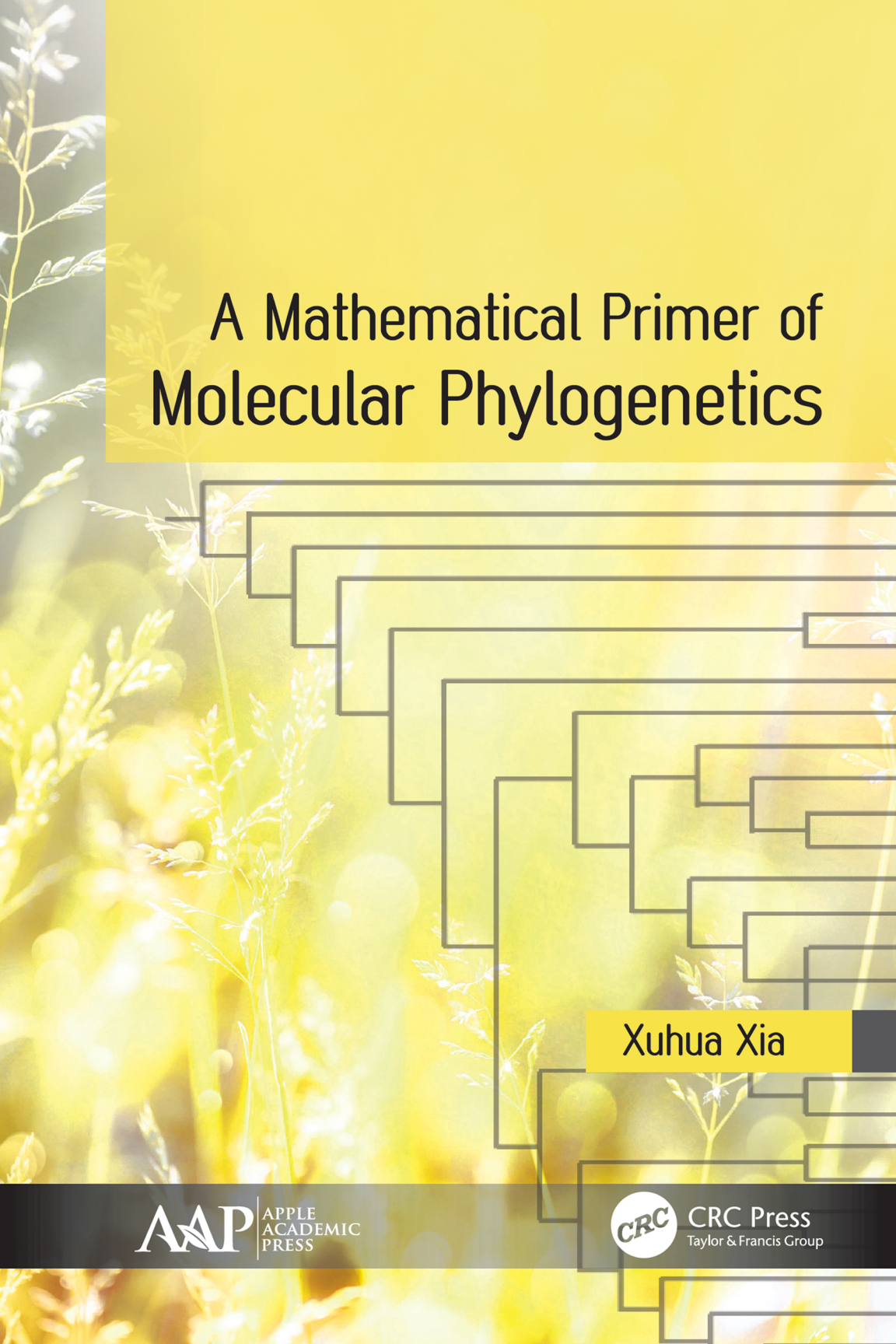
A Mathematical Primer of Molecular Phylogenetics



Xuhua Xia

AAP APPLE ACADEMIC PRESS

CRC CRC Press
Taylor & Francis Group

A MATHEMATICAL PRIMER OF MOLECULAR PHYLOGENETICS

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

A MATHEMATICAL PRIMER OF MOLECULAR PHYLOGENETICS

Xuhua Xia

APPLE ACADEMIC PRESS

Apple Academic Press Inc.
4164 Lakeshore Road
Burlington ON L7L 1A4
Canada

Apple Academic Press Inc.
1265 Goldenrod Circle NE
Palm Bay, Florida 32905
USA

Exclusive worldwide distribution by CRC Press, a member of Taylor & Francis Group

International Standard Book Number-13: 978-1-77188-755-7 (Hardcover)
International Standard Book Number-13: 978-0-42942-587-5 (eBook)

Library and Archives Canada Cataloguing in Publication

Title: A mathematical primer of molecular phylogenetics / Xuhua Xia.

Names: Xia, Xuhua, 1959- author.

Description: Includes bibliographical references and index.

Identifiers: Canadiana (print) 20190163291 | Canadiana (ebook) 20190163313 | ISBN 9781771887557 (hardcover) | ISBN 9780429425875 (ebook)

Subjects: LCSH: Phylogeny—Molecular aspects—Mathematics.

Classification: LCC QH367.5 .X53 2020 | 591.3/80151—dc23

Library of Congress Cataloging-in-Publication Data

Names: Xia, Xuhua, 1959- author.

Title: A mathematical primer of molecular phylogenetics / Xuhua Xia.

Description: Oakville, ON ; Palm Bay, Florida : Apple Academic Press, [2020] | Includes bibliographical references and index. | Summary: "This volume, A Mathematical Primer of Molecular Phylogenetics, offers a unique perspective on a number of phylogenetic issues that have not been covered in detail in previous publications. The volume provides sufficient mathematical background for young mathematicians and computational scientists, as well as mathematically inclined biology students, to make a smooth entry into the expanding field of molecular phylogenetics. The book will also provide sufficient details for researchers in phylogenetics to understand the workings of existing software packages used. The volume offers comprehensive but detailed numerical illustrations to render difficult mathematical and computational concepts in molecular phylogenetics accessible to a variety of readers with different academic background. The text includes examples of solved problems after each chapter, which will be particularly helpful for fourth-year undergraduates, postgraduates, and postdoctoral students in biology, mathematics and computer sciences. Researchers in molecular biology and evolution will find it very informative as well. Key features: Provides mathematical background for young mathematicians and computational scientists to understand the expanding field of molecular phylogenetics Includes information for researchers in phylogenetics to understand the workings of existing software packages used in phylogenetics Offers a unique perspective on a number of phylogenetic issues that have not been covered in detail in previous publications Provides support via a comprehensive software package (DAMBE), written by the book's author Aims to act as a middle ground for effective interdisciplinary communication among molecular biologists, mathematics, and computational scientists"-- Provided by publisher.

Identifiers: LCCN 2019034524 (print) | LCCN 2019034525 (ebook) | ISBN 9781771887557 (hardcover) | ISBN 9780429425875 (ebook)

Subjects: LCSH: Phylogeny--Molecular aspects--Mathematical models. | Evolutionary genetics--Mathematical models.

Classification: LCC QH367.5 .X53 2020 (print) | LCC QH367.5 (ebook) | DDC 576.8/8--dc23

LC record available at https://lccn.loc.gov/2019034524

LC ebook record available at https://lccn.loc.gov/2019034525

About the Author

Dr. Xuhua Xia obtained his Ph.D. in population biology from University of Western Ontario. He became an assistant professor at University of Hong Kong in 1996, and served in 2001 as a senior scientist and head of Bioinformatics Laboratory in the newly established HKU-Pasteur Research Centre. He came back to Canada in 2002 and has been full professor at University of Ottawa since 2009. He serves as an associate editor for several journals, including ***Molecular Biology and Evolution, Scientific Reports,*** and ***Journal of Heredity***. He has published in leading journals in molecular phylogenetics and evolution, including ***Systematic Biology, Molecular Biology and Evolution, Genetics,*** and ***Global Ecology and Biogeography.*** Dr. Xia is the author of the widely used software package DAMBE which is freely available at http://dambe.bio.uottawa.ca.

Dr. Xia's current research is on (1) developing bioinformatic algorithms and software to meet the challenge of accurate analysis of high-throughput data, (2) optimization of translation and splicing machinery to facilitate mRNA processing and protein production, and (3) interaction and evolution of macromolecules over time and space to understand the origin and maintenance of biodiversity.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

Abbreviations ..... *ix*

Preface ..... *xi*

1. Introduction to Molecular Phylogenetics **1**

2. Sequence Alignment Algorithms **15**

3. Nucleotide Substitution Models and Evolutionary Distances **69**

4. Protein and Codon Substitution Models and Their Evolutionary Distances **141**

5. Substitution Rate Heterogeneity Over Sites **171**

6. Maximum Parsimony Method in Phylogenetics **185**

7. Distance-Based Phylogenetic Methods **209**

8. Maximum Likelihood Methods in Phylogenetics **267**

9. Phylogeny-Based Comparative Methods **301**

Appendix ***329***

References ***337***

Index ***355***

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Abbreviations

BH	Benjamini–Hochberg
BMM	Brown motion model
BY	Benjamini–Yekutieli
DF	degree of freedom
EM	expectation–maximization
FDR	false discovery rate
FM	Fitch–Margoliash
GE	gap extension penalty
GO	gap open penalty
HA	hemagglutinin
HMM	hidden Markov models
IE	independent estimation
IUPAC	International Union of Pure and Applied Chemistry
LGT	lateral gene transfer
lnL	log-likelihood
LRT	likelihood ratio test
LS	least-squares
ME	minimum evolution
ML	maximum likelihood
MM	match-mismatch
MP	maximum parsimony
MSA	multiple sequence alignment
mtDNA	mitochondrial DNA
NC	number of changes
NJ	neighbor-joining
OGT	optimal growth temperature
OLS	ordinary least-squares
OTUs	operational taxonomic units
PhyPA	phylogenetics with pairwise alignment
PSA	pairwise sequence alignment
rRNA	ribosomal RNA
RSS	residual sum of squares

SD	standard deviation
SE	simultaneous estimation
SE	simultaneously estimated
SE	standard error
SM	scoring matrix
SPS	sum-of-pairs score
ssu rRNA	small subunit ribosomal RNA
TL	tree length
WC	weighted contrasts
WLS	weight least-squares

Preface

Molecular phylogenetics is important, and I wish to promote it.

