2008-08-21

Biologically Relevant Multiple Sequence Alignment

Hyrum D. Carroll
Brigham Young University - Provo

BIOLOGICALLY RELEVANT MULTIPLE SEQUENCE ALIGNMENT

by

Hyrum D. Carroll

A dissertation submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

Brigham Young University

December 2008

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a dissertation submitted by

Hyrum D. Carroll

This dissertation has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	Mark J. Clement, Chair
Date	Quinn O. Snell
Date	David A. McClellan
Date	Kevin D. Seppi
Date	Daniel Zappala

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the dissertation of Hyrum D. Carroll in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date	Mark J. Clement Chair, Graduate Committee

Accepted for the Department

Date	Kent E. Seamons Graduate Coordinator

Accepted for the College

Date	Thomas W. Sederberg Associate Dean, College of Physical and Mathematical Sciences

ABSTRACT

BIOLOGICALLY RELEVANT MULTIPLE SEQUENCE ALIGNMENT

Hyrum D. Carroll

Department of Computer Science

Doctor of Philosophy

Researchers use multiple sequence alignment algorithms to detect conserved regions in genetic sequences and to identify drug docking sites for drug development. In this dissertation, a novel algorithm is presented for using physicochemical properties to increase the accuracy of multiple sequence alignments. Secondary structures are also incorporated in the evaluation function. Additionally, the location of the secondary structures is assimilated into the function. Multiple properties are combined with weights, determined from prediction accuracies of protein secondary structures using artificial neural networks.

A new metric, the PPD Score is developed, that captures the average change in physicochemical properties. Using the physicochemical properties and the secondary structures for multiple sequence alignment results in alignments that are more accurate, biologically relevant and useful for drug development and other medical uses.

In addition to a novel multiple sequence alignment algorithm, we also propose a new protein-coding DNA reference alignment database. This database is a collection of multiple sequence alignment data sets derived from tertiary structural alignments. The primary purpose of the database is to benchmark new and existing multiple sequence alignment algorithms with DNA data. The first known comparative study of protein-coding DNA alignment accuracies is also included in this work.

ACKNOWLEDGMENTS

I would like to first and foremost thank the two most important people in my life, my wife Melissa and my Heavenly Father. They have provided encouragement, faith in me and positive attitudes throughout the entire project.

I would also like to thank my father for his interest, insightful questions and love.

My advisor, Dr. Mark J. Clement and "pseudo-advisor", Dr. Quinn O. Snell, have been supportive, interested and flexible. Additionally, Dr. David A. McClellan has always provided encouraging discussions.

Finally, I am grateful for the stimulating conversations about doctoral degrees with the late Dr. James Edwin Dalley (1922–2008), my maternal grandfather, to whom this work is dedicated.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The central dogma of biology hypothesizes that DNA is converted to RNA, which is processed by the ribosome to create proteins that interact to create the physical features of an organism (see Figure 1.1). Changes happen randomly throughout the DNA of an organism. Mutations that occur in unimportant regions remain, but changes to the parts of DNA that create the active regions of a protein can cause the organism to die and keep the mutations from being passed on to descendants of the organism. When researchers find an area of DNA that is very similar (or conserved) for distantly related organisms, then there is a reason to believe that this region is important to the survival of the organism, or it would have had random mutations in that region.

Multiple sequence alignments are useful for many areas in Bioinformatics. They are used to predict the functional segments of a sequence (genes) and the areas of a protein under selective pressures (Woolley *et al.*, 2003). MSAs are also the primary input for reconstructing phylogenetic trees, or phylogenies. Phylogenies hierarchically relate evolutionary events and are used in diverse areas of research (e.g., epidemiology (Clark *et al.*, 1998; Sing *et al.*, 1992), viral transmission (Crandall, 1996; Herring *et al.*, 2007), biogeography (DeSalle, 1995) and evolutionary studies (Whiting *et al.*, 2003)). For each of these areas, the alignment is used as a foundation and the accuracy of further analysis is directly correlated with the quality of the alignment.

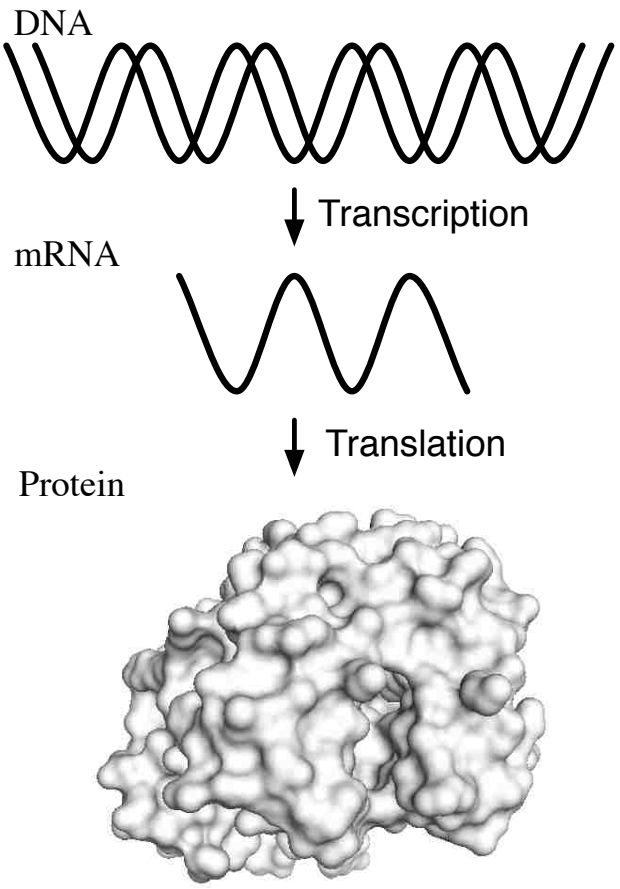


Figure 1.1: The Central Dogma of biology states that DNA is transcribed into RNA, which is translated in proteins.

Researchers also use MSAs to predict the location of a drug docking site. The conserved columns (or regions) identify locations on the sequences that have the least amount of change. These areas are projected onto the tertiary structure[1] of a protein to predict potential drug docking sites (see Figure 1.2). Biologists use this information to develop new drugs to inhibit the protein from interacting within a biological pathway. This process has lead to drugs to treat glaucoma, inhibit COX-2 and a treatment for HIV-1.

Many Bioinformatics studies begin use a multiple sequence alignment (MSA) as the foundation for their research. MSAs are a set of genetic sequences and their

[1]The tertiary structure of a protein is the three dimensional position of each of its atoms in space

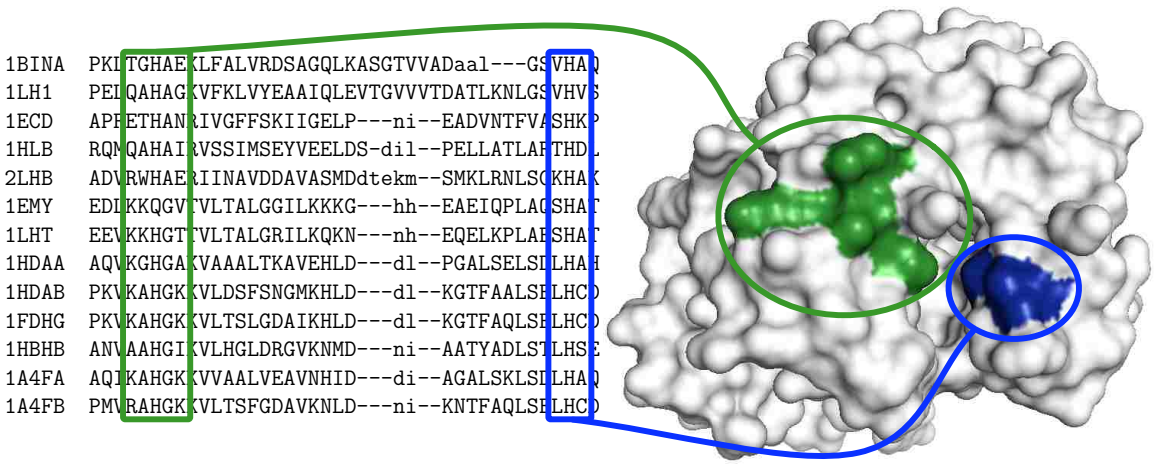


Figure 1.2: Example of using an alignment to identify potential drug docking sites. The first column lists the PDB ID of each of the sequences. The protein shown is hemoglobin (PDB ID 1A4FA). The most conserved columns are highlighted on both the ChemAlign alignment and the protein. The regions are at a possible drug docking site. ChemAlign is able to find both regions, whereas other MSA algorithms are only able to find the one of the left.

evolutionary relationship with each other. Existing MSA algorithms have the following three main deficiencies:

1. Optimization for sequence similarity
2. Ignoring secondary structure information
3. Static comparison of sequences

The first limitation of current MSA algorithms is that their optimization criteria focuses exclusively on sequence similarity. Although algorithms for calculating alignments that minimize changes in genetic characters are easier to develop and are commonplace, biologically accurate alignments minimize the change in physicochemical properties of the amino acids (e.g., hydropathy, polarity and volume) (see section 1.4). Second, most MSA algorithms ignore additional contextual information, such as protein secondary structures (α-helices, β-strands and loops) (see section 1.5). Secondary structure has long been understood to be more conserved than the primary amino acid sequence. This has been verified through a number of different experiments and reports (Gibrat *et al.*, 1996; Rost, 1999; Sander and Schneider, 1991). This more

resilient information can provide a unique component to the sequence similarity criteria. The third weakness of most current MSA approaches is that they treat every position in the sequence as the "average" position. This shortcoming is eloquently stated by Thorne *et al.* (1996):

> A problem with the Dayhoff approach is that it effectively models the replacement process at the "average" site in the "average" protein. There may be no such thing as an "average" site in an "average" protein.

Although explicitly citing the Dayhoff *et al.* (1978) approach, this limitation generalizes to the vast majority of MSA algorithms. Biologically meaningful relationships between sequences depend on the location of the amino acid in the protein. These three drawbacks lead current MSA algorithms to produce inferior alignments.

1.1 Thesis Statement

A multiple sequence alignment algorithm that optimizes for different physicochemical properties in each secondary structure can create alignments with better scores, that are more biologically relevant. A new MSA algorithm, ChemAlign, addresses the three main limitations of existing methods, producing accurate alignments that identify more biologically relevant features. First, it incorporates physicochemical properties of the amino acids (e.g., hydropathy, polarity and volume). It uses these properties as an integral part of the optimization criteria. Evaluating similarity based on these properties incorporates more information and models the criteria that nature uses. Second, ChemAlign explicitly combines secondary structure elements into the evaluation function. Incorporating this additional information aligns the secondary structures, which are typically more conserved than the amino acids themselves. Third, ChemAlign adjusts its evaluation function by calculating the relationship between the amino acids differently, based on their secondary structure. This increases specificity

of the function and provides a dynamic comparison of the sequences. ChemAlign integrates these three pieces of information to produce biologically accurate multiple sequence alignments.

Data sets with very low percent identity are particularly difficult for current MSA methods. These data sets are one of the best sources for finding drug docking sites since they contain distantly related species and therefore conserved columns are more obvious. The globin family is a good example of this. Due to its low average percent identity of 25.9%, the globin family remains difficult for existing methods to accurately align. Current algorithms align at most 38.4% of the positions correctly. Using a physicochemical property, ChemAlign correctly aligns 90.6% of the positions. Figure 1.2 shows part of a ChemAlign alignment of the globin domains and a hemoglobin protein. Conserved columns are marked on the alignment and the protein, and appear at a possible drug docking site. ChemAlign is able to find both regions, whereas other algorithms do not.

1.2 Pairwise Sequence Alignment

Initially, alignment algorithms focused on aligning two genetic sequences. Let g_1 and g_2 be genetic sequences of length l_1 and l_2, defined over the alphabet $\sum_{DNA} = \{\texttt{a},\texttt{c},\texttt{g},\texttt{t}\}$ (nucleotides) for DNA and for amino acids $\sum_{AA} = \{\texttt{A},\texttt{C},\texttt{D},\texttt{E},\texttt{F},\texttt{G},\texttt{H},\texttt{I},\texttt{K},\texttt{L},\texttt{M},\texttt{N},\texttt{P},\texttt{Q},\texttt{R},\texttt{S},\texttt{T},\texttt{V},\texttt{W},\texttt{Y}\}$. A pairwise alignment of g_1 and g_2 is the set of sequences g_1' and g_2' defined over the alphabet $\sum_{DNA} \cup \{\text{-}\}$ or $\sum_{AA} \cup \{\text{-}\}$. The character '-', a *gap*, represents an insertion or deletion (indel) caused by mutations. Furthermore, the length of g_1' and g_2' is m, where $m \geq l_1$ and $m \geq l_2$. After accurately inserting gaps, columns with a high-degree of similarity indicate functional importance for that part of the protein.

Needleman and Wunsch (1970) developed the classic dynamic programming algorithm that calculates the optimal alignment for a pair of sequences. Their al-

		K	E	D	L	K	K	Q	G	V
	0.0	-10.0	-10.2	-10.4	-10.6	-10.8	-11.0	-11.2	-11.4	-11.6
K	-10.0	**5.0**	**-5.0**	**-5.2**	**-5.4**	**-5.6**	-5.8	-10.0	-13.2	-13.4
A	-10.2	-5.0	4.0	-6.0	-6.2	-6.4	**-6.6**	-6.8	-10.0	-13.2
H	-10.4	-5.2	-5.0	3.0	-7.0	-7.2	-7.4	**-6.6**	-8.8	-13.0
G	-10.6	-5.4	-7.2	-6.0	-1.0	-9.0	-9.2	-9.4	**-0.6**	-10.6
K	-10.8	-5.6	-4.4	-8.2	-8.0	4.0	-4.0	-8.2	-10.6	**-2.6**

(a)

```
KEDLKKQGV
K----AHGK
```

(b)

Figure 1.3: (a) Example Needleman-Wunsch matrix. The evaulation critera for (mis)matches is the BLOSUM62 matrix. The gap open penalty is -10.0 and the gap extension penalty is -0.2. The optimal traversal is highlighted. (b) Example pairwise alignment corresponding with an optimal traversal.

gorithm determines, for every position of every possible combination of gaps, the maximum score between 1) inserting a gap in the first sequence 2) inserting a gap into the second sequence and 3) aligning the two characters. Figure 1.3(a) illustrates a completed Needleman-Wunsch matrix for the sequences `KAHGK` and `KEDLKKQGV`. The evaluation criteria for (mis)matches is the BLOSUM62 (Henikoff and Henikoff, 1992) matrix. The gap open penalty is -10.0. Using these parameters yields the alignment shown in Figure 1.3(b).

1.3 Multiple Sequence Alignment

The natural extension of pairwise alignment algorithms are multiple sequence alignment algorithms. Let S be a set of genetic sequences $g_1, \ldots g_n$ of lengths $l_1, \ldots l_n$ defined over the alphabet $\sum_{DNA}$ or $\sum_{AA}$. A multiple sequence alignment of S is formally defined as a set S' of sequences $g'_1, \ldots g'_n$ all of length m, defined over the alphabet $\sum_{DNA} \cup \{\text{-}\}$ or $\sum_{AA} \cup \{\text{-}\}$. Finally, $\forall_i (m \geq l_i)$.

MSA algorithms exist that return optimal alignments using an n-dimensional extension of the Needleman-Wunsch method (Kececioglu and Starrett, 2004; Lipman *et al.*, 1989). However, they are limited to all but the smallest data sets since the problem is NP-Complete (Kececioglu and Starrett, 2004; Wang and Jiang, 1994). Therefore, heuristic methods are used to calculate alignments.

1.3.1 Metrics

Central to algorithmic development of MSA algorithms are evaluation metrics. This section describes the most commonly employed metrics used to evaluate MSAs. They are presented in part to aid in explaining the MSA algorithms themselves in the following chapter.

Self Sum Of Pairs

One of the earliest metrics of MSA is the *self sum of pairs* (Carrillo and Lipman, 1988). The self sum of pairs score for an alignment is the percentage of pairs of characters that match:

$$\text{self sum of pairs} = \frac{2}{n^2 - 1} \sum_{i}^{n} \sum_{j \neq i}^{n} \sum_{k}^{m} \delta(g_i(k), g_j(k)) \tag{1.1}$$

$$\delta(x, y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases} \tag{1.2}$$

Unfortunately, such a simple scoring metric does not necessarily reflect biological accuracy.

Reference Sum Of Pairs Score

Recently, researchers created several reference amino acid databases (Edgar, 2004b; Letunic *et al.*, 2004; Mizuguchi *et al.*, 1998; Raghava *et al.*, 2003; Subramanian *et al.*,

2005; Thompson *et al.*, 2005; Van Walle *et al.*, 2005). Most of these databases leverage secondary and tertiary structural alignments to provide a suite of "gold standard" alignments. Calculated alignments are evaluated by comparing against them. They have been well accepted by the scientific community and used in numerous studies to compare the quality of amino acid alignments generated by MSA algorithms (Do *et al.*, 2005; Edgar, 2004a,b; Karplus and Hu, 2001; Lassmann and Sonnhammer, 2002, 2005a; Thompson *et al.*, 1999b; Van Walle, 2004). These amino acid alignment benchmarks are limited to the evaluation of amino acid alignment algorithms.

One of the most commonly applied metrics for multiple sequence alignment algorithms is the *reference sum of pairs score*. It is calculated in a similar manner to the self sum of pairs score, except that each position of a calculated alignment is compared to the corresponding position in a reference alignment. Let $r_1, \ldots r_n$ be a sequences of a reference alignment, each of length p. Let $q = min(m, p)$.

$$\text{reference sum of pairs} = \frac{1}{nq} \sum_{i}^{n} \sum_{k}^{q} \delta(g_i(k), r_i(k)) \tag{1.3}$$

This metric is generally preferred to the self sum of pairs score since it evaluates how close an alignment is to the "gold standard" alignment.

Column Score

Another often used metric is the *column score* (Karplus and Hu, 2001). The column score is more conservative than the reference sum of pairs score in that it is the percentage of columns of a calculated alignment that completely match a reference alignment. Let $g_i(k)$ be the k^{th} genetic character of the i^{th} sequence:

$$\text{column score} = \frac{1}{q} \sum_{k}^{q} \sigma(k)$$

$$\sigma(k) = \begin{cases} 1 & \forall_t \left(g_t(k) = r_t(k)\right) \\ 0 & \text{otherwise} \end{cases}$$

Physicochemical Properties Difference Score

In addition to existing MSA metrics, the Physicochemical Properties Difference (PPD) score is presented in this dissertation. The score is calculated as follows for a single physicochemical property p:

$$\text{PPD score} = \frac{1}{nq} \sum_{i}^{n} \sum_{k}^{q} D^{p}_{s_i(k), r_i(k)} \tag{1.4}$$

$$D^{p}_{i,j} = 1 - \frac{2 * |p[i] - p[j]|}{argmax_x(p[x]) - argmin_y(p[y])} \tag{1.5}$$

Here, $p[i]$ is the value of p for amino acid i. D^p is the normalized difference matrix of p. The values of D^p range from -1.0 for the most dissimilar pair of amino acids to 1.0 for identical amino acids. PPD scores range from -1.0 to 1.0. In general, a negative PPD score means that the average amino acid pairing in an alignment is worse than the average difference in the physicochemical property values. A score of 1.0 means the calculated alignment is the same as the reference alignment. This score takes a step beyond sequence similarity and measures characteristics of the amino acids to provide a more biologically relevant metric. It can be adapted to account for multiple physicochemical properties by incorporating multiple D^p matrices into a single matrix with weights.

1.4 Physicochemical Properties

Several researchers are using the structural and biochemical characteristics of the 20 amino acids (Goldman and Yang, 1994; McClellan *et al.*, 2005; Xia and Li, 1998).

BLOSUM62		Hydropathy	
seq1:	-Y	seq1:	Y-
seq2:	FG	seq2:	FG
(a)		(b)	

Figure 1.4: Example amino acid alignments using BLOSUM62 (a) and hydropathy (b) for the evaluation function. The hydropathy alignment detects that tyrosine (Y) and phenylalanine (F) are more similar than Y and glycine (G) and therefore should be aligned together.

These physicochemical properties, such as hydropathy (Kyte and Doolittle, 1982), polarity (Grantham, 1974) and volume (Bigelow, 1967), better represent the molecular forces impacting the system. The genetic code seems to have evolved to minimize differences in physicochemical properties (Xia and Li, 1998) and consequently, researchers have been quantifying properties for amino acids. Repositories, such as AAindex: Amino Acid Index Database (Kawashima *et al.*, 1999, 2008; Tomii and Kanehisa, 1996), catalog such properties.

The value of using physicochemical properties for multiple sequence alignment is illustrated in the following example. Consider the alignment of a single tyrosine (Y) with either a phenylalanine (F) or glycine (G). Using the BLOSUM62 substitution matrix for the evaluation function yields the alignment shown in Figure 1.4(a). The evaluation function returns a cost of -1.0 for changing from a Y to a F and 3.0 for Y to G (higher values denotes more similar). On the other hand, using hydropathy for the function detects that Y and F are more similar than Y and G and should be aligned together (see Figure 1.4(b)). The hydropathy alignment is therefore preferred to the BLOSUM62 alignment, especially for segments of the sequences that are known to conserve this property (e.g., on the exterior of a protein). Although this example deals with only a few residues, similar evaluations are often made thousands of times to calculate an alignment. At other locations in the protein, another physicochemical property could be more important. A biologically accurate alignment algorithm weights the properties based on the their location in the structure.

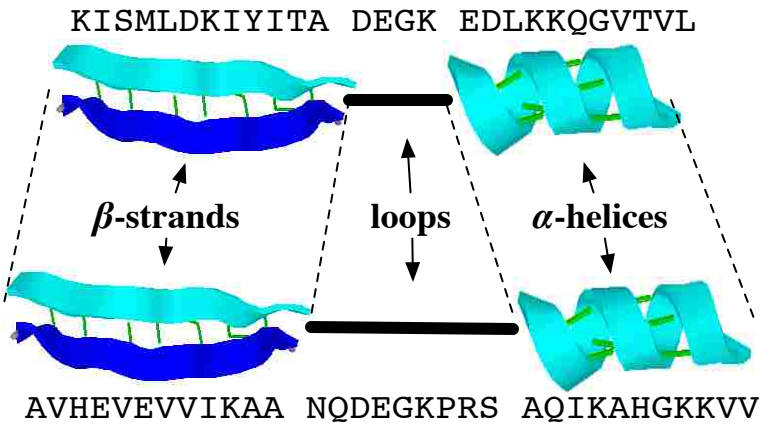


Figure 1.5: Amino acid alignment example of the three protein secondary structures elements. Each sequence has a β-strand, a loop (indicated with a solid thick line) and then an α-helix. Regions marked by the dotted lines should be aligned together.

1.5 Protein Secondary Structures Elements

Protein secondary structures elements (SSEs) are contiguous strings of α-helices, β-strands or loops. They are usually determined by the hydrogen bonds of the amino acids (the primary structure) using the DSSP definitions (Kabsch and Sander, 1983). Therefore, each amino acid has an accompanying SSE. Figure 1.5 illustrates the three SSEs and how they can be used to aid alignments of amino acids. In the figure, dotted lines indicate regions that should be aligned together by inserting gaps.

1.6 Contributions

The main contribution of this work is a new multiple sequence alignment algorithm, ChemAlign, that incorporates three novel pieces of information: physicochemical properties, protein secondary structures and the location of the secondary structures. This algorithm achieves accuracies higher than existing MSA algorithms for some of the most difficult reference alignments benchmarks. Furthermore, the alignments have been shown to be more biologically relevant.

Additionally, a novel MSA metric, the Physicochemical Properties Difference score is included in this work. This score measures the amount of similarity of one or more physicochemical properties in an alignment. It provides a more biologically accurate perspective than existing metrics.

Furthermore, this work also introduces a protein-coding DNA reference alignment database (Carroll *et al.*, 2007). This database is a collection of 3,545 MSA data sets derived mostly from tertiary structure alignments. Its primary purpose is to quantitatively benchmark the accuracy of several MSA algorithms using DNA data. The first known performance analysis of these DNA databases is included.

In summary, the main contributions of this work include the following:

- A novel algorithm that incorporates physicochemical properties to produce biologically relevant multiple sequence alignments
- Developing an evaluation function for multiple sequence alignments that includes secondary structures (both the structures themselves and their location)
- A biologically sensitive multiple sequence alignment metric, the Physicochemical Properties Difference score
- Reference protein-coding DNA multiple sequence alignment databases
- First known performance analysis of alignment accuracies for protein-coding DNA

1.7 Dissertation Outline

The remainder of the dissertation is as follows. Chapter 2 covers related work, detailing multiple sequence alignment algorithms and physicochemical properties. Chapters 3–5 are journal papers that are detailed below. Finally, concluding remarks are given in Chapter 6.

Chapter 3 is the journal paper, *DNA Reference Alignment Benchmarks Based on Tertiary Structure of Encoded Proteins*. Oxford University Press published it on August 8, 2007 in the 23^{rd} volume, 19^{th} issue of *Bioinformatics*. This journal enjoys an impact factor of 5.039. This work has already been cited by at least eight different papers (Agrawal and Huang, 2008; Hall, 2007, 2008a,b; Katoh and Toh, 2008; Sundberg *et al.*, 2007, 2008; Wilm *et al.*, 2008).

The first part of this paper introduces the reference protein-coding DNA alignment benchmarks. It briefly explains how they are derived from reference protein alignment databases. Statistics of the quality of the conversion are also given. These databases are extremely useful in evaluating the quality of DNA alignments generated by existing and forthcoming MSA techniques since there are no known equivalent benchmarks.

The second part of the paper is the published Supplementary Material and appears directly after the main paper. It is the first known performance comparison of alignment algorithms for both amino acids and DNA. Eight of the most common MSA algorithms are benchmarked and ranked according to their accuracy and execution time. The case study reveals two general points about the accuracy ranks. First, the amino acid benchmarks generally have higher accuracy scores than the DNA benchmarks. Second, and more importantly, the results show that certain algorithms that achieve high accuracy scores on amino acid sequences tend to have low ranks for DNA sequences. This is important new information for biologist using existing algorithms to align protein-coding DNA.

Chapter 4 is the journal paper, *ChemAlign: Biologically Relevant Multiple Sequence Alignment Using Physicochemical Properties*, submitted to *Bioinformatics*. This paper introduces ChemAlign and details how it incorporates a physicochemical property, secondary structures and their location to produce biologically accurate alignments. Additionally, an in-depth analysis of alignments of the globin domain

is presented, including using them for predicting drug docking sites. Moreover, the PPD score is introduced and included in analysis of ChemAlign.

Using a single physicochemical property, ChemAlign calculates alignments that are as high as 499.3% more accurate than other methods. Additionally, ChemAlign earns the highest PPD scores. These higher accuracies translate into more biologically correct alignments, as is shown with an example of identifying potential drug docking sites. The improvements in accuracies of ChemAlign over existing methods using these two metrics are statistically significant according to the Friedman rank test, with p-values $\ll 0.001$.

