BIOINFORMATICS

Vol. 16 no. 10 2000
Pages 923–931

The reduction of large molecular profiles to informative components using a Genetic Algorithm

F. M. Stefanini[1,*] and A. Camussi[2]

[1]Department of Statistics 'G.Parenti', University of Florence, Viale Morgagni 59, 50134, Firenze, Italy and [2]Department of Agricultural Biotechnology, University of Florence, P.le delle Cascine, 24, 50144, Firenze, Italy

Received on August 17, 1999; accepted on May 10, 2000

Abstract

Motivation: *Molecular profiles (DNA fingerprints) may be used to allocate an individual of unknown membership to one among the known groups of a reference population. Time and costs of profile assessment may be reduced by identifying informative profile components (markers).*

Results: *A genetic algorithm (GA) is proposed to identify promising candidate markers from a pilot experiment in which observations are supposed to be without measurement error. The analysis of simulated datasets suggests reasonable values for GA parameters and confirms that the GA finds components of the profile showing association with the considered groups. Our GA may be used to perform a first screening of candidate markers to be included in subsequent experiments.*

Availability: *The 32-bit executable (Windows 95, 98 and NT) is available at http://www.ds.unifi.it/~stefanin/bioinformatics.htm.*

Contact: *stefanin@ds.unifi.it*

Supplementary Information: *The algorithm is implemented for research purposes, i.e. a limited amount of input filtering and error messages are provided.*

Introduction

Individual-specific fingerprints of human DNA were introduced by Jeffreys *et al.* (1985) and are now extensively used in epidemiology, forensic science, molecular and population genetics. In crop science, DNA fingerprint technology offers new tools to improve procedures of pedigree analysis, genetic selection and genetic mapping (Beckman and Osborn, 1992; Phillips and Vasil, 1994).

This work deals with the use of molecular profiles to improve the ability of allocating an unknown individual to one of the groups in a biological reference population.

In typical experimental setups, DNA fragments are assayed as an array of dark bands on the electrophoretic gel. The array is called the molecular profile of the individual (DNA fingerprint). If a laboratory protocol includes DNA amplification, thousands of observable bands may be surveyed for each individual sampled. A wide amount of potentially useful information is readily accessible but costs can rise quickly.

The spread of molecular profile technology in routine applications requires methods to perform feature extraction (Devroye *et al.*, 1996, chap. 32), that is the reduction of large profiles to small informative components: the markers of biological groups. If the biological populations are stable regarding their DNA features, then group membership may often be successfully evaluated through informative profile components.

Algorithms based on evolutionary computation have been developed in forecasting (Packard, 1990), pattern recognition problems (Pei *et al.*, 1995a,b), molecular marker identification (Stefanini and Camussi, 1997) and to analyze high dimensional datasets in time series analysis (Packard and Meyer, 1991). Genetic algorithms (GAs) have been used to search for homologies in genetic sequence classification (Chuzhanova *et al.*, 1998) and to perform combinatorial optimization in molecular modelling and ligand design (Willet, 1995).

Recently, a GA optimizing a fitness function based on Bayesian predictive distributions has been proposed to analyze molecular profiles from pilot experiments (Stefanini, 1998). The approach differs from that in Stefanini and Camussi (1997) because a simpler and faster algorithm takes into account sampling variability. Moreover, it introduces Bayesian arguments to build the fitness function, a component that strongly determines GA performances.

Here we study the minimum number of generations (MNG) that the GA requires in order to find profile components associated to groups in the reference population, before the pre-assigned maximum number of generations is accomplished.

*To whom correspondence should be addressed.

The presentation opens with the problem abstraction in which measurement errors are assumed to be negligible. Then, an objective function based on Bayesian predictive distributions is developed and used to define a fitness function.

The main features of the implemented GA are explained in the last part of the Section **System and methods**. Two simulation experiments using the proposed fitness function are performed. The first (see the Section *Experiment 1: the identification of GA and fitness function parameters*) deals with the full factorial experiment in which MNG is recorded for each simulation repeated on the same dataset but with different choice of parameter values in the GA. The analysis of results from experiment 1 were used to select the parameter values used in experiment 2. In the second experiment (see the Section *Experiment 2: some performances of the GA*) datasets in which profile components have a known degree of association with groups are used as training data for the GA.

The analysis of MNG from experiment 2 revealed a dependence on the amount of information carried by profile components associated to groups. The **Discussion** focuses on the issues that might arise in the analysis of actual data.

Our algorithm may be used without additional tools only if data are not affected by measurement error. Moreover, the algorithm has the purpose of finding all the components of the profile that are associated to groups in the reference population, whether they are fake markers due to sampling noise or not. Confirmatory experiments are required to validate candidate markers (informative profile components).

System and methods

The objective function

Let $\mathcal{O} = \{O_1, \ldots, O_j, \ldots, O_K\}$ be the set of locations on electrophoretic gels (observable bands) that may show DNA bands. Let $\mathcal{P}(\mathcal{O})$ be the power set of $\mathcal{O}$. Then, a profile component is an element $\mathcal{S}$ of $\mathcal{P}(\mathcal{O})$. Let $\mathbb{B} = \{0, 1\}$ and $\mathbb{B}^K$ be its Cartesian product. Then, the band pattern of an individual is an element of $\mathbb{B}^K$, that is his *molecular profile* is represented by a vector of length K, whose element j is equal to 1 if a DNA band is present at location j on the gel, 0 otherwise. Let $\Omega_Y = \{0, 1, \ldots, M-1\}$ be the set of integers used to label the reference population that is made up of M subpopulations (groups). The random vector of $K+1$ elements associated with the random sampling of one individual from the reference population is

$$(Y, \underline{X}) = (Y, X_1, X_2, \ldots, X_K) \text{ on } \Omega_Y \times \mathbb{B}^K, \quad (1)$$

where X_j, $j = 1, \ldots, K$, refers to $O_j \in \mathcal{O}$. A realization $(y, x_1, x_2, \ldots, x_K)$ of (1) indicates the cell of the two-way contingency table 'group by band pattern' whose count has to be increased by one. Therefore, a saturated multinomial model for such a table is obtained by associating a parameter $\theta_{i,j}$ with each cell i, j where $j = 1, \ldots, 2^K$ and $i = 0, \ldots, M-1$. The parameter $\theta_{i,j}$ represents the probability of observing the band pattern in column j in the individual sampled from group i, with $\theta = (\theta_{0,1}, \theta_{0,2}, \ldots, \theta_{i,j}, \ldots)$ the vector of parameters and $\sum_{i,j} \theta_{i,j} = 1$.

