Bioinformatics - Tree Comparison
Project 1

Peter Gabrielsen 20114179
Christoffer Hansen 20114637

November 4, 2015

Introduction

This project deals with implementing Day's algorithm for computing the RF distance between two unrooted trees over the same set of species. We present experiments where we compare the evolutionary trees constructed using the Neigbor Joining (NJ) methods implemented as `QuickTree` and `RapidNJ` on different datasets which were computed using align methods `Clustal Omega`, `Kalign`, `MAFFT` and `MUSCLE`.

Everything works as expected.

Code can be found at `https://dl.dropboxusercontent.com/u/8990890/2015Q2_AiBST_20114179_20114637_Project1.zip`.

Implementation

We implemented `rfdist` using Day's algorithm as presented in class. The only place we deviate from the algorithm is on sorting the collected intervals. We decided to do this using `QuickSort` instead of `RadixSort`. We are aware that this design choice makes an expected $\mathcal{O}(n \log n)$ algorithm, but as `QuickSort` is shown to perform well in practice, we did not consider this as being a problem.

The implementations is done in `c++`. It compiles using *make rfdist* and runs using *./rfdist <tree1> <tree2>*, where *<tree1>* and *<tree2>* should be valid paths to files containing trees in well-formed `Newick`-format.

Correctness was tested on the provided *testdata.zip* example.

Experiment 1

For each alignment method ClustalW2 (1), Kalign (2), MAFFT (3), MUSCLE (4), we build a NJ tree using QuickTree and RapidNJ, and compute the RF-distance between each combination of these eight tree. The outcome of our experiment, an 8x8 table showing the RF-distance between each pair of constructed trees, is presented in figure 1.

	(1)NJ	(2)NJ	(3)NJ	(4)NJ	(1)QT	(2)QT	(3)QT	(4)QT
(1)NJ	0	222	192	242	230	258	256	284
(2)NJ	222	0	164	212	228	198	222	268
(3)NJ	192	164	0	210	194	226	200	254
(4)NJ	242	212	210	0	246	248	260	192
(1)QT	230	228	194	246	0	158	114	198
(2)QT	258	198	226	248	158	0	128	176
(3)QT	256	222	200	260	114	128	0	182
(4)QT	284	268	254	192	198	176	182	0

Figure 1: Results of experiment 1

We are pleased to see the table is symmetric, as this is what we would expect.

Experiment 2

The first experiment is done on the 395 input sequences in patbase_aibtas_permuted.fasta. This yields another 8x8 table presented in figure figure 2.

	(1)NJ	(2)NJ	(3)NJ	(4)NJ	(1)QT	(2)QT	(3)QT	(4)QT
(1)NJ	0	210	172	246	238	264	252	280
(2)NJ	210	0	176	252	250	230	246	276
(3)NJ	172	176	0	238	268	268	244	286
(4)NJ	246	252	238	0	216	242	232	208
(1)QT	238	250	268	216	0	138	108	158
(2)QT	264	230	268	242	138	0	124	172
(3)QT	252	246	244	232	108	124	0	148
(4)QT	280	276	286	208	158	172	148	0

Figure 2: Results of experiment 2

Again, we are are pleased to see the table is symmetric. This gives us confidence that our implementation is correct.

Experiment 3

Compute the RF-distance between the trees produced in 'Experiment 1' and 'Experiment 2' using the same alignment and tree reconstruction method. This yields 8 distances presented in figure 3.

(1)NJ	(2)NJ	(3)NJ	(4)NJ	(1)QT	(2)QT	(3)QT	(4)QT
160	200	184	216	68	62	42	158

Figure 3: Results of experiment 3

We conclude that aligning using `Clustal Omega` seems to perform the best on `RapidNJ` and that `MAFFT` seems to perform the best on `QuickTree` on the test data we have been experimenting on.

Experiment 5

We test running time by computing perfectly balanced binary `Newick`-trees with height in the interval $k \in [2, \cdots, 23]$ giving a total number of nodes equal to $2^{k+1} - 1$. As we double the number of nodes on each test-run we expect to see a straight line in a log-log plot. We are pleased that this is indeed the case, as presented in figure 4.

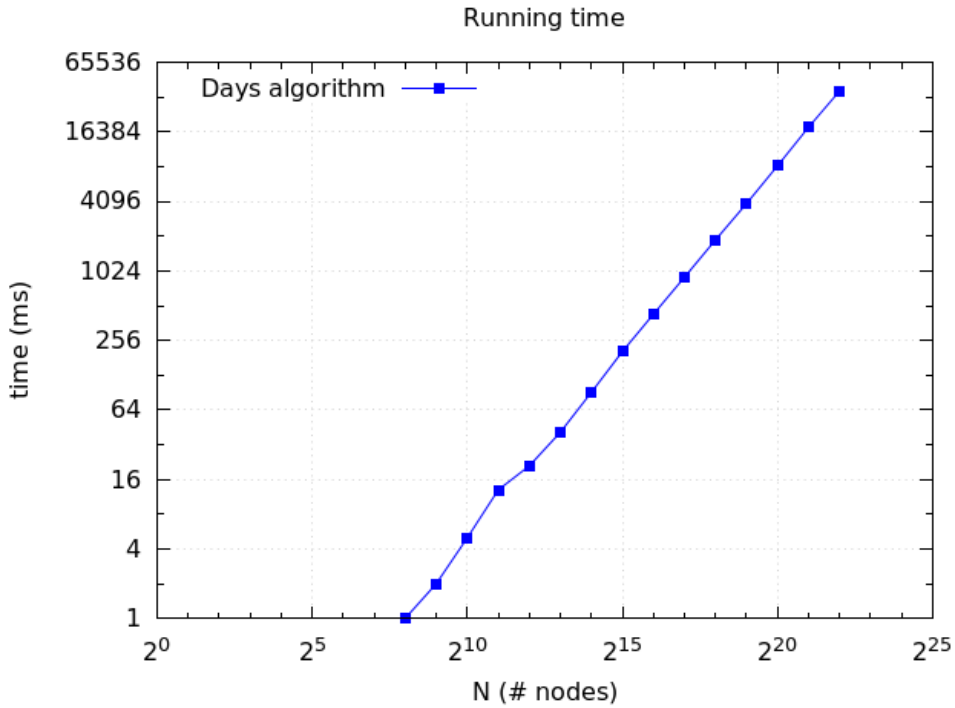

Figure 4: Running time of rfdist