

Contents

Chapter 1	Why Phylogenomics Matters	1
Chapter 2	The Biology of Linear Molecules: DNA and Proteins	15
Chapter 3	Evolutionary Principles: Populations and Trees	35
Chapter 4	Databases	57
Chapter 5	Homology and Pairwise Alignment	71
Chapter 6	Multiple Alignments and Construction of Phylogenomic Matrices	93
Chapter 7	Genome Sequencing and Annotation	103
Chapter 8	Tree Building	117
Chapter 9	Robustness and Rate Heterogeneity in Phylogenomics	147
Chapter 10	A Beginner's Guide to Bayesian Approaches in Evolution	163
Chapter 11	Incongruence	185
Chapter 12	Adapting Population Genetics to Genomics	201
Chapter 13	Detecting Natural Selection: The Basics	219
Chapter 14	Refining the Approach to Natural Selection at the Molecular Level	233
Chapter 15	Genome-Level Approaches in Population Genetics	247
Chapter 16	Genome Content Analysis	263
Chapter 17	Phylogenomic Perspective of Biological Diversity: Tree of Life, DNA Barcoding, and Metagenomics	277
Chapter 18	Microarrays in Evolutionary Studies and Functional Phylogenomics	303
Index		325

Detailed Contents

Chapter 1 Why Phylogenomics Matters

Phylogenomics and Bioinformatics

- A microarray is a simple concept with powerful applications
- The Human Genome Project was a watershed event in DNA sequencing
- Bioinformatics tools enable data analysis and identification of patterns in biological experiments

The Rise of Phylogenomics

- Functional phylogenomics employs common ancestry to infer protein function
- Pattern phylogenomics gathers information from branching patterns of a group of organisms

The Phylogenomic Toolbox

- Sequence alignment programs and databases are essential computational tools in phylogenomics
- Statistical analysis enables comparison of biological sequences and generation of phylogenetic trees
- Parametric statistics are derived from the parameters of a given statistical distribution
- Nonparametric statistical analysis relies on resampling techniques to determine significance
- Maximum likelihood and Bayesian analysis are additional statistical methods used in phylogenomics

Key Attributes of Phylogenomicists

Summary

Discussion Questions

Further Reading

Chapter 2 The Biology of Linear Molecules: DNA and Proteins

Nucleic Acids

- DNA is a perfect molecule for transmitting information
- DNA is synthesized by specific pairing
- DNA can mutate and impart heritable information important in understanding descent with modification

Polymerase chain reaction is a milestone development

Proteins

- Proteins are linear polymers of amino acids
- Proteins have multiple levels of structure
- Translation of the information in DNA is accomplished by the genetic code
- A single nucleic acid sequence has multiple reading frames

The DNA Data Explosion

- Sequencing methods for linear molecules have become more powerful
- Next-generation sequencing allows small genomes to be sequenced in a day
- Next-generation sequencing leads to practical applications

Alternatives to Whole Genome Sequencing: Detecting Variation in Populations

- Single-nucleotide polymorphisms differ at one position
- Methods that take advantage of DNA hybridization can be used to examine single-nucleotide polymorphisms

Analyzing Gene Expression

- Northern blots assess RNA production on a small scale
- Microarrays allow an entire transcriptome to be examined
- Expressed sequence tags use reverse transcription to study RNA production

Summary

Discussion Questions

Further Reading

Chapter 3 Evolutionary Principles: Populations and Trees

Darwin and Evolutionary Theory

- Four major contributions can be credited to Darwin
- Darwin's research lacked a valid genetic or hereditary component

