EAAP Uppsala 2005 Session 32

Bioinformatics tools in analyzing molecular genetics data for breeding and genetics

Luc Janss, Wageningen-UR Animal Sciences Group
(new address ETH Zurich, Luc.Janss@inw.agrl.ethz.ch)

QTL mapping, gene identification

As a starting point, it is taken here that it will generally be interesting to identify the gene and causative mutation underlying a QTL. This information is not purely of fundamental interest and should also interest practical breeders who wish to apply QTL mapping: selection on QTL's increases the pace of selection, therefore involving higher risks of unwanted negative responses or surprises of evolving gene-gene and gene-environment interactions. Fundamental knowledge from identifying the underlying gene will help to understand the underlying biology and therefore reduce such risks of adverse selection. Also, generalizability across species is expected to increase, because exact gene-effects may be less replicable, whereas effects of pathways are.
In identifying genes underlying QTL there is still a gap to be filled after applying the latest fine mapping tools. With linkage analyses, QTL regions can be identified of roughly half a chromosome (about 1000 genes), which could be reduced further to regions of some cM by fine mapping (about 100 genes). Thereafter, however, tools based on associations with markers will generally cease to bring further clarification. Hence, extra tools and extra information are needed, where we could consider omics data (transcriptomics, proteomics), and bioinformatics tools and data.

Statistical complexity

QTL results are often little comparable between experiments. There are a number of reasons for this, some genetical (e.g. the mutation is not present in another experiment), some related to a too narrow focus (e.g. focus on pathways could give already a more consistent picture), and some statistical. In the latter category, current QTL mapping experiments are likely suffering from large false negative error rates (genes being there are not detected), due to the application of very stringent test thresholds. The statistical problems will be further increased by the sketched trend to need to combine more sources of data and to combine different tools. Often this is done in "pipeline" constructions where errors made in one step are not taken into account in the following step, therefore accumulating errors which may lead to artifacts. In order to handle multiple sources of information and multiple tools in a sensible way, leading to the proper detection of QTL's, will require adoption of new statistical tools and paradigms.

False Discovery Rate

A first "paradigm" shift needed in statistics will be wider use of False Discovery Rates (FDR). Statistical testing is a problem of balancing two types of errors: false positive and false negative errors. The common approach is to steer on false positives, accepting virtually no chance to make any false positive error. However, especially when many tests are being made, this requires the use of very stringent thresholds, and therefore a very high false negative error rate (many real genes are being missed). A paradigm shift here is the acceptance of much higher false positive rates, realizing that this comes with also detecting more true positive results. An example from a microarray analysis for instance shows the following results in which 20 more real genes are found by relaxing the thresholds (estimating the number of false positives using the SAM technique):

Threshold	Postives (of which false)	Extra real genes detected
Stringent	25 (1)	
Relaxed	46 (6)	+20

For QTL mapping, application of FDR could also be interesting, especially when combined with a multi-QTL model in which a QTL is modeled in every bracket (Meuwissen, GSE, 2005). In this type of approach, a prior could be applied to assign high probability to have many genes of small effect, with only a few genes of large effect, for instance in the form of an exponential distribution. Hayes and Goddard (2001) showed that such an experimental distribution fits actual data on QTL effects quite well.

Accumulating errors in pipelines

Common approaches for analysis of microarray data imply the sequential application of a larger number of statistical estimation and correction procedures. For microarray analysis this involves steps like: image analysis, intensity estimation from the image, background correction, correction for dye bias, correction for heterogeneity of variance over intensity range, corrections of level and heterogeneity of variance between slides, before to proceed to a general statistical analysis of gene effects based on multiple slides. Also in QTL mapping a sequence of steps is performed, notably the making of marker maps, allele scoring and data pre-adjustment is done in separate steps. It is becoming evident now that these sequential procedures can introduce artifacts, for instance markers wrongly placed on linkage maps can produce erroneous QTL results. Also bioinformatics data that may be used to ultimately annotate genes have many sources of potential error: it is the largest databases that are the least curated (so in general poor data is overwhelming the good data), and making comparative links between species (on which animal breeding will largely rely) can err because of inaccuracies in comparative maps, and mistakes caused by gene duplications. Also, the matching of traits between species can be risky, and in order to do that properly, good trait ontologies should be developed.

Two solutions to the problem of accumulating errors in pipelines are to perform more integrative analyses (which may be feasible in some areas, but not in all), and to assign levels of confidence to information, which could be used in further steps. Hereto, meta-analysis tools or Bayesian modeling could be used to sequentially update knowledge and uncertainty.

More integrative solutions: supervised clustering to use pathway information

Common “unsupervised” clustering techniques (e.g. k-means, PCA) are generally not fully rewarding as the pure statistical association brings little biological soundness to the clusters being made. More useful clusterings can be obtained using supervised clustering techniques. This can for instance be applied to combine gene expression data and various bioinformatics data sources into gene identification and QTL mapping tools. In this approach, several sources of data (e.g. from pathway databases, sequences and literature) are used as “priors” in clustering gene expressions, so adding information and cause-effect relationships which would otherwise not be available from pure statistical association.

Conclusions. In order to fill the last steps to ultimately identify genes underlying QTL, a large array of tools and (genomics) data will have to be combined and streamlined. Some statistical problems (and directions for solutions) are indicated to help in this. In general more relaxed (in using FDR) and better (in pipelines) approaches to handle errors should become in use. The ultimate identification of genes, but also the bioinformatics information on e.g. pathways, will ultimately also help breeders to better understand gene effects, to make QTL results more generalizable across populations, and so to devise more robust selection programs based on molecular data.