Listy Biometryczne - Biometrical Letters
Vol. 36(1999), No. 1, 47-59

The choice of molecular profile components based on a quantitative evaluation of convenience

Federico Mattia Stefanini[1] and Alessandro Camussi[2]

[1]Department of Statistics 'G.Parenti', University of Florence,
Viale Morgagni 59, 50134 Firenze, Italy
stefanin@ds.unifi.it

[2]Genetic Unit, Institute of Silviculture, University of Florence,
via S. Bonaventura 13, Firenze, Italy
gene_agr@cesit1.unifi.it

SUMMARY

Molecular profiles obtained by electrophoresis may be exploited in classification problems by using molecular profiles of unknown individuals to predict their group membership. In a reference population, the uncertainty about group membership is reduced if one or more informative profile components are identified through pilot experiments. The size of a molecular profile is virtually infinite because an experiment could include a huge number of different setups (e.g., pairs enzyme-probe) so that from the amplified DNA of an individual thousands of bands could be obtained. Time and cost of assessment put a constraint on the size of molecular profiles utilized in actual experiments.

In this paper, pilot experiments are analyzed from the standpoint of statistical decision theory to find informative profile components that should be included in future experiments. The trade-off between cost of the experiment and decrease of uncertainty due to the profile assessment is defined to evaluate the convenience of different actions in a quantitative way. The proposed framework is based on a well known body of knowledge, but the analysis of actual pilot experiments with molecular profiles has to cope with some open problems.

KEY WORDS: Bayesian inference, decision theory, molecular profiles.

*The paper was submitted on the occasion of 70-th birthday of Professor Tadeusz Caliński.

1. Introduction

Recent techniques of molecular biology (e.g., Beckmann and Osborn, 1992; Philips and Vasil, 1994) have introduced the possibility of surveying 'molecular phenotypes' that are closely related to the DNA level. Each experimental unit (individual) is scored for the presence or the absence of DNA fragments (bands) on lines of electrophoretic gels. The experimental setup, namely the restriction enzymes, the DNA probes and the bio-chemical protocol, defines the features of DNA fragments detected on a gel. Thus, the researcher has access to a potentially wide amount of molecular information.

The polymorphism found in many populations of economic interest may be useful not only to study QTLs and to build genetic maps, but also to face discriminant problems. A few examples from crop science make the statement clear:

– Identification of potato clones carrying an allele that determines tolerance to a pathogen within a reference population made up of tolerant and susceptible groups;

– Early stage individual forecast of phenotypic value for a late maturation trait in *Zea mais*;

– Classification of *Zea mais* cultivars for tolerance to stresses without the destruction of analyzed plants and without using expensive/dangerous physicochemical reactions.

In all these situations, a pilot experiment is performed to search for molecular markers that are informative about group membership.

Pilot experiments typically deal with molecular profiles of large size. Moreover, many of its components may be non-informative and expensive to assess. Therefore, in future experiments the researcher would like to utilize a subset of the profile surveyed in the pilot experiment.

The choice of profile component should be performed by means of the quantitative evaluation of convenience. We propose a model based on statistical decision theory to identify the best action in future experiments.

This paper opens with the notation and the description of pilot experiments dealing with molecular profiles. Then, the state of nature is defined as a parameter representing the association between band patterns and groups in the reference population. Inference about the unknown state of nature is performed following the Bayesian setting (Bernardo and Smith, 1994). The utility of an action (choice of a profile component) is defined as a trade-off between the information gain and the cost of assessment, thus the best action is found by maximizing the expected utility (Berger, 1985). Finally, the formalization of the decision problem reveals a few open problems and suggests future developments to perform the analysis of actual pilot experiments.

2. Information, actions and utility

Let $\mathcal{G}$ be the reference population made up of M known groups (subpopulations)

$$\mathcal{G} = \{G_0, \ldots, G_i, \ldots, G_{M-1}\}, \tag{1}$$

where it is assumed that the subpopulation size for each $G_i \in \mathcal{G}$ is unknown.

