Genet Program Evolvable Mach (2007) 8:395–411
DOI 10.1007/s10710-007-9039-5

ORIGINAL PAPER

Genomic mining for complex disease traits with "random chemistry"

Margaret J. Eppstein · Joshua L. Payne · Bill C. White · Jason H. Moore

Received: 15 November 2006 / Revised: 14 June 2007 / Published online: 4 October 2007

Abstract Our rapidly growing knowledge regarding genetic variation in the human genome offers great potential for understanding the genetic etiology of disease. This, in turn, could revolutionize detection, treatment, and in some cases prevention of disease. While genes for most of the rare monogenic diseases have already been discovered, most common diseases are complex traits, resulting from multiple gene–gene and gene-environment interactions. Detecting epistatic genetic interactions that predispose for disease is an important, but computationally daunting, task currently facing bioinformaticists. Here, we propose 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 to detecting small auto-catalytic sets of molecules from within large sets. Although the algorithm is conceptually straightforward, its success hinges upon the creation of a fitness function able to discriminate large sets that contain subsets of interacting genetic features from those that don't. Here, we employ an approximate and noisy fitness function based on the ReliefF data mining algorithm. We establish

M. J. Eppstein (✉)
Departments of Computer Science and Biology, University of Vermont, Burlington, VT 05405, USA
e-mail: Maggie.Eppstein@uvm.edu

J. L. Payne
Department of Computer Science, University of Vermont, Burlington, VT 05405, USA
e-mail: Joshua.Payne@uvm.edu

B. C. White · J. H. Moore
Computational Genetics Laboratory, Dartmouth College, Lebanon, NH 03756, USA

B. C. White
e-mail: Bill.C.White@dartmouth.edu

J. H. Moore
e-mail: Jason.H.Moore@dartmouth.edu

proof-of-concept using synthetic data sets, where individual features have no marginal effects. We show that the resulting algorithm can successfully detect epistatic pairs from up to 1,000 candidate single nucleotide polymorphisms in time that is linear in the size of the initial set, although success rate degrades as heritability declines. Research continues into seeking a more accurate fitness approximator for large sets and other algorithmic improvements that will enable us to extend the approach to larger data sets and to lower heritabilities.

Keywords Evolutionary algorithms · Epistasis · Single nucleotide polymorphisms · Data mining · Genome-wide association studies · Complex traits · Feature selection

1 Introduction

The successful sequencing of an entire "representative" human genome [7, 29] and development of methods for rapid and affordable genotyping [26] have stimulated large-scale research efforts in identifying human genetic variability, and millions of single nucleotide polymorphisms (SNPs) have now been identified [6, 8, 10]. Such genomic information carries with it the potential for improved understanding regarding the genetic etiologies of disease, and may revolutionize our abilities to detect, treat, and even prevent disease [4, 13, 21]. Indeed, linkage studies have already been highly successful in identifying the genes responsible for most of the known rare monogenic diseases (i.e., those that are caused by mutations in a single gene), including cystic fibrosis, Alzheimer's disease, Hirschsprung disease, and phenylketonuria [3, 13, 21]. But monogenic diseases are the exception, not the rule. Our growing understanding of the complex interconnectedness of genetic [22, 28] and metabolic [23] regulatory networks is spawning a new appreciation for the ubiquity of non-linear epistatic genetic interactions in predisposing individuals for disease [14]. In fact, most common diseases, including heart disease, obesity, cancer, diabetes, and schizophrenia, are caused by complex interactions between many genetic and environmental factors [3, 13, 30]. Because these complex diseases do not follow Mendelian inheritance patterns, linkage studies are ineffective in identifying which genetic variations are associated with these diseases [4]. Consequently, researchers are seeking new methods for conducting genome-wide association studies to detect non-linearly interacting SNPs that are associated with disease [4, 27, 30]. Detecting which handfuls of SNPs (from among hundreds, thousands, or even hundreds of thousands of genotyped candidate SNPs) exhibit non-linear epistatic interactions that predispose for disease is a computationally daunting combinatorial optimization task, because individual SNPs may have little or no detectable ill effects [17]. Solution of this important problem is further exacerbated by low disease heritability, small sample sizes, and a lack of information regarding how many, if any, SNPs interact. New methods, such as multifactor dimensionality reduction [15, 16, 24], show promise for detecting epistatic interactions, yet are still dependent on exhaustive search [12]. However, exhaustive searches are not computationally feasible for large-scale association studies, and the optimal strategy for genome-wide analysis is still an open question. Various statistical, data mining, and machine learning strategies have been applied to the problem [5, 11, 20], and while several of these methods have shown encouraging results for small to moderate problems (typically a few hundred SNPs or less), none have emerged that are suitable for really large-scale genome-wide association studies. There have been some recent attempts to apply evolutionary algorithms to this problem [18, 31]. However, these approaches attempt to grow and recombine small, high fitness building blocks into complete solutions. Consequently, if there are no detectable effects until the correct combination of SNPs is found, such approaches will be no better than random search [18, 31]. It has been suggested that additional expert knowledge could be provided to bias the search towards including more promising SNPs [18], but such knowledge may not be available and also carries the risk of incorrectly biasing the search.

