Poster M-12

An open-source software platform for the management and analysis of genotyping data

Authors:
Corne Schriek (*Bioinformatics and Computational Biology Unit, University of Pretoria*)

Short Abstract: An open-source web-based genotyping platform for general purpose management, analysis and visualisation of genotyping data such as SNPs, AFLPs and microsatellites. This software ties scattered genotyping technologies and standards together with cutting-edge programming technologies to form one coherent extensible module that slots into a larger functional genomic information management system.

Long Abstract:

Introduction

Genotyping refers to testing that reveals the specific alleles inherited by an individual. The collection of these inherited alleles genotyped at different loci is called the genetic fingerprint of an organism.

There are three main areas in which genotyping can be useful: it can be used to determine the lineage/heredity of an individual, to prove the identity of an individual, and to estimate the probability that an individual may express some phenotypic trait, given its genotype.

The first main area of application for genotyping is that of determining the ancestry of an individual. It is based on the premise that diploid organisms receive DNA from both parents. Specifically, for every locus of chromosomal DNA (with two alleles) in the diploid organism, only two of the many possible alleles for that locus can be inherited – one allele is inherited from the male parent, and the corresponding allele on the homologous chromosome is inherited from the female parent. Thus, one can compare the alleles of a child with those of its possible parents and generate a probability of relatedness. This area of genotyping has applications in paternity testing, and population and evolutionary genetics.

The second main area of application for genotyping is that of proving the identity of an individual. Generally speaking, each organism can be uniquely identified by its own specific collection of alleles or DNA fingerprint. The most common use for this kind of genotyping is in crime forensics, where tissue samples found at a crime scene can be matched to a suspect's DNA. This is also applicable in the fingerprinting of plant breeding germplasm for variety protection and for the maintenance of clonal identity.

Lastly, estimating the probability that an individual may express some phenotypic trait by looking at its DNA has important applications in medicine and agriculture. This field is generally called association genetics. By looking at differences in the DNA of individuals in a species and associating these differences with certain phenotypes, e.g. susceptibility to a certain disease, one can determine which alleles contribute toward the expression of that phenotypic trait. Thus, the knowledge gained by genotyping can be very useful in finding the origins of illnesses and how to diagnose, prevent or treat them. The knowledge may also

help to advance agriculture by enabling the breeding of organisms with superior genotypes for certain traits such as natural immunity to diseases, or larger and better fruits.

DNA-based marker genotyping clearly has many commercial, academic, social, and agricultural applications. Although many sophisticated proprietary genotyping software packages are available, as well as a some scattered open-source packages that focus on solving specific problems or only run on certain platforms, no general-purpose platform independent open-source genotyping software solutions exist, especially for the integrative analysis of different kinds of genotyping data.

Description

The Bioinformatics and Computational Biology Unit at the University of Pretoria has undertaken a project aimed at creating an open-source Functional Genomic Information Management System (FunGIMS). This open-source system is being designed to enable the archiving and annotation of different functional-genomics experimental data types, and to enable data integration. Included in the system will be different specialised modules, each focusing on a certain data type, but integrating into the larger system to enable powerful integrative data analysis. It is hoped that the integrative analysis of the data will allow for the formulation of systems biology-level research questions.

The module described in this poster forms one of a number of modules comprising the Open-Source Functional Genomic Information Management System. This specific project focuses on the development of a module for the archiving, management and analysis of genotyping data. An integrated interface is being developed to construct a database of marker alleles, allele sizes and allele fingerprints. Calculation and recalculation of allele frequencies for fingerprinting projects will be enabled. The aim isn't to duplicate functionality of existing packages, but interface layers are created to export data to a series of commonly-used mapping, phylogenetics and fingerprinting analysis packages. The project caters for data from AFLP, Microsatellite, SNP and other projects, and will provide facilities for the management of group-based projects, storage of experimental methods, annotation of results, and linking to other experimental data types in the Functional Genomic Information Management System with time/user stamp validation of all records in the system. The project further focuses on the development of an allele fingerprint matching tool for matching unknown subjects to known individuals in a database, as well as paternity, maternity, and sibling matching.

This genotyping module will be superior in certain respects to existing software because it will cater for the specific needs of a broad base of researchers. Research teams will be provided with a central workbench that contains all their project data and that may be accessed from any web browser. Additionally, researchers will no longer have to install specific analysis software themselves (provided that the required tools are included in this genotyping module), or convert their data into the correct input format before running an analysis tool, as this will be handled by the system as gracefully as possible. The system will be extensible, thus allowing supplementary analysis tools and database specifications to be added easily in the future.

Technology

This module is being written using the Jakarta Struts framework, which follows a

Model-View-Controller paradigm for system design. The view component is constructed using Java Server Pages and Javascript, as well as AJAX (Asynchronous Javascript and XML) to further enrich the user interface. The controller component is written in Java and utilizes functionality from BioJava and Jemboss as well as XML-RPC and CORBA to incorporate the different analysis and visualisation tools, and to expose local functionality. The model component uses Hibernate technology to facilitate object-relational mapping to a PostgreSQL database. The data model being implemented is derived from the OMG SNP and genotype model developed by the LSR. This OMG SNP model is a platform independent XMI (XML metadata interchange) model, aimed to become a standard interchange format between genotyping organisations like HapMap, dbSNP, HGVBase, ALFRED, and this module.