Codon Evolution

Mechanisms and Models

EDITED BY

Gina M. Cannarozzi
University of Bern, Switzerland

Adrian Schneider
University of Utrecht, The Netherlands

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published in the United States
by Oxford University Press Inc., New York

First published 2012

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Library of Congress Control Number: 2011944051

Typeset by SPI Publisher Services, Pondicherry, India
Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

ISBN 978–0–19–960116–5

1 3 5 7 9 10 8 6 4 2

CHAPTER 6

Detecting and understanding natural selection

Maria Anisimova and David A. Liberles

6.1 Selective mechanisms operating on gene sequences

More than 150 years after the publication of Darwin's *Origin of species*, natural selection continues to be the dominant explanation for the phenotypic variety of living organisms on earth. In recent years, the wealth of comparative and functional genomic studies demonstrated a number of different ways in which natural selection operates on gene sequences. Molecular evolutionary processes are driven by mutations in a single individual in a population. They occur in genomic regions of different functions, from those that code for proteins to those that regulate the expression of proteins and other regions. Mutations themselves can range from single nucleotide changes, insertion or deletion events to gene and chromosome duplication or rearrangement events. As these events occur in an individual, their net selective effects, including the selective effects of linked changes, may increase or decrease an individual's fitness compared to other individuals in a population, which then dictates the probability of fixation of a new mutation given the population size (see Chapter 7).

At the molecular level, several types of selection may be distinguished (Figure 6.1). *Positive* selection acts upon advantageous mutations (with selection coefficient $s > 0$), reflecting the preferential fixation of mutations with a higher probability compared to the random expectation for a given population size. For inter-species data, positive selection that favours recurrent fixation of amino acid changes is known as *diversifying* selection. Diversifying selection is often the molecular mechanism to avoid host recognition. For example, the evolutionary arms race drives diversifying changes in poliovirus PV1 to keep the recognition of the host's receptor, which in turn mutates to avoid binding (Figure 6.1b; Zhang *et al.*, 2008). *Directional* selection eliminates variation within populations, increasing the frequency of the beneficial mutation and leading to its fixation. Environmental adaptation in bats is one such example (Figure 6.1a; Tellgren-Roth *et al.*, 2009). In population data, positive selection may manifest itself through *balancing* or frequency-dependent selection, which increases variability within a population due to a fitness advantage to maintain a polymorphism. Classic examples include balancing selection in immune system molecules (e.g. mammalian Major Histocompatibility Complex; Hughes and Nei, 1988), disease resistance loci (e.g. human genes associated with malaria resistance (Kwiatkowski, 2005), R proteins in plants (Van der Hoorn *et al.*, 2002), and in the sex locus in honey bees (Figure 6.1d; Cho *et al.*, 2006). These are cases where diversity across the population is favoured and rare alleles gain in frequency until they are no longer rare. In the case of the immune system, rare alleles are less likely to have generated neutralizing resistance mutations from pathogens in the evolutionary arms race. In honey bees, to prevent inbreeding, drones with rare alleles are more likely to find queens with different alleles (Cho *et al.*, 2006). Another type of positive selection that affects populations is the selective sweep, whereby a new advantageous mutation reduces variation in linked neutral sites (known as the *hitchhiking effect*) as it increases in frequency and is fixed in the population. One famous example of a selective sweep is the development of lactose tolerance in humans in response to dietary change (Figure 6.1e; Tishkoff *et al*., 2007). In contrast, *negative* or *purifying* selection acts against low-fitness changes ($s < 0$), most often conserving the amino acid sequence. This type of selection is most common and affects the majority of proteins. Purifying selection was implicated in the conservation of the protein sequence of the melanocortin 1 receptor (MC1R) locus in human African populations. This ensured that dark skin colour was maintained, as it was important for survival with prolonged sunlight exposure after the loss of body hair (Figure 6.1g; Rogers *et al*., 2004).

When developing new models and methods to detect selection, features of selective forces on proteins need to be considered. Most proteins have either solely a binding function or can both bind and catalyse a reaction on at least one of the bound entities (enzymes). Proteins bind mostly to either large biological macromolecules, like other proteins and nucleic acids, or to small molecules. The rules of binding to proteins and to small molecules appear to be different. For protein–protein interactions, affinity tends to derive from hydrophobic patches on the surface, while specificity derives from localized charged residues (Pechmann *et al*., 2009). There are of course, exceptions to this. In binding small molecules the rules are less clear, where van der Waals' interactions are important for affinity, and a number of factors, including charge and steric fit, affect specificity. There are larger level concerns governing the degree of specificity. For example, a hydrophobic patch without charge is expected to be fairly non-specific in its interactions. Further, the kinetic flexibility of a binding pocket will also affect specificity (DePristo *et al*., 2005). Disordered regions reflect an extreme case of this, where refolding upon binding can give specificity by deriving energy from the conformational shift from a lower energy disordered state. This can also enable allosteric coupling of binding events mediated by disorder (Hilser and Thompson, 2007). Ultimately, specificity of binding appears to be an important part of biological selection, where there is selective pressure not only on what to bind, but also on what not to bind (Liberles *et al*., 2011).

Within this opaque rule structure, positive selection acting upon a binding partner of a protein may affect its function in several ways. For example, the change of lysine to aspartic acid in a binding pocket can be predicted to have an effect on the affinity of potentially bound molecules. The classic case in enzyme specificity involves the modulation of pocket size and charge in the trypsin/chymotrypsin/elastase gene family, where trypsin prefers positively charged amino acids, while elastase prefers small amino acids.

In addition to folding and binding, selection also occurs on catalysis for enzymes. However, it appears to be easier to shift substrate (binding partner) than enzymatic reaction class or mechanism. In fact, enzyme specificity appears to be difficult to achieve, often with 'moonlighting' reactions (secondary reactions that are carried out at lower enzymatic efficiencies). Gene duplication is one process that enables optimization of a secondary reaction while maintaining a paralog that catalyses the original reaction. Copley (2009) has suggested that this process is a common mechanism by which bacteria evolve the capacity to metabolize anthropogenic compounds. In this case, multiple enzymes may be co-opted in the process of linking up metabolites to existing pathways in the species. This may be a mechanism by which new pathways are established. Classically, two hypotheses have been presented for the formation of new pathways. In the retrograde evolution model, enzymes evolve by changing catalytic mechanism, while maintaining binding to a substrate that becomes a product (Horowitz, 1945). Under this mechanism, pathways are built up backwards with substrate depletion conferring a selective advantage to individuals that can now produce the substrate. In the patchwork model (consistent with Copley's examples), enzymes maintain catalytic mechanism, but carry out reactions on new substrates (Jensen, 1976). In a systematic study of *E. coli* metabolism, Light and Kraulis (2004) suggest that the patchwork mechanism dominates.

Currently, selection is typically studied at the level of the individual gene or protein. Ultimately, however, selection acts on the inputs and outputs of pathways. Ardawatia and Liberles (2007) have examined average selective pressures across pathways in mammals based on d_N/d_S estimates (see Chapter 2) for gene families. It is not clear that selection needs to act on multiple members

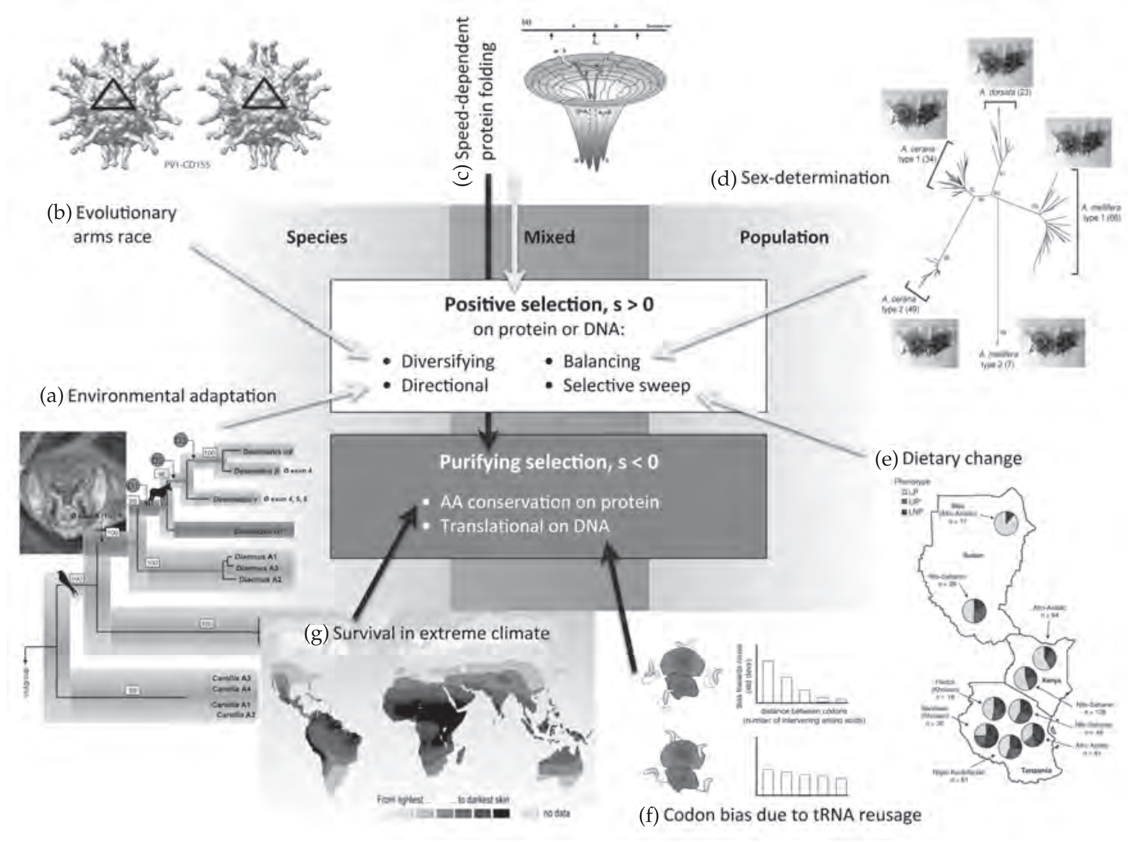
PV1-CD155
(c) Speed-dependent protein folding
(d) Sex-determination
A. dorsata (23)
A. cerana type 1 (34)
A. mellifera type 1 (66)
A. cerana type 2 (49)
A. mellifera type 2 (7)
(b) Evolutionary arms race
Species
Mixed
Population
Positive selection, s > 0
on protein or DNA:
• Diversifying
• Directional
• Balancing
• Selective sweep
(a) Environmental adaptation
Purifying selection, s < 0
• AA conservation on protein
• Translational on DNA
(e) Dietary change
Sudan
Kenya
Tanzania
(g) Survival in extreme climate
From lightest ... to darkest skin
no data
distance between codons (number of intervening amino acids)
(f) Codon bias due to tRNA reusage

Figure 6.1 Types of natural selection at the molecular level.

Natural selection is an important biological force shaping genetic patterns in molecular data. Distinct selective mechanisms are often responsible for morphological and behavioural evolution, the origin of evolutionary innovations, emergence of competition, environmental adaptations, and the evolution of complexity. Depicted are examples of different types of selection.

