Tatra Mt. Math. Publ. **26** (2003), 307–321

TATRA MOUNTAINS
Mathematical Publications

BIOINFORMATICS: SOME CHALLENGING STOCHASTIC PROBLEMS

Pranab Kumar Sen

ABSTRACT. Bioinformatics is the emerging field of science growing from the application of mathematics, statistics, and information technology, including the computers and the theory surrounding them, to study and analysis of very large biological, and in particular genetic data sets. The field has been fueled by the increase in the DNA data generation. At the present, gene scientists can not scramble fast enough to keep up with the genomics, emerging at a furious rate and in astounding detail, while inspired by computational aspects and demand for numerical evidence, data mining algorithms are on the rampage. Yet stochastic evolutionary undercurrents make it essential to incorporate statistical methods based on probability theory and stochastic processes. Along with a broad review of the basics of bioinformatics, some of these challenging statistical problems are outlined, and their resolutions are appraised.

1. Introduction

With the advent of modern information technology as well as scientific curiosity in genomics, we are at the dawn of bioinformatics, though it is not precisely known what really constitutes its core domain. Pharmaceutical researchers, molecular biologists, computer scientists, (bio-)mathematicians, and statisticians (at large) all differ in their definitions and interpretations, albeit to a much lesser extent than in their basic emphasis on methodological versus computational aspects. As such, there is a need to formulate the core domain of bioinformatics and appraise the stochastic aspects thoroughly. In computational biology, and bioinformatics, in general, principles of *molecular genetics* provide room for *stochastics* to resolve the basic differences between *biological diversity* and *mathematical exactness*. Typically, enormously large dimensional purely qualitative (categorical) data models with rather imprecise (stochastic) dependence patterns are encountered here. As such, even *discrete multivariate analysis* may encounter roadblocks of various types, and (partial-, profile-,

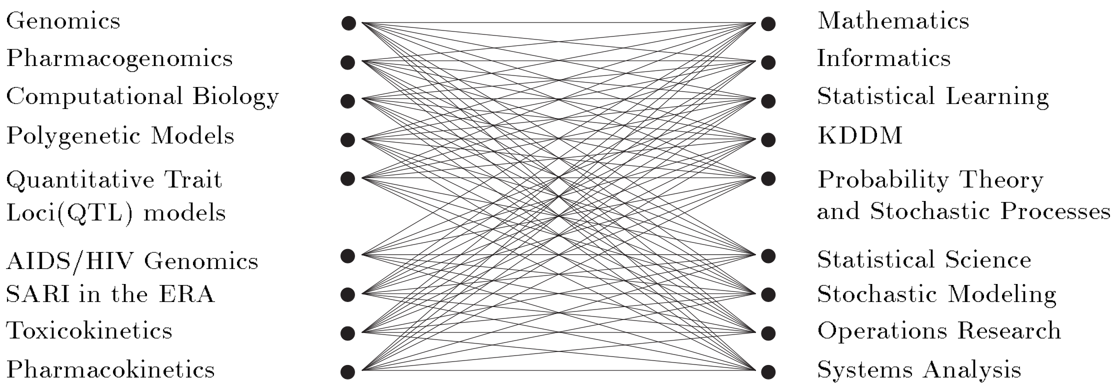
2000 Mathematics Subject Classification: 62F99, 62G99, 62P10.
Keywords: biostochastics, CSA, DNA, genomics, high-dimensional models, KDDM, large biological system, likelihoods, polygenics.

pseudo-, conditional, or quasi-) likelihoods may not be very appropriate for statistical modeling and analysis. Consequently, encountered with the immense computational aspects and the need for computational algorithms, *knowledge discovery and data mining* (KDDM) or *statistical learning* approaches have so far dominated the field ([5]). Compared to the likelihood based approaches, the limitations of the KDDM approaches are no less apparent, and there are some challenging stochastic problems in this interdisciplinary field; it is our aim to appraise some of them in a systematic manner.

In Section 2, we provide a broad outline of the domain of bioinformatics and biostochastics in relation to other allied sciences. Section 3 is devoted to the stochastic aspects of the contemplated bioinformatics setups. Section 4 deals with some of the challenging problems, and in the concluding section some general remarks and observations are made.

2. Bioinformatics vs. biostochastics

To accommodate the diverse disciplines that pertain to the core of bioinformatics, we could at best draw the following network that illustrate the need for quantification of information in biological systems.



(Here SARI stands for structure activity relationship information and ERA for environmental (health) risk assessment [14].)

