ADEG1_1          -------.....M.TS.PL...TL...SQ..RALAL.........KT...
FIBP_ADE05_248        QGNMQLN.....V.AG.GL...RI...DSqnRRLIL.........DV...
FIB2_ADEG1_59         DGLLNVR.....L.TA.PL...VIirqSN.gNAIGV.........KT...
FIB1_ADEG1_156        PNTMQVN.....T.GP.--.....-...SG..GMLAV.........KLk.s
FIB2_ADEG1_96         ALQIGIS.....T.AG.PL...TT...TA..NGIDL.........NI...
FIBP_ADECT_110        ENTVSLA.....L.GD.GL...ED...EN..GTLKV.........TFptp
FIBP_ADE31_84         LTTTNTK.....V.LE.PL...PH...TS..QGLTL.........SW...
FIBP_ADEB3_617        QHGLTLR.....V.GS.GL...QM...RD..GILTVtpsgtpiepRL...
FIB1_ADEG1_253        ----TIS.....A.SP.PL...TY...TN..GQIGL.........SI...
FIBP_ADE15_130        VNTLVVL.....T.GK.GLgtdTT...DN.gGSIRV.........RVg.e
FIBP_ADE09_130        RNTLVVL.....T.GK.GIgteST...DN.gGTVCV.........RVg.e
FIB1_ADEG1_208        SGTIALT.....TdTQ.TM...QV...NS..NQLAV.........KLk.t
                                    •
                                    •
                                    •
```

Figure 2-12: `Pfam` **Adenovirus Fiber Output.** A sample of the output alignment from the `Pfam` Adenovirus Fiber model (`PF00608`). In this figure, residues in caps correspond to `Match` positions in the HMM, residues in lower case correspond to `Insert` positions, and dashes correspond to `Delete` positions. Note the large inserts into β-strand regions. For the the total number of "repeats" that this search finds, we give double the total number of hits because this HMM was built with a double-repeat.

Two general comments about these results are in order. First, like `PSI-BLAST`, it is difficult to reconstruct individual repeat elements from this search. Although `HMMER` does utilize position specific gap and insertion penalties, it permits gaps and insertions at any point in the profile. This means that the final output contains many non-physical insertions – that is, there are many insertions inside of β-strands. We could, of course, go back through these results and hand-correct these insertions, but this is arduous and error-prone. It would also give a less optimal final score, though a more intuitive physical one. Second, although the results of this `HMMER` search are clearly more selective than `PSI-BLAST`, this selectivity comes at a price. This approach discovers fewer potential homologs than `PSI-BLAST`, it requires more effort to set up initially, and it is more difficult to iterate. These shortcomings can be addressed – by varying the training sequences, for instance – and in the next section we will examine ways to do this.

Reovirus Fiber – PF01664

The curators of the Pfam database constructed the profile HMM Pfam entry for the Reovirus fiber fold by hand-selecting and aligning all three known strains of Reovirus (Dearing, Jones, Lang) along their entire lengths. That is, this entry included both the fiber and knob domains for Reovirus, and was not split up into repeats. The profile HMM from this alignment contains 474 positions, and is appropriate for identifying other Reoviruses, but not for picking out individual structural repeat elements.

When this profile HMM is used to search Swiss-Prot it finds only one sequence not included in the original training set: The Lysozyme domain for Bacteriophage SF6 (LY_BPSF6). Although the structure of this sequence is not known, it does have over 98% sequence identity to the Dearing strain of the Reovirus, and is almost certainly also a Triple β-Spiral fold [65].

Given the length of the initial training alignment for this HMM, it is unfortunate but not surprising that this search does not discover any of the Adenovirus fiber sequences in Swiss-Prot.

2.3.4 Model 4: Single Repeat Profiles

Although the HMMs in the Pfam database identify all Adenovirus and Reovirus orthologs, these results are unsatisfying for two reasons. First, a large number of sequences were used to train the two HMMs. Second, neither model used an individual repeat element as its base unit.

A simple remedy to both of these problems is to create a HMM using only the repeats from the known Triple β-Spiral fiber sequences (see Figure 1-5). This approach is appealing for three reasons. First, it uses only sequences with known structure, so we can be certain that we are training our HMM with true Triple β-Spiral repeats. Second, it uses a single repeat element, so the output of our HMM will also be single repeat elements. Third, it uses only a single sequence to train our HMM, so we avoid pre-biasing our results in favor of a set of viral orthologs.

We constructed two HMMs from the hand-aligned repeats in Figure 2-8 using

Fiber	Adenovirus	Reovirus
FIBP_ADE02	X	
FIBP_ADE03	X	
FIBP_ADE04	X	
FIBP_ADE05	X	
FIBP_ADE07		
FIBP_ADE08	X	
FIBP_ADE09	X	
FIBP_ADE12	X	
FIBP_ADE15	X	
FIBP_ADE31	X	
FIBP_ADE1A	X	
FIBP_ADE1P	X	
FIBP_ADEB3	X	
FIBP_ADECC	X	
FIBP_ADECG	X	
FIBP_ADECR	X	
FIBP_ADECT	X	
FIBP_ADEP3	X	
FIBP_ADEM1	X	
FIB1_ADE40	X	X
FIB2_ADE40	X	
FIB1_ADE41	X	X
FIB2_ADE41	X	
FIB1_ADEG1	X	
FIB2_ADEG1		
VSI1_REOVD		X
VSI1_REOVL		
VSI1_REOVJ		
1QIU (PDB Rank)	1	25
1KKE (PDB Rank)	207	1
Number of Repeats	216	7
Additional Hits	1	2

Table 2.9: **Model 4: Single Repeat HMM.** This table presents Hits from the HMM constructed by using only the Ad2 and Rσ1 fiber sequence alignments.

the HMMER package. (HMMER is the same package used to create HMMs for the Pfam database.) We then searched the Swiss-Prot and PDB databases using these HMMs. Table 2.9 summarizes our results.

Although we did not identify as many Adenovirus and Reovirus sequences in this approach as were identified by the Pfam searches, this much simpler training alignment did identify most of the Triple β-Spiral fibers in the Swiss-Prot database. In addition, using only the Rσ1 fiber to train, we identified more Adenovirus fibers than the Pfam HMM that was trained with all three Reovirus Fibers.

When used to search the PDB, these two simple HMMs did somewhat better than the HMMs in the Pfam database. The HMM trained with only the Ad2 sequence ranked the Rσ1 fiber higher (207) than the Pfam HMM trained with 14 fiber sequences (245). The HMM trained with the Rσ1 fiber repeats identified the Ad2 fiber in the PDB with a rank of 25. The Pfam HMM trained using the Reovirus sequences did not identify the Ad2 fiber in the PDB at all. In short, for such a simple initial training alignment, this model performs remarkably well.

2.3.5 Model 5: Strict Repeat Profiles

One rather unfortunate aspect of all of the profile models (2, 3, and 4) that we have discussed up to this point is that they all produce non-physical results. That is, these methods produce results with insertions into the β-strand regions of potential Triple β-Spiral repeat elements. Although these insertions are, of course, possible, the solved structure of the Triple β-Spiral leads us to think that it is far more likely that insertions occur in the solvent exposed loop and β-turn regions of the Triple β-Spiral fold.

One simple way to eliminate these insertions in the β-strand regions of the Triple β-Spiral is to modify our profile HMM from the previous section (Model 4) to eliminate insertions at every profile position except position 7. In this approach, deletions (gaps) at any position in the profile are permitted, but the only insertions that are permitted are in the solvent-exposed loop region. We can eliminate all insertions in an HMM profile by modifying the HMMER input file to set all transitions to the

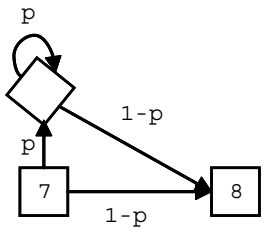

Figure 2-13: **Modified Transition Probabilities.** A schematic representation of a portion of the HMM for Model 5. Note that we set p so that it produces an expected insertion of length equal to the average insertion during the previous iteration. The insertions occur according to a binomial process.

`Insert` state to probability 0 (represented by a $*$ character in the `HMMER` model file) at positions 1-6 and 8-15.

At position 7 in the profile we are still left with the problem of how to set the transition probability into the `Insert` state. Although we could leave this transition probability at its default value, we instead choose to modify it slightly to coincide with our understanding of the variable nature of the solvent-exposed loop. Figure 2-13 shows a portion of a profile HMM specifically designed to discover complete Triple β-Spiral repeats. Note that insertions occur at position 7 according to a binomial process. If we make the simplifying assumption that $P(m \rightarrow i) = P(i \rightarrow i) = p$ then the expected number of residues in this insert is equal to $p/(1-p)$. We can adjust the value of p so that the expected length of the inserted loop is equal to the average solvent-exposed loop length in our input training set. In practice, for an average solvent-exposed loop length a, we set

$$p = \max \begin{cases} \frac{a}{a+1} \\ \\ 0.5 \end{cases}$$

We set a floor for p of 0.5 so that our expected insert does not fall below one residue in length.

Figure 2-14 shows a portion of the modified `HMMER` input file for this model. Table 2.10 gives the results of searching the `Swiss-Prot` and `PDB` databases with these

Fiber	Adenovirus	Reovirus
FIBP_ADE02	X	
FIBP_ADE03	X	
FIBP_ADE04	X	
FIBP_ADE05	X	
FIBP_ADE07		
FIBP_ADE08	X	
FIBP_ADE09	X	
FIBP_ADE12	X	X
FIBP_ADE15	X	
FIBP_ADE31	X	
FIBP_ADE1A	X	
FIBP_ADE1P	X	
FIBP_ADEB3	X	X
FIBP_ADECC	X	
FIBP_ADECG	X	
FIBP_ADECR	X	
FIBP_ADECT	X	
FIBP_ADEP3	X	X
FIBP_ADEM1	X	X
FIB1_ADE40	X	X
FIB2_ADE40	X	
FIB1_ADE41	X	X
FIB2_ADE41	X	
FIB1_ADEG1	X	X
FIB2_ADEG1		
VSI1_REOVD		X
VSI1_REOVL		
VSI1_REOVJ		
1QIU (PDB Rank)	1	63
1KKE (PDB Rank)	514	1
Number of Repeats	232	7
Additional Hits	0	2

Table 2.10: **Model 5: Single Repeat HMM with Restricted Insertions.** Hits from the HMM constructed by only using the Ad2 and Rσ1 fiber sequence alignments. Insertions are only permitted at position 7.

```
      m->m   m->i   m->d   i->m   i->i   d->m   d->d   b->m   m->e
6       -3      *  -9046   -894  -1115   -701  -1378      *      *
7    -1000  -1000      *  -1000  -1000      0      *      *      *
8       -3      *  -9046   -894  -1115   -701  -1378      *      *
```

Figure 2-14: **Modified HMM Transition Probabilities.** The transition probabilities for an HMM that gives equal probability to insertions at position 7 but vanishing probability to insertions at all other positions. Note that we only show a portion of the HMMER file.

custom profile HMMs. Note that we do not find any more Adenovirus fibers with the this model than we found with Model 4, but we do find 16 more total repeats. We suspect that we find more repeats in this model because of the slight bias that we have introduced toward longer solvent-exposed loop insertions. Although we find more repeats for the Adenovirus model, the Rσ1 sequence actually ranks lower in the PDB, indicating that this model is slightly less sensitive than Model 4.

For the Rσ1-trained model, this method discovers several more Adenovirus fiber sequences than the previous models. Although we are not entirely sure why this is the case, we suspect that by trimming our Rσ1 alignment to 15 residues and then allowing an insertion only at position 7, we have actually imposed the Adenovirus structural model on the Rσ1 sequence. This may account for the greater sensitivity of the Rσ1-trained model for Adenovirus sequences.

Figure 2-15 shows a portion of the output from our method. In contrast to previous approaches, it is simple to reconstruct Triple β-Spiral repeats from the output of this method.

2.3.6 Model 6: Iterated Strict Repeat Profiles

The similarity of the input and output sequences for the modified HMM from Model 5 suggests that we should be able to iterate these results. Iteration seems like a natural extension to our modified HMMER method because unlike PSI-BLAST and traditional HMMER this method finds domains that exactly match our expected Triple β-Spiral

```
FIBP_ADE02_379  AGLSFDNS------GAITIGNK
FIBP_ADE03_52   NPLTTAS-------GSLQLKVG
FIBP_ADE03_67   SGLTVDTTD-----GSLEENIK
FIBP_ADE03_86   TPLTKSN-------HSINLPIG
FIBP_ADE03_101  NGLQIEQ-------NKLCSKLG
FIBP_ADE03_116  NGLTFDSS------NSIALKNN
FIBP_ADE31_37   PPFTSSNAFQEKPPGVLSLNYK
FIBP_ADE31_59   DPIVTEN-------GSLTLKLG
FIBP_ADE31_74   NGIKLNSQ------GQLTTTNT
FIBP_ADE31_93   EPLPHTS-------QGLTLSWS
FIBP_ADE31_108  APLSVKA-------SALTLNTM
                     •
                     •
                     •
```

Figure 2-15: **Alignments Generated by Model 5.** A portion of the alignment generated by Model 5. The insertions after position 15 in the model have been truncated. There are no non-physical insertions – i.e. insertions at positions other than 7.

repeat. Because we specifically do not allow insertions at non-physical locations, iterating these results should improve the sensitivity of our search without significantly biasing our results with false positives.

We created a simple program to re-train HMMER based upon the results of a previous HMMER run. At each step of the iteration, as in Model 5, we modify the HMMER model to assign vanishing transition probabilities to all insertions except at position 7 in the profile. We train our initial HMM with the Ad2 and Rσ1 fiber sequences as in previous models, but after the first iteration, all of the hits from the previous iteration are incorporated into the new alignment. We stop iterating when no new domains are found in successive iterations.

Table 2.11 gives the results of this iterated model. The HMM converged after 7 iterations for the Ad2-seeded HMM and after 8 iterations for the Rσ1-seeded HMM. Note that starting with only the two known solved structures, we were able to find all but one of the Reovirus fibers and all of the Adenovirus fibers in Swiss-Prot. This method also provides more repeat sequences than any previous method. Figure 2-16 shows the average loop length per iteration in the two folds.

We used the converged HMM model trained using Swiss-Prot sequences to search the PDB. It is interesting that this method does better than all previous methods at

Fiber	Adenovirus	Reovirus
FIBP_ADE02	X	X
FIBP_ADE03	X	X
FIBP_ADE04	X	X
FIBP_ADE05	X	X
FIBP_ADE07	X	X
FIBP_ADE08	X	X
FIBP_ADE09	X	X
FIBP_ADE12	X	X
FIBP_ADE15	X	X
FIBP_ADE31	X	X
FIBP_ADE1A	X	X
FIBP_ADE1P	X	X
FIBP_ADEB3	X	X
FIBP_ADECC	X	X
FIBP_ADECG	X	X
FIBP_ADECR	X	X
FIBP_ADECT	X	X
FIBP_ADEP3	X	X
FIBP_ADEM1	X	X
FIB1_ADE40	X	X
FIB2_ADE40	X	X
FIB1_ADE41	X	X
FIB2_ADE41	X	X
FIB1_ADEG1	X	X
FIB2_ADEG1	X	X
VSI1_REOVD	X	X
VSI1_REOVL	X	
VSI1_REOVJ		
1QIU (PDB Rank)	1	1
1KKE (PDB Rank)	8	11
Number of Repeats	336	7
Additional Hits	11	12

Table 2.11: **Model 6: Iterated Single Repeat HMM with Restricted Insertions.** This table presents hits from the HMM constructed by only using the Ad2 and Rσ1 fiber sequence alignments. Insertions are only permitted at position 7. The HMM was iterated with the output from each iteration serving as the training alignment for the next iteration. The model converged after 7 iterations for Ad2 and after 8 iterations for Rσ1.

