Genome bioinformatics at SCRI: engineering the datastream

D.F. Marshall & L. Cardle

One of the main tasks of the new Bioinformatics & Information Technology Research Unit (BITR) at SCRI is to provide the computational and data-handling infrastructure that is required to underpin plant genomics research. This essentially breaks down into three main areas.

- the development of efficient procedures for handling and validating data, in particular the large data volumes that modern genomics research can generate.
- the development of novel software tools for data analysis and visualisation.
- the development of an appropriate framework for storage and retrieval of data.

The data resources that are being generated at SCRI must also be continually reviewed, not only in a local context, but also in a national and international context to maximise the value of our data.

Modern PCR-based techniques for molecular marker genotyping and recent developments in sequencing technology have dramatically increased both the rate at which genotype and sequence data can be generated and the volumes of data that must be efficiently handled. The critical rate-limiting step in genotype and sequencing projects is now no longer the data generation phase but rather the rate at which the data can be captured, validated, fully analysed and finally placed in context. This means that we now need to develop new data handling and storage procedures.

One of the major factors we have to deal with is the fact that, though there are a wide range of software tools, many of them public domain, for molecular biology, often they are only available on certain hardware platforms or operating systems. For example, much of the software that is available for directly handling automated sequencing or genotyping is specific to the Apple Macintosh, whereas the majority of major analysis tools are available on Unix systems. Finally, the most important factor is that, no matter what software tools and hardware platforms we use along the way, the final results must be available for the research scientist to browse from his or her desktop computer.

Ideally, we would like to write all the software components used from scratch to provide a uniform software platform. In many cases this may well be our ultimate goal. However, we need to take a more pragmatic approach in the shorter term. An ideal computer language to help implement this approach is Perl, which was originally developed on Unix systems, but is now available on Apple Macintoshes and PCs running Windows 95 or Windows NT. Larry Wall originally conceived Perl as a ‘glue’ language. It has superb text processing capabilities and can operate as a command language running external commands and programs. These features make it ideal for our purposes. We can automate our data handling by the use of Perl programs or ‘scripts’ to run one or more software tools and process the output from each for input into the next step in the process. It is also possible to write new modules entirely in Perl, which, despite the fact that it is an interpreted language, is remarkably fast and efficient for many sequence and genotype data handling tasks.

The need for such new approaches to data handling can best be illustrated by a couple of simple examples.

SSR development and genotyping

In the last few years at SCRI, we have developed over 300 barley simple sequence repeat (SSR) genetic markers and are continually adding to this total. The first stage in the process is the development of genomic libraries enriched at the pre-cloning stage for an appropriate microsatellite repeat, usually either $(CA)_n$ or $(GA)_n$. This enrichment process maximises the rate at which SSR-containing clones can be extracted from the library. The next stage is to sequence SSR-containing clones in order to design suitable PCR primers that amplify a particular SSR sequence. It is at this stage that bioinformatics can begin to make an impact. If we are dealing with a single sequence, or even 10 sequences, it is a relatively straight forward matter to take each sequence in turn, identify the SSR sequence and, if possible, design suitable PCR primers for subsequent testing. However, if rather than 10 we have several hundred sequences to analyse, the situation is more complex. If we break the problem down into its component stages, we can identify steps at which we can improve the efficiency of the entire process by either implementing existing software tools or developing our own software.

Sequencing Currently, we sequence putative SSR-containing clones using an ABI 377 with supporting software running on an Apple Macintosh. The major task at the initial stage is to check that the sequence is of sufficient quality and to remove any sequence that has come from the cloning vector. We then need to transfer the sequence data, via our local area network, to a Sun Unix Workstation for further analysis.

SSR location The next step is to confirm that the clones containing an SSR sequence have sufficient flanking sequence on either side to design suitable PCR primers, to precisely locate the SSR repeat and then mask it to avoid spurious database matches.