Based on this composite network, it may be quite difficult to outline precisely the fundations of bioinformatics; we follow Ewens and Grant [4] and quote the following: *We take bioinformatics to mean the emerging field of science growing from the application of mathematics, statistics, and information technology, including the computers and the theory surrounding them, to the study and analysis of very large biological, and particularly genetic data sets. The field has been fueled by the increase in* DNA *data generation.*

Bioinformatics does not aim to lay down the fundamental mathematical laws that govern biological systems parallel to those laid down in physics. There is, at this stage, mathematical utility in the creation of tools that investigators can use to analyse data (e.g., statistical assessment of the similarity between two or more DNA or Protein sequences, finding genes in genomic DNA, and estimating differences in how genes are expressed in different tissues etc.). Such tools involve statistical modeling of biological systems, and this in turn, calls for development of statistical reasoning to meet such a challenging task. We may also refer to Waterman [17] for an account of statistical methods in computational biology.

At the current stage, gene scientists (even borrowing strength from modern information technology) cannot scramble fast enough to keep up with the science behind genomics, emerging at a furious pace and in astounding detail. It would be improper to jump to scientific conclusions based on data analysis alone (even under the engulfing environment of knowledge discovery and data mining (KDDM)). There is a genuine need to examine the *stochastic evolutionary forces* underlying genomics, and in that way, to have a better grasp of the interdisciplinary genomic science. Primarily driven by this motivation, we use the following terminology:

Bioinformatics: Application of mathematics, statistics and informatics to study and analysis of very large biological, and particularly genetic, data sets.

Note that all large biological systems may not (at least, up to the present time) have an overwhelming genetic undercurrent, and yet, they may need more refined statistical tools. Neuronal spatio-temporal models, toxicokinetics and aftermaths of human exposures to toxic and other harmful environmental substances, etc., are notable examples of this type. As such, we place side by side the following terminology (Sen [12]):

Biostochastics: Stochastic modeling and analysis (i.e., stochastics) in very large biological (including genetic and genomic) data sets.

From a statistical perspective, it would be our advantage to view bioinformatics as a subdiscipline of biostochastics (albeit from a molecular biology perspective, the picture could be different).

In order to comprehend the basic differences between the two setups, we consider some current problems in bioinformatics and bring the relevance of biostochastics more explicitly.

(I) DNA*, the carrier of genetic information.* Gene, in the Mendelian setup, is the basic unit of inheritance. Genes occur at definite sites, or *loci*, on *chromosomes* which are strings of *Deoxyribonucleic acid* (DNA), the basic genetic material in a cell. There are two other molecules, namely *Ribonucleic acid* (RNA) having a single strand, and *Proteins*, composed of 20 amino acids, which are intricately interrelated. DNA is a double helical molecule, a polymer, made up of nucleotides which are 4 in number, designated by the 4 bases A (adenine), C (cytosine), G (guanine) and T (thymine). DNA contained in the cell is known as the *Genome*. Both RNA and proteins are made from instructions in the DNA, and new DNA molecules are made from copying existing DNA molecules. In DNA, A pairs with T and G pairs with C; A and G are *purines*, while C and T are *pyrimidines*. Unlike proteins (whose 2- and 3-dimensional structures depend critically upon the composition and order of amino acids), DNA has, under usual conditions, a quite regular and nearly identical structure regardless of the composition or order of the 4 bases: A, C, G, T.

If we accept the basic hypothesis that DNA is the blueprint for a living organism then it becomes natural to conclude that molecular evolution is directly related to changes in DNA; during the course of molecular evolution substitutions occur. The substitutions A $\leftrightarrow$ G or C $\leftrightarrow$ T are called transitions., while A $\leftrightarrow$ C, A $\leftrightarrow$ T, G $\leftrightarrow$ C, and G $\leftrightarrow$ T are called transversions. Also, note that amino acids are encoded by triplets of nucleotides of DNA called codons. Let us define $\mathcal{N}_R = \{A, C, G, U\}$ and set $\mathcal{C} = \big\{(x_1, x_2, x_3) \colon x_j \in \mathcal{N}_R ,\ j = 1, 2, 3\big\}$ (be the codon). Finally, let $\mathcal{A}$ be the set of amino acids and termination codon. Then the genetic code can be defined as a map:

$$g \colon \mathcal{C} \to \mathcal{A}\,, \quad g \in \mathcal{G}\,,$$

so that $\mathcal{G}$ is the set of all genetic codes $\{g\}$.

