

Estimating ClustalW pairwise alignment parameters for non-Coding DNA sequences

Meraj Aziz

Dr Rosenberg's Lab

Computational Biosciences Program

Outline

- Vocabulary
- Overview Sequence Alignment
- Variety of alignment software
- ClustalW
- Project background
- Project goal
- Software Used
- Work Flow
- Parameters
- Results
 - Gap opening and extension cost
 - Transition weights
- Conclusion
- Future work

Vocabulary

- **Rate:** determines how many substitutions occur on each branch of the phylogenetic tree and allows for rescaling of the divergence between the sequences
- **Kappa:** transition/transversion ratio
- **Gamma:** parameter for rate variation.
- **Indel:** Insertion/deletion
 - **Indel Exponent:** determines the length of Indels.
 - **Indel Frequency/Rate:** how often Indels occur in a sequence.
- **Transweights:** ClustalW parameter indication Transition weights. Range between 0→1.
- **HKY + Γ :** Model of substitution with gamma (for rate variation)
- **“True” sequences:** Sequences generated by MySSP. (benchmark)
- **Aligned sequences:** Sequences aligned with ClustalW

Genome Sequencing

- 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.

There are several advantages to sequencing the genomes of organisms:

- First-- provides an opportunity to compare genomes and to find similarities/differences among different organisms.
- 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.

Sequence Alignment

- As a result of Genome Sequencing, a substantial amount of genetic data has been generated.
- Very short or very similar sequences can be aligned by hand;
- Interesting problems require the alignment of lengthy, highly variable or extremely numerous sequences that cannot be aligned solely by human effort.

- **Solution**

Construct algorithms to produce high-quality sequence alignments

Variety of Alignment Software

- Many Sequence alignment programs are available on line:
 - Multiple Sequence Alignment (MSA)
 - T-Coffee
 - MUSCLE
 - MAVID
 - kalign
 - MAFFT
 - ClustalW

ClustalW Web Interface

- **ClustalW**: command line interface
- **ClustalX**: graphical user interface.

The image shows a screenshot of the ClustalW web interface, which is a graphical user interface for the ClustalW command-line program. The interface is organized into several sections, each with a title and a set of controls:

- YOUR EMAIL**: A text input field.
- ALIGNMENT TITLE**: A text input field containing "Sequence".
- RESULTS**: A dropdown menu set to "interactive".
- ALIGNMENT**: A dropdown menu set to "full".
- KTUP (WORD SIZE)**: A dropdown menu set to "def".
- WINDOW LENGTH**: A dropdown menu set to "def".
- SCORE TYPE**: A dropdown menu set to "percent".
- TOPDIAG**: A dropdown menu set to "def".
- PAIRGAP**: A dropdown menu set to "def".
- MATRIX**: A dropdown menu set to "def".
- GAP OPEN**: A dropdown menu set to "def", circled in red.
- END GAPS**: A dropdown menu set to "def".
- GAP EXTENSION**: A dropdown menu set to "def", circled in red.
- GAP DISTANCES**: A dropdown menu set to "def", circled in red.
- OUTPUT**:
 - OUTPUT FORMAT**: A dropdown menu set to "aln w/numbers".
 - OUTPUT ORDER**: A dropdown menu set to "aligned".
- PHYLOGENETIC TREE**:
 - TREE TYPE**: A dropdown menu set to "none".
 - CORRECT DIST.**: A dropdown menu set to "off".
 - IGNORE GAPS**: A dropdown menu set to "off".

Background

- One of the issues with Sequence alignment parameters is with Initial starting parameters.
- Different DNA sequences have different evolutionary nature (mammalian or primate sequences may be different from those for yeast).
- Each require different ClustalW parameters to result in optimal alignment.
- Usually it's a matter of trial and error.
- Or user can use the **default values**.
- It may be useful to **customize** one's own parameters.

Background

- In order to customize the user needs to be aware of the evolutionary nature of the sequences and the corresponding ClustalW parameters.



Project Goal

Approximating ClustalW parameters for pairwise non-coding DNA sequences

Assumption

- User has some prior knowledge of the evolutionary nature of their sequences.

Software Used

The screenshot displays the MySSP (1.1.1.7) software interface. A central window titled "Tree" is open, showing a blank tree structure. The main interface is divided into several sections:

- Model Tree:** Contains a "Load" button with a green checkmark and the filename "mytree.txt". Below it are "Display" and "List Branches" buttons.
- Number of Genes:** A text box containing "1" and a "Create" button with a green checkmark.
- Gene #:** A dropdown menu showing "1".
- Initial Sequence:** A section with a "# of Sites" field containing "1000".
- Substitution Model:** A group box containing radio buttons for "Jukes-Cantor", "Kimura's Two-Parameter", "Equal Input", "HKY" (selected), and "General Reversible".
- Rate and Kappa:** Input fields for "Rate" (0.5) and "Kappa" (3.600000).
- Gamma Distribution:** A checked checkbox and an "Alpha" field (0.50000).
- Equilibrium Frequencies:** A table with columns for A, G, C, T and rows for AC, AG, AT, CG, CT, GT, all with values of 1.000.
- Power Model Parameters:** Two "Power" radio buttons are selected. Each has a "Constant" field (4.00000) and an "Exponent" field (-4.00000).
- Include Deletions:** A checked checkbox.
- Size Distribution:** A group box with a "Poisson" radio button selected.
- Rate (per substitution):** A field containing "48.00000".
- Buttons:** "Clear", "Run" (with green checkmark), "Batch" (with refresh icon), and "Close" (with window icon) buttons are visible.