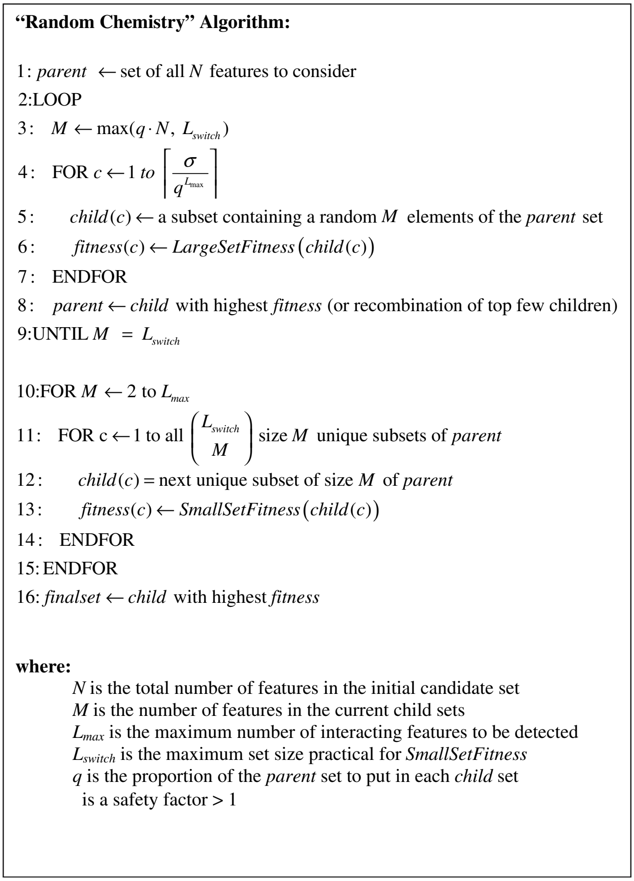
We propose a new type of evolutionary algorithm, inspired by the "random chemistry" procedure outlined by Kauffman to identify small sets of auto-catalytic molecules [9]. In this approach, we tackle this non-linear feature selection problem by hill-climbing from larger to smaller sets of SNPs, rather than vice versa. The proposed algorithm necessitates the creation of a different type of fitness metric that (a) yields higher fitness for large sets of SNPs that contain all members of a target subset of interacting SNPs, as compared to same-sized sets that do not contain all of the members of the interacting subset, and (b) yields higher fitness for smaller vs. larger sets of SNPs, both of which contain all members of a target interacting subset. We propose two such fitness metrics, an accurate but computationally costly fitness metric for evaluating small sets of SNPs, and a noisy but computationally tractable fitness metric for evaluating large sets of SNPs. The former requires building cross-validation directly into the fitness metric, and the latter requires noise compensation techniques to be incorporated into the random chemistry algorithm. Although non-deterministic, the resulting algorithm requires only $\theta(\log N)$ fitness evaluations to detect a small subset of interacting SNPs from an initial set of N candidate SNPs. We establish proof-of-concept using synthetic data sets of up to 1,000 candidate SNPs, and with heritabilities of both 0.4 and 0.1, where heritability is the proportion of disease cases attributable to genetic effects.

2 Random chemistry algorithm

Kauffman [9] outlined a simple procedure for detecting small auto-catalytic sets of L molecules from a large number of N candidate molecules, as follows. Consider putting a random half of the molecules in a test tube. Clearly, any given molecule has a 0.5 probability of ending up in this tube, and the probability of all L desired molecules being in this tube is nearly 0.5^L. Thus, it is expected that one out of almost every 2^L such tubes will contain all L interacting molecules. By doubling the number of tubes to 2^{L+1}, the chance that at least one tube will be "positive" for all L molecules is also doubled. Now, assuming that the L molecules interact in some detectable way (e.g., by forming a by-product), one could simply screen for a "positive" tube, and repeat the process on the molecules in that tube, until only the correct L molecules remain. Note that this process will require only $2^{L+1}\log_2 N/L$ screening tests to pick out the L interacting molecules. This process can be viewed as feature selection, where the features Kauffman sought were non-linearly interacting molecules.

Our goal is to detect sets of L features that are epistatically interacting SNPs. Unfortunately, we will not know *a priori* how big L is. However, the sample size will provide an upper limit on the maximum degree of epistasis L_{max} that can feasibly be detected with a given statistical power, so final subsets can be restricted to be of size L_{max} or less. In Fig. 1, we generalize the "random chemistry" algorithm for detecting small sets of up to L_{max} epistatically interacting features from among N candidate features.

Note that lines 1–9 use the random chemistry approach to exponentially reduce the candidate set size from N to L_{switch}, and so require $\theta(\log N)$ fitness evaluations with a noisy fitness approximation function "*LargeSetFitness*". The first few



"Random Chemistry" Algorithm:

```
1: $parent \leftarrow$ set of all $N$ features to consider
2:LOOP
3:   $M \leftarrow \max(q \cdot N,\ L_{switch})$
4:   FOR $c \leftarrow 1$ to $\left\lceil \frac{\sigma}{q^{L_{\max}}} \right\rceil$
5:       $child(c) \leftarrow$ a subset containing a random $M$ elements of the $parent$ set
6:       $fitness(c) \leftarrow LargeSetFitness\left(child(c)\right)$
7:   ENDFOR
8:   $parent \leftarrow child$ with highest $fitness$ (or recombination of top few children)
9:UNTIL $M = L_{switch}$

10:FOR $M \leftarrow 2$ to $L_{max}$
11:   FOR c $\leftarrow 1$ to all $\binom{L_{switch}}{M}$ size $M$ unique subsets of $parent$
12:       $child(c) =$ next unique subset of size $M$ of $parent$
13:       $fitness(c) \leftarrow SmallSetFitness\left(child(c)\right)$
14:   ENDFOR
15:ENDFOR
16:$finalset \leftarrow child$ with highest $fitness$
```

where:

N is the total number of features in the initial candidate set
M is the number of features in the current child sets
L_{max} is the maximum number of interacting features to be detected
L_{switch} is the maximum set size practical for *SmallSetFitness*
q is the proportion of the *parent* set to put in each *child* set
is a safety factor > 1

Fig. 1 Pseudo code for a generalized "random chemistry" algorithm for detecting small sets of up to L_{max} epistatically interacting features from among N candidate features

