BIOINFORMATICS

Vol. 17 no. 5 2001
Pages 389–390

# Editorial

## BIOINFORMATICS NEEDS TO ADOPT STATISTICAL THINKING

Until a couple of years ago a typical question from my colleagues on the biological workbench would be phrased something like: "I have this sequence here and cannot find out anything about it. Can you help?". From today's standpoint two things are remarkable about this question. First, it deals with only one sequence. Today, the question might be "I have got 2500 sequences...". Secondly, even when using updated terms, the question is rarely asked. More typically, I now get asked questions like: "I have 17 hybridisations of such-and-such material versus an array of 10 000 genes. Can you help me interpret the data?".

This new type of question reflects the change of paradigm in molecular biology. When I was at the European Molecular Biology Laboratory as a doctoral student with a background in mathematics, among the first things I had to learn was to think in terms of experiments. I was extremely impressed by the care taken in design and set-up of experiments and their controls. However, the typical experiment would result in a very concise output, for example a particular gel showing the binding of two components, or a microscopy image showing co-localisation of a molecule with some marker. Today's experiments are fundamentally different. The genome of an organism is sequenced with the goal of generating information to answer more than one question. But the sequence in itself is not the answer—not even the human sequence. Likewise, a micro-array experiment is not typically done with the goal of proving that a particular gene is up-regulated under certain conditions. Instead, the experiment generates a wealth of information that awaits interpretation.

The transition from 'small science' to 'big science' has to bring about fundamental changes in our way of thinking about biological data. On a very general level, one can reason about hypothesis-driven research in contrast to hypothesis-free data generation in genomics. But it is only the entry point for the hypothesis that has changed. We now pose questions to this pool of data and these questions constitute our hypothesis. The fact that bioinformatics tools allow us to ask many different questions in a relatively short time does not make this type of research hypothesis free. We should, however, appreciate how different this procedure is in light of the tradition of molecular biology. Many experimentalists experience considerable frustration because the result of an experiment does not tell them anything—at least not immediately. A result will only materialise through data analysis (if at all). The burden of data analysis, of course, has to be shouldered by bioinformatics.

No doubt bioinformatics has contributed greatly to the genomic revolution. There would be no large-scale sequencing without bioinformatics, no databases to store and hopefully retrieve the data, no functional predictions for new sequences. However, analysing large data sets poses some new kinds of problems. Some of these are purely computational. Manipulating the sequence of a human chromosome is not a trivial task and the delight with, for example, the use of Hidden Markov Models for simultaneous homology search and gene prediction soon fades when one tries to routinely apply this to a real genomic data set. In addition, large data sets may pose statistical problems, though, as is the case in the analysis of micro-array generated gene expression data. Computationally, these are just large arrays of data that are not difficult to handle while their interpretation leads to serious statistical and data analytical problems.

Indeed, probability theory and statistics have contributed significantly to the tools we are using today. Computation of statistical significance for sequence alignment and database searching has had tremendous influence on the development of genome analysis. Without it, it would not be possible to design automatic methods that extract sets of significant hits from a database search, and thus no iterative searching, no genome comparisons, etc. The success of Hidden Markov Models in describing intricate biological structures has made it utterly clear that probabilistic models are a valuable and powerful tool in bioinformatics. Molecular evolution has made heavy use of maximum likelihood estimation for phylogeny construction.

But the large-scale data sets we are confronted with now pose many new problems, mostly statistical in nature. In gene expression analysis, technological problems and biological variation make it difficult to distinguish signal from noise. Once we obtain reliable data, we look for patterns and need to determine their significance. The experiments are expensive and it is unclear how to set up the most informative experiment given the constraints on the number of hybridisations that can be performed. Protein expression levels will pose some problems similar to gene expression and will lead to interesting further questions pertaining to the relationship between these two kinds of data. Furthermore, the increasing overlap between genomics and medical research prompts many new questions. A map of human SNPs will bring about a plethora of statistical questions when researchers start to look for associations between phenotypes and genetic diversity. All of these areas require statistical methods, because the paradigm change in biology gives us very large data sets to analyse.

However, only rarely are statisticians at hand to ask all these questions. It is the local bioinformatics person or group that one can turn to, but depending on the individual background they may lack the experience in these new problems. Many obstacles will need to be overcome to open up the communication channels to statistics, though. It would seem that language barriers between computer science and biology are just slightly larger than those between computer science and statistics, or bioinformatics and statistics for that matter. The scientific cultures of bioinformatics and statistics have remained fairly distinct in spite of this ever-increasing need for interaction. In fact, I think that the major upcoming challenge for the bioinformatics community is to adopt a more statistical way of thinking and to interact more closely with statisticians. Whilst geneticists traditionally have had much tighter links to statistics, the paradigm change in biology dictates that statistical analysis of data be included much more prominently in the biology–computer science merger that has generated bioinformatics.

Eventually, bioinformatics and statistics will become essential in planning experiments. This implies that we understand the biological question and consult with the lab biologist about what is technically possible or feasible. The lab biologist and theoretician need to make a concerted effort to design experiments that *can be realised and analysed*. Bioinformaticians are predestined for this role because they have learned to bridge the communication barriers and they know the available data. But most of us need to improve the statistical know-how or learn to efficiently interact with statisticians. The consequence of all this is that we need to get back to school and learn more statistics. Not so much with the goal of mastering all of statistics but with the goal of sufficiently educating ourselves in order to pull in statisticians.

Martin Vingron
vingron@molgen.mpg.de