Computational intelligence for genetic association study in complex diseases: review of theory and applications

Arpad Kelemen

Department of Organisational Systems and Adult Health,
University of Maryland,
655 W. Lombard. St., Rm 475A,
Baltimore, MD, 21201, USA
Fax: 410-706-3289
E-mail: akele001@umaryland.edu

Athanasios V. Vasilakos

Department of Computer and Telecommunications Engineering,
University of Western Macedonia,
50100 Kozani, Greece
E-mail: vasilako@ath.forthnet.gr

Yulan Liang*

Department of Family and Community Health,
University of Maryland,
655 W. Lombard. St., Rm 404K,
Baltimore, MD, 21201, USA
Fax: 410-706-0730
E-mail: ylian001@umaryland.edu
*Corresponding author

Abstract: Comprehensive evaluation of common genetic variations through association of SNP structure with common complex disease in the genome-wide scale is currently a hot area in human genome research thanks for the recent development of the Human Genome and HapMap Projects. Computational science, which includes computational intelligence, has recently become the third method of scientific enquiry besides theory and experimentation. There have been fast growing interests in developing and applying computational intelligence in disease mapping using SNP and haplotype data. This review provides coverage of recent developments of theory and applications in computational intelligence for complex diseases in genetic association study.

Keywords: computational intelligence; SNPS; single nucleotide polymorphisms; epistasis; common complex diseases.

Reference to this paper should be made as follows: Kelemen, A., Vasilakos, A.V. and Liang, Y. (2009) 'Computational intelligence for genetic association study in complex diseases: review of theory and applications', *Int. J. Computational Intelligence in Bioinformatics and Systems Biology*, Vol. 1, No. 1, pp.15–31.

Biographical notes: Arpad Kelemen is an Associate Professor in the Department of Organisational Systems and Adult Health at the University of Maryland, Baltimore. He received his PhD from the University of Memphis and his MS Degree from the University of Szeged, Hungary, both in Computer Science. His research interests include computational intelligence, artificial intelligence, intelligent agents, machine learning, neural networks, bioinformatics, neuroimaging, medical imaging, and computational biology. He published two books and more than 50 papers in international journals and conferences. He received several awards and honors.

Athanasios V. Vasilakos is a Professor at the Department of Computer and Telecommunications Engineering, University of Western Macedonia, Greece; and a Visiting Professor at the Graduate Programme, Department of Electrical and Computer Engineering, National Technical University of Athens, Greece. He serves on the editorial board of several international journals, including *Int. J. Adaptive and Autonomous Communications Systems*; and *Int. J. Arts and Technology*. He has published more than 150 papers in international journals and conferences. He has co-authored a number of books, including Ambient Intelligence, Wireless Networking, Ubiquitous Computing, Computational Intelligence in Telecommunication Networks, Arts and Technologies, and more.

Yulan Liang is an Associate Professor in the Department of Family and Community Health and Senior Biostatistician at the Research Office of the University of Maryland, Baltimore. She was an Assistant Professor at the Department of Biostatistics at the University at Buffalo from 2002 to 2008. She received her PhD and MS Degrees in Applied Statistics from the University of Memphis. Her research interests include Bayesian inference, categorical data modelling, multivariate statistics, time series, data mining, neural networks, statistical learning, statistical pattern recognition. Her application areas include statistical genomics and bioinformatics, epidemiology, neuroimaging, medical imaging and clinical trials.

1 Introduction

DNA sequencing, the process of determining the exact order of the 3 billion chemical building blocks (called DNA base pairs or nucleotides and abbreviated A, T, C, and G) that make up the DNA of the 24 different human chromosomes, was the greatest technical challenge in the Human Genome Project. Achieving this goal has helped reveal the estimated 20,000–25,000 human genes within our DNA as well as the regions controlling them. The resulting DNA sequence maps are being used by 21st century scientists to explore human biology and other complex phenomena. In May 2006, Human Genome Project (HGP) researchers announced the completion of the DNA sequence for the last of the 24 human chromosomes. Many small regions of DNA that vary among individuals (called polymorphisms) were also identified during the HGP, mostly Single Nucleotide Polymorphisms (SNPs). For the most part, chromosome composition is similar between two random persons. The term polymorphism' refers to the set of possible genetic configurations (alleles) at a specific location (locus). About 90% of genetic variation in humans is attributed to differences in single bases of DNA (Collins et al., 1998). These single nucleotide polymorphisms have been dubbed 'SNPs' in the literature. SNPs are single base pair positions in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals in some population(s), wherein the least frequent allele has an abundance of 1% or greater (Brookes, 1999). In practice, the term 'SNP' is used more loosely. Most SNPs are without physiological effect, although a minority contributes to the delightful and beneficial diversity of humanity. A much smaller minority of polymorphisms affect an individual's susceptibility to disease and response to medical treatments.

Identification of genetic polymorphisms involved in the etiology of human disorders relies on the tools of linkage and association mappings. Linkage mapping searches for markers cosegregating with disease within family or pedigree, which have been developed for decades. Association mapping looks for association between disease phenotype and genetic polymorphism, which can be obtained either through family based study or general population. In order to increase the efficiency of analysis processes, association mapping has increasingly focused on genetic variations (polymorphisms) that occur naturally in the population. Availability of improved detail and resolution of genetic maps also resulted in increased use of association mapping because it is more powerful than linkage methods.

The HapMap Project has collected genotypes of millions of SNPs from populations with ancestry from Africa, Asia and Europe and makes this information freely available in the public domain (The International HapMap Consortiu, 2003, 2004, 2005). While millions of SNPs have been identified, with an estimated two common missense variants per gene, there is a great need, conceptually as well as computationally, to develop advanced robust algorithms and analytical methods for characterising genetic variations that are non-redundant and identify the target SNPs that are most likely to affect the phenotypes and ultimately contribute to disease development. The recent extensive interest in genome polymorphism signifies a development in human genetics research that will have a major impact upon population genetics, drug development, forensics, cancer and genetic disease research (Cardon and Bell, 2001; Risch, 2000; Risch and Merikangas, 1996; Stephens and Donnelly, 2000).