Different people have different ways of promoting their ideas and beliefs. Many would use dramatic or witty titles such as "The Communist Manifesto" or "A pain in the torus." Some would resort to incendiary, but typically audacious or even mendacious claims, such as "What is true in *E. coli* is also true in the elephant, only truer," "Nothing in biology makes sense except in the light of evolution," or "All science is either physics or stamp collecting." Occasionally, some rare authors would adopt even more extreme but less acceptable ways of imposing their views on others, such as Ted Kaczynski the Unabomber.

But I am not capable of dramatizing, and English as my second language prevents me from being witty. Making incendiary or audacious claims seems repugnant to me and to those around me. In particular, I do not wish to impose my views on others just as I am not fond of having others imposing their views on me. So how am I going to promote molecular phylogenetics when all these options are unavailable?

Evangelical preachers often promote their religion by linking their belief to famous people of the past, with the implication that, if such great people have adopted their religion, then you should, too. This has resulted in the creation of fables such as Darwin repenting in his last days and Einstein being God-fearing and deeply religious. Phylogeneticists have often taken the same approach, by highlighting two historical observations. The first is a quote from Aristotle that "He who sees things from the very beginning has the most advantageous view of them," and the second is that the single figure in the 1859 book by the old man in evolutionary biology represents a phylogenetic tree. If bright people such as Aristotle and Darwin were so fond of tracing natural history back to its beginning, then surely you should, too, shouldn't you? Being an empiricist, I have tried this trick multiple times in multiple situations. Unfortunately, it did not work magic.

Some authors, confident in their eloquence and passionate about their beliefs, will simply issue a directive: "Please read the book." This is indeed a simple sentence, but I found it hard and heavy to articulate when my passion is not buttressed by eloquence.

So I will just paraphrase what A. W. F. Edwards said in his lovely little book entitled *Likelihood*. Molecular phylogenetics has been a fertile land for me. I have toiled on it and reaped the harvest. Although I would not claim myself to be a great farmer, I did have the privilege of meeting many great and productive ones on the land who have helped me to settle down comfortably. Given my own positive experience, I have no hesitation to invite you to join me in growing your crops here. This book is the best fruit of my harvest, produced in collaboration with Sandra Sickels and Ashish Kumar of Apple Academic Press. I am presenting it to you, for you to enjoy and to be convinced that it is good fruit from good earth. The book contains many useful advices on how you can grow, and improve upon, the existing crops.

If you are a young mathematics-inclined student interested in phylogenetics, this book is exactly for you. It provides not only a mathematical conceptual framework for molecular phylogenetics, but also algorithmic details and programming tips. However, I wish to take this opportunity to warn you that molecular phylogenetics is not easy and would demand two prerequisites from you. First, you need to have faith in yourself that you can learn molecular phylogenetics well. Second, you should never underestimate the difficulty in gaining proficiency in molecular phylogenetics.

While I generally do not cite religious books in teaching molecular phylogenetics, there happen to be two excellent examples in the Bible to illustrate the paramount importance of the two prerequisites. In the first example, Moses led the Israelites to the edge of their promised land flowing with milk and honey. In order to gather information to facilitate an attack, Moses sent 12 spies to survey the enemy territory. While two spies (Joshua and Caleb) came back with a united voice in favor of an attack, the other 10 were terrified by the giants inhabiting the territory and lost their faith in winning the battle. Their fear quickly spread out of control and the Israelites fled without a fight. Many biology students came to the land of molecular phylogenetics, surveyed its fertile land flowing with milk and honey, but became terrified by just a few symbols and equations that loomed large like giants, and fled without making an effort to gain an entry. Failure is guaranteed when one loses faith in oneself. In the second example, the Israelites came to the edge of Ai, a Canaanite royal city. This time they had in themselves a great deal of faith built up over 40 years of overcoming trials and tribulations. However, they committed the sin of underestimating the difficulty of conquering the city—they sent only

about 3000 half-hearted soldiers into the battle against the well-prepared enemy and consequently got beaten and slaughtered. They did learn the lesson and eventually took the city by mobilizing more than 30,000 mighty warriors and careful deployment of their forces. The bottom line is that you should never underestimate the difficulty you are facing. Many students came to the land of molecular phylogenetics half-prepared, thinking that they could master the subject by just going to the class and listening to lectures. This is equivalent to sending 3000 half-hearted soldiers when 30,000 mighty warriors are required. So have faith in yourself and try your best to mobilize the 30,000 mighty warriors in you. Don't run away and wander for another 40 years before circling back.

This book aims to serve three purposes. First, it is a personal invitation to you from a phylogeneticist. I hope that it will spin an invisible link between you and me so that I can be your personal guide. Please contact me whevever you have issues with my presentation of phylogenetic algorithms and applications. Second, it serves as a self-contained textbook that paves the way to ease your entry into the terrain of phylogenetics. Third, it represents a token of appreciation for the logistic support from University of Ottawa, the research grant from the Natural Science and Engineering Research Council of Canada and, in particular, the love and care I received from my wife and my children. One of the most emotionally voiced phrases by Christians is "God of our fathers." I wish that my children and their generation would someday come to explore this rugged terrain of science, and speak softly and emotionally "This land of our parents."

None of these purposes would be well served without your holding the book in your hand. Thank you for reading.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

CHAPTER 1

Introduction to Molecular Phylogenetics

ABSTRACT

Molecular phylogenetics has two key objectives: (1) to elucidate the branching pattern of speciation and gene duplication events, and (2) to date the speciation and gene duplication events with a molecular clock. Molecular phylogenetics is instrumental in the discovery of the three domains of life, in providing crucial evidence for the hypothesis of endosymbiosis for the origin of mitochondria and plastids, and in offering a new perspective in a variety of biological research. I illustrate the success of molecular phylogenetics with a few classic and not-so-classic examples.

Although there is only one phylogenetic tree in Charles Darwin's book (Darwin, 1859), that tree has proliferated over years and spawned a jungle of mathematics and computational algorithms. This chapter does not plunge you right into this jungle. Instead, it will just share with you a few legends and landmarks that may entice you to see more of the jungle in subsequent chapters.

1.1 GENETIC MARKERS ARE IDEAL FOR GENEALOGICAL RELATIONSHIPS

Molecular phylogenetics uses genetic markers as building blocks. The most fundamental genetic marker is the genome responsible for the manifestation of life in any living entity. These genetic markers have been imprinted a history of life, its origin, and its diversification on earth. Molecular phylogenetics aims to reconstruct this history of life from these genetic markers.

I once attended, as a graduate student at Western in the 1980s, a seminar by Dr. Shiva Singh on genetic markers after his sabbatical. Shiva showed photos of a number of places he had visited during his sabbatical, and

nobody knew where they were. Finally Shiva flashed a picture of Eiffel Tower and everyone knew that he was in Paris.

"You won't get lost if you know the landmarks," Shiva asserted, "and geneticists won't get lost if they have genetic markers," and he proceeded to offer a nice presentation on the development and application of genetic markers in solving practical biological and biomedical problems.

Genetic markers go way beyond science. I came across a story of a farmer and his two sons who lived in Germany in early 1970s. The older son spent most of his time working as a farmer like his father. They are both muscular and robust with copper-colored skin. The younger son, in contrast, disliked manual labor. He was not muscular, had pale skin, and did not look quite manly. The morphological difference between the German farmer and his younger son was so obvious that the father decided to go to court to disinherit his younger son, believing that the boy must have resulted from an extramarital affair. At that time, there was no DNA available, but the method of allozyme electrophoresis and immune responses allowed forensic scientists to reach a rather dramatic and surprising conclusion. The younger son was definitely the biological son of the German farmer, but the older one was quite doubtful. The story highlighted how lost the German farmer was without the guidance of genetic markers.

In retrospect, we can see that the morphological similarity between the German farmer and his older son is a consequence of morphological convergence resulting from working hard in the farm. Such similarities are weak for tracing human relationships. Other examples of convergence include the morphological similarities between certain placental mammals and their marsupial counterpart, e.g., between the placental wolf (*Canis lupus*) and the marsupial Tasmanian wolf (*Thylacinus cynocephalus*), between the placental cat (*Felis catus*) and the marsupial tiger quoll (*Dasyurus maculatus*), and between the placental mouse (*Mus musculus*) and the marsupial fat-tailed mouse opossum (*Thylamys elegans*). All these examples of morphological convergence that are not related to true genetic affinity, together with the peculiar phenomenon of mimicry, remind us of nature's tendency to hide true geological relationships from us.