Stochastic evolutionary forces act on genomes, and hence, probability theory and statistical reasoning are very pertinent in genomic studies, though there is more emphasis on computational aspects. Moreover, it should be noted that genes are not simple. As a matter of fact, the very basic high-dimensionality and yet unknown nature of the battery of activities of the genes, especially in the evolutionary phase, have created an enormous task for molecular biologists and geneticists in the years to come. As the human genome project is heading for completion, there are other formidable statistical tasks that are surfacing into research efforts to fathom out the mystery of the genetic code. For example, locating the transcription regulatory sites in the genome needs a large number of genome sequence comparisons. Elucidating the human transcriptional regulatory network is indeed a challenging task, and statistical reasoning is as much important as computational algorithms. In the next section, some of these points would be elaborated in a more general framework.

(II) *Proteins: The key to unlocking human biology.* Protein is not only a vital part of human diet, it is in the forefront of biotechnology and drug research. The human genetic map shows there are some 35,000 genes responsible for instructing proteins, which come in a far greater number of variations. By identifying these proteins, doctors would also have a way to detect disease or know ahead of time whether a certain patient will respond to a drug. Tracking and sorting the thousands of proteins in our bodies is a task of awesome technical complexity, requiring a special combination of biological and computational skills.

DNA microarrays, or biochips, are readily available for sorting tens of thousands of genes simultaneously, but there is still no dominant technology for identifying or characterizing proteins, which consists of myriad patterns of 20 amino acids folded into delicate 3-dimensional shapes. The shapes determine their functions. DNA chips and protein chips have the same fundamental concept, but the underlying technology is very different. DNA can be treated almost harshly, but proteins are very delicate. Drug researchers are working on high-throughput protein chips which would differentiate the proteins, once there is a clue. Large scale biology, emphasizing proteomics, is focussed primarily on developing contents for chips. The protein chips could also be used to determine quickly whether a trial patient responds to treatment. This would not only improve the quality of trials but also increase the number of trials that can be conducted. Thus, proteins are likely to be used in microarray formats and have applications in a wide range of other formats.

(III) *Inflammatory Bowel Disease* (IBD) and its genetic undercurrents. IBD or the Crohn disease is a chronic inflammation that shreds the lining of the digestive tract. Scientists in USA and France have pinpointed a genetic abnormality in humans—one originally discovered in plants- that increases susceptibility to IBD. A mutation to the gene on chromosome 16 increases a person's chances of developing an inherited form of the IBD by at least twenty-five percent (Nature, May 31, 2001). Prior to this discovery, the general belief used to relate certain environmental factors (toxins) to IBD: though this belief can still be sustained, the emphasis has been shifted to finding which factors induce or influence this mutation to the gene on chromosome 16 and also to assess whether this impact is particularly due to this specific gene or shared by others in the same or other chromosomes. In this way mapping of a disease gene and mutation are on the research task list, and so, we are in bioinformatics! Nevertheless, the assessment of susceptibility and chances are clearly in the domain of stochastics, so that biostochastics may be equally pertinent in this respect. Also, relating such a genetic abnormality to suitable environmental factors needs a lot of toxicologic as well as stochastic risk assessment studies wherein biostochastics occupies a focal point.

(IV) *Toxicology and Bioinformatics*: Developmental toxicology deals with adverse effects in the developing organism due to exposures to either parent before the actual conception or during the prenatal or the postnatal development. These include still-birth, birth defects, teratogenesis, and structural abnormality, altered growth, and abnormal functions. Reproductive toxicology deals with adverse effects on male and female resulting from exposures to specific toxicants that may lead to reduced fertility, alterations in sexual behavior, adverse pregnancy outcomes, and alterations of other functions relating to the reproductive system (including depression). Neurotoxicology deals with the impact of (mostly, environmental) toxicity on human nervous system which is controlled by the (cortex based) *central nervous system* (CNS). A less than ideal level of functioning of the CNS and the cortex may not only result in serious diseases or disorders but also affect the personality and sense of security. Thus, taken together, various subdisciplines of toxicology are to be judged by their toxicokinetic structures, mode of uptake, and *in vivo* activities following their intake, the latter phase being essentially related to bioinformatics to a greater extent (as they relate to large biological systems where genetic aspects can not be ignored). When it comes to such *structure activity relationship information* (SARI) we are dragged into the domain of biostochastics, even if we do not pursue the genetic aspects to a greater extent ([14]). The evolution of the interdisciplinary field of *toxicogenomics* has its genesis in this field, and is likely to play a very significant role in genomics research from pharmacological and environmetal stressors points of view.

(V) *Neuronal Spatio-Temporal Models* (NSTM) relate to large neurobiological systems (not yet totally engulfed by the bioinformatics evolution) where neurophysiology provides the basic biological background (structure) and thereby nurtures the stochastics (activity), resulting in some less precise biologic hypotheses amidst stochastic clouds. The cortex field having millions of nerve cells for each of which we have a spike train model. However, for observing neuronal spike train, one needs to contact the cell electronically which may destroy the cell. Therefore, it may neither be feasible or advisable to get the realization of spike trains for the entire cortex manifold, so that one ends up with a gridded data set of finite number of grid points. Therefore, from a statistical perspective, there is the basic problem of selecting a number of such grid points in an optimal way so as to draw effective conclusions for the entire cortex field based on such a discretized subset. In this respect, conventional stochastic modeling and analysis may not work out, and we may need to explore biostochastics to a greater extent (Sen [15]).

