

New methods for estimating amino acid replacement rates

Lars Arvestad¹

1 Introduction.

Anyone performing a simple homology search does not worry too much about deciding on a scoring matrix. The community has chosen to standardize on the PAM and BLOSUM series of matrices. With that perspective, estimating amino acid replacement rate matrices is not very interesting. However, there is a recent and growing literature [1, 2] where increasingly detailed models of proteins are made, with replacement rates estimated depending on local protein structure. It is well known that depending on whether a site is in an α helix or in a loop, whether it is buried or exposed, and so on, both the residue distributions and replacement rates differ substantially. Applications are for example in phylogenetics, ancestral sequence reconstruction, and secondary structure prediction, and the quality of replacement is very important. In particular, it is important to do well with very little data.

2 Results.

Recently, Müller and Vingron [3] introduced the *resolvent method* (MVR) for estimating replacement rate matrices, which they showed to perform better than the original Dayhoff method and comparable to a maximum likelihood method [4].

We propose two new methods and show that they both give better results than MVR. The first method (denoted BW) is using techniques for estimating DNA substitution matrices [5]. The second method (BR) is a combination of BW and MVR. It is using an improved method for estimating the resolvent matrix R .

Figure 1 shows results from comparing the estimation methods. For each length $L = 100, \dots, 1000$, 100 comparisons were made. In each test, 30 synthetic alignments of length L were generated on PAM distances drawn from $N(150, 90)$, and the BW, BR, and MVR methods were applied. MVR exhibited a surprising behaviour due to how an essential parameter in the method was estimated, and as an improvement a method MVRf (where f is for 'fixed parameter') was also considered. As can be seen, BW perform substantially better than MVR and MVRf, and BR is comparable to MVRf except for the case of short sequences, when the data provides a bad sample of the PAM distances.

BR is efficient enough to handle very large sets of data, and can produce a rate matrix based on 10,113 sequence pairs in just over two hours. The method is also robust and can do well on a small number of sequence pairs. Contrary to other frequency-based methods BR handles data in the form of short alignments (less than 200 residues) very well.

¹Stockholm Bioinformatics Center and Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology (KTH), Sweden. E-mail: arve@nada.kth.se

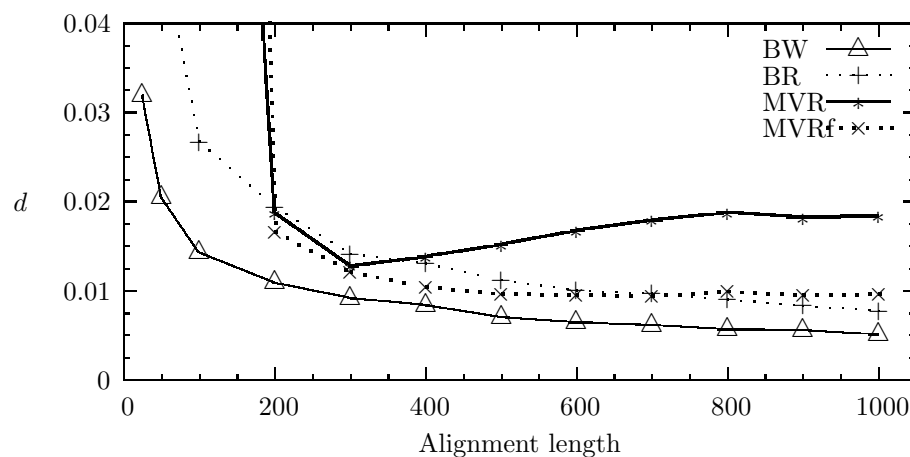


Figure 1: Performance of rate matrix estimation methods. d is the average Frobenius norm of the differences between rate matrix estimate \tilde{Q} and true Q .

References

- [1] Koshi, J. M. and R. A. Goldstein. 1995. Context-dependent optimal substitution matrices. *Protein Eng* 8(7), 641–645.
- [2] Liò, P., N. Goldman, J. L. Thorne, and D. T. Jones. 1998. PASSML: combining evolutionary inference and protein secondary structure prediction. *Bioinformatics* 14(8), 726–733.
- [3] Müller, T. and M. Vingron. 2000. Modeling amino acid replacement. *J Comput Biol* 7(6), 761–776.
- [4] Müller, T., R. Spang, and M. Vingron. 2002. Estimating amino acid substitution models: a comparison of Dayhoff’s estimator, the resolvent approach and a maximum likelihood method. *Mol Bio Evol* 19(1), 8–13.
- [5] Arvestad, L., W.J. Bruno. 1997. Estimation of reversible substitution matrices from multiple pairs of sequences. *J Mol Evol* 45(6), 696–703.