Database searches We then need to take each SSR-containing sequence and compare it against our local database of existing sequences to ensure that it is unique. It is also sensible at this stage to compare the sequence against all publicly available sequences by BLAST searching against GenBank or other appropriate databases. This gives us the opportunity to identify any possible sequence matches and identify corresponding predicted function(s) associated with sequences flanking the SSR.

Primer design If the SSR-containing sequence is unique, we need to design appropriate PCR primers using the program Primer. This enables us to design primers with appropriate position and amplification conditions. Indeed, it is possible at this stage to automatically order the synthesis of the best primers by E-mail.

We have designed a series of Perl scripts that enable us to automate most of the analytical steps in this process. These scripts run each sequence through a series of analytical tools, pre-processing the input or post-processing the output from each analysis. The remote database searching is also automated, with the results of each BLAST analysis being returned as formatted HTML files for subsequent analysis. We have also adapted this approach for the analysis of sequences from a number of Expressed Sequence Tag (EST) programs currently underway at SCRI. In this case, the primary aim of the analysis is to assign a function to each unknown cDNA sequence through database homology to characterised sequences. Since the number of characterised sequences in the major international databases is rapidly expanding, we need to repeat this process at regular intervals to find matches for still unknown sequences or to improve the quality of the matches and thereby our confidence in the assigned function. We are now processing several thousand EST sequences in this way from a range of crops, their pests and pathogens. We are also examining ways to handle the large body of information that is returned from these searches so that we can efficiently mine it for relevant information and build suitably structured indexes. We hope to be able to provide automatic notification of significant changes to the BLAST scores of individual sequences that result from each new pass of database searching.

Quick and dirty (QAD) mapping
As there are now good quality genetic maps in of all of the major crop species, we are frequently faced, not with the problem of how to generate a new genetic map from scratch, but rather how to efficiently add more loci onto existing saturated maps. Conventional linkage analyses, using *standard* programs such as JoinMap or Mapmaker, are relatively slow and inefficient for such purposes.

We have developed an alternative approach, Quick and Dirty (QAD) Mapping, which is based on a Perl module that uses simple pattern-matching to map each new SSR or other locus by comparison with the genotypes of existing mapped markers. For a saturated map the best matching locus gives us a good location for our unknown locus. This process is illustrated in Figure 1 which shows a pattern-matching scan of an unknown locus across the entire barley genome with markers in linkage group order. A clear peak in the ‘Quality of fit’ shows the location of the unknown locus. We are currently generalising this QAD mapping approach to a range of barley, potato, *Arabidopsis* and *Brassica* populations. The QAD tool can also be used to place newly mapped markers in map order

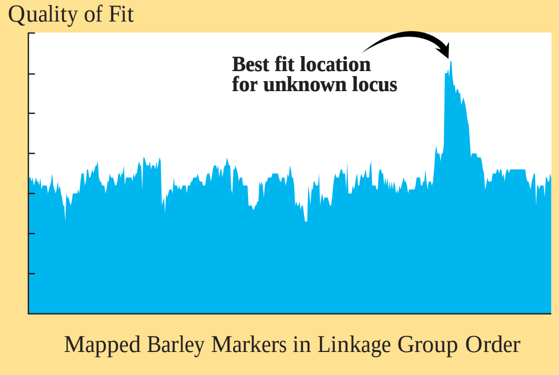


Figure 1 A QAD mapping pattern matching scan of an unknown locus against the entire barley genome. The most likely map position of the unknown locus is at the 'quality of fit' peak.

within the existing raw data files. QAD mapping is particularly suitable for adaptation as a WWW-based mapping service linked to HTML forms controlled through a web browser. We are developing this aspect for use on our own intranet, but, in future, we hope to be able to offer this via our extranet Web-server as a service to rapidly map new loci in our target crops using public domain mapping populations and datasets.

