

---

## Contents

---

### Part I Molecules: Proteins and RNA

---

<b>1 Modeling Conformational Flexibility and Evolution of Structure: RNA as an Example</b>	
<i>P. Schuster and P.F. Stadler</i> .....	3
1.1 Definition and Computation of RNA Structures .....	3
1.1.1 RNA Secondary Structures .....	4
1.1.2 Compatibility of Sequences and Structures .....	8
1.1.3 Sequence Space, Shape Space, and Conformation Space .....	11
1.1.4 Computation of RNA Secondary Structures .....	14
1.1.5 Mapping Sequences into Structures .....	15
1.1.6 Suboptimal Structures and Partition Functions .....	18
1.2 Design of RNA Structures .....	19
1.2.1 Inverse Folding .....	19
1.2.2 Multiconformational RNAs .....	20
1.2.3 Riboswitches .....	22
1.3 Processes in Conformation, Sequence, and Shape Space .....	23
1.3.1 Kinetic Folding .....	23
1.3.2 Evolutionary Optimization .....	25
1.3.3 Evolution of Noncoding RNAs .....	30
References .....	32
<b>2 Gene3D and Understanding Proteome Evolution</b>	
<i>J.G. Ranea, C. Yeats, R. Marsden, and C. Orengo</i> .....	37
2.1 Protein Family Clustering .....	42
2.1.1 SYSTERS .....	42
2.1.2 ProtoNet .....	42
2.1.3 ADDA .....	42
2.1.4 ProDom .....	43

2.2	The PFscape Method .....	43
2.3	The NewFams .....	44
2.4	Describing the Proteome .....	45
2.5	Superfamily Evolution and Genome Complexity .....	46
2.6	Superfamily Evolution and Functional Relationships .....	48
2.7	Limits to Genome Complexity in Prokaryotes .....	50
2.8	The Bacterial Factory .....	52
2.9	Conclusions .....	53
	References .....	54

### 3 The Evolution of the Globins: We Thought We Understood It

<i>A.M. Lesk</i> .....	57
3.1 Introduction .....	58
3.2 Coordinates and Calculations .....	59
3.3 Results .....	59
3.3.1 Description of Secondary and Tertiary Structure of Full-Length (~150-Residue) Globins .....	59
3.3.2 Description of Secondary and Tertiary Structure of Truncated Globins .....	60
3.3.3 Alignment .....	60
3.4 Helix Contacts .....	62
3.4.1 Geometry of Inter-Helix Contacts .....	62
3.4.2 Pairs of Helices Making Contacts .....	63
3.4.3 Structures of Helix Interfaces in Truncated Globins, Compared to Those in Sperm Whale Myoglobin .....	65
3.4.4 The B/G Interface .....	65
3.4.5 The A/H Interface .....	66
3.4.6 The B/E Interface .....	67
3.5 Patterns of Residue-Residue Contacts at Helix Interfaces .....	68
3.5.1 The G/H Interface .....	69
3.6 Haem Contacts .....	72
3.7 The Tunnel .....	72
3.8 Conclusions .....	72
References .....	73

### 4 The Structurally Constrained Neutral Model of Protein Evolution

<i>U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo</i> .....	75
4.1 Aspects of Population Genetics .....	76
4.1.1 Population Size and Mutation Rate .....	76
4.1.2 Natural Selection .....	77
4.1.3 Mutant Spectrum .....	78
4.1.4 Neutral Substitutions .....	80
4.1.5 Beyond the Small $M\mu$ Regime: Neutral Networks ....	81