Knowledge of the interplay between genetics and environmental factors is central to the understanding of multifactorial disease processes (Ioannidis et al., 2006; Chatterjee 2006; Cordell 2004; Hunter 2005; Zondervan 2004; Azevedo et al., 2006). The biological interest is in how polymorphic genes interact with each other and with environmental factors to influence susceptibility and outcome in common, complex diseases. The risks of major common diseases such as cancer, cardiovascular disease, mental illness, auto- immune states, and diabetes, are expected to be heavily influenced by the patterns of SNPs one possesses in certain key susceptibility genes yet to be identified. The term 'complex disease' refers to the scenario in which these diseases are contributed to the susceptibility by many genes, many environmental factors that interact in a hierarchical fashion with nonlinear, polygenic, epistasis (the interaction between alleles at different loci) effects (Moore, 2003; Jasnos and Korona, 2007; Martin et al., 2007). By disregarding these interactions, effect sises of relative risk for individual genetic variants are expected to be small. Disregarding gene-environment interaction also weakens exposure-disease and gene-disease associations.

During the last few years, there have been fast growing interests in developing and applying computational and statistical approaches in disease mapping in genetic association study using SNPs and haplotype data (Chapman et al., 2003; Liu and Lin, 2005; Halldrsson et al., 2004; Howie et al., 2006; Ke and Cardon, 2003; Akey et al., 2001; Gabriel et al., 2002; Daly et al., 2003; Zhang et al., 2002a, 2002b; Meng et al., 2004; Anderson and Novembre, 2003; Mannila et al., 2003; Beckmann et al., 2005; Levin et al., 2005; Schaid et al., 2002; Nothnagel et al., 2002; Hampe et al., 2003; Zhao et al., 2005; Zaykin et al., 2002; Ott, 2001, 2004; He and Zelikovsky, 2006; Durrant et al., 2004; Baker, 2005; Tzeng et al., 2006; Burkett et al., 2004; Greenspan and Geiger, 2004, 2006;Thomas et al., 2003; Schwender and Ickstadt, 2006; Verzilli et al., 2006; Clark et al., 2005; Lam et al., 2000; Chang et al., 2006; Li and Jiang, 2005; Lin and Altman, 2004; Horne and Camp, 2004; Ao et al., 2005; Kooperberg et al., 2001; Moore, 2007; Liu et al., 2001; Molitor et al., 2003; Toivonen et al., 2000; Ritchie et al., 2003). These computing approaches can be roughly categorised into several groups:

- Statistical measure and testing based approaches. Examples including Mantel statistic; Scan Statistic; Score statistic; Minimum description length; Entropy-based measure (Chapman et al., 2003; Anderson and Novembre, 2003; Mannila et al., 2003; Zhang et al., 2002; Beckmann et al., 2005; Levin et al., 2005; Schaid et al., 2002; Nothnagel et al., 2002; Hampe et al., 2003; Zhao et al., 2005; Zaykin et al., 2002; Ott, 2004; He and Zelikovsky, 2006; Durrant et al., 2004; Baker, 2005; Tzeng et al., 2006; Burkett et al., 2004; Greenspan and Geiger, 2004, 2006; Thomas et al., 2003; Schwender and Ickstadt, 2006).

- Unsupervised learning algorithms, such as clustering and graph methods; Principal Component analysis; Haplotype Pattern Mining. tagSNP selections partially belong here (Li and Jiang 2004; Lin and Altman, 2004; Horne and Champ, 2004; Ao et al., 2005; Kooperberg et al., 2001; Moore, 2007; Molitor et al., 2003; Toivonen et al., 2000).

- Supervised learning algorithms and statistical models which require clinical outcomes such as disease status to guide the SNPs selection: Random forests; Multifactor Dimensionality Reduction (MDR); support vector machine; Multiple linear regressions; logistic-regression; haplotype trend regression. Some of these supervised approaches were also developed for modelling gene-gene interaction analysis such as logic regression and MRD.

- Machine learning and statistical learning approaches (Ott, 2001; Clark et al., 2005; Lam, et al., 2000; Chang et al., 2006); Computational intelligence approaches including neural networks; genetic algorithms; genetic programming; evolutionary trees; hybrid systems: genetic programming neural networks (Clark et al., 2005, Lam et al., 2000, Ritchie et al., 2003a, 2003b; Banzhaf et al., 2006; Moore and White, 2006; Foster, 2001; Motsinger et al., 2006a, 2006b). Several surveys relating these approaches to disease mapping have been provided (Onkamo and Toivonen, 2006; Salem et al., 2005; Molitor et al., 2004; Shah and Kusiak, 2004; McKinney et al., 2006). For instance, Onkamo and Toivonen (2006) provided a survey of data mining approaches of disease mapping in bioinformatics (Onkamo and Toivonen, 2006); McKinney et al. reviewed a number of different machine learning methods that have been applied to detect gene-gene interactions (McKinney et al., 2006). Although the number of publications of SNP-disease association study using computational intelligence approaches is lower than statistical modelling based approaches, some recent studies have demonstrated promises and future impacts of this field, especially when applied to complex disease association analysis (Clark et al., 2005; Lam et al., 2000, Ritchie et al., 2003; Banzhaf et al., 2006; Moore and White, 2006a; Ritchie et al., 2003; Foster, 2001; Motsinger et al., 2006a, 2006b). This review focuses on recent developments in computational intelligence for diseases mapping in genetic association study. At the end of the review we uncover some areas which may be potential future directions in computational intelligence for genomic studies in complex diseases.

2 Some computational challenges of association study in common complex diseases

An important challenge that faces molecular association study in the post genomic era is to understand the inter-connections from a network of genes and their products that are initiated and mediated by a variety of environmental changes. The variety of phenotype definitions lead to a multiplicity of tests and also involve large number of comparisons, which often result in less power. Non-reproducibility of many reported significant associations in subsequent studies has led to criticism of association studies.

For SNP and haplotype data in common complex diseases, in addition to being large, redundant, diverse and distributed, there are three important characteristics that pose challenges for data analysis and modelling:

- complexity
- heterogeneity
- a constantly evolving nature.

It is heterogeneous, in the sense that it involves a wide array of data types, including categorical, continuous, sequence data, as well as temporal data, incomplete and missing data. It is large with a lot of redundancy in SNP and haplotype databases. It is very dynamic and continuously evolving – both the data and the schema, which means that it requires special knowledge when design the modelling techniques. Finally but mostly importantly, SNP and haplotype data is complex with intrinsic features and subtle patterns, in the sense that it is very rich in associated complex phenotype traits or common multifactor disease.

In complex diseases, it is likely that a combination of genes predisposing for the disease and environmental factors aggravate the impact of these genes are jointly responsible for disease development in populations (known as epistasis or epistatic effects). In addition, environmental factors which seem to have only a moderate impact at the population level might have larger relative risks in subpopulations with certain genetic predispositions. There are major methodological challenges in the study of gene-gene and gene-environment interactions. The other challenge is through these high-dimensional datasets to identify combinations of interacting SNPs that are predictive of common diseases. There is a need for useful and expeditious methods for analysing massive SNP data in common complex diseases beyond that of traditional statistical approaches.