Chapter 5 is the journal paper, *Relative Importance of Physicochemical Properties of Amino Acids for Multiple Sequence Alignment*, submitted to *Nucleic Acids Research*. This paper details extending the evaluation function of ChemAlign to incorporate multiple physicochemical properties to increase accuracy of generated alignments. Several properties are combined using an exponentially decaying function. The weights for each property are based on the accuracies of artificial neural networks trained to predict protein secondary structures using that property. The specificity of the evaluation function is further increased by allowing gap penalties to be set for each of the different secondary structures.

The accuracies of the alignments are evaluated on thirteen of the largest reference amino acid data sets. The improved version of ChemAlign performs as well as 121.3% better on average across these data sets than other methods, and 15.8% better than the original ChemAlign. Additionally, ChemAlign achieves the highest average PPD score. It earns scores between a score as high as 105.3% better than the other methods, average across several reference data set. Again, the differences in these scores are statistically significant with a p-value $\ll 0.001$.

Chapter 2

Related Work

Researchers use multiple sequence alignment algorithms to detect conserved regions in genetic sequences, which are used to identify drug docking sites for drug development. While the base algorithms in the field have been known for decades, there has been a continually increasing interest in development of better algorithms. Initially, pairwise alignment algorithms were developed. A heuristic of one of these algorithms is one of the most widely used tools in Bioinformatics (Altschul *et al.*, 1990). More recently, several new multiple sequence alignment algorithms have been proposed. In this chapter, they are detailed and their accuracies are compared. Additionally, a review of research incorporating phylogenetic properties of amino acids is presented The algorithms detailed in this chapter are characterized according to major algorithmic classifications in Table 3.3. The table references local and global alignments. A *local alignment* of two sequences is the alignment of a contiguous segment of each of the sequences, where the length is shorter than the longest sequence. Smith and Waterman (1981) presented the classic local alignment variant of the Needleman-Wunsch method (see section 1.2). Their approach is different in that negative cell values are replaced with zeros and the highest scoring local alignment is chosen. Alternatively, *global alignments* include all of the characters from both sequences.

Table 2.1: Categorization of Sequence Alignment Algorithms

Algorithm	Pairwise/MSA	Progressive	Iterative	Global	Local
ClustalW	MSA	X		X	
DIALIGN	MSA		X		X
Kalign	MSA	X		X	
MAFFT-GINSI	MSA	X	X	X	
MAFFT-LINSI	MSA	X	X		X
MAFFT-NS1	MSA	X		X	
MAFFT-NSI	MSA	X	X	X	
MUSCLE	MSA	X	X	X	
T-Coffee	MSA	X		X	X
SAM[1]	MSA			X	
ProbCons[1]	MSA	X	X	X	
SAGA[2]	MSA		X	X	
POY[3]	MSA	X	X	X	
Gonnet and Lisacek	pairwise			X	
Gupta *et al.*	pairwise			X	
Lüthy *et al.*	pairwise			X	
PRALINE	MSA	X	X	X	
Jennings *et al.*	MSA	X		X	
HSA[4]	MSA		X	X	
PROMALS[1]	MSA	X		X	

[1]Uses a hidden Markov model, [2]Uses a genetic algorithm, [3]Employs Optimization Alignment, [4]Graph-based approach

2.1 Progressive Multiple Sequence Alignment

The majority of MSA algorithms can be classified into two areas: progressive and/or iterative (see Table 3.3). The *progressive multiple sequence alignment* method (PMSA) (Corpet *et al.*, 1988; Feng and Doolittle, 1987, 1990) is one of the most common heuristics to an n-dimensional Needleman-Wunsch. The algorithm has two main phases. First, a distance matrix is calculated from similarity scores for every pair of sequences. Often the Wilbur and Lipman algorithm (Wilbur and Lipman, 1984) is used to calculate the scores. These similarity scores are only very general approximations, but work as a starting point (Wilbur and Lipman, 1984). The similarity scores are hierarchically clustered together, usually with the UPGMA or the

Neighbor-Joining algorithm (Saitou and Nei, 1987), thereby producing a guide tree. The second phase consists of a recursive traversal of the guide tree, starting at the root node. The base case of the traversal is a node that only contains two leaf nodes. The sequences associated with those nodes have a higher sequence identity to each other than to any other sequence, and are therefore aligned first. During the post-order traversal phase of the recursion, an alignment of alignments is calculated until all sequences are included in the alignment.

2.1.1 ClustalW

The most commonly used implementation of the PMSA algorithm is ClustalW (Larkin *et al.*, 2007; Thompson *et al.*, 1994).

2.1.2 T-Coffee

T-Coffee (Notredame *et al.*, 2000) is another MSA algorithm that uses the progressive alignment approach with two distinguishing features. First, instead of ignoring the global pairwise alignments produced in the first phase, T-Coffee uses a library consisting of a combination of global and local pairwise alignments (see Figure 2.1) in its progressive alignment phase. By default, the library is populated initially by both global and local pairwise alignments (generated with ClustalW and Lalign (Pearson and Lipman, 1988) respectively), and a weight is assigned to each pair of aligned residues. The global and local alignments are merged into a primary library, giving the pairs that match in both alignments a greater weight and creating new entries for those pairs that do not match. T-Coffee extends the primary library by comparing triplets of aligned residues with every entry in the library. Starting with version 2.00 of T-Coffee, if the tertiary structure is known for one or more sequences, then an alignment generated using a 3D structural alignment algorithm (e.g., SAP (Taylor and Orengo, 1989b), DALI (Holm and Sander, 1993) or Fugue (Shi *et al.*, 2001)) can

Figure 2.1: Flow chart for T-Coffee. Note, square boxes are procedures and rounded boxes are data structures. Graphic from (Notredame *et al.*, 2000).

be incorporated into the library. The second distinguishing feature of T-Coffee is that it does not use gap penalties during the progressive phase. Instead, gap positions are determined by considering the weights in the library for all of the possible pairs of characters in the two sequences. Due to these features, T-Coffee has been shown to give high accuracy scores on the amino acid benchmarks. Unfortunately, this comes at a great cost in computational time, and alignments of large datasets with long sequences is very time consuming.

2.1.3 Kalign

The initial step of pairwise alignment in the progressive alignment strategy is the most computationally intensive. Many algorithms use the k-mer counting method to speed

up the process of finding the initial distance scores, but this method is less accurate. Kalign (Lassmann and Sonnhammer, 2005b) follows the progressive strategy but uses the Wu-Manber string-matching algorithm (Wu and Manber, 1992) to find the initial distance scores, which is faster than pairwise alignment and more accurate than k-mer counting. In the Wu-Manber algorithm, two sequences have a distance score equal to the number of mismatches or indels that can be applied to one sequence in order for it to match the other. Matches are found by searching three residues at a time along the sequences. These scores are used to produce the initial distance matrix that the guide tree is created from. Traditional progressive alignment proceeds and sequences are clustered according to the branch order of the guide tree. Kalign is one of the fastest MSA algorithms and shows comparable accuracy to MAFFT and MUSCLE on amino acid benchmarks (Carroll *et al.*, 2007). Like DIALIGN, Kalign is shown to be more accurate than many MSA methods on amino acid sequences with low sequence identity.

2.2 Iterative Refinement of Multiple Sequence Alignments

Iterative refinement of the MSA algorithm has been around for a number of years (Sankoff *et al.*, 1976). While several of the most recently proposed algorithms build upon a progressive approach with iterative refinement (see Table 3.3), DIALIGN just uses iterative refinement.

2.2.1 DIALIGN

The MSA algorithm DIALIGN (Morgenstern, 1999; Morgenstern *et al.*, 1998; Subramanian *et al.*, 2005) builds an alignment from pairwise local alignments (see Figure 2.2). Initially, all pairwise local alignments are calculated. The algorithm does not align segments of the sequences that are not statistically similar to other sequences in the alignment. Next, a greedy set of the best scoring consistent local alignments is

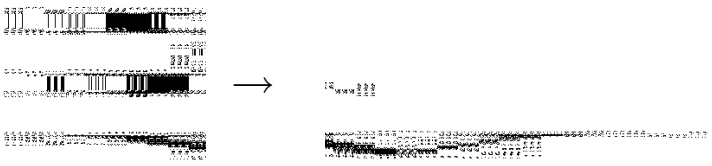


Figure 2.2: Example DIALIGN alignment using pairwise local alignments. Graphic from (Morgenstern *et al.*, 1998).

determined from the initial alignments. Iterations of these two steps continue until all local alignments are found. If the sequences only share local segments of similarity, the algorithm returns a local alignment with the unrelated segments untouched. In this way, DIALIGN can accurately align sequences with different degrees of similarity separated by unrelated sequences. Otherwise, DIALIGN finds local alignments that cover the entire length of the sequences and returns a global alignment. DIALIGN has been shown to be more accurate than T-Coffee in aligning amino acid sequences with low identity, but it is generally less accurate than T-Coffee in amino acid alignments of high sequence identity (Lassmann and Sonnhammer, 2002).

2.2.2 MUSCLE

One of the most noteworthy recent algorithms is MUSCLE (Edgar, 2004a,b). Edgar developed MUSCLE by first applying a progressive MSA phase and then an iterative phase. In Figure 2.3, phases 1.1 to 2.3 are the progressive portion and phases 3.1 to 3.4 are the iterative refinement part of the algorithm. During the iterations, an edge of the guide tree is removed, creating two trees. The alignments, or *profiles*, of these two trees are realigned. Iterations continue until the self sum of pairs score stops improving. MUSCLE typically produces a reasonable balance between speed and accuracy.

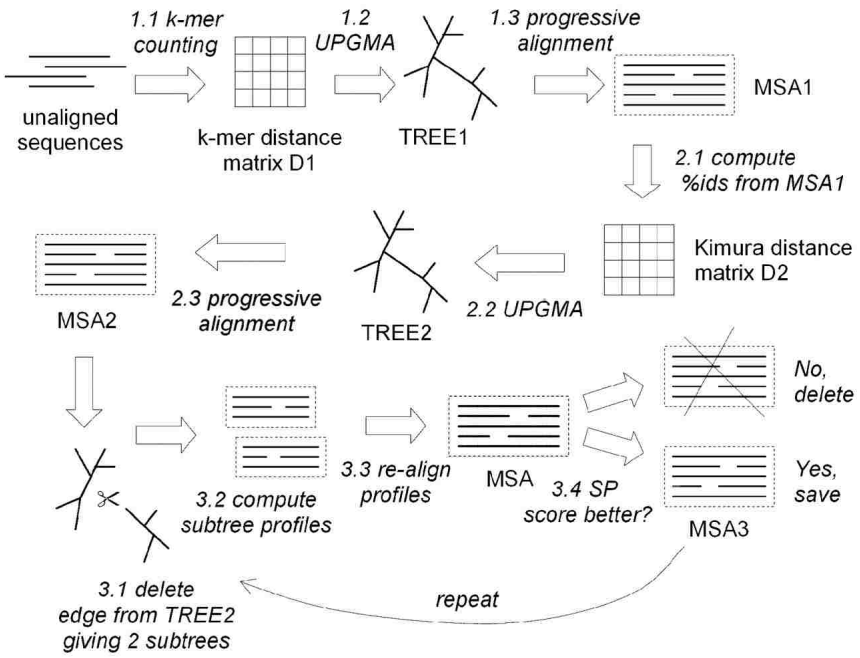


Figure 2.3: Flow chart for MUSCLE. Phases 1.1 to 2.3 are the progressive portion and phases 3.1 to 3.4 are the iterative refinement part of the algorithm. Graphic from (Edgar, 2004b).

2.2.3 MAFFT

The MSA algorithm MAFFT (Katoh *et al.*, 2002, 2005) uses a fast Fourier transform (FFT), to reduce computational time without a reduction in accuracy. FFT analysis is used to quickly find peaks of similarity throughout the sequences (see Figure 2.4). MAFFT has options to allow the user to do iterative refinement similar to MUSCLE and ProbCons. MAFFT provides the user with different strategies to choose from, ranging in speed and accuracy. MAFFT has been shown to be very accurate on DNA data sets (Carroll *et al.*, 2007).

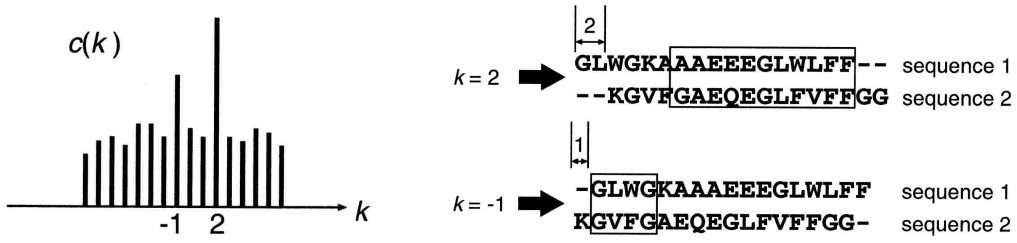


Figure 2.4: (Left) Example results from the FFT used in MAFFT. (Right) Positioning of sequences that correspond with k values. Graphic from (Katoh *et al.*, 2002).

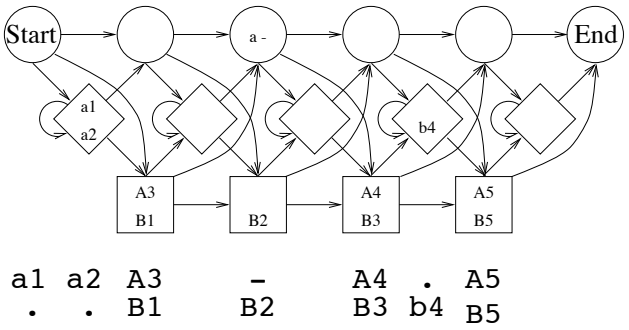


Figure 2.5: A linear hidden Markov model with each node corresponding to a column in the alignment. Each sequence uses a match state (square), an insert state (diamond) or a delete state (circle) for every column. Also, an example alignment of sequences A and B is also shown. Graphic from the SAM manual.

2.3 Hidden Markov Models

2.3.1 SAM

MSA algorithms have been an active area of research for several years. One of the earliest MSA algorithm is SAM (Sequence Alignment and Modeling System) (Krogh *et al.*, 1994). In 1994, Krogh *et al.* successfully used a hidden Markov model (HMM) to produce global MSAs. Their algorithm, SAM, has been used to aid in secondary structure prediction (Karplus *et al.*, 1998) and is still actively maintained. The states in their model represent the different columns in a MSA (see Figure 2.5). Transitions are added to allow for gaps. The models are trained on a data set of sequences using

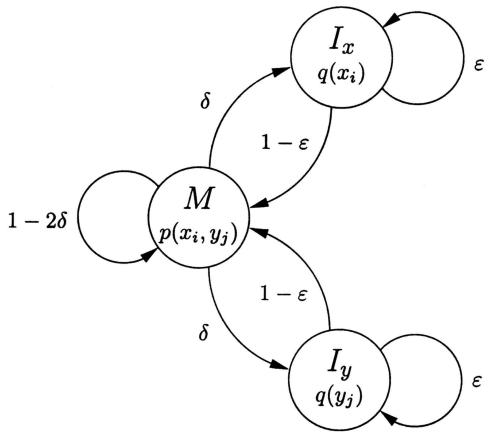


Figure 2.6: ProbCons' three-state pair-HMM for alignment of sequences x and y. State M emits two letters from each sequence. State I_x emits the letter in sequence x (and a gap for the other sequence), and state I_y does the opposite. Graphic from (Do *et al.*, 2005).

an expectation-maximization algorithm. Once the model has been trained, it can either be used to produce an alignment or search a database for similar sequences. Due to SAM's popularity, other HMMs have been introduced (see (Eddy, 1998) for a review).

2.3.2 ProbCons

ProbCons (Do *et al.*, 2005) combines techniques from HMMs, progressive and iterative refinement methods. Initially, ProbCons calculates posterior probabilities of nucleotide substitution values from a simple three-state pair-HMM (see Figure 2.6). It then uses these values in a Needleman-Wunsch matrix to calculate a pairwise alignment. A probabilistic value is calculated for each alignment and a guide tree is produced through a greedy clustering method. Next, ProbCons uses a standard progressive alignment approach, aligning the sequences in the order dictated by the guide tree. Then it follows the same procedure as MUSCLE and MAFFT to iteratively refine the alignment with a series of bipartitions in the guide tree and re-alignment

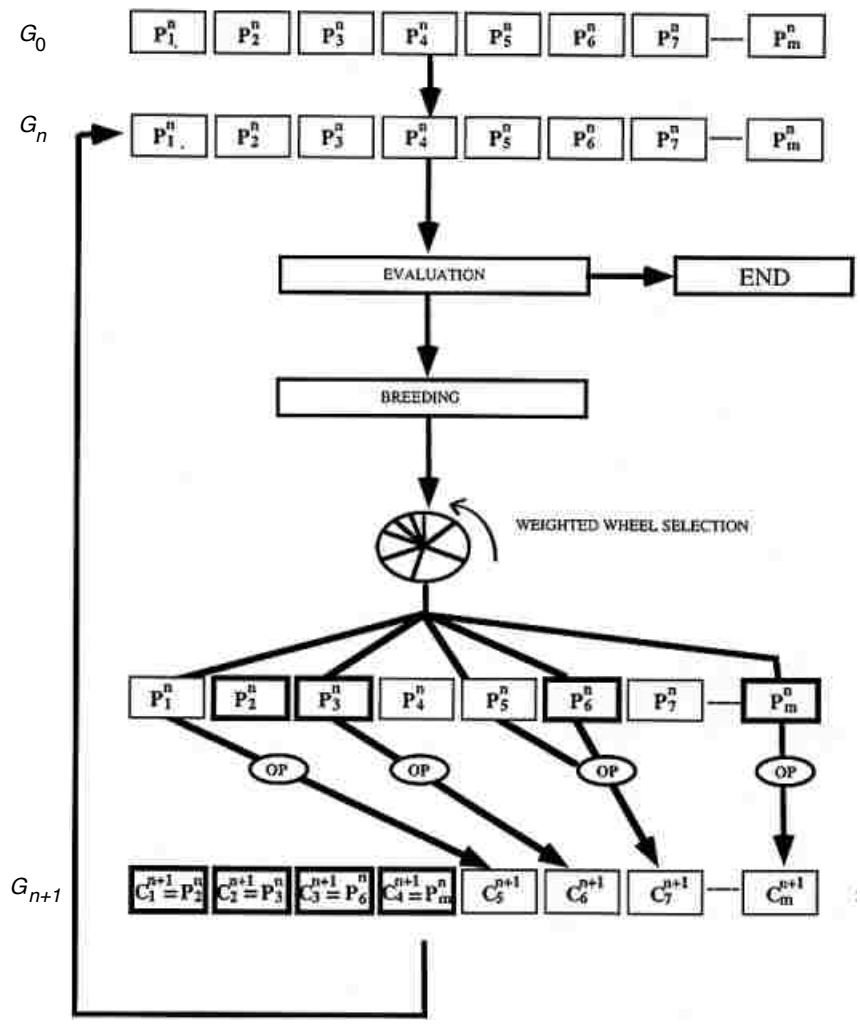


Figure 2.7: The basic structure of SAGA. See text for details. Graphic from (Notredame and Higgins, 1996).

of the two groups of sequences. ProbCons has been shown to give accuracy scores comparable to T-Coffee (Do *et al.*, 2005).

2.4 Genetic Algorithms for Multiple Sequence Alignment

2.4.1 SAGA

In 1996, Nortedame and Higgins developed SAGA (Notredame and Higgins, 1996), a global MSA algorithm that optimizes the self sum of pairs objective function with

a genetic algorithm. They evolved populations of sequences in a quasi-evolutionary manner using 22 different operators (comprised of several block shuffling operators, two crossover types, block searching, gap insertion and local rearrangement). These operators are dynamically scheduled, starting with uniform probabilities. Figure 2.7 illustrates the basic structure of SAGA. The initial population of alignments is indicted by G_0. Subsequent generations are G_n. A *parent* of the n^{th} generation is denoted as P_i^n. *Children* of those parents are similarly noted. Both parents and children are alignments. Breeding is determined with by a *weighted wheel selection* technique (selection without replacement). *OP* refers to a randomly chosen operator. While SAGA can use any objective function, using the self sum of pairs it has been shown to produce comparable results (Notredame and Higgins, 1996). Since the development of SAGA, other MSA algorithms that use a genetic algorithm have been published (Szustakowski and Weng, 2000; Zhang and Wong, 1997).

2.5 Optimization Alignment

2.5.1 POY

POY (Wheeler *et al.*, 2003) uses a completely different approach to MSA. It uses *Optimization Alignment*, a process that creates a phylogenetic tree without requiring a multiple sequence alignment as input. In POY, the tree is created and then the alignment inferred from the tree is calculated. Therefore, the alignment is a means to the end and not the goal itself. Calculating an alignment for every tree analyzed is very time consuming. Finally, POY only infers phylogenies for DNA sequences.

2.6 Benchmarking Results

To provide more insight into how these different algorithms compare, eleven of the above MSA algorithms were recently benchmarked in terms of their execution times

Table 2.2: Reference Sum of Pairs Score and CPU Time Ranks

MSA Algorithm	Reference Sum of Pairs Rank (↓)	CPU Time Rank
ProbCons	**7.68**	7.44
MAFFT-LINSI	7.26	7.28
MAFFT-GINSI	6.96	7.67
MUSCLE-Default	6.67	5.44
MAFFT-NSI	6.56	7.01
T-Coffee	6.26	10.39
ClustalW	5.59	3.87
Kalign	5.49	**1.43**
MUSCLE-Fast	5.34	4.06
MAFFT-NS1	4.66	5.86
DIALIGN	3.52	5.56

For each category, the ranks according to the Friedman test are given. Results are the aggregates of 3,541 alignment data sets. For the Reference Sum of Pairs scores, the higher the rank indicates higher accuracy. For the times, a lower rank indicates better performance in comparison to other algorithms. The alignment algorithms that ranked the best in each column are presented in bold face. The results are statistically significant with a P-value $< 2.2 \times 10^{-16}$ (using a Chi-square test). Data from (Carroll *et al.*, 2007).

and reference sum of pairs scores (see Table 2.2). For this comparison, the BAliBASE (Thompson *et al.*, 2005), OXBench (Raghava *et al.*, 2003), PREFAB (Edgar, 2004b) and SMART (Letunic *et al.*, 2004) databases are used. The algorithms chosen were selected for their popularity and availability. The performance of each algorithm on each data set is ranked. An exclusively better algorithm would have a rank of eleven for reference sum of pairs (and one for CPU time rank). Interestingly, the ordering of the performance of algorithms in terms of CPU time is much more discriminatory than that of the reference sum of pairs scores. Kalign nearly universally calculates MSAs faster then other algorithms, and T-Coffee almost always takes the most amount of time.

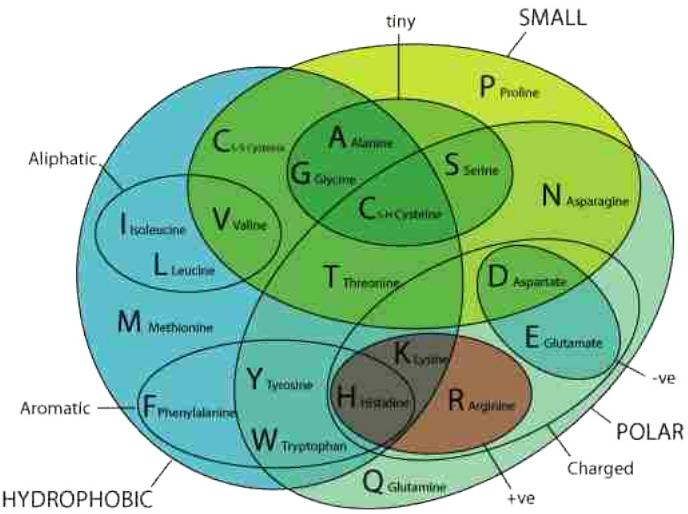


Figure 2.8: Venn diagram of some of the properties of the 20 amino acids. Amino acids are indicated by both their single letter abbreviations and their full names.

2.7 Physicochemical Properties

In an effort to better model nature in bioinformatics analysis, several researchers are using the structural and biochemical characteristics of the 20 amino acids (Goldman and Yang, 1994; Xia and Li, 1998) (see Figure 2.8). Sneath published values for 134 physicochemical properties for each of the amino acids (Sneath, 1966). These properties, such as volume, weight and hydropathy tendencies, represent the molecular forces impacting the sequences. Slightly more recently, Grantham argued for using chemical properties for amino acid exchanges (Grantham, 1974). Since then, Xia and Li have studied ten amino acid properties and their effects on the evolution of the genetic code (Xia and Li, 1998). Their studies include a multiple sequence alignment of sequences and a corresponding evolutionary tree. To determine selection for a physicochemical property, they calculate the mean of the property's values for all the pairwise combinations of amino acids, and compare it to empirical data. Their results suggest that the genetic code has minimized polarity and hydropathy. Furthermore, Woolley *et al.* (2003) use their algorithm, TreeSAAP, to calculate the

difference between physicochemical property values of two amino acids to determine selection.

2.7.1 Pairwise Sequence Alignment

While researchers are using physicochemical properties for various processes, few have incorporated them into sequence alignment. Gonnet and Lisacek (2001, 2002) used the physicochemical property hydrophobicity along with secondary structures α-helices and β-strands to build regular expressions to find similar genetic sequences in protein databases. Comparing these regular expressions against other sequences is a form of pairwise alignment.

Gupta *et al.* (2005) developed a similarity scoring method using the FFT algorithm to find subsequences with high similarity of a single physicochemical property, but not character similarity. The authors suggest that it "is suited for detailed analysis of sequences in a locality and can be wrapped over by other global alignment tools" (Gupta *et al.*, 2005). This scoring metric has only been used to perform pairwise alignments.

2.7.2 Multiple Sequence Alignment

The most notable use of physicochemical properties in MSA is ClustalW's modification of the gap open penalty. The penalty is reduced by one third for any position within a stretch of five or more hydrophilic amino acids[1] without a gap (Thompson *et al.*, 1994). These stretches usually indicate regions with a loop where gaps are more likely. While this is a step in the right direction, it does not account for the multitude of other characteristics that can be explained with physicochemical properties. The improvement in accuracy seen with a minimal incorporation of physicochemical prop-

[1] ClustalW conservatively defaults to considering the following amino acids as hydrophilic: {D,E,G,K,N,Q,P,R,S}

erties in ClustalW reinforces an overall strategy using physicochemical properties for multiple sequence alignment.

2.8 Secondary Structures

Some sequence alignment algorithms incorporate secondary structure elements (SSEs) (α-helixes, β-strands and loops). Of these, the depth of incorporation varies from modifying the gap penalties to algorithms built explicitly for using the secondary structure assignments.

Lüthy *et al.* (1991) were the first known group to use different substitution matrices based on the secondary structures. They applied this to database searching by extending their "profile method" (Gribskov *et al.*, 1987). The profile method determines if a sequence in a database belongs to a family of proteins by aligning it to an existing alignment, or profile. While their method is reported to find more related sequences in a database than other methods (with less false positives) (Lüthy *et al.*, 1991), it has not been shown to be effective for pairwise or multiple sequence alignment. In fact, the profile (alignment) used in the database searching was not produced by their method.

