Sharing paths of exploration to support collaborative reasoning in genomic data analysis

David Hoyle[a] Peter Crowther[b] Mark Delderfield[a] Lee Kitching[a] Gareth Smith[a] Iain Buchan[a]

[a] *North West Institute for BioHealth Informatics, University of Manchester, Manchester, UK*
[b] *Melandra Ltd, Manchester, UK.*

Abstract and Objective

A significant proportion of data analysis of modern large-scale genomic data sets takes place in an investigative, exploratory mode of working that helps to refine hypotheses. This typically proceeds by reviewing annotation meta-data associated with statistical results, with a view to using the annotation data to suggest refinements to the statistical models used. The researcher's exploratory path through the annotations also represents part of the overall evidence chain, and as such should be captured. We identify three distinct main aspects to the capture problem, i) automated capture of the images and meta-data of the viewed annotations, ii) replay of the captured images and meta-data, iii) a record of insights with discoveries shared with collaborators. We report a prototype capture tool, built around a Workbench for the analysis of large-scale genetic data sets that uses Rich Internet Application technologies to visually replay the captured exploratory paths taken.

Keywords:

Genome-wide association studies, Case-control studies.

Building an Exploratory Path Capture Tool

An exploratory path will consist of reviewing results from many statistical analyses, along with sources of data or information that might assist in the interpretation of those results. The sources may be local or online databases that provide initial information (annotations) about the raw data; typically online databases that provide further annotations based on the earlier retrieved information; and follow-on navigation within these systems. For example, when reviewing the results of a statistical computation within the Shared Genomics Workbench, initial sources might be Google searches or NCBI database lookups that map a particular Single Nucleotide Polymorphism (SNP) to gene information; further information might come from mapping the gene to biomedical publication and pathway information; and all of these annotations might be presented as structured data or as Web pages, with active areas that allow the user to review more detailed content or navigate around. We believe the path taken by the user through the annotations is commonly not captured.

Tools could address this issue by serving several functions: i) provide automated persistence of the annotations visited during a piece of research, rather than relying upon investigator recollections, ii) provide an electronic method of reviewing annotation sessions and enable a user to recall an original annotation in its entirety, iii) enable a user to record their own comments next to annotations which can they can share with their collaborators. Humans have a remarkable memory for pictures, and can scan through many images quickly for one that "looks about right". We use this premise to direct User Interface design in our methods for reviewing annotations.

Implementation

We have implemented a prototype exploratory path capture tool using our existing Workbench, a Windows Forms application based on Microsoft .Net 3.5. We have also developed a Rich Internet Application component to support the replaying and reviewing of these paths. Replay of images is as a CoverFlow [see http://intersoftpt.com/WebCoverFlow/]. The following features have been implemented and are being reviewed.

- Automated capture of annotations.
- User directed capture of importance and comments.
- Replay of time ordered annotation images with description through a CoverFlow. Users can click on an image during the review process and be taken to a 'live' copy.
- Sharing of annotation sessions with collaborators.

Conclusion

The tools we have built address the problem of a researcher losing the path they took when exploring the biological annotations associated with their statistical analysis results from large-scale genetic data sets. Through the use of a CoverFlow interface we believe we have developed a novel way of replaying annotations to a researcher in a structured way with the ability to mark importance and add comments. This potentially enables an easier exchange of ideas between collaborating researchers, but also makes explicit the translation from the model-based statistical analysis results to the clinician's experience of a particular patho-physiological condition.