Towards a deeper understanding of protein sequence evolution

A Thesis submitted for the degree of
PhilosophiæDoctor.

Candidate:
Francesca Rizzato

Supervisor:
Alessandro Laio

October, 20^{th} 2016

Molecular and Statistical Biophysics Group
Ph.D. curriculum in Physics and Chemistry of Biological Systems

Table of Contents

Abstract

Most bioinformatic analyses start by building sequence alignments by means of scoring matrices. An implicit approximation on which many scoring matrices are built is that protein sequence evolution is considered a sequence of Point Accepted Mutations (PAM) (Dayhoff *et al.*, 1978), in which each substitution happens independently of the history of the sequence, namely with a probability that depends only on the initial and final amino acids. But different protein sites evolve at a different rate (Echave *et al.*, 2016) and this feature, though included in many phylogenetic reconstruction algorithms, is generally neglected when building or using substitution matrices. Moreover, substitutions at different protein sites are known to be entangled by coevolution (de Juan *et al.*, 2013).

This thesis is devoted to the analysis of the consequences of neglecting these effects and to the development of models of protein sequence evolution capable of incorporating them. We introduce a simple procedure that allows including the among-site rate variability in PAM-like scoring matrices through a mean-field-like framework, and we show that rate variability leads to non trivial evolutions when considering whole protein sequences. We also propose a procedure for deriving a substitution rate matrix from Single Nucleotide Polymorphisms (SNPs): we first test the statistical compatibility of frequent genetic variants within a species and substitutions accumulated between species; moreover we show that the matrix built from SNPs faithfully describes substitution rates for short evolutionary times, if rate variability is taken into account. Finally, we present a simple model, inspired by coevolution, capable of predicting at the same time the along-chain correlation of substitutions and the time variability of substitution rates. This model is based on the idea that a mutation at a site enhances the probability of fixing mutations in the other protein sites in its spatial proximity, but only for a certain amount of time.

Chapter 1

Introduction

Proteins: sequence, structure, function and evolution

Proteins are the main toolbox of cells and it is among them that we can find most of the instruments that carry out the fundamental tasks of life. Scissors, radio antennae, switches, sensors, hand-carts, fabric, torches, traffic lights and much more: a very well equipped toolbox. Indeed, some proteins provide structural support to cells and tissues, others promote chemical reactions, or carry messages between cells, or transport cellular cargoes around. Other specialized proteins act as antibodies, hormones or luminescence generators, and so on with all the small and big tasks that must be carried out in an organism.

The variety of functions accomplished by proteins may seem surprising when realizing that the vast majority of them are assembled from a set of only twenty different amino acid types joined one after another to form a chain. But, as a set of Lego bricks, these twenty amino acids duly combined can produce the rich variety of three-dimensional shapes that is necessary to perform all the aforementioned tasks. In fact, despite being constrained by a linear backbone, a chain of amino acids still maintains a big rotational and conformational freedom and so a big variety of three-dimensional structures may be attained. Moreover, even though all amino acid types share a common backbone by which they interlock, they all differ in the side chain and this difference provides them with precise physico-chemical properties and determines a higher or lower propensity to make specific chemical bonds and to interact with water (Alberts *et al.*, 2013).

So, given a sequence of amino acids, some three-dimensional conformations will be favored and other disfavored depending on the possibility to perform stabilizing

