

Internship Report

Estimating ClustalW pairwise alignment parameters for non-Coding DNA sequences

Meraj Aziz
Computational Biosciences Program
Arizona State University

Committee Members

Michael S Rosenberg
Martin Wojciechowski
Rosemary Renaut

Internship:
Summer 2007
Report Number:
07-07
Not Confidential

Table of Contents

Abstract.....	3
Introduction.....	4
Methods.....	6
• <i>Defining the data pipeline</i>	10
• <i>How is the Ratio Computed?</i>	12
• <i>Parameters</i>	13
Results and Discussion.....	15
• <i>Gap Extension and Gap Opening Cost</i>	15
• <i>Transition Weights</i>	18
Conclusion and Future Work.....	20
Acknowledgements.....	22
References.....	23
Appendices.....	24
• <i>Appendix A</i>	24
• <i>Appendix B</i>	26
• <i>Appendix C</i>	35

Abstract

Sequence alignment programs such as ClustalW are widely used in aligning protein and DNA sequences in order to determine regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. In order to align sequences the user has to define initial parameters for the alignment program or simply use the default parameters declared in the programs. Prior knowledge of the molecular evolutionary nature of the sequences may be vital in determining the best sequence alignment parameters. These initial parameters lay the foundation for the sequence alignment process and it is important that we input the right parameters in order to get the best possible alignment. This report discusses a preliminary effort to determine the best ClustalW parameters based on a known evolutionary model of DNA. In order to achieve our goal we simulated pairs of DNA sequences and aligned them with ClustalW multiple times with different parameters. The results were later analyzed by comparing ClustalW aligned sequences with the original simulated DNA sequences. Our results indicate that for most of our simulated DNA sequences it is better to use a significantly lower gap extension cost than the default value. Moreover we also found that using a transition weight less than 0.1 would give a better sequence alignment than the default option.

Introduction

DNA sequencing has come a long way since the Human Genome Project was proposed in 1990. Since 1995, the genomes of more than 180 organisms have been sequenced (4). As a result, a substantial amount of genetic data has been generated.

There are several advantages to sequencing the genomes of organisms. First, it provides an opportunity to compare genomes and to find similarities/differences among different organisms. This could lead to characterizing protein families, to finding homologous regions between new sequences and existing families of sequences, and also help predict the secondary and tertiary structures of new sequences (1). Second, genomic similarities/differences may infer evolutionary relationships between different organisms and help to answer questions about the origins of life on earth. Finally, we can also compare genomes of individuals of the same species to observe differences and learn how changes in the genomes may explain disease vulnerability and reaction to therapeutic drugs, toxic substances, and other environmental factors (5).

In order to compare DNA or protein sequences, efficient computational algorithms can be used to align the sequences and find homologous regions in sequences, or even to discover any regions of the genome where DNA insertion/deletion may have happened through the course of time. The computational power of today's computers has enabled the development and use of sophisticated alignment algorithms which were deemed computationally infeasible twenty years ago. Today, there are several sequence alignment programs available for use. Some examples include ClustalW (1), Multiple Sequence

Alignment (MSA) (8), T-Coffee (9), MUSCLE (10), MAVID (11), kalign (12) and MAFFT (13). For our research we chose the ClustalW sequence alignment program because it is one of the most widely used and easy to run programs.

ClustalW uses the progressive multiple sequence alignment algorithm to align sequences (1). This alignment technique aligns two closely related sequences and successively adds the next closest sequence to the alignment produced in the previous step. In order to align sequences the user has to input the values for the initial parameters into the ClustalW program. Some of the significant parameters are Gap opening cost, Gap extension cost, Gap distance and the Transition weights. The optimal values for aligning sequences, i.e., the parameter values which lead to the most accurate alignments will vary depending on the exact nature of the sequences. For example the actual values of the parameters for gap opening and gap extension cost for mammalian or primate sequences may be different from those for yeast. This is because mammalian or primate DNA sequences evolve under different conditions than those for yeast. Currently, there is no standard to indicate for which set of DNA sequences one would choose what values of the ClustalW parameters. Having reasonable values for the initial parameters, leads to a more accurate sequence alignment. The goal of this study is to determine approximate ClustalW parameters for pairwise non-coding DNA sequences under the assumption that the user has some prior knowledge of the evolutionary nature of their sequences.

Methods

Simulating DNA sequences using molecular sequence evolution has become an important part of comparative genomics. It is a good method for testing the efficiency of alignment programs such as ClustalW. Therefore, DNA sequences for this project were simulated using the software MySSP (2). The user interface for MySSP is shown in Figure1. The MySSP non-stationary evolutionary sequence simulation software is designed and developed by Dr. Michael S Rosenberg at Arizona State University. This software simulates sequences based on a user defined phylogenetic tree. In our case the tree, depicted in Figure2, is very simple and has just two branches of equal length. Initially we are only interested in pairwise alignment. MySSP has many features which are not readily available in other DNA sequence simulation programs. It provides the user with the flexibility of generating DNA sequences with insertion and deletion rates and also the option of choosing a Poisson or power distribution for the Indel size distribution. Another important feature of this software is that the output sequences have gaps in them. These gaps are included such that the output sequences are aligned correctly and the aligned sites across sequences represent true homologies (2). Here, we denote the simulations generated by MySSP “true” alignments. The “true” alignment will be used to compare the hypothesized alignments obtained from ClustalW.

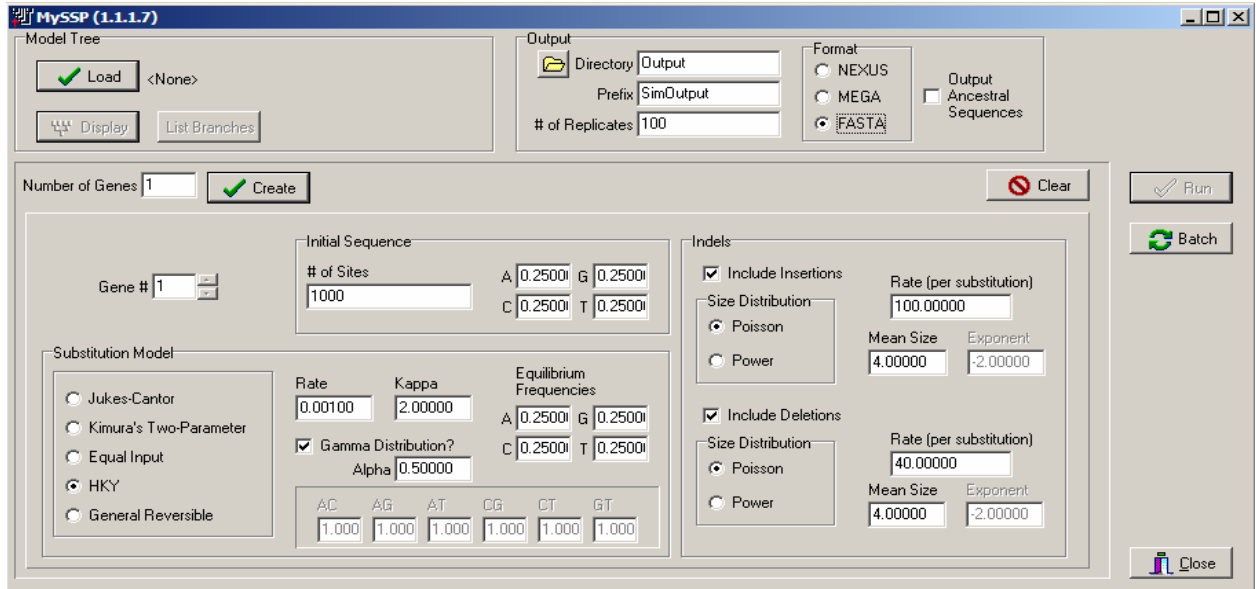


Figure1. MySSP Graphical User Interface

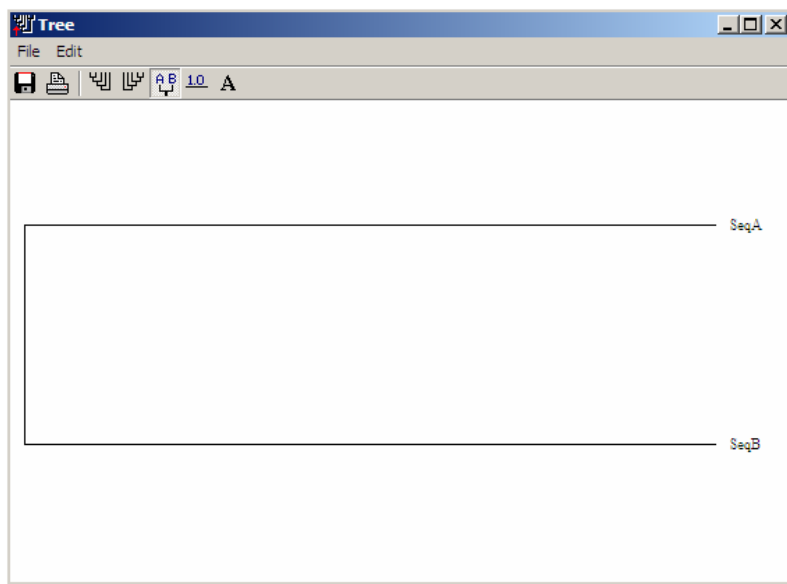


Figure2. Sample tree with two branches

MySSP user interface provides the user with all the parameters available in the program that can be used for the DNA sequence simulation. These parameters can be changed by the user to generate DNA sequences based on their requirements. Initial MySSP parameters were carefully chosen after consulting the literature extensively. Initial models were chosen to mimic the evolutionary process observed in model organisms,

especially mammals. Initial values reflect those obtained from empirical analysis of these organisms. Selecting reasonable values for the initial parameters is important because DNA sequences should not be simulated with random parameters. In particular, the simulations should be as realistic and close to the DNA molecular evolution model of the species which are currently used as model organisms.

Table1 summarizes the initial parameters used as a baseline for simulating non-coding DNA sequences in MySSP. All simulations used HKY + Γ (7) model of substitution.

Default MySSP Parameters used				
Sequence Length : 1000		Substitution Model		HKY
#Replicates : 100		kappa		3.6
Frequency of nucleotides :		Rate		0.5
A	0.2	Gamma Distribution		0.5
C	0.3	Insertions		Rate 12
G	0.3			Distribution power
T	0.2			Constant 0.42
				Exponential -2
		Deletions		Rate 12
				Distribution power
				Constant 0.42
				Exponential -2
Default Clustalw Parameters used				
Transweights	0	Gap Open		15
		Gap Extension		7
DNA matrix	IUB	Gap dist		4

Table1. Initial Default parameters used

Random sequences of length 1000 nucleotides were generated with initial nucleotide frequencies of C = G = 0.3, T = A = 0.2. Kappa is the transition to transversion rate ratio. The choice kappa =3.6 mimics that found empirically in neutral sites in mammals (14). The rate parameter determines how many substitutions occur on each branch of the

phylogenetic tree and allows for rescaling of the divergence between the sequences (e.g. a rate of 0.25 would approximate the divergence between humans and mice).