There are numerous other examples of this type, and for our purpose, we shall bypass some of them, and illustrate the main features with motivations from the above illustrations.

3. Stochastics in bioinformatics

Among the various important problems in bioinformatics, and biostochastics in general, we may mention the following:

(i) Modeling DNA or Protein sequence data.
(ii) Analysis of one or multiple DNA or Protein sequences.
(iii) Alignment of two sequences.
(iv) Comparison of two aligned sequences.
(v) Evolutionary models, QTL models.
(vi) Phylogenetic tree estimation.
(vii) PBPK and PBTK models.
(viii) Carcinogenic evolutionary models.
(ix) Pharmacokinetics, Pharmacogenetics and Pharmacogenomics.
(x) Hidden Markov models, MCMC, etc.

All these models and associated problems can be viewed from stochastic perspectives in a common vein so as to have a better perception of the underlying stochastic complexities as well as the challenging task resulting therefrom.

First and foremost, the challenging task is due to the excessively large dimension of observational data sets, and this has been stressed in the preceding section. Secondly, in the context of (i), (ii), (iii), (iv), (ix) and (x), we may have a large number of purely qualitative response variables without much structure so that their joint distribution and dependence patterns may not be precisely known or well anticipated. If a model is assumed purely from mathematical convenience point of view, even if that is not too different from the true model, because of the high-dimensionality and unstructured dependence patterns, statistical conclusions based on the assumed model may not be very reliable; thus, validity and robustness considerations are very important. Thirdly, in statistical modeling and analysis, there is a need to import some basic regularity assumptions to focus stochasticity in a manageable way, and in the light of that, to formulate suitable likelihoods that would yield appropriate statistical inference tools. Clearly, in many such high dimensional nonstandard data models, a precise formulation of a likelihood function, or some of its variants, may encounter roadblocks, and this in turn, may raise questions about optimality or desirable properties of statistical tools to be employed for stochastic analysis. We refer to [3] for an introduction of computational algorithms that have been proposed in this respect.

Most of the problems in bioinformatics can be characterized in the form of causal-effect identification or structure-activity relationship; in this way, we have a more general setup than conventional regression models comprising linear, nonlinear and generalized linear models, including possibly random and mixed-effects. In large biological systems, the structure is often too complex and too implied to be formulated explicitly in a simple manageable form. On top of that, latent effects and enormous variability in the reaction (activity) phase make it even harder to consider simple regression type statistical tools to analyse such data. The very characterization of the inherent stochastics and their impact on the outcome variables may be a challenging task. Yet, this task lies at the very foundation of any quantification of biological response systems. In bioinformatics, in particular, because of the prevalence of purely qualitative variables, it is often more difficult to import the classical measures of dependence, association and regression in a manageable way. Therefore, a basic task is to formulate suitable statistical measures which can withstand, to a greater extent, the irregularities arising in such large and less precisely quantified systems. Of course, there are some bioinformatic models that are more related to pure biological systems of large dimensions, and as such any conventional approach to study dependence, association or regression may become too naive to accomplish any real purpose. Incorporation of latent biological (genetic, kinetic etc.) information is therefore essential in any stochastic analysis of such systems. A key factor to the success of stochastic modeling and analysis of conventional data models is the replication of the study; an adequately large sample of presumably comparable data sets provides good estimate of the underlying errors, and thereby provides more confidence in detecting the assignable variations which go beyond the level of chance variation. As the dimensionality increases, the requirement of an adequately large sample to yield conclusions with a predictable level of errors becomes increasingly stringent; in fact, it may require an exponentially increasing sample size with the (linear) increase in the dimension to match the level of precision, and this is not likely to be the case in a majority of bioinformatics problems. The information technology and the simulation worlds have nurtured countless algorithms for exploratory data analysis. However, there is no guarantee that such exploratory data analysis (even under the fancy name of KDDM or statistical learning) would lead us to design and biologically consistent stochastics. This is what we would explore more specifically in some of the problems in the broader area of bioinformatics, and biostochastics in general.

