BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 16 2007, pages 2113–2120
doi:10.1093/bioinformatics/btm317

Genetics and population analysis

Evaporative cooling feature selection for genotypic data involving interactions

B.A. McKinney[1,*], D.M. Reif[2], B.C. White[3], J.E. Crowe, Jr.[4] and J.H. Moore[3]

[1]Department of Genetics, University of Alabama School of Medicine, Birmingham, AL 35294, [2]National Center for Computational Toxicology, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711, [3]Computational Genetics Laboratory, Department of Genetics, Dartmouth Medical School, Lebanon, NH 03756 and [4]Program in Vaccine Sciences, Departments of Microbiology and Immunology, and Pediatrics, Vanderbilt University Medical Center, Nashville, TN 37232, USA

Received and revised on May 2, 2007; accepted on June 6, 2007

Advance Access publication June 22, 2007

Associate Editor: Keith Crandall

ABSTRACT

Motivation: The development of genome-wide capabilities for genotyping has led to the practical problem of identifying the minimum subset of genetic variants relevant to the classification of a phenotype. This challenge is especially difficult in the presence of attribute interactions, noise and small sample size.

Methods: Analogous to the physical mechanism of evaporation, we introduce an evaporative cooling (EC) feature selection algorithm that seeks to obtain a subset of attributes with the optimum information temperature (i.e. the least noise). EC uses an attribute quality measure analogous to thermodynamic free energy that combines Relief-F and mutual information to evaporate (i.e. remove) noise features, leaving behind a subset of attributes that contain DNA sequence variations associated with a given phenotype.

Results: EC is able to identify functional sequence variations that involve interactions (epistasis) between other sequence variations that influence their association with the phenotype. This ability is demonstrated on simulated genotypic data with attribute interactions and on real genotypic data from individuals who experienced adverse events following smallpox vaccination. The EC formalism allows us to combine information entropy, energy and temperature into a single information free energy attribute quality measure that balances interaction and main effects.

Availability: Open source software, written in Java, is freely available upon request.

Contact: brett.mckinney@gmail.com

1 INTRODUCTION

Evaporative cooling (EC) was first proposed as an experimental technique for cooling a collection of atoms (Hess, 1986). The method consists of the selective removal of atoms in the high energy tail of the thermal distribution and the collisional equilibration of the remaining atoms. The combination of atom selection and collisions increases the phase–space density and can greatly reduce the temperature of a gas. This method turned out to be a powerful technique and, when combined with laser cooling, played a key role in reaching the ultra-cold temperatures needed to achieve Bose–Einstein condensation and Fermi degeneracy of trapped atomic alkali gases.

The development of genome-wide capabilities for genotyping has led to the problem known in data mining and machine learning as the curse of dimensionality, which refers to the rapid increase in search space volume as the dimensionality (e.g. the number of genetic variants) in the data set becomes large compared to the number of samples. Just as EC of an atomic gas increases the phase–space density of the gas by removing the most energetic atoms, the purpose of EC feature selection is to increase the feature–space density by removing the genetic variants with the most noise. In machine learning, the term noise refers to attributes (e.g. genetic variants) that are irrelevant to the classification of samples in a data set. Feature selection is the separation of the DNA variations that are functionally related to the phenotype from noisy variations, usually for the purpose of classification. Thus, the goal of EC feature selection is to increase the feature–space density by evaporating the least informative attributes from the data set.

Essential to the EC algorithm is our definition of an attribute quality score that couples mutual information (MI) and Relief-F into an information score analogous to a thermodynamic free energy. MI is a function based on information entropy that measures the mutual dependence between two variables, in this case, between a DNA variant and a phenotype. One potential drawback of this correlation measure occurs in the presence of strong attribute interactions, or epistasis. In the most extreme case of an exclusive-OR interaction, the association of a variant with the phenotype can only be detected in the presence of one or more other variants (McKinney *et al.* 2006a). In this case, MI would not accurately discriminate the functional genetic variants from noise variables because MI assumes independence between the attributes. Relief-F, on the other hand, is a heuristic attribute quality measure that is able to identify functional attributes in data sets that include interactions with other attributes that influence their dependence on the phenotype. However, Relief-F is sensitive to the presence of

*To whom correspondence should be addressed.

noise attributes, which when added to a data set cause Relief-F scores of some relevant attributes to worsen (Draper *et al.*). To overcome the bias caused by the context of noise attributes, iterative strategies are necessary (Draper *et al.*; Moore and White, 2007). Motivated by information theory and thermodynamics, we leverage the advantages of MI and Relief-F by coupling them by an information temperature into a single information free energy measure as a part of the feature evaporation method.

To achieve ultra-cold temperatures in inhomogeneous atomic gases, experimenters gradually lower the trap depth, which increases the elastic collision rate and accelerates evaporation. In the EC method, the attributes are in a fictitious potential energy trap whose depth is decreased so that the most energetic attributes (i.e. those attributes with the least relevance to the phenotype) evaporate from the gas of interacting attributes according to the classification accuracy of the remaining attributes. After the gas of attributes is allowed to re-equilibrate by re-calculating the information relevance scores, the algorithm yields a higher probability that the gas has a lower information free energy and temperature. This process of forced evaporation is repeated until a collection of attributes with the optimum information free energy and information temperature is obtained.