Rates are of equal duration and we can set different rates of nucleotide substitution along each branch, if we use the batch processing command. MySSP software allows users to set Indel gap size distributions. A power distribution for Indel sizes was selected. Indel sizes are linearly related to their frequency on a log-log scale. Therefore Indel gap sizes follow a power law (6). The power distribution with an Indel distribution rate of 12 was chosen to match that empirically observed in humans (3). An Indel rate of 12 means that an insertion/deletion event occurs, on average, every 12 substitutions. A larger Indel rate, such as 100, results in fewer substitutions so that each event occurs only every 100 substitutions. Also, note that insertions and deletions are modeled separately. So a rate of 12 for each means that an Indel is actually occurring approximately every 6 substitutions. The length of an individual Indel event is controlled by the exponential value of the power curve, capped at a maximum length of 100 sites. A value of -1 means a broader Indel size distribution and a value of -4 means that the Indel sizes are almost all equal to 1. The Indel size distribution generated by these parameters in MySSP is illustrated in the histograms in Appendix A.

As far as ClustalW parameters are concerned the focus here is on Transition Weights, Gap Opening cost, Gap Extension Cost and Gap Distance parameters. The default ClustalW parameter values for DNA are shown in Table1. The IUB DNA matrix was used. This matrix assigns a score of 1.9 for matches and 0 for mismatches. These parameters are recommended as default at the ClustalW website.

	A	C	G	T
A		Transversion	Transition	Transversion
C	Transversion		Transversion	Transition
G	Transition	Transversion		Transversion
T	Transversion	Transition	Transversion	

Table2. Transitions Transversions illustrated.

(Source: <http://www.life.umd.edu/labs/delwiche/bsci348s/lec/Phylogenetics1.html>)

The transition weight parameter ranges between 0 and 1. A 0 indicates that transitions are scored as mismatches while a weight of 1 means that $A \leftrightarrow G$ or $C \leftrightarrow T$ are treated as matches. Intermediate values treat transitions as partial values. It has been observed that for distantly related DNA sequences, the transition weight should be near to a value of 0 and for closely related sequences it should be greater than 0.5

(http://web.mit.edu/clustalw_v1.83/README).

Previous simulation studies (14) have shown that the default simulation and alignment parameters used here should lead to alignment accuracies of about 65 to 70%; this was viewed as a good starting point from which to alter parameters so as to improve/decrease accuracy.

Defining the Data and Pipeline

In order to make statistically significant simulations with minimum variability, one hundred replicates in FASTA format for a given set of MySSP parameters are generated. The files are transferred to an input folder and a Perl Script (RemoveGaps.pl) is used in order to remove all the gaps from the sequences of the replicates. Once gaps are removed the Perl script Automater.pl runs the ClustalW program on each replicate. The results

from Automater.pl are put into the results folder. The Perl script Analyzer.pl executes and generates a summary file for the hundred replicates. This summary file consists of a value for each replicate which represents the proportion of aligned sites in the hypothesized alignments which are correct, i.e., which are found in the “true” alignments. The higher the ratio, the closer the aligned sequence to the “true” alignment. These values are averaged across the 100 replicates to produce the “alignment accuracy score”. After averages are calculated results are plotted using Excel. The work flow of the process explained above can be seen in Figure 3.

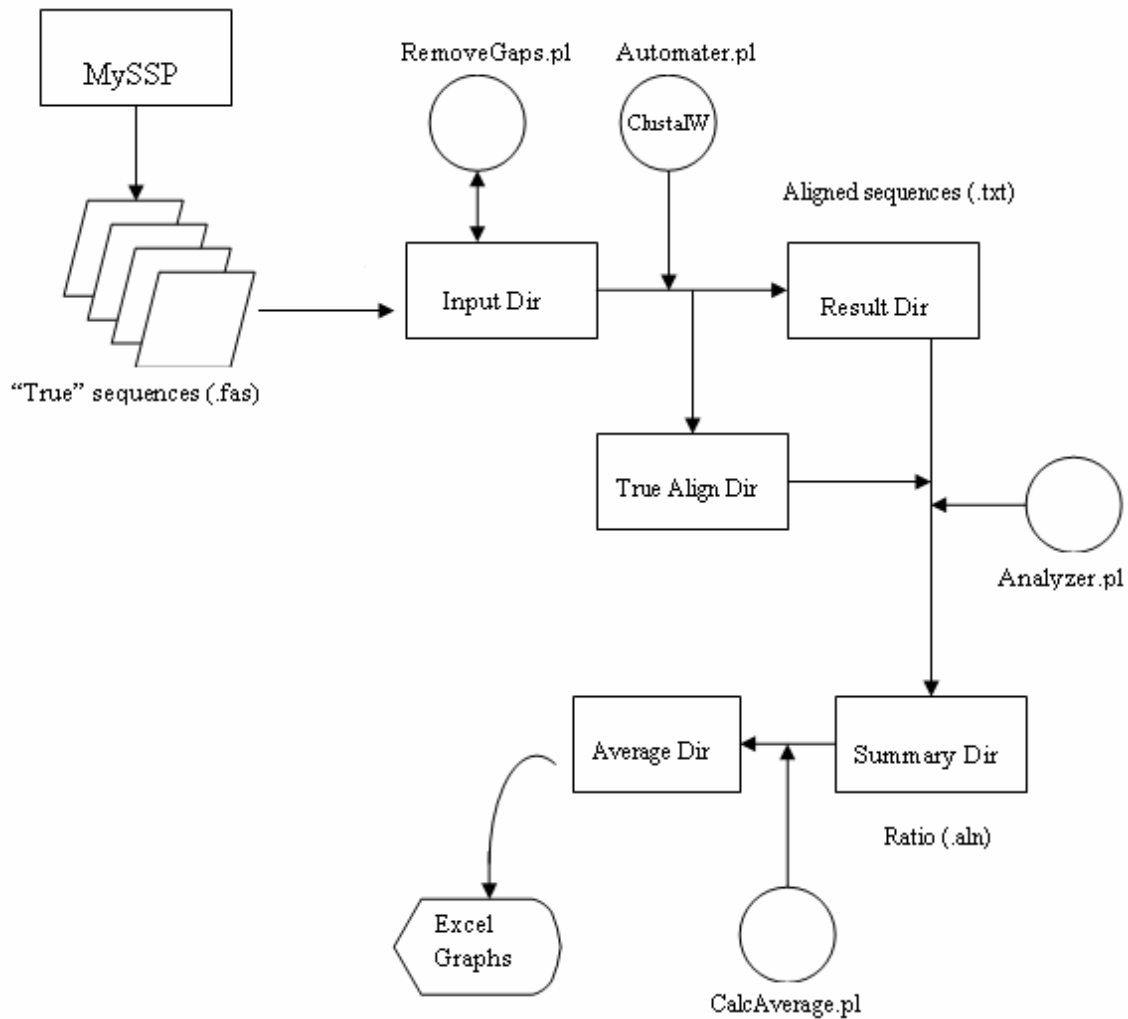


Figure3. Work Flow

How is the Ratio Computed?

We are measuring alignment accuracy as the proportion of aligned sites in the hypothesized alignment which are aligned identically in the true alignment (14). Figure 4 shows a simple example. The hypothesized alignment is 12 sites long, but we are only interested in un-gapped sites, of which there are 10 in the alignment. Each of these 10 sites is compared to the true alignment. If the same site in sequence 1 is aligned to the same site in sequence 2 in both alignments, this is considered to be a correctly aligned site; if they are different it is an incorrectly aligned site. In the example, the first five aligned sites in the hypothesized sequence are identical to those in the true alignment (we do not count the A/- site because of the gap). The sixth aligned site in the hypothesized alignment is a C/C. A C/C is also the sixth aligned site in the true alignment, but these are not the same: in the true alignment the 7th nucleotide of the 1st sequence is aligned with the 7th nucleotide of the 2nd sequence, while in the hypothesized alignment the 7th nucleotide of the 1st sequence is aligned with the 6th nucleotide of the 2nd sequence. This is an incorrect alignment, even though the state of the characters in the alignment is identical in both cases (C/C). The next 3 sites are also incorrectly aligned (incorrect sites marked with *'s), with only the final site correctly aligned. Thus of the 10 aligned sites in the hypothesized sequence, 6 are correct, leading to an alignment accuracy measure of 6/10 or 0.6.

True alignment

1: TACCAT-CAGGG
 2: TCCG-TCCAGAG

Hypothesized alignment

1: TACCATCAGG-G
 2: TCCG-TCCAGAG
 * * * *

Figure4.

There are a number of possible strategies for including sites with gaps in the measure, but these can be more complicated to estimate and have been shown to be highly correlated with the measure presented above (Rosenberg and Ogden, in prep).

Parameters

****Default parameters are marked in red**

MySSP Parameters		Range	
insertion/deletion frequency		{6, 12, 24, 48, 96, 192}	
Rate		{0.125, 0.25, 0.5, 0.75}	
Kappa		{1.8, 3.0, 3.2, 3.4, 3.6, 3.8, 4.0, 4.2, 4.4, 4.6, 4.8, 5.0, 7.2, 15.0}	
Power exponential rate(indel)		{-1, -2, -4}	
Clustal parameters		Range	
transweights		0.0--> 1.0	and later zoomed in 0.0 to 0.4
gap open		0, 5, 10, 15, 20, 25, 30	
gap ext		0, 1, 3, 5, 7, 15	
DNA matrix		All matches score 1.9; mismatches for IUB symbols score 0	

Table3.

The value of the default parameters was used as a starting point for the study. Table3 gives the parameter ranges used in this study. We not only changed the ClustalW parameters but also the evolutionary DNA sequence parameters. Numerous combinations were considered, see Table4. The left four columns (in blue heading Table4) denote MySSP parameters and the right four columns (in orange heading Table4) denote the values of the ClustalW parameters. An “x” in the Table4, denotes that the 42

combinations for gap open and gap extension costs were used for a particular set of MySSP parameters. These combinations are illustrated in Table5.

MySSP			Clustalw		
Frequency	Rate	Kappa	Power exp	Gap cost	Transweights
6	0.5	3.6	-1	x	
			-2	x	
			-4	x	
12	0.125	1.8			0.0--> 1.0 (0.1 incr)
		3.6			0.0--> 1.0
		7.2			0.0--> 1.0
		15			0.0--> 1.0
	0.25	1.8			0.0--> 1.0
		3			0.0-->0.4 (0.25 incr)
		3.2			0.0-->0.4
		3.4			0.0-->0.4
		3.6		0.0--> 1.0	0.0-->0.4
		3.8			0.0-->0.4
		4			0.0-->0.4
		4.2			0.0-->0.4
		4.4			0.0-->0.4
		4.6			0.0-->0.4
		4.8			0.0-->0.4
		5			0.0-->0.4
		7.2		0.0--> 1.0	
		15		0.0--> 1.0	
	0.5	1.8			0.0--> 1.0
		3			0.0-->0.4
		3.2			0.0-->0.4
		3.4			0.0-->0.4
		3.6	-1	x	0.0--> 1.0
			-2	x	0.0--> 1.0
			-4	x	0.0--> 1.0
		4			0.0-->0.4
		4.2			0.0-->0.4
		4.4			0.0-->0.4
		4.6			0.0-->0.4
		4.8			0.0-->0.4
		5			0.0-->0.4
		7.2		0.0--> 1.0	
		15		0.0--> 1.0	
	0.75	1.8			0.0--> 1.0
		3.6			0.0--> 1.0
		7.2			0.0--> 1.0
		15			0.0--> 1.0
24	0.5	3.6	-1	x	
			-2	x	
			-4	x	
48	0.5	3.6	-1	x	
			-2	x	
			-4	x	
96	0.5	3.6	-1	x	
			-2	x	
			-4	x	
192	0.5	3.6	-1	x	
			-2	x	
			-4	x	

Table4. For example, if we focus on the high lighted (yellow) section of the table, first column Frequency is the Indel frequency (12), second column is the rate of substitutions (0.5), third column is the kappa (ranges from 1.8 to 15), the fourth column is exponential column (range -1 to -4) indicating Indels sizes. The 'x' in column five denotes 42 pairs of gap extension and gap cost (0 1, 0 3, 0 5...) as shown in Table4. Column 6 denotes the transition weights used, ranging from 0 to 1 with an increment of 0.1. Column 7, we zoomed in the range 0 to 0.4 transition weights.