(a) ***Environmental adaptation***: directional selection in plasminogen activators of vampire bats due to the transition in feeding behaviour from bird to mammalian blood (phylogeny picture from Tellgren-Roth *et al.* (2009), copyright © Springer; used with permission). The vampire bat picture was provided by Daniel Abram from Rancho Transylvania (New Mexico, USA). **(b)** ***Evolutionary arms race***: diversifying positive selection in poliovirus PV1 to maintain the recognition of the host receptor, which in turn mutates to escape binding. From Zhang *et al.* (2008), © 2008 National Academy of Sciences, USA; used with permission. **(c)** ***Speed-dependent protein folding***: the folding pathways are drawn as step-by-step arrows on the simplified folding funnel surface. Without a pause at S1, fragment B folds before fragment A; then, fragment A folds on fragment B with an A1 conformation. On the other hand, with a synonymous mutation in S1, the pause enables A2 to fold first, and fragment B follows. The folding branches due to a pause in sequential folding, eventually lead to the bottom of the funnel with a minor conformational change between them. The figure and the description are reprinted from Tsai *et al.* (2008) © 2008 Elsevier; used with permission. **(d)** ***Sex-determination***: balancing selection in honey bees. To prevent inbreeding, drones with rare alleles are more likely to find queens with different alleles (Cho *et al.* 2006). The phylogeny is from Cho *et al.* (2006) © 2006, Cold Spring Harbor Laboratory Press; used with permission. Bees photograph taken by Zachary Huang (Michigan State University; http://www.beetography.com) and kindly provided by Soochin Cho (Creighton University). **(e)** ***Dietary change***: lactose persistence due to positive selection on LCT gene in African populations (adapted from Tishkoff *et al.* (2007) and reproduced by permission from Macmillan Publishers Ltd: Nature Genetics, © (2007). (f) Codon bias due to tRNA re-usage: codons using the same tRNA tend to be re-used for the same amino acids in close proximity. From Cannarozzi *et al.* (2010), adapted and reproduced with permission from Elsevier. Depicted are the diffusion and channeling models suggested by the authors. The tRNA diffusion away from the ribosome is slower than translation, and some tRNA channeling takes place at the ribosome. The ribosomal shape is drawn based on the crystal structure of the bacterial ribosome with tRNA to scale (Schuwirth *et al.*, 2005). **(g)** ***Survival in adverse climate conditions***: the MC1R locus was affected by strong purifying selection in African populations, since maintaining dark skin color was important for survival with prolonged exposure to sunlight after the loss of human body hair earlier along the human lineage (Rogers *et al.* 2004). Depicted is the skin color map for indigenous people predicted from multiple environmental factors—produced by Emmanuelle Bournay, UNEP/GRID-Arendal (http://maps.grida.no/go/graphic/skin-colour-map-indigenous-people). Data source: Chaplin G., Geographic distribution of environmental factors influencing human skin coloration. American Journal of Physical Anthropology 125:292–302, 2004; map updated in 2007.

of a pathway to achieve an effect, as changes to activities of key members of pathways can alter metabolic fluxes. Modularity emerges, perhaps non-adaptively, in simulations of pathway evolution (Soyer and Bonhoeffer, 2006). Such evolutionary dynamics can have important downstream effects on evolvability and adaptive potential (for discussion see Teufel *et al*., 2012). An important future direction will be to improve our understanding of selective pressures at the pathway level.

Gene duplication itself provides increased opportunities for selection. This has been reviewed extensively elsewhere (Liberles *et al*., 2010; Roth *et al*., 2007). The duplicates themselves can be under selection. For example, if increased expression of a product is desired, selective pressure will act to retain a duplicate. An interesting example of this involves the amylase protein in the human population (Perry *et al*., 2007). Conversely, the dosage-balance model provides selective pressure to eliminate duplicates that are not co-duplicated with interacting partners (Hughes *et al*., 2007). One molecular mechanism underpinning this may be dominant negative effects. Further, most models of gene duplication assume that the initial duplication event is neutral. A calculation from Wagner (2010) suggests that the metabolic cost of extra gene expression is, in fact, deleterious at a level that would show effects reaching to small population size organisms. This suggests that any retained duplicates were kept through sufficiently strong positive diversifying selection.

While in most cases of reported selection the change is observed *at the protein level*, both positive and negative selection may also be observed *at the DNA level* (either on silent codons and non-coding sites). Negative selection is known to cause strong codon bias, which works to optimize translational accuracy, efficiency, and robustness, as well as to provide control for optimal gene expression (e.g. Duret, 2002). Codons are known to be unequal in their usage in organisms, although the bias is not universally conserved across species. Codon usage bias correlates with tRNA concentration, where the more common codons have a higher concentration of cognate tRNA (e.g. Rocha, 2004). It has further been shown that codons using the same tRNA tend to be re-used for the same amino acids in close proximity (Figure 6.1f; Cannarozzi *et al*., 2010). This is due to an increased effective local concentration of the tRNA, even if it is not the tRNA with the highest cellular concentration. Genes expressed at high levels tend to use more common codons, presumably to reduce the waiting time for a tRNA to occupy the site. This serves three purposes: to increase the rate of expression, to decrease the rate of mistranslation due to the occupancy of an open site by a non-cognate tRNA, and to control kinetic processes of protein folding during translation. In this last case, it has been demonstrated that a synonymous change can affect the ultimate folded structure of the protein through this process, by not ending up in the kinetically trapped wild-type structure (Figure 6.1c; Tsai *et al*., 2008). In one example, a silent polymorphism in the Multidrug Resistance 1 gene affects the substrate specificity of P-glycoprotein (Kimchi-Sarfaty *et al*., 2007). Evidence for co-translational folding suggests that the speed of translation may affect the eventual 3D structure and the function of the protein (Komar, 2009), with rare codons used to slow down translation to obtain optimal folding. In this case, for certain codons (e.g. between distinct structural domains) a diverse choice of non-optimal codons may be favoured (if ribosomal pausing has increased fitness benefits), driven by positive selection on the DNA. In a systematic analysis of mammalian orthologs, positive selection on synonymous sites was inferred for 12% of the analysed genes, and was found to correlate with lower predicted mRNA stability compared to genes with negative selection on synonymous sites (Resch *et al*., 2007). Thus mRNA destabilization (affecting mRNA levels and translation) could be another important factor driving positive selection on synonymous sites. It should be noted that most discussion on selection focuses on primary selection, where there is a direct selective effect of the substitution. However, there have been several important suggestions of secondary selection, where the selective pressure provides a buffer against deleterious mutation, acting by selecting for processes that either prevent or buffer the effects of deleterious mutational events. Secondary selection is expected to be strongest in organisms with high mutation rates and large population sizes, where there is an increased chance of specific deleterious mutations and where the power to select for mechanisms to prevent them is greatest (Elena *et al*., 2007; Forster *et al*., 2006). One important mechanism of secondary selection that has emerged is the hypothesis of selection for more stable proteins to prevent mistranslation-induced misfolding (Drummond and Wilke, 2008; Wilke and Drummond, 2010). This involves selection on synonymous sites not only for the speed, but also the accuracy of translation and the mutation induced by mistranslation.

We continue by discussing a range of statistical methods used to study selection in molecular sequences based on inter-specific comparisons or within a population. Several methodological challenges are outlined, and common misconceptions of the statistical modelling are discussed in the context of detecting selection. Finally, we briefly review recent conclusions from large-scale genomic studies and their applications in the emerging field of the evolutionary medicine.

6.2 Brief overview of statistical methodologies for detecting positive selection

Statistical methods for detecting selection may be roughly classified according to the type of data under consideration. Population genetic samples typically consist of very similar sequences, where most suitable methods study the frequency spectrum of mutations, including neutrality tests or methods explicitly based on population genetic models. Maximum likelihood (ML) and Bayesian methods based on codon models are most appropriate at the intermediate divergence ranges and are typically applied to intra-specific samples (e.g. Anisimova *et al*., 2001). On the other edge of the evolutionary spectrum are datasets of homologous genes separated by deep evolutionary times. Here methods based on codon models lose their appeal since d_N/d_S becomes ineffective as d_S reaches saturation (but see Seo and Kishino, 2008, 2009). Because of multiple hits, the ability to accurately measure d_S decreases above values of 2–3 expected substitutions per site. When this occurs, the most powerful methods to measure selection become methods that detect rate shifts during the evolutionary history of a sample.

6.2.1 Neutrality tests based on frequency spectrum

With the rise of the neutral theory (see Chapter 2), tests for neutrality became very prominent and continue to be widely used. Tajima's test calculates statistic D as a scaled difference between the estimates of population-scaled mutation rate θ, one from the number of pairwise differences and another from the number of segregating sites in a sample (Tajima, 1989). Selection, demographic changes, genetic hitchhiking, and other violations of the neutral model will affect the two estimates differently, causing significant deviations of D from 0. However, the test may be rejected for various reasons. Estimates of $D < 0$ may indicate negative selection, including selective sweeps but also population expansion, while estimates of $D > 0$ are consistent with balancing selection, as well as a population bottleneck. Similar to Tajima's test, other neutrality tests contrast different estimates of θ from the site-frequency spectrum of a sample (Fay and Wu, 2000; Fu and Li, 1993). The stronger the contrast between the two estimates in presence of selection is, the more powerful the test of selection is. Note that the power of a neutrality test may be increased by the use of an outgroup, which helps to distinguish ancestral and derived states for polymorphism data, but may also be problematic due to inaccuracies of ancestral inference (Baudry and Depaulis, 2003).

6.2.2 Neutrality tests based on variability within and between species

The neutral hypothesis may be assessed by comparing the variability within and between species for two or more loci. The popular HKA test evaluates whether levels of polymorphism and divergence are proportional to the mutation rate, resulting in a constant ratio of polymorphism to divergence (Hudson *et al*., 1987). For example, the HKA test was one of tests used to demonstrate balancing selection in the honey bee in Figure 6.1d (Cho *et al*., 2006). Like other tests based solely on simple summary statistics (e.g. Tajima's *D*, Fu and Li's *D* and *F*, Fay and Wu's *H* tests), the HKA test is sufficiently powerful to reject the strictly neutral model, but is sensitive to demographic assumptions, failing to distinguish the demographic processes from selective forces. Conducting neutrality tests separately on nonsynonymous and synonymous differences may provide some additional insight into the interplay of the forces operating on the protein-coding level. In particular, the MK test (based on the idea of HKA test) compares the ratio of nonsynonymous to synonymous differences within and between species, which should be the same in absence of selection (McDonald and Kreitman, 1991). This test is more robust to demographic assumptions, since the effect of the demographic model is expected to be the same for both nonsynonymous and synonymous sites (Nielsen, 2001). Modifications of the MK test were proposed to differentiate between the types of selection (Akashi 1995, 1999b; Templeton, 1996). For example, Akashi examined the frequency distribution of observed synonymous and nonsynonymous changes compared with the neutral expectation. The power of this test is low when selection is weak or with only few adaptive mutations. Moreover, deviations from neutrality may be equally attributed to changes in population size (Eyre-Walker, 2002; Smith and Eyre-Walker, 2002). While the demographic process affects all genomic loci, selection affects only some. Many genome-wide studies used this argument to separate the effects of selection and demography (e.g. Thornton *et al*., 2007).

6.2.3 Poisson random-field models (PRF)

Unlike neutrality tests, PRF models explicitly include mutation and selection parameters under various population genetics scenarios (Akashi, 1999a; Hartl *et al*., 1994; Sawyer and Hartl, 1992). The strength of selection is estimated from the observed deviation of site-frequency distribution (including entries of MK tables with synonymous and nonsynonymous counts) from the expectation under neutrality. On the downside, the assumption of site independence means that selection on linked sites is ignored, biasing estimates from PRF models (Bustamante *et al*., 2001). A composite likelihood (CL) approach allows the inclusion of recombination and the relaxation of the assumption of site independence (Zhu and Bustamante, 2005). The composite likelihood ratio (CLR) test for selection showed good power to detect recurrent directional selection and was relatively robust in estimating the bias of the local recombination rate but not of population growth or a recent bottleneck. However, accounting for a suitable demographic model makes the selection test more robust to basic assumptions about demography (e.g. Williamson *et al*., 2007). Moreover, PRF-based methods are more powerful for multiple loci, since they provide more information about species' divergence time and population sizes, which is common for all loci. Because HKA, MK tests, and PRF models make the infinite-sites assumption (where each new mutation is observed at a different site), they are only appropriate for samples of low divergence.

