J Hum Genet (2006) 51:751–759
DOI 10.1007/s10038-006-0016-2

ORIGINAL ARTICLE

An algorithm for model construction and its applications to pharmacogenomic studies

Kung-Hao Liang · Yuchi Hwang · Wan-Ching Shao · Ellson Y. Chen

Received: 22 March 2006 / Accepted: 12 May 2006 / Published online: 10 August 2006

Abstract A model depicts the relationship between clinical phenotypes and genotypes on a set of genetic polymorphisms. After the model is constructed and validated, it may be used to predict clinical phenotypes such as traits of complex diseases. A pharmacogenomic model is used to predict the efficacies or adverse drug reactions of a medication. The construction of a model is a challenging task. This is because a single-locus polymorphism does not contain enough information to stratify patients in general, given the complex biological mechanisms involved. An exhaustive search for the correct combination of genotypes across multiple loci is, however, computationally infeasible. We are, thus, motivated to propose a novel algorithm for the construction of models using the multiple single-nucleotide polymorphism (SNP) information in diplotype forms. This algorithm utilizes the techniques of genetic algorithms and Boolean algebra (GABA). The proposed algorithm is tested on simulated data, as well as real genotype datasets of chronic hepatitis C patients treated with interferon-combined therapy. A model for predicting the treatment efficacy is constructed and validated. The results showed that the proposed algorithm is very effective in deriving models comprising multiple SNPs.

K.-H. Liang (✉) · Y. Hwang · E. Y. Chen
Vita Genomics, Inc., 7F, No. 6, Sec. 1, Jungshing Rd.,
Wugu Shiang, Taipei 248, Taiwan
e-mail: kunghao.liang@vitagenomics.com

W.-C. Shao
National Taiwan Normal University, Taipei, Taiwan

Keywords Prediction model · Epistasis · Multiple SNP · Boolean algebra · Complex disease · Pharmacogenomics

Introduction

Polymorphisms of the human genome are responsible for the causations of many genetic-linked phenotypes, including the traits of complex diseases, as well as the efficacies of medications addressed in pharmacogenomic studies. A model depicts the association between clinical phenotypes and multiple genetic information, such as single-nucleotide polymorphisms (SNPs) in either the haplotype or diplotype forms, or short tandem repeats (STRs) (Cordell and Clayton 2002; Yang et al. 2003). A model may even include physical information (such as the age, weight, diet, life style, and state of health) and clinical information (such as biochemical measurements, or the viral type for viral infection diseases). Once the associations between the clinical endpoints and the multi-locus genotypes are found and validated, the model could serve as the basis of new clinical prediction or prognostic methods. This paper presents a methodology for constructing models for a case–control study by using multiple SNP information in diplotype forms.

The rapid advance of genotyping techniques, exemplified by the recent Affymetrix GeneChip Human Mapping 500K Array, enables association studies with extensive genes and SNPs (Rabbee and Speed 2006). The immediate hurdle is the lack of an adequate strategy for multi-locus association and model construction from the vast amount of data. A case–control study is generally used for association studies, where the cases and controls refer to two distinct clinical groups (Cordell and Clayton 2002). Statistical tests based on the contingency tables (e.g., the χ^2 test) are commonly applied to the examination of associations for each screened diplotype polymorphisms (Pritchard and Rosenberg 1999). The polymorphisms with the *p* values smaller than a predefined threshold (commonly set at 0.05 and then adjusted with respect to multiple-comparison considerations) are declared significant in association, implying that the polymorphism is statistically associated to, or even functionally responsible for, a particular trait of interest.

Statistical tests of association are adequate particularly for single-gene diseases. There are many such diseases in the Online Mendelian Inheritance in Man database (OMIM 2000). However, the prediction of common multifactorial diseases is greatly improved by considering multiple alleles concurrently (Yang et al. 2003). Most pharmacogenomic studies also require an adequate combination of multi-loci information, given the complex pharmacokinetic and pharmacodynamic mechanisms involved. Gene–gene interactions need to be considered; therefore, a prediction model needs to characterize the complex roles of genes which lead to the phenotype.

Hence, we propose an algorithm which evaluates a set of SNPs simultaneously for model construction. A model is represented by a Boolean expression, facilitating biological interpretations. The amount of prior assumptions underlying the model is minimized. For example, the number of SNPs in the model is automatically determined by the algorithm, based on the dataset. A variety of models can be presented in Boolean expressions which reflect various forms of gene–gene interactions.

The search space for an adequate model is linearly proportional to the number of samples. It is, however, exponentially proportional to the number of screened SNPs when all of the combinations of SNPs need to be enumerated and calculated. Hence, an exhaustive search is prohibited, even for studies on hundreds of SNPs, let alone the whole-genome screening studies. To address this, the genetic algorithm is employed to systematically explore the vast choices of models described by Boolean expressions. The proposed algorithm is, thus, referred to as the genetic algorithm with Boolean algebra (GABA).

