# Alignment and tree inference

Peter Beerli

December 5, 2005

## 1 Evolution of sequences

In several chapters we did talk about mutation models. These mutation models are unfortunately only looking at point mutations, other processes such as duplication, gene conversion, inversions, insertions and deletions are ignored. Typically we run a two-part analysis by first aligning the sequences, and then do the real analysis and we ignore everything except the point mutation and the insertion/deletion pattern. Only recently procedures are developed that take into account

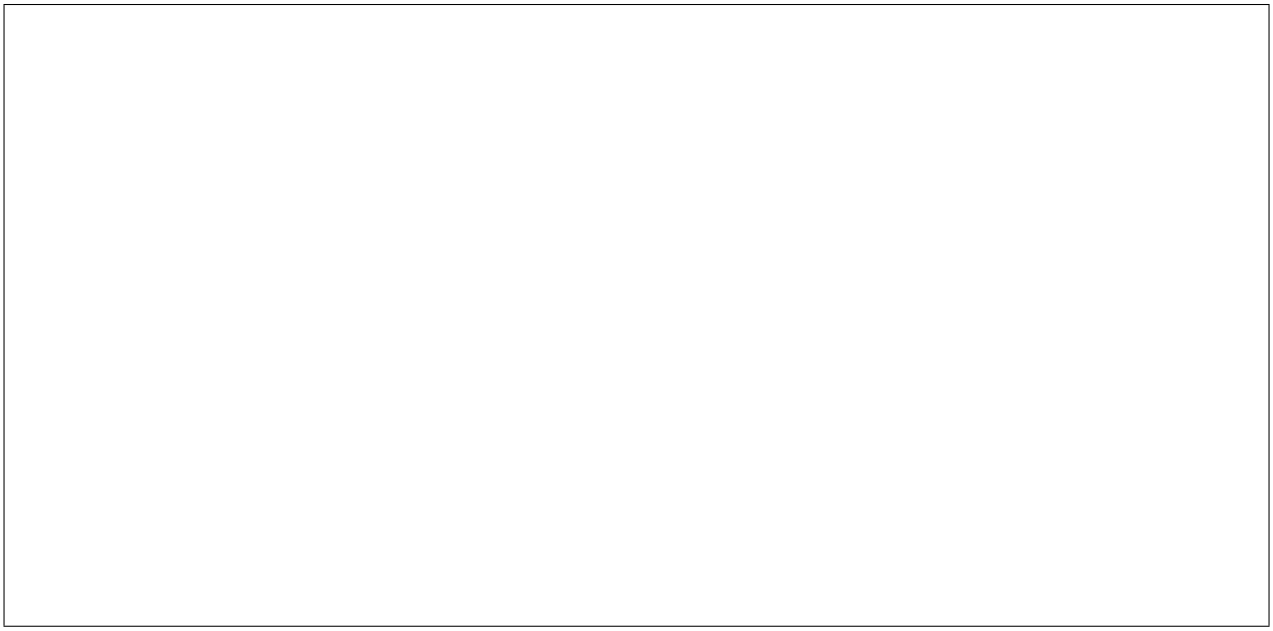

Figure 1: Processes that change a sequence: Point mutation, inversion, duplication, conversion, insertion, deletion.

multiple processes in a joint estimation process, as none of these processes is independent of the phylogeny or from each other. In this chapter we will only explore the complication introduced by insertion and deletion patterns.

In a most simple procedure we produce genealogies or species phylogenies using already aligned sequences. Aligned sequences are the product of an alignment algorithm where we minimize the changes among several sequences. For example look at this fragment:

```
AATACAATTC GACACGGGGG CCACTCACGA AAATTAGAAG AAGAGC
AATACCATTC GACACGGGGC CACTCACTAA AACTAGAAGA AAAGC
AATACCAATT CGACACGGGC CACTCACTAA AACTAGAAGA AGAGC
AATACAATTC GACACCGGGC CACTCACTTA AAACTAGAAG AAAAGC
AATACAATTC AACACTAAAA TTAGAAGAAA AGC
GAATCATTCC GACACCGGGC CACTCACTAA AACTAGAAGA AAAGC
```

We can see that the sequence does not match at all between the individual sequences. We could align this by hand and try to minimize changes by inserting “gaps” so that nucleotides match up; and after lots of work we would end up with something like the sequence below.

```
AATAC-AATT CGACACGGGG GCCACT-CAC G-AAAATTAG AAGAAGAG-C
AATAC-CATT CGACACG-GG GCCACT-CAC T-AAAACTAG AAGAAAAG-C
AATACCAATT CGACAC--GG GCCACT-CAC T-AAAACTAG AAGAAGAG-C
AATAC-AATT CGACACC-GG GCCACT-CAC TTAAAACTAG AAGAAAAG-C
AATAC-AATT CAACAC---- ---------- T-AAAATTAG AAGAAAAG-C
GAATC-ATTC CGACACC-GG GCCACT-CAC T-AAAACTAG AAGAAAAGC-
```

Of course alignment by hand is tedious and often problematic, although most current alignment programs have their own little problems and tweaking by hand is often needed when working with real data.