Other researchers have also developed algorithms that use secondary structures for database searches and pairwise alignment (Fontana *et al.*, 2005; Ginalski *et al.*, 2003, 2004; Jeong *et al.*, 2006; Soding, 2005; Sturrock and Dryden, 1997; Taylor and Orengo, 1989a). While using SSEs has improved these approaches, their algorithms have not been extended to multiple sequence alignments.

2.8.1 PRALINE

PRALINE (Heringa, 1999; Simossis and Heringa, 2003, 2004, 2005) is a multiple sequence alignment algorithm that incorporates secondary structure predictions to choose substitution matrices. It uses Lüthy's (1991) substitution matrices when the

two amino acids have the same SSE. It also uses different gap open and gap extension penalties for the different substitution matrices. Finally, PRALINE iterates between alignment and predicting the secondary structure elements. The first alignment does not use secondary structure information. While PRALINE uses different substitution matrices for different secondary structure elements, it does not account for the physicochemical properties of the sequences. Additionally, PRALINE is only available through an interactive website and therefore requires substantial amounts of human interaction for large-scale use or testing.

2.8.2 Jennings' Method

Jennings *et al.* (2001) approached the problem with the attitude that "It was considered important that the computational tools employed in this work were readily available in the public domain and that the implementation should be within the grasp of scientists in the area" (Jennings *et al.*, 2001). To this end, they modified the substitution matrix supplied to ClustalW to incorporate a degenerate set of amino acids and a secondary structure element. For example, one of the cells of the matrix holds a value for the cost to align any of the aromatic residues (`H,W,F,Y`) that are in an α-helix with any of the polar residues (`Q,N,S,T`) that are in a β-strand. The amino acids were clustered so that the 20 by 20 substitution matrix could incorporate secondary structure elements. While this approach incorporates secondary structure into amino acid alignment, it does so at the expense of the specificity of the substitution costs.

2.8.3 Horizontal Sequence Alignment (HSA)

Zhang and Kahveci (2005, 2006) use a graph-based approach to calculate multiple sequence alignments. They call their method Horizontal Sequence Alignment (HSA). Additionally, they incorporate secondary structure information by modifying the edge

weights based on the secondary structures of the two nodes. While they publish good performance on eight BAliBASE datasets, the time complexity of their algorithm is $O(W^K N + K^2 M^2)$ where K is the number of sequences, W is the sliding window size, N is the sequence length and M is the number of fragments in a protein sequence (Zhang and Kahveci, 2006). This suggests that their algorithm is only suitable for very small data sets.

2.8.4 PROMALS

PROMALS (Pei and Grishin, 2007) uses HMMs and probabilistic consistency-based scoring (in a similar manner as ProbCons) to perform alignments. It uses a simple HMM just to calculate alignments of closely related sequences ($\geq$ 60% identity). For the rest of the sequences, it runs PSI-BLAST (Altschul *et al.*, 1997) and PSIRED (Jones, 1999) to get homologous sequences and the secondary structures. The HMM for these sequences emits both an amino acid and a SSE. This second HMM requires about 30 minutes for 24 sequences (Pei and Grishin, 2007). Comparatively, ClustalW executes in a matter of seconds for a data set of the same size.

2.9 Characterization of MSA Algorithms

For multiple sequence alignment algorithms to produce biologically meaningful results, there are three main characteristics that are essential:

1. Minimizes changes in physicochemical properties
2. Incorporates secondary structure information
3. Utilizes a dynamic evaluation function

Table 2.3 characterizes the MSA algorithms discussed above in terms of these attributes. First, biologically relevant alignments minimize changes in physicochemical properties. None of the existing algorithms fully incorporate physicochemical properties. As mentioned earlier, ClustalW does adjust the gap open penalty for stretches

of hydrophilic amino acids. Most of the optimization criteria for these algorithms are either sequence similarity (self sum of pairs) or minimizing the summation of substitution matrix values. Second, biologically accurate alignments account for the contextual information found in protein secondary structures. Secondary structures are more conserved than the amino acid sequences (Gibrat *et al.*, 1996; Rost, 1999; Sander and Schneider, 1991). This more resilient information reflects natural forces. Unfortunately, most of the algorithms do not use secondary structures. Third, evolutionary forces vary depending upon the context of the amino acid. PRALINE is the only algorithm that uses a dynamic evaluation function to align sequences. For all of the other algorithms, a static evaluation function is used over the entire length of the sequences. This treats every position in the sequence as if it is in the "average" position. A new MSA algorithm is needed that incorporates these three pieces of information to produce biologically accurate alignments.

Table 2.3: Characterization of MSA Algorithms

MSA Algorithm	Physico-chemical Properties	Secondary Structures	Dynamic Evaluation Function	Optimization Criterion
ClustalW	GOP	No	No	Sub Mat
DIALIGN	No	No	No	Sub Mat
Kalign	No	No	No	Sub Mat
MAFFT-GINSI	No	No	No	SSofP/Consistency
MAFFT-LINSI	No	No	No	SSofP/Consistency
MAFFT-NS1	No	No	No	Sub Mat
MAFFT-NSI	No	No	No	SSofP
MUSCLE	No	No	No	Sub Mat
POY	No	No	No	Phylogeny Score
ProbCons	No	No	No	SSofP
SAM	No	No	No	EM
SAGA	No	No	No	SSofP
T-Coffee	No	No	No	Consistency
PRALINE	No	Yes	Yes	Sub Mat
Jennings *et al.*	No	Yes	No	Sub Mat
HSA	No	Yes	No	SSofP
PROMALS	No	Yes	No	Consistency

Abbreviations: GOP = Gap open penalty; SSofP = Self sum of pairs; Sub Mat = Minimization of substitution matrix values; Consistency = COFFEE like consistency between multiple sequence alignment and pairwise alignments (Notredame *et al.*, 1998); EM = Expectation Maximization.

Chapter 3

DNA Reference Alignment Benchmarks Based on Tertiary Structure of Encoded Proteins

Hyrum Carroll, Wesley Beckstead, Timothy O'Connor, Mark Ebbert,

Mark Clement, Quinn Snell and David McClellan

Published in Bioinformatics

(Carroll *et al.*, 2007)

Abstract

Motivation: Multiple sequence alignments (MSAs) are at the heart of bioinformatics analysis. Recently, a number of multiple protein sequence alignment benchmarks (i.e., BAliBASE, OXBench, PREFAB and SMART) have been released to evaluate new and existing MSA applications. These databases have been well received by researchers and help to quantitatively evaluate MSA programs on protein sequences. Unfortunately, corresponding DNA benchmarks are not available, making evaluation of MSA programs difficult for DNA sequences.

Results: This work presents the first known multiple DNA sequence alignment benchmarks that are 1) comprised of protein-coding portions of DNA 2) based on biological features such as the tertiary structure of encoded proteins. These reference DNA databases contain a total of 3,545 alignments, comprising of 68,581 sequences. Two versions of the database are available: mdsa_100s and mdsa_all. The mdsa_100s version contains the alignments of the data sets that TBLASTN found 100% sequence identity for each sequence. The mdsa_all version includes all hits with an E-value score above the threshold of 0.001. A primary use of these databases is to benchmark the performance of MSA applications on DNA data sets. The first such case study is included in the supplementary material.

Availability: The databases, further details and the supplementary material are publicly available at http://csl.cs.byu.edu/mdsas/.

3.1 Introduction

Multiple sequence alignments (MSAs) provide the foundation for much of the analysis in bioinformatics. They are the first step for everything from annotation of genomes to evolutionary studies. Because of this, it is crucial for automated alignment programs to generate highly accurate and biologically meaningful MSAs to ensure accuracy in subsequent steps in the research process.

Recently, a number of protein sequence databases have been presented to provide a benchmark for alignment algorithms: BAliBASE (Thompson *et al.*, 2005), OXBench (Raghava *et al.*, 2003), PREFAB (Edgar, 2004b), and SMART (Ponting *et al.*, 1999). These databases leverage structural alignments to provide a suite of "gold standard" alignments. They are assumed to be the "true" alignments, and calculated alignments are evaluated by comparing against them. They have been well accepted by the scientific community and used in numerous studies to compare the quality of protein alignments generated by MSA programs (Do *et al.*, 2005; Edgar, 2004a,b; Karplus and Hu, 2001; Lassmann and Sonnhammer, 2002, 2005a; Thompson *et al.*, 1999b; Van Walle, 2004). These multiple protein sequence alignment (MPSA) benchmarks are limited to the evaluation of protein alignment applications.

Rarely is a novel alignment technique assessed for its ability to align nucleotide data accurately. The shortage of assessments of MSAs with DNA data may be due to the lack of DNA reference alignments. Applications that work well on amino acid sequences may not be as accurate on DNA data sets. One solution to this problem would be to compare calculated nucleotide alignments against reference nucleotide alignments that are based on the biological features used in protein benchmarks.

Work has been done to address this lack of reference DNA alignments. Pollard *et al.* (2004) created a benchmarking tool for the alignment of non-protein coding DNA using simulated data. While this benchmark gives researchers a starting point

to evaluate DNA alignments, the degree to which the simulated sequences reflect those in nature is uncertain.

A "gold standard" benchmark of DNA alignments that is 1) comprised of protein-coding portions of DNA and 2) based on biological features such as the tertiary structure of encoded proteins can help researchers assess the quality of DNA alignment algorithms. This paper presents the first known collection of protein-coding DNA benchmark alignments that meet this criteria. A computational tool, MPSA2MDSA, was developed and utilized to convert the following MPSAs into multiple DNA sequence alignment (MDSAs): BAliBASE, OXBench, PREFAB, and SMART.

3.2 Materials and Methods

Estimating a MDSA from a MPSA is a straight forward procedure that requires three steps. The first step is to find the best corresponding DNA sequence (hit) from a protein sequence (query). We queried the September 2006 version of GenBank's nt database (Benson *et al.*, 2005) with each of the protein sequences using the TBLASTN algorithm (Altschul *et al.*, 1990). TBLASTN provides the accession number of the best hit. The DNA sequences are then retrieved from the nt database with fastacmd, an NCBI tool. The second step is to account for the occasional gaps introduced by the similarity search. The final step is to apply the alignment from the MPSA to the MDSA. This is done by inserting gaps that correspond with the gaps in the protein alignment. This step is important to preserve the alignment features obtained by higher order methods (e.g., secondary and tertiary structure or chemical properties) or in other words, to preserve the higher order benchmark alignment. By preserving the biological information, the DNA alignment can be considered a reference alignment. Each step is covered in more detail in the supplementary material.

Two versions of each database are publicly available. The first version, mdsa_100s, includes only those data sets with all perfect matches (100% sequence

identity). This version ensures the highest level of integrity in the conversion. The second version, mdsa_all, includes all hits with an E-value score above the threshold of 0.001. This version retains more of the MPSAs and aids in comparison with the original MPSAs.

For any heuristic, it is important to quantify the accuracy. Here the accuracy can be measured by the sequence identity of the hit sequence. In general, as the sequence identity increases, so does the likelihood that the two sequences share the same tertiary structure. For this work, sequences that share 100% sequence identity are assumed to have the same tertiary structure. Sequences with the same tertiary structure will have the same alignment.

Using the nt database, 97.4% of the protein queries found a match with an E-value score above the threshold of 0.001. Furthermore, 69.0% of these hits have 100% sequence identity with the query. While the tool finds a high percentage of exact matches with a current database, databases are growing at an exponential rate, thereby increasing the number of hits of protein queries.

In total, 3,545 DNA reference alignments, comprising of 68,581 sequences and 35,600,958 bases are publicly available at http://csl.cs.byu.edu/mdsas/.

To illustrate the usefulness of the reference DNA databases, a case study of the performance and ranks of alignment programs on DNA data sets is included in the supplementary material (see also Table 3.1). Alignments and their respective scores were calculated for seven different multiple sequence alignment applications for each of the 3,545 alignments.

3.3 Conclusion

In this work, the first known databases of reference protein-coding DNA alignments are presented. These databases are constructed by leveraging the popular BLAST program to find DNA sequences corresponding to those found in multiple protein

Table 3.1: Q Score, TC Score and CPU Time Ranks

	DNA Data Sets			Amino Acid Data Sets		
Program	Q Score Rank	TC Score Rank	CPU Time Rank	Q Score Rank	TC Score Rank	CPU Time Rank
CLUSTALW	6.35	5.94	4.81	5.99	5.59	3.87
DIALIGN	4.35	4.71	6.03	3.53	3.52	5.56
Kalign	6.77	6.05	**1.69**	6.04	5.49	**1.43**
MAFFT-GINSI	9.22	9.59	7.12	6.30	6.96	7.67
MAFFT-LINSI	**9.31**	**9.73**	6.27	6.74	7.26	7.28
MAFFT-NS1	7.70	6.96	4.77	4.84	4.66	5.86
MAFFT-NSI	8.68	8.73	6.08	5.84	6.56	7.01
MUSCLE-Default	6.63	7.43	6.02	6.66	6.67	5.44
MUSCLE-Fast	5.22	5.06	4.37	5.94	5.34	4.06
POY	4.03	3.94	9.52	-	-	-
ProbCons	6.24	6.38	9.96	**7.64**	**7.68**	7.44
T-Coffee	3.51	3.52	11.36	6.47	6.26	10.39

For each category, the ranks according to the Friedman test are given. For the Q and TC scores, the higher the rank indicates higher accuracy. For the times, a lower rank indicates better performance in comparison to other programs. The alignment programs that ranked the best in each column are presented in bold face.

sequence alignments. The alignments of the protein sequences (which reflect higher-order information) are applied to the DNA sequences to qualify them to be reference alignments. High quality hits were obtained from public databases. Over two-thirds of the queries found a perfect match in the nt database. Two versions of the converted databases are available, the first only contains hits that perfectly matched the query, and the comprehensive second version includes all hits above the cut-off threshold. These DNA reference alignment databases are publicly available. This benchmark will be extremely useful in evaluating the quality of DNA alignments generated by existing and forthcoming MSA techniques. Finally, the first case study of DNA alignments evaluated by these reference alignments is included in the supplementary material.

Acknowledgments

We would like to thank Keith Crandall for his review and critiques of this work. This material is based upon work supported by the National Science Foundation under Grant No. 0120718.

3.4 Supplementary Material

3.4.1 Introduction

Protein-coding (exonic) DNA alignments are useful for several applications, especially where more information and sensitivity is desired than what amino acid alignments offer. Murphy *et al.* (2007) have reported that "Protein-coding alignments have the advantage of being more reliable for establishing sequence alignment orthology than noncoding alignments". As another example, selection studies primarily use protein-coding DNA alignments (Chamala *et al.*, 2007; Marques *et al.*, 2006; Porter *et al.*, 2007; Zhang *et al.*, 2004). Furthermore, gene prediction for sequences without strong homology to known amino acid sequences can also use on DNA alignments (Mathé *et al.*, 2002). With all of these examples, and many others, the underlying principle is that accurate protein-coding DNA alignments are essential for accurate analysis.

Starting with Thompson and her group publishing the BAliBASE database in 1999 (Thompson *et al.*, 1999a), several amino acid reference benchmarks have been released in the past few years (Edgar, 2004b; Letunic *et al.*, 2004; Ponting *et al.*, 1999; Raghava *et al.*, 2003; Subramanian *et al.*, 2005; Thompson *et al.*, 2005; Van Walle *et al.*, 2005). Most of these databases leverage structural alignments to provide a suite of "gold standard" alignments. These benchmarks have been well accepted by the community to provide evaluations of the accuracy of multiple sequence alignment (MSA) programs (Do *et al.*, 2005; Edgar, 2004a,b; Karplus and Hu, 2001; Lassmann and Sonnhammer, 2002, 2005a; Subramanian *et al.*, 2005; Thompson *et al.*, 1999b;

Van Walle, 2004). Unfortunately, researchers assessing the accuracy of MSA algorithms have focused almost exclusively on amino acid alignments. This focus is due primarily to a lack of reference DNA data sets.

Previous work addresses this lack of reference DNA alignments. Pollard *et al.* (2004) created alignments of non-protein coding DNA using simulated data. While this benchmark gives researchers a starting point to evaluate DNA alignments, the degree to which the simulated sequences reflect those in nature is uncertain. Akin to the alignments used in this study, but serving a different purpose, are those constructed by Gardner *et al.* (2005, 2004, 2007). Their database, BRAliBase, facilitates the assessment of aligning structural RNAs.

We recently published four protein-coding DNA reference multiple sequence alignment databases (Carroll *et al.*, 2007). The databases allow researchers to quantitatively evaluate multiple sequence alignments using DNA in the same manner as is done with amino acid sequences. In this paper, we provide the details of the process of conversion and provide a study analyzing the accuracy of multiple sequence alignment programs on multiple amino acid (protein) sequence alignments (MPSAs) and these multiple DNA sequence alignments (MDSAs).

3.4.2 Methods

We have included each of the amino acid databases in Table 3.2 in our benchmark. BAliBASE (Benchmark Alignment dataBASE) contains reference alignments that have been manually refined and validated by superposition of known tertiary structures (Thompson *et al.*, 2005). OXBench (from the University of Oxford) contains automated amino acid alignments that were benchmarked using tertiary structure associations (Raghava *et al.*, 2003). PREFAB (Protein REFerence Alignment Benchmark) contains amino acid alignments based on pairs of amino acid sequences that have been structurally aligned and supplemented with as many as 50 homologs found

Table 3.2: Reference Protein Alignment Benchmark Suites

Name	Version	# of Alignments
BAliBASE (Thompson *et al.*, 2005)	3.0	498
OXBench (Raghava *et al.*, 2003)	1.3	672
PREFAB (Edgar, 2004b)	4.0	1682
SMART (Ponting *et al.*, 1999)	June 7, 2006	701

by PSI-BLAST (Edgar, 2004b). SMART (Simple Modular Architecture Research Tool) alignments were also manually refined with structure comparisons, but where no structure was available, automated alignment techniques were used (Letunic *et al.*, 2004).

DNA Benchmark Alignments

We estimate a DNA benchmark alignment from an amino acid alignment in three steps (see Figure 3.1):

1. Similarity searching
2. Reconciling inconsistencies
3. Applying the multiple amino acid sequence alignment

First, we use TBLASTN (Altschul *et al.*, 1990) to perform a similarity search of an amino acid sequence to get a corresponding DNA sequence. Second, we reconcile any inconsistencies in the hit sequence, in terms of length or introduced gaps, by inserting gaps or ambiguous characters respectively. Finally, we insert the gaps dictated by the MPSA into the MDSA to reflect biological accuracy. We implement these three steps in a computer program called MPSA2MDSA. The details of these steps are covered in the remainder of this section.

Step 1: Similarity Search The first step in building a multiple DNA sequence alignment involves finding DNA sequences that correspond to the amino acid sequences in the MPSA. Corresponding sequences can be determined by similarity

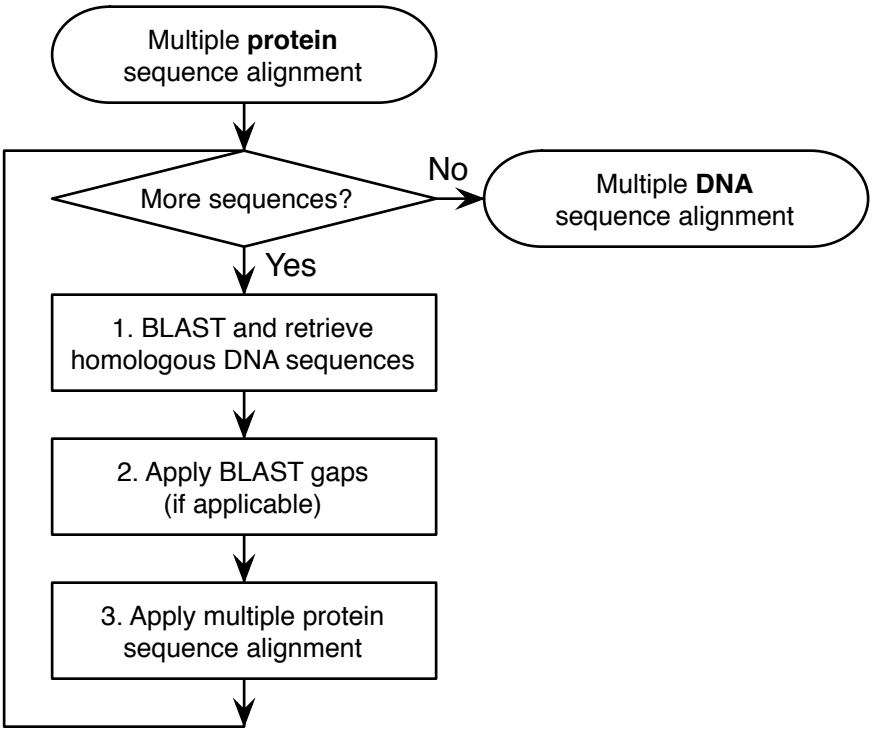


Figure 3.1: Flow chart for MPSA2MDSA.

searches when an appropriate statistical test is used as the metric or scheme (Karlin and Altschul, 1990). We use the Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990) for the similarity search algorithm. We choose BLAST for its statistical scoring metric, performance, ease of use and speed (McGinnis and Madden, 2004). TBLASTN is a BLAST derivative, which translates nucleotide databases into amino acid sequences in all six reading frames, then identifies the most statistically probable sequences as hits (Altschul *et al.*, 1997). The input to TBLASTN is an amino acid sequence (the query), a database of nucleotide sequences and a cut-off threshold for the E-value. For this work, similarity searches are performed on the September 2006 version of the nt GenBank (Benson *et al.*, 2005) database, which has 16.9 billion base pairs in 3.8 million sequences. The second parameter to TBLASTN is a cut-off threshold value. In this study, matches with an E-value larger than 0.001 are ignored. This threshold is interpreted as there being 0.001 matches with a similar score or better due to chance in the current database. The output of TBLASTN

is the translated sequences with the lowest E-value and corresponding identification information. As it is the DNA sequences that we are interested in, we use the NCBI tool fastacmd to retrieve the corresponding DNA sequences from the nt database.

Step 2: Reconcile Inconsistencies The second step to building a MDSA is to account for the occasional gaps introduced by the similarity search. BLAST, like other similarity search programs, uses a pairwise alignment criteria for matches. Adding gaps into the hit sequence (which account for insertions/deletions) can improve the calculated likelihood that the query and the modified hit sequence correspond. This produces two sources of gaps in the hit sequence: terminal gaps and interior gaps. Terminal gaps occur when the matching portion of the hit sequence is shorter than the query sequence (it either does not start early enough and/or it is not long enough). The user can choose to account for interior gaps by either ignoring them or adding additional gaps into the MDSA. Finally, if a hit sequence does not provide the DNA for a section of the query (due to gaps), the least ambiguous characters possible are inserted to account for the missing data. For example, if the amino acid in the query sequence is tyrosine, then the first two nucleotides are known to be thymine and adenine respectively, and the most resolution that the third character can have is a pyrimidine (thymine or cytosine).

Step 3: Apply Multiple Amino Acid Sequence Alignment In the last step to produce a MDSA, MPSA2MDSA applies the alignment from the MPSA to the hit sequences. This step is important to preserve the alignment features obtained by higher order methods (e.g., secondary and tertiary structure and chemical properties) or in other words, to preserve the higher order benchmark alignment. For each gap in the MPSA, the program inserts three gaps into the MDSA at the respective location.

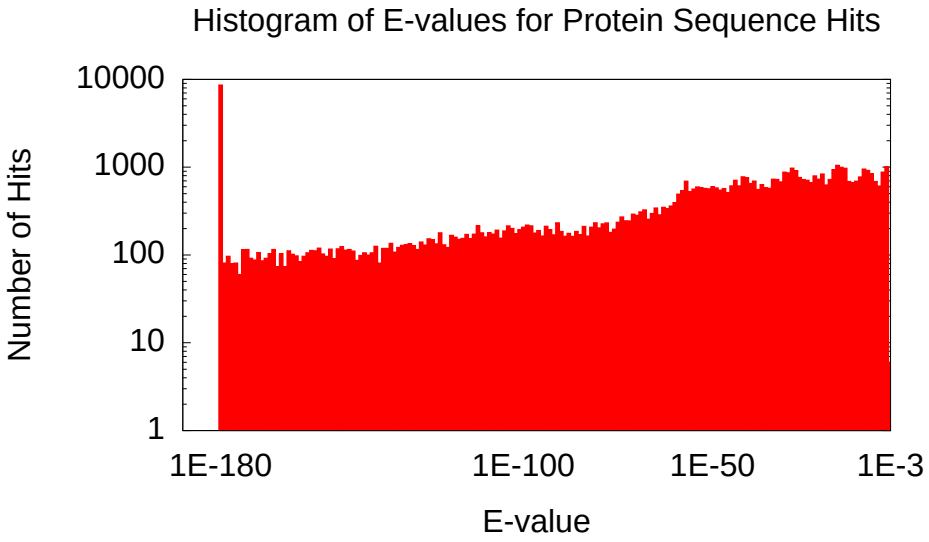


Figure 3.2: Histogram of the E-values from the hits of the protein sequences. (Note: To conservatively correct for scores reported by BLAST to have an E-value of 0.0, scores less than or equal to 1E-180 are reported as 1E-180.)

DNA Benchmark Alignment Conversion

In general, MPSA2MDSA finds good matches in the nt database in terms of sequence identity and E-value. The majority (69.0%) of amino acid sequences have matches in the database that have 100% sequence identity with the translated DNA sequences. Furthermore, another 3.9% of the hits have only one mismatched amino acid with the amino acid query. In terms of E-values, 98.3% of the amino acid sequences found a DNA sequence in the database with a score of 0.001 or better. A lower E-value (for a given length) indicates higher similarity between the query and the hit sequences. Figure 3.2 illustrates all the E-values for the amino acid sequence hits. In the graphic, E-values with a score better than or equal to 1E-180, are conservatively displayed at the 1E-180 location to accommodate a logarithmic axis. This adjustment accounts for the E-values that BLAST reports as 0.0 (8,567 of them, or 12.5% of all hits).