Evolution is now divided into microevolution and macroevolution	36	Discussion Questions	69
Microevolution	37	Further Reading	69
Microevolution is studied by population genetics	37	Chapter 5 Homology and Pairwise Alignment	71
Advances in molecular techniques led to new thinking in microevolution	37	Homology of Genes, Genomic Regions, and Proteins	71
Knowledge of codon changes and usage can provide evolutionary insights	38	Genomes can diverge by speciation and by duplication	71
Modern microevolutionary study relies on computational approaches and models	39	Alignment is a proxy for homology at the sequence level	72
Macroevolution	40	Nucleic acid sequence alignments can be evaluated manually	73
Macroevolution is studied by systematics	40	A paired protein alignment can be evaluated manually	75
The key to systematics is hierarchical arrangements	41	Dynamic Programming and Sequence Alignment	76
Competing philosophies and phylogenies form the basis of phylogenetics	41	Initialization sets up the matrix for alignment accounting	76
Modern phylogenetics is dominated by trees and tree thinking	42	Filling of the matrix is guided by preset rules	77
Modern phylogenetics aims to establish homology	43	Traceback uses the filled matrix to obtain the alignment	80
Species	46	Database Searching via Pairwise Alignments: The Basic Local Alignment Search Tool	82
The definition of species is a subject of debate	46	BLAST can identify inexact matches of high similarity	83
A simple solution is to define species phylogenetically	48	BLAST is optimized for searching large databases	83
Modern Challenges to Darwinian Evolution	50	There are several variations of BLAST for nucleotide and amino acid sequences	86
Punctuated equilibrium suggests that not all evolution is gradual	50	Performing a BLAST search is straightforward	86
Epigenetic changes caused by outside influences can be inherited	51	Whole genome alignments can also be performed	90
Markov chain models	51	Summary	91
Summary	54	Discussion Questions	91
Discussion Questions	55	Further Reading	92
Further Reading	56	Chapter 6 Multiple Alignments and Construction of Phylogenomic Matrices	93
Chapter 4 Databases	57	Multiple Sequence Alignment	93
Databases and Phylogenomics	57	Changing Alignment Parameters	97
DNA sequences are stored in large international databases	58	Multiple optimal alignments may exist	97
Specific data sets may be held in special repositories	58	Culling and elision are ways to explore the alignment space	98
Databases offer free access and availability for scientific inquiry	59	Choosing an Alignment Program	99
Information Retrieval from the NCBI Database	59	Automated alignment results are frequently adjusted "by eye"	99
Publications are archived in the PubMed database	60	Alignment programs can be compared by use of benchmark data sets	100
Sequences are collected in the GenBank database	60	Dynamic Versus Static Alignment	101
Whole genomes are collected on the Genome Page	66		
Other databases such as Online Mendelian Inheritance in Man (OMIM) can be used to obtain information	68		
Summary	69		

Summary	101	Rescoring is a simple way to weight characters in parsimony	129
Discussion Questions	102		
Further Reading	102		
Chapter 7 Genome Sequencing and Annotation	103		
Genome Sequencing	103		
Small viral genomes were the first to be sequenced	103		
Bacterial artificial chromosome-based sequencing employs a "divide and conquer" strategy for larger genomes	103		
The Human Genome Project	104		
Statistical Aspects of DNA Sequencing	105		
Lander-Waterman statistics tell how many sequencing reads are needed to cover a genome	105		
Sequence Assembly	106		
Next-Generation Sequencing	107		
Gene Finding and Annotation	109		
Gene finding can be accomplished via extrinsic, <i>ab initio</i> , and comparative approaches	109		
Gene functional annotation helps to determine gene function	110		
Gene ontology facilitates the comparison of genes	111		
A Phylo-View of Genomes Sequenced to Date	113		
Summary	114		
Discussion Questions	115		
Further Reading	116		
Chapter 8 Tree Building	117		
Distances, Characters, Algorithms, and Optimization	117		
The number of trees grows with each additional taxon	118		
Trees can be rooted by several methods	119		
Characters and Weighting	120		
Character states in molecular data may include the presence of genes and the sequence of nucleotides or amino acids	120		
Some discrete and numerical character states are ordered	121		
Characters can be weighted relative to one another	121		
Which characters should be used?	122		
A matrix of DNA sequences is used to illustrate different tree-building approaches	123		
Basics of Maximum Parsimony Analysis	124		
Fitch's algorithm uses set theory	125		
		Distance Methods	130
		Corrections for multiple hits may be introduced	131
		Corrections can be made by using evolutionary models	132
		Neighbor joining is a stepwise-based approach to tree building	134
		Minimum evolution uses minimal distance as a criterion to choose the best solution among multiple trees	134
		Basics of the Maximum Likelihood Approach	138
		Transformation and Probability Matrices	140
		Transformation cost matrices address the probability of character state changes in parsimony analysis	140
		Generalized probability matrices incorporate probabilities rather than integral weighting values in likelihood analysis	143
		Summary	144
		Discussion Questions	145
		Further Reading	145
		Chapter 9 Robustness and Rate Heterogeneity in Phylogenomics	147
		So Many Trees, So Little Time	147
		Tree space allows trees to be grouped by optimality	147
		Selection of a starting tree is the first challenge	149
		Peaks in tree space can be reached by branch swapping	149
		Moving from local optimality peaks to the true optimal tree is the most critical step in exploring tree space	150
		Rate Heterogeneity	151
		Rate heterogeneity can be included in likelihood models by use of a γ distribution	152
		The likelihood ratio test can be used to determine which models are more likely or probable than others	154
		Several programs can rapidly compare models	154
		Simple Metrics as Measures of Consistency in a Parsimony Analysis	155
		Tree length evaluates the degree of parsimony within a given solution	156
		Consistency index estimates convergence in a data set	156
		Retention index assesses degree of agreement with the maximum parsimony tree	157
		Rescaled consistency index addresses some shortcomings of retention and consistency indices	158

