INTRODUCTION

A COMPUTATIONAL NECESSITY

C. Fred Fox*

Department of Microbiology, Immunology and Molecular Genetics
University of California Los Angeles, Los Angeles, CA 90095-1489

The acceleration towards complete sequencing of the human genome, as well as the genomes of several invertebrates and a host of microorganisms, can be attributed to new technology in both genomics and bioinformatics. Though production sequencing, finishing, and assembly are far from complete for the human genome, many in our community have begun to address "post-genomic" issues. This is appropriate because a recognition of the most interesting problems will also reveal deficiencies in technologies required to address them effectively, and bioinformatics stands out as the technology area which may provide the greatest benefit from the quickest fix. Presently available bioinformatics tools meet all too few of the challenges that must be confronted by scientists who perform genome-wide screens.

Companies and universities alike are scrambling for solutions. An obvious one is to hire qualified bioinformaticians to lead initiatives in research and/or instruction. In the six months from October 1999 through March of this year, over two hundred job advertisements for Ph.D. level bioinformaticians/genomicists appeared in "Science" alone. Coming from one of those acquainted with recent bioinformatics recruitment efforts: Good luck. Federal agencies anticipated this challenge a few years late, but the phenotype of not responding to problems before they hit us squarely between the eyes has been firmly established for our species. Both the National Institutes of Health and the National Science Foundation are now providing substantial resources for training in bioinformatics and related areas of computational biology, and organizations such as the Wellcome and Sloan Foundations are predictably playing important leadership roles. University administrations are responding by providing faculty groups engaged in grass roots efforts to mount bioinformatics programs with the resources necessary to get the job done effectively.

Two years ago, Dr. Susanne Huttner, who leads the University of California Biotechnology Research Education and BioSTAR Programs and I visited or talked extensively with scientists at over twenty California companies about their preparation in bioinformatics to meet future challenges. These companies ranged from the small, with a single information technologist at the B.A. level who could load software and maintain it in functional condition, to the mid-size and large, with bioinformatics departments employing from fifty to over a hundred scientists. In this latter group, bioinformatics is a very much a team sport, with players drawn from computer science, mathematics, statistics and molecular biology. It was no surprise that very few of them had received formal training in bioinformatics. A substantial number who had migrated from the aerospace industry had found that sound fundamental training in engineering and physics had provided them with quantitative tools that are applicable to problems in genomic biology. Based on information gained from this survey and with cooperation from scientists encountered during the process, a workshop was organized jointly by the University of California BioSTAR Program and the Lake Tahoe Symposia. This assembled University of California faculty and graduate students collaborating in the creation of bioinformatics training programs with industry scientists who had recognized bioinformatics challenges and could express their needs for a new breed of scientist prepared to meet them. The following group of Prospects provides examples of how bioinfor-

*Correspondence to: C. Fred Fox, Ph.D., Department of Microbiology, Immunology and Molecular Genetics, University of California Los Angeles, Los Angeles, CA 90095-1489. E-mail: fredbox@microbio.ucla.edu matics can be applied in support of genomics and proteomics.

Farlow et al., (this issue, p.171), have described the use of microarrays to monitor differential gene expression in models that simulate regeneration, differentiation, or recovery from injury. Bioinformatics is an obligatory tool for extracting large sets of gene expression data from microarrays, and the development and refinement of strategies for analysis of data sets such as these to reveal targets for drug discovery is a major challenge in bioinformatics. An example of analysis of gene sequence to support annotation for gene function is described by Bingham et al., (this issue, p.181), who define steps for elucidating which genes encode protein kinases in the fully sequenced, but relatively small, 100 megabase *C. elegans* genome. Their computational framework for analysis, which is applicable to protein families generally, has revealed that protein kinases are the second most prevalent protein family in *C. elegans* and consitute up to 2.4 per cent of its genome. The goal of achieving precise, meaningful annotation of all expressed genes for function will be a bioinformatics challenge for at least the immediate future for even the simplest genomes.

High throughput methods for genome-wide expression of proteins to support functional genomics is described by Albala et al., (this issue, p.187), who have developed a recombinant baculovirus array system subject to automation in 96 well titer plate format. This system is scaleable through automation to support a wide range of genome-wide applications to study protein function, structure or recognition, and an integrated bioinformatics platform has been created for both tracking and analysis.

Microarrays are subject to a wide variety of errors that reduce both the specificity and sensitivity of array to array comparisons in analysis of gene expression data. Schadt et al., (this issue, p.192), have developed low level analytical strategies that account for many sources of error and they devised procedures that can substantially improve the quality of gene expression data within arrays or in array to array comparisons. Improvements in low level analysis can reduce the number of replicate arrays required to validate results. This is especially important when applied to problems where the available amount of biological material may be limited, e.g., small tumors.

Extension of expression array analysis to functional genomic studies of plant to plant parasitisim are described by Torres et al., (this issue, p.203), who identified several hundreds of cDNA transcripts that are much more abundant in roots of parasitic plants that respond to host manufactured factors that induce organs associated with host invasion. Arrays of these cDNA's have been used to interrogate transcription in parasitic and non-parasitic plants to identify candidate genes for plant root signal transduction associated with parasitism. A root transcript database that will be accessible on the Internet is being developed by these authors.

The management and analysis of massive data sets accumulated in genome annotation and genome-wide gene expression studies requires sophisticated bioinformatics tools, and their development requires extensive mathematical and computational skill in addition to knowledge of biology. Though a number of universities are establishing training programs in bioinformatics, we are yet to see a consensus view on how students should prepare for bioinformatics study. What are the prerequisites? Professor Wing Wong, a co-author in this set of papers, recommends in addition to calculus, a minimum of one firm course in both probability and applied linear algebra. We have also included in this issue, the self-directed preparatory experiences of three young bioinformatics scientists, two of whom are also co-authors of papers in this set. If you wish to prepare yourself for entry to this new field, you can learn much from their experiences.