E-value scores are calculated from the length and similarity of the query and hit sequences. Figure 3.3 shows the correlation between the length and the E-value of the hits found in the nt database. The data here suggests that as a database

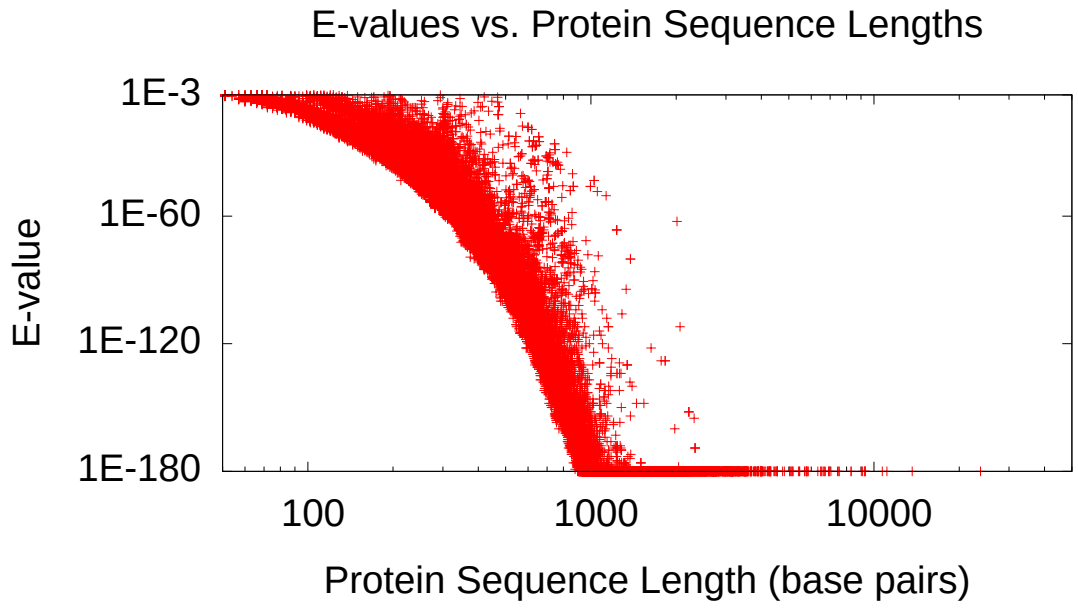


Figure 3.3: E-values of all the hits plotted against the length of the protein sequence query. (Note: To conservatively correct for scores reported by BLAST to have an E-value of 0.0, scores less than or equal to 1E-180 are reported as 1E-180.)

increases (i.e., as longer and more corresponding sequences are included), BLAST will find hits with greater similarity to the query. While MPSA2MDSA already finds a high percentage of quality matches, as databases continue to grow at an exponential rate, more and higher quality hits will match the amino acid queries.

In total, the DNA databases derived from BAliBASE, OXBench, PREFAB and SMART contain 3,545 reference alignments, comprising of 68,581 sequences and 35,600,958 bases. These reference alignments are publicly available at http://dna.cs.byu.edu/mdsas/.

Experimental Setup

We perform an alignment study by testing several of the leading alignment programs on these MDSAs. The purpose of this study is to test the empirical performance of commonly used alignment programs on protein-coding DNA. These programs have

previously been tested on amino acid benchmarks by others and shown to be effective at aligning amino acid sequences. Even so, their performance on DNA sequences is virtually unknown, hence benchmarking these programs on DNA reference data sets is important. In addition to running each alignment program on reference MDSAs, we also run them on the reference MPSAs. This provides for a uniform method of assessing each alignment algorithm on amino acid sequences and comparing these results to the accuracy of each alignment algorithm on the corresponding DNA sequences that are found in our DNA alignments. All test alignments and accuracy measures were executed with the supercomputers in the Ira and Mary Lou Fulton Supercomputing Laboratory at Brigham Young University, using Dual-core Intel Xeon EM64T processors (2.6GHz) with 8 GB of memory.

Alignment Programs We chose eight different alignment programs to benchmark: ClustalW (Thompson *et al.*, 1994), DIALIGN (Morgenstern *et al.*, 1998), Kalign (Lassmann and Sonnhammer, 2005b), MAFFT (Katoh *et al.*, 2005), MUSCLE (Edgar, 2004a), POY (Wheeler *et al.*, 2003), ProbCons (Do *et al.*, 2005), and T-Coffee (Notredame *et al.*, 2000). These programs use a variety of strategies to construct a multiple sequence alignment, such as progressive alignment, iterative refinement, probabilistic alignment etc. (see Table 3.3). They are widely used in biology and bioinformatics. For each alignment program, we used default parameters, unless noted otherwise in Table 3.4).

Alignment Benchmarks The alignment programs are evaluated with the following MPSAs: BAliBASE, OXBench, PREFAB, and SMART, as well as their respective MDSAs. The one exception is POY, which in the version tested restricts its analysis to DNA sequences. For BAliBASE, OXBench and SMART, we did not consider alignments that have over 100 sequences in order to make the test manageable for the

Table 3.3: Categorization of Multiple Sequence Alignment Programs

Program	Progressive	Iterative	Local
CLUSTALW	X		
DIALIGN		X	X
Kalign	X		
MAFFT-GINSI	X	X	
MAFFT-LINSI	X	X	X
MAFFT-NS1	X		
MAFFT-NSI	X	X	
MUSCLE-Default	X	X	
MUSCLE-Fast	X		
POY	X	X[1]	
ProbCons/ProbConsRNA	X[2]	X	
T-Coffee	X		X

The progressive column indicates programs that use progressive alignment algorithm (Feng and Doolittle, 1987). Iterative refers to programs to refine the multiple sequence alignment. Programs that incorporate local alignment (in addition to global alignment) have a mark in the local column. [1]Optimization Alignment, [2]Markov model

slower programs. In addition, we discard alignments that did not complete within two weeks for one or more MSA programs.

We use reference sets 1–5 of BAliBASE for assessing each alignment algorithm on DNA sequences. Reference sets 6–8 contain repeats, inversions and transmembrane helices. We exclude these reference sets because none of the chosen alignment programs are designed to handle these cases. MPSA2MDSA, converts all of the amino acid alignments in reference sets 1–5 to DNA alignments. We exclude eight of these alignments because they contain more than 100 sequences, allowing 378 DNA alignments to be included in the study for BAliBASE. To test each alignment algorithm on amino acid sequences, we use the 378 corresponding amino acid alignments in BAliBASE.

For OXBench, MPSA2MDSA successfully converts all 672 MPSAs to MDSAs. We discard four alignments that were over 100 sequences and four alignments that aborted or did not finish after two weeks while being analyzed. In total, we include

Table 3.4: Arguments Used For Multiple Sequence Alignment Programs

Program	Version	Arguments
ClustalW	1.83	defaults
DIALIGN	2.2.1	defaults
Kalign	2.0	defaults
MAFFT-GINSI	5.861	–maxiterate 1000 –globalpair
MAFFT-LINSI	5.861	–maxiterate 1000 –localpair
MAFFT-NS1	5.861	–maxiterate 0 –retree 1
MAFFT-NSI	5.861	–maxiterate 1000
MUSCLE-Default	3.6	-stable
MUSCLE-Fast	3.6	-stable -maxiters 1 -diags
POY	3.0.11	-replicates 10 -repintermediate
ProbCons/ProbConsRNA	1.10	-ir 1000
T-Coffee	4.58	defaults

664 DNA alignments in the study for OXBench. For analyzing each alignment algorithm on amino acid sequences, we use 668 corresponding amino acid alignments in OXBench.

MPSA2MDSA converts 1676 of the 1682 amino acid alignments in PREFAB to DNA alignments. We use these alignments and all 1682 amino acid alignments.

For the SMART database, we use the June 7, 2006 version. MPSA2MDSA converts 698 of the 701 MPSAs in SMART to MDSAs. We exclude 108 alignments that either contain over 100 sequences or did not complete within two weeks for all programs. This gives a total of 590 MDSAs and 592 MPSAs from SMART.

Accuracy Measurement and Statistical Analysis To ascertain the accuracy of the alignments generated by each program we use a variety of scoring metrics that compare a calculated multiple sequence alignment to a reference alignment. In general, we use the scoring metrics that are provided by or suggested for each respective database. These scoring metrics are all forms of the Q (Quality) and TC (Total Column) scores. The Q score, previously termed as the developer score (Sauder *et al.*, 2000) or SPS (Sum of Pairs Score) (Thompson *et al.*, 1999b), is defined as the

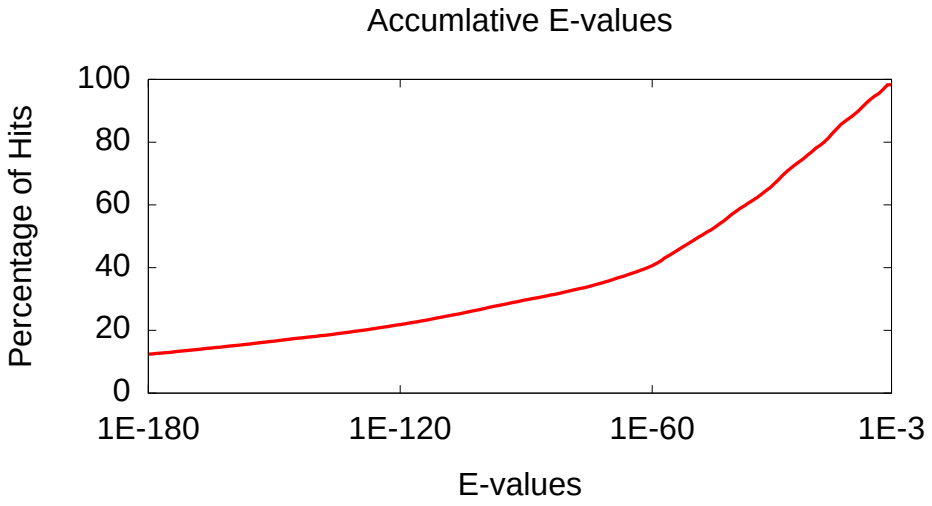


Figure 3.4: Aggregates of the number of hits plotted against E-values. (Note: To conservatively correct for scores reported by BLAST to have an E-value of 0.0, scores less than or equal to 1E-180 are reported as 1E-180.)

number of correctly aligned residue pairs in the generated alignment divided by the number of residue pairs in the reference alignment. The TC score, also known as the CS score (Karplus and Hu, 2001), is the number of correctly aligned columns in the generated alignment divided by the number of columns in the reference alignment. The TC score is the same as the Q score in the case of pairwise alignment. For BAliBASE, OXBench and SMART we use the Q and TC scores. We use only the Q score for PREFAB since the alignments in these databases are pairwise.

For an individual database, we average each score across all of the alignments. To measure statistical significance in the accuracy differences between alignment programs, we perform a Friedman rank test with the accuracy scores (Friedman, 1937). This test is more conservative than the Wilcoxon test, which has also been used to determine statistical significance in past alignment studies (Edgar, 2004b).

3.4.3 Results

To assess the accuracies of several multiple sequence alignment programs on protein-coding DNA, we use the protein-coding DNA benchmark alignments created by Carroll *et al.* (2007). The Friedman ranks of accuracy and CPU times for the converted BAliBASE, OXBench, PREFAB and SMART databases are given in Tables 3.5–3.8. The differences in ranks are statistically significant (p-value $\ll 0.0001$). Furthermore, the differences are relevant in that the average values differ significantly. For comparison purposes, the Friedman ranks for accuracy and CPU times on the original amino acid databases are given in Tables 3.9–3.12. The average CPU times are dramatically lower on the amino acid alignments than on the respective DNA alignments since the amino acid sequences are one third the length of the corresponding DNA sequences. As is the case with the ranks of the DNA alignments, the differences in ranks for the amino acid alignment benchmarks are both statistically significant (p-value $\ll$ 0.0001) and relevant.

Two general points about the accuracy ranks are worth noting as an overview. First, the amino acid benchmarks generally have higher accuracy scores than the DNA benchmarks. Furthermore, the range of scores cover a smaller interval for the amino acid databases. The accuracies for the OXBench and SMART databases exemplify this well. The average accuracy scores for amino acid data sets in OXBench only vary between 0.82 and 0.86 (Table 3.10) and the SMART scores range from 0.76 to 0.87 (Table 3.12). For the DNA alignments of OXBench, the corresponding scores range from 0.69 to 0.80. The range varies even more for the DNA SMART database (0.44 to 0.83). While the inherent difference in the length of the two types of data is a likely reason for the improvement, the primary factor is unknown. Possibly, these higher accuracy scores are due to adjustments encouraged from MPSAs during development of these algorithms.

Table 3.5: DNA BAliBASE scores, times, and ranks

Program	Q Score		TC Score		CPU Time	
	Avg.	Rank	Avg.	Rank	Avg.	Rank
MAFFT-GINSI	.617	**11.21**	.277	**9.91**	58.1	5.87
MAFFT-LINSI	.607	10.75	.275	9.78	47.9	6.28
MAFFT-NSI	.559	9.57	.207	8.30	21.7	4.02
MUSCLE-Default	.516	8.02	.198	7.90	188.0	8.02
ProbCons	.452	6.78	.124	6.39	3228.9	10.10
MAFFT-NS1	.459	6.68	.141	6.44	2.3	2.25
ClustalW	.445	5.78	.120	5.45	52.0	6.53
Kalign	.408	5.27	.105	5.40	1.7	**1.44**
MUSCLE-Fast	.291	4.42	.099	5.05	6.3	3.11
DIALIGN	.389	4.34	.099	5.24	169.7	7.80
POY	.305	2.59	.045	3.79	26364.4	11.14
T-Coffee	.308	2.56	.071	4.35	10453.7	11.45

The average Q scores, TC scores, and times (in seconds) for the DNA alignments of BAliBASE. The ranks according to the Friedman test are given for each category. A higher Q and TC score rank indicate better accuracy in comparison with other programs. For the CPU times, a lower rank indicates better performance. The best rank for each category appears in boldface. On this database, MAFFT-GINSI achieves a Q score rank higher than that of any applications on any of the other database. Furthermore, POY requires a longer average CPU time here than for any other application on any of the other database.

Table 3.6: DNA OXBench scores, times, and ranks

	Q Score		TC Score		CPU Time	
Program	Avg.	Rank	Avg.	Rank	Avg.	Rank
MAFFT-LINSI	.795	**9.11**	.699	**8.72**	1.4	7.29
MAFFT-GINSI	.789	8.50	.687	8.07	1.2	7.67
MAFFT-NSI	.789	8.33	.687	8.03	0.5	6.12
MAFFT-NS1	.782	7.97	.677	7.70	0.8	5.16
ClustalW	.766	7.46	.671	7.61	1.6	4.89
MUSCLE-Default	.755	6.75	.660	7.02	1.4	6.47
Kalign	.756	6.30	.645	5.88	0.2	**1.57**
MUSCLE-Fast	.743	6.02	.643	6.40	0.4	3.51
ProbCons	.741	5.57	.626	5.34	8.2	9.58
DIALIGN	.696	4.57	.604	5.50	1.5	5.23
POY	.694	3.85	.574	3.91	79.4	8.95
T-Coffee	.692	3.56	.577	3.80	49.0	11.55

Column descriptions and other details are as in Table 3.5. These runs have the smallest difference between the highest and lowest average Q score accuracies, yet the Q score ranks still distinguish between an clear ordering.

Table 3.7: DNA PREFAB scores, times, and ranks

	Q Score		CPU Time	
Program	Avg.	Rank	Avg.	Rank
MAFFT-LINSI	.380	**8.39**	1.1	7.41
MAFFT-NS1	.376	8.15	0.7	5.44
MAFFT-GINSI	.376	8.15	1.0	7.66
MAFFT-NSI	.375	8.03	0.8	6.70
Kalign	.344	7.19	0.33	**1.88**
ClustalW	.351	6.88	0.34	3.94
ProbCons	.298	5.82	4.5	10.11
MUSCLE-Fast	.297	5.73	0.4	3.80
MUSCLE-Default	.297	5.73	1.5	6.44
POY	.254	4.74	2.4	8.59
DIALIGN	.248	4.70	0.77	4.90
T-Coffee	.254	4.50	7.2	11.14

Column descriptions and other details are as in Table 3.5. TC scores are omitted since the data sets only have two sequences each. Here, the average Q score accuracies are the lowest of any of the database, yet the same general ordering of applications is still preserved.

Table 3.8: DNA SMART scores, times, and ranks

Program	Q Score		TC Score		CPU Time	
	Avg.	Rank	Avg.	Rank	Avg.	Rank
MAFFT-GINSI	.833	**11.08**	.468	**10.81**	9.6	6.19
MAFFT-LINSI	.812	10.67	.460	10.65	0.5	2.55
MAFFT-NSI	.790	9.93	.415	9.70	8.2	6.00
MUSCLE-Default	.700	7.71	.331	7.48	1.3	3.15
ProbCons	.701	7.49	.301	7.38	544.4	9.86
Kalign	.687	7.27	.294	6.68	0.5	**1.52**
MAFFT-NS1	.673	7.07	.288	6.62	4.2	4.60
ClustalW	.577	4.41	.224	4.68	10.7	5.61
MUSCLE-Fast	.550	3.79	.194	3.78	19.3	7.48
POY	.555	3.51	.200	4.07	4163.6	11.15
DIALIGN	.515	3.29	.183	3.58	37.3	8.23
T-Coffee	.444	1.77	.146	2.67	2507.1	11.66

Column descriptions and other details are as in Table 3.5. For all of the DNA databases, MAFFT-GINSI achieves the highest average Q score accuracy here than any other application. Furthermore, not all of the applications are able to achieve high accuracies for this database as is shown by the largest difference between the highest and lowest Q score ranks for any of the DNA or amino acid database.

The second point about the accuracy ranks, and more important of the two, is that the results show that certain programs that achieve high accuracy scores on amino acid sequences tend to rank low for DNA sequences. T-Coffee and ProbCons, for example, rank very high on amino acid benchmarks but they are the least accurate of all the alignment methods for many of the DNA databases. Conversely, other alignment algorithms achieve higher ranks on the DNA databases then on the amino acid databases. The MAFFT strategies (MAFFT-LINSI, MAFFT-GINSI, and MAFFT-NSI) have lower accuracies than ProbCons and MUSCLE on the amino acid benchmarks, but achieve the highest accuracy scores on every DNA benchmark. These two points indicate that there is room for improvement of the existing multiple sequence alignment algorithms for protein-coding DNA data.

Table 3.9: Amino Acid BAliBASE scores, times, and ranks

	Q Score		TC Score		CPU Time	
Program	Avg.	Rank	Avg.	Rank	Avg.	Rank
ProbCons	.866	**9.06**	.620	**8.76**	250.0	9.86
MAFFT-LINSI	.860	8.59	.616	8.20	6.8	6.56
MAFFT-GINSI	.847	7.88	.586	7.51	8.8	6.83
MAFFT-NSI	.832	6.79	.571	6.87	3.8	5.32
T-Coffee	.815	6.74	.557	6.68	150.6	10.46
MUSCLE-Default	.828	6.43	.550	6.33	7.0	6.65
Kalign	.811	5.37	.526	5.16	0.3	**1.21**
MUSCLE-Fast	.776	4.37	.473	4.52	1.4	3.27
MAFFT-NS1	.784	4.20	.484	4.33	0.6	3.15
ClustalW	.755	3.59	.447	4.22	5.1	4.94
DIALIGN	.743	2.97	.435	3.42	20.3	7.75

Column descriptions and other details are as in Table 3.5 except that here the average Q and TC scores only cover the core blocks. Two general differences between the performance of the alignment applications on the amino acid and corresponding DNA databases are evident here: First, the alignment applications achieved higher average accuracy scores for the amino acid databases. Second, the general ordering of the applications, in terms of their accuracies, is significantly different.

Table 3.10: Amino Acid OXBench scores, times, and ranks

	Q Score		TC Score		CPU Time	
Program	Avg.	Rank	Avg.	Rank	Avg.	Rank
MUSCLE-Default	.861	**6.78**	.775	**6.88**	0.81	6.33
ClustalW	.861	**6.78**	.772	6.78	0.93	3.68
MUSCLE-Fast	.859	6.66	.772	6.74	0.79	4.90
ProbCons	.859	6.47	.768	6.21	0.91	5.95
MAFFT-LINSI	.852	6.42	.766	6.36	0.57	7.25
T-Coffee	.856	6.29	.767	6.31	4.45	10.24
Kalign	.854	6.18	.766	6.25	0.05	**1.19**
MAFFT-GINSI	.853	5.67	.760	5.52	0.51	7.53
MAFFT-NSI	.852	5.64	.760	5.56	0.83	7.78
MAFFT-NS1	.847	5.06	.752	5.07	0.50	6.67
DIALIGN	.823	4.04	.733	4.31	0.58	4.48

Column descriptions and other details are as in Table 3.5. The runs for this database have the smallest range of average accuracy scores, and not surprisingly the smallest range of ranks too. This suggests that the accuracies of the alignment applications is less distinguishable here than for other databases.

Table 3.11: Amino Acid PREFAB Q scores, and ranks

Program	Q Score		CPU Time	
	Avg.	Rank	Avg.	Rank
ProbCons	.590	**7.18**	0.43	6.54
ClustalW	.585	6.99	0.59	3.64
T-Coffee	.583	6.69	1.76	10.21
MUSCLE-Fast	.584	6.62	0.38	4.32
MUSCLE-Default	.584	6.62	0.32	4.53
Kalign	.588	6.53	0.19	**1.58**
MAFFT-LINSI	.571	5.89	0.64	7.80
MAFFT-GINSI	.558	5.14	0.65	8.27
MAFFT-NS1	.558	5.14	0.49	7.15
MAFFT-NSI	.558	5.14	0.41	7.58
DIALIGN	.513	4.07	0.73	4.38

Column descriptions and other details are as in Table 3.5. TC scores are omitted since the data sets only have two sequences each. This also contributes to this database having the fastest average CPU times of the databases.

Table 3.12: Amino Acid SMART scores, times, and ranks

Program	Q Score		TC Score		CPU Time	
	Avg.	Rank	Avg.	Rank	Avg.	Rank
ProbCons	.873	**8.87**	.550	**8.32**	39.49	9.32
MAFFT-GINSI	.871	8.57	.549	7.95	2.25	6.95
MAFFT-LINSI	.858	7.78	.533	7.46	2.04	6.57
MAFFT-NSI	.853	7.04	.534	7.29	1.34	6.10
MUSCLE-Default	.851	6.80	.520	6.72	1.86	5.89
T-Coffee	.836	5.90	.490	5.91	78.03	10.91
Kalign	.830	5.23	.478	5.00	0.27	**1.45**
MUSCLE-Fast	.823	4.73	.461	4.59	0.52	3.18
ClustalW	.819	4.55	.481	5.43	1.31	3.85
MAFFT-NS1	.818	4.38	.460	4.49	0.53	3.90
DIALIGN	.766	2.15	.395	2.84	6.18	7.87

Column descriptions and other details are as in Table 3.5. Here, MAFFT-GINSI has the highest average Q score for any application on any of the databases.

MSA Program Discussion

In this section, each of the alignment applications benchmarked in this study are discussed in alphabetical order.

ClustalW Even though ClustalW (Thompson *et al.*, 1994) is the oldest alignment application tested, it consistently produced alignments with high accuracies for the amino acid database. In fact, for the OXBench database, it achieved the highest Q score rank (shared with MUSCLE-Default). For the DNA alignments however, its rank is typically in the middle of all of the programs. An exception of this is on the DNA OXBench database, in that it achieves the next best rank after all of the MAFFT strategies.

DIALIGN DIALIGN (Morgenstern *et al.*, 1998) is consistently the least accurate on amino acid sequences with an overall rank of 3.53 (Carroll *et al.*, 2007). Using DNA data sets, DIALIGN does better in the rankings (4.35 (Carroll *et al.*, 2007)). DIALIGN is not particularly fast either. On amino acid and DNA sequences, DIALIGN ranges in rank from the third to the ninth fastest alignment program. It is worth noting that the benchmarks are global alignments and DIALIGN calculates local alignments. DIALIGN only truly calculates a global alignment if the local alignment spans the entire length of all of the sequences.

Kalign Kalign (Lassmann and Sonnhammer, 2005b) is extremely fast and consistently ranks number one in execution time on all databases. The longest average time for Kalign on a database is only 1.7 seconds (BAliBASE MDSAs). This is inordinately fast, considering T-Coffee averages over 7,000 seconds and ProbCons averages 3,900 seconds on the same database. This would seem to indicate that Kalign takes a great reduction in accuracy in order to achieve this type of speed, but the results suggest otherwise. Kalign consistently takes first place in CPU time while maintain-

ing moderately high accuracy scores on DNA and amino acid sequences and a decent "middle ground" ranking according to Q and TC scores. This is important to many biologists who are interested in aligning large data sets quickly without taking a large reduction in accuracy.

MAFFT MAFFT (Katoh *et al.*, 2005) does well on amino acid sequences, but is surpassed in accuracy ranks in many instances by ProbCons and MUSCLE. On the DNA benchmarks, MAFFT maintains its high accuracy scores. MAFFT-GINSI, MAFFT-LINSI and MAFFT-NS1 rank first, second and third respectively on all DNA benchmarks. In the case of the DNA alignments from PREFAB, all four MAFFT strategies do better than any other alignment method. MAFFT does this without a significant loss in execution time, generally ranking around fifth or sixth. For these reasons, MAFFT is a good choice for any biologist interested in aligning either DNA or amino acid sequences in a decent amount of time.

MUSCLE The MUSCLE (Edgar, 2004a) strategies (MUSCLE-Fast and MUSCLE-Default) consistently rank well on the amino acid benchmarks. Even MUSCLE-Fast, which does not include iterative refinement, does better than many alignment programs. MUSCLE retains its accuracy on DNA but is surpassed by the MAFFT strategies.

POY POY (Wheeler *et al.*, 2003) was chosen in order to assess the quality of DNA alignments that are produced as it performs optimization alignment and creates a tree without the use of a MSA as input. The accuracy of the DNA alignments produced by POY to build a tree has been virtually unknown due to the lack of DNA benchmarks in the past. The results show that POY has low accuracy scores compared to most of the other alignment methods tested. POY consistently ranks second or third to last in accuracy. The goal of the POY analysis is to eliminate errors produced by preliminary

alignment programs. It does this by producing the alignment in conjunction with the phylogenetic tree. Though this is a worthy goal, these results suggest that the alignments of POY are not as accurate as other alignment programs, and this may affect the resulting tree that is produced. POY's lower accuracy ranks may be due in part to it focusing on building a refined phylogeny for non-coding DNA, while we tested it with protein-coding DNA.

ProbCons ProbCons (Do *et al.*, 2005) does very well in the alignment of amino acid sequences. It ranks first in accuracy on three of the four amino acid databases. ProbCons requires large amounts of time to accomplish this, making it one of the slowest methods tested. For DNA data sets, ProbCons drops in the rankings and in general places around seventh in terms of accuracy. This suggests that ProbCons has been optimized for amino acid sequences but it may not be the best choice for aligning DNA sequences.

T-Coffee T-Coffee (Notredame *et al.*, 2000) also does well on amino acid benchmarks but at a great cost in time. T-Coffee is ranked last on amino acid benchmarks in the rankings according to CPU time. When tested on DNA, T-Coffee, like ProbCons, drops in accuracy and consistently ranks the lowest. It also gets the lowest rank for execution time on the DNA benchmarks. These results suggest that the current version of T-Coffee is an undesirable choice for the alignment of DNA even though it does well in aligning amino acid sequences.

3.4.4 Conclusion

The results of this study show that many alignment programs appear to be optimized and/or trained on amino acid sequences, but vary greatly in accuracy when applied to DNA sequences. Not only are accuracies generally lower for the DNA databases, but the most accurate applications for the amino acid database are not the most

accurate for the DNA databases. The MAFFT-LINSI, MAFFT-GINSI, and MAFFT-NSI strategies are the most accurate on DNA sequences while T-Coffee, DIALIGN and POY are the least accurate.