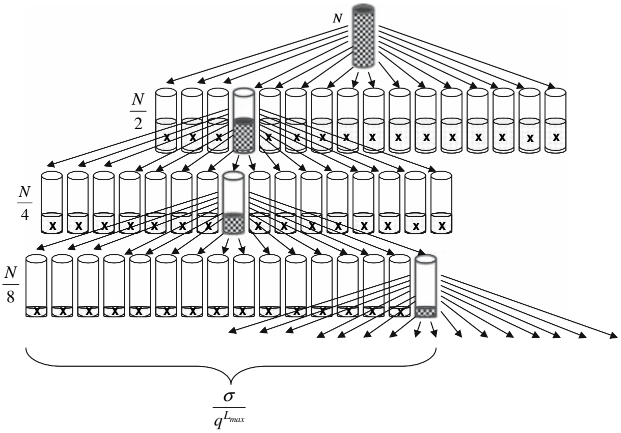


Fig. 2 An illustration of the first few iterations of the "random chemistry" algorithm (Fig. 1) for a set initially containing N features, illustrated here using a proportion $q = 0.5$, $L_{max} = 3$, and $\sigma = 2$. Thus, in each iteration, an independently selected random 50% of the features are placed into each of 16 "test tubes". One (of ~ 2 expected) of the 16 tubes that screens "positive" for the desired interaction is selected (black stippling) and the others are discarded (black X's). This is repeated for $\theta(\log_2 N)$ levels

iterations of the loop starting on line 2 are illustrated in Fig. 2. The size of the child sets are designated to be not more than some proportion q of the size of the parent sets, but not smaller than L_{switch}. Noise in the *LargeSetFitness* function can be partially compensated for by lowering selection pressure. This can be accomplished by saving and recombining the top few children as indicated on line 8, rather than simply selecting the fittest child, as described in more detail in Sect. 4. Here, we simply choose L_{switch} to be the size of the largest set that can be practically computed with a more accurate, but more computationally intensive, fitness function "*SmallSetFitness*", with the caveat that $L_{switch} \geq L_{max}$. Once the best subset of size L_{switch} has been identified, lines 10–16 perform a final exhaustive search on all possible subsets of size 2 to L_{max}, thus requiring $\theta\left(\sum_{M=2}^{L_{max}} \binom{L_{switch}}{M}\right)$ fitness evaluations with the function "*SmallSetFitness*".

In the general random chemistry algorithm (Fig. 1), the ratio of child set size to parent set size is the proportion q. Thus, the expected number of "positive" child sets in a given iteration (i.e., those containing all of the desired interacting features) is:

$$E[\#\text{ positive child sets}] = \frac{(qN - L)!N!}{(qN)!(N - L)!} \tag{2.1}$$

For large N and small L, Eq. (2.1) approaches $1/q^L$. Since the creation of child sets is non-deterministic, we allow the user to specify a safety factor σ, which specifies the approximate number of the child sets that are expected to be positive. Thus, the number of child sets we generate during a given iteration is the ceiling of σ/q^{Lmax}, as indicated on line 4 of Fig. 1. The larger the proportion q, the fewer the number of child sets that must be generated in order to expect σ positive child sets, but the more iterations that will be required in order to reduce the feature set size to L_{switch}. Thus, q can be optimally determined once the runtime requirements of the *LargeSetFitness* approximator are known as a function of the size of the set being evaluated, in order to minimize the overall time required for the loop on lines 2–9, as long as qN does not exceed the maximum set size for which *LargeSetFitness* can distinguish "positive" from "negative" sets. However, since research continues regarding the best algorithm for the *LargeSetFitness* function, we simply set $q = 0.5$, as suggested by Kauffman [9], for the experiments reported here. For realistically sized data sets, L_{max} will be bounded by a small constant and qN will be bounded by a relatively large constant, depending on the power of the *LargeSetFitness* function.

With our current implementations of the two fitness functions (described in the following sections), the function *LargeSetFitness* scales linearly with M (the size of the set being evaluated), while the function *SmallSetFitness* scales exponentially with M. Consequently, we set $L_{switch} = 8$, and the time complexity of the loop on lines 10–15 always requires a constant amount of time, independent of N. Thus, the time complexity of the overall algorithm is governed by the $\theta(\log N)$ fitness evaluations with the noisy fitness approximator. The time complexity of the fitness function is obviously a function of both the initial set size N and the sample size. However, we are most interested in how the algorithm scales with the number of features N in the initial set for a given sample size, and so treat sample size as a constant for the purposes of this analysis. Since the current implementation of *LargeSetFitness* is $\theta(M)$, where M starts at $N/2$ and decreases exponentially for each of the $\theta(\log N)$ iterations, the overall time complexity of the current implementation is $\theta(N)$.

The random chemistry algorithm can be considered as an evolutionary algorithm with a variable sized (but strictly decreasing) genome and a ($\mu = 1$, $\lambda = \sigma/q^{Lmax}$) generational strategy with truncation child selection, a random halving mutation operator, and a recombination operator that is described in Sect. 4. However, unlike a typical evolutionary algorithm, the random chemistry algorithm is bounded in time, as described above. The success of this algorithm hinges on the ability of the fitness functions to accurately predict the likelihood of whether or not a given set of features is positive (contains all of the features in the interacting subset) or negative (doesn't contain all of the features in the interacting subset). In the next two sections, we describe the two fitness functions currently employed.

3 Assessing the fitness of small feature sets

How does one evaluate whether a small set of SNP loci (the features in question) contain a subset that interact to influence susceptibility to a given disease? This is a non-trivial question, especially when heritability is low and epistatic interactions are such that different genotypes at the same loci exhibit different penetrance values for the same disease.

We borrow and modify an idea commonly employed in medical decision making; i.e., the receiver operating characteristic (ROC) curve [1], described below.

