Université de Montréal

Phylogenetic structural modeling of molecular evolution

par
Nicolas Rodrigue

Département de biochimie
Faculté de médecine

Thèse présentée à la Faculté de médecine
en vue de l'obtention du grade Ph.D.
en bio-informatique

Décembre, 2007

Université de Montréal
Faculté de médecine

Cette thèse intitulée:
Phylogenetic structural modeling of molecular evolution

présentée par
Nicolas Rodrigue

a été évaluée par un jury composé des personnes suivantes:

Sabin Lessard
président-rapporteur

Hervé Philippe
directeur de recherche

David Bryant
codirecteur

Franz Lang
membre du jury

Jeffrey L. Thorne
examinateur externe

François Major
représentant du doyen de la FES

Résumé

Le domaine de la biologie moléculaire computationnelle n'en est qu'à ses débuts. En dépit des technologies modernes permettant de produire et d'archiver de grandes quantités de données, les modèles tentant d'expliquer ces données sont encore bien loin d'un niveau de réalisme acceptable. Par exemple, la plupart des modèles phylogénétiques d'évolution moléculaire reposent sur l'hypothèse que chaque position (site) d'une protéine évolue indépendamment des autres positions. Cette simplification est évoquée pour des raisons de calcul, bien qu'elle soit biologiquement infondée.

Dans cette dissertation, nous explorons différentes techniques computationnelles pour l'étude de modèles phylogénétiques avec interdépendance entre les acides aminés d'une protéine, ou entre les codons du gène associé. Ces modèles prennent en compte les interdépendances résultant de la structure tertiaire de la protéine, utilisant des représentations structurales simplifiées en combinaison avec des potentiels statistiques, eux-mêmes dérivés d'une base de données de protéines ayant des structures connues. Dans ce contexte, les potentiels statistiques procurent une estimation de la compatibilité d'une séquence d'acides aminés dans une structure donnée. Ainsi, le critére de compatibilité de l'ensemble de la séquence avant et après un évènement de substitution aura une influence sur la probabilité d'un scénario évolutif. Nous appliquons une analyse Bayesienne de sélection et d'évaluation de modèle—par l'entremise de calculs numériques de vraisemblances marginales, et de vérification prédictive—étendu sur plusieurs types de modèles d'évolution, avec et sans critère de compatibilité structurale. En y considérant deux niveaux d'interprétation des données (soit focalisé sur des séquences d'acides aminés, ou bien sur des séquences nucléotidiques codantes), nous proposons le concept de *référence phénoménologique*, comme moyen d'évaluer et de dégager des pistes de modélisation méchanistique.

Notre analyse sur des données réelles nous indique que les modèles incorporant des

considérations de compatibilités structurales apportent toujours une amélioration de la vraisemblance marginale. Par contre, l'usage d'un potentiel statistique en soi n'explique pas des caractéristiques bien connues de l'évolution moléculaire, tel que l'hétérogénité des taux de substitutions entre sites, ou l'interchangeabilité de paires d'acides aminés. D'autres études seront nécessaires afin d'établir si de meilleurs potentiels statistiques, ou d'autres mesures, peuvent arriver à reproduire ces caractéristiques. Pour l'instant, les meilleurs modèles sont ceux qui combinent un potentiel statistique avec une formulation sous-jacente suffisamment riche et bien construite. Nous proposons plusieurs pistes de recherche, menant à un cadre qui pourrait éventuellement avoir des répercussions sur l'inférence phylogénétique, la détection et la caractérisation de pressions sélectives, la prédiction de structure, l'interaction protéine-protéine, et le dessin de séquences protéiques.

Mots clés : évolution moléculaire; phylogénie; structure protéique tertiaire; potentiel statistique; chaîne de Markov Monte Carlo; statistique Bayesienne; modélisation phénomenologique; modélisation méchanistique.

Abstract

The field of computational molecular biology is at an early stage. Despite major advances in producing and gathering large quantities of molecular data, the actual development of models capable of adequately explaining such data are still a far cry from a suitable level of realism. For instance, most phylogenetic models of molecular sequence evolution assume that each position of an alignment evolves independently of all other positions—a computationally motivated simplification well-known to be biologically unsound.

In this work, we explore different computational methods for the study of phylogenetic models that allow for a general interdependence between the amino acid positions of a protein, or between the codons of the associated gene. The models are focused on site-interdependencies resulting from sequence-structure compatibility constraints, using simplified molecular structure representations in combination with a set of statistical potentials, which are themselves derived from a protein database of resolved structures. This structural compatibility criterion defines a *sequence fitness* concept, and the methods developed can incorporate different site-interdependent sequence fitness measurements. We apply Bayesian methods of model selection and assessment—based on numerical calculations of marginal likelihoods, and posterior predictive checks—to evaluate evolutionary models encompassing the site-interdependent framework. Through our consideration of different levels of data interpretation (either focusing on amino acid sequences only, or focusing on coding nucleotide sequences), we propose the concept of *phenomenological benchmarking*, as a means of guiding and assessing mechanistic modeling strategies.

Our applications of these methods on real data indicates that considering sequence-structure compatibility requirements, as done here, leads to an improved model fit for all datasets studied. Yet, we find that the use of potentials alone does not suitably ac-

count for across-site rate heterogeneity or amino acid exchange propensities, and more work is needed to establish if richer forms of potentials, or other type of sequence fitness concepts, might better capture such features. In the meantime, the most favored models combine the use of statistical potentials with a suitably rich and well-posed site-independent model. We propose several avenues meriting further investigation, leading to a research expanse with possible impacts on phylogenetic inference, the detection and characterization of selective features, protein structure prediction, protein-protein interactions, and computational protein design.

Key words: molecular evolution; phylogeny; protein tertiary structure; statistical potential; Markov chain Monte Carlo; Bayesian statistics; phenomenological modeling; mechanistic modeling.

Contents

List of Tables

List of Figures

List of Abbreviations

AAP: Amino acid preference;

BAS: Bastolla;

CP: Codon preference;

CPU: Central processing unit;

DA: Data augmentation;

DNA: Deoxyribonucleic acid;

DP: Dirichlet process (on nonsynonymous rate factors);

EM: Expectation maximization;

GHz: Giga Hertz;

GTR: General time reversible;

GY: Goldman and Yang;

JTT: Jones, Taylor, and Thornton;

MCG: Monte Carlo gradient;

MCEM: Monte Carlo expectation maximization;

MCMC: Markov chain Monte Carlo;

MG: Muse and Gaut;

MH: Metropolis Hastings;

MJ: Miyazawa and Jernigan;

ML: Maximum likelihood;

PDB: Protein data bank;

PX: Parameter expansion;

RNA: Ribonucleic acid;

SC: Stucturally constrained;

WAG: Whelan and Goldman.

Pour mes parents, Laurence et André

Preface

In the fall of 2003, I had the good fortune of beginning studies in the bioinformatics programme at the Université de Montréal, joining a vibrant research group led by Hervé Philippe. Well aware of the potential of computational methods for studying the growing banks of biological data, the group's research activities could be categorized along two main axes: first, the use of mathematical models of molecular evolution for inferring the relatedness of species, or *phylogenies*; and second, the development of new evolutionary models, which, on the one hand, exhibit robustness in phylogenetic analysis per se, and, on the other hand, elucidate patterns of the underlying substitution process. With then post-doctoral fellow Nicolas Lartillot, several projects along these lines were initiated, of which I had the chance to participate.

The present work is my recapitulation of research endeavors along one of these projects, which has been the focus of my doctoral studies. The work could be summarized as an exploration of computational methods for implementing richer descriptions of molecular evolution, with the specific objective of incorporating explicit protein structure considerations within different models of sequence evolution.

The dissertation is organized in three parts. The first briefly overviews the historical settings for the recent emergence of the field, and presents the core methodological framework adopted throughout the text. The second part applies the framework to revising modeling assumptions at different levels of interpretation. The bulk of this second part has been the subject of published or forthcoming articles. These are the

following:

Rodrigue, N., Lartillot, N., Bryant, D., and Philippe, H. (2005). Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene, 347:207-217.

Rodrigue, N., Philippe, H., and Lartillot, N. (2006). Assessing site-interdependent phylogenetic models of sequence evolution. Molecular Biology and Evolution, 23:1762-1775.

Kleinman, C. L., **Rodrigue, N.**, Bonnard, C., Philippe, H., and Lartillot, N. (2006). A maximum likelihood framework for protein design. BMC Bioinformatics, 7:326.

Rodrigue, N., Philippe, H. and Lartillot, N. (2007). Exploring fast computational strategies for probabilistic phylogenetic analysis. Systematic Biology, 56:711-726.

Rodrigue, N., Philippe, H. and Lartillot, N. (in press). Uniformization for sampling realizations of Markov processes: Applications to Bayesian implementations of codon substitution models. Bioinformatics.

Rodrigue, N., Lartillot, N., and Philippe, H. (submitted). Mechanistic modeling of amino acid or codon preferences for protein-coding nucleotide sequence evolution. Submitted to Genetics.

Rodrigue, N., Philippe, H., and Lartillot, N. (in preparation). Sampling methods for computing Bayes factors across site interdependent codon substitution models. Planned for Journal of Computational Biology.

The presentation does not, however, follow the "dissertation by articles" format, where a set of articles would be included untouched. Rather, I have tried to re-work the material into a more unified whole. This has considerably reduced the length of the document, while allowing for a homogenization of the mathematical notation, and a clearer emphasis on the main themes of the thesis. Portions of text, figures, and tables from published articles appear with permission from the respective journals, as well as all co-authors. Finally, the third part of the dissertation describes specific calculations

of interest for future work, as well as several modeling extensions and variations meriting further exploration.

The work was very much a collaborative project, with Hervé Philippe and Nicolas Lartillot contributing much of the theoretical ideas throughout. The work was formally documented by myself, including fine details of methods and implementations, as well as initials drafts of manuscripts, which were then collaboratively revised and improved. This is with the exception of chapter 5, which is modified from the work of the third article listed above. The article in question was primarily written by Claudia Kleinman and Nicolas Lartillot, and therefore, for the purpose of this dissertation, I have significantly abbreviated the material, including only that which is directly pertinent to the main developments of this work. Much of the future work discussed in the final part is also a resultant from the collaboration with Claudia, Hervé, and Nicolas.

Following this preface, I have opted to retain the first-person plural throughout the text, as this best reflects the collaborative effort of all those involved (although any errors that may be contained are of my own doing). I use footnotes to clarify the meaning of jargon, or to include non-essential information that may nevertheless be of relevance to the reader. Appendices are used for descriptions of data sets, for lengthy mathematical developments, and to outline the use of a computer package implementing the methods developed in the main body of the document.

I am very grateful for the financial support and encouragement of many organizations. The first three years of my studentship were funded by Génome Québec. The biT bursaries and fellowships for excellence (a Canadian Institute of Health Research strategic training program grant in bioinformatics) provided supplementary funding in 2005, and provided the main funding for my final year. The Faculté des études supérieures provided supplementary funding for attending conferences and meetings; the activities of the Robert Cedergren Center also provided additional funding opportunities; and the

60ème commission franco-québécoise funded trips to Montpellier, to work with Nicolas.

I owe immense thanks to Hervé Philippe for the numerous opportunities he provided throughout my studentship. Hervé's passion for science is invigorating, and his drive contagious. He has tactfully given me both positive encouragement and constructive criticism out of a sincere desire to stimulate better scientific work. He has been a key contributor to scientific orientations, and choice of experiments, and played an active role in the writing of scientific texts. He has also arranged for me to present at numerous scientific meetings, for which he patiently helped me prepare. All the while, Hervé is never but a moment away from the next laugh, going about with a bon-vivant enthusiasm. He has also shown to be incredibly understanding with regards to my personal affairs, for which I owe him my deepest gratitude. Merci Hervé.

I also owe immense thanks to Nicolas Lartillot, without whom I could not have done this work. Nicolas is a true polymath. He is also a very patient and strategic teacher. Although barely my senior, he has been a tremendous source of ideas, techniques, and hands-on help and guidance with computer programs as well as the analyses themselves. Nicolas was also a key contributor in the preparation of scientific texts, and carefully read all of my technical communications. He also responded to my countless emails with detailed ideas and explanations. Beyond these teachings, Nicolas has also stimulated my interest in statistical theory and philosophy of knowledge, all of which he did with good-natured camaraderie, and kind hospitality during my stays at Montpellier. Merci Nico.

I thank the thesis jury members, as well all my teachers, colleagues and class-mates, who have listened to my presentations, re-read parts (or all) of the work, or otherwise positively contributed to my studentship: David Bryant, Claudia Kleinman, Henner Brinkmann, Béatrice Roure, Fabrice Baro, Yan Zhou, Olivier Jeffroy, Frédéric Delsuc, Denis Baurain, Naiara Rodriguez-Ezpeleta, Cécile Bonnard, Wafae El Alaoui,

Guy Larochelle, Gertraud Burger, Marie Robichaud, Elaine Meunier, Miklós Csürös, François-Joseph Lapointe, Pascale Legault, Sabin Lessard, Franz Lang, François Major, and Jeff Thorne. I'm also grateful to John Huelsenbeck, Asger Hobolth, Paul Lewis, Rod Page and all anonymous reviewers for their communications regarding parts of the work presented here.

I would not have completed this dissertation without the support of my loving spouse Rachel Gouin, whose enduring encouragement I have cherished for years. To boot, Rachel also helped me proof-read the entire dissertation. I thank my sweet daughter Kaya Gouin for her encouragement as well, and for her gifts adorning my work desk. Many thanks to my parents-in-law, Pierrette and Jacques Gouin, for their support, and for caring for Kaya on numerous occasions. I also thank my brother François Rodrigue and sister-in-law Ivana Kostic for their love and encouragement. Finally, I thank my dear parents Laurence Drouin and André Rodrigue, for whom I cannot suitably express my gratitude, and to whom, rather, I dedicate this dissertation.

—NR, December, 2007.

Part I

Foundations

Chapter 1

Historical background

> If, during the long course of ages and under varying conditions of life, organic beings vary at all in the several parts of their organization, and I think this cannot be disputed; if there be, owing to the high geometric powers of increase of each species, at some age, season, or year, a severe struggle for life, and this certainly cannot be disputed; then, considering the infinite complexity of the relations of all organic beings to each other and to their conditions of existence, causing an infinite diversity in structure, constitution, and habits, to be advantageous to them, I think it would be a most extraordinary fact if no variations useful to any organic being do occur, assuredly individuals thus characterized will have the best chance of being preserved in the struggle for life; and from the strong principle of inheritance they will tend to produce offspring similarly characterized. This principle of preservation, I have called, for the sake of brevity, Natural Selection.

—CHARLES DARWIN, *Origin*, p. 127

1.1 Introduction

The objective of evolutionary biology is to propose a mechanistic and historical explanation for the intricate attributes and similarities of different living things. The core of this explanation is commonly associated with Charles Darwin (1809–1882), and his famous book *On The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, commonly contracted to the simple title *Origin* (Darwin, 1859). Beyond its purely biological implications, Darwin's main

message, summarized in the opening citation, has had a profound impact on conceptualizations of self, and of mind, stimulating an ongoing revolution of the general world view. In contrast with the Copernican revolution, which had not attracted wide-spread interest until the scientific details had been resolved, the Darwinian revolution was engaged with important pieces of the theory still missing, and altogether devoid of precise mathematical characterization.

In this first chapter, we outline the main historical and philosophical developments contributing to evolutionary thought, beginning with a brief overview leading to the *Darwinian core*, and the *Evolutionary Synthesis* that subsequently emerged among biologists in the 1930s and 1940s. We next introduce early discoveries in molecular biology, and set forth the modern enterprise of computational evolutionary biology. The literature on the rise of evolutionary thought is vast. For accessible and engaging accounts of the movement, see Burrow (1966), Oldroyd (1980), or Dennett (1995). The present overview merely sketches in contour the main turning points leading to the subject of this thesis, and provides a schematic description of phenomena to be mathematically modeled in later chapters.

1.2 The Darwinian core and the Evolutionary Synthesis

The basic concept of biological evolution can be traced back to the pre-Socratic Greeks. Anaximander (610–546 BC) is thought to be one of the earliest proponents of evolutionary thinking, proposing the first speculations to an aquatic origin of life. However, the *fixist*[1] influence of Plato (428/427–348/347 BC) and Aristotle (384–322 BC) has dominated most occidental cultures. Judeo-Christian cultures in particular have vehiculed

[1] A *fixist* views organismal forms as static, or fixed over time.

the fixist perspective, and only after the Enlightenment would truly *transformist*[1] views take firm rooting in wider discourses. Nonetheless, the concept of *species relatedness* was pervasive in some form or another throughout history, which had allowed organisms to be classified into formal groups. The most influential figure in developing such a classification was Carolus Linnaeus (1707–1778). Yet even Linnaeus did not offer a causative explanation for the relatedness of organisms, subscribing to the prevailing fixist view.

The work of Jean-Baptiste Lamarck (1744–1829) was a noteworthy turning point. Lamarck's transformist theory proposed that organisms are somehow intrinsically driven toward complexification—the giraffe stretching and strengthening its neck to attain higher leaves—and that traits acquired by individuals are passed on in the next generation—having put such efforts into stretching and strengthening its neck, the giraffe's offspring, so goes the theory, have longer and stronger necks. Lamarck's proposal was viewed with skepticism. Although Darwin himself accepted the possibility that acquired traits might be passed on, he considered the appeal to the "drive" of organisms of weak explanatory power. Ultimately, Lamarckian transformists received their final blow from the work of August Weisnamm (1833–1914), who observed the distinction between germ line and soma, and generally excluding the possibility of somatically derived characteristics being passed on to offspring. Nonetheless, the ideas set forth by Lamarck were important in inspiring truly transformist theories.

The field of geology was also a burgeoning science in the 19th century. In particular, the principle of *uniformitarianism*[2] was expounded by Charles Lyell (1797–1875), who, based on observations of erosion rates, determined the Earth to be at least millions of years old. These observations were influential in setting a new time-frame for interpreting the diversity of organisms, and Darwin is said to have brought Lyell's then recently

[1] A *transformist* acknowledges that organismal forms are subject to transformation over generations.

[2] The principle of *uniformitarianism* states that basic forces acting in the geological past are the same as those acting in the present.

published *Principles of Geology* on his five year voyage around the globe.

Another key contribution came many years earlier from political economics, with the work of Thomas Malthus (1766–1834). In his *Essay on the Principle of Population* (Malthus, 1798), Malthus pointed out that exponential population growth—displayed under plentiful conditions of existence—must eventually be kept in check, ultimately by the limited resources of a finite world. Malthus' examples were anthropomorphic, but nonetheless had a profound impact on Darwin, helping him crystallize the concept of natural selection:

> In October 1838, that is, fifteen months after I had begun my systematic inquiry, I happened to read for amusement Malthus on Population, and being well prepared to appreciate the struggle for existence which everywhere goes on from long-continued observation of the habits of animals and plants, it at once struck me that under these circumstances favorable variations would tend to be preserved, and unfavorable ones to be destroyed. The results of this would be the formation of a new species. Here, then I had at last got a theory by which to work. (From Darwin's autobiography, retrieved online at http://onlinebooks.library.upenn.edu/.)

Darwin was well aware of the epistemological implications of his theory, which diverged markedly with the main stream theological, social and political agendas of his day. For years he remained reluctant to openly come forward with his ideas, until he received a letter from another naturalist, Alfred R. Wallace (1823–1913), expounding the basic elements of the theory. Wallace later recounted how he formed the theory:

> Something led me to think of the positive checks described by Malthus in his essay on population. These checks—war, disease, famine, and the like—must act on animals as well as on man. While pondering vaguely on this fact there suddenly flashed upon me the idea of the survival of the fittest—that the individuals removed by these checks must be on the whole inferior to those that survived. (From *Alfred Russel Wallace: Letters and Reminiscences*, retrieved online at http://manybooks.net/)

Upon reading Wallace's first correspondence on the theory, Darwin wrote:

> I never saw a more striking coincidence. If Wallace had my M.S. sketch written out in 1842 he could not have made a better short abstract! Even his terms now stand as Heads of my chapters (Darwin, 1858, in a letter to Charles Lyell, from *The Correspondence of Charles Darwin*, retrieved online at http://www.darwinproject.ac.uk/).

Wallace had been influenced by the same body of work as Darwin, and their independent convergence on the concept of natural selection testifies to it being a mature free-floating rational at the time, ripe for articulation and serious consideration. Indeed, the main ideas of Darwin and Wallace had already been presented in outline, in 1813 by William Charles Wells (1757–1817), and again (independently) in 1831 by Patrick Mathew (1790–1874). Despite the anticipations of Wells and Mathew, and the convergence of Darwin and Wallace, the theory has historically been attributed to Darwin, mainly due to the breadth of his treatise.

The basic elements of the theory, which we refer to here as the *Darwinian core*, can be broken down into the following argument (modified from Gould, 2002, p. 125):

- Super-fecundity: Organisms tend to produce more offspring than can survive.
- Variation: Organismal forms tend to vary, so that each individual bears distinguishing features.
- Heredity: An organism's offspring tend to be characterized similarly to it.
- Natural selection: Organismal forms endowed with variations well-suited to the conditions of existence will tend to be more successful in producing offspring than ill-suited variations; well-suited variations thus come to dominate the population.

Despite debates among theoreticians from the early-20^{th} century onward, this basic argument is not put into question. In the modern literature, the Darwinian core argument is often taken for granted, for instance, appearing only in footnote in Gould's big book *The Structure of Evolutionary Theory* (2002).

When the theory was first proposed, however, several outstanding questions remained. By far the most important of these was the question of inheritance: how are organismal attributes transmitted to offspring? In 1865, six years after Darwin's *Origin*, Gregor Mendel (1822–1884) published a work demonstrating the existence of discrete

heritable determinants, now called genes, which can be passed on largely unchanged over generations. Unfortunately, Mendel's work went unnoticed, and debates regarding the basic workings of inheritance continued until 1900, when Mendel's work was rediscovered.

The discovery of Mendelian inheritance had the surprising (in hindsight) effect of increasing skepticism for the Darwinian core. In particular, the *mutationist* school, lead by Hugo de Vries (1848–1935), William Bateson (1861–1926) and others, considered that most of the variation in organismal forms could be explained as arising by mutation, without needing to invoke the principle of natural selection. This view was vehiculed by the leading geneticists of the early 20$^{\text{th}}$ century, and eventually lead to the opinion that the Darwinian core had been refuted:

> Modern critics have often asked themselves how it is that a hypothesis like Darwin's, based on such weak foundations, could all at once win over to its side the greater part of contemporary scientific opinion. If the defenders of the theory refer with this end in view to its intrinsic value, it may be answered that the theory has long ago been rejected in its most vital points by subsequent research (Nordenskiold, 1928, p. 477).

Over the years 1918 to 1931, biometric analysis synthesized the seemingly disparate concepts of mutationists and selectionists. In another case of largely convergent theorizing, Ronald A. Fisher (1890–1962), John B. S. Haldane (1892–1964), Sewall G. Wright (1889–1988) and Sergei. S. Chetverikov (1880–1959) proposed mathematical models integrating Mendelian inheritance, mutation, and other biometrical knowledge, with the Darwinian core. Their work was soon corroborated and expanded by biologists at the time, in particular Theodosius Dobzhansky (1900–1975), Ernst W. Mayr (1904–2005), Julian S. Huxley (1887–1975) and George G. Simpson (1902–1984). With this firm theoretical foundation, and empirical substantiations, a general consensus about the workings of biological evolution emerged, which came to be known as *The Evolutionary Synthesis* (or, more simply, the *synthesis*).

The basic elements of the synthesis are well described in modern terms by Futuyma (1986):

> The major tenets of the evolutionary synthesis, then, were that populations contain genetic variation that arises by random (i.e. not adaptively directed) mutation and recombination; that populations evolve by changes in gene frequency brought about by random genetic drift, gene flow, and especially natural selection; that most adaptive genetic variants have individually slight phenotypic effects so that phenotypic changes are gradual (although some alleles with discrete effects may be advantageous, as in certain color polymorphisms); that diversification comes about by speciation, which normally entails the gradual evolution of reproductive isolation among populations; and that these processes, continued for sufficiently long, give rise to changes of such great magnitude as to warrant the designation of higher taxonomic levels (genera, families, and so forth). (p. 12)

As these tenets gained broader acceptance, another revolution was under way with the rise of molecular biology, which would introduce a new kind of data, lending itself to a new level of mathematical analysis.

1.3 Molecular biology

In 1869, Friedrich Miescher (1844–1895) isolated a phosphate-rich chemical he called "nuclein", since it was found in the nuclei of while blood cells. The chemical was later isolated from many other cell types, and was renamed *nucleic acid*. The biological function of nucleic acids remained elusive, however, for many years. In 1944, after over ten years of experimental study, Oswald Avery and colleagues (Avery et al., 1944) gave the first clue that nucleic acids are responsible for the transmission of genetic information. In the ensuing years, their results were expanded, in particular by Hershey and Chase (1952), who demonstrated that deoxyribonucleic acid (DNA) alone is the hereditary material.

In the late 1940s and 1950s, the structure of nucleic acids was worked out in detail: with few exceptions, nucleic acids are polymers of *nucleotides*; each nucleotide is constituted of a nitrogenous base (of which there are four types in DNA), a pentose

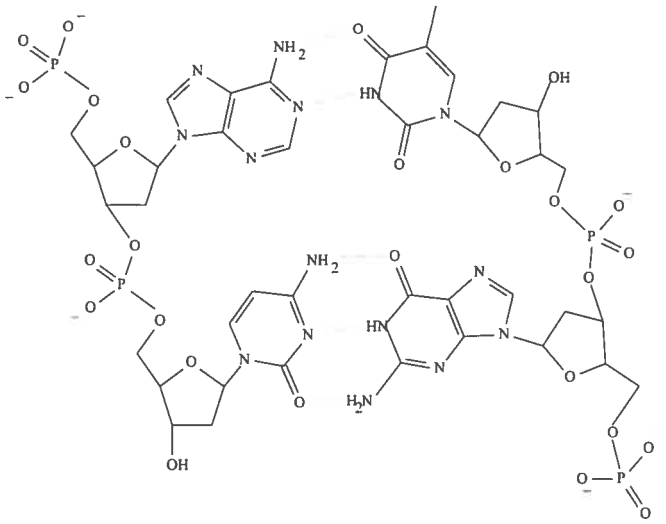

Figure 1.1. Double stranded DNA. The left strand, consisting of A (top) and C (bottom), forms hydrogen bonds (dotted lines) with an anti-parallel strand, consisting of T (top) and G (bottom). The figure was drawn using ChemTool, availible online at http://ruby.chemie.uni-freiburg.de/.

sugar (deoxyribose), and a phosphate group. The sugar and phosphate group form the phosphate-deoxyribose backbone linking nucleotides into a strand. In addition, the nitrogenous bases adenine (A) and thymine (T), as well as guanine (G) and cytosine (C), were found in the same proportions in DNA isolates, a property now called *Chargaff's rule* in honor of its discoverer Erwin Chargaff (1905–2002). This property is explained by the structural pairing of bases, which in turn relates to the overall structure of the DNA molecule: the bases A and T, as well as G and C, are said to be complementary, interacting through hydrogen bonds[1]; complementary anti-parallel single strands of nucleotide polymers interact through such bonds, playing a central role in the formation of double stranded DNA. The chemical arrangement of the components of DNA are displayed in figure 1.1[2]. The chemical structure of DNA was found to induce a

[1]Many exceptions to such pairings have since been established.

[2]The similar ribonucleic acid (RNA), which differs from DNA only through the oxidized ribose sugar, follows the same arrangement. Also, the enzymes involved in RNA synthesis have a much higher affinity for a variant of thymine that lacks a methyl group, called uracil (U). As such, in RNA, T is replace by U, which nonetheless forms hydrogen bonds with A.

coiling pattern in the double strand, in a manner exposing the hydrophilic backbone, while burying and piling the hydrophobic nitrogenous bases into a central core. The now-famous double-helical three-dimensional structure of DNA, as first described by Watson and Crick (1953), is displayed in figure 1.2.

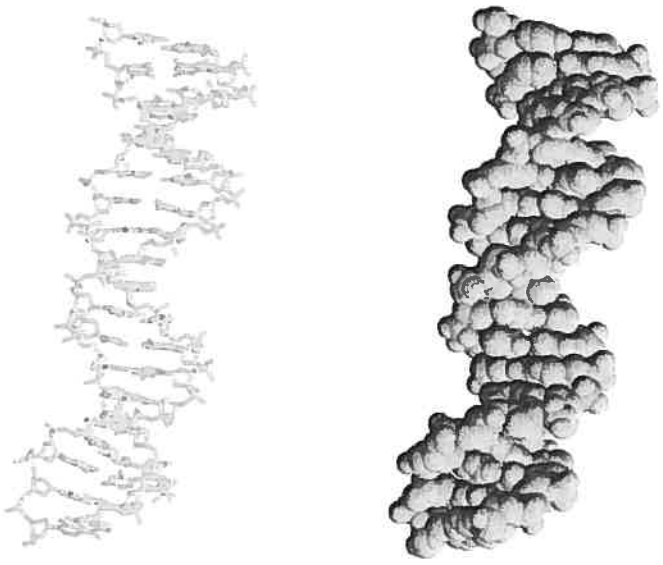

Figure 1.2. Structure of DNA (PDB code 1D66) in "stick" (left) and "spacefilling" (right) representations. The figures were generated using RasMol, availible online at http://www.umass.edu/microbio/rasmol/.

For cells to proliferate, DNA must be replicated, and it is mainly in this replication process that "errors" are made; enzymes and regulatory factors involved in DNA replication may occasionally lead to slight differences in sequences, referred to as mutations. Other factors may also lead to differences in sequences, such as recombination (e.g., chromosomal crossover), or even non-replicative changes (e.g., cytosine deamination), and altogether these different factors induce genetic variability.

Before the sweeping advances of the 1940's and 50's, it was thought that DNA was too simple to be able to specify the complex features of organisms. Proteins,

however, in their enormous varieties, seemed more likely candidates as the carriers of genetic information. Proteins are chains of amino acids, of which there are 20 types in most organisms. An amino acid consists of a carboxylate and an amino group, each attached to a central carbon referred to as the α-carbon. The α-carbon's tetra-valence is completed by a hydrogen and one of 20 organic substituents, or side chains. The full chemical structure of the twenty amino acids is displayed in figure 1.3.

A condensation reaction between the carboxyl group of one amino acid and the amino group of another forms a peptide bond between the two. The chemical structure of the three amino acid chain, or *tripeptide*[1], is displayed in figure 1.4. The sequence of an amino acid chain is referred to as its *primary structure.* The *secondary structure* refers to the manner in which a chain coils (e.g., α-helices) or folds over to form lateral interactions with itself (e.g., β-sheets), whereas the tertiary structure refers to its overall three-dimensional configuration (figure 1.5), formed through networks of interaction between amino acids. Finally, the *quaternary structure* refers to the multimeric assemblage of different protein subunits.

At the time of the publication by Watson and Crick (1953), the relation between the sequence of amino acid chains and DNA was not known. In the subsequent years, a flurry of research in molecular biology produced the modern consensus referred to as the *central dogma* of information flow in the cell (figure 1.6). Through the concerted action of several enzymes and regulatory factors, double stranded DNA momentary "unzips" (hydrogen bonds between A and T, as well as C and G, are disrupted), and one of the strands serves as a template for the *transcription* of a "messenger" RNA, or simply mRNA, with a sequence of ribo-nucleotides of matching base complementarity; by convention, a gene's nucleotide sequence corresponds to that of the mRNA, such that the opposing DNA strand actually serves as the template. The mRNA itself is

[1]Referring to a three amino acid chain as a tripeptide is a misnomer, because such a chain involves only two peptide bonds.

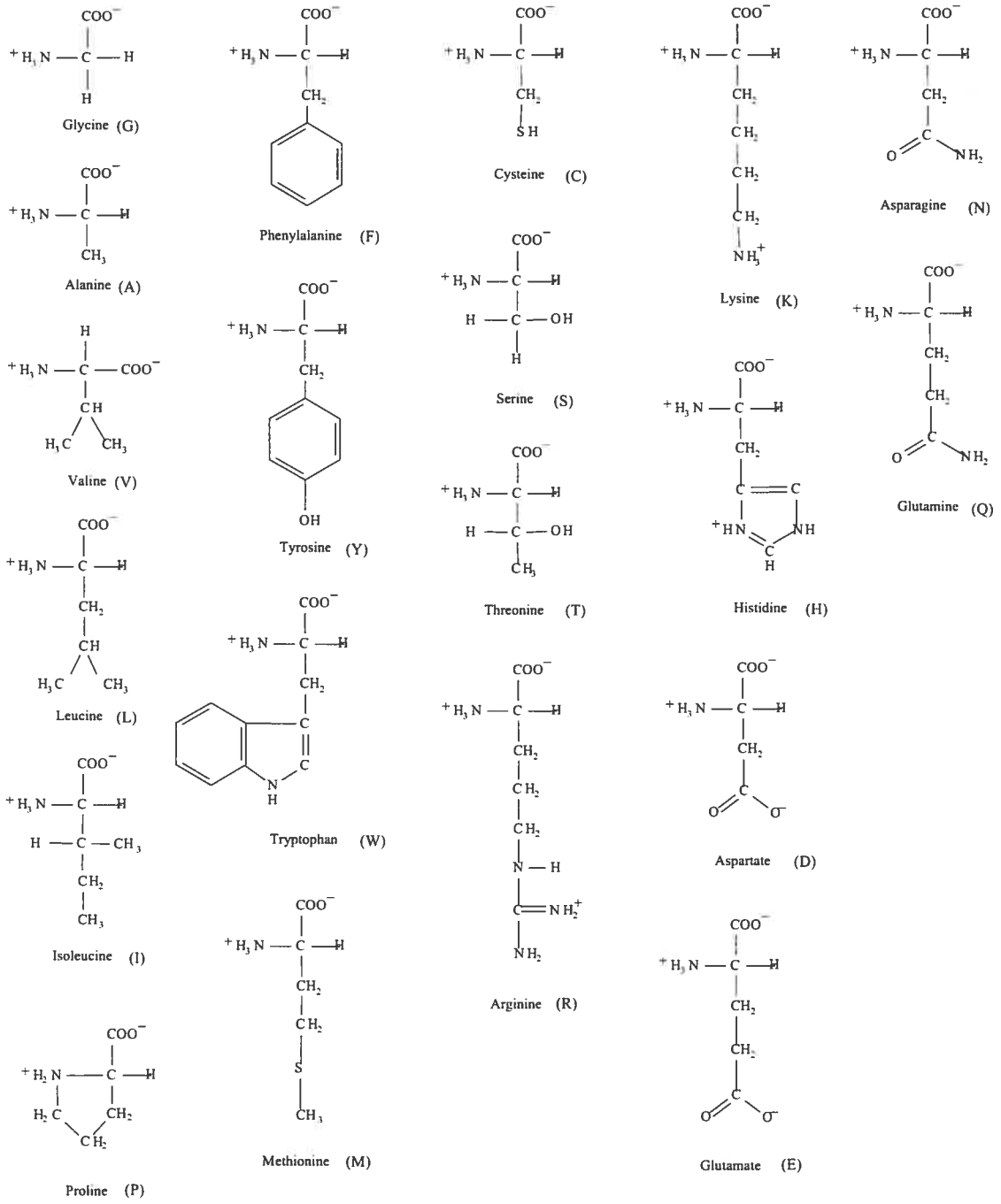

Figure 1.3. Chemical structure of the twenty amino acids (with single letter abbreviations), drawn using ChemTool.

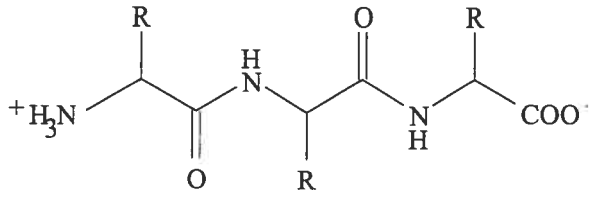

Figure 1.4. Structure of a three amino acid chain, or *tripeptide*. Here, R stands for one of the twenty possible side chains, which characterizes each of the 20 amino acids. The figure was drawn using ChemTool.

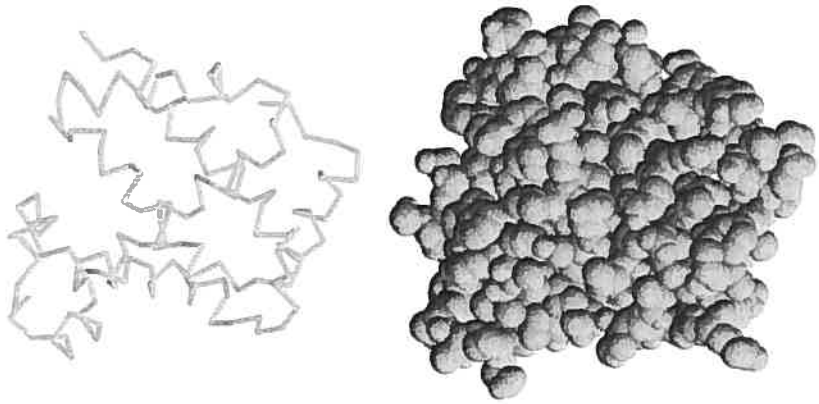

Figure 1.5. Structure of β-globin (PDB code 4HHB), in backbone "stick" (left) representation, and "spacefilling" (right) representation. The figure was generated using RasMol.

a transient macromolecule, which, following some post-transcriptional modifications, interacts with a molecular machinery for *translation* of its sequence into an amino acid sequence. Each nucleotide triplet along the mRNA codes for a specific amino acid; since there are 64 possible triplets, or codons, and 20 amino acids, the code is degenerate (table 1.1). The matching of codon to amino acid is accomplished via adapter RNA molecules called "transfer" RNA, or simply tRNA. A tRNA molecule has three key features: 1) it binds a particular amino acid; 2) it has an affinity for a specific codon, via a complementary nucleotide triplet or *anticodon*; and 3) it binds to the ribosome, which coordinates the overall process of translation. The ribosome is the multi-unit

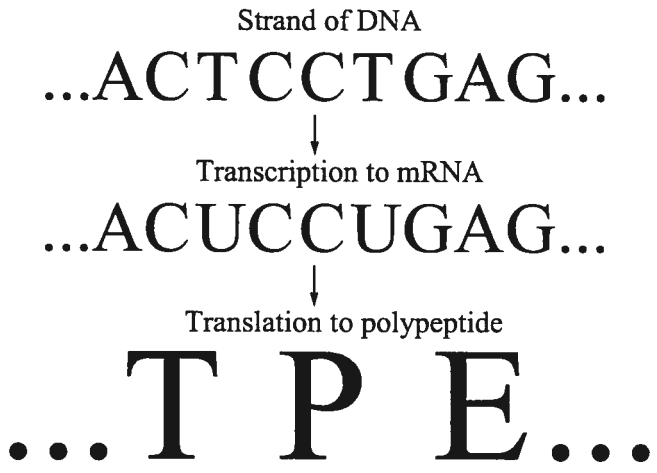

Figure 1.6. The central dogma of molecular biology. Figure made using Gnuplot, availible online at http://gnuplot.info/.

macromolecular scaffold (mainly constituted of RNA) over which the mRNA is passed, one nucleotide triplet at a time, with the complementary anticodon of the appropriate tRNA binding to each successive codon, enabling the formation of peptide bonds in the sequence order originally specified by the DNA. The amino acid chain folds into a particular three-dimensional configuration, and takes on some operational role in the cell[1].

A mapping from DNA sequence to amino acid sequence was thus made possible in the early second half of the 20^{th} century, and many researchers eventually turned to the much more daunting problem of producing a mapping from amino acid sequence to tertiary structure and to protein function. This, as we shall see, constitutes one of the central endeavors of modern molecular biology.

[1]Of course, exceptions to the central dogma abound (e.g., reverse-transcriptase), and the role of RNA in the cell goes well beyond what is described here, as much recent research has shown.

Table 1.1. The standard or "universal" genetic code

TTT,	Phe,	F	TCT,	Ser,	S	TAT,	Tyr,	Y	TGT,	Cys,	C
TTC,	Phe,	F	TCC,	Ser,	S	TAC,	Tyr,	Y	TGC,	Cys,	C
TTA,	Leu,	L	TCA,	Ser,	S	TAA,	stop		TGA,	stop	
TTG,	Leu,	L	TCG,	Ser,	S	TAG,	stop		TGG,	Trp,	W
CTT,	Leu,	L	CCT,	Pro,	P	CAT,	His,	H	CGT,	Arg,	R
CTC,	Leu,	L	CCC,	Pro,	P	CAC,	His,	H	CGC,	Arg,	R
CTA,	Leu,	L	CCA,	Pro,	P	CAA,	Gln,	Q	CGA,	Arg,	R
CTG,	Leu,	L	CCG,	Pro,	P	CAG,	Gln,	Q	CGG,	Arg,	R
ATT,	Ile,	I	ACT,	Thr,	T	AAT,	Asn,	N	AGT,	Ser,	S
ATC,	Ile,	I	ACC,	Thr,	T	AAC,	Asn,	N	AGC,	Ser,	S
ATA,	Ile,	I	ACA,	Thr,	T	AAA,	Lys,	K	AGA,	Arg,	R
ATG,	Met,	M	ACG,	Thr,	T	AAG,	Lys,	K	AGG,	Arg,	R
GTT,	Val,	V	GCT,	Ala,	A	GAT,	Asp,	D	GGT,	Gly,	G
GTC,	Val,	V	GCC,	Ala,	A	GAC,	Asp,	D	GGC,	Gly,	G
GTA,	Val,	V	GCA,	Ala,	A	GAA,	Glu,	E	GGA,	Gly,	G
GTG,	Val,	V	GCG,	Ala,	A	GAG,	Glu,	E	GGG,	Gly,	G

Note.—Each codon is followed by the three letter and single letter abbreviations of the amino acids they encode. Stop codons correspond to a termination of the translation process.

1.4 Computational evolutionary biology

In the early 1960s, the idea that homologous[1] bio-molecules (DNA or amino acid sequences), sampled from different species, could be analyzed to infer their evolutionary history was gaining ground (e.g., Zuckerkandl and Pauling, 1962, 1965). This idea, coupled with the rise of information technologies enabling the automation of such analyses, eventually lead to the modern field of computational evolutionary biology. Indeed, before the advent of molecular data, evolutionary analyses were typically based on

[1] In evolutionary biology, the term homologous refers to similarities between given features that are a result of shared ancestry.

morphological features, subjectively defined into characters, and taking on subjectively defined states. Molecular sequences, on the other hand, lend themselves to a natural discretization: characters are defined as nucleotide, amino acid, or codon sites along the sequence, with each site taking on one of 4, 20, or 61 (excluding stop codons) states respectively. To the mathematically inclined biologist, such data raised numerous questions that could be addressed through direct calculation.

One such biologist was Motoo Kimura (1924–1994). With a strong background in population genetics, Kimura stayed in tune with the developments of molecular biology, calculating the implications of the new data coming out from the field. In 1968, Kimura published calculations claiming that the rate of molecular evolution is much higher than expected under the assumed strength of selection (Kimura, 1968). His conclusion was that many residue changes must be selectively neutral. This idea, later known as the *neutral theory of molecular evolution*, would form the hallmark of much of the rest of his career.

Plainly stated, the neutral theory asserts that many different versions of a molecule are selectively equivalent in a population. In other words, selection is indifferent to these different versions, since each, for whatever reason, performs its biological role equivalently. This idea was not novel. Darwin himself had stated: "variations neither useful nor injurious would not be affected by natural selection [...]" (*Origin*, p. 108). Nonetheless, the neutral theory sparked intense debate regarding the relative importance of neutral drift versus selection. To onlookers, the debate was unfortunately viewed as casting doubt on the validity of natural selection, which was of course not the case. In his later book, Kimura attempts to clarify:

> The neutral theory is not antagonistic to the cherished view that evolution of form and function is guided by Darwinian selection, but it brings out another facet of the evolutionary process by emphasizing the much greater role of mutation pressure and random drift at the molecular level (Kimura, 1983, p. ix).

With some toning down, namely by striking "much greater", this statement would likely be endorsed by most of today's molecular evolutionists.

Another prominent figure in the early years of computational evolutionary biology is Margaret O. Dayhoff (1925–1983). Besides her interests in cataloging and organizing molecular data, she and her co-workers proposed the first empirical model of amino acid sequence evolution, now famously known as the Dayhoff substitution matrix (Dayhoff et al., 1972, 1978). Using the amino acid sequences available at the time, Dayhoff and colleagues devised a counting approach to construct a 20×20 matrix of substitution probabilities over a short evolutionary distance (of, say, 0.01 changes per amino acid site). Their procedure involved several *ad hoc* choices, for accommodating the sparse data sets of the day, and for reducing the possibility that inferred single amino acid replacements may be the result of several unobserved replacements. Nonetheless, the ideas proved inspiring, and highly useful to those interested in inferring evolutionary relations, or *phylogenies*[1].

1.5 Conclusions

Kimura and Dayhoff are but two (arbitrary) examples of the type of research and evolutionary analysis made possible by molecular biology. From about the 1980's onward, novel molecular techniques have made it possible to sequence far greater amounts of DNA and amino acid sequences (fig. 1.7a). Technical advances have also made it possible to resolve three-dimensional molecular structures much more easily and quickly (fig. 1.7b). Over the same years, the capabilities of computing machines have experienced a similar growth trend, sending most practitioners into an ever-lasting overhaul of information technology infrastructures.

[1]The term *phylogeny* comes from the Greek *phyle*, meaning "tribe" or "race", and *genetikos*, meaning "birth".

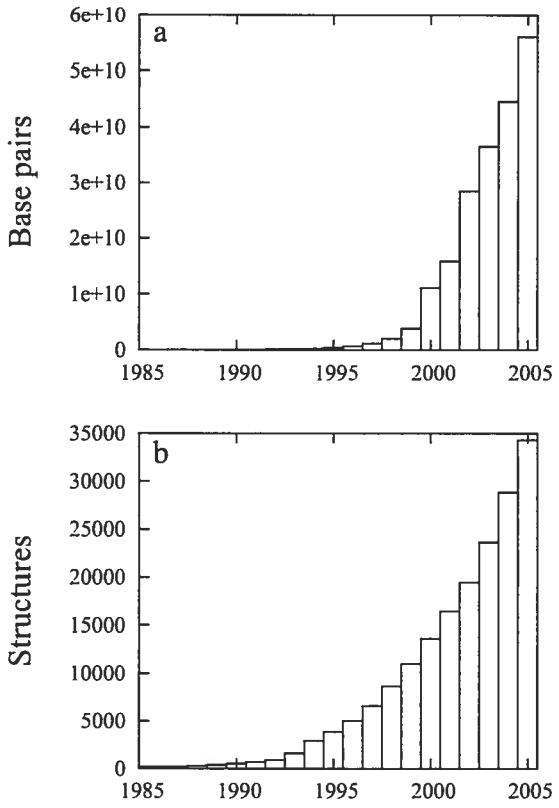

Figure 1.7. Increasing amounts of molecular data. In panel a), the number of base-pairs in GenBank (http://www.ncbi.nlm.nih.gov/Genbank) is displayed over 20 years, spanning 1985–2005. Panel b) shows the number of structures in the PDB (http://www.pdb.org/) over the same period. This figure was made using Gnuplot, as were all other quantitative figures.

Altogether, these developments have driven evolutionary biology into the so-called *genomic revolution*, where questions about the underlying evolutionary process, or about phylogenetic relations, can be addressed based on massive amounts of molecular level data. There has also been a movement away from *ad hoc* methodologies, with many researchers now attempting to devise richer mathematical models of molecular evolution within a sound probabilistic framework. The details of such a framework are the subject of the next chapter.

Chapter 2

Probabilistic phylogenetic analysis

> Suppose you're on a game show and you're given the choice of three doors. Behind one door is a car; behind the others, goats. The car and the goats were placed randomly behind the doors before the show. The rules of the game show are as follows: After you have chosen a door, the door remains closed for the time being. The game show host, Monty Hall, who knows what is behind the doors, now has to open one of the two remaining doors, and the door he opens must have a goat behind it. If both remaining doors have goats behind them, he chooses one randomly. After Monte Hall opens a door with a goat, he will ask you to decide whether you want to stay with your first choice or to switch to the last remaining door. Imagine that you choose Door 1 and the host opens Door 3, which has a goat. He then asks you "Do you want to switch to Door Number 2?" Is it to your advantage [if you wish to maximize the probability of wining the car] to change your choice?

—Re-statement of the Monte Hall problem, from Krauss and Wang, 2003, p. 25.

> Yes, you should switch. The first door has a 1/3 chance of winning, but the second door has a 2/3 chance.

—Marilyn vos Savant, columnist, responding a reader posing the Monte Hall problem.

2.1 Introduction

The famous Monte Hall problem, described in the opening citations, caused a wave of astonishment in the 1990's as a vivid example of people's deficiency in logically

handling uncertainties. Even in this simple case, with a clearly defined system to evaluate, most people's intuition, including that of trained statisticians, fails to correctly account for all relevant information when making a choice, and columnist Marilyn vos Savant had to push through a surprising lengthy series of responses before her readers correctly recognized the solution[1]. When the situation under consideration is more complicated, and when the system under study is not clearly understood, intuition becomes particularly untrustworthy.

Probability theory offers a natural framework for making decisions or inferences, which reduces to deductive logic in cases of complete information (Jaynes, 2003). The Bayesian paradigm in particular is considered a full probabilistic framework, in the sense that it forces the investigator to explicitly state all assumptions during an analysis. This view is based on interpreting probabilities as expressions of our state of knowledge. Gelman et al. (2004) succinctly summarize the framework:

> The process of Bayesian data analysis can be idealized by dividing it into the following three steps:
>
> 1. Setting up of *full probability model*—a joint probability distribution for all observable and unobservable quantities in a problem. The model should be consistent with knowledge about the underlying scientific problem and the data collection process.
> 2. Conditioning on observed data: calculating and interpreting the appropriate *posterior distribution*—the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data.
> 3. Evaluating the fit of the model and the implications of the resulting posterior distribution: does the model fit the data, are the substantive conclusions reasonable, and how sensitive are the results to the modeling assumptions in step 1? If necessary, one can alter or expand the model and repeat the three steps. (p. 3)

The first step is a creative process. Indeed, the creative nature of this step implies

[1]The Monte Hall problem has become a useful tool in cognitive psychology as a means of elucidating mental strategies to problem solving. Krauss and Wang (2003), for instance, point out that the correct solution, changing your choice, can be derived from Bayes' theorem, and that the solution is readily seen when approaching the question from Monte Hall's perspective: "[...] the change from the contestant's perspective to Monte Hall's perspective corresponds to a change from non-Bayesian to Bayesian thinking." (p. 7)

that no general method is available for constructing the basic form of a model. Choices at this step are necessarily arbitrary, and must be evaluated retrospectively in the third step of the Bayesian framework. Given a model M, specified by some high-dimensional parameter vector $\theta \in \Theta$, the second step is purely technical, and can be formalized as an update of our state of knowledge about the hypothesis vector θ before observing any data, the *prior probability*, written as $p(\theta \mid M)$, to our state knowledge after observing the available data, the *posterior probability*, written as $p(\theta \mid D, M)$, and calculated according to Bayes' theorem:

$$p(\theta \mid D, M) = \frac{p(D \mid \theta, M)p(\theta \mid M)}{p(D \mid M)}, \tag{2.1}$$

where $p(D \mid \theta, M)$ is the *likelihood function*, and where

$$p(D \mid M) = \int_{\Theta} p(D \mid \theta, M)p(\theta \mid M)d\theta \tag{2.2}$$

is a normalizing constant, also called the *marginal likelihood* or the *prior predictive probability*. The distribution given by (2.1) is the focus of the second step of the Bayesian framework, whereas the quantity in (2.2) is of interest in the third step.

Until relatively recently, adopting full probabilistic approaches was computationally prohibitive in most contexts. Over the last decade, however, Markov chain Monte Carlo (MCMC) computational techniques have permeated across several disciplines as general and unifying approaches to addressing many of these practical difficulties. The evolutionary analysis of molecular data has greatly benefited from these advances, which have sparked several research programs in population genetics and phylogenetics.

In this chapter, we present the evolutionary context and models in greater mathematical detail, focusing on two different levels of interpretation: 1) the amino acid level, with data consisting solely of aligned amino acid sequences, and 2) the codon level, with data consisting of aligned protein coding nucleotide sequences. We lay out the com-

putational methods used for approximating the probabilities of interests, and offer a few practical examples, of prior structure explorations, of parameter expansion-based MCMC methods, and of simple means of displaying posterior distributions.

2.2 Data

In the present work, the data available consist of aligned sequences of either nucleotides or amino acids, sampled across different species. A phylogenetic tree is used as an account of the evolutionary paths relating the sequences (fig. 2.1). Upon considering such data, one might first object that evolution occurs over populations, and that evolutionary analyses should be based on markers sampled across members of the same species, so as to characterize the variation and evolution of these markers over time. This is indeed a central motivation of population genetics. However, the motivations of phylogenetic analysis can be considered as encompassing those of population genetics, by studying variation across a broader range of genetic diversity, and thus attempting to uncover high-order evolutionary features or patterns that might be too subtle to detect from population level data. In most cases, phylogenetic analyses make the assumption of nul polymorphism[1], and are based on defining a substitution as the fixation of a mutation in the population. Phylogenetic models thus consider a substitution as the elementary event. Loosely speaking, the models focus on long-term evolutionary patterns, by attempting to describe an evolutionary process which, over time, could plausibly have panned out to produce the aligned set of extant sequences. Furthermore, for some types of evolutionary models considered here the link with population genetic theory can be made mathematically explicit (e.g., Thorne et al., 2007), as a consequence of their distinct parameters bearing on mutational features and selective constraints,

[1] In the present context, a polymorphism refers to the occurrence of several different versions of bio-molecule in a given population.

which are combined multiplicatively to specify the overall substitution process.