6.2.4 Methods based on population differentiation

Increased levels of subdivision in natural populations may be caused by selection. For example, if geographical barriers cause population structure, advantageous mutations may arise only in a subpopulation, or the fitness of existing allele changes during the migration event in response to a new environment. Thus, unusually high levels of genetic population differentiation at one locus, compared to other loci, may be interpreted as evidence for positive selection (Lewontin and Krakauer, 1973). Several neutrality tests measure the population differentiation using the F_{ST} statistic and its variants (Hudson *et al*., 1992; Shriver *et al*., 2004; Weir *et al*., 2005). Levels of population differentiation may be modelled and estimated using a sophisticated Bayesian framework (Beaumont and Balding, 2004). A recent human genome study detected selection from patterns of allelic differentiation between two populations (Nielsen *et al*., 2009), where the demographic model was first estimated and then used to obtain the expected neutral frequency spectrum. Locus-specific outliers were considered to have been affected by selection. Chen *et al*. (2010) suggested a more rigorous treatment of allelic differences within both neutral and selection models. Based on patterns of allelic differences in two populations, they used the CLR method to test for selective sweeps. Using allelic differences makes the method more robust to the ascertainment bias (sampling bias caused by the process of the SNP discovery), which affects all other methods based on frequency spectra, population differentiation, and linkage disequilibrium (Nielsen *et al*., 2005).

6.2.5 Methods based on linkage disequilibrium (LD) and haplotype structure

Genomic regions with polymorphisms under balancing selection, or due to selective sweeps, may increase (or reduce) the correlation between alleles from different loci, known as LD. An ongoing incomplete selective sweep (advantageous mutation has not yet fixed in the population) leaves a special signature in the haplotype structure—the presence of a high-frequency haplotype with high LD. This is because there was little time for recombination to occur during this rapid spread of a haplotype containing the advantageous mutation. Popular tests based on LD and haplotype structure (Andolfatto *et al*., 1999; Depaulis and Veuille, 1998; Hudson *et al*., 1994) now include more recent additions: the relative extended haplotype homozygosity (rEHH; Sabeti *et al*., 2002), the integrated haplotype test (iHS; Voight *et al*., 2006), and the LD decay test (LDD; Wang *et al*., 2006). Related test statistics detect geographically restricted selection (Kimura *et al*., 2007; Sabeti *et al*., 2007; Tang *et al*., 2007). However, once the sweep is complete, there remains little variation to study LD patterns. In addition, methods based on LD also heavily rely on assumptions about recombination rates, as well as the demographic model. Note that selective sweeps and LD can be explicitly included in a population genetic model using the CL method (Kim and Nielsen, 2004; Kim and Stephan, 2002).

6.2.6 Methods based on detecting rate shifts

For divergent inter-specific samples, a popular strategy is to detect substitution rate shifts during the evolutionary history of a sample. Rather than normalizing one rate by another rate that is expected to be neutral, these measures look for site-specific shifts in substitution rates along a branch. The drawback compared to codon model-based approaches (Chapter 2), is that selection is not modelled explicitly. Also at the amino acid level, there is not a stringent criterion to statistically confirm positive selection without pursuing further functional and structural studies on detected sites. While instances of rate changes may be often caused by selection, they can be a result of other forces, including compensatory covariation driven by protein structural constraints (Fukami-Kobayashi *et al*., 2002; Philippe *et al*., 2003). Ultimately, rate-shift models (like other phylogenetic models) assume site-independence to model a process that is inherently site inter-dependent, generating a model that diverges from underlying biological processes and also requiring higher order Markovian models that are computationally hard. Signatures of rate shifts may indicate that a gene has been affected by variable selective pressures during evolution. When positive selection occurs, one expects an increase in the number of sites with rate shifts. In particular, rate shifts at generally conserved positions are good predictors of functional divergence (Philippe *et al*., 2003), most likely due to positive selection. Indeed, a study of covarion-like rate shifts in Ensembl homologs found that sites with rate shifts were frequently detected, but they were not as often associated with positive selection (detected based on codon models) compared to generally conserved sites with rate shifts (Studer and Robinson-Rechavi, 2010). Note also that a shift in a substitution rate is not a necessary condition that describes genes affected by positive selection. In many genes, boundaries of functionally constrained regions are relatively well conserved through time, such as in immune genes, where positive selection affects mostly the hypervariable antigenic regions (as is in the example of the MHC mentioned above; Hughes and Nei, 1988).

Numerous methods for detecting functional divergence exist (for more detail see Anisimova and Liberles, 2007). Similar to branch-site and clade models (Chapter 2), they search for a lineage-specific change (Blouin *et al*., 2003; Gaucher *et al*., 2002; Landau *et al*., 2005; Lockhart *et al*., 1998; Miyamoto and Fitch, 1995; Penny *et al*., 2001;

Siltberg and Liberles, 2002). For example, site-specific profiles based on a hidden Markov model may be used to identify residues responsible for functional differences between gene clusters (Gu 2001, 2006). Alternatively, with *a priori* partitioning of a tree, rate estimates in distinct subtrees may be compared using a LRT (Pupko and Galtier, 2002). Covarion models of rate switching may be formulated via a Markov-modulated processes (Galtier, 2001; Galtier and Jean-Marie, 2004; Huelsenbeck, 2002; Wang *et al*., 2007). While most tests for rate shifts assume *a priori* partitioning of sequences into groups with potential differences, several approaches can infer specific sites and lineages where rate shifts occurred (Dorman, 2007; Penn *et al*., 2008).

Note that the power of methods to detect rate shifts is low for sequences of insufficient divergence or an insufficient magnitude of rate shifts. For sites detected to have undergone a rate shift, adaptive substitutions affecting the function have to be discriminated from neutral or those due to compensatory changes based on further structural and experimental studies. Apart from detecting candidate genes under positive selection, predicting a functional shift from sequence data alone can be useful for large-scale protein annotation (Abhiman and Sonnhammer, 2005a, 2005b; Krishnamurthy *et al*., 2006).

6.2.7 Detecting selection based on d_N/d_S with Markov codon models

Unlike amino acid or nucleotide-based methods for detecting selection, at the codon level the ability to discriminate between synonymous and nonsynonymous substitutions provides us with an objective way to measure selection and to differentiate between positive and negative selection. Methods based on estimating d_N/d_S and codon models were discussed in Chapter 2. Clearly, they are the most successful at detecting recurrent positive selection in inter-specific samples, as they distinguish nonsynonymous and synonymous changes based on the structure of genetic code. Such methods may allow variation of selective pressure among sites and during the evolution. Thus these methods can be very informative about specific locations in the protein affected by recurrent changes and can detect lineages that were affected by selection during certain episodes of time.

The effectiveness (the power) of methods based on the d_N/d_S measure depends on the signal-to-noise ratio present in data, which is defined not only by divergence (Anisimova *et al*., 2001, 2002) but also by the fraction of residues with the potential to impact function. This depends upon the precise protein fold, the binding-site size, and the surface-area-to-volume ratio of the protein. This includes the contact-density hypothesis describing functional selection based upon the fraction of residues required for protein–protein interactions (for example see Zhou *et al*., 2008).

More intricate details of evolutionary specifics have been added recently to the toolbox for selection studies provided by standard codon models (Chapter 2). For example, better model fit may be achieved by including empirically estimated parameters that capture exchangeability patterns between codons (Chapter 3). Using different amino acid fitness profiles for sites, or including content dependency, should make models more reliable (Robinson *et al*., 2003; Rodrigue *et al*., 2010; Stern and Pupko, 2006; Yap *et al*., 2010). Such so-called semi-parametric models should increase the accuracy of inferences of selection. More recent codon models may be used to study positive and negative selection not only at the protein level, but also on synonymous substitutions (Yang and Nielsen, 2008; Zhou *et al*., 2010; see also Chapter 14).

To accommodate positive selection acting upon a binding partner of a protein and so affect protein function (see 6.6.1), several strategies were proposed in order to integrate related aspects of protein function into a codon model. Biophysical parameters can be integrated into codon models, explicitly characterizing the energetics of protein folding and binding interactions. The field has moved from modelling proteins as lattices (Williams *et al*., 2001) to forward (Rastogi *et al*., 2006) and backward (Kleinman *et al*., 2010) parameterization of codon models for coarse-grained approximations of real proteins. Another class of models that is computationally simpler involves gross analysis of biophysical properties (McClellan and Ellison, 2010; Woolley *et al*., 2003) that can easily be extended from amino acid models to codon models to also include types of synonymous substitutions.

Binding interactions can be predicted using the mirror tree method. This method looks for correspondence of evolutionary rates between sets of orthologous (or paralogous) proteins to identify interacting partners. In the most common implementation, a distance matrix is built for potential (orthologous) interacting partners and covariance of the rates in the matrix are assessed. Because there is an underlying species' tree to the gene tree evolution, one improvement involves the use of expected correlated branch lengths based upon the species tree. Another approach is to examine regions of a protein that may interact, where one expects the signal to be stronger, rather than examining an average over the protein as a whole. A recent improvement to this method uses selective pressures based upon codon models rather than rates to evaluate correspondence (Clark and Aquadro, 2010).

As the d_N/d_S measure and its related modifications continue to be widely used for evaluating selective pressures on protein-coding genes, here we continue by discussing several details of the d_N/d_S interpretation and possible pitfalls.

6.3 The utility and the interpretation of the d_N/d_S measure

Throughout this book it can be seen that the ω-ratio is the most-widely used measure of selective pressures on protein. It is often thought that the assumption of neutrality at synonymous sites is necessary for the measure to be meaningful. The concerns are caused by a possibility of selection acting on codon usage, which would reduce d_S, resulting in elevated ω-values, and possible corrections were suggested (Hirsh *et al.*, 2005). Reports from Drosophila studies demonstrated that synonymous sites could be affected by selection for translational efficiency (Akashi, 2001; Akashi and Eyre-Walker, 1998; Duret, 2002; Kreitman *et al*, 1995). More recent evidence from high-profile experimental studies shows other cases when synonymous substitutions may not be considered neutral as they influence translation, splicing, gene regulation, mRNA stability, protein abundance, and even protein folding (Carlini and Genut, 2006; Chamary *et al.*, 2006; Kimchi-Sarfaty *et al.*, 2007; Komar, 2007, 2009; Parmley *et al.*, 2006; Tsai *et al.*, 2008).

Currently it is unclear how often estimates of ω are biased due to reduced d_S. In a large-scale study of human-mouse orthologs, Zhang and Li (2004) found no trend for increased ω for lower values of d_S. Yet, one recent study suggests that ignoring among-site synonymous variability may cause an elevated level of false-positive inferences of positive selection (Rubinstein *et al.*, 2011). Consequently, modelling variation of synonymous rates (as well as nonsynonymous) may be desirable, so as to avoid the possible negative effects of the d_S underestimation (for example, as is suggested by: Kosakovsky Pond *et al.*, 2010; Rubinstein *et al.*, 2011).

Nevertheless, the neutrality of d_S is generally not required for the ω-ratio to be an effective measure of selection on protein, so long as the evolutionary forces apply equally to synonymous and nonsynonymous sites (Yang, 2006). Since doubts about the d_S neutrality assumption recently re-occurred in the literature, here we briefly review the arguments evoked by Yang (2006).