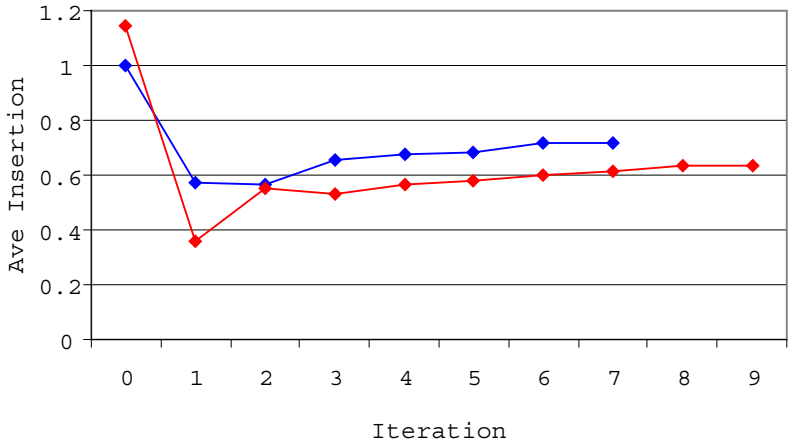

Figure 2-16: **Average Insertion Lengths.** The average insertion length at position 7 for each iteration of the HMM. The Ad2-seeded HMM is shown in red and the Rσ1-seeded HMM is shown in blue.

ranking the Ad2 and Rσ1 sequences. It is also interesting that the model seeded with the Rσ1 fiber actually ranks the Ad2 fiber sequence (`1QIU`) above the Rσ1 fiber sequence (`1KKE`). This is reasonable when we examine the final results of the Rσ1-seeded model. In early iterations, the Rσ1-seeded model finds many Adenovirus fiber repeats. By the time that this model converges, the Adenovirus sequences dominate the Reovirus sequences. This is because of their greater prevalence in the `Swiss-Prot` database.[4]

Figure 2-16 shows the average length of the solvent-exposed loop residues for the Ad2 and Rσ1 HMMs at each iteration. Note that the average loop length is usually below one residue, which means that the $p = 0.5$ floor that we imposed is binding. This suggests that either the solvent-exposed loop is not especially long, or that we should have used a statistic other than the average to set p.

One rather strange outcome of this model is that the HMM seeded with the Ad2 sequence finds more Reovirus fibers (`VSI1_REOVJ` and `VSI1_REOVL`) than the HMM seeded with the Rσ1 fiber sequence. This is especially strange in light of the fact that

[4]In theory, this should not occur because `HMMER` probabilistically weights the sequences in its profile.

both models find all of the Adenovirus fiber sequences in Swiss-Prot.

In spite of the success of this model, however, there is still clearly some room for improvement, as the Adenovirus and Reovirus fiber sequences are not definitively identified as Triple β-Spirals in the PDB.

2.3.7 Model 7: Iterated Profiles with Custom Insertions

We have seen that one advantage of HMMER over other profile methods (like PSI-BLAST) is that HMMER gives position-specific scores (transition probabilities) to residue insertions and deletions. Once a model has entered an Insert state, however, the residue emissions from this state are the same in every Insert state. That is, HMMER does not recognize that different positions in the profile might have differentially expressed preferences for the insertion of different types of residues. This is an especially relevant point for the Triple β-Spiral fold because we have observed that insertions in the Triple β-Spiral repeats occur primarily in the solvent-exposed loop region.

We can modify our previous model (Model 6) to incorporate specific residue preferences in the Insert state at position 7 in the profile. This amounts to creating a set of custom residue emission probabilities based upon the observed residue preferences at this point in the profile. To do this, we assign residue emission probabilities according to Equation 1.1.

To calculate these probabilities, we use residue counts from the solvent-exposed region in the Ad2 and Rσ1 sequences during the first iteration, and then reset the emission probabilities at each iteration based upon the residues in the solvent-exposed loop during the previous iteration. We do not distinguish between individual positions in the insert region, but treat all residues equally.

Table 2.12 shows the results of searching Swiss-Prot and the PDB using this HMM. Figure 2-17 gives the final residue composition of the solvent-exposed loop region for the Adenovirus and Reovirus folds, and compares these to the generic HMMER Insert state emission scores. Note that we do not find any new sequences with this approach, but we do find 4 more Adenovirus repeats. This suggests that biasing the insertions in the solvent-exposed loop region to polar and charged residues does slightly improve

Fiber	Adenovirus	Reovirus
FIBP_ADE02	X	X
FIBP_ADE03	X	X
FIBP_ADE04	X	X
FIBP_ADE05	X	X
FIBP_ADE07	X	X
FIBP_ADE08	X	X
FIBP_ADE09	X	X
FIBP_ADE12	X	X
FIBP_ADE15	X	X
FIBP_ADE31	X	X
FIBP_ADE1A	X	X
FIBP_ADE1P	X	X
FIBP_ADEB3	X	X
FIBP_ADECC	X	X
FIBP_ADECG	X	X
FIBP_ADECR	X	X
FIBP_ADECT	X	X
FIBP_ADEP3	X	X
FIBP_ADEM1	X	X
FIB1_ADE40	X	X
FIB2_ADE40	X	X
FIB1_ADE41	X	X
FIB2_ADE41	X	X
FIB1_ADEG1	X	X
FIB2_ADEG1	X	X
VSI1_REOVD	X	X
VSI1_REOVL	X	
VSI1_REOVJ		
1QIU (PDB Rank)	1	1
1KKE (PDB Rank)	5	6
Number of Repeats	340	7
Additional Hits	18	7

Table 2.12: **Model 7: Iterated Single Repeat HMM with Restricted Insertions and Custom Loop Residues.** Hits from the HMM constructed by only using the Ad2 and Rσ1 fiber sequence alignments. Insertions are only permitted at position 7. The HMM was iterated with the output from each iteration serving as the training alignment for the next iteration. The model converged after 8 iterations for Ad2 and after 9 iterations for Rσ1. Insertion emission probabilities at position 7 are calculated from the loop residue composition in the training alignment.

PDB ID	SCOP Classification	Adenovirus Model Rank	Reovirus Model Rank
1GYT	Leucine aminopeptidase (Aminopeptidase A) N-terminal	2	2
1LAP	Leucine aminopeptidase (Aminopeptidase A) N-terminal	3	3
1LCP	Leucine aminopeptidase (Aminopeptidase A) N-terminal	4	4
1MPR	Transmembrane beta-barrel		5
2MPR	Transmembrane beta-barrel		6

Table 2.13: **Model 7 False Hits from the PDB.** False positives from Model 7 from the PDB database for the Ad2 and Rσ1 trained models. Note that the false hits are basically identical because of the convergence when the two models are iterated.

the sensitivity of the profile search.

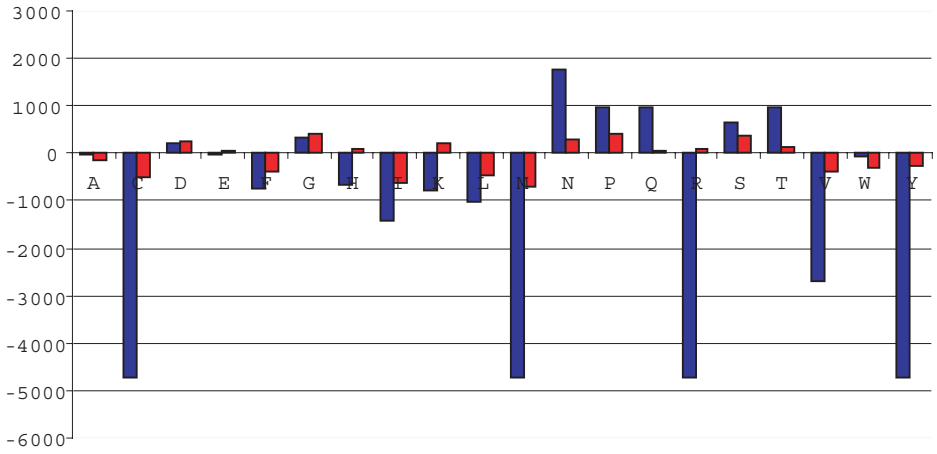

Figure 2-17: **Final Inserted Residue Scores.** The final residue scores for an iterated HMM that learns its loop residue probabilities from its training alignment. Note the relatively high scores given to polar and charged residues, as we would expect in the solvent-exposed region.

One especially felicitous outcome of model 7 is that it outperforms all previous models in ranking the Ad2 and Rσ1 sequences in the PDB. As occurred for Model 6, the Rσ1-seeded HMM actually ranked the Ad2 fiber sequence above the Rσ1 fiber sequence in this case. For this model, however, there were only 3 false positives for the Ad2-trained HMM and only 5 false positives for the Rσ1-trained HMM. Table 2.13 lists the PDB entries that outscored the Ad2 and Rσ1 fibers, and Figure 2-18 shows these structures.

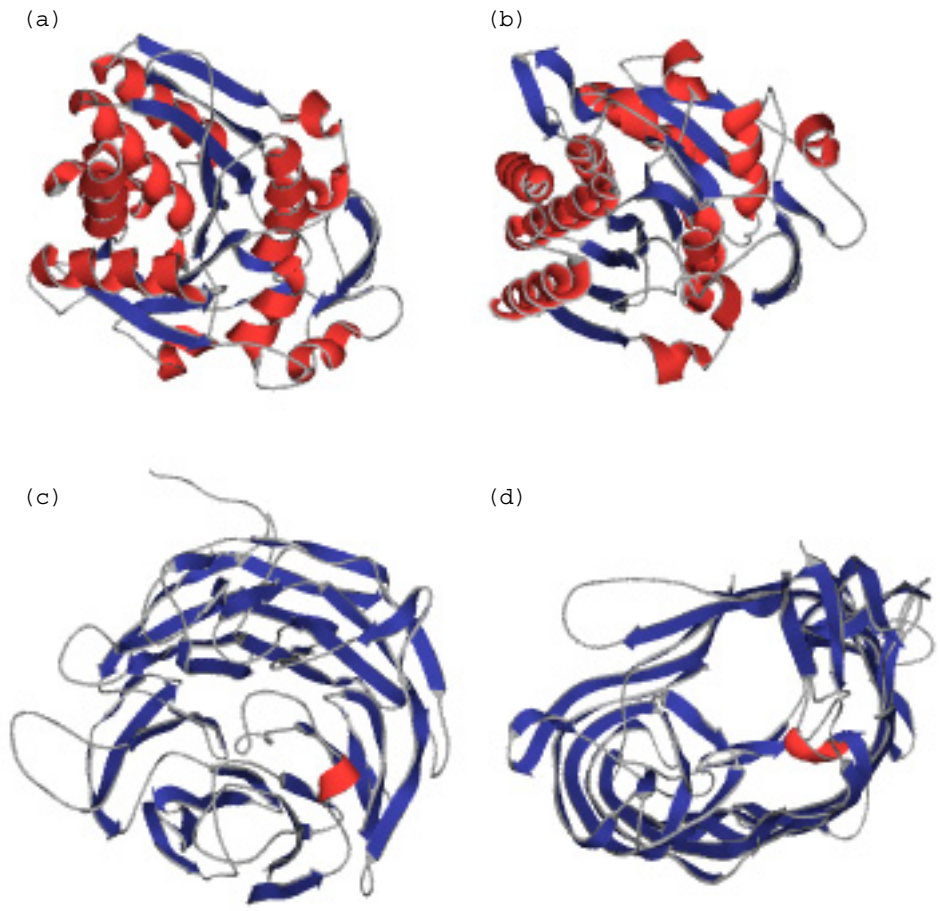

Figure 2-18: **False PDB Hits.** Pictured are the (a) 1LAM (b) 1GYT (c) 1K3I, and (d) 1MPR. Together with 2MPR (which is almost identical to 1MPR), these proteins are the false positives for Model 7 in the PDB database.

Table 2.14 compares model 7 to all of the previous methods that we have applied in this chapter. Notice that in addition to outperforming all previous models on the PDB, model 7 also detects most Adenovirus and Reovirus sequences while finding relatively few additional sequences in the Swiss-Prot database. In addition, model 7 outperforms all previous models in terms of the number of sequence repeats that it detects for both Adenovirus and Reovirus fibers.

Adenovirus Fiber Repeats

The results of Models 6 and 7 suggest that we may finally be close to a fully automated method for the discovery of repeat elements in the Adenovirus fiber sequences.

Fiber	1		2-con		2-un		3		4		5		6		7	
	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b
FIBP_ADE02	X		X		X		X		X		X		X	X	X	X
FIBP_ADE03			X		X		X		X		X		X	X	X	X
FIBP_ADE04	X		X		X		X		X		X		X	X	X	X
FIBP_ADE05	X		X		X		X		X		X		X	X	X	X
FIBP_ADE07			X		X		X						X	X	X	X
FIBP_ADE08	X		X		X		X		X		X		X	X	X	X
FIBP_ADE09			X		X		X		X		X		X	X	X	X
FIBP_ADE12	X		X		X		X		X		X	X	X	X	X	X
FIBP_ADE15	X		X		X		X		X		X		X	X	X	X
FIBP_ADE31	X		X	X	X	X	X		X		X		X	X	X	X
FIBP_ADE1A			X		X		X		X		X		X	X	X	X
FIBP_ADE1P			X		X		X		X		X		X	X	X	X
FIBP_ADEB3	X	X	X	X	X	X	X		X		X	X	X	X	X	X
FIBP_ADECC	X		X		X		X		X		X		X	X	X	X
FIBP_ADECG	X		X		X		X		X		X		X	X	X	X
FIBP_ADECR	X		X		X		X		X		X		X	X	X	X
FIBP_ADECT	X		X		X		X		X		X		X	X	X	X
FIBP_ADEP3	X		X		X		X		X		X	X	X	X	X	X
FIBP_ADEM1	X		X	X	X	X	X		X		X	X	X	X	X	X
FIB1_ADE40	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X
FIB2_ADE40			X	X	X	X	X		X		X		X	X	X	X
FIB1_ADE41	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X
FIB2_ADE41			X	X	X	X	X		X		X		X	X	X	X
FIB1_ADEG1	X	X	X		X		X		X		X	X	X	X	X	X
FIB2_ADEG1			X		X		X						X	X	X	X
VSI1_REOVD		X		X	X	X		X		X		X	X	X	X	X
VSI1_REOVL	X				X			X					X		X	
VSI1_REOVJ	X				X			X								
1QIU (PDB Rank)	208	-	1	-	1	-	1	-	1	25	1	63	1	1	1	1
1KKE (PDB Rank)	2845	35	34	1	34	1	245	1	207	1	514	1	8	11	5	6
Number of Repeats	114	6	NA	NA	NA	NA	288	-	216	7	232	7	336	7	340	7
Additional Hits	11965	213	52	9	114	9	0	4	1	2	0	2	11	12	18	7

Table 2.14: **Comparison of Models 1 through 7.** The information from models 1 through 7. Note that Model 2 (PSI-BLAST) has both converged and unconverged results.

Figure 2-19 gives a sample of the final results of Model 7 for the final iteration of the Ad2-trained and iterated model.[5] Although we find many Adenovirus repeats, there are some low-homology repeat regions that the method fails to discover.

