Some background on the program *Saté*

Alex Landy and Peter Beerli

***Abstract*—A short overview of the theory used in the program *Saté*. *Saté* estimates the sequence aligment and the phylogenetic tree in a divide and conquer approach. It generates new alignment of subtrees of the whole phylogeny, combines the alignments and than reestimates the maximum likelihood of the tree. This approach is repeated until convergence is achieved.**

Simultaneous Alignment and Tree Estimation (*Saté*, Liu et al. 2012) is a software package which is used for the alignment of DNA sequence data and for the estimation of phylogenetic trees. The original *Saté* program was written to use an iterative procedure to compute a series of alignments and tree pairs by running both RAxML and MAFFT. *Saté* was upgraded to *Saté*-II in 2012 to incorporate additional options for both the sequence alignment and sequence mergers. These additional options include the packages ClustalW, MAFFT, MUSCLE, OPAL, and PRANK. In general OPAL was found to be the best merger software. *Saté*-II was also upgraded to include both RAxML and FastTree for the construction of phylogenetic trees.

I. Decomposition and divide-and-conquer

Saté uses a start tree or then constructs a tree from a first alignment of the whole data set. The maximum likelihood tree is then decomposed into subtrees. Subtrees are specified by the longest internal branch – splitting the tree into 2. This splitting procedure continues until only a small number of samples remain in a subtree, for example 3. The samples of each subtree are then aligned. New alignments are then combined into groups. if there are gaps between groups then the whole group will receive these gaps (Figure 1) Once

```
group1 ATGCACA-CTA
group1 ATGCACATCTA
group2 ATG--CATCTC
group2 ATG--CATCTC
```

Figure 1. Alignment of two subsets with gaps. Group 2 has a gap to allow to align to group 1

the new alignment is complete, the maximum likelihood tree is found using a RaXML or another phylogenetic inference method based on likelihood. This procedure is repeated until the likelihood does not improve anymore. This approach is greedy and may be trapped by local maxima. *Saté* has options to allow less greedy tree searches: by allowing an arpbitrary number of cycles during which the likelihoods does not need improve.

In addition to new alignment and tree building options *Saté*-II also three variants of the basic algorithm. These include *Saté*-II simple, *Saté*-II ML, and *Saté*-II fast. These variants differ mainly in the number of iterations performed and in the criterion used to select a tree/alignment pair generate from the *Saté*-II algorithm. *Saté* works well with larger number of taxa (>200), with smaller numbers it is about as accurate as the standard method of generating one multiple sequence alignment and one one maximum likelihood analysis.

II. Biliography

Liu, K. et al. 2012. *Saté* II: Very Fast and Accurate Simultaneous Estimation of Multiple Sequence Alignments and Phylogenetic Trees. Systematic Biology 61: 90-106.

III. Disclaimer