Introduction to bioinformatics

2

2.1 A primer to molecular biology

To understand bioinformatics, it is necessary to have a rudimentary grasp of biology. This section gives a brief introduction to some basic concepts of molecular biology that are relevant to the bioinformatics problems discussed in later chapters.

The *cell* is the basic unit of life. Despite of the diversity of cells, they all have a life cycle: they are born, eat, replicate, and die. During the life cycle, a cell makes different decisions through the manifestation in pathways.

Three types of basic molecules are present in a cell: deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins. Intuitively, DNA, RNA, and proteins can be viewed as strings. DNA is a very long molecule that is composed of four types of bases: adenine (A), thymine (T), guanine (G), and cytosine (C). Similar to DNA, there are four bases in RNA as well. The major difference is that the T base is replaced by the base uracil (U) in RNA. Each protein is a string sequence consisting of 20 types of amino acids.

DNA carries the genetic information of a cell and is composed of thousands of genes. Every cell contains the genetic information so that the DNA is duplicated before a cell divides (*replication*). When proteins are needed, the corresponding genes are transcribed into RNA (*transcription*). Therefore, RNA's primary responsibility is to synthesize the particular protein according to the protein-encoding information within the DNA (*translation*). Proteins are responsible for performing biochemical reactions, sending signals to other cells, and forming the body's major components.

Overall, DNA makes RNA, and then RNA makes proteins. This information flow,

$$\textbf{DNA} \rightarrow \textit{transcription} \rightarrow \textbf{RNA} \rightarrow \textit{translation} \rightarrow \textbf{protein},$$

is often referred to as the *central dogma in molecular biology*.

2.2 What is bioinformatics?

Bioinformatics is an interdisciplinary field that develops computational methods and software packages for analyzing biological data. As an interdisciplinary field of science, bioinformatics combines technologies from computer science, statistics, and optimization to process biological data. The ultimate goal of bioinformatics is to discover new biological insights through the analysis of biological data.

Currently, a general pipeline for addressing a science problem in bioinformatics is as follows [1]:

1. Wet labs design experiments and prepare samples.
2. Large amount of biological data are generated.

Data Mining for Bioinformatics Applications. http://dx.doi.org/10.1016/B978-0-08-100100-4.00002-8

3. Existing (or new) computational and statistical methods are applied (or developed).
4. Data analysis results are further validated by wet lab testing.
5. If necessary, the procedure of 1–4 is repeated with refinements.

However, the bioinformatics research often reflects a two-sided problem [1]: (1) Researchers in computer science and other related fields just regard bioinformatics as one specific application of their theories and methods due to the inability to provide precise solutions to complex molecular biology problems. (2) Biologists focus on hypothesis testing of wet labs so that bioinformatics serves as a tool for analyzing the biological data generated from their experiments.

It is not difficult to see that both sides have their own limitations. Computational scientists need to have a good understanding of biology and biomedical sciences, whereas biologists need to better understand the nature of their data analysis problem from an algorithmic perspective. Therefore, the lack of integration of these two sides not only limits the development of life science research, but also limits the development of computational methods in bioinformatics.

2.3 Data mining issues in bioinformatics

In recent years, rapid developments in genomics and proteomics have produced a large amount of biological data from wet labs in different formats. Data mining approaches are ideally suited for data-rich bioinformatics applications, where enormous amounts of data need to be analyzed to solve the puzzles of life. Drawing scientific conclusions from the biological data requires the application and development of data mining techniques in a right way.

However, it is vital to examine what are the important research issues in bioinformatics. There is no doubt that the objective of research is to address scientific challenges. Therefore, the real challenges in bioinformatics are how to solve the scientific issues rather than focusing too heavily on collecting and analyzing biological data. In other words, bioinformatics research ought to be science driven instead of data driven.

Because of the complexity of life's organization, there are numerous challenging research issues in bioinformatics. Thus, it is very hard to provide a complete categorization of bioinformatics problems. Generally, data analysis-related problems in bioinformatics can be divided into three classes according to the type of biological data: *sequences*, *structures,* and *networks*.

2.3.1 Sequences

The sequences of DNA, RNA, and protein are of primary importance in life science. Many bioinformatics problems that focus on the studies of sequences, which can be roughly classified into three categories: (1) *the analysis and comparison of multiple sequences*, (2) *sequence identification from experimental data,* and (3) *sequence classification and regression*.

2.3.1.1 The analysis and comparison of multiple sequences

The analysis and comparison of multiple biological sequences has become a fundamental bioinformatics issue in many different domains in modern molecular biology, whose applications range from evolutionary studies to the prediction of molecular functions and intermolecular interactions, among others. The key issue to be addressed is to find patterns or structures from a given set of biological sequences. There are many variants of such sequence analysis and comparison problems. Among these problems, multiple sequence alignment and motif discovery are probably two of the most important bioinformatics problems in the literature.