```
FIBP_ADE02_379  AGLSFDNS-------GAITIGNK
FIBP_ADE03_52   NPLTTAS--------GSLQLKVG
FIBP_ADE03_67   SGLTVDTTD------GSLEENIKVN
FIBP_ADE03_86   TPLTKSN--------HSINLPIG
FIBP_ADE03_101  NGLQIEQ--------NKLCSKLG
FIBP_ADE03_116  NGLTFDSS-------NSIALKNN
FIB2_ADEG1_53   GPLYSTD--------GLLNVRLT
FIB2_ADEG1_68   APLVIIRQSNG----NAIGVKTD
FIB2_ADEG1_87   GSITVNAD-------GALQIGISTA
FIB2_ADEG1_105  GPLTTTA--------NGIDLNIDP
FIB2_ADEG1_121  KTLVVDGSSGK----NVLGVLLKGQ
FIB2_ADEG1_142  GALQSSA--------QGIGVAVD
FIB2_ADEG1_157  ESLQIVD--------NTLEVKVDAA
FIB2_ADEG1_174  GPLAVTA--------AGVGLQYD
FIBP_ADE31_37   PPFTSSNAFQEKPP-GVLSLNYK
FIBP_ADE31_59   DPIVTEN--------GSLTLKLG
FIBP_ADE31_74   NGIKLNSQ-------GQLTTTNTKVL
FIBP_ADE31_93   EPLPHTS--------QGLTLSWS
FIBP_ADE31_108  APLSVKA--------SALTLNTM
FIBP_ADE31_123  APFTTTN--------ESLSLVTAPPITVEASQLGLASCSTSKLRGGGNLGFHLP
FIBP_ADE31_169  APFVVPSS-------NALTLSAS
FIBP_ADE31_185  DPLTVNS--------NSLGLNIT
                                •
                                •
                                •
```

Figure 2-19: **Alignments Generated by Model 7.** A portion of the alignment generated by Model 7. The intervening residues between successive repeats have been added to show both those repeats that the model finds and those that it fails to find.

We implemented an iterated, profile-based algorithm to place the remaining repeats. The algorithm proceeded as follows:

1. Find a region with over 15 residues in the gap between successive repeats.
2. Run a profile search on this region using our custom HMM from Model 7.
3. Place this repeat in the alignment.
4. Return to step 1.

Using this method, we reconstructed every repeat from all of the Adenovirus fiber sequences. These repeats are available on our website.

[5]The full results are available at `http://theory.lcs.mit.edu/~eben/mthesis`.

2.4 Analysis of Significant Hits

Given the careful annotation of the Swiss-Prot database and the relatively small number of sequences that it contains, we think that it is unlikely that this database contains a large number of unidentified Triple β-Spiral sequences. That is why throughout most of this chapter, we treated hits in the Swiss-Prot database as probable false positives. That is also why we were gratified that our final model (Model 7) found most of the Adenovirus and Reovirus fibers in Swiss-Prot but did not find a great number of additional sequences.

Table 2.15 lists the top 50 hits that we found in the Swiss-Prot database and which models found these hits. We ranked hits by the number of times that they are identified by different models.

Among these top hits, the Bacteriophage SF6 Lysozyme (LY_BPSF6) almost certainly contains the Triple β-Spiral fold, as it has over 98% sequence identity to the VSI1_REOVD Triple β-Spiral sequence. Figure 2-20 shows an alignment of these two sequences. This is quite an interesting result, because the Bacteriophage SF6 is a virus that infects bacteria, whereas Adenovirus and Reovirus are both eukaryotic viruses. This provides further support to the hypothesis that the Triple β-Spiral is an ancient protein fold. If this is indeed the case, then we expect there to be more instances of the Triple β-Spiral fold in other highly divergent organisms.

```
CLUSTAL W (1.81) multiple sequence alignment


LY_BPSF6        APLSIRNNRITMGLNDGLTLSGNNLAIRLPGNTGLNIQNGGLQFRFNTDQFQIVNNNLTL
VSI1_REOVD      APLSIRNNRMTMGLNDGLTLSGNNLAIRLPGNTGLNIQNGGLQFRFNTDQFQIVNNNLTL
                *********:**************************************************

LY_BPSF6        KTTVFDSINSRIGATEQSYVASAVTPLRLNSSTKVLDMLIDMSTLEINSSGQLTVRST
VSI1_REOVD      KTTVFDSINSRIGATEQSYVASAVTPLRLNSSTKVLDMLIDSSTLEINSSGQLTVRST
                **************************************** ***************
```

Figure 2-20: **Alignment of LY_BPSF6 and VSI1_REOVD Fibers.** An alignment generated by ClustalW for the fiber portion of Rσ1 and its homologous region in the Lysozyme domain of Bacteriophage SF6.

Among the remaining hits, there is a clear identification of two families of probable

Swiss-Prot ID	Protein Name	1		2a		2b		3		4		5		6		7	
		a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b
LY_BPSF6	Lysozyme		x		x	x	x		x		x		x	x	x	x	x
PMP9_CHLPN	Probable outer membrane protein	x		x		x									x	x	x
OMPB_RICPR	Outer membrane protein B	x				x								x	x	x	
TROP_HUMAN	Trophinin	x				x								x	x		x
BIGA_SALTY	Putative surface-exposed virulence	x		x		x										x	
OMPB_RICRI	Outer membrane protein B	x												x	x	x	
YDBA_ECOLI	Hypothetical protein ydbA	x		x		x										x	
Z236_HUMAN	Zinc finger protein 236				x	x	x										
FHAB_BORPE	Filamentous hemagglutinin	x												x	x		
N145_YEAST	Nucleoporin NUP145	x		x		x											
N189_SCHPO	Nucleoporin nup189	x		x		x											
OMPB_RICCN	Outer membrane protein B	x												x		x	
OMPB_RICJA	Outer membrane protein B	x												x		x	
OMPB_RICTY	Outer membrane protein B	x												x		x	
P5CS_CAEEL	Probable delta 1-pyrroline-..	x		x		x											
PM11_CHLPN	Probable outer membrane protein	x		x		x											
PM14_CHLPN	Probable outer membrane protein	x		x		x											
PM15_CHLPN	Probable outer membrane protein	x		x		x											
PM16_CHLPN	Probable outer membrane protein	x		x		x											
PM18_CHLPN	Probable outer membrane protein	x		x		x											
PM19_CHLPN	Probable outer membrane protein	x		x		x											
PMP6_CHLPN	Probable outer membrane protein	x		x		x											
PMP7_CHLPN	Probable outer membrane protein	x		x		x											
PMP8_CHLPN	Probable outer membrane protein	x		x		x											
PMPA_CHLTR	Probable outer membrane protein	x		x		x											
PMPB_CHLTR	Probable outer membrane protein	x		x		x											
PMPC_CHLMU	Probable outer membrane protein	x		x		x											
PMPC_CHLTR	Probable outer membrane protein	x		x		x											
7SBG_SOYBN	Basic 7S globulin [Precursor]				x		x										
AMA2_SHEON	Probable cytosol aminopeptidase 2														x		x
AMPA_VIBCH	Cytosol aminopeptidase														x		x
AMPA_VIBVU	Probable cytosol aminopeptidase														x		x
COXZ_BRAJA	Cytochrome c oxidase assembly				x		x										
ELT1_CAEEL	Transcription factor elt-1													x		x	
FWDC_METTM	Formylmethanofuran dehydrogenase			x		x											
FWDC_METWO	Formylmethanofuran dehydrogenase			x		x											
IGA_NEIGO	IgA-specific serine endopeptidase			x		x											
IM23_SCHPO	Mitochondrial import..				x		x										
NU98_HUMAN	Nuclear pore complex protein			x		x											
OGP_MESAU	Oviduct-specific glycoprotein			x		x											

Table 2.15: **The 50 Top Scoring Swiss-Prot Hits.** Hits were ranked according to the number of models that identified them.

outer membrane proteins (prefixed by PM and OMP). Although the structure of these proteins is not known, they are believed to be transmembrane β-barrels, which would coincide well with our false positives from the PDB [64, 29].

One interesting hit on this list is the surface exposed virulence protein from salmonella typhimurium (BIGA_SALTY). This is a 1953 amino acid long protein with suspected links to pathogenesis [10]. BetaWrap identifies this sequence as a Right-Handed Parallel β-Helix with p-value 0.00010, so it is likely that this sequence is actually a Right-Handed Parallel β-Helix and not a Triple β-Spiral [10]. There are also several other likely Right-Handed Parallel β-Helix structures in our list, including FHAB_BORPE, which is the attachment protein for filamentous haemagglutinin [39].

2.5 The Swiss-Prot/TrEMBL Database

Given the success of Model 7, a reasonable next step is to run this model on the Swiss-Prot/TrEMBL database. The Swiss-Prot/TrEMBL database contains many more sequences than Swiss-Prot and unlike Swiss-Prot it contains sequences that are not well-annotated. Many of these sequences may contain the Triple β-Spiral fold. Table 2.16 lists several of the interesting hits from this search. Appendix A lists all of the hits.

Among these top hits, we find many Adenovirus and Reovirus fiber proteins. We also discover several suspected trans-membrane β-barrels (the OMPB prefixed entries) and a few Right-Handed Parallel β-Helix proteins (FHAB_BORPE, Q880E1, etc.) There are also several other proteins that are likely additional instances of the Triple β-Spiral fold. We will discuss these proteins in more detail at the end of the next chapter.

2.6 Discussion

In this chapter we have presented an analysis of the Adenovirus and Reovirus fiber sequences. Although we did not explicitly make use of the three-dimensional structures of these proteins, we did choose our sequence repeats based upon the known structure

ID	Protein Name	Source
Q8QZQ6	Hypothetical protein 443R	Chilo iridescent virus (CIV)
Q8P942	YapH protein	Xanthomonas campestris
Q88RG2	Surface adhesion protein, putative	Pseudomonas putida (strain KT2440)
Q8GDL9	Orf2	Photorhabdus luminescens
Q8PKM0	YapH protein	Xanthomonas axonopodis (pv. citri)
Q9F285	YapH protein	Yersinia pestis
Q8ZHA1	Putative autotransporter protein	Yersinia pestis
Q8CZU2	Putative autotransporter adhesin	Yersinia pestis
Q8PF72	YapH protein	Xanthomonas axonopodis (pv. citri)
Q8QZQ8	261R	Chilo iridescent virus (CIV)
Q98E20	Hypothetical protein mll4444	Rhizobium loti
Q9XC47	Outer membrane protein A	Rickettsia australis
Q9I120	Hypothetical protein PA2462	Pseudomonas aeruginosa
TROP_HUMAN	Trophinin	Homo sapiens (Human)
LY_BPSF6	Lysozyme	Bacteriophage SF6
Q8GD27	Adhesin FhaB	Bordetella avium

Table 2.16: **Hits from Swiss-Prot/TrEMBL.** This table presents several of the more interesting hits from running model7 on the Swiss-Prot/TrEMBL database. The complete results are in the appendix.

of the Triple β-Spiral fold. We also chose to treat residues in the solvent-exposed loop differently from other residues in the sequence repeat. We did this to accomodate the variable length and biased residue composition of this region.

Using simple profile based homology-search methods with some modifications we were able to use the Ad2 fiber sequence to discover the Rσ1 fiber sequence in Swiss-Prot and we were likewise able to use the Rσ1 fiber sequence to discover the Ad2 fiber sequence. This is not surprising given the weak but significant sequence identity in Table 2.2.

In spite of the sequence similarity between the Ad2 and Rσ1 fiber sequences, none of the methods that we tested were able to definitively identify the two fiber sequences in cross-validation tests in the PDB. This suggests that homology-modeling methods are not by themselves sufficiently specific to model the complexity of the Triple β-Spiral fold. We therefore turn our attention to tools that incorporate the three-dimensional structure of the Triple β-Spiral fold in the next chapter.

Chapter 3

Structure Analysis

3.1 Introduction

In the previous chapter of this thesis we presented an analysis of the Triple β-Spiral protein fold based solely on amino acid sequence. In this chapter we continue to analyze the Triple β-Spiral fold, but we enhance our analysis by incorporating structural information into our computational methods. There are many different structure prediction algorithms that we could explore, but in this chapter we focus our attention on the three-dimensional counterparts of the profile methods presented in the previous chapter. We focus on these methods to maintain the continuity of our analysis.

At first glance, the Triple β-Spiral fold does not seem to be an appealing candidate for structural analysis. There are, after all, only two solved instances of the fold. Even when we include all of the Adenovirus and Reovirus orthologs in the `Swiss-Prot` database, this number rises only to 28 total known Triple β-Spiral fibers. This should be compared, for example, to the TIM-Barrel fold, which has 157 solved structures, comprising 19 distinct representative classes [21].

As we look more closely at the Triple β-Spiral fold, however, we note two aspects that make it more appealing from a structural standpoint. First, since the Triple β-Spiral fold contains a structural repeat, we can envision analyzing each of these repeats separately, thereby increasing by many times the number of available structures. Second, since the Triple β-Spiral consists of three identical protein chains, there is a

rich set of implied inter-chain interactions in a putative Triple β-Spiral protein chain. This second point is rather subtle. Since the Triple β-Spiral occurs as a homotrimer, there is a set of implied interactions from the chain to other chains, but since these three chains are identical, the implied interactions can be analyzed as interactions from a chain to itself. Figure 2-3 in Chapter 2 demonstrates this point. Of course, it is possible that there are Triple β-Spiral folds that occur as heterotrimers. In this thesis, however, we make the working hypothesis that other Triple β-Spiral folds also occur as homotrimers.

3.2 Simulated Annealing

In this chapter, we will employ simulated annealing to incorporate structural information about the Triple β-Spiral fold into our previous analysis [48]. Although we will apply simulated annealing to a variety of different models, our basic approach will be the same in each case. We begin with a set of target sequences and a scoring function. The scoring function should peak for Triple β-Spiral sequences.

We employ simulated annealing to optimally place four Triple β-Spiral repeat elements along the target sequence. We place the four repeats according to the following constraints:

1. There cannot be more than 8 residues in the solvent exposed loop region between β-strands within the same repeat.

2. There cannot be more than 6 residues inserted between successive repeats.

The loop and insertion lengths were determined empirically from the Adenovirus and Reovirus fiber sequences. With these values we should find four contiguous repeats. This differs from the methods in Chapter 2, which searched for individual repeat elements rather than four contiguous repeats.

With these constraints, the simulated annealing algorithm proceeds as follows:

1. An initial annealing temperature is set.

2. The four repeat units are placed along the target sequence subject to the constraints outlined above.

3. A score (σ_1) for this initial configuration is calculated.

4. A "move set" of allowable moves is generated by testing whether each of the eight β-strands in the model can be moved one residue to the left or right according to the constraints outlined above. In this way, a set of 0 to 16 possible moves is generated. One of the moves in the "move set" is performed at random and the model enters a new configuration.

5. The score of this new configuration (σ_2) is compared to the score of the previous configuration. If the score of the new configuration is higher than the previous configuration, then this new configuration is kept. If the score of the new configuration is lower than the previous configuration, then this new configuration is rejected with probability $\exp((\sigma_2 - \sigma_1)/T)$ and kept otherwise.

6. Steps 4 and 5 are repeated a fixed number of times. The temperature is then decreased and steps 3 and 4 are repeated again. This procedure is continued until some final annealing temperature is achieved.