Chapter 4

ChemAlign: Biologically Relevant Multiple Sequence Alignment Using Physicochemical Properties

Hyrum D. Carroll, Mark J. Clement, Quinn O. Snell and David A. McClellan

Submitted to Bioinformatics

Abstract

Motivation: Physicochemical properties (e.g., polarity, hydropathy, etc.) quantitatively characterize the 20 amino acids. Researchers use these properties to understand the underlying mechanisms influencing amino acid exchanges. We present a new algorithm, ChemAlign, that uses physicochemical properties to achieve biologically relevant multiple sequence alignments.

Results: ChemAlign achieves higher accuracies (reference sum of pairs scores) than the other programs analyzed (ClustalW, MAFFT, ProbCons and PRALINE) for two different classes of data sets. First, we consider some of the largest data sets in the BAliBASE, HOMSTRAD, OXBench and SMART databases. Second, we include data sets in the "Midnight Zone" (very low sequence identity ($< 25\%$)). These two classes represent the major challenges for current alignment programs. Additionally, we introduce the Physicochemical Property Difference (PPD) score. This score is the normalized difference of physicochemical property values between a calculated and a reference alignment. It takes a step beyond sequence similarity and measures characteristics of the amino acids to provide a more biologically relevant metric. ChemAlign earns the highest PPD scores for both classes of data sets. These higher accuracies translate into more biologically correct alignments, as is shown with an example of identifying potential drug docking sites.

Availability: ChemAlign is implemented in the PSODA package. PSODA is open source, free and available for Mac OS X, Linux, Windows and other operating systems at `http://dna.cs.byu.edu/psoda`.

4.1 Introduction

Multiple sequence alignments (MSAs) are at the heart of several bioinformatics research areas. For example, alignments are used to identify conserved regions, which are crucial to finding drug docking sites. Current methods can miss biologically relevant features such as these because they only consider sequence similarity. Most of them are further limited because they do not incorporate secondary structure information. Particularly difficult for these methods are data sets with very low percent identity. These data sets are one of the best sources for finding drug targets since they contain distantly related species and therefore conserved regions are more obvious. The globin family is a good example of this. Even though myoglobin was the first protein to have its structure determined (Kendrew *et al.*, 1958), the globin family remains difficult for existing methods to align correctly. The HOMSTRAD database (Mizuguchi *et al.*, 1998) includes a data set with globin domains. This data set is at the bottom of the "Twilight Zone" (Doolittle, 1994) with an average percent identity of 25.9%. Of the algorithms tested here, the best one only aligns 38.4% of the positions correctly. On the other hand, using a physicochemical property, ChemAlign correctly aligns 90.6% of the positions. Figure 4.1 shows an example hemoglobin protein with marked conserved regions. The regions are determined from an alignment using ChemAlign, and appear at a possible drug docking site. ChemAlign is able to find both regions, whereas other algorithms do not.

ChemAlign uses physicochemical properties to produce biologically relevant alignments. Researchers have used these properties in other areas (Goldman and Yang, 1994; Grantham, 1974; Xia and Li, 1998). The AAindex database (Kawashima *et al.*, 2008) has numerical values for each of the amino acids for over 500 properties. These properties include volume (Bigelow, 1967), polarity (Grantham, 1974) and hydropathy (Kyte and Doolittle, 1982). The purpose of the properties is to quantitatively capture the differences between the amino acids. They have been used to iden-

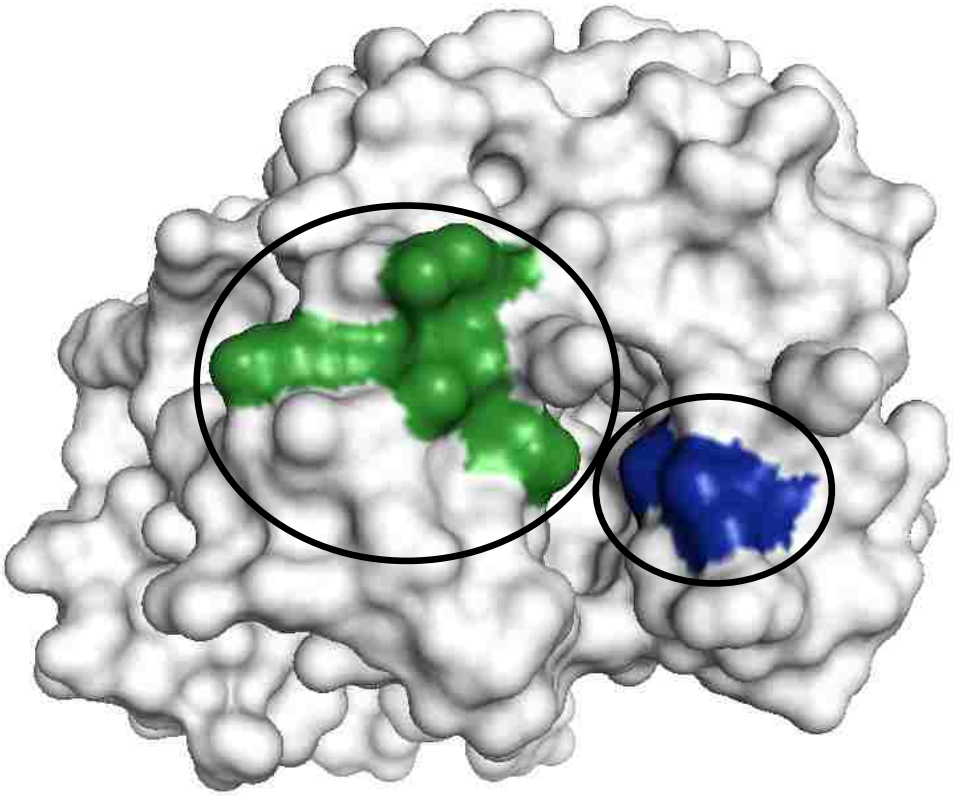


Figure 4.1: Hemoglobin (1A4FA) protein with highlighted conserved regions determined by ChemAlign. The regions are at a possible drug docking site. ChemAlign is able to find both regions, whereas other algorithms are only able to find the one of the left.

tify sites under selection (Woolley *et al.*, 2003), in phylogeny reconstruction (Thorne *et al.*, 1996), to correlate and approximate the most commonly used substitution matrices (Méndez *et al.*, 2008; Pokarowski *et al.*, 2007; Rudnicki and Komorowski, 2005), to identify characteristics of alignments (Afonnikov and Kolchanov, 2004; Thorvaldsen *et al.*, 2005; Wrabl and Grishin, 2005), for protein secondary structure prediction (Lim, 1974; Periti *et al.*, 1967), pairwise alignment (Gonnet and Lisacek, 2002; Gupta *et al.*, 2005) and to identify particular proteins (Kim *et al.*, 2000).

Physicochemical properties have a varying effect depending on the secondary structure where they occur. ChemAlign incorporates knowledge of the secondary structure elements (SSEs) (α-helices, β-strands and loops) to capitalize on this and

address one of the problems with current alignment techniques as stated by Thorne *et al.* (1996):

> A problem with the Dayhoff approach is that it effectively models the replacement process at the "average" site in the "average" protein. There may be no such thing as an "average" site in an "average" protein.

Each amino acid in a protein belongs to one of the SSEs. Typically they are determined from tertiary structure information, if it is known (Kabsch and Sander, 1983), or are predicted (e.g., using PSIRED (Jones, 1999)). Protein secondary structure has long been understood to be more conserved than the amino acid sequence. This has been verified through a number of different experiments and reports (Gibrat *et al.*, 1996; Rost, 1999; Sander and Schneider, 1991). Using this more resilient information has improved the accuracy of sequence alignments (Heringa, 1999; Jennings *et al.*, 2001; Lüthy *et al.*, 1991; Sturrock and Dryden, 1997; Zhang and Kahveci, 2006).

In this paper, we explore the hypothesis that using physicochemical properties and secondary structures produces biologically relevant multiple sequences alignments. To do so, we introduce ChemAlign, which incorporates both physicochemical properties and secondary structures.

The remainder of this paper is as follows. First, we discuss alignment methods related to ChemAlign. Next, we detail how physicochemical properties and secondary structures are used in alignment. This is followed by comparisons of accuracy measurements of ChemAlign and other programs for several data sets, with an in-depth look at the globin domain family. We close this paper with some concluding remarks and future directions of this work.

4.2 Related Work

Multiple sequence alignment is an active area of research (Edgar and Batzoglou, 2006). Related to ChemAlign are sequence alignment algorithms that fit into three categories:

1. Uses primary sequence information
2. Incorporates secondary structure elements
3. Integrates physicochemical properties

Each of these categories are reviewed in this section.

4.2.1 Alignment Using Primary Sequence Information

Among the primary sequence alignment applications, MAFFT (Katoh *et al.*, 2005) and ProbCons (Do *et al.*, 2005) deserve special attention. In a benchmarking study performed previously (Carroll *et al.*, 2007), these two applications performed the best. The defining characteristic of MAFFT is that it uses a fast Fourier transformation (FFT) to quickly find peaks of similarity throughout the sequences. ProbCons on the other hand combines techniques from hidden Markov models, progressive and iterative refinement methods. While both ProbCons and MAFFT obtain high accuracies scores from benchmark testing, they only use sequence similarity for alignment. Data sets with low sequence identity are difficult for these algorithms to align correctly since they do not leverage physicochemical properties and secondary structures. Furthermore, regions of the alignment that are governed by a physicochemical property more than sequence similarity will be missed using sequence information alone.

4.2.2 Alignment Using Secondary Structure

ChemAlign builds upon the success of other algorithms that use secondary structure information to improve the biological relevance of alignments. These alignment algorithms either modify the gap penalties based on the secondary structure or ex-

plicitly incorporate the elements. ClustalW(Thompson *et al.*, 1994) allows the user to input secondary structure information only when aligning two sub-alignments (profile alignments). It uses this information to adjust gap penalties, based on the SSEs. Unfortunately, the secondary structures are not used for multiple sequence alignment.

Lüthy *et al.* (1991) were the first known group to use different substitution matrices based on the SSEs. They calculated these matrices by gathering data sets with known tertiary structure, and partitioned them according to their secondary structures. Other researchers have also developed algorithms that use secondary structures for database searches and pairwise alignment (Fontana *et al.*, 2005; Ginalski *et al.*, 2003, 2004; Jeong *et al.*, 2006; Soding, 2005; Sturrock and Dryden, 1997; Taylor and Orengo, 1989a). While using SSEs has improved these approaches, their algorithms have not been extended to multiple sequence alignments.

PRALINE (Heringa, 1999) on the other hand is a MSA algorithm that incorporates secondary structure. First, it builds an alignment without secondary structure information, then uses that alignment to predict the SSEs. PRALINE continues by iterating between alignment and predicting the SSEs. Once it has the structures, it uses Lüthy's substitution matrices when the two amino acids have the same SSE. Unfortunately, the SSEs are not incorporated in the initial alignment. PRALINE is also subject to the same limitation as the primary sequence alignment algorithms—that of not being able to correctly produce alignments governed by physicochemical properties. Additionally, PRALINE is only available through an interactive website and therefore requires substantial amounts of human interaction for large-scale use or testing.

4.2.3 Alignment Using Physicochemical Properties

While researchers are using physicochemical properties for various analyses, few have incorporated them into sequence alignment. Those that do, use them:

1. In pairwise alignments
2. To find matching subsequences
3. To adjust gap penalties

ChemAlign extends these ideas to produce multiple sequence alignments. First, Gonnet and Lisacek (2002) used both the physicochemical property hydrophobicity and secondary structure assignments to build regular expressions (motifs). They use these motifs to find similar genetic sequences in protein databases. Second, Gupta *et al.* (2005) developed a pairwise similarity scoring method using a FFT algorithm to find subsequences with high similarity for a single physicochemical property. Third, the most notable use of physicochemical properties in MSA is ClustalW's modification of the gap open penalty. The penalty is reduced by one third for any position within a stretch of five or more hydrophilic amino acids without a gap (Thompson *et al.*, 1994). These stretches usually indicate regions with a loop where gaps are more likely. While this is a step in the right direction, it does not account for the multitude of other characteristics that can be accounted for with physicochemical properties. The improvement in accuracy seen with a minimal incorporation of physicochemical properties in ClustalW reinforces an overall strategy using physicochemical properties with secondary structures.

4.3 Methods

ChemAlign is a multiple sequence alignment algorithm that uses the physicochemical property values and secondary structures of amino acids. It employs a traditional dynamic programming (Needleman and Wunsch, 1970) approach during both the pairwise and the progressive phases. After calculating all of the pairwise "distances" between sequences, ChemAlign clusters them to produce a guide tree (Saitou and Nei, 1987). This tree directs the order that sequences and alignments of sequences are aligned in the progressive stage (Feng and Doolittle, 1987). ChemAlign also uses

affine gap penalties. Instead of using a substitution matrix based solely on log-odds probabilities from an amino acid database, ChemAlign combines amino acid exchange counts with normalized differences of a physicochemical property. Additionally, different substitution matrices are employed for different SSEs. In the rest of this section, we explain ChemAlign's use of physicochemical properties and secondary structures, how it calculates gap costs and PSODA, the package that ChemAlign is implemented in.

4.3.1 Substitution Matrices

ChemAlign uses a hybrid substitution matrix comprised of both observed amino acid exchanges and differences between physicochemical properties. First, to obtain the observed amino acid exchanges, we build a reference database of alignments with their secondary structures. We combined the OXBench database (Raghava *et al.*, 2003) with the respective secondary structures from the RCSB Protein Data Bank (PDB) (Berman *et al.*, 2000). To avoid adding noise, only those sequences in OXBench that have an exact match in the PDB are retained. We count the number of each set of amino acid pairs (for each column in the alignment), according to their SSEs (i.e., both α-helices, both β-strands, both loops or mismatch) producing four matrices of observed amino acid exchanges $O^{\alpha}, O^{\beta}, O^{l}$, and O^{m}. We calculate a normalized difference matrix D^{p} for a physicochemical property p using Equation 4.1.

$$D^{p}_{i,j} = 1 - \frac{2 * |p[i] - p[j]|}{argmax_{x}(p[x]) - argmin_{y}(p[y])} \tag{4.1}$$

Here, $p[i]$ is the value of a physicochemical property for amino acid i. The values of D^{p} range from -1.0 for the most dissimilar pair of amino acids to 1.0 for identical amino acids. For this work, we use the Effective Partition Energy (Miyazawa and Jernigan, 1985) for its aggregate characteristics as an illustrative physicochemical property (see Figure 4.2). This property includes hydrophobic, hydrogen bonding

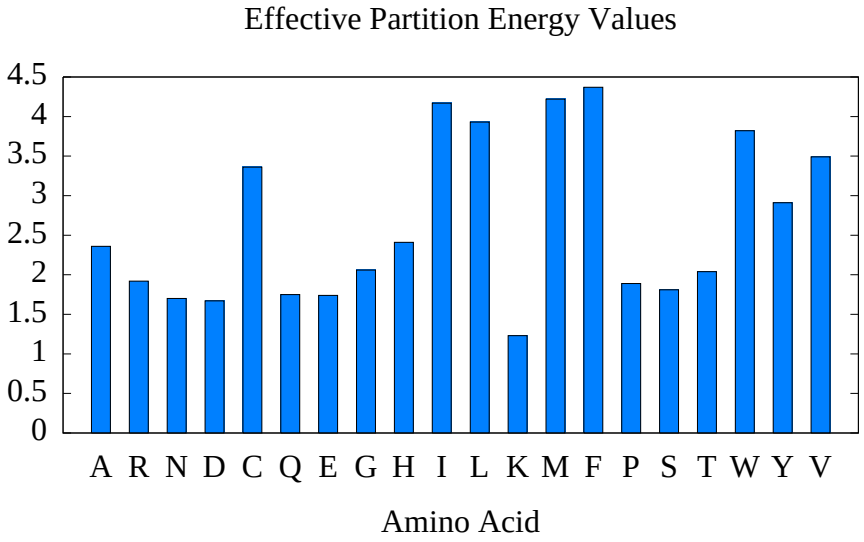


Figure 4.2: Values for the physicochemical property effective partition energy (Miyazawa and Jernigan, 1985).

and electrostatic energies. Each of the O matrices are multiplied element-wise with D^p to get M^α, M^β, M^l, and M^m, for α-helices, β-strands, loops and mismatches. Combining the O matrices with D^p aggregates the benefits of each. Finally, as is commonly done elsewhere (e.g., BLOSUM (Henikoff and Henikoff, 1992)), the log-odds probabilities of the values in each of the M matrices are calculated to get the substitution matrices S^α, S^β, S^l and S^m:

$$S_{i,j} = \frac{1}{\lambda} log\left(\frac{l_{i,j}}{f_i f_j}\right) \tag{4.2}$$

Here, $l_{i,j}$ is the likelihood that amino acids i and j appear aligned in the database and f_i is the background frequency of amino acid i. Also, λ allows for scaling the matrix. For each of the S matrices, λ is set to one. This results in four substitution matrices: S^α, S^β, S^l and S^m. These matrices are significantly different from each other. Figure 4.3 reports the similarity distance for each pair of the S matrices, BLOSUM62 (Henikoff and Henikoff, 1992) and GONNET80 (Gonnet *et al.*, 1992)(two of the most commonly used substitution matrices). The distances are calculated as

a)

	S^{α}	S^{β}	S^{l}	S^{m}	B62
S^{β}	0.14				
S^{l}	0.13	0.18			
S^{m}	0.07	0.05	0.07		
B62	0.14	0.21	0.15	0.15	
G80	0.10	0.16	0.13	0.11	0.05

b)

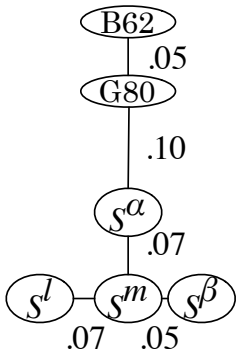


Figure 4.3: a) Distances between BLOSUM62 (B62), GONNET80 (G80) and the ChemAlign substitution matrices for α-helices (S^{α}), β-strands (S^{β}), loops (S^{l}) and mismatches (S^{m}). Distances are calculated as one minus the inner product. b) The minimum spanning tree for the similarity scores. BLOSUM62 and GONNET80 are more similar to each other than to any other matrix. Also, it is not surprising that the S matrices cluster around the mismatch substitution matrix S^{m}. Note, GONNET80 is the default substitution matrix for ClustalW.

	α-helix	β-strand	loop
α-helix	7.11		
β-strand	-12.81	2.97	
loop	-2.42	-3.33	1.95

Figure 4.4: Secondary structure scoring matrix N. The values are log-odd ratios based on observed counts in the OXBench-PDB database.

one minus the inner product of the two respective matrices. The figure also shows the minimum spanning tree for these distances. Of all the matrices, BLOSUM62 and GONNET80 are the most similar. Additionally, the S matrices cluster around the mismatch substitution matrix S^{m}. These observed differences in substitution matrices shows the importance of customizing the matrices for each of the secondary structures.

4.3.2 Incorporating Secondary Structure

ChemAlign uses a straightforward approach to incorporate protein secondary structures into both pairwise and progressive alignment. The secondary structure influences the alignment in two ways:

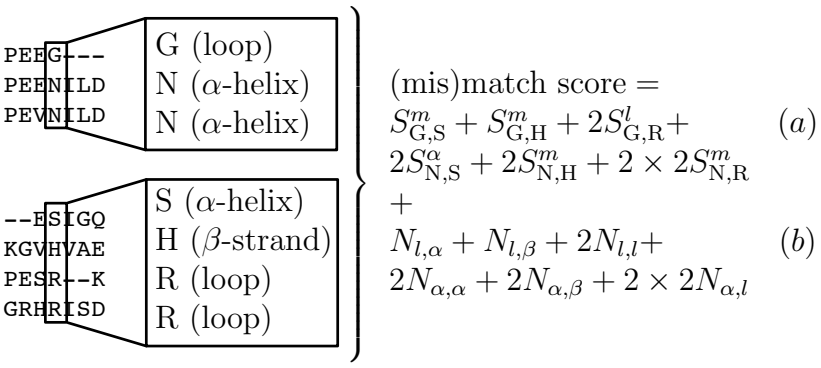


Figure 4.5: Example calculation of the (mis)match score using secondary structure elements during the progressive phase for two columns (GNN and SHRR) of two sub-alignments. Part (a) uses the S substitution matrices and part (b) the log-odds secondary structure matrix N.

1. The choice of a substitution matrix
2. An additional score for (mis)matching of the SSEs

First, it uses a substitution matrix according to the SSEs of the two amino acids currently being considered. If they are the same (either both α-helices, both β-strands or both loops), then the S^α, S^β or S^l matrix is used. If the SSEs are not the same, then the mismatch matrix, S^m, is used. ChemAlign also incorporates secondary structures by adding a (mis)match score for the SSEs to the (mis)match score for the amino acids. The SSE scores are specified by the matrix $N_{c,d}$ (where c and d are SSEs) as shown in Figure 4.4. N is the log-odds ratios of the observed matches of the SSEs in the OXBench-PDB database. For example, if the SSEs for two amino acids are an α-helix and a β-strand, then 12.81 is subtracted from the (mis)match score. Alternatively, if both SSEs are α-helices, then 7.11 is added. Incorporating N into the Needleman-Wunsch values aligns secondary structures, which are typically more conserved than the amino acids themselves.

An example calculation of the (mis)match score for two columns from two sub-alignments using the S and N matrices is illustrated in Figure 4.5. In the figure, each amino acid in the highlighted column of the first sub-alignment is matched with each amino acid in the column of the other sub-alignment. For example, the first

term in the calculation, $S^m_{\mathrm{G,S}}$, is the value for aligning the glycine (G) in the upper sub-alignment with the serine (S) in the lower sub-alignment. The S^m matrix is used because the SSEs of this glycine and serine are not the same. Furthermore, the third term, $2S^l_{\mathrm{G,R}}$, accounts for aligning the same glycine with the two arginines (R) in the lower sub-alignment. Both the glycine and the set of arginines are in loop regions, so the S^l matrix is used. The scaler in this term illustrates a performance optimization. In ChemAlign, identical amino acid-SSE pairs are treated as a single pair with a weight. This is a very common scenario, resulting in large reductions in execution times.

4.3.3 Gap Penalties

ChemAlign implements affine gap penalties with a user specified gap open penalty (GOP) and gap extension penalty (GEP). Additionally, a user may specify a gap distance penalty (GDP) (Thompson *et al.*, 1994). This penalty increases the weight of opening a gap as follows:

$$\text{gap penalty}(d) = \begin{cases} GEP & d = 0 \\ GOP\left(4 - \frac{2d}{GDP}\right) & 1 \leq d \leq GDP \\ GOP & d > GDP \end{cases} \tag{4.3}$$

where d is the number of amino acids in a sequence since the last gap. Using the gap distance penalty promotes alignments with gaps that are not close to other gaps, or in other words, more biologically agreeable alignments.

4.3.4 PSODA

ChemAlign is implemented in the software package PSODA (Carroll *et al.*, 2008a). PSODA is a comprehensive alignment and phylogenetic search package. It includes analysis under both parsimony and maximum likelihood, visualization and analysis

tools. PSODA uses a NEXUS-based format file for commands and sequence data. It is open source, free and available for Mac OS X, Linux, Windows and other operating systems. Inside of PSODA, the method `ssalign` is used to invoke ChemAlign (see section 4.4.1 for a more complete example).

4.4 Results

To quantitatively assess the performance of ChemAlign, its accuracy is compared with that of ClustalW, MAFFT, ProbCons and PRALINE. These programs are chosen for their ubiquity and performance (Carroll *et al.*, 2007). Both the reference sum of pairs score and the Physicochemical Property Difference (PPD) scores are used in our evaluation. An analysis of several data sets and an in-depth look at the globin domain family are presented. In summary, ChemAlign achieves higher accuracy scores and a more biologically meaningful alignment than the other programs tested.

4.4.1 Experimental Setup

To analyze the accuracy of ChemAlign, we look at two different classes of data sets. The first class consists of thirteen of the largest data sets in the BAliBASE (Thompson *et al.*, 2005), HOMSTRAD, OXBench and SMART (Letunic *et al.*, 2004) databases (see Table 4.1). For the other class, we collect fourteen data sets with very low sequence identity ($< 22\%$) from the HOMSTRAD and SMART databases (see Table 4.1). This range is commonly referred to as the "Midnight Zone" for sequence alignment (Rost, 1997). While the two classes overlap in definition, each is considered to explicitly address the two most difficult scenarios for MSAs. We combine each of these data sets with the SSEs from the PDB. Sequences that did not have a perfect sequence match in the PDB are filtered out.

The arguments we use for testing the different programs are given in Table 4.2. For ChemAlign, we use the default gap distance penalty of four (see section 4.3.3).

Table 4.1: Data Sets

Class	Database	Data Set	Num. Taxa	Ave. Chars	% ID
Large	BAliBASE	BBS20008	17	90.2	20.6
	BAliBASE	BBS30025	19	89.8	17.5
	BAliBASE	BBS20036	20	71.4	35.2
	BAliBASE	lrr_ref6_centre	115	23.6	20.0
	HOMSTRAD	az	27	232.7	22.4
	HOMSTRAD	globin	29	113.1	25.9
	HOMSTRAD	sermam	41	146.1	25.9
	OXBench	12	52	98.0	27.7
	OXBench	22	87	112.5	19.7
	SMART	FN3	37	83.2	16.2
	SMART	IG	41	94.5	12.5
	SMART	RRM	44	71.7	19.6
	SMART	WD40	54	41.0	17.3
Midnight Zone	HOMSTRAD	Acetyltransf	6	224.0	16.0
	HOMSTRAD	ABC_tran	6	351.8	15.0
	SMART	AAI	5	93.4	16.1
	SMART	C2	5	108.2	21.2
	SMART	CBS	12	50.0	19.1
	SMART	CHROMO	8	60.3	15.0
	SMART	CYCLIN	9	87.7	15.3
	SMART	HRDC	5	80.8	19.9
	SMART	HTH_XRE	10	56.2	21.4
	SMART	HX	11	44.8	18.5
	SMART	PUA	5	76.2	17.9
	SMART	Pumilio	8	36.3	21.1
	SMART	SANT	7	52.2	17.9
	SMART	SPEC	8	103.1	16.0

Additionally, a range of GOP and GEPs are considered. For the GOPs, we explore ten values from one tenth of the maximum value in S^m to the maximum value. For each GOP, the GEPs range from one tenth of the GOP to the GOP itself. The next set of arguments specify files containing the substitution matrices S^α, S^β, S^l and S^m. The last argument is a file containing the SSEs defined by DSSP (Kabsch and Sander, 1983). For the other programs, we use the default arguments, as is commonly

Table 4.2: Arguments Used For Alignment Programs

Program	Version	Arguments
ChemAlign	1.0	ssalign(gapdist=4 gapopen=<GOP> gapext=<GEP> subMatA=S^{α} subMatB=S^{β} subMatL=S^{l} subMat=S^{m} ss=<SSE file>)
ClustalW	2.06	<defaults>
MAFFT	6.240	<defaults>
ProbCons	1.12	<defaults>
PRALINE	-	secondary structure prediction: PSIPRED

done. The exception to this is PRALINE, in which the secondary structure prediction program PSIPRED is used.

