Ant Colony Optimisation for Exploring Logical Gene-Gene Associations in Genome Wide Association Studies

Emmanuel Sapin, Ed Keedwell, and Tim Frayling

University of Exeter, College of Engineering, Mathematics and Physical Sciences, Harrison Building, Exeter, England
{e.sapin,e.c.keedwell}@exeter.ac.uk
Genetics of Complex Traits, Institute of Biomedical and Clinical Science, Peninsula Medical School, Magdalen Road, Exeter, England
{tim.frayling}@pms.ac.uk

Abstract. In this paper a search for the logical variants of gene-gene interactions in genome-wide association study (GWAS) data using ant colony optimisation is proposed. The method based on stochastic algorithms is tested on a large established database from the Wellcome Trust Case Control Consortium and is shown to discover logical operations between combinations of single nucleotide polymorphisms that can discriminate Type II diabetes. A variety of logical combinations are explored and the best discovered associations are found within reasonable computational time and are shown to be statistically significant.

Keywords: Genome Wide Associations, Type II diabetes, Single Nucleotide Polymorphisms

1 Introduction

The human genome is written in base pairs of the four single nucleotides (Adenine, Cytosine, Guanine and Thymine) and variations of these nucleotides exist within a population of individuals called single-nucleotide polymorphisms (SNPs). These SNPs can determine phenotypic traits of individuals (e.g. height, BMI) and the propensity to suffer from diseases such as Type 1 and Type 2 Diabetes. Each SNP has three possible genotypes (e.g. CC, GG, CG) due to the diploid nature of human genomes. As sequencing genomes becomes cheaper, the sequencing of thousands of genomes individuals are now possible and this has opened a door to new types of wide-ranging studies. It is indeed now possible to search for the relationship between the genome and diseases across a population of individuals and for a large number of SNPs. This major challenge is represented by a set of studies known as Genome-Wide Association Studies (GWAS).

Type II diabetes (T2D), affected hundreds of millions people over the world [6] and is characterized by insulin resistance. The heritability of this disease has

been proven widely [4]. Insight into the genetic architecture of T2D detecting replicated diabetes association signals is provided in [2] and associations for T2D using a family-based design to control for population stratification are reported in [3]. In [7], the role of genetic variants to confirm if they increase risk of T2D is evaluated and in [8] a common genetic variant in T2D mellitus that does not appear to be in a coding region is identified. In [9], the role of the quantitative contribution of insulin resistance and impaired insulin secretion as genetic factors is studied. However identifying genetics risk in Type II diabetes has met with only limited success [10] and then remains as a formidable challenge [11].

From a computational perspective GWAS present a significant challenge as there are hundreds of thousands SNPs (variables) per individual and in the majority of studies these are recorded for thousands of individuals creating a database of large proportions (almost 2.5bn elements in the experiments described below). Any computational approaches used to analyse this data therefore must be scalable in the face of this large-scale data. The task becomes even more complex when two or more SNPs are investigated for association, where a number of logical relations might exist between the SNPs. Here we consider the full range of two SNP logical associations with an ant-colony optimisation approach and report the most promising of these. Therefore, this paper presents a stochastic approach to the analysis of full-scale genome-wide association studies data with the aim to find combinations of SNPs that have association with T2D across of a population of thousands individuals.

2 Data

This research uses the database of The Wellcome Trust Case Control Consortium (WTCCC) which is a collection of GWAS studies relating to a variety of diseases including Type II Diabetes [1]. In this database, a total of 5003 human genomes are provided, with ∼500,000 SNPs recorded for each individual in the database. Each SNP represents a small change in the genome and consists of two alleles (Adenine, Cytosine, Guanine and Thymine). Due to the diploid nature of human genomes, there are three possible genotypes for each SNP (e.g. CC, GG, CG).

GWASs data has to undergo a series of quality control tests before it can be used. Therefore there are some exclusion criteria that must be applied to ensure the quality of the remaining data. The SNPs that were kept are those that met these four conditions in the 3,004 samples of genome of individuals without T2D. Readers are asked to refer to the GWAS literature for more information on these criteria [5].

- Hardy-Weinberg equilibrium Exact Test $> 10^{-4}$
- Minor allele frequency $> 1\%$ for these 3,004 individuals
- Studywise missing data proportion $< 5\%$ for these 3,004 individuals
- Studywise minor allele frequency $> 5\%$ for these 3,004 individuals or studywise missing data proportion $< 1\%$ for these 3,004 individuals

and meeting these two conditions in the 1,999 samples of genomes of individuals with T2D:

– Hardy-Weinberg equilibrium Exact Test $> 10^{-4}$
– Minor allele frequency $> 1\%$ for these 1,999 individuals.

