1

INTRODUCTION

COPYRIGHTED MATERIAL

Majority of organisms on Earth, though diverse, share a significant biological similarity. There is an abundance of biological sequence data showing that any two mammals can have as many as 99% genes in common. Humans and fruit flies are two very different species that share at least 50% common genes. These striking facts have been discovered largely through biological sequence analysis.

Multiple sequence alignment is a fundamental task in bioinformatics and sequence analysis. In the early 1970s, deoxyribonucleic acid (DNA) sequences were obtained using laborious methods based on 2D chromatography. Thus, the number of sequences is limited and often being studied and annotated individually by scientists. By the late 1970s, Gilbert [1] and Sanger and Coulson [2] proposed DNA sequencing by chemical degradation and enzymatic synthesis, respectively. Their works earned a Nobel Prize in chemistry in 1980. Later, sequences are obtained by many newer methods such as dye-based methods [3], microarrays, mass spectrometry, X-ray, ultracentrifugation, and so on. Since the development of Sanger's method, the volume of sequences being identified and deposited is enormous. The current commercial sequencing such as "454 sequencing" can read up to 20 million bases per run and produce the sequences in hours. With this vast amount of sequences, manually annotating each sequence is infeasible. However, we need to categorize them by family, analyze them, find features that are common between them, and so on. The main step to solve this problem is finding the best way to start with the sequence fundamentals, thus leading the readers to the most

Multiple Biological Sequence Alignment: Scoring Functions, Algorithms and Applications, First Edition. Ken Nguyen, Xuan Guo, and Yi Pan.
© 2016 John Wiley & Sons, Inc. Published 2016 by John Wiley & Sons, Inc.

modern and practical alignment techniques that have been proven to be effective in biological sequence analysis.

1.1 MOTIVATION

There are two popular trends in sequence analysis. One trend focuses primarily on applying rigorous mathematical methods to bring out the optimal alignment of the sequences, thus leading to revelation of possible hidden biological significance between sequences. The other trend stretches on correctly identifying the actual biological significance between the sequences, where some or all biological features may have already been known. These two trends emerge from specific tasks that bioinformatics scientists are dealing with. The first trend relates to prediction of the sequence structures and homology, evolution of species, or determination of the relationship between sequences in order to categorize and organize sequence databases. The second trend is to perform a daily task in which scientists want to arrange similar known features of the sequences into the same columns to see how closely they resemble each other. Thus, the second trend can be seen in evolution analysis, in sequence structure and functional analysis, or in drug design and discovery. In the later case, for each specific virus sequence, drug designers search for possible drug-like compounds from libraries of simple sequence models annotated with functional sites and specific drug-like compounds that can bind [4, 5] . Hence, aligning a sequence obtained from a new virus against the library of sequences may lead to a manageable set of sequences and compounds to work with.

Consequently, these two distinctive perspectives lead to different approaches to sequence alignment, and the development of sequence alignment algorithms, in turn, allows scientists to automate these tremendous and time-consuming tasks.

1.2 THE ORGANIZATION OF THIS BOOK

Multiple sequence alignment study involves many aspects of sequence analysis, and it requires broad and significant background information. Therefore, we present each aspect as a chapter starting with existing methodologies and following by our contributions.

The rest of this chapter provides basic information on biological sequences.

Chapter 2 provides fundamentals in pairwise sequence alignment.

Chapter 3 describes popular existing quantitative models that have been designed to quantify multiple sequence alignment along with their analysis and evaluations.

Chapter 4 describes practical clustering techniques that have been used in multiple sequence alignment.

Chapter 5 describes, characterizes, and relates many multiple sequence alignment models.

Chapter 6 describes how traditional phylogenetic trees have been constructed and how available sequence knowledge bases can be used to improve the accuracy of reconstructing phylogeny trees.

Chapter 7 describes the latest methods developed to improve the run-time efficiency of multiple sequence alignment. A large section of this chapter is devoted to parallel alignment model on reconfigurable networks.

Chapter 8 describes several popular existing multiple sequence alignment server and services.

Chapter 9 describe several multiple sequence alignment techniques that have been developed to handle short sequences (reads) produced by the next-generation sequencing technique (NSG).