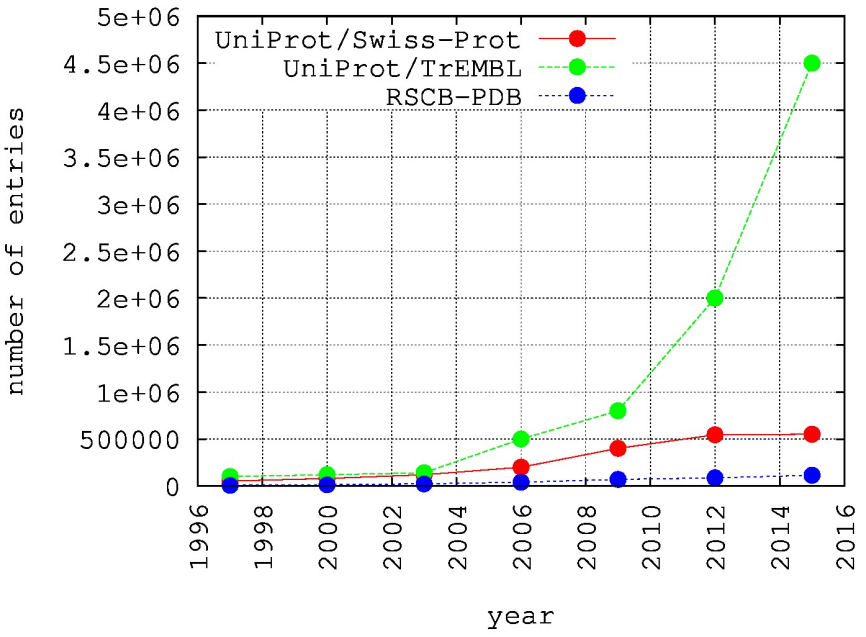

Figure 1.1: *Number of resolved sequences in UniProt (red for UniProt/Swiss-Prot, which is non-redundant and manually annotated, and green for the more general UniProt-TrEMBL) compared with the number of resolved structures in the Protein Data Bank (blue) in the last 20 years.*

chemical bonds between the side chains (e.g. hydrogen bonds) and on whether the hydrophobic residues are hidden from water or not. In the 1960s Anfinsen *et al.* (1961) proved that the most stable conformation of a protein in physiological conditions, often called *native conformation*, depends only on its sequence. Interpreting this finding from a physical perspective, we can say that each sequence folds into the conformation characterized by the minimal free energy. In the last 30 years many scoring functions, either based on physical principles (Srinivasan and Rose, 1995; Huang *et al.*, 1995) or on statistical inference (Miyazawa and Jernigan, 1985; Cossio *et al.*, 2012) have been developed to search for the native fold of a given protein sequence, but the results are still not completely satisfactory. And, if it is difficult to predict theoretically the structure of a protein sequence, determining it experimentally is not simpler. Among the techniques used for this purpose, two proved particularly successful over the years: X-ray cristallography (Kendrew *et al.*, 1958) and nuclear magnetic resonance (NMR) spectroscopy (Wuthrich, 2001), but unfortunately they are both expensive and time-consuming.

In this framework of incertitude it is interesting to notice that some proteins with

different sequences fold into the same structure. So, by learning how to predict which sequences share a common native conformation, we can transfer the available information on structure and function of the known cases to many others. This operation can increase our knowledge considerably and almost for free, sequencing being much faster and cheaper than resolving a structure experimentally. The large impact that this procedure may have becomes evident when comparing the size of the two main databases respectively for protein sequences, UniProt (The UniProt Consortium, 2015), and for protein structures, Protein Data Bank or PDB (Berman *et al.*, 2000), in the last 20 years (see figure 1.1): up to now the PDB contains 115000 structures while more than 40 millions of sequences are available on UniProt.

But what do sequences folding into the same structure have in common? Can we detect them a priori? It is well known that sequence similarity is a hallmark of structural similarity, but often almost identical protein structures differ significantly in their sequence. In figure 1.2, for example, two proteins of the family of globins are structurally compared (Roberts *et al.*, 2006) proving to be highly superimposable even if their sequence identity is less than 30%. Indeed, the vast majority of the sequences resulting in the same structure are related by evolution. In other words, they derive from a common ancestral protein whose sequence is unknown but whose structure probably resembled much the present ones. So, structures seem to be conserved by evolution much more than sequences, probably because structure is intrinsically connected with function. Indeed, if the structure is disrupted by a harmful mutation, the protein will not be able to perform its task any more, generally leading to the death of the organism, while, if the sequence changes keeping the same structure, the function may not be seriously affected. In this way, many mutations can cumulate and be carried along evolutionary lineages. The groups of proteins that share common structure and function, due to a common ancestor, are generally known as *protein families* (Finn *et al.*, 2015). Or, if we want to see them the other way around, the proteins related by evolution, which consequently share structure and function, can also be called *homologous proteins*.

Developing a reliable approach capable of detecting these homologous protein sequences is an issue that has been object of investigation for at least four decades. Unfortunately, this task is far from trivial. These proteins, in fact, are connected to each other through family relationships that cannot be observed directly, since, in most cases, only the present generation is known. So, the reconstruction of their family tree, which goes under the name of *phylogenetic tree*, is the other side of the coin with respect to the search for homologues.

