OWEN-SCRIPT – AN EXTENDED TOOL FOR PAIRWISE GENOME ALIGNMENT

Ogurtsov A.Yu.[*1]***, Vasilchenko A.N.***[2]***, Vlasov P.K.***[3]***, Shabalina S.A.***[1]***,***
Kondrashov A.S.[1]***, Roytberg M.A.***[*2]
[1] National Center for Biotechnology Information, NIH, 45 Center Drive, Bethesda, MD 20892-6510, USA; [2] Institute of Mathematical Problems in Biology, Pushchino, Moscow Region, 142290, Russia;
[3] Institute of Molecular Biology, Moscow, 117234, Russia
[*] Corresponding authors: e-mail: Ogurtsov@ncbi.nlm.nih.gov, Roytberg@impb.psn.ru

Key words: pairwise alignment, genomes, hierarchical approach

SUMMARY

Motivation: A genome alignment is an important instrument of post-genomic computational biology. The commonly available tools (LAGAN, BLAT, YASS, etc) designed for command line mode and thus tend to loose some similarities without any possibility for user to learn about this. In contrast, the OWEN is an interactive tool, allowing user to control the alignment process and to be sure that no interesting events were lost. However, one may need tools to store some alignment protocols that are suitable for a class of similar situations and then implement the protocol automatically.

Results: We propose OWEN-SCRIPT, an extension of the OWEN program thet al.lows to perform OWEN based scripts. The commands of the scripts correspond to the actions of interactive OWEN. Examples of protocols obtained from alignment human and mouse genomes are also available.

Availability: Program OWEN-SCRIPT is available on request from the authors.

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). Unlike other popular tools (e.g. LAGAN (Brudno *et al.*, 2003), YASS (Noe, Kucherov, 2005), etc.) OWEN is an interactive tool and allows human intervention at every step of the alignment process. This makes the user sure that (s)he did not miss any essential similarity. Constructing a detailed alignment usually takes 5–15 iterative steps; each of steps consists of constructing and editing local similarities and with resolving conflicts between them.

However an alignment protocols invented during the interactive work can be adequate for a series similar cases. Because of this we have implemented in the OWEN a script option. The script commands are in almost one-to-one correspondence with the interactive actions and thus any alignment protocol can be represented with a proper script; the script then can be used to align automatically proper genome pairs.

METHODS AND ALGORITHMS

OWEN actions: an overview. OWEN session starts with the determination of input data. During the session OWEN stores a set of local alignments. All alignments can be divided in two classes: those that are in conflict with some other alignments, and the non-conflicting ones, i.e. those are collinear to any other alignment (two alignments are collinear if segments involved in one of them precede in both sequences segments involved in the other, and are in conflict otherwise). The aligning with OWEN consists mainly of creating, editing, and deleting local alignments; the corresponding actions are listed under CONSTRUCT, CONFLICT and FILTER items of the main menu.

Actions listed under CONSTRUCT create new and modify the already present alignments, e.g. *Align* creates new alignments in areas defined by the present alignments (so that the new alignment cannot be in conflict with any of the present ones); *Expand* extends existing alignments. Actions listed under CONFLICT resolve conflicts between alignments by trimming conflicting alignments (*Reconcile*) or by completely deleting some of them (*Greedy, Optimal*, and *Kill*). Actions listed under FILTER can create, update, and delete the filter. A filter is a list of segments in both sequences that (i) are annotated as repeats, and/or (ii) are aligned with several segments in the other sequence, and/or (iii) have low complexity. Segments included in the filter can be masked when actions *Align* and *Expand* are performed.

The ultimate goal of a session is to construct the best (from the user's point of view) chain of non-conflicting alignments, then to fill the gaps between them by the algorithm of global sequence alignment and thus to obtain the global alignment of the given sequences. However, the user can produce and save different global alignment and/or save intermediate sets of local alignments that possibly contain conflicts.

