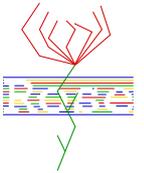


# Generating Benchmarks for Multiple Sequence Alignments and Phylogenetic Reconstructions



J. Stoye, D. Evers, F. Meyer

## Purpose

The simulation of evolutionary processes on the molecular sequence level has a long tradition. Starting with the model of Jukes and Cantor [1], several generalizations and alterations have been presented. But in none of the methods the length of the sequences is altered by insertion and/or deletion (*indels*) of subsequences making these approaches impracticable for several applications. We have added indels and "sequence motifs" (patterns in a family of related sequences) to the so-called HKY-model [2] to create more realistic sequence families. The data created by our tool *rose* (random-model of sequence evolution) has been extensively tested with our *Divide-and-Conquer Alignment*[3] and *GeneFisher*[4] software packages.

## Approach

We simulate an evolutionary process by iterated mutation of a "common ancestor sequence" following the edges of a given "mutation guide tree". This way, the topology of the tree induces the relationships of the sequences. The mutations are performed by **insertion**, **deletion**, and **substitution** of single characters or whole subsequences of the ancestor sequence. In addition to knowing the exact evolutionary *distance* of the sequences, our approach provides us with their whole evolutionary *history*. Therefore, in contrast to biological applications, it is easily possible to verify predictions about phylogenetic relationships drawn from the sequences simply by comparing the predicted phylogeny to the tree that was used in the creation process. Figure 1 sketches the creation process of a family of four sequences.

## Model

### Input

#### Alphabet

e.g. DNA, RNA, protein

#### Root Sequence/Average Sequence Length

if no root sequence is specified, a random sequence of average length is generated

#### Character Frequencies

used for insertions and creation of root sequence

#### Mutation Matrix

used for substitutions

#### Insertion/Deletion Probability Function

probability of an indel event and indel length function

#### Mutation Guide Tree

if no tree is entered, a binary mutation guide tree of user defined average pairwise sequence distance is created

#### Sequence Motifs

allows to specify regions of high/low mutation rate

### Output

#### Family of Sequences

containing sequences with average length and average pairwise evolutionary distance

#### Multiple Sequence Alignment

of the sequences that is optimal with respect to the creation process

#### Relatedness Tree

representing the phylogenetic relationship of the created sequences

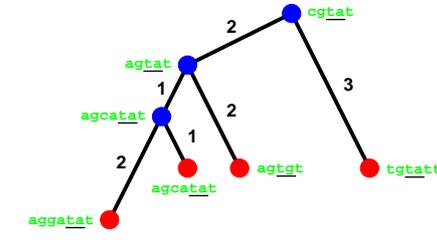


Figure 1: Example of a creation process of four sequences from a common ancestor *cgtat*. The underlined part denotes a sequence motif with much smaller substitution probability.

## Example 1: A Protein Sequence Family

Figures 2 and 3 show a family with  $k = 4$  sequences of average length  $n = 50$  created with the default settings of *rose*: A uniform binary mutation guide tree of depth  $m = 9$  and uniform edge length  $R = 18 \text{ PAM}^*$ . The probability for insertions and deletions is set to  $p_{ins} = p_{del} = 0.3\%$  and the insertion and deletion length functions are exponentially decreasing with a maximal length value of 10. The average sequence distance is  $d_{av} = 250 \text{ PAM}^*$ .

```
(a) FSAEAALVSPGKGDDEQVFNKDKCVYBGRKDKRMMVKTPTTGPLVGVYBQ
    YEGANEVGATCESSYCVYKQEAIQVKESQECTDFAARBEVKSFRGVPKLTLEIIPVPL
    YGAABPVGDPKIKLGLFLNBYESKQHTAAMCLLQMKTELEIEPIEYQA
    SGVTEPPVPPVPTGKLDKVTREENCLQMLCMQMGPPMVTIGEVGI

(b) FSAEAALVSP-----GKGDDEQVFNKDKCVYBGRKDKRMMVKTPTTGPLVGVYBQ
    YEGANEVGATCESSYCVYKQEAIQVKESQECTDFAARBEVKSFRGVPKLTLEIIPVPL
    YGAABPVGDP-----IKLGLFLNBYESKQHTAAMCLLQMKTELEIEPIEYQA
    SGVTEPPVPP-----VPTGKLDK--YTRREENCLQMLCMQMGPPMVTIGEVGI

(c) FSAEAALVSP-----GKGDDEQVFNKDKCVYBGRKDKRMMVKTPTTGPLVGVYBQ
    YEGANEVGATCESSYCVYKQEAIQVKESQECTDFAARBEVKSFRGVPKLTLEIIPVPL
    YGAABPVGDP-----IKLGLFLNBYESKQHTAAMCLLQMKTELEIEPIEYQA
    SGVTEPPVPP-----VPTGKLDK--YTRREENCLQMLCMQMGPPMVTIGEVGI
```

Figure 2: (a) Sample family of random sequences obtained with *rose* for  $n = 50$  and  $k = 4$ ; (b) "true" alignment of these sequences; (c) an optimal alignment according to PAM250 substitution matrix and gap function  $g(l) = 8 + 12l$  computed with the program MSA. While the overall optimal alignment is correct, the exact location of the gaps does not coincide in all cases.

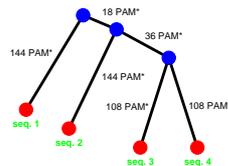


Figure 3: Relatedness tree for the sequences shown in Figure 2 (a). The third and the second sequence are closer related to each other than to the other ones which is also reflected in the alignments of Figure 2 (b) and (c).

