Chapter 8.4
Solving Complex Problems in Human Genetics Using Nature-Inspired Algorithms Requires Strategies which Exploit Domain-Specific Knowledge

Casey S. Greene
Dartmouth College, USA

Jason H. Moore
Dartmouth College, USA

ABSTRACT

In human genetics the availability of chip-based technology facilitates the measurement of thousands of DNA sequence variations from across the human genome. The informatics challenge is to identify combinations of interacting DNA sequence variations that predict common diseases. The authors review three nature-inspired methods that have been developed and evaluated in this domain. The two approaches this chapter focuses on in detail are genetic programming (GP) and a complex-system inspired GP-like computational evolution system (CES). The authors also discuss a third nature-inspired approach known as ant colony optimization (ACO). The GP and ACO techniques are designed to select relevant attributes, while the CES addresses both the selection of relevant attributes and the modeling of disease risk. Specifically, they examine these methods in the context of epistasis or gene-gene interactions. For the work discussed here we focus solely on the situation where there is an epistatic effect but no detectable main effect. In this domain, early studies show that nature-inspired algorithms perform no better than a simple random search when classification accuracy is used as the fitness function. Thus, the challenge for applying these search algorithms to this problem is that when using classification accuracy there are no building blocks. The goal then is to use outside knowledge or pre-processing of the dataset to provide these building blocks in a manner that enables the population, in a nature-inspired framework,

DOI: 10.4018/978-1-61350-456-7.ch8.4

to discover an optimal model. The authors examine one pre-processing strategy for revealing building blocks in this domain and three different methods to exploit these building blocks as part of a knowledge-aware nature-inspired strategy. They also discuss potential sources of building blocks and modifications to the described methods which may improve our ability to solve complex problems in human genetics. Here it is argued that both the methods using expert knowledge and the sources of expert knowledge drawn upon will be critical to improving our ability to detect and characterize epistatic interactions in these large scale biomedical studies.

INTRODUCTION

Nature-inspired algorithms are a natural fit for solving problems in biological domains, not just because of the connection between method and application but also because many of the problems natural systems solve are common to biological data. Biological organisms evolve in a noisy environment with a rugged fitness landscape. Many of the interesting problems in human genetics also likely involve a rugged fitness landscape where models that contain some but not all of the relevant attributes may not have an accuracy greater than that of the surrounding noise. In addition these data are frequently noisy, in the sense that two individuals with the same values at the relevant attributes may have different disease states. In this context it is no surprise that we look to natural systems for inspiration when designing algorithms which succeed in this domain. Wagner discusses the role of robustness and evolvability in living systems (Wagner, 2005). We must design and use algorithms that, like living systems, are both robust to the noise in the data and evolvable despite the rugged fitness landscape. We briefly discuss the Relief family of machine learning methods which are useful for separating signals from noise in this type of data and then focus on approaches that exploit this information. The nature-inspired methods we examine here are genetic programming (GP), a computational evolution system (CES), and ant colony optimization (ACO).

GP is an automated computational discovery tool inspired by Darwinian evolution and natural selection (Koza, 1992, 1994; Koza, Andre, Bennett & Keane, 1999; Koza, 2003; Banzhaf, Nordin, Keller & Francone, 1998; Langdon & Koza, 1998; Langdon & Poli, 2002). The goal of GP is to evolve computer programs which solve problems. This is accomplished by generating programs composed of the building blocks needed to solve or approximate a solution and then iteratively evaluating, selecting, recombining, and mutating these programs to form new computer programs. This process repeats until a best program or set of programs is identified. Genetic programming and its many variations have been applied successfully to a wide range of different problems including data mining, knowledge discovery (Freitas, 2002), and bioinformatics (Fogel & Corne, 2003). Despite the power of this method, there remain a number of challenges that GP practitioners and theorists must address before this computational discovery tool becomes a standard in the modern problem solver's toolbox. Yu et al. list 22 such challenges (Yu, Riolo & Worzel, 2006). We discuss here methods that address some of these challenges, in particular those related to practice. We specifically discuss methods that use information from pre- and post- processing, methods for handling large high dimensional datasets, and methods for integrating domain knowledge. We argue that these methods will be critical if we are to successfully and reliably analyze these genetic data for epistasis.

Spector, as part of an essay regarding the roles of theory and practice in genetic programming, discusses the push towards biology by GP practitioners (Spector, 2003). Banzhaf et al. propose the transformation of overly simplistic and