A profile component $\mathcal{S} \subseteq \mathcal{O}$ is informative for Y if its observable bands reduce the uncertainty about Y (Stefanini, 1998). An informative profile component is a marker only if the association with groups of the reference population is not due to sampling noise. Thus, a validation experiment is required to assess if an informative profile component really marks the reference population. In this case the information is stable over replicated experiments.

Let $(Y, \underline{X}_{\mathcal{S}}, \underline{X}_{\overline{\mathcal{S}}})$ be a partitioned representation of (1) for a given $\mathcal{S}$ in which $\underline{X}_{\mathcal{S}}$ contains those X_j with $O_j \in \mathcal{S}$. Let $\theta_{\mathcal{S}} = t(\theta_{\mathcal{S},0,1}, \theta_{\mathcal{S},0,2}, \ldots, \theta_{\mathcal{S},i,j'}, \ldots)$ be the vector of parameters associated to the contingency table Y by $\underline{X}_{\mathcal{S}}$ and $\theta_{\overline{\mathcal{S}}} = (\theta_{\overline{\mathcal{S}},1}, \theta_{\overline{\mathcal{S}},2}, \ldots, \theta_{\overline{\mathcal{S}},j''}, \ldots)$ be the vector of parameters related to the band patterns $1, 2, \ldots, j'', \ldots$ due to $\underline{X}_{\overline{\mathcal{S}}}$. Let $p(Y \mid \underline{X}_{\mathcal{S}}, \theta_{\mathcal{S}}, \mathcal{S})$ be the probability mass function of Y conditional on the parameter $\theta_{\mathcal{S}}$ and the molecular information carried by $\mathcal{S}$. The band pattern indicated by j in the whole table is split into a component indicated by j' in the table Y by $\underline{X}_{\mathcal{S}}$, and a component indicated by j'' in the vector of band patterns due to $\underline{X}_{\overline{\mathcal{S}}}$.

Let $p(Y \mid \underline{X}_{\mathcal{S}}, \underline{X}_{\overline{\mathcal{S}}}, \theta, \mathcal{S})$ be the the conditional distribution of Y given $\underline{X}$, where $\underline{X}$ is partitioned according to $\mathcal{S}$. Assume that conditional independence holds, i.e.

$$p(Y \mid \underline{X}_{\mathcal{S}}, \underline{X}_{\overline{\mathcal{S}}}, \theta, \mathcal{S}) = p(Y \mid \underline{X}_{\mathcal{S}}, \theta_{\mathcal{S}}, \mathcal{S}), \quad (2)$$

then the conditional distribution of Y given a component of the band pattern indicated by j' does not change its shape whatever the component of the band pattern indicated by j''. In other words, the information content of the conditional distribution does not change after assessing the component of the band pattern due to observable bands in $\overline{\mathcal{S}}$.

We follow the Bayesian paradigm for the unknown parameters in θ by summarizing the uncertainty through a probability distribution. The information on θ is updated by conditioning on data from a pilot experiment. From the reference population, $N_t = M \cdot N$ individuals are sampled, N for each of the M subpopulations. Let $\underline{n}_{\mathcal{S}} = (n_{0,1}, \ldots, n_{i,j}, \ldots)$ be the value of the sufficient statistics (cell counts) that are calculated using the training dataset on the marginal two-way table induced by $\mathcal{S}$. Let the prior distribution be a Dirichlet probability density function

with parameters $\lambda_{i,j} = \text{const}$, $\lambda = \sum_{i,j} \lambda_{i,j} = 1$. The posterior distribution $p(\theta_{\mathcal{S}} \mid \underline{n}_{\mathcal{S}}, \mathcal{S})$ of $\theta_{\mathcal{S}}$ given $\underline{n}_{\mathcal{S}}$ and $\mathcal{S}$ is Dirichlet with updated parameters $\widetilde{\lambda}_{i,j} = \lambda_{i,j} + n_{i,j}$. The predictive distribution $p\left(Y \mid \underline{x}_{\mathcal{S}}, \underline{n}_{\mathcal{S}}, \mathcal{S}\right)$ obtained by integrating with respect to $\theta_{\mathcal{S}}$ captures the uncertainty about Y for the next individual that carries the molecular profile component equal to $\underline{x}_{\mathcal{S}}$.