Genetic markers have been used not only to identify paternity in humans but also in studying multiple paternity in animals. Circumstantial evidence suggests that the white-footed mouse, *Peromyscus leucopus,* may be promiscuous because (1) males do not provide any sort of parental care (Xia and Millar, 1988) based on observations in semi-natural enclosures,

and (2) males gather around females in the field only when females are in estrus (Xia and Millar, 1989). When pregnant females were brought back and allowed to give birth to her young, and when the genotype of both mother and offspring were assessed, one obtains clear evidence of multiple paternity in single litters (Xia and Millar, 1991). For example, a mother's genotype at one autosome locus is AA, but three offspring genotypes are AA, AB, and AC. This implies that alleles A, B, and C are from males. Because each male has only two alleles, at least two males must have contributed to the litter of offspring.

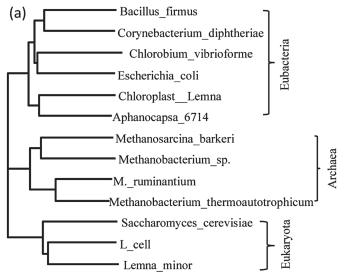
Genetic markers can also contribute to resolving national conflicts. Between Canada and the United States, there has been a long-term dispute on the management and harvest of Pacific salmon, in particular the allocation of fishing quotas (Emery, 1997). The most fundamental principle of allocation, the equity principle, is to "ensure that each country receives benefits equivalent to the production of salmon originating in its waters." This principle, rational in its articulation, has one major difficulty in its implementation. That is, how would one know if a salmon caught in the Pacific originated in Canadian or US waters? Fish biologists in the past have studied differences in morphological characters, parasite loads, and many other traits of salmon sampled in Canadian and American rivers, but these traits provide poor resolution for discrimination. Fortunately, sea-type salmons are philopatric and migrate to their natal place to breed. This implies genetic differentiation among salmon populations between Canadian and US river systems. Identification of salmon at species, population, or even individual level is now possible with well-developed DNA markers (Beacham et al., 2017).

1.2 SUCCESS STORIES IN THE APPLICATION OF DNA AND RNA AS GENETIC MARKERS

There are many success stories in the application of molecular phylogenetics, some well-known and some little known. I will present two well-known stories as well as two little known ones, partly because of my conviction that many biological Cinderellas deserve a better fate in real life, and partly because all these stories illustrate the unique insights we can gain only through molecular phylogenetics.

1.2.1 THE DISCOVERY OF THREE KINGDOMS OF LIFE

One of the landmark discoveries in molecular phylogenetics is the discovery of three domains of life (Eubacteria, Archaea, and Eukarya) in 1977. Prior to that discovery, we have two domains, prokaryotes without a cell nucleus and eukaryotes with a nucleus. Based on a similarity index (S_{AB}) derived from sharing of RNA fingerprints between taxa A and B, with $S_{AB} = 2N_{AB}/(N_A+N_B)$, Woese and Fox (1977) showed that the three domains of life are roughly of equal distance from each other. No phylogenetic tree was constructed in that paper, but one can readily derive distances (D) from S_{AB} values and reconstruct a tree (Fig. 1.1a). As the maximum of S_{AB} is 1, we may simply have $D_{AB} = 1 - S_{AB}$. I have replicated such a distance matrix in Figure 1.1b. The resulting distance matrix, when analyzed by a distance-based phylogenetic method such as FastME (Desper and Gascuel, 2002; 2004) which is also implemented in DAMBE (Xia, 2013b; 2017a), would generate the tree in Figure 1.1a with the representatives of the three kingdoms.



(b)

S_cerevisiae											
L_minor	0.71										
L_cell	0.67	0.64									
E_coli	0.95	0.9	0.94								
C_vibrioforme	0.94	0.95	0.94	0.76							
B_firmus	0.92	0.94	0.93	0.75	0.78						
C_diphtheria	0.91	0.9	0.93	0.72	0.78	0.66					
Aphanocapsa_6714	0.89	0.91	0.91	0.74	0.8	0.74	0.77				
Chloroplast_(Lemna)	0.92	0.89	0.94	0.79	0.81	0.8	0.79	0.69			
M_thermoautotrophicum	0.89	0.9	0.9	0.89	0.94	0.89	0.88	0.89	0.86		
M._ruminantium	0.89	0.9	0.9	0.88	0.93	0.87	0.88	0.89	0.88	0.49	
Methanobacterium_sp.	0.92	0.87	0.91	0.93	0.94	0.94	0.91	0.9	0.9	0.75	0.75
Methanosarcina_barkeri	0.92	0.93	0.93	0.88	0.91	0.88	0.9	0.9	0.88	0.7	0.76	0.68

FIGURE 1.1 A distance-based tree (a) with a distance matrix (b) derived from S_{AB} values in Table 1 of Woese and Fox (1977).

While the result in Figure 1.1 by itself is not strong evidence for the three-kingdom trichotomy, it is sufficient to stimulate further investigation by scientists. This has eventually resulted in empirical substantiation of the three-kingdom classification. In today's world with almost every corner of the earth being accessible by human, it would have been quite remarkable to discover just a single new species. Imagine how electrified biologists were when a whole new domain of life was discovered!

The effort to trace history back to its origin has gradually shaped a new scientific consensus of cenancestor (Xia and Yang, 2013), the common ancestor of all forms of life. The cenancestor is neither a single cell nor a

single genome, but is instead an entangled bank of heterogeneous genomes with relatively free flow of genetic information. Out of this entangled bank of frolicking genomes arose probably many evolutionary lineages with horizontal gene transfer gradually reduced and confined within individual lineages. Only three of these early lineages (Archaea, Eubacteria, and Eukarya) have known representatives survived to this day.

1.2.2 ORIGIN OF MITOCHONDRION AND PLASTIDS (e.g., CHLOROPLASTS)

One of the most fundamental questions in evolutionary biology is the origin of species, but the origin of species ultimately involves the origin of new traits, especially landmark traits such as the origin of mitochondria and chloroplasts. Mitochondria are powerhouses in eukaryotic cells, and chloroplasts allow life on earth to harvest the energy of the sun. How did eukaryotes gain these fantastic organelles?

The endosymbiosis theory, originally proposed by the Russian botanist Konstantin Mereschkowski (1905) but promoted most vigorously by Lynn Margulis (Lynn Sagan) since 1967 (Margulis, 1970; Sagan, 1967) stipulates that mitochondria and plastids in eukaryotic cells represent formerly free-living prokaryotes engulfed by other prokaryotes and reduced in the process of endosymbiosis, around 1.5 billion years ago. However, there was no direct evidence supporting the theory when Margulis articulated the theory, and her paper was rejected about 15 times before its final appearance in the *Journal of Theoretical Biology* (Margulis, 1995).

The most convincing evidence supporting the endosymbiosis theory came from phylogenetic analysis of conserved segments of small subunit (ssu) ribosomal RNA (rRNA) sequences (Gray, 1989a; 1989b; 1992; 1993) from bacteria, archaea, mitochondria of plants, fungi and animals, and chloroplasts of plants. Mitochondrial sequences from eukaryotic species appear to be monophyletic and their common ancestor clustered with Alphaproteobacteria. Similarly, chloroplast sequences clustered with cyanobacteria. The significant sequence homology represents undisputable coancestry between mitochondria and Alphaproteobacteria, and between chloroplasts and cyanobacteria.

The mitochondrial genome (mtDNA) that perhaps best represents the ancestral state is that of *Reclinomonas americana*, a heterotrophic flagellate. *R. americana* has a large mtDNA of 69,034 bp and 97 genes,

including 4 genes specifying a multisubunit eubacterial-type RNA polymerase. Almost all extant mtDNA lineages can be viewed as containing a subset of its genes. It is reasonable to infer that the proto-mtDNA is very similar to that of *R. americana,* and that extant mtDNA lineages were subsequently derived from this proto-mtDNA and now exist as various degenerated forms.

The phylogenetic relationship between the mtDNA in *R. americana* and bacterial species can be reconstructed by using small and large subunit of rRNA. rRNA genes have been termed universal yardstick (Olsen and Woese, 1993) because they are shared among all living organisms and therefore can facilitate the quantification of their phylogenetic relationship. I have reconstructed such a tree based on aligned ssu rRNA from *R. americana* mtDNA and from the genome of a diverse array of bacterial species (Fig. 1.2). It is interesting to note that *R. americana* mtDNA forms a monophyletic group with Rickettsiales, an order of bacteria that are intracellular endosymbionts or pathogens of eukaryotic cells. They all exhibit genome degeneration due to the endosymbiont or parasitic lifestyle. Thus, a mitochondrion is just an extremely degenerated intracellular endosymbiont.

While the original hypothesis of mitochondrial origin by endosymbiosis does not preclude multiple mitochondrial origins through multiple endosymbiotic events, the common consensus is that the protomitochondrion originated only once, through the internalization of a Rickettsia-like bacterium into a host cell which is more likely a prokaryotic cell than a eukaryotic one (Gray, 2012; Lane and Martin, 2010). This ancestral host with the protomitochondrion subsequently diverged into numerous eukaryotic lineages. Only the *R. americana* lineage still has a mitochondrial genome retaining many of the protomitochondrial states.