4.2	Structural Aspects of Molecular Evolution .....	83
4.2.1	Neutral Theory and Protein Folding Thermodynamics .....	83
4.2.2	Structural Conservation and Functional Changes in Protein Evolution .....	84
4.2.3	Models of Molecular Evolution with Structural Conservation .....	85
4.3	The SCN Model of Evolution .....	87
4.3.1	Representation of Protein Structures .....	88
4.3.2	Stability Against Unfolding .....	88
4.3.3	Stability Against Misfolding .....	89
4.3.4	Calculation of $\alpha(\mathbf{A})$ .....	89
4.3.5	Sampling the Neutral Networks .....	91
4.3.6	Fluctuations and Correlations in the Evolutionary Process .....	91
4.3.7	Substitution Process .....	93
4.4	Site-Specific Amino Acid Distributions .....	97
4.4.1	Vectorial Representation of Protein Sequences .....	98
4.4.2	Vectorial Representation of Protein Folds .....	99
4.4.3	Relation Between Sequence and Structure .....	99
4.4.4	The PE as a Structural Determinant of Evolutionary Conservation .....	100
4.4.5	Site-Dependent Amino Acid Distributions .....	101
4.4.6	Sequence Conservation and Structure Designability ..	104
4.4.7	Site-Specific Amino Acid Distributions in the PDB ..	105
4.4.8	Mean-Field Model of Mutation plus Selection .....	107
4.5	Conclusions .....	109
	References .....	109
 <b>5 Towards Unifying Protein Evolution Theory</b>		
	<i>N.V. Dokholyan and E.I. Shakhnovich</i> .....	113
5.1	Two Views on Protein Evolution .....	113
5.2	Challenges in Functionally Annotating Structures .....	113
5.3	The Importance of the Tree of Life .....	115
5.4	Building the PDUG .....	116
5.5	Properties of the PDUG: Power Laws on Very Different Evolutionary Scales .....	117
5.6	Functional Flexibility Score: Calculating Entropy in Function Space .....	118
5.7	Lattice Proteins and Its Random Subspaces: Structure Graphs ..	119
5.8	Divergence and Convergence Explored: What Power Laws Tell Us about Evolution .....	120
5.9	Context Is Important .....	122
5.10	Not All Functions Are Created Equal and Neither Are Structures .....	122

XII Contents

5.11	Concluding Remarks .....	124
	References .....	124

---

**Part II Molecules: Genomes**

---

<b>6</b>	<b>A Twenty-First Century View of Evolution: Genome System Architecture, Repetitive DNA, and Natural Genetic Engineering</b>	
	<i>J.A. Shapiro</i> .....	129
6.1	Introduction: Cellular Computation and DNA as an Interactive Data Storage Medium .....	129
6.2	Genome System Architecture and Repetitive DNA .....	130
6.3	Genomes and Cellular Computation: <i>E. coli lac</i> Operon .....	132
6.4	New Principles of Evolution: The Lessons of Sequenced Genomes .....	135
6.5	Natural Genetic Engineering .....	136
6.6	Conclusions: A Twenty-First Century View of Evolution .....	141
6.7	Twenty-First Century Directions in Evolution Research .....	143
	References .....	144
<b>7</b>	<b>Genomic Changes in Bacteria: From Free-Living to Endosymbiotic Life</b>	
	<i>F.J. Silva, A. Latorre, L. Gómez-Valero, and A. Moya</i> .....	149
7.1	Introduction .....	149
7.2	Genetic and Genomic Features of Endosymbiotic Bacteria .....	153
	7.2.1 Sequence Evolution in Endosymbionts .....	153
	7.2.2 Reductive Evolution: DNA Loss and Genome Reduction in Obligate Bacterial Mutualists .....	158
	7.2.3 Chromosomal Rearrangements Throughout Endosymbiont Evolution .....	160
7.3	Conclusions and Prospects .....	162
	References .....	163

---

**Part III Phylogenetic Analysis**

---

<b>8</b>	<b>Molecular Phylogenetics: Mathematical Framework and Unsolved Problems</b>	
	<i>X. Xia</i> .....	169
8.1	Introduction .....	169
8.2	Substitution Models .....	170
	8.2.1 Nucleotide-Based Substitution Models and Genetic Distances .....	171

8.2.2	Amino Acid-Based and Codon-Based Substitution Models	176
8.3	Tree-Building Methods	178
8.3.1	Distance-Based Methods	178
8.3.2	Maximum Parsimony Methods	181
8.3.3	Maximum Likelihood Methods	182
8.3.4	Bayesian Inference	185
8.4	Final Words	187
	References	187