Decay index or Bremer index estimates robustness of a node	158	Character congruence uses total evidence or concatenation to create supermatrices	188
Determining the Robustness of Nodes on a Phylogenetic Tree by Resampling Techniques	158	Statistical methods assess incongruence to decide whether information should be concatenated	188
Bootstrapping analyzes a phylogenetic matrix by resampling with replacement	159	The incongruence length difference test uses parsimony to determine whether two gene partitions are incongruent	189
Jackknifing analyzes a phylogenetic matrix by resampling without replacement	160	Likelihood ratio tests compare likelihoods to determine whether two gene partitions are incongruent	191
Parametric bootstrapping applies a distribution model to the data	161	Fork indices provide measures of tree similarity	191
Summary	162	The Gene Tree/Species Tree Problem	191
Discussion Questions	162	Phylogeography adds a new dimension to population-level analysis	193
Further Reading	162	Lineage sorting was first observed in mtDNA	194
Chapter 10 A Beginner's Guide to Bayesian Approaches in Evolution	163	Three examples illustrate lineage sorting in studies of closely related taxa	194
Bayesian Inference	163	Genome-level examples of lineage sorting have also been documented	195
Generating a distribution of trees is an important application of the Bayesian approach	164	Coalescence offers a partial solution to the gene tree/species tree problem	196
MCMC is critical to the success of Bayesian analysis	167	Horizontal Transfer	197
Bayesian Analysis in a Phylogenetic Context	169	Programs That Consider Nonvertical Evolution and Lineage Sorting to Infer Phylogeny	197
Model selection can be utilized on any biologically meaningful partition	169	Coalescence programs use both gene trees and species trees as input	197
Selection of priors may involve default values, but priors can be adjusted manually	170	Programs that consider horizontal gene transfer generate nets and webs	198
Setting the MCMC parameters has greater impact for larger data sets	170	Summary	198
Increasing ngen in MCMC leads to better distribution of trees but at increased computational cost	172	Discussion Questions	199
An Example Using MrBayes	173	Further Reading	199
MrBayes can be run with default settings	174	Chapter 12 Adapting Population Genetics to Genomics	201
The MrBayes run can be interpreted by use of several parameters	176	Modernizing Population Genetics by Use of High-Throughput Methods	201
Models can be changed in MrBayes	177	Kimura and Lewontin contributed important new ways to think about genes in nature	202
Priors can be changed in MrBayes	179	The Hardy-Weinberg theorem has been extended in modern population genetics	202
Ending a Bayesian analysis involves an assessment of the run's efficiency	180	DNA Variation among Individuals	202
Summary	183	Single-nucleotide polymorphisms can be used to differentiate members of the same species	203
Discussion Questions	183	Microsatellites provide another analytical tool for species where SNPs are less abundant	204
Further Reading	183	Extending Fundamental Population Genetics	206
Chapter 11 Incongruence	185	Tajima's <i>D</i> distinguishes between sequences evolving neutrally and those evolving non-neutrally using allele frequencies	207
Incongruence of Trees	185		
Taxonomic congruence achieves consensus by construction of supertrees	187		

F statistics measure the degree of isolation of entities	208	Expanding the Search for Natural Selection at the Molecular Level	234
There are two approaches to estimating population-level statistics	210	Statistical tests of significance are required at various levels	234
FST and related measures have four major uses in evolutionary biology	211	Natural selection is variable across protein components and across time	236
Imputation	212	Examples of nonuniformity are seen in <i>Drosophila</i> and in the <i>BRCA1</i> gene	236
Population-Level Techniques: Mismatch Distribution Analysis and STRUCTURE Analysis	212	Maximum likelihood approaches are implemented in selection studies at the molecular level	238
Mismatch distribution analysis compares haplotype data of populations	212	Statistical tests are used to choose the best model for estimating dN and dS	239
STRUCTURE analysis reveals substructure and genetic cross talk	214	There are two caveats about detecting selection at the molecular level	240
Summary	215	Codon Selection Bias	241
Discussion Questions	215	Codon selection bias can be calculated manually or by various analytical methods	241
Further Reading	217	Codon usage bias usually occurs in cellular housekeeping genes and varies among species	243
Chapter 13 Detecting Natural Selection: The Basics	219	Summary	245
Analyzing DNA Sequences for Natural Selection	219	Discussion Questions	245
DNA sequences can be examined for silent and replacement changes	219	Further Reading	246
Several variables affect the detection of natural selection at the genomic level	221	Chapter 15 Genome-Level Approaches in Population Genetics	247
Approximate Methods of Determining dN/dS	223	Genome-Wide Association Studies	247
Basic dN and dS calculations begin with counting the observed number of changes	224	A simple example illustrates the association technique	247
Scaling for redundancy and getting the number of potential substitutions is necessary for determining dN/dS	224	The National Human Genomics Research Institute maintains a database of genome-wide association studies	249
Pathways of codon change are an important element in calculating dN/dS	226	Role of the Coalescent in Population Genetics	249
Codon change pathways can be used to account for redundancy	228	The coalescent addresses the time for an allele to coalesce and the variation in populations under drift	250
Summary	230	The coalescent in practice explores a large number and broadly representative sample of plausible genealogical scenarios	251
Discussion Questions	230	High-quality DNA sequence data from a random sample constitute the best input for a coalescence analysis	251
Further Reading	231	Importance sampling and correlated sampling are used to generate a collection of simulated genealogies	252
Chapter 14 Refining the Approach to Natural Selection at the Molecular Level	233	Programs for coalescence analysis include BEAST and LamarC	252
Accounting for Multiple Hits in DNA Sequences for dN/dS Measures	233	Genetic Hitchhiking and Selective Sweeps	252
The Jukes-Cantor conversion corrects for multiple hits	233	Selective sweeps are detected in four basic ways	253
Estimating natural selection requires adjusting the calculation of sequence changes	234	Empirical examples of selective sweeps include boxers, flies, and humans	255