The remainder contains 405,139 SNPs.

3 Combinations

The combinations of two SNPs that are usually considered is the following. An individual is positive if and only if a first SNP ($snp1$) takes a specific value ($value1$) and a second SNP ($snp2$) takes a specific value ($value2$). This implements the logical operation AND between two SNPs and is the standard GWAS method to consider the combination of SNPs associated with a disease. In this study however, the complete set of logical operations between two SNPs is considered. The logical operations that are considered are based on a first SNP ($snp1$) taking a specific value ($value1$) and a second SNP ($snp2$) taking a specific value ($value2$). The two following logical expressions can take the two values 0 and 1:

$Snp1 = value1$

$Snp2 = value2$

Thus there are 4 (2^2) possibilities to consider and for each possibility, a combination can take the values 0 and 1 thus there are 16 (4^2) combinations to consider as described in the table 1 with XOR a type of logical disjunction on two operands that results in a value of true if exactly one of the operands has a value of true:

Combination1	Always 0
Combination2	$snp1 = value1$ AND $snp2 = value2$
Combination3	$snp1 = value1$ AND NOT($snp2 = value2$)
Combination4	$snp1 = value1$
Combination5	$snp2 = value2$ AND NOT($snp1 = value1$)
Combination6	$snp2 = value2$
Combination7	$snp1 = value1$ XOR $snp2 = value2$
Combination8	$snp1 = value1$ OR $snp2 = value2$
Combination9	NOT($snp1 = value1$ OR $snp2 = value2$)
Combination10	NOT($snp1 = value1$ XOR $snp2 = value2$)
Combination11	NOT($snp2 = value2$)
Combination12	NOT($snp2 = value2$ AND NOT($snp1 = value1$))
Combination13	NOT($snp1 = value1$)
Combination14	NOT($snp1 = value1$ AND NOT($snp2 = value2$))
Combination15	NOT($snp1 = value1$ AND $snp2 = value2$)
Combination16	Always 1

Table 1. The complete set of logical operations from $snp1 = value1$ and $snp2 = value2$ that are considered.

However, a number of combinations among the 16 combinations do not need to be considered:

- The combinations 1, 4, 6, 11, 13 and 16 do not involve the two SNPs.
- The combinations 12 and 5 are respectively the same as the combinations 14 and 3 with swapping the two SNPs with each other and the two values with each other.
- The combinations 9, 10, 14 and 15 are the negations of the combinations 8, 7, 3 and 2.

Thus, only four logical operations need to be considered:

- An individual is positive if and only if the first SNP takes a specific value and the second SNP takes a specific value. (Combination2: AND)
- An individual is positive if and only if the first SNP or the second SNP takes their specific values. (Combination8: OR)
- An individual is positive if and only if the first SNP takes a specific value and the second SNP does not take a specific value. (Combination3: AND NOT)
- An individual is positive if and only if exactly one of the two SNPs takes its specific value. (Combination7: XOR)

4 Methodology

A permutation-based ant colony optimisation (ACO) approach is used to search for combinations of SNPs that can discriminate T2D. A value is associated with every SNP that represents how likely the SNP could be good at discriminating T2D. This value P(n) is called the amount of pheromone of the SNP n. The amounts of pheromone P is used to select SNPs for new combinations thanks to a tournament selection inspired by [18]. The algorithm can be described as follow :

```
1 Initialise pheromone on each SNP
2 Repeat
3    For all the 100 ants:
4       Select two SNPs via tournament selection of size 50
5       Calculate the fitness of the combination
6    End
7    Updated pheromone of the two SNPs with the best fitness
8    For all SNPs: apply evaporation rate 1%
9 End
```

5 Results

The best results of one hundred algorithm runs over 10000 generations with 100 ants and 50 items in the tournament selection are saved. With these parameters, a generation of the algorithm lasts in average 1.28 seconds so a run lasts

an average of 3 hours and 32 minutes using a machine with a 1TB 7200RPM hard-drive and an Intel Core i7-2600 CPU @3.40GHz. The ACO algorithm found good results for combinations of two SNPs. Each combination is shown in tables 2 and 3 with the rs identification number and then chromosome number and position on the genome in brackets. Each combination has a calculated p-value to determine the likelihood of this association being discovered by chance, hence smaller values are better. These p-values can be compared to those drawn from a sample of randomly generated AND associations that resulted in a best p-value of 7×10^{-13}.