Let $\mathcal{O}$ be the set of observable bands included in the study

$$\mathcal{O} = \{O_1, \ldots, O_j, \ldots, O_K\}, \tag{2}$$

so that the observable band O_j, $j = 1, \ldots, K$, corresponds to a specific location on a electrophoretic gel that can be surveyed for the presence of DNA fragments. Then, each experimental unit is an individual from group $G_i \in \mathcal{G}$, and his digested DNA is found (presence=1) or not (absence=0) in O_j for $j = 1, \ldots, K$: the string of zeros and ones is called the *molecular profile* of the individual.

Let $U = \bigcup_{i=0}^{M-1} U_i$ be the set of sampled units, where $U_i = \{u_{i,1}, u_{i,2}, \ldots, u_{i,N_i}\}$ is the set of individuals from group G_i. In the adopted simple random sampling, the total number of experimental units N is fixed, and the number of units sampled from each group $G_i \in \mathcal{G}$ is not fixed in advance of the experiment. Besides, exchangeability is assumed for individuals sampled from the same group.

Let $\mathbb{B} = \{0, 1\}$ be a binary set in which 0 and 1 are, respectively, labels for presence and absence of DNA, and $\mathbb{B}^K$ be the Cartesian product of $\mathbb{B}$. Let $\Omega_Y = \{0, 1, \ldots, M-1\}$ be the set of integers used to label groups. A random vector

$$(Y, X_1, X_2, \ldots, X_K) \text{ on } \Omega_Y \times \mathbb{B}^K \tag{3}$$

is associated to each experimental unit, where X_j is the random variable related to O_j for $j = 1, \ldots, K$. A realization of the random vector (3) is indicated as $(y, x_1, x_2, \ldots, x_K)$. The vector $(X_1, X_2, \ldots, X_K)$ is also marked as $\mathbf{X}$ and a generic realization of $\mathbf{X}$ is $\mathbf{x} = (x_1, \ldots, x_K)$. A subscript z is used to indicate a specific band pattern $\mathbf{x}_z$ (realization of the random variable $\mathbf{X}$). The binary notation discloses the band pattern that corresponds to each value z in the set $\{0, 1, ..., 2^K - 1\}$. For example, the band pattern (0,1,1,0) is associated to $z = 6$, thus $\mathbf{x}_6 = (0, 1, 1, 0)$.

The set of actions $\mathcal{A}$ specifies how future experiments may be performed, that is the list of observable bands that will be surveyed. The information provided by the profile component $\mathcal{S} \subset \mathcal{O}$ about Y is not necessarily smaller than the information provided by $\mathcal{O}$, and if $\mathcal{S}$ is a proper subset then its cost of assessment may be smaller than the cost of assessing $\mathcal{O}$.

The set $\mathcal{A}$ is formally defined as

$$\mathcal{A} = \{a : a = 0, \ldots, 2^K - 1\} \tag{4}$$

where $a = 0 \equiv (0, \ldots, 0)$ means that no observable band will be assessed, that is $\mathcal{S} \equiv \varnothing$, the empty set; $a = 1 \equiv (0, \ldots, 0, 1)$ means $\mathcal{S} \equiv \{O_K\}$; $a = 2^K - 1 \equiv (1, \ldots, 1)$ represents the whole set of observable bands, that is $\mathcal{S} \equiv \mathcal{O}$. In other words, $\mathcal{A}$ is one-to-one with the power set $\mathcal{P}(\mathcal{O})$ built on $\mathcal{O}$, and the notation $\mathbf{X}_a \equiv \mathbf{X}_\mathcal{S}$ stands for the vector of random variables that includes only those X_j for $O_j \in \mathcal{S}$. A generic realization of $\mathbf{X}_a$ is indicated by $\mathbf{x}_a$.

The subscript z defined above has to be modified for indexing the band patterns of $\mathbf{X}_a$, but the rules to build it are unchanged. The symbol $\mathbf{x}_{a,z}$ indicates the specific band pattern that corresponds to z when the index z is built using $\mathbf{X}_a$ instead of $\mathbf{X}$. That is, $z \in Z(a)$ is the full notation for the dependence of the subscript values by a, with $Z(a)$ the set of values assumed by z. An implicit redefinition of z is chosen to keep the notation simple.

