Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W613–W616***
doi:10.1093/nar/gkl288

AliWABA: alignment on the web through an A-Bruijn approach

Neil C. Jones, Degui Zhi and Benjamin J. Raphael*

Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA

Received February 14, 2006; Revised March 1, 2006; Accepted April 5, 2006

ABSTRACT

Multiple sequence alignment programs are an invaluable tool in computational biology. A-Bruijn Alignment (ABA) is a method for multiple sequence alignment that represents an alignment as a directed graph and has proved useful in aligning nucleotide and amino acid sequences that are composed of repeated and shuffled subsequences. AliWABA is a web server that provides tools to generate alignments with ABA, visualize the resulting ABA graphs and extract subsequences from ABA graphs. AliWABA greatly simplifies the problem of analyzing multiple sequences for local similarities that may be reordered, as is common with the domain architectures of proteins. To facilitate the analysis of protein domains, AliWABA provides direct querying of the Conserved Domain Database. Availability: http://aba.nbcr.net/

INTRODUCTION

Aligning and comparing multiple biosequences is a fundamental task in molecular biology. However, almost all algorithms that have been designed to align two or more sequences enforce a linearity constraint on the resulting alignments. That is, suppose we are given the following two 'sequences of blocks': ABCDE and ABECD. All global alignment algorithms, even those that purportedly solve the global alignment problem exactly—will fail to recognize the complete set of similarities between these two sequences (Figure 1). This failure is not a result of a deficiency in the programs themselves, but rather because of a limitation in the alignment representation: namely, only similarities in the same order and orientation are recognized. In many contexts, such as in the analyses of multi-domain proteins, alternative splicing or repeat families (1), this limitation obscures relationships that would otherwise shed light on the biological origin or function of a particular set of sequences (2). Whenever a biological sequence is best modelled as a collection of independently-acting domains that may be shuffled, reversed or reordered, one would prefer an alignment algorithm that could reveal all similarities without the constraint of linearity of context.

A few algorithms have been proposed to address the more general alignment problem hinted at above, for example TBA (3), POA (2), and ABA (4). ABA, or A-Bruijn Alignment relies on the formalism of the A-Bruijn graph (5) which is generated from a set of input sequences and a set of pairwise local alignments between the sequences. ABA produces a simplified A-Bruijn graph where each edge in the graph corresponds to either a substring of one input sequence or to a local alignment among two or more sequences. Each edge in the ABA graph has a consensus sequence and a multiplicity, which is the number of substrings that align to that consensus sequence. Each input sequence has a corresponding start node and end node in the ABA graph. Tracing a path from an input sequence's start node to its end node and concatenating the substrings that make up each edge in the path will reconstruct that input sequence. Edges within such a path with a multiplicity of 2 or more represent a similarity either to some portion of another sequence or to some other portion of the same sequence. Cycles in the ABA graph represent complex non-linear alignments. In the case of DNA sequences, the ABA graph can be constructed on a set of sequences and their reverse complements to yield modular relationships that span the forward and reverse strands of a DNA molecule. We note that the generation of an A-Bruijn graph is dependent on some other algorithm that can identify local similarities, and it is acceptable for this alignment algorithm to enforce the linearity constraint. In that respect, ABA acts as sophisticated 'glue' to form a more comprehensive global alignment from individual local alignment blocks.

The AliWABA web server at aba.nbcr.net provides an implementation of the ABA algorithm described in (4) and includes a comprehensive set of tools for visualizing, manipulating and analyzing the resulting ABA graphs.

*To whom correspondence should be addressed. Tel: +1 858 534 9989; Fax: +1 858 534 7029; Email: braphael@cs.uscd.edu
*Correspondence may also be addressed to Neil C. Jones. Email: ncjones@cs.uscd.edu