Hard and soft sweeps produce different effects in the genome	256	Yeast and <i>Drosophila</i> represent examples of whole genome approaches	284
Genome-Wide Scans to Address Population Genetic and Evolutionary Questions	257	Specific Problems Caused by Bacterial Phylogenomics	286
Phylogenetic shadowing identifies regulatory elements in DNA sequences	259	Does a tree of life really exist for bacteria?	287
Regions of the human genome experience accelerated evolution	260	Can all genes be used in a concatenated analysis?	287
Regions that are both strongly conserved and rapidly deleted are of interest	260	Tree-Building Methods	288
Summary	261	Search-based innovations avoid the problem of being trapped in a suboptimal solution	288
Discussion Questions	261	Coalescence can partially solve the problem of incongruence	289
Further Reading	262	The supertree grafting approach subdivides a problem into a series of smaller nonoverlapping problems with less taxa	289
Chapter 16 Genome Content Analysis	263	The matrix representation approach recodes the information in multiple trees	290
Genome Biology Databases	263	The supertree divide-and-conquer approach subdivides a problem into a series of smaller overlapping problems by using many analyses with less taxa	291
Comparative Genomics Approaches	264	DNA Barcoding	291
Venn diagrams, EDGAR, and Sungear visualize the overlap of genes from two or more genomes	264	Taxonomy and species involve the study of species delimitation	292
Generating Genome Content Phylogenies	266	DNA taxonomy is very different from classical taxonomy	292
Clusters of orthologous groups is a method that enables identification of orthologs of genes across multiple species	267	Character-based or distance-based approaches to DNA barcoding result in identification of species	294
Single linkage clustering compares genes in a cross-species context based on sequence	268	Is there enough information in a single gene to do DNA barcoding?	296
A presence/absence matrix is constructed via single linkage clustering	268	Potential new species are flagged by DNA barcoding	297
Genome content analysis was first demonstrated for bacterial genomes	272	Metagenomics	297
Genome content analysis is currently performed by computer software packages	273	Summary	300
Summary	274	Discussion Questions	300
Discussion Questions	275	Further Reading	301
Further Reading	275	Chapter 18 Microarrays in Evolutionary Studies and Functional Phylogenomics	303
Chapter 17 Phylogenomic Perspective of Biological Diversity: Tree of Life, DNA Barcoding, and Metagenomics	277	Transcription-Based Approaches	303
A Hierarchy of Sequencing Approaches for Biodiversity Studies	277	Microarrays resemble high-throughput Southern and northern blots	304
In shallow targeted sequencing, a small number of interesting genes are sequenced in as many taxa as feasible	278	Microarrays are used for class comparison, prediction, and discovery	305
Expressed sequence tag approaches use expressed genes to amass a collection of gene sequences for analysis	281	Microarray intensity data are transformed for use in dendrograms and other techniques	306
		Next-generation approaches are applied to transcription analysis	309
		Microarrays are useful in evolutionary and phylogenomic studies	310

Corrections are needed in cross-species microarray studies	311
Protein-Protein Interactions	311
Various approaches are used to examine protein-protein interactions	311
Mendelian phenotypes in humans and model organisms are studied by Web-based approaches	312
Phylogenetic Approaches to Understanding Function	312
Phylogenomic gene partitioning is used to explore function	313
A gene presence/absence matrix was employed to compare 12 <i>Drosophila</i> genomes	316
Expressed sequence tags and phylogeny can be used to study plant function	320
Gene clustering in <i>Caenorhabditis elegans</i> was determined from RNA interference phenotype	321
Summary	322
Discussion Questions	323
Further Reading	323
Index	325