The directed divergence (Kullback, 1968, p. 7) quantifies the information gain about Y provided by the data $\underline{n}_{\mathcal{S}}$ for a given profile component $\mathcal{S}$ and observed band pattern $\underline{x}_{\mathcal{S}}$, which is

$$I(p_Y \mid \underline{x}_{\mathcal{S}}, \underline{n}_{\mathcal{S}}, \mathcal{S}) = \sum_{i=0}^{M-1} p(Y \mid \underline{x}_{\mathcal{S}}, \underline{n}_{\mathcal{S}}, \mathcal{S}) \times \log\left(\frac{p(Y \mid \underline{x}_{\mathcal{S}}, \underline{n}_{\mathcal{S}}, \mathcal{S})}{p_Y(i)}\right). \quad (3)$$

The evaluation of the information gain depends on the amount of uncertainty for the same subpopulation before observing the band pattern $\underline{x}_{\mathcal{S}}$, i.e. the denominator $p_Y(i)$. If the probability value $p_Y(i)$ is close to one, the gain is small even if $p(Y = i \mid \underline{x}_{\mathcal{S}}, \underline{n}_{\mathcal{S}}, \mathcal{S})$ is equal to one. The information gain associated to a band pattern $\underline{x}_{\mathcal{S}}$ is maximum if the probability mass function $p(Y \mid \underline{x}_{\mathcal{S}}, \underline{n}_{\mathcal{S}}, \mathcal{S})$ is degenerate, that is if it is concentrated on one subpopulation that prior to observing the band pattern had a probability value close to zero. If the numerator of the logarithm in equation (3) is zero, we set $0 \cdot \log(0) = 0$. In experiments 1 and 2 the denominator is equal to the constant M^{-1}, because the probability mass function $p_Y(y)$ is uniform. This choice maximizes the entropy of the distribution and it is suited to circumstances of weak priori information about Y. We will not discuss here the possibility of modifying $p_Y(y)$ according to experimental data.

The value of a band pattern in a future experiment is unknown, thus different results $\underline{x}_{\mathcal{S}}$ correspond to different values of equation (3). We propose a weighted average of equation (3) according to the probability of observing $\underline{x}_{\mathcal{S}}$. The expected information gain $I_E(p_Y \mid \underline{n}_{\mathcal{S}}, \mathcal{S})$ with respect to $\underline{X}_{\mathcal{S}}$ is an averaged gain with respect to the band patterns defined through $\mathcal{S}$, that is

$$I_E(p_Y \mid \underline{n}_{\mathcal{S}}, \mathcal{S}) = \mathbf{E}[I(p_Y \mid \underline{X}_{\mathcal{S}}, \underline{n}_{\mathcal{S}}, \mathcal{S})]. \quad (4)$$

The expected information gain approaches the maximum value if different subpopulations do not share band patterns, that is if equation (3) is maximized for each band pattern.

It follows that a profile component that is very large, will determine a huge number of band patterns, and each one will be observed one or zero times. Moreover, it is likely those zeros are due to sampling (small sampling size), thus the expected information gain does not capture interesting features if large profile components are investigated with small samples (as expected). This is the main reason to constraint the maximum size of the candidate markers.

Two assumptions make the evaluation of the expected information gain meaningful:

(1) the variability of band patterns 'between' and 'within' subpopulations is not negligible; if this assumption does not hold then the probabilistic description of the uncertainty loses its purpose;

(2) if some informative profile components are contained in the profile, their size is small thus the number of elements of $\mathcal{O}$ contained in $\mathcal{S}$ is small; it is difficult to specify a maximum number, although it is clear from the discussion above that the maximum number of observable bands is related to the sample size.

We anticipate here that in our simulations, the maximum number of elements contained in a set $\mathcal{S}$ is three.

For the goal of our analysis, the dependence of equations (3) and (4) over M is not interesting here, because the search for informative profile components is performed in the preassigned reference population, that is conditional to a given M.

From the equations above, the expected information gain is conditional to the observed data. A profile component could be evaluated as highly informative because of sampling noise. The large number of observable bands in pilot experiments suggests that a validation experiment including only informative profile components is the best approach to distinguish fake from true markers. Statistical tests might be used to perform some kind of noise filtering, but the statistical power in detecting signals might be so low to make pilot experiments useless.

The unnormalized fitness function

The unnormalized fitness function $g_E(\mathcal{S}, q_1, q_2)$ for a given training dataset $\underline{n}$ is obtained from the expression (4):

$$g_E(\mathcal{S}, q_1, q_2) = \{I_E(p_Y \mid \underline{n}_{\mathcal{S}}, \mathcal{S}) \cdot h(\mathcal{S})\}^{q_1} + q_2, \quad (5)$$

where q_1 and q_2 are two tuning constants that affect the way the expected information gain is mapped in the fitness function. Thus a rich class of functions is available .

The function $h(\mathcal{S})$ penalizes the profile components that exceed the minimum number of observable bands required to discriminate among M groups. In our simulations it is defined by

$$h(\mathcal{S}) = 1 - \frac{W + W_l}{\widetilde{W}},$$

where M_l is the least integer not smaller than $\log_2(M)$, $W = \text{card}(\mathcal{S})$ and $\widetilde{W} = \max(\text{card}(\mathcal{S}))$. In our simulations the maximum size of an informative profile components is $\widetilde{W} = 3$, and $M_l = 1$ (two subpopulations).

Numerical investigations (results not shown) were performed on simulated datasets to check if the equivalence classes induced by (5) on $\mathcal{P}(\mathcal{O})$ were suited to the goal and if the values shown for pairs of equivalence classes were reasonable. Nevertheless, the uncertainty about the choice of q_1 and q_2 was still considerable even after the numerical inspection.

Moreover, it was not clear if the choice of GA parameters should be performed one at a time, that is by ignoring the value of parameters within $g_E(\cdot)$, or if parameters should be jointly selected because of the relevance of their interaction.

We designed the Monte Carlo experiment 1 to answer the above questions, because the analytical solution related to GA dynamics seemed unfeasible.

Algorithm and implementation

The GA used in experiments 1 and 2 resembles the simple GA described in Stefanini and Camussi (1993), but the chromosome structure is different (Stefanini, 1998).

