Example Topics of Bioinformatics

With the quick review of some warm-up biological knowledge, far from being adequate for building any solid background, we quickly move to introduce some typical examples of bioinformatics research. Some of the topics will be deeply covered by following chapters. The purpose of this section is to give readers an overall feeling about what types of questions can be and should be answered by bioinformatics and computational biology.

Examples of Algorithmatic Topics

The birth and growth of bioinformatics as a discipline was accompanied by the generation and accumulation of data in molecular biology. When data size increases, even a simple task in manipulating the data may become non-straightforward. We need special algorithms to do that. "Algorithm" means a step-by-step computational procedure for solving a problem.

For example, when we get many DNA sequence data, storing them in computer databases seems trivial. However, when we find a new sequence segment (called a query) and ask whether this sequence has already been deposited in the database, it becomes less trivial, or even challenging, when the database is huge. When we are looking not only at the sequence segments that are exactly the same as the query but at those sequences that look similar with the query, it becomes even more difficult. This is what sequence analysis is about.

The sequence database query problem can be boiled down to finding the best local alignment between two sequences, or two strings in computer science jargon. Figure 1.12 presents a very simple illustration of the problem. It was a breakthrough when the dynamic programming approach for such problems was proposed by Temple Smith and Michael Waterman in 1981, although at that time only few people realized how important the work was. Sequence alignment is a very basic problem in bioinformatics. The question has many variations, and it is the foundation for many other topics.

From the brief introduction of the shotgun sequencing method, we could realize that assembling the long sequence from many short reads is a challenging task for algorithms. It is like to solve a huge jigsaw puzzle problem. With the availability of massive deep sequencing data, a related problem is how to efficiently map the short sequence reads back to the genome.

Multiple sequence alignment brings another dimension of complexity to the problem. Comparative genomics is based on multiple sequence alignment. The genomes of multiple organisms can be compared to infer the evolutionary history of the species. Building the phylogenetic tree is an important challenge for algorithms.

Besides sequence-related problems, there are also many other types of algorithmatic problems in bioinformatics, such as finding hidden patterns in a noisy microarray matrix, inferring the amino acid sequences from possible combinations, and analyzing biological network graphs.

Examples of Statistical Topics

When we search for a short query sequence in a long genome segment, we need to design powerful algorithms to find matched targets efficiently. But when we want to infer biological conclusions from the searching result, we need to ask questions like "what is the probability of finding the matched targets in the candidate sequence under a certain biological context?" This is one type of questions that statistics help to answer.

From data point of view, there are two types of bioinformatics tasks: one is the processing of the data themselves, and the other is inferring answers to biological questions from the data. Most, if not all, biological data can be viewed as noisy sample generated by some underlying probabilistic rules. Statistical inference is a discipline to infer the underlying rules from data. The key concept is the so-called *p*-value, which gives an estimation of the probability to have the observed data when a hypothesized rule does not apply. For example, when a particular short sequence pattern (called motif) is found in the promoters of a set of genes that tend to express in a coordinated manner, one will ask the probability of observing the multiple appearance of such a sequence pattern by chance. The question can be answered with some statistical models about the DNA sequences. If the probability is small enough, then one will tend to believe that the sequence motif has some role in the coordination of the genes. This example is a simplified description of the motif discovery problem which plays a key role in many bioinformatics and functional genomics study.

In microarray study of cancers, a basic question is which genes are differentially expressed between cancer and normal samples. This is a typical statistical question, and many standard statistical methods can be applied. However, due to the special characteristics of microarray data, new methods are also needed.

The complexity nature of many types of biological data raises many new challenges to established statistical models. How to build proper statistical models based on biological knowledge and make inferences from data is a key question in many bioinformatics research. For example, in gene recognition tasks, scientists have built very sophisticated hidden Markov models that incorporate existing knowledge about gene structure.

Machine Learning and Pattern Recognition Examples

Building statistical models is one way to describe the data and make predictions. Another approach is to build a prediction machine directly from the data. This approach is called machine learning, which is an important topic in the field of intelligent information processing. When the target to be predicted is discrete classes, the task is called pattern recognition or pattern classification.

