

Contents

1 Basics in Molecular Evolution	1
1.1 Evolutionary distance of DNA sequences	1
1.1.1 Jukes and Cantor's model: a tutorial	1
1.1.2 Models of nucleotide substitution	3
1.1.3 One-parameter method	6
1.1.4 Kimura's two-parameter method	6
1.1.5 The general stationary and time-reversible model	7
1.1.6 Estimation of d under variable rates	8
1.1.7 The LogDet distance	9
1.2 Evolutionary distance between protein-encoding sequences	12
1.2.1 Poisson distance of protein sequence	12
1.2.2 Amino acid substitution matrix	12
1.2.3 Synonymous and nonsynonymous distances	13
1.3 Phylogenetics trees: an overview	15
1.4 Distance method for phylogenetic inference	16
1.4.1 Principle: minimum-evolution (ME)	16
1.4.2 Algorithm: neighbor-joining (NJ) method	16
1.4.3 Four-point condition and NJ algorithm	18
1.4.4 The Q -score of Studier and Keppler	19
1.5 Parsimony methods for phylogenetic inference	19
1.6 Maximum-likelihood (ML) methods for phylogenetic inference	21
1.6.1 Likelihood function	21
1.6.2 Time-reversibility and the root problem	22
1.6.3 Search strategies for ML trees	22
1.7 Bayesian methods for phylogenetic inference	23
1.8 Ancestral sequence inference	24
1.8.1 The maximum parsimony approach	24
1.8.2 The probabilistic (Bayesian) approach	25
1.8.3 Deletions and insertions	27
1.9 Rate variation among sites	28
1.9.1 Number of substitutions at a site	29
1.9.2 Estimation of α	30
2 Basics in Bioinformatics and Statistics	33
2.1 Bioinformatic resources for evolutionary genomics	33
2.2 Basic statistics for homologous search	36
2.3 Sequence alignment	37
2.3.1 Pairwise alignment	38
2.3.2 Multiple alignment with a guide tree: Clustal	39

2.4	Microarrays and statistics	40
2.4.1	Types of microarray data	40
2.4.2	Sources of noises	41
2.4.3	Multiple gene problem	41
2.4.4	False discovery rate (FDR)	42
2.4.5	ANOVA analysis of many genes	44
2.5	Markov chain Monte Carlo (MCMC)	44
2.5.1	Metropolis Hastings algorithm	45
2.5.2	Calculation of posterior distribution	45
3	Functional Divergence after Gene Duplication:	
	Statistical Modeling	47
3.1	Modeling functional divergence	47
3.2	Poisson model for type-I functional divergence	49
3.2.1	Two-state model	49
3.2.2	The Poisson-gamma model for protein sequence evolution	50
3.2.3	The likelihood function	51
3.2.4	Maximum-likelihood estimation (MLE)	52
3.2.5	Predicting critical amino acid residues	54
3.2.6	Reduced rate correlation between duplicate genes: an alternative view of θ_f	57
3.3	Markov chain model for type-I functional divergence	58
3.3.1	The Markov chain model	58
3.3.2	Case study: COX (cyclooxygenase) gene family	61
3.3.3	Comparisons between the Poisson model and the Markov chain model	63
3.4	Statistical method for type-II functional divergence	63
3.4.1	Modeling type-II functional divergence	63
3.4.2	Two clusters by gene duplication	65
3.4.3	Poisson model in the late-stage	65
3.4.4	Maximum-likelihood estimation	66
3.4.5	Predicting critical amino acid residues: empirical Bayesian approach	67
3.5	A unifying model for type-I and -II functional divergences	69
4	Functional Divergence after Gene Duplication:	
	Applications and Others	73
4.1	DIVERGE-based analysis	73
4.1.1	Functional-structural basis of shifted evolutionary rates between caspases	73
4.1.2	Pseudokinase domain in Jak protein kinase is functional	77
4.1.3	Pattern of type-II functional divergence	82
4.2	Functional distance analysis	84
4.2.1	Distance of functional divergence	84
4.2.2	Three-cluster analysis	86
4.2.3	Examples: vertebrate developmental gene families	87

4.3	Other methods for type-I functional divergence	90
4.3.1	Knudsen–Miyamoto method	90
4.3.2	Gaucher–Miyamoto–Benner method	91
4.3.3	Codon-based methods	91
4.3.4	Heterotachy model	92
4.3.5	The alpha shift measure (ASM) method	93
4.3.6	Nam <i>et al.</i> 's method for detecting functional divergence of protein domains	93
5	Phylogenomic Expression Analysis between Duplicate Genes	95
5.1	Brownian-related stochastic model	95
5.1.1	Expression likelihood under phylogeny	95
5.1.2	Method of expression distance	100
5.2	Ancestral gene expression inference	101
5.3	Oakley <i>et al.</i> 's model	103
5.4	Expression divergence under stabilizing selection: the Ornstein–Uhlenback (OU) model	105
5.5	Likelihood and distance methods under experimental correlations	106
5.6	Yeast GlnS gene family: an example	107
5.7	Estimating expression divergence based on massively parallel sequencing technology	109
5.7.1	The Poisson-lognormal model	109
5.7.2	U -distance for expression divergence	110
6	Expression between Duplicate Genes: Genome-Wide Analysis	113
6.1	Coding sequence divergence vs expression divergence	113
6.2	Regulatory motif divergence vs expression divergence between duplicates	114
6.3	Gene duplication and expression diversification	115
6.4	Expression divergence and retention of duplicate genes	116
6.5	Evolutionary distance of expression divergence	118
6.6	Rate of expression divergence between yeast duplicate genes	120
6.7	Asymmetric expression evolution after gene duplications	123
6.8	Concluding remarks	125
7	Tissue-Driven Hypothesis of Genomic Evolution	127
7.1	Tissue-driven hypothesis of genomic evolution	127
7.1.1	Expression divergence under stabilizing model	127
7.1.2	Tissue-dependent rate of protein evolution	129
7.1.3	Tissue-driven hypothesis	130
7.2	Testing the tissue-driven hypothesis	132
7.2.1	Estimation of genomic distances	132
7.2.2	Tissue expression divergence between human and mouse	133
7.2.3	Correlation ($E_{ti} - D_{ti}$) between tissue expression and sequence divergence	134

