JIRSS (2003)
Vol. 2, No. 1, pp 1 - 19

Bioinformatics to Biostochastics: Statistical Perspectives and Tasks Ahead

Pranab Kumar Sen

Departments of Biostatistics and Statistics, University of North Carolina, Chapel Hill, NC 27599-7420, USA. (pksen@bios.unc.edu)

Abstract. Bioinformatics is an emerging field of science emphasizing the application of mathematics, statistics, and informatics to study and analysis of very large molecular biological (mostly, genetic and genomic) systems (data sets). In a comparatively broader setup of large biological systems without necessarily having a predominant genetic undercurrent, and having genesis in biometry to biostatistics, biostochastics has evolved as the primary vehicle for the much needed statistical reasoning. It is intended to point out the genuine need for statistical reasoning in this evolving interdisciplinary field, and in that way, to appraise the limitations of current (mostly, algorithm based) statistical resolutions.

Received: January 2003

Key words and phrases: Arsenites, biostatistics, dimension reduction, genomics, image analysis, informatics, KDDM, molecular biology, NSTM, pharmacogenomics, statistical inference, statistical learning, statistical modeling, toxicology.

1 Introduction

Computational biology, genomics, polygenic models, and in general, molecular biology (genetics) have nurtured the ongoing evolution of bioinformatics, though it is still not precisely known what really constitute the core of bioinformatics. The pharmaceutical researchers, primarily interested in drug development as well as drug-marketability, have their own way of looking into bioinformatics: better understanding of the therapeutic, toxic and genetic undercurrents in drug developments for properly unfolding the (biological as well as genetic) intricacies of various diseases and providing better health care. Molecular biologists and computer scientists, on the other hand, aim to chalk out completely the intricacies of genomics through computational sequence analyis (and computational biology in general) while the biomathematicians and statisticians at large desparately look for mathematical laws that might underlie such biological mysteries; there are genuine reasons to doubt about the existence of such precise laws, even in a broad statistical perspective. At best, statistical appraisal might separate the chaos from drifts in a meaningful way. In this vein, Ewens and Grant (2001) have provocated a nice introduction to bioinformatics that can be taken as a working rule: *We take bioinformatics to mean the emerging field of science growing from the application of mathematics, statistics, and information technology, including 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.* Waterman (1995) has a similar interpretation with more emphasis on the computational biological perspectives. Lange (1997) emphasized on the mathematical and statistical aspects of genetic analysis with emphasis on bioinformatics. Durbin et al. (1998) attempted to develop some probabilistic models for proteins and nucleic acids incorporating the so termed *hidden Markov models* (HMM), although the prime emphasis being on the computational algorithms. The developments in the past four/five years have completely reshaped this interdisciplinary field, and thereby call for a deeper appraisal of both methodologic and computational perspectives. The very basic applicability of probabilistic reasoning in bioinformatics is itself worthy of serious appraisal. Largely flanked by the spectacular advent of information and biotechnology, Genome projects involving

human as well as other organisms have been undertaken all over the World, and two authentic scientific reoprts in 2001 (in *Science*, and *Nature*), both cited here in the bibliography, relate to the current state of art with these nearly completed omnibus projects. Their generated data sets of exceedingly large dimensions have posed many challenging problems in an interdisciplinary channel of research. It has become clear that molecular biology might have stolen the limelight of the scientific developments, and yet there is ample room for stochastics to comprehend the basic difference between mathematical exactness and biological diversity / disparity (inexactness). At the current stage, gene scientists can not scramble fast enough to keep up with the genomics emerging at a furious rate and in astounding detail. It's a basic dilemma to look for any precise mathematical structure or laws that underlie such mysteries. It's equally dubious to let computers alone dictate the basic (molecular or biologic) laws unless they be substantiated by sound mathamtical and statistical modeling and analysis. At least at this stage, bioinformatics, as a discipline, does not aim to lay down fundamental mathematical laws that govern biological systems parallel to those laid down for physics or engineering sciences. Bioinformatics is not encompassed by biomathematics. There is, however, at the present, mathematical utility in the creation of tools (for example, computational schemes and algorithms) that investigators can use to analyse enormously large data sets that typically arise in bioinformatics (and more generally large biological systems) studies. Biological diversity, and underlying stochastic evolutionary forces make it equally appealing to comprehend appropriate stochastic models and incorporate probabilistic as well as statistical tools to undertake such tasks in an objective manner. However, such probabilistic and statistical tools need to be sharpened in the light of the underlying biological intricacies, and much of the difficulties lies there. In this setup, *knowledge discovery* and *data mining* (KDDM) or *statistical learning* tools are providing valuable computational algorithms, and there is a vast scope for methodlogic advances to connect them to statistical reasoning in a theoretical mold. We refer to Hastie et al. (2001) for a nice account of statistical learning and all the algorithms developed so far (including neural networks); even so, there remains the basic question of statistical methodologic support for KDDM. With the rampage of biotechnology and information technology, we are in a realm of interdisciplinary research where large