I should make a point here that a biologically appealing hypothesis, such as the endosymbiosis hypothesis for the single origin of mitochondria, is often accepted without rigorous and critical examination of relevant evidence. rRNA genes, while universal, may have difficulty even for resolving shallow phylogenies such as vertebrate phylogeny (Xia et al., 2003a). If we replace the mitochondrial ssu rRNA sequence from *R. americana* in Figure 1.2 by mitochondrial ssu rRNA sequences from other species, we do not consistently observe these other mitochondrial ssu rRNA sequences clustering with species in Rickettsiales. In fact, most of them cluster with bacterial species remotely related to Rickettsiales, that is,

evidence for monophyly of mitochondrial rRNA genes is extremely weak. We thus have at least two hypotheses. First, mitochondrial DNA, after accumulating so many substitutions eroding phylogenetic information, is no longer good genetic markers for reconstructing a deep phylogeny. Accepting this hypothesis would effectively rescind the phylogenetic support for the endosymbiosis hypothesis. Second, the diverse array of mitochondria, with many associated genetic codes, resulted from multiple origins involving multiple endosymbiosis events.

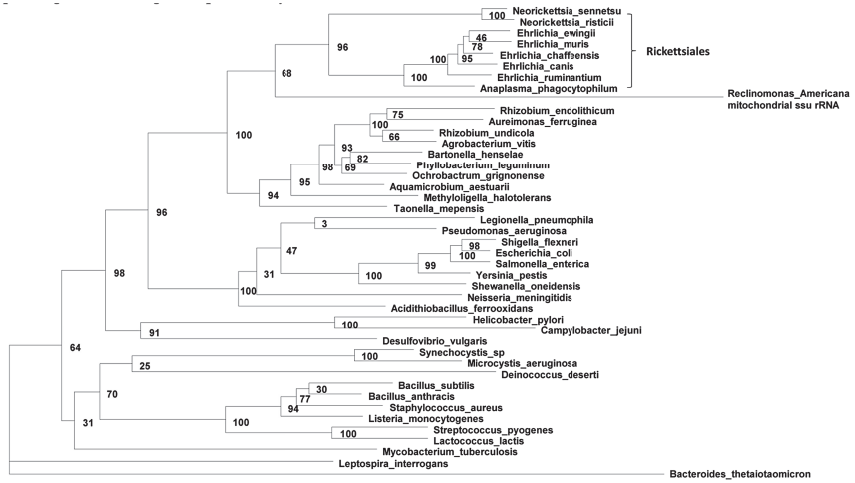


FIGURE 1.2 Phylogenetic relationship of *Reclinomonas americana* mtDNA with bacterial species, based on aligned small subunit ribosomal RNA gene sequences, reconstructed using PhyML (Guindon and Gascuel, 2003; Guindon et al., 2005) with GTR+Γ as substitution model. The values next to internal nodes are support values.

1.2.3 RESOLVING THE CONTROVERSY ON DNA METHYLATION AND CpG DEFICIENCY

CpG deficiency has been documented in a large number of genomes covering a wide taxonomic distribution (Cardon et al., 1994; Josse et al., 1961; Karlin and Burge, 1995; Karlin and Mrazek, 1996; Nussinov, 1984). DNA methylation is one of several hypotheses proposed to explain differential CpG deficiency in different genomes (Bestor and Coxon, 1993; Rideout et al., 1990; Sved and Bird, 1990). This methylation hypothesis of CpG deficiency features a plausible mechanism as follows. Methyltransferases

in many species, especially those in vertebrates, appear to methylate specifically the cytosine in CpG dinucleotides, and the methylated cytosine is prone to mutate to thymine by spontaneous deamination (Frederico et al., 1990; Lindahl, 1993). This implies that CpG would gradually decay into TpG and CpA, leading to CpG deficiency, TpG and CpA surplus, and reduced genomic GC%, which has been substantiated numerous times by genomic analysis. Different genomes may differ in CpG deficiency because they differ in methylation activities, with genomes having high methylation activities exhibiting stronger CpG deficiency than genomes with little or no methylation activity.

The seemingly well-established association between CpG deficiency and CpG-specific DNA methylation was recently challenged in a few genomic analyses (Cardon et al., 1994; Goto et al., 2000). For example, *Mycoplasma genitalium* does not have any methyltransferase and exhibits no methylation activity, yet its genome shows strong CpG deficiency. Therefore, the CpG deficiency in *M. genitalium*, according to the critics of the methylation hypothesis of CpG deficiency, must be due to factors other than DNA methylation. A related species, *M. pneumoniae*, also devoid of any DNA methyltransferase, exhibits only mild deficiency in CpG. Given the difference in CpG deficiency between the two *Mycoplasma* species, the methylation hypothesis of CpG deficiency would have predicted that the *M. genitalium* genome is more methylated than the *M. pneumoniae* genome, which is not true as neither has CpG-specific methyltransferase genes. Thus, the methylation hypothesis does not seem to have any explanatory power to account for the variation in CpG deficiency, at least in the two *Mycoplasma* species.

These criticisms are derived from phylogeny-free reasoning and its fallacy is easy to see in a phylogenetic perspective (Xia, 2003). First, several lines of evidence suggest that the common ancestor of *M. genitalium* and *M. pneumoniae* have CpG-specific methyltransferases, and should have evolved strong CpG deficiency and low genomic GC% as a result of the specific DNA methylation. Methylated m^5C exists in the DNA of a close relative, *Mycoplasma hyorhinis* (Razin and Razin, 1980), suggesting the existence of methyltransferases in *M. hyorhinis*. Methyltransferases are also present in *Mycoplasma pulmonis* which contains at least four CpG-specific methyltransferase genes (Chambaud et al., 2001). Methyltransferases are also found in all surveyed species of a related genus, *Spiroplasma* (Nur et al., 1985). These lines of evidence suggest

that methyltransferases are present in the ancestors of *M. genitalium* and *M. pneumoniae.*

Second, the methyltransferase-encoding *M. pulmonis* genome is even more deficient in CpG and lower in genomic GC% than *M. genitalium* or *M. pneumoniae,* consistent with the methylation hypothesis of CpG deficiency (Fig. 1.3). It is now easy to understand that, after the loss of methyltransferase in the ancestor of *M. genitalium* and *M. pneumoniae* (Fig. 1.3), both genomes would begin to accumulate CpG dinucleotides and increase their genomic GC%. However, the evolutionary rate is much faster in *M. pneumoniae* than in *M. genitanlium* based on the comparison of a large number of protein-coding genes (Xia, 2003). So *M. pneumoniae* regained CpG dinucleotide and genomic GC% much faster than *M. genitalium* whose slow rate of genomic evolution allows it to retain the ancestral low CpG phenotype better than *M. pneumoniae.* In short, the Mycoplasma genomic data that originally seem to contradict the methylation hypothesis actually provide strong support for the methylation hypothesis when phylogeny-based genomic comparisons are made.

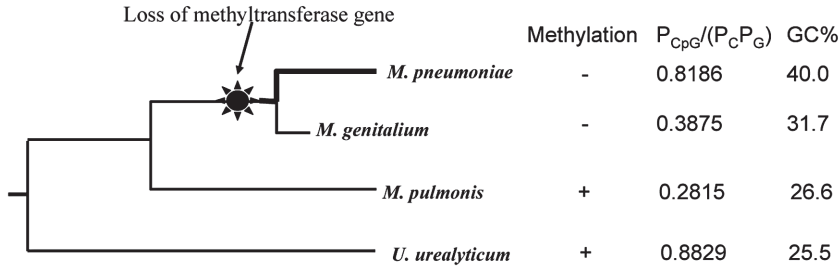


FIGURE 1.3 Phylogenetic tree of *Mycoplasma pneumoniae*, *M. genitalium*, and their relatives, together with the presence (+) or absence (−) of CpG-specific methylation, $P_{CpG}/(P_CP_G)$ as a measure of CpG deficiency, and genomic GC%. *M. pneumoniae* evolves faster and has a longer branch than *M. genitalium*. Cytosine methylation in *U. urealyticum* is not CpG specific, so it does not reduce CpG dinucleotide but does reduces GC% in the genome.

One might note that *Ureaplasma urealyticum* in Figure 1.3 is not deficient in CpG because its $P_{CpG}/(P_CP_G)$ ratio is close to 1, yet its genomic GC% is the lowest. Has its low genomic GC% resulted from CpG-specific DNA methylation? If yes, then why doesn't the genome exhibit CpG deficiency? It turns out that *U. urealyticum* has C-specific, but not CpG-specific, methyltransferase, so the genome of *U. urealyticum* is expected, and indeed observed, to have low CG% (because of the

methylation-mediated C→T mutation) but not a low $P_{CpG}/(P_C P_G)$ ratio. The methyltransferase gene from *U. urealyticum* is not homologous to those from *M. pulmonis.*

1.2.4 CHILOÉ ISLAND AND DARWIN'S FOX

Off the western coast of South America is Chiloé Island on which a special kind of fox, named Darwin's fox (*Dusicyon fulvipes*), was found. On the mainland opposite the island thrives another fox species, the gray fox (*Urocyon cinereoargenteus*). For a long time it has been thought that Darwin's fox has descended from the gray fox. In other words, during the Quaternary glaciation period with a low sea level (because much water was retained on land in the form of ice sheets), Chiloé Island, just slightly north of the northern edge of glaciation, was connected to the mainland. Gray foxes were expected to roam the Chilean Coast Range including Chiloé Island. When glaciation period ended, the ice sheets returned to the ocean and sea level rose to isolate the island from the mainland. Gray foxes that remained on the island became isolated from the mainland population and diverged independently to become Darwin's fox. Because the last glaciation ended only about 10,000–15,000 years ago, the divergence time between Darwin's fox on the island and gray fox on the mainland was thought to be just about 10,000–15,000 years. It is partly for this reason that Darwin's fox had been classified as a subspecies of the gray fox because a period of 10,000–15,000 years of isolation does not seem sufficient for the evolution of a new mammalian species (Yahnke et al., 1996).

