BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 12 2002
Pages 1703–1704

OWEN: aligning long collinear regions of genomes

Aleksey Y. Ogurtsov[1,], Mikhail A. Roytberg[2], Svetlana A. Shabalina[1] and Alexey S. Kondrashov[1]*

[1]National Center for Biotechnology Information, NIH, 45 Center Drive, Bethesda, MD 20892-6510, USA and [2]Institute of Mathematical Problems in Biology, 142290, Pushchino, Moscow Region, Russia

Received on February 4, 2002; revised on May 9, 2002; accepted on May 21, 2002

ABSTRACT

Summary: OWEN is an interactive tool for aligning two long DNA sequences that represents similarity between them by a chain of collinear local similarities. OWEN employs several methods for constructing and editing local similarities and for resolving conflicts between them. Alignments of sequences of lengths over 10^6 can often be produced in minutes. OWEN requires memory below 20 L, where L is the sum of lengths of the compared sequences.

Availability: Pre-compiled versions of OWEN for Windows, Linux for PC, SunOS for Sun, and Irix for SGI, user manual, and examples of alignments are freely available at ftp://ftp.ncbi.nih.gov/pub/kondrashov/owen.

Contact: ogurtsov@ncbi.nlm.nih.gov

INTRODUCTION

OWEN, named after a scientist who developed the concept of homology (Owen, 1848) is a software tool for aligning pairs of long sequences based on greedy paradigm (Roytberg *et al.*, 2002). Since information essential for correctly aligning sequences that are not uniformly similar to each other is often difficult to formalize, OWEN allows human intervention at every step of the alignment process. Constructing a detailed alignment usually takes 5–15 iterative steps, since extensive (e.g. between exons) and short (e.g. between conserved regulatory sites) local similarities can hardly be found simultaneously. OWEN organizes these iterations into a natural hierarchy, since new similarities are sought only between the already constructed similarities. Most of the actions performed by OWEN deal with constructing and editing local similarities and with resolving conflicts between them. An early version of OWEN has been used for aligning 100 orthologous intergenic regions from human and mouse genomes (Shabalina *et al.*, 2001).

*To whom correspondence should be addressed.

INPUT AND OUTPUT

OWEN imports sequences from FASTA or GenBank files. In the latter case, it can also utilize information on annotated repeats. Alternatively, OWEN can use its own output files, in order to refine alignments created earlier.

Output of OWEN consists of the backbone chain (Roytberg *et al.*, 2002) of local similarities and, optionally, of some extra similarities that are in conflict with this chain. The chain can be recorded either as a succession of local alignments, or as the global alignment of the whole compared sequences which includes all reliable local alignments, marked as such, alternating with regions where alignment may be unreliable.

USER INTERFACE

The root window of an OWEN session consists of a dot matrix which displays all current local similarities between the compared sequences, the list of locations of these similarities, pull-down menu (with headings: FILE, CONSTRUCT, CONFLICT, FILTER, SELECT, EDIT, WINDOW), and the information line at the bottom (Figure 1). Subordinate windows displaying alignments between only parts of the sequences can be opened and closed. Actions taken in the course of a session can be reversed and restored (Undo and Redo actions).

CONSTRUCTING LOCAL SIMILARITIES

Actions Align, Expand, Merge, and Refine construct and edit local similarities. Construction occurs only in the areas of the dot matrix which are not in conflict with the current similarities.

Action Align constructs new local similarities, treated as chains of dense ungapped segments (islands). An island must contain at least m matches within each frame of length n, and must totally contain at least k matches. Local similarities are assembled from all islands as their unexpandable chains, with a penalty for linking two islands depending on the distance between them. Only

