Information Sciences 181 (2011) 1609–1622

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Ant colony optimisation to identify genetic variant association with type 2 diabetes

Jacqueline Christmas [a], Edward Keedwell [a,*], Timothy M. Frayling [b], John R.B. Perry [b]

[a] School of Engineering, Computing and Mathematics, University of Exeter, North Park Road, Exeter EX4 4QF2EF, UK
[b] Genetics of Complex Traits, Institute of Biomedical and Clinical Science, Peninsula Medical School, Magdalen Road, Exeter EX1 2LU, UK

ARTICLE INFO

Article history:
Received 18 May 2010
Received in revised form 6 December 2010
Accepted 11 December 2010
Available online 21 December 2010

Keywords:
Ant colony optimisation
Bioinformatics
Single nucleotide polymorphisms
Genome wide association studies

ABSTRACT

Around 1.8 million people in the UK have type 2 diabetes, representing about 90% of all diabetes cases in the UK. Genome wide association studies have recently implicated several new genes that are likely to be associated with this disease. However, common genetic variants so far identified only explain a small proportion of the heritability of type 2 diabetes. The interaction of two or more gene variants, may explain a further element of this heritability but full interaction analyses are currently highly computationally burdensome or infeasible. For this reason this study investigates an ant colony optimisation (ACO) approach for its ability to identify common gene variants associated with type 2 diabetes, including putative epistatic interactions. This study uses a dataset comprising 15,309 common (>5% minor allele frequency) SNPs from chromosome 16, genotyped in 1924 type 2 diabetes cases and 2938 controls. This chromosome contains two previously determined associations, one of which is replicated in additional samples. Although no epistatic interactions have been previously reported on this dataset, we demonstrate that ACO can be used to discover single SNP and plausible epistatic associations from this dataset and is shown to be both accurate and computationally tractable on large, real datasets of SNPs with no expert knowledge included in the algorithm.

1. Introduction

According to [6] around 1.8 million people in the UK have type 2 diabetes, representing around 90% of all diabetes cases in the UK. Diabetes is the leading cause of blindness in the working age population and although type 2 diabetes is traditionally associated with ageing, it is appearing increasingly in younger adults. In 2006, clinical care for diabetic patients accounted for 5% of the NHS budget (about £10 million per day) and this is expected to rise to 10% by 2011. There is currently no cure, but the identification of genetic risk factors has implicated several new genes that are likely to be involved. It is as yet unclear how this will translate into improved prediction and prevention.

1.1. Genetic association studies

Genetic association studies aim to discover which genetic variations increase or decrease the likelihood of an individual contracting a given disease by comparing common DNA variants between affected and unaffected individuals. Recent efforts, including this study, aim to link small changes in the DNA of individuals known as single nucleotide polymorphisms (SNPs) in particular positions of the genome to increased risk of the individual developing particular diseases [14]. SNPs consist of

* Corresponding author.
E-mail address: E.C.Keedwell@ex.ac.uk (E. Keedwell).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.12.005

two alleles (two of the four bases of the genetic code, A, C, G or T) and because humans have two copies of the genome (diploid), each individual has one of three genotypes at each SNP position. For example, at an A/C SNP, individuals will be one of AA, AC or CC genotypes. These differences in genotype, although small, can have profound effects on the probability of individuals developing certain diseases.

In monogenic diseases, such as cystic fibrosis [12], the presence or absence of a single allele completely predicts the presence or absence of the disease [17]. However, for many common diseases, such as type 2 diabetes [15], an individual's susceptibility is influenced by a complex interaction of environmental and genetic factors which yields a probabilistic connection between genetic variation and the disease. In polygenic diseases, the presence of a risk allele will increase or decrease the *probability* of the disease, and there may be many different risk alleles [17] associated with a disease. The discovery of single risk alleles requires a computationally tractable search through all SNPs (~400,000 in the human genome) and their association with the disease. Whilst single risk alleles have been shown to be very informative, much of the genetic variation in a trait or disease remains uncharacterised. One of the leading contenders for this "missing heritability" is epistasis, the interaction between genes in a genome. Moore [21] asserts that epistasis, may influence predisposition to common diseases and the same author, in [20], asserts that it is likely to be ubiquitous, or at least widespread. Furthermore, epistasis has previously been implicated in insulin resistance [2], HIV [4] and Alzheimer's disease [31]. This discovery of gene–gene interactions in the same dataset (i.e. the search for pairs, triplets and higher order combinations of SNPs) yields combinatorial complexity that is no longer computationally tractable.