biological systems have been focussed in a way that used to be impassing even a few years back. *Neuronal spike-train models, brain mapping, bioenvironmental models, polygenetic models, pharmacokinetics* and *pharmacodynamics, physiologically based pharmacokinetic (PBPK) models, toxicodynamics* and *toxicokinetic (TDTK) models*, are just a few important areas under active research interest, and in each of these areas, statistical reasoning, albeit in an interdisciplinary mold, is essential. Even *cybernetics* is not outside the realm of biostochastics, although the emphasis on astrophysics and its implications are to be appraised for biological systems (Sen, 2002a). Should we therefore confine ourselves only to the realm of bioinformatics? Motivated by this diversity, we comprehend biostochastics in a broader framework (of interpretation) than bioinformatics. It includes (large as well as small) biological models that are not not necessarily genomic ones or the ones with persistent genetic undercurrents. As is the case with many macro-biological models, there are some challenging high- dimensional statistical problems, and some of these are similar to most of the ones arising in bioinformatics as well. In order to comprehend this evolutionary field, it is better to trace the origin, tracing the genesis to biometry to biostatistics to biostochastics, with special emphasis on bioinformatics. For this reason, it is desirable for us to trace the developments in three phases : (i) Biometry to biostatistics, (ii) biostatistics to bioinformatics, and (iii) bioinformatics to biostochastics. Sections 2, 3 and 4 are to deal with these three aspects thoroughly. The concluding section is devoted to a review of some of the outstanding statistical problems that have been encountered in this evolving interdisciplinary field of research. In this setup, limitations of standard statistical methodology as well as the need for novel tools are appraised.

2 Biometry to Biostatistics

Biometry, and agricultural and industrial sciences formed the basis of applied statistics, and the basic idea was to develop suitable statistical methodology that could be used as handy tools in the respective fields of application. In this respect, some of the early statistical works started with demography, anthropometry, genetics, agronomy, and biological studies. Although, in this way, biological

and agricultural sciences can claim a greater part of the genesis of modern statistical science, in view of the motivational differences and applicational emphasis, there has been a persistent undercurrent in consolidating the mathematical abstraction as well as sophistication solely under the jurisdiction of *mathematical statistics* while relegating the *data analysis* part to *applied statistics*. This trend continues even today. The evolution of biometry goes back to more than 100 years ago. The terminiology was coined in Great Britain with the appearance of the first issue of *Biometrika* in 1901. In the March 2001 issue of Biometrika, the highlights of the evolution of biomretry, as captured in this journal, are reported. Biometry not only sparked the need for looking into the intricacies of biological factors in statistical formulations, but also, nurtured much of the developments in mathematical statistics at a time where there was practically no foundation of statistics as a discipline. The salient points of diffenence between agricultural and biological fields of research have led to a divergence of biometry and agricultural statistics, and to a greater extent this has led to the evolution of biostatistics with an annexation of clinical (biomedical), and environmental sciences along with a broad field of public health disciplines. In this way, biostatistics had to cater for a greater need for statistical reasoning in a wider interdisciplinary field where controlled experimental setups are usually not tenable. Naturally, the need for data analysis arose in a more perceptible manner, and in course of this crusade, mathematical sophistications (that are characteristically associated with theoretical statistics and probability theory) were diffused to a greater extent to facilitate more effective modeling and analysis of acquired fdata sets (often from observational studies, instead of experimental ones). Admittedly, in either way, there is a persistent emphasis on analysis of experimental or observational data sets cropping out of biological/ biomedical studies, and no wonder, countless statistical packages have been developed to meet this goal. Yet there remains much to bridge the gap between sound statistical methodology and superb statistical analysis in biosciences. Usually scientific experiments in physical sciences are conducted in a controlled laboratory setup with precise instruments to record the outcomes, and the conclusions can then be advocated for more diverse setups. Generally, suitable mathematical laws underlie such studies, and conducted studies focus on their formulations. In many biological studies, such a mathematical law,

