CHAPTER 2

STRATEGY AND PLANNING OF BIOINFORMATICS EXPERIMENTS

Christian Schönbach
RIKEN Genomic Sciences Center
schoen@gsc.riken.jp

Genomics has revolutionized biology by the generation of huge data sets. However, it has also created a data analysis and interpretation bottleneck. Fancy programming techniques and newest technologies are irrelevant to dealing with this bottleneck if we do not have a proper understanding of the specific questions that a biologist is seeking answers to from these data sets.

This chapter briefly discusses the planning and execution of bioinformatics experiments.

ORGANIZATION.

Section 1. We briefly explain that a novice Bioinformatician needs to learn to identify problems of interest, and then develop a solution strategy before going into details.

Section 2. Then we provide a short example to illustrate the reasoning and overall strategy for identifying alternative splice variants from cDNA libraries for the construction of a variant expression array.

Section 3. Finally, we touch on the steps that are often necessary in the planning and execution of such a bioinformatics experiment.

1. Multi-Dimensional Bioinformatics

The discovery of the DNA-double helix 50 years ago[875] triggered a new type of biology that enabled biologists to link biological phenomena to the function of genes. The main strategies are cloning, sequencing, and the analysis of specific functions on sequence, transcript, protein, cellular, and organism level. The roadmap from gene to function has changed little except the pace. High-throughput methods facilitate the sequencing and expression of whole genomes and systematic testing of protein interactions within weeks, instead of the years that these projects used to take. While Genomics has revolutionized biology by the generation of huge data sets, it has also created a data analysis and interpretation bottleneck.

To understand the underlying biology, huge amounts of heterogeneous data on different functional levels must be acquired from their sources, integrated, queried, analyzed, and modeled. The types of data include sequence data, gene expression data, luciferase assay data for protein-protein interaction data, phenotype data from knock-out mice, and so on. The different functional levels include transcription, translation, protein-protein interactions, signaling, and so on. The "omes" in genomics, transcriptomics, proteomics, metabolomics, and physiomics have generated multi-dimensional, time-dependent, and noisy data with sometimes analog non-linear interactions. To complicate matters further, the causes and effects of individual components such as transcription, translation, and protein interactions are often unknown. Therefore, bioinformatics needs to keep up with the data growth on multiple levels of biological hierarchy and the increasing overlap of biological and biomedical subdisciplines to obtain a more holistic view of biological phenomena.

In the past five or six years, bioinformatics methods of individual components—such as sequence, microarray, and structure analyses—have improved and become scalable. However, the combination of various data types and cross component analysis towards exploring the path from a gene to a disease are still in their infancy. Bioinformatics books that explain the usage of analysis tools are not the only solution for complex problem solving. A novice Bioinformatician also needs to learn to identify problems of interest in the context above, and then develop a solution strategy before going into the details. The difficulties in successfully identifying, defining, and drafting a solution to bioinformatics problems have also an environment and education component that requires attention.

Bioinformaticans should not have the illusion to efficiently solve all questions that biologists or medical researchers ask. On the other hand, the experimental scientists should not expect Bioinformaticans to be service providers and technical problem solvers. I believe that this traditional view of bioinformatics has done a lot of damage a decade ago and has not been completely eradicated. Part of the problem arose from policy makers that did not perceive bioinformatics early enough as an independent subject that merges biology, computer science, mathematics, and physics. Therefore, Bioinformaticians that can apply mathematics, physics, and computer science to biological problems are still in the minority.

2. Reasoning and Strategy

In this book, we introduce a series of biologically relevant case studies and techniques as a hands-on reference for planning and executing bioinformatics experiments. Here, I provide a brief example to illustrate the reasoning and overall strategy for identifying alternative splice variants from cDNA libraries to construct a variant expression array and protein-protein interaction experiments.

Recent EST-based analysis by Modrek *et al.*[573] and Kan *et al.*[408] estimates that 42–55% percent of the human genes are alternatively spliced. The importance of detecting alternative splice is underlined by the frequently found single base substitutions in human mRNA splice junctions and association with genetic diseases. If we want to make a conservative minimum estimate of alternative splicing of human transcripts we need to map them to genomic sequences. Sequences that cluster and show variations with a canonical splice site motif are candidates for alternative splicing.

Scoring or threshold settings can facilitate the decision on statistically significant results that are biologically relevant. Nevertheless statistically significant results are sometimes biological irrelevant. For example, the unspliced intron of a cDNA clone will produce statistically significant results to genomics DNA sequence. The result is not of biological interest unless the cDNA sequence is an alternative splice candidate or if we want to estimate which introns are the last to be spliced out. If the sequence is an alternative splice variant, it will have statistically significant hits to ESTs and genomic sequences. To increase confidence, the ESTs should be derived from various independent library sources.

When preparing a large-scale bioinformatics experiment to detect potential transcriptional variants for designing a variant microarray or protein-protein interaction screen, it is crucial to keep the number of false positives as low as possible. Error sources to be eliminated are unmapped pieces of genomic sequences and splice candidates that are hit by only one EST.

After minimizing error sources, we need to prioritize the targets by intellectual property status and relevance to disease geno- and pheno-types. A search of the transcripts against a patent database will yield potential new candidates for patenting. Potential disease genotypes can be established by using disease genes of the OMIM database.[318] Potential disease phenotype candidates, such as organ specificity, can be extracted from the tissue and/or organ sources of matched ESTs and expression information of the source clones. Disease MeSH (Medical Subject Headings) of MEDLINE abstracts extracted with the knowledge discovery support system FACTS[592] are applicable for gross classification of mRNA transcript inferred human disease associations and decision support of biomedical experts to assist the targeting process in validating potential human disease genes

3. Planning

The example in Section 2 contains several implicit steps that are often not described in publications but are applicable to any bioinformatics experiment:

- Identify the exact nature of the problem by interviewing domain experts
- Understand the data sources
- Determine
  - problem solving strategy
  - data sources
  - methodology
  - cost and time constraints
- Minimize errors and their propagation
- Evaluate overall strategy with domain experts
- Amend strategy and methodology if necessary
- Determine how to prioritize results
- Perform a test on a characterized data set
- Evaluate test results together with a domain expert
- Amend experimental set-up if required
- Start real experiment

The understanding of data sources is crucial in deriving an appropriate strategy, and must precede the algorithm selection step. Therefore, a significant amount of time should be invested in understanding and questioning the data to identifying potential problems in the data. Understanding the data requires biological knowledge that a bioinformatican should ideally know or obtain by interviewing a domain expert. Fancy programming techniques and newest technologies are irrelevant for the problem identification and data understanding steps of a bioinformatics experiment. The opposite approach of choosing the methodology and algorithm before understanding the data may induce, in the worst case, data fitting. Another common but problematic attitude caused by casual data preparation is to ignore data that do not fit the expected outcome. If the general guidelines and the biological context are taken into account, bioinformatics experiments will produce more biologically relevant results that are of interest for the biologist and medical scientist.