Each chromosome is an ordered list of integers whose size is variable within a preassigned set (1–3 in our simulations). Chromosomes are initialized by a single random drawing from the set of integers $\{11, \ldots, K\}$, so that informative profile components located in $\{1, \ldots, 10\}$ are certainly excluded from the first generation. The probability of sampling informative components by chance during the initialization is negligible for datasets larger then those considered here, thus we introduced this feature as a correction in the evaluation of MNG. Stochastic Universal Sampling (Baker, 1987) is used in the reproduction step to diminish the selection bias.

A specialized crossing over operator and three different types of mutation are defined to work on ordered lists of integers. As regards mutation, a Monte Carlo mutation operator guarantees the global convergence of the algorithm in the limit (Geyer-Schulz, 1995, chapter 10, pp 316). In our algorithm it corresponds to the random draw without replacement of a random number of integers from the urn defined by the set $\{1, 2, \ldots, K\}$. The mutation operator called *Del* causes random deletion of size 1 in a chromosome. The mutation operator called *Sub* substitutes a single observable band j with a new observable band randomly drawn from the set $\{1, 2, \ldots, j-1, j+1, \ldots, K\}$. A mixture of these mutation operators is used with probability pm on a given chromosome. See Stefanini (1998) for details.

The probability of recombination for a chromosome is equal to pc. The partner chromosome is randomly drawn according to fitness values in the population. The recombination operator samples half of the observable bands contained in the first chromosome and half from the second, checking for duplication of observable bands in the derived list. Then, it stores the ordered list of observable bands in the newborn chromosome.

The *ran2()* generator described in Press *et al.* (1992) was used as a source of pseudo-random numbers, after embedding it in a C++ class that contains member functions to produce various other types of random deviates. The initialization may be chosen by the user or set using internal clock values.

The implementation of the GA was performed in C++, using the Borland compiler 5.01 (Borland International, Inc., 1996) in a Windows 98 environment. Simulations were performed on a Pentium II PC 200 MHz with 128 Mb Ram. Each GA run took approximately 33 s.

The 32-bit executable file runs on PC machines and uses ASCII files to read the dataset and GA parameters.

Results

Experiment 1: the identification of GA and fitness function parameters

The designed simulation plan corresponds to the Cartesian product among the following sets: $\{0.5, 1\}$ for q_1, $\{0.01, 0.001\}$ for q_2, $\{0.01, 0.05, 0.1\}$ for the probability of mutation pm (it refers to one chromosome, see the Section *Algorithm and implementation*), $\{0.1, 0.25, 0.35\}$ for the probability of crossing-over pc. For each point in the Cartesian product, 50 GA runs (replicates) were performed on a simple simulated dataset that retained a non-trivial association structure between Y and $\underline{X}_{\mathcal{S}}$. This means that there was an observable band not fully informative for the reference population if considered alone, but fully informative if jointly considered with another observable band.

Three observable bands O_1, O_2 and O_3 carried information about Y. The observable bands O_1, O_2 were fully informative for Y, and O_3 carried some information but it was redundant given O_1, O_2. Furthermore, both O_1, O_3 and O_2, O_3 were not fully informative for Y. In other words, individuals in the dataset that shared the same band pattern in O_1, O_2 belonged to the same subpopulation.

The dataset of 100 simulated molecular profiles had two subpopulations, $M = 2$, and $K = 1000$ observable bands.

The total number of GA runs was 1800, and the maximum number of generations was set to 10 000.

Observable bands in $\underline{X}_{\overline{\mathcal{S}}}$ were identical in the two subpopulations, and were independently sampled from a uniform distribution on $\{0, 1\}$, that is $(x_4, x_5, \ldots, x_{1000})_r = (x_4, x_5, \ldots, x_{1000})_{r+50}$, where $r = 1, 2, \ldots, 50$ and the $y_r = 0$, $y_{r+50} = 1$.

Generalized linear models (McCullagh and Nelder, 1989) were used to clarify the structure underlying the

Table 1. Descriptive statistics for experiment 1. Mean (top in each cell) and standard deviation (bottom in each cell) of the variable minimum number of generations (MNGs), cross-classified for parameter values of q_1, pm and pc. This table is marginalized with respect to q_2, thus in each cell there are 100 observations

pc	0.1	0.25	0.35
$q_1 = 0.5$			
pm			
0.01	4185	2485	2282
	3499	*2283*	*2500*
0.05	970	480	402
	984	*536*	*411*
0.1	445	298	220
	431	*304*	*211*
$q_1 = 1$			
pm			
0.01	5439	3671	2750
	5673	*2982*	*2452*
0.05	1465	636	615
	1209	*544*	*609*
0.1	644	349	289
	547	*320*	*277*

observed variability of MNGs (Z in equation (6)). Several quasi-likelihood models (McCullagh and Nelder, 1989) were fitted to data and the final model, obtained using a stepwise procedure, was

$$\begin{aligned} &\log(\eta_{ijk}) = \mu_{111} + \alpha_i + \beta_j + \gamma_k \\ &\mathrm{Var}(Z) = \phi \cdot \eta^2_{i,j,k} \\ &\eta_{i,j,k} = E[Z_{i,j,k,l}] \\ &i = 2;\ j = 2, 3;\ k = 2, 3;\ l = 1, \ldots, 100. \end{aligned} \tag{6}$$

The parameter μ_{111} refers to the cell in the top-left corner of Table 1, and α_i, β_j, γ_k represent respectively the main effects of q_1, pm and pc. In equation (6), ϕ is a scale parameter and η the expected value of Z. The index l is from 1–100 because the 50 replicates are multiplied by 2, the two levels of the model factor q_2 that was excluded from the final model.

Table 2 provides estimates for all terms in (6). The simple effect of q_2, as well as all interaction effects, were excluded from the final model as they did not significantly differ from 0. All the main effects retained in (6), on the other hand, differed from 0 at a significance level of 1%. Absence of interactions, in particular, allows us to select the values of q_1, q_2, pm and pc separately (that is, without having to consider joint effects on MNGs).