The GABA algorithm

Boolean algebra

Boolean algebra is a bivalent algebraic system (i.e., false and true, commonly shown as 0 and 1, respectively). It is used as the mathematical framework for representing the model. The defined operations of Boolean algebra include addition (+), multiplication (×), and negation (–) (Whitesitt 1995). They correspond to the operations of union, intersection, and complement in the set theory, respectively. The addition operation is also equivalent to the logical operation 'OR' and the multiplication operation to 'AND'. Boolean algebra has several basic algebraic properties, such as the commutative and associative laws for addition and multiplication, and so on (Whitesitt 1995).

A model, denoted as *M*, comprises a chain of polymorphisms joined together by Boolean operators. To construct a model, a training dataset *T* containing genotypes of cases and controls is required. *T* is presented as a master table, a two-dimensional table, with each rank representing a sample of a subject (person), and each column representing an SNP. Denote *l* as the number of screened SNPs, therefore, $T=\{\mathrm{SNP}_k|0\leq k<l\}$. The genotype dataset are commonly arranged in a consecutive order according to their chromosomal position. The assessment of an SNP is defined by a model element (m_i). On a typical biallelic SNP, a mode element could, for example, represent either a recessive mode of inheritance:

$$m_i:\ \mathrm{SNP}_k =' \mathrm{AA}' \tag{1}$$

or a dominant mode of inheritance:

$$m_i:\ \mathrm{SNP}_k =' \mathrm{AA} + \mathrm{AT}' \tag{2}$$

The intersection of the model element (m_i) and its complement element ($–m_i$) is an empty set. Their union is the set containing all possible diplotypes in the SNP, for example, {AA, AT, TT} at a biallelic A/T locus.

Hence, a model can be denoted succinctly as $M(m_i|1\leq i\leq n,\ n\leq l)$, where *n* is the number of SNPs in the model. The result of the assessment on an SNP locus could be true or false. The model elements are then joined together by either a multiplicative (×) or an additive (+) operator of Boolean algebra. A possible biological interpretation of the multiplicative (×) operator is that, for example, the concurrent appearance of two diplotypes activates a particular biological pathway. A possible biological interpretation for the additive (+) operator is that, for example, the mutations on any of the two SNPs in the same gene, joined by the additive operator, could result in the malfunction of this gene. It also represents the situation when the two SNPs reside in two genes of the same biological pathway, and the malfunction of either gene inactivates the pathway.

Given the genotypes of a particular subject, the computational result of M is either 0 or 1, indicating a control or a case, respectively. One negation operator (–) could be positioned at the starting position of M, converting those originally predicted cases to controls, and vice versa.

Hence, a legitimate model with four elements is exemplified as follows:

$$\begin{aligned} M(m_1,\ m_2,\ m_3,\ m_4) &= m_1 \times m_2 \times m_3 + m_4 \\ &= (\mathrm{SNP}_3 =' \mathrm{AA} + \mathrm{AT}') \times (\mathrm{SNP}_5 =' \mathrm{CC}') \\ &\quad \times (\mathrm{SNP}_7 =' \mathrm{CC}') + (\mathrm{SNP}_8 =' \mathrm{TC} + \mathrm{TT}') \end{aligned} \tag{3}$$

M is a case model that, if the computational result of M is true, then the subject is predicted as a case; otherwise, it is a control. The complement of M, denoted as M^C, is a control model. According to DeMorgan's law in Boolean algebra (Whitesitt 1995), M^C can be exemplified as:

$$\begin{aligned} &M^C((-m_1),\ (-m_2),\ (-m_3),\ (-m_4)) \\ &\quad = ((-m_1) + (-m_2) + (-m_3)) \times (-m_4) \\ &\quad = \big((\mathrm{SNP}_3 = '\mathrm{TT}') + (\mathrm{SNP}_5 = '\mathrm{AA} + \mathrm{AC}') \\ &\qquad + (\mathrm{SNP}_7 = '\mathrm{GG} + \mathrm{GC}')\big) \times (\mathrm{SNP}_8 = '\mathrm{CC}') \end{aligned} \tag{4}$$

Genetic algorithm

The genetic algorithm is a modern heuristic method for solving combinatorial optimization tasks (Holland 1998; Goldberg 1989). The task of model construction may be formulated as:

$$M_{\mathrm{opt}} = \arg\max_{M} F(T) \tag{5}$$

where F is the fitness score reflecting the prediction performance of the model M on the dataset T. We denote R as the case population and R^C as the control population; thus, T=R+R^C. The sensitivity of a model M on T is Pr(M|R), i.e., the probability of M being true within the case population, and the specificity is Pr(M^C|R^C), i.e., the probability of M^C being true (M being false) within the control population. In this paper, we defined F as:

$$F = \mathrm{Sensitivity} + \mathrm{Specificity} + \mathrm{Sensitivity} \times \mathrm{Specificity} \tag{6}$$