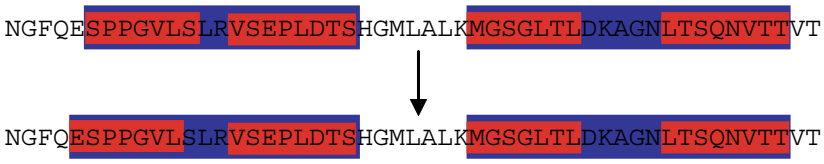

Figure 3-1: **Simulated Annealing.** In simulated annealing, repeats (in blue) are placed along a sequence and then the β-strands in the repeats (in red) are moved. As the β-strands move, the repeats travel through many different regions of the sequence until settling on the (hopefully) optimal score. Due to space constraints, we show only two repeat units instead of the four that our algorithm actually uses.

Although this method is not guaranteed to converge to the global optimum, it is extremely likely that it will do so if it is run for a sufficient number of steps with a sufficiently small temperature step. In practice, there is a trade-off between the speed

Start Temp	0.60
Final Temp	0.01
Temp Step	0.01
Iterations per Step	250,000

Table 3.1: **Parameters for Simulated Annealing.** The parameters used for each of the simulated annealing scoring methods discussed in this chapter.

of the algorithm and the optimality of the final result. For our work, we have found that the values presented in Table 3.1 provide good convergence.

3.2.1 Test Sequences

Because simulated annealing is a relatively slow algorithm, we cannot run the methods presented in this chapter on the full `Swiss-Prot` or `PDB` databases. Instead, we have chosen a set of 2200 protein sequences for testing the models that we present in this chapter. This set contains all of the `PDB` sequences that scored higher than `1QIU` and `1KKE` in Models 2 through 7. It also includes all of the Adenovirus and Reovirus sequences from `Swiss-Prot`, and approximately 1000 more randomly selected protein sequences from the `PDB`.

We have chosen decoy candidates exclusively from the `PDB` so that we can be certain that each candidate is not, in fact, a Triple β-Spiral. We do include the Adenovirus and Reovirus sequences of unknown structure from the `Swiss-Prot` database, as we feel safe in assuming that these sequences do contain the Triple β-Spiral fold.[1]

3.3 Computing Sequence Scores with Structural Models

In this section we present scores based upon an abstract structural template for the Triple β-Spiral fold. We use simulated annealing to find the optimal alignment of a sequence to this abstract structural template. There are numerous scoring models

[1] A full list of the sequences can be found at `http://theory.lcs.mit.edu/~eben/mthesis`.

for evaluating whether a given sequence folds into the Triple β-Spiral fold. One advantage of simulated annealing is that it can accomodate many different scoring schemes by simply changing its scoring function. In this section we present three different scores: one based on the Triple β-Spiral profile, one based on the β-strand interactions within each chain of the Triple β-Spiral, and one based on the inter-chain packing interactions between the three identical chains in the Triple β-Spiral.

3.3.1 Profile Score

As a basic test of our simulated annealing algorithm, the first model that we employ for predicting the occurrence of the Triple β-Spiral fold is a slight modification to the profile methods presented in the previous chapter. For this approach, we developed two profiles: one profile from the Adenovirus repeats and one profile from the Reovirus repeats. [2]

We used these profile as scoring functions for placing repeats along a target sequence. Figures 3-2 and 3-3 present the profiles that we used for scoring repeats. Figures 3-4 and 3-5 present the results of our simulated annealing algorithm using these profiles.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	1.26	-2.32	1.00	-0.36	-2.59	0.38	-1.45	-3.49	-0.30	-2.01	-4.38	1.20	1.57	-0.05	-0.86	0.48	-0.81	-4.38	-1.99	-0.29
2	-1.76	-4.38	-3.42	-4.38	-4.38	2.77	-4.38	-4.38	-4.38	-4.38	-4.38	-3.26	3.14	-3.19	-4.38	-1.62	-1.69	-4.38	-4.38	-4.38
3	-3.18	-2.32	-3.42	-3.56	0.61	-3.59	-1.45	0.56	-3.50	2.86	0.34	-3.26	-2.03	-2.11	-4.38	-2.72	-1.33	-1.57	-4.38	-4.38
4	-0.03	-1.50	-0.55	-0.36	-3.22	-0.57	-0.16	-0.60	-0.45	-1.33	-0.85	-0.71	-2.03	1.64	-0.20	0.12	2.27	-0.52	-1.14	-0.16
5	-0.96	-2.31	-3.41	-0.42	1.65	-3.08	-0.79	0.82	0.19	0.39	-0.85	-1.00	-2.76	-1.26	-4.37	-1.00	0.89	1.82	-4.37	-0.42
6	-0.93	-4.35	1.43	-0.39	-1.08	-1.58	-1.42	-0.76	-0.50	-2.13	-4.35	1.78	-2.32	0.58	-0.83	1.12	1.65	-0.65	-4.35	-2.97
7	-0.22	-4.32	1.33	0.43	-3.16	0.78	-2.64	-4.32	-1.54	-2.96	-4.32	2.52	-0.11	-0.09	-2.37	1.07	-0.13	-2.32	-4.32	-4.32
loop:	-0.70	-3.83	-0.10	0.26	-0.76	0.12	-0.24	-0.87	-0.36	-0.78	-2.19	1.44	1.07	0.83	-1.88	0.96	0.85	-1.83	-0.59	-3.83
8	-1.48	-4.38	-0.75	-1.55	-4.38	2.42	0.28	-4.38	-1.43	-2.16	-4.38	2.67	-1.34	0.66	-1.44	0.07	-1.69	-4.38	-1.14	-4.38
9	1.19	-4.38	-2.44	-1.55	-1.54	0.03	-0.55	-1.59	0.87	-2.33	-0.85	1.05	-0.64	1.32	-0.64	0.65	1.10	-0.60	-4.38	-2.31
10	-2.54	-4.38	-3.42	-4.38	-3.22	-3.59	-2.70	1.67	-3.50	2.58	-0.40	-2.21	-1.17	-4.38	-4.38	-2.72	-2.48	0.80	-4.38	-2.31
11	0.78	0.50	-1.86	-0.36	-3.22	0.49	-1.45	-0.90	-0.71	-1.74	-4.38	-0.47	-1.34	0.71	-0.64	1.09	1.94	-0.13	-4.38	-4.38
12	0.02	-4.37	-2.84	-2.10	-3.21	-3.59	-4.37	0.64	-2.95	2.13	-1.14	-2.63	-1.53	-4.37	-2.42	-1.34	0.20	1.76	-4.37	-2.99
13	0.44	-4.37	-0.36	-1.26	-2.58	-1.46	-0.79	-2.94	1.70	-0.73	-4.37	1.92	-0.63	0.13	0.15	0.86	-0.04	-0.85	-1.99	-2.30
14	-0.05	-4.35	-2.42	-2.34	-0.49	-2.39	-2.67	0.87	-2.23	1.16	-0.59	-0.44	0.08	-4.35	-2.82	-1.20	1.58	1.15	-0.19	0.95
15	-0.17	-4.32	0.30	-1.21	-2.09	2.46	-4.32	-2.89	-0.31	-2.69	-1.45	0.14	-0.82	-0.41	-2.78	1.07	0.63	-1.86	-1.93	-4.32

Figure 3-2: **Adenovirus Profile Scores.** Bit scores for each position of a profile of the Triple β-Spiral fold. This profile was compiled using all of the repeats from the Adenovirus sequences in `Swiss-Prot`. Note that the solvent-exposed loop residues are not considered part of the profile, but are treated separately.

[2]Since we used all of the Adenovirus and Reovirus repeats in these two profiles they are not strictly comparable to any of the models that we developed in Chapter 2.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	-0.32	-1.04	0.90	-1.04	-1.04	-1.04	-1.04	-1.04	0.39	-1.04	-1.04	1.14	0.58	0.79	-0.07	0.62	0.86	-0.23	-1.04	-1.04
2	-1.04	-1.04	0.50	-0.22	-1.04	1.87	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	1.32	0.15	-1.04	-0.25	-0.10	-1.04	-1.04	1.04
3	-1.04	-1.04	-1.04	-1.04	1.80	-1.04	-1.04	-0.15	-1.04	1.82	1.34	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-0.23	-1.04	-1.04
4	-0.32	-1.04	-1.04	-0.22	-1.04	0.26	-1.04	-1.04	-1.04	-1.04	-1.04	0.70	0.58	1.23	-0.07	0.25	0.86	0.29	2.21	-1.04
5	-0.32	-1.04	-1.04	-1.04	1.80	-1.04	-1.04	1.57	0.39	-0.01	-1.04	-1.04	-1.04	-1.04	-1.04	0.25	-1.04	0.29	-1.04	0.34
6	-1.04	-1.04	0.50	-1.04	-1.04	-1.04	-1.04	0.39	-1.04	-1.04	-1.04	1.74	-1.04	0.79	1.49	-0.25	0.46	0.29	-1.04	-1.04
7	-1.04	-1.04	-0.08	-0.22	-1.04	0.93	-1.04	-0.15	-1.04	-1.04	-1.04	1.74	-1.04	-1.04	-0.07	1.37	-0.10	-0.23	-1.04	-1.04
loop:	-0.13	-1.32	-1.32	-1.32	-1.32	0.64	-1.32	-1.32	-1.32	-0.29	-1.32	1.88	-1.32	-0.14	-1.32	1.27	1.15	0.00	-1.32	-1.32
8	-1.04	-1.04	-1.04	0.68	-1.04	0.93	-1.04	-1.04	-0.16	0.32	-1.04	1.97	-1.04	-1.04	0.50	-0.25	-0.10	-1.04	-1.04	-1.04
9	0.16	-1.04	-1.04	-0.22	-1.04	0.63	-1.04	-1.04	-1.04	-0.43	-1.04	0.70	-1.04	1.57	-0.07	0.25	0.46	0.66	-1.04	-1.04
10	-0.32	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	1.10	-1.04	1.72	1.83	-1.04	-1.04	-1.04	-1.04	0.25	-1.04	-1.04	-1.04	-1.04
11	-1.04	-1.04	0.50	-1.04	0.12	-1.04	-1.04	-1.04	0.39	-1.04	-1.04	0.08	-1.04	0.15	0.91	0.25	1.65	0.29	-1.04	0.34
12	-1.04	-1.04	-1.04	-1.04	0.76	-1.04	-1.04	0.39	-1.04	1.47	1.34	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	1.42	-1.04	-1.04
13	-0.32	-1.04	0.50	-1.04	-1.04	0.26	-1.04	0.39	0.78	-0.43	0.60	0.08	-0.02	-1.04	1.23	0.25	-1.04	-0.23	-1.04	-1.04
14	-1.04	-1.04	-1.04	-1.04	0.12	0.26	-1.04	0.79	-1.04	-0.43	-1.04	1.14	-1.04	-1.04	-1.04	1.17	-0.10	0.66	-1.04	1.04
15	-0.32	-1.04	1.22	-0.22	-1.04	0.26	-1.04	-1.04	-1.04	-1.04	-1.04	1.47	-1.04	1.23	0.50	-1.04	0.46	0.29	-1.04	-1.04

Figure 3-3: **Reovirus Profile Scores.** Bit scores for each position of a profile of the Triple β-Spiral fold. This profile was compiled using all of the repeats from the Reovirus sequences in `Swiss-Prot`. Note that the solvent-exposed loop residues are not considered part of the profile, but are treated separately.

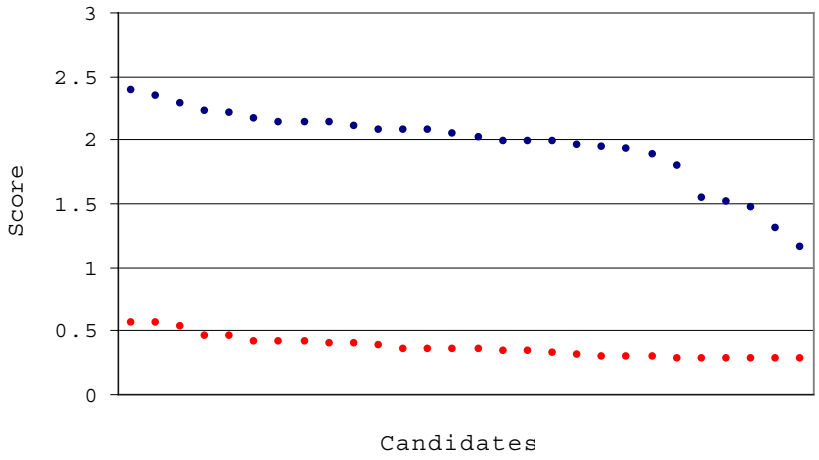

Figure 3-4: **Adenovirus Profile Scores from Simulated Annealing.** This graph shows the optimal scores from the Adenovirus repeat profile using simulated annealing. Adenovirus and Reovirus sequences are shown in blue, and the top 28 scoring decoys are shown in red.

The two simulated annealing profiles perform very well. The Adenovirus profile scores all of the Triple β-Spiral sequences from `Swiss-Prot` above all of the decoy sequences. The Reovirus profile scores 26 of the Triple β-Spiral sequences higher than all of the decoys. However, twelve decoys in this scheme score better than the bottom scoring Triple β-Spiral sequence in this scheme – `FIBP_ADE04` with a score of

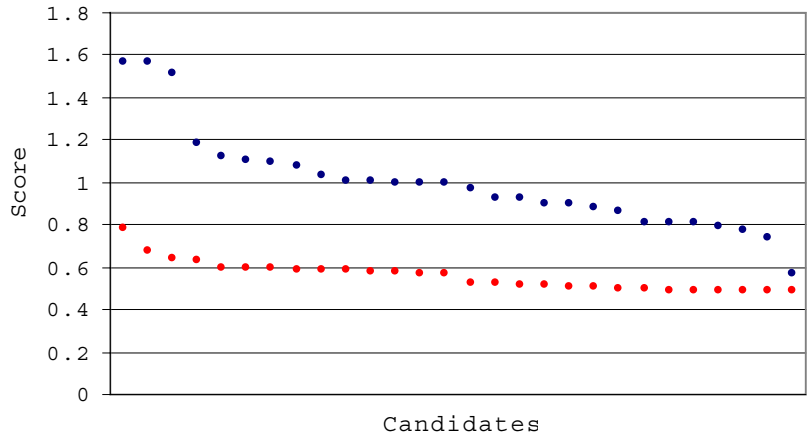

Figure 3-5: **Reovirus Profile Scores from Simulated Annealing.** This graph shows the optimal scores from the Reovirus repeat profile using simulated annealing. Adenovirus and Reovirus sequences are shown in blue, and the top 28 scoring decoys are shown in red.

0.57. Interestingly, the top scoring decoy was a Right-Handed Parallel β-Helix – `1HG8` with a score of 0.78. It is interesting that the top scoring hits were different for this method than for the HMM presented in Model 7. This probably occurred because we are only taking four repeats into account in these profile models, whereas the models in Chapter 2 could accomodate additional repeats, and did not enforce a strict limit on the gaps between repeats.

It is, of course, counterproductive to use simulated annealing to find optimal profile scores when we could employ dynamic programming to achieve a more reliable result in a fraction of the time. We present this result only to provide a basic comparison of this approach to the models presented in the previous chapter.

3.3.2 Antiparallel β-Strand Pair Score

A more interesting test of our simulated annealing algorithm is a score based upon the pairwise residue correlation probabilities in antiparallel β-strands. This approach is motivated by Lifson and Sander's discovery in 1983 that the pairs of hydrogen bonded amino acids in antiparallel β-sheets differ significantly from what we would

expect from a simple uncorrelated model [46].