3 The promise of Computational Intelligence

The theoretical framework considered in the context of this review is Computational Intelligence (CI). CI (Pedrycz, 2000; Pedrycz and Vasilakos, 2000) is a well-established paradigm that seamlessly combines three main technologies aimed at the development of intelligent systems, named granular computing, neural networks and biologically-inspired (evolutionary) optimisation. As in the design of such systems, we have to address various challenging issues, such as knowledge representation, adaptive properties, learning abilities and structural developments. CI has to a cope with each one of them. Regarding the properties of intelligent systems being supported by the paradigm of CI, we envision two general points of view. These properties can be sought as *intrinsic* to any intelligent systems or they can be *extrinsic* to them. In the first case, we are concerned with the features that are crucial to the design of the systems, which usually do not manifest externally, so by analysing the performance of the system we cannot say whether a specific technology has been utilised. Essentially, we are not concerned about that. The extrinsic properties are dominant and become of a paramount relevance when dealing with communication of intelligent systems with others or facilitating an effective interaction with human users. This aspect is extremely relevant in providing the user a sense of intelligent, user-friendly and user-centric capabilities of the systems. Here we can stress that these capabilities are much diversified and could cover a vast territory. For instance, one can envision several interesting scenarios.

- Coping with heterogeneous information. Quite often, in intelligent systems we encounter information coming not only from sensors (in which case these are numeric readings) but also from users (in the form of linguistic evaluations) or being a result of some initial aggregation or summarisation. Interestingly, these inputs are essential to the functioning of a system and cannot be ignored or downplayed. The heterogeneity of information requires special attention in the sense of the use of more advanced mechanisms of processing and representing such a mix of various pieces of evidence.
- Establishing an effective, transparent, and customised communication with the end user when presenting the results of processing completed by a system. Here the notion of generality (abstraction) or granulation of information plays a pivotal role. A suitable level of granulation of information is essential to the effective communication and acceptance of a system (in whichever role we can envision the system to be utilised). This immediately leads us to the concept of adaptive and user-driven interfaces, which become an essence to most interactive and human centric systems including tutoring architectures, decision-support systems, and knowledge-based architectures (including expert-like systems and their more advanced topologies).

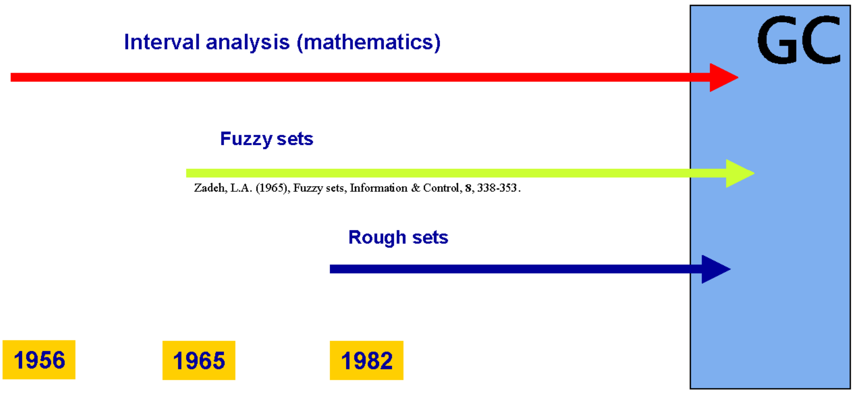
Computational intelligence is generally accepted to include evolutionary computation, fuzzy systems, neural networks, and combinations thereof. One might also extend this definition to include reaction speeds and error rates approaching human performance as an answer to Turing's comment "we may hope that machines will eventually compete with men in all purely intellectual fields" (Turing, 1956).

The term of CI being coined in the 1990s (quite commonly viewed as a synonym of soft computing) helps us establish a sound mapping between the technologies and their dominant role in meeting some specific requests of the domain. What is also very characteristic for CI today is a broad array of hybrid systems (called neuro-fuzzy systems, neuro-evolutionary systems, and genetic-fuzzy systems). They emerge as a result of an in-depth understanding of the benefits of individual technologies and their genuine complementarity.

In what follows, we briefly highlight the essence of the contributing technologies of CI, discuss their synergies and elaborate on the resulting architectures:

Granular computing. Granular information is everywhere. We granulate information all the time. We rarely reason on a basis of numbers. Our judgment is often triggered by some aggregates, which in a nutshell are a result of abstraction: a process which leads to human-like decision making. CI embraced fuzzy sets as the key vehicle of information granulation. It is worth stressing that the other fundamental environments for describing granular information are readily available and a suitable choice depends on a specific problem at hand. Figure 1 visualises the main developments in granular computing; it could help to gain a better view as to their possible linkages.

Figure 1 Main developments of granular computing (see online version for colours)

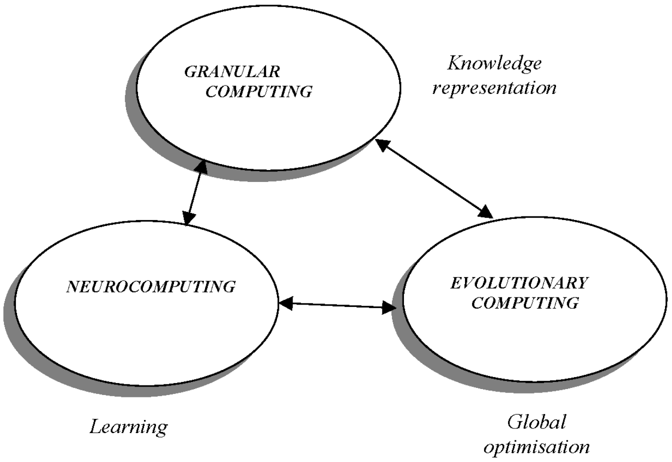

Interval analysis (mathematics)
Fuzzy sets
Zadeh, L.A. (1965), Fuzzy sets, Information & Control, 8, 338-353.
Rough sets
1956
1965
1982
GC

Neurocomputing is inherently associated with adaptive and highly flexible systems – neural networks. The learning abilities of the networks (either through supervised or unsupervised learning) are in the heart of networks. The learning is exploited when building systems that can learn from data, adapt to the nonstationary environment (including preferences of users) and help generalise to new, unknown situations. The spectrum of learning models, network architectures is impressive. Neural networks are highly distributed, which makes them fault tolerant. What has been said so far is definitely very encouraging. The drawback is the lack of transparency of the networks. The distributed character of processing can be pointed at as the most prominent reason of this deficiency. Similarly, as no prior domain knowledge could be 'downloaded' onto the network, its learning is carried out from scratch, which by itself is not the most encouraging.