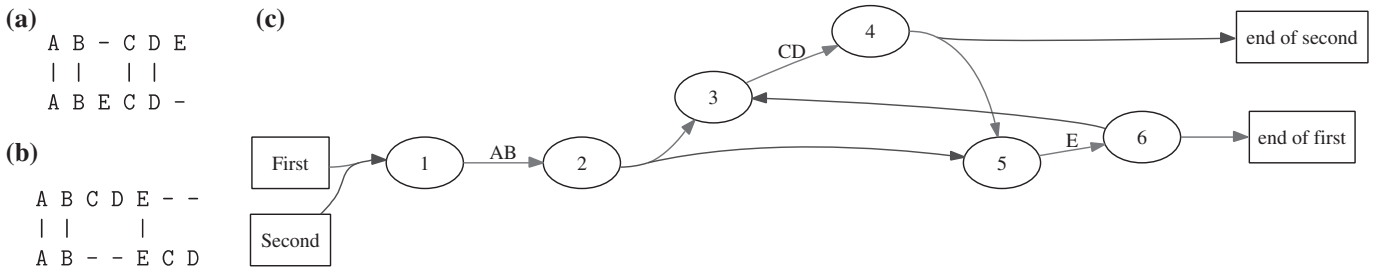


Figure 1. Existing alignment tools force linearity on the output alignments. (**a** and **b**) The 'optimal' alignments produced by traditional global alignment programs are incomplete because they do not reveal the similarity of blocks C, D and E in a single alignment. (**c**) The ABA graph representation of the same pair of sequences. In the ABA graph, each sequence is a path from a start node to an end node (but not all such paths are actually sequences) and edges correspond to substrings of a sequence of some (possibly 0) length. Note that ABA graphs are usually more complicated than this example and contain additional labelling information along the edges (Figure 2).

THE AliWABA WEB SERVER

Alignment on the web through an A-Bruijn Approach (AliWABA) starts with the submission of a Pearson (FASTA) formatted set of DNA or protein sequences. The sequences are examined for local similarities using either BLAST (for DNA or proteins) or CrossMatch (http://www.phrap.org; useful for certain kinds of DNA sequences, such as repeat libraries). Alternatively, the user can upload a set of pairwise local alignments along with the input sequence set for use in the construction of the ABA graph. AliWABA builds the ABA graph and presents it to the user after layout by GraphViz (6). AliWABA is implemented as a set of CGI scripts that operate on persistent session-based data.

The display of the ABA graph can be controlled in four main ways. The graph itself can be resized to any number of pixels and can also be downloaded in either PDF format or in DOT (6) format for editing. The display of the graph can be changed from its default stretched-out mode, where the start of each sequence is on the left hand side, and the end is on the right, to an energy-minimized format for compactness on small displays. In addition, the graph can be simplified by contracting groups of edges that are shorter than a user-provided width into a single vertex. (In AliWABA, collapsed groups of short edges are denoted as sXX where XX is an integer, and drawn as boxes instead of circles.) This feature helps overcome a complication that arises when pairs of sequences align along short segments, resulting in a large number of uninformative edges that may visually distract from the more important domain architecture. Finally, one can zoom in on any subgraph of the ABA output and view only the edges pertinent to that subgraph.