The “sensitivity” of a test is equivalent to the true positive fraction (TPF); that is, the fraction of test subjects which have the disease and yield a positive test result. On the other hand, the “specificity” of a test is simply 1 minus the false positive fraction (FPF), where the FPF is the fraction of test subjects that do not have the disease but still give a positive test result. An ROC curve is simply a plot showing the trade-off between increasing sensitivity and decreasing specificity, as we vary some cutoff criterion for when a test result is considered positive (e.g., Fig. 3). Ideally, a good test will have both high sensitivity and high specificity, so will have an ROC curve that passes close to the upper left hand corner of the plot. The area under the curve (AUC) is frequently used as a measure of the predictive power of a test. A test with no predictive power will simply have a diagonal ROC curve, with AUC = 0.5. In the case of assessing the predictive power of a set of SNPs, we can create an ROC curve as follows. First, we compute the “sample penetrance” p_g for each possible diploid genotype g at the specified SNP loci; i.e., the proportion of subjects in the sample with genotype g that exhibit the disease. The ROC curve can then be estimated by varying the cutoff of sample penetrance that is considered a positive test result. Since there are only a small finite number of possible genotypes for a given small set of loci, this is not a continuous curve, but rather is a set of discrete points. Thus, it is not meaningful to try to calculate an AUC. Instead, we use the maximum distance (*MD*) above the diagonal as a measure of the predictive power of a set of SNPs. The “sample prevalence” P is the proportion of the sample that exhibits the disease. The *MD* occurs at the point where the cutoff is P; that is, where we consider all genotypes whose sample penetrance is greater than or equal to the sample prevalence to be positively associated with the disease. Specifically

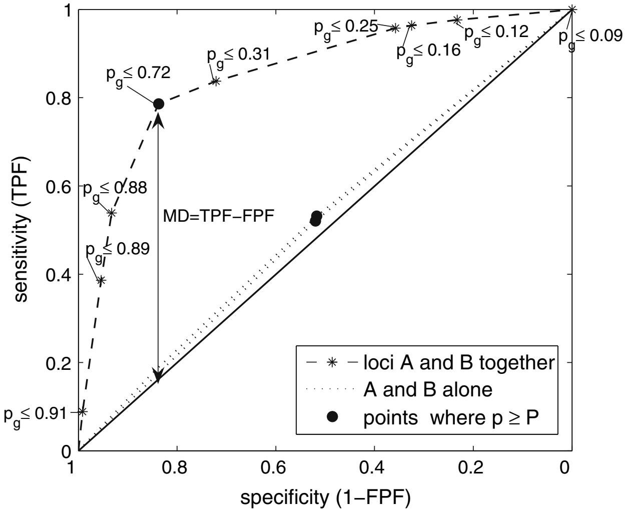


Fig. 3 ROC curve for the data shown in Table 1, with the maximum distance (MD) fitness metric shown for the two interacting A and B loci. Note that the specificity axis is shown in decreasing order

Table 1 One sample table of penetrance p_g for all nine genotypes of two loci, A and B, that interact epistatically to influence susceptibility to a disease with heritability 0.4 in the synthetic data set described in the text

Genotype	*BB*	*Bb*	*bb*	*A alone*
AA	0.09	**0.72**	**0.91**	0.52
Aa	**0.89**	0.25	0.31	0.48
aa	0.12	**0.88**	0.16	0.53
B alone	0.48	0.52	0.48	*P*=0.50

$$MD = TPF - FPF,\ \forall g \middle| p_g \geq P \tag{3.1}$$

This approach is consistent with the way in which high and low risk genotypes are lumped in multifactor dimensionality reduction [15, 16, 24]. The computation of the ROC curve and the maximum distance *D* metric is illustrated with a synthetic data set generated from the penetrance model shown in Table 1. This data set contains 1,000 SNPs, each with two possible allele values, from 1,600 individuals, half of whom have disease (so $P = 0.50$). Although none of the SNP loci are individually correlated with the disease, the alleles at $L = 2$ loci interact epistatically to affect susceptibility to the disease with 0.4 heritability. Suppose the two epistatically interacting SNP loci are locus A, with 3 possible diploid genotypes *AA*, *Aa*, and *aa*, and locus B, with genotypes *BB*, *Bb*, and *bb*. In this example, the calculated sample penetrance (p_g) for the nine possible genotype combinations is as shown inside the double lines in Table 1, with the boldface values representing those genotypes with $p_g \geq P$ that are positively associated with the disease. Note that sample penetrance values for the A and B loci taken individually (right column and bottom row, respectively) are all close to the sample prevalence of $P = 0.5$ (bottom right value), indicating that there are no marginal effects.

The ROC curve for this data is shown in Fig. 3, where the maximum distance metric *MD* occurs when individuals with genotypes *AABb*, *AAbb*, *AaBB*, and *aaBb* are considered positively associated with the disease and all others are considered negatively associated with the disease. Also shown, for comparison, are the ROC curves for the A and B loci alone, which are not significantly above the main diagonal and hence indicate that these loci have no predictive power when viewed in isolation.

Unfortunately, the *MD* metric, by itself, is insufficient for use in the random chemistry algorithm. Although *MD* is higher for “positive” sets containing the correct subsets of epistatically interacting SNPs than for “negative” sets of the same size, the *MD* metric continues to increase as additional “extra” SNPs are considered. This is illustrated in Fig. 4a for the same synthetic 1,600 member data set described above, for random sets containing both correct SNP loci (×‘s) and

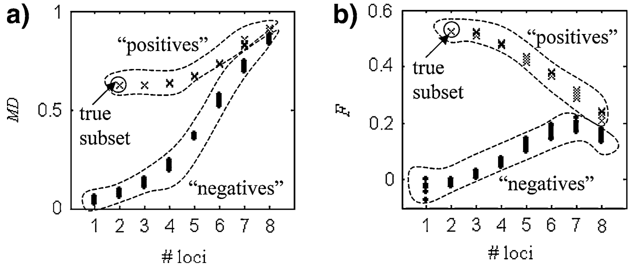


Fig. 4 (**a**) Maximum ROC distance metric *MD* from Eq. 3.1, and (**b**) fitness metric *F* from Eq. 3.2. "Positive" subsets of SNPs are shown with ×'s (top trajectories), with the minimal true subset of two interacting SNPs (circled). "Negative" sets containing 0 or 1 of the correct SNPs are shown with •'s (bottom trajectories)

random sets containing 0 or 1 of the correct loci (•'s). The reason for this is two-fold; first, adding in additional loci does not lower the predictive power of epistatically interacting loci also contained in the set, and second, as the number of loci M in a set increases, the number of genotypes increases as 3^M. As the number of possible genotypes approaches the number of samples in the data set, the MD metric simply over-fits the data and thus continues to rise.