The sensitivity and specificity are used in the fitness function because the purpose of a model is mainly for clinical prediction, where the sensitivity and specificity are two important and commonly used indexes of performance. Different clinical applications may require different weightings on the sensitivity and specificity. Although positive predictive values (Pr(R|M)) and negative predictive values (Pr(R^C|M^C)) are also clinically important, they are variable with respect to the ratio of the numbers of cases and controls and, therefore, are not used. The last term of the fitness score, i.e., 'Sensitivity×Specificity', enable the algorithm to select favorably those models with high values in both sensitivity and specificity. The optimum fitness score in this definition is three.

The fitness score F is a heuristic parameter which should be defined according to the purpose of each clinical study. The likelihood ratio (LR(M)) has been used for measuring the performance of diagnostic testing (Yang et al. 2003). It is also a function of the sensitivity and specificity:

$$\mathrm{LR}(M) = \frac{\Pr(M|R)}{\Pr(M|R^C)} = \frac{\mathrm{Sensitivity}}{1 - \mathrm{Specificity}} \tag{7}$$

It is worth noting that Akaike's information criteria (AIC) has been widely used for various model selection tasks, including biological studies and spectral analysis (Gardner 1988). AIC is solidly based on information theory, that it measures how good the model approximates the data using the likelihood function (Burnham and Anderson 2001). It also penalizes the increase of model length, i.e., the number of SNPs and their interaction terms, based on the principle of parsimony. AIC is also an adequate choice of F, where F can be defined as 1/AIC.

A parsimony constraint may be introduced to F so that a model with a smaller number of SNPs will be preferred. The parsimony constraint is to avoid overfitting of the model to the data. However, we found in our simulation that the GABA algorithm can find the model with an adequate number of SNPs automatically, without the parsimony constraint (see section on Performance evaluation, where no parsimony constraint is used). We observed that the addition of extra SNPs to the optimum model will result in poorer performance. Hence, the parsimony constraint is not adopted in F.

A random model generator is required to initiate the computation. To generate a random model, the number of n, $n \le l$, is first randomly determined. Then, a series of SNP_k, $k \le l$, are randomly chosen for model elements m_i, i={1,..., n}. Each SNP_k has four possible diplotypes in our implementation, each corresponding to various dominant and recessive modes of inheritance. For example, if SNP_k is a 'C/G' allele, then SNP_k

will be assigned as ‘CC’, ‘CC+CG’, ‘GG’, or ‘GG+CG’ with equal opportunities for m_i. The additive (+) and multiplicative (×) Boolean operators are then randomly chosen between the model elements. Finally, a negation (–) operator is randomly determined whether or not to appear in front of the entire statement.

The GABA algorithm employs mutation and cross-over operations for altering an existing model.

Mutation operations

Five different types of mutations are employed in the GABA algorithm: (1) element insertion, (2) element deletion (3) element substitution, (4) operators ×/+ swap, and (5) case/control swap. The element insertion operation introduces a new random element into the model, increasing the model length by 1. The element deletion operation removes an element from the model. The element substitution operation changes the specified genotypes in a model element, for example, from SNP_k=‘CC’ to ‘CC+CG’. The operators ×/+ swap converts a multiplication (×) into an addition (+), or vice versa. This operation changes the nonlinear relationship among the model elements. For example, if this operation modifies the model $M=m_1+m_2\times m_3\times m_4$ as $m_1+m_2+m_3\times m_4$, then the relationship between the elements is changed. Finally, the case/control swap introduces a negation operator in front of the model. If there is already a negation operator, then this operation effectively removes the original negation operator.

A mutation rate (p) is required for a mutation operation, where p percent of the model elements are mutated and (1–p) percent are not mutated. Those mutated elements are subject to one of the four mutation methods (1)–(4) with equal opportunity. In addition, the entire model is subject to mutation (5) with 50% probability.

Cross-over operations

The cross-over operation is analogous to the chromosomal recombination events occurring in meioses of cell cycles. Note that chromosomal recombination is a basic concept underlying the discipline of statistical genetics, particularly in linkage analysis, association, as well as haplotype analysis (Cardon and Bell 2001; Schaid 2004). The rationale for the cross-over operation is that, if the good performances of two models are mainly due to parts of themselves, then a cross-over operation may combine these two parts, resulting in a model which outperforms the previous two models.

