Towards Human-Human-Computer Interaction for Biologically-Inspired Problem-Solving in Human Genetics

Jason H. Moore
Dartmouth College
HB7937, One Medical Center Dr.
Lebanon, NH 03756 USA
603-653-9939
jason.h.moore@dartmouth.edu

Nate Barney
Dartmouth College
HB7937, One Medical Center Dr.
Lebanon, NH 03756 USA
603-653-9939
nate.barney@dartmouth.edu

Bill C. White
Dartmouth College
HB7937, One Medical Center Dr.
Lebanon, NH 03756 USA
603-653-9939
bill.c.white@dartmouth.edu

ABSTRACT
Genetic programming (GP) shows great promise for solving complex problems in human genetics. Unfortunately, many of these methods are not accessible to biologists. This is partly due to the complexity of the algorithms that limit their ready adoption and integration into an analysis or modeling paradigm that might otherwise only use univariate statistical methods. This is also partly due to the lack of user-friendly, open-source, platform-independent, and freely-available software packages that are designed to be used by biologists for routine analysis. It is our objective to develop, distribute and support a comprehensive software package that puts powerful GP methods for genetic analysis in the hands of geneticists. It is our working hypothesis that the most effective use of such a software package would result from interactive analysis by both a biologist and a computer scientist (i.e. human-human-computer interaction). We summarize briefly here the design and implementation of an open-source software package called Symbolic Modeler (SyMod) that seeks to facilitate geneticist-bioinformaticist-computer interactions for problem solving in human genetics. More information can be found at www.epistasis.org or www.symbolicmodeler.org.

Categories and Subject Descriptors
J.3 [**Computer Applications**]: Life and Medical Sciences – *biology and genetics.*

General Terms
Algorithms, Design, Human Factors

Keywords
Genetic Analysis, Genetic Epidemiology, Genetic Programming, Open-Source Software, Symbolic Discriminant Analysis, Symbolic Regression.

1. INTRODUCTION
Human genetics is transitioning away from the study of single-gene Mendelian diseases such as cystic fibrosis to tackling common complex diseases such as cancer and cardiovascular disease that represent the majority of the public health burden. The transition to more complex diseases and the widespread availability of high-throughput technologies for measuring genes necessitates powerful analytical methods that are able to model the relationship between multiple genetic and environmental factors and susceptibility to disease in the context of high-dimensional datasets. The ultimate goal of these endeavors is the identification and characterization of genetic risk factors that can be used to improve the detection, prevention and treatment of disease.

Genetic algorithms, genetic programming, and other biologically-inspired computational intelligence methods show great promise for solving complex biomedical problems. This especially true in human genetics where these methods have been used to identify genetic risk factors for disease. Unfortunately, many of these methods are not accessible to biologists and biomedical researchers for applied studies. This is partly due to the complexity of the algorithms that limit their ready adoption and integration into an analysis or modeling paradigm that might otherwise only use univariate statistical methods. This is also partly due to the lack of user-friendly, open-source, platform-independent, and freely-available software packages that are designed to be used by biologists for routine analysis.

Our goal was to develop a software package that would make available powerful genetic programming (GP) methods for data mining and machine learning to the human genetics community. There were several important objectives to the software design and development. First, it was important for the software to be platform-independent. Second, it was important for the software to include a user-friendly graphic-user interface (GUI) in addition to a simple command-line interface that could be scripted. Third, it was important to include publication quality graphical output in the GUI. Fourth, it was important to include a number of configuration options for the expert user. Finally, it was important for the software to be able to generate and use expert knowledge that can be used to help guide the algorithms.

There are two primary data mining methods implemented in SyMod. The first, symbolic discriminant analysis (SDA), was developed as a flexible alternative to linear discriminant analysis for modeling predicting discrete outcomes or classes. The second, symbolic regression, is similar to SDA except that the endpoint that is modeled is continuous. SyMod uses GP as a stochastic search algorithm for identifying optimal SDA and symbolic regression models.

SyMod was programmed entirely in Java and is available for download as open-source from www.epistasis.org or www.symbolicmodeler.org.

Copyright is held by the author/owner(s).
GECCO'07, July 7–11, 2007, London, England, United Kingdom
ACM 978-1-59593-697-4/07/0007.