In 1980s when molecular techniques became widely available to field biologists, researchers began to reconstruct phylogenetic trees for various fox species and to date their speciation events. They were surprised to find that the divergence time between Darwin's fox and the gray fox was about a million years, much longer than the originally hypothesized 10,000–15,000 years. This is clearly incompatible with the original hypothesis that Darwin's fox evolved from gray fox after the isolation of Chiloé Island from the mainland at the end of last glaciation period.

One possible hypothesis is that Darwin's fox had diverged from the gray fox for a long time on the mainland, long before the geographic separation of Chiloé Island from the mainland. During the last glaciation period, some Darwin's foxes, not gray foxes, remained on the island and became isolated from the mainland population of Darwin's fox. Meanwhile, the

mainland population of Darwin's fox had gone extinct in competition against the gray fox.

This is a bold hypothesis. It predicted the existence of a species on the mainland that nobody had seen. However, researchers had faith in the prediction and went on looking for historical footprints (e.g., fossils) of Darwin's fox left on the mainland. It is in search of these footprints that researchers were pleasantly surprised to find a living population of Darwin's fox on the mainland. This discovery, in my opinion, rivals the success of predicting the existence of an unseen planet based on the orbits of other visible planets in astrophysics.

You might be thinking privately that the researchers, out of desperation to support their hypothesis, might have sneaked onto the island, caught some Darwin's foxes, transported them to the mainland, and then declared "Lo and behold ..." However, the molecular clock can again come to their rescue. If they were indeed guilty of the crime, then there would be no genetic variation between the island population and the mainland population. However, if the island population has been isolated from the main population for 1,000–15,000 years, then there should be genetic variation consistent with such isolation.

The dating evidence that Darwin's fox diverged from the gray fox for about a million years immediately raised Darwin's fox not only from subspecies to species, but also to a high conservation status because the population has only about 500 individuals. Keep in mind that species conservation has two essential criteria. The first is that the species is indeed endangered. The second is genetic uniqueness. If Darwin's fox diverged from gray fox for only 10,000–15,000, then it will not be considered genetically unique enough for a high conservation status. A divergence time of a million years makes all the differences. While the exact sequence of events related to the phylogenetic research on Darwin's fox is difficult to reconstruct, the potential of molecular phylogenetics in science and in species conservation is clearly visible.

1.3 TWO KEY OBJECTIVES OF MOLECULAR PHYLOGENETICS

Molecular phylogenetics has two objectives (Fig. 1.4): (1) characterizing the branching patterns (cladogenic events, specifically the speciation and gene duplication events) in evolution and (2) dating of these speciation or gene duplication events that may help us understand origin of species

and functional divergence of duplicated genes (Vlasschaert et al., 2017; Vlasschaert et al., 2015). Phylogenetics can also help us to identify common ancestors such as the mitochondrial "Eve" (because mitochondrial genomes in mammals are maternally inherited) or the Y-chromosome "Adam" (because human Y-chromosome is passed down from father to son). The universal common ancestor for all living organisms is termed cenancestor which is assumed to exist on the basis of extensive sharing of inferred homologous characters among representatives of living cellular organisms, such as the near universal genetic code, the concordance of phylogenetic trees from different genes, the sharing of fundamental biochemical processes, and the existence of numerous transitional fossils. Cenancestor is a logical necessity if the cellular structure originated only once, and if we assume to be true the cell theory stating that new cells are created only by old cells dividing into two. One early concept of cenancestor is a genome that codes a minimal set of core genes essential for cellular life (the minimal genome) and from which all other genomes are derived. However, few genes are shared universally because a biological function can often be performed by unrelated genes. Even if such a set of core genes can be identified, the identification and dating of the cenancestor is difficult because of the lack of a universal global molecular clock and the rampant horizontal gene transfer.

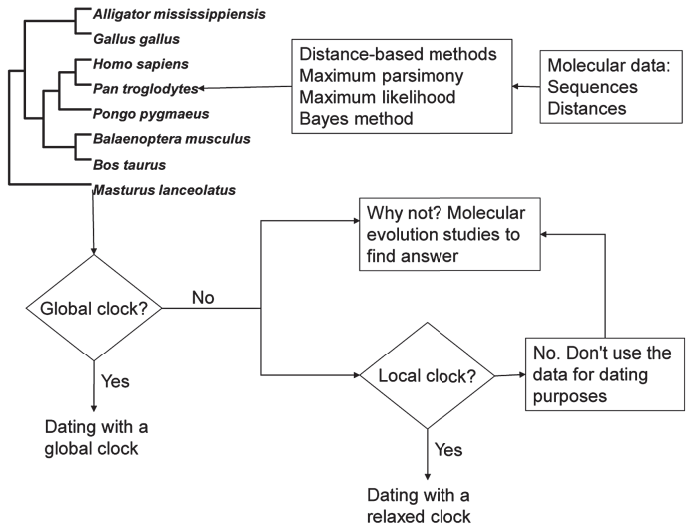


FIGURE 1.4 High-level summary of molecular phylogenetics: defining branch patterns and dating internal nodes.

Note that the first objective of molecular phylogenetics, i.e., characterizing the branching patterns (cladogenic events, specifically the speciation and gene duplication events) in evolution involves two types of genes, orthologous and paralogous genes. Paralogous genes arose by gene duplication within a lineage, e.g., α-globin and β-globin genes in a mouse genome, but orthologous genes are more difficult to define. In this book, I impose a strong definition of orthologous genes, that is, they are single-copy genes in different genomes resulting from speciation, descending from a common ancestral genome, and having never undergone gene duplication, that is, their "single-copy" is an ancestral character that does not result from gene duplication and then gene loss leading to the survival of a single copy. The *ERN1* gene from human and mouse are orthologous because the gene has no paralogues not only among mammals, but also among vertebrates, suggesting an extremely low likelihood of the gene ever being duplicated. Species trees should only be inferred from orthologous genes, while gene duplication events are inferred from paralogous genes by studying genes in a gene family (which is a collection of all homologous genes in one genome) from multiple lineages, ideally with one or more lineages that branched out before any gene duplication events and consequently retain the single-copy ancestral status.

Both species events and gene duplication events contribute to biodiversity and represent parts of natural history. Molecular phylogenetics aims to trace natural history as close as possible to the cenancestor and to reconstruct what experiments nature has performed over billions of years. Such knowledge will not only enlighten us on the origin and evolution of biodiversity, but also guide us toward a more profitable and more harmonious way of life. This last sentence I borrowed from an Evangelical preacher because it sounds very profound.

Molecular phylogenetics almost always start with compilation of homologous sequences from OTUs (optional taxonomic units such as species), alignment of the sequences to identify homologous sites and inference of phylogenetic relationships among the OTUs represented by the sequences. There could be many multiple sequence alignments and phylogenetic trees from even a single set of homologous sequences, so some criteria are always necessary for us to choose the best alignment among many alternative alignments, and the best tree among many possible trees. One may go so far as to claim that the entire field of molecular phylogenetics is about formulating, justifying, and applying these criteria. Pay particular attention to these criteria in reading the following chapters.

KEYWORDS

- **three kingdoms**
- **endosymbiosis**
- **mitochondria**
- **DNA methylation**
- **Darwin's fox**

CHAPTER 2

Sequence Alignment Algorithms

ABSTRACT

Accuracy of molecular phylogenetic analysis depends on correct identification of homologous sites. Sequence alignment serves two purposes: Global alignment is mainly for identifying site homology between sequences to facilitate the inference of ancestral-descendant relationships and local alignment mainly for identifying sequence similarities that may be due to either coancestry or convergence. I illustrate, with published data, how misalignment can distort phylogenetic signal. Sequences can be aligned in many different ways, so a criterion is needed for choosing the best alignment. Operationally, the best alignment is one with highest alignment score for a given scoring scheme. Dynamic programming algorithm guarantees to find the best alignment with the highest alignment score. It is illustrated in detail for pairwise alignment and profile alignment with both constant gap penalty and affine function gap penalty, followed by progressive multiple sequence alignment using a guide tree, and by how to align protein-coding nucleotide sequences against aligned amino acid sequences. PAM and BLOSUM matrices, which are typically derived from protein alignments, are derived from both nucleotide and amino acid sequences. The effect of mutation, selection, and amino acid dissimilarities on substitution frequencies were illustrated and discussed.

Almost all molecular phylogenetic studies start with sequence alignment of homologous sequences. Global sequence alignment (Needleman and Wunsch, 1970) and local sequence alignment (Smith and Waterman, 1981) by dynamic programming represent the core algorithms for sequence alignment. Dynamic programming algorithms constitute a general class of algorithms not only used in sequence alignment but also in many other applications. For example, the Viterbi algorithm and the forward algorithm used in hidden Markov models (HMM), which were numerically

illustrated in detail (Xia, 2018a, Chapter 7) are also dynamic programming algorithms, so are Fitch and Sankoff algorithms for maximum parsimony (MP) and the pruning algorithm for maximum likelihood (ML) reconstruction of phylogenetic trees. We will cover MP and ML algorithms in great numerical detail latter. Learning the dynamic programming algorithms used in sequence alignment paves the way for more advanced applications in later chapters.