In order to compensate for overfitting, we incorporate cross-validation directly into the fitness metric. Specifically, we divide the sample in two, and compute sample penetrance tables for each half of the data set. For each of these two tables, we determine which of the 3^M genotypes (a) were represented in both tables, (b) had sample penetrance < the sample prevalence ("negatives"), (c) had sample penetrance ≥ the sample prevalence ("positives"). We then define a cross-validation value C to be the proportion of genotypes that "agreed" between the two tables (i.e., were either both positively or both negatively associated with the disease), and we also define a metric of support S as the proportion of the 3^M genotypes that had representatives in both tables. We define MD_{min} to be the smallest MD above the ROC diagonal that is considered to be more predictive than random (we used $MD_{min} = 0.1$). If $MD > MD_{min}$, then we repeat the cross-validation test for some number of repetitions (we used 20), otherwise we do not repeat the cross-validation. We define the fitness F as follows, where averaged quantities (averaged over either the 1 or 20 repetitions) are indicated by horizontal bars:

$$F = \left(\overline{MD} - MD_{min}\right) \cdot \overline{C} \cdot \overline{S}^{0.25} \tag{3.2}$$

The average support $\overline{S}$ is raised to the 0.25 power to minimize the effects of the strong non-linearity in this metric. While none of $\overline{MD}, \bar{C}$, or $\bar{S}$ alone satisfy the criteria for our fitness function, the combined fitness metric F does, at least for sets of size $M \leq 8$ (Fig. 4b). Thus we can correctly distinguish larger sets of SNPs that contain the correct loci from those that don't, and hill-climb from larger sets to smaller sets that contain fewer "extra" loci. Function F, shown in Eq. 3.2, is what is currently implemented in the function *SmallSetFitness* (Fig. 1, line 13).

Since there are three possible diploid genotypes per each biallelic locus (the two homozygotes and the heterozygote), there are 3^M possible genotype combinations at

M loci for which the sample penetrance must be computed. Thus, computing the maximum distance MD metric for M loci has time complexity $\theta(3^M)$, which is only practical for small sets of M SNPs. Moreover, for sets of size 8 the support $\overline{S}$ is already nearly zero. For these reasons, we did not apply Eq. 3.2 to sets with more than $L_{switch} = 8$ loci.

4 Assessing the fitness of large feature sets

The most difficult implementation detail required for the random chemistry algorithm to be useful is the determination of an effective way to estimate the fitness of large sets of SNPs. This remains a non-trivial problem and we are continuing to seek better approaches. However, to date, the most effective fitness approximator we have achieved is based on the ReliefF data mining algorithm [25], which has previously been shown to have promise for a similar SNP problem [19]. The ReliefF algorithm attempts to estimate importance of weights of each locus in discriminating between two classes (e.g., healthy and diseased) as shown in Fig. 5. We define a "rough" fitness function F_r as:

$$F_r = \text{mean of top 25\% of ReliefF weights } W \tag{4.1}$$

For the results reported here, we used $k = 1{,}600$ (where we used each sample individual exactly once) and $nn = 10$ (i.e., 10 nearest neighbors). The nearest neighbors are determined by maximizing the number of loci with the same genotype as the sample R_i. For large numbers of loci N, this fitness approximator gets increasingly noisy because the ReliefF algorithm may pick the "wrong" nearest neighbors (i.e., based on matching genotypes of irrelevant SNP loci). Nonetheless, this fitness approximator works surprisingly well in distinguishing "positive" from "negative" sets up to a few hundred SNPs, although it becomes noisier as heritability decreases, as shown in Fig. 6, where asterisks denote where the average

```
"ReliefF" Algorithm:

Weights $W_j$=0,  $j$ {1..$N$} loci
FOR $i$ = 1 to $k$ ($k$ = # random samples)
  Select an individual $R_i$
  $Hits$ = $nn$ nearest neighbors from same class as $R_i$
  $Misses$ = $nn$ nearest neighbors from other class as $R_i$
  FOR $j$ = 1 to $N$ (for all loci)
    $H_j$ = proportion of the $nn$ $Hits_j$ that matched value of $R_{ij}$
    $M_j$ = proportion of the $nn$ $Misses_j$ that matched value of $R_{ij}$
    $W_j = W_j + (H_j - M_j)/k$ (estimate importance of each locus)
  ENDFOR
ENDFOR
```

Fig. 5 Pseudo-code for the ReliefF data mining algorithm [25]

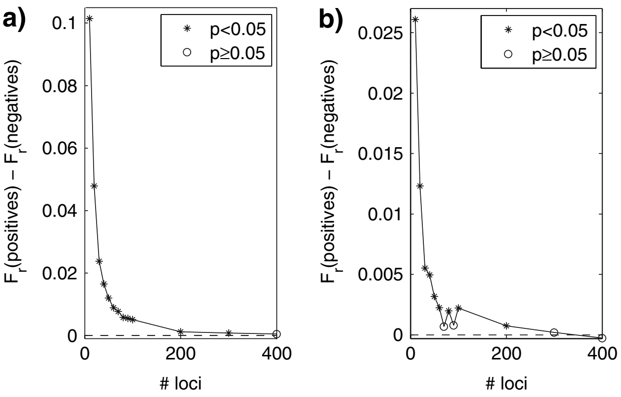


Fig. 6 The differences in mean rough fitness using the approximator F_r from Eq. 4.1 between "positive" random sets (that contain the correct two interacting SNP loci) and "negative" random sets (without the correct loci). Each data point represents the differences in the means of five repetitions using random sets of the designated size; asterisks indicate that the means were statistically significantly different ($P < 0.5$, 2-tailed t-test), whereas small circles indicate set sizes where the means were not statistically different

rough fitnesses of five trials of random "positive" sets and five trials of random "negative" sets were statistically different ($P < 0.05$, 2-tailed t-test). Function F_r, shown in Eq. 4.1, is what is currently implemented in the function *LargeSetFitness* (Fig. 1, line 6).