## 2 Pairwise alignment

Needleman and Wunsch (1970) developed the first algorithms to align sequences. this procedure does not take into account any tree structure (and therefore correlation) among sequences. The algorithm has 3 main steps,

**Algorithm 1** Needleman-Wunsch algorithm

```
Create a matrix $M$ with one sequence as row header and the other sequence as column header
Assign a 1 where the column and row site matches, zero otherwise
For each element in the matrix evaluate $m_{ij} = m_{ij} + \max(m_{k,j+1}, m_{i+1,l})$ {moving from the last element to the first}
start at the $m_{11}$ and proceed right and down the table and connect the highest scores.{Moving from the first element down and right}
```

|   | A | B | C | N | J | R | Q | C |
|---|---|---|---|---|---|---|---|---|
| A | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| J | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| C | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| J | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| N | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| R | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| C | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

|   | A | B | C | N | J | R | Q | C |
|---|---|---|---|---|---|---|---|---|
| A | 5 | 4 | 3 | 3 | 2 | 1 | 1 | 0 |
| J | 4 | 4 | 3 | 3 | 3 | 1 | 1 | 0 |
| C | 3 | 3 | 4 | 3 | 2 | 1 | 1 | 1 |
| J | 3 | 3 | 3 | 2 | 3 | 1 | 1 | 0 |
| N | 2 | 2 | 2 | 3 | 2 | 1 | 1 | 0 |
| R | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 0 |
| C | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Connecting the highest scores will result in the two alternative alignments

```
A  B  C  N  J  -  R  Q  C           A  B  C  -  N  J  R  Q  C
A  J  C  -  J  N  R  -  C     or    A  J  C  J  N  -  R  -  C
*     *     *     *     *           *     *     *     *     *
```

where the * mark the matches. Several refinement of this basic algorithm exist.

# 3 Joint adjustment of sequence alignment and tree topology

## 3.1 Parsimony method

Sankoff, Morel, and Cedergren (1973) developed a parsimony method that estimates the alignment and the topology jointly. They used a penalty function for the alignment of two sequences: penalties for substitutions, insertion/deletion. Sequences at each internal node are inferred using the penalties, and the tree with the minimal penalty score was the best tree and alignment. For each tree one would need to calculate for each internal node alignment scores as shown above, for $n$ sequences of length $L$ this results in approximately in $L^n$ computations, only feasible for small $n$. Several approximations allow to cut down the burden, but still this method will only work for small numbers of taxa. Most popular today are even more approximative methods (like the one used in the program ClustalV), where one does not revisit the alignments once they are chosen, these methods allow to align large data matrices but are less precise than the Sankoff et al algorithm.

## 3.2 Probabilistic methods

A probabilistic model that takes into account insertion and deletions of single sites was developed by Bishop and Thompson in 1986. Their methods could not handle superimposed gaps and gaps were collapsed or expanded by single steps.

### 3.2.1 The Thorne-Kishino-Felsenstein model

Jeff Thorne developed during his thesis a model that allows for multiple gaps and that is tractable (although it is still difficult to use on data sets with multiple sequences). He parametrized the model like this

$$*A \sim G \sim A \sim C \sim A \sim T \sim G \sim$$

on the left there as immortal link, need that a sequence cannot shrink to nothing, each nucleotide has an attached link. Each nucleotide can be deleted and each link can add nucleotide-link pairs. The model contains a standard substitution model and additionally two rates that control the insertion/deletion process. Each link can insert a new site with rate $\lambda$. Each site has a constant risk $\mu$ of being deleted. Since all action happens in pairs of nulceotide and link, the immortal link to the left of everything cannot be deleted. The whole sequence of length $n$ will increase with probability $(n+1)\lambda$ per unit time and will shrink with $n\mu$ per unit time. This type of birth-death process is well understood and with $\mu > \lambda$ the equilibrium distribution of the sequence lengths is a geometric distribution. And the equilibrium base composition will be the on from the substitution model (Felsenstein 2003). The parametrization allows to calculate transition probabilities where one can calculate the 3 possible events that happen to a particular sequence

$$\begin{aligned} GTACCTAC &\rightarrow \text{substitute} \rightarrow GTACTTAC \\ &\rightarrow \text{insert} \rightarrow GTACCATAC \\ &\rightarrow \text{delete} \rightarrow GTACTAC \end{aligned}$$

Given the rates and all possible events one can set up a transition probability matrix, that can be solved analytically (for more detail see Lunter et al. 2005). A specific alignment is the sum off all

Figure 2: Example of a transition from ancestral sequence to the current one (based on Lunter et al. 2005)

possible events, some of these are not visible, for example

$$*A \sim C \sim \underline{C \sim} T \sim A \sim\rightarrow * A \sim C \sim T \sim A \sim$$

$$* \underline{A \sim} C \sim T \sim A \sim\rightarrow *A \sim C \sim C \sim T \sim A \sim$$

will show no change but the end-sequence is different as the $C$ are independent of each other. The TFK91 model has the limitation that it can only insert insertion/deletion of size one which will result in unreasonable sequences of many single gap deletions etc. In 1992, Thorne et al developed an extension that can insert frames that are longer than one site.

### 3.2.2 HMM

Knudsen and Myamoto (2003) developed an insertion/deletion model that is formulated as a hidden Markov Model. Lunter et al (2005) show that their graphical model is identical to their own HMM representation of the the TKF model.

### 3.2.3 Bayesian inference

The TKF model is interesting because one can calculate analytical solutions of the transition probabilities, this facilitates fast calculations of the changes to an alignment. Several authors have implemented the TFK model or similar models in Bayesian analyses. Perhaps the most prominent is implemented in BEAST (by Rambaut and Drummond). The analyses of data sets is still a challenge with such methods, but eventually they will replace the standard practice of (1) align and (2) phylogenetic analysis.

Figure 3: HMM for the TKF model