Red circles highlight the following elements:

- The "Load" button and "mytree.txt" text.
- The "Rate" field (0.5).
- The "Kappa" field (3.600000).
- The "Gamma Distribution" checkbox.
- The "Alpha" field (0.50000).
- The "HKY" radio button.
- The "Power" radio button in the upper section.
- The "Constant" (4.00000) and "Exponent" (-4.00000) fields for the upper Power model.
- The "Rate (per substitution)" field (48.00000).
- The "Power" radio button in the lower section.
- The "Constant" (4.00000) and "Exponent" (-4.00000) fields for the lower Power model.

HKY Substitution Model

		To			
		A	T	C	G
From	A	–	βf_T	βf_C	αf_G
	T	βf_A	–	αf_C	βf_G
	C	βf_A	αf_T	–	βf_G
	G	αf_A	βf_T	βf_C	–

α is the probability of a transitional change

β is the probability of a transversional change

f_X is the expected frequency of nucleotide X

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

****Gamma distribution:** good approximation for rate variation among sites

Evolutionary Models: Rate Heterogeneity

Gamma Distribution

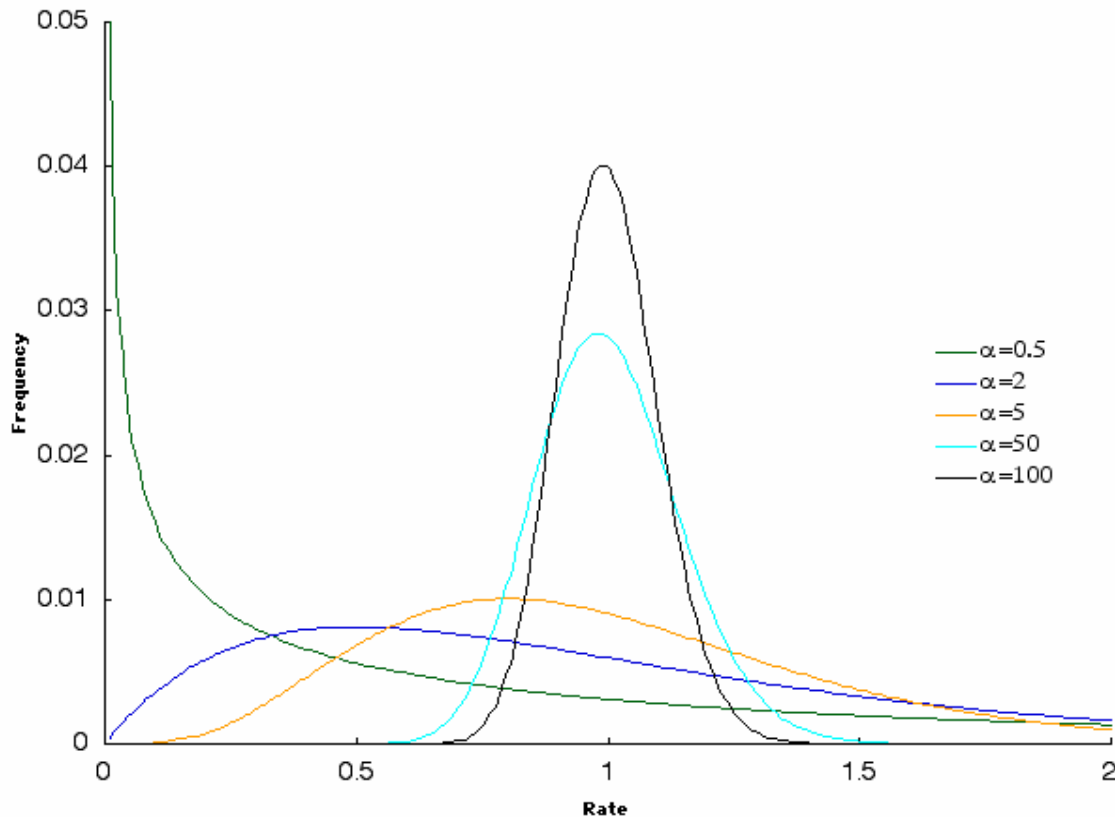
$$f(r) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta r} r^{\alpha-1}$$