Over the past 20 years, many groups have tried to use Lifson and Sander's observation to aid in structure prediction [66, 34, 56]. These efforts were largely unsuccessful until 2002, when Bradley et al. introduced the `BetaWrap` program. The `BetaWrap` program uses a scoring function based upon paired residue correlations in β-strands to successfully predict the occurrence of the Right-Handed Parallel β-Helix fold in protein sequence databases.[3]

In this thesis, we have modified the `BetaWrap` scoring function to create a scoring function for Triple β-Spiral sequences. In our approach, a score at each annealing step is generated based upon the proposed residue pairing in the β-sheets of each of the four repeats. (Recall that each β-sheet contains paired β-strands from one structural repeat.) The β-strand bit-score for a single structural repeat is calculated as

$$\beta\text{-Strand Score} = \sum_{i=1}^{5} s_\beta(i, 13-i)/5$$

where $s_\beta(i,j)$ is the score for placing residue at position i next to the residue at position j in an antiparallel β-strand. The s_β scores are shown in Figure 3-6 and are calculated as

$$s_\beta(r_1, r_2) = \log_2 \frac{p(r_1, r_2)}{b_{r_1} b_{r_2}}$$

where $p(r_1, r_2)$ is the pseudocount-weighted probability of observing residues of type r_1 and r_2 in antiparallel β-strands:

$$p(r_1, r_2) = \frac{c_{r_1 r_2} + A b_{r_1} b_{r_2}}{n + A} \tag{3.1}$$

The score for each of the four repeats is calculated in this way, and then the final

[3] `BetaWrap` also integrates expert knowledge based scores and weights that are external to this probabilistic framework.

bit-score at each annealing step is found by normalizing by the total number of repeat bit-scores.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	-0.26	1.08	-0.40	-0.16	1.41	-0.71	0.19	1.65	-0.13	1.10	0.50	-0.40	-1.31	-0.31	0.29	-0.10	0.92	2.05	1.57	1.74
C	1.08	3.94	0.08	0.15	2.11	0.54	1.37	1.94	0.61	1.49	1.72	0.16	-0.31	0.68	0.88	0.94	1.18	2.17	3.11	2.24
D	-0.40	0.08	-1.40	-0.50	0.03	-0.96	0.63	0.12	0.95	-0.48	-0.31	0.37	-2.06	0.32	0.95	0.17	1.03	0.66	0.26	0.50
E	-0.16	0.15	-0.50	-0.85	0.62	-1.51	0.72	0.80	1.56	0.20	-0.03	0.60	-1.68	0.38	1.80	0.44	1.44	1.11	0.80	1.17
F	1.41	2.11	0.03	0.62	1.82	0.48	1.30	2.39	0.68	1.92	1.86	0.01	0.12	0.92	1.01	0.78	1.22	2.82	2.64	2.70
G	-0.71	0.54	-0.96	-1.51	0.48	-0.52	-0.48	0.28	-1.71	-0.28	-0.76	-0.89	-1.72	-0.98	-1.07	-0.81	-0.25	0.74	1.17	0.67
H	0.19	1.37	0.63	0.72	1.30	-0.48	0.54	1.33	0.68	0.52	0.48	0.54	-0.32	0.80	0.90	1.03	1.74	1.54	1.72	1.73
I	1.65	1.94	0.12	0.80	2.39	0.28	1.33	1.90	0.94	2.09	1.87	-0.07	-0.74	0.92	1.19	0.76	1.50	2.93	2.22	2.62
K	-0.13	0.61	0.95	1.56	0.68	-1.71	0.68	0.94	-0.50	0.29	0.55	0.32	-1.91	0.79	0.19	0.59	1.68	1.34	1.66	1.82
L	1.10	1.49	-0.48	0.20	1.92	-0.28	0.52	2.09	0.29	0.68	1.25	-0.33	-0.80	0.52	0.57	0.14	0.75	2.33	2.07	1.88
M	0.50	1.72	-0.31	-0.03	1.86	-0.76	0.48	1.87	0.55	1.25	0.72	-0.20	-0.67	0.48	0.31	0.39	0.83	2.10	1.69	1.58
N	-0.40	0.16	0.37	0.60	0.01	-0.89	0.54	-0.07	0.32	-0.33	-0.20	-0.28	-1.39	0.62	0.38	0.70	1.24	0.57	1.16	1.07
P	-1.31	-0.31	-2.06	-1.68	0.12	-1.72	-0.32	-0.74	-1.91	-0.80	-0.67	-1.39	-4.41	-1.11	-1.07	-1.17	-0.41	-0.16	0.72	0.76
Q	-0.31	0.68	0.32	0.38	0.92	-0.98	0.80	0.92	0.79	0.52	0.48	0.62	-1.11	-0.12	1.13	0.75	1.73	1.35	1.72	1.58
R	0.29	0.88	0.95	1.80	1.01	-1.07	0.90	1.19	0.19	0.57	0.31	0.38	-1.07	1.13	-0.66	0.67	1.58	1.67	1.76	1.83
S	-0.10	0.94	0.17	0.44	0.78	-0.81	1.03	0.76	0.59	0.14	0.39	0.70	-1.17	0.75	0.67	-0.09	1.49	1.14	1.56	1.38
T	0.92	1.18	1.03	1.44	1.22	-0.25	1.74	1.50	1.68	0.75	0.83	1.24	-0.41	1.73	1.58	1.49	1.68	1.99	0.79	1.78
V	2.05	2.17	0.66	1.11	2.82	0.74	1.54	2.93	1.34	2.33	2.10	0.57	-0.16	1.35	1.67	1.14	1.99	2.37	2.43	2.84
W	1.57	3.11	0.26	0.80	2.64	1.17	1.72	2.22	1.66	2.07	1.69	1.16	0.72	1.72	1.76	1.56	0.79	2.43	2.38	2.75
Y	1.74	2.24	0.50	1.17	2.70	0.67	1.73	2.62	1.82	1.88	1.58	1.07	0.76	1.58	1.83	1.38	1.78	2.84	2.75	1.85

Figure 3-6: **β-Strand Pairwise Correlation Scores.** Bit scores for pairwise residue packing in antiparallel β-strands. The pairwise counts that were used to generate these scores were compiled from a complete search of the PDB using the same parameters for β-strand recognition as Lifson and Sander [46]. We did not use the BetaWrap database because it was reported as conditional probabilities, and we wanted to use absolute pairing probabilities.

Figure 3-7 shows the results of running this algorithm on our candidates. Note that we do not need to split our model into Adenovirus and Reovirus trained sub-models in this case because the β-correlation probabilities that we are employing was drawn from all antiparallel β-proteins and is not specific to the Triple β-Spiral fold.[4] The top 28 scoring decoys strictly dominate the Triple β-Spiral sequences in this method. This indicates that this score alone may not a good metric for identifying the Triple β-Spiral fold.

[4]Although we did not specifically omit Ad2 and Rσ1 when we constructed out β-correlation probability table, we consider the contribution of these two proteins (from among the thousands of proteins that we drew from the PDB for this effort to be insignificant.

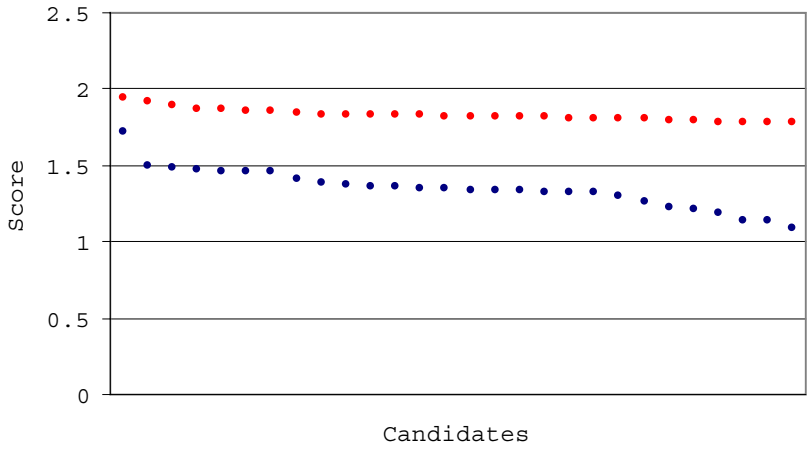

Figure 3-7: **β-Strand Scores from Simulated Annealing.** This graph shows the optimal scores from the Triple β-Spiral repeat β-strands using simulated annealing. Adenovirus and Reovirus sequences are shown in blue, and the top 28 scoring decoys are shown in red.

3.3.3 Inter-Chain Hydrogen Bonding Scores

As we have seen, the Triple β-Spiral fold contains three identical and interacting protein chains. This means that a Triple β-Spiral sequence has an implied set of interactions not only with itself (via antiparallel β-strand hydrogen bonding) but also with each of its two sister chains. This is a situation similar to homotrimeric three-stranded α-helical coiled-coils [58, 7, 6].

One reasonable method of approach, then is to create a database similar to that in Figure 3-6 but which contains the inter-chain hydrogen bonding residue correlation scores at the protein interface. These hydrogen bonds are shown in Figure 2-3. We constructed this database by listing the residue pairs in all of the Triple β-Spiral sequence in the `Swiss-Prot` database. We then calculated residue-residue pair scores according to Equation 3.1.

We realize that we are significantly biasing our results by including only the 28 Triple β-Spiral sequences. However, we proceeded in this way because there are so few representatives for this fold. Furthermore, because there are so few known Reovirus fibers we do not split our analysis into two parts but rather simply consider how well

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.01	-6.30	-1.63	-2.86	-6.30	-0.72	-1.42	1.06	-0.51	2.70	0.48	0.80	1.02	-2.17	-2.57	-1.45	-1.12	1.51	-0.50	-6.30
C	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-0.70	-6.30	-6.30
D	-1.63	-6.30	-2.06	-6.30	-6.30	-0.18	-6.30	0.88	-2.22	2.73	-0.95	0.19	0.58	-6.30	-6.30	-2.43	0.14	1.12	-6.30	-6.30
E	-2.86	-6.30	-6.30	-6.30	-1.99	-0.80	-6.30	-0.58	-6.30	1.16	-6.30	-6.30	0.49	-1.94	-6.30	-1.78	-2.40	1.03	-6.30	-6.30
F	-6.30	-6.30	-6.30	-1.99	-6.30	-6.30	-6.30	-6.30	-6.30	-1.56	-6.30	-6.30	-6.30	-1.28	-1.68	-1.11	-0.21	-0.48	-6.30	-6.30
G	-0.72	-6.30	-0.18	-0.80	-6.30	3.09	1.87	2.75	-1.07	3.37	0.23	3.76	3.92	1.61	-2.41	-0.09	0.02	1.67	-6.30	-6.30
H	-1.42	-6.30	-6.30	-6.30	-6.30	1.87	-6.30	-1.04	-1.05	0.24	-6.30	-6.30	0.80	-6.30	-6.30	-6.30	-6.30	-0.23	-6.30	-6.30
I	1.06	-6.30	0.88	-0.58	-6.30	2.75	-1.04	-6.30	2.29	0.32	-1.10	2.92	0.16	0.15	-0.67	1.96	0.98	-0.03	-6.30	-1.48
K	-0.51	-6.30	-2.22	-6.30	-6.30	-1.07	-1.05	2.29	-6.30	2.29	1.44	-1.93	0.15	-1.81	-6.30	-1.08	-1.32	2.17	-6.30	-1.50
L	2.70	-6.30	2.73	1.16	-1.56	3.37	0.24	0.32	2.29	0.16	-0.81	3.29	2.65	2.08	1.19	2.93	2.12	0.68	-0.80	1.10
M	0.48	-6.30	-0.95	-6.30	-6.30	0.23	-6.30	-1.10	1.44	-0.81	-6.30	-6.30	0.73	-6.30	-0.94	-6.30	-0.02	-6.30	-6.30	-6.30
N	0.80	-6.30	0.19	-6.30	-6.30	3.76	-6.30	2.92	-1.93	3.29	-6.30	-1.47	3.85	-6.30	-1.76	0.37	1.58	1.43	-6.30	-6.30
P	1.02	-6.30	0.58	0.49	-6.30	3.92	0.80	0.16	0.15	2.65	0.73	3.85	-1.83	1.42	0.81	2.55	0.26	1.74	1.15	-6.30
Q	-2.17	-6.30	-6.30	-1.94	-1.28	1.61	-6.30	0.15	-1.81	2.08	-6.30	-6.30	1.42	-6.30	-6.30	0.24	-0.73	1.77	-6.30	-6.30
R	-2.57	-6.30	-6.30	-6.30	-1.68	-2.41	-6.30	-0.67	-6.30	1.19	-0.94	-1.76	0.81	-6.30	-6.30	-0.90	-2.10	-0.12	-6.30	-0.34
S	-1.45	-6.30	-2.43	-1.78	-1.11	-0.09	-6.30	1.96	-1.08	2.93	-6.30	0.37	2.55	0.24	-0.90	-0.89	0.42	1.39	-6.30	-1.71
T	-1.12	-6.30	0.14	-2.40	-0.21	0.02	-6.30	0.98	-1.32	2.12	-0.02	1.58	0.26	-0.73	-2.10	0.42	-2.17	1.65	-6.30	-6.30
V	1.51	-0.70	1.12	1.03	-0.48	1.67	-0.23	-0.03	2.17	0.68	-6.30	1.43	1.74	1.77	-0.12	1.39	1.65	-1.74	-6.30	-0.11
W	-0.50	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-0.80	-6.30	-6.30	1.15	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	0.80
Y	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-6.30	-1.48	-1.50	1.10	-6.30	-6.30	-6.30	-6.30	-0.34	-1.71	-6.30	-0.11	0.80	-6.30

Figure 3-8: **Triple β-Spiral Inter-chain Hydrogen Bonding Scores.** Bit scores for pairwise hydrogen bonding pairs in the Triple β-Spiral fold. These scores were calculated from the 28 Triple β-Spiral fibers in `Swiss-Prot`.

the joint Adenovirus/Reovirus inter-chain hydrogen bonding scores do at predicting the Triple β-Spiral fold. It might have been interesting to also use the scores from Figure 3-6 or to compile inter-chain hydrogen bonding frequencies from other trimeric proteins to create this database. We plan to pursue both of these approaches in the future.

To score a single repeat in our simulated annealing algorithm, we use the following equation:

$$\text{H-bonding Score} = s_H(2,8) + s_H(13,10) + s_H(15,10) + s_{H'}(1,12)$$

where $s_H(i,j)$ is the hydrogen bonding score for residue i on the *previous* repeat and residue j on this repeat, and $s_{H'}(i,j)$ is the hydrogen bonding score for residues i and j on the same repeat.

Figure 3-9 gives the results of running our simulated annealing algorithm using this scoring scheme. Note that this scheme does better than the β-strand scoring scheme but not as well as the profile score at ranking Triple β-Spiral sequences. The best

scoring decoy (an aldehyde ferrodoxin oxireductase) outscores all of the Triple β-Spiral folds. Among the top scoring 28 decoys about half score better than the true Triple β-Spiral folds. Thus, although the inter-chain hydrogen bonding score does reasonably well (there are more than 2000 decoys that this score strictly outperforms) it does not provide good separation between the Triple β-Spiral and non-Triple β-Spiral folds.