This chapter covers (1) pairwise global and local alignment by dynamic programming with different scoring schemes, from the simplest scoring scheme with two-valued match/mismatch scores and constant gap penalties, to the more useful scoring schemes with match-mismatch (MM) matrices and affine function gap penalties, (2) detailed derivation of PAM and BLOSUM matrices, (3) profile alignment between one sequence and a set of aligned sequences which is essential for practical implementation of multiple sequence alignment (MSA), and (4) multiple alignment that is reduced to pairwise alignment and profile alignment by using a guide tree. Most textbooks on bioinformatics omit the dynamic programming algorithm using affine function gap penalty (Gotoh, 1982) and no textbook I know of includes any detailed explanation of profile alignment. This chapter is intended to fill the gap.

The objective of sequence alignment is to identify homologous sites among sequences so that functional and phylogenetic inferences can be made. For example, the multiple alignments of FoxL2 protein (Fig. 2.1) show a highly conserved and positively charged domain (the forkhead domain, with many positively charged residues, such as R and K) which should have strong electrostatic interactions with negatively charged molecules such as nucleic acids. It was found that FoxL2 is indeed a nuclear transcription factor with the forkhead domain being DNA-binding (Baron et al., 2004; Cocquet et al., 2003). Another feature standing out from the alignment is the conserved polyalanine tract of exactly 14 residues (Fig. 2.1). Indeed, lengthening the polyalanine tract is frequently associated with the blepharophimosis syndrome (De Baere et al., 2002). From a phylogenetic point of view, one can immediately see the difference between the mammalian sequences (first seven in Fig. 2.1) and the fish sequences (the last three in Fig. 2.1).

While biological insights can often be derived directly from MSA, the main objective of MSA is to build phylogenetic trees so as to make phylogeny-based inferences. Inaccurate multiple alignments can introduce

not only phylogenetic noise but also distort phylogenetic signals. This I illustrate below based on a reanalysis by Noah et al. (2020) of aligned sequences from a paper published in the journal *Nature.*

```
M AEKRLTLSGIYQYIIAKFPFYEKNKKGWQNSIRHNLSLNECFIKVPREGGGERKGNYWTLDPACEDMFEKGNYRRRRRMKRPFRP
R AEKRLTLSGIYQYIIAKFPFYEKNKKGWQNSIRHNLSLNECFIKVPREGGGERKGNYWTLDPACEDMFEKGNYRRRRRMKRPFRP
B AEKRLTLSGIYQYIIAKFPFYEKNKKGWQNSIRHNLSLNECFIKVPREGGGERKGNYWTLDPACEDMFEKGNYRRRRRMKRPFRP
C AEKRLTLSGIYQYIIAKFPFYEKNKKGWQNSIRHNLSLNECFIKVPREGGGERKGNYWTLDPACEDMFEKGNYRRRRRMKRPFRP
S AEKRLTLSGIYQYIIAKFPFYEKNKKGWQNSIRHNLSLNECFIKVPREGGGERKGNYWTLDPACEDMFEKGNYRRRRRMKRPFRP
H AEKRLTLSGIYQYIIAKFPFYEKNKKGWQNSIRHNLSLNECFIKVPREGGGERKGNYWTLDPACEDMFEKGNYRRRRRMKRPFRP
O AEKRLTLSGIYQYIIAKFPFYEKNKKGWQNSIRHNLSLNECFIKVPREGGGERKGNYWTLDPACEDMFEKGNYRRRRRMKRPFRP
F SEKRLTLSGIYQYIISKFPFYEKNKKGWQNSIRHNLSLNECFIKVPREGGGERKGNYWTLDPACEDMFEKGNYRRRRRMKRPFRP
T SEKRLTLSGIYQYIISKFPFYEKNKKGWQNSIRHNLSLNECFIKVPREGGGERKGNYWTLDPACEDMFEKGNYRRRRRMKRPFRP
D SEKRLTLSGIYQYIISKFPFYEKNKKGWQNSIRHNLSLNECFIKVPREGGGERKGNYWTLDPACEDMFEKGNYRRRRRMKRPFRP
   ************** ************************************************************************

M PPAHFQPGKGLFGSGGAAGGCGVPGAGADGYGYLAPPKYLQSGFLNNSWPLPQPPSPMPYASCQMAAAAAAAAAAAAAGPGSPG
R PPAHFQPGKGLFGSGGGAGGCGVPGAGADGYGYLAPPKYLQSGFLNNSWPLPQPPSPMPYASCQMAAAAAAAAAAAAAGPGSPG
B PPAHFQPGKGLFGAGGAAGGCGVAGAGADGYGYLAPPKYLQSGFLNNSWPLPQPPSPMPYASCQMAAAAAAAAAAAAAGPGSPG
C PPAHFQPGKGLFGAGGAAGGCGVAGAGADGYGYLAPPKYLQSGFLNNSWPLPQPPSPMPYASCQMAAAAAAAAAAAAAGPGSPG
S PPAHFQPGKGLFGAGGAAGGCGVAGAGADGYGYLAPPKYLQSGFLNNSWPLPQPPSPMPYASCQMAAAAAAAAAAAAAGPGSPG
H PPAHFQPGKGLFGAGGAAGGCGVAGAGADGYGYLAPPKYLQSGFLNNSWPLPQPPSPMPYASCQMAAAAAAAAAAAAAGPGSPG
O PPAHFQPGKGLFGAAGAAGGCGVAGAGADGYGYLAPPKYLQSGFLNNSWPLPQPPSPMPYASCQMAAAAAAAAAAAAAGPGSPG
F PPTHFQPGKSLFG--------------GDGYGYLSPPKYLQSSFMNNSWSLGQPPAPMSYTSCQMASGNVSPVN-----------
T PPTHFQPGKSLFG--------------GDGYGYLSPPKYLQSSFMNNSWSLGQPPPPMSYTSCQMASGNVSPVN-----------
D PPTHFQPGKSLFG--------------GEGYGYLSPPKYLQSGFINNSWS----PAPMSYTSCQVSSGSVSPVN-----------
  ** **********                ***** ******* * ****     * ** *.***
```

FIGURE 2.1 Partial multiple alignments of partial FoxL2 protein of 10 vertebrate species. M: *Mus musculus*, mouse (GenBank AI: AF522275); R: *Rattus norvegicus*, rat (AI: AC105826); B: *Bos taurus*, cow (AI: AY340970); C: *Capra hircus*, goat (AI: AY112725); S: *Sus scrofa*, pig (AI: AY340971); H: *Homo sapiens*, human (AI: AF301906); O: *Oryctolagus cuniculus*, rabbit (AI: AY340972); F: *Fugu rubripes*, pufferfish (AI: Scaffold_8165/Prot JGI_24134); T: *Tetraodon nigroviridis*, tetraodon; and D: *Danio rerio*, zebrafish. Shown are the part of the highly conserved and positively charged forkhead domain and its downstream polyalanine tract (in bold) which is missing in the three fish species.

2.1 POOR ALIGNMENT INTRODUCES NOT ONLY NOISE BUT ALSO PHYLOGENETIC BIAS

Reliable MSA is difficult to obtain with divergent lineages because of erosion of homology over time (Blackburne and Whelan, 2013; Edgar and Batzoglou, 2006; Herman et al., 2014; Kumar and Filipski, 2007; Lunter et al., 2008; Wong et al., 2008; Xia, 2016). A poor alignment typically leads to bias and inaccuracy in phylogenetic estimation (Blackburne and Whelan, 2013; Kumar and Filipski, 2007; Wong et al., 2008; Xia et al., 2003a). There are many publications with poor MSA. The examples of alignment errors shown here are taken from the Online Supplemental file nature08742-s2.nex in Regier et al. (2010).

2.1.1 PHYLOGENETIC NOISE INTRODUCED BY POOR ALIGNMENT

A sample of the alignment from Regier et al. (2010) is shown in Figure 2.2a, together with an alternative alignment (Fig. 2.1b) which is clearly more preferable. The phylogenetic impact of a poor alignment is often unpredictable. If a phylogenetic analysis includes the poorly aligned region in Figure 2.2a, then the evolutionary distance among the species or branch length in the tree will be overestimated. If one excludes this poorly aligned region, then the distances and branch lengths may be underestimated. Regier et al. (2010) kept the poorly aligned region in nucleotide-based phylogenetic analysis, which generated their main results in their Figure 2.1, but excluded the region in amino acid-based analysis. While the alignment in Figure 2.2b is visibly better than that in Figure 2.2a, we do need to have a sensible criterion for evaluating different alignments. We will learn to use pairwise alignment score and its derivatives to evaluate the quality of sequence alignment.

```
(a)          190       200       210
            ---|----|----|----|----|----...
FauNEOPT    GAUGUUCCACCUCCAGUA---GAAUUUU...
ApaukNEOPT GGYCGCCUCCCGGUA---------GAACUGU...
CpoNEOPT   GGCGGCAAGCAACCUGUG------GAACUGU...
PquNEOPT    AACGGUCGCGCGCCGGUC---GAGCUGU...
PamNEOPT    GACACACCACCUCCAGUG---GAAUUCU...
AdoNEOPT    AAUUUGCCACCUCCA---GUGGAGUUUU...

(b)
FauNEOPT    GAUGUUCCACCUCCAGUAGAAUUUU...
ApaukNEOPT ---GGYCGCCUCCCGGUAGAACUGU...
CpoNEOPT    GGCGGCAAGCAACCUGUGGAACUGU...
PquNEOPT    AACGGUCGCGCGCCGGUCGAGCUGU...
PamNEOPT    GACACACCACCUCCAGUGGAAUUCU...
AdoNEOPT    AAUUUGCCACCUCCAGUGGAGUUUU...
```

FIGURE 2.2 Part of multiple alignments for a subset of six species (a) taken from the supplementary file (nature08742-s2.nex) in Regier et al. (2010). Realignment by MAFFT with the optimized options is shown in (b). Note that the two codons highlighted in red (coding for amino acids Pro and Val) are identical among the six species. Dots in (b) represent nucleotides not present in (a).