4. Challenges in stochastic modeling

Let us illustrate some of these problems with specific models in mind. Consider first the DNA model with K (a large number of) positions or sites, labelled as $1, \ldots, K$. At each site we have a categorical response that takes on the four nucleotide labels A, C, G, T (or one of the 20 aminoacids, in the case of Protein models). Thus, we have a K-dimensional response vector

$$\mathbf{Y} = (Y_1, \ldots, Y_K)', \tag{4.1}$$

where each Y_j can assume one of the possible M outcomes (typically, $M = 4$ or 20). Therefore, $\mathbf{Y}$ can have M^K possible (purely) qualitative realizations without an ordering of the categories. Since the K positions may not be independent, it may not be wise to assume that the multinomial law of $\mathbf{Y}$ can be factorized into K simpler multinomial laws, each involving M categories. Viewed from this angle, the joint multinomial law of $\mathbf{Y}$ will have $M^K - 1$ parameters. The cardinality of the parameter space can be partitioned into (1) $K(M-1)$ for the main-effects of the K marginal laws, (2) $(M-1)^2\binom{K}{2}$ for the two-factor interaction effects, and in general, (r) $(M-1)^r\binom{K}{r}$ for the r-factor interactions, for $r = 1, \ldots, K$. For $M = 2$, this relates to the so called Bahadur [1] reparameterization for correlated dichotomous response variables, which has been studied in greater detail by Liang, Zeger and Qaqish [9], among other. The representation for $M \geq 3$ is of course more complex, and interpretations are less transparent. Suppose now we have N sequences, denoted by $\mathbf{Y}_i$, $i = 1, \ldots, N$, and, based on this set, we would like to draw statistical conclusions on the model based parameters. It is customarily assumed that these N vectors are stochastically independent, though in general, under phylogenetic setups such an assumption may not be true. Recall that generally K is a large positive integer, so that even if $M = 2$, the cardinality of the parameter space is quite large. As such, if a conventional multinomial is to be advocated, in view of the large dimension of the parameter space and not enough structures on the associated multinomial law, we may need an enormously large value of N to draw statistical conclusions with the usual manageable level of precision. As such, two basic questions arise in this context: (i) *Can the parameter space be reduced by suitable marginalization so as to be consistent with the available sample size* N, and (ii) *what type of likelihood formulation can be made to draw valid and reasonably efficient statistical conclusions?*

For the specific case of $M = 2$, relating to the events: whether or not there has been a *mutation* at site j, for $j = 1, \ldots, K$, the Bahadur [1] representation has been incorporated, and under (a stringent) assumption that 3-factor or higher-order interactions are all negligible, an *autologistic* model has been considered in the literature. This model involves $\binom{K+1}{2}$ model based parameters, and even has been justified in the light of some *pseudo-likelihood* formulation. We may refer to Sen [13] for some critical appraisal of such modeling and related stochastic analysis; other pertinent references are cited there. Even in this case, if K is above 100 (which is typical in Genomic studies) then $\binom{K+1}{2}$ is over 50,000, and as a result we need N to be at least 250,000 in order that conventional categorical data modeling and analysis can be made with confidence.

Pinheiro *et al.* [10] considered the general case of polychotomous variables, and developed a categorical analysis of variance procedure using alternative norms of divergence or dissimilarity. For a pair of independent random variables Y_i and Y_j, each having the same set of of possible realizations, labelled as $1, \ldots, M$, consider the function

$$I_{ij} = I(Y_i \neq Y_j), \qquad i \neq j = 1, \ldots, N. \tag{4.2}$$

If we denote the probabilities for these M realizations by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)'$, then for independent Y_i, Y_j, we have

$$E(I_{ij}) = P\{Y_i \neq Y_j\} = 1 - \boldsymbol{\pi}'\boldsymbol{\pi} \tag{4.3}$$
$$= 1 - \sum_{k=1}^{M} \pi_k^2 = I(\boldsymbol{\pi}) \,.$$

The right hand side of (4.3) is also known as the Gini-Simpson index of biodiversity (Simpson [16]); this has also been used in genetic diversity studies (Chakraborti and Rao [2]). We then define

$$U_N = \binom{N}{2}^{-1} \sum_{1\le i<j\le N} I_{ij} \tag{4.4}$$

and notice that it is a U-statistic (Hoeffding [6]). If now the N sequences $\{Y_i,\ i = 1, \ldots, N\}$ can be grouped into r groups of sizes $N_1, \ldots, N_r$ respectively (so that $N = N_1 + \cdots + N_r$), and if we denote the within group U-statistics by $U_{N_s}^{(s)}$, $s = 1, \ldots, r$, then it can be shown (Sen [11]) that U_N can be decomposed into two additive components, namely, the within-group component which is a linear combination of the $U_{N_s}^{(s)}$, and a nonnegative between- group component that vanishes only under perfect homogeneity of the r groups. We denote these two components as U_{NW} and U_{NB} respectively. Note that by definition the $U_{N_s}^{(s)}$ (and U_N) are all nonnegative random variables, so that on writing