r = rate of substitution

$$\alpha = \bar{r}^2 / V(r)$$

$$\beta = \bar{r} / V(r)$$

$$\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1} dt$$

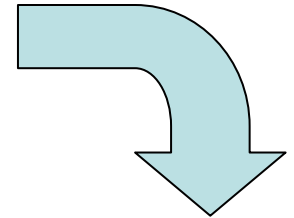
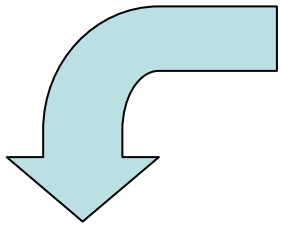
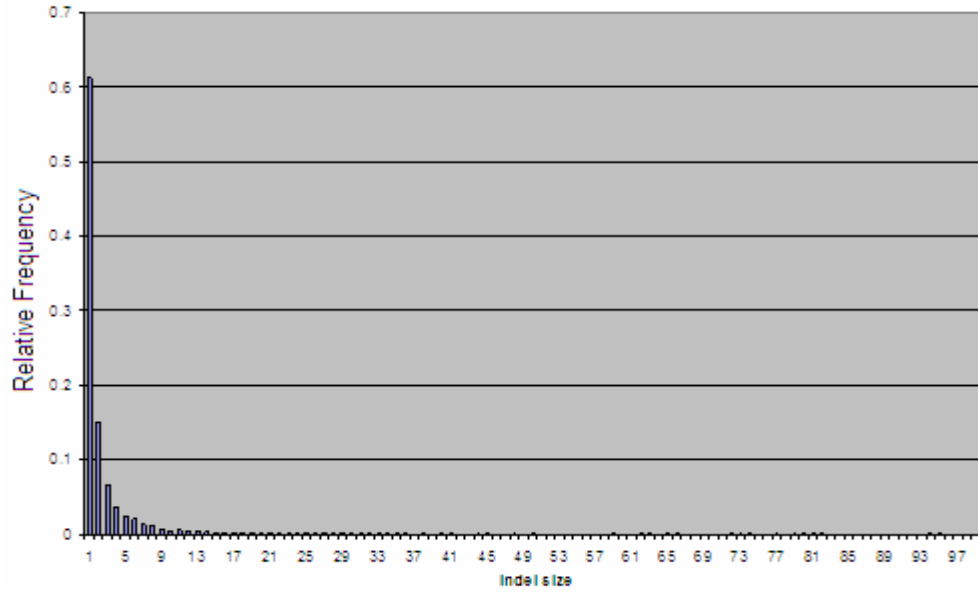


$\alpha = \infty$ rates are equal

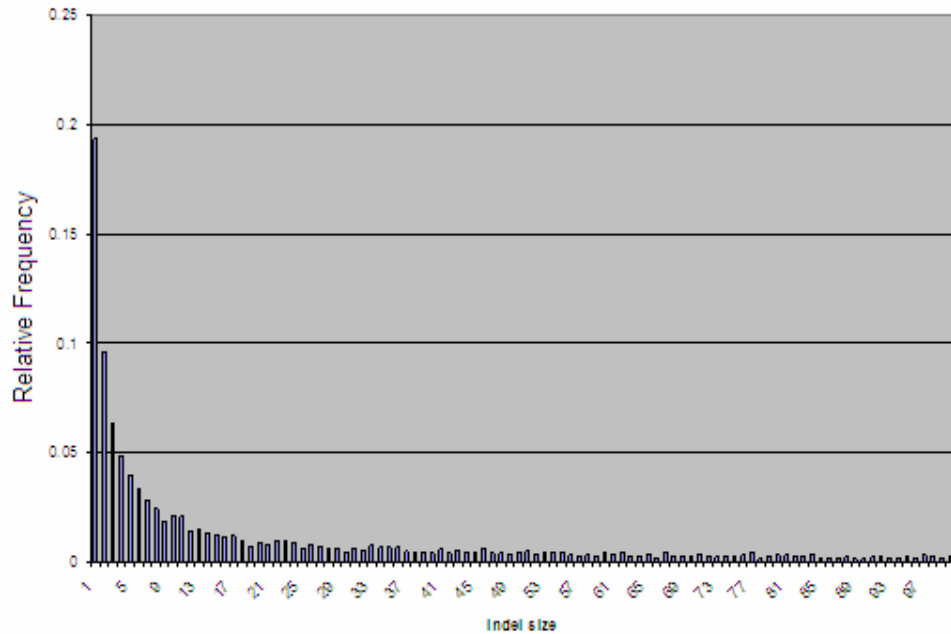
$\alpha < 1$ rates are very heterogenous