		Gap Extension Cost					
		0	1	3	5	7	15
Gap Open Cost	0	[0 0]	[0 1]	[0 3]	[0 5]	[0 7]	[0 15]
	5	[0 5]	[5 1]	[5 3]	[5 5]	[5 7]	[5 15]
	10	[0 10]	[10 1]	[10 3]	[10 5]	[10 7]	[10 15]
	15	[0 15]	[15 1]	[15 3]	[15 5]	[15 7]	[15 15]
	20	[0 20]	[20 1]	[20 3]	[20 5]	[20 7]	[20 15]
	25	[0 25]	[25 1]	[25 3]	[25 5]	[25 7]	[25 15]
	30	[0 30]	[30 1]	[30 3]	[30 5]	[30 7]	[30 15]

Table5. 42 pairs of gap extension and gap open cost

**Default parameters are marked in red

Results and Discussion

Our results are divided into two sections. First we will discuss the effects of Gap opening and extension costs and their relationship to Indel frequency and size. Some tables have been drawn to depict the results with respect to Indel sizes (Table6) and Indel frequency (Table7). Then we will discuss the ClustalW transition weight parameter and its relationship with MySSP kappa and Rate parameters.

Gap Opening and Extension Cost

	Indel Frequency	Alignment Score Range	Best Gap Open	Best Gap Extension
Indel Exponent -1	1/6	~ 0.05 -- 0.30	25	0
	1/12	~ 0.15 -- 0.45	25	0
	1/24	~ 0.28 -- 0.67	25	0
	1/48	~ 0.55 -- 0.86	30	0
	1/96	~ 0.66 -- 0.95	30	0
	1/192	~ 0.72 -- 0.97	25	0/1
Indel Exponent -2	1/6	~ 0.19 -- 0.55	10	0
	1/12	~ 0.49 -- 0.75	15	0
	1/24	~ 0.53 -- 0.85	30	0
	1/48	~ 0.76 -- 0.94	20	0/1/3
	1/96	~ 0.80 -- 0.97	15/20/25	0/1/3/5
	1/192	~ 0.80 -- 0.98	15/20/25/30	0/1/3/5/7
Indel Exponent -4	1/6	~ 0.49 -- 0.71	0	7
	1/12	~ 0.70 -- 0.84	0	15
	1/24	~ 0.86 -- 0.92	5	15
	1/48	~ 0.79 -- 0.96	5	15
	1/96	~ 0.79 -- 0.97	10	15
	1/192	~ 0.80 -- 0.98	20	0/1/5/7/15

* The default Gap Open cost is 15, Gap Extension cost is 6.66

Table6. Results grouped by Indel Exponent rates (or Indel sizes).

Table6 Summarizes results from the graphs in Appendix B. Vertical color bands group results with respect to Indel sizes. The first column gives the Indel frequency for each

Indel size. The Alignment Score Range column indicates the range which the graph spans with respect to the Indel size and frequency. Best Gap opening and extension cost columns tell us about the best ClustalW Gap opening and extension cost for a particular Indel size and rate.

Our results from Table 6 indicate that as Indel frequencies become rarer ($1/6 \rightarrow 1/192$), the range of the alignment scores improves from poor alignment results (e.g. ~ 0.05 -- 0.30 for Indel size exponent -1) to almost perfect alignment (range ~ 0.72 -- 0.97). The same is the case with Indel size exponent of -2 and -4 . Table7 shows alignment score ranges with respect to the Indel frequency. Indel frequency and Indel size both have an effect on the alignment accuracy, as is evident by looking at the rows in Table7.

		Alignment Score Range		
		Indel Exponent -1	Indel Exponent -2	Indel Exponent -4
Indel Frequency	1/6	~ 0.05 -- 0.30	~ 0.19 -- 0.55	~ 0.49 -- 0.71
	1/12	~ 0.15 -- 0.45	~ 0.49 -- 0.75	~ 0.70 -- 0.84
	1/24	~ 0.28 -- 0.67	~ 0.53 -- 0.85	~ 0.86 -- 0.92
	1/48	~ 0.55 -- 0.86	~ 0.76 -- 0.94	~ 0.79 -- 0.96
	1/96	~ 0.66 -- 0.95	~ 0.80 -- 0.97	~ 0.79 -- 0.97
	1/192	~ 0.72 -- 0.97	~ 0.80 -- 0.98	~ 0.80 -- 0.98

Table7. Alignment score ranges, w .r .t. Indel frequency.

For example, Indel exponent of -2 for Indel frequency of $1/6$, starts with a higher range (~ 0.19 -- 0.55) than that of -1 (~ 0.05 -- 0.30) and the alignment score range for Indel exponent -4 is even higher (~ 0.49 -- 0.71) than that of -2 . As Indels become smaller in size, the alignments improve and become closer to the “true” sequences. Moreover, we observe increasing alignment accuracy as the Indel frequency decreases. This could be because there are fewer shorter gaps and more nucleotides to align. Therefore it is easier for ClustalW to align such sequences rather than sequences which have long frequent gaps. Thus, alignments improve with shorter less frequent Indels, as expected.

As far as Gap opening costs are concerned, we expected there to be a dependency between Gap opening cost and Indel frequency. Table6 shows the data for gap opening cost with different Indel sizes and frequencies. For Indel size exponent of -1, we observe the Gap opening cost to increase slightly as Indels become less frequent. The same is the case with Indel size exponent of -2 (except slight inconsistency in Indel frequency of 1/48 and 1/96). For the Indel size exponent of -4, we notice very low Gap opening costs. The value is as low as 0 for Indel frequency 1/6 and 1/12. As Indels become less frequent we observe an increase in Gap opening cost. These results are consistent with our expectation i.e. as Indels become rarer, the cost of opening a gap should increase.

The best Gap extension cost is 0 for long and medium sized Indels, see Table6. This is an unexpected outcome since we were expecting a dependant relationship between the gap extension and Indel sizes. Looking at the results from Table 6 we observe that once a gap is opened, the best alignments (true vs ClustalW aligned) are obtained when there is no cost to extending the Gaps. This is true for long (exponent -1) and medium (exponent -2) sized Indels. As for short Indels (exponent -4) we see results that correlate with our expectations. The cost of opening gaps is low and cost of extending the gaps is more expensive (around 15). This result indicates that since 90% of our Indels have a length of 1, opening a gap is cheap but to discourage gap extension the cost is much higher. However, the size distribution (Appendix B) simulated with an exponent of -1 is unrealistic. Most empirical studies show Indel size distributions much closer to the exponent of -2 (3).

One possible explanation for a gap extension cost of 0 for short and medium sized indels could be that ClustalW avoids putting separated gaps close together. So instead of having two large or medium sized gaps nearby, these are merged together to form a single larger gap. Based on this line of thought we decided to shift our attention towards the ClustalW Gap distance (-gapdist) parameter. The default value for this parameter is 4. Gap distance prevents gaps from being too close to each other. Gaps less than a specified value are penalized more than other gaps. This results in less frequent, block-like appearance of the alignment (Reference ClustalW website). We tried different values for the Gap distance such as 8, 20 and 100 but we never saw any change in our sequence alignment scores. We concluded that Gap distance only plays a role in multiple sequence alignments and not pair wise alignments.

Transition weights

Besides Gap opening, extension and distance costs we also explored the effect of another ClustalW parameter called transition weight. The transition weights range between a value of 0 and 1. A value of 1 for the parameter means that the transitions are scored as matches and 0 indicates mismatches. Anything in between 0 and 1 indicates varying degrees of mismatches (http://web.mit.edu/clustalw_v1.83/README). For our simulations, we initially decided to simulate DNA sequences with four kappa parameters (1.8, 3.6, 7.2 and 15) with four different rate parameters (0.125, 0.25, 0.5, and 0.75). The term kappa refers to transition/transversion ratio and the rate parameter is a scalar of the model tree's branch length (which is 1 in our case). The rate x branch length gives us the expected number of substitutions/site (2).

Our initial results indicate that the use of transition weights does influence the alignment accuracy, but the effect is very minute. Therefore we needed to zoom into our results in the range of 0.0 to 0.4 for transition weights and instead of using the kappa rates used above, we used rates in the range of kappa 3.0 to 5.0 and rates of 0.25 and 0.5 only.

		Rate 0.25	Rate 0.5
		Best Transition weight	Best Transition weight
Kappa	3	0.1	0.1
	3.2	0.05	0.1
	3.4	0.05	0.05
	3.6	0.03	0.03
	3.8	0.03	0.03
	4	0.05	0.05
	4.2	0.03	0.05
	4.4	0.03	0.03
	4.6	0.05	0.1
	4.8	0.03	0.03
	5	0.03	0.05

Table8. Kappa rate, best transition weight pairs for rates 0.25 and 0.5.

Results are summarized in Table8. We notice a repeated occurrence of values for the transition weights ranging from 0.03 to 0.1. We conclude that using a transition weight of 0.03 \rightarrow 0.1 will result in slightly better alignments than the default value of 0.0.

Conclusion and Future Work

The goal of this study is to provide an individual with a set of ClustalW parameters which can be used to achieve better sequence alignment scores based on the evolutionary nature of the sequences. The results are summarized in Table6 and Table8. We are solely basing our results by comparing simulations (which we call “true” alignments) and ClustalW aligned sequences. We have tried to simulate sequences based on a realistic model organism (in our case human and primates). Therefore whenever we use the term “best” or “better” we mean in comparison to our simulations. Moreover, if we look at the graphs in Appendix B, we realize that each graph shows us an evolutionary model and we run ClustalW multiple times to observe which ClustalW parameters did well and which ones resulted in poor alignment accuracies. Suppose, if we have DNA sequences with Indel frequency $1/12$ and Indels size exponent of -2 , a reasonable alignment accuracy i.e. the closest alignment to the “true” sequence can be achieved if we use a Gap opening cost of 15 and Gap extension cost of 0 (Table6).

We expected there to be a dependency between Gap opening cost and Indel rate and between Gap extension cost and Indel sizes. But this was not the case as our results indicate that Indel rate and size are both influenced by Gap opening cost and are less influenced by Gap extension cost. This is an unexpected result since we will now only require Gap opening cost to achieve better sequence alignments. This point towards the possible use of Gap opening cost as the key parameter for ClustalW sequence alignment rather than using a combination of both Gap opening and extension cost. Moreover, ClustalW uses a default of 0 for transweights, we also came to the conclusion that very

low (but not zero) transition weights should be used in order to achieve better sequence alignments.

This report describes a preliminary step towards finding better ClustalW sequence alignment parameters. In the future we would like to extend the analysis to include multiple sequences and different alignment programs such as MSA, T-Coffee, MAFFT, MUSCLE, kalign and MAVID. The ultimate goal is to remove guess work from alignment parameter choices and to provide the scientist with a set of values for their particular sequences which would give them the closest hypothesized alignment. We would also like to include coding DNA sequences in our analysis. Moreover, we would like to predict ClustalW alignment parameters for any new sequences that we come across. We can implement this by developing a knowledgebase for sequence parameters and applying learning algorithms to predict any DNA sequence alignment parameter, coding or non-coding.