The data set $\mathbf{d}$ from the pilot experiment performed by the researcher is

$$\mathbf{d} = \{(y, x_1, x_2, \ldots, x_K)_u : u \in U\}. \tag{5}$$

It is used to increase the knowledge about the state of nature (see the next paragraph).

Note that a band pattern $\mathbf{x}_z$ may have a not null frequency in several (all) groups of $\mathcal{G}$, a situation that is assumed to be typical of a generic reference population.

2.1. Bayesian inference about the state of nature

A discrete probability density for the random vector (3) is defined conditional on a vector of parameters $\boldsymbol{\theta}$, the state of nature, as

$$p(Y, X_1, X_2, \ldots, X_K \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta, \tag{6}$$

where $\boldsymbol{\theta} = (\theta_{0,0}, \ldots, \theta_{i,z}, \ldots, \theta_{M-1,t})$, $\sum_{i,z} \theta_{i,z} = 1$, $0 \leq \theta_{i,z} \leq 1$, and where $t = 2^K - 1$ and 2^K is the number of band patterns. Each element $\theta_{i,z}$ represents the probability of observing the band pattern $\mathbf{x}_z$ in the subpopulation G_i, as it could be made explicit using a two-way table 'group by band pattern'. In other words, this is a saturated model for a two-way contingency table.

The likelihood $L(\mathbf{d} \mid \boldsymbol{\theta})$ is proportional to

$$L(\mathbf{d} \mid \boldsymbol{\theta}) \propto \prod_{i=0,z=0}^{M-1,t} \theta_{i,z}^{n_{i,z}}, \tag{7}$$

where $n_{i,z}$ are counts of experimental units carrying the band pattern $\mathbf{x}_z$ that belong to G_i.

Equations (6) and (7) require some comments. First, the model for the two-way table is saturated. This choice is due to the high number K of observable bands in relation to the sample size N, a typical feature of many pilot experiments, e.g.

$K = 1000$ and $N = 100$. In these situations the search space of reduced models is huge, and the support to discriminate among different models may be weak.

Second, the Bayesian paradigm (Bernardo and Smith, 1994) is adopted to specify the degree of belief about the state of nature θ in a quantitative way. In other words, the uncertainty about θ is modelled by a probability distribution, thus parameters are considered as random variables. Besides having nice theoretical features (e.g., Box, 1983; O'Hagan, 1994; Bernardo and Smith, 1994), Bayesian inference seems suited for the needs of molecular profile analysis. This point is made clear by the description of the relationship between the genetic system and the proposed model.

The experimental observations in $\mathbf{d}$ provide information about θ. It is assumed that the genetic features of $\mathcal{G}$ do not change fast enough to invalidate inferences about future experiments based on $\mathbf{d}$. This is a key point because:

– No explicit mechanism accounting for changes of $\mathcal{G}$ along generations is considered;

– Reference populations with sexual reproduction, short life cycle, high mutation rates (at least in regions of the genome related to $\mathcal{O}$) and genetic selection may show a sharp change of the association pattern among variables $(Y, X_1, \ldots, X_K)$ on a short time scale; in this situation no useful inference based on $\mathbf{d}$ is possible for future experiments;

– No abstract space is adopted to minimize the effects of population changes or to make the characterization of the dynamics easier (e.g., recombination maps).

Small changes in the state of nature over time-generations may be interpreted as realizations of a random variable in the frequentist sense. Moreover, Bayesian inferential statements based on pilot experiments are explicitly conditional on the obtained data (available experimental information), so that information updating is formally allowed. Moreover, prior information may be included in the analysis if available. Finally, the Bayesian approach allows for the smoothing of parameter values, an important feature for the analysis of pilot experiments.