Given a codon-substitution model with the instantaneous rate matrix $Q = \{q_{ij}\}$ (for examples see Chapter 2), proportions of nonsynonymous and synonymous mutations can be calculated:

$$\rho_N = \sum_{i \neq j} \pi_i q_{ij} \text{ where } i \text{ and } j \text{ are nonsynonymous,}$$
$$\rho_S = \sum_{i \neq j} \pi_i q_{ij} \text{ where } i \text{ and } j \text{ are synonymous.} \tag{6.1}$$

Then the rates of nonsynonymous and synonymous substitutions per codon between two sequences over time t are:

$$d_N = N_d/N \quad \text{and} \quad d_S = S_d/S, \tag{6.2}$$

where N_d and S_d are the numbers of nonsynonymous and synonymous substitutions per codon:

$$N_d = t\rho_N \quad \text{and} \quad S_d = t\rho_S, \tag{6.3}$$

and N and S are numbers of nonsynonymous and synonymous sites per codon:

$$N = 3\rho_N^{\omega=1} \quad \text{and} \quad S = 3\rho_S^{\omega=1} \tag{6.4}$$

with proportions $\rho_N^{\omega=1}$ and $\rho_S^{\omega=1}$ computed as in Eqn 6.1 but assuming $\omega = 1$, i.e. no selection on the protein. From Eqn 6.2–6.4 we can see that the ω-ratio evaluates the disruption of nonsynonymous and synonymous changes caused by natural selection on the protein, as it is the ratio of two ratios:

$$\omega = (\rho_N/\rho_S)/\left(\rho_N^{\omega=1}/\rho_S^{\omega=1}\right), \tag{6.5}$$

so that the observed ratio of proportions of nonsynonymous and synonymous changes is compared to a neutral expectation. The potential selection acting on synonymous sites is essentially the selection at the DNA and RNA levels, as it affects both synonymous and nonsynonymous sites equally. Whether or not the evolution at synonymous sites is neutral, it can be shown mathematically that the d_S is the average rate of change over the three codon positions before selection on the protein $d_S = \frac{t}{3}\sum_{i\neq j}\pi_i q_{ij}^{\omega=1}$, where $q_{ij}^{\omega=1}$ is calculated the same way as q_{ij} but assuming $\omega = 1$, and $d_N = \omega d_S$ is the rate of change after the selection on the protein (Yang, 2006). As a result, contrasting d_N and d_S evaluates the difference of rates before and after selection operated on the protein, whether evolution at silent sites is driven by mutation or selection. If synonymous sites evolve non-neutrally due to codon bias, mutation-selection models (Nielsen and Yang, 2003; Yang and Nielsen, 2008) may be used to also study the mutational biases or selection on synonymous codon usage. For example, in the model FMutSel of Yang and Nielsen (2008), the mutational biases and selection at the DNA level are incorporated using fitness parameters s_{ij} of each possible change, which are dependent on the effective population size.

However, forces that act differentially on synonymous and nonsynonymous sites are of concern, if they are not incorporated into a model. Xing and Lee (2006) discussed possible sources of bias, such as RNA selection pressure that is 3-nucleotide-periodic and systematically different between adjacent nonsynonymous and synonymous sites, so that the average effect on nonsynonymous and synonymous sites is distinct. Codon bias can produce such effects but may be accounted for with models like FMutSel (Yang and Nielsen, 2008). Another potential source of such unequal bias may be the *synonymous phasing* of binding sites for splicing factors or other proteins (Xing and Lee, 2006), where it may be advantageous for the binding sites to place their most constrained nucleotides in synonymous sites and avoid nonsynonymous sites. Indeed, empirical studies show that binding sites for splicing factors, such as exonic splicing enhancers, may exhibit such a behaviour (Cartegni *et al.*, 2003; Liu *et al.*, 1998). For example, if a motif SF2/ASF systematically positions its conserved nucleotide G at a synonymous site, this may reduce (by maximum 54%) the probability of a substitution at a synonymous site compared to a nonsynonymous site. However, such maximum effect is rather unlikely, since it requires a systematic positioning bias (which is not observed for every instance) and four-fold degeneracy at all synonymous sites (which is not true at all sites). In addition, short lengths of such motifs (e.g. 6 nt for SF2/ASF) means that the overall effect on the ω-ratio is likely to be negligible, since it is typically measured over much longer lengths of coding sequences (with recommended min. ~100 codons, (Anisimova *et al.*, 2001; Anisimova *et al.*, 2007)). In fact, several experimental studies showed no strong phasing effect (Dirksen *et al.*, 2000; Pollard *et al.*, 2002; Rooke *et al.*, 2003). Both bioinformatics and significant experimental effort will be necessary to evaluate whether and how often RNA regulatory motifs have a tendency to place their conserved positions in synonymous sites.

Whether or not it is rare for some biological forces to act differently on nonsynonymous and synonymous sites, can be studied by adapting existing codon models. For example, the new codon models of Zhou *et al.* (2010) distinguish conserved and non-conserved synonymous changes, unlike the standard models that assume all synonymous changes are the same (but not FMutSel of Yang and Nielsen, 2008). In the presence of codon bias, it seems more realistic to differentiate between synonymous changes that retain a preferred or non-preferred codon and those that interconvert between such codons. Based upon application of this method, it was found that purifying selection acted upon 5–10% of synonymous sites, whereas positive selection on synonymous sites was rare (Zhou *et al.*, 2010).

Another important consideration when interpreting estimates of the ω-ratio relates to the genetics of populations represented in a dataset. The ω-ratio represents the selective pressure for a particular codon site (or a set of sites) on a macro-evolutionary scale. On a shorter scale, i.e. in population genetics, the focus of study is the distribution of the selection coefficient s of new mutations (or alleles) within a population. Using Kimura's result for the fixation probability of new mutations (Kimura, 1962), Sawyer and Hartl (1992) derived the relationship between ω and s for the infinite sites model, while Nielsen and Yang (2003) used a similar reasoning to demonstrate such a relationship for the finite sites model (also see Chapter 7). The ω-ratio may be represented as a function of the effective population size and the fitness coefficients, which can be derived as a limit of an underlying Wright–Fisher population process (Fisher, 1930) or the Moran (1962) model. If all synonymous sites are assumed to be neutral and all nonsynonymous changes have the same selective coefficient s, then the relative rate of nonsynonymous vs. synonymous *fixation* events is described by:

$$\omega = f(S) = \frac{S}{1 - e^{-S}} \tag{6.5}$$

where $S = 2N_e s$ is the population-scaled selection coefficient for haploid organisms with effective population size N_e.

Other assumptions (typical for population genetics' models) include independence of sites and the fact that no more than two alleles are segregating in the population at a single site, which is realistic for low mutation rates (typical of most organisms). The interpretation of $\omega > 1$ as evidence of positive selection is theoretically supported given a Wright–Fisher model with selection (Nielsen and Yang, 2003), so that $\omega > 1$ corresponds to $s > 0$. With selection being more efficient in larger populations, the power of detecting positive or negative selection is expected to be higher for organisms with large population size. On the other hand, there will be more relaxed selection and potentially more difficulty in differentiating it from neutral evolution in species with small population sizes.

Based on Eqn 6.5, inferences about relative population sizes may be made based on estimates of ω ratios (e.g. Kosiol *et al.*, 2008). For example, if ω_1 and ω_2 are the estimates for populations represented by lineages 1 and 2 both with selection coefficient s, then the ratio of effective population sizes N_1 and N_2 may be estimated using the inverse mapping between ω and S:

$$\frac{N_1}{N_2} = \frac{N_1 S}{N_2 S} = \frac{f^{-1}(\omega_1)}{f^{-1}(\omega_2)}. \tag{6.6}$$

However, when modelling assumptions of Eqn 6.5 are not satisfied, selection coefficients will tend to be underestimated. Moreover, typical intra-specific samples include polymorphisms that segregate within populations, instead of fixed differences as in inter-specific samples. Kryazhimskiy and Plotkin (2008) derived an analytical approximation for the expected ω under a single-population Wright–Fisher model with selection, which is different from Eqn 6.5 and in contrast depends not only on the scaled selection coefficient, but also on the population mutation rate. Their computer simulations were used to study the interpretation of ω in a single population and demonstrated that the estimate of ω becomes less reliable as an indicator of selection. In particular, for large values of S, the estimates of ω are often ≤ 1. This means that the ω-based test for positive selection in a single population sample will often fail to detect selection, even if selection has operated. On the other hand, estimates of $S \leq 0$ are unlikely to have a correspondent ω estimate > 1. This is consistent with the current view that LRT for positive selection lack power to detect selection in population samples (Anisimova, 2003). A significant LRT for positive selection in a population sample may be due either to positive selection or differences in a population size. Slightly deleterious nonsynonymous mutations are more likely to be segregating in small populations than in large populations. To distinguish the two scenarios, the population size should be estimated using neutral markers.

6.4 Accounting for indels and overlapping ORFs

Most methods for detecting selection, including those based on codon models, examine simple point substitutions, but ignore insertions, deletions, overlapping ORFs, and more complex events. Positive diversifying selection acts not only on substitutions in protein coding genes, but insertions and deletions may also play an important functional role. Podlaha and Zhang (2003) have shown that positive selection can act on linker length, where the length of a loop can affect the local effective concentration of a domain on one side of a loop with a domain on the other side. If function relies upon interaction of the domains, the probability of interaction at any time will depend upon the length of the loop (and the association and dissociation constants for the interaction). This was shown in the CATSPER1 voltage gated calcium channel involved in sperm motility. Further, loops tend to be the most variable parts of proteins accumulating insertions and deletions at a higher rate, and are known to form binding pockets and interfaces for protein–protein interaction, as well as intra-molecular domain–domain interactions. In a systematic study of insertion and deletion dynamics across gene families in the PVC superphylum of bacteria, it was found that lineage-specific positive diversifying selection on indels acts at least as frequently as positive diversifying selection on substitutions (Kamneva *et al*., 2010). Examples of positive diversifying selection on insertions and deletions were detected in all secondary structural units, while occurring most frequently in looped regions. For example, specific insertions into alpha-helical regions of the *Gemmatu obscuriglobus* L17 ribosomal protein are thought to affect its interaction with 23S rRNA (Kamneva *et al*., 2010). As codon models develop, transitions between gapped and non-gapped states will need to be incorporated. The first steps toward this difficult task have been taken (Fletcher and Yang, 2010; Rivas, 2005; Rivas and Eddy, 2008; Suchard and Redelings, 2006). Another fertile direction in improving codon models concerns their ability to accommodate frameshift mutations (unlike amino acid models) and the underlying functional consequences (unlike DNA models) (Sabath and Graur, 2010; Sabath *et al*., 2008). Chapter 2 of this book discusses some solutions to address violations of other model assumptions, such as recombination and non-independence of sites.

6.5 Model-based approaches and common misconceptions

The use of sound and robust statistical methods is fundamental in any problem where inferences are made based on observed data. Model-based inference offers great advantages by explicitly incorporating parameters of interest, allowing studies of the interplay between different model features using a statistical inference framework of choice, such as maximum likelihood or Bayesian inference. Models provide an excellent foundation for hypotheses testing, prediction, and decision-making. Critics of model-based approaches point out that every model makes a number of unrealistic assumptions and thus cannot truly reflect real data. While models may be inherently incorrect in several ways, some of them can be very useful (Box, 1979). Choosing or defining a useful model is a balancing act, where only the factors reflecting major biases and features should be included, while omitting other factors that have little effect on model robustness. In the words of Einstein, the model should be ‘as simple as possible, but not simpler’. In place of model-based approaches, parsimony-motivated arguments and *ad hoc* techniques are sometimes preferred for their simplicity. However, non-model approaches also make assumptions, and their statistical properties are similar to ‘no common mechanism’ models, which are inherently too parameter-rich and never have enough data to estimate all their parameters (Holder *et al*., 2010; Tuffley and Steel, 1998). As should transpire throughout this book, robust statistical approaches based on consistent and identifiable models should always be preferred. For example, multiple *ad hoc* methods have been suggested for the estimation of d_N and d_S rates. However, ML estimation based on Markov codon models outperforms all such methods, given that the same biases are accounted for (Yang and Nielsen, 2000). Complex demographic scenarios become possible to study in a model-based framework, while *ad hoc* approaches produce very high rates of false inferences (Beaumont *et al*., 2010). Most simulation approaches require explicit models. For example, approximate Bayesian computation

(ABC) uses MCMC simulation to approximate the posterior of distributions or the likelihood surface from a population genetic model. This has been successfully used for phylogeographic inferences and testing for selection (Beaumont, 2002; Thornton and Andolfatto, 2006)). Moreover, effect of violations of fundamental assumptions may be tested in simulation, where the analysis model ignores or misplaces the important forces present in the simulation model. For example, such robustness tests were performed on LRTs for positive selection based on codon-models (Anisimova *et al.*, 2001, 2002, 2003; Anisimova *et al.*, 2007). Although simulation studies are often valuable sanity checks, simulation studies *should not be over-generalized* but provide some intuition about the properties of datasets for which the methods remain accurate. For example, optimal divergence and recombination levels that can be tolerated before resulting in excessive false-positive inferences of selection can be inferred. Computer simulations are also useful to evaluate the rate of false-positive inferences under the null hypothesis and the rate of false-negative inferences when the null does not hold. While it is naturally understandable to prefer the methods that do not make any, or very few, false positives, in practice such tests can be very conservative, as the high power of the test is achieved as *a trade-off between false positive and false-negative rates*. A method with no false positives is usually no better than a method with a low level of false positives, since it will typically be more conservative, making few true positive inferences. For example, Nozawa *et al*. (2009) criticized branch-site models of codon evolution, since they resulted in 32 cases of false positives out of 14,000 datasets simulated under the null model without selection. This is only 0.23% rate of false-positive error, which is lower than 5% defined by the significance level. At the same time, as pointed out by Yang and Goldman (Yang *et al.*, 2009), the power of parsimony-based methods promoted by Nozawa *et al*. (2009) is typically very low, whereas power of ML methods in detecting selection is often close to 100% (Wong *et al.*, 2004). Moreover, for divergent data parsimony or other counting approaches rely on reconstructed ancestral sequences as if they were observed. Even when parsimony is performed using the probability vectors of ancestral states, this will result in an under-counting of the number of mutations by failing to consider multiple mutations per site; such methods (Benner *et al*., 1998; Liberles *et al*., 2001) are better than other parsimony methods, but inferior to model-based approaches. For divergent datasets such approaches will be less accurate, while the ML method does not cause an elevated number of false positives in a LRT for positive selection (Anisimova *et al*., 2001; Anisimova *et al*., 2007).