To measure the statistical significance of the differences in accuracies between alignment algorithms, we perform a Friedman rank test (Friedman, 1937) with both the reference sum of pairs and PPD accuracy scores. This test is more conservative than the Wilcoxon test, which has also been used to determine statistical significance in other alignment studies (Edgar, 2004b).

4.4.2 Reference Sum of Pairs Score

Probably the most commonly applied metric for MSA algorithms is the reference sum of pairs score. It reports the percentage of positions in a calculated alignment that match the same character in a reference alignment. Let $s_1, \ldots s_n$ be sequences of a calculated alignment, each of length l. Let $r_1, \ldots r_n$ be sequences of a reference alignment, each of length p. Let $q = min(l, p)$.

$$\text{reference sum of pairs score} = \frac{1}{nq} \sum_{i}^{n} \sum_{k}^{q} \delta(s_i(k), r_i(k)) \tag{4.4}$$

$$\delta(x, y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases} \tag{4.5}$$

This metric is generally preferred to the self sum of pairs score (Carrillo and Lipman, 1988) (the percentage of amino acids in each column that match for all pairs of sequences within a single data set) since it evaluates how close an alignment is to the "gold standard" alignment.

ChemAlign achieves significantly better reference sum of pairs scores than the other methods tested. Figure 4.6 reports the scores for both the Large and Midnight Zone data sets. The difference in scores are statistically significant according to the Friedman rank test, with p-values of 3.7×10^{-6} and 4.2×10^{-6} for the Large and Midnight Zone classes. ChemAlign performs between 48.7–91.1% and 44.8–80.4% better on average for the Large and Midnight Zone data sets, and as much as 499.3% better on a single data set (lrr_ref6_centre). To put this in perspective, ChemAlign is able to correctly align between 11,685–16,012 and 2,380–2,964 more positions in the Large and Midnight Zone data sets respectively than the other methods.

4.4.3 Physicochemical Property Difference (PPD) Score

In addition to using the reference sum of pairs score, we also look at the normalized difference in physicochemical properties values, or the PPD score. The score is calculated as follows:

$$\text{PPD score} = \frac{1}{nq} \sum_{i}^{n} \sum_{k}^{q} D^{p}_{s_i(k), r_i(k)} \tag{4.6}$$

where D^p is the normalized difference matrix of a physicochemical property p (see section 4.3.1). PPD scores range from -1.0 to 1.0. A score of -1.0 means that all of the amino acids in the calculated alignment are in the same respective position as the amino acid that is the most dissimilar in terms of the physicochemical property. For example, using the Effective Partition Energy, this would mean that all of the sequences consist of only lysines (K) and phenylalanines (F), and that all of the lysines in the calculated alignment match up with phenylalanines in the reference

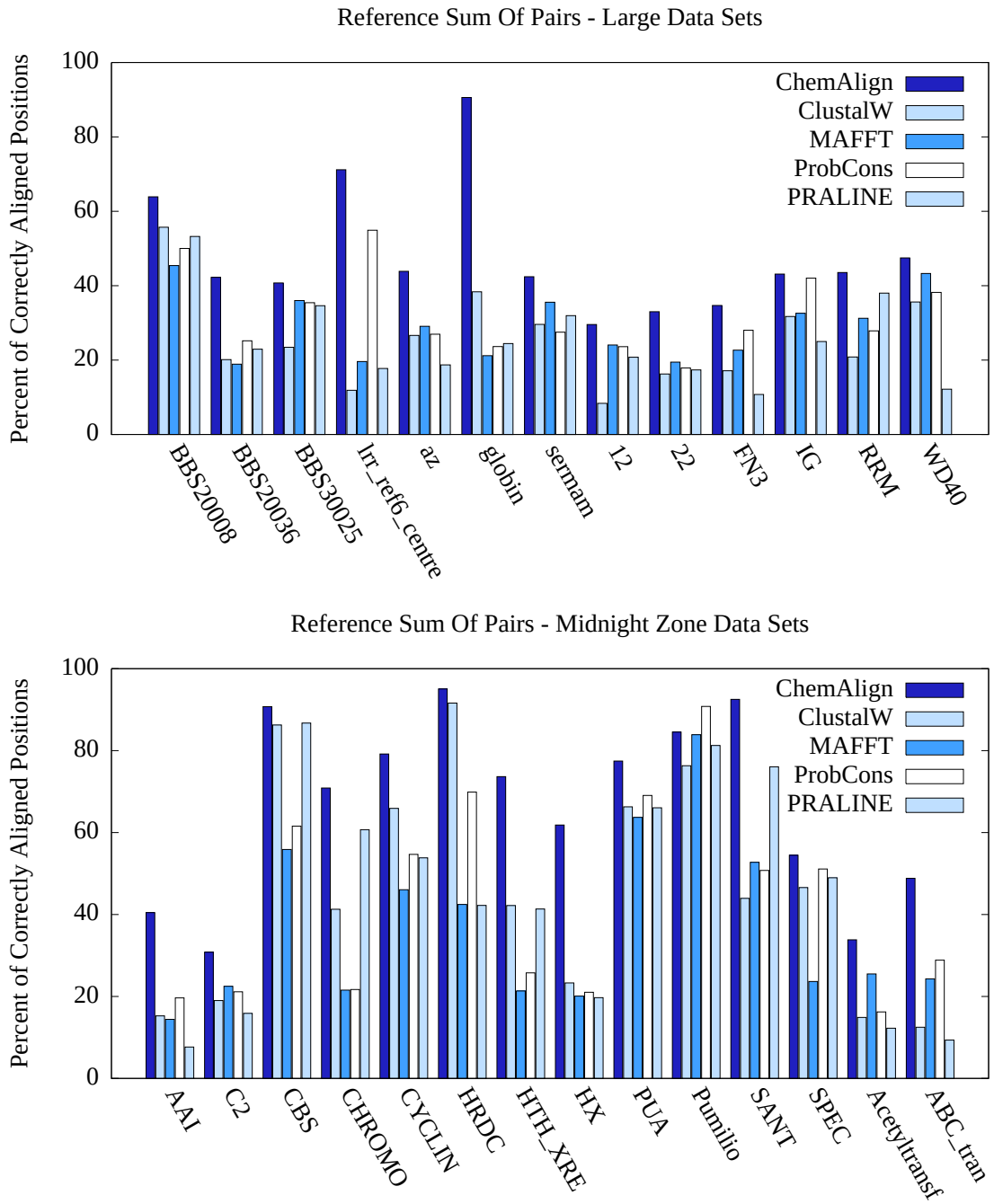


Figure 4.6: Reference Sum of Pairs Scores for the Large and Midnight Zone data sets. This score is probably the most commonly employed accuracy measurement of multiple sequence alignments. ChemAlign achieves the highest scores on all of the data sets except for Pumilio—in which it is the second highest. The differences between it and the other programs are as high as 499.3% (lrr_ref6_centre).

alignment since $D^p(\mathrm{K},\mathrm{F}) = -1.0$. In practice, this is impossible, but serves as an absolute minimum. In general, a negative PPD score means that the average amino acid pairing in an alignment is worse than the average difference in the physicochemical property values. A score of 1.0 means the calculated alignment is the same as the reference alignment. This score takes a step beyond sequence similarity and measures characteristics of the amino acids. It can be adapted to account for multiple physicochemical properties by incorporating them into D, with weights for each one.

We also evaluate the alignments generated from ChemAlign, ClustalW, MAFFT, ProbCons and PRALINE using the PPD score (with the physicochemical property Effective Partition Energy) for the Large and Midnight Zone data sets. ChemAlign achieves the highest average PPD score for each class, with the Large data sets proving more difficult (see Figure 4.7). The differences in scores are statistically significant according to the Friedman rank test, with p-values of 1.8×10^{-6} and 1.3×10^{-6} for the Large and Midnight Zone classes. ChemAlign performs between 50.8–64.0% and 26.1–76.1% better on average for the Large and Midnight Zone classes, and as high as 1,049.6% better for a single data set (Acetyltransf). Additionally, ChemAlign earns the best PPD score for each of the Large data sets, and for all but the Pumilio data set in the Midnight Zone class. While the Effective Partition Energy generally captures the forces of mutation here, researchers can also use the PPD score to evaluate additional properties (i.e., polarity or volume) affecting their alignments.

4.4.4 Globin Domain Alignment

The globin data set, used here for the purpose of example, was taken from the HOMSTRAD database, and is composed of 41 protein sequences, all of which have representative crystal structures in the PDB. Seven different categories of globin proteins are represented:

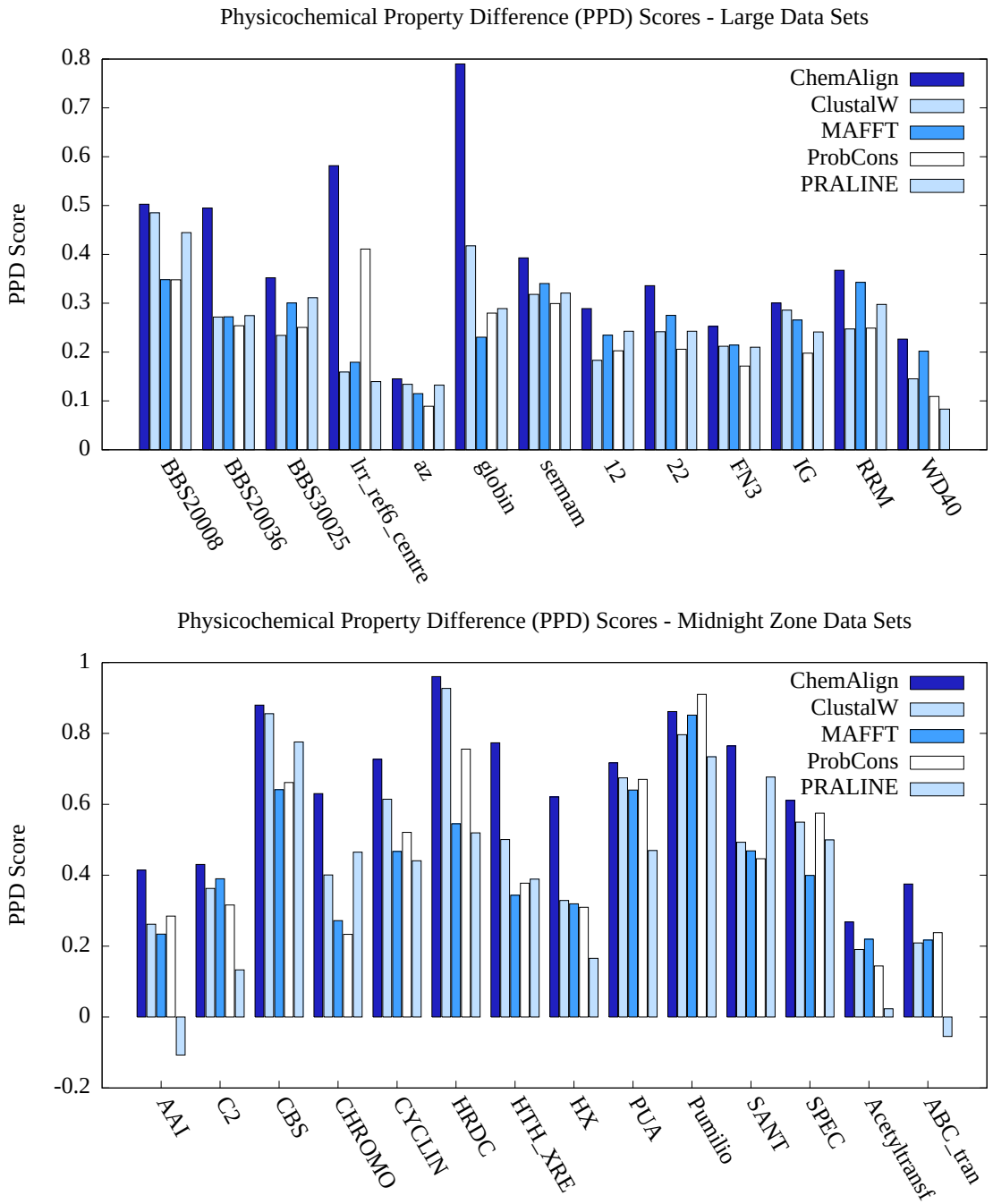


Figure 4.7: Physicochemical Property Difference Scores for the Large and Midnight Zone data sets. ChemAlign achieves the highest scores on all of the data sets except for Pumilio—in which it is the second highest. The differences between it and the other programs are as high as 1,049.6% (Acetyltransf).

- Two plant leghemoglobins (1BIN and 1LH1)
- Seven invertebrate hemoglobins (1ECD, 1HLB, 1HLM, 1ITH, 1MBA, 2HBG, and 3SDH)
- A curiously divergent lamprey hemoglobin (2LHB)
- Seven vertebrate myoglobins (1EMY, 1LHT, 1MBS, 1MYT, 1PMB, 1YMB, and 2MM1)
- Eight vertebrate α-globins (1HDA, 1HDS, 1OUT, 1PBX, 1SPB, 2HHB, 2MHB, and 2PGH)
- Nine vertebrate β-globins (1FDH, 1HBH, 1HDA, 1HDS, 1OUT, 1SPG, 2HHB, 2MHB, and 2PGH)
- Seven globins that are derived copies or adapted to an extreme habitat condition (1A4FA, 1A4FB, 1A6M, 1A9WE, 1CG5A, 1CG5B, and 1HBRA)

Such protein diversity, in terms of primary and secondary structure, as well as overall function, makes accurate alignment notoriously difficult.

As mentioned in the introduction, the globin data set has a low percent identity of 25.9%, making it difficult for current methods to correctly align. ChemAlign is able to produce an alignment with 90.6% of the positions correct, while MAFFT only achieves 21.2% of the characters correct (ClustalW: 38.4%, ProbCons: 23.6% and PRALINE: 24.4%). In terms of percentages, ChemAlign is between 135.9–328.8% better (3,727–4,951 more positions) than the other methods. ChemAlign earns a PPD score of 0.79, which is between 76.2–242.6% better than the other methods. These scores reflect that ChemAlign produces alignments with columns of higher Effective Partition Energy similarity than the other algorithms. This is a characteristic of biologically relevant alignments.

The first globin protein from each of the seven categories listed above are used to illustrate the quality of alignments produced by ChemAlign. Figure 4.8 shows the ChemAlign, ClustalW and PRALINE alignments of the first six α-helices of these

ChemAlign

SS	α_1	α_2	α_3	α_4	α_5	α_6
1BINA	EKQDALVSSSFEAFK	IPQYSVVFYTSILEK	PAAKDL	dp---	PKLTGHAEKLFALVRDSAG	VVADaal---GSVHAQ
1ECD	ADQISTVQAS-fdkv	---DPVGILYAVFKA	PSIMAK	LESIK	APFETHANRIVGFFSKIIG	i--EADVNTFVASHKP
2LHB	AAEKTKIRSAWAPVY	YETSGVDILVKFFTS	PAAQEF	ADELK	ADVRWHAERIINAVDDAVA	m--SMKLRNLSGKHAK
1EMY	DGEWELVLKTWGKVE	IPGHGETVFVRLFTG	PETLEK	EGEMK	EDLKKQGVTVLTALGGILK	h--EAEIQPLAQSHAT
1HDAA	AADKGNVKAAWGKVG	AAEYGAEALERMFLS	PTTKTY	-----	AQVKGHGAKVAAALTKAVE	l--PGALSELSDLHAH
1FDHG	EEDKATITSLWGKV-	VEDAGGETLGRLLVV	PWTQRF	ASAIM	PKVKAHGKKVLTSLGDAIK	l--KGTFAQLSELHCD
1A4FA	AADKTNVKGVFSKIS	AEEYGAETLERMFTA	PQTKTY	-----	AQIKAHGKKVVAALVEAVN	i--AGALSKLSDLHAQ
cons %	434522623284652	213283458284523	927333	21222	3255597348425541444	4	2325218524*52

ClustalW

SS	α_1	α_2	α_3	α_4	α_5	α_6
1BINA	EKQDALVSSSFEAF	ipqysv--VFYTSILEK	PAAKDL	-vdpt	-kltghaeklfalvrdsag	TVVADaal---GSVHA
1ECD	ADQISTVQASFDKV	-dpvgi----LYAVFKA	PSIMAK	lesik	apfethanrivgffskiig	ieadvntfv---ASHK
2LHB	AAEKTKIRSAWAPV	yetsgv--DILVKFFTS	PAAQEF	adelk	advrwhaeriinavddava	msmklrnLs--gkhak
1EMY	DGEWELVLKTWGKV	ipghge--TVFVRLFTG	PETLEK	egemk	edlkkqgvtvltalggilk	heaeiqpla--qshat
1HDAA	AADKGNVKAAWGKV	aaeyga--EALERMFLS	PTTKTY	-dlsh	aqvkghgakvaaaltkave	lpgalsels--dlhah
1FDHG	EEDKATITSLWGKV	vedagg--ETLGRLLVV	PWTQRF	asaim	pkvkahgkkvltslgdaik	lkgtfaqls--elhcd
1A4FA	AADKTNVKGVFSKI	aeeyga--ETLERMFTA	PQTKTY	-dlqh	aqikahgkkvvaalveavn	iagalskls--dlhaq
cons %	43452262328466	223283	458284523	927333	14222	3255597348425541444	423243185	24742

PRALINE

SS	α_1	α_2	α_3	α_4	α_5	α_6
1BINA	EKQDALVSSSFEAF	anipqysvvfytsilek	paakdl	vd--p	pkltghaeklfalvrdsag	tvvadaalgsv---ha------
1ECD	ADQISTVQASFDK-	---kgdpvgilyavfka	psimak	lesik	apfethanrivgffskiig	ieadvntfvas---hk------
2LHB	AAEKTKIRSAWAPV	Y--etsgvdilvkffTs	paaqef	adelk	advrwhaeriinavddava	tekmsmklrnlsgkha-----k
1EMY	DGEWELVLKTWGKV	adipghgetvfvrlftg	petlek	egemk	edlkkqgvtvltalggilk	heaeiqplaqs---ha-----t
1HDAA	AADKGNVKAAWGKV	ghaaeygaealermfls	pttkty	---sh	aqvkghgakvaaaltkave	lpgalselsdl---ha-----h
1FDHG	EEDKATITSLWGKV	V--edaggetlgrllvv	pwtqrf	asaim	pkvkahgkkvltslgdaik	lkgtfaqLsel---hc-----d
1A4FA	AADKTNVKGVFSKI	ghaeeygaetlermfta	pqtkty	---qh	aqikahgkkvvaalveavn	iagalsklsdl---ha-----q
cons %	43452262328465	21213283458284523	927333	22222	3255597348425541444	42324318524	95	2

Figure 4.8: ChemAlign, ClustalW and PRALINE alignments of the globin data set. The secondary structure and conservation percentage (divided by ten) are shown above and beneath each alignment. Uppercase letters denote amino acids that match the reference data set. ChemAlign is able to align the vast majority (90.6%) of the positions correctly. ClustalW's accuracy is 38.4%. PRALINE correctly aligns only the first of the eight α-helices (24.4%). The columns with colored boxes correspond with the conserved regions shown in Figure 4.1. ChemAlign is able to find both of them, while ClustalW and PRALINE only find the first one. Due to space constraints, only the first six α-helices are shown. Additionally, only seven of the 41 species are included.

sequences. ChemAlign is able to correctly align the vast majority of the amino acids throughout the data set. ClustalW aligns the first, part of the second and the third α-helices correctly. PRALINE correctly aligns only the first of the eight α-helices. Highlighted on the alignments are the most conserved regions (using a sliding window of size three). ChemAlign is able to find both regions, while ClustalW and PRALINE only find the first one. These regions correspond to the highlighted regions on the

hemoglobin protein in Figure 4.1. The positions of these regions on the protein is a potential drug docking site. Alignment methods that do not incorporate physicochemical properties and secondary structure information can obfuscate discovery of such regions.

4.5 Conclusion

Multiple sequence alignments are the foundation for several bioinformatics research areas. For example, identifying genes for drug development relies on an accurate alignment of sequences. Current methods struggle to accurately align data sets with low percent identity. ChemAlign is a new algorithm that addresses these problems by using a physicochemical property to produce biologically relevant MSAs. It also incorporates SSEs to overcome limitations employed by traditional approaches that use the "'average' site in the 'average' protein" (Thorne *et al.*, 1996). Leveraging this additional information, it is able to find more potential drug docking sites than other algorithms (see Figures 4.1 and 4.8). Furthermore, ChemAlign achieves higher accuracies for data sets with very low percent identity. It also obtains higher reference sum of pairs accuracies for the largest data sets in the BAliBASE, HOMSTRAD, OXBench and SMART databases. Additionally, we introduce the Physicochemical Property Difference (PPD) score. This score measures the average difference in values for a physicochemical property for all pairs of amino acids in an alignment. It takes a step beyond sequence similarity and measures characteristics of the amino acids. ChemAlign achieves the highest PPD scores for both classes of data sets.

ChemAlign is implemented in the open source package PSODA. PSODA is free, and available for Mac OS X, Linux, Windows and other operating systems.

4.6 Future Work

There are several directions that we are working on in regards to ChemAlign. First, we are extending the difference in physicochemical properties matrix, D, to handle multiple properties with weights. Additionally, we are looking at increasing the specificity of the substitution matrices by using different physicochemical properties for each of the secondary structures.

Chapter 5

Relative Importance of Physicochemical Properties of Amino Acids for Multiple Sequence Alignment

Hyrum D. Carroll, Kenneth A. Sundberg, Mark J. Clement, Quinn O. Snell and David A. McClellan

Submitted to Nucleic Acids Research

Abstract

ChemAlign is a multiple sequence alignment algorithm that uses a single physicochemical property (e.g., residue volume, polarity, hydropathy) and secondary structure elements (α-helices, β-strands or loops) to create biologically meaningful alignments. In this paper, alignment accuracies are dramatically improved by ranking physicochemical properties for each of the secondary structures. To establish the orderings, artificial neural networks are trained to predict protein secondary structures found in the PDB database. The orderings are based on the Q_3 scores for the default case, and the correlation coefficients for each of the secondary structures. The most important properties are used to calculate substitution matrices. Using the matrices along with an improved version of ChemAlign yields alignments with higher accuracies.

Key words: **multiple sequence alignment, physicochemical properties, protein secondary structure prediction**

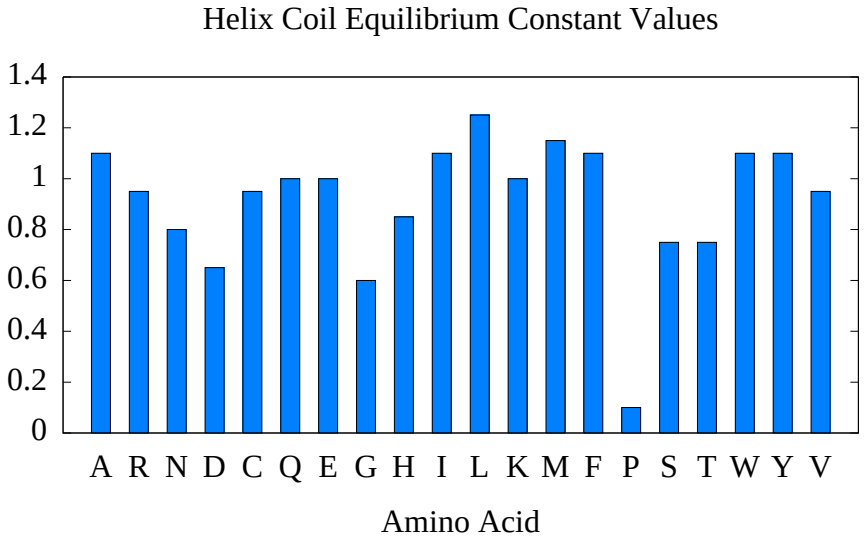


Figure 5.1: Values for the physicochemical property Helix coil equilibrium constant (Ptitsyn and Finkelstein, 1983).

5.1 Introduction

Physicochemical properties of the amino acids are important to a number of areas in bioinformatics. The AAindex database (Kawashima *et al.*, 2008) has over 500 physicochemical properties, such as volume (Bigelow, 1967), polarity (Grantham, 1974) and hydropathy (Kyte and Doolittle, 1982). Using a single property to generate multiple sequence alignments (MSAs) has significantly increased their accuracy compared to other methods (Carroll *et al.*, 2008b). A natural extension of this approach is to combine multiple properties to further improve accuracy. A hurdle to accomplishing this is knowing which properties to use. Testing and using all of these properties in an analysis is usually time consuming and is not typically feasible. Instead, what is needed is an ordering of these properties so that a subset can be used that best summarizes and aggregates the net effect of several properties. An ordering is calculated and used to produce MSAs with dramatically improved accuracies—15.8% better compared to the original version of ChemAlign, and 72.2–121.3% better than the other methods tested.

A physicochemical property is an attribute of the amino acids that is numerically quantified as a rational number (see Figure 5.1 for an example). These properties are important to a number of areas in bioinformatics research and analysis (Goldman and Yang, 1994; Grantham, 1974; Lim, 1974; Xia and Li, 1998). In particular to the area of multiple sequence alignments, researchers have used physicochemical properties in three different ways. First, they have correlated and approximated the most commonly used substitution matrices with a few physicochemical properties (Méndez *et al.*, 2008; Pokarowski *et al.*, 2007; Rudnicki and Komorowski, 2005). Second, others have used the properties to calculate pairwise and multiple sequence alignments (Carroll *et al.*, 2008b; Gonnet and Lisacek, 2002; Gupta *et al.*, 2005). Finally, other studies have focused on identifying characteristics of the calculated alignment (Afonnikov and Kolchanov, 2004; Thorvaldsen *et al.*, 2005; Woolley *et al.*, 2003; Wrabl and Grishin, 2005). Using multiple physicochemical properties for MSA builds upon the successes of these researchers.

Accuracies of MSAs can be dramatically improved by using an ordering of the physicochemical properties. To quantify the influence of each property, an artificial neural network (ANN) is developed to predict the protein secondary structure of an amino acid using that property. Instead of amino acids as inputs to the ANNs, they are encoded by the values of a physicochemical property. Using ANNs to determine the ordering of the properties is based on the idea that, with the input data constant, there is a correlation between the encoding of that data and the accuracy of the network. In more colloquial terms, this idea is known in the negative form as "garbage in, garbage out". There are two main characteristics of amino acids supporting this idea. First, secondary structure is more conserved than the primary sequence. This statement has been verified through a number of different experiments and reports (Gibrat *et al.*, 1996; Rost, 1999; Sander and Schneider, 1991). Second, the secondary structure is a reliable attribute of an amino acid. In other words, it can be determin-

istically assigned given the tertiary structure (Kabsch and Sander, 1983). Leveraging these two advantages provides a vehicle to determine the relative importance of the properties.