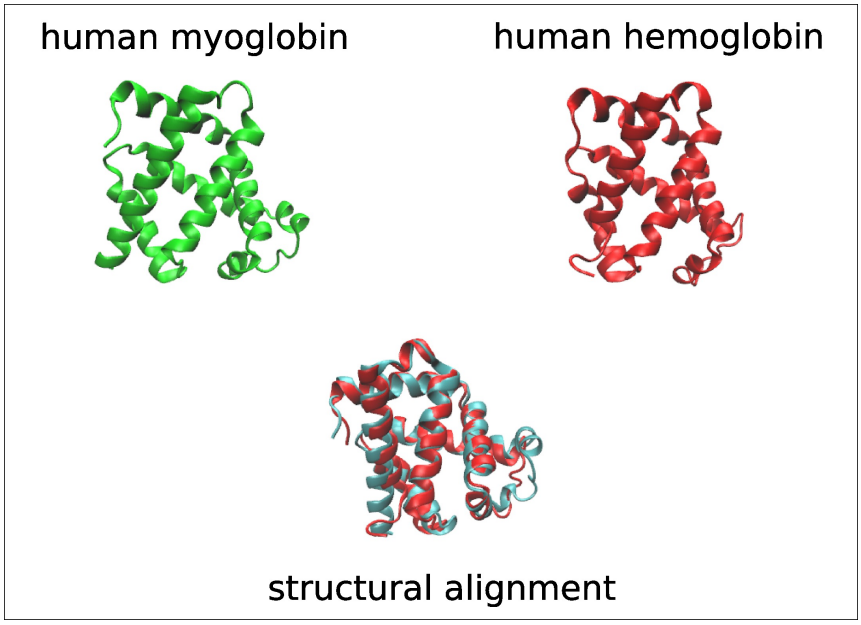

Figure 1.2: *Structural alignment (Roberts* et al.*, 2006) between human myoglobin (PDB-id: 3RGK) and human hemoglobin (chain alpha; PDB-id: 3W4U): the structure similarity is impressive even if the two proteins share less than one third of their sequences.*

To approach the problems mentioned above, the first essential ingredient is a method to align protein sequences. In particular, one needs both a scoring scheme and an alignment algorithm: the former associates to each alignment a score according to its probability of deriving from a common ancestor, and the latter, given two sequences and a scoring scheme, provides their best alignment.

Many algorithms have been developed to perform sequence alignment, each with slightly different targets. The easiest approach is to compare sequences two by two, choosing the best starting and ending point for the alignment and deciding the possible insertion of gaps into the sequences in order to maximize the score. More complex algorithms build multiple sequence alignments, analyzing many sequences at the same time, often exploiting site-specific properties. For what concerns scoring schemes, the simplest approach is based on testing whether two amino acids are identical, while more complex scores take into account the different probability of observing certain substitutions between amino acids. In fact, even if all amino acids have different physico-chemical properties, some of them are more similar than others and the impact of a mutation on the stability and functionality of a protein can be very different from case to case. Scoring matrices, containing 20x20 scores when

built for amino acid sequences, lie on the hypothesis that the knowledge of the amino acids involved in a mutation provides some information on its probability to be accepted or rejected by natural selection and so to be observed in an alignment. This is of course an approximation that averages over all protein specificities, but it is the approach used in the first steps of most bioinformatic analysis when no further specific information is available. It is therefore evident that the quality of a substitution matrix, and of the algorithm that uses, may influence all the following steps, where more precision is reached and required.

Thesis Outline

This thesis starts from a critical analysis of the theoretical framework on which substitution matrices are based and used, and proposes an attempt to overcome some of the difficulties that are intrinsic in this framework.

Many substitution matrices describe protein sequence evolution as a sequence of independent Point Accepted Mutations, whose acronym, PAM, usually refers to the pioneer matrix derived by Dayhoff *et al.* (1978). In this framework each substitution is assumed to happen independently of the history of the sequence, with a probability that depends only on the initial and final amino acids. Technically, one assumes that protein sequence evolution is a Markov process, defined by a set of transition probabilities between amino acids. Within this framework, the probability of observing a new sequence starting from a given initial sequence after a generic evolutionary time can be straightforwardly computed. Similarly, one can estimate the probability that a sequence has evolved from another one or that two sequences have evolved from a common ancestor. The transition rate matrix is generally learned from amino acid substitutions observed in alignments, often manually curated. In the original approach, the transition probabilities were inferred from alignments at high sequence identity (the lower threshold was set to 85% of sequence identity), but more advanced methods allow the use of alignments at any sequence identity, by estimating simultaneously their evolutionary time and the entries of the desired matrix. This more advanced approach is followed, for example, in ref. (Whelan and Goldman, 2001).