Since F_r is a noisy approximate fitness function, we modified the step shown in Fig. 1 Line 8 of the random chemistry algorithm to save the top t fittest children sets (rather than just the single fittest child set), where $t = \lfloor \log_4 N_p \rfloor$ and N_p is the number of SNPs in the parent set of the current iteration. These sets are then recombined in an *ad hoc* fashion by saving all of the SNPs in the fittest set along with all other SNPs that occurred in at least two of the remaining top sets

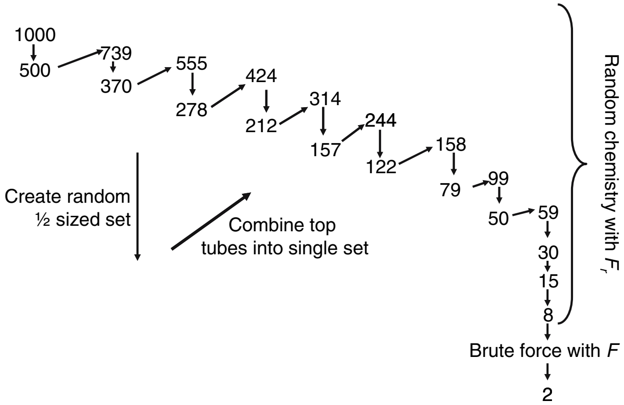

1000
500
739
370
555
278
424
212
314
157
244
122
158
79
99
50
59
30
15
8
Brute force with F
2
Create random ½ sized set
Combine top tubes into single set
Random chemistry with F_r

Fig. 7 Decrease in SNP set sizes during a representative run of the random chemistry algorithm compensated for noise, in which two epistatically interacting SNPs were correctly detected from 1,000 SNP loci, using the data set described in Sect. 2

(i.e., parent ← $S_1 \cup (S_2 \cap S_3 \cap \ldots S_t)$, where S_i represents the i^{th} fittest child set). Thus, set sizes are no longer reduced by exactly ½ during each iteration, but were only reduced by approximately ½ to ¼ (Fig. 7). In practice, this requires on the order of $\log_{1.5}N$ iterations (rather than $\log_2 N$ iterations, when the set size is strictly halved) due to the merging of sets.

5 Experimental results

We demonstrate proof-of-concept using several sets of synthetic data. Each data set had two loci generated using a distinct 2-locus epistatic penetrance table with no marginal effects, created by a stochastic method designed to achieve specific heritabilities that meet the criteria described in [2], with varying numbers of extraneous loci added in that exhibit no association with the disease. Using synthetic data sets with loci that exhibit no individual effects enables us to validate the method under the most extreme conditions of pure epistasis. There were 1,600 samples in each set with sample prevalence of 0.5, and major/minor allele frequencies of 0.6/0.4, respectively, at each locus. Some of the data sets had disease heritability of 0.4, and some had heritability of 0.1. A total of 250 test runs were performed, as follows. We ran five repetitions of data sets from five different penetrance tables with each of $N = 200$, 500, and 1,000 initial SNPs, for a total of 75 test runs at 0.4 heritability. Since F_r is so much noisier for the 0.1 heritability data than at 0.4 heritability (Fig. 6) we tested more intermediate set sizes for these lower heritability data sets. Specifically, we ran five repetitions of data sets from five different penetrance tables with each of $N = 100$, 150, 200, 300, 400, 500, and 1,000 initial SNPs, for a total of 175 test runs at 0.1 heritability. For computational efficiency in running the experiments reported here, we generated 12 child sets per iteration, corresponding to $\sigma = 3$, $q = 0.5$, and $L_{max} = 2$ on line 4 of Fig. 1, although we used $L_{max} = L_{switch} = 8$ for the loop on line 10 of Fig. 1, in order to ensure that we could hill-climb from 8 SNPs to the correct 2 SNPS with the fine fitness function In preliminary experiments, we had confirmed that using a higher L_{max} on line 4 of Fig. 1 did not change the final results for these data sets (where the true degree of epistasis is *known* to be 2), but simply increased the computation time since the number of required child sets that must be evaluated each iteration scales with $1/q^{Lmax}$. In a real data set (where the true degree of epistasis is unknown), one should use the highest L_{max} that statistical limitations of the sample size allow.

For the 0.4 heritability data, the overall percent of successful trials for all repetitions on all penetrance tables (where success is defined as correctly identifying the set of 2 epistatically interacting SNPs) declined from 68% to 40% as the size N of the initial set increased from 200 to 1,000 SNPs (Fig. 8a, solid line, filled squares). However, even with 1,000 initial SNPs, at least one of the five trials on a given data set was able to correctly identify the 2 SNPs (Fig. 8b), illustrating how repeated trials can further compensate for noise in the fitness function (resulting in 100% success for all five penetrance models with up to 1,000 SNPs, as shown in Fig. 8a, solid line, open squares). For the 0.1 heritability data, the percent of