Data validation

One of the major concerns with respect to molecular linkage map data, is that given the extremely large number of individual data items, there is a significant chance of introducing errors. Such errors can arise at several steps in the data generation, data entry and data analysis process and may significantly distort or inflate the resulting genetic maps. A major challenge, therefore, is first of all to audit these processes to minimise the risk of data errors, and, secondly, to try and develop data validation processes which can serve to indicate data items that are, to some extent, inconsistent with the main volume of data. In the former case, we can take some simple steps to ensure that our procedures for naming lines and DNA samples are robust. We also must make sure that the location of samples at every stage of the PCR, gel and auto-radiography processes is robust and auditable. Even once we have established such procedures, we have to realise that there is an appreciable probability of data errors occurring. In the case of map data, it is possible to identify a number of factors which are at least indicative of poor quality data or data inconsistencies. Segregation distortion often occurs in mapping populations, especially those generated from wide crosses or involving doubled haploid populations in diploids. However, individual loci with high levels of segregation distortion, especially those that show significantly different patterns of distortion from adjacent loci, should be treated with suspicion. A second element, often associated with data errors, is the occurrence of what are known *as single marker double recombinants*. A mistaken genotype at a single locus can often generate two spurious recombination events in a very small genetic interval. However, since chiasma interference prevents two or more cross-over events occurring close together on the same chromosome, a pair of such recombination events is likely to occur with an extremely low probability in saturated genetic maps with high marker density. Therefore, we can treat such a data item with care and closely re-examine the original raw data. We are currently evaluating a range of algorithms to objectively gauge the quality of marker data sets and identify loci, or individuals, which are in tension with the bulk of the data, for subsequent reappraisal.

Graphical genotypes and data visualisation

Single-marker double-recombinants are one of a series of problems of handing genomic data where visualisation of the data, based on a *Graphical Genotype* of either a single chromosome or the entire chromosome complement of an individual, can play a considerable rôle in evaluating what is a complex array of data. We are developing a general Graphical Genotyping tool in the platform-independent language Java. This will enable us to use a visualisation approach to investigate not only the presence of single-marker double-recom-