Indel Size Distribution

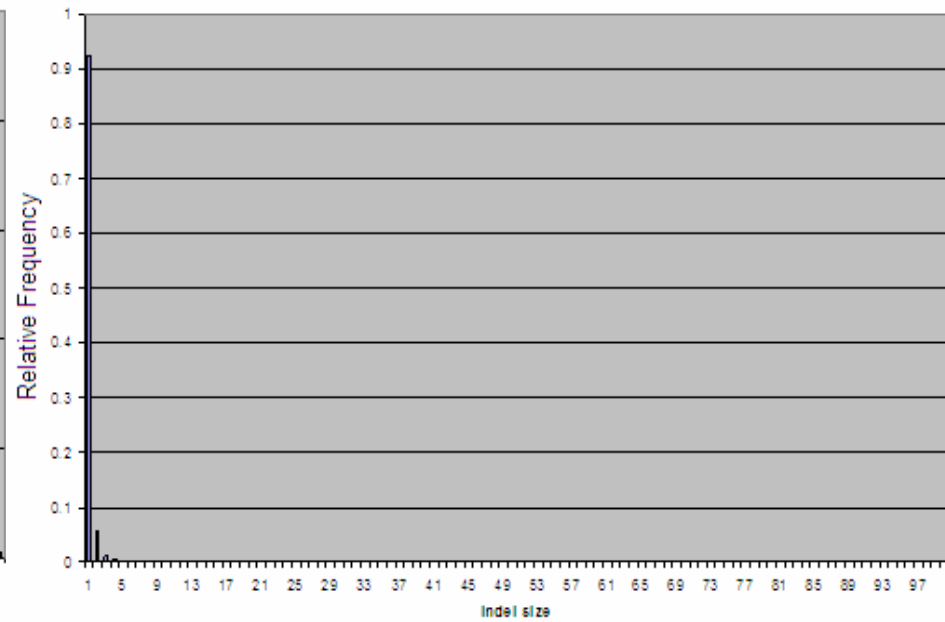
Histogram Exponential Rate -2 R0.5 k3.6



Histogram Exponential Rate -1 R0.5 K3.6



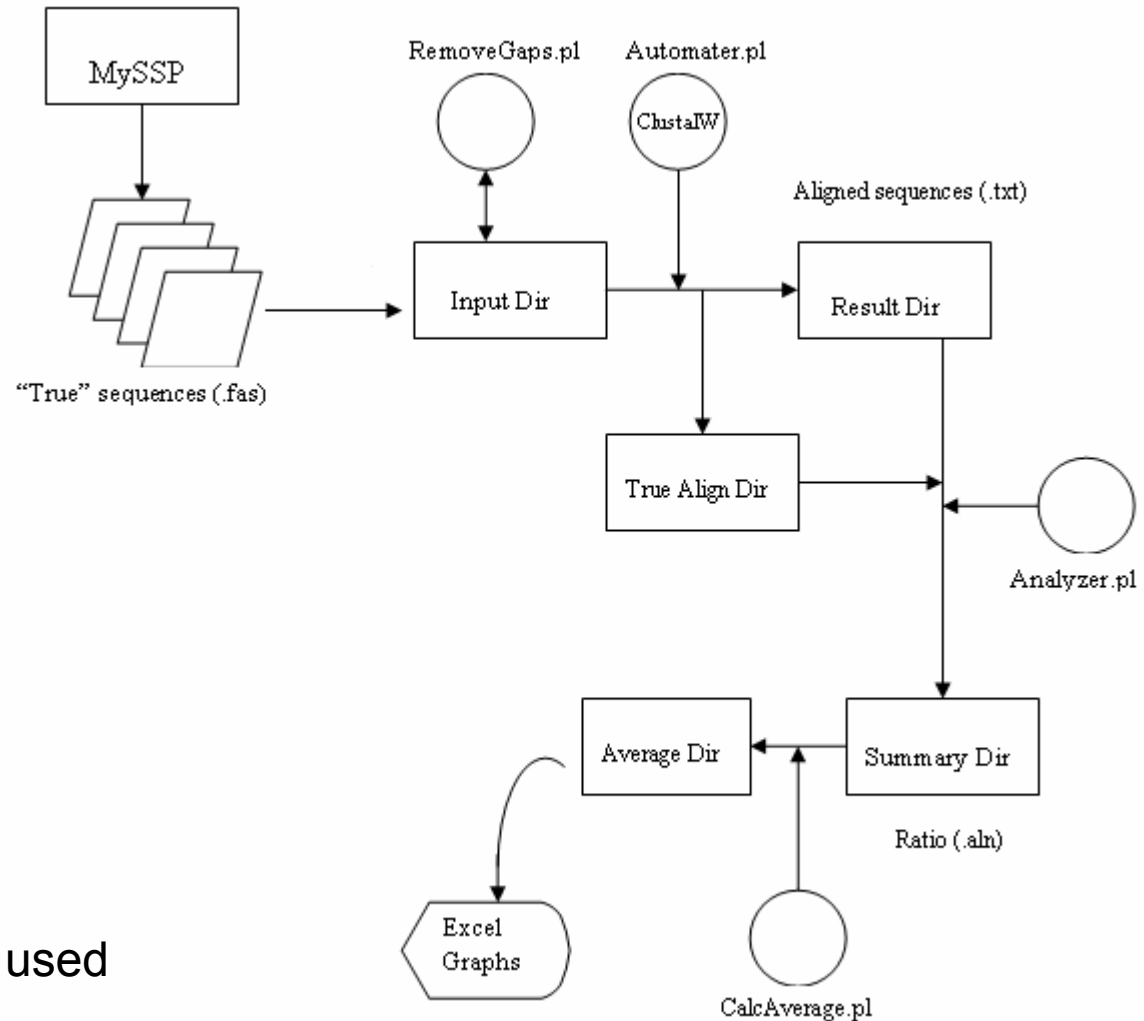
Histogram Exponential -4 R0.5 k3.6



Initial Parameters

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

Defining the data pipeline



Perl and Bash Scripting used

How is the Ratio Computed?