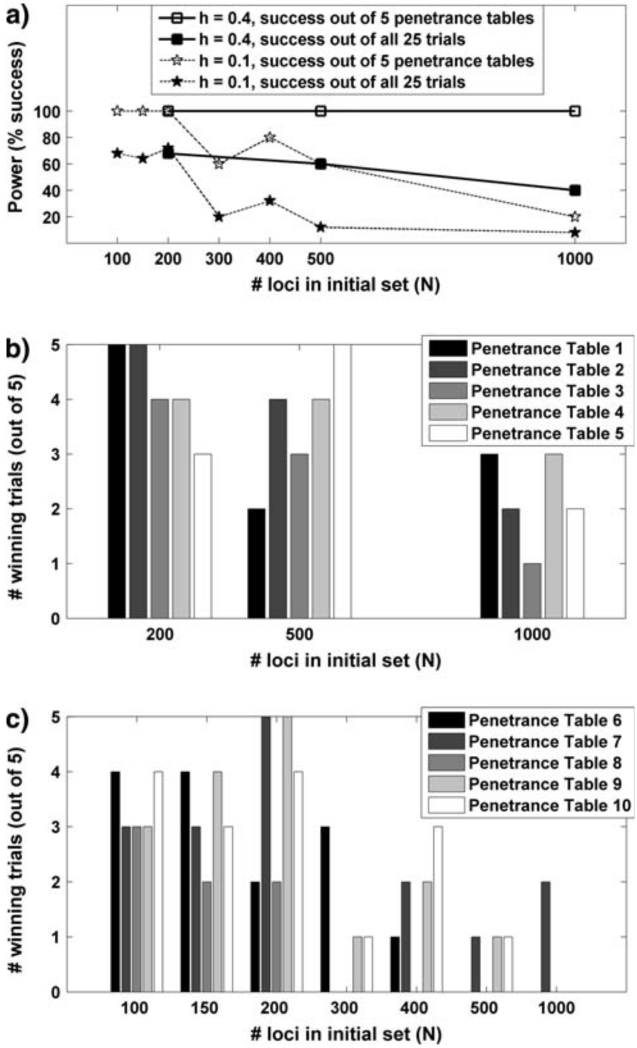


Fig. 8 (**a**) Power (% success) of the algorithm in correctly isolating the two epistatically interacting SNPs from initial sets of varying initial sizes *N* for disease heritability 0.4 (solid lines, squares) and 0.1 (dashed lines, stars), where power is either calculated based on how many of the five penetrance models had at least one successful trial (open markers), or on all 25 trials (five repetitions on each of five penetrance models, filled markers). The same data is shown for the (**b**) 0.4 and (**c**) 0.1 heritability data sets, where individual bars represent the number of successful repetitions out of five on the same data set of a given size, with color coding representing the five distinct penetrance tables used to generate the interacting SNPS in the data sets. Note that the *x*-axis is not uniformly scaled in panel c

successful trials for all penetrance models dropped more rapidly with increasing initial set size *N*, to a minimum of only 8% success with 1000 SNPs (Fig. 8a, dashed line, filled stars). For the lower heritability data, at least one of the five trials on a given data set was able to correctly identify the 2 SNPs from initial set sizes up to

200 SNPs (Fig. 8c), indicating 100% success for all five penetrance models with up to 200 SNPs (Fig. 8a, dashed line, open stars). However, for larger initial set sizes there were some penetrance models for which none of the five trials found the correct two SNPs (Fig. 8c), and at 1000 SNPs only one of the five penetrance models had any successful trials (i.e., yielding only a 20% success rate on the five data sets with 1,000 SNPs, Fig. 8a, dashed line, open stars). These results indicate that some penetrance models may be inherently more difficult than others (e.g., note in Fig. 8c that none of the five trials were successful with penetrance table 8, for initial sets of size 300 or more) and/or that more trials are warranted at lower heritabilities. Overall, these results are better than we would have predicted, given the level of noise in the rough fitness approximator.

6 Discussion and future work

Detecting small sets of epistatically interacting SNPs that can influence an individual's susceptibility to disease is a difficult but important challenge facing bioinformaticists. When smaller subsets of purely epistatically interacting sets of SNPs have no effect, it is not possible to hill-climb from smaller to larger building blocks in constructing these SNP sets. In this work we have presented proof-of-concept for an alternative non-deterministic bounded time evolutionary approach dubbed the "random chemistry" algorithm. This approach enables us to hill-climb from larger to smaller sets of SNPs in only $\theta(\log N)$ fitness evaluations, for an overall time complexity of $\theta(N)$ in the current implementation, where N is the number of candidate SNPs in the initial set. Furthermore, the algorithm is inherently parallelizeable.

Evaluating the fitness of small sets of loci is accomplished using an ROC based metric with built-in cross validation to penalize over-fitting. This fitness function, while fairly accurate, is exponential in the set size and therefore not practical for large sets. The main challenge in implementing the random chemistry algorithm is in finding a reliable and computationally efficient way to approximate which large sets contain the correct SNPs and which don't. Herein we employ an approximate and noisy fitness function based on the ReliefF data mining algorithm. Although this fitness approximator is far from ideal, using it in combination with noise compensation techniques has enabled us to pick out two epistatically interacting SNP loci from up to 1,000 candidate loci, although not surprisingly the success rate declines with declining heritability. Research continues into seeking a more accurate fitness approximator for large sets that will enable us to extend the approach to larger data sets and to lower heritabilities. For example, we are investigating the use of artificial neural networks or support vector machines as more accurate rough fitness approximators, although in preliminary studies these have not performed as well as the ReliefF-based approach.

Regardless of which approximate fitness functions are developed for large sets, there is no question that screening of large sets will be noisy. Consequently, sometimes the most "fit" set will *not* be a true "positive" set, and above some upper limit in size the approximate fitness function will not be able to distinguish "positive" from "negative" sets at all. In the random chemistry approach described earlier (Fig. 1), it is clear that, once a "negative" set is selected, the algorithm will fail, since, once removed, SNPs are never added back into a set. In our current implementation, we attempt to compensate for noisy fitness evaluation of large sets in two ways: (1) by repeating the probabilistic algorithm a number of times on the same data set, selecting the most "fit" solution of all trials, and (2) by saving and merging the top few most fit subsets. However, there are several other promising approaches that we plan to explore. For very large sets, we plan to start screening random sets of sizes near the upper bound of the region in which the approximate fitness function can begin to distinguish "positive" sets, rather than starting with a single parent set containing all the loci. While this would mean that more sets would need to be initially evaluated, the sizes of those sets would have been reduced to a size for which fitness can be more accurately assessed and in a more computationally efficient manner. Additionally, rather than always creating a fixed number of child sets, one could continue generating and testing child subsets until one is generated with a sufficient fitness increase, in an attempt to ensure that the upper (positive) trajectory of the noisy bifurcated fitness landscape is selected. One could also add in the possibility of "repair" mutations, such as a random doubling mutation operator (to try to jump from the negative to the positive trajectory) in addition to the random halving mutation (to try to hill-climb the positive trajectory, once there). Although the latter method would not be bounded in time complexity, it would facilitate working with variable sized sets that stay small enough to be evaluated relatively accurately and efficiently. This may prove to be more computationally efficient and effective for very large-scale genome-wide association studies. In all cases, it may be possible to reduce the size of the initial set of SNPs by pre-filtering based on *a priori* information (e.g., as proposed in [20]), although this approach carries the risk of precluding detection of unexpected associations.