Acknowledgements

I would like to thank Dr. Michael S Rosenberg for giving me an opportunity to work in his lab and on this project. I would also like to thank him for his guidance and help throughout this project.

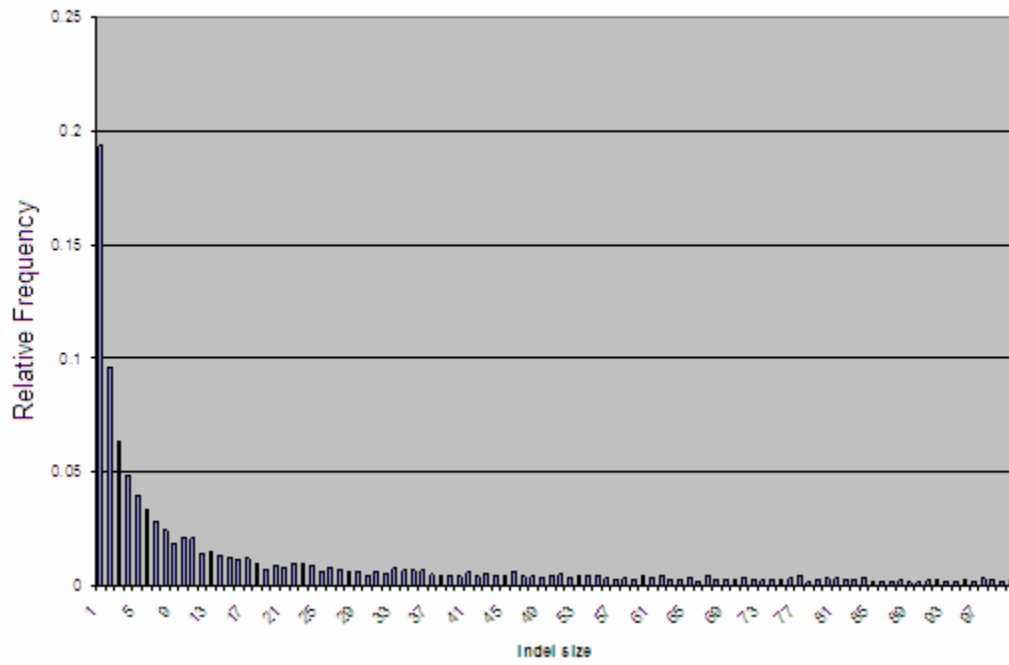
References

1. Higgins D., Thompson J., Gibson T. Thompson J. D., Higgins D. G., Gibson T. J. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." *Nucleic Acids Res* (1994): 22 4673-4680
2. Rosenberg, Michael S. "MySSP: Non-Stationary Evolutionary Sequence Simulation, Including Indels." *Evolutionary Bioinformatics Online* (2005): 1 81-83.
3. Mills, Ryan E., Christopher T. Luttig, Christine E. Larkins, Adam Beauchamp, Circe W. Tsui, Stephen Pittard, and Scott E. Devine. "A Initial Map of Insertion and Deletion (INDEL) Variation in the Human Genome." *Genome Research* 16 (2006): 1182-1190.
4. Ruder, Kate, and Edward R. Winstead. "A Quick Guide to SEQUENCED GENOME." *Genome News Network*. J. Craig Venter Institute. 8 July 2007
http://www.genomenewsnetwork.org/resources/sequenced_genomes/genome_guide_p1.shtml
5. Kate Marusina, Ph.D. "\$1,000 Genome Remains the Holiest, Innovations and Advances Move Sequencing Forward, But the Ultimate Goal Will Continue to Be Elusive." *Genetic Engineering and Biotechnology News*. 01 Feb. 2007. 7 July 2007
<http://www.genengnews.com/articles/chitem.aspx?aid=2000&chid=2>
6. Cartwright, Reed A. "Logarithmic gap costs decrease alignment accuracy" *BMC Bioinformatics* (2006) 7:527
7. Hasegawa, M., Kishino, H and Yano, T. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22:160-174.
8. Lipman DJ, Altschul SF, Kececioglu JD. A tool for multiple sequence alignment. *Proc Natl. Acad. Sci. USA*. 1989; 86:4412-4415.
9. Notredame C, Higgins DG, Heringa J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302(1):205-17
10. Edgar RC. (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5), 1792-97
11. Bray N and Pachter L, MAVID: Constrained ancestral alignment of multiple sequences, *Genome Research*, 14:693-699 (2004)
12. Kalign - an accurate and fast multiple sequence alignment algorithm. Lassmann T. and Erik L.L. Sonnhammer (2005) *BMC Bioinformatics*, 6: 298
13. Kazutaka Katoh, Kazuharu Misawa¹, Kei-ichi Kuma and Takashi Miyata
MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform *Nucleic Acids Research*, 2002, Vol. 30, No. 14 3059-3066
14. Rosenberg, M. S. (2005) Multiple sequence alignment accuracy and evolutionary distance estimation. *BMC Bioinformatics* 6:278