True alignment

1: TACCAT-CAGGG
2: TCCG-TCCAGAG

Hypothesized alignment

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

- We are measuring alignment accuracy as the proportion of aligned sites in the hypothesized alignment which are aligned identically in the true alignment.
- 12 nucleotides long
- We ignore gaps therefore 10 pairs
- Incorrect sites marked with *'s
- $6/10 = 0.6$ alignment accuracy score

Parameters Tried

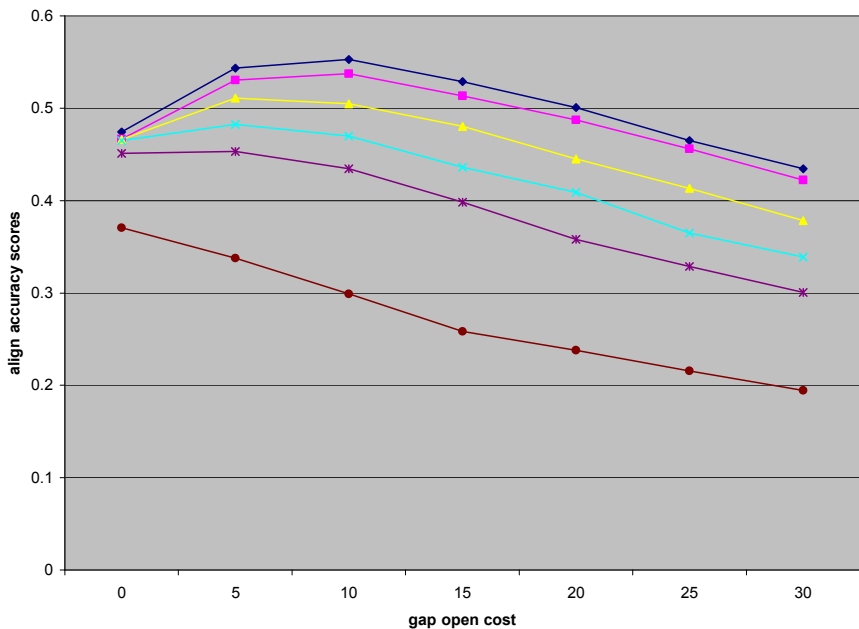
MySSP			Clustalw			
Frequency	Rate	Kappa	Power exp	Gap cost	Transweights	
6	0.5	3.6	-1	x		
			-2	x		
			-4	x		
12	0.125	1.8			0.0--> 1.0	(0.1 incr)
		3.6			0.0--> 1.0	
		7.2			0.0--> 1.0	
		15			0.0--> 1.0	
	0.25	1.8			0.0--> 1.0	
		3			0.0-->0.4	(0.25 incr)
		3.2			0.0-->0.4	
		3.4			0.0-->0.4	
		3.6		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	0.0-->0.4
			-2	x		0.0-->0.4
			-4	x		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	
	0.75	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		

		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]