## 9 Phylogenetics and Computational Biology of Multigene Families

	<i>P. Liò, M. Brilli, and R. Fani</i>	191
9.1	Introduction	191
9.2	How Do Large Gene Families Arise?	193
9.3	The Classical Model of Gene Duplication	193
9.4	Subfunctionalization Model	194
9.5	Subneofunctionalization	195
9.6	Tests for Subfunctionalization	196
9.7	Tests for Functional Divergence After Duplication	196
9.7.1	Case Study 1: Chemokine Receptors Expansion in Vertebrates	197
9.7.2	Case Study 2: The Evolution of TIM Barrel Coding Genes	199
	References	204

## 10 SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis

	<i>D. Charif and J.R. Lobry</i>	207
10.1	Introduction	207
10.1.1	About R and CRAN	207
10.1.2	About this Document	208
10.1.3	About Sequin and <b>seqinR</b>	208
10.1.4	About Getting Started	208
10.1.5	About Running R in Batch Mode	208
10.1.6	About the Learning Curve	209
10.2	How to Get Sequence Data	213
10.2.1	Importing Raw Sequence Data from Fasta Files	213
10.2.2	Importing Aligned Sequence Data	214
10.2.3	Complex Queries in ACNUC Databases	218
10.3	How to Deal with Sequence	220
10.3.1	Sequence Classes	220
10.3.2	Generic Methods for Sequences	220
10.3.3	Internal Representation of Sequences	221

XIV Contents

10.4	Multivariate Analyses .....	225
10.4.1	Correspondence Analysis .....	225
10.4.2	Synonymous and Nonsynonymous Analyses.....	230
References	.....	232

---

**Part IV Networks**

---

**11 Evolutionary Genomics of Gene Expression**

<i>I.K. Jordan and L. Mariño-Ramírez</i>	.....	235
11.1	Sequence Divergence .....	236
11.1.1	Ortholog Identification .....	236
11.1.2	Sequence Alignment .....	237
11.1.3	Sequence Distance Calculation .....	237
11.2	Gene Expression Divergence.....	240
11.2.1	Database Sources .....	241
11.2.2	Probe-to-Gene Mapping.....	241
11.2.3	Structure of the Data .....	242
11.2.4	Transformation and Normalization .....	242
11.2.5	Measuring Divergence .....	243
11.2.6	Clustering and Visualization .....	245
11.3	Integrated Analysis .....	246
11.3.1	Sequence vs. Expression Divergence .....	246
11.3.2	Neutral Changes in Gene Expression .....	247
11.3.3	Evolutionary Conservation of Gene Expression .....	250
References	.....	251

**12 From Biophysics to Evolutionary Genetics:  
Statistical Aspects of Gene Regulation**

<i>M. Lässig</i>	.....	253
12.1	Introduction .....	253
12.2	Biophysics of Transcriptional Regulation .....	254
12.2.1	Factor-DNA Binding Energies .....	255
12.2.2	Energy Distribution in the Genome.....	257
12.2.3	Search Kinetics .....	258
12.2.4	Thermodynamics of Factor Binding .....	258
12.2.5	Sensitivity and Genomic Design of Regulation .....	260
12.2.6	Programmability and Evolvability of Regulatory Networks .....	260
12.3	Bioinformatics of Regulatory DNA .....	261
12.3.1	Markov Model for Background Sequence .....	261
12.3.2	Probabilistic Model for Functional Sites .....	262
12.3.3	Bayesian Model for Genomic Loci .....	263
12.3.4	Dynamic Programming and Sequence Analysis .....	264
12.4	Evolution of Regulatory DNA .....	266
12.4.1	Deterministic Population Dynamics and Fitness .....	267