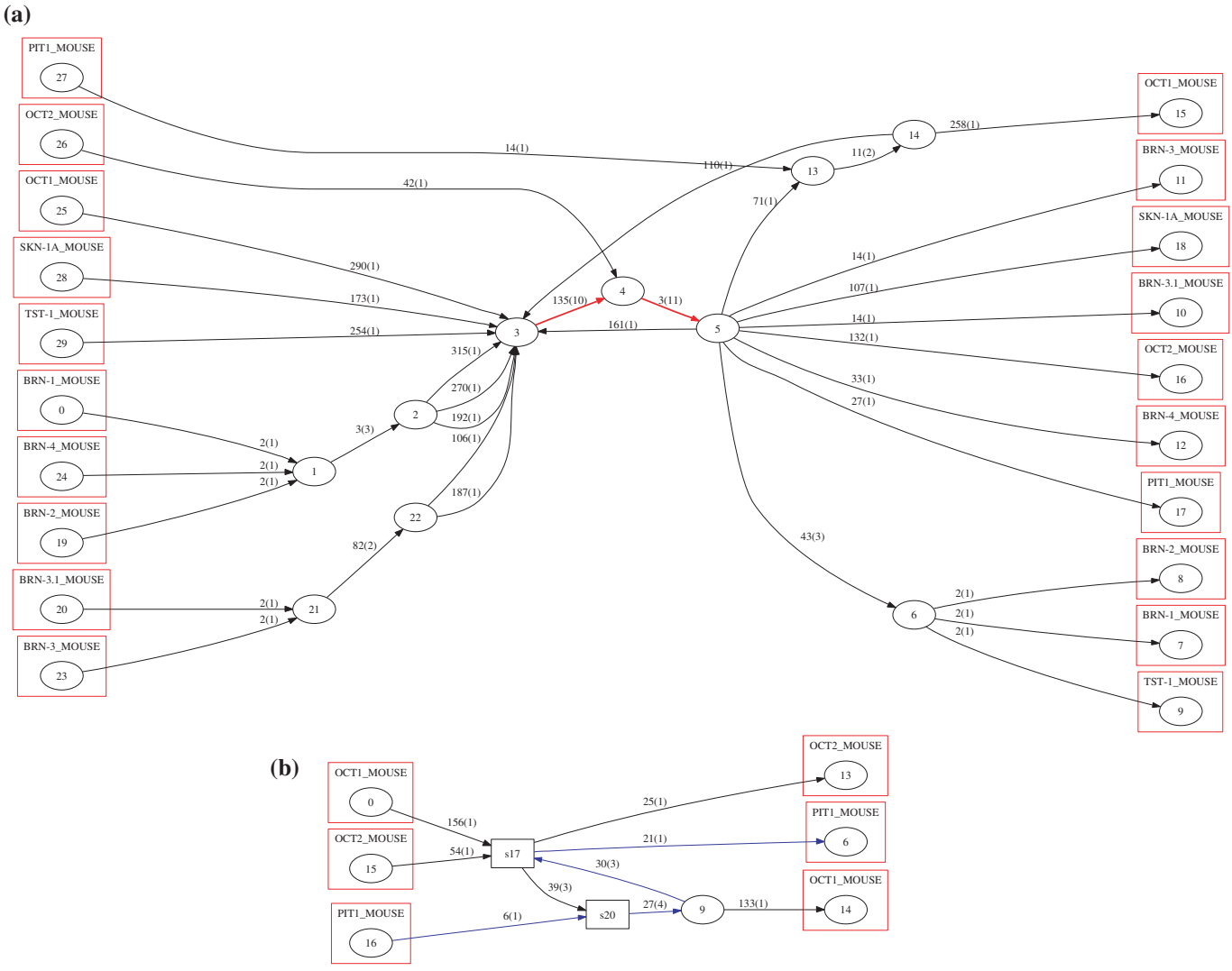
From the visual representation of the ABA graph, the user may select edges or paths in several ways: (i) explicitly, for example by listing sets of edges; (ii) by asking for all edges that obey some constraint on either multiplicity or length, e.g. all edges with more than four sequences that align or that are longer than 200 characters; (iii) by finding edges incident to a node that obeys some constraint on in- or out-degree or (iv) by extending the existing selection by some number of edges in any direction. The user may also iteratively combine selections, using unions, intersections or set differences, thereby allowing an arbitrary portion of the graph to be selected. For a given selection, the user may view information about the sequences under study. The sequence information is presented to the user in FASTA format so that subparts of the alignment may be further analyzed with other tools; for example local alignments of interesting protein domains may be created with TCoffee (7), ClustalW (8), DIALIGN-T (9), or MUSCLE (10); or domain profiles for the sequences on an edge may be created with HMMer (http://hmmer.wustl.edu). For a given graph selection, AliWABA can display a multiple sequence alignment produced by ClustalW and present either the verbatim report or an enhanced version through Jalview (11). Finally, a given subgraph can be queried against the CDD (12) database using the RPSBlast tool from the NCBI toolbox.

ALIGNMENT OF POU DOMAIN TRANSCRIPTION FACTORS

The ABA graph representation of alignments enables one to see a higher level of homology among a set of sequences by simultaneously displaying multiple compatible alignments among that set. This example, which recapitulates some knowledge of a class of transcription factors demonstrates that more insight can be gained from an ABA graph than from a traditional multiple sequence alignment.

The transcription factors Oct-1, Oct-2, Skn-1a, Tst-1, Pit-1, Brn1, Brn2, Brn3, Brn3.1 and Brn4 each contain a region described as a POU domain, which contains two parts: a POU-specific domain and a POU-homeodomain. Both subdomains are necessary for regulatory activity. It is known that POU domain factors can be organized into six main classes (13) based on a phylogeny inferred from the POU-specific subdomain. When these protein sequences are input into AliWABA, several known features of the POU family become readily apparent (Figure 2). For example, a single high-multiplicity edge, of length 135 amino acids corresponds to both the POU binding domain (with an E-value of 4×10^{-10} when the consensus sequence along this edge is searched against CDD) and a homeodomain (E-value of 2×10^{-8}). (Because both domains are separated by short stretches of amino acids, any reasonable alignment will include both.) The factors Brn3 and Brn3.1 share an 82 amino acid alignment ~200 residues before the POU-domain cluster; while this shared sequence does not correspond to any known domain in the Conserved Domain Database, it contains the POU-IV Box described in (14).