The specification of a prior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$, with $\boldsymbol{\lambda}$ a vector of constants, for the parameter $\boldsymbol{\theta}$ fully defines the model. An over-dispersed Dirichlet distribution (Bernardo and Smith, 1994) may be conveniently used to define the uncertainty about the state of nature that is typical of many experimental setups. Moreover, the advantages of a Bayesian conjugate setting are also available. The vector of constants $\boldsymbol{\lambda} = (\lambda_{0,1}, \ldots, \lambda_{M-1,t})$ that defines the prior distribution has elements

$$\lambda_{i,z} = \frac{1}{M \cdot t}, \text{ for } i = 0, \ldots, M-1; z = 0, \ldots, t, \tag{8}$$

where $\lambda = \sum_{i,z} \lambda_{i,z} = 1$. The adopted value of parameters makes the prior distribution quite non-informative, although this definition is not unique (Bernardo and

Smith, 1994, par 5.6.2). The vector of parameters related to the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{d})$ has elements

$$\widetilde{\lambda}_{i,z} = \lambda_{i,z} + n_{i,z}, \tag{9}$$

where $n_{i,z}$ is defined in (7) and $\widetilde{\lambda} = \sum_{i,z} \widetilde{\lambda}_{i,z} = N + 1$.

2.2. The utility of molecular profile components

The utility of assessing molecular profiles in samples from the reference population $\mathcal{G}$ depends on the decrease of uncertainty due to the conditioning variables $\mathbf{X}_a$, $a \in \mathcal{A}$, and on the overall monetary cost of assessment (number and type of observable bands, equipment, reagents, work of technicians, etc.).

The quantitative evaluation of the gain in information due to profile components is based on the family of conditional distributions

$$\mathcal{F} = \{p(Y \mid \mathbf{x}_{a,z}, \boldsymbol{\theta}) : a \in \mathcal{A}, z \in Z(a), \boldsymbol{\theta} \in \Theta\}. \tag{10}$$

The elements of $\mathcal{F}$ capture the uncertainty about Y given the state of nature and each band pattern that may be observed when the action a is chosen. Note that the data set $\mathbf{d}$ does not appear in the conditional distributions of equation (10).

A subset $F \subset \mathcal{F}$ is associated to each $a \in \mathcal{A}$. The information content of distributions that belong to F may be quantified using the definition of entropy of a discrete distribution (Bernardo and Smith, 1994)

$$H(p_{Y|\boldsymbol{\theta}}) = -\sum_{Y=0}^{M-1} p(Y \mid \boldsymbol{\theta}) \cdot \log\left(p(Y \mid \boldsymbol{\theta})\right) \tag{11}$$

where the marginal distribution $p(Y \mid \boldsymbol{\theta})$ of Y is chosen as an example.

The uniform distribution has the maximum amount of entropy, thus it may be used as a yardstick to quantify the gain in information due to the conditioning variables. The gain due to the knowledge of the state of nature θ and the action $a = 0$ ($\mathcal{S} \equiv \emptyset$) is

$$\mathcal{I}(p_{Y|\boldsymbol{\theta}}) = \mathcal{I}(p_{Y|\mathbf{x}_{0,z}\boldsymbol{\theta}}) = \mathcal{I}(p_{Y|\mathbf{x}_0\boldsymbol{\theta}}) = \sum_{Y=0}^{M-1} p(Y \mid \boldsymbol{\theta}) \cdot \log\left(\frac{p(Y \mid \boldsymbol{\theta})}{M^{-1}}\right), \tag{12}$$

because the state of nature does not necessarily correspond to groups of equal size in the reference population.