IMPLEMENTATION

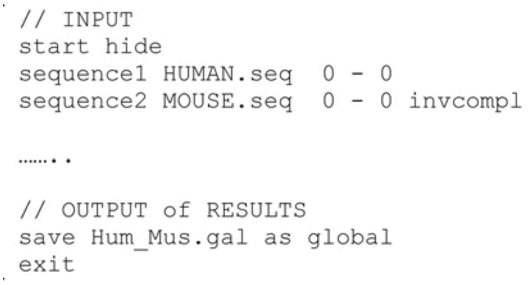
OWEN command file. The OWEN command file is a text file, each it's line is an operator, describing an action of OWEN and parameters of the action. For example, sequence1 chr6_hum.seq 1–1000000 causes an input of first million of nucleotides from a file chr6_hum.seq as the first sequence of the pair to be aligned. The operator align p = 0.000001 w = 16 5/8 = 12 nomask leads to generation of all local similarities, which have P-value below p = 0.000001, and are detectable with given parameters of the algorithm, e.g. they should contain at least 16 consecutive matches. The sequences to be analyzed should be prepared by preceding operators. All OWEN actions can be represented with proper operators. The complete list of operators (commands) and corresponding actions is given in the Manual, available at ftp://ftp.ncbi.nih.gov/pub/kondrashov/owen. The web site also contains templates of command files; using the templates one can create command files adjusted to a typical biological problems. Protocols of genome alignments and their script representations are also discussed in (Ogurtsov, 2005).

The operators of OWEN-SCRIPT command file are performed one by one, condition operators are not allowed. We have declined implementation of BASIC-like command language, because the developed tool was sufficient to solve all problems arisen in our work.

The general form of the OWEN command file is given on Fig. 1.

The command file *owen.cmd* can be executed with the command>owen owen.cmd Scripts based on OWEN command files.

A simple, but important way to extend abilities to describe alignment tasks is to utilize UNIX scripts, MS WINDOWS batch files, or analogous resources of other operating systems. This is the way, for example, to prepare a task to align a large set of sequence pairs.



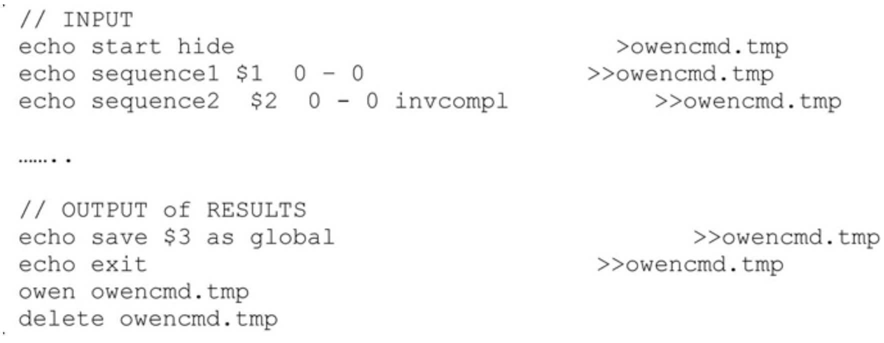
```
// INPUT
start hide
sequence1 HUMAN.seq  0 - 0
sequence2 MOUSE.seq  0 - 0 invcompl

…….

// OUTPUT of RESULTS
save Hum_Mus.gal as global
exit
```

Figure 1. General form of the OWEN command file. The file determines alignment of the sequences from files HUMAN.seq and MOUSE.seq; the latter should be invert-complemented. Results will be stored in the file Hum_Mus.gal.

Indeed, having an OWEN command file (see Fig. 1), describing an alignment protocol, one can easily create a UNIX script (see Fig. 2) that provides an alignment of given sequence according the protocol. To obtain the script one needs (1) add "echo" at the beginning of each line of the command file; (2) add ">> owencmd.tmp" at the end of each line or "> owencmd.tmp" at the end of the 1st line; (3) add two lines at the bottom of the file:

owen owencmd.tmp
delete owencmd.tmp

The obtained script will create a file owencmd.tmp, which is a copy of an initial command file, then run OWEN with the command file and delete command file. By substitution of any parameter of the script with "$1", "$2", etc. one can obtain a parameterized script (see Fig. 2).



```
// INPUT
echo start hide                             >owencmd.tmp
echo sequence1 $1  0 - 0                   >>owencmd.tmp
echo sequence2  $2  0 - 0 invcompl             >>owencmd.tmp

…….

// OUTPUT of RESULTS
echo save $3 as global                           >>owencmd.tmp
echo exit                                   >>owencmd.tmp
owen owencmd.tmp
delete owencmd.tmp
```

Figure 2. UNIX script obtained from the command file given on Fig. 1. The names of input and output files are described as parameters of the script.

