Contents

Foreword ix
Nick Goldman and Ziheng Yang

Preface xi
List of Contributors xiv

Part I: Modelling codon evolution

1: Background 3
Adrian Schneider and Gina M. Cannarozzi
1.1 Models of molecular evolution 3
1.2 Markov models 3
  1.2.1 Markov chains 4
  1.2.2 Multiple substitutions 4
  1.2.3 Continuous-time processes 4
  1.2.4 Time-reversibility 5
1.3 Maximum-likelihood estimation 5
  1.3.1 ML example 5
  1.3.2 Posterior probabilities 6
  1.3.3 Likelihood of a phylogenetic tree 6
1.4 Performance assessment 7
  1.4.1 Likelihood-based tests 7
  1.4.2 Simulations 8
  1.4.3 Empirical tests 8

2: Parametric models of codon evolution 12
Maria Anisimova
2.1 Basic Markov models of codon substitution 12
  2.1.1 From DNA substitution models to codon models 12
  2.1.2 Estimating codon frequency distribution 14
2.2 Evaluating selective pressure at the protein level 15
  2.2.1 The neutral theory and the likelihood ratio test (LRT) for positive selection 15
  2.2.2 Modelling variable selection pressure over time 16
  2.2.3 Modelling variable selection pressure among sites 19
  2.2.4 Predicting locations of sites under positive selection 20
  2.2.5 Detecting positive selection in presence of recombination 20
  2.2.6 Modelling variable selection pressure among sites and over time 22
2.3 Measuring selection on physico-chemical properties of amino acids 24
2.4 Modelling site-dependence in coding sequences
2.5 Further development of parametric models

3: Empirical and semi-empirical models of codon evolution
Adrian Schneider and Gina M. Cannarozzi

3.1 Introduction
3.2 Empirical model by Schneider et al. (2005)
  3.2.1 Methods
  3.2.2 Results and discussion
  3.2.3 Conclusion
3.3 Combined model by Doron-Faigenboim and Pupko (2007)
  3.3.1 Methods
  3.3.2 Discussion
3.4 Model by Kosiol et al. (2007)
  3.4.1 Methods
  3.4.2 Discussion
3.5 Codon test
3.6 Empirical search for the most important parameters
3.7 Summary

4: Monte Carlo computational approaches in Bayesian codon-substitution modelling
Nicolas Rodrigue and Nicolas Lartillot

4.1 Introduction
4.2 The Bayesian framework
4.3 Site-independent models of codon substitution
  4.3.1 The Muse and Gaut, and Goldman and Yang-based models
  4.3.2 Plain MCMC
  4.3.3 Thermodynamic MCMC
4.4 Site-interdependent models of codon substitution
  4.4.1 The Robinson et al.-based models
  4.4.2 Plain MCMC
  4.4.3 Thermodynamic MCMC
4.5 Other recent modelling innovations and overall rankings
4.6 Future directions

5: Likelihood-based clustering (LiBaC) for codon models
Hong Gu, Katherine A. Dunn, and Joseph P. Bielawski

5.1 Introduction
5.2 Theory for likelihood-based clustering (LiBaC)
5.3 Detecting positive selection in a large-scale analysis of real gene sequences
5.4 Objective comparison of model-based classifications
5.5 Simulation studies of model-based classification
  5.5.1 Performance of LiBaC and other methods on simulated data
  5.5.2 Tradeoffs between precision and recall under LiBaC are adjustable by the posterior probability cutoff
5.6 Recommendations for using LiBaC
6: Detecting and understanding natural selection

Maria Anisimova and David A. Liberles

6.1 Selective mechanisms operating on gene sequences
6.2 Brief overview of statistical methodologies for detecting positive selection
  6.2.1 Neutrality tests based on frequency spectrum
  6.2.2 Neutrality tests based on variability within and between species
  6.2.3 Poisson random-field models (PRF)
  6.2.4 Methods based on population differentiation
  6.2.5 Methods based on linkage disequilibrium (LD) and haplotype structure
  6.2.6 Methods based on detecting rate shifts
  6.2.7 Detecting selection based on \( d_N/d_S \) with Markov codon models
6.3 The utility and the interpretation of the \( d_N/d_S \) measure
6.4 Accounting for indels and overlapping ORFs
6.5 Model-based approaches and common misconceptions
6.6 Selection and adaptive traits
6.7 Lessons from genomic studies and implications for studies of genetic disease