Evolutionary computing. The principle of evolutionary computing cast in the setting of CI becomes a synonym of structural optimisation, reconfigurability, combinatorial optimisation, and variant selection usually completed in large and complicated search spaces. From its inception in the 1970s, evolutionary computing with all its variations of genetic algorithms, evolutionary strategies, genetic programming, etc. is aimed at the global, structural system optimisation that is carried out in presence of very limited and general information about the optimality criterion.

From the above summary, it becomes apparent that the main agendas of these technologies are different yet highly complementary leading to the scenarios in which the advantages and limitations of each one of them could be strengthened and compensated, respectively. This compensation effect is in essence a crux of the resulting synergy and helps to develop interesting and useful linkages. Figure 2 highlights the leading tendencies and identifies the ways in which the synergies have been invoked.

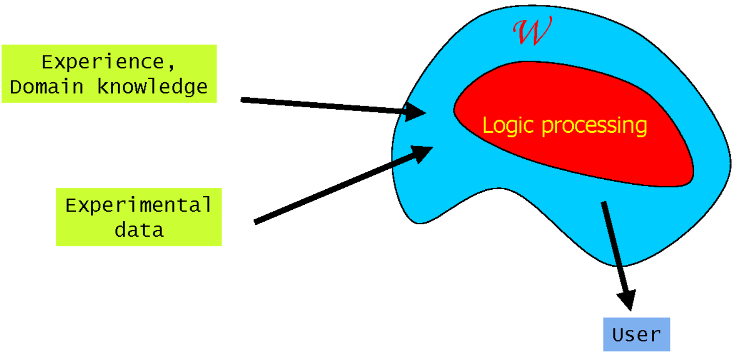
Figure 2 Main synergistic links in Computational Intelligence



As stressed, there are a significant number of possible interactions between the contributing technologies in the realm of CI. Bearing in mind the main objectives of granular computing and neural networks, we can envision a general layered type of the model in which any interaction with the external world (including users) is done through the granular interface (external layers) whereas the core computing part is implemented as a neural network or a neuro-fuzzy structure, in which case we may be emphasising the logic facet of ongoing processing faculties (see Figure 3).

Successes in computational intelligence have been forthcoming. These methods have been widely used in networking and ambient intelligence Vasilakos and Pedrycz (2006), engineering for optimisation of plant control, scheduling, for the design of small robots for locomotion and for the evolution of a human expert-level checkers player all without human expertise. As suggested at the dawn of this era "the old phrase 'the computer never knows more than the programmer' is simply no longer true". These same methods are now being applied to problems in molecular biology and bioinformatics (Fogel and Corne, 2002).

Figure 3 A layered style of CI constructs (see online version for colours)



No technological advancement has been more directly responsible for the success of molecular biology over the last 50 years than the computer. Computers have become so important in biology that it is difficult to think of any significant advancement in the last 10 years that did not have direct assistance from a computer, whether this is as a viewer for three-dimensional structures, a controller for automated robot manoeuvring of 96-well plates for PCR, a means to interpret DNA sequencing gels, microarray data, etc. The revolution of the last 50 years has resulted in such a wealth of data that our understanding of the underlying processes would be significantly reduced if computers were not at hand as our assistants with respect to 'bioinformatics'. The scale of the biological problems of interest and our understanding of those problems has closely paralleled Moore's Law. Realistically, it was not until the 1970s and 1980s when computers became truly relevant for biological information processing at a rate commensurate with the data being generated. The 1980s and 1990s heralded the internet, which has become an invaluable resource for sharing biological information. However, in parallel with molecular biology, methods of computational intelligence also share their origins in the 1950s, with refinement over time into a wide array of algorithms useful for data mining, pattern recognition, optimisation, and simulation. Today, many of these same algorithms can be said to offer 'computational intelligence' something that can handle the large size of experimental output from modern biology.

4 Computational Intelligence for SNP-disease associations

There are two major computational challenges today in genome-wide association study for complex disease: First, SNP/haplotype structure discovery, which include the selection of informative SNPs from thousands of SNPs that are associated with a disease. There are two types of approaches here

- tagSNPs, which are based on unsupervised methods with haplotype block concept and Linkage Disequilibrium (LD). These approaches intend to identify the clustering and block structure and then identify SNP blocks that are significantly associated with the disease of interest.
- Disease-associating SNPs, which are selected on clinical data using supervised approaches and therefore are block free approaches. There have been a substantial

> amount of statistical methods reported for this challenge. Second, modelling gene-gene (epistasis) and gene-environment interactions for complex diseases. Due to the complexity, there are relatively few works that are able to tackle this challenge. However, recently developed computational intelligence approaches for SNP-disease associations including genetic algorithms (Clark et al., 2005; Shah and Kusiak, 2004) neural networks (Ott, 2001; Motsinger et al., 2006a, 2006b), genetic programming (Moore and White, 2006a), evolutionary trees (Lam et al., 2000), evolutionary algorithms (Hubley et al., 2003) and various hybrid approaches, such as neural networks with genetic programming (Ritchie et al., 2003), genetic programming with multifactor dimensionality reduction (Moore and White, 2006b) and so on have demonstrated some promises, which we will summarise below.

Clark et al. (2005) developed a Genetic Algorithm (GA) to construct logic trees consisting of Boolean expressions involving blocks or strings of SNPs and applied to a candidate gene study of quantitative genetic variation. The blocks or strings of the logic trees consist of SNPs in high Linkage Disequilibrium. LD refers to the SNPs that are highly correlated with each other due to evolutionary processes. Studies showed that if high level LD occurred in the population and they are selected by the classification models, only one or two SNPs would be enough to obtain a good predictive capacity with no or only a modest reduction in power relative to direct assays of all common SNPs. In contrast, in a population, where lower levels of LD are observed at given loci, a larger number of SNPs are required to predict phenotype. Therefore, the capturing of such block structure of the gene offers the possibility of significantly reducing the number of SNPs required to completely genotype a sample with no information loss. In Clark's methods, at each generation of the GA, a population of logic trees is modified using selection, cross-over and mutation operations. Logic trees are selected for the next generation using a fitness function based on the marginal likelihood in a Bayesian regression framework. Mutation and cross-over operations use LD measures to propose changes to the trees, and facilitate the movement through the model space.