even if it exists, can be obscured due to more paramount diversity existing in the experimental units, and hence, there is a greater need of statistical considerations in unearthing such general laws as well as reasons for possible divergence from the expected laws. A similar phenomenon is observed in quantum physics and other large physical studies admitting higher degrees of data obscurity. This led to the formulation of statistical reasoning on a tripple-standing: (a) Planning, (b) Modeling, and (c) Statistical Analysis. This combination enables one to draw conclusions from conducted studies in an objective manner. R. A. Fisher's pioneering work on design of experiments laid down the fundamentals of statistical planning - a prerequisite for any objective statistical study. The three basic tools, namely, *randomisation, replication*, and *local control*, provide the necessary access to incorporate statistical methodology that has been well developed during the past seven decades. However, these concepts have mostly been developed for agricultural experiments dealing with blocks and plots, manures and fertilizers (treatment) etc. Although such terms have been adapted for biometric studies, there are some basic differences that may call for alternative formulations. In simple biometric studies, such adaptations have been made without much difficulties. However, in more complex biostatistics problems, without a controlled study plan, most of these tools need considerable modifications. For example, randomisation, that provides the very basic means for importing objective sampling schemes on which general statistical methodology rests, may not be entirely asdoptable in observational studies or in clinical trials, and at best, some restricted randomisation principles could be incorporated. In case-control studies, the design aspect may be quite different, and a different kind of randomisation principle may be needed for statistical modeling and analysis purposes. Replication in agricultural experiments is comparatively simpler; it needs additional plots or blocks to ensure that the error variances could be estimated with enough precision, adding validity and reliability of statistical conclusions. Local control, achieved either through blocking or confounding, attempts to eliminate further assignable causes of variation and thereby enhance the efficacy of the experiments. In simple biometric studies replication and local control can be made in a manner similar to that in agricultural experiments by adding additional subjects in the study and using suitable blocking by choosing the subjects in clusters of certain types. However, in

relatively more complex biostatistical problems, typically, there are a number of response variables not all of which may be continuous, and this may require, in turn, more skillful choice of subjects. Moreover, often the recruitment of an adequate number of subjects to meet the need of local control and replication may not be possible. These factors, in turn, make (statistical) modeling more complex in genuine biostatistics problems than in simple biometric ones. We shall discuss these in the next section.

3 Biostatistics to Bioinformatics

In a variety of disciplines, ranging from public opinion survey to socio-economic studies, to experimental sciences, to biomedical and clinical studies, as well as, to environmental and public health studies, in order to draw scientific conclusions in a quantitative norm, data are to be collected, analysed and incorporated in the decision making process. Because of inherent variability and uncertainty of the outcome variables in some cases, such data based decisions may not always be correct. Therefore, it is natural to plan data collection in such a way that, subject to the available resources, experimental or observational conclusions can be made in an objective (scientific) way with maximum possible precision or confidence. This objective is reachable if the bias due to possible experimental factors, measurement protocols, or inestigators can be eliminated, and an assessment of the margin of errors associated with the conclusions can be made. Ideally, one would like to reduce the bias and margin of errors as far as statistically possible. That, in turn, requires the tools randomisation, replication and local control, explained in the preceding section. Or, in other words, statistical planning is an essential task for any biometric or biostatistics study. However, whereas in mathemetical statistics, *optimal designs* have been developed on sound theoretical basis, in biostatistics, such optimal disigns may not exist mainly due to less precise setups that permit randomisation, replication and local controls only to a much limited extent. The use of human subjects in biostatistics is becoming more common with the adoption of safer conducts of experiments or observational studies involving them. *Clinical trials* have been gaining popularity in drug research and public health studies. Clinical trial relate to re-