The article is outlined as follows. In Section 2, the formalism of the EC algorithm is derived. In Section 3, we provide the thermodynamic motivation for the information free energy attribute quality measure, which is constructed in Section 3.1. In Section 4, the method is applied to simulated data then to experimentally observed single nucleotide polymorphism (SNP) data of individuals who experienced adverse events (AEs) following smallpox immunization.

2 EVAPORATIVE COOLING

EC of trapped atoms is based on the mechanism that the atoms with energies larger than the potential barrier escape and leave the atoms remaining within the trap to rethermalize. However, since the natural evaporation rate diminishes exponentially, $\exp(-U_o/kT)$, the depth U_o of the trap has to be decreased in order to force EC to continue at a reasonable rate. An evaporation step of the EC algorithm is illustrated in Figure 1. We simulate lowering of the potential well in EC by removing ΔN of the 'hottest' (i.e. least informative) attributes, represented as dark circles, from the attribute gas to the thermal (noise) fraction using classification accuracy to select which attributes to evaporate. The remaining system has a lower energy and temperature. In this section, we use a simple physical model of EC to derive an acceptance threshold and simulation temperature that adapt to the number of attributes in the cooled system.

Consider a gas of N atoms with an average energy $\langle E\rangle_N$. We assume the atoms are distinguishable and we order them by increasing energy and remove the ΔN atoms with the highest energy. For the purposes of this derivation, we consider this removal to be deterministic but for real gases the process is stochastic and this is approximated in the optimization algorithm through an acceptance threshold (Gunter and Tobias, 1990). The average energy of the atoms remaining in the colder gas after an evaporation step is given by

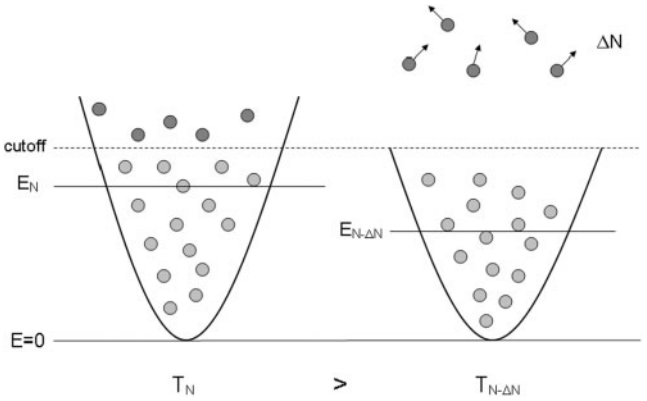

Fig. 1. Illustration of EC feature selection. Attributes are depicted as atoms (circles) interacting in an external potential that traps the attributes in the system. Evaporation of the darker, higher energy attributes is forced by lowering the height of the trap from its initial height on the left. Evaporation leaves behind a re-equilibrated system of attributes with lower energy $E_{N-\Delta N}$ and temperature $T_{N-\Delta N}$ on the right.

$$\langle E\rangle_{N-\Delta N} = \langle E\rangle_N - \frac{1}{N}\sum_{i=N-\Delta N}^{N} E_i = \langle E\rangle_N - \frac{\Delta N}{N}\langle E'\rangle_{\Delta N}, \quad (1)$$

where $\langle E'\rangle_{\Delta N}$ is the average energy of the thermal atoms that are removed from the gas by evaporation. We rearrange Equation (1) as

$$\frac{\langle E\rangle_{N-\Delta N} - \langle E\rangle_N}{\Delta N} = \frac{1}{N}\langle E'\rangle_{\Delta N'} \quad (2)$$

take the continuum limit ($N \gg 1$), introduce the average energy densities $\langle\varepsilon\rangle$ and $\langle\varepsilon'\rangle$ of the cooled gas and thermal atoms, respectively, to obtain

$$\frac{d\langle\varepsilon\rangle}{dN} = \frac{1}{N}\langle\varepsilon'\rangle. \quad (3)$$

We assume that the energy of the evaporated atoms is some multiple ($\eta > 1$) of the average energy:

$$\langle\varepsilon'\rangle = \eta\langle\varepsilon\rangle_N, \quad (4)$$

which when inserted into Equation (3) gives

$$\frac{d\langle\varepsilon\rangle}{dN} = \frac{\eta}{N}\langle\varepsilon\rangle \quad (5)$$

or

$$\frac{\ln(\langle\varepsilon\rangle)}{\ln(N)} = \eta. \quad (6)$$

Then the average energy per atom is given by

$$\frac{\langle\varepsilon\rangle_N}{\langle\varepsilon\rangle_{N_o}} = \left(\frac{N}{N_o}\right)^{\eta}, \quad (7)$$

where N_o is the number of atoms prior to an evaporation step. The quantity η is an adjustable parameter in the EC algorithm which we fix: $\eta = 2$. This parameter is related to the rate of evaporation. We choose a value corresponding to relatively slow evaporation to guard against premature convergence to a suboptimal collection of attributes.