Using the defined operations of the GABA algorithm, the models with higher fitness scores are randomly mutated and crossed over with one another so as to produce various candidate models, exploring the entire solution space in a systematic manner. Each of these models is used to predict the samples in the training dataset. The prediction performances of the models are then evaluated by their fitness scores. Models and their elements with higher fitness scores are preserved and also serve as the templates for constructing the models in the next iteration. In this way, a group of modes go through a nature selection process. As candidate models are derived from well performing models in the previous iteration, they comprise advantageous components inherited from their parents. This type of algorithm has been demonstrated to avoid prolonged search according to the theory of schemata (Holland 1998).

The algorithm

The GABA algorithm is designed to select n SNPs, from a pool of l SNPs, for building a model. It is summarized as follows:

1. Randomly generate a series of Boolean expressions as the set of candidate models, denoted as S.
2. Use each candidate model in S to predict the samples in T.
3. The fitness score of each model is calculated. Those models with better performances are defined as the set of preserved models S_p. The rest of the models, $S_\text{d}=S-S_\text{p}$, will be discarded in the next iteration.
4. The preserved models in S_p are used as templates for producing a new set of candidate models S_d'. Each model of S_d' is generated using one of the four methods:
   - (a) Randomly select one model from S_p and then apply the mutation operation
   - (b) Randomly select two different models from S_p and then proceed to apply a cross-over operation on them
   - (c) Randomly select one model from S_p and have it crossed-over with a randomly generated model
   - (d) Produce a new model using the random model generator

The four methods are selected randomly with equal opportunities. A new set of candidate models S is thus produced where $S=S_\text{p}+S_\text{d}'$.

5. Steps 2–4 are iterated until the optimum fitness score is achieved, or a maximum number of iterations

(a user-defined value) is reached where the model with the highest fitness score stays unchanged.

The number of models in S, S_p, and S_d are heuristically determined by the user. They are constant in all iterations throughout the computation.

The search space of an adequate model is linearly proportional to the number of samples, yet, exponentially proportional to the number of screened SNPs. Hence, the increase of sample sizes is encouraged, as it will enhance the probability of detecting an adequate model at the reasonable expense of time. However, as the number of SNPs increases, more computation is expected, even for a heuristic search method such as the GABA algorithm.

Comparison with other methods

Logistic regression is commonly used for the linear combination of multiple genotypes and their interactions (e.g., Cordell and Clayton 2002). However, this method usually requires an enumeration of various interaction terms, which grows rapidly as l increases. A model must reflect the underlying disease etiology or biological mechanism so as to achieve accurate prediction (Yang et al. 2003). Since the order of interaction is unknown, high-order interactions need to be considered, which will complicate the computation. In comparison, the GABA algorithm detects the order of interactions automatically. The relationship between two adjacent SNPs could be additive or multiplicative, depending on the dataset.

The multifactor dimensionality reduction (MDR) method has been proposed for the detection of high-order interactions among loci (Ritchie et al. 2001). It has been successfully used for the identification of interactions among four SNPs in the estrogen metabolism genes associated with sporadic breast cancer (Ritchie et al. 2001), as well as the three-locus epistasis model for atrial fibrillation (Tsai et al. 2004). Similar to logistic regression, the MDR method has to enumerate all combinations of genotypes. The MDR method has the following weaknesses: (1) it requires a dataset where the numbers of cases and controls are close to 1:1; (2) it is difficult to make a biological interpretation of an MDR model, which is presented in a lookup table style (compare with Ritchie et al. 2001, p 144, Fig. 2). The lookup table style may incur difficulties on the biological interpretation of the model. It will be difficult to show the model in 2D lookup tables if more than five SNPs are involved. Note that the MDR model is analogous to the truth table in Boolean algebra, and the Karnaugh map is a useful tool to convert a truth table into a Boolean expression (Whitesitt 1995). The GABA algorithm, on the other hand, presents the model in a Boolean statement, which is easier to comprehend, facilitating future biological investigations.

Performance evaluation

The GABA algorithm is tested on simulated genotype data, as well as real genotypes from chronic hepatitis C patients treated with interferon-combined therapy. In our implementation, the mutation rate (p) is set at 20%. S contains 300 models; S_p contains 90 models and S_d 210 models.

The first test is to examine the performance of the GABA algorithm using the simulated dataset. We employ datasets with 50 SNPs (i.e., l=50) and 400 samples (200 cases and 200 controls), a data size commonly used for a typical small project. The SNPs in the dataset are in Hardy–Weinberg equilibrium (where the allele distribution is 1:2:1) and are also in linkage equilibrium.

Five distinct simulation datasets are produced. These datasets are embedded with models 1–5, representing a single-locus model, as well as 2-, 3-, 4-, and 5-locus interaction models (Table 1). These models are generated randomly. We can, therefore, evaluate whether the embedded models can be detected by the GABA and MDR methods. We repetitively tested the GABA algorithm three times (i.e., tests 1–3) for each dataset, and found that the algorithm can always identify the embedded model accurately. The numbers of iterations spent for finding the embedded models are presented in Table 1. The average number of iterations increased when the embedded model involves more SNPs (Fig. 1).