Machine learning has been widely used in bioinformatics. For example, recognizing genes and other functional elements on the genome is an important topic in bioinformatics and genomics. Scientists have developed machine learning methods such as artificial neural networks and support vector machines for these types of tasks. A learning machine is actually also a model, but not necessarily a statistical one, and data reported with biological experiments are used to train the model. HMM can also be regarded as a machine learning method. It uses a sequential statistical model to describe the data, and parameters in the model also need to be trained with known data.

Another typical example is using microarray data or proteomics expression data to classify cancers. For each patient, the gene expressions measured by microarrays compose a vector. They can be viewed as the original features for classifying the samples. One can select a smaller number of genes to classify a certain type of cancer with normal cells or to classify subtypes of the cancer. It seems like a standard pattern recognition task. However, microarray data has several unique properties: the data dimension can be very high (tens of thousands of dimension), but the sample size is usually small (in hundreds or less). Some traditional machine programs cannot work in such extreme scenario. Many people developed new or improved machine methods for this type of questions.

Besides supervised machine learning problems, unsupervised machine learning also has broad application in bioinformatics. Among the many other examples, hierarchical clustering can be used to cluster genes into groups with possible function correlation according to their expression profiles and can be used to cluster samples into groups based on their gene expressions.

Basic Principles of Genetics

Up to this point, we have compiled an incomplete compendium of research areas of modern biology from bioinformaticians' perspective. One of the important areas worth a separate discussion here is genetics. As we elaborated at the beginning of this chapter, it is hard to quantify the precise scope of bioinformatics, as a result of its multidisciplinary nature. Genetics, however, has seldom been taken as a part of bioinformatics. It sounds surprising, since both fields are entrenched on the shared methodological background: statistics and algorithm. But it is understandable that while the large body of bioinformatics is focused on a single representative sequence of the genome, the principal concept of genetics is interindividual variation which makes it quite detached from the result of biology. On the other hand, we emphasize that the development of modern genetics cannot be possible without the advancement in biotechnology and aids from bioinformatics; bioinformaticians should be acquainted with the basic principles of genetics in order to better communicate with geneticists. In this section, we take a historical approach to distill the essential concepts of genetics within the context of disease gene mapping.

Mendel and Morgan's Legacy

The dawn of the modern genetics is unanimously attributed to Mendel's seminal work on pea plant. More than 140 years ago, Mendel observed that crossing purebred peas with one binary trait (e.g., yellow and green seed color) resulted in one trait (yellow seeds) rather than a mixture of two; after selfing of F1 generation, seed color (yellow/green) exhibited 3:1 ratio. Similarly when crossing two binary traits (e.g., purple or white flower color plus spherical or wrinkled seed shape), 9:3:3:1 ratio was observed among the F2 generation for all combination of traits. Mendel postulated that each individual's binary trait was controlled by a distinct factor (later called *genes*), which had two different forms (*alleles*), recessive and dominant. Genes normally occur in pairs in a normal body cell: one is maternal derived and the other paternal derived. Within an individual, if two alleles are identical, then the individual is called *homozygous* for that gene; otherwise, the individual is called *heterozygous*. Individual's appearance is determined by the set of alleles it happens to possess (*genotype*) and environment. In case of heterozygote,

dominant allele will hide the effect of recessive allele. During the formation of sex cells (*gametes*), two alleles of a gene will segregate and pass on to eggs or sperms, each of which receives one randomly chosen allele copy (law of segregation). And

alleles of different genes will pass on independently to each other to the offspring, so there is no relation between, for example, seed shape and color of flower (law of independent assortment).

The significance Mendel's work was the proposition of the concept of gene as the discrete hereditary unit whose different alleles control different traits. It took another 40 years until the importance of Mendel's idea was recognized. Soon after geneticists rediscovered Mendel's law, they found that the independent assortment for different traits was not always the case. Instead, they observed that there are groups of traits tended to be inherited together (*linked*) by the offspring rather than assorted independently (*unlinked*). The dependence of inheritance (*linkage*) led Morgan et al. to the chromosome theory of inheritance in which chromosomes were thought to harbor genetic material. In diploid organism, chromosomes come in pairs; each *homolog* comes from one parent. During *meiosis*, the process to produce gametes, one parent provides one chromosome from each homologous pair. During first round division of meiosis, several *crossover* events will take place between homologous positions of two parental chromosomes, such that the transmitted chromosome consists alternating segments from two parental alleles. Chromosome theory elucidated the biological basis for Mendel's law of segregation and also reconciled the contradiction between linked traits and the violation to law of independent assortment. It turned out that genes controlling Mendel's pea traits were either on different chromosomes or located far apart on the same chromosome where an obligatory crossover in between must occur. Chromosome theory postulated that genes are arranged linearly along the chromosomes; the combination of nearby alleles along the same chromosome (*haplotype*) tended to be transmitted jointly unless they are shuffled by crossover.