The alignments used here were selected or constructed on the basis of them being free (or virtually free) of gaps, being of relatively short length (for computational reasons), and consisting of (or encoding for) proteins with one representative having a resolved tertiary structure (for the structural evolutionary models studied in Part II). We refer to data sets using a shorthand indicating the protein type, the number of sequences, and their length, in number of amino acids (or codons, in the case of nucleotide sequences):

- MYO*20-153*: 20 amino acid sequences of tetrapod myoglobin;
- MYO*60-153*: 60 amino acid sequences of mammalian myoglobin;
- MYO*10-153*: 10 amino acid sequences of mammalian myoglobin;
- MYO*4-153*: 4 amino acid sequences of myoglobin from *Physester catodon*, *Orcinus orca*, *Graptemys geographica* and *Chelonia mydas caranigra*;
- PPK*10-158*: 10 amino acid sequences of bacterial 6-hydroxymethyl-7-8-dihydroxypterin pyrophosphokinase;
- FBP*20-363*: 20 amino acid sequences of vertebrate fructose bisphosphate aldolase;
- GLOBIN*17-144*: 17 vertebrate nucleotide sequences of the β-globin gene, described in Yang et al. (2000a);
- LYSIN*25-134*: 25 abalone sperm lysin coding nucleotide sequences, described in Yang et al. (2000b);
- HIV*22-99*: 22 human immunodeficiency virus type 1 protease coding nucleotide sequences, described in Doron-Faigenboim and Pupko (2007).

The first six of these alignments are of our own construction, and are detailed in Appendix A.

2.3 Markovian models of sequence evolution and the likelihood function

Standard phylogenetic models consider the states at the positions of an alignment as the realization of a set of independent Markov substitution processes—one for each site—running along the branches of the tree. The state space, and the definition of a site, depends on the level of interpretation adopted. For instance, when analyzing DNA sequences, the state space consists of the four possible nucleotides, and a site corresponds to a single nucleotide column, or position, in the alignment; each species' sequence constitutes a row of the alignment. For amino acid sequences, the state space consists of the 20 amino acids, and a site is again simply a single amino acid column of the alignment. When analyzing protein coding DNA sequences, and acknowledging the basic coding structure, the state space consists of the 61 sense codon (in the universal genetic code excluding stop codons), with a site defined as a nucleotide triplet (codon) along the sequence. We have the latter two contexts in mind in the following, although the description is general, for any Markov process running over an alphabet of A possible states.

Regardless of the level of interpretation, these processes can be described by a rate matrix, or Markov generator, $Q = [Q_{ab}]$, specifying the instantaneous rate of substitution from state a to state b. Rate matrices are typically constructed from two sets of parameters: a stationary probability vector, otherwise referred to as the equilibrium frequencies, written as $\pi = (\pi_b)_{1 \leq b \leq A}$, with $\sum_{b=1}^{A} \pi_b = 1$, and exchangeability parameters, written as $\rho = (\rho_{ab})_{1 \leq a,b \leq A}$, such that

$$Q_{ab} \propto \rho_{ab}\pi_b, a \neq b \tag{2.3}$$

$$Q_{aa} = -\sum_{b \neq a} Q_{ab}. \tag{2.4}$$

First consider two sequences, denoted s_j and $s_{j_{up}}$, where j_{up} signifies that the sequence is ancestral to j. The models considered in this work are all reversible, such that one may arbitrarily label any one sequence as ancestral to the other. Specifically, the models satisfy the *detailed balance*, given as:

$$\pi_a Q_{ab} = \pi_b Q_{ba}. \tag{2.5}$$

Let s_{ij} and $s_{ij_{up}}$ refer to the specific states at position i of these sequences. If separated by an evolutionary distance (or branch length) λ_j, where the Markov generator has been scaled, say, to express branch lengths as the expected number of substitutions per site, the probability of $s_{j_{up}}$ changing to s_j is computed one position at a time, under the assumption of independence, based on

$$p(s_{ij} \mid s_{ij_{up}}, \theta, M) = \left[e^{\lambda_j Q}\right]_{ab}, \tag{2.6}$$

where $a = s_{ij_{up}}$ and $b = s_{ij}$, θ is the set of parameters[1], and M represents the overall construction of the model. Multiplying (2.6) with the stationary probability π_a, and multiplying across all sites, constitutes the *likelihood* of a particular parameter configuration, under the s_j and $s_{j_{up}}$ sequence pair. Now, suppose we are given P aligned sequences, related according to a given (arbitrarily) rooted phylogenetic tree (which has $2P-3$ branches) with a set of branch lengths $\lambda = (\lambda_j)_{1 \le j \le 2P-3}$. Also, suppose we know the sequence states of each branching point (or internal node) in the tree, in addition to the states in the alignment—we will denote this set of states at position i as $\mathbf{s}_i$. Then, assuming independence between lineages, equation (2.6) can simply be

[1]We have been using θ as a generic hypothesis vector, with a dimensionality and precise configuration implied by the context of the text, and shall continue to do so throughout. Up to this point, for instance, $\theta = \{\lambda_j, \rho, \pi\}$.

expanded to

$$p(\mathbf{s}_i \mid \theta, M) = \pi_{s_{i0}} \prod_{j=1}^{2P-3} p(s_{ij} \mid s_{ij_{up}}, \theta, M), \tag{2.7}$$

where s_{i0} represents the state at position i of the sequence at the root of the tree, labeled as node 0, and $\pi_{s_{i0}}$ accounts for the stationary probability of the Markov process. In practice, however, internal node states are not generally known, and the probability of the data at the i^{th} position (D_i) is thus a sum over all possible $\mathbf{s}_i$:

$$p(D_i \mid \theta, M) = \sum_{\mathbf{s}_i} p(\mathbf{s}_i \mid \theta, M). \tag{2.8}$$

Under the assumption of independence, the overall probability of the data is then computed as a product over all positions:

$$p(D \mid \theta, M) = \prod_{i=1}^{N} p(D_i \mid \theta, M), \tag{2.9}$$

where N is the total number of sites. This also referred to as the likelihood of θ.

Under the simpler types of models considered in this chapter—all of which have been previously proposed by others—the likelihood can be calculated in closed form, exploiting matrix diagonalization routines for computing (2.6) and the pruning algorithm (Felsenstein, 1981) for computing (2.8).

2.3.1 The Dayhoff-like amino acid replacement models

Working with amino acid sequences, the simplest Markovian model treats all states as equivalent, fixing $\pi_b = 1/20$ and, say, $\rho_{ab} = 1$, which we refer to as the POISSON model. Much more commonly, however, Dayhoff-inspired settings are used, such as an updated version proposed by Jones et al. (1992b), referred to as JTT, and the maximum likelihood matrix proposed Whelan and Goldman (2001), referred to as WAG. Although

derived differently, all of these matrices follow the same general motivations: estimate a robust set of values for the parameters π and ρ via an analysis of meta-datasets, and use these parameter values in subsequent phylogenetic analyses.

This is an *empirical* modeling strategy. Several variations follow. For instance, rather than fixing π to the empirically derived values, it is common to treat these as free parameters conditional on the data under study, typically designated by adding the suffix +F to the JTT or WAG acronyms. Also, with the large data sets commonly used today, which may consist of tens of thousands of amino acid positions, the ρ parameters may also be treated as free, referred to as the *general time reversible* (GTR) model[1].

It is also common to combine the above models with the gamma distributed *rates across sites* modeling approach proposed by Yang (1993, 1994). Under this model, referred to as $+\Gamma$, the overall rates of sites are treated as random variables, drawn from a prior statistical law: the gamma distribution of mean 1, and variance α^{-1}. The likelihood function then takes the form of an integral over the statistical law, and α is treated as a free parameter governing its shape. In practice, however, integrating over the gamma distribution is not analytical, and the commonly adopted approximation procedure discretizes the law into a predefined number of classes (typically 4 or 8), reducing the integral into a weighted sum (Yang, 1994).

The central application of these types of models of amino acid sequence evolution is to infer a phylogeny from a set of homologous protein sequences. For instance, using the WAG+Γ model implemented in the PhyML program (Guindon and Gascuel, 2003), the maximum likelihood topology a of the MYO*20-153* data set is displayed in figure 2.1a. By performing a heuristic exploration of the possible topologies, associated branch lengths, and α parameter, PhyML attempts to maximize the likelihood function, and the inference is then based on this maximum likelihood (ML) estimate. Note that the

[1]The acronym GTR is more commonly associated with the nucleotide level model, consisting of six nucleotide relative exchangeabilities, and a four dimensional stationary probability vector. Distinguishing between these is obvious from the context, and so we do not give them separate designations.

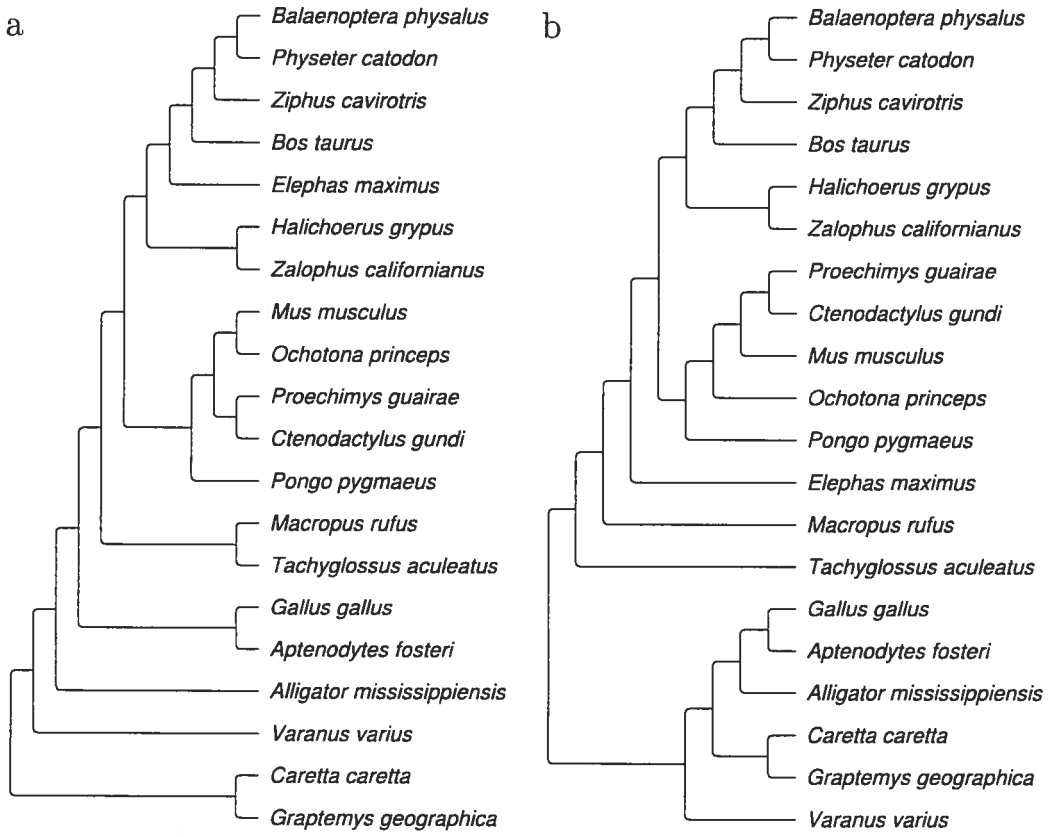

Figure 2.1. Tree topologies for MYO*20-153*. Panel a) displays the maximum likelihood topology obtained by PhyML (Guindon and Gascuel, 2003) under the WAG+Γ model. Panel b) shows a "hand-drawn" topology, which is in closer agreement with accepted groupings. The trees were drawn using TreeGraph, available online at http://www.math.uni-bonn.de/.

tree inferred from such a small data set (of 153 amino acid positions) should not be taken too seriously. Indeed, figure 2.1a has several questionable groupings, such as the position of the red kangaroo (*Macropus rufus*) with the Prototheria (*Tachyglossus aculeatus*), as opposed to a placement as a sister-group to the rest of the Theria; an alternative topology, which might be considered more reasonable, is displayed in figure 2.1b.

Although the probabilistic framework we expound here offers means of dealing with phylogenetic uncertainty, this is not the focus of the present work, and we shall always consider the tree topology as known, based on some external criteria. When studying previously assembled data sets, we use the same topologies used in the previous works, and when studying our own data sets, we use the WAG+Γ maximum likelihood topology, even if it mildly conflicts with established groupings. Our focus is on developing new Markovian models, and exploring their statistical merits; in such contexts, previous studies have found model comparisons robust to slight topological differences, so long as reasonable trees are used (e.g., Yang et al., 2000a; Sullivan and Joyce, 2005). Furthermore, using a fixed tree greatly simplifies the computational devices, for reasons that will become apparent later on.

2.3.2 The GY codon substitution models

Despite the practical success of the amino acid level models described above, the codon level of interpretation offers a theoretically more attractive framework for molecular evolutionary analysis, in actually reflecting basic biological knowledge. The core features of such models include their parameterizations of nucleotide-level mutational properties, and their distinction between synonymous substitution events (that do not imply a change in amino acid) versus nonsynonymous events (implying an amino acid replacement). By accommodating the basic information flow of the central dogma, the models

offer a wide range of relevant biological applications (Yang, 2006).

These could be qualified as *mechanistic* modeling strategies, first explored in evolutionary contexts by Goldman and Yang (1994) and Muse and Gaut (1994). In the present chapter, we shall focus on the widely used codon substitution models in the style of Goldman and Yang (1994), with the entries of the Markov generator given as

$$Q_{ab} \propto \begin{cases} \omega\kappa\pi_b, & \text{if nonsynonymous transition,} \\ \omega\pi_b, & \text{if nonsynonymous transversion,} \\ \kappa\pi_b, & \text{if synonymous transition,} \\ \pi_b, & \text{if synonymous transversion,} \\ 0, & \text{if } a \text{ and } b \text{ differ by 2 or 3 nucleotides,} \end{cases} \tag{2.10}$$

where ω is the nonsynonymous/synonymous rate ratio, κ is the transition/transversion rate ratio, and π_b is the equilibrium frequency of the target codon. This specification of π as a full 61-dimensional vector is often denoted as F61, and we therefore refer to this model as GY-F61.

The computational demands of the codon state space have prevented its wide-spread use for phylogenetic inference. This is mainly because the pruning algorithm (Felsenstein, 1981, for computing 2.8) has a computational time that increases with the square of the alphabet size, but also because the time of matrix diagonalization algorithms (for computing 2.6) increases with the cube of the alphabet size. Rather, the central application of these types of evolutionary models has been to uncover amino acid positions under positive selection, and some of the basic extensions have allowed for heterogeneous nonsynonymous rates across codon sites (e.g., Yang et al., 2000a; Huelsenbeck and Dyer, 2004; Huelsenbeck et al., 2006).

Specifically here, we shall use the *Dirichlet process* apparatus described in Huelsen-

beck et al. (2006). The idea behind this model is to assume that rate heterogeneity is a result of sites coming from a mixture of models, with each component endowed with its own nonsynonymous rate factor. However, under the Dirichlet process prior, the number of components—in this case, the number of selection coefficients—of the assumed mixture is not predetermined, but rather adjusts to the complexity of the data.

2.4 Bayesian MCMC

2.4.1 Plain MCMC

For all models of interest here, the integral in (2.2) has no analytical form. However, modern computing machines and MCMC approaches allow one to sample from the posterior distribution of parameters of interest, without knowing the marginal likelihood, which cancels out in the basic Metropolis-Hastings (MH) kernel (Metropolis et al., 1953; Hastings, 1970): given the current parameter configuration θ, generate a new parameter configuration θ' from the density $q(\theta, \theta')$, and set $\theta = \theta'$ with probability ϑ, where

$$\vartheta = min\left\{1, \frac{p(\theta' \mid D, M)}{p(\theta \mid D, M)} \frac{q(\theta', \theta)}{q(\theta, \theta')}\right\}. \tag{2.11}$$

The factor $\frac{p(\theta'|D,M)}{p(\theta|D,M)}$ is referred to as the Metropolis ratio and $\frac{q(\theta',\theta)}{q(\theta,\theta')}$ is known as the Hastings ratio, correcting for asymmetries in proposal densities. Under certain conditions, repeatedly cycling through these steps forms a Markov chain with (2.1) as its stationary distribution (see, e.g., Robert and Casella, 2004, for a more extensive exposition). In general, the first portion of the chain is discarded—the so-called *burn-in* period—and hypothesis vectors are drawn at regular intervals as the algorithm proceeds. Based on this sample, written as $(\theta^{(h)})_{1 \leq h \leq K}$, expectations are approximated from the usual

Monte Carlo relation:

$$\langle T \rangle = \int_{\Theta} T(\theta) p(\theta \mid D, M) d\theta \tag{2.12}$$

$$\simeq \frac{1}{K} \sum_{h=1}^{K} T(\theta^{(h)}) \tag{2.13}$$

where T is some test statistic of interest, and $\langle . \rangle$ stands for an expectation with respect to (2.1).

The proposal densities $q(\theta, \theta')$ are designed to be easy to implement, and are "tuned" empirically to optimize mixing kinetics[1]. We mention the previously proposed MH mechanisms that are used in this work, and present a brief tuning example in a later section. The mechanisms are:

- ADDITIVE: Treating θ as univariate for the moment, this operator proposes a new value $\theta' = \theta + \delta(U - 1/2)$, where U is a random draw on the uniform [0, 1] interval, and δ is a tuning parameter, with larger values amounting to bolder moves. The Hastings ratio is 1.

- MULTIPLICATIVE: When θ has no constraints except positivity, a new value can be proposed as $\theta' = \theta e^{\delta(U-1/2)}$. The Hastings ratio is θ'/θ.

- DIRICHLET: For multidimensional profile-like parameters, summing to 1 or some constant, this is the update procedure described in Larget and Simon (1999). For instance, updating π for an alphabet of size A would be done by drawing $\pi' = X$, where $X = (\delta\pi_1, \delta\pi_2, \ldots, \delta\pi_A)$. Note that the operator can be applied on a sub-space of π, as explained in Larget and Simon (1999). Also note that for this operator, the lower the value of the tuning parameter, the bolder the move.

[1] In theory, the use of different valid proposal densities should not influence the limiting distribution of the Markov chain. However, different tunings on proposal densities can lead to vastly different sampling behaviors, and tuning is aimed at reducing the number of cycles to "turnover".

- DIRICHLET PROCESS: This operator actually consists of a set of operators, and here is only invoked under the Dirichlet process prior modeling of nonsynonymous rate heterogeneity across sites in the codon context, as described in Huelsenbeck et al. (2006). Supposing H classes of ω factors, updating the Dirichlet process is accomplished by first drawing a set of L temporary classes from $p(\omega_l) = 1/(1+\omega_l)^2$, for $1 \leq l \leq L$; this can be sampled from $\omega_l = \ln U_1 / \ln U_2$, where U_1 and U_2 are two distinct random draws on [0, 1]. Then, taking site i, an update is performed on an auxiliary variable specifying the affiliation of the site to a particular ω class, written as y_i, and which, under the current configuration of the Dirichlet process, ranges over $1 \leq y_i \leq H$. The number of sites affiliated to the h^{th} of H classes is written as η_h. If $y_i = h$ and $\eta_h = 1$, the count of existing classes (H) is decreased by one. Otherwise, η_h is decreased by 1. Then pooling all $H + L$ classes, y_i is reset to the h^{th} class with a probability proportional to $\eta_h p(D_i \mid \theta, \omega_h, M)$, or to the l^{th} class with a probability proportional to $\frac{\tau}{L} p(D_i \mid \theta, \omega_h, M)$, where τ is the "graininess" parameter of the Dirichlet process[1]. The procedure is repeated for all sites. With a given configuration of the Dirichlet process, the values of the H different ω classes are updated based on MULTIPLICATIVE mechanisms, and the τ parameter is updated based on ADDITIVE mechanisms.

2.4.2 Thermodynamic MCMC

With these proposal mechanisms, all of the previously studied models that are included in this work can be implemented, so as to address the second step of the Bayesian framework. If our objective is to compare two models (M_0 and M_1), as part of the third step of the framework, it is interesting to evaluate the Bayes factor (B_{01}), defined as the ratio of their respective marginal likelihoods (Jeffreys, 1935; Kass and Raftery,

[1]This parameter is usually referred to as α in the statistical literature on the Dirichlet process, but using this symbol would be confusing with the shape parameter of the gamma distributed rates model.

1995):

$$B_{01} = \frac{p(D \mid M_1)}{p(D \mid M_0)}. \tag{2.14}$$

A Bayes factor greater than (less than) 1 is considered as evidence in favor of M_1 (M_0). The Bayes factor does not require models to be nested, and intrinsically penalizes for higher dimensional formulations; loosely speaking, averaging the likelihood over the prior distribution implies parameter configurations that induce very low likelihood values, which has the effect of "bringing down the average"; and higher dimensional models tend to have more of such parameter configurations leading to low likelihoods, hence producing a natural Ockham effect. Unfortunately, because the basic MCMC algorithms described above are explicitly designed to avoid computing marginal likelihoods, more elaborate methods are needed.

The model-switch thermodynamic integration method (Lartillot and Philippe, 2006) extends the advantages of MCMC sampling by devising a path linking the posterior distributions of two models. Let θ now represent the union of parameters from both models, some of which may indeed be relevant to both models, while others are only relevant to one of the two. Two models of interest can be connected by defining

$$p(D \mid \theta, M_\beta) = e^{(1-\beta)\ln p(D|\theta,M_0)+\beta \ln p(D|\theta,M_1)}, \tag{2.15}$$

$$p(\theta \mid M_\beta) = e^{(1-\beta)\ln p(\theta|M_0)+\beta \ln p(\theta|M_1)}, \tag{2.16}$$

$$p(\theta \mid D, M_\beta) = \frac{p(D \mid \theta, M_\beta)p(\theta \mid M_\beta)}{p(D \mid M_\beta)}, \tag{2.17}$$

and the Metropolis-Hastings kernel as

$$\vartheta \;=\; min\left\{1, \frac{p(\theta' \mid D, M_\beta)}{p(\theta \mid D, M_\beta)} \frac{q(\theta', \theta)}{q(\theta, \theta')}\right\}. \tag{2.18}$$

For any value $0 < \beta < 1$, the kernel given in (2.18) allows one to sample from a posterior distribution consisting of a partial "morphing" between M_0 and M_1, without knowing $p(D \mid M_\beta)$. The *quasi-static* method described in Lartillot and Philippe (2006) initially sets to $\beta = 0$, and the resulting sampler has the posterior of parameters under M_0 as its limiting distribution. Then, the value of β is regularly incremented by a small value $\delta\beta$ after a set of MCMC cycles, until $\beta = 1$; the sampler finally has the posterior under M_1 as its limiting distribution. Note that here, we do not explore models with different priors on the same parameters, and hence we can dispense with the morphing prior defined in (2.16), substituting it with $p(\theta \mid M_0, M_1)$. When calling Metropolis-Hastings operators on components of θ that are only relevant to M_0, the prior can be reduced to $p(\theta \mid M_0, M_1) = p(\theta \mid M_0)$; and likewise when calling operators on components relevant only to M_1, in which case $p(\theta \mid M_0, M_1) = p(\theta \mid M_1)$. Based on a sample collected along the entire path of posterior distributions, written as $(\theta^{(h)})_{0 \leq h \leq K}$, and with the h^{th} draw associated with β_h ($\beta_0 = 0$, $\beta_K = 1$ and $\forall h,\ 0 \leq h < K$, $\beta_{h+1} - \beta_h = \delta\beta$), the log Bayes factor between M_0 and M_1 can be estimated based on the Monte Carlo relation:

$$\begin{aligned}
\ln B_{01} \;&=\; \ln p(D \mid M_1) - \ln p(D \mid M_0) && (2.19)\\
&=\; \int_0^1 \langle \ln p(D \mid \theta, M_1) - \ln p(D \mid \theta, M_0) \rangle_\beta d\beta && (2.20)\\
&\simeq\; \frac{1}{K}\left[\frac{1}{2}\Big(\ln p(D \mid \theta^{(0)}, M_1) - \ln p(D \mid \theta^{(0)}, M_0)\Big) + \right.\\
&\qquad \Big(\sum_{h=1}^{K-1} \ln p(D \mid \theta^{(h)}, M_1) - \ln p(D \mid \theta^{(h)}, M_0)\Big) +\\
&\qquad \left.\frac{1}{2}\Big(\ln p(D \mid \theta^{(K)}, M_1) - \ln p(D \mid \theta^{(K)}, M_0)\Big)\right], && (2.21)
\end{aligned}$$

where $\langle . \rangle_\beta$ stands for an expectation with respect to (2.17). The overall precision of the method depends on a number of factors, such as the step size ($\delta\beta$), and whether the number of cycles between steps is sufficient to allow the chain to re-equilibrate to (2.17), for instance, but also on the inherent distance between the two models being compared. These issues need be explored in practice, through a progressive tuning that depends on the precise application.

In the next section, we will illustrate the properties of the basic Bayesian MCMC approaches using well-known types of models of amino acid and codon sequence evolution. In later chapters, we will return to the thermodynamic MCMC methods to evaluate these and other models.

2.5 Practical examples

2.5.1 The WAG model

As a first practical example, we applied the WAG model to the MYO*20-153* data set, assuming the tree topology given in figure 2.1a. Under such a model, the only free parameters are the branch lengths λ. Also note that for this particular model, large Monte Carlo samples of high quality are easily obtained within a few hours on a Xeon 2.4 GHz desktop computer, and so we defer the subject of tuning the MCMC to another example[1]. Instead, we go through a simple exploration, for amusement, of the effect of previously proposed prior probabilities on λ on the overall tree length.

We first tried using a uniform prior on branch lengths[2], leading to a posterior mean tree length of 2.41 ± 0.12 substitutions per site.

[1]The computational burden of some of the calculations presented in this dissertation imply over 4 months of CPU time. We mention this here in order to give the reader a general sense of what we mean by computationally "easy" versus what we consider as computationally "challenging".

[2]Strictly speaking, a uniform prior must have bounds, but it is common to explore the behavior of a sampler without such bounds, referred to as an *improper* prior, since it is not defined to integrate to 1.

We next tried the commonly used *Exponential* prior distribution on branch lengths, with a mean determined by a hyperparameter[1] ν, in turn endowed with a truncated uniform hyperprior. The tree length in this case has a posterior mean of 2.15 ± 0.12, and the mean branch length parameter ν has a posterior mean of 0.059 ± 0.011. Note, however, that the posterior mean log-likelihood is $-2,169.5$ for the exponential prior, and $-2,171.7$ for the uniform prior, a difference of only 2.2 log units. This indicates that the likelihood surface may be relatively flat with respect to branch lengths (at least in this region of branch length space). Finally, we tried a type of prior structure suggested in Yang and Rannala (2005) (and also used in Lartillot and Philippe, 2006), where we attribute an *Exponential* hyperprior of mean 1 on ν. This hyperprior does not impact on ν in practice, as we obtained essentially identical distributions as under the uniform hyperprior. This last prior structure, however, has the advantage of being both flexible and proper. Although the prior structure on branch lengths may have important effects on phylogenetic inference *per se* (Yang and Rannala, 2005), we have not found it to have any significant impacts in the fixed tree context of the present work, either in terms of estimated log-Bayes factors, or on the posterior distributions of other parameters. We shall see a case in chapter 4, however, where the prior on a specific parameter does have a significant impact on the log-Bayes factor.

2.5.2 The $+\Gamma$ model via parameter expansion

Model developments can always be cast as revisions in prior structure. For instance, in the $+\Gamma$ models, rather than fixing a prior of 1 for all sites having a rate of 1, as in the classic uniform rates model, a flexible prior structure on rate variation across sites is used, consisting of the gamma distribution of mean 1 and variance α^{-1} (Yang, 1993, 1994). As previously stated, in the ML perspective, the likelihood function takes the

[1] In general, parameters governing the prior laws are referred to as *hyperparameters*, and the priors on them are referred to as the *hyperpriors*.

form of an integral over the gamma law, and the α parameter is adjusted to maximize the likelihood function.

In the present context, one may instead invoke the concept of *parameter expansion* (PX) (Liu et al., 1998) as we explore here. Let us define a rate vector $r = (r_i)_{1 \leq i \leq N}$ specifying the overall rate at each site, with the gamma as a prior law, written as $p_\alpha(r)$ (and α now included in θ). The gamma prior is used in its continuous form, with integration over the law accomplished via MCMC sampling. To see how this works, first note that the marginal and joint probabilities on θ and $\{\theta, r\}$, are related as follows:

$$p(\theta \mid D, M) = \int_{\mathbf{r}} p(\theta, r \mid D, M) dr \tag{2.22}$$

$$\propto \int_{\mathbf{r}} p(D \mid \theta, r, M) p_\alpha(r) dr \tag{2.23}$$

The basic idea underlying parameter expansion is that if a sample $(\theta^{(h)}, r^{(h)})_{1 \leq h \leq K}$ is drawn from the joint distribution $p(\theta, r \mid D, M)$, then, the θ component of this sample, $(\theta^{(h)})_{1 \leq h \leq K}$, is distributed according to $p(\theta \mid D, M)$. Therefore, to obtain a sample from $p(\theta \mid D, M)$, first draw a sample from $p(\theta, r \mid D, M)$, and if only the parameter vector is of interest, discard the r component. This sampling approach can be written more formally, with a MH kernel defined as

$$\vartheta = min \left\{ 1, \frac{p(D \mid \theta', r', M) p_\alpha(r') p(\theta' \mid M)}{p(D \mid \theta, r, M) p_\alpha(r) p(\theta \mid M)} \frac{q(\theta', r', \theta, r)}{q(\theta, r, \theta', r')} \right\} \tag{2.24}$$

In practice, MH operators are typically applied separately on model parameters and auxiliary parameters; the basic sampling module in this case is referred to as the PX module, and is written symbolically as

$$r \mid \theta, D$$

$$\theta \mid r, D$$

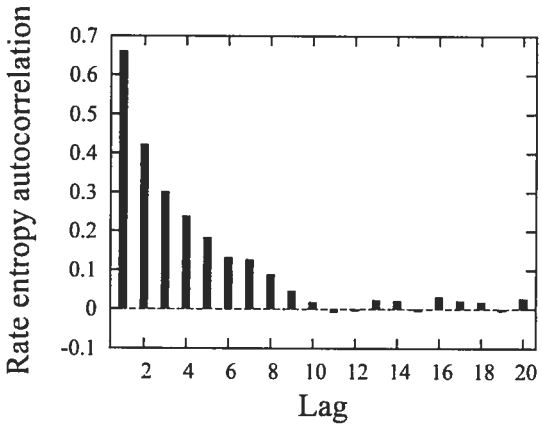

Figure 2.2. Tuning MH updating of rate vector. The figure shows the autocorrelation function of the rate entropy when sampling rate vectors under WAG+Γ.

which is to say that first, a rate vector is sampled conditional on the current parameter vector and the data, followed by a parameter vector sampled conditional on the current rate vector and the data. Also note that we may take the sample of auxiliary parameters seriously, for instance, by constructing the posterior distribution of a site-specific rates (Mateiu and Rannala, 2006).

We ran a PX-based MCMC sampler under the WAG+Γ model, assigning a uniform prior on α. Multiplicative operators are applied on rates, and additive operators are applied on α. We take this as an opportunity to illustrate one way of determining a suitable tuning of MH updates. Specifically, we begin by assuming that the rest of our sampler has already been tuned for sampling over other parameters (branch lengths, and ν), and that we now wish to incorporate sampling for the $+\Gamma$ model, and tune number of MH updates on rates so as to decorrelate successive draws.

Figure 2.2 displays the *autocorrelation function* of the rate vector entropy[1]. More precisely, we repeatedly computed the autocorrelation of the rate entropy on samples of 100 successive draws from the posterior distribution, but with an increasing number of cycles, or *lag*, between each draw. In this case, a MCMC cycle consists of one MH update to the rate of each site. Based on this plot, one would estimate that at least

[1]Writing $p_i = r_i / \sum_i r_i$, the rate entropy is $-\sum_i p_i \ln p_i$.

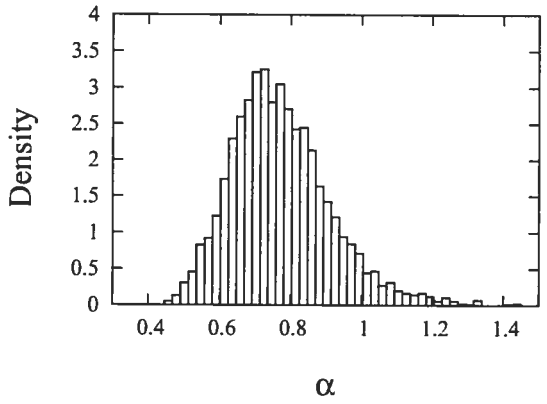

Figure 2.3. Posterior density plot of α, approximated using MH sampling.

10 cycles should be done in order to effectually be sampling independent rate vectors. Of course, tuning must also be validated by running several independent calculations, checking that the mean and variance of parameter values match closely across chains; this is part of our basic set of checks, conducted following pilot runs for tuning[1].

We collected a sample of 1,000 draws from the posterior, and display the distribution of α in figure 2.3. The posterior distribution is nearly identical when using an exponential hyperprior of mean 1 on α (not shown). The posterior distribution of α, centered around 0.7, suggests a pronounced rate heterogeneity across sites. This is a property now observed for many data sets (Yang, 1996, 2006).

2.5.3 The GY-F61 model

We next turn to the mechanistic model specified above as the GY-F61. We explored this modeling framework using the 17 vertebrate sequences of the GLOBIN*17-144* data, described in Yang et al. (2000a), as well as the tree topology used therein. We used an exponential prior on branch lengths, with a mean ν, itself endowed with an exponential hyperprior of mean 1. For κ, we used the prior $p(\kappa) = 1/(1+\kappa)^2$, the ratio of two

[1]It is still an open question whether or not it is possible to design the "holy grail" of MCMC samplers, which would not require such pilot run tuning steps.

Table 2.1. Posterior expectations under the GY-F61 model.

log-likelihood	−3656.83±6.66
tree length	7.17±0.35
ν	0.23±0.05
κ	1.52±0.15
ω	0.22±0.02
π-entropy	3.86±0.02

Note.—The tree length is in expected number of substitutions per codon site.

independent draws from an exponential distribution, as proposed in Huelsenbeck and Dyer (2004), and likewise for ω (as a constraint on the DP model described above to $H = 1$, i.e., a single global nonsynonymous rate factor for the entire alignment).

Following the usual tuning procedures, we obtained a sample of 1,000 draws from the posterior distribution, and produced summarizing statistics in table 2.1 as well as graphical displays of the substitution model parameters in figures 2.4 and 2.5. These are simply meant to illustrate the ways in which posterior distributions can be summarized.

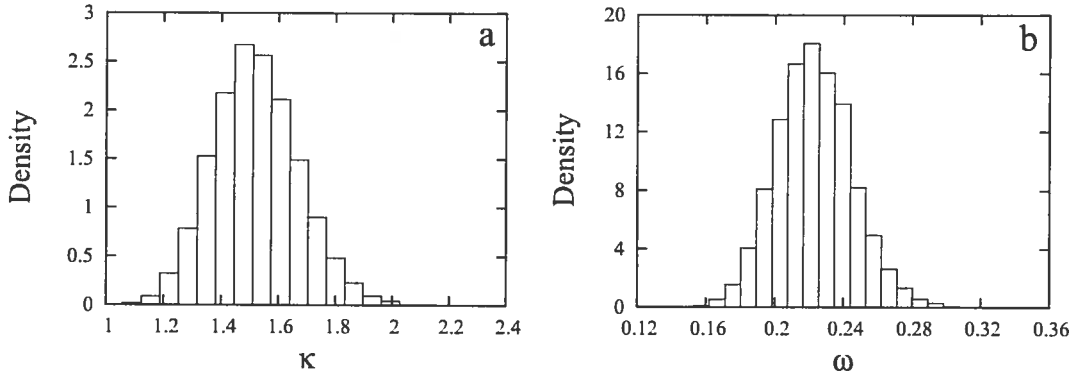

Figure 2.4. Posterior distributions of κ and ω for the GLOBIN*17-144* data set.

2.5.4 The Dirichlet process on ω

Huelsenbeck et al. (2006) have proposed a flexible extension to this model, based on

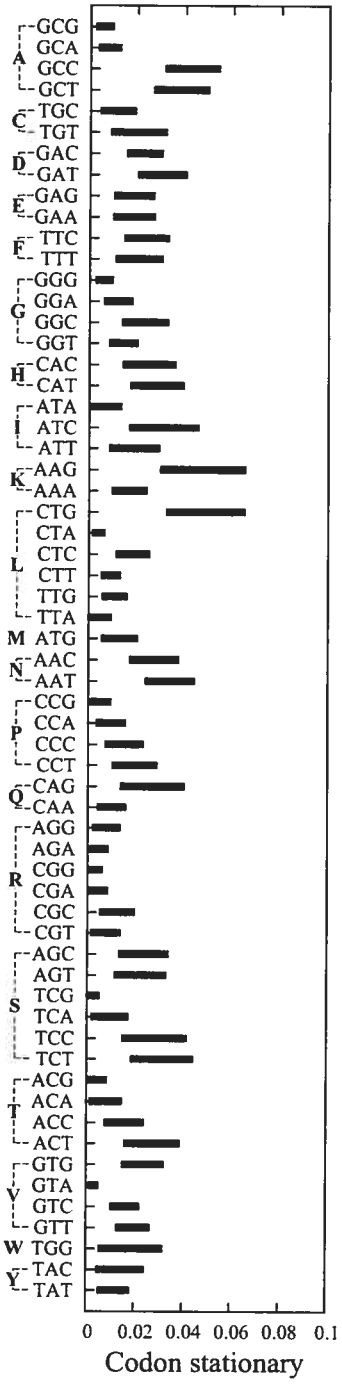

Figure 2.5. Posterior 95% credibility intervals of codon stationary probabilities for the GLOBIN*17-144* data set, sorted according to amino acids.

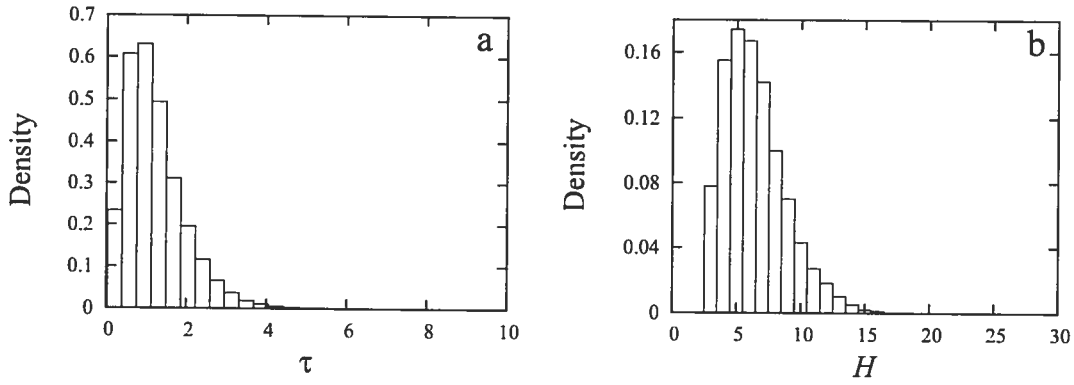

Figure 2.6. Posterior distributions of τ and H under the GY-F61-DP model, for the GLOBIN*17-144* data set.

the Dirichlet process apparatus described above. We explored this model as well, using the same prior structure as Huelsenbeck et al. (2006), except that we endowed hyperparameters τ and ν—respectively the "graininess" parameter of the Dirichlet Process and the mean of the exponential prior on branch lengths—with exponential priors of mean 1.

Huelsenbeck et al. (2006) performed analyses under the Dirichlet process prior by systematically fixing τ to predefined values. Here, given that we treat τ as a free parameter, we inspected its posterior distribution, as well as that of the number of selection coefficients (H). The results are displayed in figure 2.6. Of particular interest is the distribution of H, which is situated at relatively low values, in comparison with the overall length of the alignment, but is still at consistently higher values than the common usage of finite mixture models of the same type, which are typically fixed at $H = 3$ (Yang et al., 2000a).

Finally, as described in Huelsenbeck et al. (2006), we computed site-specific probabilities of positive selection $p(\omega > 1)$ under the Dirichlet process, simply as the proportion of draws from the posterior in which a site is found to be affiliated to a class having $\omega > 1$. Confirming previous studies (e.g., Yang et al., 2000a; Huelsenbeck et al., 2006) most positions of the GLOBIN*17-144* data set appear to be under strong purifying

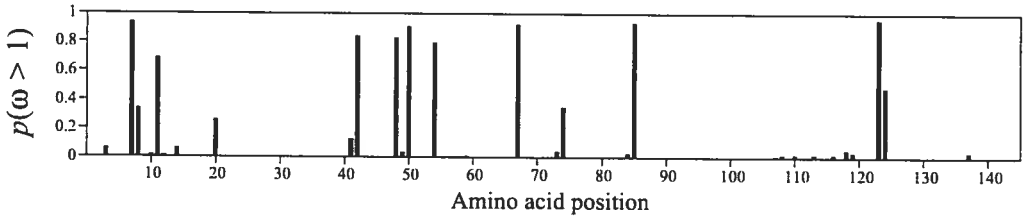

Figure 2.7. Posterior probability of each site being under positive selection for the GLOBIN*17-144* data set.

selection (fig. 2.7).

2.6 Evaluating models via phenomenological benchmarking

The practical examples presented above are mainly focused on the second step of the Bayesian framework—computing the posterior distribution. There are several ways of engaging the third step of the framework—model evaluation—and these different approaches constitute a main subject of study and debate in modern statistical theory. In this chapter, we have touched upon methods for computing Bayes factors which have recently been proposed as versatile tools for model ranking, and which we will make use of extensively in this work.

One of the reasons for this recent interest in rigorous model comparisons comes from the observation that even under conditions of very large data sets, consisting of tens of thousands of residues, phylogenetic reconstruction artifacts[1] are still observed in some cases (Philippe et al., 2005), which implies that the models used are too grossly mis-specified. An active research direction to address these issues consists in devising new evolutionary models, which more reasonably acknowledge molecular evolutionary

[1] A phylogenetic reconstruction artifact is an inferred tree topology that is obviously wrong, based on some other knowledge. Phylogenetic contradictions resulting from slightly different choices in data set construction are also referred to as reconstruction artifacts.

patterns, and thus exhibit greater robustness in wake of difficult phylogenetic questions (e.g., Buckley et al., 2001; Brinkmann et al., 2005; Lartillot et al., 2007). An important part of the assessment of such new models consists of measuring their statistical fit to real data, as we have discussed.

Beyond the applications to phylogenetic understanding, the development of better models is hoped to elucidate and quantify the importance of different aspects of molecular evolution (Pal et al., 2006). With the objective of acquiring a deeper understanding of molecular evolution comes the question of the most suitable level of interpretation to be adopted. The codon level of interpretation is far more attractive from this standpoint, in enabling one to explore parameterizations that disentangle the different factors bearing on the overall evolutionary process (Thorne, 2007). As we have mentioned, these models are often referred to as mechanistic in approach. By this, we mean that the models recognize basic biological understanding, in terms of an underlying mutational process, with selective forces acting at higher levels. Here, we also use the phrase *mechanistic modeling* to refer to attempts at formulating an account of deeper causes, which would explain some of the observed features.

Another modeling perspective is often referred to as *phenomenological*. The term phenomenological has several different meanings. In science, generally, a phenomenological model is one that is not directly derived from theory, but rather provides a preliminary account of some observed tendency or feature of the data by attributing parameters directly to the aspects in question. Phenomenological approaches are generally motivated by a lack of understanding of underlying causative relations, or by the practical applications that they enable. The approach often explicitly omits some data, or fails to incorporate basic knowledge, typically because it appears too difficult in preliminary model explorations. Empirical modeling approaches go one step further by pre-fitting parameters to large-scale observations. The WAG model used above is of

this type; it omits the nucleotide data altogether, and is derived by applying the ML principle under a large meta-data set.

Of course, in practice, models are never strictly phenomenological, since their general form is derived from some basic understanding of the case at hand. In the particular case of the WAG model, for instance, the form of the model attempts to account for biophysical similarities between amino acid pairs, through the set of exchangeability parameters ρ, while capturing global amino acid propensities through the set of stationary probabilities π. Furthermore, aspects of the codon level models used above may also be cast as phenomenological, as is most clearly the case concerning ω: we have good reason to believe that nonsynonymous rates might be mediated—at least partially—by structural constraints, such that, for instance, a given site might have a very low nonsynonymous rate, in comparison with other sites, as a result of being involved in a set of interactions crucial to establishing the functional shape of a protein. At this point, however, the phenomenological standpoint consists of ignoring this understanding, and simply provides a preliminary account of the nonsynonymous rates, with the richest model here consisting of the Dirichlet process as a so-called *infinite mixture* across sites.

We suggest that phenomenological/empirical modeling approaches can provide pertinent references in the initial exploratory stages of a new modeling approach, as part of a *phenomenological benchmarking* strategy that we propose in this dissertation. Phenomenological benchmarking is mainly concerned with the third step of the Bayesian framework, but in some sense, also ties into the first step of the framework—in subsequent cycles of development—by concretely indicating model weaknesses, and thus informing new model constructions. It is meant as an assessment of the ability of a new mechanistic model to adequately account for basic phenomenological observations. Loosely speaking, the motivation of a new mechanistic model will be to generalize ex-

isting phenomenological descriptions of some process or part of the world, under which well-known features would emerge as basic resultants. Phenomenological benchmarking is the strategy employed to evaluate if this is indeed the case. If so, we consider that progress has been made, and that we have formed a better or more precise representations of our current state of knowledge. If not, we may seek to improve the mechanistic description, or reconsider its basic form.

2.7 Conclusions

The Bayesian paradigm provides a flexible and attractive framework for formalizing phylogenetic analysis. Although our survey is not representative of all of the developments currently under way, we have described up-to-date models of molecular evolution at different levels of interpretation, and the general motivations associated with each. In the second part of this work, we build on the types of evolutionary models presented above, in an exploration of novel strategies allowing for a class of models with dependence between amino acid sites due to a protein's tertiary structure (Robinson et al., 2003). We touch all three steps of the Bayesian framework, including our phenomenological benchmarking strategies, at both amino acid and codon levels of interpretation. We begin with the amino acid level—over chapters 3 and 4—assessing how the dependence models compare with empirical amino acid replacement matrices, and with the $+\Gamma$ model. The results from chapter 4 motivate the development of a new statistical framework for relating the compatibility of an amino acid sequence with a given tertiary structure—presented in chapter 5. In chapter 6, we explore more economical computational methods for approximating posterior distributions and marginal likelihoods, which could render future calculations more tractable, and thereby enabling a pipeline of development for different forms of sequence-structure measurements. In chapter 7, we return to the biologically motivated codon level of interpretation, and apply the

framework anew to construct and evaluate a large set of models, but still assuming independence between sites. Finally, in chapter 8, we present our progress in the study of the codon site-interdependent models, which suggest that evolutionary modeling strategies could be extended along three different conceptual levels: 1) parameters describing the underlying mutational process, operating at the nucleotide sequence level; 2) parameters accommodating global codon preferences; and 3) parameters bearing on the overall compatibility of the encoded amino acid sequence with a given (coarse-grained) tertiary structure.

Part II

Revising Modeling Assumptions

Chapter 3

Site-interdependent phylogenetics

3.1 Introduction

In the last few years, several modeling advances have been made, beyond those of the phylogenetic models described up to this point. The general strategy is to propose biologically motivated parameterizations, relaxing the assumptions of standard models—such as the assumption of stationarity (e.g., Galtier and Gouy, 1998), or of homogeneity in the substitution process across sites (e.g., Lartillot and Philippe, 2004; Pagel and Meade, 2004)—without inducing computationally intractable formulations. This last condition in particular has been the main justification for the assumption of independence between sites, which persists in most models currently applied.

Obviously, the assumption of site independence is not biologically sound; as mentioned in chapter 1, different positions of an amino acid chain form complex networks of interactions, important to the overall structure adopted by a protein. Means of relaxing this assumption in evolutionary models have been pursued, usually with correlations or dependence introduced between a limited number of sites (Felsenstein and Churchill, 1996; Siepel and Haussler, 2004), or considered for a limited number of sequences (Jensen and Pedersen, 2000; Pedersen and Jensen, 2001; Robinson et al., 2003).

As previously mentioned in the last chapter, we are particularly interested in the modeling ideas of Robinson et al. (2003), who have introduced sampling techniques that allow for a general dependence between codons. With their sampling procedure, which is applicable to pairs of coding nucleotide sequences, one can consider the stochastic process underlying the evolution of a sequence as a whole, so that the probability of a given substitution, at any given time and at any site, depends, in principle, on the states at all other positions.

Robinson et al. (2003) attempt to capture site interdependencies using an empirical energy function, otherwise known as a *statistical potential*, derived in the context of protein threading (e.g., Jones et al., 1992a). Such potentials are meant to provide an estimate of the compatibility of an amino acid sequence with a given protein structure, so that the differences in compatibility, before and after inferred amino acid replacement events, influence the probability of an evolutionary scenario. This modeling approach is computationally bold, but provides an attractive mechanistic description of molecular evolution; the codon substitution process is formulated as combination of a mutational parameterization, at the DNA level, with an evaluation of the phenotypic effects of mutations, which are considered for the overall amino acid sequence. In line with the theoretical objectives of population genetics, their evolutionary model explicitly relates genotype to the fitness of the corresponding phenotype.

This is a clear example of a new mechanistic modeling strategy; as a byproduct of the explicit structural modeling, the potential could, in principle, account for observed rate heterogeneity, or account for uneven amino acid exchangeabilities, and possibly more complex features as well. However, the suitability of such a model depends on how well one can approximate the overall fitness of a given amino acid sequence. In the case of the model proposed by Robinson et al. (2003), the use of this type of potential was meant as a proof-of-concept investigation of their novel statistical methodology, and

the extent to which the potential actually captures evolutionary features remains to be explored. Indeed, there is some cause for concern: protein fold prediction potentials (e.g., Jones et al., 1992a; Bastolla et al., 2001) were designed to optimally distinguish which conformation a given sequence is likely to adopt, whereas Robinson et al. (2003) use a potential under a fixed conformation, attempting to distinguish which sequences would be suitable to it.

In this chapter, we further explore methodologies and approaches proposed by Robinson et al. (2003), re-formulating their model directly at the level of amino acids. In so doing, we relinquish the theoretically attractive description of molecular evolution at the level of nucleotide sequences. However, the amino acid-level framework will be used to investigate if statistical potentials can render expected features of amino acid sequence evolution, with rate heterogeneity and amino acid exchangeabilities constituting our two basic phenomenological benchmarks. First, however, we must set up the precise models and computational devices. Our objective here is simply to contrast the use of a statistical potential (Bastolla et al., 2001) in combination with either a flat set of amino acid exchangeability parameters (POISSON) or an empirically derived set (JTT) (Jones et al., 1992b), and explore how different combinations may impact on posterior distribution of parameters. In addition, we generalize the sampling scheme proposed by Robinson et al. (2003) to multiple sequences. We apply the methods to three data sets, and prospect the possibility of applying the approach to the comparison of different tree topologies.

3.2 Material and methods

3.2.1 Data sets, trees, and protein structures

We used the PPK*10-158*, MYO*10-153*, and MYO*4-153* data sets. We apply a simple structure representation based on a *contact map* (see below). The contact map is derived from a reference structure, determined by X-ray crystallography for one of the sequences included in the data set (PDB accession numbers 1HKA and 1MBD).

3.2.2 Site interdependent notation

A few brief notational remarks are needed for clarifying the site interdependent framework. As before, data sets (D) consist of alignments of P amino acid sequences of length N, assumed related according to a particular phylogenetic tree. The tree is rooted arbitrarily, as all models considered here are reversible. We use i to index positions of a sequence, and j to specify the nodes, with a node having the same index as the branch leading to it, with the exception of the root node, which has index 0 ($0 \leq j \leq 2P-3$). We specify the sequence at node j as s_j (with s_0 being the sequence at the root node, which we place at a leaf node, i.e., an observed sequence from the alignment), and a particular amino acid state at position i in this sequence as s_{ij}—in other words, the absence of the i index indicates that the sequence is referred to globally (considering its entire length). The sampling methods described below utilize a demarginalization, or data augmentation, method requiring the specification of a detailed substitution mappings over the entire tree. We write the set of branch specific substitution mappings as $\phi = (\phi_j)_{1 \leq j \leq 2P-3}$. The total number of substitutions along a branch is written as z_j ($z_j \geq 0$). We index substitution events as k ($k \leq z_j$) and refer to the time of an event on branch j as t_{jk}. A substitution event alters a single site of the sequence, at position σ_{jk}. When specifying the series of substitution events occurring on a branch

j, let s_{jk-1} and s_{jk} represent the sequence states before and after substitution event k. Note that when $k = 1$, we let $s_{jk-1} = s_{j_{up}}$, where j_{up} is the immediate ancestral node of j. Finally, when $k = z_j$ we let $s_{jk} = s_j$.

3.2.3 Structural fitness approximations

We used the knowledge-based protein energy function described in Bastolla et al. (2001) to estimate the structural fitness of a sequence in a given three-dimensional structure. Our use of the energy function is straightforward. Given a PDB file, one computes the distances between all atoms of all amino acids. As defined by Bastolla et al. (2001), two amino acids are said to be in contact if any of their heavy atoms (atoms other than hydrogen) are at a distance of 4.5 Å or less (contacts due to sequential proximity, within three positions or less, are ignored). As such, the structure of a protein can be represented as a contact map. The contact map of a protein structure of length N is an $N \times N$ matrix $\Delta = (\Delta_{ii'})_{1 \leq i < i' \leq N}$, with elements

$$\Delta_{ii'} = \begin{cases} 1 & \text{if amino acids at sites } i \text{ and } i' \text{ are in contact,} \\ 0 & \text{otherwise, or if } |i - i'| \leq 3. \end{cases} \tag{3.1}$$

Given the contact map, the pseudo-energy of a sequence is calculated as:

$$E_s = \sum_{1 \leq i < i' \leq N} \Delta_{ii'} \epsilon_{s_i s_{i'}}. \tag{3.2}$$

where $\epsilon = (\epsilon_{ab})_{1 \leq a,b \leq 20}$ are the coefficients of the amino acid pair potentials of Bastolla et al. (2001).

As crude first efforts, we impose the same structure over the tree by using the same contact map on all sequences, both observed and inferred.