Genetic Programming (GP) is closely related to genetic algorithms. It makes use of genetic algorithms and they are a stochastic, population based, evolutionary approaches for search and optimisation. GP uses tree based strategies to represent a solution for a problem instead of a string of variables. Recent studies has shown that GP outperforms many traditional statistical, data mining and machine learning methods, such as linear regression and support vector machines. Ritchie et al. (2003) introduced a Genetic Programming Optimised Neural Network (GPNN) as a method for optimising the architecture of a neural network to improve the identification of gene combinations associated with disease risks Ritchie et al., 2003. The strength of this approach is the ability to discover the optimal NN architecture as part of the modelling process. Motsinger et al. applied a Genetic Programming Neural Network (GPNN) approach for detecting epistasis in case-control studies for SNPs data. They evaluated the power of GPNN for identifying high-order gene-gene interactions and applied GPNN to a real data analysis in Parkinson's disease (Motsinger et al., 2006a).

Motsinger et al. (2006b) developed a Grammatical Evolution Neural Network (GENN), a machine-learning approach to detect gene-gene and gene-environment interactions in high dimensional genetic epidemiological data. GENN has been shown to be highly successful in a range of simulated data. Regarding the power of GENN to detect interesting interactions in the presence of noise due to genotyping error, missing data, phenocopy, and genetic heterogeneity, they have found that the GENN method is relatively robust to all error types, including genetic heterogeneity.

Motsinger et al. further proposed an Ensemble Learning Approach for Set-association (ELAS) to detect a set of interacting loci that predicts the complex trait. In ELAS, they first search 'base-learners' and then combine the effects of the base learners. ELAS can jointly analyse single-marker effects and two way interaction effects for many markers including genome-wide association studies. Simulation studies demonstrated that ELAS is more powerful than single-marker tests. ELAS also outperformed the other three existing multi-locus methods in almost all cases. They also applied an application to a large-scale case-control study for Type 2 diabetes. ELAS identified eleven SNPs that have a significant multi-locus effect, while none of the SNPs showed significant marginal effect and none of the two-locus combinations showed significant two-locus interaction effect.

Hubley et al. (2003) presented an evolutionary algorithm for multi-objective SNP selection, which approximates the set of optimal trade-off solutions for large problems with semi-supervised learning. This set is very useful for the design of large studies, including those oriented towards disease identification, genetic mapping, population studies, and haplotype-block elucidation. They implemented a modified version of the Strength-Pareto Evolutionary Algorithm in Java and concluded that evolutionary algorithms are particularly suited for optimisation problems that involve multiple objectives and a complex search space on which exact methods, such as exhaustive enumeration cannot be applied. They provide flexibility with respect to the problem formulation if a problem description evolves or changes. Results are produced as a trade-off front, allowing the user to make informed decisions when prioritising factors. Evolutionary algorithms are well suited for many other applications in genomics.

Banzhaf et al. (2006) employed a variety of evolutionary computing methods, such as genetic algorithms in modelling epistasis. Haplotype fine mapping by evolutionary trees describe a method that seeks to refine location by analysis of 'disease' and 'normal' haplotypes, thereby using multivariate information about linkage disequilibrium. Under the assumption that the disease mutation occurs in a specific gap between adjacent markers, the method first combines parsimony and likelihood to build an evolutionary tree of disease haplotypes, with each node (haplotype) separated by a single mutational or recombinational step (from its parent). If required, latent nodes (unobserved haplotypes) are incorporated to complete the tree. Once the tree is built, its likelihood is computed from probabilities of mutation and recombination. When each gap between adjacent markers is evaluated in this fashion and these results are combined with prior information, they yield a posterior probability distribution to guide the search for the disease mutation. They show, by evolutionary simulations, that an implementation of these methods, called 'FineMap', yields substantial refinement and excellent coverage for the true location of the disease mutation.

The detection of epistasis is an important priority in the genetic analysis of complex human diseases. The most challenging epistatic effects to model are those that do not exhibit statistically significant marginal effects. Identifying these types of nonlinear interactions in the context of genome-wide association studies is considered a needle in a haystack problem (Moore and Hahn, 2002). Given this complexity, it is unrealistic to expect that stochastic search algorithms will do any better than a simple random search. Hybrid computational intelligence approaches have been investigated for finding epistatic needles with the assistance of expert knowledge (Ritchie et al., 2003, Moore and Hahn, 2002).

Ritchie et al. (2003) proposed a neural network with genetic programming for studying SNP-SNP interaction and used expert knowledge to guide the genome-wide analysis of epistasis using stochastic genetic programming. They first developed and evaluated a genetic programming approach with one-objective fitness function with classification accuracy to SNP selection and classification and showed that GP is no better than a simple random search when classification accuracy s used as the fitness function. Then they further included pre-processed estimates of attribute quality (i.e., expert knowledge) using Tuned ReliefF (TuRF) in a multi-objective fitness function, which significantly improved the performance of GP over that of random search. Results showed that using expert knowledge to select trees performs as well as a multi-objective fitness function and can not only improve the accuracy of prediction, but also can achieve the same power at one tenth of the population size.

Moore and White (2006b) developed a hybrid genetic programming with Multifactor Dimensionality Reduction (MDR) to pick SNPs for epistasis. They also found no evidence to suggest that GP-MDR performed better than random search on the simulated genome-wide data sets. They further modified GP-MDR to select SNP combinations for virtual recombination, mutation, and reproduction by using ReliefF filter algorithm. ReliefF filter provided statistical measures and prior information about the quality of each SNP as assessed during a pre-processing analysis, which the GP-MDR was based on. They found that the expert knowledge provided by ReliefF about which SNPs might be interacting significantly improved the ability of GP-MDR to identify epistatic SNPs in the absence of marginal effects over that of a simple random search. An important advantage of this hybrid approach is that any form of expert knowledge could be used to guide the stochastic search algorithm. For example, information about biochemical pathways, protein-protein interactions, Gene Ontology (GO), or even evidence from the literature could be used in addition to statistical measures.

Moore and Hahn introduced Cellular Automata (CA) as a novel computational approach for identifying combinations of Single-Nucleotide Polymorphisms (SNPs) associated with clinical endpoints (Moore and Hahn, 2002). This alternative approach is nonparametric (i.e., no hypothesis about the value of a statistical parameter is made), model-free (i.e., assumes no particular inheritance model), and is directly applicable to case-control and discordant sib-pair study designs. Using simulated data, they demonstrated that the approach has good power for identifying high-order nonlinear interactions (i.e., epistasis) among four SNPs in the absence of independent main effects.