Some combinations are with two SNPs in similar regions on the genome are often correlated through a phenomenon known as linkage disequilibrium (LD), meaning this association is therefore likely to be an artefact, despite they are the strongest signal within this dataset. These combinations are shown table 2 whereas the combination with two SNPs that cannot be correlated through linkage disequilibrium (LD) (as they are on different chromosomes) are shown in table 3.

Combination	p-value
rs7031174(9,36651528)=CA and rs7045471(9,36603003)=CC	8×10^{-36}
rs7031174(9,36651528)=CA and rs10814425(9,36643684)=AA	7×10^{-35}
rs7031174(9,36651528)=CA and rs10973013(9,36645648)=AA	7×10^{-35}
rs7031174(9,36651528)=CA and rs2151644(9,36623898)=AA	7×10^{-35}
rs7031174(9,36651528)=CA and rs10972978(9,36566847)=CC	2×10^{-34}
rs10829495(10,130478405)=CT or rs4132670(10,114757761)=TT	1×10^{-15}
rs4512469(9,36649725)=GA xor rs7031174(9,36651528)=GA	4×10^{-28}
rs4512469(9,36649725)=AA xor rs7031174(9,36651528)=AA	1×10^{-27}
rs10733480(9,36684476)=AA xor rs7031174(9,36651528)=TT	2×10^{-26}
rs10733480(9,36684476)=AT xor rs7031174(9,36651528)=AT	1×10^{-26}
rs7031174(9,36651528)=CA xor rs4512469(9,36649725)=AA	5×10^{-18}

Table 2. Sample of the best combinations of SNPs described as rs number(chromosome, position) and their p-value that were discovered.

Table 2, as expected shows very small p-values, theoretically the strongest signals in the dataset, but linkage disequilibrium prohibits them from further analysis. The very close positions on the chromosome mean that it is likely that these SNPs are indeed correlated. However, it is interesting to note that a variety of logical operations appear in this table of lowest p-values. Table 3 shows a more biologically plausible set of results with rs7901695, rs11196205 and rs4506565 previously associated with Type 2 diabetes [20] [1]. It should be noted that rs7077039 is also closely located on the genome to rs11196205, although it has not been named as a contributor to T2D.

Combination	p-value
rs10992923(9,93688875)=GA or rs11196205(10,114797037)=GG	3×10^{-16}
rs7666328(4,116140909)=GA or rs7077039(10,114779067)=GG	5×10^{-16}
rs6449054(4,14240655)=TC or rs4506565(10,114746031)=CC	8×10^{-16}
rs11196208(10,114801306)=GG or rs10992923(9,93688875)=AG	1×10^{-15}
rs17489797(4,63836645)=TA and not(rs7901695(10,114744078)=AA)	8×10^{-18}
rs4506565(10,114746031)=AA and not(rs6449054(4,14240655)=AA)	8×10^{-16}
rs11196205(10,114797037)=GG and not(rs10992923(9,93688875)=CC)	2×10^{-15}
rs17489797(4,63836645)=TT and not(rs7901695(10,114744078)=AA)	1×10^{-15}
rs4132670(10,114757761)=AA and not(rs6449054(4,14240655)=AA)	1×10^{-15}

Table 3. Sample of the best combinations of SNPs described as rs number(chromosome, position) and their p-value that were discovered.

6 Conclusion

In this paper, an ant colony approach to the problem of discovering combinations of SNPs from large-scale GWAS data has been described. Combinations of 2 SNPs that can discriminate T2D patients from controls have been discovered by the approach. The ACO has been able to find some of the strongest signals in the dataset (although as explained above these have been ruled out on biological grounds) and has found associations that are replicated in the literature. The investigation of logical variations has shown that these provide the algorithm with greater power to express the relationship between two or more SNPs. In particular, the NOT operator allows the system to exclude one genotype in a SNP and exclude the others, an important logical distinction. Further work is required to examine these relationships in more detail and to determine if they have biological plausibility in addition to statistical significance.

The ACO method can be applied to any GWAS dataset that conforms to the standard OXSTATS format and so further trials are planned on other disease datasets from the WTCCC, including Type I Diabetes and Rheumatoid Arthritis. The algorithm is also able to discover higher order combinations of SNPs (e.g. 3+ SNPs, not shown) that would not be possible using existing methods. However, care would need to be taken with higher-order interactions as the number of possible logical operations will increase very quickly and it may not be practical to test all feasible combinations for every SNP combination.