One common mistake may be described as *data dredging*, so that a hypothesis is inferred from data and consequently validated using the same data. For example, a typical problem in evolutionary biology seeks to detect episodes of positive selection that affected one or more lineages in a protein-coding alignment. The biological insight is not often available to formulate *a priori* hypothesis for selection tests, as it is required with site-branch models (see Chapter 2). Thus, it may be tempting to apply another model, such as the free-ratio branch model to estimate the ω-ratio for each branch and then use these estimates to formulate subsequent hypotheses for the branch-test. However, letting the data influence the *a priori* hypothesis distorts the p-values of subsequent significance tests, although the parameterization of a model focused on the previously inferred lineage is still sensible. Tests that are not biased by the previous use of the data, on the other hand, are fully valid, including an analysis of a lineage where positive selection was previously detected based on a different dataset. To summarize, the use of data to formulate the *a priori* hypothesis (based on inferences) for subsequent testing biases the p-value of the test, but can still be used for model parameterization for the subsequent test.

Another problem involves the assumptions of the applied models and tests, and their power. A free-ratio model averages over all sites, but looks for lineage-specific selection, whereas a site model averages over all branches, but looks for sites that are on average under selection through evolutionary history. Tellgren *et al*. (2004) applied a free-ratios model to the myostatin gene in Artiodactyls, identifying positive selection on several lineages. Pie and Alvares (2006) applied a site model to the same dataset and did not find evidence for positive selection, claiming that it invalidated the results of Tellgren *et al.* (2004). Indeed, a simulation of sequences under the exact parameterization from the free-ratio model that was generated by Tellgren *et al.* does not show evidence for positive selection under the sites tests applied by Pie (data not shown), but this may be interpreted as a result of low power of the tests applied by Pie and Alvares (2006). Instead, branch-site models should have a better power to detect lineage-specific positive selection, as is the case with the myostatin data, where branch-site models detect only three Artiodactyl lineages affected by positive selection against a background of strong negative selection on other lineages, with no sites showing evidence for being under positive selection when the ω-ratio is averaged over all lineages.

When several tests are performed on the same or overlapping data, multiple hypotheses testing is required so that the overall false-positive error rate (known as family-wise error rate or FWER) is still below the required level. For example, if 10 tests are performed, each at the 5% significance level, then the overall error rate can be as high as $1 - (1 - 0.05)^{10}$, which is 40%. *Multiple testing correction* (e.g. Miller, 1981; Rom, 1990) is employed to reduce the FWER to the required level, but this also reduces the power of the test and causes increased levels of false negatives, especially when the number of tests is large. Since FWER may be often too stringent, the false-discovery rate (FDR) was proposed (Benjamini and Hochberg, 1995). FDR is defined as the expected proportion of false rejections among all rejected hypotheses. By definition, controlling FDR is possible when, at least for some tests, the rejection of the null is expected, and the threshold is set to indicate the tolerable (small) percentage of false rejections (for review see Manly *et al.*, 2006). For example, in the case of the multiple branch-site tests for positive selection where each test sets one branch at the foreground, the FDR may be controlled if positive selection on the dataset was already demonstrated on a gene as a whole (Anisimova *et al.*, 2007). Subsequent multiple branch-site tests will merely infer the branches likely to be under positive selection at some sites. However, corrections for multiple testing often seriously reduce the power of the test to detect true positives, especially when controlling FWER. Being an important part of biological discovery, the ability to identify lineages and sites under positive selection necessitates flexible approaches that are not only statistically viable but are also sufficiently powerful to discover episodic patterns. Since Bayesian inference methods do not require multiple testing, they appear more attractive when applied to infer loci or lineages under selection in large data, although such methods are often more computationally demanding.

Inevitably, inferences of natural selection come down to the classic problem of model selection: the model providing the best description of the data should be favoured. Both likelihood and Bayesian frameworks allow provision for model selection. Likelihood-ratio tests (or their CLR analogues) may be used to compare nested hypotheses, and so require the null hypothesis, which (in tests for selection) is typically described by a model without selection. Hierarchical LRT testing is possible for multiple hierarchically nested hypotheses, but requires multiple testing corrections and depends on the order of testing the hypotheses. The Bayesian equivalent of the LRT is to compute the Bayes factors. If the null is too simplistic and describes data poorly, it can be rejected, even if no selection is present. Likewise, if the alternative hypothesis misrepresents the phenomenon of interest (e.g., the way selection acts), the test may have low power due to poor fit compared to the null and not because the phenomenon is not present. The requirement for an alternative hypothesis was at the heart of the classic debate between the Fisherian and Neyman–Pearson statisticians. Ideally, multiple models should be formulated. These can be evaluated based on the information criteria, which intend to find a balance between maximizing the model fit and minimizing the number of parameters necessary to describe the data: AIC (Akaike, 1974), BIC (Schwarz, 1978), and DIC (Spiegelhalter *et al.*, 2002). The more recent DIC is still underutilized in bioinformatics and phylogenetics communities. While AIC and BIC are based on the maximized likelihood, DIC selects a model with the smallest deviance of the likelihood. In addition, DIC does not require the knowledge of the number of parameters describing a model, but rather estimates it based on the difference between the log-likelihood of the parameter expectation and the expectation of the log-likelihood over a sample. This may be convenient in some cases where the complexity and the model hierarchy prevent us from knowing the exact number of parameters. For example, for a given protein-coding matrix, AIC and BIC cannot be used to compare codon, nucleotide, and amino acid models, due to our inability to include the transformation of the data structure in the parameter calculation. DIC makes it possible to compare such models, relying on a Bayesian framework.

However, just like with the LRT, the properties of information criteria hold asymptotically (for large samples). Given this and the problems associated with multiple testing or defining sensible *a priori* hypotheses, a Bayesian framework for model comparison and selection may offer more elegant statistical solutions. Indeed, given a set of models (e.g., representing various selective or demographic scenarios) posterior probabilities for each model may be compared without *a priori* knowledge of most likely scenarios, with no need for multiple testing correction. The Bayesian framework has a strong potential for discovering the unknown relationships in large comparative and population genomics data, together with other probabilistic machine learning approaches. Bayesian approaches are often better at dealing with smaller samples and may incorporate more parameters compared to likelihood approaches, making them convenient for model selection among multiple complex scenarios. However, problems with formulating reasonable priors and convergence issues may pose serious setbacks (e.g., (Rannala *et al*., 2011).

Hahn (2008) argued that recent evidence from genomic analyses indicates that neutral evolution no longer constitutes a useful null hypothesis, since most predictions of the neutral theory are overwhelmingly rejected by genomic data. Assuming that the majority of genes do not evolve under selection (Cavalli-Sforza, 1966; Lewontin and Krakauer, 1973) biases the results of selection tests, where the signal from the majority of genes is equated to be neutral and is used to estimate demographic model without selection. Given the complexity of the problem, model-averaging approaches (both frequentist and Bayesian) may be helpful to estimate confidence regions of the parameters of focus. Stochastic approaches allowing variation of population models among loci (such as selection vs. neutral) may also be promising.

Nevertheless, it can be quite challenging to avoid false positives (and false negatives) in large-scale scans for selection. Besides issues stemming from model misspecifications, artefacts in genomic data (Mallick *et al*. 2009; Schneider *et al*. 2009), errors in alignment (Fletcher and Yang, 2010) or other downstream analyses, such as biases due to coupling of multiple effects or failure to correct for multiple testing, all contribute to an amount of error in the final inferences of selection. We try to minimize the systematic error at every step of the procedure hoping that the end result will provide more than just noise. A carefully conducted selection scan provides a fertile ground for further testing of the candidate genes. It is here where further dangers lie: a careful judgment is required when interpreting the results from single-gene studies to avoid fictitious 'just-so stories'.

6.6 Selection and adaptive traits

More than 30 years ago in their seminal paper, Gould and Lewontin (1979) warned against equating the observed functional differences with adaptive changes since the existence of one particular form is not sufficient to deduce its purpose. Unfortunately, Gould and Lewontin's beautifully framed discussion evaded some patches of the genomics community, resulting in several embarrassing claims of trait adaptation without the direct evidence that selection was operating on these specific traits (Nielsen, 2009).

A variety of statistical methods enable us to detect selection on specific residues and possibly pinpoint the time episode during evolution, when selection operated. Researchers then strive to demonstrate the functional effect of such specific mutations (e.g. MacCallum and Hill, 2006). Multiple well-documented cases of adaptive evolution have been published, including some of the examples shown in Figure 6.1. However, selection may act differentially on different pleiotropic effects of selected residues, making it a much harder task to relate specific mutations to the adaptation of the phenotypic traits. For fast-evolving organisms such as viruses and bacteria, experimental evolution experiments can be used to demonstrate that certain mutations go to fixation under certain environmental changes (Wichman *et al*., 2005). There are two goals here and rigorous demonstration of adaptive mechanisms for an evolutionary biology audience requires more proof than characterization of genotype-molecular phenotype links, which is a goal in itself for molecular geneticists.

6.7 Lessons from genomic studies and implications for studies of genetic disease

In the last decade, various predictions from neutral theory have been intensely tested on genomic data or large-scale SNP datasets. Selection scans focused on detecting genomic regions affected by positive or negative selection and, in particular, new advantageous mutations that recently came to fixation in populations. For population genomic data, fitting suitable demographic models became crucial to disentangling the effects of selection (Nielsen *et al*., 2009). On the other hand, estimating demography from neutral models may also introduce bias when the majority of genes are affected by selection (Hahn, 2008), invalidating the *outlier approaches* employed in many population genetic studies. Despite the theoretical difficulties, this may bring us back to the almost forgotten *nearly-neutral theory* (Ohta 1992, 2002), which allows small amounts of positive selection at the background of mostly negative selection. Alternatively, the *genetic draft models* that include repeated selective sweeps (Gillespie 2000a, 2000b, 2001) may provide a better description of the population dynamics. Several comparative genomic studies used codon models to evaluate selective pressure based on the ω-ratio distribution over genes, among sites, and/or along lineages (Anisimova *et al*., 2007; Clark *et al*., 2003; Kosiol *et al*., 2008). Such studies may be even more insightful if conducted at both species and population levels. Indeed, integrating population genetic methods with comparative species methods may be very useful, as the dynamics of molecular change could be examined simultaneously at both population and species levels, providing additional insights about the dynamics of populations and speciation.