Chapter 10 describes a bioinformatics application, genetic variant detection, using multiple sequence alignment of short reads or whole genomes as input.

Lastly, Chapter 11 provides a review of ribonucleic acid (RNA) and protein secondary structure prediction using the evolution information inferred from multiple sequence alignments.

1.3 SEQUENCE FUNDAMENTALS

DNA is the fundamental unit that characterizes a living organism and its genome, that is, its genetic information set. DNA contains thousands of genes that carry the genetic information of a cell. Each gene holds information of how to build a protein molecule, which serves as building blocks for the cell or performs important tasks for the cell functions. The DNA is positioned in the nucleus, which is organized into chromosomes. Since DNA contains the genetic information of the cell, it must be duplicated before the cell divides. This technique is called duplication. When proteins are required, the corresponding genes are transcribed into RNA (transcription), the noncoding parts of the RNA are removed, and the RNA is transported out of the nucleus. Proteins are built outside of the nucleus based on the code in the RNA (see Figure 1.1). Thus, DNA sequence determines protein sequence and its structure;

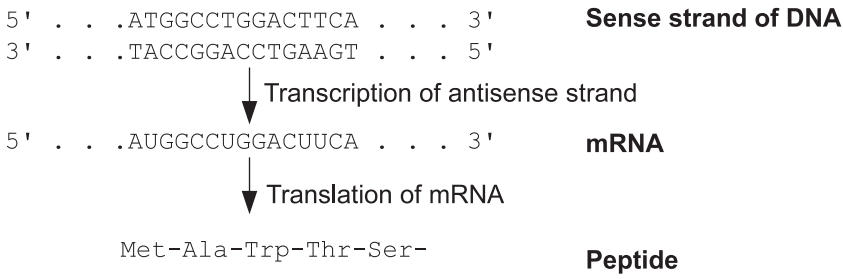
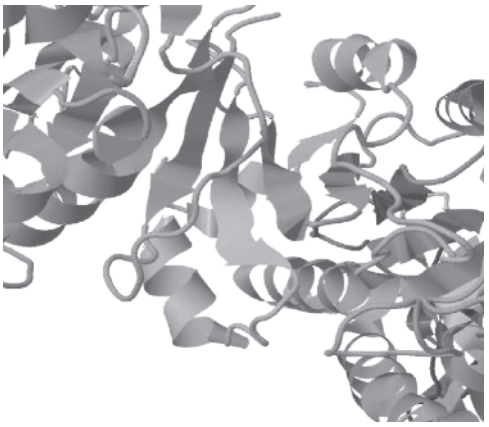


Figure 1.1 Transcription and translation processes: DNA → RNA → protein (the "central dogma" of biology).

(a)



(b) . . . PICNSRNPLELQSTTIEYITNKPVTVPPIFFFVVDLTSE . . .

Figure 1.2 The yeast Sec23/24 heterodimer 1M2V: (a) protein structure and (b) primary sequence.

TABLE 1.1 Common Amino Acids

Name	3-Letter	1-Letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V
Unknown or “other”		X

TABLE 1.2 Ambiguous Amino Acids

Ambiguous Amino Acids	3-Letter	1-Letter
Asparagine or aspartic	Asx	B
Glutamine or glutamic acid	Glx	Z
Leucine or isoleucine	Xle	J

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	–1	5																		
Asn	–2	0	6																	
Asp	–2	–2	1	6																
Cys	0	–3	–3	–3	9															
Gln	–1	1	0	0	–3	5														
Glu	–1	0	0	2	–4	2	5													
Gly	0	–2	0	–1	–3	–2	–2	6												
His	–2	0	1	–1	–3	0	0	–2	8											
Ile	–1	–3	–3	–3	–1	–3	–3	–4	–3	4										
Leu	–1	–2	–3	–4	–1	–2	–3	–4	–3	2	4									
Lys	–1	2	0	–1	–3	1	1	–2	–1	–3	–2	5								
Met	–1	–1	–2	–3	–1	0	–2	–3	–2	1	2	–1	5							
Phe	–2	–3	–3	–3	–2	–3	–3	–3	–1	0	0	–3	0	6						
Pro	–1	–2	–2	–1	–3	–1	–1	–2	–2	–3	–3	–1	–2	–4	7					
Ser	1	–1	1	0	–1	0	0	0	–1	–2	–2	0	–1	–2	–1	4				
Thr	0	–1	0	–1	–1	–1	–1	–2	–2	–1	–1	–1	–1	–2	–1	1	5			
Trp	–3	–3	–4	–4	–2	–2	–3	–2	–2	–3	–2	–3	–1	1	–4	–3	–2	11		
Tyr	–2	–2	–2	–3	–2	–1	–2	–3	2	–1	–1	–2	–1	3	–3	–2	–2	2	7	
Val	0	–3	–3	–3	–1	–2	–2	–3	–3	3	1	–2	1	–1	–2	–2	0	–3	–1	4