sponse variables ranging from survival times to drug-effects of various types. Medical ethics prevent using human subjects on a discriminatory basis, so that no subject is to alloted to a treatment which is either known to be less effective or to have serious side-effects. This results in a different class of statistical designs where too many subjects might not be recruited on cost and time considerations, and other restraints may apply in preserving randomisation to a reasonable extent. We may refer to Sen (1999, 2001c) for some of these technicalities in clinical trials with reference to statistical modeling and analysis. Bioethics and public advocades have voiced concern about clinical trial exploitation in Third World countries where the cost-benefit factor may be the primary issue; the goal is to identify *effective* as well as *affordable* regimens to suit the need of developing countries. In this respect, the 1997 *Helsinki declaration* of the World Medical Association , namely, in any medical study, every patient - including those of control group - if any, should be assured of the best proven diagnostic and therapeutic method, has raised a basic queries: How far medical ethics can be implemented in clinical trials with such diverse perspectives in mind? To what extent cost-benefit aspects can overturn the basic medical prerequisites of a clinical trial (especially, in developing countries)? It is therefore more pertinent to assess how much statistical reasoning can be imparted in such a broader set-up? There is another important issue (Temple and Ellenberg, 2000): Placebo- controlled trials (PCT) are extensively used in developing new pharmaceuticals. There are allegations that PCT are invariably unethical when known effective therapy is available for the condition being treated or studied, regardless of the condition or the consequences of deferring treatments. Based on the Helsinki declaration, patients asked to participate in a PCT must be informed of the existence of any effective therapy, must be able to explore the consequences of deferring such therapy with the investigator, and must provide fully informed consent. This would provide justification for the PCT even when effective therapy exists. Another by-product of this declaration is the formulation of active- control equivalence trials (ACET) which may show that a new therapy is superior (or not inferior) to an existing one - but may not have all the other characteristics of a PCT. This, of course, calls for more innovative statistical reasoning tools for ACET. The development of the field of clinical epidemiology and controlled clinical trials has a significant biostatis-

tics base, and as we shall see later on that pharmaceutical research in this way has gained solid footing with effective intervention from biostatistics. Case control studies are quite common in epidemiologic investigations, and there too, the design aspects may be quite different from conventional biometric or agricultural experiments. As biostatistics is moving more in large scale epidemiologic and clinical studies, more challenging problems are cropping up with statistical planning of such studies, as well as, related modeling and analysis aspects. In data acquisition, incompleteness due to missing patterns and / or censoring of various kinds is commonly encountered. In a *random censoring* scheme, it is tacitly assumed that censoring occurrs randomly, independently of the primary response variable(s); hence, it is also often called *noninformative censoring*. On the other hand, in actual practice, usually the clause of random censoring may not be tenable, and as a result, some of the conclusions based on a noninformative censoring assumption may not be tenable. *competing risk* models are also very common in many clinical studies. It is not uncommon to assume that such competing risk factors work independently, whereas in reality, this could be quite different. Equal probability sampling, as is classically taken for granted in mathematical statistics, sometimes appears to be unreasonable in clinical and biomedical studies. Moreover, the independence assumption of a set of observations may not be taken for granted in many biostatistics models and analysis schemes. More and more scientific attention is being fixed on large biological systems (with or without having persistent genetic undercurrents). Although there are remarkable similarities in such biological syatems in terms of molecular or cellular structures, there is also an enormous variation in the associated response variables that are to be statistically modeled and analysed. The situation has turned out to be more complex due to various influencing environmental factors and toxicologic interventions. Take, for example, the case of *computational sequence analysis* (CSA) relating to genomic sequences. Most problems in CSA are essentially statistical. Stochastic evolutionary forces act on genomes, Typically, there is a large number (K) of positions or sites, and in each position, there is a purely qualitative (nucleotides or amino acids) categorical response with 4 to 20 categories depending on the DNA or Protein sequence. The spatial (functional as well as stochastic) dependence (or covariation) patterns of these sites may not be generally known,