$$U_N = U_{NW} + U_{NB} \,, \tag{4.5}$$

it is quite tempting to test for the homogeneity of the r groups based on the test statistic

$$\mathcal{L}_N = U_{NB}/U_{NW} \,, \tag{4.6}$$

rejecting the null hypothesis for large values of $\mathcal{L}_N$. For this testing problem, we need to find an appropriate critical level of $\mathcal{L}_N$. A procedure for finding the asymptotic null distribution of $\mathcal{L}_N$ has been studied thoroughly in Pinheiro *et al.* [10]; unfortunately this asymptotic distribution is neither normal nor a chi square. Their result provides a means of finding out the asymptotic critical levels of the test statistic $\mathcal{L}_N$. Some alternative tests based on suitable measures of entropy (for $\boldsymbol{\pi}$) have also been discussed there.

The above procedure relates to testing for homogeneity of r groups of sequences relating to a single position. In actual genomic studies, we have sequences comprising of K positions, so that we have a set of K-vectors that we denote by $\mathbf{Y}_i$, $i = 1, \ldots, N$. As a natural generalization of the Gini–Simpson index, *Hamming distance* has been incorporated in this setup (Pinheiro *et al.* [10]). For a pair (i, j) of sequences, we define the Hamming distance as

$$d_{ij} = K^{-1} \sum_{k=1}^{K} I(Y_{ik} \neq Y_{jk}), \tag{4.7}$$

so that we can define the corresponding sample measure as

$$U_N = \binom{N}{2}^{-1} \sum_{1 \le i < j \le N} d_{ij}. \tag{4.8}$$

It is easy to see that if for the kth position we define the sample counterpart of the Gini-Simpson index by $U_N^{(k)}$, for $k = 1, \dots, K$, then

$$U_N = K^{-1} \sum_{k=1}^{K} U_N^{(k)}, \tag{4.9}$$

and as a result, the population measure of Hadamard distance is the average of the K Gini-Simpson indexes for the K marginal probability law. This formulation allows for possible stochastic association of the K positions, and the decomposition for each coordinate percolates through their average as well. Thus, virtually, a similar test as in the case of a single position can be formulated. However, bearing in mind the stochastic interdependence of the K sample Gini-Simpson indexes, we may argue that the distribution theory (even under the null hypothesis) could be much more complicated in this case.

We may appraise this testing problem from a somewhat broader perspective. At first sight, it might be tempting to consider the K positionwise Gini–Simpson indexes, denoted by $\mathbf{U}_N = \big(U_N^{(1)}, \dots, U_N^{(K)}\big)$, and consider a linear combination of this vector that would have the maximum discrimination capacity. Since the covariance matrix of these K statistics is not known, a sample estimate of that is needed to formulate such a discriminant function. Secondly, the joint distribution may not be multinormal, even asymptotically (as commented after (4.6)). That would make it difficult to use the likelihood function or its asymptotics to formulate a linear discrimination function (as in the conventional multinormal model); this difficulty is also due to the fact that the aforesaid covariance matrix may depend, in general, on the unknown parameters which are under test as well (while in the multinormal case, the covariance matrix is free from the location parameters). Further, as the basic response variables are purely qualitative (categorical), instead of a linear model approach, one could think of using a *generalized linear model* (GLM). But, a GLM rests on the choice of appropriate *link functions*, and in the multivariate case, in general, such link functions are hard to construct. One may even think of *generalized estimating* *equations* (GEE) based on suitable *quasi-likelihood* formulations. However, as the statistics are nonlinear functions of the multinomial cell frequencies, even a quasi-likelihood approach may encounter difficulties. Finally, as generally K is large while the positions may not be stochastically independent, in order to apply conventional large sample approximation to the distribution of suitable test statistics, we may generally need N to be very large compared to K. On top of that, the very assumption that the N sequences are independent and identically distributed may not be tenable in genomic studies where phylogenetic factors may affect the independence, and even stochastic exchangeability may not be true. Faced with these limitations, it seems that formulating an optimal testing procedure is indeed a challenging task.

