EST Bioinformatics

D. Marshall, L. Cardle & P. Shaw

A major change to the information landscape in which we operate for most of our key crop and pathogen species has occurred over the last 2-3 years as the result of a dramatic change in the volume of available sequence information. This is true both for organisms themselves and for the key model organisms which form a major comparative genome resource.

For this account we will focus on barley and potatoes. The major contribution has come from a number of large scale EST (Expressed Sequence Tag) sequencing programmes, though a small amount of BAC-scale genomic sequence is also becoming available in both species. This has generated some 370,000 barley and 132,000 potato ESTs (See dbEST current EST statistics at: http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). These resources are leading to a dramatic change in working patterns in molecular biology and to an increasing role for bioinformatics within the institute's research programmes.

EST sequencing programmes rely on a philosophy based on a relatively 'quick and dirty' approach which involves single pass sequencing of ESTs from a range of cDNA libraries. This is usually 5' sequence though individual clones may also be sequenced in the 3' direction. Such single pass sequences then form the template for a range of subsequent molecular biology and bioinformatics activities. The first stage after the removal of low quality sequence and vector contamination is to obtain an initial indication of function through homology searches against a range of public databases using the Blast suite of programmes. The second step is to carry out sequence assembly which uses software tools such as CAP3 to assembly the EST sequence fragments into contigs and then subsequently derive consensus sequences. The high redundancy level of EST sequences from relatively highly expressed genes and the occurrence of cDNA clones truncated at the 5' end together lead to the generation of consensus sequences which are both of surprisingly high quality and much longer than the individual single pass EST sequences.

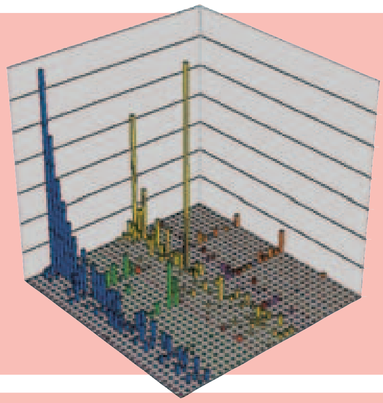

Figure 1 Data from 31 barley EST libraries chosen to show the relative abundance of the most highly abundant expressed endosperm sequences across a range of other tissues

Within the Computational Biology Programme at SCRI we have built up a considerable expertise in the process of EST sequence assembly and the use of the subsequent assemblies and consensus sequences as the substrate for a whole series of downstream activities ranging from the identification of Simple Sequence Repeat (SSR) and Single Nucleotide Polymorphism (SNP) genetic markers, expression analysis (based on microarrays, SAGE analysis and electronic Northerns - see Figure1) and for the identification of barley and potato orthologues of genes that have been characterised on other species especially in model organisms such as Arabidopsis.

The task we face in helping develop the EST sequence resource from potato and barley as a major tool in our research is confounded by the continuing generation of new sequence information as well as the rate of change in sequence annotation, especially due to the progress of functional genomics in Arabidopsis and rice as well as more directly in barley and potatoes. To meet this challenge we have adopted two main approaches. The first of these is to develop a relatively lightweight scalable database infrastructure based on a combination of public domain software tools and applications tools, including MySQL, PostgreSQL, Perl and Apache. This has now been applied to a series of database applications both for day to day support for individual research programmes and as a central resource for internal and external interaction with SCRI data resources. Examples may be seen though the SCRI Bioinformatics web server at http://bioinf.scri.sari.ac.uk/. The second major activity has been to develop strong relationships with a number key of information providers. These include the Graingenes, Gramene and BarleyBase cereal databases in North America. A key aspect of our involvement has been to highlight the importance of data quality and data validation in comparatively map and sequence analysis and the promotion of appropriate informatics standards.