nor they can be taken to be stochastically independent. On the other hand, the DNA have fairly regular and nearly identical structures, so that statistical appraisals of such genomic sequence are to be based on other variational properties which exhibit more statistical variation and possess more statistical information. If we take literally the CSA as a multivariate statistical problem, then there are numerous roadblocks. First, it is a very high-dimensional qualitative response model, so that the conventional multinormal models are not that appropriate in this setup. Even we take recourse to discrete multivariate analysis (Bishop et al., 1975) there are difficulties in modeling in a parametric framework (as the structure is not that precisely known), and more due to the *curse of dimensionality*. Typically, the number of parameters arising in conventional qualitative categorical data models could become so large in CSA that standrad statistical modeling and inference tools are not of much appeal. For this reason, alternative procedures that attach less emphasis on the likelihood approach and more on alternative approaches that takes into account the undelying (molecular) biological information to a greater exctent are to be worked out. In this respect, if we consider the HIV/AIDS case against not HIV positive case, we may apparently treat this as a classical categorical analysis of covariance model in a two-sample set-up. Howe3ver the curse of dimensionality may render a statistical test based on this modeling may be practically powerless. On the other hand, we may note that like many other retroviruses, HIV has the ability to reverse the flow of genetic information in the DNA sequence. As a result, genetic mutations are likely to be more frequent in HIV positive people than those who do not have HIV positive status. Thus, statistical modeling and analysis of HIV genomic sequence center around the genetic variability. Parametric methods are not of much use in this respect, and nonparametrics fares better. We refer to Pinheiro et al. (2000, 2003) for some work in this direction. There is a genuine need to have an easy access from biostatistics to bioinformatics wherein appropriate mehodologic bases can be developed for drawing scientific conclusions in bioinformatics. At the present, KDDM or statistical learning dominates the scenario. Though that has generated a wealth of statistical packages and algorithms, there is a pressing need for methodologic support with full statistical considerations.

4 Bioinformatics to Biostochastics

Whereas bioinformatics pays due importance to underlying molecular biologic (and genetic) structures, there are many other large biological systems where the genetic undercurrents might not be apparent or dominant in that way. Biostochastics is primarily geared to provide the desired statistical reasoning for modeling and analysing large biologicaql systems which exhibit considerable stochastic or unassignable variation. To illustrate this point, we consider the following example of a large biological system which may not have that much genetic undercurrent. This is the so called *neuronal spatio-temporal models* (NSTM) relating to large neurobiological systems (viz., the cortex (brain) in human as well as subhuman primates (Sen, 2002a)). In neuronal firing phenomenon, firing times in nerve cells (neurons) interact and form networks. Therefore, it might be tempting to incorporate conventional *point processes* to model and analyse such NSTM's. The primary difficulty in a point process approach stems from the enormously large number of nerve cells in the cortex, packed densely, with diverse activities, their inhomogeneity and spatial dependence; also the very experimental process of extracting the response (spike trains) from these nerve cells may generally be destructive. Thus, there is a need to address the following statistical perspectives: (1) Reduction of a functional (or at best, huge dimensional) data space (viz., the cortex manifold) into a manageable finite dimensional space (viz., a handful number of neurons covering the cortext), (2) data collection and monitoring(i.e., choosing the most relevant statistical information from apparently a chaos of infinite dimension), and (3) incorporation of possibly nonstandard statistical tools for modeling and analysis of the entire complex (the cortex) based on the acquired reduced dimensional data sets. The first aspect is known as the *dimension reduction* (DR) methodology, the second one comes more under modern information technology, and *knowledge discovery and data mining* (KDDM) tools are currently being used in this context. The last aspect comes under *image analysis* (IA), all these being related to some neuronal simultaneous spike train models. The *central nervous system* (CNS) occupies a focal point in this respect. Study of intra- and inter-sector stochastic dependence and association (of tens of thousands of neurons) in the cortex is indeed a challenging statistical task. Such NSTM's are typically different from the *neu-*