A first important assumption at the basis of this procedure is that substitutions are modeled at the level of amino acids. Instead, it is well know that mutations happen on the DNA sequence, so on nucleotides, and only after the process of translation affect the sequence of amino acids. This means that one should model

protein sequence evolution at the level of codons[1] rather than of amino acids, because, due to the redundancy of the genetic code, the two procedures lead to different outcomes. This issue has been addressed in ref. (Kosiol and Goldman, 2011), in which the importance of modeling protein sequence evolution on codons for building a meaningful substitution matrix was made evident.

A second important feature of protein sequence evolution, which is normally neglected in the derivation of substitution matrices, is that different protein sites evolve at a different rate (Echave *et al.*, 2016): some of them mutate with a very high probability, others hardly ever do. This effect is considered a signature of the different structural and functional importance of each protein site: residues localized in unstructured loops are generally more tolerant to mutation than others in the functional core of the protein. Rate variability is often taken into account when reconstructing phylogenies (Yang, 1994, 2007) or in some advanced tools for aligning multiple sequences (Halpern and Bruno, 1998; Pagel and Meade, 2004; Le *et al.*, 2008), but it is generally neglected when inferring scoring matrices or performing pairwise sequence alignments.

Another important feature of protein sequence evolution is that residues tend to coevolve. Namely, after the appearance of a mutation at a certain site, the residues in contact with it in the three-dimensional structure are more likely to accept mutations in order to recover the original structural and functional balance. This effect is so strong that it can be used to predict protein structures (de Juan *et al.*, 2013; Gobel *et al.*, 1994; Weigt *et al.*, 2009; Morcos *et al.*, 2011; Ekeberg *et al.*, 2013; Burger and Van Nimwegen, 2010). Even if this feature is well known, most aligning procedures consider sites to be independent from each other.

It is plausible that neglecting this effect, or the variability of the rates, can lead to systematic errors in the estimate of the evolutionary relationships and times. In this thesis we make an attempt to analyze the consequences of these features and to include them in a model of protein sequence evolution.

We will start by reviewing **(Chapter 2)** the main concepts and methods lying at the basis of our contributions. There, we introduce some notions on Markov processes and on their application to the study of protein sequence evolution; we

[1] A codon is a triplet of consecutive nucleotides which, during the process of translation of the messenger RNA into a protein, is translated to an amino acid. Each triplet of nucleotides corresponds to a single amino acid according to the map described by the genetic code. Due to the different dimension of the space of codons (64 possible states) and the space of amino acids (20 possible states) there are more codons coding for the same amino acid. These are called synonymous codons. Because of this degeneracy, the genetic code can be considered redundant.

present the most common methods for sequence alignment and for the reconstruction of phylogenies; we analyze the problem of substitution rate variability both among sites and in time and we discuss some basic notions in coevolution.

Chapters 3 and 4, which respectively deal with among site rate variability and Single Nucleotide Polymorphisms, are interlocked: actually our research was carried out on the other way around with respect to the order of presentation in this thesis. We started working on Single Nucleotide Polymorphisms (SNPs), variations of the DNA sequence within the same species involving a single nucleotide, in an attempt of building a scoring matrix for protein sequence alignments from them. Scoring matrices are rooted on the principle that knowing which amino acids are involved in a mutation can give some information on its probability of fixation and rejection. But, while substitutions are mutations already fixed by natural selection[2], polymorphisms are genetic variants present only in a fraction of the population. So, SNPs lie in principle in the no man's land after the occurrence of a random mutation and before its definitive fixation or rejection. We then selected only polymorphisms present in at least 1% of the human population and without a known clinical significance, in order to mostly collect nearly neutral polymorphisms. We then performed consistency tests aimed at proving if, and how much, the frequencies of amino acid interchanges in these polymorphisms and in substitutions from alignments at short evolutionary times are similar. In these tests, which we will describe in detail in **Chapter 4**, we found a good agreement, with no significant variation in the statistics, for interchanges that can occur by a single nucleotide variation, while, as expected by the definition itself, SNPs could not reproduce multiple nucleotide variations present in alignments. This seems to indicate that most of the frequent polymorphisms are indeed nearly-neutral on their effects on protein stability and functionality, and they can then be used to build a scoring matrix. Therefore, SNPs seem to provide a source of data that can be used to infer scoring matrices. The dataset derived from SNPs is also characterized by two interesting features with respect to standard datasets: SNPs are by definition isolated mutations and they are observed directly on the DNA sequence.
Contrary to our naïve expectations, our first scoring matrices based on the SNPs, while predicting quite accurately substitutions characterized by a single nucleotide variation, gave inaccurate predictions for substitutions associated to multiple nucleotide variations. This failure was the starting point for a deeper analysis on scoring matrices and protein sequence evolution. We realized that substitution rates, which are known to vary among sites and in time (Echave *et al.*, 2016; Gaucher *et al.*,