7: Codon models as a vehicle for reconciling population genetics with inter-specific sequence data

Jeffrey L. Thorne, Nicolas Lartillot, Nicolas Rodrigue, and Sang Chul Choi

7.1 Introduction
7.2 The importance of phenotype
7.3 The Halpern-Bruno approach
  7.3.1 The basic idea
  7.3.2 Population genetic interpretations through retrofits
  7.3.3 The Robinson model
  7.3.4 The Sella–Hirsh refinement
  7.3.5 The \( \omega \) parameter
  7.3.6 Applications and potential applications
7.4 Limitations of the Halpern-Bruno approach
  7.4.1 The stationarity assumption
  7.4.2 The low mutation rate assumption and the Hill–Robertson effect
7.5 Future directions

8: Robust estimation of natural selection using parametric codon models

Gavin A. Huttley and Von Bing Yap

8.1 Introduction
8.2 Context-dependent substitution models
8.3 Evaluating properties of dinucleotide models
  8.3.1 Analysis of simulated data
  8.3.2 Analysis of primate introns
8.4 Evaluating properties of codon models
  8.4.1 Analysis of simulated data
  8.4.2 Analysis of primate introns
8.5 Impact of model definitions on statistical power
8.6 Conclusion
9: Simulation of coding sequence evolution 126
   Miguel Arenas and David Posada
   9.1 Introduction 126
   9.2 Simulation of coding sequences
      9.2.1 Forward simulations 126
      9.2.2 Simulations of coalescent histories 127
      9.2.3 Simulation of codon substitutions 127
   9.3 Uses of simulated coding data 128
   9.4 Software implementations 130

10: Use of codon models in molecular dating and functional analysis 133
    Steven A. Benner
    10.1 Introduction 133
    10.2 The level of analysis most useful for functional biology 133
    10.3 Improving codon analysis beyond the $K_a/K_s$ and $d_N/d_S$ ratios 135
    10.4 Heuristic approaches to improve codon analysis beyond the $K_a/K_s$ and $d_N/d_S$ ratios 136
    10.5 Clocks 138
    10.6 Calibrating the TREx clock 140
    10.7 Conclusions 143

11: The future of codon models in studies of molecular function: ancestral reconstruction and clade models of functional divergence 145
    Belinda S.W. Chang, Jingjing Du, Cameron J. Weadick, Johannes Müller, Constanze Bickelmann, D. David Yu, and James M. Morrow
    11.1 Introduction 145
    11.2 Ancestral reconstruction 145
    11.3 Reconstructing synonymous evolution in vertebrate rhodopsins 148
    11.4 Clade models of functional divergence 152
    11.5 Testing for functional divergence among teleost SWS2 opsins 155
    11.6 Conclusions 158

12: Codon models applied to the study of fungal genomes 164
    Gabriela Aguileta and Tatiana Giraud
    12.1 Introduction 164
    12.2 Fungi as pathogens
       12.2.1 Adaptive evolution: characterizing functional divergence and associated selective pressure changes 164
       12.2.2 Host-pathogen evolution: detecting arms races through the evolution of R-genes, avirulence genes, as well as fungal effectors and elicitors 169
       12.2.3 Lifestyle-associated adaptations: from saprophytes to pathogens 172
    12.3 Fungi as symbionts: selective pressure to maintain symbiosis in mycorrhizae and lichens 173
    12.4 Evolution of codon usage in fungal genomes
       12.4.1 Fungi as eukaryotic models of codon usage evolution 174
       12.4.2 Codon models applied to detect codon bias in fungi: translational selection 175
       12.4.3 Fungal preferred codon uses 176
    12.5 Functional shifts: measuring the concomitant variation in selective pressure 177
12.6 Adaptive evolution of gene expression: wiring and re-wiring regulatory networks 177
12.7 Ancestral polymorphisms: maintaining allelic variants for extended periods 178
12.8 The origin of sexual chromosomes in Fungi: reduced selection efficiency and degenerative changes in preferred codon usage 180
12.9 Finding genes associated with specialization and speciation 180
12.10 Conclusion: new uses of codon models for analysing fungal genomes 181