The distance separating two genes on the same chromosome determines the frequency of their recombinant (*genetic distance*) and the probability that corresponding traits will be inherited together by offspring. By analyzing co-inheritance pattern of many linked traits from experimental crosses or family pedigrees, it is possible to place corresponding genes in order and estimate genetic distances between neighboring genes. Rigorous statistical methods were developed to construct such genetic maps. It is truly remarkable in retrospect that early-day geneticists were able to know where genes were and their relative positions even they had no idea about molecular structure of genes.

Disease Gene Mapping in the Genomic Era

The early day practice taking gene as a polymorphic landmark naturally spawned the concept of *genetic markers* (or locus) in the genomic era. Alleles giving rise to different Mendel's pea traits are just coding variants that produce different protein *isoforms* among individuals (called *non-synonymous* variants; also recall

that alternative splicing creates protein isoforms within the same individuals). There are many more types of variations whose different forms (also termed *alleles*), coding or noncoding, can be directly assayed from DNA level. While some alleles may cause changes in phenotypes, for example, increasing the risk to diseases, most are *neutral* (little phenotypic consequences) and commonly occurring within human population. Among them, two types of variations have shown greatest practical utility: single base-pair change (*single nucleotide polymorphism*, *SNP*) and short sequence of 1.6 bp repeated in tandem (*microsatellite*).

A microsatellite locus typically has tens of alleles (copy numbers of repeating unit), which can be determined via PCR amplification from unique flanking sequences. Highly variable alleles among human individuals make microsatellite the ideal markers to construct human genetic map from extended pedigrees. A map of ordered DNA markers had huge practical values. It allowed geneticists to localize *loci* (e.g., protein-coding genes and regulatory elements) whose mutations therein are responsible for the trait of our interest (e.g., diseases status and crop yield) on to the grid of prearranged genomic landmarks, a process known as *gene mapping*. The idea of *gene mapping* via *linkage analysis* is not new, inheriting the legacy from Mendel and Morgan: both DNA tags and traits loci are taken as genetic markers; and their relative orders are determined by tracing co-inheritance pattern of traits with markers in families or experimental crosses. Linkage studies using human pedigrees during the past 30 years have led to the mapping of thousands of genes within which some single mutations cause severe disorders (*Mendelian disease*), like Tay-Sachs diseases and cystic fibrosis, among others (see Online Mendelian Inheritance in Man for a complete compendium).

Encouraged by the huge success of mapping genes for rare Mendelian disease, geneticists were eager to apply the linkage analysis to common and complex diseases (like hypertension, diabetes), which also exhibit familial aggregation. But this time, they fell short of luck. At least two distinct features of common diseases are known to compromise the power of linkage analysis: first, the risk of getting the diseases for the carriers of causal variants is much lower than in Mendelian cases. Second, there may be multiple genes that, possibly through their interaction with environment, influence the disease susceptibility.

An alternative way emerged during mid-1990s. Rather than tracing the segregation patterns within families, we can pinpoint disease mutations by systematically testing each common genetic variation for their allele frequency differences between unrelated cases and controls sampled from population (*association mapping*). Aside from the practical tractability, the focus on common variants is based on the "common disease-common variants" (CDCV) hypothesis, which proposes that variants conferring susceptibility to common diseases occur commonly in population (with allele frequency >5 % as an operational criteria). While idea of association study is absolutely simple, transforming this blueprint into practices awaits for more than a decade.