If the information contained within the molecular profile is used, then the information gain given the state of nature $\boldsymbol{\theta}$ and the band pattern $\mathbf{x}_{a,z}$ is

$$\mathcal{I}(p_{Y|\mathbf{x}_{a,z},\boldsymbol{\theta}}) = \sum_{Y=0}^{M-1} p\left(Y \mid \mathbf{x}_{a,z}, \boldsymbol{\theta}\right) \cdot \log\left(\frac{p\left(Y \mid \mathbf{x}_{a,z}, \boldsymbol{\theta}\right)}{M^{-1}}\right). \tag{13}$$

The overall expected gain in information for the different values of band pattern $\mathbf{x}_{a,z}$ depends on the probability of sampling each band pattern

$$\mathcal{I}(p_{Y|\mathbf{X}_a,\boldsymbol{\theta}}) = \mathbf{E}^{\mathbf{X}_a}\left[\mathcal{I}(p_{Y|\mathbf{x}_{a,z},\boldsymbol{\theta}})\right] = \sum_z p\left(\mathbf{x}_{a,z} \mid \boldsymbol{\theta}\right) \cdot \mathcal{I}(p_{Y|\mathbf{x}_{a,z},\boldsymbol{\theta}}), \tag{14}$$

where $p\left(\mathbf{X}_a \mid \boldsymbol{\theta}\right)$ is the marginal distribution obtained integrating out from the vector (3) the random variable Y and the X components not included in the action a.

If a preference scheme can be expressed by a system of weights $\mathbf{w} = (w_1, w_2, \ldots)$, where $\sum_z w_z = 1$, then the expected weighted gain in information may be defined by

$$\mathcal{I}_{\mathbf{w}}(p_{Y|\mathbf{X}_a,\boldsymbol{\theta}}) = \mathbf{E}^{\mathbf{X}_a,\mathbf{w}}\left[\mathcal{I}(p_{Y|\mathbf{x}_a,\boldsymbol{\theta}})\right] = \sum_z p\left(\mathbf{x}_{a,z} \mid \boldsymbol{\theta}\right) \cdot w_z \cdot \mathcal{I}(p_{Y|\mathbf{x}_{a,z},\boldsymbol{\theta}}), \tag{15}$$

and a similar weighted expression $\mathcal{I}_{\mathbf{w}}(p_{Y|\boldsymbol{\theta}})$ may be also obtained, using a similar weighting of eq. (12). The eq. (15) may be used if the researcher has special interest for one or more groups, e.g. $Y = 0$ being the group of individuals that are resistant to high pesticide doses.

The second component of the utility deals with the cost of assessment for the profile component defined by $a \in \mathcal{A}$.

A utility function $\Theta \times \mathcal{A} \rightarrow \Re$ is defined to map the state of nature $\boldsymbol{\theta}$ and the action a to a real number that quantitatively expresses the value of an action for $\boldsymbol{\theta}$ known

$$u(\boldsymbol{\theta}, a) = \alpha \cdot \mathcal{I}(p_{Y|\mathbf{X}_a,\boldsymbol{\theta}}) + \beta(\mathbf{X}_a), \tag{16}$$

where α is a positive constant and $\beta(\mathbf{X}_a)$ is a function that expresses the overall cost of assessment if the action a is chosen, thus it has negative values. A positive value of u implies a monetary gain, while a monetary loss is indicated by a negative value. It is clear that $a = 0 \Rightarrow \mathcal{S} \equiv \emptyset \nRightarrow \beta(\mathbf{X}_a) = 0$ if the specification of costs is not restricted to the assessment of observable bands.

It is important to underline that the definition of utility function, i.e. α and $\beta()$ in eq. (16), is subjective and specific for the experimental setup that is considered. Nevertheless, the definition in eq. (16) may be useful in many studies, because besides its generality it also plays the role of basic definition for highly tuned and more specific mappings. A straightforward extension of eq. (16) is obtained using weighted versions of information gain.

3. The decision as a convenient action

It is assumed that a specific equation like (16) may be built using utility theory (Berger, 1985). The choice of an action $a \in \mathcal{A}$ requires more than eq. (16) because the state of nature, i.e. $\boldsymbol{\theta}$, is usually unknown.

Nevertheless, some knowledge about $\boldsymbol{\theta}$ is given by the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{d})$, that is based on the data set $\mathbf{d}$ of experimental observations.