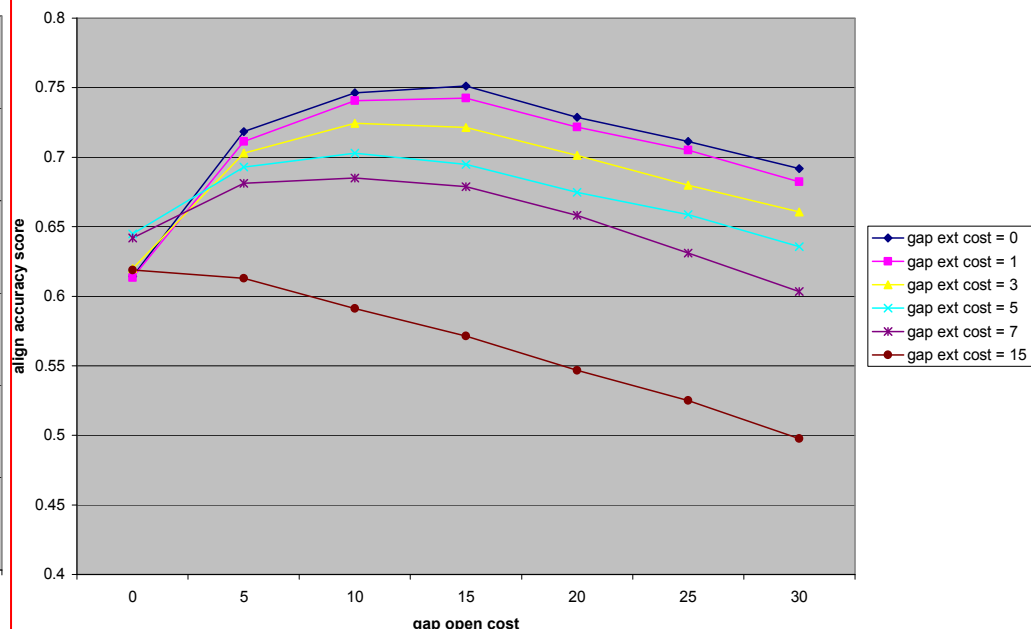
**Default parameters are marked in red

Results

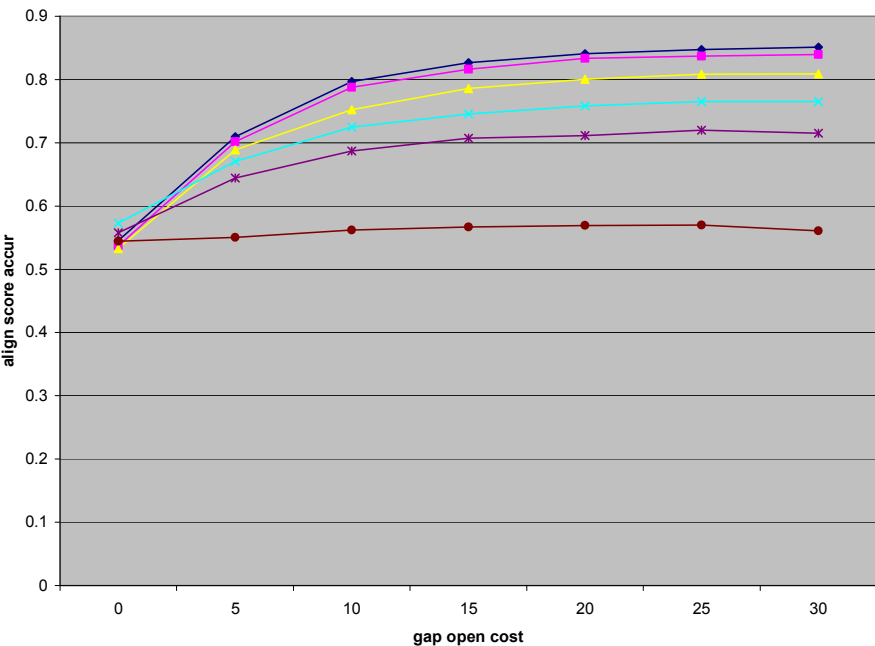
Comparisons Power Exponent -2 indel 6



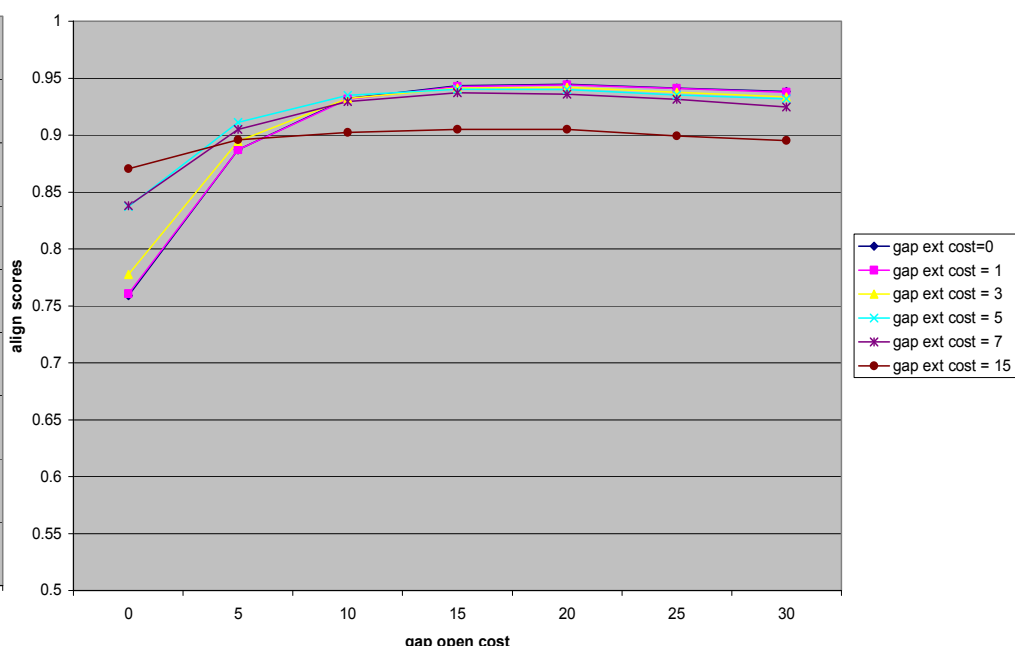
Comparisons Power Exponent -2 Indel Rate 12



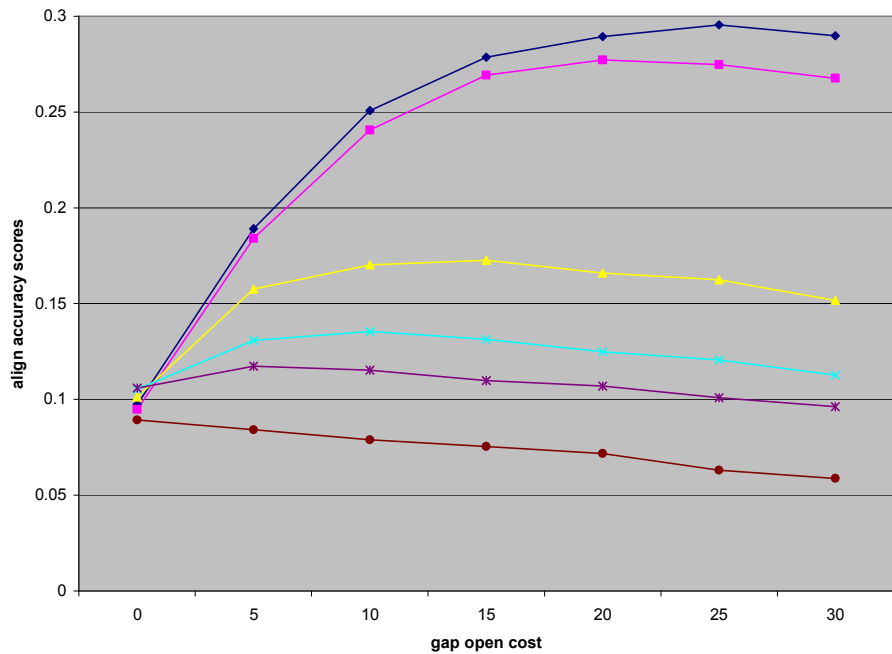
Comparisons Power Exponent -2 Indel Rate 24