Biswas and Akey (2006) evaluated the consistency of selection inferences across several different genome scans for selection. While the overlap between the identified loci is typically very modest, this seems not only due to difficulties with confounding factors such as demography, but also to the fact that different methods detect different types of selection (and have different accuracy and power), which also depends on the evolutionary scale and populations included in a study. Overall, the emerging patterns strongly suggest that positive selection plays an important role in shaping the evolution of genomes, both within humans and between different species. While the majority of protein-coding genes evolve on average under strict purifying pressure, several studies detected positive selection in genes that are involved in a variety of biological processes; see, for example, the TAED (Liberles *et al*., 2001; Roth *et al*., 2005) and Selectome databases (Proux *et al*., 2009). In humans genes with positive selection signal are related to immunity, defense, tumour suppression, apoptosis, olfaction, sensory perception, and spermatogenesis (Akey, 2009; Nielsen *et al*., 2007). Genome scans can also be used to characterize the distribution of finesses of selected species' differences, as has been done, for example, for human-chimp differences, where 10–20% were estimated to be under positive selection, while the majority were deleterious (Boyko *et al*., 2008).

In contrast, genome-wide studies of selection on indels have been rare. This may be due to the fact that computationally tractable population or evolutionary models with indels are still lagging behind. Despite this, several genomic studies have already shone some light on indel evolution. Lunter and colleagues (2006) defined a 'neutral indel' model and used it to measure selection on non-coding regions of genomes. Coding indel patterns were examined in multiple sequence alignments of human, chimp, and rhesus macaque (De la Chaux *et al*., 2007) and in the PVC superphylum of bacteria (Kamneva *et al*., 2010). De la Chaux *et al*. found that coding indels were much less frequent compared to non-coding, indicating strong purifying constraints similar to constraints acting on codon substitutions. On the other hand, human-specific small-insertion events may be driven by positive selection (Chen *et al*., 2009).

Increasingly, disease-related studies stand to profit from genomic scans. Strong associations between selection and disease have been investigated, primarily in humans, but potentially should be very useful for other organisms. Many candidate genes under positive selection are involved in cancer-related processes, defence, immunity, chemosensory perception, and reproduction (for example, Kosiol *et al*., 2008). Genes with stronger purifying selection have a greater likelihood of being involved in Mendelian diseases, which are typically due to new deleterious alleles segregating in families (Bustamante *et al*., 2005; but see Clark *et al*., 2003). Human adaptations to climate may have contributed to selective pressure on genes associated with common metabolic disorders (Hancock *et al*., 2008). Complex disease like diabetes, asthma, heart disease and bipolar disorder also exhibit footprints of selection (Corona *et al*., 2010; Ding and Kullo, 2009). Blekhman *et al*. (2008) contrasted the evolutionary forces acting in complex and simple Mendelian disorders in humans. Genes involved in complex disease showed lower evolutionary conservation and were affected by both positive and purifying selection, unlike the Mendelian disease genes that are largely under strict negative selection. Unlike disease-association mapping, macro-evolutionary and population genetic studies focus on the fitness effect of susceptibility alleles, accounting for evolutionary dynamics in ancestral lineages. It is reasonable to believe that selective pressures acting on disease susceptibility alleles change over time due to environmental or cultural changes, and several hypotheses were proposed to reflect this. One classic example is 'the thrifty genotype' hypothesis explaining the high incidence of obesity and type II diabetes in modern humans (Neel, 1962). It was postulated that as ancestral human hunter-gatherer populations were regularly subjected to seasonal periods of feast and famine, with a very efficient system for fat and carbohydrate storage, this 'thriftiness' became detrimental when food became easily available across seasons with the development of food storage and processing strategies.

Thus fitness of ancestral alleles reflecting ancient adaptations to ancestral lifestyle is better described within non-stationary evolutionary models (Di Rienzo and Hudson, 2005). Detecting signatures of positive selection with such models contributes additional valuable insights during disease mapping, as fitnesses of derived and ancestral alleles are compared (Di Rienzo, 2006).

Finally, evolutionary inferences from both comparative and population genomic data, in combination with functional and structural information, can be used to make predictions of mutations or loci most likely to have negative fitness consequences (Adzhubei *et al*., 2010; Boyko *et al*., 2008; Ng and Henikoff, 2001; Ramensky *et al*., 2002). A combination of such analyses with analyses of positive selection and genome-wise association studies opens new prospects for identifying the genetic factors underlying complex disease (Chun and Fay, 2009; Corona *et al*., 2010; Manolio *et al*., 2009). Statistical analyses of the human genome may enable applications in a clinical content (Ashley *et al*., 2010). Among current objectives is the integration of the evolutionary and population genetics models with complimentary data sources (e.g., Dimitrieva and Anisimova, 2010) within machine learning approaches for pattern discovery and integrated mining of bio-data. The fields of evolutionary and medical genomics are growing and already boast some promising results. With new mechanistic codon models and their application to a host of biological and biomedical problems (Goode *et al*., 2008; Kosakovsky Pond *et al*., 2006), the future of functional genomics looks exciting.

Acknowledgements

We thank Ashley Teufel for careful reading of the chapter. We are also grateful to Joe Bielawski and Tal Pupko for many useful comments on our manuscript, which helped to improve it. Stefan Zoller (CBRG, ETH Zurich) generously invested some time to improve the figure.

David Liberles receives funding from NSF grant DBI-0743374 that supported writing of this chapter. Maria Anisimova receives funding from Swiss National Science Foundation (award 31003A_127325) and ETH Zurich.

References

Abhiman, S. and E.L. Sonnhammer (2005a). FunShift: a database of function shift analysis on protein subfamilies. Nucleic Acids Res **33**: D197–200.

Abhiman, S. and E.L. Sonnhammer (2005b). Large-scale prediction of function shift in protein families with a focus on enzymatic function. Proteins **60**: 758–768.

Adzhubei, I.A., S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova *et al*. (2010). A method and server for predicting damaging missense mutations. Nat Methods **7**: 248–249.

Akaike, H. (1974). A new look at the statistical model identification. Automatic Control, IEEE Transactions on Automatic Control **19**: 716–723.

Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in Drosophila DNA. Genetics **139**: 1067–1076.

Akashi, H. (1999a). Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics **151**: 221–238.

Akashi, H. (1999b). Within- and between-species DNA sequence variation and the 'footprint' of natural selection. Gene **238**: 39–51.

Akashi, H. (2001). Gene expression and molecular evolution. Curr Opin Genet Dev **11**: 660–666.

Akashi, H. and A. Eyre-Walker (1998). Translational selection and molecular evolution. Curr Opin Genet Dev **8**: 688–693.

Akey, J.M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res **19**: 711–722.

Andolfatto, P., J.D. Wall, and M. Kreitman (1999). Unusual haplotype structure at the proximal breakpoint of In(2L)t in a natural population of *Drosophila melanogaster*. Genetics **153**: 1297–1311.

Anisimova, M. (2003). Detecting positive selection in protein coding genes. PhD thesis. University College London, London.

Anisimova, M. and D.A. Liberles (2007). The quest for positive election in the era of comparative genomics. Heredity **99**: 567–579.

Anisimova, M. and Z. Yang (2007). Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol **24**: 1219–1228.

Anisimova, M., J.P. Bielawski, and Z. Yang (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol **18**: 1585–1592.

Anisimova, M., J.P. Bielawski, and Z. Yang (2002). Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol **19**: 950–958.

Anisimova, M., R. Nielsen and Z. Yang (2003). Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics **164**: 1229–1236.

Anisimova, M., J. Bielawski, K. Dunn, and Z. Yang (2007). Phylogenomic analysis of natural selection pressure in Streptococcus genomes. BMC Evol Biol **7**: 154.

Ardawatia, H. and D.A. Liberles (2007). A systematic analysis of lineage-specific evolution in metabolic pathways. Gene **387**: 67–74.

Ashley, E.A., A.J. Butte, M.T. Wheeler, R. Chen, T.E. Klein *et al*. (2010). Clinical assessment incorporating a personal genome. The Lancet **375**: 1525–1535.

Baudry, E. and F. Depaulis (2003). Effect of misoriented sites on neutrality tests with outgroup. Genetics **165**: 1619–1622.

Beaumont, M. (2002). Flavouring composition prepared by fermentation with *Bacillus* spp. Int J Food Microbiol **75**: 189–196.

Beaumont, M.A. and D.J. Balding (2004). Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol **13**: 969–980.

Beaumont, M.A., R. Nielsen, C. Robert, J. Hey, O. Gaggiotti *et al*. (2010). In defence of model-based inference in phylogeography. Molecular Ecology **19**: 436–446.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Statist Soc Ser B **57**: 289–300.

Benner, S., N. Trabesinger, and D. Schreiber (1998). Post-genomic science: converting primary structure into physiological function. Adv Enzyme Regul **38**: 155–180.

Biswas, S. and J.M. Akey (2006). Genomic insights into positive selection. Trends Genet **22**: 437–446.

Blekhman, R., O. Man, L. Herrmann, A.R. Boyko, A. Indap *et al*. (2008). Natural selection on genes that underlie human disease susceptibility. Curr Biol **18**: 883–889.

Blouin, C., Y. Boucher, and A.J. Roger (2003). Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. Nucleic Acids Res **31**: 790–797.

Box, G.E. P. (1979). Robustness in the strategy of scientific model building in *Robustness in statistics*, edited by R.L. Launer and G.N. Wilkinson. Academic Press, New York.

Boyko, A.R., S.H. Williamson, A.R. Indap, J.D. Degenhardt, R.D. Hernandez *et al*. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet **4**: e1000083.

Bustamante, C.D., J. Wakeley, S. Sawyer and D.L. Hartl (2001). Directional selection and the site-frequency spectrum. Genetics **159**: 1779–1788.

Bustamante, C.D., A. Fledel-Alon, S. Williamson, R. Nielsen, M.T. Hubisz *et al*. (2005). Natural selection on protein-coding genes in the human genome. Nature **437**: 1153–1157.

Cannarozzi, G., N.N. Schraudolph, M. Faty, P. von Rohr, M. Friberg *et al*. (2010). A role for codon order in translation dynamics. Cell **141**: 355–367.

Carlini, D.B. and J.E. Genut (2006). Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. J Mol Evol **62**: 89–98.

Cartegni, L., J. Wang, Z. Zhu, M.Q. Zhang, and A.R. Krainer (2003). ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res **31**: 3568–3571.

Cavalli-Sforza, L.L. (1966). Population structure and human evolution. Proc R Soc Lond B Biol Sci **164**: 362–379.

Chamary, J.V., J.L. Parmley, and L.D. Hurst (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet **7**: 98–108.

Chen, C.H., T.J. Chuang, B.Y. Liao, and F.C. Chen (2009). Scanning for the signatures of positive selection for human-specific insertions and deletions. Genome Biol Evol **1**: 415–419.

Chen, H., N. Patterson, and D. Reich (2010). Population differentiation as a test for selective sweeps. Genome Res **20**: 393–402.

Cho, S., Z.Y. Huang, D.R. Green, D.R. Smith, and J. Zhang (2006). Evolution of the complementary sex-determination gene of honey bees: balancing selection and trans-species polymorphisms. Genome Res **16**: 1366–1375.

Chun, S. and J.C. Fay (2009). Identification of deleterious mutations within three human genomes. Genome Res **19**: 1553–1561.

Clark, N.L. and C.F. Aquadro (2010). A novel method to detect proteins evolving at correlated rates: identifying new functional relationships between coevolving proteins. Mol Biol Evol **27**: 1152–1161.

Clark, A.G., S. Glanowski, R. Nielsen, P.D. Thomas, A. Kejariwal *et al*. (2003). Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios. Science **302**: 1960–1963.

Copley, S.D. (2009). Evolution of efficient pathways for degradation of anthropogenic chemicals. Nat Chem Biol **5**: 559–566.

Corona, E., J.T. Dudley, and A.J. Butte (2010). Extreme evolutionary disparities seen in positive selection across seven complex diseases. PLoS ONE **5**.

de la Chaux, N., P.W. Messer, and P.F. Arndt (2007). DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. BMC Evol Biol **7**: 191.

Depaulis, F. and M. Veuille (1998). Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol Biol Evol **15**: 1788–1790.

DePristo, M.A., D.M. Weinreich, and D.L. Hartl (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet **6**: 678–687.

Di Rienzo, A. (2006). Population genetics models of common diseases. Curr Opin Genet Dev **16**: 630–636.

Di Rienzo, A. and R.R. Hudson (2005). An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet **21**: 596–601.

Dimitrieva, S. and M. Anisimova (2010). PANDITplus: toward better integration of evolutionary view on molecular sequences with supplementary bioinformatics resources. Trends Evol Biol **2**: e1.