The expected value of the utility function defines the expected monetary gain (or loss) due to the choice of action a and the available information about the state of nature. The conditional Bayes principle (Berger, 1985) prescribes the action a that maximizes the expected utility, that is

$$\max_a \int_\Theta u(\boldsymbol{\theta}, a) \cdot p(\boldsymbol{\theta} \mid \mathbf{d}) \cdot d\boldsymbol{\theta} \tag{17}$$

where the notation embeds some steps of algebra required to derive the posterior distribution that is suited for computations with the state of nature related to action a. In short, the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{d})$ has to be transformed by appropriate addition of those cell parameters that are not distinguished under the choice of action a.

The computation of the integral in eq. (17) may be difficult, especially for a generic utility function, so Monte Carlo integration is often required (Berger, 1985).

If eq. (17) may be (approximately) calculated for each $a \in \mathcal{A}$ then the decision may be:

(i) $a = 0$; if values are negative for each action then the experiment should not be performed;

(ii) $1 \leq a \leq 2^K - 1$; if only one positive (or one maximum) value is found then the correspondent action should be chosen, either including or not one or more observable bands;

(iii) $\widetilde{a} \in \widetilde{\mathcal{A}}$; if two or more actions in the set $\widetilde{\mathcal{A}} \subset \mathcal{A}$ have equal positive value then the randomized choice of action $\widetilde{a}$ (with equal probability over $\widetilde{\mathcal{A}}$) is required.

Nevertheless, a different approach is available when the last situation above occur. It is based on the idea that a suboptimal action (smaller expected utility then the best) may exists so that the expected utility is close to the best value, but several (all) maximizer actions are embedded in the final decision.

Let $\widetilde{\mathcal{A}} = \{a_1, a_2, \ldots\}$ be the set of actions, called generators, that have the same value of expected utility, as specified by eq. (16) and (17). Then, the correspondent set of profile components is $\widetilde{\mathcal{O}} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots\}$, with $\mathcal{S}_i$ the generic element related to a_i. The power set $\mathcal{P}\left(\widetilde{\mathcal{O}}\right)$ has generic element P. For each element $P \in \mathcal{P}\left(\widetilde{\mathcal{O}}\right)$, observable bands included in the profile components belonging to P are merged into

one profile component

$$\mathcal{S}_p = \bigcup_{\mathcal{S} \in P} \mathcal{S}, \tag{18}$$

therefore the set of generators of $\mathcal{S}_p$ is the subset $\widetilde{\mathcal{A}}_p \subset \widetilde{\mathcal{A}}$ of actions that specifies the components in P. Moreover, an action $\widetilde{a}_p$ is associated to $\mathcal{S}_p$. A reformulation of the utility function in eq. (16) is required for actions obtained using generators

$$\widetilde{u}(\boldsymbol{\theta}, \widetilde{a}_p) = \alpha \cdot \max_{a \in \widetilde{\mathcal{A}}_p} \mathcal{I}(p_{Y|\mathbf{X}_a, \boldsymbol{\theta}}) + \beta(\mathbf{X}_{\widetilde{a}_p}), \tag{19}$$

where the cost associated to $\widetilde{a}_p$ is not changed, but the information gain is given by the maximum obtained considering special subsets of the profile component, those related to the generators in $\widetilde{\mathcal{A}}_p$.

Eq. (19) may substituted in eq.(17) to identify 'almost best' actions, with as many actions $a \in \widetilde{\mathcal{A}}$ embedded in $\widetilde{a}_p$ as allowed by the increase of experimental cost.

Note that the computational burden due to the straightforward definition of the utility function on the power set of a component $\mathcal{S}$ is avoided by delaying these calculations to the situation in which they are required.

If values of expected utility found using equation (19) are far from the values defining the set $\widetilde{\mathcal{A}}$ then the randomized choice of action has to be performed.

4. Discussion

The proposed model and algorithms are just a sketch that may be followed to choose the profile component that will be included in future experiments by evaluating the convenience in a quantitative way.

In the paragraphs of this section some of the possible obstacles and pitfalls are described, together with open problems.

4.1. The set $\mathcal{O}$ of observable bands

The definition of $\mathcal{O}$ should be performed by means of criteria suited to general use.