For a physical gas of confined particles, the relationship between the average energy and the temperature depends on the details of the cooling process. If we were to treat our information system as a weakly interacting gas, to a first approximation the ratio of the temperatures at evaporation steps k and $k-1$ would be

$$\frac{T_k}{T_{k-1}} = \frac{\langle \varepsilon \rangle_{N_k}}{\langle \varepsilon \rangle_{N_{k-1}}} = \left(\frac{N_k}{N_{k-1}} \right)^{\eta}, \tag{8}$$

or since we have $N_k = N_{k-1} - \Delta N_k$ for a cooling step k, the temperature-lowering update equation would be

$$T_k = \left(1 - \frac{\Delta N_{k-1}}{N_k} \right)^{\eta} T_{k-1}. \tag{9}$$

However, since we lack mechanistic details of the evaporation of an abstract gas of attributes, we use cross-validated (CV) classification accuracy to estimate the 'information temperature' and to remove the least informative attributes in an evaporation step. Evaporation of an uninformative attribute is proposed based on the attribute's information free energy quality score (defined in the following section), and the proposed evaporation is accepted or rejected based on whether or not the classification accuracy improves for the remainder of attributes. One can think of EC as a wrapper approach to feature selection. The temperature — which couples the two terms in the quality score — is initialized to unity. For each subsequent evaporation step, a local search about the previous best temperature is performed to maximize the classification accuracy.

3 INFORMATION FREE ENERGY

The presence of a temperature in the EC formalism suggests a relation between information entropy and energy. We will more fully interpret the meaning of the energy in the next section, but for now we treat it mathematically as an information energy associated with each genetic variant. In this section, we derive the energy that combines information energy and entropy, which we will then use to develop an attribute quality score.

We begin by maximizing the information entropy of the N attributes:

$$H = -\sum_{i=1}^{N} p_i \ln p_i, \tag{10}$$

where p_i is the probability of attribute i having information energy ε_i. We use the maximum entropy principle (MEP) to find the least biased probability distribution associated with the information energy, which is the distribution that maximizes H under the constraints imposed by prior information (Jaynes, 1957). The reasoning behind maximizing the entropy is that choosing a probability with entropy lower than the maximum subject to the constraints would assume information not contained in the constraints while a probability with a higher entropy would violate the constraints imposed by the prior information.

For our simulated evaporation problem, we use the method of Lagrange multipliers to maximize the entropy H with the constraints

$$\sum_{i=1}^{N} p_i = 1 \text{ and} \tag{11}$$

$$\sum_{i=1}^{N} p_i \varepsilon_i = k T_{\text{info}}. \tag{12}$$

The first constraint ensures that the probability field is normalized and the second allows us to assign a characteristic information temperature to the system through the average energy per attribute of the energy spectrum ε_i. For simplicity, we have assumed the characteristic average energy per attribute is the ideal gas value kT_{info}, where T_{info} is the information noise temperature. We have made statistical mechanics assumptions similar to a canonical ensemble in which the particle number is constant. A more microscopically detailed treatment might consider a grand canonical ensemble in which the particle number fluctuates due to evaporation and the chemical potential is introduced. However, we are not describing the thermodynamics of evaporation, but rather, the thermodynamics of the system of attributes between discrete evaporation steps. Change in energy with respect to the number of attributes is calculated empirically.

To maximize Equation (10) subject to the above constraints, we introduce Lagrange multipliers λ_1' and λ_2' corresponding to the first two constraints and require for all i from 1 to N that

$$\frac{\partial}{\partial p_i} \left[-\sum_{n=1}^{N} p_n \ln p_n + \lambda_1' \left(\sum_{n=1}^{N} p_n - 1 \right) + \lambda_2' \left(\sum_{n=1}^{N} p_n \varepsilon_n - k T_{\text{info}} \right) \right] = 0, \tag{13}$$

which yields

$$p_i = e^{-\lambda_1} e^{-\lambda_2 \varepsilon_i}, \tag{14}$$

where we defined the new parameters $\lambda_1 = 1 - \lambda_1'$ and $\lambda_2 = -\lambda_2'$. Constraint (11) determines λ_1 so that

$$p_i = \frac{e^{-\lambda_2 \varepsilon_i}}{Z}, \tag{15}$$

where $Z = \Sigma_i \, e^{-\lambda_2 \varepsilon_i}$ is a normalization constant referred to as the partition function in statistical mechanics. Constraint (12) determines λ_2 in terms of the temperature:

$$\frac{1}{Z} \sum_{i=1}^{N} \varepsilon_i e^{-\lambda_2 \varepsilon_i} = k T_{\text{info}}. \tag{16}$$

There are various ways to relate temperature to λ_2, but rather than introduce thermodynamic potentials for this abstract information system, we simply take the continuum limit

($N \gg 1$) so that we can write the sums in Equation (16) as integrals:

$$\frac{1}{Z}\int_0^\infty \varepsilon e^{-\lambda_2 \varepsilon} d\varepsilon = kT_{\text{info}}, \tag{17}$$

where $Z = \int e^{-\lambda_2 \varepsilon} d\varepsilon$, which is easily integrated to give $\lambda_2 = 1/kT_{\text{info}}$. Finally, from Equation (15) the probability of state i is

$$p_i = \frac{e^{-\varepsilon_i / kT_{\text{info}}}}{Z}, \tag{18}$$

which may be rearranged to give the energy of state i:

$$\varepsilon_i = -kT_{\text{info}}(\ln p_i + \ln Z). \tag{19}$$

One can immediately calculate the average energy in terms of the entropy by multiplying both sides of Equation (19) by p_i and summing over i:

$$\langle E \rangle = -kT_{\text{info}}(H + \ln Z), \tag{20}$$

which can be rewritten with the definition of the free energy, $F = -kT \ln Z$, as

$$F = \langle E \rangle - kT_{\text{info}} H. \tag{21}$$

Next, we will define an information relevance score for attributes in the EC algorithm based on Equation (21). Instead of the entropy, we use a relative information entropy-based quantity, and we define our information potential energy in terms of Relief-F. We tune T_{info} by 5-fold CV classification accuracy of the collection of attributes with the best information free energy scores.