7 Acknowledgements

This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

The work contained in this paper was funded by an EPSRC First Grant (EP/J007439/1) and we acknowledge their kind support.

References

1. Wellcome Trust Case Control Consortium. Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls. Nature, 447:661-978, 2007.
2. E. Zeggini, M. Weedon, C. Lindgren, T. Frayling, K. Elliott, and Hana Lango et al. Replication of genome-wide association signals in uk samples reveals risk loci for type 2 diabetes. Science, 316:1336-1341, 2007.
3. D. Altshuler, J. Hirschhorn, M. Klannemark, C. Lindgren, M. Vohl, J. Nemesh et al. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet, 26:76-80, 2000.
4. S. Rich. Mapping genes in diabetes. genetic epidemiological perspective. Diabetes., 39(11):1315-9, 1990.
5. M. Weedon et al. Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. PLoS Medicine 3, 10:1877-1882, 2006.
6. M. Kasuga Insulin resistance and pancreatic beta cell failure. J Clin Invest, 116:1756–1760, 2006.
7. A. Gloyn, M. Weedon, K. Owen, Martina J. Turner, Br. Knight, and G. Hitman et al. Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes.. Diabetes, 52:568-572, 2003.
8. S. Grant, G. Thorleifsson, I. Reynisdottir, R. Benediktsson, A. Manolescu, and J. Sainz et al. Variant of transcription factor 7-like 2 (tcf7l2) gene confers risk of type 2 diabetes. Nat Genet, 38(3):320-3, 2006.
9. G. John. The genetic basis of type 2 diabetes mellitus: Impaired insulin secretion versus impaired insulin sensitivity. Endocrine Reviews, 19:491-503, 1998.
10. J. Kaprio, J. Tuomilehto, M. Koskenvuo, K. Romanov, A. Reunanen, J. Eriksson et al. Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in finland. Diabetologia, 35(11):1060-7, 1992.
11. L. Scott, K. Mohlke, L. Bonnycastle, C. Willer, Y. Li, and W. Duren et al. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. Science, 316(5829):1341-5, 2007.
12. J. Moore and W. White. Exploiting knowledge in genetic programming for genome-wide genetic analysis. in Lecture Notes in Computer Science, T. Runarsson, H. Beyer, E. Burke, J. Merelo-Guervs, L. Whitley and X. Yao, Eds. Springer, 4193:969-977, 2006.
13. Moore and H. Jason A global view of epistasis. Nat Genet, 37(1):13-14, January 2005.
14. C. Greene, B. White, and J. Moore. Ant colony optimization for genome-wide genetic analysis. In Marco Dorigo, Mauro Birattari, Christian Blum, Maurice Clerc, Thomas Sttzle, and Alan Winfield, editors, Ant Colony Optimization and Swarm Intelligence, volume 5217 of Lecture Notes in Computer Science, pages 37-47. Springer Berlin /Heidelberg, 2008.
15. M. Dorigo and G. Di Caro. The ant colony optimization meta-heuristic. In in New Ideas in Optimization, pages 11-32. McGraw-Hill, 1999.
16. A. Zecchin, H. Maier, A. Simpson, M. Leonard, and J. Nixon. Ant colony optimization applied to water distribution system design: Comparative study of five algorithms. In Journal of Water Resources Planning and Management, Vol. 133, No. 1, January 1., 2007.

17. T. Stutzle and M. Dorigo. Aco algorithms for the traveling salesman problem 1999. In Periaux (eds), Evolutionary Algorithms in Engineering and Computer Science: Recent Advances in Genetic Algorithms, Evolution Strategies, Evolutionary Programming, Genetic Programming and Industrial Applications. John Wiley & Sons, 1999.
18. E. Sapin and E. Keedwell. T–ACO tournament ant colony optimisation for high dimensional problems. in ECTA 2012 - 4th International Conference on Evolutionary Computation Theory and Applications, Barcelona, Spain, 5th – 7th Oct 2012.
19. J. Christmas, E. Keedwell, T. Frayling, and J. Perry. Ant colony optimisation to identify genetic variant association with type 2 diabetes. In Information Sciences., volume 181, pages 1609-1622,2011.
20. T. Hayashi, Y. Iwamoto, K. Kaku, H. Hirose, S. Maeda. Replication study for the association of TCF7L2 with susceptibility to type 2 diabetes in a Japanese population. In Diabetologia. 2007 May;50(5):980–4. Epub 2007 Mar 6.