ral network models, advocated for *cognitive sciences (artificial or machine intelligence)*, especially due to the neurophysiological undercurrents. Conventional statistical DR approaches are generally not appropriate for NSTM's (as the response variables are functional point processes). Basic theoretical (neurobiological) and experimental considerations call for a somewhat different DR formulation for statistical modeling and analysis of NSTM. We refer to Sen (2002a) for a broad review of these statistical methods. The salient point of distinction of the CSA and NSTM is the emphasis on the underlying genetic factors, although both are related to large biological systems. Consider a third illustration from environmental toxicity. Toxicity abounds in nature, environment, and in our modern life-style. Toxicology relates to the study of the intake process of such toxins by human being, their mode of propagation, biological reactions, molecular level of penetration, xenobiotics and aftermaths. Because of the latent nature of a large class of toxic substances, the extreme variability of human metabolism as well as their exposure to toxic material, yet unknown nature of many carcinogenic activities, and immense difficulties in the assessment of effective toxicity levels (especially in the environment), there is a need to have statistical appraisal at each phase. Lack of experimental control, difficulties with standard dose-response analysis, as well as, limitations of usual dosimetric studies create impasses. Thus, we have a challenging statistical task that relates to a large bioenvironmental setup, and thereby, it is more in line with biostochastics that is advocated for development of statistical methodology and modeling for such systems. In this setup, we have both toxicologic and xenobiotic factors, and we need a somewhat different appraisal of the much needed statistical task (Sen, 2003). Basically, chemical or viral structure and in vivo biological activity relationship information needs to be incorporated adequately to depict the causal cum stochastic relationship between environmental exposures (of toxins and virus) and specific health hazards; incorporation of this SARI (structure-activity relationship information) puts dosimetry in a more comprehensive stand, albeit at the cost of more complex statistical modeling and analysis. As a very notable case, let me discuss the rampage of *arsenites* in many parts of the world, even the most developed countries are not immune to the arsenic toxicity in some form or otherwise. Arsenic contamination of groundwater is due to a chemical process wherein arsenous acids (arsenic trioxide)

from buried (mostly ferric) arsenates are produced. Besides arsenic minerals, organic arsenic occurs, albeit to a lesser extent, in plants, fish, crab, human body, and other organisms. With a high moisture level and substandard hygienic or sanitation practice may magnify the problem with microbeal contaminstion of human feces etc. the use of treated water for drinking purposes may lessen the impact (from ingestion toxicology point of view), but use of untreated contaminated water for household work can trigger skin cancer or other absorption toxicologic outcomes. It is in this sense, the agricultural use of land with arsenic contamination of groundwater can lead to both ingestion and absorption toxics. Not only the picture may vary considerably from one area or region (or country) to another but also considerably from one socio-economic stratum to another. The poor people without having much access to treated water in rural areas are thereby expected to have a greater share of the bladder and liver cancer (due to ingestion of contaminated water) as well as skin cancer (due to external use of such water). As such, if a scientific study of the arsenic problem has to be made in an objective manner, the following aspects need to be addressed properly:

(a) Identification of the arsenites and allied contaminants.
(b) Statistical modeling of their prevalence levels.
(c) Determination of the *bioconcentration factor*.
(d) Toxicokinetics : intake and reaction process.
(e) (Molecular) biological reaction and in vivo activity process.
(f) Incorporation of SARI in the response pattern.
(g) Demographics for the concerned population.
(h) Formulation of dose-response regression with SARI.

It is quite clear that at each step there is a genuine need for statistical appraisal, and this can only be done with adequate conformity to the bioenvironmental factors. While (a), (b) and (c) are more in the quarters of environmetrics, biostochastics is equally relevant in this assessment task. (d), (e) and (f) are more dependent on biological undercurrents and can only be pursued with good understanding of the inherent molecular biologic and genetic background. (g) and (h) are more in line with statistical modeling and analysis; however, in order to do it effectively, it is essential to pay adequate attention to all the preceding steps. Bypassing (d), (e) and (f), and linking directly (g) and (h) to (a), (b) and (c), as is often done in an environmental epidemiologic study, could be rather misleading. In USA,