Part II: Codon usage bias

13: Measuring codon usage bias 189

Alexander Roth, Maria Anisimova, and Gina M. Cannarozzi

13.1 Introduction 189
13.2 Causes of codon usage bias 189
13.2.1 Mutational biases affecting codon usage 189
13.2.2 Selection affecting codon usage 190
13.3 Applications for indices of codon usage bias 192
13.4 Previous studies of codon usage indices 192
13.5 Measures of codon bias 193
13.5.1 Relative codon frequencies 194
13.5.2 Measures based on reference 194
13.5.3 Measures based on the geometric mean 196
13.5.4 Measures based on deviation from an expected distribution 199
13.5.5 Measures based on information theory 200
13.5.6 Measures focusing on tRNA interaction 201
13.5.7 Measures based on intrinsic properties of codon usage 202
13.5.8 Measures for total codon usage in genomes 205
13.6 Dependencies of measures 206
13.6.1 Dependence on nucleotide composition 206
13.6.2 Dependence on gene length 207
13.6.3 Dependence on the degree of codon degeneracy 207
13.6.4 Dependence on the skewness of synonymous codon usage 208
13.6.5 Dependence on amino acid discrepancy 208
13.7 Comparisons using biological data 210
13.7.1 Correlation with transcript and protein levels 211
13.7.2 Correlation with rate of protein synthesis 211
13.8 Limitations of codon usage indices 212
13.9 Conclusions 212

14: Detection and analysis of conservation at synonymous sites 218

Nimrod D. Rubinstein and Tal Pupko

14.1 Introduction to conservation 218
14.2 Classical view regarding synonymous mutations as neutral 218
14.3 Conservation due to translational optimization 219
14.4 Conservation due to mRNA structure 220
14.5 Conservation due to overlapping genes 222
14.6 Conservation to maintain splicing signals 223
14.7 Application of codon models to the detection of conserved synonymous sites 223
14.8 Other cis-encoded elements responsible for synonymous conservation 224
14.9 Concluding remarks 225
15: Distance measures and machine learning approaches for codon usage analyses
Fran Supek and Tomislav Šmuc
15.1 Causes of biased codon usage
15.2 Methods for quantifying codon biases
   15.2.1 Unsupervised methods
   15.2.2 Supervised methods
15.3 Application to bacterial and archaeal genomes
   15.3.1 Rationale behind using classifiers to control for background nucleotide composition
   15.3.2 An example application of supervised machine learning in codon usage analysis
   15.3.3 Proportion of genomes subject to translational selection and correlations with gene functional categories
   15.3.4 Distribution of codon-optimized genes within specific gene functional categories and relationship to microbial lifestyle
   15.3.5 mRNA expression levels and codon preferences of genes subject to translational selection

16: The application of population genetics in the study of codon usage bias
Kai Zeng
16.1 Introduction
16.2 Theory
   16.2.1 The reversible mutation model and the infinite sites model
   16.2.2 Parameter estimation and data preparation under the RM model
   16.2.3 Parameter estimation and data preparation under the IS model
16.3 Some recent theoretical developments
   16.3.1 Methods that take account of the effects of recent changes of population size
   16.3.2 A multi-allele model with reversible mutation
   16.3.3 The effects of linkage on parameter estimation
16.4 Conclusion

17: Structural and molecular features of non-standard genetic codes
Maria do Céu Santos and Manuel A. S. Santos
17.1 Overview
   17.1.1 Genetic code diversity: mitochondrial and nuclear
   17.1.2 Neutral and non-neutral mechanisms
17.2 How are non-neutral genetic code changes selected?
   17.2.1 Selenocysteine
   17.2.2 Pyrrolysine
   17.2.3 The CUG case in Candida spp.
17.3 Cellular and molecular consequences of non-neutral genetic code alterations
   17.3.1 Consequences at proteome level
   17.3.2 Consequences at genome level
   17.3.3 Consequences at phenotypic level
17.4 Conclusions and perspectives

Index