Recently, Eppstin et al. (2007) described a random chemistry approach for detecting epistasis in genome-wide association studies and proposed a new evolutionary approach that attempts to hill-climb from large sets of candidate epistatic genetic features to smaller sets, inspired by Kauffman's 'random chemistry' approach. As the authors pointed out although the algorithm is conceptually straightforward, its success hinges upon the creation of a fitness function to discriminate large sets that contain subsets of interacting genetic features from those that don't. They employed an approximate and noisy fitness function based on the ReliefF data mining algorithm and applied to synthetic data sets with individual features having no marginal effects. Results show that the algorithm can successfully detect epistatic pairs from up to 1000 candidate SNPs. However, more accurate fitness approximator for large data sets with lower heritabilities.

5 Discussions

Designing, developing and implementing computational intelligence methods for identifying genetic triggers and components responsible for complex diseases, such as diabetes, cancer, cardiovascular disease, etc. is one of the new and challenging areas in human genetics and bioinformatics. Computational intelligence is a side branch of artificial intelligence, where well-crafted algorithms are being developed that solve complex, computationally expensive problems that are believed to require intelligence. Computational Intelligence is one of the most promising tools today to attack the remaining hard problems in bioinformatics and human genetics. This review covered some theories and applications of computational intelligence for SNP-disease association study. We demonstrated the promises and the importance of computational intelligence for today's common complex diseases with SNP-haplotype data, especially focusing on gene-gene and gene-environment interactions and the notorious 'curse of dimensionalit' problem. Success in identifying SNPs and haplotypes conferring susceptibility or resistance to common diseases will provide a deeper understanding of the architecture of the disease, the risks and offer a more powerful diagnostic tool and predictive treatment.

References

Akey, J., Jin, L. and Xiong, M. (2001) 'Haplotypes vs. single marker linkage disequilibrium tests: What do we gain?', *Eur. J. Hum. Genet*., Vol. 9, No. 4, pp.291–300.

Anderson, E.C. and Novembre, J. (2003) 'Finding haplotype block boundaries by using the minimum-description-length principle', *American Journal of Human Genetics*, Vol. 73, pp.336–354.

Ao, S., Yip, K., Ng, M., Cheung, D., Fong, P.Y., Melhado, I. and Sham, P.C. (2005) 'CLUSTAG: hierarchical clustering and graph methods for selecting tag SNPs', *Bioinformatics*, Vol. 21, No. 8, pp.1735–1736.

Azevedo, L., Suriano, G., van Asch, B., Harding, R.M. and Amorim, A. (2006) 'Epistatic interactions: How strong in disease and evolution?', *Trends Genet*., 12 August.

Baker, S.G. (2005) 'A simple loglinear model for haplotype effects in a case-control study involving two unphased genotypes', *Statistical Applications in Genetics and Molecular Biology*, Vol. 4, No. 1, p.14.

Banzhaf, W., Beslon, G., Christensen, S., Foster, J.A., Kepes, F., Lefort, V., Miller, J.F., Radman, M. and Ramsden, J.J. (2006) 'Guidelines: from artificial evolution to computational evolution: a research agenda', *Nat. Rev. Genet*., Vol. 7, No. 9, September, pp.729–735.

Beckmann, L., Thomas, D.C., Fischer, C. and Chang-Claude, J. (2005) 'Haplotype sharing analysis using Mantel statistics', *Human Heredity*, Vol. 59, pp.67–78.

Brookes, A.J. (1999) 'Review: the essence of SNPs', *Gene*, Vol. 234, pp.177–186.

Burkett, K., McNeney, B. and Graham, J. (2004) 'A note on inference of trait associations with SNP haplotypes and other attributes in generalized linear models', *Human Heredity*, Vol. 57, pp. 200–206.

Cardon, L.R. and Bell, J.I. (2001) 'Association study designs for complex diseases', *Nat. Rev. Genet*., Vol. 2, pp.91–99.

Chang, C., Huang, Y. and Chao, K. (2006) 'A greedier approach for finding tag SNPs', *Bioinformatics*, Vol. 22, No. 6, pp.685–691.

Chapman, J.M., Cooper, J.D., Todd, J.A. and Clayton, D.G. (2003) 'Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power', *Hum. Hered.*, Vol. 56, pp.18–31.

Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U. and Wacholder, S. (2006) 'Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions', *American Journal of Human Genetics*, Vol. 79, No. 6, pp.1002–1016.

Clark, T.G., De Iorio, M., Griffiths, R.C. and Farrall, M. (2005) 'Finding associations in dense genetic maps: a genetic algorithm approach', *Human Heredity*, Vol. 60, pp.97–108.

Collins, F., Brooks, L. and Chakravarti, A. (1998) 'A DNA polymorphism discovery resource for research on human genetic variation', *Genome Research*, Vol. 8, No. 12, pp.1229–1231.

Cordell, H.J., Barratt, B.J. and Clayton, D.G. (2004) 'Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects', *Genetic Epidemiology*, Vol. 26, No. 3, pp.167–185.

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., Lander, E.S., Phillips, M.S., Lawrence, R., Sachidanandam, R., Morris, A.P., Balding, D.J., Donaldson, M.A., Studebaker, J.F., Ankener, W.M., Alfisi, S.V., Kuo, F-S., Camisa, A.L., Pazorov, V., Scott, K.E., Carey, B.J., Faith, J., Katari, G., Bhatti, H.A., Cyr, J.M., Derohannessian, V., Elosua, C., Forman, A.M., Greeco, N.M., Hock, C.R. Kuebler, J.M., Lathrop, J.A., Mockler, M.A., Nachtman, E.P., Restine, S.L., Varde, S.A., Hozza, M.J., Gelfand, C.A., Broxholme, J., Abecasis, G.R., Boyce-Jacino, M.T. and Cardon, L.R. (2003) 'Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots', *Nature. Genet.*, Vol. 33, pp.382–387.

Durrant, C., Zondervan, K.T, Lon, R., Cardon, L., Hunt, S., Deloukas, P. and Morris, A.P. (2004) 'Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes', *Am. J. Hum. Genet.*, Vol. 75, pp.35–43.

Eppstein, M.J., Payne, J.L., White, B.C. and Moore, J.H. (2007) 'Genomic mining for complex disease traits with 'random chemistry'', *Genetic Programming and Evolvable Machines*, Special Issue on Medical Applications, Vol. 8, No. 4, pp.395–411.

Fogel, G.B. and Corne, D.W. (2002) *Evolutionary Computation in Bioinformatics*, Morgan Kaufmann, San Francisco.

Foster, J.A. (2001) 'Evolutionary computation', *Nat. Rev. Genet.*, Vol. 2, No. 6, pp.428–36.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J. and Altshuler, D. (2002) 'The structure of haplotype blocks in the human genome', *Science*, Vol. 296, No. 5576, pp.2225–2229.

