

# Gap mapping: a paradigm for aligning two sequences

Matthew Bellgard,<sup>1</sup> Thomas Gamble,<sup>1</sup> Mark Reynolds,<sup>1</sup> Adam Hunter,<sup>1</sup> Ed Trifonov,<sup>2</sup> Ross Taplin<sup>1,3</sup>

<sup>1</sup>Centre for Bioinformatics and Biological Computing, School of Information Technology, Murdoch University, WA, Australia; <sup>2</sup>Genome Diversity Center, Institute of Evolution, University of Haifa, Haifa, Israel; <sup>3</sup>School of Mathematics and Statistics, Murdoch University, WA, Australia

**Abstract:** Pairwise sequence alignment is one of the most essential tools in comparative genomic sequence analysis. It is used to compare the sequences of genes and proteins with the aim of inferring structural, functional and evolutionary relationships. However, current ‘mainstream’ alignment algorithms have optimisation criteria based primarily on computational efficiency using parameters such as gap penalties, which are not biologically motivated. In addition, current alignment algorithms such as the Smith and Waterman technique provide a single alignment that could be sensitive to rather arbitrary choices in parameters such as gap penalties. This paper explores the range of properties resulting from posing the alignment problem more as a ‘mapping gaps in sequences’ exercise. We argue that this approach is intuitive and provides greater control over the number of gaps placed within an alignment. This type of approach was proposed by Sankoff (1972), but unfortunately has not received much attention. We report and discuss our findings by comparing this approach to other techniques using structurally confirmed aligned sequences from a benchmark alignment database. Interestingly, this approach consistently provides optimal and near optimal alignments and is thus a viable approach to sequence alignment.

**Keywords:** pairwise sequence alignment, bioinformatics, gap penalties

## Introduction

Sequence alignment is unquestionably the most powerful computational tool available for comparative analysis of DNA and protein sequences. Sequence alignments are used to compare the sequences of genes and proteins with the aim of inferring structural, functional and evolutionary relationships among the sequences under study. The advent of automated DNA sequencing techniques and large genome sequencing initiatives has seen the number of gene and protein sequences in the public databases grow exponentially in recent years. The ability to compare a given sequence against all nucleotide and/or protein sequences in silico provides scientists with an ideal opportunity to rapidly identify their particular gene(s) and develop meaningful hypotheses to examine gene function in the laboratory. Improvements in the speed and sophistication of sequence alignment algorithms and computer performance have enabled scientists to keep pace with the growth of sequences in the public databases. However, it is now evident that with availability of vast amounts of protein sequences, existing algorithms for sequence alignment may not be adequate.

The Smith and Waterman algorithm (S&W) (Smith and Waterman 1981) for sequence alignment is one of the most

important techniques in computational molecular biology. The ingenious dynamic programming approach was designed to reveal the highly conserved fragments by discarding poorly conserved initial and terminal sequence segments. However, the notion of similarity has raised some concerns. For example, while the S&W finds the alignment with maximal score, it is unable to find an alignment with maximum degree of similarity (Arslan et al 2001). Another concern surrounds the use of gap penalties, which are an integral component of mainstream alignment programs. Penalties restrict the initiation and extension of gaps, which attempt to model insertion and/or deletion (indels) events that have occurred in either or both sequences through the course of evolution. They can be modified in combination using a substitution matrix (Smith and Waterman 1981; Pearson 1996; Apostolico and Giancarlo 1998). The scientific literature includes numerous empirical studies that prescribe generalised gap penalty parameters for ‘typical’

Correspondence: Matthew Bellgard, Centre for Bioinformatics and Biological Computing, School of Information Technology, Murdoch University, Murdoch, WA 6150, Australia; tel +61 8 9360 6088; fax +61 8 9360 7238; email m.bellgard@murdoch.edu.au

situations (Smith and Waterman 1981; Altschul 1991, 1993; Pearson 1995; Barton 1996). Even a zero gap penalty has been suggested (Roytberg 1998; Morgenstern 1999). However, it is now widely accepted that the use of a specific set of penalty parameters is not selective and/or sensitive enough for all types of alignments (Pearson 1995; Apostolico and Giancarlo 1998; Roytberg 1998; Morgenstern 1999; Arslan et al 2001).