3.2.4 Evolutionary models

In order to build a site-interdependent model directly at the amino acid level, we first note that the independent Markov processes operating at each site, specified by a 20×20 infinitesimal generator Q, can equivalently be considered as a single Markov process, whose state space is now the set of all sequences of length N. There are 20^N such sequences, and thus, the matrix of this Markov process will be a $20^N \times 20^N$ matrix R:

$$R_{ss'} = \begin{cases} 0 & \text{if } s \text{ and } s' \text{ differ at more than one position,} \\ Q_{ab} & \text{if } s \text{ and } s' \text{ differ only at site } i,\ s_i = a \text{ and } s'_i = b, \end{cases} \tag{3.3}$$

with diagonal entries given by the negative sum of the off-diagonal entries. With the formulation of equation (3.3), it is possible to introduce a site-interdependent criterion: the pseudo-energy before and after an amino acid substitution. The new matrix R is then

$$R_{ss'} = \begin{cases} 0 & \text{if } s \text{ and } s' \text{ differ at more than one position,} \\ Q_{ab} e^{\beta(E_s - E_{s'})} & \text{if } s \text{ and } s' \text{ differ only at site } i,\ s_i = a \text{ and } s'_i = b, \end{cases} \tag{3.4}$$

where β acts as a parameter weighting the pseudo-energy difference's impact on the rate of substitution. When $\beta = 0$, the model simplifies to the usual site-independent model specified in (3.3). However, when $\beta \neq 0$, the substitution process can no longer be decomposed into a set of N independent processes, since the pseudo-energy measure considers the entire amino acid sequence[1]. We use the suffix +BAS to indicate the

[1]It is conventional practice to express branch lengths in terms of the expected number of substitutions per site. To obtain such a scaling, the rate matrix, here denoted as R, must be properly normalized. Formally, the normalizing constant is

$$Z_R = -\frac{1}{N} \sum_s p(s \mid \theta) R_{ss} \tag{3.5}$$

$$= -\frac{1}{N} \langle R_{ss} \rangle \tag{3.6}$$

model with statistical potentials ($\beta \neq 0$).

3.2.5 Priors

We treated the parameters $\pi = (\pi_b)_{1 \leq b \leq 20}$, comprised in the matrix Q, as free parameters, indicated using the +F suffix. The overall prior structure used in this chapter is $\pi \sim Dirichlet$, $\beta \sim Uniform[-5, 5]$, and $\lambda_j \sim Uniform[0, 100]$.

3.2.6 Likelihood function

As previously discussed, conventional models generally invoke pruning-based likelihood calculations (Felsenstein, 1981), and compute a finite-time transition probability matrix by rate matrix exponentiation, computing the likelihood by summing transition probabilities for all possible internal node state configurations. Here, given the order of R ($20^N \times 20^N$), an equivalent calculation is not tractable. As an alternative, Robinson et al. (2003) proposed the use of a *data augmentation* (DA) framework, based on substitution mappings. Given a hypothesis vector $\theta \in \Theta$ under model M, the probability of going from a given sequence to another over branch j, and through a specific

where the angular brackets $\langle \cdot \rangle$ represent an expectation with respect to $p(s \mid \theta)$. The sum in (3.5) is over 20^N terms, and calculating it explicitly is not tractable. It can, however, be approximated via MCMC sampling based on a sample of K sequences, written as $s^{(1)}, s^{(2)}, \cdots s^{(K)}$, drawn from $p(s \mid \theta)$ using the Gibbs sampling procedure described in Robinson et al. (2003). Given this sample, the normalizing constant can be estimated as

$$Z_R \simeq -\frac{1}{N}\frac{1}{K}\sum_{h=1}^{K} R_{s^{(h)}s^{(h)}}. \tag{3.7}$$

With this estimate of Z_R, the non-zero, non-diagonal entries in R become $\frac{1}{Z_R} Q_{ab} e^{\beta(E_s - E_{s'})}$. A simpler alternative, which we found much more convenient in practice, is to leave R unnormalized, and rather monitor the actual number of substitutions in the mappings of our sample from the posterior distribution.

substitution history ϕ_j, can be calculated as

$$p(s_j, \phi_j \mid s_{j_{up}}, \theta, M) = \left(\prod_{k=1}^{z_j} R_{s_{jk-1}s_{jk}} r_{\sigma_{jk}} e^{-(t_{jk}-t_{jk-1})\Upsilon(s_{jk-1})} \right) \times e^{-(\lambda_j - t_{jz_j})\Upsilon(s_{jz_j})}, \tag{3.8}$$

where $\Upsilon(s_{jk-1}) = \sum_{i=1}^{N} \sum_{s_i'} R_{s_{jk-1}s_i'} r_i$ represents the *rate away* from sequence s_{jk-1}, with the inner sum being over the 19 sequence states that differ with s_{jk-1} at position i, and where r_i is the rate at site i (but in this chapter, we omit this level of complication, fixing all $r_i = 1$ for all i).

The likelihood computations also require the probability of the sequence at the root of the tree:

$$p(s_0 \mid \theta, M) = \frac{1}{Z} e^{-2\beta E_{s_0}} \prod_{i=1}^{N} \pi_{s_{i0}}, \tag{3.9}$$

with Z being the associated partition function (normalizing "constant")

$$Z = \sum_{s} e^{-2\beta E_s} \prod_{i=1}^{N} \pi_{s_i}, \tag{3.10}$$

summing over all possible sequences of length N. Assuming lineages evolve independently, the product of (3.8) over all branches, along with the probability in (3.9), yields the overall *augmented likelihood* function:

$$p(D, \phi \mid \theta, M) = p(s_0 \mid \theta, M) \prod_{j=1}^{2P-3} p(s_j, \phi_j \mid s_{j_{up}}, \theta, M). \tag{3.11}$$

3.2.7 Markov chain Monte Carlo sampling

Our MCMC procedure consists of using the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) to define a Markov chain with the posterior probability as its stationary distribution, by updating both mappings and parameters; assuming a current

state (θ, ϕ), an update to a new state (θ', ϕ') is proposed according to $q(\theta, \phi, \theta', \phi')$, and accepted with a probability ϑ:

$$\vartheta = min\left(1, \frac{p(\phi', \theta' \mid D, M)}{p(\phi, \theta \mid D, M)} \frac{q(\theta', \phi', \theta, \phi)}{q(\theta, \phi, \theta', \phi')}\right). \tag{3.12}$$

Most implementations, our own included, apply MH operators separately on model parameters and data augmentations, with the DA sampling module written symbolically as

$$\phi \mid \theta, D$$

$$\theta \mid \phi, D$$

As in the case of the PX module described in chapter 2, the effect of cycling over this module is a sample of parameter vectors distributed according to $p(\theta \mid D, M)$, and is strictly equivalent to what we would obtain if we had access to the integrated (over mappings) likelihood function. We describe the MH operators in detail below.

3.2.7.1 Proposing mappings

Substitution mappings are proposed using a model that assumes independence between sites, here denoted Q^*. We set the amino acid relative exchangeability to those of the underlying site-interdependent model, and the amino acid frequencies to the empirical values observed in the alignment. Under this model, we used the method proposed by Nielsen (2002) for drawing site-specific mappings, as part of three MH operators:

- BRANCHHISTORY: This first type of move randomly selects a branch j and a set of positions, denoted collectively as ζ_{BrHis}. For each site selected, a new substitution mapping is drawn using Nielsen's method, given the states at the ends of the

branch[1]. The move is then accepted with a probability given in (3.12). The corresponding Hastings ratio is:

$$\frac{q(\phi', \phi)}{q(\phi, \phi')} = \prod_{i \in \zeta_{\text{BrHis}}} \frac{p(s_{ij}, \phi_{ij} \mid s_{ij_{up}}, \lambda_j, r_i, Q^*)}{p(s_{ij}, \phi'_{ij} \mid s_{ij_{up}}, \lambda_j, r_i, Q^*)} \tag{3.13}$$

- NODESTATE: This second move randomly selects an internal node j and a set of sites, denoted collectively as ζ_{NdSt}. The move then re-samples the amino acid states of selected sites at node j, again using the Nielsen approach. Having re-sampled the states, the move also re-samples a substitution mapping for each of the selected positions along the three branches connected to j, and acceptance of this overall update is again based on (3.12). The corresponding Hastings ratio is the same as above, but multiplied over the three branches in question.

- TREEHISTORY: This last move randomly selects a set of sites, denoted as ζ_{TrHis}, and re-samples all integral node states and branch-wise mappings. As always, the move is accepted with probability (3.12), and the Hastings ratio is the product of (3.13) over the tree.

In this chapter, however, we have only tested the BRANCHHISTORY and NODESTATE operators, updating a single site at a time as proposed in Robinson et al. (2003).

3.2.7.2 Proposing parameter updates

We applied multiplicative update operators referred to as BRANCHLENGTH: a randomly selected branch j, as well as the times of each substitution event along that branch, is multiplied by $x = e^{\delta(U-1/2)}$. The Hastings ratio is x^{1+z_j}. A Dirichlet operator, referred to here as the STATIONARY mechanism, can be applied to π. Finally, ADDITIVE

[1]We make use of Nielsen's suggestion of sampling conditional on there being at least one event in cases where the states at the ends of the branch differ.

operators can be applied to β, which we call STRUCTURE. For these last two operators, an additional level of complication arises: evaluating the MH ratio involves the ratio of (3.9) for two different parameter values, and thus the ratio of two non-analytical normalizing "constants" is needed.

Robinson et al. (2003) provide an approximation strategy re-implemented in this work. The strategy rests on sampling a set of K sequences, denoted as $(s^{(h)})_{1\leq h\leq K}$, from the stationary distribution of sequences given a third set of parameter values θ^*. These sequences can be sampled using the Gibbs sampling method described in Robinson et al. (2003)[1]. For sufficiently large values of K, the importance sampling argument of Robinson et al. (2003) can be applied to this model to yield

$$\frac{p(s_0 \mid \theta', M)}{p(s_0 \mid \theta, M)} \simeq e^{-2(\beta'-\beta)E_{s_0}} \left(\prod_i \frac{\pi'_{s_{0i}}}{\pi_{s_{0i}}}\right) \frac{\sum_{h=1}^{K} e^{-2(\beta-\beta^*)E_{s^{(h)}}} \left(\prod_i \frac{\pi_{s_{0i}}}{\pi^*_{s_{0i}}}\right)}{\sum_{h=1}^{K} e^{-2(\beta'-\beta^*)E_{s^{(h)}}} \left(\prod_i \frac{\pi'_{s_{0i}}}{\pi^*_{s_{0i}}}\right)}. \tag{3.14}$$

The approximation's quality depends on two factors: the value of K (high values improve the approximation) and the distance of θ^* to both θ and θ' (a θ^* at the midpoint between θ and θ' gives the best approximation). Robinson et al. (2003) opt to partition their parameter space into a predefined grid. They then use the grid point θ^* this is nearest to the midpoint of θ and θ'.

Our protocol is slightly different, creating new θ^*s dynamically, always at the midpoint of θ and θ'. A new θ^* is created whenever the distance (χ) between the midpoint of θ and θ', and the nearest θ^* is beyond a predefined threshold (χ_{max}). In practice, a limit is set on the number of θ^* stored in memory. Whenever this limit is reached, and a new θ^* is to be created, one simply writes over the θ^* (and the respective K sequences) that is the furthest away from the midpoint of θ and θ'. As such, one eventually has

[1]We also tried a slightly different Gibbs scheme, which performed well: rather than updating the states at sites at random, we simply perform a full sweep across the sequence. The number of sweeps is then tuned empirically, for instance, by plotting the autocorrelation function of the sequence pseudo-energy.

a "hyper-cloud" of θ^*s following the θ and θ' as the MCMC run progresses. We determined empirically the acceptable setting for χ_{max} and K, fixing $\chi_{max} = 0.01$ and $K = 1000$. However, a larger χ_{max} and a lower K can be used to obtain faster rough estimates. Also note that restraining the interval of the uniform distribution used as the prior over β serves to increase the speed of convergence; an overly wide interval could lead to initial values that are very far from those at stationarity, which would require invoking the approximation procedure for $p(s_0 \mid \theta', M)/p(s_0 \mid \theta, M)$ many times before convergence.

3.2.7.3 General settings and implementation checks

As usual, we explored the call frequency of operators empirically, and the final setting used here to define a cycle are given in table 3.1[1]. We ran the chain for 100,000 cycles,

Table 3.1. MCMC settings used here.

Operator	Call frequency	Tuning δ	Tuning ζ
BRANCHHISTORY	50	NA	1
NODESTATE	50	NA	1
STATIONARY	1	5000	5
STRUCTURE	1	0.1	NA
BRANCHLENGTH	5	1.0	NA
BRANCHLENGTH	5	0.5	NA

discarded the first 10,000 cycles as burn-in, and sub-sampled every 50 cycles from the remaining sample. The MCMC runs require 10-15 days of CPU time on a Xeon 2.4 GHz desktop computer.

When the parameter $\beta = 0$, our model simplifies to the site independent JTT+F model (or POISSON+F, depending on the exchangeability parameters chosen). We

[1] In subsequent work, we have found it much more efficient to propose substitution mappings to several sites at once (say 50), and to use the TREEHISTORY move as well.

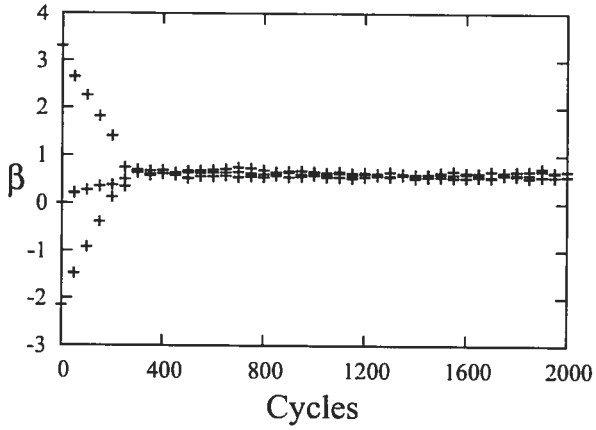

Figure 3.1. Stabilization of β (when combined with JTT+F) in three different MCMC runs for the MYO*10-153* dataset. Only the first 2000 cycles (with points every 50) are shown.

tested our implementation with $\beta = 0$, and compared the results with those obtained using the standard pruning-based sampling method, and found both to converge to essentially identical parameters and branch lengths at stationarity (not shown). We also verified that when $\beta = 0$ and $Q = Q^*$, all substitution mapping moves are accepted (since in the case, the MH ratio cancels out).

3.3 Results and discussion

3.3.1 Exchangeability parameters in relation to structural fitness considerations

We first applied our model to the MYO*10-153* data set. We performed several independent runs, starting from different initial parameter values, to explore convergence of the MCMC. Focusing on β, figure 3.1 shows its evolution over three different chains starting from different values. These runs consistently stabilize around the same values.

Additionally, β converges to positive values across all data sets (table 3.2), possi-

Table 3.2. Posterior mean (and 95% credibility intervals) of β.

Q specification	PPK*10-158*	MYO*10-153*	MYO*4-153*
POISSON+F	0.4207 (0.3300, 0.4994)	0.7005 (0.5876, 0.8164)	0.6901 (0.6141, 0.7913)
JTT+F	0.3613 (0.2759, 0.4358)	0.6273 (0.5042, 0.7386)	0.5717 (0.4804, 0.6555)

bly indicating that selection prefers sequences that maintain a good structural fitness. These results corroborate with those of Robinson et al. (2003).

Interestingly, we note that β consistently stabilizes at higher values when combining the potential to the POISSON+F model than when combining with the JTT+F model. For example, for the MYO*10-153* data set, the mean posterior values (and 95% credibility intervals) obtained are $\beta = 0.7005$ (0.5876, 0.8164) and $\beta = 0.6273$ (0.5042, 0.7386) when using POISSON+F and JTT+F respectively. Being empirically derived, the JTT matrix has a considerable amount of prior biochemical information regarding the amino acid substitution process. Accordingly, these results seem to indicate that, despite being formally site independent, the JTT matrix implicitly captures, to some extent, the average effects of dependencies between sites, measured by the potential. Hence, the potential's weighting (β) need not be as high when using the JTT matrix in comparison to that when using the naive POISSON model.

3.3.2 Amino acid stationary probabilities and branch lengths

The substitution process, as specified in (3.4), can be viewed as a composition of two layered elements: 1) a process proposing substitutions, according to Q, and 2) a process selecting substitutions, by accepting or refusing according to $e^{\beta(E_s - E_{s'})}$. Consequently, the amino acid stationary probabilities are those of the substitution process *in the absence of the* $e^{\beta(E_s - E_{s'})}$ *factor*. The potential itself will have an influence on amino

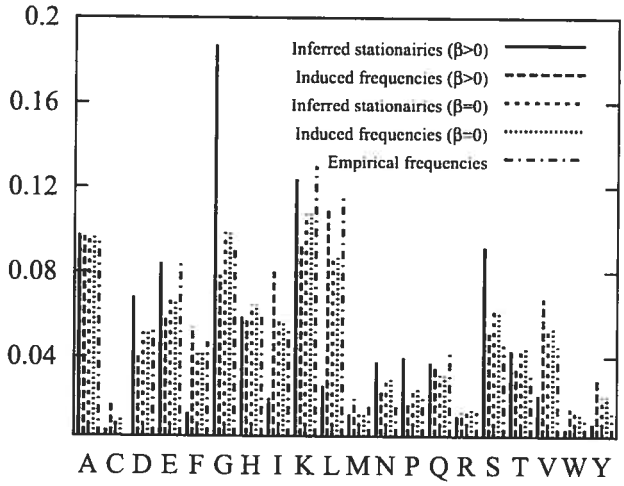

Figure 3.2. Mean values inferred for the π parameters, as well as the induced amino acid frequencies, and the empirical amino acid frequencies.

acid frequencies, thereby creating an interplay between β and π. A better measure of the true (or actual) prevalence of each of the 20 amino acids is obtained by looking at the *induced* amino frequencies in a set of sequences sampled from the stationary probability (see eqn. 3.9) implied by θ. To monitor the induced frequencies, we found it convenient to simply look at the relative frequencies of amino acids in the sequences sampled given θ^*, as this parameter vector is always in the vicinity of the θ to θ' proposal. When β is fixed ($\beta = 0$), sequences are directly sampled according to π, and the stationary probabilities and induced frequencies are necessarily equivalent (fig. 3.2). When β is a free parameter ($\beta > 0$), the π values inferred often differ widely with those when $\beta = 0$. However, we found that the induced frequencies, with $\beta > 0$ have only mild differences with those when $\beta = 0$ (or with the empirical frequencies observed in the alignment; fig. 3.2).

Likewise, branch lengths correspond to the expected number of substitutions per site proposed upstream of the selection step described above, and therefore do not reflect the true branch lengths induced by the model (i.e., the number of substitutions having actually occurred once the statistical potential has been taken into account).

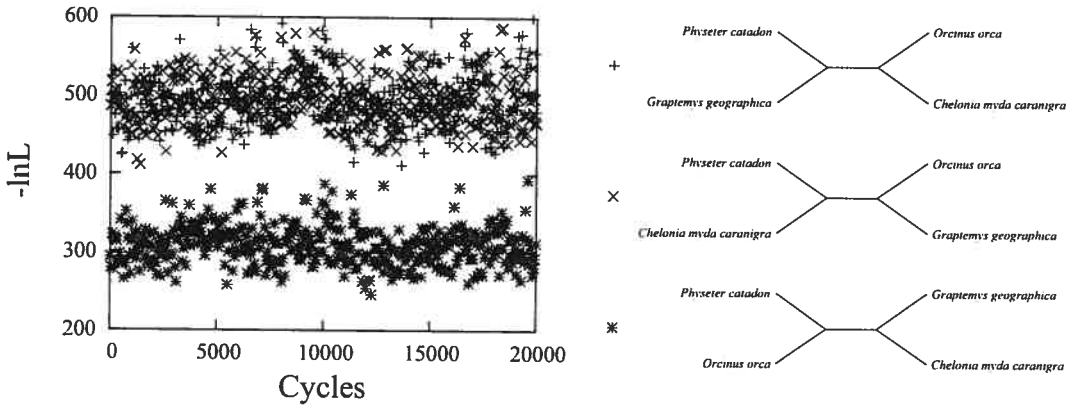

Figure 3.3. Comparison of $L = \prod_{j=1}^{2P-3} p(s_j, \phi_j \mid s_{j_{up}}, \theta, M)$ for the three possible topologies of the MYO*4-153* data set.

As expected, we found that these two measures of branch lengths do not correspond when $\beta > 0$, with the induced number of substitutions consistently lower. Using the MYO*10-153* data set as an example, we found that the tree length inferred with $\beta > 0$ was 1.0874 (0.8926, 1.3045), whereas the induced number of substitutions per site was 0.7779 (0.7255, 0.8434), a value only slightly higher to that with $\beta = 0$, at 0.7678 (0.7190, 0.8235).

3.3.3 Sensitivity to tree topology

Using the MYO*4-153* data set, we ran a MCMC under each of the three possible tree topologies. We found that parameter estimates under each topology were essentially identical (not shown). However, we did find that each tree was clearly distinguishable utilizing the MCMC methodology; figure 3.3 shows the augmented likelihood factor $L = \prod_{j=1}^{2P-3} p(s_j, \phi_j \mid s_{j_{up}}, \theta, M)$ in a window of the MCMC. The correct tree, grouping the two whales together and the two turtles together, is indeed favored.

It should be noted that these are not true likelihood-based comparisons, which would require computing the factor $p(s_0 \mid \theta, M)$, as well as the integral over all possi-

ble substitution mappings. A treatment of the tree topology as a free parameter may be technically complex. The main complication arises from the fact that a rearrangement of the tree means that the current substitution mapping may not be compatible with the newly proposed topology. This raises the difficult problem of devising update mechanisms that simultaneously change the topology and the substitution history, while having a good acceptance rate—a task that would certainly be computationally very demanding.

3.4 Conclusions

The basic model proposed here can be viewed as having two *layers*: one layer of underlying parameters that assume site-independence, specified by Q, and a second layer accounting for site interdependence, weighted using parameter β. We have found the value of β to be lower when using a more reasonable matrix Q (i.e., JTT+F) than when using a less reasonable one (i.e., POISSON), giving some indication that both layers interact in some way.

Further contrasting is needed. For instance, we have only combined the potential with JTT+F and POISSON+F, whereas several other combinations are obviously possible (e.g., combining the potential with a GTR matrix, or with $+\Gamma$ settings). Also, it would be interesting to investigate the impact of different existing statistical potentials (e.g., Miyazawa and Jernigan, 1985). In all cases, these endeavors are focused on the first step of the Bayesian framework. We now need quantitative measurements of the statistical merits of these different choices, which we treat in the next chapter.

Chapter 4

Assessing site interdependent phylogenetic models

4.1 Introduction

The numerical means of applying general site-interdependent models introduces a wide spectrum of possible model configurations; the MCMC procedures allow for a broader class of models than previously proposed methods of incorporating interdependence (e.g., Felsenstein and Churchill, 1996; Jensen and Pedersen, 2000; Pedersen and Jensen, 2001; Arndt et al., 2002; Siepel and Haussler, 2004), because the substitution process is effectively defined in the space of sequences. In other words, invoking some sequence fitness criterion could—in theory—accommodate a *total interdependence* across all sites. An ideal perspective would include full knowledge of the posited fitness landscape of the sequences under study, forming the basis of all evolutionary inferences. In practice, however, it follows that some proxies for sequence fitness may be better suited than others, and that their application may produce different results depending on the specifications of the formally site-independent components of the model. This raises the question of choosing the most relevant combination for a particular dataset.

As previously discussed, in the Bayesian paradigm, model evaluations constitute the third step of the development cycle, and the basic strategies commonly used to engage this step can be categorized along two broad axes. The first is used to compare the fit of alternative models, and, as previously mentioned, is often achieved by computing the *Bayes factor* (Jeffreys, 1935; Kass and Raftery, 1995). The second, known as *posterior predictive checking* (Rubin, 1984; Gelman et al., 1996), is used as an absolute test, characterizing discrepancies between features of true data and data simulated under the model of interest. Both strategies have become widely used for the study of phylogenetic models (Sullivan and Joyce, 2005).

In the present chapter, we explore these model evaluation strategies within the site-interdependent framework, in order to conduct our first phenomenological benchmarking of statistical potentials in this new evolutionary context. From a technical standpoint, posterior predictive checks require nothing more than posterior sampling and simulation of data replicates under the site-interdependent model. The calculation of the relative fit of different models, however, requires more elaborate methods, since the models do not allow for a closed form computation of the likelihood. Indeed, in the previous study by Robinson et al. (2003), as well as our own first explorations of the last chapter, the importance of explicit site-interdependent structural considerations was assessed based on the plausibility of associated parameter estimates. Such model assessments remain qualitative; they do not allow for selection between alternative fitness proxies, or even for a quantified comparison against site-independent models.

Here, we propose the use of a numerical technique for the evaluation of Bayes factors, yielding quantitative model comparisons under the fully site-interdependent framework originally proposed by Robinson et al. (2003). Summarized in chapter 1, the method is commonly known under the names of *thermodynamic integration*, *path sampling*, or *Ogata's method*. The technique has been used extensively in statistical physics for

evaluating (the ratio of) partition functions (for instructive reviews, see Neal, 1993; Gelman, 1998), and more recently for the study of phylogenetic models (Lartillot and Philippe, 2004, 2006). We derive an adaptation of the method, which, in combination with previously proposed techniques (Lartillot and Philippe, 2006), can provide an overall ranking of models, with or without site-interdependent criteria.

We have implemented these model assessment strategies and applied them on real protein datasets, comparing the relevance of two sets of statistical potentials (Miyazawa and Jernigan, 1985; Bastolla et al., 2001), combined with several different and well known types of models of amino acid sequence evolution. By contrasting different model configurations, we have evaluated the relative contribution of each component to the overall model fit.

4.2 Material and methods

4.2.1 Data

We used three data sets: FBP*20-363*, PPK*10-158*, and MYO*60-153*. As in the previous chapter, the contact map is derived from a reference structure, determined by X-ray crystallography for one of the sequences included in the dataset (PDB accession numbers 1ALD, 1HKA and 1MBD for FBP*20-363*, PPK*10-158* and MYO*60-153* respectively).

4.2.2 Statistical potentials

We tried the statistical potentials of Bastolla et al. (2001) and of Miyazawa and Jernigan (1985). Both are based on a contact map of the form given in (3.1). Recall that Bastolla et al. (2001) define a contact as two amino acids with any heavy atoms (atoms other than hydrogen) within 4.5 Å, whereas Miyazawa and Jernigan (1985) consider side-chain centers within 6.5 Å. Also note that Bastolla et al. (2001) ignore contacts between amino

acids within 2 positions along the sequence, while Miyazawa and Jernigan (1985) ignore contacts between immediate neighbors in the sequence. As in the previous chapter, we impose the same protein structure over the tree by applying the same contact map to all sequences considered throughout the inference.

4.2.3 Evolutionary models

We build on the previously mentioned evolutionary models, combining the potentials with site-independent amino acid formulations. In the simplest case, both equilibrium frequencies and exchangeability parameters are fixed to uniform values (referred to as POISSON). We also fixed equilibrium frequencies and exchangeability parameters to the empirically derived values of Jones et al. (1992b) (written as JTT). Other alternatives might consider equilibrium frequencies as free parameters (designated as +F), or both equilibrium frequencies and exchangeability parameters as free (indicated as GTR). We also use the $+\Gamma$ settings (Yang, 1993, 1994), based on the parameter expansion sampling methods described in chapter 2. To all of these different configurations, we apply either the potential of Bastolla et al. (2001) (indicated as +BAS) or the potential of Miyazawa and Jernigan (1985) (indicated as +MJ).

4.2.4 Priors

We used the following priors:

- $\lambda \sim$ *Exponential*, with a mean determined by a hyperparameter ν, itself endowed with an exponential prior of mean 1;
- $r \sim$ *Gamma*, with a 'shape' hyperparameter α, in turn endowed with an exponential prior of mean 1 (for notational simplicity in this chapter, we include r in the generic θ);

- $\rho \sim Dirichlet(1, 1, \ldots, 1)$;
- $\pi \sim Dirichlet(1, 1, \ldots, 1)$;
- $\beta \sim Uniform[-\beta_{max}, \beta_{max}]$, where, unless stated otherwise, $\beta_{max} = 5$.

4.2.5 Computing Bayes factors

In the present application, the thermodynamic integration method rests in defining a continuous path connecting a standard site-independent model with the model including the sequence fitness proxy, i.e. the set of statistical potentials. To do so, we make use of the fact that when $\beta = 0$, the site-interdependent model collapses to the usual site-independent model. From the partition function formalism (Appendix B), we find that for a particular value of β, the derivative of the logarithm of the marginal likelihood with respect to β gives:

$$\frac{\partial \ln p(D \mid \beta)}{\partial \beta} = \langle \frac{\partial \ln p(D, \phi \mid \beta, \theta)}{\partial \beta} \rangle, \tag{4.1}$$

where $\langle . \rangle$ represents an expectation with respect to the posterior distribution over θ and ϕ (we momentarily omit the dependence on M from the notation, considering it as implicit). Based on a sample $(\theta^{(h)}, \phi^{(h)})_{1 \leq h \leq K}$, obtained via the Metropolis-Hastings algorithm, expectations over the posterior probability distribution can be estimated for any value of β using the standard Monte Carlo relation:

$$\langle \frac{\partial \ln p(D, \phi \mid \beta, \theta)}{\partial \beta} \rangle \simeq \frac{1}{K} \sum_{h=1}^{K} \frac{\partial \ln p(D, \phi^{(h)} \mid \beta, \theta^{(h)})}{\partial \beta}. \tag{4.2}$$

Our *quasi-static* procedure then consists of sampling along a path linking the standard site-independent model, $\beta = 0$, to some arbitrary point $\beta = x$, by slowly incrementing β by a small value $\delta\beta$ after a set of MCMC cycles. The h^{th} draw of our sample, $(\theta^{(h)}, \phi^{(h)})_{1 \leq h \leq K}$, is associated with β_h, where $\beta_0 = 0$, $\beta_K = x$ and $\forall h$, $0 \leq h < K$,

$\beta_{h+1} - \beta_h = \delta\beta$. Integrating over the interval $[0, x]$ can then be estimated:

$$\ln \frac{p(D \mid \beta_K)}{p(D \mid \beta_0)} = \int_0^x \frac{\partial \ln p(D \mid \beta)}{\partial \beta} d\beta \tag{4.3}$$

$$= \int_0^x \langle \frac{\partial \ln p(D, \phi \mid \beta, \theta)}{\partial \beta} \rangle d\beta \tag{4.4}$$

$$\simeq x \times \frac{1}{K} \left[\frac{1}{2} \frac{\partial \ln p(D, \phi^{(0)} \mid \beta_0, \theta^{(0)})}{\partial \beta} + \sum_{h=1}^{K-1} \frac{\partial \ln p(D, \phi^{(h)} \mid \beta_h, \theta^{(h)})}{\partial \beta} + \frac{1}{2} \frac{\partial \ln p(D, \phi^{(K)} \mid \beta_K, \theta^{(K)})}{\partial \beta} \right]. \tag{4.5}$$

Equation (4.5) provides an estimate of the logarithm of the Bayes factor for the model including statistical potentials, with $\beta = x$, over the site-independent model, $\beta = 0$. The value of x is arbitrary. However, with this procedure, we can monitor the Bayes factor anywhere we choose along the dimension of β. Also note that, using the same sample, $\ln p(D \mid \beta_{K'}) - \ln p(D \mid \beta_0)$ can be computed for any value K' $(0 \leq K' \leq K)$. In other words, the curve of the log marginal likelihood along β can be estimated (fig. 4.1). In practice, since the high-likelihood region is restricted to a very small proportion of the admissible values of β, the integration procedure can be constrained to a small and specific interval; one can consider that outside this specific interval the marginal likelihood given β is ~ 0. Thus, exponentiating and integrating this curve yields the overall Bayes factor between the model with statistical potentials (M_1) against the model assuming independence (M_0), with the Monte Carlo estimate derived as

$$B_{01} = \frac{\int p(D \mid \beta) p(\beta) d\beta}{p(D \mid \beta_0)} \tag{4.6}$$

$$= \int \frac{p(D \mid \beta)}{p(D \mid \beta_0)} p(\beta) d\beta \tag{4.7}$$

$$\simeq \sum_{h=1}^{K} \frac{p(D \mid \beta_h)}{p(D \mid \beta_0)} \times \frac{\delta\beta}{I}, \tag{4.8}$$

where I is the interval size of the uniform prior on β, and hence $\delta\beta/I$ is the density of the prior contained between each successive $\delta\beta$ step of the quasi-static procedure.

The analogy with thermodynamics here is that the inverse of β can be thought of as a "site-interdependence temperature", with $\beta = 0$ effectively "melting" out all structural information. Alternatively, when $\beta > 0$ the models can be said to be "annealed" into site-interdependence. From this perspective, plain MCMC runs are in fact sampling the appropriate temperature for the particular sequence fitness proxy[1].

We also use this analogy in referring to our tuning of the thermodynamic integration, which we explore by applying the procedure in different directions. Specifically, *annealing integrations* work by first equilibrating a MCMC with $\beta = 0$, followed by a slow and progressive increase to $\beta = x$. If the value of β is increased too quickly, the MCMC run will not have sufficient time to equilibrate, always dragging behind configurations from preceding cycles with each increment of β. Conversely, *melting integrations* work by equilibrating a MCMC at $\beta = x$ and slowing decreasing to $\beta = 0$. Performing a bi-directional check, i.e. both annealing and melting integrations, forms the basis of our empirical exploration of the MCMC settings needed for refining the estimation procedure (fig. 4.1).

Obviously, obtaining precise integrations is computationally more challenging when applying the statistical potentials to models with greater degrees of freedom. For example, using MYO*60-153*, figure 4.1a shows that the annealing and melting integrations, applied under JTT+BAS, are very similar for fast runs ($\delta\beta = 0.005$ and $K = 100$) requiring about 2 hours of CPU time on a Xeon 2.4 GHz desktop computer. Slower runs ($\delta\beta = 0.0001$ and $K = 5,000$), requiring about 2 days of CPU are essentially indistinguishable (fig. 4.1b). When applying the integration under JTT+F+BAS, however, a clear discrepancy is observed between fast (approx. 30 hours, $\delta\beta = 0.001$ and

[1]This is also the reason we use the same notation for this parameter as we do for the model-switch thermodynamic integration morphing parameter, under site-independent models as described in chapter 2.

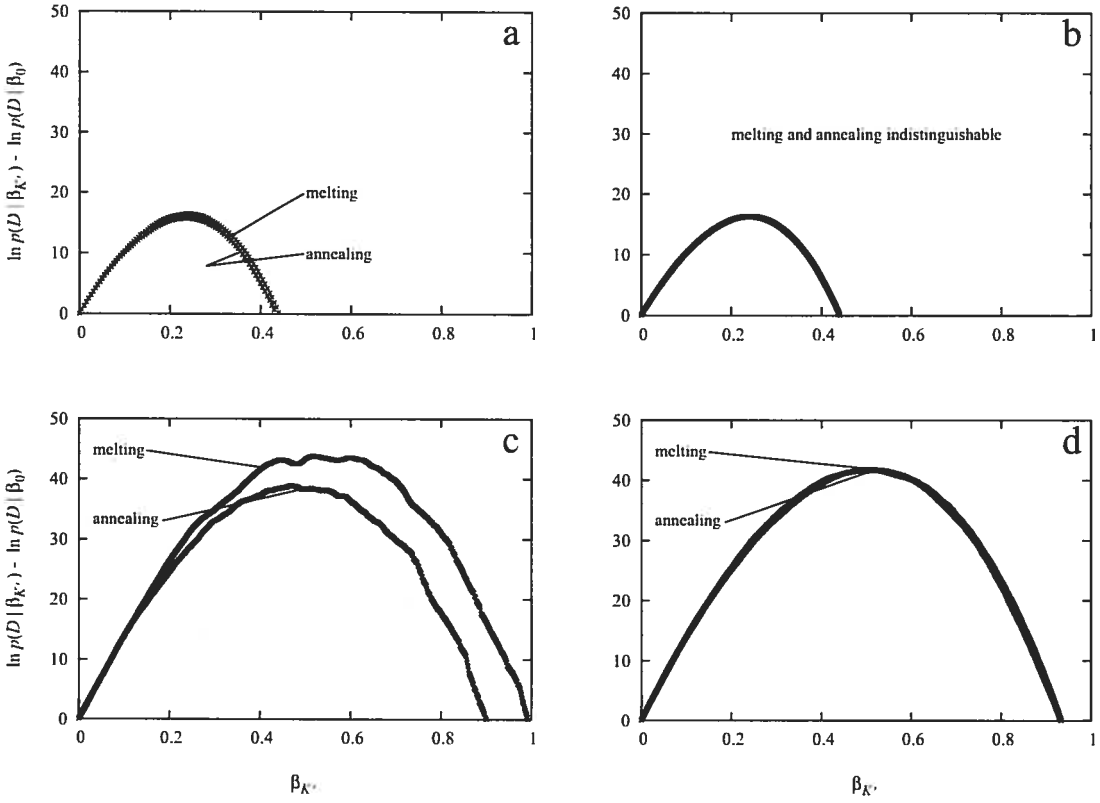

Figure 4.1. Bi-directional integrations along β for JTT+BAS (a and b) and JTT+F+BAS (b and d) performed with 'fast' (a and c) and 'slow' (b and d) settings using the MYO*60-153* dataset. The trace plots illustrate the empirical tuning of the thermodynamic MCMC sampling, which is more challenging for the model with greater degrees of freedom (bottom).

$K = 1,000$) annealing and melting runs (fig. 4.1c). Nevertheless, by tuning the call frequency of the various Monte Carlo operators, the step size of the quasi-static scheme, and the number of cycles between each increment, the integration settings can be adjusted ($\delta\beta = 0.0005$ and $K = 20,000$) to obtain precise Bayes factors estimates within about 15 days (fig. 4.1d).

Our integration scheme along β allows us to compute the Bayes factor between a site-interdependent model and its site-independent counterpart. We also need to compute Bayes factors between site-independent models, which we do using the *model-switch* integration method described in Lartillot and Philippe (2006), and summarized in chapter 2. For example, in assessing the model GTR+BAS, we first perform the integration along β, giving the log Bayes factor of GTR+BAS against GTR. Then, applying the model-switch method, we compute the log Bayes factor between GTR and POISSON. With both estimates at hand, we calculate the log Bayes factor of GTR+BAS against POISSON, simply using the additive quality of logarithms:

$$\ln \frac{p(D \mid \text{GTR+BAS})}{p(D \mid \text{POISSON})} = \ln \frac{p(D \mid \text{GTR+BAS})}{p(D \mid \text{GTR})} + \ln \frac{p(D \mid \text{GTR})}{p(D \mid \text{POISSON})}. \tag{4.9}$$

In this way, it is possible to observe the overall ranking of models for a given dataset, by having all Bayes factors against the simplest POISSON model. Note that error of the integration procedures is cumulative in equation 4.9; for succinct comparisons of models, we report the mean of the highest and lowest values obtained using bi-directional checks (table 4.1). For the simpler models, the error can be reduced to less than one natural log unit, whereas the more challenging models can lead to an error $\sim \pm 4$.

The following protocol summarizes:

- For a particular model setting, run a quasi-static thermodynamic integration, estimating the log marginal likelihood curve along β (applying the Monte Carlo estimate given by eqn. 4.5);

- Exponentiate and integrate the resulting curve to estimate the overall Bayes factor between the site-interdependent model and the underlying site-independent model (applying the Monte Carlo estimate given by eqn. 4.8);

- Given the marginal likelihood comparisons between site-independent models, estimated using the model-switch scheme described in Lartillot and Philippe (2006) (as well as in chapter 2), compute all Bayes factor with respect to POISSON (applying relations analogous to eqn. 4.9).

4.2.6 Posterior predictive re-sampling

The sampling techniques used here are particularly well suited to performing posterior predictive checks, as described in Nielsen (2002). A posterior predictive scheme is based on a simulation procedure, which consists of drawing a sequence from the stationary probability written in (3.9) under a given $\theta \in \Theta$, and simulating a substitution mapping on the branches of the tree to generate a replication of the data—in other words, these mappings are unconstrained to any states at the leaves of the tree (Nielsen, 2002). The simulation procedure is repeated on each successive parameter values of the initial MCMC sampling performed on the true data.

Given a statistic of interest, posterior predictive checks then consist in comparing the value of the statistic observed on the data, with the distribution obtained on the replicates; a discrepancy indicates that the model does not adequately account for the phenomena summarized by the statistic. Here, our statistics are not exactly computed on the data, but on mappings sampled from their posterior distribution. We refer to the substitution histories obtained from simulations as *predictive mappings*, in contrast with what we call the *"observed" mappings*, which are compatible with the true observed data. Note, of course, that these latter mappings are not actually observed, but rather constitute the data augmentation step of the MCMC methods.

To explore whether a model can explain the level of rate heterogeneity of a given dataset, we compared the variance in number of substitutions across sites, calculated based on the number of substitutions counted at each site in predictive and observed mappings. This particular statistic was used by Nielsen (2002) as an example demonstrating the utility of a mapping-based framework.

Also, in order to observe how well a model captures amino acid exchange propensities, we counted each of the 190 possible types of exchange in mappings to generate what we refer to as the residue *exchange distribution.* We then computed the Euclidean distance between predictive and observed exchange distributions for each sample point from the posterior distribution.

4.3 Results and discussion

4.3.1 Bayes factors

We applied the thermodynamic integration procedures to all datasets, and for all model combinations described in this chapter. The resulting Bayes factors, computed against the simplest model (POISSON), are reported in table 4.1.

4.3.1.1 Overall fit of site-independent models

The most favored site-independent model is JTT+Γ for FBP*20-363* and PPK*10-158*, and JTT+F+Γ for MYO*60-153*. This is somewhat expected. The POISSON-based models are obviously unrealistic, since the exchangeability between amino acids is clearly not uniform, hence giving support to JTT-based models. Also, allowing for rate heterogeneity is known to nearly always improve the model fit (Yang, 1996; Buckley et al., 2001; Posada and Buckley, 2004), as is the case here. The equilibrium frequencies of JTT appear to be suitable for the two smaller datasets, in as much as the dimensionality

Table 4.1. Natural logarithm of the Bayes factor for all models studied in this chapter, with POISSON used as a reference (the best site-independent models for each dataset are emphasized in italics, whereas the best overall models are emphasized in bold).

Model	FBP*20-363*	PPK*10-158*	MYO*60-153*
POISSON	0	0	0
POISSON+BAS	10	16	24
POISSON+MJ	6	7	18
POISSON+F	103	34	70
POISSON+F+BAS	158	78	142
POISSON+F+MJ	144	65	129
POISSON+Γ	135	53	138
POISSON+Γ+BAS	138	69	162
POISSON+Γ+MJ	137	58	156
POISSON+F+Γ	238	89	207
POISSON+F+Γ+BAS	296	139	280
POISSON+F+Γ+MJ	285	122	267
JTT	380	144	368
JTT+BAS	391	155	382
JTT+MJ	386	150	379
JTT+F	365	137	389
JTT+F+BAS	397	159	427
JTT+F+MJ	389	145	417
JTT+Γ	*529*	*195*	499
JTT+Γ+BAS	540	206	512
JTT+Γ+MJ	535	200	508
JTT+F+Γ	513	186	*513*
JTT+F+Γ+BAS	**546**	**216**	**551**
JTT+F+Γ+MJ	539	203	537
GTR	310	102	347
GTR+BAS	346	139	394
GTR+MJ	338	121	383
GTR+Γ	434	147	466
GTR+Γ+BAS	471	185	512
GTR+Γ+MJ	462	168	501

penalty renders a specific adjustment of these parameters unreliable. For MYO*60-153*, however, such a dataset-specific adjustment of equilibrium frequencies seems worthwhile. The GTR matrix is always rejected over the JTT-based models, most likely since the data sets considered are much too small to reliably infer the 189 additional free parameters introduced by this model. Note, however, that the GTR-based models are still far better than POISSON-based models.

4.3.1.2 Overall fit of site-interdependent models

Models including statistical potentials are always favored over their site-independent counterparts, under all configurations explored here. This being the case for all three proteins studied suggests that such an improvement in fit is general. Nevertheless, the improved fit observed when including statistical potentials is mild, when compared to the overall fit of rich site-independent models. Specifically, the use of an empirical amino acid replacement matrix and a gamma distributed rates model both outperform the sole use of statistical potentials.

4.3.1.3 Interplay between model configurations

Interestingly, the relative improvement brought about by the potentials is very much a function of the site-independent components of the models. In particular, the amelioration in model fit when applying statistical potentials, as well as the equilibrium value of β under plain MCMC sampling (table 4.2), is noticeably lower when the π-vector is fixed, which is the case irrespective of the other site-independent settings. This is perhaps best understood by observing the stationary probability distribution written in (3.9). While the stationary distribution is given by π under the standard notation of continuous-time Markov chains, under the site-interdependent models studied here it is given by a combination of π and the exponentiated pseudo-energy factor. This

Table 4.2. Equilibrium values of β. Mean posterior values (with 95% credibility intervals) under all model combinations described in the text.

Model	FBP*20-363*	PPK*10-158*	MYO*60-153*
POISSON+BAS	0.107 (0.0521, 0.162)	0.249 (0.176, 0.321)	0.268 (0.203, 0.330)
POISSON+MJ	0.0074 (0.0001, 0.0150))	0.0279 (0.0163, 0.0395)	0.0423 (0.0307, 0.0539)
POISSON+F+BAS	0.402 (0.335, 0.474)	0.462 (0.378, 0.549)	0.637 (0.543, 0.725)
POISSON+F+MJ	0.0658 (0.525, 0.0787)	0.0724 (0.0553, 0.0905)	0.1086 (0.0890, 0.1295)
POISSON+Γ+BAS	0.0989 (0.0509, 0.158)	0.268 (0.197, 0.350)	0.239 (0.169, 0.332)
POISSON+Γ+MJ	0.0058 (0.0001.0139)	0.0296 (0.0164, 0.0423)	0.0397 (0.0202, 0.0501)
POISSON+F+Γ+BAS	0.439 (0.373, 0.511)	0.564 (0.463, 0.665)	0.717 (0.601, 0.825)
POISSON+F+Γ+MJ	0.0811 (0.0611, 0.0953)	0.0983 (0.0786, 0.1198)	0.1406 (0.1147, 0.1663)
JTT+BAS	0.176 (0.126, 0.224)	0.264 (0.193, 0.332)	0.240 (0.167, 0.318)
JTT+MJ	0.0231 (0.0144, 0.0316)	0.0368 (0.0234, 0.0499)	0.0423 (0.0273, 0.0560)
JTT+F+BAS	0.305 (0.232, 0.369)	0.378 (0.228, 0.464)	0.501 (0.409, 0.598)
JTT+F+MJ	0.0449 (0.0335, 0.0577)	0.0722 (0.0562, 0.0894)	0.0816 (0.0630, 0.1014)
JTT+Γ+BAS	0.177 (0.130, 0.231)	0.277 (0.206, 0.346)	0.244 (0.170, 0.321)
JTT+Γ+MJ	0.0234 (0.0149, 0.0320)	0.0391 (0.0254, 0.0531)	0.0424 (0.0276, 0.0561)
JTT+F+Γ+BAS	0.333 (0.264, 0.413)	0.478 (0.368, 0.582)	0.575 (0.465, 0.685)
JTT+F+Γ+MJ	0.0541 (0.0407, 0.0688)	0.0724 (0.0519, 0.0943)	0.0975 (0.0715, 0.1197)
GTR+BAS	0.433 (0.351, 0.508)	0.511 (0.412, 0.608)	0.625 (0.505, 0.745)
GTR+MJ	0.0680 (0.0504, 0.0854)	0.0777 (0.0574, 0.0999)	0.1148 (0.0901, 0.1402)
GTR+Γ+BAS	0.440 (0.362, 0.513)	0.546 (0.442, 0.649)	0.679 (0.563, 0.792)
GTR+Γ+MJ	0.0791 (0.0607, 0.0910)	0.0929 (0.719, 0.1171)	0.1228 (0.0961, 0.1590)

forces a re-interpretation of the usual meaning given to π: rather than representing the amino acid equilibrium frequencies, these parameters should be viewed as "chemical potentials" associated to each residue, and whose effect is combined to that of the statistical potentials in the final amino acid equilibrium frequencies, as discussed in the previous chapter. From this perspective—related to *random energy* approximations (Shakhnovich and Gutin, 1993; Sun et al., 1995; Seno et al., 1998)—fixing the values of π, to uniform values (in the case of POISSON) or to the JTT values, effectively prevents the model from compensating for the coupling to the exponentiated pseudo-energy factor, and thus leads to a low support for the site-interdependent models. Indeed, while the +F settings were rejected in favor of JTT for FBP*20-363* and PPK*10-158* under site-independence, when invoking the statistical potentials, this increased parameterization seems favored.

Also of interest, we find that the relative improvement brought about by the potentials is more important when using POISSON-based models than when using a JTT-based models. This is consistent with the fact that the JTT matrix inherently accounts for protein structure features, by assigning greater exchange propensities between amino acids sharing various physico-chemical properties. In other words, explicitly accounting for site-interdependencies due to tertiary structure requirements is more important when using the naive POISSON-based model than when using the more informed JTT-based model.

When invoking the GTR configuration, the potentials give a greater improvement in fit than when applying the JTT settings. Nevertheless, site-interdependent GTR-based models are still poorer for these small datasets than the JTT-based models.

The use of a $+\Gamma$ model seems to give an essentially additive improvement in model fit, with little, or no interaction with other model configurations. Since the statistical potentials could impact directly on site-specific rates, this result is unexpected; the lack

of interaction in itself may be indicative that the potentials do not, in fact, acknowledge significant rate heterogeneity.

4.3.1.4 Comparison of statistical potentials

We find that for these applications the potentials of Bastolla et al. (2001) and Miyazawa and Jernigan (1985) receive similar support, with +BAS models mildly favored over +MJ. The comparable merit of these potentials is somewhat expected; both work with a similar contact-based protein structure representation. The fact that +MJ models receive lower support than +BAS models may be a consequence of the over-simplified *quasi-chemical* approximation used in the derivation of the potentials of Miyazawa and Jernigan (1985), or to differences in the contact definition itself.

4.3.1.5 Sensitivity to the prior on β

It is common practice, when assessing a new class of models, to evaluate the influence of the prior on the resulting model fit (Kass and Raftery, 1995). Here, we focus on the distinguishing feature of our model: the prior on β. Note that the trace plots shown in figure 4.1 display, up to an additive constant, the marginal likelihood of the model with β successively fixed to each value along the integration procedure. Treating β as a free parameter requires that we define a proper prior probability distribution, over which these curves are averaged (eqn. 4.8). Since little is known regarding the usage of statistical potentials in this context, we follow the practice of assigning a bounded uniform prior, and testing empirically that the posterior distribution of β is well within these bounds (Robinson et al., 2003).

It should be noted that the two sets of potentials studied here are not scaled equivalently, which leads to different temperature factors at equilibrium—the potentials of Bastolla et al. (2001) lead to higher values of β (table 4.2). This means that ap-

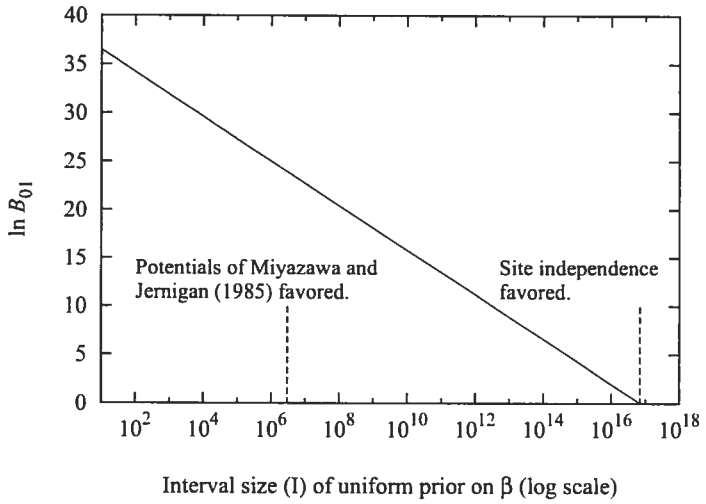

Figure 4.2. Influence of the interval size (I) of the uniform prior distribution for β on the calculated Bayes factor. Here, the models being compared are JTT+F+Γ+BAS against JTT+F+Γ, applied to MYO*60-153*. Two thresholds are marked on the graph. The first (leftmost) indicates the point beyond which JTT+F+Γ+MJ (with prior on $\beta \sim$ [-5, 5]) is favored over JTT+F+Γ+BAS. The second indicates the point beyond which JTT+F+Γ is favored over JTT+F+Γ+BAS.

plying the same uniform prior on β under +BAS and +MJ models amounts to giving favor to the potentials of Bastolla et al. (2001); loosely speaking, the differences in scaling make the space of admissible values for β "appear" larger to +MJ models. To illustrate this problem, we performed a simple exploration of the influence of the size of the interval (I) of the uniform prior on β. Using the same sample, the Monte Carlo approximation given by (4.8) can be re-computed for different interval sizes. For example, figure 4.2 shows the log Bayes factor comparing JTT+F+Γ+BAS and JTT+F+Γ as a function of the interval size I. As I increases, the density of the prior contained in each increment of the quasi-static procedure decreases, leading to a lower support for JTT+F+Γ+BAS. When I reaches an order of magnitude around 10^6, the JTT+F+Γ+MJ model, with prior on $\beta \sim$ [-5, 5], becomes favored over JTT+F+Γ+BAS. Moreover, when I reaches an order of magnitude $\sim 10^{17}$, the JTT+F+Γ becomes favored over the site-interdependent model. This illustrates a fun-

damental theoretical consequence of the Bayesian paradigm: model rankings can change by redefining the space of admissible parameter values (the prior). In the present case, this means that no matter how strong the signal for site-interdependence, their exists an interval size I for the uniform prior on β such that the site-independent model is favored, an example directly related to the so-called *Jeffreys-Lindley paradox* (Lindley, 1957; Bartlett, 1957; Lindley, 1980).

In practice, the resulting difference in dimensionality penalty does not appear problematic in the present case; the potentials do not differ drastically in scaling, and the maximum marginal likelihood along β was always greater for the potentials of Bastolla et al. (2001) than for those of Miyazawa and Jernigan (1985). For example, for MYO*60-153* under the model JTT+F+Γ+BAS, the maximal point along the marginal likelihood curve gives a log Bayes factor of ~ 553, whereas under JTT+F+Γ+MJ the maximal point gives ~ 540.

For this particular comparison, one simple alternative would be to re-normalize the potentials to an equivalent scaling. Yet this solution would still not be applicable when comparing sequence fitness proxies based a fundamentally different rationales. In the longer run, non-uniform priors could be used, particularly as more datasets are analyzed; Lempers (1971), for example, suggested setting aside some datasets for constructing proper priors to be used in subsequent analyses. Along these lines, we are currently devising other forms of statistical potentials, with each having the same overall temperature scaling (Kleinman et al., 2006).