Multiple sequence alignment: A sequence alignment is a way of arranging the biological sequences to identify similar regions that may be a consequence of functional, structural, or evolutionary relationships among the sequences. Aligned sequences are typically represented as rows within a matrix, where gaps are inserted so that identical or similar characters are aligned in successive columns. There are numerous methods for aligning multiple sequences, such as Clustal W, Muscle, T-Coffee, Dialign, Mafft, DCA, and ProbCons. Figure 2.1 presents an example of multiple sequence alignment, where five biological sequences are aligned together.

Motif discovery: Motif discovery is an important bioinformatics problem with numerous applications. Generally, a sequence motif is a nucleotide or amino-acid sequence pattern that is widespread and has a biological significance. Given a set of biological sequences, the motif discovery is to find a set of motifs, where each motif satisfies the given criteria. There are different problem formulations for motif discovery in different domains, ranging from regulatory DNA motif to posttranslational modification (PTM) motif of proteins. The discovery of phosphorylation motif, which is one particular and important PTM motif, is discussed in detail in Chapter 3.

2.3.1.2 Sequence identification from experimental data

It is an important bioinformatics problem to determine the biological sequences such as DNA sequences and protein sequences from the data generated by wet lab experiments. Because of the significant differences among technologies and methods used for generating experimental data, the sequence identification problem has several variants that are totally different from the computational viewpoint. For the purpose of illustration, the identification problems for DNA sequences and protein sequences are briefly discussed.

DNA sequence identification from DNA sequencing data: To identify DNA sequences, the so-called DNA sequencing technology is widely used. In DNA

```
- A G G C T A T C A C C T G
T A G - C T A C C A - - - G
C A G - C T A C C A - - - G
C A G - C T A T C A C - G G
C A G - C T A T C G C - G G
```

Figure 2.1 An example of the alignment of five biological sequences. Here "–" denotes the gap inserted between different residues.

sequencing, many copies of original DNA sequence are cut into millions of fragments. Each copy is cut in a different way, so a fragment from one copy may overlap fragments from another. Given a set of fragments, the sequence assembly is to align and merge fragments to reconstruct the original DNA sequence.

Protein sequence identification from mass spectrometry data: In the identification of protein sequences, a protein is first digested into peptides by proteases such as trypsin. Then, the tandem mass spectrometer breaks peptides into even smaller fragments and records the mass of each fragment in a mass spectrum. The peptide sequence identification problem is to derive the sequence of a peptide given its mass spectrum. The identified peptide sequences are further assembled to infer protein sequences in the so-called protein inference procedure. The problem of protein inference is discussed in Chapter 5.

2.3.1.3 Sequence classification and regression

Many problems in biology require the accurate prediction of certain properties of biological sequences. For instance, the primary DNA sequence is highly correlated with the functional and structural protein properties. However, such relationships are still not fully understood. Therefore, computational prediction methods have to be used for advancing our understanding. Likewise, the accurate prediction of candidate PTM sites is of crucial importance for reducing the cost and time of wet lab validation. Overall, all of these bioinformatics problems can be essentially modeled as a sequence classification or regression problem. Here two representative bioinformatics applications are used for illustration purposes.

Sequence classification for phosphorylation site prediction: Phosphorylation is one of the most important and widely studied PTM events. The experimental determination of phosphorylation sites by wet labs is time consuming and labor intensive. Alternatively, one feasible approach is to collect a set of known phosphorylated peptides and another set of unphosphorylated peptides to construct a training data set. Based on the training data, a classification model is built from the labeled peptide sequences for predicting unknown phosphorylation events. The phosphorylation site prediction problem is presented in detail in Chapter 4.

Sequence regression for peak intensity prediction: Mass spectrometry is an important technique in proteomics for analyzing complex protein samples. To perform accurate protein quantification, one key problem is to predict peak intensities from peptide sequences. This is actually a sequence regression problem.

2.3.2 Structures

Structural biology is a branch of molecular biology, which is mainly concerned with the "tertiary structure" of biological macromolecules and complexes. The tertiary structure of a macromolecule is its three-dimensional structure. Macromolecules carry out most of the functions of cells on the basis of precise tertiary structures. Therefore, many bioinformatics researches focus on the study of tertiary structures of macromolecules such as RNA and proteins. Similar to the sequence-central research, bioinformatics problems

with respect to tertiary structures of macromolecules can be divided into three different classes: (1) *multiple structure analysis*, (2) *structure prediction,* and (3) *structure-based prediction*.

2.3.2.1 Multiple structure analysis

The comparison of multiple structures and the discovery of common patterns from a set of structures have numerous applications in life science. Protein structure patterns can be used for characterizing families of proteins that are functionally or structurally related. Such patterns can reveal the relationships between sequences, structures, and functions of proteins. Two representative bioinformatics problems in this category are structural alignment and structural motif discovery.

Structural alignment: The goal of structural alignment is to establish the homology between two or more structures based on their three-dimensional conformation. This procedure is usually applied to protein tertiary structures, which can transfer information about a well-known protein to unknown proteins that can be structurally aligned to it.

Structural motif discovery: One structural motif is a recurring set of residues spatially close in three dimensions, but not necessarily adjacent in the sequence. Such motifs are useful for revealing interesting evolutional and functional relationships among proteins when the sequence similarity between proteins is very low.