In this paper, we propose an alignment approach that does not employ a gap penalty, and refer to it as gap mapping. For a given  $k$  gaps, the alignment algorithm gives us the alignment with the best score with up to  $k$  gaps. That is, if we allow, for example, a maximum of four gaps in the sequences, the algorithm will find the best location to map these gaps to maximise the alignment score. This alignment score is calculated using a substitution matrix as in current algorithms.

This algorithm is a modification of the S&W's dynamic programming approach, in which a matrix of best partial match scores is populated via a recurrence relation. However, in gap mapping, we use an extra dimension on the matrix to allow us to constrain the number of gaps introduced. To match a sequence of length  $m$  against one of length  $n$ , allowing up to  $k$  gaps, we use an  $m$  by  $n$  by  $k$  matrix. The entry we eventually place in position  $a,b,c$  ( $a$  at most  $m$ ,  $b$  at most  $n$ , and  $c$  at most  $k$ ) is the best score obtainable by matching the length  $a$  prefix of the first sequence with the length  $b$  prefix of the second sequence, allowing up to  $c$  gaps. From this and neighbouring entries we can find the corresponding score at  $a+1, b, c$  and  $a, b+1, c$  and  $a, b, c+1$  etc. The recurrence relation is straightforward to specify albeit more complicated than that in S&W.

This algorithm is a variant of one proposed by Sankoff (1972). Another independent study has implemented a similar algorithm (Roytberg 1998). From now on we refer to our implementation as the SANKOFF method. Previously, there have been attempts to determine  $k$  (the number of gaps) to obtain a single optimal alignment (Sankoff and Cedergren 1973; Elleman 1978; Zhu et al 1997). Our approach is not to focus on trying to determine  $k$  but rather to provide several alignments, and let the biologist determine the appropriate number of gaps. In practice, we show that in many cases this does not result in a large number of alignments required to be reviewed.

To assess the performance of the SANKOFF method, we use a database of confirmed structurally aligned

sequences: BALiBASE (<http://www-igbmc.u-strasbg.fr/BioInfo/BALiBASE/>). The BALiBASE benchmark alignment database has been constructed to facilitate the unbiased, objective comparison of (multiple sequence) alignment programs. BALiBASE has been used in many recent studies evaluating alignment programs (Thompson et al 1999a; Notredame et al 2000; Karplus and Hu 2001). We discuss our findings.

## Material and methods

For this study, we mainly wish to compare the SANKOFF approach to the Smith and Waterman algorithm, SSEARCH (Smith and Waterman 1981). However, we also compare Fasta (Pearson and Lipman 1988), BLAST® (Altschul et al 1990), ClustalW (Thompson et al 1994), and our new SANKOFF algorithm using the BALiBASE database (Thompson et al 1999b). The gap mapping algorithm was implemented in Java. The programs were tested using local copies from the command-line (rather than web-based) implementations on a 686 class machine with dual Athlon MP 1533 MHz processors running RedHat Linux 7.2.

Currently, BALiBASE contains 143 reference alignments. It is split into five different categories, of which we used category 1. It contains more than 80 alignments of sequences that have a similar length. Within category 1 there also exists a number of subcategories: short sequences (60–130 residues); medium sequences (200–320 residues); and long sequences (415–1000 residues). Each of these groups is further divided on the basis of identity: less than 25%; 20%–40%; and greater than 35%.

As our study dealt with pairwise alignment only, the database was modified to extract the first pair of sequences from each set of sequences in BALiBASE. Thus, our test set consisted of between 7 and 12 pairs of sequences in each of the 9 groups identified above. Each alignment method was used to align each of the test cases, and the alignment was then compared to the reference alignment. Clearly, the way this comparison is quantified is of great importance. The simplest method is to simply compare the sequence identity for each attempted alignment with the reference alignment. However, this method is flawed as it does not take into account chance matching and, more importantly, does not look at the quality of the alignment. This issue was addressed by Elofsson (2002) and also earlier by Sauder et al (2000). Two measures were derived,  $F_m$  and  $F_d$ , that measure alignment quality by comparing the aligned residues with the residues in the reference alignment. However, the former

caters for local alignment techniques and the latter for global alignment techniques. That is,  $Fm$  only compares the number of correctly aligned residues derived from the local alignment, whereas  $Fd$  evaluates the global alignment with reference to the number of residues in the reference alignment. Clearly, local alignment methods such as Fasta, BLAST and SSEARCH produce higher  $Fm$  values than  $Fd$  values. Both these measures were used in this study. The programs and results are available at <http://www.cbbc.murdoch.edu.au>.