4.3.1.6 Permutations checks

Overall, the pairwise contact potentials studied here appear inadequate; given the choice between the sole use of statistical potentials and the standard site-independent models, one would opt for the latter. Yet, a signal for site-interdependence is clearly detected.

Perhaps the simplest check that can be done when constructing a model accounting for a particular signal, is the evaluation of the model's performance when deliberately removing that signal from the data. Following Telford et al. (2005), we explore this through simple permutation tests, whereby we swap the positions of a percentage of random pairs of columns in the alignment. Such permutations have the effect of blurring the structural signal. Indeed, the tests can be viewed as a randomization of the contacts in the contact map. We defined four levels of randomization, swapping the position of 25, 50, 75, and 100% of columns. For each randomization, we computed the Bayes factor in favor of the site-interdependent model. Given the computational burden, we performed only three replicates for each randomization level.

We performed these permutation checks using the MYO*60-153* dataset, comparing the log Bayes factor of POISSON+F+Γ+BAS against POISSON+F+Γ (this is the case giving the greatest improvement in model fit when applying the sequence fitness proxy). As expected, the support for site-interdependent considerations is a decreasing function of the percentage of randomization, essentially dropping to zero for a fully permuted column ordering (fig. 4.3). Also note that each replicate randomization gives slightly different results; evidently, the interdependencies between different positions of a protein are not all equivalent.

This test plainly illustrates the distinguishing feature of the models in simplistic terms: site-interdependent models give meaning to the order of amino acid columns in the alignment.

4.3.2 Posterior predictive re-sampling

Two of the most fundamental patterns of amino acid sequence evolution are 1) the heterogeneity of substitution rates across sites and 2) the heterogeneity of amino acid exchange propensities. Both of these heterogeneities could be effects induced by struc-

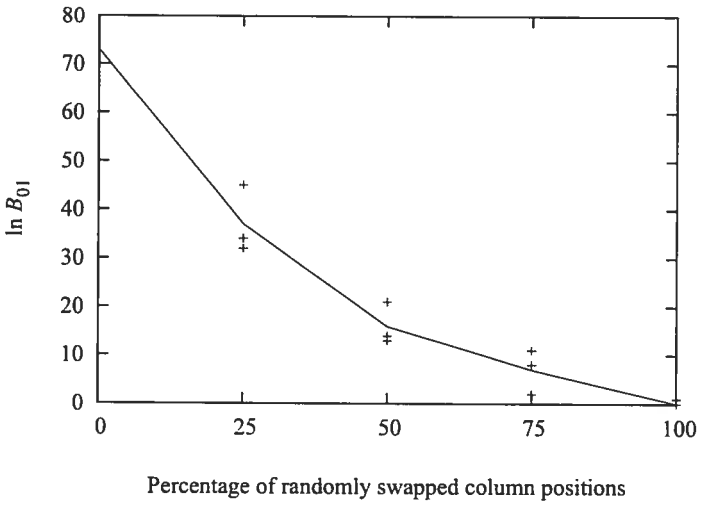

Figure 4.3. Permutation checks randomizing the order of columns in the alignment. The log Bayes factor is estimated between POISSON+F+Γ+BAS and POISSON+F+Γ, for three replicates at each randomization level. A line joining the mean values at each randomization level is drawn as a visual aid.

tural constraints, and, hence, could be accounted for—at least in part—by the sequence fitness proxy. However, accommodating rate-across sites variations (+Γ) and using an empirical amino acid replacement matrix (JTT) also accounts for these heterogeneities. As such, the best model obtained for all three datasets (JTT+F+Γ+BAS) seemingly corresponds to a redundant configuration. To further explore this point, we have applied simple posterior predictive checks, as described in the material and methods.

4.3.2.1 Rate heterogeneity

Under a model assuming uniform rates across sites, and if there is rate variation in the dataset considered, the observed rate variance is likely to depart significantly from the predictive rate variance; by the definition of the model, the predictive rate variance will tend to be very low. This is indeed the case, as can be seen from figure 4.4a. The extreme discrepancy between observed and predictive rate variance is in itself enough to reject the uniform rates model (Nielsen, 2002). Comparing figures 4.4a and 4.4c shows that using the potentials of Bastolla et al. (2001) essentially leaves the observed

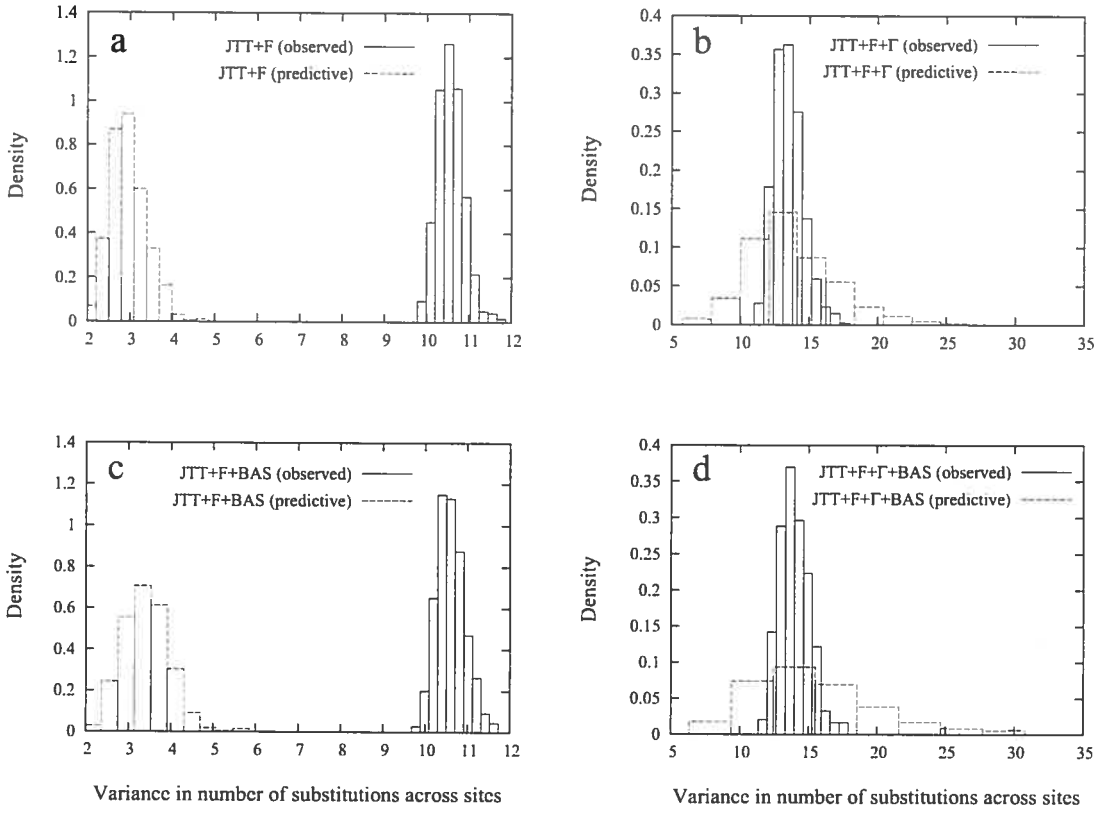

Figure 4.4. Posterior density plots of the variance in the number of substitution across sites obtained in predictive mappings and observed mappings of our sample from the posterior distribution, under the JTT+F (a), JTT+F+Γ (b), JTT+F+BAS (c) and JTT+F+Γ+BAS (d) models (using MYO*60-153*).

rate variance unchanged, and the predictive rate variance is only slightly higher than the simple model assuming uniform rates—the mean predictive rate variance increases from 2.95 in 4.4a to 3.40 in 4.4c.

In contrast, (fig. 4.4b), under the $+\Gamma$ model, the observed rate variance is even greater than under the uniform rates model. As can be appreciated graphically, and according to the calculated Bayes factors, an explicit treatment of rate variation ($+\Gamma$) gives a better correspondence between model and data, with the predictive distribution centered on the observed (fig. 4.4b and 4.4d).

Thus, on one hand, the $+\Gamma$ model accommodates rate heterogeneity across sites very well, but does not explain it, i.e., it is phenomenological. On the other hand, the +BAS model, which was hoped to explain this heterogeneity on mechanistic grounds, essentially fails at doing so.

Note that predictive distributions tend to have a greater spread than observed distributions. This is a result of predictive distributions comprising two levels of uncertainty: the fundamental uncertainty associated with the inferred parameter values of the model (the posterior distribution)—an uncertainty which tends to be greater for higher dimensional models—and the uncertainty associated to the data replication (the simulation procedure). Indeed, this effect is displayed in the more pronounced spread in rate variance under the more complex +BAS model (comparing 4.4b and 4.4d).

4.3.2.2 Amino acid exchange propensities

Figure 4.5 is a comparison of the Euclidean distance between predictive and observed exchange distributions, as explained in the material and methods. In principle, a model yielding a lower distance between observed and predictive amino acid exchange distributions would be favored.

In figure 4.5a, the distance between predictive and observed distributions under the

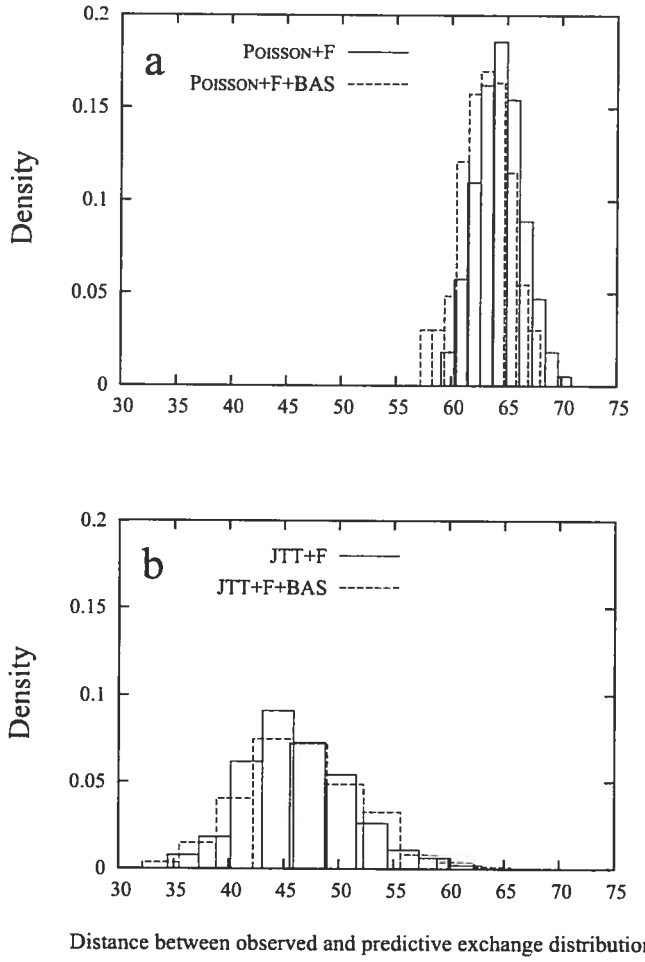

Figure 4.5. Posterior density plot of the Euclidean distance between predictive and observed substitution type distributions (see material and methods). In a), the models used are POISSON+F and POISSON+F+BAS. In b), the models are JTT+F and JTT+F+BAS.

POISSON+F is high, and is only slightly reduced when applying the potentials of Bastolla et al. (2001)—the mean distance goes from 63.92 under POISSON+F to 62.51 under POISSON+F+BAS. In the case of JTT (fig. 4.5b), the distance between predictive and observed distributions is much lower. This is indicative that a much better adequation is obtained between the types of substitutions of mappings conditioned on the data, with those predicted under the model when using the empirical amino acid exchange propensities of JTT, even when applying the potentials of Bastolla et al. (2001).

4.4 Conclusions

The results of the different model assessment strategies converge to the same fundamental conclusion: while an improved model fit is observed when applying the statistical potentials, the improvement does not justify abandoning the successful techniques previously developed for modeling complexities such as across-site rate heterogeneity, or variations in amino acid exchange propensities. In other words, the model fails to attain the simple phenomenological benchmarks of interest. It would indeed have been surprising to see such a simple 0/1 contact map, with potentials devised for other purposes, supplanting all strategies developed under the assumption of independence. Also note that the mild improvement in model fit brought about by the use of statistical potentials comes at a high computational cost. Indeed, the total CPU time for the present study is estimated at about 1000 days on a Xeon 2.4 GHz computer.

Two directions for further research are thus pressing. The first is the design of richer potentials, specifically adapted to the evolutionary framework at hand. The second is the development of faster computational methods. We explore the first direction in chapter 5, and return to the second direction in chapter 6.

Chapter 5

Devising statistical potentials for phylogenetic analysis

5.1 Introduction

The direct use of statistical potentials as done in the model proposed by Robinson et al. (2003), as well as those studied in the last two chapters, should be considered as a preliminary step to exploring a novel class of models. As previously mentioned, currently available statistical potentials may not be ideal for the evolutionary context that interests us, since they have generally been optimized in the context of protein fold recognition, i.e., for maximizing the rate of correct structure prediction, given the sequence. In an evolutionary perspective, and assuming that the protein's structure is well conserved over the time span in consideration, we would like to make a reciprocal prediction: what are the sets of possible sequences relating an alignment of observed sequences, that are compatible with a given structure? We might simplify the question by removing the phylogenetic component, and asking: what are the sequences that stably fold into a pre-specified conformation? This is more generally known as the *inverse folding* problem, or *protein design* (Drexler, 1981; Pabo, 1983; Ponder and

Richards, 1987).

Several approaches have been proposed for protein design, consisting in maximizing the z-score between the energy of the native sequence on the target conformation and its energy on a set of decoy sequences (Chiu and Goldstein, 1998), or, alternatively, in applying a mean-square criterion on the values taken by the pseudo-energy score on each structure-sequence pair of a database (Seno et al., 1998). However, these methods have thus far only been tested in cubic lattice protein models. In addition, they lack a firm theoretical basis. In particular, it would be interesting to guarantee optimal predictive power of any given form of potential, and to have a robust methodology for exploring the merits of different functional forms.

In this chapter[1], we set out a general protein-design framework, deriving a new set of statistical potentials from a database of sequences of known three dimensional structure. In effect, the framework focuses on the stationary distribution of a site-interdependent Markov process, but treating the coefficients of the potential as free parameters, which we adjust to their ML estimates from a large set of sequence-structure pairs. Reformulated in this way, the method maximizes the predictive power of the potential, now in the structure-seeks-sequence direction. By construction, it yields the optimal parameter values that can be obtained for a given form of potential. In addition, different functional forms can be devised, and compared based on the likelihood obtained on a test data set, distinct from the learning data set, in a procedure known as *cross-validation* (Stone, 1974). The overall ML framework could also be extended to a full Bayesian approach, but the ML approach relieves certain computational difficulties, and thus provides a practical first avenue to investigate the modeling framework. We explore the same functional form as Miyazawa and Jernigan (1985), which we also supplement with a solvent-accessibility potential, of a form chosen via cross-validation. Finally, we re-

[1]We mention here again that this chapter reproduces results from Klienman et al. (2006), the third paper mentioned in the preface. The material has been considerably shortened for the purpose of this dissertation.

inject these potentials into the phylogenetic context to compare them to the potentials used in the last two chapters.

5.2 Material and methods

5.2.1 Data

We used proteins culled from the entire PDB according to structure quality (resolutions better or equal to 2.0 Å) and with less than 25% mutual sequence identity (Wang and Dunbrack, 2003). Two subsets of approximately equal size were obtained by partitioning proteins randomly: DS1, 449 proteins, 100,077 sites, and DS2, 465 proteins, 99,894 sites.

5.2.2 Structure representation

We used the contact representation of Miyazawa and Jernigan (1985), as in the preceding chapter. The accessible surface of a residue is defined as the atomic accessible area when a probe of the radius of a water molecule is rolled around the Van der Waal's surface of the protein (Lee and Richards, 1971). We used the program NACCESS (Hubbard and Thornton, 1993) to make this calculation. When treating PDB files with multiple chains, solvent accessibility was calculated taking into account all molecules in the structure. The accessibility classes (percentage relative to the accessibility in Ala-X-Ala fully extended tripeptide) were defined so as to generate W equal-sized subsets of sites.

5.2.3 Model

Let us consider a sequence $s = (s_i)_{1 \leq i \leq N}$, of length N, and of conformation c. By Bayes' theorem, we write the probability of a sequence conditional on the conformation, (and

the model M) as

$$p(s \mid c, M) = \frac{p(c \mid s, M)p(s \mid M)}{\sum_s p(c \mid s, M)p(s \mid M)}, \tag{5.1}$$

where the sum in the denominator is over all possible sequences of length N. Given a statistical potential $E(s, c)$, the conformational probability can be expressed as a Boltzman distribution:

$$p(c \mid s, M) = \frac{1}{Z} e^{-E(s,c)/kT} \tag{5.2}$$

$$= e^{-(E(s,c)-F(s))/kT}, \tag{5.3}$$

where,

$$Z = \sum_c e^{-E(s,c)/kT} \tag{5.4}$$

is a normalizing constant, summing over all possible conformations, and

$$F(s) = -\ln Z. \tag{5.5}$$

Here, k and T represent the Boltzman constant and absolute temperature respectively. Without loss of generality, it is possible to rescale the potential so that $kT = 1$, which we will do in the following.

By defining the *inverse potential*

$$G(s, c) = E(s, c) - F(s), \tag{5.6}$$

and assuming a uniform prior $p(s \mid M)$, the conditional probability of a sequence reads as

$$p(s \mid c, M) = \frac{1}{Y} e^{-G(s,c)}, \tag{5.7}$$

where

$$Y = \sum_{s} e^{-G(s,c)} \tag{5.8}$$

is the normalizing factor, summing over all possible sequences of length N. We used a statistical potential made of two terms:

$$E(s,c) = \sum_{1 \leq i < i' \leq N} \Delta_{ii'} \epsilon_{s_i s_{i'}} + \sum_{1 \leq i \leq N} \Xi^{v_i}_{s_i}, \tag{5.9}$$

where Δ is the 0/1 contact map according to the definition of Miyazawa and Jernigan (1985), $\epsilon = (\epsilon_{ab})_{1 \leq a,b \leq 20}$ is the set of energy coefficients associated with pairwise amino acid contacts (ϵ is entirely specified from 209 parameters), and where $\Xi = (\Xi^w_a)_{1 \leq a \leq 20, 1 \leq w \leq W}$ is the set of energy coefficients associated with observing each amino acid in each of the W possible solvent accessibility classes.

Deriving the inverse potential requires the calculation of $F(s)$, which is already entirely specified, from a sum over all conformations. However, this computation is difficult in practice. As an alternative, we can give it a simple phenomenological form, inspired from the random energy model (Shakhnovich and Gutin, 1993; Seno et al., 1998; Sun et al., 1995):

$$F(s) = - \sum_{1 \leq i \leq N} \Sigma_{s_i}. \tag{5.10}$$

where $\Sigma = (\Sigma_a)_{1 \leq a \leq 20}$ is a set of free parameters analogous to the chemical potential of each amino acid. Note that in chapters 3 and 4, chemical potentials were given the +F form, so as to relate more closely to the substitution models studied.

Altogether, our parameter vector is made of three components ($\theta = \{\epsilon, \Xi, \Sigma\}$), and

the inverse potential reads as

$$G(s,c) = \sum_{1\le i<i'\le N} \Delta_{ii'}\epsilon_{s_i s_{i'}} + \sum_{1\le i\le N} \Xi^{v_i}_{s_i} + \sum_{1\le i\le N} \Sigma_{s_i}. \tag{5.11}$$

Note that the probability (5.7) is invariant under the following transformation:

$$\Sigma'_a = \Sigma_a + J_1 \tag{5.12}$$

$$\epsilon'_{ab} = \epsilon_{ab} + J_2 \tag{5.13}$$

$$\Xi'^w_a = \Xi^w_a + J_3, \tag{5.14}$$

where J_1, J_2 and J_3 are arbitrary real constants. Therefore, to ensure identifiability of our model, we enforce the following constraints:

$$\sum_a \Sigma_a = 0, \tag{5.15}$$

$$\sum_{ab} \epsilon_{ab} = 0, \tag{5.16}$$

$$\sum_a \Xi^w_a = 0, 1 \le w \le W. \tag{5.17}$$

Finally, we assume that all sequence-structure pairs in our database are independent, and multiply the probability in (5.7) over all sequence-structure pairs, based on the same values of the potential. In the following, however, we retain the single sequence-structure pair notation for simplicity.

5.2.4 Optimizing the potentials by gradient descent

In the present context, the Bayesian approach of conditioning θ on the data is likely to be computationally demanding, due to the intractable normalizing constant Y. Rather, we shall adjust our parameters so as to maximize the (log) probability in (5.7), which, in the ML perspective, is view as the likelihood function. We will work with the negative

log of (5.7), defining: $\ell = -\ln p(s \mid c, M)$, turning the problem into a minimization of ℓ.

The derivative of ℓ with respect to the parameters of the potential reads as

$$\frac{\partial \ell}{\partial \theta} = \frac{\partial G(s,c)}{\partial \theta} + \frac{\partial \ln Y}{\partial \theta}. \tag{5.18}$$

As in the previous chapter, one can applying the partition function formalism (Appendix B) to Y to express the second term as

$$\frac{\partial \ln Y}{\partial \theta} = -\langle \frac{\partial G}{\partial \theta} \rangle, \tag{5.19}$$

where $\langle . \rangle$ stands for an expectation with respect to (5.7). As before, this expectation can be approximated from a sample of sequences $(s^{(h)})_{1 \leq h \leq K}$, drawn according to (5.7). This sample can be obtained using the same Gibbs sampling procedure used in previous chapters.

The derivatives with respect to ϵ can be expressed as:

$$\frac{\partial \ell}{\partial \epsilon_{ab}} = -\left[n_{ab} - \langle n_{ab} \rangle \right], \tag{5.20}$$

where n_{ab} is the number of contacts between amino acids a and b observed in the data base, and where $\langle n_{ab} \rangle$ is approximated from the Monte Carlo average, in a sample of sequences, of the number of contacts between a and b, with the sequences drawn under the current values of the energetic coefficients. Formula (5.20) thus leads to an intuitive characterization of the maximum likelihood estimate $\hat{\epsilon}$: it is the value of ϵ such that the average number of each type of contact predicted by the potential matches the number observed in the database. Following a similar derivation, we have the relation

for solvent accessibility coefficients as

$$\frac{\partial \ell}{\partial \Xi_a^w} = -\left[l_a^w - \langle l_a^w \rangle\right], \tag{5.21}$$

where l_a^w is the number of amino acids of type a in solvent accessibility class w. Finally, for the chemical potentials, we have

$$\frac{\partial \ell}{\partial \Sigma_a} = -\left[m_a - \langle m_a \rangle\right], \tag{5.22}$$

where m_a is the number of amino acids of type a.

The above relations allow us to approximate the gradient of ℓ, which we may follow using standard gradient descent: the n^{th} iteration updates θ according to

$$\theta^n = \theta^{n-1} - \delta\theta \frac{\partial \ell}{\partial \theta}, \tag{5.23}$$

where $\delta\theta$ is a pre-defined step vector. The gradient steps are repeated until the gradient vanishes. In practice, the values of $\delta\theta$ are tuned empirically, allowing for three degrees of freedom for ϵ, Ξ and Σ.

5.2.5 Evaluating the log-likelihood using thermodynamic integration

We would like to evaluate the fit of different models based on the log-likelihood, but due to the intractable normalizing factor Y, we need to invoke more elaborate numerical techniques. We do this using a similar thermodynamic integration method to that described by Lartillot and Philippe (2006), and summarized in the second chapter.

First, for $0 \leq \beta \leq 1$, we define:

$$G_\beta(s, c) = \beta \left(\sum_{1 \leq i < i' \leq N} \Delta_{ii'} \epsilon_{s_i s_{i'}} + \sum_{1 \leq i \leq N} \Xi_{s_i}^{v_i} \right) + \sum_{1 \leq i \leq N} \Sigma_{s_i}. \tag{5.24}$$

The associated probability distribution is

$$p_\beta(s \mid c, M) = \frac{1}{Y_\beta} e^{-G_\beta(s,c)}, \tag{5.25}$$

with

$$Y_\beta = \sum_s e^{-G_\beta(s,c)}. \tag{5.26}$$

What we are looking for is $\ln Y_1$. As for $\ln Y_0$, it factors out, and can be computed directly:

$$\ln Y_0 = N \ln \left(\sum_a e^{-\Sigma_a} \right). \tag{5.27}$$

We can thus equivalently evaluate the difference $\ln Y_1 - \ln Y_0$, given by:

$$\ln Y_1 - \ln Y_0 = \int_0^1 \frac{\partial \ln Y}{\partial \beta} d\beta. \tag{5.28}$$

The thermodynamic approximation is obtained by starting a Gibbs sampling of sequences with $\beta = 0$. Following a series of cycles, the value of β is incremented by a small value $\delta\beta$, until $\beta = 1$. Based on the sample of sequences over the entire run, written as $(s^{(h)})_{0 \leq h \leq K}$, the approximation reads as

$$\ln Y_1 - \ln Y_0 \simeq \frac{1}{K} \left[\frac{1}{2} \frac{\partial G(s^{(0)})}{\partial \beta} + \sum_{h=1}^{K-1} \frac{\partial G(s^{(h)})}{\partial \beta} + \frac{1}{2} \frac{\partial G(s^{(K)})}{\partial \beta} \right]. \tag{5.29}$$

In the present conditions, $K = 1,000$ is sufficient to obtain an estimate of $\ln Y_1 - \ln Y_0$ with an error less than one natural log unit.

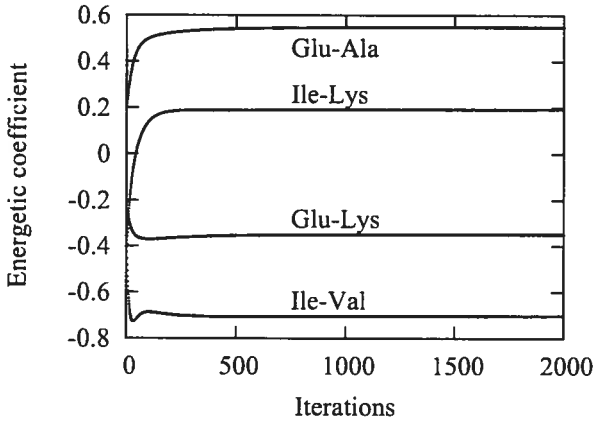

Figure 5.1. Stabilization of pairwise energetic coefficients over a gradient descent optimization.

5.3 Results and discussion

5.3.1 Optimization of the statistical potential

We first performed an optimization of the pure contact potential ($\epsilon + \Sigma$) on each of two data sets. The evolution of a few contact energy coefficients over the course of the gradient optimization are displayed in figure 5.1. The coefficients converge within a few hundred iterations of the gradient descent optimization. We started several optimizations from different initial values and found convergence to essentially identical values (not shown), indicating that the method does not become trapped in local minima. The values also appear to be biologically reasonable, attributing negative energies to known favorable pairwise interactions (e.g., isoleucine-valine), and positive values to known unfavorable pairwise interactions (e.g., glutamate-alanine).

We also compared the values obtained on the two different data sets. Figure 5.2 displays the contact energy coefficients obtained from one data set against those from a the second data set. The correlation is high (0.96), providing a first indication that the data sets are large enough for the learning procedure to reach stability.

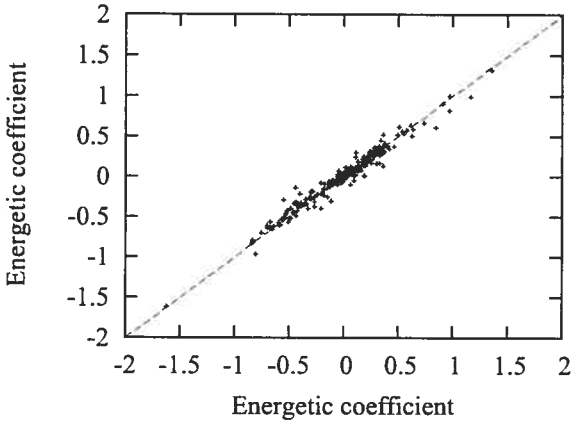

Figure 5.2. XY-plot of pairwise contact energy parameters obtained from the 2 data sets.

5.3.2 Refining functional forms

We next explored different functional forms of statistical potentials, beginning with a pure solvent-accessibility potential. This form is based on classifying amino acid sites into one of several solvent accessibility classes. Here, several choices are possible for the number of classes, but the log-likelihood scores obtained under the different choices cannot be directly compared, since the models do not have the same dimensionality. We thus applied a 2-fold cross-validation procedure, consisting of learning the potential on DS1, and evaluating the log-likelihood using these ML parameter values on DS2 (and vice-versa). Note that since this is a blind test, evaluating the fit of the potential based on data never "seen" by the model, differences in dimensionality are intrinsically accounted for in the assessment. Also note that the cross validation score reported is actually the log-likelihood obtained from the flat potential (based solely on the chemical potential component) minus that under the potential of interest, and multiplied by -1 to make to score positive (the higher the score, the better the model). For the pure solvent potential, figure 5.3 displays the cross-validation score as a function of the number of classes. When W increases, the fit of the model improves, until a point is reached ($W = 16$) where the penalization for model dimensionality starts to dominate

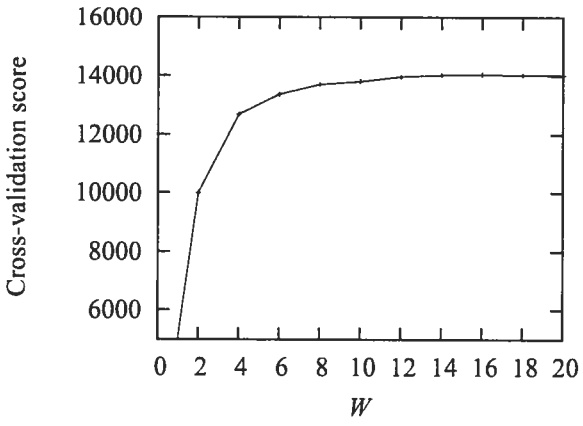

Figure 5.3. Cross-validation score as a function of the number of solvent accessibility classes.

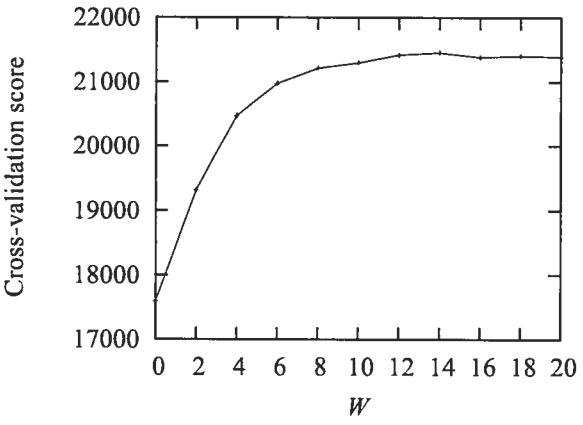

Figure 5.4. Cross-validation score as a function of the number of solvent accessibility classes, with a potential also based on pairwise contacts.

the score.

We applied the same approach utilizing the contact potential as well, displayed in figure 5.4. The model fit displays a similar trend as with the pure solvent potential, reaching an optimal setting at $W = 14$.

We compared the fit of the different forms of potentials, taking the average of the 2-fold CV score: for the pure solvent potential (but always including the chemical potential component) we obtained a score of 14,394; the pure contact potential performs better, with a score of 17,798; and the combined contact and solvent potential performs

best, with a score of 21,058. We also tried using the contact potential of Miyazawa and Jernigan (1985), in which case we include a scaling factor; the β parameter in front of the contact component is optimized by ML, along with the chemical potentials. The resulting potential was the poorest of all, with a score of 11,236. The fact that our potential has a significantly better predictive power than that of Miyazawa and Jernigan (1985) is trivially expected, by construction of the ML potential, and the much larger data set used to derive it. What is more surprising is that the Miyazawa and Jernigan (1985) potential is less fit than a site-independent solvent accessibility profile. A possible explanation would be that their potential is based on the quasi-chemical approximation, which is now known to be somewhat drastic (Godzik et al., 1995; Thomas and Dill, 1996; Skolnick et al., 1997), as it neglects correlations between observed pairing frequencies due to chain connectivity and multiple contacts. Alternatively, this poor fit could mean that potentials optimized for folding are really not suited for protein design purposes. Testing other pairwise contact potentials, in particular those that do not rely on the quasi-chemical approximation (e.g., Maiorov and Crippen, 1992; Tobi and Elber, 2000; Bastolla et al., 2001; Tiana et al., 2004), would be a way to address this issue.

5.3.3 Phylogenetic comparisons

Finally, we computed Bayes factors, as done in the last chapter, but utilizing the newly derived potential. As previously discussed, when using potentials such as those of Miyazawa and Jernigan (1985) in the phylogenetic context, the +F configuration plays the role of a chemical potential. As such, to set up an equivalent dimensionality, we dispensed with the chemical component of the potential, and used the +F configuration in this case as well. Using the MYO*60-153* data set, recall that the log Bayes factor in favor of POISSON+F over POISSON obtained in chapter 4 was 70 natural log units. As a first contrast, we combined our pure contact potential with the POISSON+F config-

uration and obtained a log Bayes factor of 158 (always using the flat POISSON model as a reference). This is already better than the MJ potential in the same combination, which resulted in a log Bayes factor of 129. Thus, using the exact same protein structure representation and parametric form, the method developed in this chapter produced an amelioration of 29 natural log units. Using the richer form of potential, based on both contact and solvent accessibility components, with the same underlying POISSON+F configuration, yields a Bayes factor of 208 natural log units. This is encouraging; using a more refined description of protein structure leads to a better model fit. Note, however, that much work remains, as the overall fit is still much poorer than even the simple rigid JTT matrix, which yields a log Bayes factor of 368. We may be a long way to attaining our basic phenomenological benchmarks with this type of model, and much more work is needed in this direction.

5.4 Conclusions

The central idea of the present chapter is to reformulate the problem of devising statistical potentials for protein design as a statistical inference problem. This formulation, based on the ML principle, led us naturally to a gradient descent method, with the only additional aspect being that the gradient to follow is itself estimated by Monte Carlo averaging. The main advantage of this ML framework is that it guarantees an optimal predictive power of the resulting potential. In addition, it is very general, and can in principle be applied to any form of statistical potential. In particular, it is not restricted to coarse grained descriptions of proteins, and it could also be applied at the atomic level.

In general, the present methodology could be used to investigate many other forms of potentials. As the last sub-section suggests, it would be interesting to extend the overall framework into a pipeline of comparisons within the phylogenetic context as

well. The main hurdle for this last objective is computational, and further algorithmic developments are needed to reduce CPU time. We explore avenues to address this issue in the next chapter.

Chapter 6

Exploring computational strategies

6.1 Introduction

The MCMC methods used in chapters 3 and 4 may be viewed as brute-force approaches to approximating posterior distributions and marginal likelihoods. By this we mean that whenever some density is unavailable by analytical means, we pose a new MH kernel, and revise sampling modules with new update operators, always sampling over the full posterior distribution, or a full path in the space of posterior distributions linking two models. Such brute-force sampling is computationally costly. Alternatives to brute-force sampling, however, are commonly applied in the statistical literature, with first approximations often based on assumptions of normality about a dominant posterior mode (Gelman et al., 2004; Robert and Casella, 2004). These techniques have been adapted to cases of non-analytical models, for instance using MCMC-based optimization schemes (see Robert and Casella, 2004, chapter 5), and while they may fail when the posterior is not normally distributed, or if modes of significant density are missed or ignored, they may still serve as guides for constructing distributions (e.g., as posterior mode finders), and can provide rough estimates of Bayes factors (Gelman et al., 2004).

In this chapter, we further explore the use of different MCMC-based approaches for statistical computation in a phylogenetic context, with the objective of enabling more tractable applications of evolutionary models that are too complex to be manipulated using conventional methods. Utilizing previously presented MH operators, we first illustrate MCMC-based optimization algorithms, which can be used to estimate ML parameter values, even under non-analytical models. The methods are directly related to those developed in the previous chapter. In a second step, MCMC approaches are applied in conjunction with normal developments around the optimal point in order to approximate the posterior distribution. We further combine these different MCMC schemes and normal approximations into a Bayes factor estimator, based on a variant of the Laplace method (Raftery, 1996). The approaches are applied under a fixed tree topology, using three different types of models of amino acid sequence evolution, and the resulting approximations are compared with those obtained under previously available brute-force MCMC methodologies.

6.2 Material and methods

6.2.1 Models

In this chapter, we use the POISSON and WAG models, with the $+\Gamma$ extension in each case. Borrowing the nomenclature of Parisi and Echave (2001), we also use a *structurally constrained* (SC) model, based on the optimal potential derived in chapter 5. Recall that the potential is formulated in terms of a pseudo-energy score of a sequence given a conformation c, written here as $G(s,c)$, and having the following form:

$$G(s,c) = \sum_{1 \leq i < i' \leq N} \Delta_{ii'} \epsilon_{s_i s_{i'}} + \sum_{1 \leq i \leq N} \Xi_{s_i}^{v_i} + \sum_{1 \leq i \leq N} \Sigma_{s_i}. \tag{6.1}$$

As discussed in chapter 5, the first term in (6.1) is the contact component, the second term is the solvent accessibility component, and the last term accounts for compositional effects. Note that here, because we use the chemical component of the potential, we do not need to (but could) invoke the +F configurations. We thus focus on a model entirely based on the potential. As before, the continuous-time Markov chain under this model is specified as a sequence-wide process, with the infinitesimal generator being a $20^N \times 20^N$ matrix R with off entries

$$R_{ss'} = \begin{cases} 0 & \text{if } s \text{ and } s' \text{ differ at more than one position,} \\ e^{\beta[G(s,c)-G(s',c)]} & \text{if } s \text{ and } s' \text{ differ only at one site,} \end{cases} \tag{6.2}$$

where β is a parameter weighting the impact of $G(s, c)$ on the rate of substitution, and where diagonal entries are given from the negative sum of off-diagonal entries. Note that here, because the potentials used have been pre-optimized so as to maximize the stationary probability of the Markov process (on the meta-data set) but with a scaling that implies $\beta = 1/2$, we may fix the β parameter as such, and we refer to the model simply as SC. In others cases, it may be worthwhile to treat β as a free parameter (SC+β), in order to give some flexibility to the model, or if the scaling of the potential is unclear. In addition, it may be pertinent to combine SC and SC+β models with a direct account of rate heterogeneity (+Γ), as was suggested from chapter 4.

6.2.2 Priors

Unless specified otherwise, we use the following default priors:

- $\lambda \sim Exponential$, with a mean determined by a hyperparameter fixed at 0.1;
- $\alpha \sim Exponential$, with mean 1;
- $\beta \sim Uniform[-\beta_{max}, \beta_{max}]$, where $\beta_{max} = 5$.

6.2.3 Alignment, tree, and protein structure

For illustrative purposes, we apply the techniques to the MYO*20-153* data set. The contact map and solvent accessibility profile are derived from PDB accession number 1MBD.

6.2.4 Normal approximation methods

6.2.4.1 Posterior distributions

An alternative to brute-force sampling is to assume that the posterior is normally distributed, and attempt to estimate its mean and variance. For simplicity, we focus on estimating the mean and variance of the posterior of a particular component of the parameter vector, and assume, for now, that the rest of the parameter vector is known (i.e., θ is now univariate). We begin by estimating the mean, and will assume that the prior on θ is uniformly distributed over some interval. Under these conditions, the mean of the posterior distribution corresponds to the maximum likelihood parameter estimate.

First, under analytical models, it is possible to apply the simulated annealing technique proposed by Kirkpatrick et al. (1983). Drawing on an analogy with thermodynamics, the method consists in *heating* the MCMC sampler, by introducing a parameter τ, mediating the *temperature* $(1/\tau)$ of the chain. Setting uniform a prior, the MH kernel becomes

$$\vartheta \;=\; min\left\{1, \left[\frac{p(D \mid \theta', M)}{p(D \mid \theta, M)}\right]^{\tau} \frac{q(\theta', \theta)}{q(\theta, \theta')}\right\}. \tag{6.3}$$

As $\tau \rightarrow \infty$, the chain *freezes*, since an update leading to a lower likelihood has a progressively lower probability of acceptance. The algorithm is thus hoped to converge to the ML point.

An important aspect of simulated annealing is the *cooling schedule*, which is typically explored empirically (Nourani and Andresen, 1998). We explored two types of simple cooling schedules here. The first, which we refer to as *proportional cooling*, updates the value of τ at iteration n according to

$$\tau^n = \tau^{n-1} \times \delta, \tag{6.4}$$

where $\delta > 1$ serves to tune the cooling scheme. In another cooling schedule, referred to as *linear cooling*, τ is updated according to

$$\tau^n = \tau^{n-1} + \delta, \tag{6.5}$$

now with $\delta > 0$, again serving to adjust the cooling rate.

The simulated annealing optimization may be useful in a variety of situations, but may nevertheless be unsuitable when the likelihood is unavailable in closed form. For such situations, however, we may rely on latent state methodologies. Note that when working with a non-analytical model, for instance relying on a DA scheme, the gradient of the log-likelihood is given by

$$\frac{\partial \ln p(D \mid \theta, M)}{\partial \theta} = \langle \frac{\partial \ln p(D, \phi \mid \theta, M)}{\partial \theta} \rangle, \tag{6.6}$$

where $\langle . \rangle$ stands for an expectation over the distribution of latent states. In practice, the gradient can be approximated by

$$\langle \frac{\partial \ln p(D, \phi \mid \theta, M)}{\partial \theta} \rangle \simeq \frac{1}{K} \sum_{h=1}^{K} \frac{\partial \ln p(D, \phi^{(h)} \mid \theta, M)}{\partial \theta}, \tag{6.7}$$

where $(\phi^{(h)})_{1 \leq h \leq K}$ is a set of sampled augmentations, drawn using the first element of the DA module (derivatives are given in Appendix A). This gradient approximation can then be embedded within classical optimization methods, for example, following

the gradient according to an iterative updating, with cycle n given by

$$\theta^n = \theta^{n-1} + \nabla\theta^{n-1}, \tag{6.8}$$

where

$$\nabla\theta^n = \delta\theta \frac{1}{K} \sum_{h=1}^{K} \frac{\partial \ln p(D, \phi^{(h)} \mid \theta^n, M)}{\partial \theta^n}, \tag{6.9}$$

with $\delta\theta$ being a pre-defined step parameter. The iterative cycling between augmentation steps and gradient steps can be repeated until the gradient vanishes, thus declaring the maximum likelihood estimate $\hat{\theta}$. We refer to this algorithm as *Monte Carlo gradient* (MCG) optimization.

It is often also possible to apply the *expectation maximization* (EM) algorithm (Dempster et al., 1977) in conjunction with data augmentation schemes (Wei and Tanner, 1990), using (6.7) as the expectation (E-step) estimate, followed by a maximization (M-step):

$$\theta^n = \underset{\theta}{\text{argmax}} \langle \ln p(D, \phi \mid \theta^{n-1}, M) \rangle \tag{6.10}$$

$$= \underset{\theta}{\text{argmax}} \frac{1}{K} \sum_{h=1}^{K} \ln p(D, \phi^{(h)} \mid \theta^{n-1}, M) \tag{6.11}$$

This inner maximization step is often analytical, but can otherwise be accomplished using gradient or Newton-like methods (see Appendix D). We refer to this algorithm as *Monte Carlo EM* (MCEM) optimization.

Once the mean of the posterior (here, equivalent to the maximum likelihood estimate) has been found, we may estimate the variance at this point as

$$Var(\hat{\theta}) \simeq - \left[\frac{\partial^2 \ln p(D \mid \theta, M)}{\partial \theta^2} \right]^{-1}. \tag{6.12}$$

The second derivative of the log-likelihood may be expressed as

$$\begin{aligned}\frac{\partial^2 \ln p(D \mid \theta, M)}{\partial \theta^2} &= \langle \frac{\partial^2 \ln p(D, \phi \mid \theta, M)}{\partial \theta^2} \rangle \\ &+ \langle \left[\frac{\partial \ln p(D, \phi, \mid \theta, M)}{\partial \theta} \right]^2 \rangle \\ &- \left[\langle \frac{\partial \ln p(D, \phi, \mid \theta, M)}{\partial \theta} \rangle \right]^2 . \end{aligned} \tag{6.13}$$

and the Monte Carlo evaluation is given by

$$\begin{aligned}\frac{\partial^2 \ln p(D \mid \theta, M)}{\partial \theta^2} &\simeq \frac{1}{K} \sum_{h=1}^{K} \frac{\partial^2 \ln p(D, \phi^{(h)} \mid \theta, M)}{\partial \theta^2} \\ &+ \frac{1}{K} \sum_{h=1}^{K} \left[\frac{\partial \ln p(D, \phi^{(h)}, \mid \theta, M)}{\partial \theta} \right]^2 \\ &- \left[\frac{1}{K} \sum_{h=1}^{K} \frac{\partial \ln p(D, \phi^{(h)}, \mid \theta, M)}{\partial \theta} \right]^2 . \end{aligned} \tag{6.14}$$

Analogous schemes for estimating the mean and variance under PX and PX-DA contexts can be devised, and can be extended for joint applications over many parameters (see Appendices C and D).

6.2.4.2 Bayes factors

The Laplace method for estimating the marginal likelihood is given as (see, e.g., Tierney and Kadane, 1986):

$$p(D \mid M) \simeq (2\pi)^y |\tilde{H}|^{1/2} p(\tilde{\theta} \mid M) p(D \mid \tilde{\theta}, M), \tag{6.15}$$

where y is dimension of the model, $\tilde{\theta}$ is the parameter vector maximizing the posterior probability, and $\tilde{H}$ is minus the inverse Hessian matrix (of second derivatives) evaluated at $\tilde{\theta}$. An important variant on (6.15), suggested by Raftery (1996), consists of substituting $\tilde{\theta}$ with $\hat{\theta}$ and $\tilde{H}$ with $\hat{H}$ (the inverse of $\hat{H}$ is otherwise referred to as the

Fisher information matrix). This variant slightly simplifies the mathematical developments, and has the advantage of potential applicability with any maximum likelihood implementation. As such, based on the maximum likelihood parameter vectors of two models ($\hat{\theta}_0$, $\hat{\theta}_1$), we will use the following Laplace approximation to the Bayes factor:

$$\ln B_{01} \simeq \frac{1}{2}(y_1 - y_0)\ln(2\pi) + \frac{1}{2}\ln\left(\frac{\hat{H}_1}{\hat{H}_0}\right) + \ln\frac{p(D \mid \hat{\theta}_1, M_1)}{p(D \mid \hat{\theta}_0, M_0)} + \ln\frac{p(\hat{\theta}_1 \mid M_1)}{p(\hat{\theta}_0 \mid M_0)}. \quad (6.16)$$

The second term in (6.16) can be calculated from the developments in Appendix A. If the models are not analytical, the third term is calculated using a thermodynamic integration method, which, for a given tree configuration, computes the log-likelihood difference under two different models. These calculations are restricted versions of the more general thermodynamic integrations methods for evaluating differences of log marginal likelihoods between pairs of models.

6.3 Results and discussion

6.3.1 MCMC-based optimization: an analytical example

Before applying the methods developed above to non-analytical models, we first explore the properties of MCMC-based optimizations under a simpler case, where comparisons can be made with other implementations. In particular, we apply different approaches to maximizing the likelihood with respect to branch lengths for a given topology under the WAG model.

First, under this model, the simulated annealing method can be applied. Figure 6.1a shows the evolution of the overall tree length during the first 100 iterations of a simulated annealing run, based on a proportional cooling schedule, with the initial $\tau = 1$ increased to $\tau > 10,000$ according to (6.4), with $\delta = 1.1$. As can be seen, the chain begins with a somewhat erratic behavior, oscillating around, yet progressively

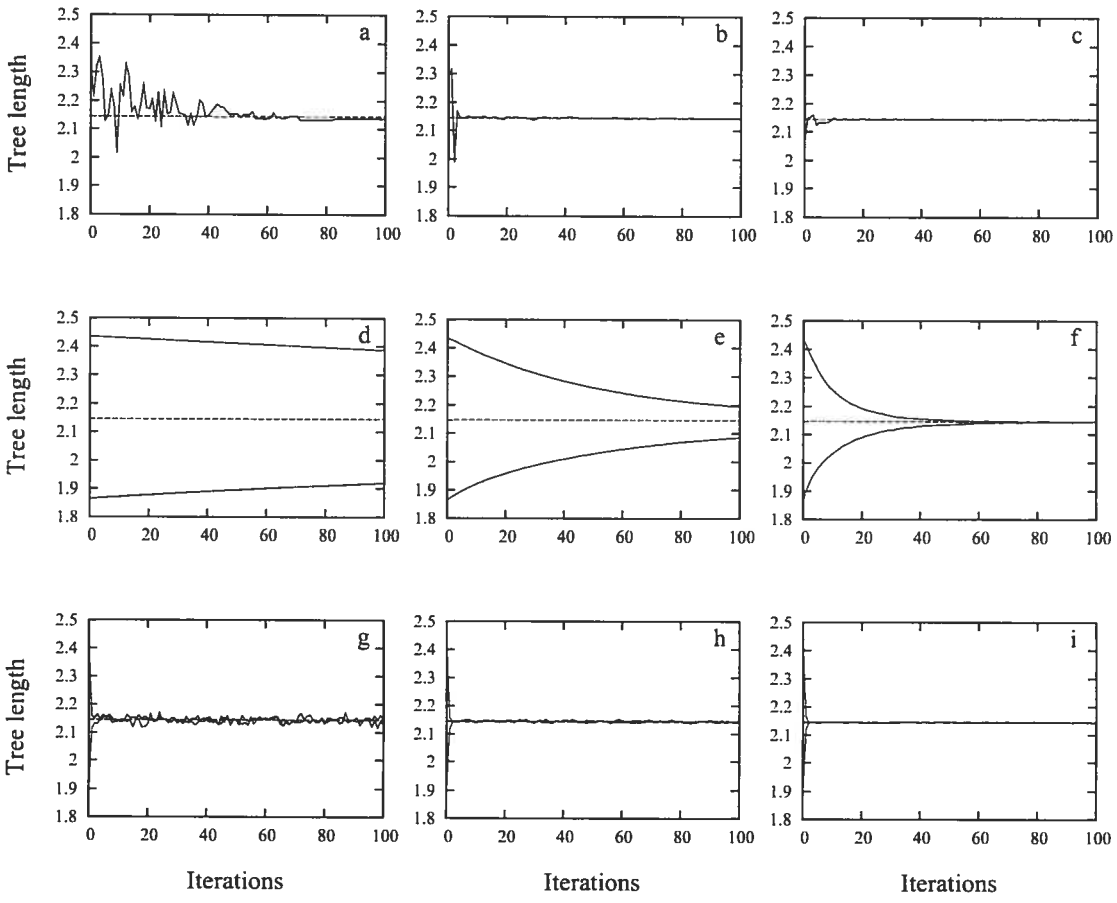

Figure 6.1. Markov chain Monte Carlo maximum likelihood estimation of the tree length. In a), b), and c) simulated annealing is used. In d), e) and f) we use MCG based on a sample of 100 mappings. In g), h), and i) we use MCEM, based on 10 (g), 100 (h), and 1000 (i) mappings. In each panel, a dashed line is drawn for the tree length returned by PAML (Yang, 1997).

gravitating towards, the tree length obtained using the PAML package (Yang, 1997). Ultimately, however, after 100 iterations, the chain slightly misses the mark.

We found the linear cooling scheme easier to adjust than proportional cooling, and less likely to become trapped in sub-optimal configurations as the chain approaches the freezing point. In figure 6.1b, we started from $\tau = 1$, and updated according to (6.5), with $\delta = 100$. The chain converges to essentially identical branch length values as returned by PAML in about 35 iterations. Tuning $\delta = 500$ (fig. 6.1c), the maximum likelihood branch lengths were obtained in about 18 iterations.

We next explored the MCG algorithm, as a first latent state optimization scheme. Nielsen (2002) has proposed a straightforward DA method, which, under models like WAG, allows for a direct sampling of substitution mappings. We used Nielsen's method to draw a sample of mappings for estimating the log-likelihood gradient, as written in (6.7), in a MCG optimization of branch lengths. The needed derivatives are given in the Appendix A. As illustrated in figure 6.1d, e and f, a significant amount of trial-and-error tuning of the gradient optimization method can be important for reducing CPU time. In this case, expectations were estimated based on a sample of 100 mappings, and only the step parameters ($\delta\lambda_j$) were adjusted. As crude explorations, we set the same value for each branch length step parameter throughout the run, with $\delta\lambda_j = 0.000001$ in 6.1d, $\delta\lambda_j = 0.00001$ in 6.1e, and finally $\delta\lambda_j = 0.00005$ in 6.1f.

We also tried the MCEM algorithm in the present example. We once again relied on Nielsen's method, drawing samples of substitution mappings for estimating expectations, followed by the maximizations step given in (6.11). In this case, the precision of the algorithm depends solely on the sample size used to estimate the expectation, since the maximization step is analytical (see Appendix D). Using a sample of 10 mappings, significant fluctuations of the overall tree length are observed from one MCEM iteration to the next (fig. 6.1g). Fluctuations are reduced using 100 mappings (fig. 6.1h), and

become negligible (± 0.001 natural log-likelihood units) using 1000 mappings (fig. 6.1i).

This corroboration across methods, as well as with the PAML package, is a useful check, and helps in getting a sense of the general behavior of the MCMC methods. In this particular case, we give preference to the MCEM algorithm, if only for the fact the tuning is exclusively based on sample size for the E-step. In fact, the sample size can be increased "online", for instance, by a factor of 10 every 10 iterations, or according to any other scheme. It should be noted, however, that the Monte Carlo error only decreases with the square root of the sample size, and that the MCEM is not necessarily the best choice for all contexts in terms of computational requirements, as we illustrate in an example below.

6.3.2 MCMC-based optimization: non-analytical examples

In the preceding subsection, we applied Monte Carlo techniques for parameter optimization to a case where such methods are unnecessary. In this section, we explore non-analytical models for which standard optimization techniques are not directly possible.