Let us consider another related problem treated in Karnoub *et al.* [7]. Consider the same setup as above. There are K positions where at each position a *mutation* may or may not take place. Thus, if there are N sequences, we could attempt to test for the hypothesis that these positions undergo mutation independently of each other. The basic difficulty is to ascertain whether or not a mutation has taken place in a specific site. For the case of $K = 2$, Karnoub *et al.* [7] attacked this problem by defining a *consensus pair* (i.e., a combination of the chemical letters, one in each position from the set (A, C, G, T)) to be that which has the maximum frequency in the sample. Given this consensus pair of chemical words, we can then define a 2×2 table of no mutation (at either site), one mutation (either at the first or the second site) and double mutation (at both sites); the resulting sample frequencies are denoted by $N_{11}, N_{12}, N_{21}, N_{22}$ respectively. It is therefore tempting to treat this as the usual 2×2 contingency table, and apply the conventional tests. However, there are certain limitations of such a conventional approach. First, note that by definition, N_{11} is the maximum among the 4 entries N_{ij}, $i, j = 1, 2$, and hence, we need to condition on N_{11} and N, instead of the marginals $N_{1.} = N_{11} + N_{12}$ and $N_{.1} = N_{11} + N_{21}$. This altered conditioning feature also alters the probability structure of the observed 2×2 table, and hence, Fisher's exact (conditional) test or the large sample chi square test needs to be examined critically. Karnoub *et al.* [7] considered a conditional binomial test (conditioned on N_{11} and N), and studied its asymptotic properties. Even for the two position case, if we have a general $r \times c$ table, such a test becomes quite complicated. A more pertinent point is to extend such conditional tests to the general case of K positions. If we intend to find a consensus K-tuple of categories, denoted by $(A_1, \ldots, A_K)$, the first question may arise: Whether or not such a consensus cell exists with probability close to one? Secondly, rearranging the K-fold table in such a way that the left-hand topmost vertex is the maximum frequency cell, may not lead to a simple resolution for conditioning on a subset of other cells, and base the test on the remaining cell frequencies. With such a large number of cells, unless the expected frequency in each cell is at least moderately large, a large sample test may not work out well. Tests based on marginal and joint entropy functions are more complicated to handle for such high-dimensional tables. We may think of some other alternative ways of test construction to avoid some of these problems.

Consider the set of all possible $\binom{K}{2}$ pairs of sites, and for each pair, define a consensus pair of categories, and given that and N, construct a test statistic. Let these test statistics be denoted by $\mathcal{L}_{Nkq}$, $1 \le k < q \le K$. Note that $\mathcal{L}_{Nkq}$ is designed to test for the hypothesis of independence of mutation in the sites k and q. Thus, if we frame the null hypothesis (H_0) as the hypothesis of independence of mutation for all possible pairs of sites, we may base an overall test on the set $\mathcal{L}_{Nkq}$, $1 \le k < q \le K$. One possibility is to choose an overall test statistic as

$$\mathcal{L}_N^* = \sum_{1 \le k < q \le K} \mathcal{L}_{Nkq}\,; \tag{4.10}$$

another possibility is to choose

$$\mathcal{L}_N^o = \max\bigl\{|\mathcal{L}_{Nkq}|\colon 1 \le k < q \le K\bigr\}\,. \tag{4.11}$$

Recall that the $\mathcal{L}_{Nkq}$ may not be stochastically independent for different pairs, and the consensus pairing conditioning makes their dependence structure even more complex. As such, the (exact as well as asymptotic) distribution theory of such overall test statistics (even under the null hypothesis) may not be of simple form. Moreover, the rationality of such overall test statistics needs to be appraised with a view to having some asymptotic optimality properties. In general, we need to probe into high-dimensional purely qualitative categorical data models with constraints framed by some (random) sample functions, and subject to such constraints, we need to develop statistical inference methodology. In this generality, it is indeed a challenging task.

Let us consider a standard polygenic model of biometrical genetics, motivated by considering a quantitative trait determined by a large number of loci, assumed to be acting *independently* and *additively*. Let $X_{i,k}$ be the contribution of locus k to person i whose trait value is taken as

$$X_i = \sum_{k=1}^{N} X_{i,k}\,, \qquad i = 1, \ldots, m\,. \tag{4.12}$$

If the $X_{i,k}$ are all independent, then we have

$$\mathrm{Cov}(X_i, X_j) = \sum_{k=1}^{N} \mathrm{Cov}(X_{i,k}, X_{j,k})\,, \qquad i \ne j = 1, \ldots, m\,. \tag{4.13}$$

In the last equation, kinship and other factors can be called upon to structure the covariance pattern more vividly. It is usually assumed that the $X_{i,k}$ are all normally distributed (Lange [8]), and thereby mixed-effects models with variance component analysis are adopted. From a statistical perspective it might be desirable to probe into robustness aspects due to possible departure from the assumed normality, additivity and independence. Actually, this needs to be done in the general context of multivariate data models with structures on the covariance pattern but without possibly imposing independence, additivity and normality assumptions. There are other genetics models in bioinformatics which also call for critical appraisal.