[2]A mutation is said to be fixed when it is present in all the individuals of a certain species.

2001; Lopez *et al.*, 2002), are implicitly assumed to be uniform when dealing with substitution matrices. By investigating the consequences of releasing this assumption, we observed that, when substitution rates are allowed to vary across the protein sites, the evolution of full protein sequences becomes effectively non-Markovian even if the single protein site still evolves by a Markov process. In **Chapter 3** we discuss this topic in detail and we quantify the effects of rate variability in a realistic case, showing its impact on the prediction of substitution probabilities and on the estimate of scoring matrices from alignments. In the light of these findings we could understand the failure of our initial attempt to build scoring matrices from the SNPs. We show that SNPs, when combined with an adequate treatment of the among-site substitution rate variability, can be successfully used to learn scoring matrices and score alignments at high sequence identity (80-100%) with performances comparable to the best scoring matrices used nowadays, with correct predictions also for the substitutions associated with multiple nucleotide variation. The results of this research can be found in **Chapter 4**. We also show that at medium and low sequence identities our matrices perform worse than the standard ones, indicating that, in the medium-low sequence identity range, the information embedded in the alignments becomes important for building a reliable model. Clearly, in this range, other effects start playing a role. For example, at medium sequence identity one can no longer assume that rates are constant during the whole evolutionary time.

We then focused more thoroughly on substitution rates and their variability in time and along the protein chain. Inspired by the work of Fitch and Markowitz (1970), we focused on the relationship between rate variability and coevolution, showing that these ingredients alone can already account for some observable features. We then developed a simple model to describe the time evolution of the substitution rates along the protein chain, explicitly including spatial and temporal interdependence among sites. The main idea is that the probability of fixing a substitution at a site is enhanced if this site is structurally in contact with other recently mutated sites, mimicking the mechanism of compensatory mutations. Importantly, and at variance with standard models of coevolution, we assume that this mechanism acts only for a finite time: if compensatory mutations do not appear in a few generations the effect of the initial substitution is effectively forgotten. With this model we accurately reproduce the experimental patterns of along-chain conditional probability of observing a substitution in the proximity to another one, in a wide range of sequence identity. Moreover, the number of substitutions per site produced by our model is well described by a negative binomial distribution, as generally found in

phylogenetic analysis. The shape parameter estimated by our model grows with the sequence identity, a feature also observed in real protein families. This model predicts that substitutions take place in *avalanches* localized not only in three-dimensional space, as commonly predicted by coevolution, but also in time. We found qualitative signatures of this phenomenon in the sequence evolution of the viral protein Influenza Hemagglutinin. These results seem to foster the hypothesis that the variability of substitution rates is strongly connected with coevolution and that a mutation triggers indeed the acceptance of other mutations in nearby sites for a limited time. This research is described in detail in **Chapter 5**.

Chapter 2

Methodological background

In this chapter we review the main concepts and contributions which served as starting points for our original work. One first section gives some notions on Markov processes and their use in the modeling of protein sequence evolution, which are fundamental for the comprehension of chapters 3 and 4. The second section concerns more directly the core of protein sequence analysis: alignments and trees. The third section addresses the problem of the variability of substitution rates both among sites and in time, which are the starting points respectively for chapters 3 and 5. The last section discusses some basic concepts of coevolution, which will be useful for the comprehension of chapter 5.

2.1 Markov models of protein sequence evolution

Since the first quantitative studies on the evolution of DNA and proteins, Markov processes provided evolutionary biologists with a strong statistical framework for embedding their models. In this section we are going to briefly describe them in the formalism that we will adopt in the rest of the thesis.