## Example 2: A DNA Sequence Family with Motif

Figure 4 demonstrates the simple use of motifs in sequence families created by *rose*. Within the motif sequence mutability is set to zero, outside it remains normal.

```
(a) AGC----ATTATAATGAGTCAACA--TAGAAAGCC
    CGC----AGTATAATGAGTCAACA--TAGAAAGCC
    AGT-CTACACTATAATAGG----AAG--AAGCC
    GGTTCCTTAAGTATAATAGG----AAG--AAGCC
    GGT-CTATAATGATAATGTT----ACT--AAGCC

(b) AG--C--A--TTATAATGAGTCACTAGAAAGCC
    CG--C--A--GTATAATGAGT-AGCA--AAGC
    AGT-CT-ACAATAAT-AGG-AGGAC--AAGC
    GGTTCCTTA-AGTATAATG-AG-AGGAA--AAGC
    GGT-CT-ATAATAATAATG-GT-ACTAG--AAGC
```

Figure 4: (a) A "true" alignment of a sample family of 5 DNA sequences which contains a conserved TATAAT motif obtained with *rose* using a mutation vector disallowing mutations within the motif; (b) an optimal alignment with MSA. The overall length of the alignment is shorter than the "true" alignment. The parsimony objective underlying the sum of pairs scoring of MSA fails here.

## Example 3: A Protein Sequence Family with Varying Mutation Rate

Figure 5 shows an example with varying sequence mutability along the sequence.

```
ILGPAAGVSVGLVWGGIG-----KLSGGYNAESLQDMFLVLPPTQTQYF----SBFTID
CISPSDRKSGAKGAWGIGG-----RAAGEYTTQELARMLFVYPTAKTYF----ABRRLS
PKPPGGGNGVKAADKVE-----DRMGGAHAGALIELYLQFPTLKYF----ERFK--
VKWTQDRAKTLWQKDG-----EPKGGPSASLABFYMKLPPTKYF----TVGLS
ITFPADDTIKAGWQVYATSLCKWEGGATPEAAMRFLKYPPTKYFVATL--YBIS
CLSPAERTNKQSWKAGEA----ABDGGSDGVKERRFLKQPTTKTPF----SQFALK

DRAKVEKGRKXVLDALVHVARMDLTPALGALSLLBARWLRKNDPTNKLKLLTCLLLML
RSTAVYKGRKQKSEDAQSSAQDQRMQMPQEMPAISDLRGRVKNRSDWIKLIDSCLLMRY
STDTVEKGRGTVERDSWTSGAASLQSPPEALCALSDFRBARKRNVIDWIKLIDSCLLMRY
RIDEVYKGRKQVYVGTSAQSHHVIQ--LGLKDLRE--LAEVYALRLIRKCLLVTL
RDSQVYKGRKQVYVPAALTSVLRHAGVGGVLAWLKDLRM--LNSLPVALRDNWCLLVTL
RTHQVYKGRGAEVYSAJLGRVYCEEDSSPMATTALRDKGRA--LQVDDMNVLSRCLQVKS

VAQLPIDVTPSVHASEDKFLPSVGRKIT----EKYS
SABAPOISTPGYHASEDKFLRHWVTVGR----VKYR
DCLPLGEDLEAVBASVYDKLAASIMGAESTIPVLSKRY
LPHLEAFTLLAQAHRDVFTRKQETALTGVY--SQTR
CBLRANTFKLDDGALDKFRTRQSTGLPGBA--SQTR
AD-YATVLDVYVABLDTLVYASTSKZEGG--SQTR
```

Figure 5: "True" alignment of six protein sequences created with *rose* using human hemoglobin  $\alpha$  as root sequence. The histogram above the alignment shows the mutation probability along the sequence allowing a higher mutation rate between the  $\alpha$  helices than within.

## Summary

The data sets created by *rose* are the first artificially created sequence families that contain both *indels* and *motifs*. The evaluation of multiple sequence alignment and phylogenetic reconstruction methods is now possible with the benchmarks created.

## Future Work

Both the model and the implementation were designed to allow future extension. We are currently working on the inclusion of varying mutation rates in different branches.

## Availability

The software is accessible on our Bioinformatics WebServer *BiBiServ* under the following address: <http://bibiserv.TechFak.Uni-Bielefeld.DE/rose/>

## Acknowledgements

The authors wish to thank Robert Giegerich for his helpful comments on an earlier version of this poster. The work was partially supported by the German Ministry for Education and Sciences (BMBF), the Ministry of Science of North Rhine Westfalia (MWF-NRW) and the German Research Council Graduate Program (DFG-GK) on Structure Formation.

## References

- [1] T. H. Jukes and C. R. Cantor. Evolution of Protein Molecules. In Munro, H. N., editor, *Mammalian Protein Metabolism*, volume 3, pages 21–132. Academic Press, New York, NY, USA, 1969.
- [2] M. Hasegawa, H. Kishino, and T. Yano. Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA. *J. Mol. Evol.*, 22:160–174, 1985.
- [3] J. Stoye. DCA: Divide and Conquer Multiple Sequence Alignment. <http://bibiserv.TechFak.Uni-Bielefeld.DE/dca/>, 1996.
- [4] F. Meyer and C. Schlieiermacher. GeneFisher. <http://bibiserv.TechFak.Uni-Bielefeld.DE/geneFisher/>, 1996.