2.1.2 POOR ALIGNMENT CAN INTRODUCE PHYLOGENETIC BIAS

The original alignment in Regier et al. (2010) in the left panel of Figure 2.3 suggests a phylogenetic similarity between the first nine species and the last two species, with the two Archeognatha species (PsaARCHEO for *Pedetontus saltator* and MbaARCHEO for *Machiloides banksi*) and a copepod (A369COPE for *Acanthocyclops vernalis*) being different. However, the last codon in red (Fig. 2.3) is a lysine codon in all sequences, and the second last is a threonine codon in all but one sequence (A369COPE). The evidence of homology is strong among these codon sites that they should really be aligned as shown in the right panel of Figure 2.3. Thus, the difference of the three species (PsaARCHEO, MbaARCHEO, A369COPE) from the rest in the original alignment (left panel of Fig. 2.3) is an alignment artifact. Of course, if these three species happen to be phylogenetically more closely related to each other than to the rest, then the wrong alignment will in fact be more efficient in recovering the true tree, just as the MP method will be more efficient in recovering the true tree if two sister lineages happen to have long branches. However, as I emphasized before (Xia, 2014), such efficiency is purchased with illegal phylogenetic currency.

A similar situation is shown in the top panel of Figure 2.4 where the alignment from Regier et al. (2010) introduced an alignment artifact increasing the distance between the first pycnogonid species (TorPYCNO for *Tanystylum orbiculare*) and the three other pycnogonid species. The 3-nt deletion in the first sequence (TorPYCNO) is misplaced, with the alignment in the bottom of Figure 2.4 having high alignment scores by any reasonable scoring scheme. The "big data" approach is disastrous for science because authors often do not have enough resources for data validation, neither do reviewers.

```
PamNEOPT    AGAACACGAGUUACCAAA---AUGUUGUGCAU    PamNEOPT    ...AGAACACGAGUUACCAAAAUGUUGUGCAU
MayEPHEM    AGAUCUCGCGUCACCAAA---AUGUUAUGUCA    MayEPHEM    ...AGAUCUCGCGUCACCAAAAUGUUAUGUCA
EinEPHEM    AGAACCAGAGUUACCAAA---AUUUUAUGUAU    EinEPHEM    ...AGAACCAGAGUUACCAAAAUUUUAUGUAU
IveODONAT   AGAAGGACUCUCACUAAA---AUGCUUUGUAU    IveODONAT   ...AGAAGGACUCUCACUAAAAUGCUUUGUAU
LlyODONAT   CGGAGGAAUAUAACUAAG---AUGCUUUGUUU    LlyODONAT   ...CGGAGGAAUAUAACUAAGAUGCUUUGUUU
StuREMI     AGGAAAAGACUUACCAAA---AUGCUGUGUAU    StuREMI     ...AGGAAAAGACUUACCAAAAUGCUGUGUAU
CliZYGEN    AGGACGAGAGUCACUAAA---AUGCUUUGCAU    CliZYGEN    ...AGGACGAGAGUCACUAAAAUGCUUUGCAU
NmeZYGEN    AGAUCAAGGGUCACAAAG---AUGUUGUGUAU    NmeZYGEN    ...AGAUCAAGGGUCACAAAGAUGUUGUGUAU
JapDIPLUR   AGGACGACAGUGACCAAG---CUCCUGUGCCA    JapDIPLUR   ...AGGACGACAGUGACCAAGCUCCUGUGCCA
PsaARCHEO   GCCAGAACAAGAGUAACAAAAAUGCUGUGUAU    PsaARCHEO   GCCAGAACAAGAGUAACAAAAUGCUGUGUAU
MbaARCHEO   GCCAGAACGAGAGUAACAAAAAUGUUGUGUAU    MbaARCHEO   GCCAGAACGAGAGUAACAAAAAUGUUGUGUAU
A369COPE    AGCGUAACCAGGCGGAGCAAGCUGUUGUGCAA    A369COPE    AGCGUAACCAGGCGGAGCAAGCUGUUGUGCAA
DtyMYSTACO  AGGAGAAGGUGCACCAAA---CUACUCUGUCA    DtyMYSTACO  ...AGGAGAAGGUGCACCAAACUACUCUGUCA
NamDIPLO    AGGAAAAGAUUUACAAAA---UUAUUAUGCCA    NamDIPLO    ...AGGAAAAGAUUUACAAAAUUAUUAUGCCA
```

FIGURE 2.3 Poor alignment can distort phylogenetic signals. The left alignment, taken from the supplementary file (nature08742-s2.nex) in Regier et al. (2010), confers undue similarity between the first nine and the last two sequences (DtyMYSTACO and NamDIPLO). An alternative alignment is shown at right, which is better by any alignment criterion.

```
TorPYCNO    GCTGTTTTAGGTAAGGTAGCAGCCGAAAAA---TGGGCTGATGTGGTCATTGCT
AeliPYCNO   TCTATAATAGGAAAAGTTTCT---TCTGAAAAATGGGCAGATGTTGTAATTGCA
AhiPYCNO    GCCGTTACCGGAAAGGTTTCT---TCCGATAAGTGGGCAGATGTTGTCATTGCA
Col2PYCNO   GCAATAATTGGTAAGATTCCA---GATAGCAAGTGGAGTGAAGTTGTCCTTGCA

TorPYCNO    GCTGTTTTAGGTAAGGTAGCAGCCGAAAAATGGGCTGATGTGGTCATTGCT
AeliPYCNO   TCTATAATAGGAAAAGTTTCTTCTGAAAAATGGGCAGATGTTGTAATTGCA
AhiPYCNO    GCCGTTACCGGAAAGGTTTCTTCCGATAAGTGGGCAGATGTTGTCATTGCA
Col2PYCNO   GCAATAATTGGTAAGATTCCAGATAGCAAGTGGAGTGAAGTTGTCCTTGCA
```

FIGURE 2.4 Poor alignment at the top, taken from the supplementary file (nature08742-s2.nex) in Regier et al. (2010), unnecessarily increase the distance between TorPYCNO and the three other Pycnogonid species, with the improved alignment at the bottom.

2.1.3 POOR ALIGNMENT LEADS TO UNNECESSARY LOSS OF PHYLOGENETIC SIGNALS

Because of the poor alignment illustrated above, some parts in the MSA were deemed unalignable by Regier et al. (2010) and removed from the translated amino acid sequences for phylogenetic analysis based on amino acid sequences. For example, the shaded segment in Figure 2.5a was deleted. This deletion is unnecessary because sequence homology is identifiable as shown in Figure 2.5b. Deleting phylogenetically significant signals reduces the phylogenetic resolution. However, the deletion of unalignable segments by Regier et al. (2010) is not consistent. While the shaded segment in Figure 2.5a is deleted, the undesirable alignment in Figure 2.2a remains in their degenerated sequence file (nature08742-s3Degen1.nex) used to generate their main phylogenetic results in their Figure 1.

I finally wish to make two points. First, the data set with the many alignment problems is still often incorporated into still larger data set without realignment, a common practice in today's phylogenetic reconstruction that erodes the credibility of published phylogenetic results and degrades this branch of science that used to be more rigorous. Second, scientists are deprived of the responsibility of being the custodians of their own science by academic journals, so fewer and fewer scientists really care much about their academic home. What matters today is to create a data set that is big, really big, and so big that reviewers will never be able to find time to check the quality of the data or details of the analysis.