In general, Markov processes are characterized by the assumption that, at any time, the future evolution depends only on the present state and not on the previous history. In particular, a continuous-time Markov process on a finite set of states is defined by the $N \times N$ instantaneous substitution[1] matrix $\boldsymbol{Q}$, where N is the number of possible states (Cox and Miller, 1977). When protein evolution is modeled at the amino acid level, the possible states are the 20 amino acids and $N^{AA} = 20$, while

[1] As commonly done in literature, in all the thesis we will call *mutation* the random occurrence of a change in the nucleotide (resp amino acid) sequence from one generation to the other, while we will call *substitution* this mutation once it has become fixed by the natural selection, with the majority of the individuals of a given species characterized by the new nucleotide (resp. amino acid) variant.

when the framework of codons is chosen, the possible states are the 61 codons coding for amino acids[2] and $N^c = 61$.[3]

Each off-diagonal entry of $\boldsymbol{Q}$, $Q_{i,j\neq i}$, represents the instantaneous substitution rate from state i to state j and, in evolutionary models, it is often assumed to be constant in time, attaining a time-homogeneous Markov process. The diagonal entries are defined as $Q_{ii} = -\sum_{j\neq i} Q_{ij}$ and account for the instantaneous probability of escaping from each state. Often $\boldsymbol{Q}$ is normalized so that $\sum_i \sum_{j\neq i} (\pi_i Q_{ij}) = 1$, where π_i is the equilibrium probability of state i. When this normalization is used, the time is measured in units of expected substitutions per site. For example, $t = 0.01$ corresponds to a typical rate of substitution of 1%, constant along the protein chain.

Given a matrix $\boldsymbol{Q}$ describing a time-homogeneous Markov process, the transition probability[4] from state i to state j in a time interval t is given by:

$$P_{ij}(t) = \left[e^{t\cdot\boldsymbol{Q}}\right]_{ij} \tag{2.1}$$

where the exponential of matrix $t \cdot \boldsymbol{Q}$ is defined by the exponential power series.

For a homogeneous-time Markov processes defined on a discrete space, such as the set of 61 codons, the Chapman-Kolmogorov equation (Papoulis and Pillai, 2002) says that

$$\boldsymbol{P}(t+s) = \boldsymbol{P}(t) \cdot \boldsymbol{P}(s) \tag{2.2}$$

where the symbol $\cdot$ must be intended as the matrix product. This allows the deduction of the transition probabilities after M time steps from the transition probability matrix at one step $\mathcal{P}$ by a simple exponentiation:

$$\boldsymbol{P}_M = (\mathcal{P})^M \tag{2.3}$$

Adherence to the Chapman–Kolmogorov equation is a necessary and sufficient condition to verify that a process satisfies the Markov assumption. We will see in

[2]The remaining three are stop codons, which are neglected in the dynamics by assuming that any substitution between a stop codon and a coding codon would be detrimental and would be consequently always rejected.

[3]In all the thesis the superscript ^{c} (resp. ^{AA}) will be reserved to codon related (resp. amino acid related) quantities.
All matrices will be indicated in bold, such as matrix $\boldsymbol{Q}$, when the symbol refers to the full matrix. We will indicate them in normal math mode, such as Q_{ij}, when we refer to specific entries of the matrix.

[4]Note that we will call transition probability (or probability of substitution) from state i to state j the sum of the probabilities of every evolutionary path starting from state i at time 0 and finishing in state j at time t, independently from all what happens in between.

section 2.2.1 how this property was exploited in the definition of PAM matrices and in chapter 3 we will employ it to test if a process is Markovian.

2.1.1 Non-Markovian evolution at the level of amino acids

In a recent paper, Kosiol and Goldman (2011) have shown with the aid of the Chapman-Kolmogorov equation that, if the sequence evolution at the level of codons is described by a Markov process, then the evolution at the level of amino acids violates the Markov assumption.
Indeed, if the sequence evolution is described by transition probabilities defined by equation 2.1 on the set of coding codons $C = \{AAA, AAC, \ldots, TTG, TTT\}$, each amino acid can be modeled as an aggregate state containing the codons coding for it, and the observable process at that level can be called an aggregated Markov process[5]. The transition probabilities from an amino acid A to another amino acid B in a time interval t can then be computed by

$$P_{A,B}(t) = \frac{\sum_{\{c1 \in A\}} \sum_{\{c2 \in B\}} f_{c1} P_{c1,c2}(t)}{f_A} \tag{2.4}$$

