EC_oligos: automated and whole-genome primer design for exons within one or between two genomes

Shaolin Liu\textsuperscript{1}, Nicholas A. Tinker\textsuperscript{2}, Stephen J. Molnar\textsuperscript{2} and Diane E. Mather\textsuperscript{1,*}

\textsuperscript{1}Department of Plant Science, McGill University, Ste-Anne-de-Bellevue, Montreal, QC, Canada and \textsuperscript{2}Agriculture and Agri-Food Canada, Bioproducts and Bioprocesses, Central Experimental Farm, Ottawa, ON, Canada

Received on April 26, 2004; revised on July 2, 2004; accepted on July 6, 2004
Advance Access publication July 22, 2004

ABSTRACT

Summary: EC_oligos designs oligonucleotides (oligos) from exons of annotated genomic sequence information. It can automatically and rapidly select oligos that are conserved between two sets of sequence data, and can pair up oligos for use as PCR primers. It can do this on a whole-genome scale and according to user-defined criteria.

Availability: The source code, executable program and user manual are available at ftp://ftp.ebi.ac.uk/pub/software/dos/EC_oligos/.

Contact: diane.mather@mcgill.ca

Oligonucleotide (oligo) design is a critical step in the application of technologies relying upon PCR or oligo-based microarrays. Early tools for oligo design, written mainly to design primers for PCR and probes for DNA hybridization (e.g. Bridges, 1990; Rozen and Skaletsky, 2000) or large-scale sequencing (e.g. Li \textit{et al.}, 1997), typically take small numbers of DNA sequences as input. Some recently developed tools use whole-genome data to design oligos for microarray experiments (e.g. Rouillard \textit{et al.}, 2002; van Hijum \textit{et al.}, 2003).

EC_oligos has several unique features that are not available in other oligo design programs: It can (1) partition annotated GenBank sequences and design oligos from exons; (2) partition results returned by RepeatMasker to exclude repetitive regions and identify microsatellite motifs flanked by oligo pairs; (3) design oligos that are conserved between two sets of sequence data; and (4) conduct the above functions rapidly and automatically for large, whole-genome datasets.

Our initial motivation in developing EC_oligos was to automatically and efficiently develop large numbers of molecular markers for comparative genomics. We sought markers that would be conserved in sequence between rice (\textit{Oryza sativa} L.), for which complete genome sequence data are available, and barley (\textit{Hordeum vulgare} L.), for which many expressed sequence tag (EST) sequences are available. EC_oligos can also be used for other organisms, either within one or between two genomes. We chose to design oligos from within exons because they are more conserved than introns.

INPUT AND OUTPUT

EC_oligos can use three types of input files: (1) sequences in GenBank default format, (2) sequences in FASTA format and (3) results returned by RepeatMasker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker). Depending on the options chosen, one, two or all three of these inputs may be required. The program has command-line arguments allowing users to choose options and parameters, for which default values are also provided. EC_oligos outputs all results in text files and generates log files to facilitate troubleshooting.

DESIGN AND IMPLEMENTATION

The program was written in standard C++ using Microsoft Visual C++ 6.0. It was compiled and tested on Windows machines, but its source code could be ported to other platforms supporting standard C++. Its object-oriented design allows all components and functions to be easily re-used or extended for other applications.

Figure 1 shows the data-processing flow chart for designing primer pairs that are conserved between two genomes, one represented by genomic sequence (genome A) and the other represented by EST sequences (genome B). For each annotated CDS (coding sequence) in genome A, the program extracts the GenBank accession number and the position of each exon, along with annotation tags. The program extracts qualified oligos from the GenBank records or from RepeatMasker output files by sliding a window of the desired oligo length one nucleotide at a time within each exon. A

\textsuperscript{*}To whom correspondence should be addressed.
**EC_oligos: automated whole-genome primer design**

Genome A

- FASTA format GenBank-records
- RepeatMasker
- Extract CDS annotations
- Extract qualified oligos
- Search index for conserved oligos
- Pair up oligos

Genome B

- FASTA format GenBank-records
- Default format GenBank-records
- Extract sequences
- Index sequences

**Fig. 1.** Flow chart of EC_oligos. The whole figure shows a full-featured run while the black boxes with bold arrows indicate a minimum run, i.e. design of oligos for CDS in one genomic sequence.

To obtain pairs of primers suitable for PCR, oligos can be paired up based on restrictions, including: (1) a range of product lengths; (2) a maximum number of complementary nucleotides at the 3' end (to avoid possible dimer formation); and (3) a maximum difference of \( T_m \). EC_oligos counts and reports the number of introns flanked by a primer pair. If results of RepeatMasker have been used as input, EC_oligos can count and report microsatellite motifs flanked by each primer pair.

**PERFORMANCE**

We tested EC_oligos on sequences of two genomes: (1) the annotated rice genomic sequences downloaded from GenBank Entrez (http://www.ncbi.nlm.nih.gov/Entrez/) using the query 'oryza sativa'[Organism] AND 'cds'[Feature key] and (2) barley tentative consensus sequences (TCs) downloaded from TIGR Barley Gene Indices (http://www.tigr.org/tdb/tgi/hvgi/). The cut-off date was May 15, 2003 for both sets of data. Default values for all the parameters were used. The machine was a Dell Dimension 8250 (3.06 GHz Intel Pentium 4 CPU, 1.0 GB RAM) running Microsoft Windows XP. The program generated 7 710 265 20 bp oligos for 26 277 CDS from 3903 rice sequences in <10 min. It finished running each oligo against 21 050 barley TCs on both strands in ∼40 min. The final pairing up of conserved oligos took <10 min.

**ACKNOWLEDGEMENT**

This work was funded by the Natural Sciences and Engineering Research Council of Canada.

**REFERENCES**