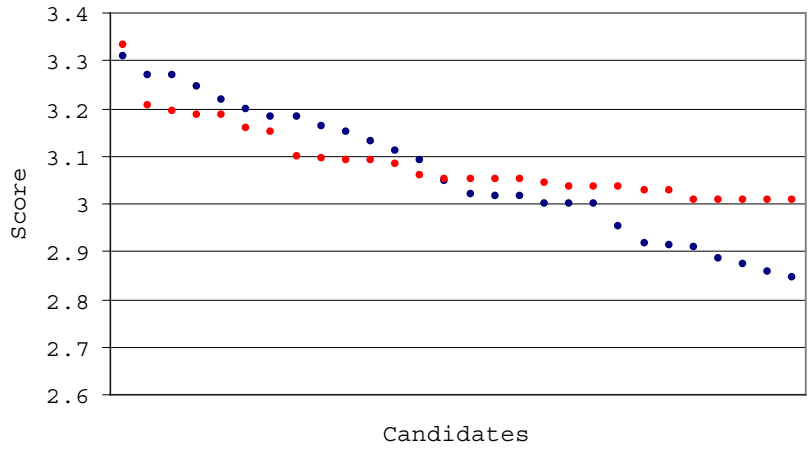

Figure 3-9: **Pairwise Hydrogen Bonding Scores from Simulated Annealing.** This graph shows the optimal scores from the Triple β-Spiral repeat hydrogen bonding pattern using simulated annealing. Adenovirus and Reovirus sequences are shown in blue, and the top 28 scoring decoys are shown in red.

3.4 Discussion

In this chapter we developed a general simulated annealing algorithm for optimizing protein structure scores. Although the parameterization of this algorithm was somewhat arbitrary – we chose to place four repeats with a maximum loop insertion of eight residues and a maximum inter-repeat insertion of six residues – the algorithm itself is comletely general and could accomodate other parameters and even other structural templates. The main drawback to this approach is that it is slower than other structure prediction algorithms (like dynamic programming) and therefore not appropriate for searching large sequence databases.

Using this simulated annealing algorithm we tested three different scoring functions for the Triple β-Spiral fold. First, we tested a profile-based score that was similar to the methods developed in Chapter 2. Second, we tested a β-strand correlation probability score that ranked structures based on the β-strand pairings in each structural repeat. Third, we tested a score based upon the inter-chain residue correlation frequencies. This score took advantage of the homotrimeric partnering in the hydrophobic core of the Triple β-Spiral fold.

In general, the profile-based method outperformed the other two. The Adenovirus-trained model correctly scored all of the Triple β-Spiral sequences above all of the decoys, but the Reovirus-trained model performed worse than the Reovirus-trained HMM in Chapter 2.

Although the profile-based method performed better than the other two, this does not mean that the β-strand and inter-chain hydrogen bonding scores are not useful. In fact, in the next chapter we will devleop a method to integrate all of these diverse scores into a single framework.

Chapter 4

Score Integration

In the previous two chapters we investigated a series of increasingly complex scoring methods for discovering the Triple β-Spiral from primary sequence data. It might be reasonable to expect that a true Triple β-Spiral fold will score well in each of these models whereas a good decoy might do well in one model but poorly in another. The question that we now turn to is whether we can integrate these scores into a single scoring scheme that outperforms any of the individual scoring schemes.

4.0.1 Integration Framework

In the previous chapter we developed three different scores for a four-repeat parse of a potential Triple β-Spiral sequence. These three scores were:

1. a profile-based score that modeled the residue frequencies at each position in the repeat,

2. a β-strand score that modeled the residue packing frequencies in each repeat unit, and

3. an inter-chain packing score that modeled the likelihood of structural repeats joining together in a trimeric fiber shaft.

If we are given a sequence that may or may not be a Triple β-Spiral, then we can place four consecutive repeats along this sequence and evaluate each of these three

scores simultaneously.

To integrate these three scores into a single framework, we employ logistic regression [27]. To perform this regression, we first develop a training set that contains both true Triple β-Spiral sequences, and a set of decoys. We use logistic regression to model the log-odds probability that a sequence is a Triple β-Spiral given the three scores that we developed in the previous chapter. The model takes the form:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_B s_B + \beta_H s_H + \beta_P s_P$$

where p is the probability that a sequence is a Triple β-Spiral β_0 is the intercept, and the $(\beta_B, \beta_H, \beta_P)$ are the coefficients for the β-strand pairing, inter-chain hydrogen bonding, and profile scores (s_B, s_H, s_P) respectively.

We used the SAS software package to perform the logistic regression using a training set of Triple β-Spiral and non-Triple β-Spiral sequences that we describe below.

4.1 Creation of a training set

To create the training set for our logistic regression, we first created a set of 500 randomly selected decoy sequences from the `PDB`. These 500 training sequences are different from the 2200 decoy sequences used to test the three scoring schemes in the previous chapter, as we wanted to reuse the decoy sequences for testing our score-integration framework. To these 500 training sequences, we added another 32 decoys from among the 2200 test sequences. We added these 32 decoy sequences so that we would have a few good candidates among our training set.

We created a set of true Triple β-Spiral sequences by splitting the Triple β-Spiral sequences from the `Swiss-Prot` database into sequences containing exactly four fiber repeats. Thus, for example, the `FIBP_ADE02` sequence was split into 18 separate sequences. There were 253 total Triple β-Spiral sequences when compiled in this way.

To place four repeats along the 532 decoy and 253 genuine Triple β-Spiral se-

(a)

Parameter	Estimate	Std. Err.	p-Score
b0	-3.1284	1.6946	0.0649
b1	-3.8234	2.5342	0.1314
b2	1.6166	0.8950	0.0709
b3	9.9866	2.7050	0.0002

(b)

Parameter	Estimate	Std. Err.	p-Score
b0	-11.6644	14.2366	0.4126
b1	-3.5336	20.3906	0.8624
b2	0.2319	4.7293	0.9609
b3	18.5516	13.4706	0.1685

Table 4.1: **Logistic Regression Coefficients.** The logistic regression coefficients for (a) Adenovirus-trained, and (b) Reovirus-trained models. The coefficients are for the the optimal log-odds linear model of Triple β-Spiral sequences.

quences, we created a profile HMM from an alignment of four-repeat Triple β-Spiral segments from Adenovirus and Reovirus fibers. We then simultaneously scored each of the resulting sequences using the three scoring schemes from Chapter 3. Figure 4-1 provides two-dimensional slices of the three-dimensional scores using the Adenovirus profile. Figure 4-2, provides two-dimensional slices of the three-dimensional scores using the Reovirus profile. Note that as a joint score using both profile and hydrogen bonding information seems to provide greater separation than either of the two methods alone, whereas the β-strand score does not seem to provide much information. These results are consistent with our experiences using the simulated annealing algorithm with these scoring schemes in the previous section.

4.2 Logistic Regression Results

We peformed two logistic regressions. In the first, we used an Adenovirus profile and scores only from the Adenovirus fiber data set. In the second, we used a Reovirus profile and scores only from the Reovirus fiber data set. In both regressions, we modeled the event that the sequence was a Triple β-Spiral.

The model developed from the Adenovirus data set achieved good convergence

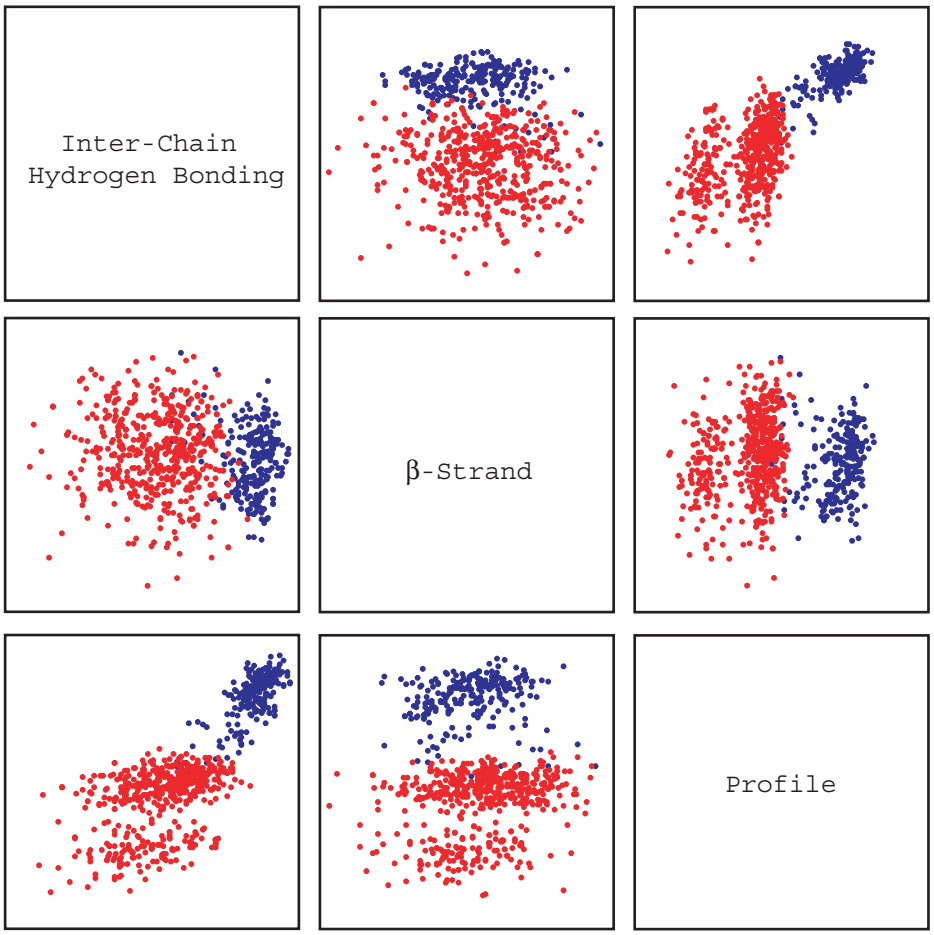

Figure 4-1: **Adenovirus-Based Scores.** Two-dimensional representation of the inter-chain hydrogen bonding, β-strand, and profile scores for the 536 decoy and 253 bona fide Triple β-Spiral folds. These scores were developed using profile information from the Adenovirus fiber.

and a model p-score of less than 0.0001, indicating that the model fit the data quite well. The model developed from the Reovirus data set did not converge – probably because there were so few Reovirus data points that complete separation of the Triple β-Spiral sequences from the decoys was possible. Table 4.1 summarizes the output from the logistic regression. Appendix B gives the complete SAS output.

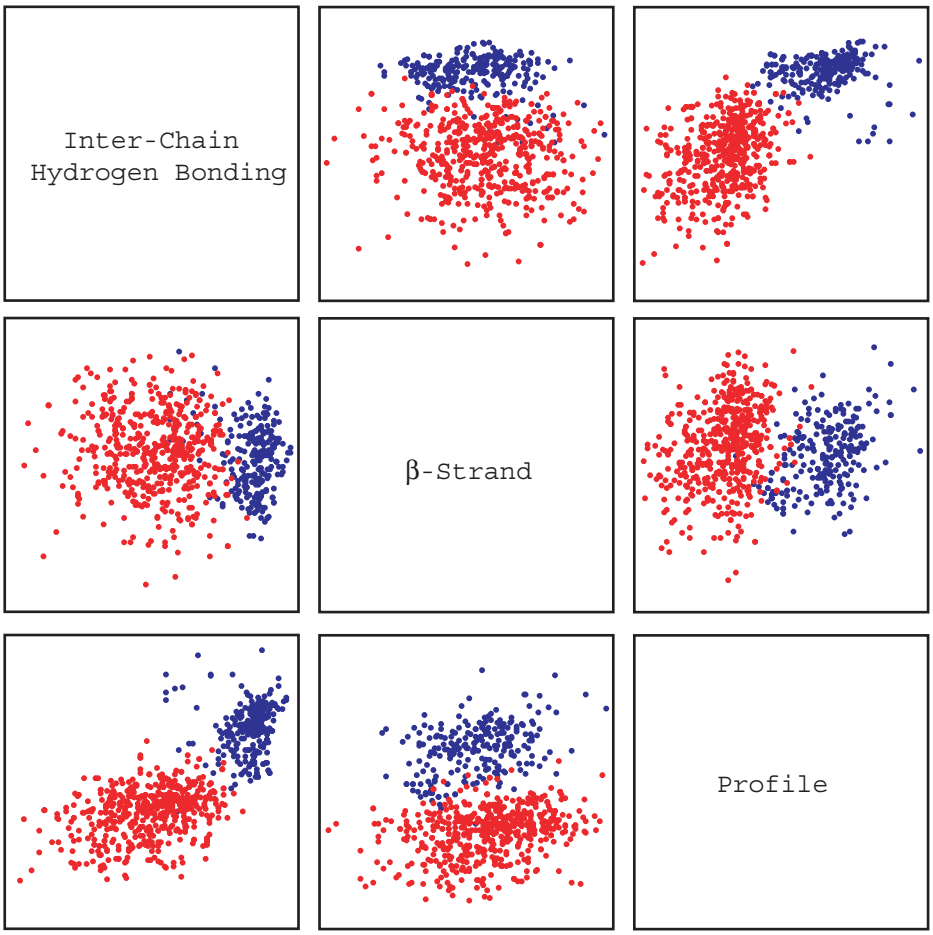

Figure 4-2: **Reovirus-Based Scores.** Two-dimensional representation of the inter-chain hydrogen bonding, β-strand, and profile scores for the 536 decoy and 253 bona fide Triple β-Spiral folds. These scores were developed using profile information from the Reovirus fiber.

4.3 Simulated Annealing

We can use the logistic regression coefficients from the Adenovirus and Reovirus models to create a simple score integration framework for simulated annealing. In this framework, a potential Triple β-Spiral sequence is given a score according to the linear coefficients of the logistic regression. The goal of the simulated annealing algorithm is to maximize the log-odds score, and thereby maximize the probability that a sequence is, in fact, a Triple β-Spiral.

Figure 4-3 presents a comparison of the 28 Triple β-Spiral sequences and the top scoring 28 decoys under this scoring scheme. Note that both of these models

perform extremely well, with the Adenovirus-trained model scoring 27 of the Triple β-Spiral sequences above all of the decoys, and the Reovirus-trained model scoring all of the Triple β-Spiral sequences above all of the decoys. In short, this integrated model outperforms all of the previous models that we have developed, and its success suggests that an integrated scoring scheme may be a good choice for prediction of other structures.

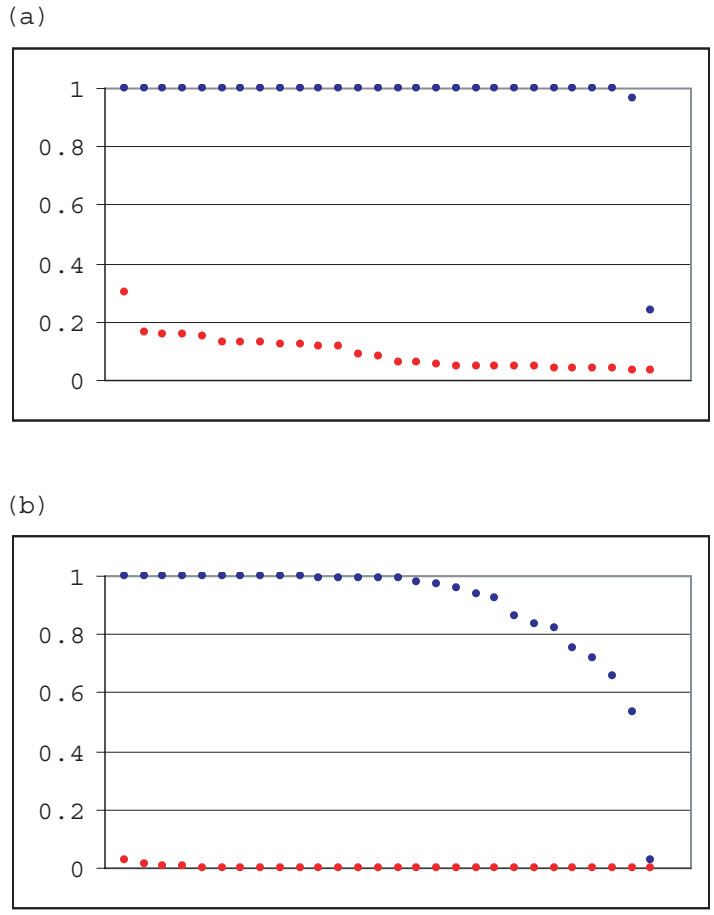

Figure 4-3: **Simulated Annealing Logistic Scores.** Scaled probabilities for the 28 Triple β-Spiral sequences in `Swiss-Prot` (in blue) and the top 28 scoring decoys (in red). The scores are created using the Logistic Regression coefficients for the (a) Adenovirus and (b) Reovirus models.

