

Improving alignment of hypervariable regions for *varDB*

C. Nelson Hayes¹

nelson@kuicr.kyoto-u.ac.jp

Diego Diez¹

diez@kuicr.kyoto-u.ac.jp

Nicolas Joannin²

nicolas.joannin@ki.se

Minoru Kanehisa¹

kanehisa@kuicr.kyoto-u.ac.jp

Mats Wahlgren²

mats.wahlgren@ki.se

Craig E. Wheelock³

craig.wheelock@ki.se

Susumu Goto¹

goto@kuicr.kyoto-u.ac.jp

¹ Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

² Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Box 280, SE-17177 Stockholm, Sweden

³ Division of Physiological Chemistry II, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-17177 Stockholm, Sweden

Abstract

Proteins sequences involved in antigenic variation are difficult to align due to the presence of one or more hypervariable (HV) regions in between conserved flanking regions. We used GBLOCKS to partition existing alignments into conserved and HV blocks, followed by clustering and realignment using BlastClust and MAFFT, and then we used the MAFFT profile tool to merge the partial alignments into a final alignment. This tool is available as part of the *varDB* database at <http://www.vardb.org/vardb/alignments/hvalign.html>

Keywords: multiple sequence alignment, antigenic variation

1 Introduction

When sequences are very similar, most current alignment methods can produce a high quality multiple sequence alignment, but when the sequences are more distantly related, results depend heavily on the assumptions and parameters of the method used. Many methods assume that the sequences represent a single homologous domain and that a single substitution model is appropriate for the entire sequence. Sequences involved in antigenic variation may violate these assumptions, however. Antigenic variation is a general strategy used by some pathogens to avoid recognition by the immune system. One mechanism involves using paralogous gene families to encode surface proteins involved in host-pathogen interactions [1]. Because of their role in immune evasion, these proteins usually contain one or more HV regions, flanked on both sides by relatively conserved regions. Aligning these sequences proves difficult because the presence of highly conserved blocks biased towards selection of a conservative substitution matrix based on the degree of sequence similarity. Aligning HV regions with a strict scoring matrix tends to result in an excess of gaps, however, making the alignment difficult to interpret. HV regions may be under diversifying selection and may code for low complexity loops that are not essential for protein structure, for which a more lenient substitution matrix may be more appropriate. New members of these rapidly evolving gene families may be added or modified as a result of ectopic recombination and gene conversion, and some members may diverge into subfamilies, differing with respect to e.g., receptor binding affinity. Each of these factors is likely to influence the accuracy of alignment methods. The method proposed here attempts to improve alignment of antigenic variation proteins by partitioning to reduce bias due to global parameters.

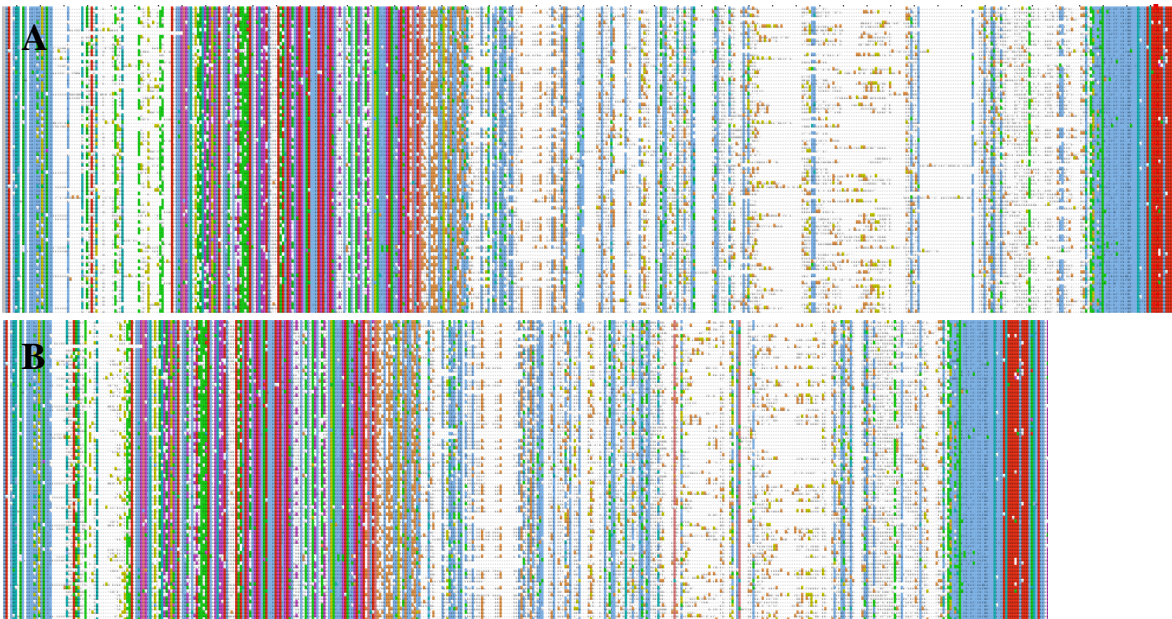


Figure 1. A) Alignment generated using MAFFT with default parameters. B) the same alignment following refinement to reduce the number of gaps.

2 Method and Results

This method assumes that input sequences are already roughly aligned; however, unaligned sequences will be aligned using the E-INS-i option in MAFFT [2]. This option assumes that the sequences contain more than one conserved motif surrounded by unalignable sequences, providing good alignments with few gaps in the HV regions but lacking sensitivity to semi-conserved patterns. This method uses GBlocks [3] to identify conserved blocks and partition the initial alignment into several sub-alignments. The conserved blocks are realigned using a strict substitution matrix, and the variable blocks are clustered into subgroups using BlastClust [4]. Each subgroup is then re-aligned separately using a relatively lenient substitution matrix. The subgroups and any unclustered sequences are then merged using MAFFT's profile alignment tool. The sub-alignments are reassembled to create the final alignment. This approach improves detection of semi-conserved sites within subgroups and results in alignments with fewer gaps. An implementation of this tool is available at <http://www.vardb.org/vardb/alignments/hvalign.html> [5].

3 Discussions

The method presented here is one of many possible approaches to clarifying the relationships between members of a paralogous gene family by revealing semi-conserved positions. The time complexity varies considerably depending on the nature of the initial alignment and the number and length of conserved and HV regions. This method assumes that conserved blocks can be correctly identified in the overall alignment, but if this is not the case, then similar sequences can be grouped using BlastClust first. The method could be further improved by using a custom substitution matrix, parameter optimization, and inclusion of nucleotide sequence data.

References

- [1] Kaviratne, M. et al. Antigenic variation in *Plasmodium falciparum* and other *Plasmodium* species. Academic Press, Amsterdam; Boston. 2003.
- [2] Katoh, K., Kuma, K., Toh, H. and Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment, *Nucleic Acids Res*, 33, 511-518. 2005.
- [3] Talavera, G., and Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* 56, 564-577. 2007.
- [4] Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research* 25:3389-3402. 1997.
- [5] Hayes, C.N. et al., varDB: a pathogen-specific sequence database of protein families involved in antigenic variation, *Bioinformatics*, (In press), 2008.