For example, suppose, that the file Align-1.sh contains a copy of the script from Fig. 2. Then for any files seq_A.txt and seq_B.txt the script Align-1.sh seq_A.txt seq_B.txt Result_AB will provide the alignment of the sequences from the files according the protocol of Fig. 1 and output results to the file Result_AB. The script Align-1.sh, in turn, can be called from another script, etc. Examples of scripts can be found at ftp://ftp.ncbi.nih.gov/pub/kondrashov/owen.

Basic tools, environment and architecture. The OWEN-SCRIPT's source is portable. It is written on ANSI C++, the total volume is ~ 10 000 lines. Graphic interface is based on the Fox-toolkit (see http://www.fox-toolkit.org/). All libraries are linked as static, this guaranties that executable module can be downloaded and run *per ce* on user's computer with the same processor type.

OWEN's architecture can be represented as a finite automaton. Receiving an input signal (user's click in interactive mode or command line in a batch mode), it performs a corresponding action. The list of actions is given in the Manual. The data structures are mainly same as in previous version of OWEN (Ogurtsov *et al.*, 2002). The main data type is *a box*, i.e. a pair of fragment U[a1, a2] and V[b1, b2] of given sequences U and V. For each box we remember a non-conflicting chain of local similarities ("backbone chain", see (Roytberg *et al.*, 2002)), its score and some other values. The boxes are arranged in 3 trees, which support quick search by both coordinates of a block and its score.

CONCLUSION

OWEN-SCRIPT is a powerful tool and have been used in many works (see e.g. (Bazykin *et al.*, 2004; Ogurtsov *et al.*, 2004; Shabalina *et al.*, 2004). The main advantage of the tool is its ability to fit the specificity of the data and then reproduce the obtained procedure of analysis.

ACKNOWLEDGEMENTS

This work was supported by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS. Roytberg M.A. acknowledges financial support by the Russian Foundation for Basic Research (project Nos 03-04-49469, 02-07-90412), by grant from the RF Ministry for Industry, Science, and Technology (20/2002, 5/2003), NWO, ECO-NET, NIH (grant TW005899, co-PIs V. Tumanyan and M. Borodovsky) and by the program of RF Ministry of Science and Education (contract No. 02.434.11.1008).

REFERENCES

Bazykin G.A., Kondrashov F.A., Ogurtsov A.Y., Sunyaev S., Kondrashov A.S. (2004) Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. *Nature*, **429**(6991), 558–562.

Brudno M., Do C.B., Cooper G.M., Kim M.F., Davydov E., Green E.D., Sidow A., Batzoglou S. (2003) NISC Comparative Sequencing Program. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*. **13**(4), 721–731.

Noe L., Kucherov G. (2005) YASS: enhancing the sensitivity of DNA similarity search. *Nucl. Acids Res.*, **33**(Web Server issue), W540–543.

Ogurtsov A.Iu. (2005) A protocol of automatic alignment of genome sequences using the program OWEN. *Biofizika*, **50**(3), 475–479. (In Russ.).

Ogurtsov A.Y., Roytberg M.A., Shabalina S.A., Kondrashov A.S. (2002) OWEN: aligning long collinear regions of genomes. *Bioinformatics*, **18**, 1703–1704.

Ogurtsov A.Y., Sunyaev S., Kondrashov A.S. (2004) Indel-based evolutionary distance and mouse-human divergence. *Genome Res.*, **14**(8), 1610–1616.

Owen R. (1848) On the archetype and homologies of the vertebrate skeleton. London, John Van Voorst.

Roytberg M.A., Ogurtsov A.Y., Shabalina S.A., Kondrashov A.S. (2002) A hierarchical approach to aligning collinear regions of genomes. *Bioinformatics*, **18**, 1673–1680.

Shabalina S.A, Ogurtsov A.Y., Rogozin I.B., Koonin E.V., Lipman D.J. (2004) Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. *Nucl. Acids Res.*, **32**(5), 1774–1782.