Our first non-analytical example consists of optimizing the shape parameter α for the $+\Gamma$ model, still using the WAG matrix, and, for now, with fixed branch lengths (as obtained under WAG). Figure 6.2 shows the progression of α as a function of the MCEM iterations, with two different initial values. Once again, the MCEM algorithm converges quickly—within about 20 iterations—and the fluctuations of the estimate progressively decreases as the sample size used in each iteration increases from 10 (fig. 6.2a) to 100 (fig. 6.2b), to 1000 (fig. 6.2c). The final value reached is $\hat{\alpha} = 0.73$. Although this estimate is not directly comparable with the discrete gamma models, we ran PAML using different numbers of categories. In general, the estimates are quite similar; using 4 categories, PAML returns $\hat{\alpha} = 0.72$, 8 categories gives $\hat{\alpha} = 0.69$, 16 categories gives $\hat{\alpha} = 0.68$, and

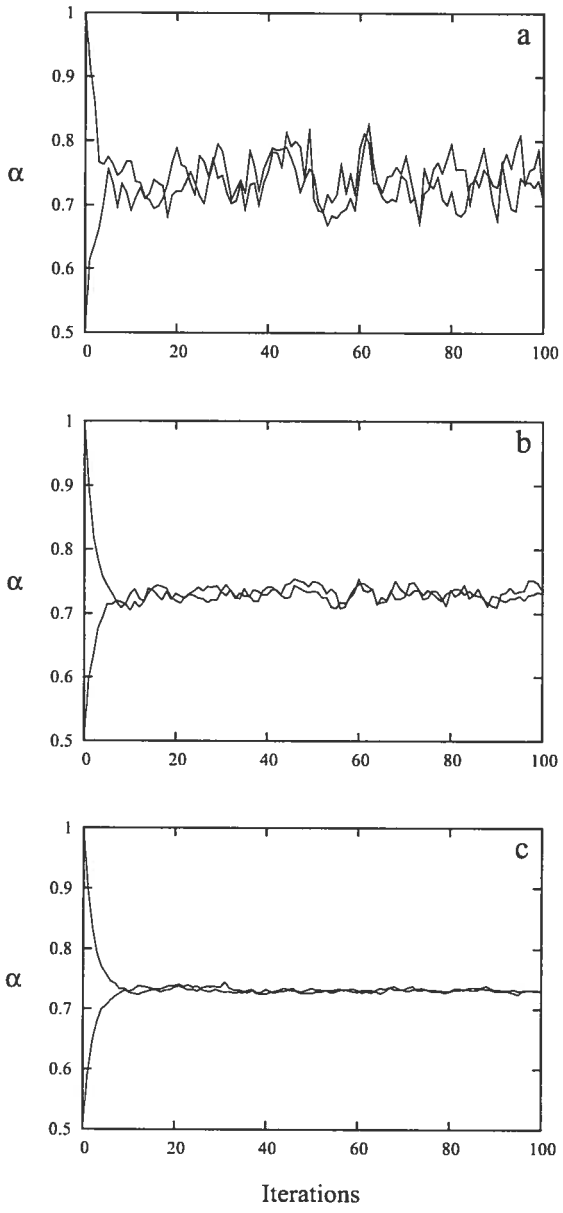

Figure 6.2. MCEM algorithm for estimating $\hat{\alpha}$. The E-step of the algorithm—Monte Carlo estimate of the expectation—is done with 10 (a), 100 (b) and 1000 (c) draws.

finally, using 32 categories, the estimate is $\hat{\alpha} = 0.70$. These mild fluctuations illustrate how the number of categories used alters the gamma approximation, and while the discrete approximation may be suitable for many practical applications, PX methods for continuous distributions could have several advantages (Mateiu and Rannala, 2006), particularly when discretization procedures are in doubt (e.g., Yang et al., 2000a; Susko et al., 2003; Mayrose et al., 2005), or when site-specific random variables are multivariate (e.g., Lartillot and Philippe, 2004; Kosakovsky Pond and Muse, 2005).

Our next non-analytical example concerns the SC+β model, where we wish to optimize β, still based on a fixed set of branch lengths. We first ran an MCEM optimization using a sample of 100 mappings, and 100 sequences (for the approximation given in (C.13)). Figure 6.3a shows the first 20 iterations of the MCEM, which displays a jagged behavior in attempting to adjust the value of β so as to cancel out two key components of the derivative of the log-likelihood function (see eqn. C.9). In contrast, the MCG optimization under the same sample size conditions is much more efficient, converging with 5 iterations (fig. 6.3b).

In both of these examples, it is interesting to note that while we have adjusted parameters so as to maximize the log-likelihood, we have not computed the log-likelihood itself. This decoupling between log-likelihood optimization and log-likelihood calculation is a key feature of latent state methodologies, and is analogous to the property allowing us to sample from the posterior without having a closed form likelihood.

6.3.3 Normal approximations of posterior distributions

The use of normal approximations in Bayesian analysis often serves as a first step to constructing posterior distributions under new statistical models (Gelman et al., 2004). We consider +Γ and SC-type models here, and focus on their distinguishing parameters (α and β respectively).

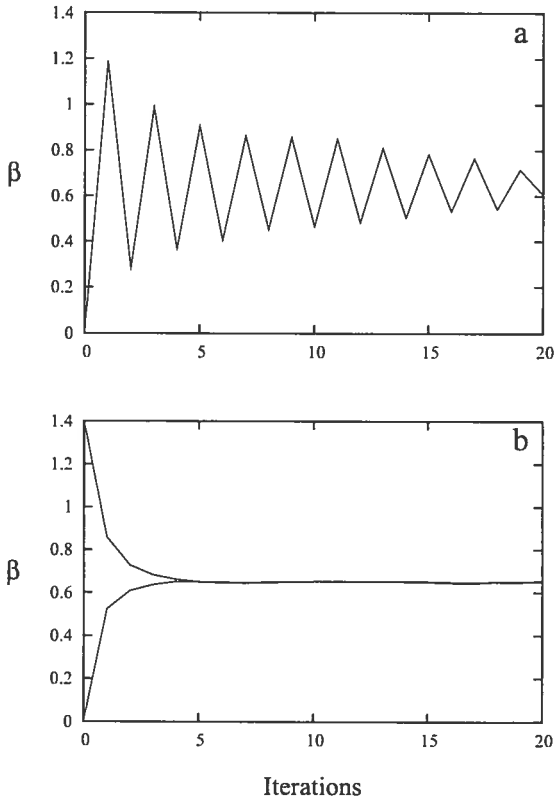

Figure 6.3. Monte Carlo estimation of $\hat{\beta}$. In a), the MCEM is used, with the Monte Carlo estimate of the expectation based on 100 draws. In b), the MCG is used with 100 draws.

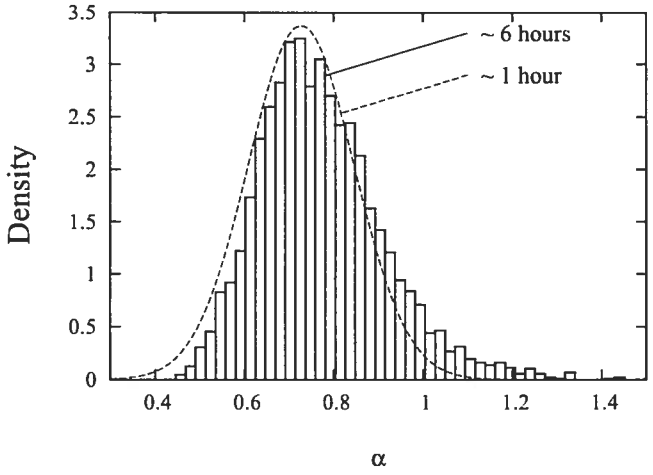

Figure 6.4. Posterior density plot of α, approximated using full MH sampling (histogram) and a normal approximation (dashed line).

First, under the WAG+Γ model, we marginalized over branch lengths using a PX sampling module, while optimizing with respect to α (here given a uniform prior) using the MCEM algorithm. Doing so simplifies the example, in that it remains univariate, while allowing us to focus on the full posterior of α. The final MCEM iterations were based on a sample of 100 sets of branch lengths and rate vectors, as was the variance estimate (referring to eqn. 6.12). We used these estimates as the mean and variance for tracing a normal probability density function, and compared this trace to the density histogram obtained using the PX module sampling branch lengths and α (fig. 6.4). The two different density plots are reasonably similar, although the histogram appears skewed to the right, particularly when $\alpha > 1$. Indeed, in this range, the shape of the gamma distribution does not undergo dramatic changes with small variations in α, which leads to a flattened out likelihood surface. This illustrates an important point: the full posterior may differ from a normal distribution, and such approximations are only meant to give a general sense of location and diffuseness for a parameter of interest.

Our second example concerns the SC+β model, where alternatives to constructing posterior distributions are of particular interest. Under these models, the MH kernel

includes the ratio $p(s_0 \mid \theta', M)/p(s_0 \mid \theta, M)$, which requires the evaluation of the ratio of normalizing constants given by

$$Y = \sum_s e^{-2\beta G(s,c)} \tag{6.17}$$

The normalizing constant in (6.17), however, is not tractable, and the MH kernel itself must therefore be approximated. Previous works investigating SC-type models have relied on an importance sampling approximation proposed by Robinson et al. (2003). Adapted to the present context, the approximation reads as

$$\frac{Y_\beta}{Y_{\beta'}} = \frac{\sum_s e^{-2\beta G(s,c)}}{\sum_s e^{-2\beta' G(s,c)}} \tag{6.18}$$

$$\simeq \frac{\sum_{h=1}^{K} e^{-2(\beta-\beta^*)G(s^{(h)},c)}}{\sum_{h=1}^{K} e^{-2(\beta'-\beta^*)G(s^{(h)},c)}}, \tag{6.19}$$

where $(s^{(h)})_{1\leq h\leq K}$ is a set of sequences sampled using a Gibbs sampling approach discussed in chapter 3, and where β^* is chosen to be as close as possible to the middle of β and β'. We used our variation given in chapter 3 for choosing β^* during the MCMC, and ran a full sampling over branch lengths and β. In another run, we marginalized over branch lengths using a DA module, while optimizing β using the MCG algorithm. We relied a sample of 100 sets of branch lengths (and mappings) and 100 sequences (see Appendix C, eqn. C.13 and C.30) for the final iterations of the MCG method, and for the subsequent variance estimate. As shown in figure 6.5a, the normal probability density function based on the mean and variance estimates matches well with the density histogram obtained using the full MCMC, although the normal approximation comparatively underestimates the variance to a small degree. Given that the full MCMC sampling of β is based on the approximation in (6.18), which could breach the conditions of Markov chain convergence theorems, we performed a third run, using the thermodynamic integration method described in chapter 4. This last method has

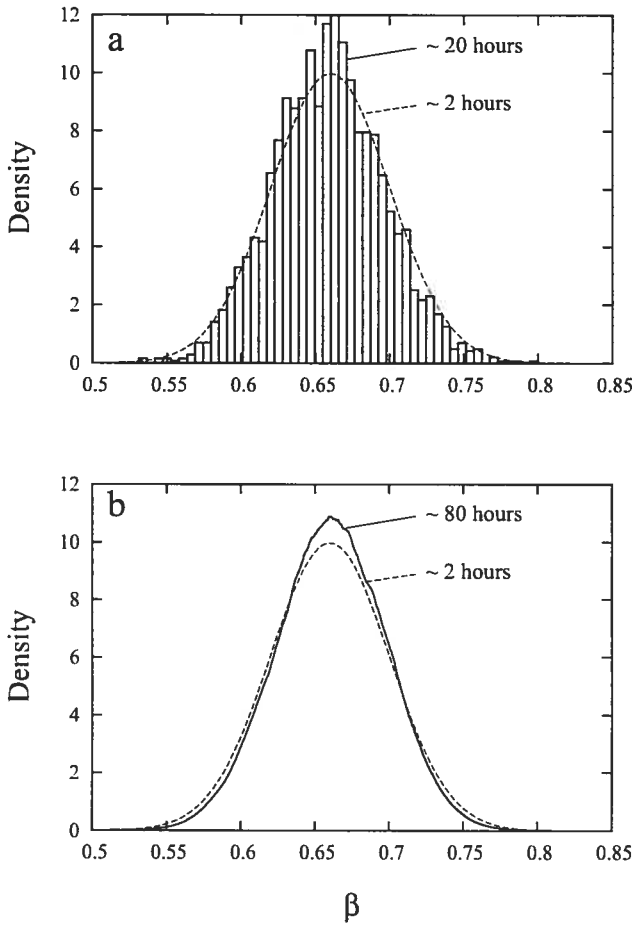

Figure 6.5. Posterior density plot of β. In panel a) a histogram was generated using a full MH sampling. Panel b) shows a density trace generated using thermodynamic integration, as presented in chapter 4. In both panels, the normal approximation is shown (dashed line).

the advantage of arbitrary accuracy, at the cost of CPU time, and the slight ruggedness of the posterior density trace (fig. 6.5b) gives a qualitative sense of the Monte Carlo fluctuations over the course of the integration. Here again, though, the posterior density of β obtained using the thermodynamic method matches well with the normal approximation, providing a reasonable corroboration across all methods. On the other hand, the normal approximation requires only a fraction of the CPU time of either of the two other methods. This may prove useful when the main interest is the posterior distribution of β or analogous parameters (Robinson et al., 2003), particularly when approximating posteriors over several different data subsets (Yu and Thorne, 2006).

6.3.4 Normal approximation of Bayes factors

Finally, we applied the Laplace normal approximation approach to estimate Bayes factors across all models mentioned herein, as well as the thermodynamic integration methods. As mentioned previously, the thermodynamic method can be tuned to any desired accuracy, and we use the results under this approach as our reference values. Our crude strategy here consisted in running triplicates of each type of calculation, progressively tuning the MCMC samplers such that, when rounding to the nearest natural log unit, identical results are obtained for all three runs. We then compared accuracy and CPU time of the two methods.

For the Laplace method, we first maximized the log-likelihood with respect to branch lengths and, as applicable, α and β. For all but SC-type models we used the MCEM algorithm for the overall optimization. For SC+β-type models, however, we used a combined MCEM-MCG algorithm, which, at each iteration, performs an M-step on branch lengths (and α, if applicable) and a gradient step on β. In all cases, the final expectation estimates for optimization and for the Laplace approximation were based on samples of 10,000 substitution mappings and rate vectors. For a particular configu-

Table 6.1. Natural logarithm of the Bayes factor for models considered, with POISSON used as a reference.

Model	Thermodynamic	Laplace
POISSON	0	0
POISSON+Γ	81 (11)	81 (2)
WAG	294 (14)	294 (2)
WAG+Γ	373 (26)	372 (5)
SC	162 (65)	162 (4)
SC+Γ	253 (131)	253 (6)
SC+β	167 (129)	167 (5)
SC+$\beta+\Gamma$	268 (197)	269 (8)

Note.—Numbers in parenthesis indicate approximate CPU time in hours.

ration, computing log-likelihood differences between an analytical and a non-analytical model was done using a constrained thermodynamic method.

The resulting Bayes factors are remarkably accurate, when compared to full thermodynamic estimates, with at most one log unit difference (table 6.1). Importantly, however, the Laplace approximation required much less CPU time. The reasons for such a reduced computational time are a combination of several factors. First, as opposed to a full MCMC sampling over all admissible parameter settings, optimizations are directed toward a single optimal point. If convergence to this point is fast, far fewer likelihood function evaluations will be needed than would a full-blown sampling from the posterior. Also, the algorithms can be used with very small samples (of say 10) to obtain crude parameter estimates to be used as the starting point of a more refined MCEM or MCG runs, and so on. Indeed, in our analyses, we always preceded the final iterations of MCEM or MCG with such crude estimations, which could be

obtained within minutes. Next, the MCEM and MCG algorithms, and the constrained thermodynamic method considerably reduce the overall sampling, as marginalization *via* MCMC is focused on latent states. Lastly, the Laplace approximation for Bayes factors makes the assumption of normality around the optimal point, and makes use of an estimate of the curvature of the likelihood surface; loosely speaking, the full thermodynamic method must effectively obtain this information using brute-force sampling.

The model rankings obtained using Bayes factors give favor to the WAG+Γ model. Note, however, that the pure SC model outperforms the POISSON+Γ, although it is in turn outperformed by the pure WAG model. This ranking is reasonably encouraging for SC-type models, and additional work is needed to determine if more sophisticated statistical potentials can achieve, or surpass, the performance of the best site-independent models.

6.4 Conclusions

Complementing MCMC methods and normal approximations considerably reduces the needed computational resources for conducting Bayesian calculations. Also, we stress here that while common Bayesian discourses often describe MCMC methodologies as alternatives giving non-analytical modeling flexibility (e.g., Paap, 2002; Brooks, 2003; Beaumont and Rannala, 2004), such features are not exclusive to Bayesian contexts. As we have shown here, MCMC techniques can also be used to instantiate the ML principle, such that they may be viewed general, and independent of any particular probabilistic paradigm.

However, as evolutionary models increase in sophistication, and as the needed sampling schemes become more elaborate, or based on additional levels of approximation, the difficulties commonly associated with MCMC devices (e.g., assessment of convergence and mixing behavior) are likely to be exacerbated, and the methods should be

approached with caution. Also, the choice among possible MCMC methods can be bewildering (see, e.g., Gelman et al., 2004; Robert and Casella, 2004), and it may be difficult to know beforehand which overall scheme gives sufficiently accurate estimates in reasonable compromise with computational effort. As as we illustrated for optimizations of β, sampling and algorithmic choices will likely need to be explored empirically for each new context.

The approaches employed here could also be adapted and reconfigured in several ways. For instance, here, for estimating Bayes factors, we used thermodynamic MCMC to integrate over latent states and the Laplace method to integrate over parameter space. However, if greater accuracy were needed, or if assumptions of normality no longer hold, we could also extend the thermodynamic method over any sub-set of parameters and apply the Laplace method over the remaining parameter(s). In addition, other approaches to the Laplace method have been proposed, several of which do not require computing derivatives. Referring to equation (6.15), $\tilde{\theta}$ and $\tilde{H}$ could be approximated based on the output of a plain MCMC run, using the component-wise posterior mean or median and the posterior variance-covariance matrix; other choices are also given in Lewis and Raftery (1997).

These sorts of approaches could enable a larger scale empirical project, in order to compare a broader set of models, and in particular, models based on the gamut of forms of statistical potentials.

Chapter 7

Comparing codon models of substitution

7.1 Introduction

In chapter 3 we initiated the first and second steps of the Bayesian framework: setting up a full probability model, and conditioning on true data. The third step of assessing model fit was addressed in chapter 4. Our phenomenological benchmarking approach of measuring the fit of site-interdependent models based on the simple forms of statistical potentials has suggested that the modeling approach poorly anticipates basic properties of the substitution process, while inducing heavy computational demands. Nonetheless, the models do show some promise; they always lead to *some* improvement in fit. To address the problems raised in chapter 4, we have proposed a framework for ameliorating the form of statistical potentials in chapter 5, and we have investigated possible computational alternatives in chapter 6.

As previously stated in chapter 3, however, in performing this exercise at the amino acid level only, we have relinquished the more attractive codon level interpretation of molecular evolution. Indeed, evolutionary models at the amino acid level should not

be viewed as reflecting basic biological knowledge. Rather, we have proposed they should be viewed as phenomenological benchmarks. Loosely speaking, in this particular modeling context, we would like to see basic evolutionary properties such as rate heterogeneity, and plausible amino acid exchangeabilities, *emerge as a result of our explicit structural modeling.* These are but the first evolutionary structural modeling attempts, and as progress is made in this direction it will be of interest to return to the codon-based mechanistic modeling, as proposed by Robinson et al. (2003), in order to construct a more realistic description of molecular evolution, which better formalizes basic biological understanding.

We presented one form of site-independent codon model in chapter 2, based on the formulation of Goldman and Yang (1994) (GY). However, another formulation was proposed by Muse and Gaut (1994) (MG), and both formulations have since been extended and modified in numerous ways. Many of these codon substitution models have not been assessed in the Bayesian framework we adopt in this work, and, more importantly, several do not appear to be based on a logical mechanistic modeling construction. These issues raise many questions, forestalling the development of structural models based on statistical potentials in the codon context. We expand these questions below.

Recall that the traditional nucleotide-level of interpretation surmises the data as arising from a continuous-time Markov process running along the branches of the phylogeny, with a state space consisting of the four different nucleotides. In its general form (e.g., Lanave et al., 1984), the model is specified from six relative exchangeability parameters, for each possible pair of nucleotides, and four stationary probabilities, or nucleotide propensities, and is often referred to as the *general time reversible* (GTR) model. Taking this model as a starting point in the case of protein-coding sequences, a first step to mechanistically acknowledging the coding nature of the data is to suppose a strong purifying selection against stop codons, and to re-formulate the process

in a state space consisting of nucleotide triplets, but now omitting triplet states corresponding to stop codons. In effect, such a model is equivalent to the same GTR-type of model applied at the nucleotide level, but with the constraint that the nucleotide sequence must encode some full length amino acid sequence (one third the length of the nucleotide sequence). This is the rational of the MG-style of codon substitution model. From this point, a further model construction step in the MG-style is to distinguish between synonymous and nonsynonymous events, for instance utilizing the parameterization presented in the original work, or the more compact representation of fixing the synonymous rate factor at one, and treating the nonsynonymous rate factor as a free parameter.

In contrast with the MG-style of model formulation, which the authors first described as having entries of the Markov generator proportional to "the equilibrium frequency of the *target nucleotide*" (Muse and Gaut, 1994, p. 717), the GY models have entries of the Markov generator proportional to the stationary probability of the *target codon*. The contrast between the two formulations can be made very subtle. Indeed, a GY-style model can be specified from the same six nucleotide relative exchangeability and four nucleotide propensity parameters used in the MG-style model above: codon stationary probabilities are approximated as proportional to the product of the three propensity parameter values associated with the nucleotides at the three codon positions. However, such a model entails peculiar properties. For instance, in a mutational context prone to events leading to A or T, a substitution from codon CGC to CTC would have a lower instantaneous rate than a substitution from codon ATA to AGA; the rate of an event involving the second codon position depends on the nucleotide states at the first and third positions, which, in this case, leads to the higher rate for the substitution going against the mutational bias. From the mechanistic model construction described above, however, there are no obvious reasons for linking a change

at the second position to the states at the first and third positions, unless this is mediated by selective effects at the codon level (e.g., stop codons). Accordingly, for this same instance, the MG model displays the reverse situation, with the CGC to CTC substitution having a higher instantaneous rate than the ATA to AGA substitution in a manner consistent with the mutational bias.

Another widely used modeling idea, adopted in both MG and GY formulations, has been to assign a separate set of nucleotide propensity parameters to each of the three codon positions. The distinction with the previously mentioned models is commonly referred to as F1×4 versus F3×4, reflecting the use of a single versus three vectors of dimension 4. From the mechanistic standpoint, however, the F3×4 configuration stands only as a phenomenological account of how the coding structure of the data induces a periodic pattern at each codon position. There is no natural interpretation to modeling features induced by the coding nature of the sequences via an expanded parameterization at the nucleotide level. Differences observed at each of the three codon positions are most likely the result of factors bearing on amino acid or codon preferences, or other high-order features, and should logically be modeled as such.

A further option available in the GY-style is based on a full 61-dimensional (assuming the universal genetic code) vector of codon stationary probabilities (indicated as F61, e.g., Huelsenbeck and Dyer, 2004; Huelsenbeck et al., 2006; Yang, 2006). The GY-F61 approach has been suggested as important in giving "more freedom for the model to explain the data by modifying substitution rates using codon frequencies" (Huelsenbeck and Dyer, 2004, p. 670). This may be the case, but the GY-F61 model again has no natural mechanistic interpretation; nucleotide propensities have no direct parameterization in this formulation, but are only implicitly modeled, in manner confounded with other effects inducing uneven codon stationary probabilities.

The impacts of the MG versus GY formulations, and the F1×4, F3×4 and F61 (in

the GY context) configurations, on overall model fit have not yet been explored within a single encompassing probabilistic framework. Most works promoting the GY-F61 formulations (e.g., Huelsenbeck and Dyer, 2004) are based on qualitative inspections of parameter values, but without any quantitative model comparison measurements. Perhaps even more surprisingly, the majority of codon-based model explorations have focused on distinctions other than those of the GY versus MG approach, and the few notable exceptions to this trend (Kosakovsky Pond, 2005; Kosakovsky Pond and Frost, 2005; Ren et al., 2005; Aris-Brosou and Bielawski, 2006) have only considered the F3×4 configurations. In such contexts, Kosakovsky Pond and Muse (2005) have concluded that the GY versus MG distinction "[...] leads to small (but typically negligible) differences [...]" (p. 2375). Also based on results from the F3×4 versions of the GY and MG formulations, Aris-Brosou and Bielawski (2006) have suggested that the optimal choice may often vary with the data considered, and have called for "[...] more effort devoted to understanding and carefully modeling the relationship between mutation process acting on protein coding genes and the precise parameterization of equilibrium frequencies in codon substitution models." (p. 63)

In this chapter, we construct a Bayesian ranking of codon substitution models, based on the evaluation of Bayes factors (Jeffreys, 1935; Kass and Raftery, 1995). To this analysis, we incorporate new models in the MG-style, which allow for a flexible account of either global amino acid preferences or global codon preferences, and which subscribe more closely to the mechanistic standpoint of separating mutational and selective features of the overall evolutionary process. We also include in our analysis all of the above-mentioned GY and MG-style models, comparing the F1×4, F3×4 and F61 (in the GY context) configurations, and contrasting each case with a modeling of nonsynonymous rate heterogeneity using the *Dirichlet process* apparatus (Huelsenbeck et al., 2006). Using three real data sets, our findings indicate that alternative configu-

rations of the GY and MG-style models can lead to considerable differences in overall model fit, to an extent sometimes greater than the contrast between homogeneous and heterogeneous (across sites) nonsynonymous rates.

7.2 Material and methods

7.2.1 Data

We used the GLOBIN*17-144*, LYSIN*25-134*, and HIV*22-99* data sets, and in all three cases used the same topologies as those used in the works cited for each data set (see chapter 2).

7.2.2 Models

In the next subsections, we describe the model components in detail, constructing the entries of rate matrix Q following modeling approaches inspired from Muse and Gaut (1994) and Goldman and Yang (1994). The models are not identical to those presented in these original works, but correspond to flexible generalizations, while allowing us to focus on the distinguishing features of interest.

7.2.2.1 MG-style models

We begin with the mechanistic modeling standpoint proposed by Muse and Gaut (1994), with a Markov generator given by

$$Q_{ab} \propto \begin{cases} \varrho_{a_c b_c} \varphi_{b_c}, & \text{if } \mathcal{A}, \\ \omega \varrho_{a_c b_c} \varphi_{b_c}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{7.1}$$

where

$\mathcal{A}$: a and b are synonymous, and differ only at codon position c;

$\mathcal{B}$: a and b are nonsynonymous, and differ only at codon position c;

and where

- $\varrho = (\varrho_{lm})_{1 \leq l,m \leq 4}$ is a set of (symmetrical) nucleotide relative exchangeability parameters, with the (arbitrary) constraint $\sum_{1 \leq l < m \leq 4} \varrho_{lm} = 1$;
- $\varphi = (\varphi_m)_{1 \leq m \leq 4}$, with $\sum_{m=1}^{4} \varphi_m = 1$, represents a set of global nucleotide equilibrium propensities;
- and ω is the coefficient bearing on nonsynonymous rates, for now treated as a global parameter.

When $\omega = 1$, this model corresponds to the well-known GTR model invoked for nucleotide level interpretations, but with the purifying constraint against all stop codons. Here, however, ω is always treated as a free parameter, and the model is referred to as MG-F1×4.

Following in the MG-style, one way of modeling factors bearing on codon preferences is given as

$$Q_{ab} \propto \begin{cases} \varrho_{a_c b_c} \varphi_{b_c} \left(\frac{\psi_b}{\psi_a}\right)^{\frac{1}{2}}, & \text{if } \mathcal{A}, \\ \omega \varrho_{a_c b_c} \varphi_{b_c} \left(\frac{\psi_b}{\psi_a}\right)^{\frac{1}{2}}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{7.2}$$

where $\psi = (\psi_b)_{1 \leq b \leq 61}$, with $\sum_{b=1}^{61} \psi_b = 1$, represents a set of 61 codon preference parameters, and where the exponent $\frac{1}{2}$ ensures reversibility (see Appendix E). Entries corresponding to substitutions from an unpreferred codon to a preferred codon ($\frac{\psi_b}{\psi_a} > 1$) will thus be higher than entries corresponding to substitutions from a preferred to an unpreferred codon ($\frac{\psi_b}{\psi_a} < 1$), and in this way, an explicit account of global codon preference

(CP) is included, while maintaining an account of background nucleotide propensities. We refer to this model as MG-F1×4-CP.

Note that the codon preferences captured by ψ can be the result of several factors, including, for instance, global amino acid preferences. One way of assessing whether the CP model is capturing effects beyond those of global amino acid preferences is to compare it with a simplified version of the CP formulation, which accounts only for such features as given by

$$Q_{ab} \propto \begin{cases} \varrho_{a_c b_c} \varphi_{b_c}, & \text{if } \mathcal{A}, \\ \omega \varrho_{a_c b_c} \varphi_{b_c} \left(\frac{\varpi_{f(b)}}{\varpi_{f(a)}} \right)^{\frac{1}{2}}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise}, \end{cases} \tag{7.3}$$

where $\varpi = (\varpi_k)_{1 \leq k \leq 20}$ is a 20-dimensional vector associated with amino acid preferences (AAP), and where $f(a)$ returns an index corresponding the amino acid encoded by codon a. As in the case of the CP model, entries corresponding to substitutions from an unpreferred amino acid to a preferred amino acid ($\frac{\varpi_{f(b)}}{\varpi_{f(a)}} > 1$) will thus be higher than entries corresponding to substitutions from a preferred to an unpreferred amino acid ($\frac{\varpi_{f(b)}}{\varpi_{f(a)}} < 1$). We refer to this model as MG-F1×4-AAP model.

Finally, despite departing from the mechanistic modeling perspective, we also investigate the F3×4 configurations for the models defined in (7.1), (7.2), and (7.3), by substituting φ appropriately with codon position specific nucleotide propensity parameters, written as $\varphi^{(c)} = (\varphi_m^{(c)})_{1 \leq m \leq 4}$, where $\forall c,\ 1 \leq c \leq 3$, $\sum_{m=1}^{4} \varphi_m^{(c)} = 1$. The MG-F3×4 model is thus given by

$$Q_{ab} \propto \begin{cases} \varrho_{a_c b_c} \varphi_{b_c}^{(c)}, & \text{if } \mathcal{A}, \\ \omega \varrho_{a_c b_c} \varphi_{b_c}^{(c)}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise}, \end{cases} \tag{7.4}$$

the MG-F3×4-CP model by

$$Q_{ab} \propto \begin{cases} \varrho_{a_c b_c} \varphi_{b_c}^{(c)} \left(\frac{\psi_b}{\psi_a}\right)^{\frac{1}{2}}, & \text{if } \mathcal{A}, \\ \omega \varrho_{a_c b_c} \varphi_{b_c}^{(c)} \left(\frac{\psi_b}{\psi_a}\right)^{\frac{1}{2}}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{7.5}$$

and the MG-F3×4-AAP model by

$$Q_{ab} \propto \begin{cases} \varrho_{a_c b_c} \varphi_{b_c}^{(c)}, & \text{if } \mathcal{A}, \\ \omega \varrho_{a_c b_c} \varphi_{b_c}^{(c)} \left(\frac{\varpi_{f(b)}}{\varpi_{f(a)}}\right)^{\frac{1}{2}}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise.} \end{cases} \tag{7.6}$$

7.2.2.2 GY-style models

The models in the style proposed by Goldman and Yang (1994) have Markov generators specified as

$$Q_{ab} \propto \begin{cases} \varrho_{a_c b_c} \pi_b, & \text{if } \mathcal{A}, \\ \omega \varrho_{a_c b_c} \pi_b, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{7.7}$$

where $\pi = (\pi_b)_{1 \leq b \leq 61}$, with $\sum_{b=1}^{61} \pi_b = 1$, represents a 61 dimensional vector of codon stationary probabilities (distinct from ψ).

Several options for π are available. First, it can be based on a set of global nucleotide propensity parameters according to

$$\pi_a \propto \varphi_{a_1} \varphi_{a_2} \varphi_{a_3}. \tag{7.8}$$

We refer to this model as GY-F1×4. Another similar choice is to base π on codon-

position-specific nucleotide equilibrium frequencies:

$$\pi_a \propto \varphi^{(1)}_{a_1}\varphi^{(2)}_{a_2}\varphi^{(3)}_{a_3}, \tag{7.9}$$

in which case we refer to the model as GY-F3×4. Note that the GY-F1×4 and MG-F1×4 models, as well as the GY-F3×4 and MG-F3×4, are respectively constructed from the exact same parameters; they also have the same stationary distributions, and hence differ only in terms of their transient specifications (further details on this point are given in Ren et al., 2005, as well as in the Appendix E). Finally, we consider the case where π is directly free, conditioning the full 61-dimensional vector to the data, which we refer to as GY-F61.

The limiting distributions of all models are given in full in Appendix E, along with further details specific to our implementation.

7.2.3 Priors

Our prior on branch lengths is *Exponential*, with a mean determined by a hyperparameter ν, itself endowed with an *Exponential* prior of mean 1. Adopting the approach presented by Huelsenbeck et al. (2006), our most general prior on nonsynonymous rate factors of the models is the *Dirichlet process* (DP)—as an infinite mixture across sites—with hyperparameter α, modulating the assumed "graininess" of selection coefficients; α is endowed with an *Exponential* prior of mean 1. The Dirichlet process prior also utilizes a base measure, defining the probability distribution of each component; as in Huelsenbeck et al. (2006), we use $p(\omega) = 1/(1+\omega)^2$, the probability density of the ratio of two identically distributed draws from an *Exponential*. This same base prior is used when dispensing with the DP framework, with the model based on a single global ω factor. All other parameters have flat *Dirichlet* priors on their respective state space.

7.2.4 Model comparisons

We used the model-switch thermodynamic integration framework described in Lartillot and Philippe (2006), and summarized in chapter 2, to evaluate Bayes factors across all codon substitution models described above. Recall that the overall precision of the method depends on a number of factors, such as the step size of the model morphing parameter ($\delta\beta$), and whether the number of cycles between steps is sufficient to allow the chain to re-equilibrate to the intended posterior distribution (see eqn. 2.17), for instance; but also on the inherent distance between the two models being compared. With a large set of candidate models, a reasonable traversal across the space of all models must be designed for efficient computation. In the following sub-sections, we describe a set of model-switch thermodynamic integrations linking together all models under study.

7.2.4.1 GY-MG-switch

The first model-switch scheme links together the GY-F1×4 and the MG-F1×4 models. This particular thermodynamic integration represents the ideal case, where all parameters are involved in both models; parameters are always sampled from the posterior distribution of one model or the other (or the partially morphed posteriors along the path). The GY-MG-switch is also applied to link GY-F3×4 and MG-F3×4 models.

7.2.4.2 F1×4-F3×4-switch

The F1×4-F3×4-switch is only used in the GY context, although it could be used in the MG context as well; here, only one of the two contexts need be calculated to link all models together. For this model-switch, the single nucleotide frequency vector of the GY-F1×4 model is also used as the first codon position nucleotide vector under the GY-F3×4 model. As such, at one end of the path, this set of nucleotide frequencies

corresponds to the single-nucleotide-vector-approximation of codon frequencies under the GY-F1×4 model, whereas at the other end, it corresponds to the first position vector of the three-vector-approximation of codon frequencies under the GY-F3×4 model. As for the other two nucleotide vectors associated with the the GY-F3×4 model, they are effectively sampled from the prior at one end of the path, and the posterior at the other end of the path. All other parameters are relevant to both models.

7.2.4.3 F1×4-F61-switch

This model-switch is only pertinent to the GY context, connecting the GY-F1×4 model and the GY-F61 model. At one end of the path, the vector of nucleotide frequencies used to approximate codon frequencies is sampled from the posterior under the GY-F1×4 model, while sampling from the prior of a (distinct) full 61-dimensional codon frequency vector. At the other end of the path, the vector of nucleotide frequencies used to approximate codon frequencies is sampled from the prior, whereas the 61-dimensional codon frequency vector is sampled from the posterior. All other parameters are relevant to both models.

7.2.4.4 CP-switch and AAP-switch

The CP-switch is only pertinent in the MG context, linking the MG-F1×4 and MG-F1×4-CP models. One end of the path samples the codon preference parameters from the prior, whereas the other samples these parameters from the posterior. All other parameters are relevant to both models. The CP-switch is also used to link the the MG-F3×4 and MG-F3×4-CP models. The AAP-switch is analogous to the CP-switch, but involving the amino acid preference parameters instead.

7.2.4.5 DP-switch

This last model-switch links together a model with a single ω factor and a model based on the Dirichlet process prior modeling heterogeneous ω factors across sites. At one end of the path, the sampler draws from the posterior of a model with a single ω factor, and from the prior (and hyper-prior) of the Dirichlet process. At the other end of the path, sampling is under the full posterior of the Dirichlet process, and the prior of the global ω factor. As before, all other parameters are relevant to both models. The DP-switch scheme is applied separately to each underlying GY and MG-style model.

7.2.4.6 Overall model ranking

From the set of model-switch methods described above, we can evaluate all models by computing Bayes factors with respect to a common reference. We use GY-F1×4 as the reference model here, which implies that as many as four different sets of model-switch schemes may be involved in reporting a particular Bayes factor. For instance, taking the example from the main text, the (log) Bayes factor between MG-F3×4-CP-DP and GY-F1×4 is assembled from four separate calculations:

$$\begin{aligned}\ln \frac{p(D \mid \text{MG-F3}\times\text{4-CP-DP})}{p(D \mid \text{GY-F1}\times\text{4})} &= \ln \frac{p(D \mid \text{MG-F3}\times\text{4-CP-DP})}{p(D \mid \text{MG-F3}\times\text{4-CP})} + \\ &\quad \ln \frac{p(D \mid \text{MG-F3}\times\text{4-CP})}{p(D \mid \text{MG-F3}\times\text{4})} + \\ &\quad \ln \frac{p(D \mid \text{MG-F3}\times\text{4})}{p(D \mid \text{GY-F3}\times\text{4})} + \\ &\quad \ln \frac{p(D \mid \text{GY-F3}\times\text{4})}{p(D \mid \text{GY-F1}\times\text{4})}, \end{aligned} \tag{7.10}$$

where the first term is computed using the DP-switch, the second using the CP-switch, the third using the GY-MG-switch, and the fourth using the F1×4-F3×4-switch. This approach can be viewed as a way of parallelizing the computation of $\ln p(D \mid \text{MG-F3}\times\text{4-CP-DP}) - \ln p(D \mid \text{GY-F1}\times\text{4})$, as opposed to performing a single

long integration directly between the two models. Note, however, that the log Bayes factor for each intermediate model is computed along the way; when building the entire set of Bayes factors against GY-F1×4, this model space traversal will result in much less overall CPU usage than would performing integrations from each model directly to GY-F1×4. The procedure also implies a level of error, which we explore empirically by running each calculation in duplicate, using the quasi-static bi-directional method detailed in Lartillot and Philippe (2006) and discussed below. Each pair of model-switch integrations produces two values, reported as an interval, and giving a sense of the precision of the Monte Carlo settings.

7.3 Results and discussion

7.3.1 Empirical explorations of thermodynamic integrations

We performed several pilot runs to tune each type of model-switch thermodynamic integration. Incorporating the bi-directional approach described in Lartillot and Philippe (2006), each model-switch scheme was explored by running integrations in duplicates, one with the morphing parameter β going from 0 to 1, and another with β going from 1 to 0. We report both values obtained from the bi-directional approach as an interval throughout. Figures 7.1 and 7.2 display examples of this tuning process in two cases. Each panel in these figures plots the values $\ln p(D \mid \theta, M_1) - \ln p(D \mid \theta, M_0)$ collected during bi-directional quasi-static runs. Graphically, the log Bayes factor corresponds to the area between the curve and the abscissa (negative below the abscissa, and positive above it), and is estimated using the relation given in (2.21).

Using the GLOBIN*17-144* data set, figure 7.1 corresponds to a case that we qualify as computationally easy: the GY-MG-switch, linking GY-F1×4 and MG-F1×4. These two models have the exact same parameters, and only differ in how parameters are

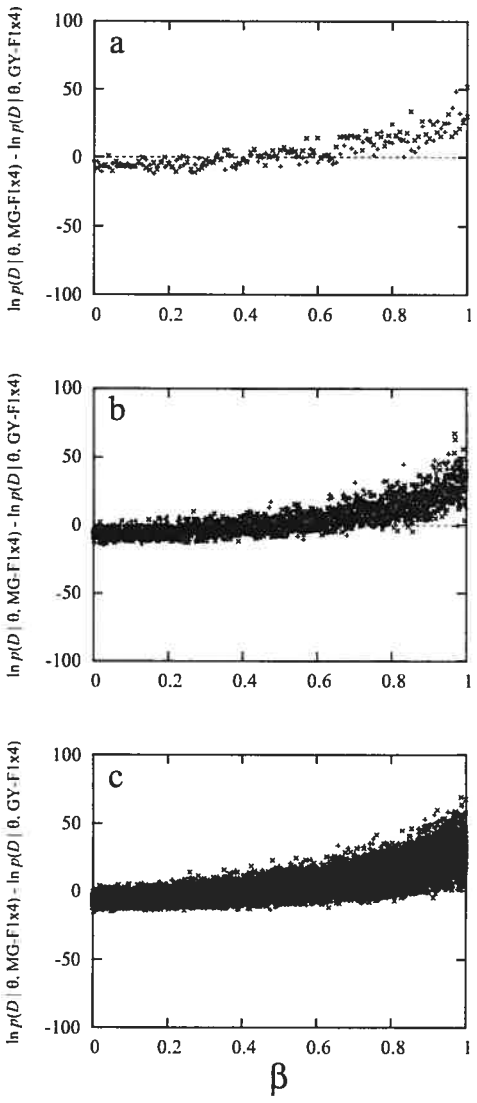

Figure 7.1. Log-likelihood differences recorded during GY-MG-switch thermodynamic integrations linking GY-F1×4 and MG-F1×4. Two integrations are plotted in each panel, one with β going from 0 to 1 (+), and another with β going from 1 to 0 (×). The collection of $K+1$ values is used to approximate the log Bayes factor according to (2.21). Panel a) displays "fast" runs, with $K = 100$, panel b) displays "medium" runs, $K = 1,000$, and panel c) displays "slow" runs, with $K = 10,000$.

assembled to specify the final model. At one end of the path ($\beta \sim 0$), the plot displays the difference in log-likelihood between MG-F1×4 and GY-F1×4, when the parameters from the posterior under GY-F1×4 are "imposed upon" the MG-F1×4 model. Reciprocally, at the other end of the path ($\beta \sim 1$), the plot displays the difference in log-likelihood when the parameters of the posterior under MG-F1×4 are "imposed upon" GY-F1×4 model. Based on the $K + 1$ draws along the path, the approximation given in (2.21) for $K = 100$ (fig. 7.1a), $K = 1,000$ (fig. 7.1b), and $K = 10,000$ (fig. 7.1c), is [2.9 ; 5.4], [3.7 ; 4.1] and [3.8 ; 3.9] respectively. These two models are quite close to each other, in terms of overall fit, but the model-switch integration procedure nonetheless allows for a very precise estimation in this case, because the models can be connected through a very short overall path. In this case, the final runs ($K = 10,000$) each required about 6 days of CPU time on an Intel P4 3.2 GHz computer node.

Still using the GLOBIN*17-144* data set, figure 7.2 corresponds to a case that we qualify as computationally challenging: the F1×4-F61-switch, linking GY-F1×4 and GY-F61. In contrast with the GY-MG-switch, in which all parameters were involved in both models, this thermodynamic integration has a set of parameters in each model that are irrelevant to the other. When $\beta \sim 0$, the plots display the difference in log-likelihood between GY-F61 and GY-F1×4 when the 61-dimensional vector of codon frequencies attributed to GY-F61 is sampled from the prior, and other parameters are those "imposed by" the posterior under GY-F1×4. Such a sampler will induce very poor log-likelihood values under GY-F61, and indeed the plots display negative values at this end of the path. At the other end of the path ($\beta \sim 1$), the plots display the difference in log-likelihood between GY-F61 and GY-F1×4 when the 61-dimensional vector of codon frequencies is sampled from the posterior under GY-F61, the single-nucleotide-vector-approximation of codon frequencies is sampled from the prior under GY-F1×4, but with other parameters being those "imposed" by the posterior under GY-

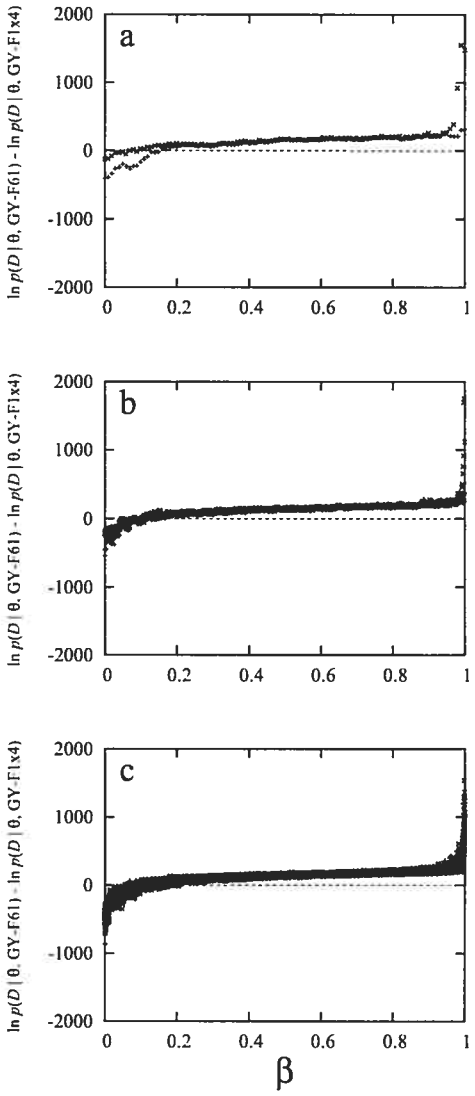

Figure 7.2. Log-likelihood differences recorded during F1×4-F61-switch thermodynamic integrations linking GY-F1×4 and GY-F61. Two integrations are plotted in each panel, one with β going from 0 to 1 (+), and another with β going from 1 to 0 ($\times$). The collection of $K+1$ values is used to approximate the log Bayes factor according to (2.21). Panel a) displays "fast" runs, with $K = 100$, panel b) displays "medium" runs, $K = 1,000$, and panel c) displays "slow" runs, with $K = 10,000$.

F61. Thus, this other end of the path will induce very poor log-likelihood values under GY-F1×4, and indeed the log-likelihood difference displayed in the plots become highly positive. As can be appreciated graphically, the tail-ends of the integrand represent the main source of error in this model switch. The interval obtained from bi-directional quasi-static runs with $K = 100$ (fig. 7.2a) is extremely broad, at [84.9 ; 175.6]. Also note that in this case the sampler does not appear to include a sufficient number of cycles between steps to decorrelate successive draws. Several tuning options could be explored, but here, we simply increase the overall sample size (or equivalently, decrease the step size $\delta\beta$). With $K = 1,000$ (fig. 7.2b) the interval obtained is [113.4 ; 126.2], and finally, the longest runs ($K = 10,000$, fig. 7.2c), each requiring about 20 days of CPU time, produce the tightest interval, at [115.6 ; 117.6].

However, when computing log Bayes factors for the more complex models, involving several distinct model-switch schemes, the interval of the overall log Bayes factor against GY-F1×4 is constructed conservatively (to produce the broadest possible interval), and in some cases an entirely unambiguous model ranking is not possible. For instance, with the GLOBIN*17-144* data, the log Bayes factor of MG-F3×4-CP-DP against GY-F1×4, and the log Bayes factor of MG-F1×4-CP-DP against GY-F1×4, overlap with each other (table 7.1), which thus prevents us from clearly distinguishing the two models. Similarly, for the LYSIN*25-134* data set, four models are ambiguously top ranking, as are three for the HIV*22-99* data set (see bold emphasis in table 7.1). In the present context, obtaining the required level of precision for distinguishing between log-marginal-likelihoods that differ by a few units is relatively uninteresting, and not worth the computational investment that would be needed when utilizing the present methods. Our objective here is rather to map out the main effects of different formulations in terms of overall model fit.

7.3.2 Bayes factors

The series of log Bayes factors reported in table 7.1 reveals considerable differences in model fit, indicating the importance of performing a careful examination of alternative parametric choices. We note that the MG-F1×4-CP-DP model is among the top ranking models for all three data sets. This result is somewhat expected. First, nonsynonymous rate heterogeneity has now been observed across numerous data sets (Yang, 2006), and it thus seems reasonable to anticipate an improved model fit under the Dirichlet process (DP) framework proposed by Huelsenbeck et al. (2006). The other specifications of this top model are also reassuring, in the sense that adhering closely to the mechanistic perspective of teasing apart mutational features and selective constraints produces, at worst, a model of roughly equivalent fit to models lacking such a natural interpretation. In addition, all three data sets suggest uneven codon preferences (CP), although such preferences appear to go well beyond amino acid preferences (AAP) only in the case of the GLOBIN*17-144*.

We next note that under the simpler settings of MG-style models, suppressing AAP or CP parameters, the F3×4 configuration is generally preferred over the F1×4 configuration for all three data sets. The periodic pattern of codon-position-specific nucleotide propensities is a feature expected from the structure of the genetic code. Such an interpretation, however, is not accurately represented by expanding the nucleotide level parameterization. Indeed, with the richer models, including the CP parameters in particular, the F3×4 configurations are only mildly preferred over the F1×4 configuration, and when invoking the Dirichlet process, modeling heterogeneous nonsynonymous rates, the numerical error no longer allows for a clear distinction between these two configurations (expect for the HIV*22-99* data set, which gives preference to the F1×4 configuration).

The GY-style of models based on the F1×4 and F3×4 configurations are generally

Table 7.1. Natural logarithm of the Bayes factor for models considered, with GY-F1×4 used as a reference.

Model	GLOBIN*17-144*	LYSIN*25-134*	HIV*22-99*
GY-F1×4	-	-	-
GY-F3×4	[69.4 ; 70.3]	[-4.7 ; -4.2]	[11.7 ; 12.0]
GY-F61	[115.6 ; 117.6]	[28.9 ; 31.4]	[24.9 ; 26.2]
MG-F1×4	[3.8 ; 3.9]	[3.0 ; 3.2]	[11.7 ; 11.8]
MG-F3×4	[45.8 ; 47.0]	[3.6 ; 4.4]	[17.9 ; 18.3]
MG-F1×4-AAP	[42.0 ; 43.8]	[46.3 ; 47.7]	[24.5 ; 25.4]
MG-F3×4-AAP	[83.3 ; 85.4]	[50.9 ; 53.1]	[20.6 ; 22.2]
MG-F1×4-CP	[125.9 ; 127.7]	[65.9 ; 68.4]	[26.4 ; 28.0]
MG-F3×4-CP	[128.1 ; 130.7]	[69.6 ; 73.3]	[22.3 ; 23.9]
GY-F1×4-DP	[102.3 ; 104.2]	[183.7 ; 185.9]	[54.7 ; 55.1]
GY-F3×4-DP	[166.7 ; 169.5]	[176.6 ; 179.8]	[65.5 ; 66.8]
GY-F61-DP	[218.5 ; 222.0]	[213.8 ; 219.1]	**[76.8 ; 78.3]**
MG-F1×4-DP	[106.0 ; 108.1]	[187.1 ; 190.0]	[69.0 ; 70.3]
MG-F3×4-DP	[148.6 ; 152.3]	[186.7 ; 189.8]	[74.3 ; 76.0]
MG-F1×4-AAP-DP	[166.0 ; 170.2]	**[240.0 ; 245.4]**	**[77.4 ; 79.3]**
MG-F3×4-AAP-DP	[206.8 ; 211.5]	**[240.0 ; 245.9]**	[74.5 ; 76.5]
MG-F1×4-CP-DP	**[240.3 ; 244.9]**	**[240.6 ; 246.9]**	**[78.0 ; 80.0]**
MG-F3×4-CP-DP	**[237.0 ; 242.7]**	**[240.7 ; 248.1]**	[74.6 ; 76.7]

Note.—Values given are the upper and lower estimates obtained from bi-directional thermodynamic integrations. Top models are emphasized in bold.

disfavored over their MG-style counterparts (except for the GLOBIN*17-144* data set, which gives favor to GY-F3×4 over MG-F3×4). Surprisingly, for the LYSIN*25-134* data set, the simpler GY-F1×4 model is slightly preferred over the GY-F3×4 model. However, for all three data sets, the GY model based on F61 configurations outperforms the other GY-style models, as well as the simpler MG-style models. In the case of the GLOBIN*17-144* data set, the contrast of the F61 configuration is even greater than that observed between homogeneous and heterogeneous models of nonsynonymous rates; for instance, the log Bayes factor of GY-F61 against GY-F1×4 is [115.8 ; 117.4] whereas for GY-F1×4-DP against GY-F1×4 is [102.3 ; 104.2]. These results for GY-F61 model are also indicative of uneven codon preferences. However, as previously mentioned, the codon preferences accounted for in this GY formulation are confounded with other features, including the background of nucleotide propensities, making the model less attractive on interpretive grounds. Accordingly, when contrasted with the richer MG formulations accounting for codon (or amino acid) preferences, the GY-F61 model is less attractive on quantitative grounds (except for HIV*22-99*, in which case it matches the top MG-style models).

7.3.3 Posterior distributions

Here, we display posterior distributions (obtained using plain MCMC sampling) for parameters of the MG-F1×4-CP-DP model. Our main focus is on the distinguishing features of the model, namely, the combination of background nucleotide propensities with global codon preference parameters. To illustrate certain features, we also contrast the distributions with those obtained under the MG-F3×4-CP-DP model, as well as under the simpler models suppressing CP parameters.