Interestingly, Pit-1 and Oct-1 have a shared segment that appears before the POU-domain cluster in Pit-1, but after



(c)

```
OCT1_MOUSE:522-552      -GTTDSTSNNNTATVISTAPPASSAVTSPSL 30
PIT1_MOUSE:31-60        --AECLPASNHATNVMSTATGLHYSVPSCHY 29
OCT2_MOUSE:419-450      GGGGGRALPLNSIPSVTPPPPATTNSTNPSP 31
                         .    .   ::   ::...      ..
```

Figure 2. (**a**) The ABA graph generated by AliWABA on a set of ten POU-domain transcription factors of varying biological function. Edges are labeled *l(m)* where *l* is the length of the alignment on the edge and *m* is its multiplicity (the number of sequences in the alignment). The POU-specific and POU-homeodomains are captured in the edge (3,4), which has length 135 and multiplicity 10. (**b**) When Oct-1, Oct-2, and Pit-1 are aligned separately, a previously unknown shared domain of length 30 becomes apparent. This domain is shuffled in Pit-1 (path traced in blue) versus Oct-1 and Oct-2. (**c**) The alignment between Oct-1 and Pit-1 that corresponds to the edge (9, *s*17). This example can be seen at http://aba.nbcr.net/aba-pou.html.

the cluster in Oct-1. Upon closer inspection of these two protein products, it appears that they might share a longer shuffled domain (62 amino acids), but the sequence similarity along this region (40%) is low enough that it is not obviously informative. In any case, if a linear multiple alignment were enforced on this set of biologically important sequences, at least one of these alignments would not be discovered.

ACKNOWLEDGEMENTS

BJR is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. Funding to pay the Open Access publication charges for this article was provided by the UC Foundation/Ronald R. Taylor Chair ITCS/PEVZNER.

Conflict of interest statement. None declared.

REFERENCES

1. Zhi,D., Raphael,B., Price,A., Tang,H. and Pevzner,P. (2006) Identifying repeat domains in large genomes. *Genome Biol.*, **7**, R7.
2. Lee,C., Grasso,C. and Sharlow,M.F. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–464.

3. Blanchette,M., Kent,J., Riemer,C., Elnitski,L., Smit,A., Roskin,K., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E. *et al*. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*., **14**, 708–715.
4. Raphael,B., Zhi,D., Tang,H. and Pevzner,P. (2004) A novel method for multiple alignment of sequences with repeated and shuffled elements. *Genome Res*., **14**, 2336–2346.
5. Pevzner,P.A., Tang,H. and Tesler,G. (2004) De novo repeat classification and fragment assembly. *Genome Res*., **9**, 1786–1796.
6. Koutsofios,E. and North,S.C. (1993) Drawing graphs with dot. Technical Report, AT&T Bell Laboratories, Murray Hill, NJ.
7. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol*., **302**, 205–217.
8. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T., Higgins,D. and Thompson,J. (2003) Multiple sequence alignment with the CLUSTAL series of programs. *Nucleic Acids Res*., **31**, 3497–3500.
9. Subramanian,A., Weyer-Menkhoff,J., Kaufmann,M. and Morgenstern,B. (2005) DIALIGN-T, an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, **6**, 66.
10. Edgar,R. (2004) Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.
11. Clamp,M., Cuff,J., Searle,S. and Barton,G. (2004) The Jalview Java alignment editor. *Bioinformatics*, **12**, 426–427.
12. Marchler-Bauer,A., Anderson,J., Cherukuri,P., DeWeese-Scott,C., Geer,L., Gwadz,M., He,S., Hurwitz,D., Jackson,J., Ke,Z. *et al*. (2005) CDD: a conserved domain database for protein classification. *Nucleic Acids Res*., **33**, D192–D196.
13. He,X., Treacy,M., Simmons,D., Ingraham,H., Swanson,L. and Rosenfeld,M. (1989) Expression of a large family of POU-domain regulatory genes in mammalian brain development. *Nature*, **340**, 35–41.
14. Gerrero,M., McEvilly,R., Turner,E., Lin,C., O'Connell,S., Jenne,K., Hobbs,M. and Rosenfeld,M. (1993) Brn-3.0: A POU-domain protein expressed in the sensory, immune, and endocrine systems that functions on elements distinct from known octamer motifs. *Proc. Natl Acad. Sci. USA*, **90**, 10841–10845.