In some studies, an observable band is included in $\mathcal{O}$ only if one sampled individual shows that DNA band, thus

$$\mathcal{O} \equiv \mathcal{O}(\mathbf{d}). \tag{20}$$

If the sample size is small, say 100, this definition of $\mathcal{O}$ causes a severe bias in the quantitative evaluation of convenience, because the true sample space is larger then considered.

The set $\mathcal{O}$ should be considered a feature of the experimental setup that does not depend on observed data. Some elements of the setup that influence the definition of $\mathcal{O}$ are: the type of restriction enzymes and of DNA probes used in the reactions, the bio-chemical protocol, the electric equipments required to assess the presence of bands.

Nevertheless, difficulties in the assessment of bands may also arise in some specific experimental setup (e.g., Navidi and Arnheim, 1994), especially if a too small quantity of DNA is amplified. In those situations, the resolution power of the equipment is not clearly defined, therefore the definition of $\mathcal{O}$ may be uncertain.

Extensions of our model to deal with errors-uncertainty in measurements should be investigated.

4.2. The computational burden

Equation (17) may be evaluated by exhaustive calculation if the number of observable bands K is relatively small, say 50. For $K = 1000$ or more, straightforward calculations are unfeasible, so that simpler expressions are required.

A first strategy deals with the simplification of the data set $\mathbf{d}$ before it is used in a decision-based framework. Stefanini and Camussi (1997) investigated the use of Genetic Classifier Systems to decrease the number of observable bands K to a smaller number by evaluating the information content of each observable band averaged on the set of highly useful band patterns identified by the algorithm. The computational burden and the lack of a smoothing mechanism discouraged further investigations on this line.

A different approach is obtained by simplifying equation (17). Recently, Stefanini (1998) proposed the optimization of a quantity based on the predictive distributions

$$\mathcal{P} = \{p(Y \mid \mathbf{x}_{a,z}, \mathbf{d}) : a \in \mathcal{A}, z \in Z(a)\}, \tag{21}$$

where

$$p(Y \mid \mathbf{x}_{a,z}, \mathbf{d}) = \int p(Y \mid \mathbf{x}_{a,z}, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta} \mid \mathbf{d}) \cdot d\boldsymbol{\theta}, \tag{22}$$

that may be obtained in a closed form, so that the computation of eq.(17) is avoided. The search for optimal actions is performed using a Genetic Algorithm specifically designed for this purpose. Promising results were obtained using simple simulated data sets.

Besides Evolutionary Computation, other methods to simplify equation (17) may be investigated. As example, if the budget for the next experiment is highly constrained to a maximum b then it might be possible to define a maximum number of observable bands $c = h(b)$ that can be assessed, so that whatever the information gain expected from actions involving $c' > c$ observable bands, the utility is set to the

value of a big loss. In other words, the set of actions $\mathcal{A}$ is reduced to the subset

$$\mathcal{A}_c = \{a : \#(a) < c, a \in \mathcal{A}\} \tag{23}$$

where the notation $\#(a)$ indicates the number of observable bands defined by a. Equation (17) requires less computation because it is evaluated on $\mathcal{A}_c \subset \mathcal{A}$.

4.3. Multiband enzymes and probes

In the discussion above, it was assumed that each observable band is generated by a feature of the experimental setup that may be changed independently of the others, e.g. each observable band is associated to a specific restriction enzyme. In this situation, the definition of $\mathcal{A}$ in eq. (4) is suited to the features of the experimental setup.

If strong dependence is present among profile components then the set $\mathcal{A}$ has to be modified. For example, if the enzyme $E_h, h \in H$ (defined by the experimental setup) generates several observable bands in each individual, then a profile component can not be changed by excluding only one observable band from the set generated by E_h.

In future experiments, the choice of a profile component implies the inclusion of all the observable bands associated to one or more useful observable bands, due to experimental constraints (dependence). Thus the set of actions $\mathcal{A}$ is transformed into $\mathcal{A}/H$, the set made up of equivalence classes induced by the dependencies of H.