Figure 1.3 BLOSUM62 substitution matrix.

and the protein structure dictates the protein function. In this section, we present the fundamentals of sequences and their basic components.

1.3.1 Protein

Proteins (also known as polypeptides) are organic compounds consisting of amino acids linked with peptide bonds forming a linear polypeptide chain and folded into a globular form. The linear sequence of amino acids in the protein molecule represents its primary structure. The secondary structure refers to regions of local regularity within a protein fold such as α-helices, β-turns, or β-strands. The size of a protein sequence ranges from a few to several thousand residues. Figure 1.2 shows both the structure and the primary sequence of the yeast 1M2V protein. Tables 1.1 and 1.2 list the amino acids and their common abbreviations. Figure 1.3 shows a substitution matrix to quantify the probable substitution between two amino acids in a protein molecule.

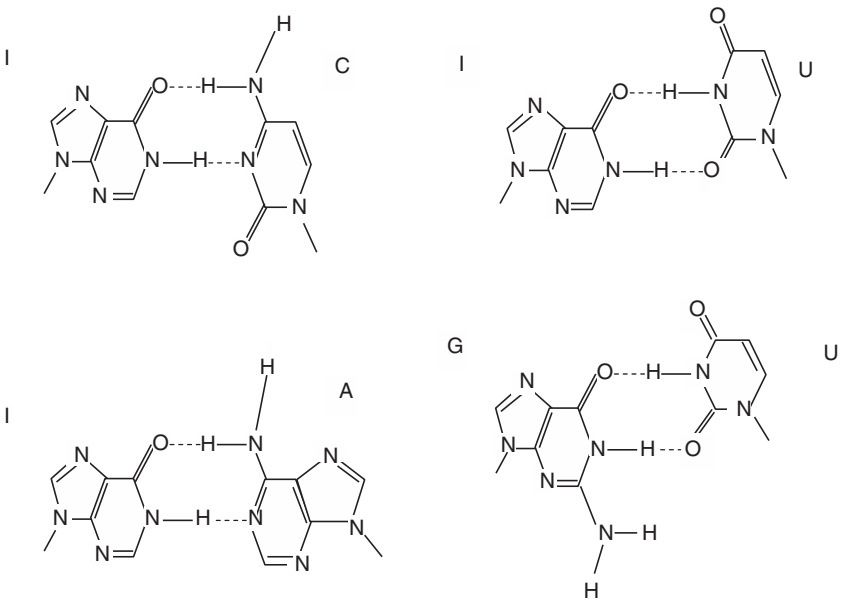


Figure 1.4 Four fundamental wobble base pairs.

1.3.2 DNA/RNA

Both DNA and RNA contain three basic components:(i) a five-carbon sugar, which could be either ribose (as in RNA) or deoxyribose (as in DNA), (ii) a series of chemical groups derived from phosphoric acid molecules, and (iii) four different nitrogen compounds having the chemical properties of bases. DNA has four bases namely adenine (A), thymine (T), guanine (G), and cytosine (C), while RNA has adenine (A), uracil (U), guanine (G), and cytosine (C) bases. Adenine and guanine are double-ring molecules known as purine, while other bases are single-ring molecules known as pyrimidine (see Figure 1.4 for an example).

DNA is a linear, double-helix structure and is composed of two intertwined chains made up of nucleotides. Unlike DNA, RNA is a single-stranded structure and contains ribose as its sugar. There are three types of RNA: messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). rRNA and tRNA are parts of protein synthesizing process, and mRNA is a template for protein synthesis. During the process of producing an amino acid chain, the nucleotide sequence of an mRNA is read from one end to the other in a group of three successive bases. These groups are called codons. Each codon is either a coding for an amino acid or a translation termination signal. There are 64 ($4 \times 4 \times 4$) possible codons for the bases.