primarily through the efforts of National Institutes of Health and its sister organizations (e.g., National Cancer Institute, National Institutes of Environmental Health Sciences) considerable attention has been paid to *dosimetry* or animal study where subhuman primates are mostly used in a relatively more controlled laboratory setup to explore the SARI in a quantitative form. This information is then tried to be incorporated for human exposure and reaction to similar toxics. *Meta analysis* has evolved as the principal statistical tool in transmitting the statistical evidence acquired from animal studies to human being. Yet, there are serious roadblocks for routine adoption of meta analysis tools for this extrapolation task. A subhuman primate and a human being may differ drastically in their metabolism, biological activities, exposure to toxics, and in many other factors. Thus, in order to validate meta analysis, it is essential to appraise all these factors thoroughly before using statistical packages to routinely transmit the statistical results across the species. Indeed, this is a genuine biostochastic task. It would not be out of the way to appraise the biostochastics undercurrents in bioenvironmental disasters and phenomena. There could not be a better illustration of this deleterious impact than the Persian Gulf War in 1991 and its revival 12 years later. Not only the burning oil fields are impounding atmospheric toxicants all over the surrounding countries but also the shower of bombs and other arsenal aside killing people is bringing in disastrous toxicity in groundwater, air and all over. The impact of such toxics in the children born after 1991 in that area as well as the population exposed during the war in 1991 strongly suggest that genetic effects in a highly complex setup have permeated all over. Conventional biostatistics tools are of very limited use in appraising such a complex and catastrophic phenomenon. The Atomic explosion aftermaths in Hiroshima and Nagasaki (Japan) have also revealed significant genotoxicity effects over generation. In the same way, it was in Vietnam, and it is likely to be the same in the present situation. A complete statistical appraisal of this genocide is beyond the means of our statistical knowledge. Nevertheless, it would be desirable to keep track of the immense biological as well as xenobiotic impacts and to grasp the aftermaths in a more objective manner (rather than summaritative statements of mortality and morbidity associated with such cruel acts).

Biostochastics plays a focal role in the evolving field of pharma-

cogenomics, pharmacokinetics, pharmacodynamics and phharmacogenetics. This is indeed where bioinformatics and biostatistics have merged to pave the way for biostochastics. It may be the microbial universe against large molecular biological systems where thousands of genes are at work. How these genes are expressed, how to find the disease gene(s), how to study the gene-environment interaction? All these are basic questions in the devlopment of new drugs and therapeutic means to better human health. Statistical considerations are so overwhelming in this context, and yet they are so fundamentally different from conventional statistical reasoning that it is indeed a challenge to come up with appropriate biostochastic tools to have satisfactory resolutions.

5 Biostochastics : The Task Ahead

It is quite clear that in bioinformatics, and large biological systems in general, there is a genuine need for statistical reasoning in every phase. Also, biostochastics should be the custodian of statistical thinking in this broad area of interdisciplinary research. However, because of deep molecular biological intricacies, biostochastics must take into account the biological factors in its approach to statistical resolutions. This delicate task in not simple or it can be accomplished in a conventional or routine way. The impasses are mainly due to the following factors:

- Hugh-dimensional, if not functional data models. For genomic sequences, it is an enormously large dimensional data set, while for the NSTM, it is actually a functional data set relating to the cortex manifold as a whole.
- Significant spatio-temporal patterns. In the NSTM, the sector of the corex and the advent of external stimulus give rise to this phenomenon. In the arsenite problem, the spatio-temporal pattern is quite evident.
- Lack of stationarity and homogeneity. This refers to both temporal and spatial variations.
- Noncontinuous (discrete / count), and often, purely qualitative categorical data models. For NSTM's it's a multidimensional

point process, while for the genomic sequence, it's purely qualitative data model. For the arsenite problem, the input variables may be continuous, but the response variables are not.

- Spatial / temporal topolgy may not be properly defined or understood. In genomic sequence, for example, adjacent positions may not have greater association in a statistical sense. In the arsenic problem, the underground arsenite distribution may not be known that well, and moreover, it might depend a lot on the flow channels for underground water or moisture.

- Sans (multi-)normality assumption, standard multivariate statistical analysis and modeling may not be appropriate. Similar criticisms may also be labeled for generalized linear mixed models (GLMM).

- Variogram, kriging, serial correlation etc. (Lawson and Cressie, 2000), resting on spatial homogeneity considerations and (almost) continuous variables, may not be generally suitable.

- In view of the high-dimensionality and biological intricacies, parametric (statistical) models are hard to justify, and as a result, classical likelihood approaches (and ramifications) may have generally serious limitations.

- Detection and elimination of outliers could be a big problem, especially with nonstandard response patterns. High-dimensionality adds more complications in this respect.

- Change-point perspectives are quite imminent, although their formulation could be much harder.

- Semiparametrics, though mathematically glittering, may have severe limitations due to the complex biological undercurrents.

- Bayesian (empirical and hierarchical Bayesian) methods are of good promise (Datta et al., 2000), although the priors should be carefully chosen so as to match the biological intricacies and to provide meaningful interpretations.

- Nonparametrics could provide a better resolution, though there is a need to appraise the data-size requirements (as the dimension could be indefinitely large).