## Results and discussion

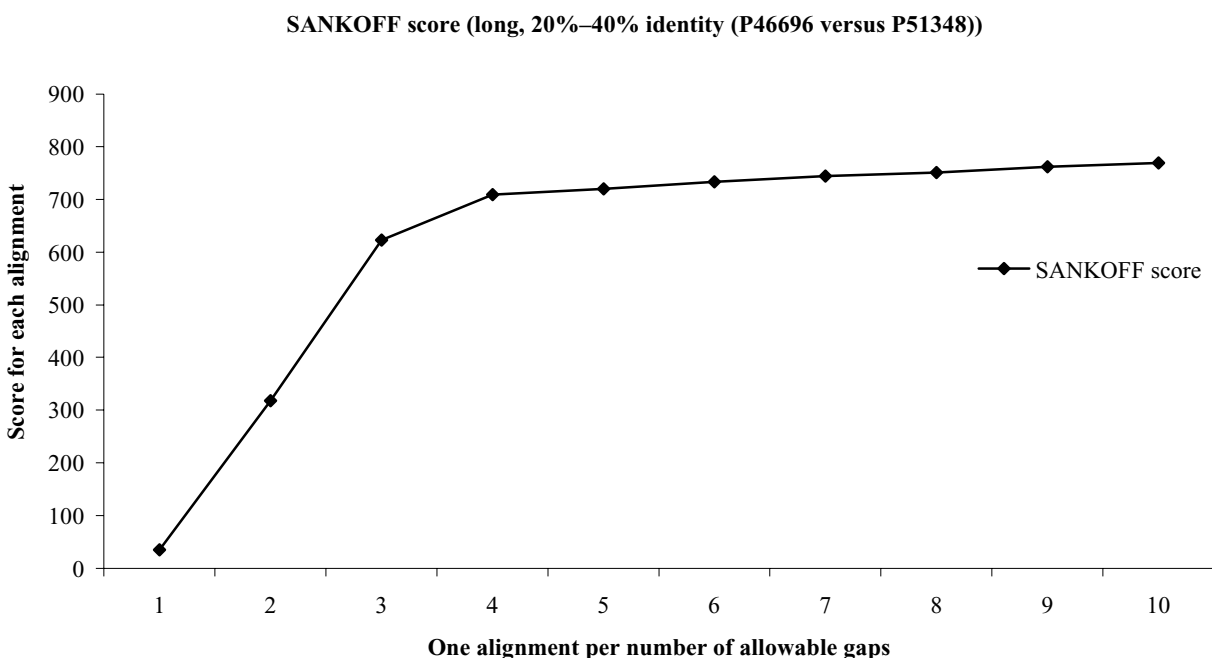
We have implemented a version of the SANKOFF algorithm. We have compared the pairwise protein alignment quality of Fasta, BLAST, ClustalW, SSEARCH and our implementation of the SANKOFF algorithm using the BALiBASE database. Default parameters and substitution matrices were used initially, although it is our intention to extend this work to include an analysis of the effect of parameter change on the quality of the alignment.

Figure 1 shows the results of the SANKOFF algorithm for a pair of sequences, each greater than 400 amino acids in length and known to be between 20% and 40% identical allowing no more than 11 gaps. Along the  $x$ -axis is an alignment constrained by the number of maximum gaps.