As a first step toward this goal, great efforts were made in parallel with human genome project to compile a comprehensive catalog of sequence variations and map them to the reference genome backbone. SNPs are the most abundant form

of variants. In contrast to high variability of microsatellites, they typically have two alleles at each locus which can be measured by hybridization (*genotyping*). Two homologous chromosomes within an individual differ on average 1 in every 1,000 bases in their aligned regions (*heterozygosity*). And more than 95 % of those heterozygous loci will have >5 % *minor allele frequencies* within population. Up to now, it has been estimated that more than 70 % of total 10 million common SNPs have been discovered and deposited in the public databases. Other forms of variations including those altering copy numbers of large DNA chunks have also been mapped in an accelerated pace recently. Nevertheless, high abundance and easy to genotype make SNPs the primal choice for association study. Meanwhile, off-the-shelf SNP genotyping microarrays nowadays can simultaneously genotype more than half million SNPs in one individual with more than 99 % accuracy. With both genomic resources and cutting-edge technologies at hand, genome-wide association study seemed tantalizing.

But question remained: do we really need to type all the variants in the genome-wide association study (which is still infeasible)? Even provided that we could type all common SNPs, but if the disease-causing variant is not SNPs, are we still able to find them? To answer these questions, we need to take on an evolutionary perspective.

Variations do not come out of nowhere. All the variations that we observe in the current day population result from historical mutations that happen on the chromosomes that are passed on to the next generation. Each SNP is typically biallelic due to a unique point mutation event earlier in the human history (because point mutation rate is very low, 10^{-8} per site per generation, recurrent mutation is negligible). As we mentioned above, most of the variation is *neutral*, so the frequencies of newly arisen alleles will subject to random fluctuation because population size is finite (*genetic drift*). As time goes by, most of the newly arisen alleles will be removed from the population, while some of them will happen to spread across the entire population (*fixation*). So the polymorphisms we observe are those old mutations that have neither become extinct nor reached fixation until today. Some of the mutations can influence individual's fitness to the environment, for example, causing severe disorder in the early age. In such cases, the probability for this allele being transmitted to the next generation will be reduced, since the carrier may unlikely to survive until reproductive age. The frequencies of such deleterious alleles, including those causing Mendelian diseases, will be kept low as a consequence of *purifying selection*. Most common diseases, however, have only mild impact on individual's reproduction. So the variants that predispose individuals to common diseases can rise to moderate frequencies, consistent with but not proving the CDCV hypothesis.

Variations do not come alone. Whenever a new allele was born, it must be embedded on the particular background of a specific combination of existing alleles (*haplotype*) at that time. In subsequent generations, the haplotype background of that specific allele will be reshuffled by the meiotic crossovers. Because nearby markers undergo fewer crossovers, alleles of closely linked loci (be it SNPs, indels, copy number variations, etc.) exhibit allelic associations with each other (termed

linkage disequilibrium, abbreviated as *LD*). It suggests that even if the disease-causing mutations are not directly typed and tested for association, they can still be "tagged" by the alleles of nearby SNPs. And by properly selecting markers based on the LD patterns of human population, genome-wide association studies can be made in a cost-effective way. Both the marker selection and result interpretation therefore require the knowledge about the interrelationship between variants.

International HapMap Project has been completed to achieve this goal, with the priority given to the common SNPs. We now know that there are regions of tens or even hundreds of kilobases long, where diversity of SNP haplotypes is limited. These "haplotype blocks" are separated by sharp breakdown of LD as a result of punctuated distribution of crossover events (with ~80 % of crossovers happen within *recombination hotspots*). Within blocks, a reduced number of common SNPs can serve as a proxy to predict allelic status of remaining common SNPs or even other common genetic variations (like copy number gain or loss). Half million SNPs can provide adequate power in association study to test most of the common SNPs in East Asia and European populations. These findings, together with the maturity of technology and statistical methodology, have paved the way for the first wave of association study during the past 2 years. More than a hundred loci have now been identified to be bona fide reproducibly associated with common forms of human diseases.

Never satisfied by the initial success, geneticists want to extend the power of the association mapping to rare variants. To this end, they call for a map that catalogs and describes the relationships among almost all the variants, be it common and rare. Armed with cutting-edge sequencers, the 1000 Genomes Project has been launched with this ambition. Geneticists and expertise from other disciplines are now working in an ever closer manner.

References

1. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145
2. Fields S (2007) Site-seeing by sequencing. Science 316(5830):1441–1442