

# Local Pairwise Statistical Alignment (LPSA)

1.10.08

**Background and motivation: Statistical Alignment and Local Alignment.** Substitutions, insertions and deletions are random evolutionary events that should be modelled by a stochastic process. For a long time this was totally ignored for insertions and deletions, which were treated by an *ad hoc* method. Using a stochastic model for these events defines a class of methods called statistical alignment.

The first model for statistical alignment was proposed for in 1986 by Bishop and Thompson. In 1991 the model was proposed that provides the basis for most statistical alignment today by Thorne, Kishino and Felsenstein. Although statistical alignment for a long time had the reputation of being very difficult, it can actually be solved by an algorithm that is very similar to the traditional optimisation alignment.

The most famous method for finding local alignment was published by Smith and Waterman in 1981 (SW81). It is based on a similarity maximizing method proposed by Needleman and Wunch in 1970 that aligned complete sequences. Smith and Waterman added the possibility of initiating and stopping the regions of interest within a longer sequence by a simple modification to the Needleman-Wunch recursions. SW81 is too slow for database searches and genomic comparisons and has for practical purposes been superseded by the BLAST family of programs (Altschul et al., 1990)

The advantage of a statistical alignment approach to local alignment would be that it would give a probability distribution of the borders of the homologous regions that would depend on the content of the sequences. If the homologous sequences were very similar, then the borders would be very well determined, if they were very different the border positions would get quite a wide distribution. Being able to do local is a real and practical issue that appears in much data analysis. Very often we want to investigate regions around a gene for for instance regulatory signals and then we need an automated way of defining cut-off's.

Some LPSA was implemented in Lunter et al. (2006), but considerably more can be done.

**Research Idea.** There are several possible ways to model this. One could have a large genome that was subject to large scale insertions-deletions and then left small homologous islands behind. This would create a two-scale alignment problem. This could be handled for instance within the framework of Miklos, Lunter and Holmes (2003), that allows a distribution on insertion-deletion lengths. This is a possibility but would most likely be slow as the given regions would be large. One should add to this framework, that the regions given are not complete genomes, but chosen by the researcher by some criteria. This implies that the borders of the regions would have to have a different insertion-deletion process than the internal region. This problem could also be addressed using the framework of Satija, Pachter and Hein (2008) that allows regions to evolve by different rates. Unrelated regions would correspond to regions with rate infinity or their probabilities can be calculated easily since they are independent. This latter method would be faster, but might still be too slow. If it were given large regions with small homologous regions it would spend much computation on aligning regions to declare them unalignable (unrelated). It could be that some pre-processing step should be used to define regions for further exploration.



Originally a chromosome existed and if two descendant chromosomes were chosen just after this, they would be comparable in their complete length. However, as time progress, large regions will be inserted and deleted. Now only smaller regions will be comparable – here shown in red. How does one determine the borders of these regions and simultaneously align them?

Statistical alignment can be solved using Hidden Markov Models (HMM) that emit alignments according to the TKF91 process. If we only search for one surviving homologous region it is possible to propose a reasonable model, if one is willing to accept certain approximations. One issue is: What is the length equilibrium distribution of one of the homologous segments? It could be geometric if assuming like TKF91 that it is always there, even if it is temporarily has length zero. Additionally, one would have to assume that the homologous islands do not cross the end of the complete sequences, ie that the homologous regions have been observed in their entirety. One could assume that there was only random non-homologous sequences only to the left or to the right of the homologous islands.

An HMM for LPSA could look like this: If we wanted a structure to the overall alignment like non-homologous1 → homologous → non-homologous2, it could be made with 7 states: for the non-homologous1 regions:  $(\#, -)_1$  and  $(-, \#)_1$ .  $(\#, -)$  means a nucleotide in the first sequence, nothing in the second.  $(\#, \#)$  means two matched nucleotides and

one evolved into the other substitutions. Correspondingly for the non-homologous regions.  $(\#, -)_1$  means that a nucleotide has been emitted in the first sequence and so forth. For the homologous regions 3 states are needed  $(\#, \#)$ ,  $(\#, -)$ , and  $(-, \#)$ . The convention can be followed that only  $(\#, -)_i \rightarrow (-, \#)_i$  is possible, simplifying the overall HMM. The probability table for the 3 homologous states can for instance be found in Hein (2001):