- Stochastic partial differential equations (SPDE), as sometimes advocated for TDTK or PBPK models, may have similar limitations due to possibly inappropriate structural assumptions.
- Data mining approaches need more methodologic support, and their inherent tendency for overfitting should be critically appraised.

Let me conclude with an optimistic note. If we stick to the biological undercurrents and direct biostochastics in that avenue, most of these anomalies can be resolved to a satisfactory extent. Once this task is sorted out properly, more progress can be made with sound computational facilities provided by the modern information technology. As statisticians, we should not give up our base and migrate totally to the wanderlands of KDDM. Rather, let us try to have the KDDM guiding us in the right direction for the much needed statistical methodologic supports.

References

Agresti, A. (1990), Categorical Data Analysis. New York: John Wiley.

Bishop, Y. V. V., Fienberg, S. E. and Holland, P. W. (1975), Discrete Multivariate Analysis. MIT Press, Cambridge, Mass.

Datta, G., Ghosh, M. and Waller, L. A. (2000), Hierarchical and empirical Bayes methods for environmental risk assessment. In Handbook of Statistics, Vol. 18, Bioenvironmental and Public Health Statistics, Eds. P. K. Sen and C. R. Rao. North Holland, Amsterdam, pp 223-245.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998), Biological Sequence Analysis: Probabilistic Models for Proteins and Nucleic Acids. Cambridge Univ. Press, UK.

Ewens, W. J. and Grant, G. R. (2001), Statistical Methods in Bioinformatics: An Introduction. New York: Springer-Verlag.

Hastie, T., Tibshirani, R. and Friedman, J. (2001), The Elements of Statistical Learning; Data Mining, Inference and Prediction. New York: Springer-Verlag.

International Human Genome Sequencing Consortium (2001), Initial sequencing and analysis of the human genome. Nature, **409**, 860-921.

Lange, K. (1997), Mathematical and Statistical Methods for Genetic Analysis. New York: Springer-Verlag.

Lawson, A. B. and Cressie, N. (2000), Spatial statistical methods for envoronmental epidemiology. In Handbook of Statistics, Vol. 18, Bioenvironmental and Public Health Statistics. Eds. P. K. Sen and C. R. Rao. North Holland, Amsterdam, pp 357-395.

Pinheiro, H., Pinheiro, A. and Sen, P. K. (2003), A comparison of genomic sequences using the Hamming distance, to appear.

Pinheiro, H., Seillier-Moiseiwitsch, F. and Sen, P. K. (2000), Genomic sequence analysis and quasi-multivariate CATANOVA. In Handbook of Statistics, Vol. 18, Bioenvironmental and Public Health Statistics. Eds. P. K. Sen and C. R. Rao. North Holland, Amsterdam, pp 713-746.

Sen, P. K. (1999), Multiple comparisons in interim analysis. Journal of Statistical Planning and Inference, **82**, 5-23.

Sen, P. K. (2001a), Excursions in Biostochastics: Biometry to Biostatistics to Bioinformatics, Lecture Notes, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan.

Sen, P. K. (2001b), Toxicology: Statistical perspectives. Current Science, **80**, 1067 - 1074.

Sen, P. K. (2001c), Survival analysis: Parametrics to semiparametrics to pharmacogenomics. Brazilian Journal of Probability and Statistics, **15**, 201 - 220.

Sen, P. K. (2002a), Neuronal spatio-temporal models: High-dimensional implications and statistical perspectives. Scientie Mathematicae Japonicae, **56**, 613 - 648.

Sen, P. K. (2002b), Computational sequence analysis: Genomics and statistical controversies. In Recent Advances in Statistical Methods, Ed. Y. P. Chaubey. World Scien. Publ. UK, pp 274-289.

Sen, P. K. (2003), Structure-activity relationship information incorporation in environmental risk assessment. Environmetrics, **14**, 223-234.

Temple, R. and Ellenberg, S. S. (2000), Placebo-controled trials and active controled trials in the evaluation of new treatments, I : Ethical and scientific issues. Annals of Internal Medicine, **133**, 455 - 463.

Venter, J. C. et al. (2001), The sequence of human genome. Science, **291**, 1304-1351.

Waterman, M. S. (1995), Introduction to Computational Biology: Maps, Sequences and Genomes. UK: Chapman-Hall.