Applied Computational Tools for Crop Genome Research

David Edwards

Plant Biotechnology Centre, Department of Primary Industries, La Trobe University, Bundoora, 3084, Australia

Summary

A major goal of agricultural biotechnology is the discovery of genes or genetic loci which are associated with characteristics beneficial to crop production. This knowledge of genetic loci may then be applied to improve crop breeding. Agriculturally important genes may also benefit crop production through transgenic technologies.

Recent years have seen an application of high throughput technologies to agricultural biotechnology leading to the production of large amounts of genomic data. The challenge today is the effective structuring of this data to permit researchers to search, filter and importantly, make robust associations within a wide variety of datasets.

At the Plant Biotechnology Centre in Melbourne, Australia, we have developed a series of tools and computational pipelines to assist in the processing and structuring of genomic data to aid its application to agricultural biotechnology research. These tools include a sequence database, ASTRA, for the processing and annotation of expressed sequence tag data. Tools have also been developed for the discovery of simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) molecular markers from large sequence datasets. Application of these tools to *Brassica* research has assisted in the production of genetic and comparative physical maps as well as candidate gene discovery for a range of agronomically important traits.

Sequence management using the ASTRA database system

The ASTRA annotation pipeline is a modular series of PERL scripts which act as wrappers for sequence processing, annotation and database management. Trace files are batch processed using phred and crossmatch (Ewing *et al.*, 1998) to call and quality score each base and screen for vector contamination. Resulting FASTA format sequences are stored within a MySQL database.

Sequences within the database are annotated by comparison to DNA and protein databases GenBank and SwissProt using BLAST (Altschul *et al.*, 1997). The FASTA headers for the ten most significant BLAST matches are parsed to the MySQL database along with HTML format files for each alignment. HTML NCBI links are maintained enabling direct, remote access to the NCBI annotation. Sequences are assembled using TGICL (Pertea *et al.*, 2003). Cluster ID, cluster members and assembled sequence alignments are then stored in the MySQL database. The database is queried through an intuitive web interface managed by a series of modular PERL scripts. The database may be searched using key words, or by sequence identity using BLAST.

The flexibility and modular design of the ASTRA system enables the incorporation and expansion of further data analysis and annotation modules. The *Brassica* ASTRA database currently incorporates modules for Gene Ontology annotation, comparative genome analysis with Arabidopsis as well as SSR discovery and PCR primer design. Further modules to incorporate gene expression data are under development.

Molecular marker discovery from large sequence data sets.

Single Nucleotide Polymorphisms (SNPs) and Simple Sequence Repeats (SSRs) are valuable molecular markers for genetic analysis. They are used routinely in agriculture as

markers in breeding programs and have many uses in human genetics, such as the detection of alleles associated with genetic diseases and the identification of individuals. SSRs and SNPs are invaluable for genome mapping, offering the potential for generating very high density genetic maps and genetic diversity analysis for the understanding of genome evolution. Traditional methods for molecular marker discovery have been lab based. However, the availability of large sequence data sets enables highly efficient computer based marker discovery.

We have developed a computer based method to identify candidate SNPs (Single Nucleotide Polymorphisms) and small indels (insertions/deletions) from expressed sequence tag (EST) data. The program uses TGICL to cluster and align EST sequences (Pertea *et al.*, 2003). Using a redundancy based approach, valid SNPs are distinguished from erroneous sequence by their representation multiple times in an alignment of sequence reads. For each candidate SNP, two measures of confidence are calculated, the redundancy of the polymorphism at a SNP locus and the co segregation of the candidate SNP with other SNPs in the alignment.

We have developed a second molecular marker discovery tool that integrates SPUTNIK, an SSR repeat finder (Abajian, 1994), with Primer3, a PCR primer design program (Rozen and Skaletsky, 2001), into one pipeline, SSR Primer. On submission of multiple FASTA formatted sequences the script screens each sequence for SSRs using SPUTNIK. Results are parsed to Primer3 for locus specific primer design. The script makes use of a web based interface enabling remote use. A FASTA file of 397 673 wheat EST sequences (183 MB) was processed to design PCR primer pairs for 70 705 SSRs (5720 dinucleotide, 46 508 trinucleotide, 10 895 tetranucleotide and 7 582 pentanucleotide). A second FASTA file of 300 870 *Brassica oleracea* genomic sequences (192 MB) was processed to design PCR primer pairs for 46 949 SSRs (18 194 dinucleotide, 14 096 trinucleotide, 6 252 tetranucleotide and 8 407 pentanucleotide). These and further processed datasets representing vertebrate, fungal and plant genomes are available at 'http://hornbill.cspp.latrobe.edu.au/'.

References

Abajian, C (1994) SPUTNIK, http://abajian.net/sputnik/.

Altschul SF, Madden TL, Sch_ffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25, 3389-3402.

Ewing B, Hillier L, Wendl MC, and Green P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research, 8, 175-185.

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J and Quackenbush J. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics, 19 (5), 651-2.

Rozen S. and Skaletsky H.J. (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S. (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp 365-386.
http://www-genome.wi.mit.edu/genome_software/other/primer3.html