Greenspan, G. and Geiger, D. (2004) 'Model-based inference of haplotype block variation', *J. Comp. Biol.*, Vol. 11, pp. 493–504.

Greenspan, G. and Geiger, D. (2006) 'Modeling haplotype block variation using Markov chains', *Genetics*, Vol. 172, No. 4, pp.2583–2599.

Halldrsson, B.V., Bafna, V., Lippert, R., Schwartz, R., De La Vega, F.M., Clark, A.G. and Istraili, S. (2004) 'Optimal Haplotype Block-Free Selection of Tagging SNPs for Genome', *Wide Association Studies Genome Res.*, Vol. 14, pp.1633–1640.

Hampe, J., Schreiber, S. and Krawczak, M. (2003) 'Entropy-based SNP selection for genetic association studies', *Hum. Genet.*, Vol. 114, pp.36–43.

He, J. and Zelikovsky, A. (2006) 'MLR-tagging informative SNP selection for unphased genotypes based on multiple linear regression', *Bioinformatics*, Vol. 22, No. 20, pp.2558–2561.

Horne, B.D. and Camp, N.J. (2004) 'Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation', *Genetic Epidemiology*, Vol. 26, No. 1, pp.11–21.

Howie, B.N., Carlson, C.S., Rieder, M.J. and Nickerson, D.A. (2006) 'Efficient selection of tagging single-nucleotide polymorphisms in multiple populations', *Human Genetics*, Vol. 120, No. 1, pp.58–68.

Hubley, R.M., Zitzler, E. and Roach, J.C. (2003) 'Evolutionary algorithms for the selection of single nucleotide polymorphisms', *BMC Bioinformatics*, Vol. 4, No. 30, pp.30–39.

Hunter, D.J. (2005) 'Gene-environment interactions in human diseases', *Nature Reviews Genetics*, Vol. 6, pp. 287–298.

Ioannidis, J.P., Gwinn, M., Little, J., Higgins, J.P., Bernstein, J.L., Boffetta, P., Bondy, M., Bray, M.S., Brenchley, P.E., Buffler, P.A., Casas, J.P., Chokkalingam, A., Danesh, J., Smith, G.D., Dolan, S., Duncan, R., Gruis, N.A., Hartge, P., Hashibe, M., Hunter, D.J., Jarvelin, M.R., Malmer, B., Maraganore, D.M., Newton-Bishop, J.A., O'Brien, T.R., Petersen, G., Riboli, E., Salanti, G., Seminara, D., Smeeth, L., Taioli, E., Timpson, N., Uitterlinden, A.G., Vineis, P., Wareham, N., Winn, D.M., Zimmern, R. and Khoury, M.J. (2006) 'Human genome epidemiology network and the network of investigator networks, A road map for efficient and reliable human genome epidemiology', *Nature Genetics*, Vol. 38, No. 1, pp.3–5.

Jasnos, L. and Korona, R. (2007) 'Epistatic buffering of fitness loss in yeast double deletion strains', *Nat. Genet.*, Vol. 39, No. 4, pp.550–554.

Ke, X. and Cardon, L.R. (2003) 'Efficient selective screening of haplotypes tag SNPs', *Bioinformatics*, Vol. 19, pp.287, 288.

Kooperberg, C., Ruczinski, I., LeBlanc, M. and Hsu, L. (2001) 'sequence analysis using logic regression', *Genetic Epidemiology*, Vol. 21, pp.626–631.

Lam, J.C., Roeder, K. and Devlin, B. (2000) 'Haplotype fine mapping by evolutionary trees', *Am. J. Hum. Genet.*, Vol. 66, No. 2, pp.659–673.

Levin, A.M., Ghosh, D., Cho, K.R. and Kardia, S.L.R. (2005) 'A model-based scan statistics for identifying extreme chromosomal regions of gene expression in human tumors', *Bioinformatics*, Vol. 21, pp.2867–2874.

Li, J. and Jiang, T. (2005) 'Haplotype-based linkage disequilibrium mapping via direct data mining', *Bioinformatics*, Vol. 21, pp.4384–4393.

Lin, Z. and Altman, R.B. (2004) 'Finding haplotype tagging SNPs by use of principal components analysis', *Am. J. Hum. Genet.*, Vol. 75, pp.850–861.

Liu, J.S., Sabatti, C., Teng, J., Keats, B.J. and Risch, N. (2001) 'Bayesian analysis of haplotypes for linkage disequilibrium mapping', *Genome Research*, Vol. 11, No. 10, pp.1716–1724.

Liu, Z. and Lin, S. (2005) 'Multilocus LD measure and tagging SNP selection with generalized mutual information', *Genet. Epidemiol.*, Vol. 29, pp.353–364.

Mannila, H., Koivisto, M., Perola, M., Varilo, T., Hennah, W., Ekelund, J., Lukk, M., Peltonen, L. and Ukkonen, E. (2003) 'Minimum description length block finder, a method to identify haplotype blocks and to compare the strength of block boundaries', *Am. J. Hum. Genet*, Vol. 73, pp.86–94.

Martin, G., Elena, S.F. and Lenormand T. (2007) 'Distributions of epistasis in microbes fit predictions from a fitness landscape model', *Nat. Genet.*, Vol. 39, No. 4, pp.555–560.

McKinney, B.A., Reif, D.M., Ritchie, M.D. and Moore, J.H. (2006) 'Machine learning for detecting gene-gene interactions', *Applied Bioinformatics*, Vol. 5, pp.77–88.

Meng, Z., Zaykin, D.V., Xu, C.F., Wagner, M. and Ehm, M.G. (2004) 'Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes', *Am. J. Hum. Genet.*, Vol. 73, pp.115–130.

Molitor, J., Marjoram, P. and Thomas, D. (2003) 'Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques', *Am. J. Hum. Genet.*, Vol. 73, pp.1368–1384.

Molitor, J., Marjoram, P., Conti, D. and Thomas, D. (2004) 'A survey of current bayesian gene mapping methods', *Human Genomics*, Vol. 1, No. 5, pp.371–374 (4).

Moore, J.H. (2003) 'The ubiquitous nature of epistasis in determining susceptibility to common human diseases', *Hum. Hered.*, Vol. 56, Nos. 1–3, pp.73–82.

Moore, J.H. (2007) 'Genome-wide analysis of epistasis using multifactor dimensionality reduction: feature selection and construction in the domain of human genetics', in Zhu, X. and Davidson, I. (Eds.): *Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data*, IGI, pp.17–30.