4.4 Discussion

Although the integrated scoring scheme that we presented in this chapter did not achieve perfect separation between Triple β-Spiral and non-Triple β-Spiral sequences in cross-validation tests, it did perform well – with only one decoy scoring above a single known Triple β-Spiral. This success suggests that we can improve homology-modeling methods by incorporating additional structural information.

Although our results are promising, in hindsight it is evident to us that we should have developed and tested these methods with a fold other than the Triple β-Spiral. The next fold that we attempt should have at least 10 representatives with less than 25% sequence identity. This will allow us to test our methods on candidates that share the same fold but are not well-predicted by homology modeling.

In the future, we would like to integrate additional types of scores into our framework. These scores might include metrics like:

Secondary Structure Prediction There are several good tools for predicting the secondary structure of a protein from primary sequence data [36]. This could supplement our methods by indicating the likely secondary structure at each simulated annealing step.

Sequence-Structure Predictions Sequence-structure tools like `Trilogy` [11] and `Conservatism-of-Conservatism` [49] detect distant structural correlations and could supplement existing profile-based methods.

Three-dimensional Threading Because our methods imply a three-dimensional threading of an amino-acid sequence along a structural model, we might improve the sensitivity of our search by incorporating solvent and steric effects as is done in existing threading methods [59].

We did not incorporate these metrics into our current scoring scheme because each of these is computationally intensive, and therefore not appropriate for an iterative algorithm like simualted annealing. In our future work we may explore other optimization methods in order to address this issue.

Appendix A

Model 7 Hits

This is a complete list of hits from the Swiss-Prot/TrEMBL database for Model 7 from Chapter 2. The list is sorted based upon the higher E-value (Adenovirus or Reovirus seeded-model) that it attained. If only one of the two models found the sequence, then there is a - in the column for the model that did not find the sequence.

Swiss-Prot	Protein Name	Source	Ad E-val	Re E-val
FIBP_ADEB3	Fiber protein	Bovine adenovirus...	1.60E-89	1.90E-93
Q64787	Long fiber	Avian adenovirus...	5.40E-52	1.60E-53
FIBP_ADE12	Fiber protein	Human adenovirus...	1.90E-51	5.60E-52
FIB1_ADEG1	Fiber protein 1	Avian adenovirus...	6.30E-51	3.10E-52
FIB1_ADE41	Fiber protein 1	Human adenovirus...	1.20E-44	7.50E-46
Q83893	Fiber	Ovine adenovirus...	1.80E-43	3.10E-45
FIB1_ADE40	Fiber protein 1	Human adenovirus...	3.40E-39	3.00E-41
FIBP_ADE31	Fiber protein	Human adenovirus...	9.50E-36	9.50E-38
FIBP_ADECT	Fiber protein	Canine adenovirus...	1.20E-35	1.90E-37
FIBP_ADE05	Fiber protein	Human adenovirus...	2.40E-35	9.50E-33
FIBP_ADECC	Fiber protein	Canine adenovirus...	1.70E-32	2.60E-34
FIBP_ADECG	Fiber protein	Canine adenovirus...	1.70E-32	2.60E-34
FIBP_ADECR	Fiber protein	Canine adenovirus...	1.70E-32	2.60E-34

Q8QVG3	Fiber	Bovine adenovirus...	4.10E-34	3.90E-32
Q9DLD8	Fiber protein	Bovine adenovirus...	4.20E-32	3.70E-32
FIBP_ADE02	Fiber protein	Human adenovirus...	2.40E-34	1.70E-31
Q911A1	Fiber protein	Human adenovirus...	3.50E-34	1.80E-31
Q8B4M4	Fiber protein...	Bovine adenovirus...	1.00E-31	1.40E-30
Q96590	Partial fiber...	Human adenovirus...	4.40E-33	4.30E-30
Q911A0	Fiber protein	Human adenovirus...	1.10E-32	4.80E-30
Q910N0	Fiber protein	Human adenovirus...	1.00E-32	5.70E-30
O11424	Fiber protein	Duck adenovirus 1	2.60E-29	5.70E-27
Q9E8G1	Fiber	Porcine adenovirus 5	4.20E-30	7.90E-27
P87656	Fiber	Duck adenovirus 1	7.80E-29	1.90E-26
Q994D4	Fiber protein	Porcine adenovirus 5	3.60E-30	3.60E-26
FIBP_ADEP3	Fiber protein	Porcine adenovirus...	9.70E-26	2.40E-27
FIB2_ADE40	Fiber protein 2	Human adenovirus...	6.40E-26	1.70E-25
FIB2_ADE41	Fiber protein 2	Human adenovirus...	8.10E-26	2.10E-25
FIBP_ADEM1	Fiber protein	Mouse adenovirus...	8.60E-26	5.50E-25
Q997H2	Fiber	Bovine adenovirus 4	8.30E-25	2.90E-26
Q98XR5	Fiber protein	Odocoileus...	5.20E-25	1.00E-24
Q8QZQ6	Hypothetical...	Chilo iridescent...	3.30E-18	3.10E-19
Q9IIG6	Fiber protein	Frog adenovirus 1	5.20E-20	4.30E-18
Q8P942	YapH protein	Xanthomonas...	4.30E-31	2.50E-17
Q88RG2	Surface...	Pseudomonas putida...	-	7.10E-17
Q8GDL9	Orf2	Photorhabdus...	6.90E-14	3.40E-13
Q64855	Fiber protein	Human adenovirus...	2.30E-12	4.20E-12
Q8PKM0	YapH protein	Xanthomonas...	2.70E-22	6.40E-12
Q96731	PVIII, fiber,...	fowl adenovirus	9.20E-12	9.10E-13
Q9F285	YapH protein	Yersinia pestis	1.40E-14	5.50E-11
Q8ZHA1	Putative...	Yersinia pestis	9.90E-15	5.50E-11
Q8CZU2	Putative...	Yersinia pestis	9.90E-15	5.50E-11
Q8PF72	YapH protein	Xanthomonas...	1.10E-12	8.60E-11

Q8QZQ8	261R	Chilo iridescent...	5.20E-11	1.50E-10
Q96739	Fiber	Avian adenovirus...	2.60E-10	7.80E-11
Q8UY68	PIV	Simian adenovirus 25	6.90E-11	6.40E-10
FIBP_ADE1A	Fiber protein	Human adenovirus...	2.30E-09	2.40E-09
Q8V791	Fiber protein	Human adenovirus...	2.20E-09	2.70E-09
FIBP_ADE03	Fiber protein	Human adenovirus...	3.30E-09	6.60E-09
Q83122	Fiber	Mastadenovirus	3.30E-09	6.60E-09
FIBP_ADE08	Fiber protein	Human adenovirus...	6.90E-09	8.50E-10
Q880E1	Filamentous...	Pseudomonas syringae...	9.30E-14	1.00E-08
Q9YYQ4	FIBRE homolog	Avian adenovirus...	1.30E-08	4.10E-09
Q9PWU3	Fiber protein	Human adenovirus...	2.00E-08	1.10E-08
Q9YUQ4	Fiber protein	Turkey adenovirus 3	2.20E-08	2.50E-09
O56261	Fiber protein	Turkey adenovirus 3	2.20E-08	2.50E-09
Q67711	Serotype 16...	Human adenovirus...	2.80E-08	4.50E-09
Q9QL91	Fiber protein	Human adenovirus...	3.00E-08	2.10E-08
O55281	Short fiber...	Avian adenovirus...	3.20E-08	4.00E-08
Q67714	Fiber protein	Human adenovirus...	1.10E-08	5.50E-08
Q67733	Fiber	Human adenovirus...	1.10E-08	5.50E-08
Q91CL7	Fiber protein	Human adenovirus...	1.10E-08	5.50E-08
FIBP_ADE1P	Fiber protein	Human adenovirus...	7.50E-08	8.00E-08
Q67713	Fiber protein	Human adenovirus...	1.20E-07	1.00E-07
Q80IV3	Fifth late...	Human adenovirus...	1.40E-07	1.30E-08
FIBP_ADE09	Fiber protein	Human adenovirus...	1.70E-07	8.20E-09
FIBP_ADE07	Fiber protein	Human adenovirus...	1.30E-07	2.00E-07
O56784	Fiber protein	Human adenovirus...	7.70E-07	1.20E-06
Q98E20	Hypothetical...	Rhizobium loti...	2.40E-12	1.20E-06
Q64823	Fiber protein	Human adenovirus...	1.40E-06	7.30E-07
FIBP_ADE15	Fiber protein	Human adenovirus...	1.60E-06	1.20E-07
Q9XC47	Outer membrane...	Rickettsia australis	3.30E-06	-
FIB2_ADEG1	Fiber protein 2	Avian adenovirus...	1.70E-05	3.30E-06

Q64790	Short fiber	Avian adenovirus...	1.70E-05	3.30E-06
Q86843	Fiber	Human adenovirus...	1.90E-05	2.00E-05
Q67712	Fiber protein	Human adenovirus...	5.10E-06	4.20E-05
Q85684	Viral...	Reovirus sp.	8.20E-05	1.80E-05
Q9I120	Hypothetical...	Pseudomonas...	7.10E-06	0.00018
TROP_HUMAN	Trophinin	Homo sapiens (Human)	3.10E-05	0.00026
Q840U5	Outer membrane...	Rickettsia...	0.0004	-
Q840U6	Outer membrane...	Rickettsia...	0.00046	-
Q85697	Viral...	Reovirus sp.	0.0006	8.90E-05
Q64822	Fiber protein	Human adenovirus...	0.00067	0.00044
Q9KKA1	OmpB [Fragment]	Rickettsia slovaca	0.00067	-
O56783	Fiber protein	Human adenovirus...	0.00056	0.0011
Q9KKA4	OmpB [Fragment]	Rickettsia sp. S	0.0015	-
Q87AN1	...	Xylella fastidiosa...	0.0011	0.0026
Q85686	Viral...	Reovirus sp.	0.0027	6.20E-05
LY_BPSF6	Lysozyme	Bacteriophage SF6	0.0028	1.60E-05
VSI1_REOVD	Sigma 1 protein...	Reovirus (type 3 /...	0.0029	6.20E-05
Q85683	Viral...	Reovirus sp.	0.0029	6.20E-05
P90215	Viral...	Reovirus sp....	0.0029	6.20E-05
Q86337	Viral...	Reovirus sp.	0.0029	6.20E-05
Q86331	Viral...	Reovirus sp.	0.0029	6.20E-05
Q86335	Viral...	Reovirus sp.	0.0029	6.20E-05
Q86333	Viral...	Reovirus sp.	0.0029	6.20E-05
P90216	Viral...	Reovirus sp....	0.0029	6.20E-05
Q86329	Viral...	Reovirus sp.	0.0029	6.20E-05
P90214	Viral...	Reovirus sp....	0.0029	6.20E-05
Q8GD27	Adhesin FhaB	Bordetella avium	0.0034	-
FIBP_ADE04	Fiber protein	Human adenovirus...	0.00023	0.0036
Q8XPU1	Putative...	Ralstonia...	0.0036	-
Q8V2D0	Fiber protein	Human adenovirus...	0.0039	0.0016

OMPB_RICJA	Outer membrane...	Rickettsia japonica	0.004	-
Q9KKB1	OmpB [Fragment]	Rickettsia japonica	0.004	-
Q99PM6	Cell adhesion...	Mus musculus (Mouse)	3.30E-05	0.0045
Q924G8	Trophinin	Mus musculus (Mouse)	3.30E-05	0.0045
Q9WUN1	Trophinin	Mus musculus (Mouse)	3.30E-05	0.0045
Q8R564	Trophinin	Mus musculus (Mouse)	3.30E-05	0.0045
Q99PB3	Mage-d3	Mus musculus (Mouse)	3.30E-05	0.0045
Q80TJ5	MKIAA1114...	Mus musculus (Mouse)	2.20E-05	0.0045
Q85690	Viral...	Reovirus sp.	0.0076	0.00086
Q9KKB8	OmpB [Fragment]	Rickettsia africae	0.0077	-
Q8XPU7	Probable...	Ralstonia...	0.0078	-
Q9WF20	Fiber	Human adenovirus...	0.004	0.0081
Q8BSK0	Melanoma...	Mus musculus (Mouse)	7.20E-05	0.01
Q9KKA9	OmpB [Fragment]	Rickettsia...	0.012	-
Q85695	Viral...	Reovirus sp.	0.013	0.003
Q8X4H5	Putative RTX...	Escherichia coli...	0.011	0.015
Q8X2T1	Hypothetical...	Escherichia coli...	0.011	0.015
Q9JP78	Adhesin	Bordetella...	0.0021	0.019
Q87ID8	Hypothetical...	Vibrio...	0.02	-
Q8WXI7	Ovarian cancer...	Homo sapiens (Human)	0.02	0.021
Q85694	Viral...	Reovirus sp.	0.023	0.0033
Q9E8F7	ORF5 [Fragment]	Porcine adenovirus 5	0.034	0.019
Q879S6	...	Xylella fastidiosa...	0.0019	0.037
Q9KKA2	OmpB [Fragment]	Rickettsia sibirica	0.037	-
Q9KKB6	OmpB [Fragment]	Astrakhan rickettsia	0.064	-
Q9KKA7	OmpB [Fragment]	Rickettsia parkeri	0.071	-
Q9KKB9	OmpB [Fragment]	Rickettsia...	0.088	-
Q85693	Viral...	Reovirus sp.	0.1	0.0016
Q8KQM9	Hemagglutinin	Moraxella...	0.1	0.054
P71401	Hsf protein	Haemophilus...	0.11	0.064

OMPB_RICPR	Outer membrane...	Rickettsia...	0.007	0.12
Q8XQZ5	Probable...	Ralstonia...	0.0054	0.12
Q85692	Viral...	Reovirus sp.	0.13	0.0022
Q8KQM8	Hemagglutinin	Moraxella...	0.13	0.11
Q9I5N6	Hypothetical...	Pseudomonas...	0.0011	0.13
Q98LN6	Hypothetical...	Rhizobium loti...	0.049	0.16
Q9ZKS9	Putative...	Helicobacter pylori...	0.18	-
Q48031	Adhesin	Haemophilus...	0.12	0.2
Q9KKA5	OmpB [Fragment]	Rickettsia...	0.2	-
Q89Q78	Blr3252 protein	Bradyrhizobium...	0.2	-
Q8FXA7	Outer membrane...	Brucella suis	0.2	-
Q9KKA6	OmpB [Fragment]	Rickettsia...	0.011	0.21
FHAB_BORPE	Filamentous...	Bordetella pertussis	0.0027	0.22
Q8VV99	FHA protein	Bordetella pertussis	0.0027	0.22
Q45365	Filamentous...	Bordetella pertussis	0.0027	0.22
Q85688	Viral...	Reovirus sp.	0.27	0.0063
Q8P9Q5	Filamentous...	Xanthomonas...	0.00057	0.33
Q8XUK0	Putative...	Ralstonia...	0.43	-
Q8FFF8	YapH homolog	Escherichia coli O6	0.43	-
Q9KKB4	OmpB [Fragment]	Rickettsia sp. Bar29	0.44	-
P94772	...	Erwinia chrysanthemi	0.0082	0.46
Q9F289	YapD protein	Yersinia pestis	0.46	-
VG37_BPT2	Long tail fiber...	Bacteriophage T2	0.51	-
OMPB_RICTY	Outer membrane...	Rickettsia typhi	0.58	-
Q9KKB0	OmpB [Fragment]	Rickettsia massiliae	0.62	-
Q8PLI3	Filamentous...	Xanthomonas...	0.65	-
Q89CB5	Bll7882 protein	Bradyrhizobium...	0.095	0.8
Q89C73	Bll7924 protein	Bradyrhizobium...	0.92	-
Q9F0P7	Outer membrane...	Rickettsia...	0.088	1.1
Q9F0P6	Outer membrane...	Rickettsia...	0.088	1.1

Q9I791	Probable...	Pseudomonas...	0.19	1.1
Q9XCJ4	ShdA	Salmonella...	1.3	-
Q8ZN57	Similar to the...	Salmonella...	1.3	-
Q89MY6	Blr4056 protein	Bradyrhizobium...	1.8	-
VSI1_REOVL	Sigma 1 protein...	Reovirus (type 1 /...	1.8	-
Q8V5E3	Cell attachment...	Ndelle virus	2	0.14
Q45364	Hemagglutinin...	Bordetella pertussis	2	-
Q93QW9	OmpB	Rickettsia felis...	2.1	-
Q8YB31	Adhesin AIDA-I	Brucella melitensis	2.5	-
Q98JH8	Serine...	Rhizobium loti...	2.7	1
OMPB_RICRI	Outer membrane...	Rickettsia...	0.052	2.7
Q9F0P9	Outer membrane...	Rickettsia sp....	2.8	-
Q9RNI2	HmwA	Haemophilus...	0.39	3.8
Q8ZBY3	Putative...	Yersinia pestis	3.9	-
Q8GF46	Hypothetical...	Zymomonas mobilis	0.00012	4
OMPB_RICCN	Outer membrane...	Rickettsia conorii	5.1	-
Q91CJ7	Fiber protein...	Human adenovirus...	5.2	1.3
Q9KKA0	OmpB [Fragment]	Rickettsia honei	0.36	5.7
Q9KKB3	OmpB [Fragment]	Rickettsia honei	0.36	5.7
Q8XYI3	Probable...	Ralstonia...	0.00081	6.1
ELT1_CAEEL	Transcription...	Caenorhabditis...	6.3	-
Q84X82	Flagella...	Chlamydomonas...	-	6.4
Q9PEY9	...	Xylella fastidiosa	0.21	6.8
Q8FUS1	Outer membrane...	Brucella suis	0.003	7.9
O25579	Toxin-like...	Helicobacter pylori...	0.056	7.9
Q8XQ42	Putative...	Ralstonia...	8.3	-
Q9KKA8	OmpB [Fragment]	Rickettsia montana	8.6	-