The cases:controls ratio of the datasets are 1:1, fulfilling the assumption of the MDR method. Hence, the MDR method is also tested on models 1–5 for comparison purposes. The open-source MDR software v1.0.0rc1 is used, which is downloaded from SourceForge.net. The default setting of the MDR software is adopted, except that the attribute count range (i.e., the range of n in the MDR model) is changed from 1:4 to 1:5 for testing model 5. We found that the MDR method can also detect models equivalent to models 1–5 successfully. To illustrate the lookup table format of MDR models, the MDR result for model 3 is presented in Fig. 2. It is not difficult to envision that the MDR models involving more SNPs are very difficult to present and interpret.

Table 1 The models used for the simulation, as well as the number of iterations computed for each test of the GABA algorithm. The last column presents the cross-validation consistency (CVC) of the multifactor dimensionality reduction (MDR) method

	Boolean statement of cases	Test 1	Test 2	Test 3	Average	MDR (CVC)
Model 1	SNP20=‘CC’	15	4	2	7	10/10
Model 2	–[(SNP24=‘TT’+‘TC’)×(SNP36=‘AA’)]	210	646	516	457.3	10/10
Model 3	–[(SNP3=‘AA’)×(SNP8=‘AA’)+(SNP38=‘GG+AG’)]	6,670	340	8,226	5,078.7	10/10
Model 4	–[(SNP20=‘CC’)×(SNP21=‘GG’)+(SNP17=‘GG’) ×(SNP39=‘AA+TA’)]	13,562	28,095	31,928	24,528.3	10/10
Model 5	[(SNP29=‘TT+TG’)+(SNP32=‘CC’)+(SNP40=‘CC’+‘AC’) +(SNP43=‘AA’)+(SNP47=‘GG’+‘TG’)]	88,182	18,521	9,394	38,699	9/10

The MDR method detected the five SNPs of model 5 correctly, with a cross-validation consistency of 0.9 (Table 1). In one of the cross-validation tests, the MDR method detects SNP0 instead of SNP43. The orders of appearance of SNPs in various tests for model 5 are summarized in Table 2. The computational process of test 1 is presented in terms of the highest fitness score (Fig. 3) and the number of SNPs detected (Fig. 4). The fitness score increases along with the number of computations, showing a continuous improvement. The length of the model varies between 2 and 5. The number of correctly detected SNPs increases gradually. At iteration 14,000, four SNPs have been corrected detected, which results in a fitness score of 2.99. The performance does not change until the fifth SNP is incorporated at iteration 88,182.

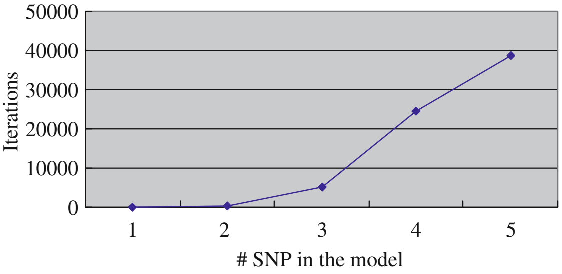

Fig. 1 The average numbers of iterations computed for detecting the embedded models 1–5

The GABA algorithm is then applied to the construction of a prediction model for the interferon-combined therapy. Interferon-α combined with ribavirin is a standard treatment for patients infected by chronic hepatitis C viruses (HCV). The training dataset comprises genotypes of 381 chronic hepatitis C patients. These patients are from National Taiwan University Hospital, Kaohsiung Medical University Hospital, Kaohsiung Municipal Hsiaokang Hospital, and Tri-Service General Hospital in Taiwan, and the samples were collected between years 2002 and 2004. Informed consent and the medical records related to the history of the disease were collected for each subject. All patients had received interferon-α (3–6 MU/dosage, three times per week) and ribavirin (1,000–1,200 mg/day) for 6 months and then followed up for 6 months after the termination of treatment. Patients with concurrent hepatitis B or D infection were excluded from the study. The responsiveness of the treatment is determined by the detection of serum HCV RNA at the end of the follow-up period. Among the 381 patients, 243 are clinically diagnosed as responders (i.e., no HCV RNA detected) and 138 as non-responders.