Table 1 provides some summary statistics (mean and standard deviation of MNGs) classified by values of q_1, pm and pc. Each cell refers to 100 replicates, since q_2 is neglected. It can be seen that q_1 (minimum fitness value assigned to non-informative markers) is inversely related to the average number of MNGs. Also pm and pc (crossing-over probability in the GA) are inversely related to the average number of MNGs.

Table 2. GLM model for experiment 1. Estimated values and standard errors of model parameters. Standard errors are almost all equal due to the selected variance function

Parameter	Value	Standard error
μ_{111}	8.398	0.055
α_2	0.307	0.045
β_2	−1.565	0.055
β_3	−2.240	0.055
γ_2	−0.577	0.055
γ_3	−0.759	0.055
ϕ	0.912	

Based on the above considerations, we selected parameter values for experiment 2 as follows:

- q_2 is not significant, thus we just picked $q_2 = 0.005$ (an average of the two values considered in experiment 1),
- q_1, pm and pc should be large to reduce MNGs and to avoid the premature convergence to relative maxima. At the same time, large values of pm and pc are undesirable as they make the GA less capable of finding the absolute maximum. Thus, striking a balance among the values considered in experiment 1, we picked $q_1 = 1$, $pm = 0.01$ and $pc = 0.25$.

These values provided reasonable time performances in our simulations, but this could clearly be the object of further investigations.

Experiment 2: some performances of the GA

In the second set of simulations, a GA with parameter values $q_1 = 1$, $q_2 = 0.005$, $pm = 0.01$ and $pc = 0.25$ was used to analyze 10 000 simulated datasets. Each sampled dataset had size 200, $K = 1000$ and $M = 2$. The *maximum* number of generations was set to 10 000.

The joint distribution from which the datasets were sampled is specified by a hierarchy of distributions. The informative profile component, if present, always contains observable bands 9 or 8 and 9. The probability mass function for such observable bands is

$$\begin{aligned} p(Y, X_9, X_{10} \mid \gamma_{11}, \gamma_{21}, \alpha_{11}, \alpha_{21}, \beta_{11}, \beta_{21}) \\ = \alpha_{11} \cdot \mathbf{I}_{(0,0,0)}(y, x_9, x_{10}) \\ + (\gamma_{11} - \alpha_{11}) \cdot \mathbf{I}_{(0,1,0)}(y, x_9, x_{10}) \\ + \beta_{11} \cdot \mathbf{I}_{(0,0,1)}(y, x_9, x_{10}) \end{aligned}$$

$$
\begin{aligned}
&+((1-\gamma_{11})-\beta_{11})\cdot \mathbf{I}_{(0,1,1)}(y, x_9, x_{10})\\
&+\alpha_{21}\cdot \mathbf{I}_{(1,0,0)}(y, x_9, x_{10})\\
&+(\gamma_{21-}\alpha_{21})\cdot \mathbf{I}_{(1,0,1)}(y, x_9, x_{10})\\
&+\beta_{21}\cdot \mathbf{I}_{(1,0,1)}(y, x_9, x_{10})\\
&+((1-\gamma_{21})-\beta_{21})\cdot \mathbf{I}_{(1,1,1)}(y, x_9, x_{10})
\end{aligned}
$$

where parameters are distributed as

$$
\begin{aligned}
\gamma_{11} &\sim \frac{1}{0.39}\cdot \mathbf{I}_{[0.60,0.99]}(\gamma)\\
\gamma_{21} &\sim \frac{1}{0.49}\cdot \mathbf{I}_{[0.01,0.50]}(\gamma)\\
(\alpha_{11} \mid \gamma_{11}) &\sim \frac{2}{\gamma_{11}}\cdot \mathbf{I}_{[\gamma_{11}/2,\gamma_{11}]}(\alpha)\\
(\alpha_{21} \mid \gamma_{21}) &\sim \frac{2}{\gamma_{21}}\cdot \mathbf{I}_{[\gamma_{21}/2,\gamma_{21}]}(\alpha)\\
(\beta_{11} \mid \gamma_{12}) &\sim \frac{2}{\gamma_{12}}\cdot \mathbf{I}_{[\gamma_{12}/2,\gamma_{12}]}(\beta)\\
(\beta_{21} \mid \gamma_{22}) &\sim \frac{2}{\gamma_{22}}\cdot \mathbf{I}_{[\gamma_{22}/2,\gamma_{22}]}(\beta).
\end{aligned}
$$

The observable bands from 1–8 are sampled from uniform distributions on the cartesian product $B_8 = \{0, 1\}^8$:

$$
\begin{aligned}
(X_1, \ldots, X_8)_r &\sim \frac{1}{256}\cdot \mathbf{I}_{B_8}(x_1, \ldots, x_8)\\
r &= 1, 2, \ldots, 200.
\end{aligned}
$$

The observable bands from 11–1000 in the 100 profiles which refer to $Y = 0$ are sampled from

$$
\begin{aligned}
\big(X_{ij} \mid Y = 0\big) &\sim \frac{1}{2}\cdot \mathbf{I}_{\{0,1\}}(x)\\
i &= 11, \ldots, 1000;\ \ j = 1, \ldots, 100,
\end{aligned}
$$

and are duplicated in the 100 profiles, which refer to $Y = 1$, that is

$$
\begin{aligned}
(x_{ij+100} \mid Y = 1) &= (x_{ij} \mid Y = 0)\\
i &= 11, \ldots, 1000;\ \ j = 1, \ldots, 100.
\end{aligned}
$$

The class of datasets that can be generated using the proposed distribution is very rich. Anyway, the maximum number of observable bands in the informative profile components is always three, and the informative components are always built among the first 10 observable bands. This feature does not limit the validity of the results as the performances of the optimization do not depend on the order of observable bands in the dataset. Note that unplanned informative components may arise by random sampling among observable bands located in a position smaller than 9.