2.3.2.2 Structure prediction

Due to the limitation of instruments and technologies, it is still very difficult to obtain the tertiary structure of every RNA and protein through wet lab experiments. Therefore, the computational prediction of the three-dimensional structure of a protein or RNA from its primary sequence is a complementary approach for biologists. Taking the protein structure prediction as an example, a large number of prediction tools have been developed during the past 20 years. These tools adopt different principles such as homology modeling, protein threading, and *ab initio* methods.

2.3.2.3 Structure-based prediction

The structure information is the main determinant of functions in a cell. Therefore, the three-dimensional information of proteins is critical to many bioinformatics applications, such as the prediction of protein–protein interactions (PPIs), protein functions, and drug targets. This family of bioinformatics problems can be unified under the umbrella of "structure-based prediction."

2.3.3 Networks

Complex biological systems are generally represented and analyzed as networks, where vertices represent biological units and edges represent the interactions between the units. The different types of biological networks include gene regulatory networks, gene coexpression networks, metabolic networks, signaling networks, PPI networks, and more. The large number of network-related bioinformatics problems can be

categorized into several classes: (1) *network analysis*, (2) *network inference*, and (3) *network-assisted prediction*.

2.3.3.1 Network analysis

To gain insight into the organization and structure of the large biological networks, various topological structures and properties, dynamic properties, and functionality–topology relationships are expected to be addressed by the analysis of networks. Network analysis is becoming the key methodology for studying complex biological systems. There are many biological network analysis tasks. Here the *network comparison* problem is used as a representative network analysis problem.

Network comparison: This is the process of contrasting two or more biological networks from different species, conditions, or interaction types. If the target networks are PPI networks, many fundamental biological questions at the protein level can be addressed. For instance, the comparison results may tell which protein interactions are likely to have equivalent functions across species. Furthermore, it is also possible to reveal the underlying evolution of proteins, networks, and even the whole species.

Generally, three types of computational methods are available for network comparison: network alignment, network integration, and network querying. Network alignment identifies regions of similarity by globally comparing two networks, which is typically applied to detect subnetworks that are conserved across species. Network integration is to integrate networks of different types from the same species to gain a more comprehensive understanding on the overall biological system under study. The integration is achieved by merging different network types into a single network with multiple types of interactions over the same set of elements. Network querying searches a network to find subnetworks that are similar to a given subnetwork.

2.3.3.2 Network inference

It is often impossible or expensive to determine the network structure by experimental validation of all interaction pairs between biological units. A more practical approach is to infer the network structure from the indirect evidence hidden in the biological experimental data. Network inference is the process of making inferences and predictions about underlying network by analyzing the experimental data.

The topic of biological network inference is of great interest and has been extensively investigated. There are many methods for inferring different types of biological networks such as gene regulatory networks and PPI networks. For example, the combination of affinity purification with mass spectrometry analysis has become one of the leading methods for PPI network construction. To derive the underlying network structure between proteins, many computational methods have been developed, which we present in Chapters 6 and 7.

2.3.3.3 Network-assisted prediction

The available large-scale networks of molecular interactions within the cell make it possible to study many bioinformatics problems in the context of a network. Typical applications include the prediction of protein functions, disease genes, and

drug–target interactions. The basic idea of such network-assisted prediction is to use the correlation information between biological entities in the network to improve the prediction accuracy.

2.4 Challenges in biological data mining

With the development of life sciences, large-scale biological data sets are generated at various levels: genome, transcriptome, epigenome, proteome, metabolome, molecular imaging, different population of people, and clinical records. To analyze these large amounts of biological data, new bioinformatics tools and techniques should be developed to overcome many challenges.

Due to the complexity of biological systems and current limitations of instruments, the majority of biological data sets are quite noisy and highly complicated. Therefore, the development of effective data preprocessing methods is critical to the success of biological data analysis.

Data mining probably is the most popular computational tool in molecular biology. Many bioinformatics problems can be cast as standard data mining problems so that existing methods can be applied. However, some bioinformatics problems cannot be modeled as existing data mining tasks, making it necessary to develop new data mining techniques and solutions.

In addition, it is still very challenging to provide sustained performance estimates for some bioinformatics algorithms. This problem will become more serious when there are no benchmark data sets or the underlying ground truth is still lacking. For example, the entire PPI networks for most species are still not established, making it difficult to accurately evaluate the performance of PPI network inference algorithms. Therefore, more research should be devoted to the development of effective validation algorithms for assessing the data mining results in bioinformatics applications.

2.5 Summary

Bioinformatics is a field that is still advancing rapidly, making it impossible to cover all the contents of bioinformatics even within a book. In this chapter, some bioinformatics and related data analysis tasks are introduced. For further reading, I would like to recommend two popular bioinformatics textbooks [2,3].

References

[1] X. Huang, et al., No-boundary thinking in bioinformatics research, BioData Min. 6 (2013) 19.

[2] N.C. Jones, P. Pevzner, An Introduction to Bioinformatics Algorithms, MIT Press, Cambridge, MA, 2004.

[3] J.M. Claverie, C. Notredame, Bioinformatics for Dummies, John Wiley & Sons, New York, 2011.