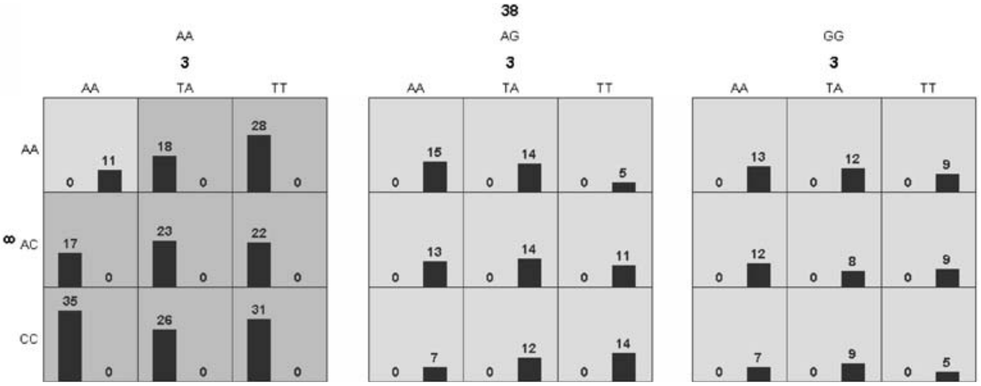

Fig. 2 The detected MDR model equivalent to model 3

Table 2 The order of SNPs appearing in the model construction process of model 5

MDR	SNP29, SNP40, SNP47, SNP32, SNP43
Test 1 of the GABA algorithm	SNP29, SNP40, SNP47, SNP32, SNP43
Test 2 of the GABA algorithm	SNP29, SNP40, SNP47, SNP43, SNP32
Test 3 of the GABA algorithm	SNP29, SNP40, SNP47, SNP43, SNP32

The training dataset comprises 24 SNPs of eight genes involved in interferon signaling and immunomodulating pathways (Table 1). The SNPs were genotyped using either direct sequencing or TaqMan methods. Direct sequencing was conducted with the ABI Prism 3700 instruments (Applied Biosystems) and the data was analyzed using the Phred and PolyPhred programs (Ewing et al. 1998; Nickerson et al. 1997). The TaqMan method was carried out on an ABI Prism 7900 instrument and genotypes were called using the SDS software (Applied Biosystems) supplemented with manual curation. The gene symbols were provided according to the HUGO gene nomenclature committee (Povey et al. 2001).

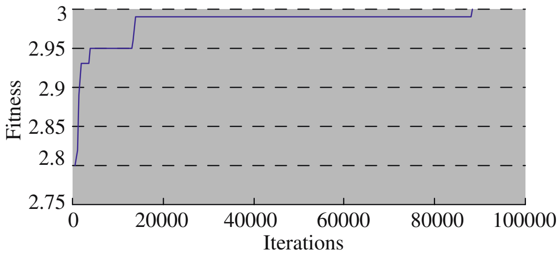

Fig. 3 The performance, shown as the fitness score, is improved gradually after more iterations during the detection of a five-SNP model. The fitness score for a correct model is 3.0, representing 100% sensitivity and specificity

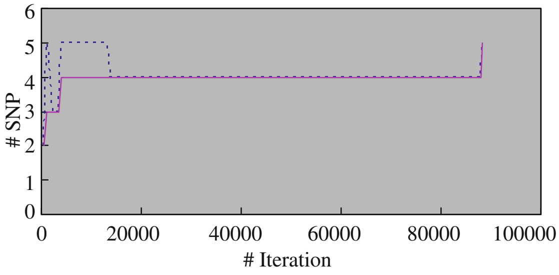

Fig. 4 The number of SNPs per iteration during the detection of model 5. The *dashed line* is the number of SNPs in the model and the *solid line* represents the number of correct SNPs compared with model 5. It shows that the entire set of five SNPs are correctly detected after 88,182 iterations

The eight genes of our study are hypothesized to be influential to the treatment efficacy of interferon-combined therapy (Hwang et al. 2006). *ADAR* is induced by interferon alpha or gamma for its antiviral effect. *ICSBP*1 is a member of the interferon regulatory factor family. It is a negative regulator on an interferon-stimulated response element (ISRE). *IFI*44 has an interferon-stimulated response element in its promoter region. *TAP* is a transporter involved in antigenic peptides transfer into the endoplasmic reticulum. *TGFBRAP*1 binds to Smad4 protein, which is involved in many signaling pathways. *CASP*5 has a central role in apoptosis. *PIK*3*CG* is the phosphoinositide-3-kinase, catalytic, gamma polypeptide. *FGF*s have mitogenic and cell survival activities and are involved in liver organogenesis. During the progression of CHC, the level of *FGF* is elevated.

Prior to model construction, the 24 SNPs are assessed individually for differences of allele and genotype frequencies between the case and control groups using standard χ^2 statistics for contingency tables (Schlesselman 1982). The level of significance is set at 0.05. The genotype comparison employs a 3×2 contingency table, comparing three diplotypes at two conditions (i.e., responders and non-responders). Since 24 SNPs are assessed simultaneously, the issue of multiple comparisons is considered and the threshold on the p value (after Bonferroni correction) is 0.0021. None of the SNPs could be declared as significant according to the allele and genotype tests (Table 3).