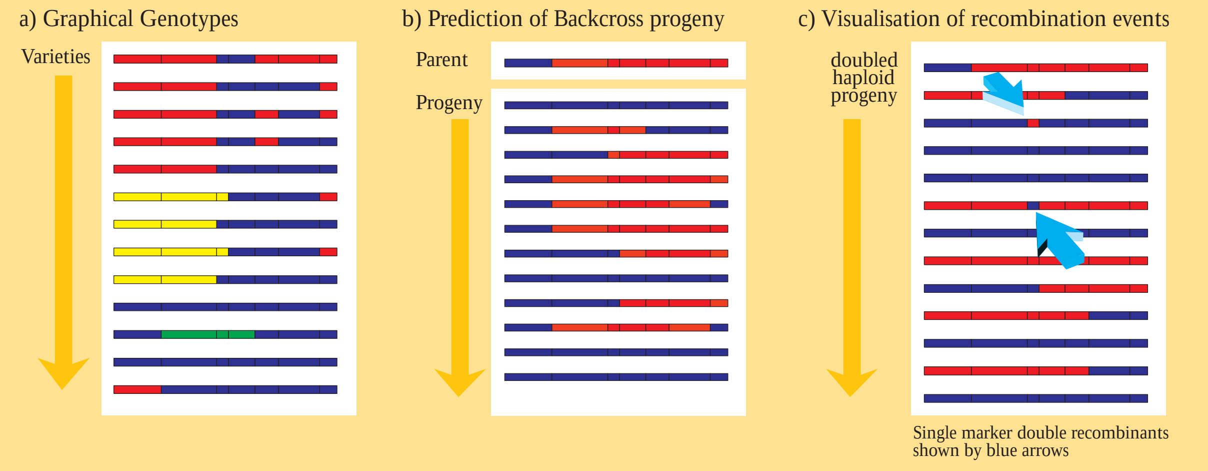
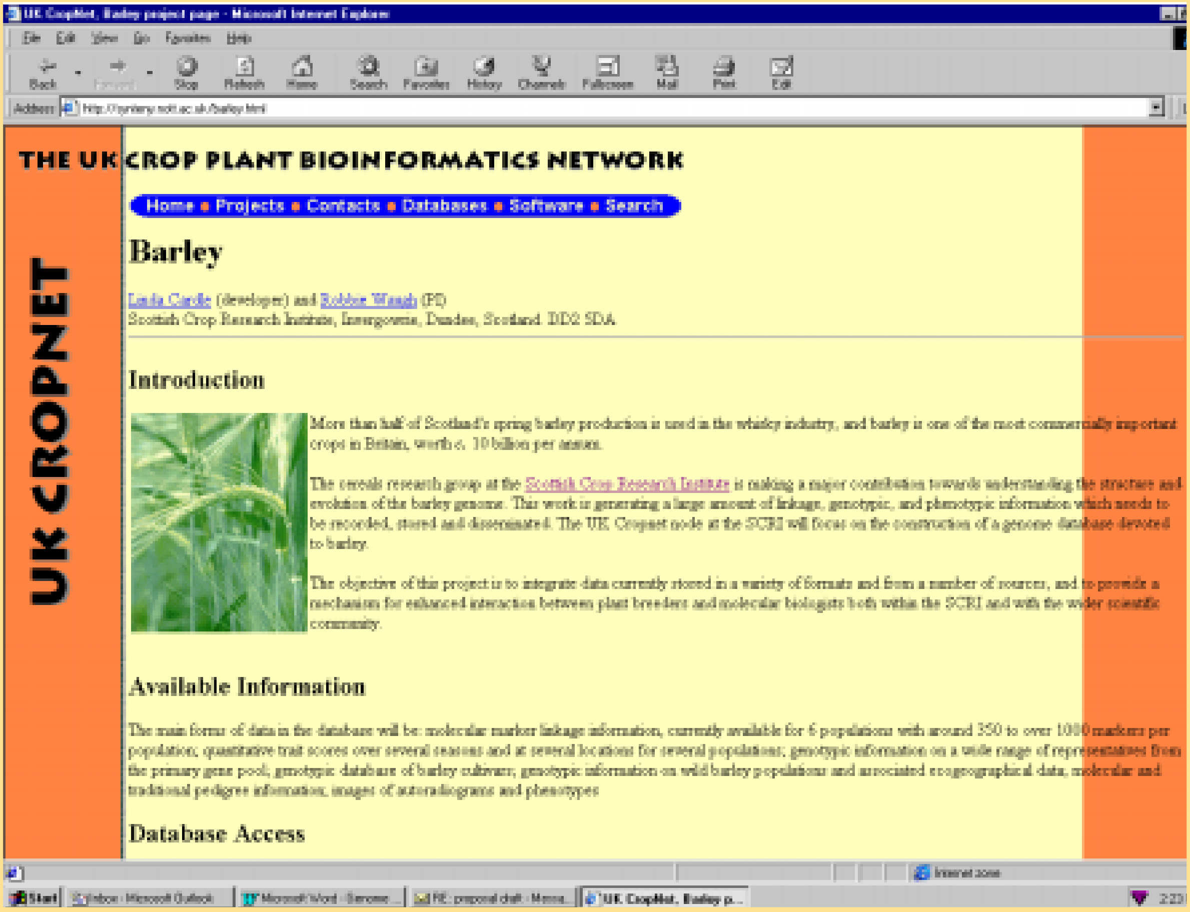

a) Graphical Genotypes
Varieties
b) Prediction of Backcross progeny
Parent
Progeny
c) Visualisation of recombination events
doubled haploid progeny
Single marker double recombinants shown by blue arrows

Figure 2 Three applications of the Graphical Genotypes interface in a) diversity analysis; b) simulation of the outcome of backcross conversion experiments; and c) the visualisation of errors in mapping data.

binants but also to evaluate diversity data (generated, for example, by the SSR analysis of a wide range of barley germplasm). We will also be able to monitor and predict the size of introgressed chromosomal tracts in directed backcrossing programmes to introduce disease resistance genes from exotic germplasm into advanced breeding lines. These graphical genotyping scenarios are illustrated in Figure 2. The combination of such a visual approach with the estimation of appropriate genetic parameters and/or the simulation of the outcomes of projected experiments, will provide a valuable 'Visual Genetics Workbench' to support a broad range of marker-enhanced breeding activities at SCRI.