Ding, K. and I.J. Kullo (2009). Evolutionary genetics of coronary heart disease. Circulation **119**: 459–467.

Dirksen, W.P., X. Li, A. Mayeda, A.R. Krainer and F.M. Rottman (2000). Mapping the SF2/ASF binding sites in the bovine growth hormone exonic splicing enhancer. J Biol Chem **275**: 29170–29177.

Dorman, K.S. (2007). Identifying dramatic selection shifts in phylogenetic trees. BMC Evol Biol **7 Suppl 1**: S10.

Drummond, D.A. and C.O. Wilke (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell **134**: 341–352.

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev **12**: 640–649.

Elena, S.F., C.O. Wilke, C. Ofria, and R.E. Lenski (2007). Effects of population size and mutation rate on the evolution of mutational robustness. Evolution **61**: 666–674.

Eyre-Walker, A. (2002). Changing effective population size and the McDonald-Kreitman test. Genetics **162**: 2017–2024.

Fay, J.C. and C.I. Wu (2000). Hitchhiking under positive Darwinian selection. Genetics **155**: 1405–1413.

Fisher, R.A. (1930). *The genetical theory of natural selection*. Dover Press, New York.

Fletcher, W. and Z. Yang (2010). The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol Biol Evol **27**: 2257–2267.

Forster, R., C. Adami, and C.O. Wilke (2006). Selection for mutational robustness in finite populations. J Theor Biol **243**: 181–190.

Fu, Y.X. and W.H. Li (1993). Statistical tests of neutrality of mutations. Genetics **133**: 693–709.

Fukami-Kobayashi, K., D.R. Schreiber, and S.A. Benner (2002). Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. J Mol Biol **319**: 729–743.

Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol **18**: 866–873.

Galtier, N. and A. Jean-Marie (2004). Markov-modulated Markov chains and the covarion process of molecular evolution. J Comput Biol **11**: 727–733.

Gaucher, E.A., X. Gu, M.M. Miyamoto, and S.A. Benner (2002). Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci **27**: 315–321.

Gillespie, J.H. (2000a). Genetic drift in an infinite population. The pseudohitchhiking model. Genetics **155**: 909–919.

Gillespie, J.H. (2000b). The neutral theory in an infinite population. Gene **261**: 11–18.

Gillespie, J.H. (2001). Is the population size of a species relevant to its evolution? Evolution **55**: 2161–2169.

Goode, M., S. Guindon, and A. Rodrigo (2008). Modelling the evolution of protein coding sequences sampled from Measurably Evolving Populations. Genome Inform **21**: 150–164.

Gould, S.J. and R.C. Lewontin (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proc R Soc Lond B Biol Sci **205**: 581–598.

Gu, X. (2001). Maximum-likelihood approach for gene family evolution under functional divergence. Mol Biol Evol **18**: 453–464.

Gu, X. (2006). A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. Mol Biol Evol **23**: 1937–1945.

Hahn, M.W. (2008). Toward a selection theory of molecular evolution. Evolution **62**: 255–265.

Hancock, A.M., D.B. Witonsky, A.S. Gordon, G. Eshel, J.K. Pritchard *et al*. (2008). Adaptations to climate in candidate genes for common metabolic disorders. PLoS Genet **4**: e32.

Hartl, D.L., E.N. Moriyama, and S.A. Sawyer (1994). Selection intensity for codon bias. Genetics **138**: 227–234.

Hilser, V.J. and E.B. Thompson (2007). Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. Proc Natl Acad Sci USA **104**: 8311–8315.

Hirsh, A.E., H.B. Fraser, and D.P. Wall (2005). Adjusting for selection on synonymous sites in estimates of evolutionary distance. Mol Biol Evol **22**: 174–177.

Holder, M.T., P.O. Lewis, and D.L. Swofford (2010). The akaike information criterion will not choose the no common mechanism model. Syst Biol **59**: 477–485.

Horowitz, N.H. (1945). On the evolution of biochemical syntheses. Proc Natl Acad Sci USA **31**: 153–157.

Hudson, R.R., M. Kreitman, and M. Aguade (1987). A test of neutral molecular evolution based on nucleotide data. Genetics **116**: 153–159.

Hudson, R.R., M. Slatkin, and W.P. Maddison (1992). Estimation of levels of gene flow from DNA sequence data. Genetics **132**: 583–589.

Hudson, R.R., K. Bailey, D. Skarecky, J. Kwiatowski, and F.J. Ayala (1994). Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. Genetics **136**: 1329–1340.

Huelsenbeck, J.P. (2002). Testing a covariotide model of DNA substitution. Mol Biol Evol **19**: 698–707.

Hughes, A.L. and M. Nei (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**: 167–170.

Hughes, T., D. Ekman, H. Ardawatia, A. Elofsson, and D.A. Liberles (2007). Evaluating dosage compensation as a cause of duplicate gene retention in Paramecium tetraurelia. Genome Biol **8**: 213.

Jensen, R.A. (1976). Enzyme recruitment in evolution of new function. Annu Rev Microbiol **30**: 409–425.

Kamneva, O., D.A. Liberles, and N. Ward (2010). Genome-wide analysis of insertion and deletion substitutions in organisms of the PVC superphylum. Genome Biol Evol **2**: 870–886.

Kim, Y. and R. Nielsen (2004). Linkage disequilibrium as a signature of selective sweeps. Genetics **167**: 1513–1524.

Kim, Y. and W. Stephan (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160**: 765–777.

Kimchi-Sarfaty, C., J.M. Oh, I.W. Kim, Z.E. Sauna, A.M. Calcagno *et al*. (2007). A 'silent' polymorphism in the MDR1 gene changes substrate specificity. Science **315**: 525–528.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. Genetics **47**: 713–719.

Kimura, R., A. Fujimoto, K. Tokunaga, and J. Ohashi (2007). A practical genome scan for population-specific strong selective sweeps that have reached fixation. PLoS ONE **2**: e286.

Kleinman, C.L., N. Rodrigue, N. Lartillot, and H. Philippe (2010). Statistical potentials for improved structurally constrained evolutionary models. Mol Biol Evol **27**: 1546–1560.

Komar, A.A. (2007). Genetics. SNPs, silent but not invisible. Science **315**: 466–467.

Komar, A.A. (2009). A pause for thought along the co-translational folding pathway. Trends Biochem Sci **34**: 16–24.

Kosakovsky Pond, S.L., S.D. Frost, Z. Grossman, M.B. Gravenor, D.D. Richman *et al*. (2006). Adaptation to different human populations by HIV-1 revealed by codon-based analyses. PLoS Comput Biol **2**: e62.

Kosakovsky Pond, S.L., K. Scheffler, M.B. Gravenor, A.F. Poon, and S.D. Frost (2010). Evolutionary fingerprinting of genes. Mol Biol Evol **27**: 520–536.

Kosiol, C., T. Vinar, R.R. da Fonseca, M.J. Hubisz, C.D. Bustamante *et al*. (2008). Patterns of positive selection in six Mammalian genomes. PLoS Genet **4**: e1000144.

Kreitman, M. and Akashi H. 1995 Molecular evidence for natural selection. Annu Rev Ecol Syst **26**: 403–422.

Krishnamurthy, N., D.P. Brown, D. Kirshner, and K. Sjolander (2006). PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. Genome Biol **7**: R83.

Kryazhimskiy, S. and J.B. Plotkin (2008). The population genetics of dN/dS. PLoS Genet **4**: e1000304.

Kwiatkowski, D.P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet **77**: 171–192.

Landau, M., I. Mayrose, Y. Rosenberg, F. Glaser, E. Martz *et al*. (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res **33**: W299–302.

Lewontin, R.C. and J. Krakauer (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74**: 175–195.

Liberles, D.A. (2001). Evaluation of methods for determination of a reconstructed history of gene sequence evolution. Mol Biol Evol **18**: 2040–2047.

Liberles, D.A., D.R. Schreiber, S. Govindarajan, S.G. Chamberlin, and S.A. Benner (2001). The adaptive evolution database (TAED). Genome Biol **2**: RESEARCH0028.

Liberles, D.A., G. Kolesov, and K. Dittmar (2010). Joining biochemistry and population genetics to understand gene duplication in *Evolution after gene duplication*, edited by K. Dittmar and D.A. Liberles. Wiley, New York.

Liberles, D.A., M.D. Tisdell and J.A. Grahnen (2011). Binding constraints on the evolution of enzymes and signalling proteins: the important role of negative pleiotropy. Proc Biol Sci B **278**(1714): 1930–1935.

Light, S. and P. Kraulis (2004). Network analysis of metabolic enzyme evolution in *Escherichia coli*. BMC Bioinformatics **5**: 15.

Liu, H.X., M. Zhang, and A.R. Krainer (1998). Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev **12**: 1998–2012.

Lockhart, P.J., M.A. Steel, A.C. Barbrook, D.H. Huson, M.A. Charleston *et al*. (1998). A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol Biol Evol **15**: 1183–1188.

Lunter, G., C.P. Ponting, and J. Hein (2006). Genome-wide identification of human functional DNA using a neutral indel model. PLoS Comput Biol **2**: e5.

MacCallum, C. and E. Hill (2006). Being positive about selection. PLoS Biol **4**: e87.

Mallick, S., S. Gnerre, P. Muller, and D. Reich D (2009). The difficulty of avoiding false positives in genome scans for natural selection. Genome Res **19**(5): 922–933.

Manly, K.F., D. Nettleton, and J.T. G. Hwang (2006). Genomics, prior probability, and statistical tests of multiple hypotheses. Genome Res **14**: 997–1001.

Manolio, T.A., F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff *et al*. (2009). Finding the missing heritability of complex diseases. Nature **461**: 747–753.

McClellan, D.A. and D.D. Ellison (2010). Assessing and improving the accuracy of detecting protein adaptation with the TreeSAAP analytical software. Int J Bioinform Res Appl **6**: 120–133.

McDonald, J.H. and M. Kreitman (1991). Adaptive protein evolution at the Adh locus in *Drosophila*. Nature **351**: 652–654.

Miller, R.G. J. (1981). *Simultaneous statistical inference*. Springer-Verlag, New York.

Miyamoto, M.M. and W.M. Fitch (1995). Testing the covarion hypothesis of molecular evolution. Mol Biol Evol **12**: 503–513.

Moran, P.A. P. (1962). *The statistical processes of evolutionary theory*. Clarendon Press, Oxford.

Neel, J.V. (1962). Diabetes mellitus: a 'thrifty' genotype rendered detrimental by 'progress'? Am J Hum Genet **14**: 353–362.

Ng, P.C. and S. Henikoff (2001). Predicting deleterious amino acid substitutions. Genome Res **11**: 863–874.

Nielsen, R. (2001). Statistical tests of selective neutrality in the age of genomics. Heredity **86**: 641–647.

Nielsen, R. (2009). Adaptionism-30 years after Gould and Lewontin. Evolution **63**: 2487–2490.

Nielsen, R. and Z. Yang (2003). Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol Biol Evol **20**: 1231–1239.

Nielsen, R., S. Williamson, Y. Kim, M.J. Hubisz, A.G. Clark *et al*. (2005). Genomic scans for selective sweeps using SNP data. Genome Res **15**: 1566–1575.

Nielsen, R., I. Hellmann, M. Hubisz, C. Bustamante, and A.G. Clark (2007). Recent and ongoing selection in the human genome. Nat Rev Genet **8**: 857–868.

Nielsen, R., M.J. Hubisz, I. Hellmann, D. Torgerson, A.M. Andres *et al*. (2009). Darwinian and demographic forces affecting human protein coding genes. Genome Res **19**: 838–849.

Nozawa, M., Y. Suzuki, and M. Nei (2009). Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc Natl Acad Sci USA **106**: 6700–6705.

Ohta, T. (1992). Theoretical study of near neutrality. II. Effect of subdivided population structure with local extinction and recolonization. Genetics **130**: 917–923.

Ohta, T. (2002). Near-neutrality in evolution of genes and gene regulation. Proc Natl Acad Sci USA **99**: 16134–16137.