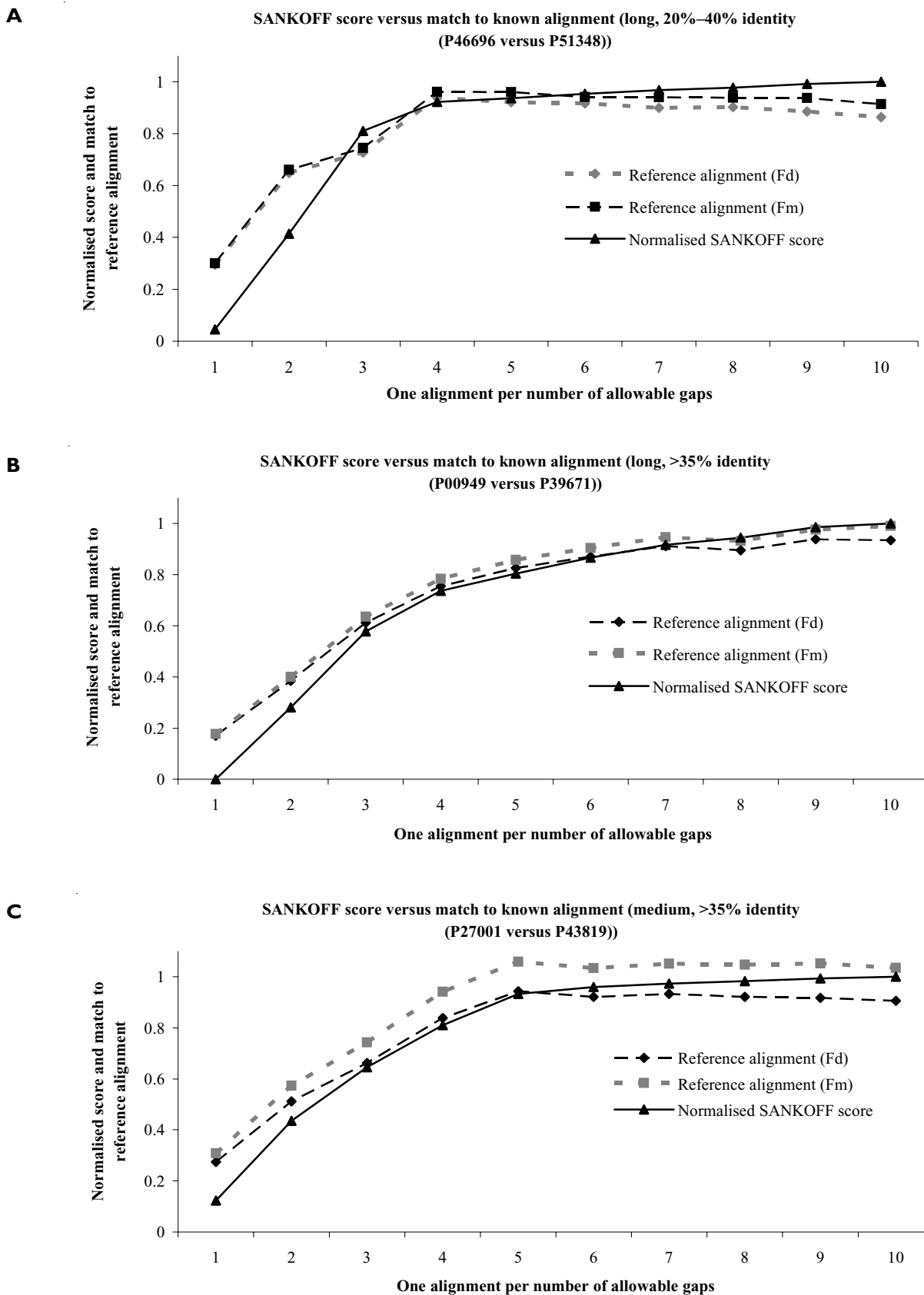
Thus, it is possible that each alignment could be different from each other. The  $y$ -axis plots the score obtained for each alignment using the BLOSUM50 substitution matrix. As can be seen, the score increases as more gaps are allowed into the alignment before the score asymptotes. We have found this to be a typical feature of the SANKOFF method.

Figure 2 shows the SANKOFF alignment score plotted against the performance of the alignment algorithm versus the reference alignment from BALiBASE. To do this, we simply obtain a normalised score (using the maximum and minimum score achieved for that alignment) and obtain a percentage using the  $Fd$  measure. As can be seen, where the SANKOFF typically asymptotes, its performance is close to optimal against the reference sequence alignment. The performance appears to drop off as the number of gaps is increased. This does not always imply that the alignment is completely wrong, as the criteria we used for comparing it to the reference alignment is very strict (Elofsson 2002).

For all the pairs of sequences examined, the SANKOFF method did not perform any worse than the other algorithms tested (results not shown). In most cases, it found an optimal/near-optimal alignment. In practice, there are not many competing alignments that need to be explored, as the optimal score is near the asymptote score produced by SANKOFF.



**Figure 1** The SANKOFF alignment scores for a given pair of sequences as the maximum allowable gaps increase. This pair of sequences is selected from BALiBASE.



**Figure 2** Three examples (A), (B) and (C) of the performance of the SANKOFF algorithm against the reference alignment from BALIbase. As the SANKOFF score asymptotes, the alignment approximates the reference alignment.