The results of table 7.1 suggest that disparities in nucleotide propensities at the first, second, and third positions could be reduced to codon (or amino acid) preferences. To

investigate this point, we first inspect the posterior distributions of nucleotide propensity parameters under various model configurations. Figure 7.3 displays the 95% credibility intervals of the global nucleotide propensity parameters for each data set. The full lines correspond to the interval obtained under MG-F1×4-DP, whereas the dashed lines are obtained under MG-F1×4-CP-DP. The distributions are far more diffuse under the CP version, although their general locations appear similar. Inversely, one can interpret that without the CP parameters, the posterior nucleotide propensity distributions are misleadingly overconfident. Figure 7.4 explores this same behavior under the F3×4 configurations. First note that without the CP parameters (full lines), the three positions show striking differences in overall distributions, and that the magnitude of the credibility intervals are much greater than under the F1×4 configuration. When the CP parameters are introduced (dashed lines), several credibility intervals considerably shift and increase in magnitude. Also note that under the CP settings, the distributions of each position tend to overlap. To show this more vividly, we reconstituted the results displayed in figures 7.3 and 7.4 into a single figure for the GLOBIN*17-144* data (fig. 7.5). Figure 7.5a displays the global nucleotide propensity parameter values obtained under the the MG-F1×4-DP model (full line) as well as each of the three nucleotide propensity parameter values under the MG-F3×4-DP (in progressively finer dashed lines for position 1, 2, and 3). In this case, the disparity between the different distributions is high. When including the CP parameters (fig. 7.5b), however, the disparity is much lower, suggesting the redundancy of the F3×4 configuration in combination with the CP parameters. We note that some values are still markedly divergent (e.g., 3rd position A and 2nd position T), indicating that other model violations may be at play. In other words, codon and/or amino acid preferences seem to explain (albeit not entirely) the observed disparities of nucleotide equilibrium frequencies at the three codon positions.

We next inspect the posterior distributions of codon preference parameters. Figure

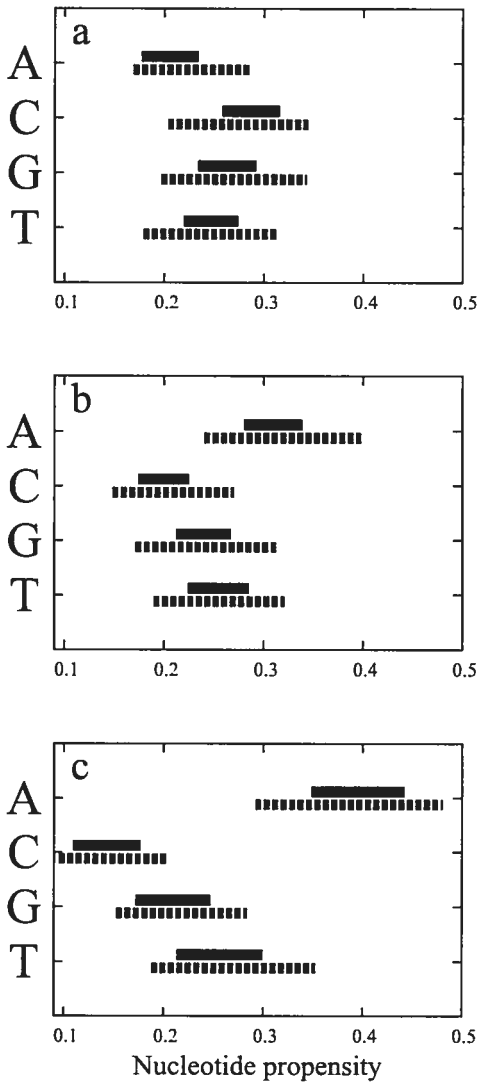

Figure 7.3. 95% credibility intervals of global nucleotide propensity parameters obtained under MG-F1×4-DP (full lines) and under MG-F1×4-CP-DP (dashed lines). The top panel (a) refers to the GLOBIN *17-144* data set, followed by LYSIN *25-134* (b),and HIV *22-99* (c).

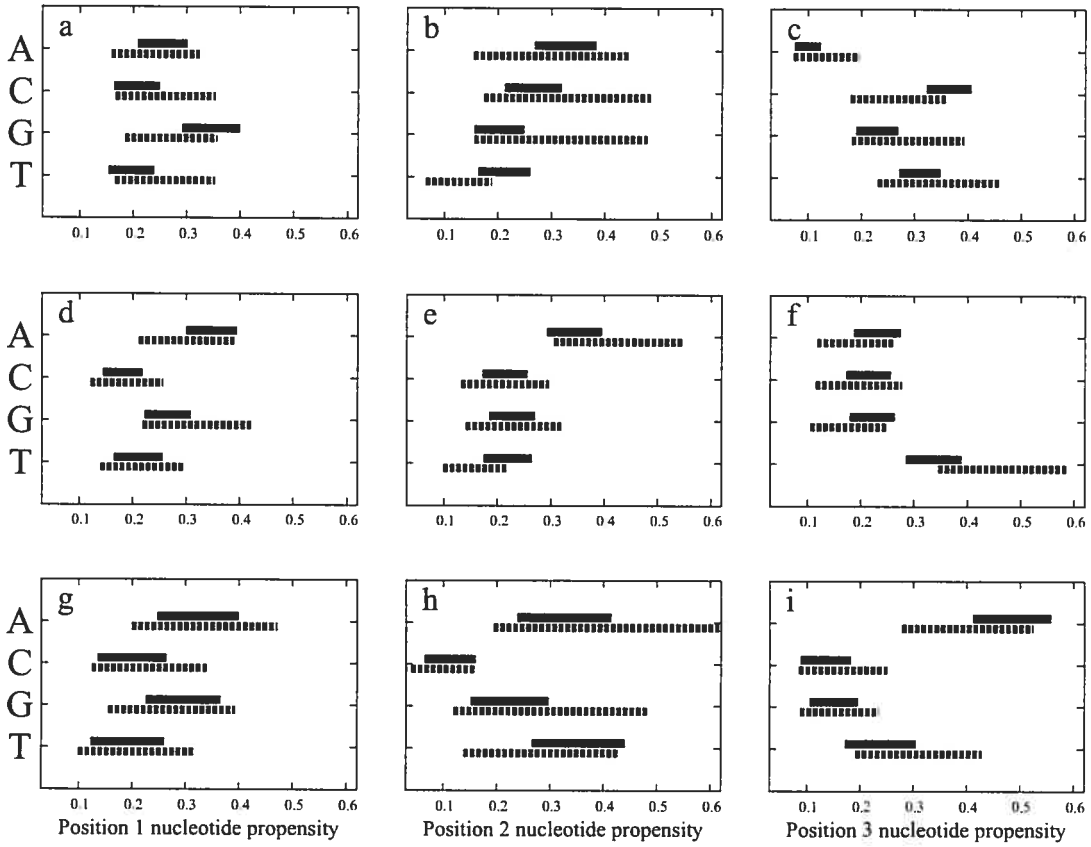

Figure 7.4. 95% credibility intervals of position-specific nucleotide propensity parameters obtained under MG-F3×4-DP (full lines) and under MG-F3×4-CP-DP (dashed lines). The three panels (a, b, c) refer to the GLOBIN*17-144* data set, followed by LYSIN*25-134* (d, e, f), and HIV*22-99* (g, h, i).

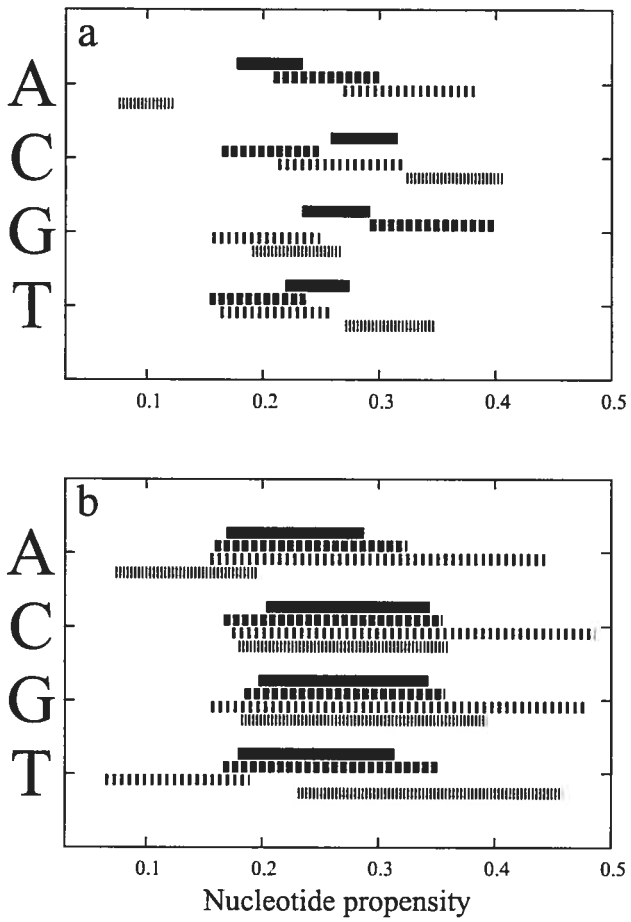

Figure 7.5. A composite from figures 7.3 and 7.4 for the GLOBIN*17-144* data set. Panel a displays the 95% credibility intervals of global nucleotide propensity parameters under the MG-F1×4-DP model (full line) as well as the 95% credibility of the three nucleotide propensity parameters under the MG-F3×4-DP (with progressively finely-dashed lines for position 1, 2, and 3 respectively). Panel b displays the 95% credibility interval for same parameters but now, under the MG-F1×4-CP-DP and MG-F3×4-CP-DP models.

7.6 displays the 95% credibitility intervals of codon preference parameters under the F1×4 (full line) and F3×4 (dashed line) configurations. The parameters appear to be only moderately sensitive to the F1×4/F3×4 choice, although a few notable shifts and increases in magnitude of credibility interval are observed. This suggests that the although the F3×4 configuration is impertinent with the CP parameters, it is not too costly in terms over-parameterization, which corroborates with the results from table 7.1. The overall CP distributions suggest pronounced overall codon preferences for the GLOBIN*17-144*, but milder preferences for LYSIN*25-134* and HIV*22-99*. This also corroborates well with our computed Bayes factors, which indicate that for LYSIN*25-134* and HIV*22-99*, the improvement brought about by the CP parameters is less important than for the GLOBIN*17-144* data. Observing the distributions for the GLOBIN*17-144* data set in detail, we find that that the parameter values appear to capture long observed tendencies of codon preferences on similar data, such as the elevated use of CTG for encoding leucine, GTG for valine, or GGC for glycine; indeed, these were some of the first observations stimulating research into the causes of codon preferences (e.g., Fitch, 1980; Modiano et al., 1981; Kimura, 1983).

7.3.4 Detection of positive selection

Finally, we contrasted the conclusions of the GY-F61-DP, MG-F1×4-DP and MG-F1×4-CP-DP models with regards to the amino acid positions inferred to have undergone positive selection. Under the DP settings, the posterior probability of a site being under positive selection can be computed from the proportion of draws from a sample (obtained via plain MCMC sampling) found to be in a class $\omega > 1$, as described in Huelsenbeck et al. (2006). We first note that for the GLOBIN*17-144*, focusing on posterior probabilities at 0.9, 0.95, and 0.99 cutoff levels, the MG-F1×4-DP and MG-F1×4-CP-DP models infer sites under positive selection at each level, whereas the

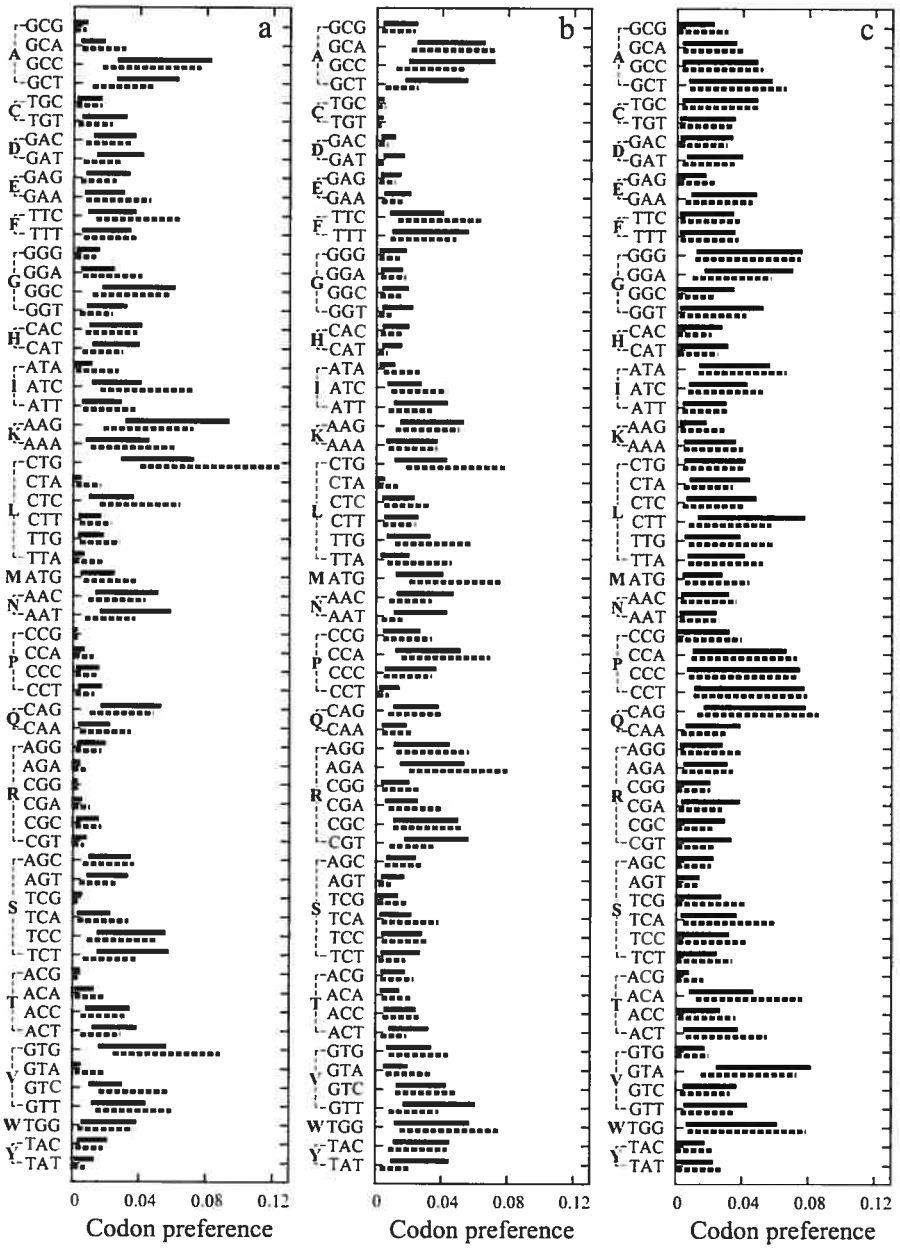

Figure 7.6. 95% credibility intervals of codon preference parameters, sorted according to amino acids. The full lines are values under MG-F1×4-CP-DP, whereas the dashed lines are values under MG-F3×4-CP-DP. The leftmost panel (a) refers to the GLOBIN*17-144* data set, followed by LYSIN*25-134* (b).and HIV*22-99* (c).

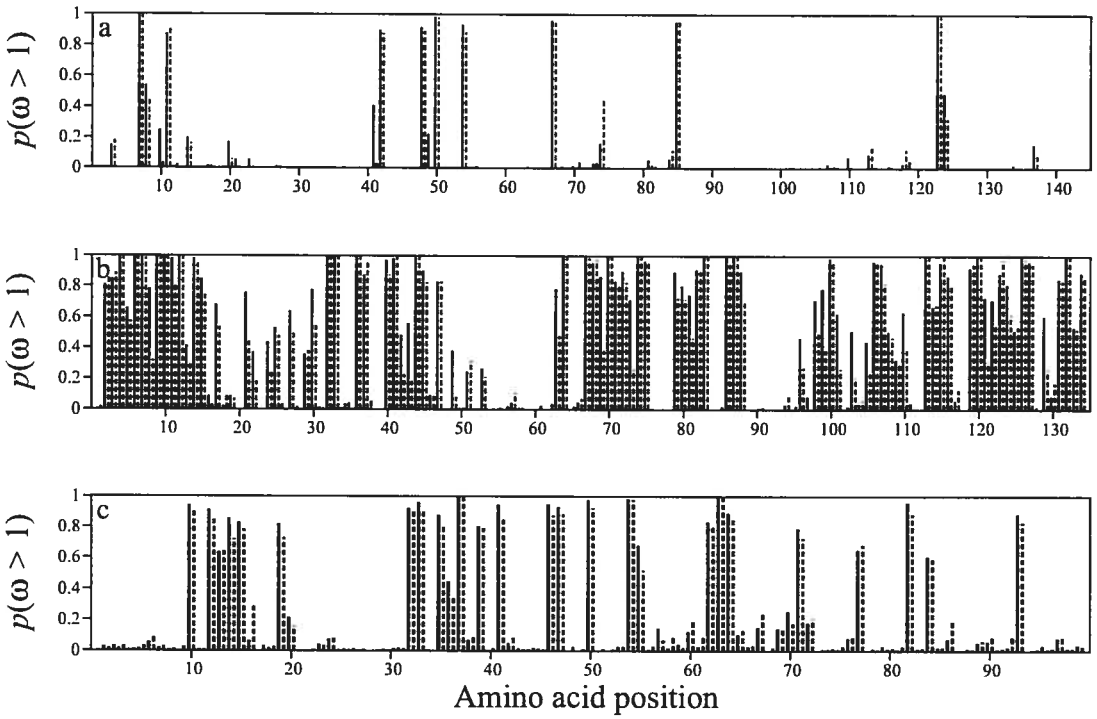

Figure 7.7. Posterior probability of each site having $p(\omega > 1)$ under MG-F1×4-DP (full lines), and MG-F1×4-CP-DP (dashed lines). The top panel (a) refers to the GLOBIN*17-144* data set, followed by LYSIN*25-134* (b), and HIV*22-99* (c).

GY-F61-DP model infers no sites at either level (table 7.2). The list of sites under positive selection under the three models considered also differs for the other two data sets (table 7.2). Comparing the MG-F1×4-DP and MG-F1×4-CP-DP models specifically, figure 7.7 displays the values $p(\omega > 1)$ across all sites. Overall, the spike patterns have the same general aspect, although the CP parameters appear to attenuate the $p(\omega > 1)$ values. There are a number of exceptions, however, and it will be important to conduct a broader empirical study of the impacts of these and other parametric choices on the detection of positive selection.

Table 7.2. Amino acid sites under positive selection.

Data	Model	Sites
	GY-F61-DP	-
GLOBIN*17-144*	MG-F1×4-DP	**7**, *48*, 50, *54*, 67, 85, **123**,
	MG-F1×4-CP-DP	**7**, *11*, 50, *67*, 85, 123
	GY-F61-DP	*2*, *3*, **4**, 6, **7**, **9**, **10**, 11, **12**, 14, **32**, **33**, **36**, *37*, **41**, **44**, 64, **67**, **68**, **70**, **74**, **83**, 86, **87**, 100, *106*, *107*, 113, **115**, *116*, **120**, *123*, **126**, **132**
LYSIN*25-134*	MG-F1×4-DP	**4**, 6, **7**, **9**, **10**, 11, **12**, 14, **32**, **33**, **36**, 40, 41, **44**, *45*,, 64, **67**, 68, **70**, **74**, 75, *82*, **83**, **86**, **87**, 100, 106, *107*, **113**, 115, 119, **120**, **126**, 127, **132**
	MG-F1×4-CP-DP	**4**, **6**, 7, **9**, **10**, **12**, *14*, **32**, **33**, **36**, 37, **41**, **44**, **64**, 67, 68, **70**, **74**, 75, **83**, *86*, **87**, 100, 106, **113**, **115**, 119, **120**, 123, 126, *127*, **132**
	GY-F61-DP	54, **37**, **63**
HIV*22-99*	MG-F1×4-DP	*10*, *12*, *32*, 33, **37**, 41, 46, *47*, 50, 54, **63**, 82
	MG-F1×4-CP-DP	*10*, *32*, *33*, **37**, *50*, 54, **63**

Note.—Numbers in *italic* font are at the 0.9 level, in plain font at the 0.95 level and in **bold** font at 0.99 level.

7.4 Conclusions

The distinction between GY and MG-style models has generally been considered as relatively subtle, and most researchers have chosen to explore modeling extensions from one of the two perspectives (e.g., Nielsen and Yang, 1998; Yang et al., 2000a; Sainudiin et al., 2005; Huelsenbeck et al., 2006; Wong et al., 2006; Kosakovsky Pond and Muse, 2005; Schaldt and Lange, 2002). In light of all of these recent developments, the study in this chapter effectively takes a step back, to re-assess the core motivations underlying codon-based models: the formulation of a biologically meaningful and readily interpretable parameterization. We have argued that the MG-style models, with the extensions studied here, subscribe most closely to these motivations. From the Bayesian standpoint, sorting the importance of different model formulations becomes an empirical issue, explored here by evaluating Bayes factors. Results confirm that a careful modeling in the MG-style, so as to acknowledge amino acid or codon preferences, tends to surpass, or at least match, the optimal GY-style model. Furthermore, the top GY and MG-style models reach different conclusions with regards to amino acid sites under positive selection, with the top-performing GY-style model comparatively over-estimating selective factors bearing on nonsynonymous substitution rates. We recommend future modeling investigations to consider incorporating any extensions in the MG context specifically, and to monitor how these alternative choices compare, in terms of model fit, but also in terms of logical interpretation (Thorne, 2007).

Chapter 8

Evaluating structural models of codon substitution

8.1 Introduction

Having now studied codon substitution models assuming independence, we are in a position to re-introduce the statistical potential into a set of different model formulations. Recall that the practical complications of the model presented by Robinson et al. (2003) led these authors to propose a set of MCMC techniques based on two different forms of auxiliary variable methods: 1) a data-augmentation system, providing a numerical means of integrating over detailed substitution mappings; and 2) an importance sampling argument, providing an approximation of the ratio of two intractable normalizing constants. Together, these approaches provided the first proof-of-concept that such models could be implemented.

In this chapter, we revise both forms of MCMC schemes for the study of site-interdependent models in the codon context, and suggest the use of flexible sampling approaches that can be more readily expanded to accommodate richer statistical potentials, as well as higher dimensional parameterizations bearing on the stationary dis-

tribution of the site-interdependent Markov process. Specifically, we first describe a method for producing data-augmentations under a proposal density designed to be as close as possible to the target site-interdependent density. The procedure is based on a definition of site-specific codon substitution matrices, and utilizes a *uniformization* technique previously used by Fearnhead and Sherlock (2006) to explore the occurrence of rare DNA motifs. Next, we adapt recent techniques derived for approximating posterior distributions involving intractable normalizing factors in the likelihood function (Murray et al., 2006). Our focus is on embedding these different techniques within thermodynamic integration methods, as described in chapter 4, to evaluate Bayes factors for different codon model versions, and to present preliminary analyses in this context. We also present preliminary posterior predictive checks, displaying how different models render features of nonsynonymous rate heterogeneity, and amino acid exchanges. Altogether, the methods proposed here amount to setting up another phenomenological benchmarking, now at the codon level of interpretation.

8.2 Material and methods

8.2.1 Data

We used the GLOBIN*17-144* data set, with PDB code 4HHB chain B used as a reference structure.

8.2.2 Evolutionary models

We again borrow the nomenclature of Parisi and Echave (2001), and refer to the models as *structurally constrained* (SC), utilizing the combined contact and solvent accessibility potential developed in chapter 5. Recall the form of the potential, with the pseudo-

energy score of as sequence s given by:

$$G(s) = \sum_{1 \le i < i' \le N} \Delta_{ii'} \epsilon_{s_i s_{i'}} + \sum_{1 \le i \le N} \Xi_{s_i}^{v_i} + \sum_{1 \le i \le N} \Sigma_{s_i}. \tag{8.1}$$

The first term in (8.1) is a contact potential, the second is a solvent accessibility potential, and the last term accounts for compositional effects, inspired from the random energy approximation (Shakhnovich and Gutin, 1993; Sun et al., 1995; Seno et al., 1998).

Let $G_i(a)$ represent the pseudo-energy associated with observing amino acid a at site i, but without consideration of the contact component; with the present form of potential, $G_i(a) = \Xi_a^{v_i} + \Sigma_a$. In the most general case studied in this chapter, we begin by constructing site-specific codon substitution matrices of the form

$$Q_{ab}^{(i)} = \begin{cases} \varrho_{a_c b_c} \varphi_{b_c} \left(\frac{\psi_b}{\psi_a}\right)^{\frac{1}{2}}, & \text{if } \mathcal{A}, \\ \omega \varrho_{a_c b_c} \varphi_{b_c} \left(\frac{\psi_b}{\psi_a}\right)^{\frac{1}{2}} e^{\beta(G_i(a) - G_i(b))}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise.} \end{cases} \tag{8.2}$$

Note that as written above, the model would imply 14 different Q matrices, for the 14 solvent accessibility classes as derived in chapter 5. However, when invoking the Dirichlet process prior on ω, it is more practical to assemble site-specific matrices as above, but with the ω factor coming from the current "pool" of ω factors, according to the configuration of the Dirichlet process.

Now let $G_\Delta(s)$ be the contact energy of sequence s, i.e., $G_\Delta(s) = \sum_{1 \le i < i' \le N} \Delta_{ii'} \epsilon_{s_i s_{i'}}$. Then, we construct the overall sequence process by specifying off-diagonal entries of the

Markov generator as

$$R_{ss'} = \begin{cases} Q^{(i)}_{s_i s'_i} e^{\beta(G_\Delta(s) - G_\Delta(s'))}, & \text{if } s \text{ and } s' \text{ differ by one codon at position } i, \\ 0, & \text{otherwise,} \end{cases} \tag{8.3}$$

and with diagonal entries given by the negative sum of off-diagonal entries. We can see that from this construction, nonsynonymous rates will be proportional to $e^{\beta(G(s)-G(s'))}$, i.e., the measure based on the overall statistical potential. Here, β can be treated as a free parameter, with a uniform prior on $[-5, 5]$, which we again indicate as $+\beta$. However, we may also fix $\beta = 1/2$, given that the potential was originally derived with this scaling. Fixing $\beta = 0$ recovers the MG-F1×4-CP-DP model studied in the previous chapter—we use the same prior structure on all other parameters as we did in the last chapter.

An attractive aspect of the model is that it acknowledges that the evolutionary process producing the different protein-coding nucleotide sequences involves several distinct features, specifically bearing on mutational tendencies at the nucleotide level, on global codon preferences, on nonsynonymous heterogeneity, as well as constraints operating at the level of the overall amino acid sequence; and we can explore the relative importance of these features by measuring the fit of specials cases, which can be recovered by suppressing (or prefixing) certain parameters.

8.2.3 Data augmentation

Although we still utilize the BRANCHHISTORY, NODESTATE, and TREEHISTORY operators in our data augmentation-based sampler, we use a different scheme for generating proposal mappings in this chapter. The mappings are proposed from the site-specific Q matrices, which include all aspects of the model, but not the contact component of the potential. The mappings are then accepted or rejected, according to the full

site-interdependent MH rule. The idea here is that as the potentials used become more sophisticated, it will become increasingly important to generate mappings from a model that is as "close" as possible to the target model, in order to have high acceptance rates and good mixing kinetics.

Thus far, we have relied on the method proposed by Nielsen (2002) to generate mappings under site-independent models. Recall that the algorithm proceeds in two basic steps: 1) sample internal node states from their joint distribution, conditional on the data and the parameters of the Markov process; and 2) sample the series of substitution events along each branch, conditional on parameters of the Markov process, and the states at both ends such as determined in step 1). The second step of the algorithm is an *accept/reject* approach: starting from the state at the ancestral node, run the Markov process—sampling the timing and nature of events to the end of the branch—and accept the resulting substitution mapping if the last event is consistent with the state of the descendant node; if not, reject the mapping and start over, until a consistent mapping is drawn. "The simulation scheme is efficient assuming that the rates of change between all nucleotides [states] is large [...]," (Nielsen, 2002, p. 732) and a provision is made to enforce the sampling of at least one event in cases where the ancestral and descendant states differ. However, the site-specific codon substitution models of interest here have instantaneous rates of change of 0 between states differing by 2 or 3 nucleotides. This has the effect of inducing stricter constraints on the possible coherent mappings, and under some conditions, Nielsen's second step stalls, entering a prolonged *while*-loop in attempting to sample an acceptable mapping.

Instead, in such problematic cases, we use a procedure is based on a *uniformization* technique. The uniformization procedure (see, e.g., Jensen, 1953; Gross and Miller, 1984; Mateiu and Rannala, 2006; Lartillot, 2006; Fearnhead and Sherlock, 2006) transforms the process defined by Q (we shall omit the site index i in the developments that

follow) into a process allowing for *virtual events*, or *self-substitutions* (from a to a). Let $P = [P_{ab}]$ be the matrix of this process, obtained from

$$P = \frac{1}{\mu}Q + I, \tag{8.4}$$

where $\mu \geq max\{-Q_{aa}\}$ is the *uniformization rate*, and I is the identity matrix. Note here that the sum of each row in R equals 1; this is also called a *stochastic matrix*. Under the uniformized process, the waiting time until an event no longer depends on the current state, and the probability of having n events (including self-substitutions) over a branch length λ (we drop the j index below) is given by a Poisson distribution:

$$p(n \mid \lambda) = e^{-\mu\lambda}\frac{(\mu\lambda)^n}{n!}. \tag{8.5}$$

Also, taking powers of the matrix P yields the probability of starting in state a and ending at state b after n events:

$$p(b \mid n, a) = P^n_{ab}. \tag{8.6}$$

We will suppose that we now want to draw a mapping along a branch, and that the states at the ends of the branch (a and b, for the beginning and ending states respectively) have already been sampled (using Nielsen's first step). The overall method can be summarized as a three stage progressive demarginalization: 1) sample the number of events (always including virtual events) marginalized over their nature and timing; 2) sample the nature of events in order, marginalized over their exact timing; and 3) sample the timing of events.

We first begin by drawing the number of events from the distribution $p(n \mid a, b, \lambda)$,

which can be calculated from (8.5) and (8.6) according to Bayes' theorem:

$$p(n \mid a, b, \lambda) = \frac{p(b \mid n, a)p(n \mid \lambda)}{p(b \mid a, \lambda)}. \tag{8.7}$$

Note that the denominator in (8.7) can be developed as follows:

$$\begin{aligned}
p(b \mid a, \lambda) &= \sum_{n=0}^{\infty} p(b \mid n, a)p(n \mid \lambda) && (8.8)\\
&= \sum_{n=0}^{\infty} P_{ab}^{n} \frac{(\mu\lambda)^n e^{-\mu\lambda}}{n!} && (8.9)\\
&= \left[e^{-\mu\lambda} \frac{\sum_{n=0}^{\infty} (\mu\lambda P)^n}{n!} \right]_{ab} && (8.10)\\
&= \left[e^{-\mu\lambda I} e^{\mu\lambda P} \right]_{ab} && (8.11)\\
&= \left[e^{\mu\lambda (P-I)} \right]_{ab} && (8.12)\\
&= \left[e^{\lambda Q} \right]_{ab}. && (8.13)
\end{aligned}$$

This development highlights the fact that Q and P are different representation of the same underlying process, and hence both representations may be exploited in the overall sampling scheme. In particular, the form in (8.13) can be calculated employing a matrix diagonalization routine for matrix exponentiation, and thus used to draw the number of events: first sample $g = p(b \mid a, \lambda) \times U$, where U is a random number on the unit interval; and next, starting from $n = 0$, cumulate $p(b \mid n, a)p(n \mid \lambda)$ over successive values of n, until surpassing g; the final n is thus the number of events sampled.

Having sampled the number of events n, we now wish to sample the specific series of events leading from a to b. The state after the first event (s_1) is sampled from

$$s_1 \sim p(s_1 = l \mid s_0 = a, s_n = b) \propto P_{al} P_{lb}^{n-1}. \tag{8.14}$$

Then, having sampled the state after first event, the state after the second event (s_2)

is sampled from

$$s_2 \sim p(s_2 = m \mid s_1 = l, s_n = b) \propto P_{lm} P_{mb}^{n-2}, \tag{8.15}$$

and so on, until the n events have been sampled. Note that the second factor on the right hand side of (8.14) and (8.15) ensures that the state sampled will not "trap" the mapping into a state s_k which could not lead to $s_n = b$ in $n - k$ events. For instance, if $n = 3$, $s_2 = m$, and m differs with b by two nucleotides, then $P_{mb}^{n-2} = 0$, and thus this particular state m could not have been sampled in (8.15).

Finally, we can draw n values uniformly distributed on $[0, \lambda]$, and sort the values to obtain the timing of events. Virtual events can then be removed so as to obtain a substitution history directly sampled from the posterior distribution under the site-specific model, which constitutes the proposed mapping for that site.

Note that the uniformization technique for sampling mappings is computationally demanding. Calculation of the successive powers of stochastic matrices is the rate-limiting step of the overall operation. Always setting $\mu = max\{-Q_{aa}\}$ as the uniformization rate, we sometimes observed cases with up to 100 virtual events, without any bona-fide events, which nonetheless implies as many powers of the stochastic matrix. As such, we have set our sampler to only use the uniformization technique when the states at both ends of a branch differ by 2 or 3 nucleotides, and to use Nielsen's method when the states are identical, or differ by only one nucleotide[1].

Altogether, this scheme enables us to propose mappings for any number of sites, from a proposal density that only differs with the target density by the contact component of potential.

[1]We have never observed Nielsen's method to stall in such conditions.

8.2.4 Updating model parameters

The same types of update operators as used previously can be applied in the present context to approximate the posterior distribution, based on the site-interdependent MH rule. However, as in previous chapters, for parameters bearing on the stationary distribution of the substitution process, the ratio of two intractable normalizing factors appears in the MH ratio, requiring a more elaborate approach. The importance sampling method proposed by Robinson et al. (2003) for approximating this ratio would involve an extensive design and tuning phase in the high-dimensional context of interest here. Instead, we used the *single variable exchange* method recently proposed by Murray et al. (2006), as we describe below.

The stationary distribution of full site-interdependent codon model given above reads as

$$p(s_0 \mid \theta, M) = \frac{1}{Z_\theta} e^{-2\beta G(s_0)} \prod_{i=1}^{N} \left(\psi_{s_{i0}} \prod_{c=1}^{3} \varphi_{s_{i0_c}} \right), \tag{8.16}$$

where Z_θ is the normalizing factor:

$$Z_\theta = \sum_{s} e^{-2\beta G(s)} \prod_{i=1}^{N} \left(\psi_{s_i} \prod_{c=1}^{3} \varphi_{s_{i_c}} \right), \tag{8.17}$$

with the sum being over all 61^N possible sequences. Of course, this sum is not tractable. When proposing new values for any of the parameters implicated in the stationary distribution, the ratio of two of these terms appears. For simplicity, let $f(s_0, \theta)$ be the unnormalized density:

$$f(s_0, \theta) = e^{-2\beta G(s_0)} \prod_{i=1}^{N} \left(\psi_{s_{i0}} \prod_{c=1}^{3} \varphi_{s_{i0_c}} \right). \tag{8.18}$$

Expanding the MH rule for the present context, we have

$$\vartheta = min\left\{1, \frac{p(D, \phi \mid s_0, \theta', M)p(\theta' \mid M)f(s_0, \theta')q(\theta', \theta)Z_\theta}{p(D, \phi \mid s_0, \theta, M)p(\theta \mid M)f(s_0, \theta)q(\theta, \theta')Z_{\theta'}}\right\}, \tag{8.19}$$

where we have written the complicating factors at the end of the ratio for emphasis. Applying the single variable exchange method given in Murray et al. (2006) to the present problem, we draw an *auxiliary sequence* ς from the distribution induced by θ' using the Gibbs sampling method used in previous chapters. Then, the MH kernel is expanded to

$$\vartheta = min\left\{1, \frac{p(D, \phi \mid s_0, \theta', M)p(\theta' \mid M)f(s_0, \theta')f(\varsigma, \theta)q(\theta', \theta)Z_\theta Z_{\theta'}}{p(D, \phi \mid s_0, \theta, M)p(\theta \mid M)f(s_0, \theta)f(\varsigma, \theta')q(\theta, \theta')Z_{\theta'} Z_\theta}\right\}, \tag{8.20}$$

where all intractable factors at the end of the ratio cancel.

The validity of this MH kernel rests on having truly sampled ς from the stationary probability induced by θ'. As always, we explored empirically the properties of our Gibbs sampler, and devised our implementation to follow a simple procedure: upon starting the overall MCMC, the sequence ς is initialized by performing a random draw from the 61 possible codons at each site; when calling an operator on a parameters bearing on the stationary distribution of the substitution process, 5 Gibbs sweeps across the positions of ς are performed; subsequent calls on parameters bearing on the stationary distribution start from the current ς, and again perform 5 Gibbs sweeps across the sequence. In this manner, we avoid performing a long burn-in of this inner (Gibbs) MCMC when calling parameter updates in the main (MH) MCMC, since at each cycle, ς was previously updated conditional on a parameter vector that was "not too far" in parameter space. Of course, the assumption that the ς sequence is drawn from the intended distribution forms part of the Monte Carlo approximations.

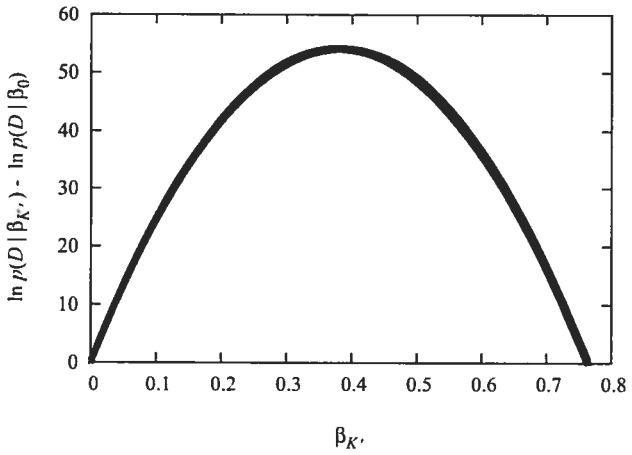

Figure 8.1. Quasi-static thermodynamic integration along β for the MG-F1×4-SC model.

8.2.5 Thermodynamic integrations

The same thermodynamic integration methods described in chapter 4 can be applied here, to contrast a model including the statistical potential with its non-structural counterpart. Recall that the procedure from chapter 4 first produces a trace of the marginal log-likelihood along the β parameter (displayed in fig. 8.1 using the MG-F1×4 as the underlying model), and in the case of SC+β settings, this is followed by an exponentiation and averaging of the curve over the prior distribution (see eqn. 4.8). When using the more rigid SC settings, the value at the $\beta = 1/2$ point along the curve shown in figure 8.1 is the log Bayes factor in favor of the structural model.

Again as in chapter 4, these methods, in conjunction with the model-switch thermodynamic methods proposed in the last chapter, allow for an overall ranking of models based on Bayes factors.

8.3 Results and discussion

8.3.1 Bayes factors

We ran the thermodynamic method combining the statistical potential with a select set of underlying models. These underlying models were chosen based on the results from chapter 7, showing that the MG-F1×4-CP-DP was the preferred model; other models considered suppress either the codon preference parameters or the Dirichlet process on ω factors (in which case a single global ω factor is used), or suppress both of these model settings. The log Bayes factors computed are displayed in table 8.1.

Table 8.1. Natural logarithm of the Bayes factor for models considered, with MG-F1×4 used as a reference.

Model	Non-structural	SC	SC+β
MG-F1×4	0	[48.5; 49.2]	[49.2; 49.6]
MG-F1×4-CP	[122.1; 123.8]	[184.8; 188.3]	[180.3; 183.7]
MG-F1×4-DP	[102.2; 104.2]	[185.7; 188.4]	[180.9; 183.8]
MG-F1×4-CP-DP	[236.5; 241.0]	[316.4; 321.5]	[313.0; 317.7]

Note.—Values given are the upper and lower estimates from bi-directional thermodynamic integrations.

In all cases, the use of the statistical potential provides an increased model fit. We note that for the results to date, considering β as a free parameter has a very mild effect; indeed, the posterior distributions of β, when it is considered as free, do not depart too drastically from $\beta = 1/2$ (fig. 8.2).

Interestingly, we observe a synergistic interplay between the CP, DP and SC configurations. For instance, had the improvement in model fit by combining CP and SC settings been additive, we would have obtained a log Bayes factor of ~ 172 in favor of the MG-F1×4-CP-SC over the reference model. Instead we find a log Bayes factor of [184.8; 188.3]. The synergy between DP and SC settings appears even greater than that

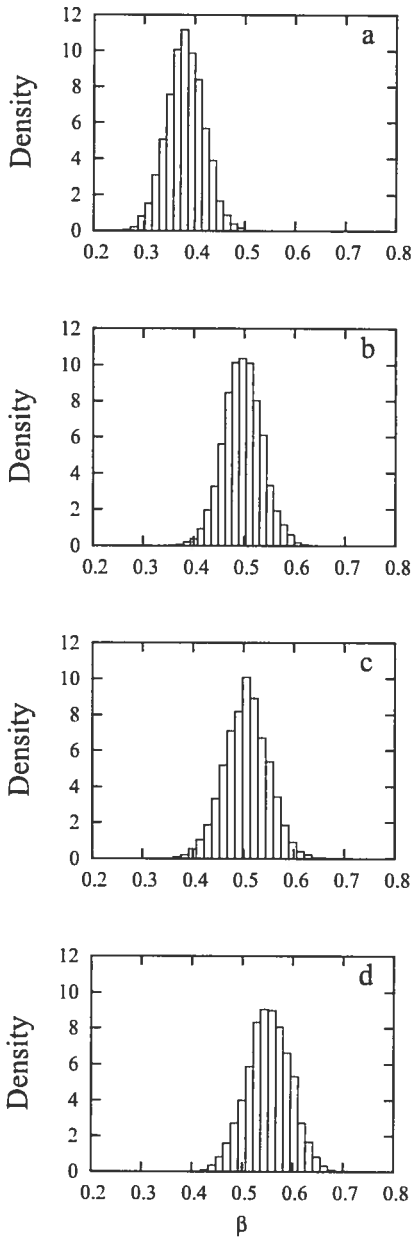

Figure 8.2. Posterior distribution of β for MG-F1×4-SC+β (a), MG-F1×4-DP-SC+β (b), MG-F1×4-CP-SC+β (c), and MG-F1×4-CP-DP-SC+β models.

between CP and SC. The combined CP-DP-SC configurations also display a synergistic effect on model fit, although it appears roughly equivalent in magnitude to the synergy between DP and SC alone. Note that since ω factors do not appear in the stationary distribution of the Markov process, the interaction of the DP and SC settings suggests that the potential has an effect on the transient properties of the substitution process.

We are currently implementing graphical displays of posterior distributions based on "heat" maps, which should allow a visual display of the shifts induced by the SC ($+\beta$) framework (forthcoming!). We note, however, that the use of the statistical potential, without CP or DP settings, produces a model of much poorer fit than the site-independent formulations including these components. Of course, the CP parameters can capture features that the potential cannot. Focusing on the DP settings as a phenomenological account of nonsynonymous rate heterogeneity, the Bayes factors indicate that the potential in itself fails to attain this benchmark. However, we note that the use of a statistical potential is likely to be focused on negative selection, and is unlikely to acknowledge much, if any, positive selection. In contrast, with the DP settings, positive selection is flexibly accounted for, and thus more apt to accommodate a high variance in nonsynonymous rates across sites. In this sense, the DP model does not constitute an entirely fair benchmark. We explore this in greater detail in the next subsection.

8.3.2 Posterior predictive checking

We re-visit the types of posterior predictive checks performed in chapter 4, but now for the codon substitution models. First, using the MG-F1×4-CP model, we computed the variance in number of nonsynonymous substitutions across the codon sites of the alignment, for both *predictive mappings* (simulations of the Markov process over the tree, all the way to the tips of the leaf nodes without any constraints) and "*observed*" *mappings*

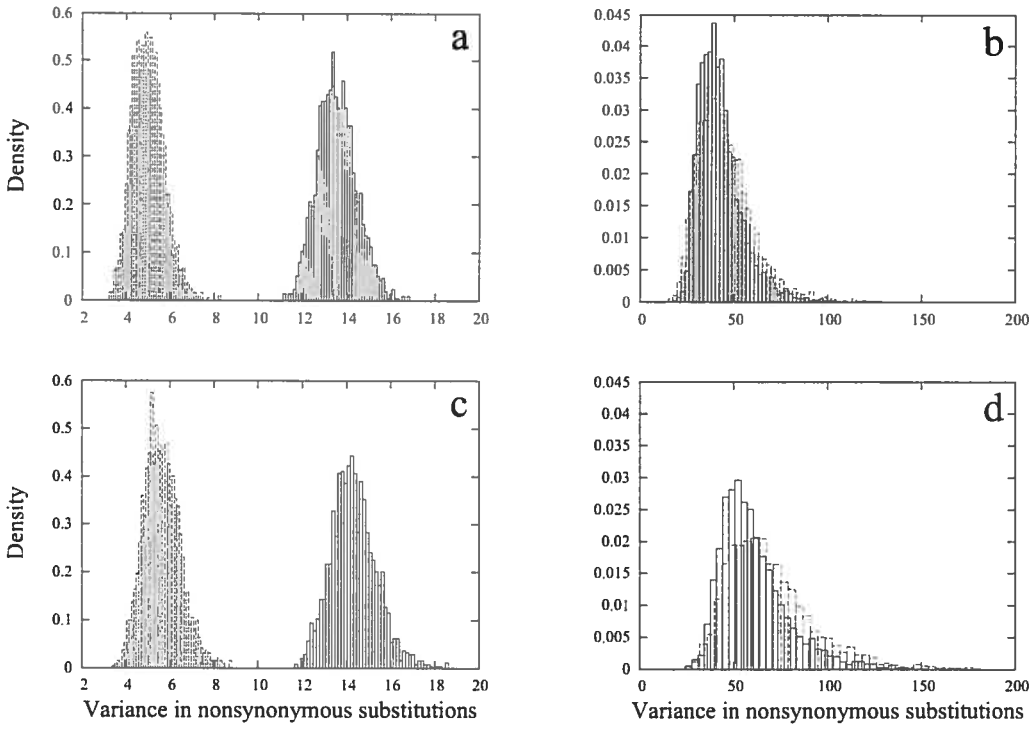

Figure 8.3. Posterior (full line) and posterior predictive (dashed line) variance in the number of nonsynonymous substitutions across the codon positions of the alignment. Panel a) corresponds to the MG-F1×4-CP model, whereas panel b) also includes the DP settings to this model as well. Panel c) corresponds to the MG-F1×4-CP-SC+β, and panel d) includes the DP settings as well.

(instantiations of the Markov process that are compatible with the true observed alignment). These statistics were computed conditional on parameters from 2,000 draws evenly inter-spaced across our MCMC sample from the posterior. As expected, the variance for the predictive mappings tends to be low, by definition of the homogeneous model, whereas the variance for observed mappings, owing to the constraints induced by the data, tends to be higher (fig. 8.3a). Also as expected, when invoking the Dirichlet process modeling nonsynonymous rate heterogeneity across sites (MG-F1×4-CP-DP), the observed mappings have a much higher variance, and the predictive mappings follow a well-matching distribution (fig. 8.3b).

As previously mentioned, the statistical potential used here could, in principle, induce nonsynonymous rate heterogeneity across sites—albeit likely focused on negative

selection. However, as displayed in figure 8.3c for the MG-F1×4-CP-SC+β, nonsynonymous rate heterogeneity is very low in practice, leading a broad discrepancy between distributions. The combined model (MG-F1×4-CP-DP-SC+β) also produces reasonably well-matching observed and predictive distributions (fig. 8.3d), and according to our calculated Bayes factors, the best overall model fit as well.

We next performed a preliminary analysis computing the relative frequency of the different amino acid replacements implied by the observed and predictive mappings. Figure 8.4 displays the mean distribution over the sample, scaled such that the total area of all circles is equivalent across the different panels. The first striking feature of this figure is that several amino acid pairs never exchange with one another. This is of course the effect of considering all substitutions as arising from point mutations, and as originally pointed out by Zuckerkandl and Pauling (1965), single base differences very often lead either to a synonymous codon, or to an amino acid of similar physico-chemical properties.

Among the remaining amino acid pairs that can undergo replacement, we first examined the distribution for observed mappings under the MG-F1×4-CP-DP (fig. 8.4a) and find, as expected, that the mappings suggest uneven exchangeabilities. The corresponding predictive mappings (fig. 8.4b) lead to slightly more even exchangeabilities, but nonetheless already display a surprisingly reasonable skewness in inducing higher values for well-known amino acid pairs (e.g., A-V, A-T, A-S, D-E, I-V). This may be explained by the Zuckerkandl-Pauling effect discussed above. Using the MG-F1×4-CP-DP-SC+β model, the observed distribution (fig. 8.4c) is very similar as observed in figure 8.4a, and the predictive distribution tends to induce slightly higher values for amino acid pairs well-known to being readily exchangeable (fig. 8.4d) than the predictive mappings under the non-structural counterpart (fig. 8.4b), although this is very mildly discernible. This may be one way in which the statistical potential brings an

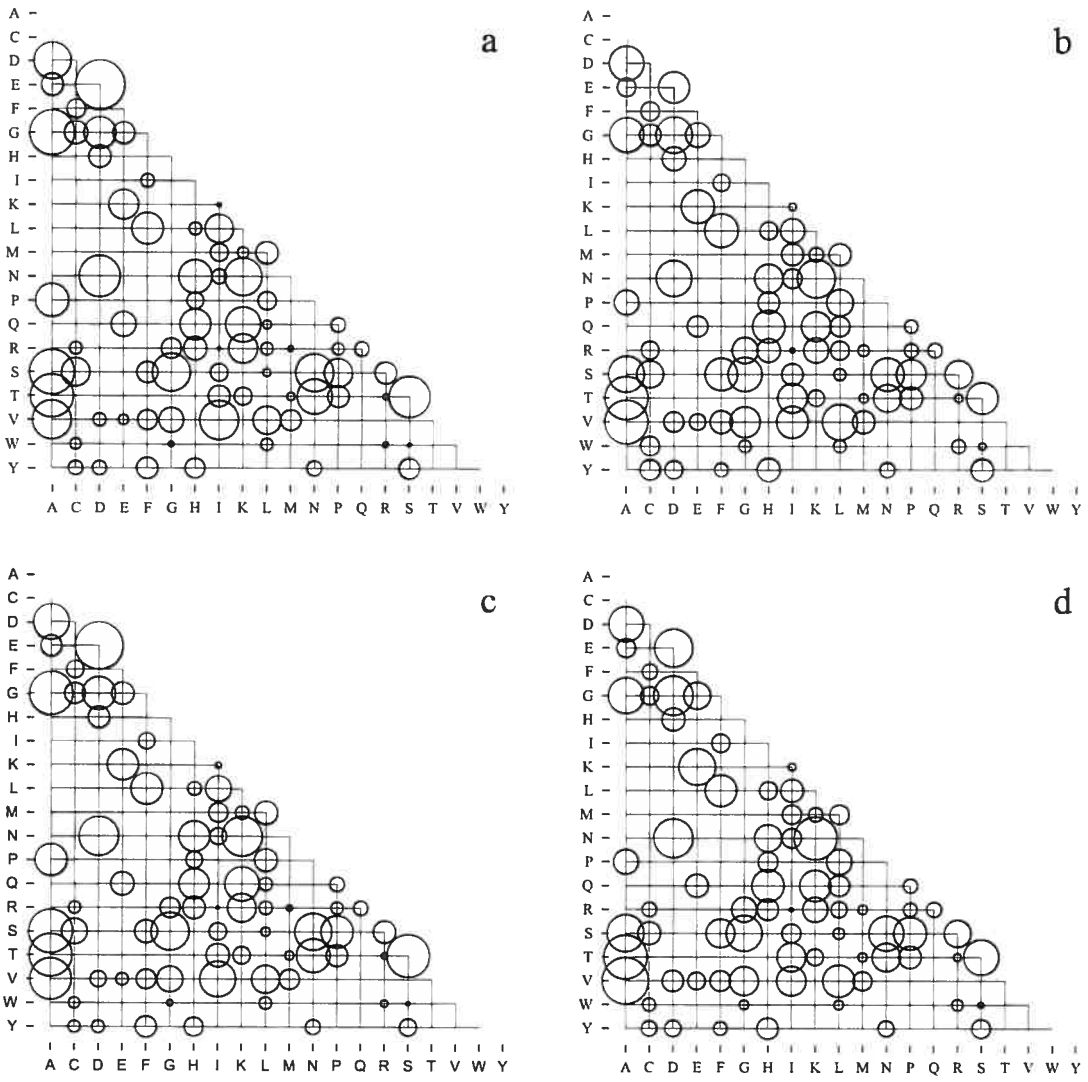

Figure 8.4. Mean amino acid exchange distributions. Panel a) corresponds to that obtained from the observed mappings under the MG-F1×4-CP-DP model, whereas panel b) corresponds to that obtained from the predictive mappings, under the same model. Panel c) is obtained from the observed mappings under the MG-F1×4-CP-DP-SC+β model, and panel d) is obtained from the predictive mappings.

improvement, but more work is needed to clarify this question.

8.4 Conclusions

The set of techniques presented in this chapter completes the site-interdependent methodological context in the case of the codon-based level of interpretation, providing a means for quantitative comparisons of several possible phylogenetic model configurations. While the methods are computationally manageable when coupling the potential to a homogeneous model (with thermodynamic integrations requiring about one week on Xeon 2.4 GHz desktop computer), they become very demanding when invoking the Dirichlet process model (with thermodynamic runs requiring up to four months). The Dirichlet process framework was designed for the context where the data (in this case codon columns) are all considered independent. As presented in chapter 2, the MCMC operators for updating the configuration of the Dirichlet process are applied one codon site at a time, and under the assumption of independence, the likelihood can be calculated for that site only within the operators; for the site-interdependent models, all likelihood calculations are sequence-wide, and thus very costly. It may be possible to design update mechanisms for the Dirichlet process that simultaneously update several sites, thereby "making the most" of each call to the site-interdependent likelihood calculation.

Use of the Dirichlet process approach on nonsynonymous rate factors across sites in combination with the use of the statistical potential could be said to constitute a phenomenological supplement, capturing those features which are beyond the current capabilities of potentials. However, this has the less attractive property of confounding modeling approaches. In any case, we stress here again that judging the relevance of a new class of models should be explored by contrasting such new approaches with existing model forms. In the present case, such a contrasting reveals the importance

combining different modeling strategies into a single framework, with parameterization at the nucleotide, codon, and protein level all meriting further explorations. We propose concrete future research perspectives that could be explored in the last part of this dissertation.