Parmley, J.L., J.V. Chamary, and L.D. Hurst (2006). Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol **23**: 301–309.

Pechmann, S., E.D. Levy, G.G. Tartaglia, and M. Vendruscolo (2009). Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. Proc Natl Acad Sci USA **106**: 10159–10164.

Penn, O., A. Stern, N.D. Rubinstein, J. Dutheil, E. Bacharach *et al*. (2008). Evolutionary modeling of rate shifts reveals specificity determinants in HIV-1 subtypes. PLoS Comput Biol **4**: e1000214.

Penny, D., B.J. McComish, M.A. Charleston, and M.D. Hendy (2001). Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J Mol Evol **53**: 711–723.

Perry, G.H., N.J. Dominy, K.G. Claw, A.S. Lee, H. Fiegler *et al*. (2007). Diet and the evolution of human amylase gene copy number variation. Nat Genet **39**: 1256–1260.

Philippe, H., D. Casane, S. Gribaldo, P. Lopez, and J. Meunier (2003). Heterotachy and functional shift in protein evolution. IUBMB Life **55**: 257–265.

Pie, M.R. and L.E. Alvares (2006). Evolution of myostatin in vertebrates: is there evidence for positive selection? Mol Phylogenet Evol **41**: 730–734.

Podlaha, O., and J. Zhang (2003). Positive selection on protein-length in the evolution of a primate sperm ion channel. Proc Natl Acad Sci USA **100**: 12241–12246.

Pollard, A.J., A.R. Krainer, S.C. Robson, and G.N. Europe-Finner (2002). Alternative splicing of the adenylyl cyclase stimulatory G-protein G alpha(s) is regulated by SF2/ASF and heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1) and involves the use of an unusual TG 3′-splice Site. J Biol Chem **277**: 15241–15251.

Proux, E., R.A. Studer, S. Moretti, and M. Robinson-Rechavi (2009). Selectome: a database of positive selection. Nucleic Acids Res **37**: D404–407.

Pupko, T. and N. Galtier (2002). A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc Biol Sci **269**: 1313–1316.

Ramensky, V., P. Bork, and S. Sunyaev (2002). Human non synonymous SNPs: server and survey. Nucleic Acids Res **30**: 3894–3900.

Rannala, B., T. Zhu, and Z. Yang, in press. Tail paradox, partial identifiability and influential priors in Bayesian branch length inference. Mol Biol Evol.

Rastogi, S., N. Reuter and D.A. Liberles (2006). Evaluation of models for the evolution of protein sequences and functions under structural constraint. Biophys Chem **124**: 134–144.

Resch, A.M., L. Carmel, L. Marino-Ramirez, A.Y. Ogurtsov, S.A. Shabalina *et al*. (2007). Widespread positive selection in synonymous sites of mammalian genes. Mol Biol Evol **24**: 1821–1831.

Rivas, E. (2005). Evolutionary models for insertions and deletions in a probabilistic modeling framework. BMC Bioinformatics **6**: 63.

Rivas, E. and S.R. Eddy (2008). Probabilistic phylogenetic inference with insertions and deletions. PLoS Comput Biol **4**: e1000172.

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 **20**: 1692–1704.

Rocha, E.P. (2004). Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res **14**: 2279–2286.

Rodrigue, N., H. Philippe, and N. Lartillot (2010). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc Natl Acad Sci USA **107**: 4629–4634.

Rogers, A.R., D. Iltis, and S. Wooding (2004). Genetic variation at the MCIR locus and the time since loss of human body hair. Current Anthropology **45**: 105–108.

Rom, D.M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. Biometrika **77**: 663–665.

Rooke, N., V. Markovtsov, E. Cagavi, and D.L. Black (2003). Roles for SR proteins and hnRNP A1 in the regulation of c-src exon N1. Mol Cell Biol **23**: 1874–1884.

Roth, C., M.J. Betts, P. Steffansson, G. Saelensminde, and D.A. Liberles (2005). The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics. Nucleic Acids Res **33**: D495–497.

Roth, C., S. Rastogi, L. Arvestad, K. Dittmar, S. Light *et al*. (2007). Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. J Exp Zool B Mol Dev Evol **308**: 58–73.

Rubinstein, N.D., A. Doron-Faigenboim, I. Mayrose, and T. Pupko, 2011. Evolutionary models accounting for layers of selection in protein coding genes and their impact on the inference of positive selection. Mol Biol Evol.

Sabath, N. and D. Graur (2010). Detection of functional overlapping genes: simulation and case studies. J Mol Evol **71**: 308–316.

Sabath, N., G. Landan, and D. Graur (2008). A method for the simultaneous estimation of selection intensities in overlapping genes. PLoS ONE **3**: e3996.

Sabeti, P.C., D.E. Reich, J.M. Higgins, H.Z. Levine, D.J. Richter *et al*. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature **419**: 832–837.

Sabeti, P.C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter *et al*. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature **449**: 913–918.

Sawyer, S.A. and D.L. Hartl (1992). Population genetics of polymorphism and divergence. Genetics **132**: 1161–1176.

Schneider, A., A. Souvorov, N. Sabath, G. Landan, G.H. Gonnet, and D. Graur (2009). Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biol Evol **1**: 114–118.

Schwarz, G. (1978). Estimating the dimension of a model. Ann. Statist. **6**: 461–464.

Seo, T.K. and H. Kishino (2008). Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. Syst Biol **57**: 367–377.

Seo, T.K. and H. Kishino (2009). Statistical comparison of nucleotide, amino acid, and codon substitution models for evolutionary analysis of protein-coding sequences. Syst Biol **58**: 199–210.

Shriver, M.D., G.C. Kennedy, E.J. Parra, H.A. Lawson, V. Sonpar *et al*. (2004). The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics **1**: 274–286.

Siltberg, J. and D.A. Liberles (2002). A simple covarion-based approach to analyse nucleotide substitution rates. Journal of Evolutionary Biology **15**: 588–594.

Smith, N.G. and A. Eyre-Walker (2002). Adaptive protein evolution in *Drosophila*. Nature **415**: 1022–1024.

Soyer, O.S. and S. Bonhoeffer (2006). Evolution of complexity in signaling pathways. Proc Natl Acad Sci USA **103**: 16337–16342.

Spiegelhalter, D.J., N.G. Best, B.P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**: 583–639.

Stern, A. and T. Pupko (2006). An evolutionary space-time model with varying among-site dependencies. Mol Biol Evol **23**: 392–400.

Studer, R.A. and M. Robinson-Rechavi (2010). Large-scale analysis of orthologs and paralogs under covarion-like and constant-but-different models of amino acid evolution. Mol Biol Evol **27**: 2618–2627.

Suchard, M.A. and B.D. Redelings (2006). BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics **22**: 2047–2048.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585–595.

Tang, K., K.R. Thornton, and M. Stoneking (2007). A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol **5**: e171.

Tellgren, A., A.C. Berglund, P. Savolainen, C.M. Janis, and D.A. Liberles (2004). Myostatin rapid sequence evolution in ruminants predates domestication. Mol Phylogenet Evol **33**: 782–790.

Tellgren-Roth, A., K. Dittmar, S.E. Massey, C. Kemi, C. Tellgren-Roth *et al*. (2009). Keeping the blood flowing-plasminogen activator genes and feeding behavior in vampire bats. Naturwissenschaften **96**: 39–47.

Templeton, A.R. (1996). Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase II gene in the hominoid primates. Genetics **144**: 1263–1270.

Teufel, A.I., J.A. Grahnen, and D.A. Liberles (2012). Modeling proteins at the interface of structure, evolution, and population genetics in *Computational modeling of biological systems: from molecules to pathways*, edited by Dokholyan. Springer-Verlag, New York.

Thornton, K. and P. Andolfatto (2006). Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. Genetics **172**: 1607–1619.

Thornton, K.R., J.D. Jensen, C. Becquet, and P. Andolfatto (2007). Progress and prospects in mapping recent selection in the genome. Heredity **98**: 340–348.

Tishkoff, S.A., F.A. Reed, A. Ranciaro, B.F. Voight, C.C. Babbitt *et al*. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet **39**: 31–40.

Tsai, C.J., Z.E. Sauna, C. Kimchi-Sarfaty, S.V. Ambudkar, M.M. Gottesman *et al*. (2008). Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. J Mol Biol **383**: 281–291.

Tuffley, C. and M. Steel (1998). Modeling the covarion hypothesis of nucleotide substitution. Math Biosci **147**: 63–91.

Van der Hoorn, R.A., P.J. De Wit, and M.H. Joosten (2002). Balancing selection favors guarding resistance proteins. Trends Plant Sci **7**: 67–71.

Voight, B.F., S. Kudaravalli, X. Wen, and J.K. Pritchard (2006). A map of recent positive selection in the human genome. PLoS Biol **4**: e72.

Wagner, A. (2010). On the energy and material cost of gene duplication in *Evolution after gene duplication*, edited by K. Dittmar and D.A. Liberles. Wiley, New York.

Wang, E.T., G. Kodama, P. Baldi, and R.K. Moyzis (2006). Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Natl Acad Sci USA **103**: 135–140.

Wang, H.C., M. Spencer, E. Susko, and A.J. Roger (2007). Testing for covarion-like evolution in protein sequences. Mol Biol Evol **24**: 294–305.

Weir, B.S., L.R. Cardon, A.D. Anderson, D.M. Nielsen, and W.G. Hill (2005). Measures of human population structure show heterogeneity among genomic regions. Genome Res **15**: 1468–1476.

Wichman, H.A., J. Millstein, and J.J. Bull (2005). Adaptive molecular evolution for 13,000 phage generations: a possible arms race. Genetics **170**: 19–31.

Wilke, C.O. and D.A. Drummond (2010). Signatures of protein biophysics in coding sequence evolution. Curr Opin Struct Biol **20**: 385–389.

Williams, P.D., D.D. Pollock, and R.A. Goldstein (2001). Evolution of functionality in lattice proteins. J Mol Graph Model **19**: 150–156.

Williamson, S.H., M.J. Hubisz, A.G. Clark, B.A. Payseur, C.D. Bustamante *et al*. (2007). Localizing recent adaptive evolution in the human genome. PLoS Genet **3**: e90.

Wong, W.S., Z. Yang, N. Goldman, and R. Nielsen (2004). Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics **168**: 1041–1051.

Woolley, S., J. Johnson, M.J. Smith, K.A. Crandall, and D.A. McClellan (2003). TreeSAAP: selection on amino acid properties using phylogenetic trees. Bioinformatics **19**: 671–672.

Xing, Y. and C. Lee (2006). Can RNA selection pressure distort the measurement of Ka/Ks? Gene **370**: 1–5.

Yang, Z. (2006). *Computational molecular evolution*. Oxford University Press, Oxford.

Yang, Z. and R. Nielsen (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol **17**: 32–43.

Yang, Z. and R. Nielsen (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol **25**: 568–579.

Yang, Z., R. Nielsen, and N. Goldman (2009). In defense of statistical methods for detecting positive selection. Proc Natl Acad Sci USA **106**: E95; author reply E96.

Yap, V.B., H. Lindsay, S. Easteal, and G. Huttley (2010). Estimates of the effect of natural selection on protein-coding content. Mol Biol Evol **27**: 726–734.

Zhang, L. and W.H. Li (2004). Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol **21**: 236–239.

Zhang, P., S. Mueller, M.C. Morais, C.M. Bator, V.D. Bowman *et al*. (2008). Crystal structure of CD155 and electron microscopic studies of its complexes with polioviruses. Proc Natl Acad Sci USA **105**: 18284–18289.

Zhou, T., D.A. Drummond, and C.O. Wilke (2008). Contact density affects protein evolutionary rate from bacteria to animals. J Mol Evol **66**: 395–404.

Zhou, T., W. Gu, and C.O. Wilke (2010). Detecting positive and purifying selection at synonymous sites in yeast and worm. Mol Biol Evol **27**: 1912–1922.

Zhu, L., and C.D. Bustamante (2005). A composite-likelihood approach for detecting directional selection from DNA sequence data. Genetics **170**: 1411–1421.