Until recent technological advances made it possible to investigate a large proportion of variation in the human genome, genetic association studies were limited to looking only at small numbers of genes where research had identified possible biological reasons for their involvement in disease phenotypes. By the end of 2006, the Wellcome Trust Case Control Consortium had performed one of the first and largest genome wide association studies using data from 400,000 SNPs in 17,000 individuals. The WTCCC investigated seven diseases including type 2 diabetes [35] and many other genome wide association studies (GWAS) have since been reported. This explosion of information has triggered the search for methods that are able to analyse such highly dimensional data for statistical correlations with disease status, leading some to apply artificial intelligence (AI) techniques (e.g. [3,22–27]).

The task for these AI techniques in genetic association studies is to discover a small number of SNPs (feature selection) that are informative to the disease status and then establish how the alleles for those SNPs combine to classify an individual as to their susceptibility to the disease.

1.2. Previous statistical and artificial intelligence research

A number of studies have established type 2 diabetes as a polygenic disease [9,28–30,32–34]. The analysis of each SNP individually, even across the entire genome, is feasible as it simply requires the testing of each SNP in turn (around 400,000–1 million tests in the case of the WTCCC and many GWAS). Research in [2,11,36] has found some evidence for epistasis but there has been little robust evidence for gene–gene interaction in type 2 diabetes and many other complex diseases/traits. Searching for these interactions requires exhaustive testing of every possible combination of *2 or more* SNPs, transforming the problem from one of manageable complexity to one where the number of possible combinations is too large to compute on the fastest machines for even modest numbers of combinations. A search for all pairs within a database of 400,000 SNPS would require 8×10^{10} tests, rendering a full enumeration highly burdensome and the investigation of higher order interactions (i.e. triplets and above) intractable.

This problem of high dimensionality and the recent explosion in the number of SNP-disease associations datasets available for study has caused some researchers to turn to AI techniques that have previously been shown to give good results for other highly dimensional problems.

Neural networks [24,26,5] and support vector machines [3] are popular, although [24] asserts that neural networks are unable to handle incomplete data. GWAS data will often include samples and SNPs with missing genotype information, which limits the usefulness of neural network approaches. Additionally, the model which results from these techniques is essentially a complex matrix of weights and biases and is therefore difficult to understand. Algorithms that are able to produce more transparent results, for example, in rules of the form "IF <conditions> THEN <class prediction>", are preferable for this type of task, especially where the end-users such as clinicians and biologists are not computer experts. Popular algorithms with the required transparency include decision tree algorithms [25] which make use of a greedy heuristic but whilst this may be good for detecting polygenic associations, it is unlikely to be beneficial for detecting epistasis. This is due to the use of a greedy heuristic which effectively implements an additive model where features (in this case SNPs) are evaluated in order of their individual classification effect on the dataset. However, there is some evidence that these techniques in ensemble can be used for such studies. Researchers in [18] used random forests (an ensemble of decision trees) to discover two-way and three-way interactions in the Age-related Macular Degeneration dataset, although they do note that none of the interactions is significant when corrected for multiple tests, this demonstrates the potential for machine learning to be used in the context of genetic association studies. Bayesian belief networks have also been applied to genetic association studies [27], though only to 20 SNPs and the high dimensionality of real-world datasets precludes the use of this exhaustive search method.

Evolutionary algorithms use parallel stochastic searches of the solution space to avoid the trap of finding locally optimum solutions and have proven particularly good at searching across highly dimensional space. Using a genetic programming algorithm against a 1000-dimensional set of artificially generated epistatic data, Moore and White [22,23] conclude that ex-