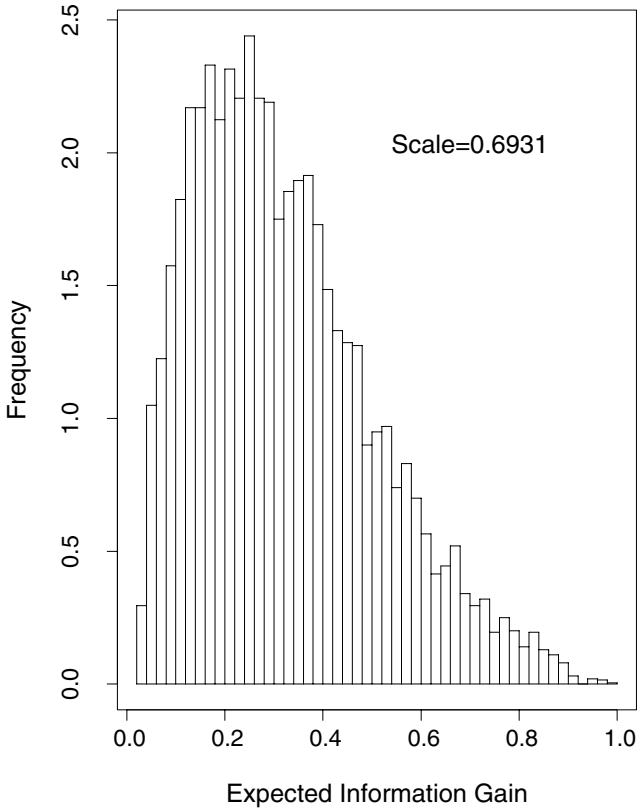

Fig. 1. Information gain in experiment 2. Histogram of the expected information gain for the best chromosome in 10 000 datasets.

A summary of the main feature of these datasets is given in Figure 1. The approximate distribution of the expected information gain based on 10 000 simulated datasets is shown after rescaling the maximum value to 1. The datasets generating process matches our belief about the amount of information contained in datasets from a pilot experiment, given that at least one profile component is informative.

The fittest chromosome after 10 000 generations of the GA was examined for each simulated dataset. The best chromosome in a dataset is known by construction, therefore it was possible to classify the fittest chromosomes in four subsets: full match (FM) with the best chromosome, partial match (PM) with the best chromosome, no match and higher fitness value (NMH), no match and smaller fitness value (NMS, Table 3). The best chromosome is found by the GA in 10 000 generations only in 90% of the datasets. Although the probability of finding the fittest chromosome is one in the limit for a number of generations that goes to infinite, there is not a strict rule to choose the maximum number of generations (see the **Discussion**). In 1.4% of the simulations, the fittest chromosome has no match and smaller fitness value than the known best. Some

Table 3. Descriptive statistics for experiment 2.The fittest chromosomes (after 10 000 generations) are cross-classified according to: full match (FM) and partial match (PM) with the best chromosome, no match but higher fitness value (NMH), and no match and smaller fitness value (NMS)

Class	Number	%
FM	9003	90.0
NMH	646	6.5
PM	212	2.1
NMS	139	1.4
Total	10 000	100.0

observable bands of the best chromosome are contained in the fittest chromosome for the 2.1% of the GA runs. If the number of generations is increased by a few hundreds generations, PM chromosomes became FM, but the percentage of PM chromosomes does not reach zero as fast (data not shown): the NMS chromosomes start moving to the PM class, thus the choice of maximum number of generations in our simulations might be about how many generations will give 100% of FM chromosomes.

We anticipate the discussion here by underlining that in actual applications no known best is available, thus other rules based on output diagnostics must be used (see the **Discussion**). In the analysis of actual data, the number of generations required might be greater than a few hundred due to higher level of sampling noise in the dataset.

Finally, the GA found fittest chromosomes that have a higher fitness value than the best chromosomes in 6.5% of the datasets (Table 3). This result suggests that, in the analysis of actual data, a relevant number of informative profile components found by the GA might just contain sampling noise instead of useful information. This is a feature that depends on the design of pilot experiments, typically constrained by high costs, not on the GA.

At the end of a computer run, the best chromosome is examined and the generation in which it was hit is recorded as value of MNG.

In Figure 2 (main window), the histogram of MNGs shows a long right tail. The dependence of MNGs on the fitness value of the best chromosome was investigated by fitting a smoothed surface to the two-way histogram 'fitness by MNGs' and displaying it in a contour plot (Figure 2, overlaid plot). In the simulation output, a fitness value equal to 0.1836 splits the MNGs values into two groups. The first, defined by values larger than 0.1836, always shows MNGs values smaller than 1200 generations; the second is characterized by MNGs values that cover the entire range 1–10 000 generations. An explanation of these findings is that the GA can establish a subpopulation of chromosomes that are almost-optimal