Q8ZRF8	Flagellar...	Salmonella...	0.96	8.9
Q8FKQ3	Putative member...	Escherichia coli O6	0.15	9.8
Q82WF7	Hypothetical...	Nitrosomonas...	9.8	-

Appendix B

SAS Results

B.1 Adenovirus Model

The LOGISTIC Procedure

Model Information

Data Set	WORK.AD
Response Variable	tbs
Number of Response Levels	2
Number of Observations	804
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	tbs	Total Frequency

1	1	241
2	0	563

Probability modeled is tbs=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	983.931	41.218
SC	988.620	59.976
-2 Log L	981.931	33.218

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	948.7129	3	<.0001
Score	625.3452	3	<.0001
Wald	19.9314	3	0.0002

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.1284	1.6946	3.4082	0.0649
betadb	1	-3.8234	2.5342	2.2762	0.1314
packing	1	1.6166	0.8950	3.2629	0.0709
profile	1	9.9866	2.7050	13.6296	0.0002

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
betadb	0.022	<0.001	3.138
packing	5.036	0.872	29.098
profile	>999.999	108.285	>999.999

Association of Predicted Probabilities and Observed Responses

Percent Concordant	99.9	Somers' D	0.997
Percent Discordant	0.1	Gamma	0.997
Percent Tied	0.0	Tau-a	0.419
Pairs	135683	c	0.999

B.2 Reovirus Model

The LOGISTIC Procedure

Model Information

Data Set	WORK.RE
Response Variable	tbs
Number of Response Levels	2
Number of Observations	574
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	tbs	Total Frequency
1	1	11
2	0	563

Probability modeled is tbs=1.

Model Convergence Status

Complete separation of data points detected.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	110.792	8.205
SC	115.145	25.615
-2 Log L	108.792	0.205

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	108.5872	3	<.0001
Score	205.8036	3	<.0001
Wald	2.2213	3	0.5278

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-11.6644	14.2366	0.6713	0.4126
betadb	1	-3.5336	20.3906	0.0300	0.8624
packing	1	0.2319	4.7293	0.0024	0.9609
profile	1	18.5516	13.4706	1.8967	0.1685

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
betadb	0.029	<0.001	>999.999
packing	1.261	<0.001	>999.999
profile	>999.999	<0.001	>999.999

Association of Predicted Probabilities and Observed Responses

Percent Concordant	100.0	Somers' D	1.000
Percent Discordant	0.0	Gamma	1.000
Percent Tied	0.0	Tau-a	0.038
Pairs	6193	c	1.000

Bibliography

[1] M. M. Altamirano, J. M. Blackburn, C. Aguayo, and A. R. Fersht. Directed evolution of new catalytic activity using the α/β-barrel scaffold. *Nature*, 403:617–621, 2000.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[3] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleid Acids Research*, 25(17):3389–3402, 1997.

[4] Z. Bao and S. R. Eddy. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research*, 12:1269–1276, 2002.

[5] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer. The Pfam protein families database. *Nucleid Acids Research*, 30(1):276–280, 2002.

[6] B. Berger and M. Singh. An iterative method for improved protein structural motif recognition. *Journal of Computational Biology*, 4(3):261–273, 1997.

[7] B. Berger, D. B. Wilson, E. Wolf, T. Tonchev, M. Milla, and P. S. Kim. Predicted coiled coils by use of pairwise residue correlations. *Proceedings of the National Academy of Sciences USA*, 92:8259–8263, 1995.

[8] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Wessig, I. S. dya lov, and P. Bourne. The protein data bank. *Nucleid Acids Research*, 28:235–242, 2000.

[9] B. Boeckmann, A. Bairoch, R. Apweiler, M. Blatter, A. Estreicher, E. Gasteiger, M. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The Swiss-Prot protein knowledgebase and its supplement TrEMBL. *Nucleid Acids Research*, 31:365–370, 2003.

[10] P. Bradley, L. Cowen, M. Menke, J. King, and B. Berger. Betawrap: Successful prediction of parallel β-helices from primary sequence reveals an association with many microbial pathogens. *Proceedings of the National Academy of Sciences USA*, 98(26):14819–14824, 2001.

[11] P. Bradley, P. S. Kim, and B. Berger. Trilogy: Discovery of sequence-structure patterns across diverse proteins. *Proceedings of the National Academy of Sciences USA*, 99(13):8500–8505, 2002.

[12] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, second edition, 1999.

[13] S. E. Brenner, P. Koehl, and M. Levitt. The ASTRAL compendium for protein structure and sequence analysis. *Nucleid Acids Research*, 28(1):254–256, 2000.

[14] C. Bystroff and D. Baker. Blind predictions of local protein structure in CASP2 targets using the I-sites library. *PROTEINS: Structure, Function, and Genetics*, Supplement 1:167–171, 1997.

[15] C. Bystroff and D. Baker. Prediction of local structure in proteins using a library of sequence-structure motifs. *Journal of Molecular Biology*, 281:565–577, 1998.

[16] J.-M. Chandonia, N. S. Walker, L. L. Conte, P. Koehl, M. Levitt, and S. E. Brenner. ASTRAL compendium enhancements. *Nucleid Acids Research*, 30(1):260–263, 2002.

[17] J. D. Chappell, A. E. Prota, T. S. Dermody, and T. Stehle. Crystal structure of the reovirus attachment protein $\sigma 1$ reveals evolutionary relationship to adenovirus fiber. *The EMBO Journal*, 21(1 and 2):1–11, 2002.

[18] D. Chivian, T. Robertson, R. Bonneau, and D. Baker. Ab initio methods. *Methods of Biochemical Analysis*, 44:547–557, 2003.

[19] J. Chrobozcek, R. Ruigrok, and S. Cusack. Adenovirus fiber. *Current Topics in Microbiological Immunology*, 199:163–200, 1995.

[20] E. Coward and F. Drabløs. Detecting periodic patterns in biological sequences. *Bioinformatics*, 14(6):498–507, 1998.

[21] C. Dodge, R. Schneider, and C. Sander. Th HSSP database of protein structure-sequence alignments and family profiles. *Nucleid Acids Research*, 26(1):313–315, 1998.

[22] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 2000.

[23] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[24] P. Edman. Title. *Acta Chem Scand*, 4:283–299, 1950.

[25] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. Sigrist, K. Hoffman, and A. Bairoch. The PROSITE database, its status in 2002. *Nucleid Acids Research*, 30:235–238, 2002.

[26] N. Green, N. Wrigley, W. Russell, S. Martin, and A. McLachlan. Evidence for a repeating cross-β sheet structure in the adenovirus fibre. *The EMBO Journal*, 2(8):1357–1365, 1983.

[27] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Publishing, 2001.

[28] A. Heger and L. Holm. Rapid automatic detection and alignment of repeats in protein sequences. *PROTEINS: Structure, Function, and Genetics*, 41:224–237, 2000.

[29] I. Henderson, F. Navarro-Garcia, and J. Nataro. The great escape: Structure and function of the autotransporter proteins. *Trends in Microbiology*, 6(9):370–378, 1998.

[30] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences USA*, 89:10915–10919, 1992.

[31] S. Henikoff and J. G. Henikoff. Position-based sequence weights. *Journal of Molecular Biology*, 243:574–578, 1994.

[32] S. Henikoff and J. G. Henikoff. Embedding strategies for effective use of information from multiple sequence alignments. *Protein Science*, 6:698–705, 1997.

[33] K. Höfling, S. Tracy, N. Chapman, K.-S. Kim, and J. S. Lesler. Expression of an antigenic adenovirus epitope in a group b coxsackievirus. *Journal of Virology*, 74(10):4570–4578, 2000.

[34] T. J. Hubbard and J. Park. Fold recognition and ab initio structure predictions using hidden markov models and β-strand pair potentials. *PROTEINS: Structure, Function, and Genetics*, 402:3398–402, 1995.

[35] D. T. Jones. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology*, 287:797–815, 1999.

[36] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292:195–202, 1999.

[37] D. T. Jones and M. B. Swindells. Getting the most from PSI-BLAST. *TRENDS in Biochemical Sciences*, 27(3):161–164, 2002.

[38] D. T. Jones, W. Taylor, and J. M. Thornton. A new approach to protein fold recognition. *Nature*, 358:86–89, 1992.

[39] A. V. Kajava, N. Cheng, R. Cleaver, M. Kessel, M. N. Simon, E. Willery, F. Jacob-Dubuisson, C. Locht, and A. C. Steven. Beta-helix model for the filamentous haemagglutinin adhesin of bordetella pertussis and related bacterial secretory proteins. *Molecular Microbiology*, 42(2):279–292, 2001.

[40] S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences USA*, 87:2264–2268, 1990.

[41] K. Karplus, C. Barrett, and R. Hughey. Hidden markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856, 1998.

[42] J. C. Kendrew, G. Bodo, H. Dintzis, R. Parrish, H. Wyckoff, and D. Phillips. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, 181:662–666, 1958.

[43] E. S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[44] R. Lathrop. The protein threading problem with sequence amino acid interaction preference is NP-complete. *Protein Engineering*, 7:1059–1068, 1994.

[45] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: A gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

[46] S. Lifson and C. Sander. Antiparallel and parallel β-strands differ in amino acid residue preferences. *Nature*, 282:109–111, 1979.

[47] S. Lifson and C. Sander. Specific recognition in the tertiary structure of β-sheets of proteins. *Journal of Molecular Biology*, 139:627–639, 1980.

[48] L. A. Mirny and E. I. Shakhnovich. Protein structure prediction by threading. why it works and why it does not. *Journal of Molecular Biology*, 283:507–526, 1998.

[49] L. A. Mirny and E. I. Shakhnovich. Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics, and function. *Journal of Molecular Biology*, 291:177–196, 1999.

[50] A. Mitraki, S. Miller, and M. J. van Raaij. Review: Conformation and folding of novel beta-structural elements in viral fiber proteins: The triple beta-spiral and the triple beta-helix. *Journal of Structural Biology*, 137:236–247, 2002.

[51] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

[52] S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[53] L. Pauling and R. B. Corey. The structure of proteins: Two hydrogen bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences USA*, 37:205–211, 1951.

[54] L. Pauling, R. B. Corey, and H. R. Branson. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. *Proceedings of the National Academy of Sciences USA*, 37:729–741, 1951.

[55] B. Rost. Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2):85–94, 1999.

[56] I. Ruczinski, C. Kooperberg, R. Bonneau, and D. Baker. Distributions of beta sheets in proteins with applications to structure prediction. *PROTEINS: Structure, Function, and Genetics*, 48:85–97, 2002.

[57] A. A. Schäffer, L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, and S. F. Altschul. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleid Acids Research*, 29(14):2994–3005, 2001.

[58] M. Singh, B. Berger, P. S. Kim, J. M. Berger, and A. G. Cochran. Computational learning reveals coiled coil-like motifs in histidine kinase linker domains. *Proceedings of the National Academy of Sciences USA*, 95:2738–2743, 1998.

[59] M. J. Sippl and H. Flöckner. Threading thrills and threats. *Structure with Folding and Design*, 4:15–19, 1996.

[60] P. F. Stouten, C. Sander, R. W. Ruigrok, and S. Cusack. New triple-helical model for the shaft of the adenovirus fibre. *Journal of Molecular Biology*, 226:1073–1084, 1992.

[61] J. Thompson, D. Higgins, and T. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleid Acids Research*, 22:4673–4680, 1994.

[62] A. Šali, E. Shakhnovich, and M. Karplus. Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. *Journal of Molecular Biology*, 235:1614–1636, 1994.

[63] M. J. van Raaij, A. Mitraki, G. Lavigne, and S. Cusack. A triple β-spiral in the adenovirus fibre shaft reveals a new structural motif for a fibrous protein. *Nature*, 401:935–938, 1999.

[64] E. Veiga, E. Sugawara, H. Nikaido, V. de Lorenzo, and L. Fernandez. Export of autotransported proteins proceeds through and oligomeric ring shaped by c-terminal domains. *The EMBO Journal*, 21(9):2122–2131, 2002.

[65] P. R. Weigele, E. Scanlon, and J. King. Homotrimeric, β-stranded viral adhesins and tail proteins. *Journal of Bacteriology*, 185(14):4022–4030, 2003.

[66] H. Zhu and W. Braun. Sequence specificity, statistical potentials and three-dimensional structure prediction with self-correcting distance geometry calculations of β sheet formation in proteins. *Protein Science*, 8:326–342, 1999.