The utility function has to be changed so that the information gain is the maximum on the set of actions obtained by ignoring some of the observable bands that will be observed due to dependencies but that are not useful.

Finally, the definition of $\mathcal{A}/H$ may be hard because it is based on the results of the experimentation, thus being subject to sampling noise.

4.4. Missing values

In several situations, the data set $\mathbf{d}$ contains missing values. If many missing values are sparse in the data set, the trivial strategy of ignoring observable bands carrying missing information is unfeasible (curse of dimensionality).

Bayesian methods of imputation (Shafer, 1997) handle these situations, but the computational burden may be unpractical. Simpler procedures (Stefanini, 1998) may offer approximated computations suited to data sets with large number of observable bands.

5. Conclusions

In this paper, a procedure based on statistical decision theory is developed to choose the profile component that should be included in future experiments regarding the reference population studied in a pilot experiment. The leading criterion is a trade-off between the gain in information about group membership and the experimental cost of assessment.

The approach is sketched in its main components, so that the discussion of some open problems is not vanishing. Therefore, the analysis of actual data should address points that here are only mentioned. First, the prior distribution in equation (8) is suited for the situation of ignorance about θ. If some knowledge is available due to other sources of information, then it should be properly used. Second, the choice of a specific utility function may involve extra-monetizing considerations, as it happens when pure scientific purposes are related to the experimentation. In those situations costs may still be expressed on a monetary scale, but the gain in information is less tractable on that scale.

The choice of molecular profile components also involves topics from the field of experimental design if the pilot experiment is not performed at the time of the analysis. It may also be the case that the reference population is not clearly defined, so that special extensions to the proposed methods should be developed.

We hope to have raised the interest of many researchers on this subject, whose economic implications are far from being immaterial.

Acknowledgments

The 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. We thank the reviewers for their helpful comments and Ludovico Piccihàto for his constructive criticism of the manuscript.

References

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

Berger J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, Berlin.

Bernardo J.M., Smith A.F.M. (1994). *Bayesian Theory.* Wiley, New York.

Box G.E.P. (1983). An apology for ecumenism in Statistics. In: G.E.P. Box, T. Leoneard and C. Wu (Eds.), *Scientific Inference*, Data Analysis and Robustness, Wiley, New York.

Navidi W., Arnheim N. (1994). Analysis of genetic data from the polymerase chain reaction. *Statistical Science* **9**(3), 320-333.

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

O'Hagan A. (1994). *Bayesian Inference.* Kendall's Advanced Theory of Statistics, Edward Arnold, London.

Shafer J.L. (1998). *Analysis of Incomplete Multivariate Data.* Chapman & Hall, New York.

Stefanini F.M., Camussi A. (1997). Information in molecular profile components evaluated by a Genetic Classifier System: a case study in Picea abies Karst. *Genetical Research Cambridge* **70**, 205-213.

Stefanini F.M. (1998). Identification of highly informative molecular profile components using Genetic Algorithms. Working paper 98-05-042, Santa Fe Institute, New Mexico.

Received 12 September 1998; revised 4 March 1999

Wybór składników profilu molekularnego na podstawie ilościowej definicji przydatności

STRESZCZENIE

Profile molekularne wyznaczane poprzez elektroforezę mogą być wykorzystane do klasyfikacji osobników. Niepewność co do przynależności grupowej może być zmniejszona jeżeli pewne informatywne profile są zidentyfikowane poprzez doświadczenie pilotażowe. W pracy analizuje się takie doświadczenia z punktu widzenia statystycznej teorii decyzji. Celem jest znalezienie takich składników profili molekularnych które powinny być uwzględnione w dalszych eksperymentach. Definiuje się kompromis pomiędzy kosztem doświadczenia a niepewnością w ocenie profilu jako ilościową ocenę przydatności różnych poczynań. Analiza bazuje na znanych metodach, jednak doświadczenia pilotażowe z profilami molekularnymi wskazują na wiele praktycznych, otwartych problemów.

SŁOWA KLUCZOWE: wnioskowanie Bayesowskie, teoria decyzji, profile molekularne.