We are currently establishing collaborations with clinicians who are collecting large-scale case-control SNP data sets for various diseases. In addition, a number of such data sets are expected to become publicly available over the next few years. We look forward to applying the random chemistry algorithm to these real data sets. While motivated by the problem of detecting epistatically interacting SNPs that predispose for disease, the proposed random chemistry algorithm could also be applied to any non-linear feature selection problem. In the context of discovering the etiology of various diseases, we plan to extend the algorithm to accommodate candidate epistatic factors other than SNPs, including sex, race, diet, exposure to toxins, or other environmental variables that may affect susceptibility to disease, as these data become available.

Acknowledgments This work was supported, in part, by a pilot award and a graduate research assistantship, funded by DOE-FG02-00ER45828 awarded by the US Department of Energy through its EPSCoR Program and by National Institutes of Health grants AI59694 and LM009012. We thank Joshua Gilbert for his aid in creating the synthetic data sets.

References

1. Barrett, H.H., Myers, K.J.: Foundations of Image Science. John Wiley & Sons, Inc., New Jersey (2004)
2. Culverhouse, R., Suarez, B.K., Lin, J., Reich, T.: A perspective on epistasis: Limits of models displaying no main effect. Am. J. Hum. Genet. **70**, 461–471 (2002)
3. Glazier, A.M., Nadeau, J.H., Aitman, T.J.: Finding genes that underlie complex traits. Science **298**, 2345–2349 (2002)
4. Hirschhorn, J.N., Daly, M.J.: Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. **6**, 95–108 (2005)
5. Hoh, J., Wille, A., Ott, J.: Trimming, weighting, and grouping SNPs in human case-control association studies. Gen. Res. **11**, 2115–2119 (2001)
6. International HapMap Consortium: The international HapMap project. Nature **426**, 789–796 (2003)
7. International human genome sequencing consortium: Initial sequencing and analysis of the human genome. Nature **409**, 860–921 (2001)
8. International SNP map working group: A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409**, 928–933 (2001)
9. Kauffman, S.: At Home in the Universe: The Search for the Laws of Self-Organization and Complexity. Oxford Univ. Press, USA (1996)
10. Kruglyak, L., Nickerson, D.A.: Variation is the spice of life. Nat. Genet. **27**, 234–236 (2001)
11. Lucek, P.R., Ott, J.: Neural network analysis of complex traits. Gen. Epidem. **14**, 1101–1106 (1997)
12. McKinney, B.A., Reif, D.M., Ritchie, M.D., Moore, J.H.: Machine learning for detecting gene-gene interactions. Appl. Bioinformatics **5**, 77–88 (2006)
13. Merikangas, K.R., Low, N.C.P, Hardy, J.: Understanding sources of complexity in chronic diseases—the importance of integration of genetics and epidemiology. Int. J. Epidemiol. **35**, 590–592 (2006)
14. Moore, J.H.: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum. Hered. **56**, 73–82 (2003)
15. Moore, J.H.: Computational analysis of gene-gene interactions in common human diseases using multifactor dimensionality reduction. Expert Rev. Mol. Diagn. **4**, 795–803 (2004)
16. Moore, J.H., Gilbert, J.C., Tsai, C.T., Chiang, F.T., Holden, T., Barney, N, White, B.C.: A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J. Theor. Biol. **241**, 252–261 (2006)
17. Moore, J.H., Ritchie, M.D.: The challenges of whole-genome approaches to common diseases. JAMA **291**, 1642–1643 (2002)
18. Moore J.H., White B.C.: Genome-wide genetic analysis using genetic programming: The critical need for expert knowledge. In: Riolo, R.L., Soule, T., Worzel, B. (eds.) Genetic Programming Theory and Practice IV. Springer, New York (2006)
19. Moore J.H., White B.C.: Tuning ReliefF for genome-wide genetic analysis. In: Rajapakse, J.C. et al. (eds.) Lecture Notes in Computer Science, 4447, pp. 166–175, Springer, New York (2007)
20. Ott, J., Hoh, J.: Statistical multilocus methods for disequilibrium analysis in complex traits. Hum. Mut. **17**, 285–288 (2001)
21. Peltonen, L., McKusick, V.A.: Dissecting human disease in the postgenomic era. Science **291**, 1224–1229 (2001)
22. Proulx, S.R., Phillips, P.C.: The opportunity for canalization and the evolution of genetic networks. Am. Nat. **165**, 147–162 (2005)
23. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabasi, A.-L.: Hierarchical organization of modularity in metabolic networks. Science **297**, 1551–1555 (2002)
24. Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., Moore, J.H.: Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. Am. J. Hum. Gen. **69**, 138–147 (2001)
25. Robnik-Sikonja, M., Konenenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn. **53**, 23–69 (2003)
26. Syvanen, A.C.: Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat. Rev. Genet. **2**, 930–942 (2001)
27. Thornton-Wells, T.A., Moore, J.H., Haines, J.L.: Genetics, statistics and human disease: analytical retooling for complexity. Trends Genet. **20**, 640–647 (2004)

28. Tong, A.H. et al.: Global mapping of the yeast genetic interaction network. Science **303**, 808–813 (2004)
29. Venter, J.C., et al.: The sequence of the human genome. Science **291**, 1304–1351 (2001)
30. Wang, W.Y., Barratt, B.J., Clayton, D.G., Todd, J.A.: Genome-wide association studies: theoretical and practical concerns. Nat. Rev. Genet. **6**, 109–118 (2005)
31. White, B.C., Gilbert, J.C., Reif, D.M., Moore, J.H.: A statistical comparison of grammatical evolution strategies in the domain of human genetics. In: Corne, D. et al (eds.) Proc. of the IEEE Congress on Evol. Computing pp. 676–682. IEEE Press, Edinburgh, UK, (2005)