5. Concluding remarks

It is not surprising that in its infancy, bioinformatics is thriving with quantification of its mostly qualitative data sets, and it is quite natural to start with simple statistical tools for data modeling, collection and drawing conclusions. However, in doing so, no matter how many statistical packages and algorithms are at our disposal, we should pay due attention to the validity, robustness and efficiency aspects of such conventional statistical tools. The field is indeed quite vast, complex and fully loaded with nontrivial problems that deserve more statistical scrutiny and mentoring as well. In doing that one must look into the molecular, biological and genetic backgrounds of specific problems, and in the light of these, develop appropriate statistical tools. This task, though it appears to be rather challenging, is within the reach of modern statistical science, and with the cooperation of information technology, statistics should be the methodology custodian of bioinformatics.

Acknowledgements. Part of this research was supported by the C. C. Boshamer Endowment Professorships Foundation at the University of North Carolina at Chapel Hill. The author is grateful to the reviewer for his very critical reading of the manuscript that eliminated a large number of typos and improved its presentation, too.

REFERENCES

[1] BAHADUR, R. R.: *A representation of the joint distribution of responses to n dichotomous items*, in: Stud. Item Anal. Prediction (H. Solomon, ed.), Stanford Univ. Press, California, 1961, pp. 158–176.

[2] CHAKRABORTI, R.—RAO, C. R.: *Measurement of genetic variation for evolutionary studies*, in: Handbook of Statistics, Vol. 8: Statistical Methods in Biological and Medical Sciences (C. R. Rao; R. Chakraborti, ed.), Elsevier, Amsterdam, 1991, pp. 271–316.

[3] DURBIN, R.—EDDY, S.—KROGH, A.—MITCHISON, G.: *Biological Sequence Analysis: Probabilistic Models for Proteins and Nucleic Acids*, Cambridge Univ. Press, England, 1998.

[4] EWENS, W. J.—GRANT, G. R.: *Statistical Methods in Bioinformatics: An Introduction*, Springer-Verlag, New York, 2001.

[5] HASTIE, T.—TIBSHIRANI, R.—FRIEDMAN, J.: *The Elements of Statistical Learning*, Springer-Verlag, New York, 2001.

[6] HOEFFDING, W.: *On a class of statistics with asymptotically normal distribution*, Ann. Math. Statist. **19** (1948), 293–325.

[7] KARNOUB, M. C.—SEILLIER-MOISEIWITSCH, F.—SEN, P. K.: *A conditional approach to the detection of correlated mutations*, in: IMS Monogr. Ser., Vol. 33, 1999, pp. 221–235.

[8] LANGE, K.: *Mathematical and Statistical Methods for Genetic Analysis*, Springer-Verlag, New York, 1997.

[9] LIANG, K.—ZEGER, S. L.—QAQISH, B. H.: *Multivariate regression analysis for categorical data (with discussion)*, J. Roy. Statist. Soc., Series B, **54** (1992), pp 3–40.

[10] PINHEIRO, H.—SEILLIER-MOISEIWITSCH, F.—SEN, P. K.—ERON, J.: *Genomic sequence analysis and quasi-multivariate* CATANOVA, in: Handbook of Statistics, Vol. 18: Bioenvironmental and Public Health Statistics (P. K. Sen; C. R. Rao, eds.), Elsevier, Amsterdam, 2000, pp. 713–746.

[11] SEN, P. K.: *Utility-oriented Simpson-type indexes and inequality measures*, Calcutta Statist. Assoc. Bull. **49** (1999), 1–21.

[12] SEN, P. K.: *Excursions in Biostochastics: Biometry to Biostatistics to Bioinformatics*, Academia Sinica (Taipei) Institute of Statistical Science, Lecture Notes, 2001.

[13] SEN, P. K.: *Computational sequence analysis: Genome and Statistical controversies*, in: Some recent developments in Statistics and probability theory (Y. P. Chaubey, ed.), World Scientific Press, 2002, pp. 274–289.

[14] SEN, P. K.: *Structure-activity relationship information incorporation in environmental risk assessment*, Environmetrics **14** (2003), 223–234.

[15] SEN, P. K.: *Neuronal spatio-temporal models: High-dimensional implications and statistical perspectives*, Scientiae Mathematicae Japonicae **56** (2002), 613–648.

[16] SIMPSON, E. H.: *The measurement of diversity*, Nature **168** (1949), 688.

[17] WATERMAN, M.: *Introduction to Computational Biology. Maps, Sequences and Genomes*, Chapman-Hall/CRC, Boca Raton, 2000.

Received April 2, 2002

Department of Biostatics and Statistics
University of North Carolina at Chapel Hill
NC 27599-7420
U. S. A.

E-mail: pksen@bios.unc.edu