3.1 Evaporative cooling attribute quality measure

Recursive elimination of features (Relief) is a heuristic machine-learning method for estimating the quality of attributes according to their ability to separate samples into classes (phenotypes). There is a family of Relief algorithms, so we will describe the original version introduced in Kira and Rendell (1992) but in the context of genotypic data. Consider a set of genetic variants (e.g. SNPs) $\{G\}$, where each genetic variant g_i in this set can be in one of the states {0, 1, 2} corresponding to the homozygous dominant, homozygous recessive and heterozygous genotypes. In Relief, the weight of each attribute g_i is initially set to zero ($W[g_i] = 0$) and for randomly selected samples (or for all samples if desired) the nearest hit and miss are computed with the chosen distance function and W[g_i] is recursively updated according to how well the attribute can separate near hits and misses. The nearest hit and miss calculations are what allow Relief to handle attribute interactions because the calculation of nearest neighbors involves distances in the space of all attributes as opposed to just one attribute at a time. The selection of the *nearest* hit/miss is also crucial to the success of finding strong attribute dependencies. For a given sample S with nearest hit H and nearest miss M, the following equation is used to update the weight of each SNP g_i:

$$W[g_i] = W[g_i] - \text{diff}(g_i, S, H)/m + \text{diff}(g_i, S, M)/m. \tag{22}$$

This is repeated for m samples selected randomly or exhaustively. For SNP g_i, the difference function between samples S_j and S_k is

$$\text{diff}(g_i, S_j, S_k) = \begin{cases} 0, & \text{if genotype}(g_i, S_j) = \text{genotype}(g_i.S_k) \\ 1, & \text{otherwise} \end{cases}, \tag{23}$$

where genotype (g, S) means the genotype of SNP g for sample S. Division by m in Equation (22) ensures that the weight of each attribute lies between −1 and 1.

Equation (22) rewards attributes that yield a large separation between the given sample and its nearest sample from the other class (misses) and penalizes attributes that give large separations between the given sample and the nearest sample from the same class (hits). For example, if the separation of a sample from its nearest hit is the same as its separation from its nearest miss then the contribution to the attribute's weight is zero because it does not contribute to the classification of the sample. In our algorithm, we use Relief-F, an extension of Relief that enables it to handle noisy and incomplete data sets and to deal with multi-class problems. For more details on Relief-F, see Kononenko, (1994). We use a Relief-based feature weighting algorithm in our objective function because of its ability to handle attribute interactions and, for reasons discussed next, because the weights can be interpreted as estimates of probabilities.

Information entropy is the amount of uncertainty in a discrete probability distribution. The probabilities must be defined with respect to some quantity, which in our case will be an information energy spectrum ε_i. Information energy, introduced in mathematical statistics by Onicescu and Stefanescu (1979) (also called the Gini impurity function) for a discrete probability field (p_1, p_2, …, p_k) is

$$\langle E \rangle = \sum_{i=1}^{k} p_i^2. \tag{24}$$

The similarity between the Relief weight and the Gini index has been discussed previously (Kononenko and Robnik-Sikonia, 1996). If we equate Equation (24) with the average energy of an attribute $\langle E \rangle = \Sigma_i p_i \varepsilon_i$, we observe that the microstate information energy of attribute i is the attribute's probability: $\varepsilon_i = p_i$. However, beyond the probability of finding a particular genotype at a SNP locus, we are interested in the ability of a SNP to separate samples into classes. And, in fact, the Relief weight approximates this through a difference in probabilities:

$$\begin{aligned} W[g_i] = {} & P(\text{equal value of } g_i | \text{same class}) \\ & - P(\text{equal value of } g_i | \text{different class}). \end{aligned} \tag{25}$$

Further, we want to define our information energy ε_i of attribute i in such a way that information is gained by evaporating noise attributes, thereby lowering the average information noise of the remaining attributes. An energy that satisfies this requirement is the following relative Relief quantity:

$$\varepsilon_i = \frac{W_{\max} - W[g_i]}{W_{\max}}, \tag{26}$$

where $W[g_i]$ is the Relief-F weight of SNP i. This transformation, involving the largest Relief-F weight $W_{\max}$ in the attribute ensemble, makes the information energy ε_i lie between 0 and 1.