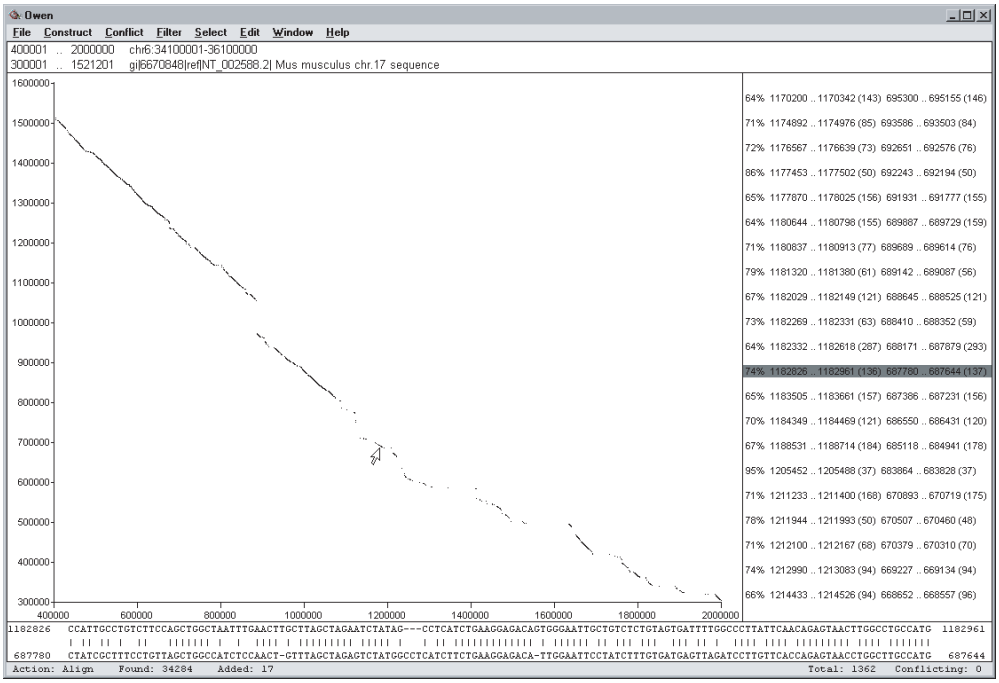

Fig. 1. Root window of an OWEN session aligning murine sequence (GenBank GI: 6670848) with the orthologous region of human chromosome 6. The present chain of local similarities consists of 1362 elements.

similarities having *P*-values below some threshold are kept.

This algorithm runs in time proportional to the area of the dot-matrix where similarities are sought. On an average PC, constructing all local similarities between two sequences of length 10^5 takes $\sim$5 minutes. However, there is a possibility to speed Align up, by seeking only those similarities that contain at least h consecutive matches. Finding all such similarities between sequences of length 10^6 takes only 3 seconds if $h = 16$.

Action Expand extends the already constructed similarities using less stringent parameters. Action Merge constructs unified similarities from those that come close to each other. Action Refine realigns loose regions of local similarities using the Needleman–Wunsch algorithm.

RESOLVING CONFLICTS

Actions Reconcile, Greedy, Optimal, and Kill resolve conflicts between local similarities. Action Reconcile resolves unessential conflicts by trimming the ends of local similarities that overlap only slightly. Action Greedy resolves, starting from similarities with low *P*-values, each pairwise conflict by deleting the similarity with a higher *P*-value (Roytberg *et al.*, 2002). Action Optimal keeps local similarities that together form the optimal chain (Zhang *et al.*, 1998) and deletes all others. Action Kill deletes all conflicting similarities.

Despite this variety of options, it is often advisable, especially at the final stages of constructing a detailed alignment, to resolve conflicts by deleting some of the conflicting similarities manually. Deleted similarities can be recorded as footnotes to the backbone chain.

FILTERING

Repeats complicate aligning and, thus, should be masked, at least until all meaningful local similarities of unique sequences are constructed. OWEN uses filters of three kinds which mask: (i) segments annotated as repeats, (ii) segments which, after a set of local similarities has been constructed, were found to align with more than one segment on the other sequence, and (iii) segments of low complexity.

REFERENCES

Owen,R. (1848) *On the Archetype and Homologies of the Vertebrate Skeleton*. Van Voorst, London.

Roytberg,M.A., Ogurtsov,A.Y., Shabalina,S.A. and Kondrashov,A.S. (2002) A hierarchical approach to aligning collinear regions of genomes. *Bioinformatics*, this issue.

Shabalina,S.A., Ogurtsov,A.Y., Kondrashov,V.A. and Kondrashov,A.S. (2001) Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.*, **17**, 373–376.

Zhang,Z., Berman,P. and Miller,W. (1998) Alignments without low-scoring regions. *J. Comput. Biol.*, **5**, 197–210.