Plant genomic databases

SCRI forms one node of the UK CropNet consortium of Plant Genome databases. Other members of UK CropNet include the John Innes Centre, Nottingham University and IGER, Aberystwyth. With funding from the BBSRC Plant and Animal Genome Analysis Initiative, CropNet has established a series of genome databases for plant species relevant to UK agriculture. At SCRI, we have responsibility for establishing and curating databases for barley and potatoes. The first of these, BarleyDB, is now publicly available through the UKCropNet web site at http://synteny.nott.ac.uk/ (see Fig. 3). At the end of 1998 a much more extensive version of this database will be released. Development work is now also underway on the potato database, SPUDB, and we plan to have an initial version of this database available on the CropNet Webserver in early 1999. These database projects serve not only as a focus for handling our genome data locally but, increasingly, will provide a national resource for the storage, maintenance of and access to data from publicly funded work in these species. We are currently in discussion with groups in Europe and North America to build a co-ordinated international framework for plant genomic databases. Genomic databases are becoming an increasingly significant resource. It is crucially important that funding agencies provide the appropriate levels and continuity of funding for stable database development and data curation.


UK CropNet, Barley project page - Microsoft Internet Explorer
File
Edit
View
Go
Favorites
Help
Back
Forward
Stop
Refresh
Home
Search
Favorites
History
Channels
Fullscreen
Mail
Print
Edit
Address
http://synteny.nott.ac.uk/barley.html
THE UK CROP PLANT BIOINFORMATICS NETWORK
UK CROPNET
Home • Projects • Contacts • Databases • Software • Search
Internet zone
Start
Inbox - Microsoft Outlook
Microsoft Word - Genome ...
RE: proposal draft - Messa...
UK CropNet, Barley p...
2:23

Barley

Linda Cardle (developer) and Robbie Waugh (PI)
Scottish Crop Research Institute, Invergowrie, Dundee, Scotland. DD2 5DA

Introduction

More than half of Scotland's spring barley production is used in the whisky industry, and barley is one of the most commercially important crops in Britain, worth c. 10 billion per annum.

The cereals research group at the Scottish Crop Research Institute is making a major contribution towards understanding the structure and evolution of the barley genome. This work is generating a large amount of linkage, genotypic, and phenotypic information which needs to be recorded, stored and disseminated. The UK Cropnet node at the SCRI will focus on the construction of a genome database devoted to barley.

The objective of this project is to integrate data currently stored in a variety of formats and from a number of sources, and to provide a mechanism for enhanced interaction between plant breeders and molecular biologists both within the SCRI and with the wider scientific community.

Available Information

The main forms of data in the database will be: molecular marker linkage information, currently available for 6 populations with around 350 to over 1000 markers per population; quantitative trait scores over several seasons and at several locations for several populations; genotypic information on a wide range of representatives from the primary gene pool; genotypic database of barley cultivars; genotypic information on wild barley populations and associated ecogeographical data; molecular and traditional pedigree information; images of autoradiograms and phenotypes

Database Access

Figure 3 The SCRI Barley page on the UK CropNet Webserver.