## Conclusion

We propose that the SANKOFF method for pairwise sequence alignment is a viable alternative to other mainstream alignment approaches. As it does not use a gap penalty, it is possible to a priori constrain the number of allowable gaps in the alignment. This means that it is possible to map (model) the number of gaps for a given pair of sequences using prior knowledge of the sequences such as evolutionary information. In addition, it could prove to be useful when aligning genomic sequences with coding DNA sequences where the number of gaps (introns) is known.

We intend to explore the properties of the SANKOFF algorithm further, as well as explore various criteria (different alignment algorithms, varying penalties, different substitution matrices etc) to produce robust pairwise alignments. To do this, we have developed what we refer to as a criteria alignment matrix (CAM). Investigation of this CAM will provide valuable information about the alignments, such as whether one alignment performs well under several criteria. While the alignment that maximises any given criteria (such as penalties for gaps) can be considered 'best' under the assumptions used to derive the criteria, other alignments that almost maximise the criteria may also be of interest. These might be: (a) near-optimal alignments that are biologically interesting; (b) near-optimal alignments that are very different to the optimal alignment; or (c) near-optimal alignments that assist development of novel optimality criteria.

## Acknowledgements

We would like to take this opportunity to acknowledge Michael Waterman and David Sankoff for their invaluable comments and suggestions via personal communications. We also wish to acknowledge Mr Yasuyuki Nozaki for revising and correcting the figures in this manuscript.

## References

- Altschul SF. 1991. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol*, 219:555–65.
- Altschul SF. 1993. A protein alignment scoring system sensitive at all evolutionary distances. *J Mol Evol*, 36:290–300.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*, 215:403–10.
- Apostolico A, Giancarlo R. 1998. Sequence alignment in molecular biology. *J Comput Biol*, 5:173–96.
- Arslan A, Egecioglu O, Pevzner P. 2001. A new approach to sequence comparison: normalized sequence alignment. *Bioinformatics*, 17:327–37.
- Barton GJ. 1996. Protein sequence alignment and database scanning. In Sternberg MJE, ed. *Protein structure prediction – a practical approach*. Oxford: IRL Pr.
- Elleman TC. 1978. A method for detecting distant evolutionary relationships between protein or nucleic acid sequences in the presence of deletions or insertions. *J Mol Evol*, 11:143–61.
- Elofsson A. 2002. A study on protein sequence alignment quality. *Proteins*, 46:330–9.
- Karplus K, Hu B. 2001. Evaluation of protein multiple alignments by SAM-T99 using the BALiBASE multiple alignment test set. *Bioinformatics*, 17:713–20.
- Morgenstern B. 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15:211–18.
- Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302:205–17.
- Pearson W. 1995. Comparison of methods for searching protein sequence databases. *Protein Sci*, 4:1145–60.
- Pearson W. 1996. Effective protein sequence comparison. *Methods Enzymol*, 266:227–58.
- Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA*, 85:2444–8.
- Roytberg MA. 1998. Sequence alignment without gap penalties [online]. Proceedings of the First International Conference on Bioinformatics of Genome Regulation and Structure. 1998 Aug 24–31; Novosibirsk, Altai Mountains, Russia. Accessed 2 Oct 2003. URL: <http://www.bionet.nsc.ru/bgrs/thesis/85/index.html>
- Sankoff D. 1972. Matching sequences under deletion/insertion constraint. *Proc Natl Acad Sci USA*, 69:4–6.
- Sankoff D, Cedergren R. 1973. A test for nucleotide sequence homology. *J Mol Biol*, 77:159–64.
- Sauder JM, Arthur JW, Dunbrack RL Jr. 2000. Large-scale comparison of protein sequence alignment algorithms with structural alignments. *Proteins*, 40:6–22.
- Smith T, Waterman M. 1981. Identification of common molecular subsequences. *J Mol Biol*, 147:195–7.
- Thompson JD, Higgins DG, Gibson TJ. 1994. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22:4673–80.
- Thompson JD, Plewniak F, Poch O. 1999a. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*, 27:2682–90.
- Thompson JD, Plewniak F, Poch O. 1999b. BALiBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 15:87–8.
- Zhu J, Liu J, Lawrence C. 1997. Bayesian adaptive alignment and inference. *Proc Int Conf Intell Syst Mol Biol*, 5:358–68.