In the remainder of this paper, details of constructing and training the ANNs are given. Next, the metrics for determining the orderings are specified, followed by our method to construct the substitution matrices for alignment. The metrics for the accuracy of a MSA are also presented. This is followed by the results of the secondary structure predictions—the ordering of the physicochemical properties—and the accuracies of the MSAs. We close this paper with some concluding remarks.

5.2 Methods

This section details the six steps to calculate multiple sequence alignments using four orderings of the physicochemical properties (see Figure 5.2):

1. Secondary Structure Prediction
2. Q_3 and Correlation Coefficients Orderings
3. Normalized Physicochemical Property Difference Matrices
4. Observed Amino Acid Exchanges
5. Physicochemical Property Substitution Matrices
6. Weighted Physicochemical Property Difference Matrices
7. Multiple Sequence Alignments

Each step is addressed in a subsection.

5.2.1 Secondary Structure Prediction

With the goal of calculating an ordering for all of the physicochemical properties, ANNs are employed to predict the secondary structure of protein sequences. ANNs are machine learning algorithms inspired by the human nervous system. In an ANN, nodes (representing neurons) are connected together, and weights are assigned to

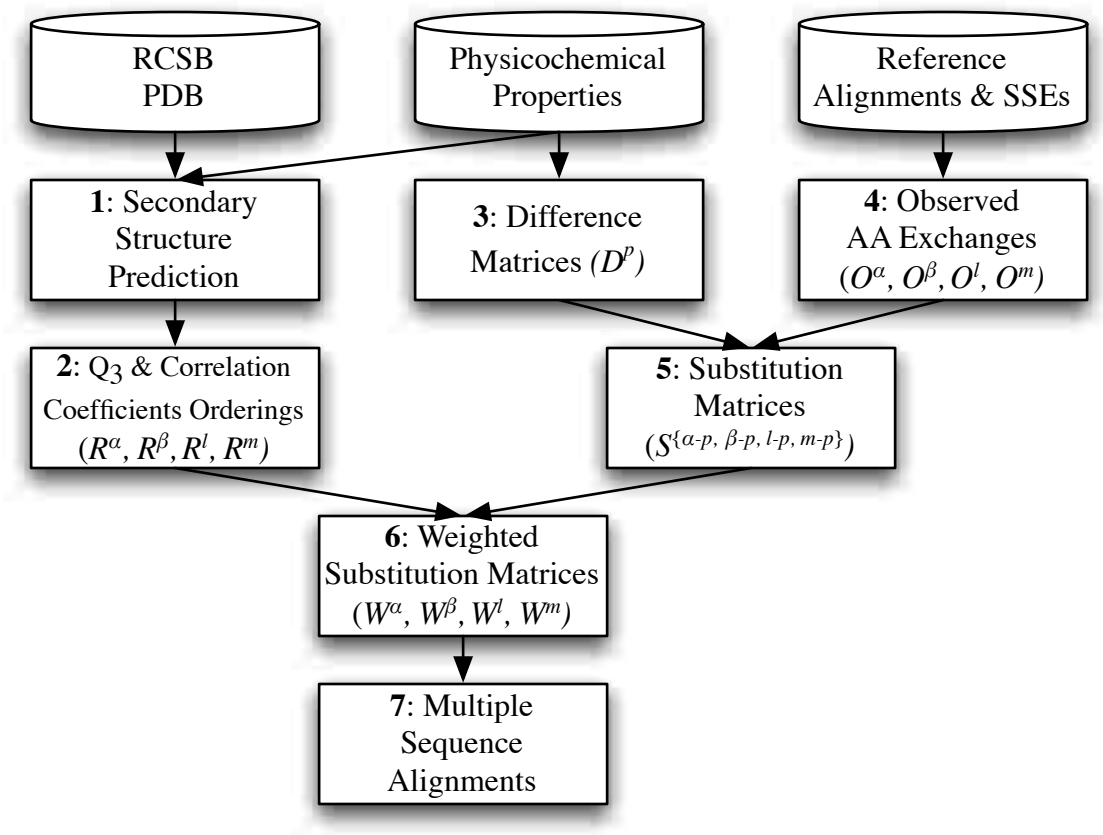
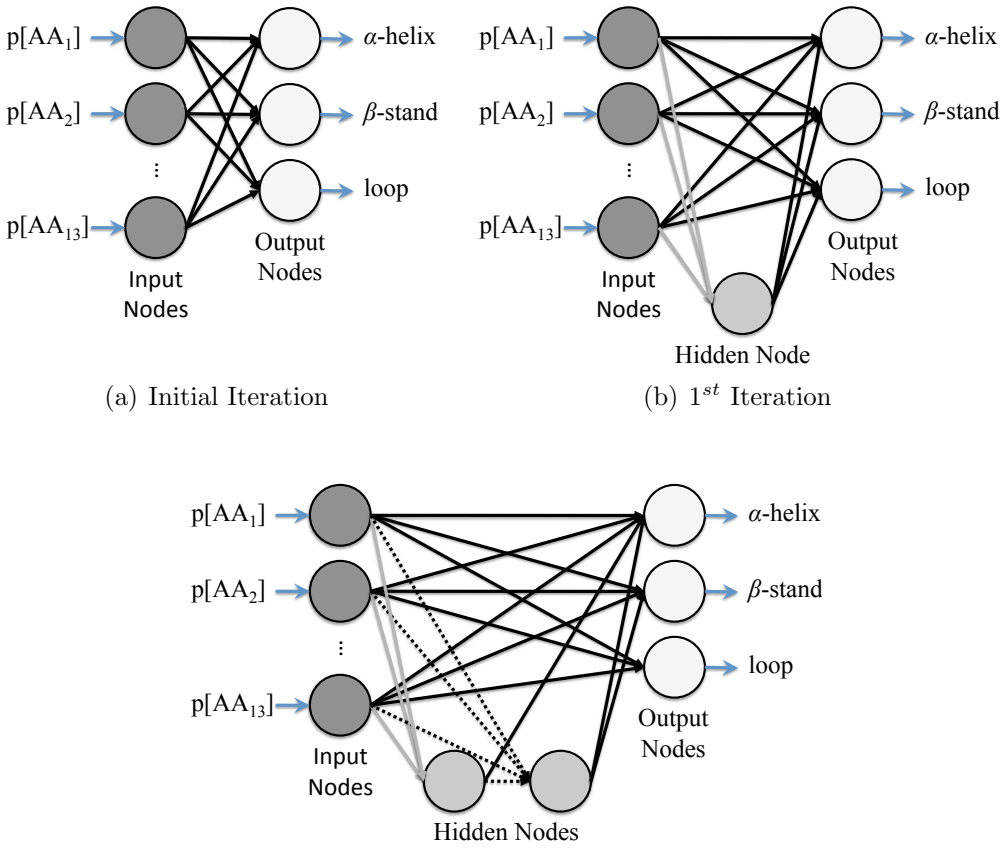


Figure 5.2: The process to calculate multiple sequences alignments using the four orderings of physicochemical properties. Numbers subsections within section 5.2 that detail the given step. Weights and matrix variables are shown in parenthesis.

the edges between the nodes. These weights are adjusted, or learned, by looking at training data. ANNs typically have one or more layers of *hidden nodes* in between the input and output node layers. After a network is trained, the weights are fixed and the network is used to calculate predicted output for new data.

For the secondary structure predictions, cascade-correlation artificial neural networks (Fahlman and Lebiere, 1990) (CCANNs) are used. CCANNS typically learn the underlying function from the training data in less time than traditional ANNs. Additionally, they learn the size and topology of the network instead of assuming a fixed architecture (see Figure 5.3). A CCANN begins with a minimal network topology of just input and output nodes. The weights are set to maximize the accuracy of the output for the training data. In the second iteration, a hidden node is added,



(a) Initial Iteration

(b) 1^{st} Iteration

(c) 2^{nd} Iteration

Figure 5.3: Graphical representation of the first three training iterations of a cascade-correlation artificial neural network. The inputs are a window of the physicochemical property values for thirteen amino acids. The output is either an α-helix, β-strand or a loop. Solid dark edges between nodes are repeatedly trained. Dashed edges are trained only in the current iteration. Gray lines are fixed.

and all of the weights are adjusted. At the end of this iteration, the weights to the hidden node are fixed. For the third iteration, an additional hidden node is added, and all but the fixed weights are adjusted. This process continues until the error is reduced to a specified amount or for a specified number of iterations or time.

The ANNs are trained using the sequences in the RCSB Protein Data Bank (PDB) (Berman *et al.*, 2000) that are annotated with secondary structures. A testing set of 96 sequences from the CASP7 (Trapane and Lattman, 2007) competition are

removed. For each amino acid in either the training or the test set, a *window* of amino acid positions is included, adding the closest neighbors on each side (if present). A window size of thirteen is employed, as is commonly done for secondary structure predictions (e.g., for PHD (Rost and Sander, 1993)). After removing ambiguous instances from the training set, additional instances are randomly filtered out so that there is a even distribution of each of the secondary structures. In all, 14,220 training instances are used.

Crucial to the success of an ANN is the encoding of data for input. For each of the training and test instances, instead of using the amino acid, the physicochemical property value is used. Typically, predictors use an orthogonal encoding of the amino acids, requiring 20 inputs for each position. Using the physicochemical properties reduces the number of inputs to be learned during training since they are already rational numbers. The ANNs are trained to predict one of the three secondary structure elements (SSEs) (α-helix, β-strand or loop).

5.2.2 Q_3 and Correlation Coefficients Orderings

To calculate the ordering of the physicochemical properties for the default case, the Q_3 score is used to evaluate the predictions on the CASP7 test sequences. In general, the Q_3 score is the percentage of predictions that are correct. Let i and j be the predicted and actual SSE respectively. Let B_{ij} be the number of occurrences of i and j. The score is calculated as in Equation 5.1.

$$Q_3 = 100 \frac{\sum_{i=1}^{3} B_{i,i}}{\sum_{i=1}^{3} \sum_{j=1}^{3} B_{i,j}} \tag{5.1}$$

The scores are sorted by their rank and assigned to R^m (the m stands for *mismatch*, since these rankings are used when the SSEs do not match) (see Figure 5.2).

While the Q_3 score aggregates the predictions of α-helices, β-strands and loops, the Pearson correlation coefficients is used to order the physicochemical properties for each SSE. For secondary structure predictions, it is often referred to as the Matthews correlation coefficient (Matthews, 1975) since he was the first to apply it to this area. It is calculated as in Equation 5.2 where T_e^P, T_e^N, F_e^P and F_e^N are the true positive, true negative, false positive and false negative values for the SSE e.

$$C_e = \frac{T_e^P T_e^N - F_e^P F_e^N}{\sqrt{(T_e^P + F_e^N)\,(T_e^P + F_e^P)\,(T_e^N + F_e^P)\,(T_e^N + F_e^N)}} \tag{5.2}$$

This metric measures how strongly the average prediction correlates with the given assignment. Again, the orderings are sorted by rank and assigned to R^α, R^β and R^l respectively.

In addition to the applicability of these ordering to MSAs, they can also be used in many of the other areas of bioinformatics that leverage physicochemical properties.

5.2.3 Physicochemical Property Difference Matrices

A normalized difference matrix, D^p, is calculated for each physicochemical property p, as in Equation 5.3.

$$D^p_{i,j} = 1 - \frac{2\,|p[i] - p[j]|}{argmax_x(p[x]) - argmin_y(p[y])} \tag{5.3}$$

Here, $p[i]$ is the value for amino acid i for the physicochemical property p. The values of D^p range from -1.0 for the most dissimilar pair of amino acids to 1.0 for identical amino acids. As an example, $D^{\text{Helix coil equilibrium constant}}_{\text{D,G}} = 0.91$, since aspartic acid (D) and glycine (G) are very similar in terms of the property Helix coil equilibrium constant (see Figure 5.1).

5.2.4 Observed Amino Acid Exchanges

To obtain the observed amino acid exchange counts, a reference database of alignments is built with their secondary structures. The OXBench database (Raghava *et al.*, 2003) is combined with the respective secondary structures from the PDB. To avoid adding noise, only those sequences in OXBench that have an exact match in the PDB are retained. The number of each set of amino acid pairs (for each column in the alignment) is tabulated, according to their SSEs (i.e., both α-helices, both β-strands, both loops or mismatch) producing four matrices of observed amino acid exchanges: $O^{\alpha}, O^{\beta}, O^{l}$, and O^{m}.

5.2.5 Physicochemical Property Substitution Matrices

To calculate the substitution matrices, the D^{p} is multiplied element-wise with each of the $O^{\alpha}, O^{\beta}, O^{l}$ and O^{m} matrices to produce $M^{\alpha\text{-}p}, M^{\beta\text{-}p}, M^{l\text{-}p}$, and $M^{m\text{-}p}$, for α-helices, β-strands, loops and the mismatch case (see Figure 5.2). Combining the physicochemical properties difference matrix and the $O^{\alpha}, O^{\beta}, O^{l}, O^{m}$ matrices, merges a theoretical and data-driven approach. Furthermore, the combined matrices achieve more accurate MSAs than using either component individually. Next, the log-odds ratios of $M^{\alpha\text{-}p}, M^{\beta\text{-}p}, M^{l\text{-}p}$, and $M^{m\text{-}p}$ are calculated (see Equation 5.4).

$$S_{i,j}^{\text{SSE-}p} = \frac{1}{\lambda} log \left(\frac{l_{i,j}^{\text{SSE-}p}}{f_i^{\text{SSE-}p} f_j^{\text{SSE-}p}} \right) \tag{5.4}$$

Here, $l_{i,j}^{\text{SSE-}p}$ is the likelihood that amino acids i and j appear aligned in the database, and are both in the same secondary structure, SSE. Also, $f_i^{\text{SSE-}p}$ is the background frequency of amino acid i, for the same criteria. Both $l_{i,j}^{\text{SSE-}p}$ and $f_i^{\text{SSE-}p}$ are derived from the $M^{\text{SSE-}p}$ matrix.

Taking the log-odds ratios aids the alignment process to identify amino acids that are less common, and therefore are more likely to be aligned together. This prac-

tice is employed to calculate the most commonly used current substitution matrices (Dayhoff *et al.*, 1978; Gonnet *et al.*, 1992; Henikoff and Henikoff, 1992).

5.2.6 Weighted Substitution Matrices

Weights, w_p, are assigned for each physicochemical property p, for each of the four orderings, R^α, R^β, R^l and R^m. The weight is determined by the exponentially decaying function: $\left(\frac{1}{2}\right)^n$, with n being the rank of p. This function combines multiple properties and favors the best performing ones. It performs better than other analyzed functions (i.e., weights proportional to the rank, etc.) (data not shown). Due to the inherent reduction in the values for the weights, only the top ten properties in each ordering are given weights.

Weighted substitution matrices are the summation of the product of the top ten weights and the respective physicochemical property substitution matrix (see Equation 5.5).

$$S_{i,j}^{\text{SSE}} = \sum_{k=1}^{10} w_{p_k} S_{i,j}^{\text{SSE-}p_k} \tag{5.5}$$

Here, p_k is the k^{th} ranked physicochemical property. This step is repeated for each of the orderings, producing S^α, S^β, S^l, and S^m (see Figure 5.2).

5.2.7 Multiple Sequence Alignments

ChemAlign version 1.4 (ChemAlign-weights) uses the S^α, S^β, S^l, and S^m substitution matrices to produce multiple sequence alignments of thirteen of the largest data sets in the BAliBASE (Thompson *et al.*, 2005), HOMSTRAD (Mizuguchi *et al.*, 1998), OXBench and SMART (Letunic *et al.*, 2004) databases. These data sets range from having 17–115 taxa. The average length of the sequences in each data set ranges from 23.6–232.7 amino acids. Also, the percent identity is very low with all of the data sets ranging from 12.5–27.7%, except for one that has 35.2%. It incorporates

physicochemical properties and secondary structures to align multiple sequences of amino acids. It uses a different substitution matrix when the secondary structures of the sequences match, and a default matrix when they do not. Additionally, affine gap penalties are used for each of the secondary structures and the default case.

The reference sum of pairs score is used to evaluate the accuracy of the alignments. It reports the percentage of positions in a calculated alignment that match the same character in a reference alignment. Let $s_1, \ldots s_n$ be sequences of a calculated alignment, each of length l. Let $r_1, \ldots r_n$ be sequences of a reference alignment, each of length p. Let $q = min(l, p)$.

$$\text{reference sum of pairs score} = \frac{1}{nq} \sum_{i}^{n} \sum_{k}^{q} \delta(s_i(k), r_i(k)) \tag{5.6}$$

$$\delta(x, y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases} \tag{5.7}$$

This metric is generally preferred to the self sum of pairs score (Carrillo and Lipman, 1988). The self sum of pairs measures the percentage of amino acids in each column that match for all pairs of sequences within a single data set. The reference sum of pairs score is favored because it evaluates how close an alignment is to the "gold standard" alignment.

In addition to the the reference sum of pairs scores, the Physicochemical Properties Difference (PPD) score (Carroll *et al.*, 2008b) is calculated. This score is extended to measure the average difference in values for multiple physicochemical properties for all pairs of amino acids in an alignment using the weighted physicochemical property difference matrices for mismatches:

$$\text{PPD score} = \frac{1}{nq} \sum_{i}^{n} \sum_{k}^{q} \sum_{k=1}^{10} w_{p_k} D^{p_k}_{s_i(k), r_i(k)} \tag{5.8}$$

It takes a step beyond sequence similarity and measures characteristics of the amino acids, making it more biologically relevant. PPD scores range from -1.0 to 1.0, with higher scores meaning that there is more similarity in the alignment in terms of the physicochemical properties.

5.3 Results

5.3.1 Secondary Structure Predictions

To obtain an ordering of physicochemical properties that improves multiple sequence alignments, cascade-correlation artificial neural networks are employed with amino acid sequences encoded with the values of a physicochemical property. The Q_3 score and the correlation coefficients are used to obtain four orderings of the properties: one for a default case and one for each of the secondary structures. Table 5.1 reports the rank and the property used to build the best ten predictors for each case. Almost all of the best predictors of α-helices include properties that explicitly capture helical properties. The same is true for β-strands and, to a slightly lesser degree, loops. This serves as an informal validation of the prediction process. Some of the properties with the highest Q_3 scores are the same properties that achieve the best correlation coefficients. This is to be expected.

5.3.2 Multiple Sequence Alignment Accuracies

The substitution matrices calculated from the ordered lists of physicochemical properties are used to generate multiple sequence alignments using ChemAlign-weights. The alignments achieve higher reference sum of pairs scores than other algorithms tested, including ChemAlign. Figure 5.4 illustrates these scores for ChemAlign-weights, ChemAlign(Carroll *et al.*, 2008b), ClustalW (Larkin *et al.*, 2007), and PRALINE (Heringa, 1999). ClustalW is included because it is probably the most commonly used

Table 5.1: Physicochemical Properties Of The Best Secondary Structure Predictors

	Rank	Physicochemical Property	AAindex ID
α-helices	1	Free energy in α-helical region	MUNV940102
	2	Weights for α-helix at the window position of -1	QIAN880106
	3	Free energy in α-helical conformation	MUNV940101
	4	Helix coil equilibrium constant	PTIO830101
	5	Normalized positional residue freq. at helix termini C1	AURR980115
	6	α-helix propensity of position 44 in T4 lysozyme	BLAM930101
	7	Normalized positional residue freq. at helix termini C4	AURR980112
	8	Information measure for middle helix	ROBB760103
	9	Side chain angle θ(AAR)	LEVM760103
	10	Normalized positional residue freq. at helix termini C3	AURR980113
β-strands	1	Thermodynamic β-sheet propensity	KIMC930101
	2	Average relative probability of β-sheet	KANM800102
	3	8 Å contact number	NISK800101
	4	β-coil equilibrium constant	PTIO830102
	5	Average surrounding hydrophobicity	MANP780101
	6	Surrounding hydrophobicity in folded form	PONP800101
	7	Normalized frequency of β-sheet with weights	LEVM780102
	8	Free energy in β-strand conformation	MUNV940103
	9	Normalized frequency of β-sheet from CF	PALJ810104
	10	Average relative probability of inner β-sheet	KANM800104
Loops	1	Helix coil equilibrium constant	PTIO830101
	2	α-helix propensity of position 44 in T4 lysozyme	BLAM930101
	3	Information measure for coil	ROBB760112
	4	Weights for coil at the window position of 1	QIAN880134
	5	Smoothed υ steric parameter	FAUJ880102
	6	Thermodynamic β-sheet propensity	KIMC930101
	7	Weights for coil at the window position of -1	QIAN880132
	8	δ G values for the peptides extrapolated to 0 M urea	ONEK900101
	9	Weights for coil at the window position of -2	QIAN880131
	10	Normalized frequency of reverse turn, with weights	LEVM780103
Default	1	Helix coil equilibrium constant	PTIO830101
	2	Thermodynamic β-sheet propensity	KIMC930101
	3	Weights for coil at the window position of 0	QIAN880133
	4	δ G values for the peptides extrapolated to 0 M urea	ONEK900101
	5	α-helix propensity of position 44 in T4 lysozyme	BLAM930101
	6	Zimm Bragg parameters at 20°C	SUEM840101
	7	Information measure for C-terminal helix	ROBB760104
	8	Weights for coil at the window position of -1	QIAN880132
	9	Information measure for middle helix	ROBB760103
	10	Helix formation parameters (δ δ G)	ONEK900102

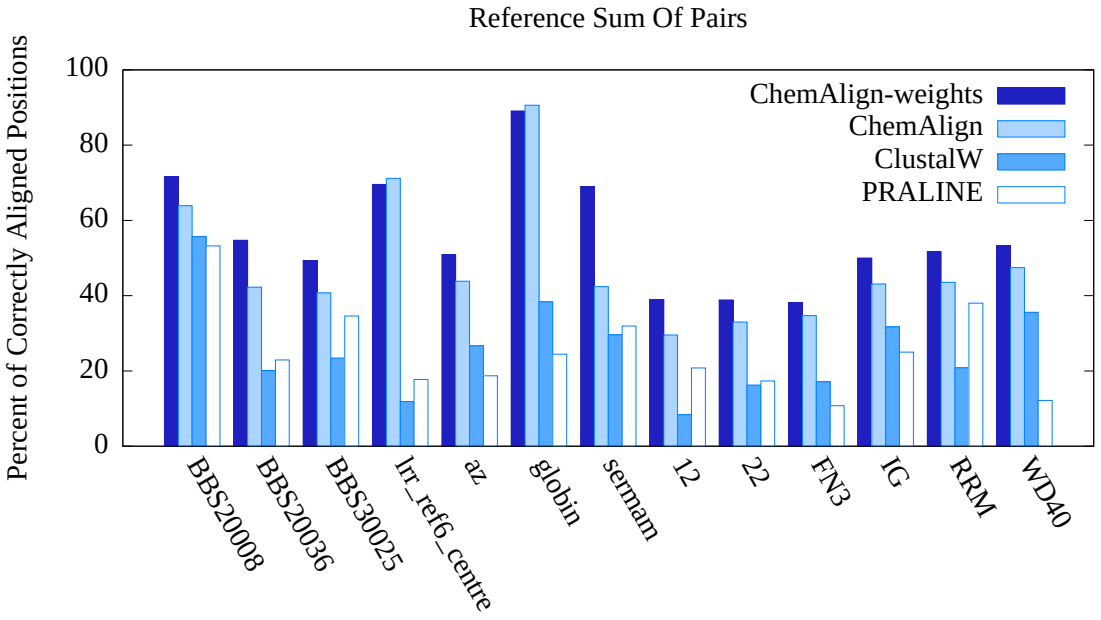


Figure 5.4: Reference Sum of Pairs scores. Combining several physicochemical properties into a different substitution matrix for each of the secondary structure elements and another one for the default case yields more accurate alignment using ChemAlign.

MSA program, and PRALINE because it incorporates SSEs into MSAs. ChemAlign-weights performs between 72.2–121.3% better on average across these data sets than these methods, and 15.8% better than ChemAlign. To help put this in perspective, ChemAlign-weights correctly aligns between 17,588–21,916 more amino acids than the other algorithms for these data sets, and 5,904 more amino acids than ChemAlign. While aligning one additional amino acid correctly can change the conclusions of an analysis using a multiple sequence alignment, certainly 21,000 more amino acids can have a larger impact. The differences in scores are statistically significant according to the Friedman rank test (Friedman, 1937), with a p-value $\ll 0.001$.

ChemAlign-weights performs better than ChemAlign in all but two cases. For the llr_ref6_centre data set, ChemAlign-weights adds an additional column of gaps near the end of the sequences, in one of the later progressive phases. For the globin data set, differences in the alignments are mostly in the mis-alignment of a single

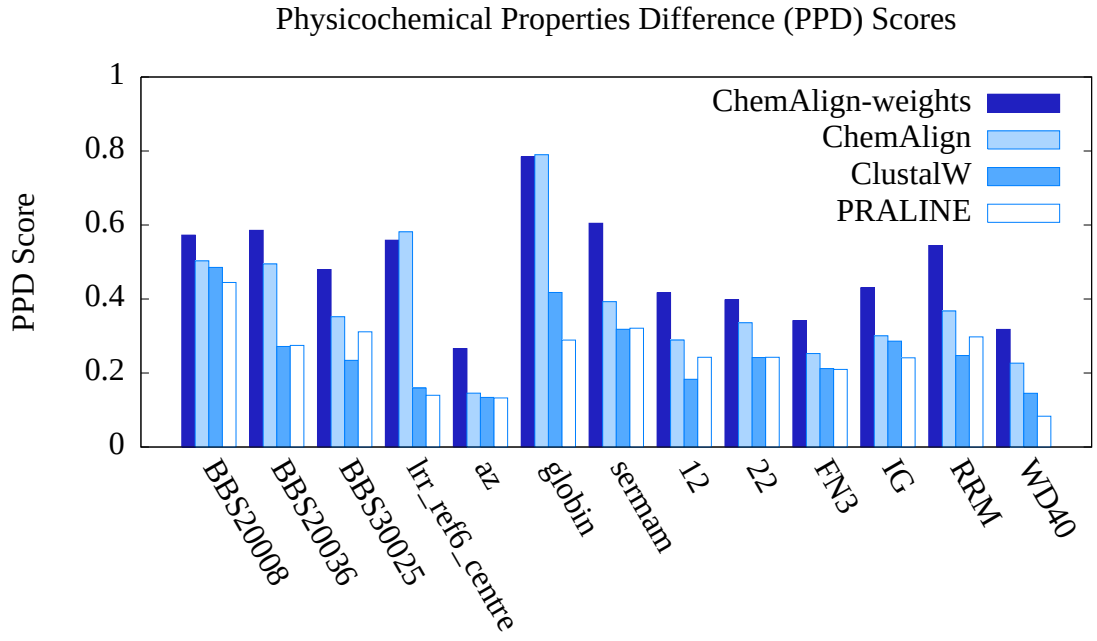


Figure 5.5: PPD scores. Combining several physicochemical properties into a different substitution matrix for each of the secondary structure elements and another one for the default case yields more accurate alignment using ChemAlign.

column in between the first and second α-helices. However, ChemAlign-weights is backwards compatible with ChemAlign, allowing for previous techniques to be used, producing the alignments with the higher accuracies.