1.3.3 Sequence Formats

Protein/DNA/RNA sequences are represented in many different formats such as AB1, ACE, CAF, EMBL, FASTA, FASTAQ, GenBank, PHD, SCF, Nexus, GFF,

Stockholm, Swiss-Prot, and so on. In general, the sequence formats can be grouped into four groups: sequencing specific, bared sequence, sequence with features, and alignment sequence. The sequencing specific formats such as AB1 from Applied Biosystems or ACE are used by companies that perform sequencing. The sequence data are obtained in fragments, called reads, and are assembled together. Many sequence formats are specifically designed for different sequencing technologies.

The bared sequence format often represents the sequence residues themselves along with an identification or sequence name. The most classic sequence format is FASTA format (or Pearson format), a text-based format where each sequence is preceded with its name and comments, and the sequence base pairs or amino acids are represented using single-letter codes. Multiple sequences can be included in a file, where each line should be fewer than 120 characters. A comment line starts with character ";" and should only be intended for human. The sequence name should starts with > character, and an asterisk "*" marks the end of a sequence and can be omitted. Each sequence should be separated by a new line. The following is an example of sequences in FASTA format:

Example 1.1 *>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken*
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEIDGDGQVNYEEFVQMMTAK

>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIEWIWG
GFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTI

In the early 1990s, the National Center for Biotechnology Information (NCBI), which houses the GenBank and genome sequencing data, defined a standard to uniquely identify the sequence. The header of the sequence that contains the sequence name should include a unique sequence identification. Figure 1.5 shows some sequence headers being used by different groups.

Sequences with feature formats allow feature data of the sequence to be included. Figure 1.6 shows a segment of sequence represented in GenBank sequence format. It includes almost everything about the sequence such as the authors of the sequence, the sequence itself, the journal in which it is published, and so on.

The alignment sequence formats such as GCG, MSF, Clustal, Phylis, and so on are used to represent sequence alignments. These formats preserve the actual alignment of the sequences making it easier for visual inspections. For example, Figure 1.7 shows two sequences in Phylis and Clustal formats.

1.3.4 Motifs

Motifs are short and conserved subsequences of amino acids that characterize a specific biochemical function. Motifs are capable of regulating particular function of a protein or can determine a substructure of a protein. Some sequence motifs are continuous such as the C2H2 zinc finger motifs. However, some motifs are discontinuous and the order in which they occur may be completely different. Thus, identifying these motifs from the sequence is not a simple task. An example of these

```
GenBank                                gi|gi-number|gb|accession|locus
EMBL Data Library                      gi|gi-number|emb|accession|locus
DDBJ, DNA Database of Japan            gi|gi-number|dbj|accession|locus
NBRF PIR                               pir||entry
Protein Research Foundation            prf||name
SWISS-PROT                             sp|accession|name
Brookhaven Protein Data Bank (1)       pdb|entry|chain
Brookhaven Protein Data Bank (2)       entry:chain|PDBID|CHAIN|SEQUENCE
Patents                                pat|country|number
GenInfo Backbone Id                    bbs|number
General database identifier            gnl|database|identifier
NCBI Reference Sequence                ref|accession|locus
Local Sequence identifier              lcl|identifier
```

Figure 1.5 NCBI's sequence formats.

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
ORGANISM    Saccharomyces cerevisiae

REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
     gene            complement(3300..4037)
                     /gene="REV7"
     CDS             complement(3300..4037)
                     /gene="REV7"
                     /db_xref="GI:1293616"
                     /translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
                     FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
                     KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
```

Figure 1.6 Illustration of GenBank sequence format.

motifs is the GCM (glial cells missing) motif, found in humans, mice, and fruit flies, which is identified by Akiyama et al. [6]. The GCM motif has the following form: WDIND*.*P..*...D.F.*W***.**.IYS**...A.*H*S*WAMRNTNNHN, where each (.) represents a single amino acid or a gap, and each * indicates one member of a closely related family of amino acids.

Since homologous sequences tend to maintain sequence similarity within core domains [7] and active sites [8], homologous sequences can have low overall