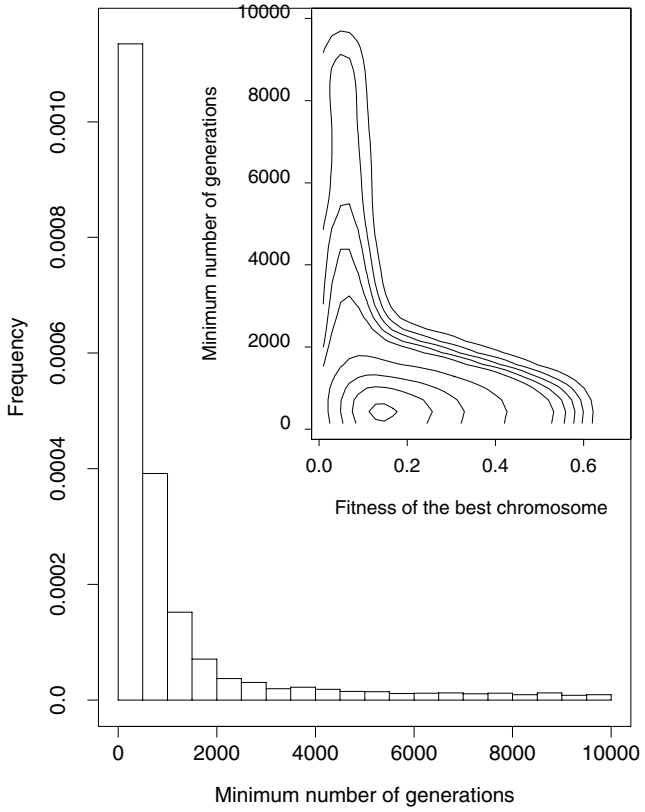

Fig. 2. Relationship between information gain and minimum number of generations. The histogram shows the minimum number of generations (MNGs) required to obtain the fittest chromosome. Overlaid on top right, the contour plot of the smoothed surface based on the two-way histogram 'fitness value by MNGs' emphasizes the dependence of MNGs on the fitness value.

if the fitness is above a threshold. If the fitness value is not large enough, then the chance to exploit neighborhoods of almost-optimal chromosomes is small. Nevertheless, such exploitation is the main reason why we use the GA instead of simple random sampling, thus a rescaling of the fitness function might be useful in these cases. This is an issue open to investigation.

Discussion

Feature extraction in datasets of actual pilot experiments is a formidable computational task due to the huge number of candidate markers. The straightforward use of machine learning algorithms may fail (no or wrong feature extraction) due to the presence of many sampling zeros in the two-way contingency table 'groups by band pattern'. Thus we proposed the use of Bayesian techniques to smooth out observed values by using predictive distributions. The definition of expected information gain is the core of the objective function optimized by the GA. Nevertheless, the

choice of a map from values of the expected information gain to fitness values is not trivial. Experiment 1, using a simulated data set, shedded some light on the role of the parameters q_1, q_2, pm and $pc : q_2$, as well as all the interaction effects, were found to be of little relevance. Thus, the parameter values for experiment 2 were chosen independently. In particular, we selected the smallest values p_m and p_c that made the computational burden feasible in experiment 2. In this way GA robustness and tractability were achieved. Experiment 2 has shown that for the (large) class of datasets generated by a specified prior distribution, the proposed GA is effective.

Several questions remain unanswered, due to the high complexity of the problem domain, especially as regards the analysis of experimental results. Comments and recommendations to perform the analysis of actual datasets are listed below, from general aspects to algorithmic details.

If there is no variability within groups and they do not share the same band pattern, then the straight comparison of bands suffices to find molecular markers.

The analysis may be performed by setting the maximum number of observable bands within informative profile components to values larger than 3. Nevertheless, the sample size sets an upper limit because a large contingency table will contain too many sampling zeroes. If the study of large profile components is essential, full scale experiments should be designed.

Researchers should save part of the resources to perform confirmatory experiments on candidate markers. The GA identifies profile components associated to groups, but the association could be due to sampling noise. Simultaneous statistical tests based on pilot experiments seem inadequate to account for noise in collected data due to small statistical power of the overall testing procedure. For similar reasons the optimization over the space of non saturated models for the two-way contingency table was not considered.

The class of datasets generated using the prior distribution described in the Section *Experiment 2: some performances of the GA* is wide as regards the informative part of the profile. We expect that actual datasets may show larger variability in the non informative part of the profile. For these reasons, our findings about MNG obtained by Monte Carlo simulation act as an expected minimum, because the fitness landscape might be less flat than the one for simulated data for the non informative part of the profiles (higher sampling noise).

In any case, the features of the DNA in the biological populations should be 'stable' with respect to time, otherwise the information obtained from data processing is useless. For example, (natural or artificial) vegetative reproduction or short rate of sexual mating should make DNA sequences stable in the population.

The analysis of actual data may take advantage of algorithmic improvements. The beneficial effects of changing the probability of mutation and crossing over along the GA run to change the trade-off existing between exploitation and exploration are well known. In any case, the number of GA chromosomes should not be smaller than 100. A small population of chromosomes makes the GA faster, but the exploitation of promising neighborhoods is almost absent (Stefanini, 1998). If the collection of data is highly expensive it might be reasonable to demand stronger experimental evidence about a candidate marker before including it in the confirmatory experiment. This goal might be achieved using a value of λ greater than 1 in the prior distribution of the probability values in the contingency table (see the Section *The objective function*). Finally, if prior information is available about some observable bands, it might be included by a suitable choice of values $\lambda_{i,j}$.

The choice of values for the GA parameters should not depend on specific association structures between groups and band patterns. Robustness and generality of the GA could be compromised to improve the performances in a relatively small subset of association structures. Moreover, pilot experiments are typically performed because it is impossible to restrict the search space to specific association structures, thus the advantage expected by such an elaborate choice of parameters vanishes.

In the study of an actual experiment, several GA runs should be performed on the same dataset to increase the confidence that the chromosome found by the GA is optimal. It is not excluded that several profile components might show similar values of expected information gain. Thus several GA runs also provide evidence of multiple optima. We suggest a simulation plan including less than five long runs and some more short runs, so that a balance between the effect of the starting population and the need for long runs may be achieved. Quantitative rules based on empirical models for output analysis are still to be developed to relate the chance of capturing the true best chromosome and the number of computer runs. A different approach based on fitness sharing is discussed in (Stefanini, 1998), but, at this time, it does not seem convenient in terms of computational load.