Part III

Perspectives

Chapter 9

Further calculations

9.1 Introduction

The study of site-interdependent models is at an early stage. Most of the work to date has been geared to developing the basic computational techniques needed to instantiate the Bayesian cycle of model development in such contexts. More work is now needed to study the properties of the new models, on much more data. Choi et al. (2007) recently presented a large scale analysis focusing solely on the limiting distribution of the Markov process; by considering data sets consisting of single sequence-structure pairs, the phylogenetic factor of the likelihood function is eliminated, which implies that the data-augmentation-based MCMC device is no longer needed. They find that the modeling approach nearly always improves the model fit. In our own analyses of real data, which include phylogenetic factors, we have also found the approach to show promise in terms of model fit, but from the results of these initial studies, it remains unclear exactly which aspects of the data are better explained by the model. Future cycles of the Bayesian framework will hopefully clarify these issues.

In this chapter we suggest a few more calculatory methods to include in this iterative model development process. We focus on the question of the contact map representa-

tion, and outline a crude yet relatively fast means of assessing the complete set of *nearest neighbor contact maps*. We next describe thermodynamic schemes to evaluate different forms of potentials in a computationally sensible manner. Finally, we propose simple approaches to explore if the use of potentials plays a more important role in ameliorating the stationary probability of the site-interdependent Markov process or the transient attributes of the process.

9.2 Contrasting nearest-neighbor contact maps

In our present framework, it would be interesting to compute the marginal likelihood of alternative contact maps. In the short-term, this could highlight the sensitivity of the model to mildly different contact maps, and in the long-term one might envisage approaching the protein folding problem with a phylogenetic component. We have already done crude explorations along these lines in chapter 4, where we deliberately made the contact map "worse" than the true native one, and, as expected, found support for the native contact map. However, we would like a more sensible exploration, by evaluating, for instance, the sensitivity of each contact map entry (0 or 1). In this subsection, we describe a crude procedure and an importance sampling argument that could be used to approximate the Bayes factor of the nearest neighbor contact maps (with respect to the native contact map), although it could perhaps be applied to more diverging contact maps. We present the developments under the site-interdependent model allowing for gamma distributed rates across sites; codon-based developments are a special case of the developments that follow.

First, recall that the marginal likelihood can be decomposed as

$$p(D \mid M) = \int_{\Theta} p(D \mid \theta, M) p(\theta \mid M) d\theta \tag{9.1}$$

$$= \int_{\Theta} \int_{\mathbf{r}} \int_{\Phi} p(D, \phi \mid \theta, r, M) p_{\alpha}(r) p(\theta \mid M) d\phi dr d\theta. \tag{9.2}$$

Our objective is to compare the marginal likelihood obtained under two different contact maps Δ and Δ', which differ only by one entry (changed to 1 if the native contact map entry of interest is 0, or changed to 0 if the native contact map is 1). However, to be able to do a relatively quick exploration, we could ease-up on the outermost integral of equation 9.2. Specifically, we will make two approximating assumptions: 1) let us suppose that the values of all parameters involved in the stationary probability, which we refer to as θ_{stat}, are known (the reasons for this will become apparent shortly), and 2) these values are the same under both Δ and Δ' (the importance sampling approximation). Under these two conditions, we write the ratio of contact map log-marginal-likelihoods as:

$$\begin{aligned} \ln \frac{p(D \mid \theta_{stat}, \Delta')}{p(D \mid \theta_{stat}, \Delta)} &= \Big(\ln p(s_0 \mid \theta_{stat}, \Delta') - \ln p(s_0 \mid \theta_{stat}, \Delta) \Big) \\ &\quad + \left(\ln \int_{\Theta} \int_{\mathbf{r}} \int_{\Phi} p(D, \phi \mid \Delta', s_0, \theta, r) p_{\alpha}(r) p(\theta) d\phi dr d\theta \right. \\ &\quad \left. - \ln \int_{\Theta} \int_{\mathbf{r}} \int_{\Phi} p(D, \phi \mid \Delta, s_0, \theta, r) p_{\alpha}(r) p(\theta) d\phi dr d\theta \right) \end{aligned} \tag{9.3}$$

where the integration over Θ is now limited to integrating over parameters not involved in the stationary distribution, and where we have dropped the dependence on M from the notation. This approximation can be developed separately for the terms at the root (in the first set of parenthesis) and the terms over the tree (in the second set of parenthesis).

First, for the terms at the root we get:

$$\begin{aligned}
\ln \frac{p(s_0 \mid \theta_{stat}, \Delta')}{p(s_0 \mid \theta_{stat}, \Delta)} &= \ln p(s_0 \mid \theta_{stat}, \Delta') - \ln p(s_0 \mid \theta_{stat}, \Delta) && (9.4)\\
&= \left(-2\beta G_{s_0|\Delta'} - \ln Z_{\Delta}\right) - \left(-2\beta G_{s_0|\Delta} - \ln Z_{\Delta'}\right) && (9.5)\\
&= \left(2\beta G_{s_0|\Delta} - 2\beta G_{s_0|\Delta'}\right) - \ln \frac{Z_{\Delta'}}{Z_{\Delta}} && (9.6)\\
&= \left(2\beta G_{s_0|\Delta} - 2\beta G_{s_0|\Delta'}\right) - \langle 2\beta G_{s|\Delta} - 2\beta G_{s|\Delta'} \rangle && (9.7)
\end{aligned}$$

where Z_{Δ} is the normalizing factor under Δ, and where $\langle . \rangle$ represents an expectation with respect to the stationary probability under Δ. This expectation can be estimated based on a sample of sequences $(s^{(h)})_{1 \leq h \leq K}$ obtained using the Gibbs sampling procedure we have been employing:

$$\langle 2\beta G_{s|\Delta} - 2\beta G_{s|\Delta'} \rangle \simeq \frac{1}{K} \sum_{h=1}^{K} 2\beta G_{s^{(h)}|\Delta} - 2\beta G_{s^{(h)}|\Delta'} \tag{9.8}$$

We will refer to (9.8) as the *root importance sampling* approximation.

Similarly, for the terms over the tree, we get:

$$\begin{aligned}
&\ln \frac{\int_{\Theta} \int_{\mathbf{r}} \int_{\Omega} p(D, \phi \mid \Delta', s_0, \theta, r) p_{\alpha}(r) p(\theta) d\phi dr d\theta}{\int_{\Theta} \int_{\mathbf{r}} \int_{\Phi} p(D, \phi \mid \Delta, s_0, \theta, r) p_{\alpha}(r) p(\theta) d\phi dr d\theta}\\
&\simeq \frac{1}{K} \sum_{h=1}^{K} \left(\ln p(D, \phi^{(h)} \mid \Delta', s_0, \theta^{(h)}, r^{(h)}) - \ln p(D, \phi^{(h)} \mid \Delta, s_0, \theta^{(h)}, r^{(h)}) \right) \qquad (9.9)
\end{aligned}$$

where $(\theta^{(h)}, \phi^{(h)}, r^{(h)})_{1 \leq h1 \leq K}$ is a sample of parameters with no bearing on the stationary distribution of the Markov process, as well as mappings and rates, obtained using the Metropolis-Hastings algorithm. We will refer to (9.9) as the *tree importance sampling* approximation.

From the developments explained above, the difference in log-marginal-likelihood between two nearest neighbor contact maps is computed in two parts, applying the root importance sampling approximation and the tree importance sampling approximation

separately, and based on two distinct MCMC sampling schemes. The following protocol summarizes:

- Run a MCMC sampling over parameters not involved in the stationary distribution, and with those that are pre-fixed to sensible values; one crude approach is to first run a full MCMC sampling over all parameters, as in previous chapters, and then run a second MCMC fixing all parameters involved in the stationary probability to their mean posterior values from the first run.

- Run a Gibbs sampling MCMC to obtain a sample of sequences from the stationary probability, induced from the relevant pre-fixed parameter values.

- Make a modification to the contact map, and, based of the two samples above, apply the root and tree importance sampling approximations to compare the contact map log-marginal-likelihoods.

- Repeat for each contact map of interest, always based on the same samples.

This last point, using the same sample for each contact map of interest, is what should make these preliminary explorations reasonably fast. Note, in particular, that had the parameters involved in the stationary distribution also been integrated over, a new sample of sequences would be needed for the root importance sampling approximation for each parameter vector sampled. This would undoubtedly slow down the procedure, although it is not entirely unfeasible either.

Finally, it should also be noted that the importance sampling arguments suggested here are best when our second approximating assumption (the assumed equivalent parameter values under both Δ and Δ') is reasonable. This assumption is likely to become markedly erroneous if the two contact maps compared are significantly different, which is the reason for constraining our preliminary analyses on nearest neighbor contact maps.

9.3 Contrasting different structural representations

In chapter 5, we performed a simple contrast via Bayes factors of the potentials of Miyazawa and Jernigan (1985) to our own potential of the same form, and found our potential to have a better fit. Furthermore, we found the combined potential, including contact map and solvent accessibility components, to outperform all of these. Several other forms of potentials are of interest. Even among those already derived, we have not yet evaluated the pure solvent accessibility potential in the phylogenetic context. Note that this last potential does not lead to site-interdependence, so that the thermodynamic methods can be based on a sum (of log terms) across sites, without additional MCMC sampling for intractable normalizing factors, and therefore computed much more quickly than under the full sequence-wide framework. In theory, such models could even be manipulated using the traditional pruning-based likelihood calculations, although this is unadvised, because data-augmentation-based schemes yield much faster samplers (not shown, but see Lartillot, 2006).

Different forms of site-independent potentials could be evaluated in this way, and when re-introducing the contact map component, the resulting potential can of course be evaluated directly, based on the methods expounded in this work. However, a slightly different approach might be more efficient when working with the SC-type models with $\beta = 1/2$. Taking the contact and solvent accessibility potential as an example, the approach first defines

$$G(s) = \beta_{dep}\left(\sum_{1\leq i<i'\leq N} \Delta_{ii'}\epsilon_{s_i s_{i'}}\right) + \beta_{indep}\left(\sum_{1\leq i\leq N} \Xi^{v_i}_{s_i} + \sum_{1\leq i\leq N} \Sigma_{s_i}\right). \tag{9.10}$$

In a first step, we set $\beta_{dep} = 0$, and perform a site-independent data-augmentation-based thermodynamic integration from $\beta_{indep} = 0$ to $\beta_{indep} = 1/2$. This provides the log Bayes factor in favor of the model including the site-independent components

of the potential over the underlying non-structural model. Then, in a second step, we set $\beta_{indep} = 1/2$, and perform a site-interdependent thermodynamic integration from $\beta_{dep} = 0$ to $\beta_{dep} = 1/2$. This provides the log Bayes in favor of the model including the contact component over the model without it. The overall log Bayes factor, in favor of the overall potential, is then simply the sum of both log Bayes factors. The advantage of separating the calculation into two steps is that the first step is very efficient, whereas the costly site-interdependent thermodynamic integration is differed to a path in the space of posterior distributions that is as short as possible. The computational advantages need be assessed in practice, and our implementation is already equipped to do so.

Also note that β_{indep} and β_{dep} could be treated as free parameters[1]. In this case, a similar two-stage procedure could be applied. First run a site-independent data-augmentation-based thermodynamic integration along the dimension of β_{indep} (with $\beta_{dep} = 0$), tracing log-marginal-likelihood curve as a function of β_{indep}. Then exponentiate and average this curve over the prior, as in chapter 4. This provides the log Bayes factor in favor of the model including the site-independent components of the potential, but now with β_{indep} treated as a free parameter of the model. Then, in a second thermodynamic run, apply plain MH operators on β_{indep}, and trace the log-marginal-likelihood curve as a function of β_{dep}. Exponentiating and averaging this curve provides the log Bayes factor in favor of the model including the contact component over the model without it, but now with β_{dep} being a free parameter. The sum of both these log Bayes factors again provides the overall log Bayes factor in favor of the full structural model, with both β_{indep} and β_{dep} treated as free parameters.

[1]The parameter β_{indep} could also be further subdivided into two parameters: one in front of the solvent component, and one in front of the chemical component.

9.4 Evaluating transient versus stationarity contributions to model fit

The use of statistical potentials within evolutionary models induces differences in both transient and stationary probabilities under the Markov process. This can be plainly seen from equation (C.9), in Appendix C, where the derivative of the augmented log-likelihood with respect to β involves both the stationary probability and the transient probability, in two distinct terms. It would be interesting to evaluate if the amelioration in model fit is mainly a result of a greater stationary probability, or a greater transient probability. However, in spite of the overall log Bayes factor being invariant to the position of the root node, the relative contribution of each term is not.

A very simple exploration of this question would start by analyzing pairs of sequences, and repeating calculations twice, taking each sequence in turn as the root. The results of calculations under both rootings should provide a first indication of stationary versus transient ameliorations. This should probably be explored for different levels of evolutionary divergence, and the natural extension of such an analysis would be applied to multiple sequence alignments. Our current implementation is based on rooting the tree at a leaf node (an observed sequence). As such, we could first try repeating calculations with a different leaf node rooting in each instance (perhaps a random subset, of say 10 leaf nodes, could suffice for these first explorations). The relative contribution of each term could then be averaged over these instances. Characterizing these properties in practice, under different model configurations, and for different data sets, should help clarify strengths and weaknesses of different choices in terms of stationary and transient amelioration.

These sorts of evaluations need not be restricted to the structural models studied in the present dissertation. We stress that the distinction between stationary and transient

probabilities should not be overlooked in the development of phylogenetic models, and that models focused solely on parameterizations of transient aspects (e.g., Kosiol et al., 2007), may be ill-suited under certain data set conditions, due to their inability to anticipate sequence saturation (Lartillot et al., 2007).

9.5 Conclusions

We have focused here on calculations of interest that require little, or no further developments within our implementation. Several other calculations are also of interest, including many other possible statistics for posterior predictive analysis (e.g., Dimmic et al., 2005; Lartillot et al., 2007), and assessments of other structural features. We hope to set up a pipeline of analysis, applied to numerous data sets, incorporating these calculations.

Also note that the different evaluations outlined in this chapter can be intersected; one might speculate that uncovering better site-independent components for structural models—utilizing the site-independent thermodynamic integrations discuss above—could "take charge" of certain structural features, in a sense "freeing" the contact component to focus on actual correlations—which might be reflected in our assessments of nearest neighbor contact maps. The impact of stationary and transient probabilities could also be investigated within assessment of nearest neighbor contact maps, by focusing either on the root or tree importance sampling approximation, perhaps averaged over root placements.

Eventual studies could also expand the research pipeline to include several other types of models, as we discuss in the next chapter.

Chapter 10

Model variations and extensions

10.1 Introduction

Several of the models studied in the present work remain relatively rudimentary. For instance, the codon models proposed in chapter 7 are all based on global amino acid or codon preferences. However, other modeling strategies have already shown, at least for amino acids, that such preferences are markedly heterogeneous across sites (e.g., Lartillot and Philippe, 2004, 2006). As far as the structural models are concerned, several crude simplifications are relied upon, such as the simple, static contact map representation. It would have been quite surprising, in fact, to find such a representation constituting an adequate description of amino acid interactions, and computationally simple means of enriching the basic contact map approach are of pressing interest. It would also be interesting to give greater flexibility to the coefficients of some or all components of the statistical potential directly within the phylogenetic context.

In this chapter, we describe these modeling themes in greater detail, as examples of some of the possible extensions. Our focus is on the codon-based models, and we present specific models addressing issues mentioned above.

10.2 Dirichlet process modeling

A broad range of model extensions are evident from the AAP and CP approaches proposed in chapter 7: given that these richer MG-style models lead to an improved overall fit, models based on mixtures of AAP or CP parameters could also be of merit, so as to capture site-specific preferences rather that global effects. To this end, the Dirichlet process prior, applied to model nonsynonymous rate heterogeneity across sites (Huelsenbeck et al., 2006), could also be applied to the AAP parameters, or to the CP parameters. Indeed, the necessary MCMC operators for manipulating such models, as well as models incorporating heterogeneities along the tree, have all been described previously (Lartillot and Philippe, 2004; Blanquart and Lartillot, 2006).

As an initial specific example, let us specify a model incorporating two independent Dirichlet processes: one acting on overall nonsynonymous rate heterogeneity across sites, and another acting on amino acid preferences across sites[1]. First, let us suppose that the current configuration of the Dirichlet process on ω consists of H classes, and let $y = (y_i)_{1 \leq i \leq N}$ be the allocation vector for *omega* classes, with y_i giving the index of the ω factor affiliated to site i. Next, let $\varpi_{z_i} = (\varpi_{z_i k})_{1 \leq k \leq 20}$ be the amino acid preference parameters currently affiliated to site i, where $z = (z_i)_{1 \leq i \leq N}$ is the allocation vector on amino acid vectors. Then, site-specific codon substitution matrices are given by:

$$Q_{ab}^{(i)} \propto \begin{cases} \varrho_{a_c b_c} \varphi_{b_c}, & \text{if } \mathcal{A}, \\ \omega_{y_i} \varrho_{a_c b_c} \varphi_{b_c} \left(\frac{\varpi_{z_i f(b)}}{\varpi_{z_i f(a)}} \right)^{\frac{1}{2}}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise.} \end{cases} \tag{10.1}$$

Note that omitting the ω factor (i.e., fixing $\omega = 1$ across all sites) would constitute a negative-selection model; a site could have a very low effective nonsynonymous rate

[1]The extension of the AAP model described here could also be applied to the CP model, to account for heterogeneous codon preference across positions.

by being affiliated to an amino acid preference class having most of its mass on a single amino acid, but a high effective nonsynonymous rate, say higher than the overall synonymous rate, would not be possible. In other words, such a model selects against certain amino acids, by attributing little mass to these. With the two Dirichlet processes, on both amino acid preference parameters and nonsynonymous rate factors, therefore, a re-interpretation of the traditional meaning given to ω, as the ratio of the nonsynonymous to synonymous rates, will need to be formulated; negative selection does not strictly correspond to $\omega < 1$ in this case, since, as described above, the model can accommodate negative selection via amino acid preference parameters. Indeed, we foresee possible identifiability problems between ω factors and amino acid preference parameters.

These models in themselves are of much interest, providing a more flexible approach than using pre-determined amino acid preferences (e.g., Sainudiin et al., 2005; Wong et al., 2006), and a less drastic alternative to using a distinct set of amino acid parameters at each site (Halpern and Bruno, 1998). However, our motivation here is to combine such a model with a contact potential. The contact potential can be incorporated by applying the same re-formulation of the Markov process given in equation (8.3). We speculate that with well-defined site-specific amino acid preferences, the contact potential could more clearly recognize pairwise correlations, since for a given pairwise contact, the possible amino acid state interactions become much more restricted.

The model proposed here poses significant computational challenges, in particular with regards to Dirichlet process update operators under site-interdependence. Further technical investigations are needed to address these difficulties in practice. Also, the model should be viewed as a phenomenological supplement, as part of an exploratory stage of development, particularly given the interpretative difficulties that it poses.

10.3 Multiple protein structures, and interdependence across genes

As previously mentioned, the single, fixed contact map structure representation utilized in this work is very crude. While it may be pertinent to design a class of model where a contact map, or some other representation, is allowed to change over the tree, we first consider a few simple ideas, which attempt to acknowledge that a protein has some level of structural flexibility in the course of its own half-life.

Let us first build upon the basic contact potential, with the form

$$G(s) = \sum_{1 \leq i < i' \leq N} \Delta_{ii'} \epsilon_{s_i s_{i'}} + \sum_{1 \leq i \leq N} \Sigma_{s_i}. \tag{10.2}$$

The contact map Δ is derived from a single reference structure. One may also consider that many proteins, in performing their biological function, exist in two or more structural states. Structural states may be difficult to characterize in a clear-cut manner, but some cases offer natural discretizations, such as the well-studied oxy- and deoxy-myoglobin structures. In this specific instance we could define a pseudo-energy score with the form

$$G(s) = \sum_{1 \leq i < i' \leq N} \Delta^{oxy}_{ii'} \epsilon_{s_i s_{i'}} + \sum_{1 \leq i < i' \leq N} \Delta^{deoxy}_{ii'} \epsilon_{s_i s_{i'}} + \sum_{1 \leq i \leq N} \Sigma_{s_i}, \tag{10.3}$$

where Δ^{oxy} and Δ^{deoxy} are the contact maps derived from the resolved structures of oxy- and deoxy-myoglobin respectively. The idea here is that an amino acid sequence should fit well with all structural states of the protein.

Along similar lines, it might be interesting to explore whether it is possible to account for the fact that an amino acid sequence must adhere to the conformational constraints of a particular stand-alone structure as well as the conformational constraints imposed

by interactions with other proteins. Data sets have already been constructed with this in mind, in the context of the protein-docking problem (e.g., Mintseris et al., 2005). More specifically, let s^{AB} represent the joint sequence state of proteins A and B (with their respective lengths written as N^A and N^B, and with the joint length as N^{AB}). Focusing on the contact components, the potential is given by

$$\begin{aligned} G(s^{AB}) &= \sum_{1 \le i < i' \le N^{AB}} \Delta^{AB}_{ii'} \epsilon_{s^{AB}_i s^{AB}_{i'}} + \sum_{1 \le i < i' \le N^{A}} \Delta^{A}_{ii'} \epsilon_{s^{A}_i s^{A}_{i'}} \\ &+ \sum_{1 \le i < i' \le N^{B}} \Delta^{B}_{ii'} \epsilon_{s^{B}_i s^{B}_{i'}} + \sum_{1 \le i \le N^{AB}} \Sigma_{s^{AB}_i} \end{aligned} \tag{10.4}$$

where Δ^{AB} is the contact map of the complex composed of protein A and B, and Δ^A and Δ^B are the contact maps of each individual subunit. Perhaps a weighting scheme should be applied to each of the subunits and the resulting complex.

Note that such a model can lead to interdependencies across sites of two different genes. One might even imagine studying a multi-gene data set, modeling networks of interaction across multiple gene sites, within the phylogenetic context.

10.4 Coefficients of the potential as free parameters

We have stressed that relying too heavily on a statistical potential leads to a model of poor fit. It remains unclear as to whether this is due to the basic form of the structural representation and potentials investigated here, or if it is a result of the fact that the potentials themselves were not derived in a true evolutionary framework. One way of addressing this question would be to construct a large data set, and treat the coefficients of the potential as free parameters. Such a data set might consist of a single multi-gene alignment, with each gene encoding for a protein of known structure. Alternatively, we could use several single gene data sets, each with their own tree structure, but with the coefficients of the potential acting as global parameters, over the entire meta-data set.

Once again, the computational challenge of such a model is significant. The single variable exchange algorithm employed in chapter 8, or other approaches discussed in Murray et al. (2006), however, could provide useful avenues in this direction. If, under a data set (or meta-data set) of sufficient size, leading to well-focused posterior distributions, the model still does not induce any significant rate heterogeneity, or any other expected evolutionary feature, one could reasonably speculate that the form of the potential simply *cannot* produce such features, and that other forms of structural representations—or other sequence fitness proxies in general—should be considered.

10.5 Conclusions

Numerous modeling extensions can be envisaged based on the ideas already discussed in the present dissertation. These models, when passed through the pipeline of analysis discussed in the previous chapter, should better inform further instances of the Bayesian model development cycle. In particular, we are very interested in engaging this development cycle with richer structural descriptions than the simple contact/solvent accessibility versions used and discussed in this work, in order to quantify the relative importance of different structural features, and how these relate to each other in the overall evoutionary process.

Afterword

All models are based on a blending of both phenomenological and mechanistic approaches. This makes the concepts somewhat difficult to grasp. Perhaps models themselves should not be viewed as being either *phenomenological* or *mechanistic*, and that these terms should be restricted to the *process of model development*: phenomenological modeling consists of drawing up a preliminary sketch of the most blatant features suggested from the data, whereas mechanistic modeling aims to provide a generalization or a synthesis of such a preliminary sketch, by attempting to describe the underlying causes that would lead to the observed features. Note that a description of the underlying causes may itself be based on a phenomenological interpretation.

We have seen examples of this modeling process in the present work. Working with non-structural models, we encountered an example in chapter 7 of a phenomenological modeling approach (the MG-F3×4 formulation, designed to accomodate the periodic pattern of nucleotide propensities at the three positions a codon) reappraised mechanistically (into the MG-F1×4-CP model, which considers the periodic pattern of nucleotides as arising from the coding nature of the data). Nonetheless, the CP parameters themselves constitute a phenomenological account. The structural modeling approaches studied here also subscribe directly to this modeling process; we have attempted to use an explicit protein structure description and statistical potential to mediate nonsynonymous rates of substitution, but the potential is itself based on phenomenologcial interpretations, and goes one step further in fixing parameter values

according to empirical observations. In loose terms, all models have a phenomenological lining, that could potentially be revised mechanistically. For instance, rather than simply attributing a free set of nucleotide exchangeability parameters, we could focus on encoding structural aspects of nitrogenous bases, or perhaps a modeling of deamination tendencies. Nucleotide propensity parameters could perhaps be replaced with a modeling of bio-energetic costs of de novo nitrogenous base synthesis. Codon preference parameters could be elaborated, into a modeling of tRNA abundance and/or translational accuracy. The stability of mRNA could also be considered, possibly using similar formulations to those used herein.

Many other developments can be envisioned. Each of these developments would undoubtedly rest on further phenomenological strategies. However, as in the case of the structural models studied here, such efforts are geared to *"pushing down" the phenomenological line of interpretation.* It would seem naive at this point to strive for some sort of ultimate mechanistic floor (as in the traditional aspirations of physics), particularly given the relatively rudimentary forms of current models; even the richest phenomenological modeling approaches studied in this dissertation are quite crude, completely ignoring the possibility of recombination, insertions or deletions, or any other high-order events. In other words, much more preliminary sketching is needed in order to stimulate further mechanistic strategies; the pursuit of both phenomenological and mechanistic modeling approaches, in combination with quantitative and qualitative probabilistic assessments aimed at determining if mechanistic approaches compare with phenomenological schemes, constitutes what we have called phenomenological benchmarking.

As computational biology enters its pubescent phase as a discipline, we can think of at least three main advantages to pursuing phenomenological benchmarking. The first is that it provides a concrete framework for attempting to formalize our current

state of understanding. This is well illustrated with the last of the codon-based models studied here, which, to the best of our knowledge, are the first to parameterically recognize the possibility of evolutionary influences coming from each level of the central dogma of molecular biology. The fact that no previous evolutionary model has ever explicitly recognized this basic biological understanding highlights the infancy of the field. The second advantage is that when mechanistic approaches appear weak, they can at least be combined with a phenomenological supplement, as a pragmatic short-term alternative. The third and most important advantage is that it generally leads to models that incorporate seemingly disparate data within an encompassing probabilistic framework; in sharp contrast with our reductionist heritage, it offers the means of integrating different biological sub-disciplines into a broad evolutionary framework. The present work highlights this advantage: from the methods we have expounded, and the future calculations and modeling extensions suggested, a vast research landscape emerges, in which the distinctions between structural and evolutionary biology become artificial. With the growing banks of data coming from the numerous domains of the life-sciences, sound probabilistic approaches that offer means of merging sub-disciplines will be essential to building a strong scientific structure, and to deepening our basic understanding of evolutionary biology.

Bibliography

Aris-Brosou, S., and J. P. Bielawski. 2006. Large-scale analyses of synonymous substitution rates can be sensitive to assumptions about the process of mutation. *Gene* 378:58–64.

Arndt, P. F., C. B. Burge, and T. Hwa. 2002. DNA sequence evolution with neighbor-dependent mutation. In *Proceedings of the sixth annual internation conference on computational biology*, ed. G. S. Myers, S. Hannenhalli, S. Istrail, P. Pevzner, and M. Waterman, 32–38. Association for Computing Machinery.

Avery, O. T., C. M. MacLeod, and M. McCarty. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.* 79:137–158.

Bartlett, M. S. 1957. A comment on D. V. Lindley's statistical paradox. *Biometrika* 44:533–534.

Bastolla, U., J. Farwer, E. W. Knapp, and M. Vendruscolo. 2001. How to guarantee optimal stability for most representative structures in the protein data bank. *Proteins* 44:79–96.

Beaumont, M. A., and B. Rannala. 2004. The Bayesian revolution in genetics. *Nat. Rev. Genet.* 5:251–261.

Blanquart, S., and N. Lartillot. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.* 23:2058–2071.

Brinkmann, H., M. van der Giezen, Y. Zhou, G. Poncelin de Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst. Biol.* 54:743–757.

Brooks, S. P. 2003. Bayesian computation: a statistical revolution. *Phil. Trans. R. Soc. Lond. A* 361:2681–2697.

Buckley, T. R., C. Simon, and G. K. Chambers. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: Effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. *Syst. Biol.* 50:67–86.

Burrow, J. W. 1966. *Evolution and society: A study o Victorian social theory*. Cambridge University press.

Chiu, T. L., and R. A. Goldstein. 1998. Optimizing potentials for the inverse folding problem. *Protein Eng.* 11:749–752.

Choi, S. C., A. Hobolth, D. M. Robinson, H. Kishino, and J. L. Thorne. 2007. Quantifying the impact of protein tertiary structure of molecular evolution. *Mol. Biol. Evol.* 24:1769–1782.

Darwin, C. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray. Retrived at `http://darwin-online.org.uk/`.

Dayhoff, M.O., R. V. Eck, and C. M. Park. 1972. A model of evolutionary change

in proteins. In *Atlas of protein sequence and structure*, ed. M.O. Dayhoff, 88–89. National Biomedical Research Foundation, Washington, DC.

Dayhoff, M.O., R.M. Schwartz, and B.C. Orcutt. 1978. A model of evolutionary change in proteins. In *Atlas of protein sequence and structure*, ed. M.O. Dayhoff, 345–352. National Biomedical Research Foundation, Washington, DC.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* 39:1–22.

Dennett, D. C. 1995. *Darwin's dangerous idea: evolution and the meanings of life*. Simon and Schuster.

Dimmic, M. W., M. J. Hubisz, C. D. Bustamante, and R. Nielsen. 2005. Detecting coevolving amino acid sites using Bayesian mutational mapping. *Bioinformatics* 21:S126–S135.

Doron-Faigenboim, A., and T. Pupko. 2007. A combined empirical and mechanistic codon model. *Mol. Biol. Evol.* 24:388–397.

Drexler, K. E. 1981. An approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. USA* 78:5275–5278.

Fearnhead, P., and C. Sherlock. 2006. An exact Gibbs sampler for the markov-modulated Poisson process. *J. R. Statist. Soc. B* 68:767–784.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17:368–376.

Felsenstein, J., and G. A. Churchill. 1996. A hidden makov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* 13:93–104.

Fitch, W. M. 1980. Estimation the total number of nucleotide substitutions since the common ancestors of a pair of homologous genes: comparison of several mehtods and three beta hemoglobin messenger RNA's. *J. Mol. Evol.* 16:153–209.

Futuyma, D. J. 1986. *Evolutionary biology*. Sunderland, Massachusetts: Sinauer Associates. 2nd ed.

Galassi, M., J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F Rossi. 2003. *Gnu scientific library: reference manual; 2nd edition*. Network Theory. Ltd.

Galtier, N., and M. Gouy. 1998. Infering pattern and process: maximum likelihood implementation of a non homogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* 15:871–879.

Gelman, A. 1998. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science* 13:163–185.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, A., X. L. Meng, and H. Stern. 1996. Posterior predicive assessment of model fitness via realised discrepancies. *Statistica Sinica* 6:733–807.

Godzik, A., Kolinski A., and J. Skolnick. 1995. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci.* 4:2107–2117.

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11:725–736.

Gould, S. J. 2002. *The structure of evolutionary theory*. London: Belknap Press.

Gross, D., and D. R. Miller. 1984. The randomization technique as a modeling tool and solution procedure for transient Markov processes. *Oper. Res.* 32:343–361.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52:696–704.

Halpern, A. L., and W. J. Bruno. 1998. Evolutionary distances for protein-coding sequences: modeling site- specific residue frequencies. *Mol. Biol. Evol.* 15:910–917.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.

Hershey, A. D., and M. Chase. 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* 1:39–56.

Hubbard, S. J., and J. M. Thornton. 1993. *Naccess*. London: University Colledge, Depart. of biochem. and mol. biol.

Huelsenbeck, J. P., and K. A. Dyer. 2004. Bayesian estimation of positively selected sites. *J. Mol. Evol.* 58:661–672.

Huelsenbeck, J. P., S. Jain, S. W. D. Frost, and S. L. K. Pond. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc. Natl. Acad. Sci. U.S.A.* 103:6263–6268.

Jaynes, E.T. 2003. *Probability theory; the logic of science*. Cambridge University Press.

Jeffreys, H. 1935. Some tests of significance, treated by the theory of probability. *Proc. Camb. Phil. Soc.* 31:203–222.

Jensen, A. 1953. Markoff chains as an aid in the study of Markoff processes. *Skandinavisk Aktuarietidskriff* 36:87–91.

Jensen, J. L., and A.-M. K. Pedersen. 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv. Appl. Prob.* 32:499–517.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992a. A new approach to protein fold recognition. *Nature* 358:86–89.

Jones, D. T., W. R. Taylor, and J. M Thornton. 1992b. The rapid generation of mutation data matrices from protein sequences. *CABIOS* 8:275–282.

Kass, R.E., and A.E. Raftery. 1995. Bayes factors and model uncertainty. *J. Am. Stat. Assoc.* 90:773–795.

Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* 217:624–626.

Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220:671–680.

Kleinman, C. L., N. Rodrigue, C. Bonnard, H. Philippe, and N. Lartillot. 2006. A maximum likelihood framework for protein design. *BMC-Bioinformaics* 7:326.

Kosakovsky Pond, S. D., S.L. Frost. 2005. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol. Biol. Evol.* 22:478–485.

Kosakovsky Pond, S. L., and S. D. Frost. 2005. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* 22:1208–1222.

Kosakovsky Pond, S. L., and S. V. Muse. 2005. Site-to-site variation of synonomous substitution rates. *Mol. Biol. Evol.* 22:2375–2385.

Kosiol, C., I. Holmes, and N. Goldman. 2007. An empirical codon model for protein sequence evolution. *Mol. Biol. Evol.* 24:1464–1479.

Krauss, S., and X. T. Wang. 2003. The psychology of the Monte Hall problem: discovering psychological mechanisms for solving a tenacious brain teaser. *J. Exp. Psychol. Gen.* 132:3–22.

Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* 20:86–93.

Larget, B., and D.L. Simon. 1999. Markov chain monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16:750–759.

Lartillot, N. 2006. Conjugate Gibbs sampling for Bayesian phylogenetic models. *J. Comput. Biol.* 13:1701–1722.

Lartillot, N., H. Brinkmann, and H. Philippe. 2007. Suppression of long branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* 7:S4.

Lartillot, N., and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* 21:1095–1109.

Lartillot, N., and H. Philippe. 2006. Computing Bayes factors using thermodynamic integration. *Syst. Biol.* 55:195–207.

Lee, B., and M. Richards. 1971. The interpretation of protein structure: estimation of static accessibility. *J. Mol. Biol.* 55:379–400.

Lempers, F. B. 1971. *Posterior probabilities of alternative linear models*. Rotterdam University Press.

Lewis, S. M., and A. E. Raftery. 1997. Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *J. Am. Stat. Assoc.* 92:648–655.

Lindley, D. V. 1957. A statistical paradox. *Biometrika* 44:187–192.

Lindley, D. V. 1980. L. j. savage—his work on probability and statistics. *The Annals of Statistics* 8:1–24.

Liu, C., D. B. Rubin, and Y. N. Wu. 1998. Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* 85:755–770.

Maiorov, V., and G. Crippen. 1992. Contact potential that recognizes the correst folding of globular proteins. *J. Mol. Biol.* 227:876–888.

Malthus, T. R. 1798. *An essay on the principle of population, as it affects the future improvement of society with remarks on the speculations of Mr. godwin, M. condorcet, and other writers*. London: J. Johnson. Retrived at `www.esp.org/books/malthus/population/malthus.pdf`.

Mateiu, L., and B. Rannala. 2006. Inferring complex DNA substitution processes on phylogenies using uniformization and data augmentation. *Syst. Biol.* 55:259–269.

Mayrose, I., N. Friedman, and T. Pupko. 2005. A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics* 21-S2:ii151–ii158.

Metropolis, S., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculation by fast computing machines. *J. Chem. Phys.* 21:1087–1092.

Mintseris, J., K. Wiehe, B. Pierce, R. Anderson, R. Chen, J. Janin, and Z. Weng. 2005. Protein-protein docking benchmark 2.0: An update. *Proteins* 60:214–216.

Miyazawa, S., and R. L. Jernigan. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* 18:534–552.

Modiano, G., G. Battistuzzi, and A. G. Motulsky. 1981. Nonrandom patters of codon usage and nucleotide substitutions in human alpha- and beta-globin genes: an evolutionary strategy reducing the rate of mutations with drastic effects? *Proc. Natl. Acad. Sci. U.S.A.* 78:1110–1114.

Murray, I., Z. Ghahramani, and MacKay D. J. C. 2006. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conferences on Uncertainty in Artificial Intelligence*.

Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitutions, with applications to chloroplast genome. *Mol. Biol. Evol.* 11:715–724.

Neal, R. M. 1993. Probabilistic inference using Markov chain Monte Carlo methods. Technical report CRG-TR-93-1, University of Toronto.

Nielsen, R. 2002. Mapping mutations on phylogenies. *Syst. Biol.* 51:729–739.

Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936.

Nordenskiold, E. 1928. *The history of biology: a survey*. New York: Tudor Publishing Co. Translated by L. B. Eyre.

Nourani, Y., and B. Andresen. 1998. A comparison of simulated annealing cooling strategies. *J. Phys. A: Math. Gen.* 31:8373–8385.

Oldroyd, D. R. 1980. *Darwinian impacts: An introduction to the darwinian revolution*. Humanities press.

Paap, R. 2002. What are the advantages of MCMC based inference in latent variable models. *Stat. Neerl.* 56:2–22.

Pabo, C. 1983. Molecular technology: designing proteins and peptides. *Nature* 301:200.

Pagel, M., and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.* 53:561–581.

Pal, C., B. Papp, and M. J. Lercher. 2006. An integrated view of protein evolution. *Nat. Rev. Genet.* 7:337–348.

Parisi, G., and J. Echave. 2001. Structural constraints and emergence of sequence patterns in protein evolution. *Mol. Biol. Evol.* 18:750–756.

Pedersen, A.-M. K., and J. L. Jensen. 2001. A dependent rates model and MCMC based methodology for the maximum likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.* 18:763–776.

Philippe, H., F. Delsuc, H. Brinkmann, and N. Lartillot. 2005. Phylogenomics. *Annu. Rev. Ecol. Evol. Syst* 36:541–562.

Ponder, J. W., and F. M. Richards. 1987. Tertiary templates for proteins. use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* 193:775–791.

Posada, D., and T. R. Buckley. 2004. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Baysian approaches over likelihood ratio tests. *Syst. Biol.* 53:793–808.

Raftery, A. E. 1996. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* 83:251–266.

Ren, F., H. Tanaka, and Z. Yang. 2005. An empirical examination of the utility of codon substitution models in phylogeny reconstruction. *Syst. Biol.* 54:808–818.

Robert, C. P., and G. Casella. 2004. *Monte Carlo statistical methods*. Springer.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. *Mol. Biol. Evol.* 18:1692–1704.

Rubin, D. B. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* 4:1151–1172.

Sainudiin, R., W.S.W. Wong, K. Yogeeswaran, J. Nasrallah, Z. Yang, and R. Nielsen. 2005. Detecting site-specific physicochemical selective pressures: applications to the class-I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *J. Mol. Evol.* 60:315–326.

Schaldt, E., and K. Lange. 2002. Codon and rate variation models in molecular phylogeny. *Mol. Biol. Evol.* 19:1534–1549.

Seno, F., C. Micheletti, and A. Martian. 1998. Variational approach to protein design and extraction of interaction potentials. *Phys. Rev. Lett.* 81:2172–2175.

Shakhnovich, E. I., and A. M. Gutin. 1993. Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. U.S.A.* 90:7195–7199.

Siepel, A., and D. Haussler. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* 21:468–488.

Skolnick, J., L Jaroszewski, A. Kolinski, and A. Godzik. 1997. Derivation and testing of pair potentials for protein folding. When is the quasi-chemical approximation correct? *Protein Sci.* 6:676–688.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. *J. R. Statist. Soc. B* 36:111–147.

Sullivan, J., and P. Joyce. 2005. Model selection in phylogenetics. *Ann. Rev. Ecol. Evol. Syst.* 36:445–466.

Sun, S., R. Bren, R. Chan, and K. Dill. 1995. Designing amino acid sequences to fold with good hydrophobic cores. *Protein, Eng.* 8:1205–1213.

Susko, E., C. Field, C. Blouin, and A. J. Roger. 2003. Estimation of rates-across-sites distributions in phylogenetic substitution models. *Syst. Biol.* 52:625–636.

Telford, M. J., M. J. Wise, and Y. Gowri-Shankar. 2005. Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: examples from the bilateria. *Mol. Biol. Evol.* 22:1129–1136.

Thomas, P. D., and K. A. Dill. 1996. Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* 257:457–469.

Thorne, J. L. 2007. Protein evolution constraints and model-based techniques to study them. *Curr. Opin. Struct. Biol.* 17:337–341.

Thorne, J. L., S. C. Choi, J. Yu, P. G. Higgs, and H. Kishino. 2007. Population genetics without intraspecific data. *Mol. Biol. Evol.* 24:1667–1677.

Tiana, G., M. Colombo, D. Provasi, and Broglia R. A. 2004. Deriving amino acid contact potentials from their frequency of occurence in proteins: a lattice model study. *J. Phys. Condens. Matter* 16:2551–2564.

Tierney, L., and J. B. Kadane. 1986. Accurate approximations for posterior moments and marginal distributions. *J. Amer. Statist. Assoc.* 81:82–86.

Tobi, D., and R. Elber. 2000. Distance-dependent, pair potential for protein folding: Results from linear optimization. *Proteins* 41:40–46.

Wang, G., and R. L. J. Dunbrack. 2003. PISCES: a protein sequence culling server. *Bioinformatics* 19:1589–1591.

Watson, J., and F. Crick. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171:737–738.

Wei, G. C. G., and M. A. Tanner. 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Stat. Assoc.* 85:699–704.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* 18:691–699.

Wong, S. W., R. Sainudiin, and R. Nielsen. 2006. Identification of physicochemical selective pressure on protein encoding nucleotide sequences. *BMC Bioinformatics* 7:148.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10:1396–1401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39:306–14.

Yang, Z. 1996. Among site variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* 11:367–370.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in BioSciences* 13:555–556.

Yang, Z. 2006. *Computational molecular evolution*. Oxfors Series in Ecology and Evolution.

Yang, Z., R. Nielsen, N. Goldman, and A-M. K. Pedersen. 2000a. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.

Yang, Z., and B. Rannala. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst. Biol.* 54:455–470.

Yang, Z., W. J. Swanson, and V. D. Vacquier. 2000b. Maximum likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol. Biol. Evol.* 17:1446–1455.

Yu, J., and J. L. Thorne. 2006. Dependence among sites in RNA evolution. *Mol. Biol. Evol.* 23:1525–1537.

Zuckerkandl, E., and L. Pauling. 1962. Molecular disease, evolution, and genic heterogeneity. In *Horizons in biochemistry: Albert Szent-Györgyi dedicatory volume*, ed. M. Kashsa and B Pullman, 189–225. New York: Academic Press.

Zuckerkandl, E., and L. Pauling. 1965. Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins*, ed. V. Bryson and H. J. Vogel, 97–166. New York: Academic Press.

Appendix A: Data sets

MYO*60-153*

This is a set 60 amino acid sequences of mammalian myoglobin: *Orcinus orca* (P02173), *Delphinus delphis* (P02172), *Globicephala melas* (P02174), *Phocoenoides dalli* (P02176), *Inia geoffrensis* (P02181), *Balaenoptera acutorostrata* (P02179), *Balaenoptera physalus* (P02180), *Megaptera novaeangliae* (P02178), *Eschrichtius robustus* (P02177), *Physeter catodon* (P02185), *Kogia simus* (P02184), *Mesoplodon carlhubbsi* (P02183), *Ziphius cavirostris* (P02182), *Halichoerus grypus* (CAA23743.1), *Phoca sibirica* (P30562), *Bos taurus* (BAA00311.1), *Cervus elaphus* (P02191), *Ovis aries* (P02190), *Elephas maximus* (P02186), *Loxodonta africana* (P02187), *Lepilemur mustelinus* (P02169), *Equus burchelli* (P02188), *Oryctolagus cuniculus* (P02170), *Otolemur crassicaudatus* (P02168), *Nycticebus coucang* (P02167), *Perodicticus potto* (P02166), *Meles meles* (P02157), *Lycaon pictus* (P02159), *Otocyon megalotis* (P02158), *Vulpes chama* (P02160), *Zalophus californianus* (P02161), *Rattus norvegicus* (AAF05848.1), *Mus musculus* (CAA27994.1), *Spalax ehrenbergi* (P04248), *Ochotona princeps* (P02171), *Sus scrofa* (XM14433_AAA31073.1), *Tupaia glis* (P02165), *Orycteropus afer* (P02164), *Erinaceus europaeus* (P02156), *Ctenodactylus gundi* (P20856), *Proechimys guairae* (P04249), *Lagostomus maximus* (P04250), *Homo sapiens* (CAA25109.1), *Pan troglodytes* (P02145), *Pongo pygmaeus* (P02148), *Hylobates syndactylus* (P02146), *Gorilla gorilla* (P02147), *Presbytis entellus* (P02149), *Macaca fascicularis* (P02150), *Callithrix jacchus* (P02152), *Aotus trivirgatus* (P02151),

Lagothrix lagotricha (P02154), *Saimiri sciureus* (P02155), *Cebus apella* (P02153), *Rousettus aegyptiacus* (P02163), *Didelphis virginiana* (P02193), *Macropus rufus* (P02194), *Ornithorhynchus anatinus* (P02196), *Tachyglossus aculeatus* (P02195), *Lutra lutra* (P11343).

MYO*20-153*

This is a set 20 amino acid sequences of tetrapod myoglobin: *Balaenoptera physyalus* (P02180), *Physeter catodon* (P02185), *Ziphus cavirostris* (P02182), *Bos taurus* (BAA00311.1), *Halichoerus grypus* (CAA23743.1), *Zalophus californianus* (P02161), *Proechimys guairae* (P04249), *Ctenodactylus gundi* (P20856), *Mus musculus* (CAA27994.1), *Ochotona princeps* (P02171), *Pongo pygmaeus* (P02148), *Elephas maximus* (P02186), *Macropus rufus* (P02194), *Tachyglossus aculeatus* (P02195), *Varanus varius* (P02203), *Gallus gallus* (P416292), *Aptenodytes fosteri* (P02199), *Alligator mississippiensis* (P02200), *Caretta caretta* (P56208), *Graptemys geographica* (P02201).

MYO*10-153*

This is a set 10 amino acid sequences of mammalian myoglobin: *Physester catodon* (P02185), *Orcinus orca* (P02173), *Bos taurus* (BAA00311.1), *Rattus norvegicus* (AAF05848.1), *Mus musculus* (CAA27994.1), *Nannospalax ehrenbegi* (P04248), *Homo sapiens* (CAA25109), *Gorilla gorilla* (P02147), *Ornithorhynchus anatinus* (P02196), *Tachyglossus aculeatus* (P02195).

MYO*4-153*

This is a set 4 amino acid sequences of myoglobin: *Physester catodon* (P02185), *Orcinus orca* (P02173), *Chelonia mydas caranigra* (P56208), *Graptemys geographica* (P02201).

FBP*20-363*

This is a set of 20 amino acid sequences of vertebrate fructose bisphosphate aldolase: *Canis familiaris* (P536914), *Oryctolagus cuniculus* (P00883), *Mus musculus* (P05064), *Rattus norvegicus* (AAA40714), *Xenopus tropicalis* (NP001005643), *Xenopus tropicalis* (NP989131), *Xenopus laevis* (BAA19524), *Xenopus laevis* (AAH84132), *Danio rerio* (AAH65847), *Danio rerio* (AAH44379), *Lethenteron japonicum* (P53446), *Sparus aurata* (P53447), *Tetraodon nigroviridis* (CAG06274), *Gallus gallus* (AAA48587), *Mus musculus* (Q91Y97), *Rattus norvegicus* (AAH81697), *Oryctolagus cuniculus* (P79226), *Pongo pygmaeus* (CAI29598), *Homo sapiens* (P04075), *Macaca fascicularis* (BAB84033).

PPK*10-158*

This is a set 10 amino acid sequences of bacterial 6-hydroxymethyl-7-8-dihydroxypterin pyrophosphokinase: *Escherichia coli* (BAB96719), *Shigella flexneri* (AAP15678), *Salmonella typhimurium* (AAL19147), *Salmonella enterica* (AAO67923), *Photorhabdus luminescens* (CAE13168), *Yersinia pestis* (AAS60560), *Erwinia carotovora* (CAG76218), *Vibrio vulnificus* (BAC95526), *Vibrio cholerae* (AAF93760), *Photobacterium profundum* (CAG21480760).

Appendix B: Partition function formalism

Here, we apply the principles of cumulant development of the log of a partition function to derive the first and second moment identities, which are needed for the Monte Carlo approximations used in this dissertation.

Suppose some unnormalized density $f(\theta)$, formulated according to some high-dimensional parameterization $\theta \in \Theta$. The normalized probability density is given by

$$p(\theta) = \frac{1}{Z} f(\theta) \tag{B.1}$$

where

$$Z = \int_{\Theta} f(\theta) d\theta \tag{B.2}$$

is the normalizing factor, which ensures that the total probability equals 1.

The derivative of the logarithm of (B.2), with respect to a particular parameter θ_i

of the parameter vector, is developed as follows:

$$\frac{\partial \ln Z}{\partial \theta_i} = \frac{1}{Z}\frac{\partial Z}{\partial \theta_i} \tag{B.3}$$

$$= \frac{1}{Z}\frac{\partial}{\partial \theta_i}\int_{\Theta} f(\theta)d\theta \tag{B.4}$$

$$= \frac{1}{Z}\int_{\Theta} \frac{\partial f(\theta)}{\partial \theta_i}d\theta \tag{B.5}$$

$$= \int_{\Theta} \frac{1}{f(\theta)}\frac{\partial f(\theta)}{\partial \theta_i}\frac{f(\theta)}{Z}d\theta \tag{B.6}$$

$$= \int_{\Theta} \frac{\partial \ln f(\theta)}{\partial \theta_i}p(\theta)d\theta \tag{B.7}$$

$$= \langle \frac{\partial \ln f(\theta)}{\partial \theta_i} \rangle \tag{B.8}$$

where $\langle . \rangle$ stands for an expectation with respect to (B.1). We refer to (B.8) as the *first moment identity*. Following a similar derivation, the second derivative of the log of (B.2), is expressed as:

$$\frac{\partial^2 \ln Z}{\partial \theta_i \partial \theta_j} = \frac{\partial}{\partial \theta_i}\frac{1}{Z}\int_{\Theta} \frac{\partial f(\theta)}{\partial \theta_j}d\theta \tag{B.9}$$

$$= \left[\frac{\partial}{\partial \theta_i}\frac{1}{Z}\right]\int_{\Theta} \frac{\partial f(\theta)}{\partial \theta_j}d\theta + \frac{1}{Z}\frac{\partial}{\partial \theta_i}\int_{\Theta} \frac{\partial f(\theta)}{\partial \theta_j}d\theta \tag{B.10}$$

$$= \left[-\frac{1}{Z^2}\int_{\Theta} \frac{\partial f(\theta)}{\partial \theta_i}d\theta\right]\int_{\Theta} \frac{\partial f(\theta)}{\partial \theta_j}d\theta + \frac{1}{Z}\frac{\partial}{\partial \theta_i}\int_{\Theta} f(\theta)\frac{\partial \ln f(\theta)}{\partial \theta_j}d\theta \tag{B.11}$$

$$= -\left[\frac{1}{Z}\int_{\Theta} \frac{\partial f(\theta)}{\partial \theta_i}d\theta\right]\left[\frac{1}{Z}\int_{\Theta} \frac{\partial f(\theta)}{\partial \theta_j}d\theta\right] + \frac{1}{Z}\int_{\Theta} f(\theta)\frac{\partial \ln f(\theta)}{\partial \theta_i}\frac{\partial \ln f(\theta)}{\partial \theta_j}d\theta + \frac{1}{Z}\int_{\Theta} f(\theta)\frac{\partial^2 \ln f(\theta)}{\partial \theta_i \partial \theta_j}d\theta \tag{B.12}$$

$$= -\langle \frac{\partial \ln f(\theta)}{\partial \theta_i} \rangle \langle \frac{\partial \ln f(\theta)}{\partial \theta_j} \rangle + \langle \frac{\partial \ln f(\theta)}{\partial \theta_i}\frac{\partial \ln f(\theta)}{\partial \theta_j} \rangle + \langle \frac{\partial^2 \ln f(\theta)}{\partial \theta_i \partial \theta_j} \rangle \tag{B.13}$$

$$= \left[\langle \frac{\partial \ln f(\theta)}{\partial \theta_i}\frac{\partial \ln f(\theta)}{\partial \theta_j} \rangle - \langle \frac{\partial \ln f(\theta)}{\partial \theta_i} \rangle \langle \frac{\partial \ln f(\theta)}{\partial \theta_j} \rangle\right] + \langle \frac{\partial^2 \ln f(\theta)}{\partial \theta_i \partial \theta_j} \rangle \tag{B.14}$$

We refer to (B.14) as the *second moment identity*. Note that the terms within [.] of equation (B.14) correspond to the variance-covariance matrix.

Appendix C: Derivatives of the augmented log-likelihood

Here, we outline first and second derivatives needed for the thermodynamic integration along β, as well as for the Monte Carlo optimizations and Laplace approximations of chapter 6. We present the developments under the site-interdependent models (allowing for gamma-distributed rates) with the understanding that the equations can be easily factored out under models assuming independence.

For a data set D, composed of an alignment of P amino acid sequences, and given a tree topology and parameters θ, the demarginalized likelihood function is given as:

$$p(D, \phi \mid \theta, r) \;=\; p(s_0 \mid \theta) \prod_{j=1}^{2P-3} p(s_j, \phi_j \mid s_{j_{up}}, \theta, r), \tag{C.1}$$

where the dependence on M has been dropped out from the notation. Each factor in (C.1) is detailed here.