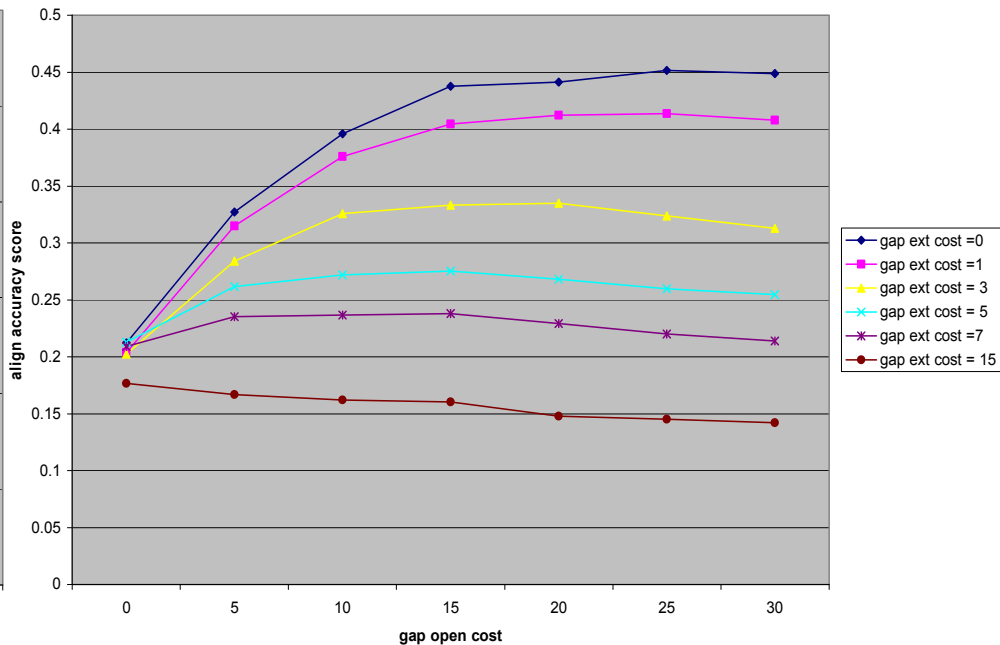
Comparisons Power Exponent -2 Indel Rate 48