	a1	-	#	#	E
	a2	#	#	-	E
a1	*	$\lambda\beta$	$\lambda/\mu(1-\lambda\beta)e^{-\mu}$	$\lambda/\mu(1-\lambda\beta)(1-e^{-\mu})$	$(1-\lambda/\mu)(1-\lambda)$
a2	*	$\lambda\beta$	$\lambda/\mu(1-\lambda\beta)e^{-\mu}$	$\lambda/\mu(1-\lambda\beta)(1-e^{-\mu})$	$(1-\lambda/\mu)(1-\lambda)$
a1	-	$\lambda\beta$	$\lambda/\mu(1-\lambda\beta)e^{-\mu}$	$\lambda/\mu(1-\lambda\beta)(1-e^{-\mu})$	$(1-\lambda/\mu)(1-\lambda)$
a2	#	$\lambda\beta$	$\lambda/\mu(1-\lambda\beta)e^{-\mu}$	$\lambda/\mu(1-\lambda\beta)(1-e^{-\mu})$	$(1-\lambda/\mu)(1-\lambda)$
a1	#	$\frac{(1-e^{-\mu}-\mu\beta)}{1-e^{-\mu}}$	$\frac{\lambda\beta e^{-\mu}}{1-e^{-\mu}}$	$\lambda\beta$	$\frac{\beta(\mu-\lambda)}{1-e^{-\mu}}$
a2	-				

The \*\* should here be interpreted as the first non-homologous region and EE as the second non-homologous region. The transition probabilities can be obtained by elementary calculations from the TKF91 process.

It is easy to make variations on this, where there can be more homologous regions with a certain probability. The emission probabilities for the different states would be:

$e(\#, \#) = f(N_1, N_2)$ , where  $N_1$  and  $N_2$  are the nucleotides observed in the sequences at that position and  $f$  is the probability of choosing  $N_1$  and that it then evolves into  $N_2$  according to the chosen substitutional process. If it was the Jukes-Cantor (JC69) process (Yang, 2006 chapt 1), then the function would be  $P(N_1=N_2) = .25(1+3e^{-3\alpha t})$  and  $P(N_1 \neq N_2) = .75(1-e^{-3\alpha t})$ . The remaining states emit exactly 1 nucleotide and would have the probability of the nucleotide in the equilibrium distribution of the chosen substitutional process - .25 if it was JC69.

### Plan.

- Read Thorne et al. (1991), Lunter et al. (2004) and Lunter et al. (2006) [methods section]
- Make simulator of sequence pair under the TKF91 model embedded in large unrelated sequences.
- Implement the basic TKF91
- Implement local version of TKF91
- Simulate sequences under local TKF91 model and see how well the new algorithm recovers the boundaries.
- Compare boundary recovery with BLAST and SW81 and compare performance.
- Apply the algorithm to pairs of 5' ends of a-globin genes from human and a variety of other species (for instance chimp, orangutan, mouse, echidna, chicken) and see where the algorithm chooses cut-off in front of the genes.

### References

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389-3402.
- Bishop, M.J. and Thompson, E.A. (1986). Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.* 190:159-165.
- Hein, J., C. Wiuf, B. Knudsen, Møller, M., and G. Wibling (2000): Statistical Alignment: Computational Properties, Homology Testing and Goodness-of-Fit. (*J. Molecular Biology* 302:265-279)
- Hein, J. (2001): A generalisation of the Thorne-Kishino-Felsenstein model of Statistical Alignment to  $k$  sequences related by a binary tree. (*Pac. Symp. Biocompu.* 2001 p179-190 (eds RB Altman et al.))
- Lunter, G.A., Drummond, A.J., Miklos, I. & Hein, J. (2004) Statistical Alignment: Recent Progress, New Applications and Challenges. In *Statistical Methods in Molecular Evolution* (Nielsen, R., ed.)
- Lunter G, Ponting CP, Hein J. Genome-Wide Identification of Human Functional DNA Using a Neutral Indel Model. *PLoS Comput Biol.* 2006 Jan 13;2(1):
- Miklós, I., Lunter, G.A. & Holmes, I. (2004) A 'long indel' model for evolutionary sequence alignment. *Mol. Biol. Evol.* 21(3), 529-540
- Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequences of two proteins." *J. Mol. Biol.* 48: 443-453.
- Satija, R., Pachter, L. & Hein, J. (2008) Combining statistical alignment and phylogenetic footprinting to detect regulatory elements. *Bioinformatics* 24,1236-1242
- Satija, R., Novak, A., Miklos, A., Lyngsø, R. & Hein, J. (2008) BiGFoot: Bayesian alignment and phylogenetic footprinting for multiple sequences with MCMC. Submitted to *BMC Bioinformatics*
- Thorne JL, Kishino H, Felsenstein J (1991) An evolutionary model for the maximum likelihood alignment of DNA sequences. *J Mol Evol* 33:114-12
- Yang, Z (2006) *Computational Molecular Evolution* OUP

**Further research.** This only describes this problem for a pair of sequences. Going to more sequences would be the natural next step and could use the approach used by Satija, Miklos, Novak, Lyngsø and Hein (2009) that used MCMC and could go to 10-15 sequences.