For a specific branch j, the *augmented transition probability* is given as

$$p(s_j, \phi_j \mid s_{j_{up}}, \theta, r) = \left(\prod_{k=1}^{z_j} R_{s_{jk-1}s_{jk}} r_{\sigma_{jk}} e^{-(t_{jk}-t_{jk-1})\Upsilon(s_{jk-1})} \right) \times\; e^{-(\lambda_j - t_{jz_j})\Upsilon(s_{jz_j})}, \tag{C.2}$$

where,

- s_j represents the sequence at node j (a node has the same index as the branch leading to it), and $s_{j_{up}}$ is the sequence at the node ancestral to j;
- ϕ_j represents the substitution mapping along branch j;
- λ_j is the length of branch j;
- z_j stands for the total number of substitution events along branch j;
- t_{jk} represents the timing of substitution event k on branch j;
- s_{jk-1} and s_{jk} represent the amino acid sequence states before and after substitution event k—the states before the first and after the last substitution leading to node j are equivalent to the states at the ends of the branch, written symbolically as $s_{j0} = s_{j_{up}}$ and $s_{jz_j} = s_j$;
- σ_{jk} is the site of substitution k along branch j;
- $\Upsilon(s_{jk-1}) = \sum_{i=1}^{N} \sum_{s_i'} R_{s_{jk-1}s_i'r_i}$ represents the *rate away* from state s_{jk-1}, with the inner sum being over the 19 sequence states that differ with s_{jk-1} at position i.

The stationary distribution of the Markov process, appearing in (C.1), is given by:

$$p(s_0 \mid \theta) = \frac{1}{Y} e^{-2\beta G(s_0,c)}, \tag{C.3}$$

where s_0 represents the sequence at the root node (labeled as node 0), and Y is the associated normalizing constant

$$Y = \sum_{s} e^{-2\beta G(s.c)}, \tag{C.4}$$

summing is over all 20^N sequences.

Let y be the dimension of θ, and let v and w index the entries in θ (i.e. $1 \leq v, w, \leq y$). The first derivative of the actual log-likelihood function is written as

$$\frac{\partial \ln p(D \mid \theta)}{\partial \theta_v} = \langle \frac{\partial \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_v} \rangle, \tag{C.5}$$

and the second derivative as

$$\begin{aligned} \frac{\partial^2 \ln p(D \mid \theta)}{\partial \theta_v \partial \theta_w} &= \langle \frac{\partial \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_v} \times \frac{\partial \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_w} \rangle \\ &- \langle \frac{\partial \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_v} \rangle \times \langle \frac{\partial \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_w} \rangle \\ &+ \langle \frac{\partial^2 \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_v \partial \theta_w} \rangle. \end{aligned} \tag{C.6}$$

Each of the expectations in (C.5) and (C.6) can be estimated based on samples drawn using the first two elements of the PX-DA module. The last term in (C.6) requires that we compute the second derivatives. The first derivative is a vector, written as

$$\frac{\partial \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta} = \begin{bmatrix} \frac{\partial \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_1} \\ \frac{\partial \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_2} \\ \vdots \\ \frac{\partial \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_y} \end{bmatrix}. \tag{C.7}$$

The second derivative therefore yields a matrix, where, for each entry in (C.7), the derivative is taken once again with respect to each parameter:

$$\frac{\partial^2 \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta^2} = \begin{bmatrix} \frac{\partial^2 \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_1^2} & \frac{\partial^2 \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_2 \partial \theta_1} & \cdots & \frac{\partial^2 \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_y \partial \theta_1} \\ \frac{\partial^2 \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_y \partial \theta_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_1 \partial \theta_y} & \frac{\partial^2 \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_2 \partial \theta_y} & \cdots & \frac{\partial^2 \ln p(D, \phi \mid \theta, r) p_\alpha(r)}{\partial \theta_y^2} \end{bmatrix} \tag{C.8}$$

Observing the structure of (C.7) and (C.8), we compute the necessary derivative combinations in the following subsections.

Computing $\frac{\partial}{\partial\beta}$

The first derivative of the augmented log-likelihood with respect to β involves two types of terms:

$$\frac{\partial \ln p(D,\phi\mid,\theta,r)}{\partial\beta} = \frac{\partial \ln p(s_0 \mid \theta)}{\partial\beta} + \sum_{j=1}^{2P-3} \frac{\partial \ln p(s_j,\phi_j \mid s_{j_{up}},\theta,r)}{\partial\beta}. \tag{C.9}$$

The first term in (C.9) is given by

$$\frac{\partial \ln p(s_0 \mid \theta)}{\partial\beta} = -\frac{\partial 2\beta G(s_0,c)}{\partial\beta} - \frac{\partial \ln Y}{\partial\beta} \tag{C.10}$$

$$= -2\left[G(s_0,c) - \langle G\rangle\right] \tag{C.11}$$

where $\langle . \rangle$ stands for an expectation with respect to (C.3). This expectation can be estimated based on a sample of sequences $(s^{(h)})_{1\leq h\leq K}$, drawn from (C.3) using the Gibbs sampling procedure described in Robinson et al. (2003):

$$\langle G\rangle = \sum_s G(s,c)p(s \mid \theta) \tag{C.12}$$

$$\simeq \frac{1}{K}\sum_{h=1}^{K} G(s^{(h)},c). \tag{C.13}$$

The second term in (C.9) is given as:

$$\frac{\partial \ln p(s_j,\phi_j \mid s_{j_{up}},\theta,r)}{\partial\beta} = \left(\sum_{k=1}^{z_j} \frac{\partial \ln R_{s_{jk-1}s_{jk}} r_{\sigma_{jk}}}{\partial\beta} - \frac{\partial\left(t_{jk}-t_{jk-1}\right)\Upsilon\left(s_{jk-1}\right)}{\partial\beta}\right) - \frac{\partial\left(\lambda_j - t_{jz_j}\right)\Upsilon\left(s_{jz_j}\right)}{\partial\beta}, \tag{C.14}$$

which can be calculated from

$$\frac{\partial \ln R_{ss'}}{\partial\beta} = G(s,c) - G(s',c) \tag{C.15}$$

and

$$\frac{\partial R_{ss'}}{\partial \beta} = [G(s,c) - G(s',c)]\, R_{ss'}. \tag{C.16}$$

Computing $\frac{\partial}{\partial \lambda_j}$

For computing derivatives with respect to branch lengths, it is more practical to re-parametrize (C.2) using the following change of variables:

$$p(s_j, u_j \mid s_{j_{up}}, \theta, M) = \frac{\partial \phi_j}{\partial u_j} p(s_j, \phi_j \mid s_{j_{up}}, \theta, M), \tag{C.17}$$

with $u_j = (u_{jk})_{k \leq z_j}$ defined as

$$u_{jk} = \frac{t_{jk}}{\lambda_j}. \tag{C.18}$$

In equation (C.17), the factor $\frac{\partial \phi_j}{\partial u_j}$ can be developed as

$$\frac{\partial \phi_j}{\partial u_j} = \begin{bmatrix} \frac{\partial t_{j1}}{\partial u_{j1}} & \frac{\partial t_{j2}}{\partial u_{j1}} & \cdots & \frac{\partial t_{jz_j}}{\partial u_{j1}} \\ \frac{\partial t_{j1}}{\partial u_{j2}} & \frac{\partial t_{j2}}{\partial u_{j2}} & \cdots & \frac{\partial t_{jz_j}}{\partial u_{j2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial t_{j1}}{\partial u_{jz_j}} & \frac{\partial t_{j2}}{\partial u_{jz_j}} & \cdots & \frac{\partial t_{jz_j}}{\partial u_{jz_j}} \end{bmatrix} = \begin{bmatrix} \lambda_j & 0 & \cdots & 0 \\ 0 & \lambda_j & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_j \end{bmatrix}, \tag{C.19}$$

such that an alternative to the augmented transition probability can be written as

$$p(s_j, u_j \mid s_{j_{up}}, \theta, r) = \lambda_j^{z_j} \left(\prod_{k=1}^{z_j} R_{s_{jk-1} s_{jk}} r_{\sigma_{jk}} e^{-\lambda_j (u_{jk} - u_{jk-1}) \Upsilon(s_{jk-1})} \right) \times e^{-\lambda_j (1 - u_{jz_j}) \Upsilon(s_{jz_j})}. \tag{C.20}$$

In logarithmic form, the derivative is thus given by

$$\frac{\partial \ln p(s_j, u_j \mid s_{j_{up}}, \theta, M)}{\partial \lambda_j} = \frac{\partial \ln \lambda_j^{z_j}}{\partial \lambda_j} - \left(\sum_{k=1}^{z_j} \frac{\partial \lambda_j (u_{jk} - u_{jk-1}) \Upsilon(s_{jk-1})}{\partial \lambda_j} \right) - \frac{\partial \lambda_j (1 - u_{jz_j}) \Upsilon(s_{jz_j})}{\partial \lambda_j}, \tag{C.21}$$

which can be evaluated based on

$$\frac{\partial \ln \lambda_j^{z_j}}{\partial \lambda_j} = \frac{z_j}{\lambda_j} \tag{C.22}$$

and

$$\frac{\partial \lambda_j (u_{jk} - u_{jk-1})}{\partial \lambda_j} = u_{jk} - u_{jk-1}. \tag{C.23}$$

Computing $\frac{\partial}{\partial \alpha}$

Computing the derivative with respect α only involves the prior on rates:

$$p_\alpha(r) = \left[\frac{\alpha^\alpha}{\Gamma(\alpha)} \right]^N \prod_{i=1}^{N} r_i^{\alpha-1} e^{-\alpha r_i}. \tag{C.24}$$

The derivative is thus given by

$$\frac{\partial \ln p_\alpha(r)}{\partial \alpha} = N \left[\ln(\alpha) + 1 - \Psi(\alpha) \right] + \sum_{i=1}^{N} \ln r_i - r_i, \tag{C.25}$$

where $\Psi(\alpha) = \frac{\partial}{\partial \alpha} \ln \Gamma(\alpha)$ is known as the *digamma* function, for which estimating routines are available (Galassi et al., 2003).

Computing $\frac{\partial^2}{\partial\beta^2}$

Computing this second derivative requires two terms. First, we have the following:

$$\frac{\partial^2 R_{ss'}}{\partial\beta^2} = \frac{\partial}{\partial\beta}\left[G(s,c) - G(s',c)\right] R_{ss'} \tag{C.26}$$

$$= \left[G(s,c) - G(s',c)\right]^2 R_{ss'}. \tag{C.27}$$

The next term involves the stationary probability:

$$\frac{\partial^2 \ln p(s_0 \mid \theta)}{\partial\beta^2} = -\frac{\partial^2 \ln Y}{\partial\beta^2} \tag{C.28}$$

$$= -2\left[\langle G^2\rangle - \langle G\rangle^2\right], \tag{C.29}$$

where the expectations can again be estimated using the Gibbs sampling method of Robinson et al. (2003), giving

$$\langle G^2\rangle \simeq \frac{1}{K}\sum_{h=1}^{K}\left[G(s^{(h)},c)\right]^2 \tag{C.30}$$

and

$$\langle G\rangle^2 \simeq \left[\frac{1}{K}\sum_{h=1}^{K} G(s^{(h)},c)\right]^2. \tag{C.31}$$

Computing $\frac{\partial^2}{\partial\beta\partial\lambda_j}$

This only requires terms already derived, as given in (C.16).

Computing $\frac{\partial^2}{\partial \lambda_j^2}$

Referring to (C.22), this second derivative requires the following term:

$$\frac{\partial^2 \ln \lambda^{z_j}}{\partial \lambda_j^2} = \frac{\partial}{\partial \lambda_j} \frac{z_j}{\lambda_j} \tag{C.32}$$

$$= -\frac{z_j}{\lambda_j^2}. \tag{C.33}$$

Computing $\frac{\partial^2}{\partial \alpha^2}$

For gamma distributed rates, we get:

$$\frac{\partial^2 \ln p_\alpha(r)}{\partial \alpha^2} = N \left[\frac{1}{\alpha} - \Psi'(\alpha) \right], \tag{C.34}$$

where $\Psi' = \frac{\partial^2}{\partial \alpha^2} \ln \Gamma(\alpha)$ can again be approximated using standard routines (Galassi et al., 2003).

Appendix D: Maximization step for the EM algorithm

In this appendix, we give details of the M-step for the MCEM algorithm used in chapter 6.

The M-step in the case of branch lengths is an example of the ideal case, where we have an analytical solution. Specifically, at iteration n of the MCEM algorithm, each branch length is updated as

$$\lambda_j^n = \frac{\langle z_j \rangle}{\langle \Lambda_j \rangle}, \tag{D.1}$$

where

$$\Lambda_j = (1 - u_{jz_j})\Upsilon(s_{jz_j}) + \sum_{k=1}^{z_j}(u_{jk} - u_{jk-1})\Upsilon(s_{jk-1}). \tag{D.2}$$

Writing $(z_j^{(h)})_{1\leq h\leq K}$ for the number of substitutions along branch j of draw h, and $(u_{jk}^{(h)})_{1\leq h\leq K}$ for the re-parameterized configuration of each mapping, the needed expectations are estimated as

$$\langle z_j \rangle \simeq \frac{1}{K}\sum_{h=1}^{K} z_j^{(h)} \tag{D.3}$$

and

$$\langle \Lambda_j \rangle \simeq \frac{1}{K} \sum_{h=1}^{K} \left[(1 - u_{jz_j}^{(h)}) \Upsilon(s_{jz_j}) + \sum_{k=1}^{z_j^{(h)}} (u_{jk}^{(h)} - u_{jk-1}^{(h)}) \Upsilon(s_{jk-1}) \right]. \tag{D.4}$$

The M-step for optimizing α and β, however, is not direct, since solving for these parameters is not possible. Nonetheless, this inner maximization can be readily done using a gradient scheme similar to that described in the main text; using the same sample, gradient steps are performed repeatedly until the maximum is reached, following which a new sample is drawn for the next MCEM cycle, and so on. In the case of α, we also tried maximization using a Newton-Raphson-like method; the M-step is accomplished through an iterative updating, with cycle m given by

$$\alpha^m = \alpha^{m-1} - \mho(\alpha^{m-1}), \tag{D.5}$$

where

$$\mho(\alpha^m) = \frac{\frac{1}{K} \sum_{h=1}^{K} \frac{\partial}{\partial \alpha} \ln p(D \mid \theta^m, r^{(h)}, M) p_{\alpha^m}(r^{(h)}))}{\frac{1}{K} \sum_{h=1}^{K} \frac{\partial^2}{\partial \alpha^2} \ln p(D \mid \theta^m, r^{(h)}, M) p_{\alpha^m}(r^{(h)})}. \tag{D.6}$$

Appendix E: Codon model specifications

In our implementation the entries of Q are based on two sets of specifications: a 61 dimensional vector of *stationary probabilities*, π, and a set of *transient specification* ρ according to

$$Q_{ab} \propto \rho_{ab}\pi_b, a \neq b \tag{E.1}$$

$$Q_{aa} = -\sum_{b \neq a} Q_{ab}. \tag{E.2}$$

In this appendix, we write out in full the stationary probabilities under the codon models, as well as the full transient specifications, and give an example of the detailed balance check.

Stationary probabilities

First, expanding (7.8) for the stationary distribution under GY-F1×4, we have

$$\pi_a = \frac{\varphi_{a_1}\varphi_{a_2}\varphi_{a_3}}{\sum_{b=1}^{61}\varphi_{b_1}\varphi_{b_2}\varphi_{b_3}}. \tag{E.3}$$

Similarly, with GY-F3×4, we have

$$\pi_a = \frac{\varphi^{(1)}_{a_1}\varphi^{(2)}_{a_2}\varphi^{(3)}_{a_3}}{\sum_{b=1}^{61}\varphi^{(1)}_{b_1}\varphi^{(2)}_{b_2}\varphi^{(3)}_{b_3}}. \tag{E.4}$$

The stationary probability under GY-F61 is already entirely specified and the models MG-F1×4 and MG-F3×4 have the same stationary distributions as (E.3) and (E.4) respectively.

Under the MG-F1×4-CP model, the stationary probability is given by

$$\pi_a = \frac{\varphi_{a_1}\varphi_{a_2}\varphi_{a_3}\psi_a}{\sum_{b=1}^{61}\varphi_{b_1}\varphi_{b_2}\varphi_{b_3}\psi_b}, \tag{E.5}$$

and under the MG-F1×4-AAP model by

$$\pi_a = \frac{\varphi_{a_1}\varphi_{a_2}\varphi_{a_3}\varpi_{f(a)}}{\sum_{b=1}^{61}\varphi_{b_1}\varphi_{b_2}\varphi_{b_3}\varpi_{f(b)}}. \tag{E.6}$$

The stationary distributions under the MG-F3×4-CP and MG-F3×4-AAP models follow directly as

$$\pi_a = \frac{\varphi^{(1)}_{a_1}\varphi^{(2)}_{a_2}\varphi^{(3)}_{a_3}\psi_a}{\sum_{b=1}^{61}\varphi^{(1)}_{b_1}\varphi^{(2)}_{b_2}\varphi^{(3)}_{b_3}\psi_b}, \tag{E.7}$$

and

$$\pi_a = \frac{\varphi^{(1)}_{a_1}\varphi^{(2)}_{a_2}\varphi^{(3)}_{a_3}\varpi_{f(a)}}{\sum_{b=1}^{61}\varphi^{(1)}_{b_1}\varphi^{(2)}_{b_2}\varphi^{(3)}_{b_3}\varpi_{f(b)}} \tag{E.8}$$

respectively.

Under the SC-type models utilizing only the site-independent components of the statistical potential, the stationary probability is a site-specific vector written as $\pi^{(i)}$.

and under a F1×4-type models is given by

$$\pi_a^{(i)} = \frac{\varphi_{a_1}\varphi_{a_2}\varphi_{a_3}e^{-2\beta G_i(a)}}{\sum_{b=1}^{61}\varphi_{b_1}\varphi_{b_2}\varphi_{b_3}e^{-2\beta G_i(b)}}. \tag{E.9}$$

Transient specifications

In the case of GY-type models, the transient specification is simply (2.10) without π_j the factor. In the case of the MG-F1×4 model, we have

$$\rho_{ab} = \begin{cases} \frac{\varrho_{a_{c'}b_{c'}}}{\varphi_{b_{c'}}\varphi_{b_{c''}}}Z, & \text{if } \mathcal{A}, \\ \frac{\omega\varrho_{a_{c'}b_{c'}}}{\varphi_{b_{c'}}\varphi_{b_{c''}}}Z, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{E.10}$$

where c' and c'' are the two constant codon positions, and Z is the normalizing factor of the stationary distribution (in this case $Z = \sum_{b=1}^{61}\varphi_{b_1}\varphi_{b_2}\varphi_{b_3}$). Note that this latter Z factor is not needed when scaling Q. Once again, substituting φ_{b_c} with $\varphi_{b_c}^{(c)}$, and the appropriate Z, yields the transient specification for MG-F3×4.

For the MG-F1×4-CP model, the transient specification is given by

$$\rho_{ab} = \begin{cases} \frac{\varrho_{a_{c'}b_{c'}}}{\varphi_{b_{c'}}\varphi_{b_{c''}}\sqrt{\psi_a\psi_b}}Z, & \text{if } \mathcal{A}, \\ \frac{\omega\varrho_{a_{c'}b_{c'}}}{\varphi_{b_{c'}}\varphi_{b_{c''}}\sqrt{\psi_a\psi_b}}Z, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{E.11}$$

and the specification of MG-F1×4-AAP by:

$$\rho_{ab} = \begin{cases} \frac{\varrho_{a_{c'}b_{c'}}}{\varphi_{b_{c'}}\varphi_{b_{c''}}\varpi_{f(b)}}Z, & \text{if } \mathcal{A}, \\ \frac{\omega\varrho_{a_{c'}b_{c'}}}{\varphi_{b_{c'}}\varphi_{b_{c''}}\sqrt{\varpi_{f(a)}\varpi_{f(b)}}}Z, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise.} \end{cases} \tag{E.12}$$

As always, substituting φ_{b_c} with $\varphi_{b_c}^{(c)}$, and the appropriate Z, yields the transient specifications for the F3×4 versions of (E.11) and (E.12).

For SC-type models, for instance under the MG-F1×4 model, the transient specification is given as

$$\rho_{ab}^{(i)} = \begin{cases} \frac{\varrho_{a_c b_c}}{\varphi_{a_{c'}} \varphi_{a_{c''}} e^{-2\beta G_i(a)}}, & \text{if } \mathcal{A}, \\ \frac{\omega \varrho_{a_c b_c}}{\varphi_{a_{c'}} \varphi_{a_{c''}} e^{-\beta G_i(a)} e^{-\beta G_i(b)}}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise.} \end{cases} \tag{E.13}$$

We have now fully specified π and ρ used in equation (E.1). We can see that upon substituting stationary and transient specifications appropriately into (E.1), the models defined in the main body of the text are obtained. For instance, for a nonsynonymous substitution under the MG-F1×4-CP model, we have

$$\rho_{ab}\pi_b = \frac{\omega \varrho_{a_c b_c}}{\varphi_{b_{c'}} \varphi_{b_{c''}} \sqrt{\psi_a \psi_b}} Z \times \frac{\varphi_{b_1} \varphi_{b_2} \varphi_{b_3} \psi_b}{Z} \tag{E.14}$$

$$= \frac{\omega \varrho_{a_c b_c} \varphi_{b_c} \psi_b}{\sqrt{\psi_a \psi_b}} \tag{E.15}$$

$$= \frac{\omega \varrho_{a_c b_c} \varphi_{b_c} \sqrt{\psi_b} \sqrt{\psi_b}}{\sqrt{\psi_a} \sqrt{\psi_b}} \tag{E.16}$$

$$= \omega \varrho_{a_c b_c} \varphi_{b_c} \left(\frac{\psi_b}{\psi_a} \right)^{\frac{1}{2}}, \tag{E.17}$$

corresponding to the entry obtained from (7.2).

Checking the detailed balance

The models studied here all satisfy the equality $\pi Q = 0$, and are time-reversible, satisfying the equality $Q_{ab}\pi_a = Q_{ba}\pi_b$. These developments are lengthy, and so we display only one example, for the detailed balance check under MG-F1×4-CP in the case where

a and b differ a one nucleotide position and implying a nonsynonymous substitution:

$$Q_{ab}\pi_a = Q_{ba}\pi_b \tag{E.18}$$

$$\frac{\omega\varrho_{a_cb_c}\varphi_{b_1}\varphi_{b_2}\varphi_{b_3}\psi_b}{\varphi_{b_{c'}}\varphi_{b_{c''}}\sqrt{\psi_a\psi_b}}\varphi_{a_1}\varphi_{a_2}\varphi_{a_3}\psi_a = \frac{\omega\varrho_{b_ca_c}\varphi_{a_1}\varphi_{a_2}\varphi_{a_3}\psi_a}{\varphi_{a_{c'}}\varphi_{a_{c''}}\sqrt{\psi_b\psi_a}}\varphi_{b_1}\varphi_{b_2}\varphi_{b_3}\psi_b \tag{E.19}$$

$$\frac{\omega\varrho_{a_cb_c}\cancel{\varphi_{b_1}\varphi_{b_2}\varphi_{b_3}\psi_b}}{\varphi_{b_{c'}}\varphi_{b_{c''}}\sqrt{\psi_a\psi_b}}\varphi_{a_1}\varphi_{a_2}\varphi_{a_3}\psi_a = \frac{\omega\varrho_{b_ca_c}\varphi_{a_1}\varphi_{a_2}\varphi_{a_3}\psi_a}{\varphi_{a_{c'}}\varphi_{a_{c''}}\sqrt{\psi_b\psi_a}}\cancel{\varphi_{b_1}\varphi_{b_2}\varphi_{b_3}\psi_b} \tag{E.20}$$

$$\frac{\omega\varrho_{a_cb_c}}{\varphi_{b_{c'}}\varphi_{b_{c''}}\sqrt{\psi_a\psi_b}}\cancel{\varphi_{a_1}\varphi_{a_2}\varphi_{a_3}\psi_a} = \frac{\omega\varrho_{b_ca_c}\cancel{\varphi_{a_1}\varphi_{a_2}\varphi_{a_3}\psi_a}}{\varphi_{a_{c'}}\varphi_{a_{c''}}\sqrt{\psi_b\psi_a}} \tag{E.21}$$

$$\frac{\omega\varrho_{a_cb_c}}{\cancel{\varphi_{b_{c'}}\varphi_{b_{c''}}\sqrt{\psi_a\psi_b}}} = \frac{\omega\varrho_{b_ca_c}}{\cancel{\varphi_{a_{c'}}\varphi_{a_{c''}}\sqrt{\psi_b\psi_a}}} \tag{E.22}$$

$$\cancel{\omega}\varrho_{a_cb_c} = \cancel{\omega}\varrho_{b_ca_c} \tag{E.23}$$

$$\varrho_{a_cb_c} = \varrho_{b_ca_c}, \tag{E.24}$$

where the array ϱ is symmetrical, satisfying the equality.

Finally, we mention here that we follow the practice proposed by Huelsenbeck et al. (2006), and (under non-structural models only) scale Q matrices such that branch length represent the expected number of synonymous substitutions per codon site, although we have also tried the model comparisons without any scaling of Q (such that branch lengths have no meaningful units) and obtained essentially identical results (not shown).

Appendix F: Implementation

The following is meant to give some entry points for using the version of the PhyloBayes package in which the developments of the present thesis were implemented. Through examples, we describe how to run several MCMC-based calculations, as well a performing posterior predictive model checking. One should keep in mind that the program is still very much an experimental tool, and that developments and modifications are continuously being made (almost daily).

Running the PhyloBayes package

Overview

PhyloBayes was developed by Nicolas Lartillot (Nico) to provide a flexible set of tools for implementing and comparing various models of amino acid replacement (Lartillot and Philippe, 2004). The package described here is an offshoot from Nico's version from around December 2003. The program has been modified and adapted in numerous ways, to handle site-interdependent models based on statistical potentials, to perform several types of thermodynamic integration, as well as maximum likelihood estimation, and Laplace approximations of marginal likelihoods. In addition, modifications have been made in order to handle nucleotide models, as well as dozens of codon models, also with specialized thermodynamic integrations. How the program threads through these

various options is controlled by a initialization file (contract to 'initfile'). Examples will follow, but first, we describe the typical work setup.

Requirements and work setup

PhyloBayes was developed in the C++ programming language, on Linux systems. To compile the program, you will need the freely available GNU g++ compiler (usually installed by default on most work stations). Also, you will need to install the GNU scientific library, which can be downloaded from http://www.gnu.org/software/gsl/.

Usually, we make a directory called 'phylobayes', which contains two sub-directories called 'data' and 'sources'. The 'sources' directory contains a makefile that will compile the programs (actually, there are several make files), placing them in the 'data' directory. Then, we make additional sub-directories within 'data', each of which contains an alignment file, and any other files that may be needed. For instance, suppose you have a dataset called `myo20`. All files pertaining to this dataset would be found in ~/phylobayes/data/myo20. From here, the programs are called one repertoire up. For example,

```
$ ../newchain <initfile> <chainname>
```

will initialize a particular calculation, whereas

```
$ ../phylobayes <chainname>
```

will launch the sampler. Other programs (`monitor`, `diagnostics`, `readthermo`,...) read and process the sample in various ways, and are called similarly.

Of course, you may prefer to arrange things otherwise, but we will assume this set-up in the following descriptions.

'Classical' Bayesian MCMC sampling under amino acid models

Although Nico has already explained the usage of PhyloBayes for phylogenetic Bayesian MCMC sampling in other texts, given the many versions now going around, it might be best to outline this again. Here is an example initfile:

```
DataFile          myo20.nex

RatePrior         GammaInv
LengthPrior       Exponential

ModeFastCompute No

MoveType          AllBranchLength          5       0.1     1
MoveType          AllBranchLength          5       0.5     1
MoveType          OneBranchLength          5       1       1
MoveType          OneBranchLength          5       1.5     1
MoveType          MeanBranchLengthMove     1       0.5     1
MoveType          Gamma                    1       1       1
MoveType          Gamma                    1       0.1     1
MoveType          Rate                     1       0.5     1
MoveType          Rate                     1       1       1
MoveType          Rate                     1       1.5     1

End

SaveEvery         50
StopAfter         -1

InitState
Tree     (PONPY:0,(((MOUSE:0.1214,OCHPR:0.0138):0.0234,(PROGU...

Nmode 1
ModeStationaries          WAG
ModeRR                    WAG

Rates                     Uniform

//
```

This will perform a sampling from the posterior distribution, integrating over branch lengths and site-specific substitution rates, under exponential and gamma laws respec-

tively, and using the WAG amino acid replacement matrix. The first line specifies the name of the file containing the alignment (`myo20.nex`). For now, this file must be in nexus format. The next two lines specify the prior laws to by applied for rates and branch lengths. Next is indicated `ModeFastCompute No`. This will bypass a recoding scheme used to accelerate likelihood calculations [explained in Lartillot and Philippe (2004)], but which is only applicable under a POISSON-based (site-independent) model.

The next series of lines specifies the set Metropolis-Hastings (MH) update operators to be applied per cycle, with their respective tunings. The three columns of numbers to the right correspond to the call frequency per cycle, the MH tuning, and, if the operator performs some type of multidimensional update, the order of the sub-space to be considered (usually put to 1 if the update is not multidimensional). More specifically, the first two operators listed, `AllBranchLength`, are applied 5 times each, but with 2 distinct tuning parameters (0.1 and 0.5). Here, the higher the tuning, the bolder the update attempt. The move `OneBranchLength` works similarly. The next operator, `MeanBranchLengthMove`, applies a MH update to the hyperparameter governing the prior law on branch lengths. Two MH operators are applied to the 'shape' parameter α for the prior law on rates[1]. Finally, three update operators are applied to the site-specific rates themselves. Note that, in this case, a call to the `Rate` operator in fact loops over all sites, performing a distinct MH update to each rate.

The `SaveEvery` line specifies how many cycles are performed between each draw saved, whereas `StopAfter` defines the total number of points you want the chain to draw; when set to -1, the chain will run indefinitely.

The keyword `InitState` indicates that the following lines will define the starting configuration of the chain. The tree topology is indicated here, and is always kept fixed for now. The next line, specifying `Nmode`, pertains to the CAT model, and should be

[1]In PhyloBayes, α is referred to as 'gamma'; the name 'alpha' is used for another type of model (CAT). In general, you should not expect names in the program to necessarily correspond to the names or symbols given in articles or other texts.

set to 1 for the classical single matrix models. The equilibrium frequencies and amino acid exchangeability parameters are specified in the next two lines; given the list of operators, these are kept fixed in this case. Finally, the site-specific rates will start out all being uniform (equal to 1).

From this initfile, it would be easy to expand or contract the model in various ways. For instance, we could add the following update operators to sample under a GTR+Γ model:

```
MoveType        ModeStationary          10      1000    5
MoveType        ModeRelativeRate        20      1000    10
```

This would include ten update attempts to amino acid equilibrium frequencies (`ModeStationary`), each update applied to 5 out of 20 (picked at random) stationary probabilities, and twenty update attempts to amino acid exchangeability parameters (`ModeRelativeRate`), each randomly picking 10 out of the set of 190. Note that both of these operators are Dirichlet type moves [see Larget and Simon (1999)]. Their MH tuning parameters work differently; in this case, the smaller the tuning the bolder the update attempt.

Alternatively, we could contract the model to a uniform rates across sites model, by simply removing update operators `Rate`, as well as `Gamma`. Other possible configurations could include setting `ModeStationaries Uniform` and `ModeRR Poisson`; in the case of the latter, you could set `ModeFastCompute Yes` to take advantage of Nico's recoding system, which can substantially increase computational performance. Overall, by playing with these different update operators and `InitState` settings, we now have the means of sampling under several common amino acid replacement models (Poisson, Poisson+F, Poisson+Γ, Poisson+F+Γ, WAG, WAG+F, etc.).

For sampling under nucleotide data, the initfile would look much the same. Only now, `ModeRR` should be initialized to `Poisson`, with calls to at least `ModeStationary` (F81), but preferably with calls to `ModeRelativeRate` as well (GTR).

'Classical' Bayesian MCMC sampling under codon models

When using codon models, the alignment should be checked carefully, to start at a clear codon break, and to be of length that is some multiple of three. Also, only nexus formats are available, which must contain the entry `datatype=codon`. The following initfile will run an MCMC to sample under the GY model specified in chapter 2, with the Dirichlet process on nonsynonymous rate factors. Albeit here, we are using a must more efficient data-augmentation-based sampling approach.

```
DataFile         bglobin.nex

Normalise        Yes
SynOnly          Yes
Uniformization   Yes
OmegaPrior       DirichletProcess
OmegaBasePrior   PairRatioOfExpOneRVs
AlphaPrior       Exponential
LengthPrior      Exponential

ModeFastCompute  No
RefFastCompute   No

MoveType         AllBranchLength          50      1.0     1
MoveType         AllBranchLength          50      0.5     1
MoveType         MeanBranchLengthMove     50      0.5     1
MoveType         MeanBranchLengthMove     50      0.1     1
MoveType         MeanBranchLengthMove     50      1.0     1
MoveType         CodonStat                50      1000    10
MoveType         CodonStat                50      5000    10
MoveType         Kappa                    25      0.1     1
MoveType         Kappa                    25      0.5     1
MoveType         Omega                    10      1.0     1
MoveType         Omega                    10      0.5     1
MoveType         Omega                    15      0.25    1
MoveType         Alpha                    50      0.5     1
MoveType         Alpha                    50      1.0     1
MoveType         Alpha                    50      1.5     1
MoveType         SwitchOmega              10      1       5
MoveType         ResampleMapping          1       1       1

End
```

```
CodonStatModelType      GY_F61
CodonRRModelType        HKY

withMappings    Yes

SaveEvery       10
StopAfter       -1

InitState
Tree    (xenlaev:0.51040247,xentrop:0.78663390,...

//
```

As before, the top line specifies the data set. The following two lines will scale all matrices as described in Huelsenbeck et al. (2006). We also indicate the use of the uniformization technique for drawing substitution mappings. The prior structure described in chapter 2 is specified in the next lines, and the sampling is reasonably self evident. Note the operator `ResampleMapping`, which draws a new mapping after the round of updates. To revert to pruning-based sampling, simply remove the operator, as well as `withMappings Yes`. To generalize the model slightly, as in chapter 7, replace operators `Kappa` with

```
MoveType        NucleotideRelRate               25      1000    4
MoveType        NucleotideRelRate               25      500     4
```

and set `CodonRRModelType GTR`. The model can also be contracted by setting `OmegaPrior Flat`, replacing the `Omega` operators with `GlobalOmega`, and removing operators `Alpha` and `SwitchOmega`.

Site-interdependent Bayesian MCMC sampling for amino acid models

Sampling under site-interdependent amino acid models involves several differences and additions to the initfile. Here is an example:

```
DataFile        myo20.nex

RatePrior       GammaInv
LengthPrior     Exponential

ModeFastCompute No
RefFastCompute No

MoveType        AllBranchLength             1       1       1
MoveType        AllBranchLength             1       2       1
MoveType        MeanBranchLengthMove        1       0.5     1
MoveType        MeanBranchLengthMove        1       1       1
MoveType        NodeSiteStateOverTreeMove   5       1       10
MoveType        NodeSiteStateMove           5       1       50
MoveType        PathSiteMove                5       1       100

End

withDependence          Yes
contactMap              1MBD.mj.cm
solventAccess           1MBD.mj.av
potential               homeMadeMJstyleWithAV
chemicalPotentials      Yes

MValue                  500
gibbsIterBtwSeqs        5
thetaStarThreshold      0.01

SaveEvery               10
StopAfter               -1

InitState
Tree    (PONPY:0,(((MOUSE:0.1214,OCHPR:0.0138):0.0234,(PROGU...

Nmode 1
ModeStationaries        Empirical
ModeRR                  Poisson

RefStationaries         Uniform
RefRR                   Poisson

Rates                   Uniform

pFactor                 0.5
```

```
//
```

First, note that the tunings of `AllBranchLength` have changed; under the conditions of the model, this operator name actually calls a different update operator, working on the basis of the data augmentation (mapping) scheme. The last three MH operators are each called five times, proposing updates to substitution mappings of a sub-set of positions: `NodeSiteStateOverTreeMove` proposes a mapping for ten (randomly picked) sites over the entire tree; `NodeSiteStateMove` proposes a mapping for fifty sites over three branches connected to a (randomly picked) internal node; `PathSiteMove` proposes a mapping to one hundred positions over a single branch.

The next lines engage the site-interdependent calculations (`withDependence Yes`), give the protein structure files (`1MBD.mj.*`), and define the potential used `homeMadeMJstyleWithAV` with chemical potentials, from chapter 5.

Stationary probabilities under the model involve an additional MCMC sampler (sometimes buried within the main chain), which draws amino acid sequences using the familiar Gibbs method. The number of sequences drawn is set using `MValue`, whereas the number of sequence-sweeping Gibbs cycles between draws is set using `gibbsIterBtwSeqs`. A cut off for an importance sampling procedure is set using `thetaStarThreshold`.

It is important to note the settings for `ModeStationaries`, `ModeRR`, `RefStationaries`, and `RefRR`. The `Mode` settings correspond to the model proposing substitution mappings, whereas `Ref` settings have a bearing on the target model. Here, we are sampling under the pure potential, and so we set `Ref` to a completely flat configuration.

Also note that `pFactor`, or β, is set to 0.5 to have the proper scaling of the energy function. However, we can give this some flexibility as well, by including the operator

```
MoveType        pfactorTypeB    1       0.1     1
```

Other models can also be combined, for instance

```
MoveType        Gamma           1       1       1
MoveType        Gamma           1       0.1     1
MoveType        Rate            20      1       10
```

will sample under a $+\Gamma$ model. Here again, the `Rate` operator actually calls a different operator than would be called under site independence; in this case, it is better to perform update attempts for several rates at once (typically 10 to 50).

Site-interdependent Bayesian MCMC sampling for codon models

The following initfile will sample from the MG-F1×4-DP-SC model:

```
DataFile        bglobin.nex

Normalise       No
SynOnly         No
OmegaPrior      DirichletProcess
OmegaBasePrior  PairRatioOfExpOneRVs
LengthPrior     Exponential
AlphaPrior      Exponential

ModeFastCompute No
RefFastCompute No

MoveType        AllBranchLength             100     1.0     1
MoveType        AllBranchLength             100     0.5     1
MoveType        MeanBranchLengthMove        50      0.5     1
MoveType        MeanBranchLengthMove        50      0.1     1
MoveType        MeanBranchLengthMove        50      1       1
MoveType        NucleotideStat              10      1000    2
MoveType        NucleotideStat              20      2000    2
MoveType        NucleotideRelRate           10      1000    4
MoveType        NucleotideRelRate           10      500     4
MoveType        Omega                       5 1.0     1
MoveType        Omega                       5       0.5     1
MoveType        Omega                       5       0.25    1
```

```
MoveType        Alpha                       50      0.5     1
MoveType        Alpha                       50      1.0     1
MoveType        Alpha                       50      1.5     1
MoveType        SwitchOmega                 1       1       5
MoveType        NodeSiteStateOverTreeMove   10      1       10
MoveType        NodeSiteStateMove           50      1       30
MoveType        PathSiteMove                50      1       50

End

CodonStatModelType      MG_F1X4
CodonRRModelType        GTR

withStructure           Yes
withDependence          Yes
potential               homeMadeMJstyleWithAV
chemicalPotentials      Yes
solventAccess           4HHBB.av
contactMap              4HHBB.cm.temp

gibbsIterBtwSeqs        5

SaveEvery               1
StopAfter               -1

InitState
Tree    (xenlaev:0.51040247,xentrop:0.78663390...

//
```

The model can be contracted as before.

Maximum likelihood parameter estimation

We focus here on Monte Carlo EM optimization, beginning with the WAG+Γ model.

```
DataFile        myo20.nex

RatePrior       GammaInv
LengthPrior     Exponential

ModeFastCompute No
RefFastCompute No
```

```
MoveType        Rate          10      1       1

End

MLmode
EMmode
decorrelate             1
reburn                  1
gradEstimateBasedOn     100

branchLengthMLtuning    1
gammaMLtuning           1

SaveEvery       1
StopAfter       -1

InitState
Tree    (PONPY:0,(((MOUSE:0.164107,OCHPR:0.0224974):0.0236,(PROGU...

Nmode 1
ModeStationaries        WAG
ModeRR          WAG

RefStationaries WAG
RefRR           WAG

Rates           Uniform

//
```

`MLmode` engages the chain through optimization cycles. The default optimization is a gradient scheme, but the EM scheme is engaged with the keyword `EMmode`. Although they are set to 1 here, `decorrelate` and `reburn` give some flexibility when making draws for each EM cycle. Here, `branchLengthMLtuning` and `gammaMLtuning` act as switches indicating that the associated components are to be optimized; in the case of gradient optimization, these are actually the step parameters (which must be tuned, hence the name). Note that while the algorithm relies on a sample of substitution mappings, under site-independence these can be drawn directly from their posterior distribution; no MH

updates over mappings are needed (but you can perform mapping-based sampling if you want...).

EM optimization under site interdependence is very similar.

```
DataFile        myo20.nex

RatePrior       GammaInv
LengthPrior     Exponential

ModeFastCompute No
RefFastCompute No

MoveType        NodeSiteStateOverTreeMove       10      1       10
MoveType        NodeSiteStateMove               25      1       50
MoveType        PathSiteMove                    25      1       100

End

withDependence          Yes
contactMap              1MBD.mj.cm
solventAccess           1MBD.mj.av
potential               homeMadeMJstyleWithAV
chemicalPotentials      Yes

MValue  500
thetaStarThreshold      0.01
gibbsIterBtwSeqs        5

MLmode
EMmode
decorrelate             1
reburn                  1
gradEstimateBasedOn     100

branchLengthMLtuning    1

SaveEvery       1
StopAfter       -1

InitState
Tree    (PONPY:0,(((MOUSE:0.2277,OCHPR:0.0283):0.0336,(PROGU...

Nmode 1
```

```
ModeStationaries        Empirical
ModeRR          Poisson

RefStationaries Uniform
RefRR           Poisson

Rates   Uniform

pFactor 0.5

//
```

This initfile will optimize branch lengths under the structural model. To optimize pFactor (β), add `pFactorMLtuning  0.0005` as well. Note that here, each cycle performs a gradient step to pFactor, and an EM step to branch lengths; this combination for optimizing pFactor was found to work best under these models. You can also combine an optimization of the shape parameter under gamma distributed rates, by simply setting `gammaMLtuning 1` and including calls to `Rate` operators.

In fact, by playing with these settings, you can imagine many ways of marginalizing over some parameters, while optimizing over others...

Thermodynamic integration for amino acid models

There are a few different reasons for using thermodynamic integration with the programs. From the WAG+Γ optimization, for example, we adjusted the parameters so as to maximize the log-likelihood, but we have not yet computed the log-likelihood itself. The following initfile will initialize a chain to make this calculation.

```
DataFile        myo20.nex

RatePrior       GammaInv
LengthPrior     Exponential

ModeFastCompute No
RefFastCompute No
```

```
QuasiStatic 0 1 0.001 10
MSMode RAS

MoveType        Rate            10      1       1

End

SaveEvery       10
StopAfter       -1

InitState
Tree    (PONPY:0,(((MOUSE:0.1265,OCHPR:0.0198):0.017,(PROGU...

Nmode 1
ModeStationaries        WAG
ModeRR                  WAG

RefStationaries         WAG
RefRR                   WAG

Rates           Uniform

Gamma   0.73

//
```

The first key line here is `QuasiStatic`. This will run the MCMC sampler across a path linking two models. The `0 1` that follow indicate the starting and ending values of the temperature or "morphing" parameter; in this case, at the beginning of the run, the sampling is with respect to the WAG model with uniform rates, and gradually switches (by steps of 0.001) to the gamma distributed rates model. The last value of this line defines a burnin; here, the sampling is equilibrated drawing ten points, before engaging the model switch. Also note that you can run the sampler from 1 to 0 (steps of -0.001) as well, to get a sense of the precision of the estimate. When the run is done, read the result calling

```
$ ../readthermo <chainname>
```

This will give you the difference in log-likelihood between uniform and gamma distributed rates, under this particular tree and with the given set of branch lengths. Note that including updates to `Gamma`, as well as branch lengths operators, you would be computing the difference in log marginal likelihood under the two models (the log Bayes factor).

The `MSMode` used here is `RAS` (for Rates Across Sites). Other settings can be used for analogous calculations across substitution matrices (`MSMode SUB`), or a straight-across path from a particular matrix with uniform rates to another matrix with gamma rates (`MSMode SUBRAS`).

Site-interdependent thermodynamic integration works similarly.

```
DataFile        myo20.nex

RatePrior       GammaInv
LengthPrior     Exponential

ModeFastCompute No
RefFastCompute No

QuasiStatic     0 0.5 0.001 10
MSMode pFactorModelSwitch

MoveType        NodeSiteStateOverTreeMove       5       1       10
MoveType        NodeSiteStateMove               15      1       50
MoveType        PathSiteMove                    25      1       100

End

withDependence          Yes
contactMap              1MBD.mj.cm
solventAccess           1MBD.mj.av
potential               homeMadeMJstyleWithAV
chemicalPotentials      Yes

MValue  500
thetaStarThreshold      0.01
gibbsIterBtwSeqs        5
```

```
SaveEvery       5
StopAfter       -1

InitState
Tree    (PONPY:0,(((MOUSE:0.2277,OCHPR:0.0283):0.0336,(PROGU...

Nmode 1
ModeStationaries        Empirical
ModeRR          Poisson

RefStationaries Uniform
RefRR           Poisson

Rates   Uniform

//
```

This will give you the log-likelihood difference between the flat model (Poisson, Uniform) and the structural model. Here, the path linking the models is defined with the pFactor parameter itself, hence the name `pFactorModelSwitch` for the `MSMode`. If you want to marginalize over other elements (e.g. to compute Bayes factors) simply include the update operators in question. For marginalizing over pFactor, run the integration across the range of interest. Calling `readthermo` here again will give you the results once the chain is done.

Thermodynamic integration for codon models

The following initfile will run the GY-MG-switch, described in chapter 7:

```
DataFile                bglobin.nex

Normalise               Yes
SynOnly                 Yes
OmegaPrior              Flat
LengthPrior             Exponential

ModeFastCompute No
RefFastCompute No
```

```
QuasiStatic 0 1 0.0005 100
MSMode SUB
CodonThermo     MG_GYSwitch

MoveType        AllBranchLength             25      0.1     1
MoveType        AllBranchLength             30      0.15    1
MoveType        AllBranchLength             35      0.05    1
MoveType        MeanBranchLengthMove        10      0.5     1
MoveType        MeanBranchLengthMove        10      0.1     1
MoveType        MeanBranchLengthMove        10      1       1
MoveType        NucleotideStat              10      1000    2
MoveType        NucleotideStat              10      2000    2
MoveType        NucleotideStat              10      3000    2
MoveType        NucleotideRelRate           5       1000    4
MoveType        NucleotideRelRate           5       500     4
MoveType        GlobalOmega                 5       0.05    1
MoveType        GlobalOmega                 5       0.1     1

End

CodonStatModelType      MG_F1X4
CodonRRModelType        GTR

SaveEvery       5
StopAfter       -1

InitState
Tree    (xenlaev:0.51040247,xentrop:0.78663390...

//
```

Numerous other thermodynamic schemes are implemented, following similar specifications, and each requiring calls to `readthermo` once the runs are completed.

Under the structural models, a thermodynamic run can be specified by:

```
DataFile        bglobin.nex

Normalise       No
SynOnly         No
OmegaPrior      Flat
LengthPrior     Exponential
AlphaPrior      Exponential
```

```
ModeFastCompute No
RefFastCompute No

QuasiStatic     0 1 0.0001 100
MSMode          pFactorJointSwitch

MoveType        AllBranchLength              100     1.0     1
MoveType        AllBranchLength              100     0.5     1
MoveType        MeanBranchLengthMove         50      0.5     1
MoveType        MeanBranchLengthMove         50      0.1     1
MoveType        MeanBranchLengthMove         50      1       1
MoveType        NucleotideStat               10      1000    2
MoveType        NucleotideStat               20      2000    2
MoveType        NucleotideRelRate            10      1000    4
MoveType        NucleotideRelRate            10      500     4
MoveType        GlobalOmega                  5       0.05    1
MoveType        GlobalOmega                  5       0.1     1
MoveType        NodeSiteStateOverTreeMove    10      1       10
MoveType        NodeSiteStateMove            50      1       30
MoveType        PathSiteMove                 50      1       50

End

CodonStatModelType      MG_F1X4
CodonRRModelType        GTR

withStructure           Yes
withDependence          Yes
potential               homeMadeMJstyleWithAV
chemicalPotentials      Yes
solventAccess           4HHBB.av
contactMap              4HHBB.cm.temp

MValue    100
gibbsIterBtwSeqs 5

SaveEvery       5
StopAfter       -1

InitState
Tree    (xenlaev:0.51040247,xentrop:0.78663390
```

Several other schemes are available here as well, with relatively few modification to this initfile.

Laplace approximation

Before running a Laplace approximation of the marginal likelihood, you should have performed a ML run and, if the model is not analytical, a thermodynamic run to compute the log-likelihood. Setting the `InitState` to the ML point, and including the log likelihood value in the initfile, a Laplace approximation run can be performed; engage the calculation using the keyword `LaplaceMode`, and indicate which components were optimized. The following is an example initfile for computing the marginal likelihood under the structural model, with pFactor treated as a free parameter.

```
DataFile          myo20.nex

RatePrior         GammaInv
LengthPrior       Exponential

ModeFastCompute No
RefFastCompute No

MoveType          NodeSiteStateOverTreeMove       5       1       10
MoveType          NodeSiteStateMove               25      1       50
MoveType          PathSiteMove                    25      1       100

End

withDependence          Yes
contactMap              1MBD.mj.cm
solventAccess           1MBD.mj.av
potential               homeMadeMJstyleWithAV
chemicalPotentials      Yes

MValue  100
thetaStarThreshold      0.01
gibbsIterBtwSeqs        5

LaplaceMode
logLikelihood           -2289.4
decorrelate             1
reburn                  1
gradEstimateBasedOn     1000
```

```
pFactorMLtuning          1
branchLengthMLtuning     1


SaveEvery        1
StopAfter        -1

InitState
Tree      (PONPY:0,(((MOUSE:0.1724,OCHPR:0.0238):0.0313,(PROGU...

Nmode 1
ModeStationaries         Empirical
ModeRR           Poisson

RefStationaries Uniform
RefRR            Poisson

Rates            Uniform

pFactor 0.64

//
```

The Laplace estimate is written in the file `<chainname>.laplace`. Other extensions for other models follow as before.

Monitoring and diagnostics

Regardless of the type of MCMC performed, it will be important to assess the general behavior of the sampler. Try the following call

```
$ ../monitor -v <chainname>
```

The `-v` option will give you additional information about the time spent in each operator, as well as the update success rates. The `monitor` program will produce many files containing the values of the hypothesis vector and associated statistics, for each sample point. Some gnuplot scripts are also made, to view multidimensional elements

more quickly (for instance `<chainname>.gnuplot_length`). Load these into gnuplot, e.g.

```
gnuplot> load "<chainname>.gnuplot_length"
```

and zoom over each value (in this case branch lengths) by hitting the carriage return.

Also, some simple statistics about the chain can be computed using

```
$ ../diagnostics <chainname>
```

Options are available to burn over a first set of points, and/or sub-sample from the chain; just make the call without arguments to see how to do this. This program will give a snapshot to the screen of the mean, variance, min, max, and autocorrelation of the components of the model. A bit more details are written into a file (`<chainame>.diagnostics`).

Posterior predictive resampling

In the previous section, we outlined how to run the MCMC-based statistical computations. While important, many of the calculations would typically be done only later in the model building cycle. Basic checks that the model is reproducing features of the data are perhaps more fundamental, and more informative. In a Bayesian context, this is known as *posterior predictive checking.*

For our purposes, posterior predictive checking works as follows. Suppose that you have run a Bayesian sampler under an evolutionary model of interest. For each draw from the posterior, simulate evolution over the tree based on the parameter values, producing a data replicate. Compute some statistic on each replicate, thereby generating the posterior distribution of the statistic. Compare with the true data. Under our specific applications, statistics may also be computed on the mappings, in which

case we compare the statistics from the unconstrained (predictive) mappings to the data-constrained ("observed") mappings.

Simulation

Simulation is simple. First, a chain sampling from the posterior of interest should have been made. Then, make the following call

```
$ ../postpredictive <chainname>
```

Use the `-progress` (or `-p`) option to count out on screen how many points have been treated. Options to burn over a first set of points, and/or to sub-sample, are also available here.

Many statistics could be explored under the posterior predictive scheme. Two examples are given below.

Rate variance

Among the files generated by the `postpredictive` program are `<chainname>.observedRateVar` as well as `<chainname>.predictiveRateVar`. As the names suggest, these statistics reflect the observed and predictive rate heterogeneity (e.g., Nielsen, 2002).

In the 'sources' directory, you will find a sub-directory called 'utilities'. The makefile found in this directory will compile some small programs, placing them in the 'data' directory, with the other programs. Among these, you can use the `makehisto` and `normhisto` to produce posterior density plots of the rate variance, which you can then use to produce figures.

Exchange distributions—bubble plots

The program `postpredictive` also produces files with an extension finishing with `ExchangeDist`. The files contain 190 values corresponding to the posterior mean pro-

portion of times an amino acid exchange occurred between a given pair.

You could compute some sort of distance between predictive and observed exchange distributions, corresponding directly to a discrepancy statistic. It may be more informative, however, to display exchange distributions graphically. If you've compiled the utility programs, try the following call

```
$ ../bubbleplot <chainname>.meanPredictiveExchangeDist <plotname>
```

This will read the `.meanPredictiveExchangeDist` file, and use gnuplot to produce a-la-Goldman bubble plots (fig. 8.4). The figure is produced as an EPS file, but the gnuplot script is also left behind for tweaking (`<plotname>.gnuplot`).

The area of each circle corresponds to the (normalized) values in the exchange distribution file (i.e. the total area of all circles equals 1).

Caveats

The programs have not yet been made user-friendly, and continue to be modified, revamped, and expanded. Likewise, the above descriptions are likely to be quickly outdated, and will require much revision before a true program release can be considered. Indeed, much more is available than has been described here, including many unpublished models. Much work remains.