Moore, J.H. and Hahn, L.W. (2002) 'A cellular automata approach to detecting interactions among single-nucleotide polymorphisms in complex multifactorial diseases', *Pac. Symp. Biocomput.*, pp.53–64.

Moore, J.H. and White, B.C. (2006a) 'Exploiting expert knowledge for genome-wide genetic analysis using genetic programming', in Runarsson, T.P., Beyer, H-G., Burke, E., Merelo-Guervos, J.J., Whitley, L.D. and Yao, X. (Eds.): *Parallel Problem Solving from Nature – PPSN IX*, Lecture Notes in Computer Science 4193, pp.969–977.

Moore, J.H. and White, B.C. (2006b) 'Detecting epistatic needles in genome-wide haystacks', *American Human Genetics Conferences*, New Orleans, Louisiana.

Motsinger, A.A., Lee, S.L., Mellick, G. and Ritchie, M.D. (2006a) 'GPNN: power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease', *BMC Bioinformatics 25*, Vol. 7, No. 1, p.39.

Motsinger, A.A., Fanelli, T.J. and Ritchie, M.D. (2006b) 'Power of grammatical evolution neural networks to detect gene-gene interactions in the presence of error common to genetic epidemiological studies', *International Genetic Epidemiology Society 15th Annual Meeting*, Tampa Bay, Florida.

Nothnagel, M., Furst, R. and Rohde, K. (2002) 'Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks', *Hum. Hered.*, Vol. 54, pp.186–198.

Onkamo, P. and Toivonen, H. (2006) 'A survey of data mining methods for linkage disequilibrium mapping', *Human Genomics*, Vol. 2, No. 1, pp.336–340 (5).

Ott, J. (2001) 'Neural networks and disease association studies', *American Journal of Medical Genetics*, Vol. 105, No. 1, pp. 60–61.

Ott, J. (2004) 'Issues in association analysis: error control in case-control association studies for disease gene discovery', *Human Heredity*, Vol. 58, pp.171–174.

Pedrycz, W. (1997) *Computational Intelligence: An Introduction*, CRC, Boca Raton, FL.

Pedrycz, W. and Vasilakos, A. (2000) *Computational Intelligence in Telecommunications Networks*, CRC, Boca Raton, FL.

Risch, N. and Merikangas, K. (1996) 'The future of genetics studies of complex human diseases', *Science*, Vol. 273, pp.1516–1517.

Risch, N.J. (2000) 'Searching for genetic determinants in the new millennium', *Nature*, Vol. 405, pp.847–856.

Ritchie, M.D., White, B.C., Parker, J.S., Hahn, L.W. and Moore, J.H. (2003) 'Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases', *BMC Bioinformatics*, Vol. 4, No. 28.

Salem, R.M., Wessel, J. and Schork, N.J. (2005) 'A comprehensive literature review of haplotyping software and methods for use with unrelated individuals', *Human Genomics*, Vol. 2, No. 1, pp.28, 39–66.

Schaid, D.J., Rowland, C.M., Tines, D.E., Jacobson, R.M. and Poland, G.A. (2002) 'Score test for association between traits and haplotypes when linkage phase is ambiguous', *Am. J. Hum. Genet.*, Vol. 70, pp.425–443.

Schwender, H. and Ickstadt, K. (2006) *Identification of SNP Interactions using Logic Regression*, http://www.sfb475.uni-dortmund.de/berichte/tr31-06.pdf, accessed on October-31-2006.

Shah, S.C. and Kusiak, A. (2004) 'Data mining and genetic algorithm based gene/SNP selection', *Artif. Intell. Med.*, Vol. 31, No. 3, pp.183–196.

Stephens, M. and Donnelly, P. (2000) 'Inference in molecular population genetics', *J. R. Stat. Soc. B.*, Vol. 62, pp.605–655.

The International HapMap Consortium (2003) 'The International HapMap Project', *Nature*, Vol. 426, pp.789–796.

The International HapMap Consortium (2004) 'Integrating ethics and science in the International HapMap Project', *Nat. Rev. Genet.*, Vol. 5, pp.467–475.

The International HapMap Consortium (2005) 'A haplotype map of the human genome', *Nature*, Vol. 437, pp.1299–1320.

Thomas, D.C., Stram, D.O., Conti, D., Molitor, J. and Marjoram, P. (2003) 'Bayesian spatial modeling of haplotype associations', *Human Heredity*, Vol. 56, pp. 32–40.

Toivonen, H.T., Onkamo, P., Vasko, K., Ollikainen, V., Sevon, P., Mannila, H., Herr, M. and Kere, J. (2000) 'Data mining applied to linkage disequilibrium mapping', *Am. J. Hum. Genet.*, Vol. 67, No. 1, pp.133–145.

Turing, A.M. (1956) 'Can a machine think?', in Newman, J.R. (Ed.): *The World of Mathematics*, Simon and Schuster, New York, Vol. 4, p.2122.

Tzeng, J., Wang, C., Kao, J. and Hsiao, C.K. (2006) 'Regression-based association analysis with clustered haplotypes through use of genotypes', *American Journal of Human Genetics*, Vol. 78, No. 2, pp.231–242.

Vasilakos, A. and Pedrycz, W. (2006) *Ambient Intelligence, Wireless Networking, Ubiquitous Computing*, ArtecHouse, MA, USA.

Verzilli, C.J., Stallard, N. and Whittaker, J.C. (2006) 'Bayesian graphical models for genomewide association studies', *American Journal of Human Genetics*, Vol. 79, No. 1, pp.100–112.

Zaykin, D.V., Zhivotovsky, L.A., Westfall, P.H. and Weir, B.S. (2002) 'Truncated product method for combining P-values', *Genet Epidemiol*, Vol. 22, pp.170–185.

Zhang, K., Calabrese, P., Nordborg, M. and Sun, F. (2002a) 'Haplotype block structure and its applications to association studies: power and study designs', *Am. J. Hum. Genet.*, Vol. 71, pp.1386–1394.

Zhang, K., Deng, M., Chen, T., Waterman, M.S. and Sun, F. (2002b) 'A dynamic programming algorithm for haplotype block partitioning', *Proc. Natl. Acad. Sci. USA*, Vol. 99, pp.7335–7339.

Zhao, J., Boerwinkle, E. and Xiong, M. (2005) 'An entropy-based statistic for genomewide association studies', *American Journal of Human Genetics*, Vol. 77, pp.27–40.

Zondervan, K.T. and Cardon, L.R. (2004) 'The complex interplay among factors that influence allelic association', *Nature Reviews Genetics*, Vol. 5, No. 2, pp.89–100.