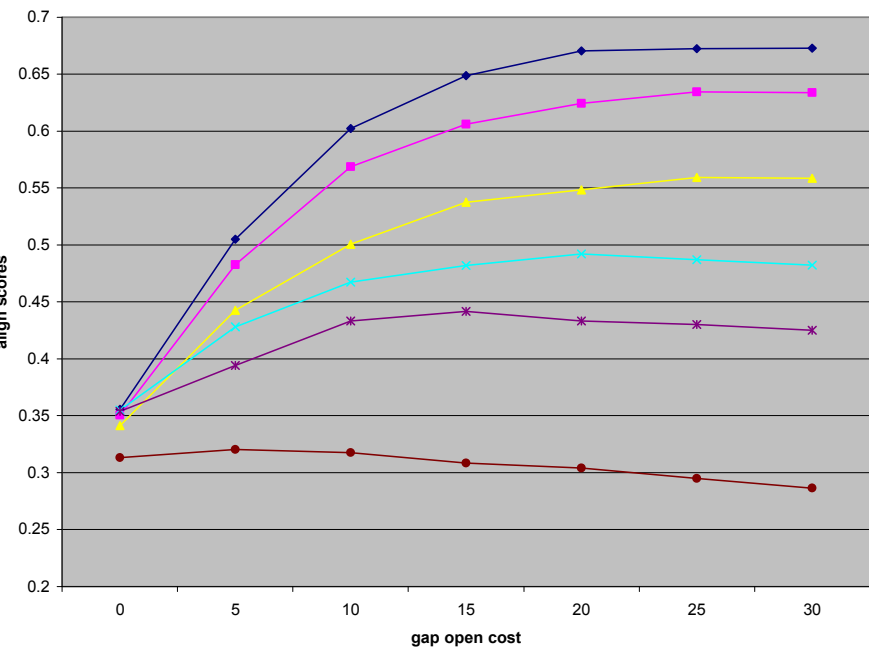
Comparisons Power Exponent -1 indel 6



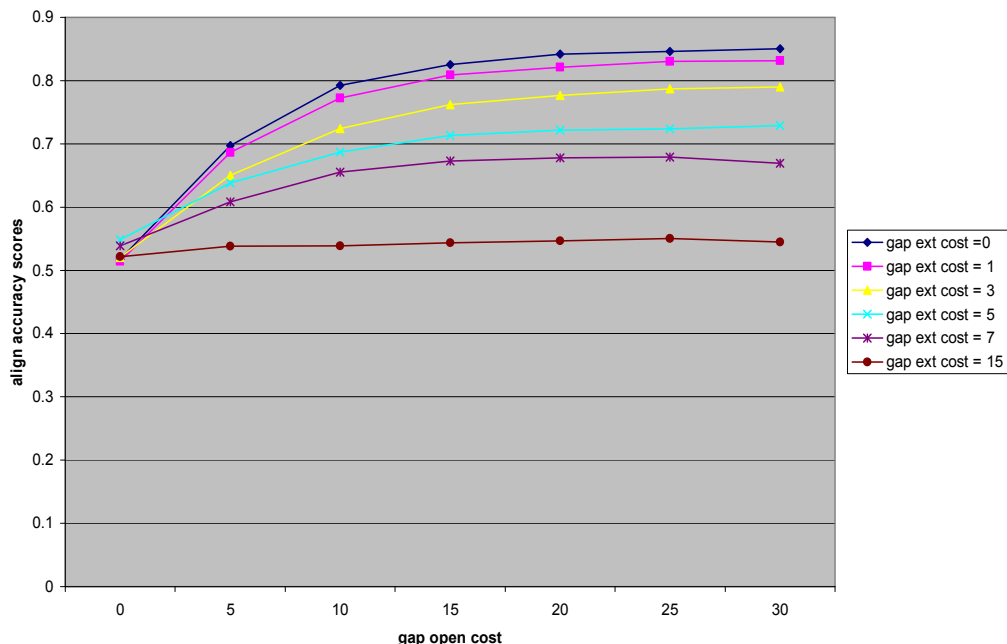
Comparisons Power Exponent -1 Indel Rate 12



Comparisons Power Exponent -1 Indel Rate 24



Comparisons Power Exponent -1 Indel Rate 48



Gap Extension and Gap Opening 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

		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

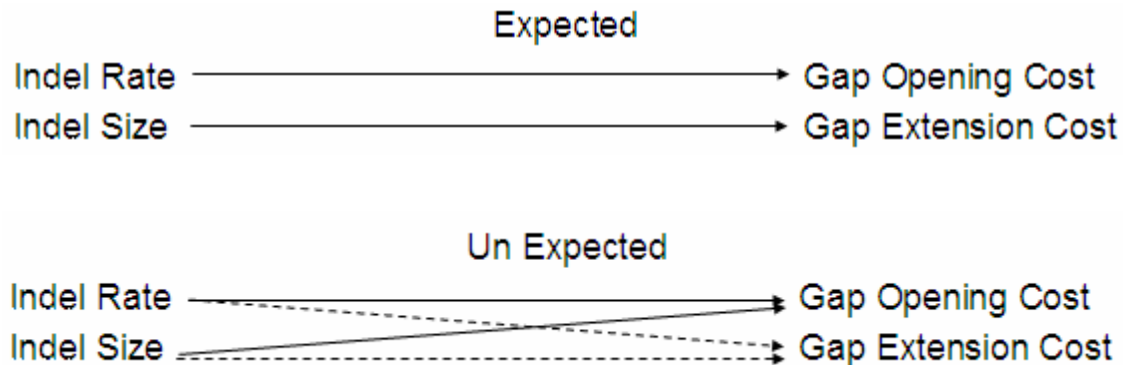
		Gap Open Cost		
		Indel Exponent -1	Indel Exponent -2	Indel Exponent -4
Indel Frequency	1/6	25	10	0
	1/12	25	15	0
	1/24	25	30	5
	1/48	30	20	5
	1/96	30	15/20/25	10
	1/192	25	15/20/25/30	20

		Gap Extension Cost		
		Indel Exponent -1	Indel Exponent -2	Indel Exponent -4
Indel Frequency	1/6	0	0	3
	1/12	0	0	15
	1/24	0	0	15
	1/48	0	0/1/3	15
	1/96	0	0/1/3/5	15
	1/192	0/1	0/1/3/5/7	0/1/5/7/15

Transition Weights

	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

Conclusion



- Implications

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

Future Work

- This report describes a preliminary step towards finding better ClustalW sequence alignment parameters.
- Future we would like to extend the analysis to include multiple sequences and different alignment programs such as
 - MSA
 - T-Coffee
 - MAFFT
 - MUSCLE
 - Kalign
 - MAVID
- The ultimate goal is to remove guess work from alignment parameter choices and to provide the users with a set of values for their particular sequences which would give them the best possible alignment.
- We would also like to include coding DNA sequences in our analysis.
- 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.

References

- 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
- Rosenberg, Michael S. "MySSP: Non-Stationary Evolutionary Sequence Simulation, Including Indels." *Evolutionary Bioinformatics Online* (2005): 1 81-83.
- 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.
- Cartwright, Reed A. "Logarithmic gap costs decrease alignment accuracy" *BMC Bioinformatics* (2006) 7:527
- 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.
- Lipman DJ, Altschul SF, Kececioglu JD. A tool for multiple sequence alignment. *Proc Natl. Acad. Sci. USA*. 1989; 86:4412-4415.
- 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
- Edgar RC. (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5), 1792-97
- [Bray N](#) and [Pachter L](#), MAVID: Constrained ancestral alignment of multiple sequences, *Genome Research*, 14:693-699 (2004)
- Kalign - an accurate and fast multiple sequence alignment algorithm. Lassmann T. and Erik L.L. Sonnhammer (2005) *BMC Bioinformatics*, 6: 298
- Kazutaka Katoh, Kazuharu Misawa, 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
- Rosenberg, M. S. (2005) Multiple sequence alignment accuracy and evolutionary distance estimation. [BMC Bioinformatics](#) 6:278