Equation (26) is the contribution to the information quality score (see below) analogous to energy whose average would be $\langle E \rangle$ in Equation (21). The contribution analogous to entropy comes from an MI-based quantity. MI is a measure of the independence of two discrete random variables X and Y. For our application, we find it appropriate to use a scaled version, known as the uncertainty coefficient (Press *et al.*, 1988):

$$\mu(X, Y) = \frac{H(X) + H(Y) - H(X, Y)}{H(X)}, \tag{27}$$

where $H(\cdot)$ and $H(\cdot\,, \cdot)$ are the entropy of the marginal and joint probability distributions, respectively. Division by $H(X)$ causes μ to lie between 0 and 1, yielding a value of 0 when X and Y are independent and 1 when X completely predicts Y. We are interested in the scaled mutual information (SMI) $\mu_i = \mu(g_i, C)$ between genetic variant $g_i \epsilon \{G\}$ and the phenotype/class variable C . The variable C can take on as many discrete states as there are classes or phenotypes. In certain contexts, Equation (27) is sometimes referred to as the information gain ratio.

The total Relief-F energy $\sum_{i=1}^{|G|} \varepsilon_i$ of a collection of attributes $\{G\}$ with cardinality $|G|$ is analogous to the heat of a thermodynamic system. In such a system, the amount of heat free to do useful work is the free energy of Equation (21), which balances the energy and entropy of the system. Borrowing similar language, we wish to estimate the amount of free information available in a system of attributes to separate samples into classes. However, instead of using energy and entropy directly, it is necessary to use relative versions of these quantities because we are dealing with a classification problem. Instead of entropy, we use a SMI which is the entropy of the joint distribution p_{XY} relative to the product distribution of $p_X \cdot p_Y$. Instead of information energy, we use the relative attribute energy $\varepsilon_{\ i}$ in Equation (26), which could be interpreted as a potential information energy. These observations motivate the following information relevance measure for a given attribute i :

$$f_i = \varepsilon_i - T_{\text{info}} \mu_i, \tag{28}$$

where we have set the information constant k (analogous to Boltzmann's constant) to unity. The information free energy for the entire collection of attributes is $F = \sum_i^{|G|} f_i$, where $|G|$ is the number of SNPs in the ensemble. At the initial evaporation step, ε_i and μ_i are equally weighted ($T_{\text{info}} = 1$), and this coupling weight is tuned as attributes are evaporated. At each EC step, a local grid search is performed in a neighborhood Δ of the best temperature from the previous step $T_{\text{info}} \pm \Delta$. We train a naïve Bayes classifier via 5-fold CV using the set of attributes with the worst attribute removed [attributes ranked by Equation (28)] for a given local trial temperature. The temperature trial that yields the highest CV accuracy results in the removal from the system of the corresponding worst attribute for that evaporation step. The search window Δ is the range $(-0.3, 0.3)$.

Table 1. Penetrance for two-locus interaction model simulations. The three models of class (A) (analysis results in Fig. 3) have a small main effect. These models are labeled by heritability (%). The models of class (B) (analysis results in Fig. 2) are strictly epistatic (i.e. no main effect) and labeled by name: XOR and ZZ. Minor allele frequencies are 0.5.

Genotype combinations ↓		Interaction model classes (A)			(B)	
		5%	10%	15%	XOR	ZZ
AA	BB	0.082	0.028	0.198	0.000	0.000
	Bb	0.004	0.030	0.081	1.000	0.000
	bb	0.287	0.401	0.000	0.000	1.000
Aa	BB	0.081	0.082	0.081	1.000	0.000
	Bb	0.082	0.083	0.087	0.000	0.500
	bb	0.081	0.031	0.003	1.000	0.000
aa	BB	0.083	0.092	0.000	0.000	1.000
	Bb	0.081	0.082	0.000	1.000	0.000
	bb	0.025	0.019	0.436	0.000	0.000

4 RESULTS

The EC algorithm was implemented in Java using the Weka machine-learning library (Witten and Frank, 2005). To assess the performance of the EC information free energy feature selection algorithm, we compare the ability of EC, χ^2, SMI and Relief-F [with and without iterative attribute removal (Draper *et al.*; Moore and White, 2007)] to identify functional loci in gene–gene interaction data simulated by the genomeSIM package (Dudek *et al.*, 2006). The data, simulated based on the interaction models in Table 1, involve 1500 SNPs in 500 cases and controls with 100 replicates for each of the five genetic models. Figures 2 and 3 show the percentage of times out of 100 simulations that each method is able to detect the two functional loci in the top 100 selected variables out of 1500. The functional loci in all simulations interact to determine the phenotype. In Figure 3, the simulations contain a weak main effect, and the heritability/odds ratio combinations are 5%/2, 10%/3 and 15%/4. In Figure 2, the simulated interactions are purely epistatic, i.e. neither functional locus has an independent (main) effect. The strictly epistatic genetic models are categorized by the names exclusive-OR (XOR) (Li and Reich, 2000) and zig-zag (ZZ) (Frankel and Schork, 1996).

EC outperforms the other methods under all conditions tested. For the detection of interactions with a small main effect (Fig. 3), SMI and χ^2 perform nearly as well as EC for high heritability models, but EC has a large advantage for lower, more realistic heritability models and purely epistatic models. This advantage originates from the removal of noise variables in the Relief-F contribution to EC because iterative Relief-F (IRF) also shows higher detection than SMI and χ^2 for lower heritability. As expected, Relief-F performs better with iteration under all conditions but does not perform as well as EC. For strictly epistatic interactions (Fig. 2), SMI and χ^2 perform poorly due to their assumption of variable independence, while EC continues to perform well by improving upon the ability of Relief-F to detect interacting attributes.