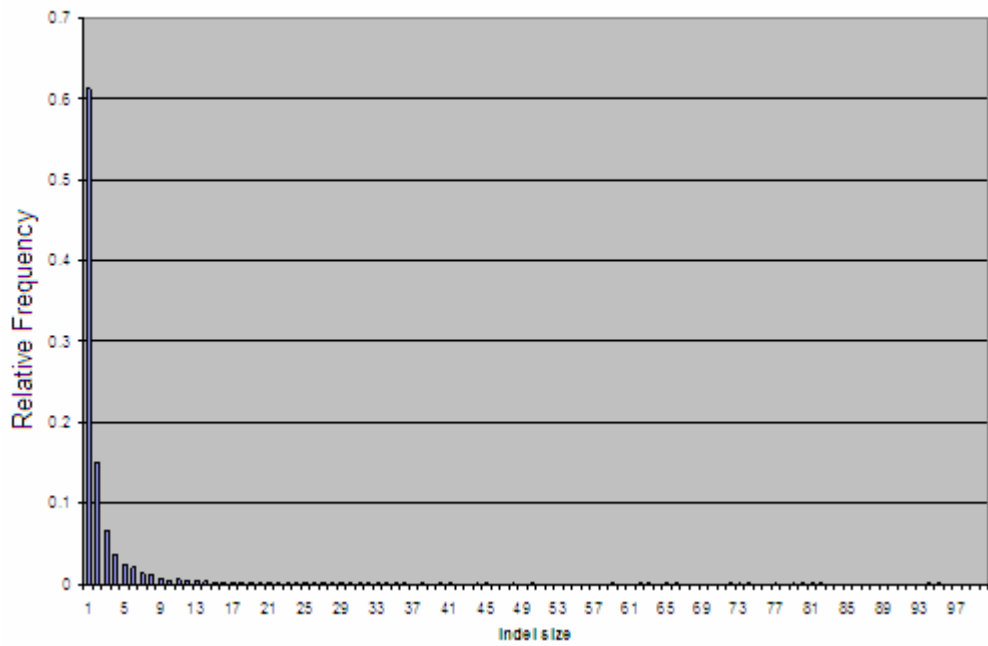
Appendices

Appendix A

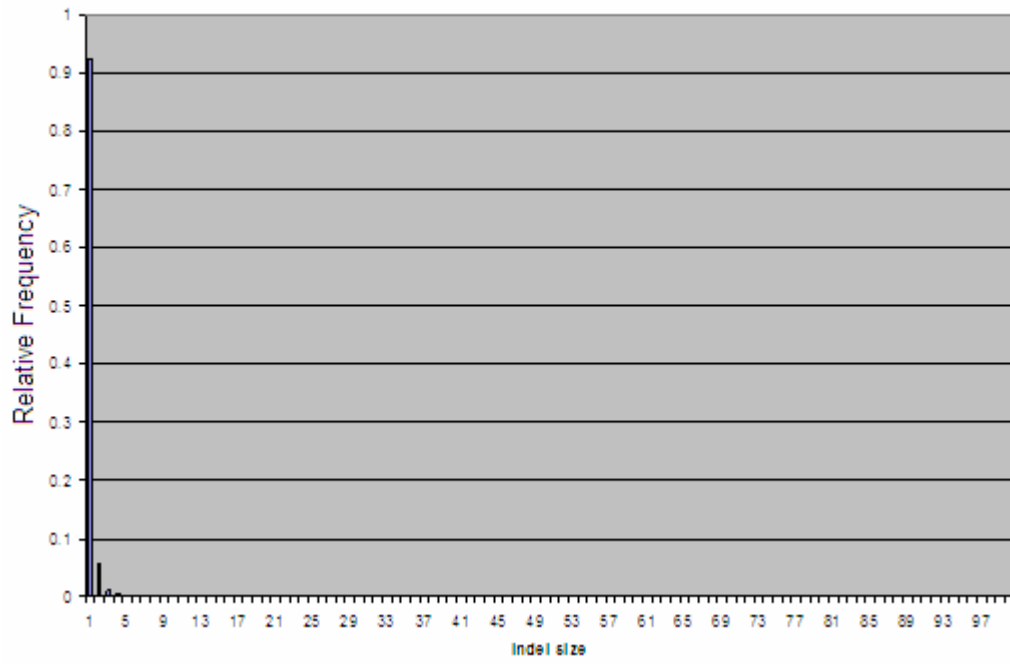
Histogram Exponential Rate -1 R0.5 K3.6



Histogram Exponential Rate -2 R0.5 k3.6

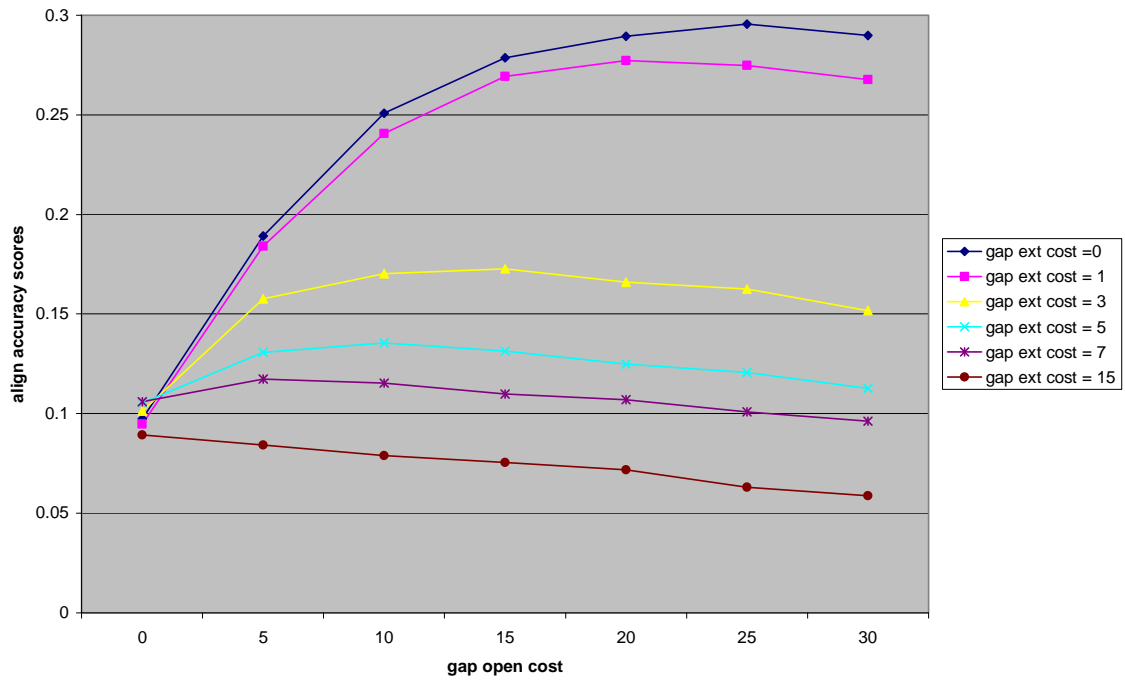


Histogram Exponential -4 R0.5 k3.6

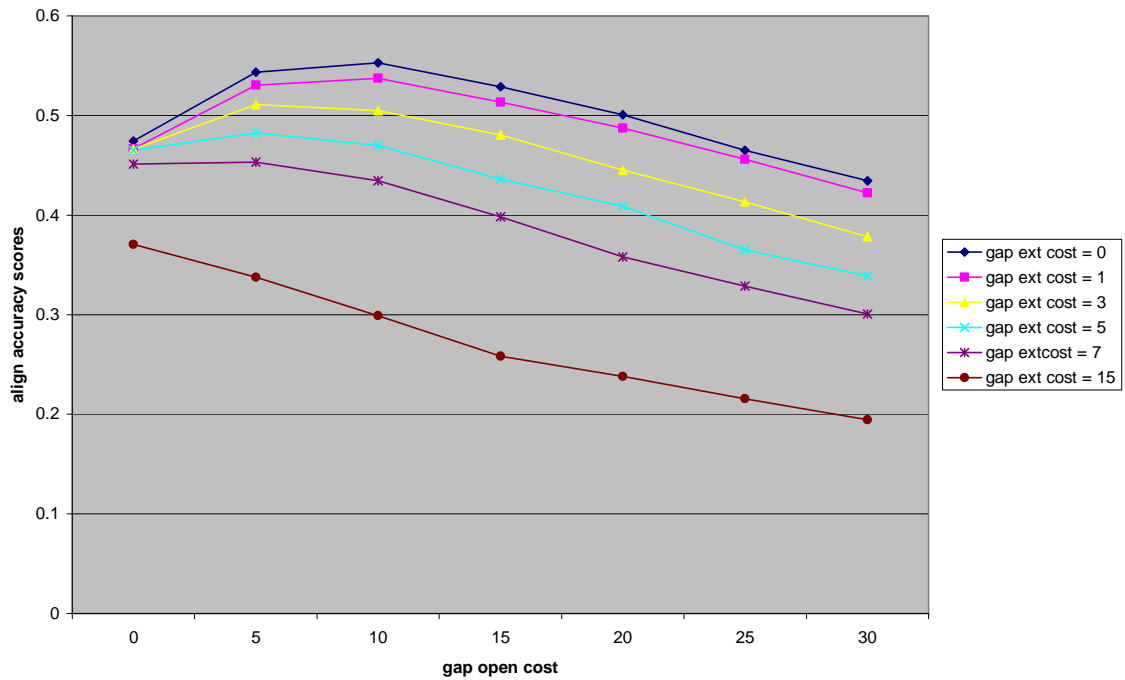


Appendix B

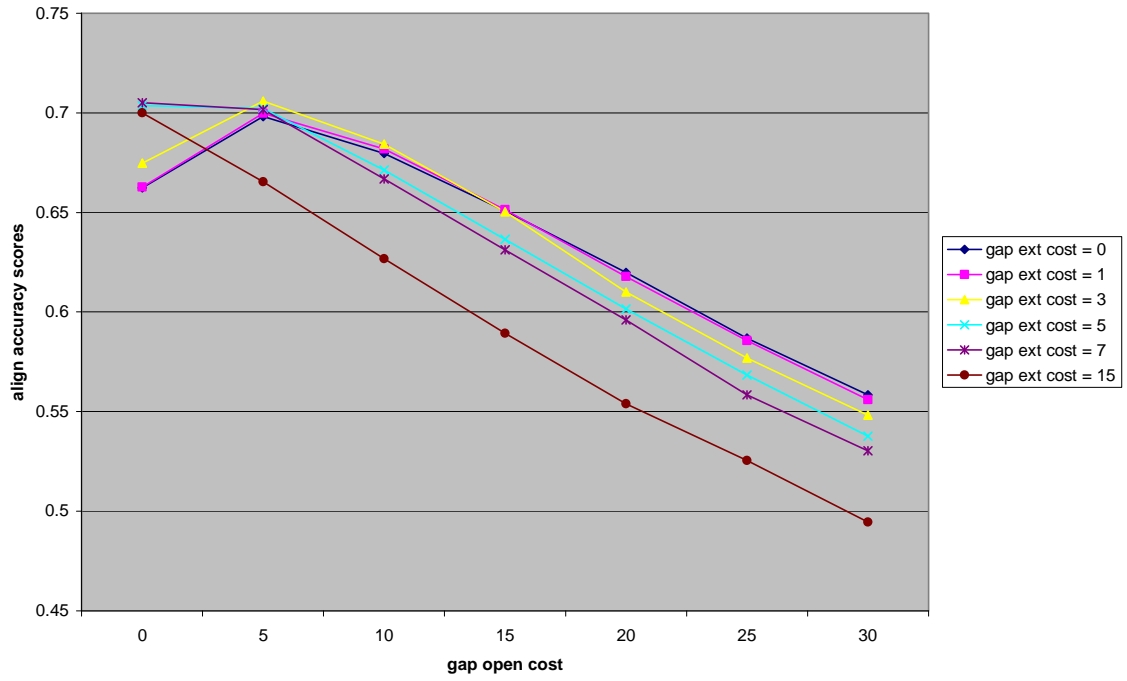
Comparisons Power Exponential -1 indel 6



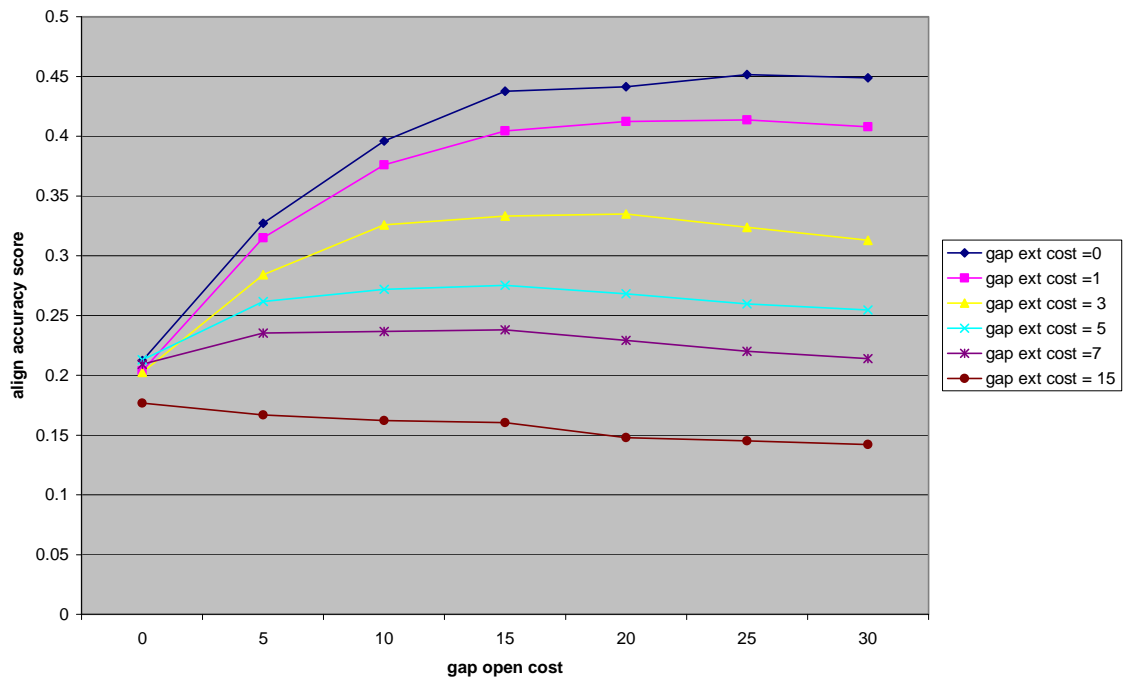
Comparisons Power Exponential -2 indel 6



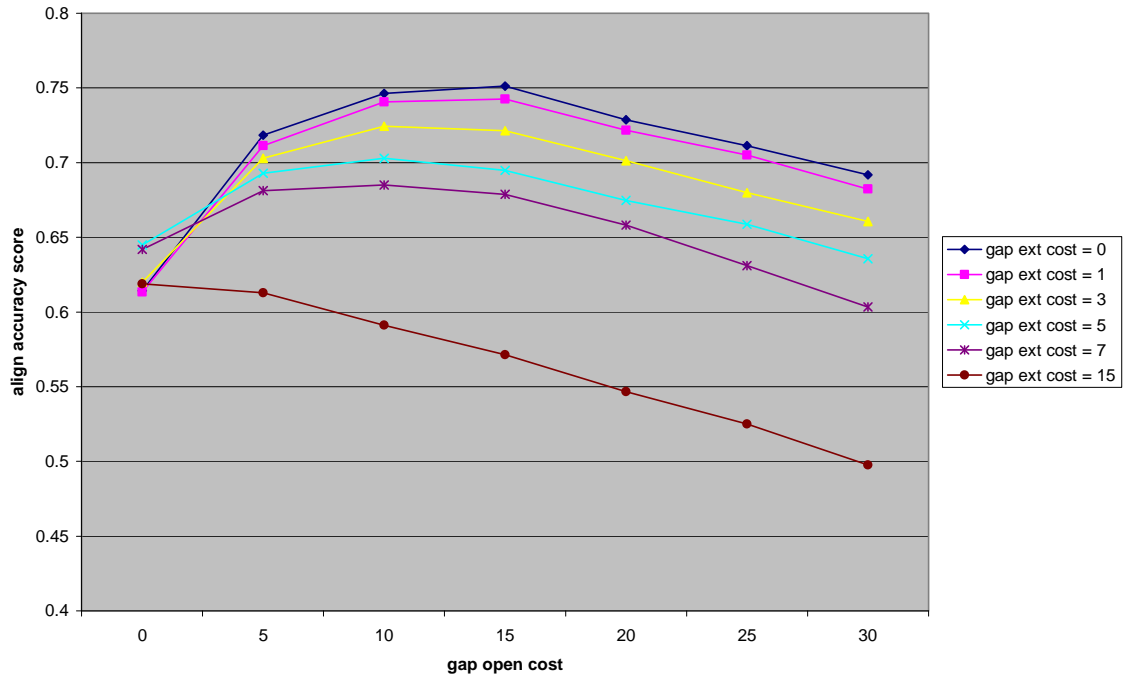
Comparisons Power Exponential -4 indel 6