7.2.4	Tissue correlation ($E_{ti} - D_{ti}$) between inter-species and duplicate expression divergence	135
7.2.5	Evolutionary rate of protein sequence under multiple tissue constraints	135
7.2.6	Some comments	136
7.3	Compound-Poisson model of expression evolution	137
7.4	Expression shifts in the human brain	138
7.4.1	Evolution of the human brain	138
7.4.2	Enard <i>et al.</i> 's analysis	138
7.4.3	Gu and Gu's analysis	138
7.4.4	Concluding comments	141
8	Gene Pleiotropy and Evolution of Protein Sequence	143
8.1	Model for protein sequence evolution	143
8.1.1	Fisher's model and molecular phenotypes	143
8.1.2	Stabilizing selection	146
8.1.3	Micro-adaptation	146
8.1.4	Distribution of mutational effects	147
8.1.5	S -distribution	148
8.2	Selection intensity and model classification	148
8.2.1	Mean of selection intensity	148
8.2.2	Model classification	150
8.3	Evolutionary rate of protein sequences	151
8.3.1	General formula	151
8.3.2	K -mode and B -mode for the rate of protein evolution	153
8.3.3	Effect of B_i -variation on the evolutionary rate	154
8.4	Estimation of gene pleiotropy and selection intensity	155
8.4.1	The second-moment of evolutionary rate	155
8.4.2	Effective gene pleiotropy (K_e)	155
8.4.3	Estimation pipeline	157
8.4.4	Effective selection intensity	157
8.4.5	Bias-corrected estimation of effectively gene pleiotropy	158
8.5	Preliminary analysis of gene pleiotropy	159
8.5.1	Extent of gene pleiotropy	160
8.5.2	Biological relevance	160
8.6	Comments on gene pleiotropy	165
9	Modeling the Genomic Evolution of Gene Contents	167
9.1	The birth-death model of gene content evolution	167
9.1.1	Joint size distribution of gene families	167
9.1.2	Genome distances and gene content information	169
9.1.3	Extended gene content and genome distance estimation	170
9.1.4	Simulations and case study	172
9.2	Likelihood of four genomes under simple gene contents	174
9.2.1	Likelihood function: case A	175

9.2.2	Likelihood function: case B	179
9.2.3	Likelihood function	180
9.3	Birth-death model with lateral gene transfer	182
9.3.1	General birth-death process considering LGT	182
9.3.2	Extended gene content under LGT	183
9.3.3	Simple gene content and LGT	184
9.3.4	Comments	185
9.4	Other Models	185
9.4.1	Blocks model	185
9.4.2	Equal birth-death rate model	186
9.4.3	Constant-birth, proportional-death model	186
10	Advanced Topics in Systems Biology and Network Evolution	187
10.1	GC mutational bias rather than adaptation driving tyrosine loss in metazoan genome evolution	188
10.2	Contribution of duplicate genes to genetic robustness	190
10.2.1	Functional compensation between duplicate genes	191
10.2.2	When an essential gene is duplicated	191
10.2.3	When a dispensable gene is duplicated	194
10.2.4	Hypothesis: duplication of dispensable genes to maintain genetic buffering	194
10.3	Evolution of gene-gene interactions	196
10.3.1	p -Value representation of gene-gene interaction	196
10.3.2	General framework	197
10.3.3	Estimation of γ_{11} , γ_{10} , γ_{01} , and γ_{00}	198
10.3.4	Comments	199
10.4	Origin of modularity and complexity	200
10.4.1	Some backgrounds	200
10.4.2	Scale-free network and modularity	201
10.4.3	Origin of modularity in a scale-free network	203
10.4.4	Protein-protein interaction data analysis	207
10.4.5	Hypothesis: random loss of interactions may shape modularity in a complex gene network	208
10.5	Network motif analysis and yeast genome duplication	211
10.6	Evolutionary kinetic (EK) analysis of duplicate genes	216
10.6.1	Reprogramming in duplicate backup circuits	216
10.6.2	Responsive backup circuits (RBC) and regulatory designs	217
10.6.3	Expression-triggered backup circuit hypothesis	219
	References	223
	Index	251