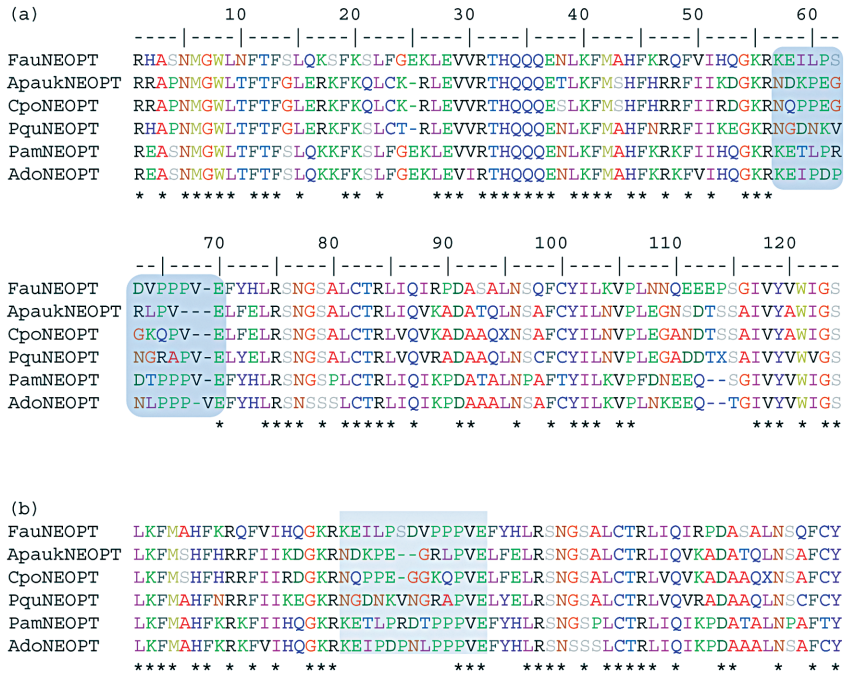


FIGURE 2.5 Unnecessary deletion of phylogenetically informative data. (a) Partial amino acid sequences translated from the codon sequences in the supplementary file (nature08742-s2.nex) in Regier et al. (2010). The shaded segments, including the amino acid Eat labeled site 70, were deemed by Regier et al. (2010) as unalignable and removed in the final amino acid sequence alignment for phylogenetic analysis. (b) Realigned sequences.

2.2 PAIRWISE ALIGNMENT

Given two strings S ($=s_1s_2...s_n$) and T ($=t_1t_2...t_m$), a pairwise alignment of S and T is defined as an ordered set of pairings of (s_i, t_j) and of gaps (s_i,−) and (−,t_j), with the constraint that the alignment is reduced to the two original strings when all gaps in the alignment are deleted. A prefix of S, specified here as S_i, is a substring of S equal to $s_1s_2...s_i$, where $i \leq n$. Figure 2.6 shows two different alignments from the same set of two sequences.

```
Alignment 1: ACCCAGGGCTTA
             ||||  ||   |
             ACCCGGGCTTAG

Alignment 2: ACCCAGGGCTTA-
             ||||  |||||||
             ACCC-GGGCTTAG
```

FIGURE 2.6 Two sequences in two different alignments implying different homology sites, e.g., A and G at the 5th site in the Alignment 1 is assumed to be homologous but the same A and G are not homologous in Alignment 2.

An optimal alignment is operationally defined as the pairwise alignment with the highest alignment score for a given scoring scheme. For this reason, an optimal alignment is meaningless without specifying the scoring scheme. A scoring scheme has two components. One is the score for the two matching characters, e.g., we may give 2 to a match nucleotide pair, e.g., A/A in the first site of the two sequences and −1 to a mismatched nucleotide pair, e.g., A and G at 5th site in Alignment 1. Thus, for a match score of 2 and a mismatch score of −1, Alignment 1, with 7 matches and 5 mismatches, would have an alignment score of $7\times2 + 5\times(-1) = 9$. The other component of a scoring scheme is gap penalty, which we need in order to obtain an alignment score for Alignment 2 in Figure 2.6. Suppose we take the simplest approach with constant gap penalty and penalize a gap with −2. With the previous match score of 2, mismatch score of −1, and a constant gap penalty of −2, the alignment score for Alignment 2, which has 11 matches, 0 mismatch, and 2 gaps, is $11\times2 + 0\times(-1) + 2\times(-2) = 18$. Thus, with the given scoring scheme, Alignment 2 is better than Alignment 1. Note that which alignment is better depends on scoring scheme. The scoring scheme we used favors Alignment 2 against Alignment 1. However, if we have a scoring scheme that penalizes gaps heavily, e.g., −7 for a gap, then Alignment 1 will have higher alignment score than Alignment 2. Therefore, when we use a criterion for making a choice, we often need to justify our criterion.

Alignment by dynamic programming guarantees that for a given scoring scheme the resulting alignment has the highest alignment score, or one of the highest alignment scores when there are equally optimal alignments. We will first illustrate the global pairwise alignment (Needleman and Wunsch, 1970) followed by a brief outline of the differences between global and local pairwise alignment (Smith and Waterman, 1981). Local sequence alignment is for searching local similarities between sequences, e.g., homeobox genes which are not similar globally but all share a very similar homeodomain motif.

Here we will first learn a simple dynamic programming algorithm for pairwise alignment using a simple scoring scheme with constant gap penalty. The simple scoring scheme is then extended in two ways, first by introducing a similarity matrix to replace match and mismatch scores and second by introducing the affine function to better approximate the origin of the insertion and deletion during sequence evolution. My experience is that an average student can understand pairwise alignment with constant gap penalty but only a very good student can understand the two extensions.

2.2.1 GLOBAL ALIGNMENT WITH CONSTANT GAP PENALTY

Suppose we want to align two sequences S and T with S = ACGT and T = ACGGCT. Practical sequence alignment typically involves sequences that are much longer, but the computation is the same. If you learned how to align these two short sequences, you know how to align sequences of any lengths.

Dynamic programming for sequence alignment needs a scoring scheme. We will use a simple one with a constant gap penalty (G) of −2, a match score (s_{ii}, where the subscript 'ii' indicates two identical nucleotides) of 2 and a mismatch score (s_{ij}, where the subscript ij indicates two different nucleotides) of −1. Global alignment with the dynamic programming approach is illustrated numerically in Figure 2.7. One of the two sequences occupies the top row and will be referred to hereafter as the row sequence (sequence S in our example). The other sequence occupies the first column and will be referred to hereafter as the column sequence (sequence T in our example).

We need to fill in two matrices. The first is the scoring matrix (SM) to obtain the alignment score, with the dimensions (n+1, m+1). A value in row i and column j in the scoring matrix ($SM_{i,j}$) is the alignment score between prefixes S_j and T_i. The second is the backtrack matrix needed to obtain the actual alignment, with the dimensions (n,m). In Figure 2.7A, the two matrices are superimposed, with the scoring matrix being the numbers and the backtrack matrix being made of arrows. The backtrack matrix is sometimes called the traceback matrix.

The first row ($SM_{0,j}$) and the first column ($SM_{i,0}$) of the scoring matrix is filled with $i \times G$ (where i = 0, 1, ..., n) and $j \times G$ (where j = 0, 1, ..., m), respectively. They represent consecutive insertion of gaps. For example,

$SM_{0,4} = -8$ (Fig. 2.7) implies the alignment of S against four consecutive gaps, so you get an alignment score of -8 (with gap penalty of −2). Similarly, $SM_{6,0} = -12$ (Fig. 2.7) implies the alignment of T with six consecutive gaps.

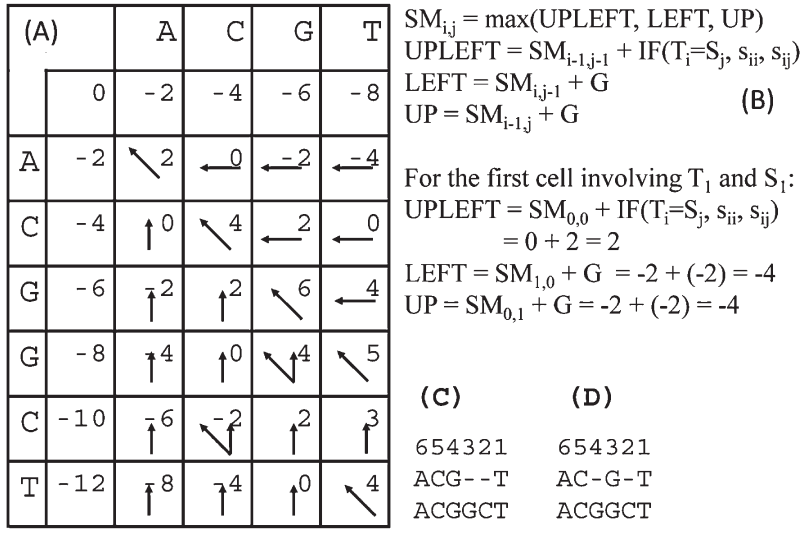


FIGURE 2.7 Aligning sequences S (Top row) and T (left column) by dynamic programming, with the score and the backtrack matrices superimposed (A). The scoring scheme has a match score of 2, mismatch score of −1 and constant gap penalty of −2. Other than the first row and first column, each cell involves computing three values and filling the cell by the maximum of the three (B). The backtrack matrix, made of all arrows in (A), is for obtaining sequence alignments (C and D), where the numbers in the first row show the order of obtaining the alignment site by site from the last to the first by backtracking.

For each of the other $SM_{i,j}$ values, we need to compute three values designated as UPLEFT, LEFT, and UP specified in Figure 2.7B. The first cell corresponds to T_1 and S_1 (Fig. 2.7B), both being A, with its UPLEFT, LEFT, and UP values being 2, –4, and –4, respectively. The maximum of these three values is UPLEFT (=2), so $SM_{1,1} = 2$, and we put an upleft arrow in the cell (Fig. 2.7). If LEFT (or UP) happened to be the maximum of the three, we would have put the LEFT (or UP) value in the cell and add a left-pointing (or up-pointing) arrow in the corresponding cell in the backtrack matrix.

The calculation of $SM_{1,2}$ is illustrated in Figure 2.8, with the maximum of the three values (UPLEFT, LEFT, UP) being 0 (Fig. 2.8A). So, we put