```
MCHU       ---------- ---------- ---------- ---------- ----------
gi|5524211 LCLYTHIGRN IYYGSYLYSE TWNTGIMLLL ITMATAFMGY VLPWGQMSFW

           ---------- ---------- ---------- ADQLTEEQIA EFKEAFSLFD
           GATVITNLFS AIPYIGTNLV EWIWGGFSVD KATLNRFFAF HFILPFTMVA

           KDGDGTITTK ELGTVMRSLG QNPTEAELQD MINEVDADGN GTIDFPEFLT
           LAGVHLTFLH ETGSNN-PLG LTSDSDKIPF HPYYTIKDFL GLLILILLLL

           MMARKMKDTD SEEEIREAFR VFDK------ DGNGYISAAE LRHVMTNLG-
           LLALLSPDML GDPDNHMPAD PLNTPLHIKP EWYFLFAYAI LRSVPNKLGG

           ---------- -----EKLTD EEVDEMIR-- ---------- ---EADIDGD
           VLALFLSIVI LGLMPFLHTS KHRSMMLRPL SQALFWTLTM DLLTLTWIGS

           GQVNYEEFVQ MMTAK----- ---------- -----
           QPVEYPYTII GQMASILYFS IILAFLPIAG XIENY
```

(a)

```
MCHU                        ----------------------------------------------------
gi|5524211|gb|AAD44166.1|   LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFW 50


MCHU                        ------------------------------ADQLTEEQIAEFKEAFSLFD 20
gi|5524211|gb|AAD44166.1|   GATVITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFAFHFILPFTMVA 100
                                                          *..    .*  .*::.

MCHU                        KDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDFPEFLT 70
gi|5524211|gb|AAD44166.1|   LAGVHLTFLHETGSNN-PLGLTSDSDKIPFHPYYTIKDFLGLLILILLLL 149
                              *      :* *:   .** .. . ::      .  *  * : :  :*

MCHU                        MMARKMKDTDSEEEIREAFRVFDK------DGNGYISAAELRHVMTNLG- 113
gi|5524211|gb|AAD44166.1|   LLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGG 199
                            ::*    *  .: : : .   ::.     :    :: * ** * .:**

MCHU                        ---------------EKLTDEEVDEMIR--------------EADIDGD 133
gi|5524211|gb|AAD44166.1|   VLALFLSIVILGLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGS 249
                                         *.:. . *:*                    *.

MCHU                        GQVNYEEFVQMMTAK-------------------- 148
gi|5524211|gb|AAD44166.1|   QPVEYPYTIIGQMASILYFSIILAFLPIAGXIENY 204
                              *:*   :    *.
```

(b)

Figure 1.7 (a) Phylis format and (b) Clustal format.

sequence similarity [9]. In sequence alignment studies, the terms motif and motifs have broader spectrum, they also refer to conserved segments of sequences that are observed in many sequences without the actual knowledge of the segments' biological functionality.

1.3.5 Sequence Databases

Today, there are many sequence databases available. Protein sequences can be found in Swiss-Prot [10], TrEMBL [10], UniProt [11], PIR [12], and PDP [13] databases. Similarly, DNA and RNA sequences are in GenBank [14], PDB, HOMSTRAD [15], and RefSeq [16] databases. As of June 16, 2011, TrEMBL contains 15,400,876 protein sequence entries comprising 4,982,458,690 amino acids. These entries

are automatically annotated and not reviewed. The shortest reported sequence is Q16047-HUMAN from humans containing four amino acids (F, P, D, and F), and the longest reported sequence is Q3ASY8-CHLCH containing 36,805 amino acids. TrEMBL's average sequence length is 322 residues. On the other hand, Swiss-Prot is manually annotated and reviewed. It contains 529,056 fully annotated sequence entries, of which 2.7% of the entries are predicted and about 0.4% of the entries are uncertain. The shortest sequence in Swiss-Prot is GWASEPOF (P83570) obtained from cuttlefish, which contains two amino acids (G and W), and the longest sequence is TITINMOUSE (A2ASS6) containing 35,213 amino acids. Swiss-Prot's average sequence length is 355 residues.