The GABA algorithm detects a model comprising eight SNPs in five genes: *ADAR*, *IFI*44, *ICSBP*1, *PIK*3*CG*, and *CASP*5. The non-responders are identified if the following statement is true:

$$\begin{aligned}&(\mathrm{SNP}_7 =' \mathrm{CC/TC}') \times (\mathrm{SNP}_{12} =' \mathrm{GG}')\\&\quad\times (\mathrm{SNP}_{14} =' \mathrm{AA/AC}') \times (\mathrm{SNP}_{20} =' \mathrm{CC/CT}')\\&\quad+ (\mathrm{SNP}_6 =' \mathrm{CC}') \times (\mathrm{SNP}_9 =' \mathrm{GG/GT}')\\&\quad\times (\mathrm{SNP}_{11} =' \mathrm{AA/AG}') \times (\mathrm{SNP}_{16} =' \mathrm{AA/AG}')\end{aligned} \tag{8}$$

According to DeMorgan's theorem (Whitesitt 1995), the responders are identified if the following statement is true:

$$\begin{aligned}&[(\mathrm{SNP}_7 =' \mathrm{TT}') + (\mathrm{SNP}_{12} =' \mathrm{AA/AG}')\\&\quad+ (\mathrm{SNP}_{14} =' \mathrm{CC}') + (\mathrm{SNP}_{20} =' \mathrm{TT}')]\\&\quad\times [(\mathrm{SNP}_6 =' \mathrm{AA/AC}') + (\mathrm{SNP}_9 =' \mathrm{TT}')\\&\quad+ (\mathrm{SNP}_{11} =' \mathrm{GG}') + (\mathrm{SNP}_{16} =' \mathrm{GG}')]\end{aligned} \tag{9}$$

Table 3 The 24 SNPs and their association test results, including the allelic comparison, the genotypic comparison of three genotypes (GG/GC/CC), and the Hardy–Weinberg test, shown in p values

SNP ID	dbSNP ID	Gene symbols	Allele	Number		p value		
				R	NR	Allele	Genotype	Hardy–Weinberg
SNP0	rs2241796	*TGFBRAP*1	T/C	242	135	0.5967	0.1837	0.2358
SNP1	rs1866040	*TGFBRAP*1	A/G	237	134	0.5692	0.6881	0.9519
SNP2	rs2576737	*TGFBRAP*1	C/T	213	116	0.3657	0.6618	0.1136
SNP3	rs518604	*CASP5*	A/G	238	131	0.1204	0.0658	0.0995
SNP4	rs2282658	*CASP5*	C/G	231	134	0.3622	0.6042	0.5344
SNP5	rs484345	*CASP5*	A/G	235	132	0.7719	0.6343	0.1957
SNP6	rs1699087	*CASP5*	C/A	233	126	0.7154	0.6206	0.3463
SNP7	rs7515339	*ADAR*	C/T	217	128	0.0598	0.0845	0.8721
SNP8	rs903323	*ADAR*	T/C	229	125	0.5916	0.5574	0.0934
SNP9	rs2148686	*IFI*44	G/T	240	130	0.0765	0.2166	0.0013
SNP10	rs2070123	*IFI*44	T/C	235	132	0.6468	0.6920	0.8029
SNP11	rs273249	*IFI*44	A/G	232	130	0.0576	0.1606	0.9005
SNP12	rs12120187	*IFI*44	G/A	232	129	0.0041	0.0096	0.3928
SNP13	rs305067	*ICSBP*1	G/C	217	117	0.9664	0.7854	0.3264
SNP14	N/A	*ICSBP*1	A/C	241	135	0.4295	0.5724	0.9529
SNP15	rs305088	*ICSBP*1	C/T	234	134	0.0737	0.1605	0.9300
SNP16	rs870614	*ICSBP*1	A/G	232	127	0.3460	0.6174	0.8828
SNP17	rs2071543	*TAP*2	G/T	230	132	0.8306	0.8882	0.0876
SNP18	rs1800453	*TAP*2	T/C	237	126	0.3302	0.5974	0.8765
SNP19	rs1526083	*PIK3CG*	G/A	233	127	0.6117	0.5488	0.4493
SNP20	rs3779501	*PIK3CG*	C/T	238	135	0.2079	0.4581	0.3740
SNP21	rs249926	*FGF*1	C/T	236	128	0.0540	0.1428	0.6239
SNP22	rs11117421	*ICSBP*1	A/G	217	112	0.9223	0.7109	0.0003
SNP23	rs305095	*ICSBP*1	C/T	240	133	0.6499	0.7775	0.9464

When the eight SNPs are combined for the classification of patients, the prediction result is as shown in Table 4. The p value of the chi-square test of Table 4 is 1.79×10^{-7}, showing a strong association in the training dataset. The performance indexes are as follows: the sensitivity is 62.1%, the specificity is 70.0%, the positive predictive value (PPV) is 79.0%, and the negative predictive value (NPV) is 50.7%.

