Hordeum Relator: a visualization tool for relating unigene datasets

Linda Cardle and David Marshall
SCRI, Invergowrie, Dundee, UK, DD2 5DA.

http://bioinf.scri.ac.uk/hordeum-relator

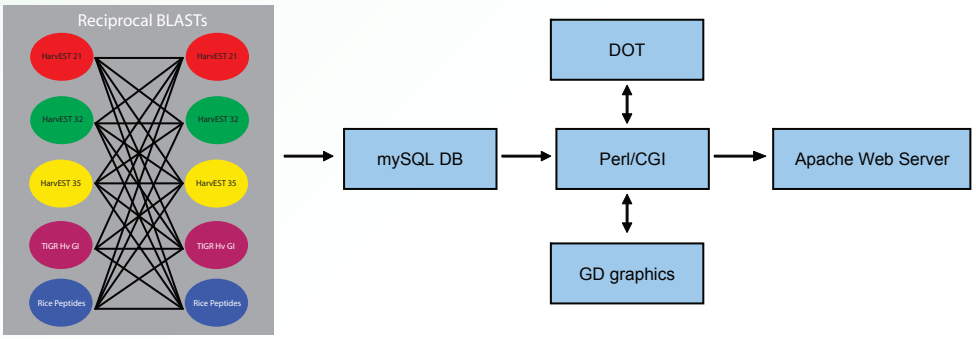
Over the past few years, large scale EST sequencing has provided a wealth of data for bioinformaticists to use sequence assembly techniques to construct unigene sets for species for which there is no full genome sequence available. Such unigene sets are viewed as the best representation of the genes present in an organism, and are used to design genetic markers and probes for mapping, genotyping and microarray chips.

In the case of barley sequence data, since batches of ESTs and other sequence data have been released at different points in time, there are several unigene sets available, all of which have been used for sequence-based genomics tools in different projects (see table below).

Assembly	HarvEST 21	HarvEST 32	HarvEST 35	TIGR GI
Date of construction	27/11/2002	03/03/2003	09/02/2007	15/09/2004
Number of included seq	615134	619084	767817	372427
Number of unigenes	53241	41923	50938	50453
Assembly parameters	stringent	relaxed	relaxed	stringent
Source	UC Riverside	UC Riverside	UC Riverside	TIGR

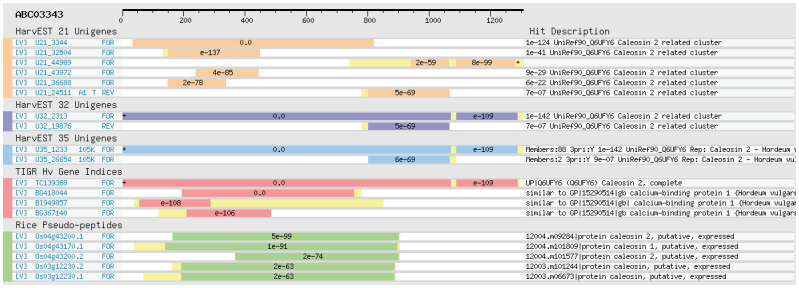
Nevertheless, if we can relate all these assemblies to one another, we will be in a position to cross-relate the experimental results from these studies. To this end, we have developed a database and web utility that visualizes the relationships between sequences from different assembly datasets in barley. This tool also compares the assemblies with an appropriate species for which there is a whole-genome sequence available; for barley the TIGR Rice Pseudo-peptides (http://www.tigr.org/tdb/e2k1/osa1) and the new Brachypodium sequence assembly (http://www.brachypodium.org).

Methods



The barley unigene datasets were sourced from the HarvEST database (http://harvest.uc.edu) and the TIGR Gene Indices (http://compbio.dfci.harvard.edu/tgi). Each dataset was BLAST searched against the others and against itself. A mySQL database was designed to hold the the BLAST results, and the data was loaded using Perl scripts. To create the schematic diagrams for the BLAST results we used a combination of Perl CGI scripts and the GD graphics library. For the graphical networks we used the C–program DOT from the Graphviz package (http://graphviz.org).

Interface and Visualizations



BLAST results are shown as a schematic diagram showing the extent of each hit against the length of the selected unigene.

Users can look up a unigene using :

a DNA sequence

a unigene ID

a set of keywords to search annotation

The relationships between the BLAST hits to the unigene can be viewed as a graphical network. Each dataset is represented by a different colour, and the best hit from each dataset is highlighted in white. The relationships between the other hits within the cut-off threshold are shown with e-values.

The display can be tailored by the user by including or excluding datasets, and by using the e-value cut-off threshold to limit the number of blast hits returned.