Actual datasets may contain missing values, especially if the amount of DNA available for the analysis is small. Bayesian imputation techniques (Schafer, 1997) should be embedded in the objective function to handle missing values. A simplified procedure was recently proposed to cut the computational burden involving missing values, so that extensive simulation plans are not compromised (Stefanini, 1998).

A different type of uncertainty is due to measurement errors (Roeder, 1994). For example, in some experiments the number of bands on the gel is uncertain because of the background noise, the low intensity of the signal and/or

the tendency of closely located bands to collapse. In this paper we assumed that this type of uncertainty is not present, but if it were, further steps would be required before running the GA. The mechanism originating the uncertainty typically depends on the protocol and equipment used, therefore we think that the identification step should be maintained separate from the algorithm to perform feature extraction. These are issues to be addressed in further research.

The definition of $\mathcal{O}$ in the Section *The objective function* is not trivial even without missing values or errors in measurement. Changes in the GA might be introduced to take advantage of multiband protocols; that is, of full linkage among sets of observable bands. In any case, we suggest associating an observable band to each gel location whether or not DNA fragments are observed in that position. Instead, a definition based on those gel locations showing variability may be used, only if the sample size is large.

Finally, further theoretical analysis of GAs might solve some of the problems described above without resorting to Monte Carlo experiments, although the high non linearity of GA dynamics makes the analysis beyond a stylized class of GAs (van Nimwegen *et al.*, 1997) very difficult. A better characterization of GA dynamics might arise from innovative hypotheses about the innermost mechanism of GA optimization (Beyer, 1997).

Acknowledgements

This work was supported by a grant from the Italian Ministry of Agricultural Policies (MIPA), in the framework of the National Project 'Biotecnologie Vegetali–Area 3'. FMS thanks the Santa Fe Institute, New Mexico, where part of this work was performed and the referees for pointing out parts of the manuscript to be clarified.

References

Baker,J.E. (1987) Reducing bias and inefficiency in the selection algorithm. In Grefenstette,J.J. (ed.), *Genetic Algorithms anf Their Applications: Proceedings of the Second International Conference on Genetic Algorithms* Lawrence Erlbaum Associates, Hillsdale.

Beckman,J.S. and Osborn,T.C. (eds) (1992) *Plant Genomes: Methods for Genetic and Physical Mapping*. Kluwer Academic Publishers, London.

Beyer,H.G. (1997) An alternative explanation for the manner in which genetic algorithms operate. *Biosystems*, **41**, 1–15.

Borland International, Inc., (1996) *Borland C++ Programmer's Guide*. Borland International, 100 Borland Way, Scotts Valley, California, USA.

Chuzhanova,N.A., Jones,A.J. and Margetts,S. (1998) Feature selection for genetic sequence classification. *Bioinformatics*, **14**, 139–143.

Devroye,L., Györfi,L. and Lugosi,G. (1996) *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

Geyer-Schulz,A. (1995) *Fuzzy Rule-Based Expert Systems and Genetic Machine Learning*. Physica, Heidelberg.

Jeffreys,A.J., Brookfield,J.F. and Semeonoff,R. (1985) Positive identification of animigration test-case human DNAfingerprint. *Nature*, **317**, 818–819.

Kullback,S. (1968) *Information Theory and Statistics*. Dover, Mineola, NewYork.

McCullagh,P. and Nelder,J.A. (1989) *Generalized Linear Models*. Chapman and Hall, New York.

Packard,N.H. (1990) A genetic algorithm for analyzing complex data. *Complex Systems*, **4**, 543.

Packard,N.H. and Meyer,T. (1991) Local forecasting in high dimensional chaotic dynamics. In Casdagli,M. and Eubank,S. (eds), *Nonlinear Modeling and Forecasting* Addison-Wesley, Reading.

Pei,M., Ding,Y., Punch,W.F. and Goodman,E.D. (1995a) Classification and feature extraction of high-dimensionality binary patterns using a ga to evolve rules. Working paper, Genetic Algorithms Research and Applications Group. Web page http://www.egr.msu.edu.

Pei,M., Goodman,E.D. and Punch,W.F. (1995b) Pattern discovery from data using genetic algorithms. Working paper for the first pacific-asia conference on knowledge discovery and data mining, Genetic Algorithms Research and Applications Group. Web page http://www.egr.msu.edu.

Phillips,R.L. and Vasil,I.K. (eds) (1994) *DNA-based Markers in Plants*. Kluwer Academic Publishers, London.

Press,W.H., Flannery,B.P., Teukolsky,S.A. and Vetterling,W.T. (1992) *Numerical recipes in C: the art of scientific computing*. Cambridge University Press, Cambridge.

Roeder,K. (1994) DNA fingerprinting: a review of the controversy. *Statistical Science*, **9**, 222–278.

Schafer,J. (1997) *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.

Stefanini,F. (1998) Identification of highly informative molecular profile components using genetic algorithms. Working paper series 98-05-042, Santa Fe Institute. Web page http://www.santafe.edu.

Stefanini,F.M. and Camussi,A. (1993) APLOGEN: an object oritented genetic algorithm performing MonteCarlo optimization. *Computer Applications in Biosciences*, **9**, 695–700.

Stefanini,F. and Camussi,A. (1997) Information in molecular profile components evaluated by a genetic classifier system: a case study in *Picea abies Karst*. *Genetical Research*, **70**, 205–213.

van Nimwegen,E., Crutchfield,J.P. and Mitchell,M. (1997) Statistical dynamics of the royal road genetic algorithm. Working paper 97-04-035, Santa Fe Institute.

Willet,P. (1995) Genetic algorithms in molecular recognition and design. *Trends in Biotechnology*, **13**, 516–521.