12.4.2	Stochastic Dynamics and Genetic Drift . . . . .	268
12.4.3	Mutation Processes and Evolutionary Equilibria . . . . .	270
12.4.4	Substitution Dynamics . . . . .	271
12.4.5	Neutral Dynamics in Sequence Space, Sequence Entropy . . . . .	273
12.4.6	Dynamics Under Selection, the Score-Fitness Relation . . . . .	274
12.4.7	Measuring Selection for Binding Sites . . . . .	275
12.4.8	Nucleotide Frequency Correlations . . . . .	276
12.4.9	Stationary Evolution of Binding Sites . . . . .	276
12.4.10	Adaptive Evolution of Binding Sites . . . . .	278
12.5	Toward a Dynamical Picture of the Genome . . . . .	278
12.5.1	Evolutionary Interactions Between Sites . . . . .	279
12.5.2	Site-Shadow Interactions . . . . .	280
12.5.3	Gene Interactions . . . . .	280
12.5.4	Evolutionary Innovations . . . . .	281
	References . . . . .	281

---

## Part V Populations

---

### 13 Drift and Selection in Evolving Interacting Systems

<i>T. Ohta</i> . . . . .	285
13.1 Hierarchy of Networks . . . . .	286
13.2 Drift and Selection, a Historical Perspective . . . . .	287
13.3 Molecular Clock and Near-Neutrality . . . . .	288
13.4 Mutants' Effects on Fitness . . . . .	291
13.5 Evolution of Form and Shape: Cooption . . . . .	294
References . . . . .	296

### 14 Adaptation in Simple and Complex Fitness Landscapes

<i>K. Jain and J. Krug</i> . . . . .	299
14.1 Basic Concepts and Models . . . . .	300
14.1.1 Fitness, Mutations, and Sequence Space . . . . .	300
14.1.2 Mutation-Selection Models . . . . .	304
14.2 Simple Fitness Landscapes . . . . .	307
14.2.1 The Error Threshold: Preliminary Considerations . . . . .	307
14.2.2 Error Threshold in the Sharp Peak Landscape . . . . .	308
14.2.3 Exact Solution of a Sharp Peak Model . . . . .	311
14.2.4 Modifying the Shape of the Fitness Peak . . . . .	312
14.2.5 Beyond the Standard Model . . . . .	317
14.3 Complex Fitness Landscapes . . . . .	321
14.3.1 An Explicit Genotype-Phenotype Map for RNA Sequences . . . . .	322
14.3.2 Uncorrelated Random Landscapes . . . . .	322
14.3.3 Correlated Landscapes . . . . .	323

XVI Contents

14.3.4	Neutrality . . . . .	326
14.4	Dynamics of Adaptation . . . . .	327
14.4.1	Peak Shifts and Punctuated Evolution . . . . .	328
14.4.2	Evolutionary Trajectories for the Quasispecies Model . . . . .	328
14.4.3	Dynamics in Smooth Fitness Landscapes . . . . .	332
14.5	Evolution in the Laboratory . . . . .	333
14.5.1	RNA Evolution In Vitro . . . . .	333
14.5.2	Quasispecies Formation in RNA Viruses . . . . .	334
14.5.3	Dynamics of Microbial Evolution . . . . .	334
14.6	Conclusions . . . . .	335
	References . . . . .	336

**15 Genetic Variability in RNA Viruses:  
Consequences in Epidemiology and in the Development  
of New Strategies for the Extinction of Infectivity**

<i>E. Lázaro</i>	. . . . .	341
15.1	Introduction . . . . .	341
15.2	Replication of RNA Viruses and Generation of Genetic Variability . . . . .	343
15.3	Structure of Viral Populations . . . . .	344
15.4	Viral Quasi-Species and Adaptation . . . . .	345
15.5	Population Dynamics of Host–Pathogen Interactions . . . . .	348
15.6	The Limit of the Error Rate . . . . .	350
15.6.1	Increases in the Error Rate of Replication. Lethal Mutagenesis As a New Antiviral Strategy . . . . .	352
15.6.2	Evolution of Viral Populations Through Successive Bottlenecks . . . . .	355
15.7	Conclusions . . . . .	359
	References . . . . .	360

<b>Index</b>	. . . . .	363
--------------	-----------	-----



<http://www.springer.com/978-3-540-35305-8>

Structural Approaches to Sequence Evolution

Molecules, Networks, Populations

Bastolla, U.; Porto, M.; Roman, E.; Vendruscolo, M.

(Eds.)

2007, XIX, 367 p., Hardcover

ISBN: 978-3-540-35305-8