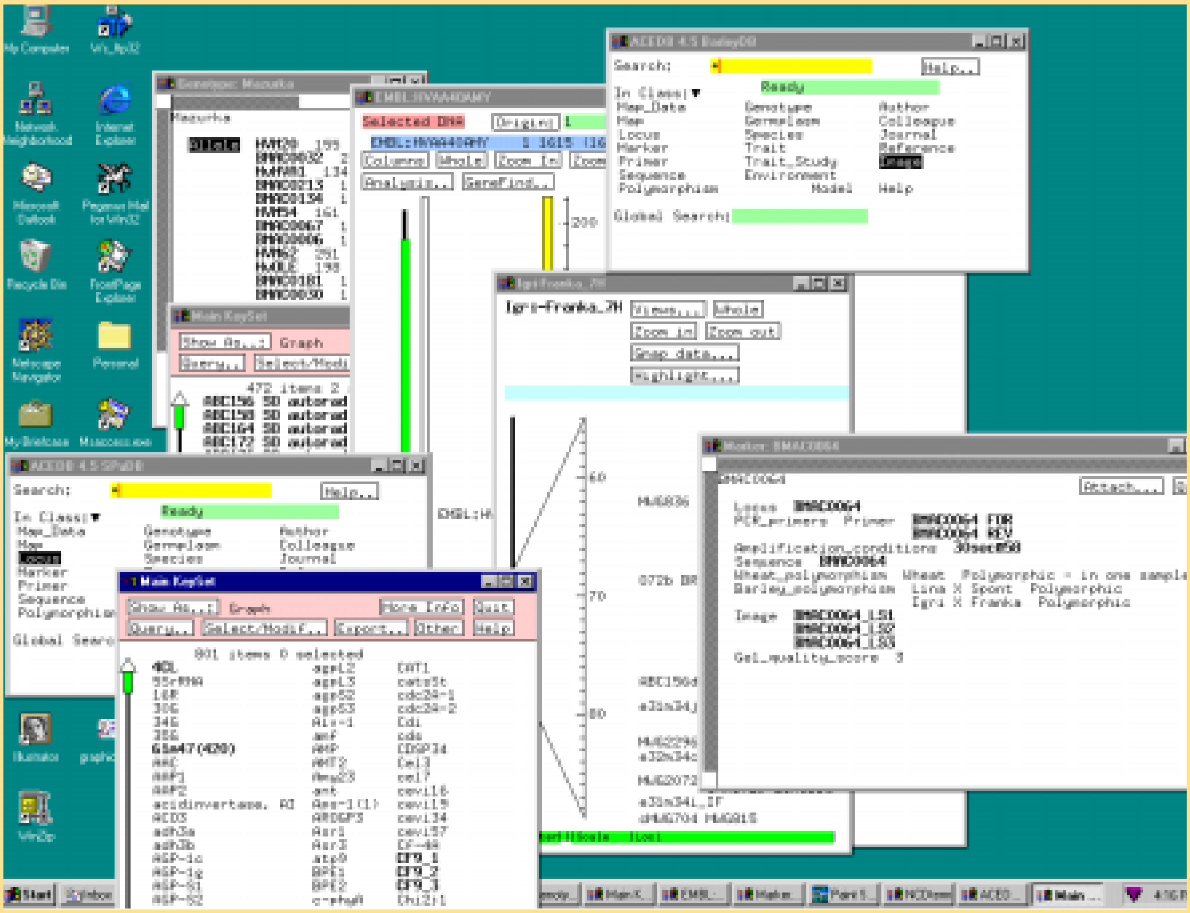


Figure 4 A Windows NT 4 Desktop showing a range of X-Windows open to both BarleyDB and SPUDB running on a SUN Unix Workstation.

A common theme which underpins most of our current software tool development is the need to provide a suitable common graphical-user interface (GUI) for laboratory and field scientists. The development of standard WorldWideWeb (WWW)-browser client software (e.g Netscape Navigator or Microsoft's Internet Explorer) for a range of hardware and operating system platforms provides us with a common environment with which most of our local users are familiar. By designing our software tools so that they can be readily accessed though WWW pages on our local intranet, we can fully exploit this familiarity. This can be achieved by a combination of HTML forms-based web pages with both client-side analytical and visualisation tools written as Java applets and server-side CGI programs on our Intranet web-server. Again, we can exploit the 'glue' language capabilities of Perl to link together otherwise disparate software components.

Overall, our aim is to provide scientists at SCRI with efficient computational tools to underpin their laboratory and field science. However, our design and implementation philosophy is a pragmatic one which gives a priority to ensuring that we can provide a series of integrated but unsophisticated tools which can then be developed in response to both user requirements and software developments. In the rapidly changing world of plant genomics, the development cycle of software tools needs to be fully integrated into that of the molecular biology.