A prediction model needs to be validated using an independent dataset, from either prospective or retrospective studies, so as to demonstrate its capability of prediction. The above model is, thus, validated using samples of 159 persons (121 responders and 38 non-responders). These samples were collected during the years 2004–2005. The validation result is shown in Table 5, where the sensitivity is 54.7%, the specificity is 71.4%, the PPV is 86.4%, and the NPV is 32.1%.

Table 4 The prediction performance of the model on the training dataset (381 patients). Only 283 patients are shown in the table. The other 98 patients have missing data in the genotypes, and, thus, were not predictable

Training		Clinical status	
		Positive	Negative
Prediction	Positive	113	30
	Negative	69	71

The p value of the chi-square test of Table 5 is 0.0067, a very small number, which shows a strong association with the validation dataset. The similarity of the sensitivity and specificity values enhances our confidence that the model has a certain degree of consistency for predicting the efficacy of interferon-combined treatment.

Conclusions

The genetic algorithm and Boolean algebra (GABA) algorithm systematically investigates multiple single-nucleotide polymorphisms (SNPs) and their adequate combinations for predicting phenotypic traits of complex diseases or pharmacogenomic studies. The GABA

Table 5 The model derived from the training dataset is then used to predict the samples in a validation dataset (159 patients). Only 152 patients are shown in the table. The other seven patients have missing data in the genotypes, and, thus, were not predictable

Validation		Clinical status	
		Positive	Negative
Prediction	Positive	64	10
	Negative	53	25

algorithm shows promising capabilities in deriving a model from a large pool of SNP genotypes. This is demonstrated by experiments on the simulated datasets, as well as a real dataset of interferon-combined treatment. A Boolean expression model detected by the GABA algorithm is easily comprehensible, interpretable, and examinable by physicians and scientists. This is a merit of the GABA algorithm compared with other methods, such as multifactor dimensionality reduction (MDR) or logistic regression.

Although we use SNP diplotypes to demonstrate the algorithm, the GABA methodology should be able to incorporate haplotypes and other physical information, provided that this information can be represent adequately as model elements m_i. This remains as our future research direction.

Acknowledgments The authors wish to express their sincere appreciation to Dr. Pei-Jer Chen of the Hepatitis Research Center, National Taiwan University, Dr. You-Chen Chao of the Tri-Service General Hospital, Dr. Ming-Lung Yu of the Kaohsiung Medical University Hospital, and Dr. Chuan-Mo Lee of the Kaohsiung Chang-Gung Memorial Hospital for the fruitful research collaborations. The efforts from the following members of Vita Genomics, Inc. are gratefully appreciated: Chun-Yi Chen, Jun-Jeng Fen, Shu-Ching Wang, Cheng-Yan Lee, and Tsang-Rong Tu.

References

Burnham KP, Anderson DR (2001) Kullback–Leibler information as a basis for strong inference in ecological studies. Wildlife Res 28(2):111–120

Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2(2):91–99

Cordell HJ, Clayton DG (2002) A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. Am J Hum Genet 70(1):124–141

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8(3):175–185

Gardner WA (1988) Statistical spectral analysis. Prentice-Hall, Englewood Cliffs, New Jersey

Goldberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Reading, Massachusetts

Holland JH (1998) Adaptation in natural and artificial systems, 5th edn. MIT Press, Cambridge, Massachusetts

Hwang Y, Su C, Chen D-S, Chen P-J (2006) Prospect of individualized medicine in chronic hepatitis C therapy by pharmacogenomics. Curr Pharmacogenom 4(2):157–167

Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucl Acids Res 25(14):2745–2751

Online Mendelian Inheritance in Man, OMIM (2000) McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University, Baltimore, Maryland and National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland. Home page at http://www.ncbi.nlm.nih.gov/omim/

Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain HM (2001) The HUGO Gene Nomenclature Committee (HGNC). Hum Genet 109(6):678–680

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65(1):220–228

Rabbee N, Speed TP (2006) A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 22(1):7–12

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69(1):138–147

Schlesselman JJ (1982) Case-control studies. Design, conduct, analysis. Oxford University Press, New York

Schaid DJ (2004) Evaluating associations of haplotypes with traits. Genet Epidemiol 27(4):348–364

Tsai CT, Lai LP, Chiang FT, Fallin D, Hwang JJ, Ritchie MD, Moore JH, Hsu KL, Tseng CD, Liau CS, Lin JL, Tseng YZ (2004) Renin-angiotensin system gene polymorphisms and atrial fibrillation. Circulation 109(13):1640–46

Whitesitt JE (1995) Boolean algebra and its applications. Dover, New York

Yang Q, Khoury MJ, Botto L, Friedman JM, Dlanders WD (2003) Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. Am J Hum Genet 72(3):636–649