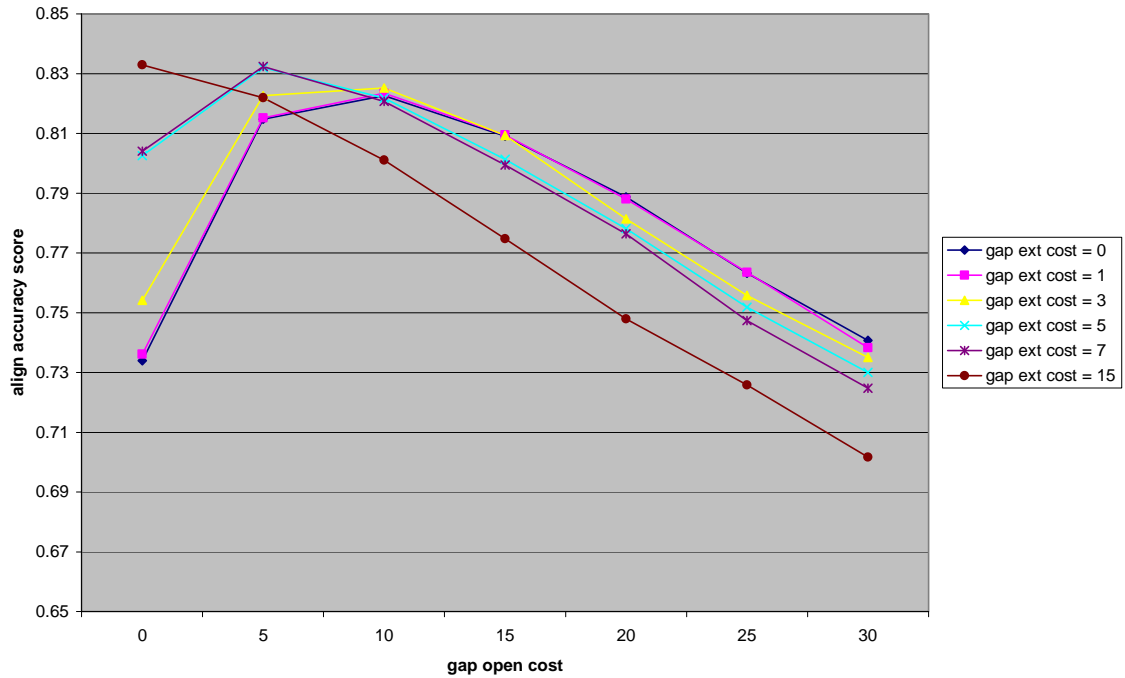
Comparisons Power Exponential -1 Indel Rate 12



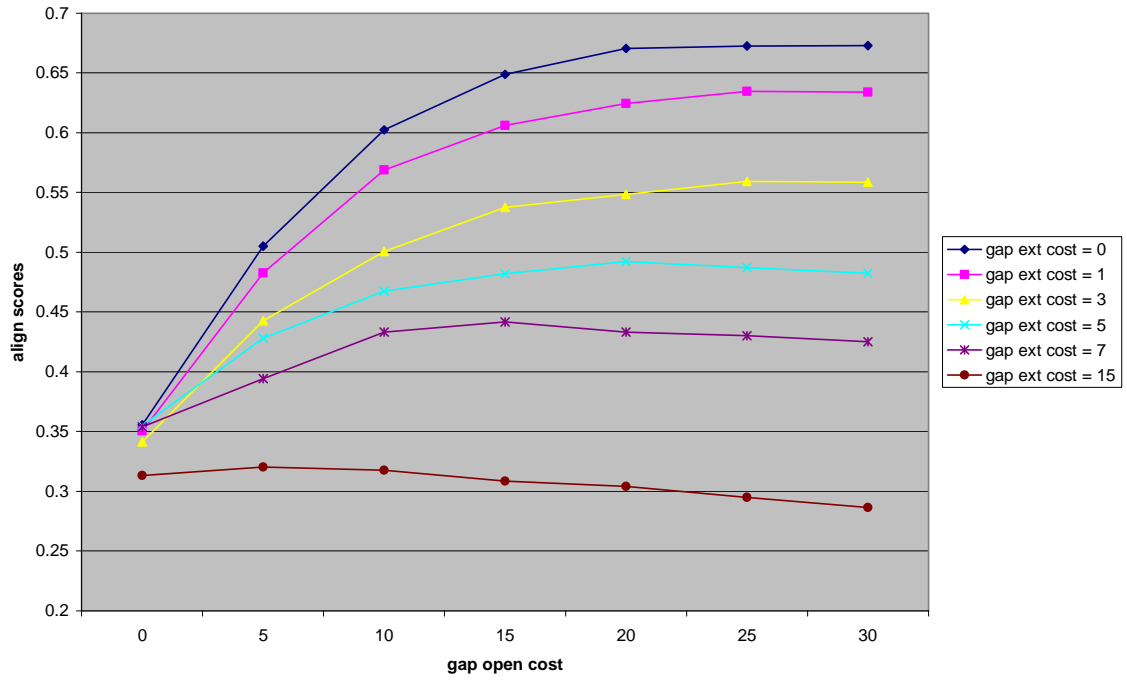
Comparisons Power Exponential -2 Indel Rate 12



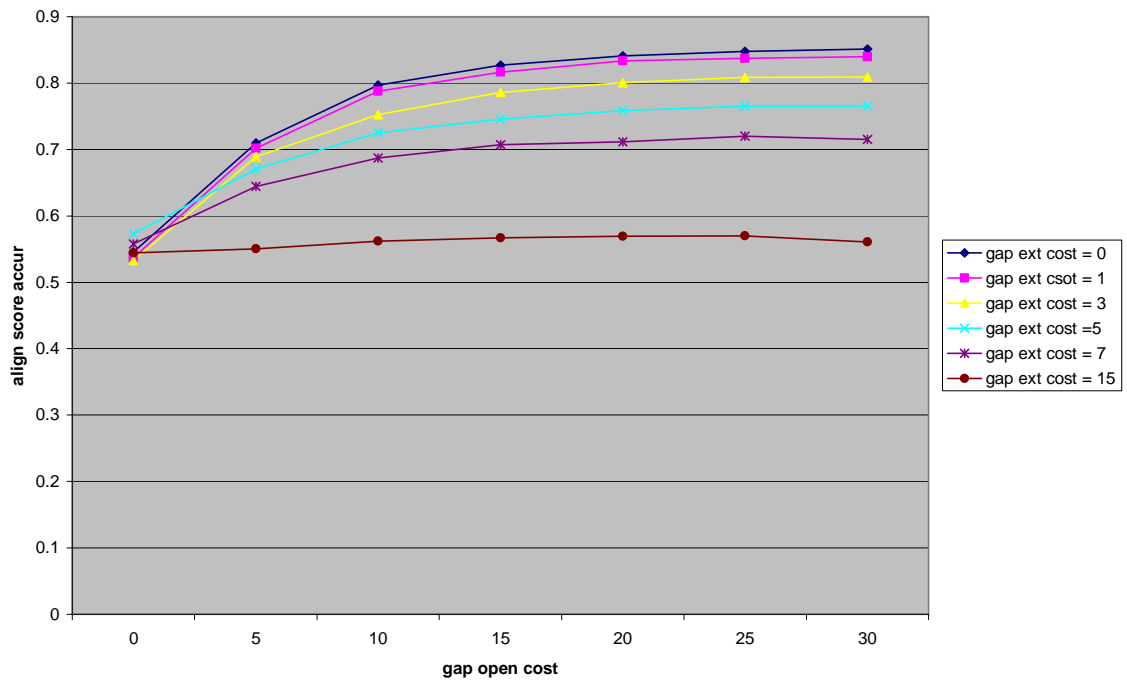
Comparisons Power Exponential -4 Indel Rate 12



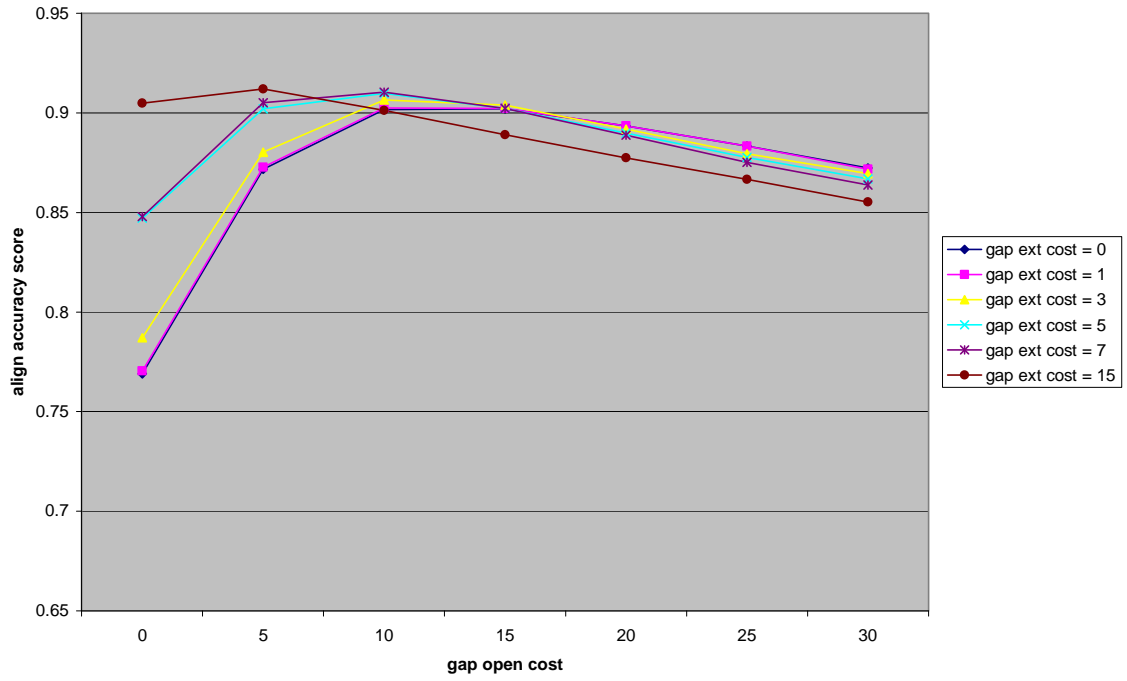
Comparisons Power Exponential -1 Indel Rate 24



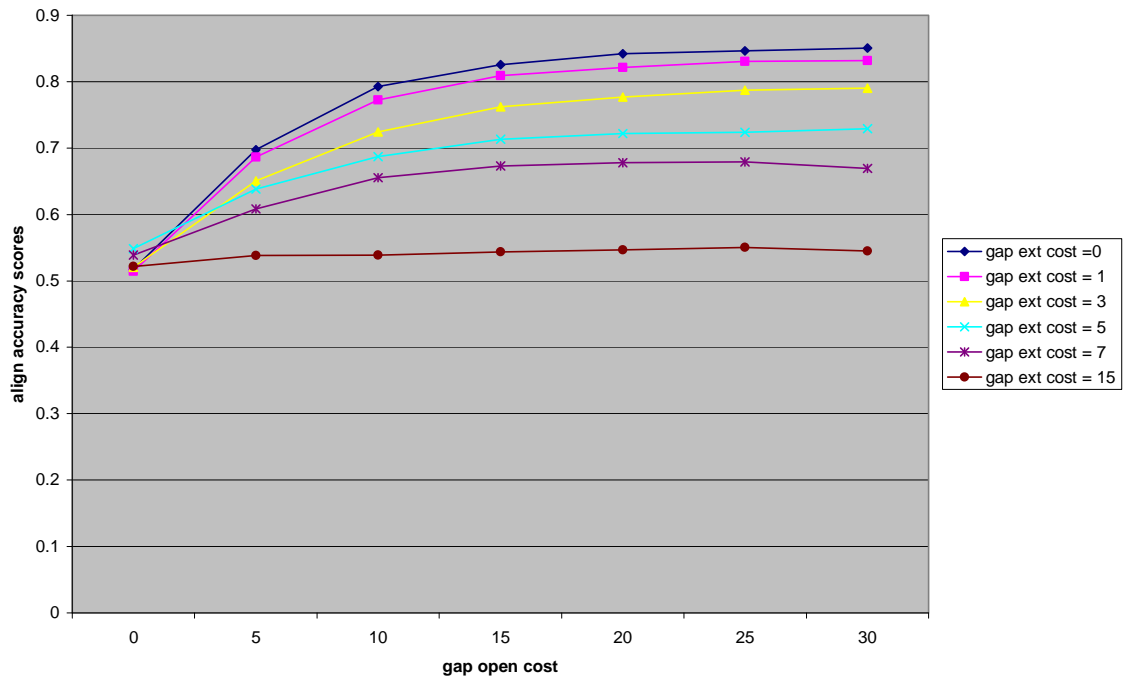
Comparisons Power Exponential -2 Indel Rate 24



