Probability in Genetics

VIth Brazilian School of Probability - Instituto do Milênio

August 2002

An excellent justification of the need of this round table, as well as a guideline for the discussions, is contained in the recent article by Richard M. Karp (Mathematical challenges from genomics and molecular biology, Notices of the AMS, vol. 49, pp. 545–53). As stated there:

Molecular biology and genomics are undergoing revolutionary changes. These changes are guided by a view of a cell as a collection of interrelated subsystems, each involving the interaction among many genes and proteins. Emphasis has shifted from the study of individual genes and proteins to the exploration of the entire genome of an organism and the study of networks of genes and proteins. As the level of aspiration rises and the amount of available data grows by orders of magnitude, the field becomes increasingly dependent on mathematical modeling, mathematical analysis, and computation.

The ultimate goal of genomics is to understand the relation between genoma and cellular processes. In its quest, the field has given rise to a number of specific problems that rely on efficient mathematical treatments. Let us mention some conspicuous examples (following the aforementioned article by Karp).

- Sequency assembling. The genoma is mapped on the basis of fragments that must be matched. There are, at present, two competing sequencing approaches: the shotgun method of the project Celera and the physical mapping of the Human Genome Project.
- Sequence comparison. In particular, comparisons are useful to relate new to known genes of the same or different species. Algorithmic

(BLAST) and probabilistic (hidden Markov models) approaches are used.

- Gene finding. In advanced organisms, genes occupy only a fraction of the genoma. The identification of the coding segments (extrons) is done by comparison among species and through hidden Markov models.

- Construction of phylogenetic trees, that is, reconstruction of the chain of divisions of species that lead to presently observed related species. Available methods rely on stochastic models combined with statistical (maximum likelihood) and combinatorial (gene rearrangement) techniques.

- Relation between polymorphism (gene variation within the same species) and disease. At present this problem is tackled via detailed statistical analysis or through decision rules determined by supervised learning techniques.

- Gene expression and cellular processes. In its present status, it involves the analysis of data obtained from DNA microarrays (classification and clustering).

The subject of the round table, and of a sizeable part of the School program, is the actual and possible contribution of probability theory to the study of these genomic problems. The table will bring together researchers from different mathematical areas working in this type of problems, as well as decision makers from Brazilians research institutions and agencies. The debate will be centered around the following rough guidelines:

- State of affairs. Present status of genomic research, in the world and within the Brazilian program. Relations between biologists and mathematicians.

- Probability and genomics today. Present use of probability theory in genomics. Advantages and limitations. Status of current research.

- Probabilty and genomics tomorrow. Areas and problems where probability theory is needed or expected to be useful, or where it has not yet lived up to its expectations.

- The mathematical priorities. Directions on which mathematical research, whichever the areas involved, should be prioritarily focused.
- The incentives. Support by agencies and institutions (grants, positions, multinational agreements). Framework for inter-departamental programs.
- The new generations. Should there be a formally recognized genomic option for math (or math-info) students? Profile?