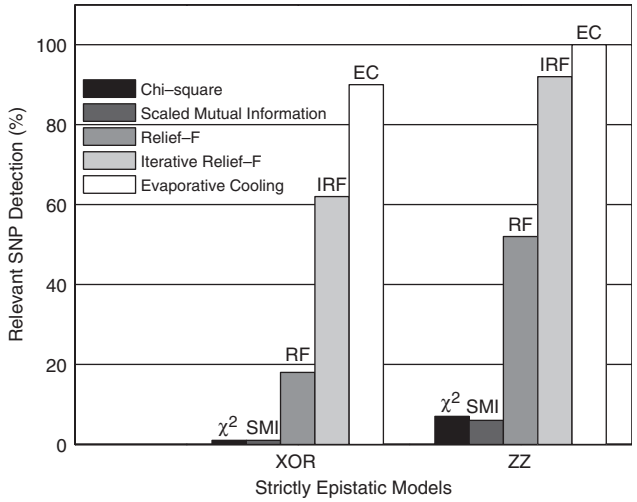

Fig. 2. Percentage of times that EC, χ^2, SMI, Relief-F (RF) and IRF detect both interacting loci relevant to the phenotype in the top 100 out of 1500 SNPs. Each method is applied to 100 replicates for both interaction models. Gene–gene interactions are purely epistatic (no main effect) using exclusive-OR (XOR) and zig-zag (ZZ) models. Simulated models are summarized in Table 1.

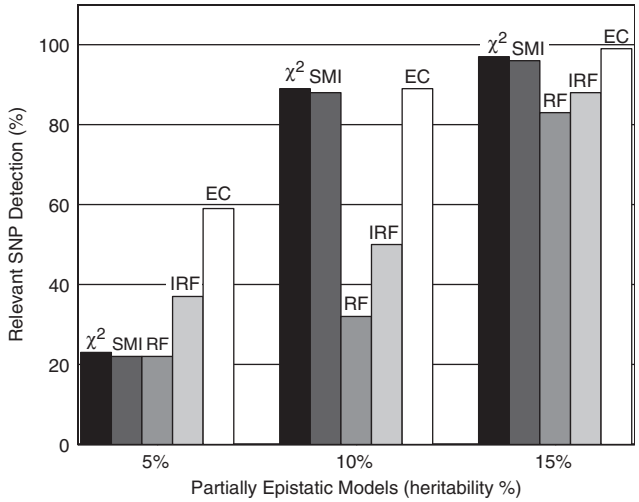

Fig. 3. Same methods compared as in Figure 2. Each method is applied to 100 replicates for each of the three interaction models with weak main effects. Three heritabilities are simulated (5, 10, and 15%) with corresponding odds ratios 2, 3 and 4. Simulated models are summarized in Table 1.

4.1 Smallpox vaccine adverse events

To demonstrate the usefulness of EC, we compare the top SNPs identified by EC in a real data set with associations of serum cytokine levels found previously in the same cohort of smallpox vaccinees some of whom experienced AEs (McKinney *et al*., 2006b). Vaccine, study subjects, study design and quantification of serum cytokine levels are described in detail in the protein array study in (McKinney *et al*. (2006b). For the present study, genotyping was performed at the core genotyping facility of the National Cancer Institute (NCI) in Gaithersburg, Maryland, USA. Genotypes were generated using the IlluminaTM GoldenGate assay technology. Of the 1536 SNPs assayed, a total of 1442 genotypes passed quality control filters. For validation purposes, we only considered subjects for whom proteomic data was also available. There were a total of 61 subjects for whom both genetic and proteomic data were gathered. Of these individuals, 16 experienced a systemic AE and 45 experience no AE. Systemic AEs included fever, generalized rash and lymphadenopathy. While other AEs were noted, only systemic AEs were considered in this study, since we expected these to be associated more strongly with serum cytokine activity than we would expect for an AE expressed only at the site of inoculation.

Table 2 shows the top 100 SNPs out of the original 1442 selected by EC for the smallpox AE phenotype. SNPs highlighted bold were also found by both SMI and Relief-F in their top 100 lists, while italicized SNPs were found by one but not both. The EC analysis identified polymorphisms in CSF-3 (G-CSF), IL-4 and IL-10, which were also found to be differentially expressed in our proteomic analysis in McKinney *et al*. (2006b). Additional cytokines were found to be significantly associated with AEs in our previous analysis, but SNPs in many of these genes were not part of the panel used in the present study. We also expect to be able to identify polymorphisms in additional genes because many genes were scanned that were not assayed in our proteomic analysis.