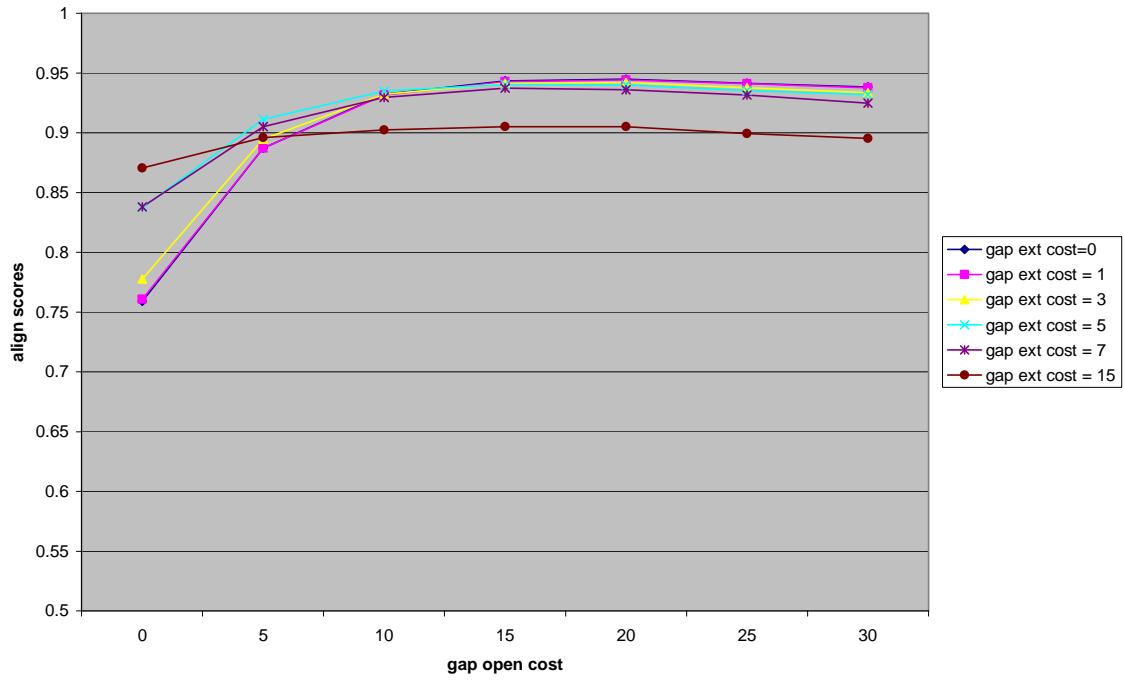
Comparisons Power Exponential -4 Indel Rate 24



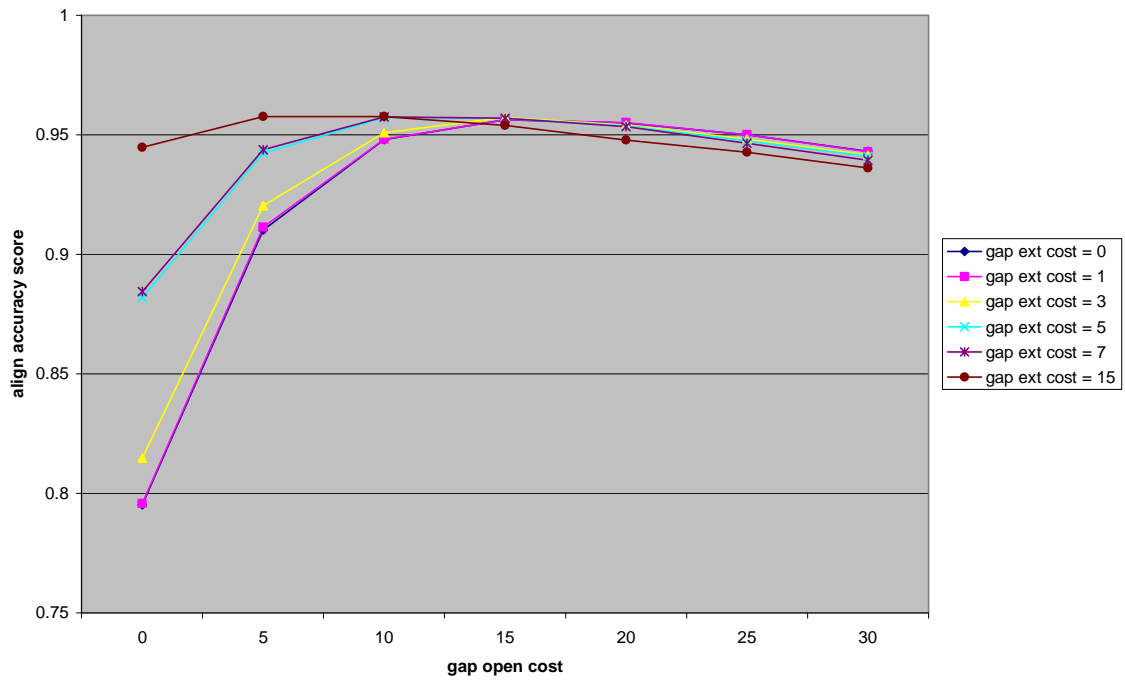
Comparisons Power Exponential -1 Indel Rate 48



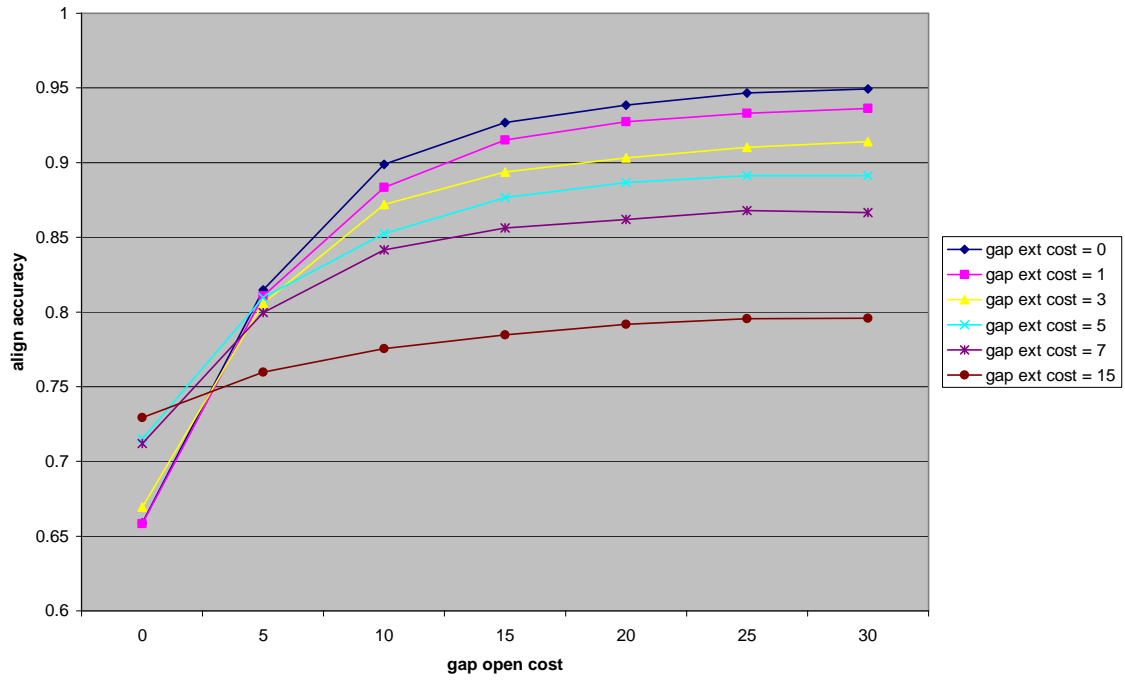
Comparisons Power Exponential -2 Indel Rate 48



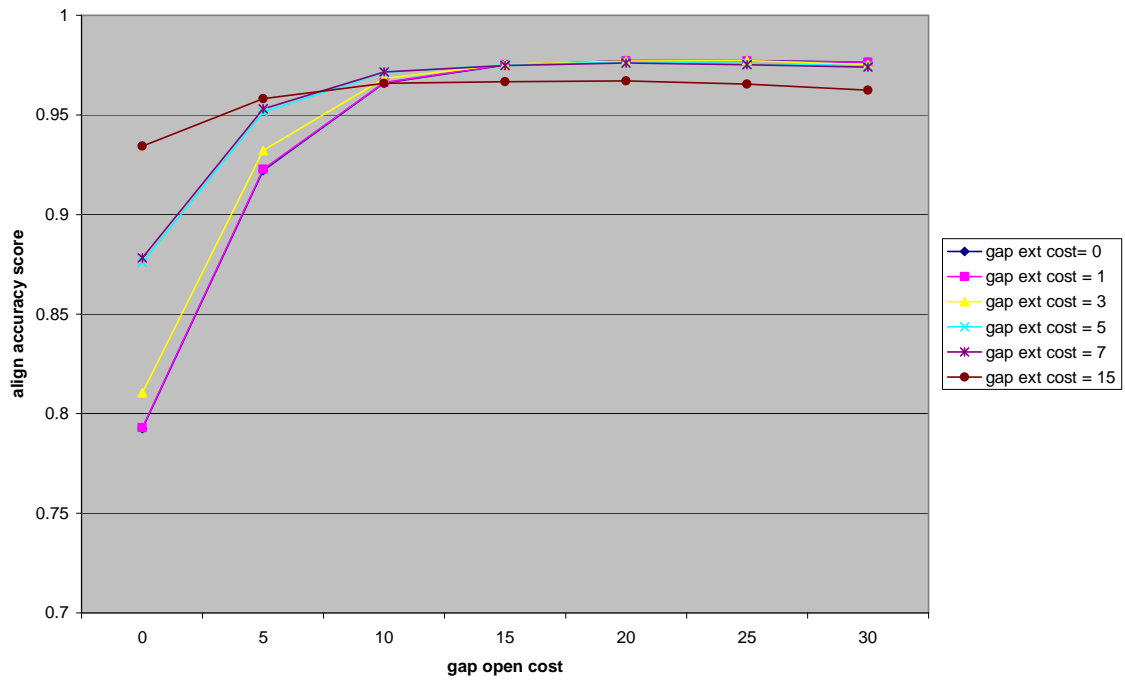
Comparisons Power Exponential -4 Indel Rate 48