The alignments generated from ChemAlign-weights and ChemAlign, ClustalW and PRALINE are also evaluated using the PPD score using the physicochemical property orderings O^m. ChemAlign-weights achieves the highest average PPD score (see Figure 5.5). It earns scores between 88.8–105.3% better on average than the other methods, and 25.2% better than ChemAlign. These score differences are statistically significant (Friedman Rank Test) with a p-value $\ll 0.001$.

5.4 Conclusion

Multiple sequence alignments generated using a single physicochemical property have been shown to be more accurate than existing methods (Carroll *et al.*, 2008b). Here,

multiple properties are combined to substantial improve accuracies. To do so, four orderings of the physicochemical properties are calculated—one for a default case and one for each of the three secondary structures. Each ordering is established by the Q_3 score and the correlation coefficients of artificial neural networks trained to predict protein secondary structures. The results presented here quantify the relative importance of over 500 chemical properties in the AAindex database. Weights are assigned to the properties according to their rank, and four substitution matrices are calculated. These matrices are used by ChemAlign-weights to align thirteen of the largest data sets in the BAliBASE, HOMSTRAD, OXBench and SMART databases. ChemAlign-weights correctly aligns as many as over 21,000 more amino acids than the other methods tested and achieves the highest average accuracy.

Chapter 6

Conclusion

Multiple sequence alignments are the foundation of analysis in bioinformatics. An example application of alignments is finding drug docking sites. Conserved columns are identified from the alignments and the corresponding areas are analyzed on the tertiary structure of a protein. When a suitable candidate is found, drugs can be developed to change the function of that protein.

There are three main deficiencies with calculating biologically relevant MSAs using current approaches:

1. Optimization for sequence similarity
2. Ignoring secondary structure information
3. Static comparison of sequences

First, current MSA algorithms are mostly based on sequence similarity and miss some conserved columns. Without the conserved columns identified, the potential drug docking site is overlooked. Second, secondary structures provides pertinent information to producing biologically accurate alignments. Additionally, different positions on a protein have different exchangabilities due to their location. Current methods use the "'average' site in the 'average' protein" (Thorne *et al.*, 1996) by using a static evaluation function, limiting their sensitivity.

6.1 ChemAlign

ChemAlign is a MSA algorithm that incorporates physicochemical properties and protein secondary structures to overcome the three main challenges of current algorithms. First, it identifies similarity with physicochemical properties. Since the exchangabilility of amino acids is based on their physicochemical properties, ChemAlign uses these properties to evaluate scores for matching amino acid pairs during alignment. It is the first known multiple sequence alignment algorithm to account for physicochemical properties. Second, ChemAlign explicitly incorporates secondary structure elements into its evaluation function. Third, ChemAlign utilizes a dynamic evaluation function, based on the secondary structures of the amino acids to account for different properties having different effects in different secondary structures. These three characteristics allow ChemAlign to produce biologically relevant alignments.

ChemAlign leverages physicochemical properties to achieve higher accuracies than existing MSA algorithms. The initial version uses a single property to earn significantly better reference sum of pairs scores than the other methods tested (see Figure 4.6). The difference in scores are statistically significant according to the Friedman rank test, with p-values $\ll 0.001$. ChemAlign performs as well as 91.1% better on average for the Large data sets, and as much as 499.3% better on a single data set. This means that it correctly aligns 16,012 more positions in the Large data sets.

Additionally, a new MSA metric, the Physicochemical Property Difference (PPD) score, is introduced that captures the average difference between physicochemical properties of a calculated alignment and a reference alignment. ChemAlign achieves the highest average PPD score (see Figure 4.7). The differences in scores are again statistically significant, with p-values $\ll 0.001$. The differences are also relevant with ChemAlign performing as well as 64.0% better on average for the Large class, and as high as 1,049.6% better for a single data set.

Furthermore, an example of using an alignment of globin domains to predict drug docking positions is included (see Figure 1.2). For this data set, ChemAlign is able to detect the two conserved columns, whereas PRALINE and ClustalW only find one of them. This illustrates just one of the many effects derived from having a more accurate alignment.

While the accuracy of the initial version of ChemAlign is impressive, incorporating multiple physicochemical properties yields further improvement in accuracies. With over 500 properties cataloged in the AAindex database, the challenge is which subset of properties to use to best summarize and aggregate the net effect of several properties. In all, 544 properties are analyzed, bringing the total number of possible combinations of ten properties to 5.76×10^{20}. Since a brute force approach is not feasible, for each property, an artificial neural network is trained to predict protein secondary structures for sequences in the PDB. Instead of using a window of amino acids for the input—as is commonly done—the numerical value of the physicochemical property is used. The Q_3 scores and correlation coefficients are sorted, and used to identify the most important properties. The normalized difference matrices for the top ten properties are combined with observed amino acid exchanges to produce four substitution matrices—one for each of the three secondary structures and one for the default case. Using these substitution matrices with an improved version of ChemAlign (v1.4) yields even higher accuracies (see Figure 5.4). ChemAlign v1.4 performs as well as 121.3% better on average than the other methods tested, and 15.8% better than ChemAlign v1.0. This corresponds to correctly aligning 21,916 more amino acids than the other algorithms, and 5,904 more than ChemAlign v1.0. Again the differences in scores are statistically significant, with a p-value $\ll 0.001$. Additionally, ChemAlign earns the highest average PPD score, which is as much as 105.3% better on average than other methods and 25.2% better than chemAlign v1.0

(see Figure 5.5). These score differences are statistically significant with a p-value $\ll 0.001$.

ChemAlign is implemented in the open source package PSODA. PSODA is free and available for Mac OS X, Linux, Windows and other operating systems at `http://dna.cs.byu.edu/psoda`.

6.2 DNA Reference Multiple Sequence Alignment Database

In addition to a novel MSA algorithm, the first known reference alignment database for protein-coding DNA has been published. Several reference amino acid alignment databases exist (e.g., BAliBASE, OXBench, PREFAB, SMART) to evaluate new and existing MSA algorithms. However, these database can only evaluate amino acid sequences.

Included in this work is the first known multiple DNA sequence alignment benchmark databases that are:

1. Comprised of protein-coding portions of DNA
2. Based on biological features such as the tertiary structure of encoded proteins

The databases contain a total of 3,545 alignments, comprising of 68,581 sequences. They are divided into two categories: mdsa_100s and mdsa_all. The mdsa_100s version contains the alignments of the data sets that TBLASTN found 100% sequence identity for each sequence. The mdsa_all version includes all hits with an E-value score above the threshold of 0.001. A primary use of these databases is to benchmark the performance of MSA applications on DNA data sets. The first such case study is included in this work. The results show that the most accurate MSA applications on protein sequences are not the most accurate for protein-coding DNA data sets. This is important information for researchers using these alignment methods for protein-coding DNA.

In conclusion, this dissertation details a multiple sequence alignment algorithm, ChemAlign, that optimizes for different chemical properties in each secondary structure. ChemAlign achieves more accurate alignments than other algorithms. Furthermore, these alignments are more biologically relevant. Additionally, this dissertation introduces a biologically sensitive multiple sequence alignment metric, the Physicochemical Properties Difference score. Finally, the first known reference protein-coding DNA multiple sequence alignment database and accompanying case study of the accuracy of several alignment algorithms using DNA are presented.

References

Afonnikov, D. and Kolchanov, N. (2004). CRASP: a program for analysis of coordinated substitutions in multiple alignments of protein sequences. *Nucleic Acids Research*, **32**(Web Server Issue), W64.

Agrawal, A. and Huang, X. (2008). DNAlignTT: Pairwise DNA alignment with sequence specific transition-transversion ratio. *IEEE International Conference on Electro/Information Technology, 2008(EIT 2008)*, pages 453–455.

Altschul, S. F., Gish, W., Miller, W., Meyers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

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

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2005). GenBank. *Nucleic Acids Research*, **33**, D34–38.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, **28**(1), 235–242.

Bigelow, C. (1967). On the average hydrophobicity of proteins and the relation between it and protein structure. *J Theor Biol*, **16**(2), 187–211.

Carrillo, H. and Lipman, D. (1988). The multiple sequence alignment problem in biology. *SIAM Journal on Applied Mathematics*, **48**(5), 1073–1082.

Carroll, H., Beckstead, W., O'Connor, T., Ebbert, M., Clement, M., Snell, Q., and McClellan, D. (2007). DNA Reference Alignment Benchmarks Based on Tertiary Structure of Encoded Proteins. *Bioinformatics*, **23**(19), 2648–2649.

Carroll, H., Teichert, A., Krein, J., Sundberg, K., Snell, Q., and Clement, M. (2008a). An open source phylogenetic search and alignment package. In press with International Journal of Bioinformatics Research and Applications (IJBRA).

Carroll, H. D., Clement, M. J., Snell, Q. O., and McClellan, D. A. (2008b). ChemAlign: Biologically Relevant Multiple Sequence Alignments Using Physicochemical Properties. Submitted.

Chamala, S., Beckstead, W., Rowe, M., and McClellan, D. (2007). Evolutionary selective pressure on three mitochondrial SNPs is consistent with their influence on metabolic efficiency in Pima Indians. *International Journal of Bioinformatics Research and Applications*, **3**(4), 504–522.

Clark, A., Weiss, K., Nickerson, D., Taylor, S., Buchanan, A., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E., and Sing, C. (1998). Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *American Journal of Human Genetics*, **63**, 595–612.

Corpet, F. *et al.* (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res*, **16**(22), 10881–10890.

Crandall, K. (1996). Multiple interspecies transmissions of human and simian t-cell leukemia/lymphoma virus type i sequences. *Molecular Biology and Evolution*, **13**, 115–131.

Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, **5**(Suppl 3), 345–352.

DeSalle, R. (1995). Molecular approaches to biogeographic analysis of Hawaiian Drosophilidae. *Hawaiian biogeography: evolution on a hot spot archipelago. Smithsonian Institution Press, Washington, DC*, pages 72–89.

Do, C. B., Mahabhashyam, M. S. P., Brudno, M., and Batzoglou, S. (2005). ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Research*, **15**, 330–340.

Doolittle, R. (1994). Convergent evolution: the need to be explicit. *Trends Biochem Sci*, **19**(1), 15–8.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**(19), 755–763.

Edgar, R. C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**(113-131).

Edgar, R. C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**(5), 1792–1797.

Edgar, R. C. and Batzoglou, S. (2006). Multiple sequence alignment. *Current Opinion in Structural Biology*, **16**, 368–373.

Fahlman, S. and Lebiere, C. (1990). The cascade-correlation learning architecture. *Advances in Neural Information Processing Systems*, **2**(2), 524–532.

Feng, D.-F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*, **25**(4), 351–360.

Feng, D.-F. and Doolittle, R. F. (1990). Progressive alignment and phylogenetic tree construction of protein sequences. *Methods Enzymol*, **183**, 375–87.

Fontana, P., Bindewald, E., Toppo, S., Velasco, R., Valle, G., and Tosatto, S. (2005). The SSEA server for protein secondary structure alignment. *Bioinformatics*, **21**(3), 393–395.

Freyhult, E., Bollback, J., and Gardner, P. (2007). Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA. *Genome Research*, **17**(1), 117.

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, **32**(200), 675–701.

Gardner, P. P. and Giegerich, R. (2004). A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**(140).

Gardner, P. P., Wilm, A., and Washietl, S. (2005). A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Research*, **33**(8), 2433–2439.

Gibrat, J., Madej, T., and Bryant, S. (1996). Surprising similarities in structure comparison. *Current Opinion in Structural Biology*, **6**(3), 377–385.

Ginalski, K., Pas, J., Wyrwicz, L., von Grotthuss, M., Bujnicki, J., and Rychlewski, L. (2003). ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acids Research*, **31**(13), 3804–3807.

Ginalski, K., von Grotthuss, M., Grishin, N., and Rychlewski, L. (2004). Detecting distant homology with Meta-BASIC. *Nucleic Acids Research*, **32**(Web Server Issue), W576–W581.

Goldman, N. and Yang, Z. (1994). A Codon-based Model of Nucleotide Substitution for Protein-coding DNA Sequences. *Molecular Biology And Evolution*, **11**(5), 725–736.

Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992). Exhaustive Matching of the Entire Protein Sequence Database. *Science*, **256**(5062), 1443.

Gonnet, P. and Lisacek, F. (2001). Aligning and Scoring Motifs Based on Secondary Characteristics. *Genome Informatics*, **12**, 376–377.

Gonnet, P. and Lisacek, F. (2002). Probabilistic alignment of motifs with sequences. *Bioinformatics*, **18**(8), 1091–1101.

Grantham, R. (1974). Amino Acid Difference Formula to Help Explain Protein Evolution. *Science*, **185**(4154), 862.

Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **84**(13), 4355–4358.

Gupta, K., Thomas, D., Vidya, S., Venkatesh, K., and Ramakumar, S. (2005). Detailed protein sequence alignment based on Spectral Similarity Score (SSS). *BMC Bioinformatics*, **6**(105).

Hall, B. G. (2007). EvolveAGene 3: A DNA coding sequence evolution simulation program. *Nature Precedings*.

Hall, B. G. (2008a). How well does the HoT score reflect sequence alignment accuracy? *Molecular Biology and Evolution*, **25**(8), 1576–1580.

Hall, B. G. (2008b). Simulating DNA Coding Sequence Evolution with EvolveAGene 3. *Molecular Biology and Evolution*, **25**(4), 688–695.

Henikoff, S. and Henikoff, J. G. (1992). Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Sciences*, **89**(22), 10915–10919.

Heringa, J. (1999). Two strategies for sequence comparison: profile-preprocessed and secondary structure induced multiple alignment. *Comput. Chem*, **23**, 341–364.

Herring, B., Bernardin, F., Caglioti, S., Stramer, S., Tobler, L., Andrews, W., Cheng, L., Rampersad, S., Cameron, C., Saldanha, J., *et al.* (2007). Phylogenetic analysis of WNV in North American blood donors during the 2003-2004 epidemic seasons. *Virology*.

Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol*, **233**(1), 123–128.

Jennings, A., Edge, C., and Sternberg, M. (2001). An approach to improving multiple alignments of protein sequences using predicted secondary structure. *Protein Engineering Design and Selection*, **14**(4), 227–231.

Jeong, J., Berman, P., and Przytycka, T. (2006). Fold classification based on secondary structure–how much is gained by including loop topology? *BMC Structural Biology*, **6**(3).

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, **292**(2), 195–202.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**(12), 2577–2637.

Karlin, S. and Altschul, S. F. (1990). Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes. *Proceedings of the National Academy of Sciences*, **87**(6), 2264–2268.

Karplus, K. and Hu, B. (2001). Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. *Bioinformatics*, **17**(8), 713–720.

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

Katoh, K. and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**(14), 3059–3066.

Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, **33**(2), 511–518.

Kawashima, S., Ogata, H., and Kanehisa, M. (1999). AAindex: Amino Acid Index Database. *Nucleic Acids Research*, **27**(1), 368–369.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research*, **36**(Database issue), D202.

Kececioglu, J. and Starrett, D. (2004). Aligning alignments exactly. In *Proceedings of the 8th ACM Conference on Research in Computational Molecular Biology*, pages 85–96.

Kendrew, J., Bodo, G., Dintzis, H., Parrish, R., Wyckoff, H., and Phillips, D. (1958). A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature*, **181**, 662–666.

Kim, J., Moriyama, E., Warr, C., Clyne, P., and Carlson, J. (2000). Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties. *Bioinformatics*, **16**(9), 767–775.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. (1994). Hidden Markov Models in Computational Biology. *Journal of Molecular Biology*, **235**, 1501–1531.

Kyte, J. and Doolittle, R. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, **157**(1), 105–132.

Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, P., McWilliam, H., Valentin, F., Wallace, I., Wilm, A., Lopez, R., Thompson, J., Gibson, T., and Higgins, D. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**(21), 2947–2948.

Lassmann, T. and Sonnhammer, E. L. L. (2002). Quality assessment of multiple alignment programs. *FEBS Letters*, **529**, 126–130.

Lassmann, T. and Sonnhammer, E. L. L. (2005a). Automatic assessment of alignment quality. *Nucleic Acids Research*, **33**(22), 7120–7128.

Lassmann, T. and Sonnhammer, E. L. L. (2005b). Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**(298).

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P., and Bork, P. (2004). SMART 4.0: towards genomic data integration. *Nucleic Acids Research*, **32**, D142–D144.

Lim, V. I. (1974). Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. *J Mol Biol*, **88**(4), 873–94.

Lipman, D. J., Altschul, S. F., and Kececioglu, J. D. (1989). A Tool for Multiple Sequence Alignment. *Proceedings of the National Academy of Sciences*, **86**(12), 4412–4415.

Lüthy, R., McLachlan, A. D., and Eisenberg, D. (1991). Secondary Structure-Based Profiles: Use of Structure-Conserving Scoring Tables in Searching Protein Sequence Databases for Structural Similarities. *Proteins: Structure, Function, and Genetics*, **10**, 229–239.

Marques, A., Antunes, A., Fernandes, P., and Ramos, M. (2006). Comparative evolutionary genomics of the HADH2 gene encoding Aβ-binding alcohol dehydrogenase/17β-hydroxysteroid dehydrogenase type 10 (ABAD/HSD10). *BMC Genomics*, **7**(202).

Mathé, C., Sagot, M., Schiex, T., and Rouzé, P. (2002). Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research*, **30**(19), 4103–4117.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, **405**(2), 442–51.

McClellan, D. A., Palfreyman, E. J., Smith, M. J., Moss, J. L., Christensen, R. G., and Sailsbery, J. K. (2005). Physicochemical Evolution and Molecular Adaptation of the Cetacean and Artiodactyl Cytochrome b Proteins. *Molecular Biology and Evolution*, **22**(3), 437–455.

McGinnis, S. and Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, **32**, 20–25.

Méndez, J., Falcón, A., Hernández, M., and Lorenzo, J. (2008). *Innovations in Hybrid Intelligent Systems*, volume 44, chapter Discovering the Intrinsic Dimensionality of BLOSUM Substitution Matrices Using Evolutionary MDS, pages 369–376. Springer.

Miyazawa, S. and Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**(3), 534–552.

Mizuguchi, K., Deane, C. M., Blundell, T. L., and Overington, J. P. (1998). HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Science*, **7**(11), 2469–2471.

Morgenstern, B. (1999). DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**(3), 211–218.

Morgenstern, B., Frech, K., Dress, A., and Werner, T. (1998). DIALIGN: Finding local similarities by multiple sequence alignment. *Bioinformatics*, **14**(3), 290–294.

Murphy, W. J., Pringle, T. H., Crider, T. A., Springer, M. S., and Miller, W. (2007). Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Research*, **17**(4), 413–421.

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

Notredame, C. and Higgins, D. G. (1996). SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Research*, **24**(8), 1515–1524.

Notredame, C., Holm, L., and Higgins, D. G. (1998). COFFEE: an objective function for multiple sequence alignments. *Bioinformatics*, **14**(5), 407–422.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**(1), 205–217.

Pearson, W. and Lipman, D. (1988). Improved Tools for Biological Sequence Comparison. *Proceedings of the National Academy of Sciences of the United States of America*, **85**(8), 2444–2448.

Pei, J. and Grishin, N. V. (2007). Promals- towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, **23**(7), 802–808.

Periti, P. F., Quagliarotti, G., and Liquori, A. M. (1967). Recognition of alpha-helical segments in proteins of known primary structure. *Jounral of Molecular Biology*, **24**(2), 313–22.

Pokarowski, P., Kloczkowski, A., Nowakowski, S., Pokarowska, M., Jernigan, R., and Kolinski, A. (2007). Ideal amino acid exchange forms for approximating substitution matrices. *Proteins*, **69**(2), 379–93.

Pollard, D. A., Bergman, C. M., Stoye, J., and Celniker, S. E. (2004). Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics*, **5**(6).

Ponting, C., Schultz, J., Milpetz, F., and Bork, P. (1999). SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Research*, **27**(1), 229–232.

Porter, M., Cronin, T., McClellan, D., and Crandall, K. (2007). Molecular Characterization of Crustacean Visual Pigments and the Evolution of Pancrustacean Opsins. *Molecular Biology and Evolution*, **24**(1), 253.

Ptitsyn, O. and Finkelstein, A. (1983). Theory of protein secondary structure and algorithm of its prediction. *Biopolymers*, **22**, 15–25.

Raghava, G. P. S., Searle, S. M. J., Audley, P. C., Barber, J. D., and Barton, G. J. (2003). OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**(47).

Rost, B. (1997). Protein structures sustain evolutionary drift. *Folding and Design*, **2**, 19–24.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering Design and Selection*, **12**(2), 85–94.

Rost, B. and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Jounral of Molecular Biology*, **232**(2), 584–99.

Rudnicki, W. and Komorowski, J. (2005). *Intelligent Information Processing and Web Mining*, chapter Soft Computing Approach to the Analysis of the Amino Acid Similarity Matrices, pages 663–670. Advances in Soft Computing. Springer.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.

Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Genetics*, **9**(1), 56–68.

Sankoff, D., Cedergren, R., and Lapalme, G. (1976). Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA. *Journal of Molecular Evolution*, **7**(2), 133–149.

Sauder, J., Arthur, J., and Dunbrack Jr, R. (2000). Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins Structure Function and Genetics*, **40**(1), 6–22.

Shi, J., Blundell, T., and Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology*, **310**(1), 243–257.

Simossis, V. and Heringa, J. (2003). The PRALINE online server: optimising progressive multiple alignment on the web. *Computational Biology and Chemistry*, **27**(4-5), 511–519.

Simossis, V. and Heringa, J. (2004). Integrating protein secondary structure prediction and multiple sequence alignment. *Current Protein and Peptide Science*, **5**, 249–266.

Simossis, V. and Heringa, J. (2005). PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Research*, **33, Web Server Issue**, W289–W294.

Sing, C., Haviland, M., Zerba, K., and Templeton, A. (1992). Application of cladistics to the analysis of genotype-phenotype relationships. *European Journal of Epidemiology*, **8**, 3–9.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**(1), 195–197.

Sneath, P. H. (1966). Relations between chemical structure and biological activity in peptides. *Journal of Theoretical Biology*, **12**(2), 157–95.

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**(7), 951–960.

Sturrock, S. S. and Dryden, D. T. F. (1997). A prediction of the amino acids and structures involved in DNA recognition by type I DNA restriction and modification enzymes. *Nucleic Acids Research*, **25**(17), 3408–3414.

Subramanian, A. R., Weyer-Menkhoff, J., Kaufmann, M., and Morgenstern, B. (2005). DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, **6**(66).

Sundberg, K., O'Connor, T., Carroll, H., Clement, M., and Snell, Q. (2007). Using parsimony to guide maximum likelihood searches. In *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering*, volume II, pages 774–779.

Sundberg, K., O'Connor, T., Carroll, H., Clement, M., and Snell, Q. (2008). Parsimony accelerated Maximum Likelihood searches. *International Journal of Computaional Biology and Drug Design*, **1**(1), 74–87.

Szustakowski, J. and Weng, Z. (2000). Protein structure alignment using a genetic algorithm. *Proteins Structure Function and Genetics*, **38**(4), 428–440.

Taylor, W. and Orengo, C. (1989a). A holistic approach to protein structure alignment. *Protein Engineering Design and Selection*, **2**(7), 505–519.

Taylor, W. and Orengo, C. (1989b). Protein structure alignment. *J Mol Biol*, **208**(1), 1–22.

Thompson, J., Plewniak, F., and Poch, O. (1999a). BAliBASE: A benchmark alignments database for the evaluation of multiple sequence alignment programs. *Bioinformatics*, **15**(1), 87–88.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Thompson, J. D., Plewniak, F., and Poch, O. (1999b). A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research*, **27**(13), 2682–2690.

Thompson, J. D., Koehl, P., Ripp, R., and Poch, O. (2005). BAliBASE 3.0 latest developments of the multiple sequence alignmnet benchmark. *Proteins: Structure, Function, and Bioinformatics*, **61**(1), 127–136.

Thorne, J. L., Goldman, N., and Jones, D. T. (1996). Combining protein evolution and secondary structure. *Molecular Biology and Evolution*, **13**(5), 666–673.

Thorvaldsen, S., Flå, T., and Willassen, N. (2005). *Biological and Medical Data Analysis*, volume 3745, chapter Extracting Molecular Diversity Between Populations Through Sequence Alignments, pages 317–328. Springer.

Tomii, K. and Kanehisa, M. (1996). Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Engineering Design and Selection*, **9**(1), 27–36.

Trapane, T. and Lattman, E. (2007). Seventh Meeting on the Critical Assessment of Techniques for Protein Structure Prediction. *Proteins*, **69**(8), 1–2.

Van Walle, I. (2004). Align-m–a new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics*, **20**(9), 1428–1435.

Van Walle, I., Lasters, I., and Wyns, L. (2005). SABmark–a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, **21**(7), 1267–1268.

Wang, L. and Jiang, T. (1994). On the complexity of multiple sequence alignment. *Journal of Computational Biology*, **1**(4), 337–48.

Wheeler, W. C., Gladstein, D. S., and Laet, J. D. (1996–2003). POY: Version 3.0.

Whiting, M. F., Bradler, S., and Maxwell, T. (2003). Loss and recovery of wings in stick insects. *Nature*, **421**(6920), 264–267.

Wilbur, W. and Lipman, D. (1984). The context dependent comparison of biological sequences. *SIAM Journal Applied Mathematics*, **44**, 557–567.

Wilm, A., Higgins, D., and Notredame, C. (2008). R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Research*, **36**(9), e52.

Woolley, S., Johnson, J., Smith, M. J., Crandall, K. A., and McClellan, D. A. (2003). TreeSAAP: Selection on Amino Acid Properties using phylogenetic trees. *Bioinformatics*, **19**(5), 671–672.

Wrabl, J. and Grishin, N. (2005). Grouping of amino acid types and extraction of amino acid properties from multiple sequence alignments using variance maximization. *Proteins*, **61**(3), 523–534.

Wu, S. and Manber, U. (1992). Fast text searching: allowing errors. *Communications of the ACM*, **35**(10), 83–91.

Xia, X. and Li, W.-H. (1998). What amino acid properties affect protein evolution? *Journal of Molecular Evolution*, **47**, 557–564.

Zhang, C. and Wong, A. (1997). A genetic algorithm for multiple molecular sequence alignment. *Bioinformatics*, **13**(6), 565–582.

Zhang, X. and Kahveci, T. (2005). A New Approach for Multiple Sequence Alignment. In *The Ninth Annual International Conference on Research in Computational Molecular Biology (RECOMB)*. Poster.

Zhang, X. and Kahveci, T. (2006). A new approach for alignment of multiple proteins. *Proceedings of the 11th Pacific Symposium on Biocomputing*, pages 339–350.

Zhang, Y., Zheng, N., Hao, P., and Zhong, Y. (2004). Reconstruction of the most recent common ancestor sequences of SARS-CoV S gene and detection of adaptive evolution in the spike protein. *Chin Sci Bull*, **49**, 1311–1313.