5 DISCUSSION

One of the outstanding challenges in the genomic era is to identify genetic variations associated with a given phenotype, such as disease or adverse event status. This challenge is made more difficult by the presence of noise, heterogeneity and attribute interactions (McKinney *et al*., 2006a; Moore, 2003; Thornton-Wells *et al*., 2004). To deal with these problems, machine learning and other heuristics are needed as traditional statistical approaches are often lacking. However, machine-learning classifiers may perform poorly on genome-wide data due to the large number of genetic variations that are irrelevant to the classification of the phenotype. Thus, feature selection wrappers and filters are needed to amend the performance of classifiers. For example, the multifactor dimensionality reduction (MDR) classifier for detecting gene–gene interactions in genotypic data should benefit from a feature selection filter when applied on a genome-wide scale (Moore, 2006). In addition, a useful feature selection method must account for interactions; otherwise, it may filter out genetic variations that are important to classification of the phenotype. Our simulation results suggest that in certain situations it is better to use mutual information (e.g. when there is a main effect) while at other times better results are achieved with Relief-F (e.g. when there is stronger attribute interaction or low heritability). For real data, one does not know which is the case; however, both are likely important. Using information theory and a thermodynamic heuristic, we combined Relief-F and mutual

Table 2. Top 100 SNPs (out of 1442) identified by EC feature selection as relevant to smallpox vaccine-related adverse events. SNPs are ranked by their final evaporated information free energy scores Equation (28). Bold SNPs were identified in the top 100 by SMI and Relief-F in separate analyses, whereas italicized SNPs were identified by one but not both.

Rank	SNP	Rank	SNP	Rank	SNP
1	**MBL2_03**	34	EGF_04	67	*NFKB1_01*
2	**SLC6A3_14**	35	*AKR1C3_19*	68	FANCA_02
3	CTNNB1_16	36	*EPHX1_18*	69	MTRR_22
4	CTNNB1_03	37	*VCAM1_38*	70	*FASLG_01*
5	CTNNB1_17	38	ERCC3_04	71	CYP1A1_81
6	CTNNB1_21	39	*PIN1_21*	72	CYP24A1_05
7	**CCR2_06**	40	OPRD1_03	73	CETP_08
8	CTNNB1_15	41	**CYP2E1_02**	74	**TSG101_36**
9	**CCR2_02**	42	ERCC3_02	75	*NFKB1_21*
10	CTNNB1_11	43	**NFKB1_14**	76	IGF2AS_04
11	CTNNB1_07	44	*APOB_21*	77	MSH3_29
12	CTNNB1_14	45	RAD51_01_2	78	*AKR1C3_35*
13	CTNNB1_08	46	*AKR1C3_01*	79	*ABCA1_12*
14	CTNNB1_02	47	**TSG101_07**	80	BRIP1_05
15	**BLM_02**	48	**TSG101_40**	81	**IL2_01**
16	**CTH_13**	49	*CTH_07*	82	HTR1B_07
17	CTNNB1_19	50	*GATA3_46*	83	GHR_21
18	**APAF1_04**	51	**APOA4_02**	84	ATP1B2_04
19	**IL2_03**	52	**MSH3_02**	85	GPX3_28
20	**IL4_10**	53	**MSH3_07**	86	NQO1_15
21	**CTH_03**	54	CDK5_16	87	DIO1_05
22	**FZD7_20**	55	RERG_47	88	SCUBE2_02
23	**IL4_01**	56	*IL10_01*	89	FANCA_03
24	**IL4_11**	57	*IGF1_46*	90	**TSG101_28**
25	*RXRA_01*	58	RB1CC1_40	91	FANCA_28
26	*ERCC5_01*	59	RXRB_11	92	AKR1C3_31
27	**BLM_25**	60	RNASEL_02	93	BRIP1_02
28	GHR_79	61	KRAS_11	94	**SHBG_01**
29	*CTLA4_25*	62	*CDK4_01*	95	CYP1B1_28
30	**IL4_03**	63	*MTRR_05*	96	STK6_02
31	**CTH_10**	64	SLC6A18_13	97	IGF1_24
32	**HSD17B4_19**	65	CYP1A1_78	98	CSF3_02
33	INSR_07	66	MYNN_01	99	GSK3B_08
				100	GSK3B_20

information into a composite attribute importance measure that compensates for a variety of effects.

Our composite attribute importance measure, which we refer to as information free energy, helps to identify interacting attributes as well as main-effect attributes. However, feature selection methods that can detect interacting attributes are potentially susceptible to the presence of noise in data sets with a large number of attributes. To remove noise while minimizing the loss of functional attributes, we introduced a feature selection algorithm called EC based on a mechanism analogous to cooling a cup of coffee by blowing on its surface. This mechanism involves repeated forced evaporation, or removal, of the least relevant attributes to maximize the classification accuracy of the final collection of selected features. We can interpret the information temperature by way of another thermodynamic analogy, namely, the resistance to the flow of ions (information) through a gas of neutral atoms. If we apply an electric field to this gas, the ions will begin to accelerate toward one wall of the container, but the presence of neutral atoms (noise attributes) in the gas will cause collisions with the ions that degrade the transmission of information flow. Increasing the temperature of the gas will cause more random collisions and an increase in the resistance to the transmission of the information in our data. Thus, information temperature measures the amount of noise in our collection of attributes.

The agreement of the SNP associations found by EC with our previous proteomic analysis (McKinney *et al.*, 2006b) adds validation to the selected genetic variants in Table 2. The proteomic data acts as a validation set because a SNP that is identified by EC is more likely to be a true positive if that SNP is contained in a gene that encodes a protein previously found to be associated with the phenotype. This proteomic data set is limited as a validation set because additional cytokines were found to be significantly associated with AEs in our previous analysis whose genes were not part of the SNP panel used in the present study. Several of the findings in Table 2 are of potential biological significance, but upcoming genotypic studies on a larger population will serve to amend the findings of the current analysis. The clustering of many SNPs in the same gene observed in Table 2 warrants further investigations into the effect of linkage disequilibrium on analytical methods such as EC.