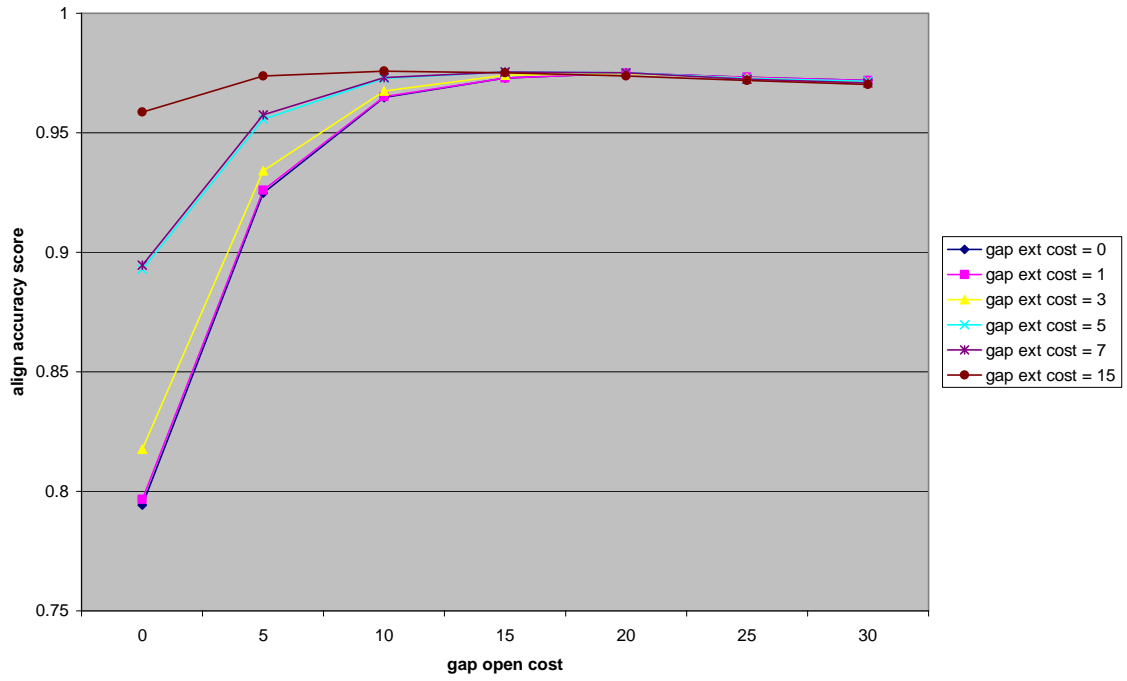
Comparisons Power Exponential -1 Indel Rate 96



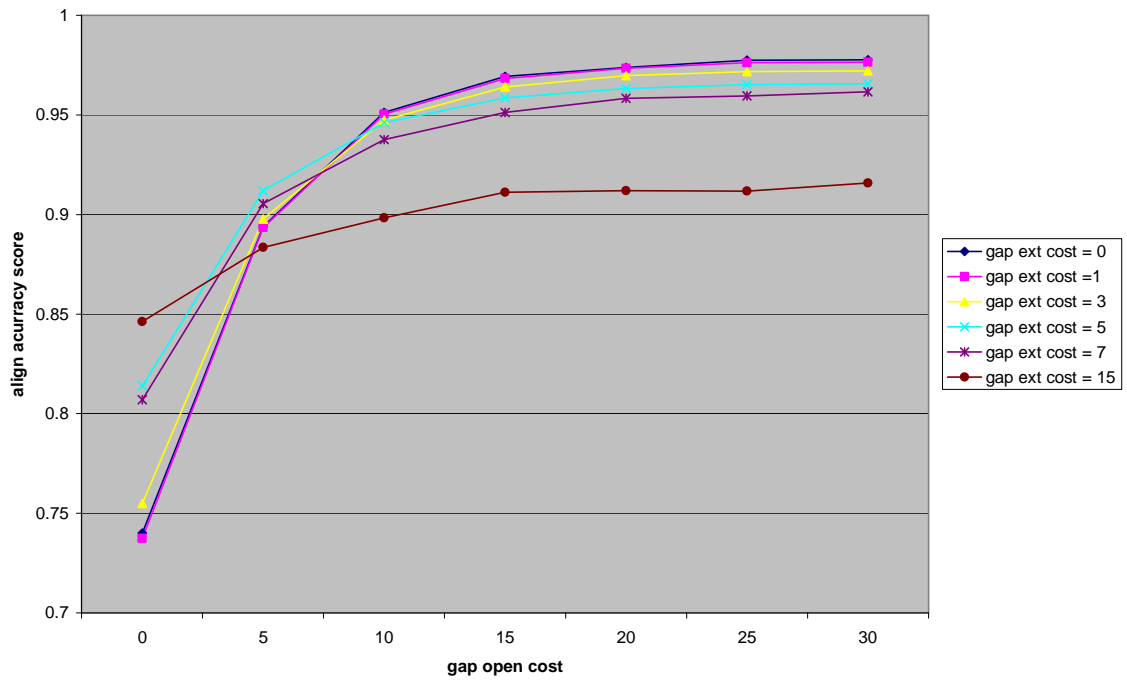
Comparisons Power Exponential -2 Indel Rate 96



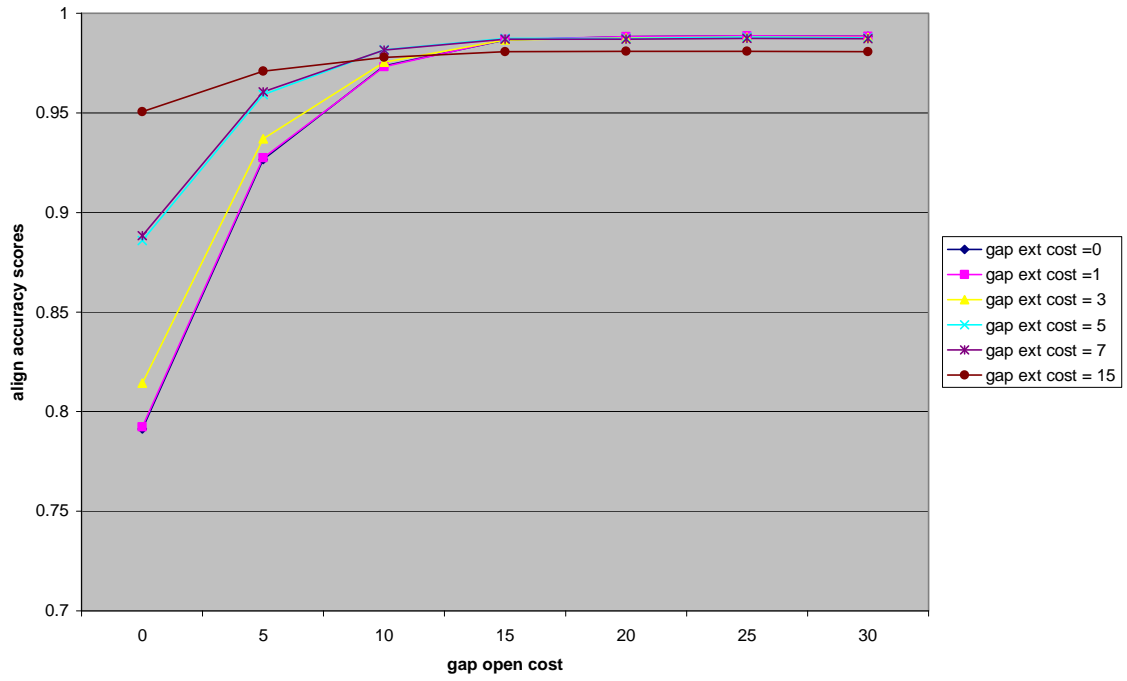
Comparisons Power Exponential -4 Indel Rate 96



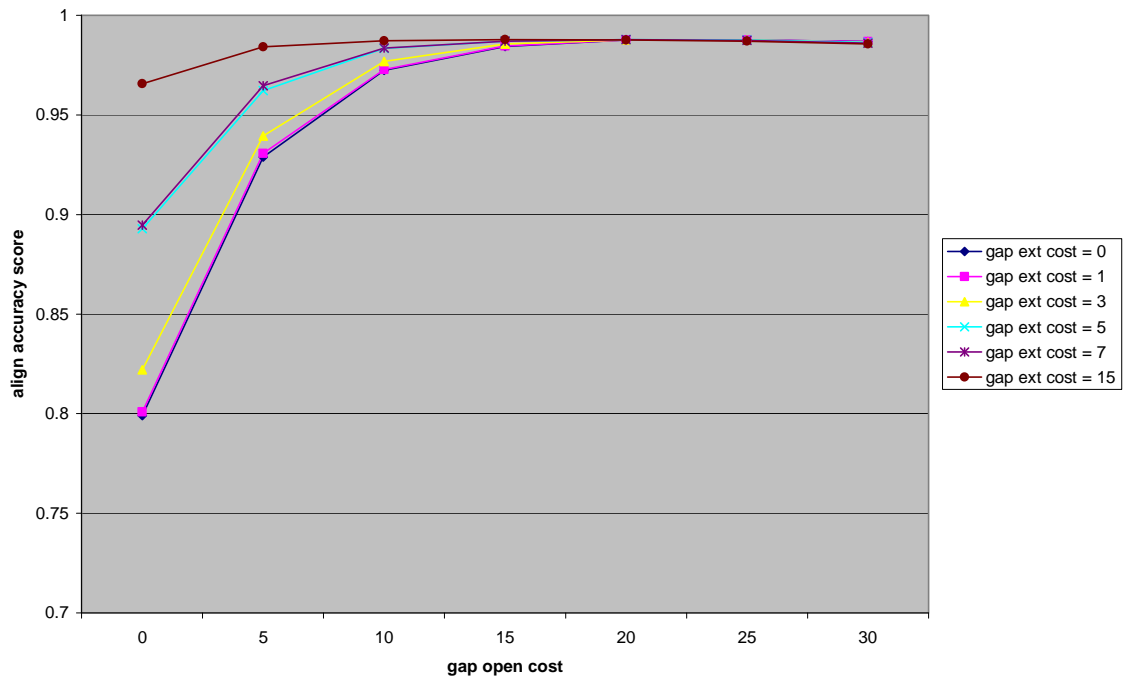
Comparisons Power Exponential -1 Indel Rate 192



Comparisons Power Exponential -2 Indel Rate 192



Comparisons Power Exponential -4 Indel Rate 192



Appendix C

```
sub calcAccr
{
    my($true1, $true2, $algn1, $algn2, $ext) = @_ ;

    $true1 =~ tr/a-z/A-Z/;
    $true2 =~ tr/a-z/A-Z/;
    $algn1 =~ tr/a-z/A-Z/;
    $algn2 =~ tr/a-z/A-Z/;

    my(@trueHom, $ratio, $i, $s1, $s2, $saCount, $totalCount, $sidenCount, $sp1Count, $sp2Count, $sqCount);
    $s1 = $s2 = $saCount = $totalCount = $sidenCount = $sp1Count = $sp2Count = $sqCount = 0;
    #working on the true sequences
    for ($i=0 ; $i < length $true1 ; $i++)
    {
        if (substr($true1, $i, 1) ne '-')
        {
            $s1++;
        }
        if (substr($true2, $i, 1) ne '-')
        {
            $s2++;
        }
        if ((substr($true1, $i, 1) ne '-') && (substr($true2, $i, 1) ne '-'))
        {
            $trueHom[$s1] = $s2;
        }
    }
    $s1 = $s2 = 0;
    #working on the aligned sequences
    for ($i=0 ; $i < length $algn1 ; $i++)
    {
        if (substr($algn1, $i, 1) ne '-')
        {
            $s1++;
        }
        if (substr($algn2, $i, 1) ne '-')
        {
            $s2++;
        }
        if ((substr($algn1, $i, 1) ne '-') && (substr($algn2, $i, 1) ne '-'))
        {
            $saCount++;
            if ( "$trueHom[$s1]" eq "$s2" )
            {
                $totalCount++;
            }
        }
    }
    if ( $saCount == 0 )
    {
        $ratio = -999;
    }
    else
    {
        $ratio = $totalCount / $saCount ;
    }
    return ($ratio);
}
```