The current version of EC can handle categorical data with multiple classes. To handle data with numeric attributes such as gene expression data, one may use probability density estimation or possibly an entropy-based discretization procedure when computing mutual information. The EC feature selection method introduced in this article may be improved and extended in other ways as well. For example, it may be possible to improve the chances of reaching a global classification accuracy maximum by resampling collections of evaporated attributes or through a more global search mechanism like a genetic algorithm. The current search is local in the sense that the effect is observed of removing the single worst attribute at a time. A less biased, though more computationally expensive, search strategy would be to evaluate the collection of attributes resulting from the removal of random *subsets* of the worst attributes.

Additional cooling may be achieved by allowing attributes from the thermal fraction to collide and eject random attributes from the cold fraction. This is analogous to mutation in evolutionary algorithms, which would help prevent the permanent loss of informative attributes. One motivation for collisional cooling is that the presence of many noise attributes during early stages of evaporation may cause a few quality attributes to be evaporated erroneously. The erroneously eliminated attributes contain information relevant to the phenotype that was masked by irrelevant attributes. These relevant attributes may have better scores in the context of the cooler gas of atoms when returned to the gas by collisions.

EC was motivated by the concinnity with which nature balances energy and entropy to achieve an equilibrium configuration via the thermodynamic free energy. A system of *non-interacting* particles is in equilibrium for the lowest possible energy (e.g. when all particles are frozen at the bottom of the

potential well of Fig. 1). However, a system of *interacting* particles is in equilibrium when low energy and high entropy are balanced. Additional insight into the way nature achieves this balance will likely improve the EC information free energy feature selection method. Validation on real and simulated data suggest that EC may be a powerful feature selection filter for genom-wide genotypic studies. Future simulation studies for a wide variety of conditions will shed more light on the advantages and limitations of EC and information free energy.

ACKNOWLEDGEMENTS

This work was supported by K25 AI-064625, R01 AI-57661, AI-59694, LM-009012 and N01 AI-25462. The authors would like to thank Alison Motsinger and Scott Dudek for assistance with genomeSIM and Stephen Chanock for genotyping services.

Conflict of Interest: none declared.

REFERENCES

Draper,B. *et al.* (2003) Iterative relief. In *Workshop on Learning in Computer Vision and Pattern Recognition*. Madison, WI.

Dudek,S.M. *et al.* (2006) Data simulation software for whole-genome association and other studies in human genetics. *Pac. Symp. Biocomput.*, **11**, 499–510.

Dueck,G. and Scheuer,T. (1990) Threshold accepting: a general purpose optimization algorithm appearing superior to simulated annealing. *J. Comput. Phys.*, **90**, 161–175.

Frankel,W.N. and Schork,N.J. (1996) Who's afraid of epistasis? *Nat. Genet.*, **14**, 371–373.

Hess,H. (1986) Evaporative cooling of magnetically trapped and compressed spin-polarized hydrogen. *Phys. Rev. B*, **34**, 3476–3479.

Jaynes,E.T. (1957) Information theory and statistical mechanics. *Phys. Rev.*, **106**, 620–630.

Kira,K. and Rendell,L.A. (1992) A practical approach to feature selection. In *Proceedings of the Tenth International Conference on Machine Learning*. Amherst Morgan Kaufmann, Amherst, MA.

Kononenko,I. (1994) Estimation of attributes: analysis and extensions of relief. In *European Conference on Machine Learning*. Springer-Verlag, Catana, Italy.

Kononenko,I. and Robnik-Sikonia,M. (1996) Relieff for estimation and discretization of attributes in classification, regression and ilp problems. In *Artificial Intelligence: Methodology, Systems, Applications*. IOS Press, Amsterdam, The Netherlands.

Li,W. and Reich,J. (2000) A complete enumeration and classification of two-locus disease models. *Hum. Hered.*, **50**, 334–349.

McKinney,B.A. *et al.* (2006) Machine learning for detecting gene-gene interactions. *Appl. Bioinformatics*, **5**, 77–88.

McKinney,B.A. *et al.* (2006) Cytokine expression patterns associated with systemic adverse events following smallpox immunization. *J. Infect. Dis.*, **194**, 36092.

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

Moore,J.H. (2006) Genome-wide analysis of epistasis using multifactor dimensionality reduction: feature selection and construction in the domain of human genetics. In *Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data*, page in press.

Moore,J.H. and White,B.C. (2007) Tuning relieff for genome-wide genetic analysis. *Lecture Notes in Computer Science*, in press.

Onicescu,O. and Stefanescu,V. (1979) *Elements of informational statistics with applications*. Tehnica, Bucharest.

Press,W.H. *et al.* (1988) *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge.

Thornton-Wells,T.A. *et al.* (2004) Genetics, statistics and human disease: analytical retooling for complexity. *Trends Genet.*, **20**, 640–7.

Witten,I.H. and Frank,E. (2005) *Data Mining: Practical machine learning tools and techniques*. 2nd ed. Morgan Kaufmann, San Francisco.