where f_{c1} is the equilibrium frequency of codon $c1$, $f_A = \sum_{c1 \in A} f_{c1}$ is the equilibrium frequency of amino acid A and $\{c1 \in A\}$ stands for the set of codons coding for amino acid A.
It can then be easily proved that this class of transition probabilities on amino acids does not satisfy the Chapman-Kolmogorov equation because, for instance, $P_{A,B}(t) \neq [P_{A,B}(t_0)]^{t/t_0}$ and so the evolution at the level of amino acids is not Markovian. In other words, the transition probability from amino acid A to B does not depend only on the amino acid state A, but also on the hidden state corresponding to the present codon. For example the transition from lysine, coded by AAA and AAG, to isoleucine, coded by ATA, ATC and ATT, depends on the initial hidden codon of lysine: if it is AAA then a single nucleotide substitution is enough to transform it into isoleucine while if it is AAG two nucleotide substitutions will be necessary, with a large reduction of the transition probability for short evolutionary times.

This fact strongly encourages modelers to rather describe evolution at the codon level and is the main reason why in the next chapters this procedure will be preferred.

[5] Aggregated Markov processes are a class of stochastic processes which belong to the class of hidden Markov models

2.2 Protein sequence alignment and phylogenetic tree reconstruction

Protein sequence alignment represents the first step of most bioinformatic analysis and one of its most interesting applications is phylogenetic tree reconstruction.

The process of aligning sequences aims at recognizing regions of similarity due to evolutionary relationships and providing the best rearrangement of the sequences by determining which positions should be paired. In this process gaps in each sequence are generally allowed, provided that a penalty is paid in the score. For example, a possible alignment between the two sequences $PEGRWDD$ and $HEGGQWE$ looks like this:

$$PE-GRWDD$$
$$HEGGQWE-$$

In this alignment, some residues are conserved (such as E_2 or G_4)[6], some others are substituted by chemically similar residues (e.i. D_7 with E_7) and others are deleted or inserted (G_3 and D_8).

To align sequences one needs a scoring function and an aligning algorithm. The former associates each alignment with a score according to the ratio between its probability to come from the evolution of a common ancestor and its probability to have arisen by chance; the latter, given the sequence and a scoring function, provides their best alignment. We will now quickly review these two topics and then describe the methods used to reconstruct phylogenies.

2.2.1 Scoring alignments

Given two aligned sequences, we need a scoring function capable to associate them a score. In particular we need a function associating a score to each pair of aligned amino acids and another function deciding a penalty for the appearance of gaps in the alignment. The easiest way to solve the first problem is by learning average substitution probabilities between amino acids, building scoring matrices upon them and using them as a scoring function for amino acid substitutions. This is quite a rough scoring method because neglects all kinds of site-specificity, but it is the

[6]Here the subscript labels the position of the letter in the alignment, so E_2 is the letter E found in position 2

only available one when no a priori information is available on the studied system. Therefore it enters in the first steps of almost all kinds of analysis, included the most precise ones. We will now describe in detail the building process of such scoring matrices, then sketch more complicated scoring systems and briefly discuss the scoring of gaps.

Scoring matrices

Scoring matrices can be computed on amino acids, codons or nucleotides and, for each pair of states (i, j), they provide a score $s_{i,j}$ that grows with the relative probability that i and j are aligned because they descend from a common ancestor rather than by chance.

The scoring matrices on amino acids are the most common to score protein evolution and they can roughly be divided into two main categories: PAM-like matrices and BLOSUM-like matrices. The first ones treat evolution as a Markov chain of independent Point Accepted Mutations (from which the acronym PAM) and put all their effort in the description of either one-step transition probabilities or, equivalently, of instantaneous rates of substitutions between amino acids, formalized respectively in a $\mathcal{P}$ or in a $\boldsymbol{Q}$ matrix in the notation of section 2.1. BLOSUM matrices, on the contrary, do not make any assumption on the evolutionary process behind a substitution and are directly learned from the probabilities of observing a certain amino acid interchange in a learning set of multiple sequence alignments. In the work where BLOSUMs were introduced, Henikoff and Henikoff (1992) analyzed *blocks* (Henikoff and Henikoff, 1991) of conserved sequences in multiple sequence alignments, where in each block the segments with a sequence identity above a certain threshold were clustered, to reduce the impact of closely related sequences. This threshold determines how much the impact is reduced and is an important feature of the matrix. BLOSUM62, for example, is the matrix obtained when the segments sharing more than 62% of their sequences are clustered and so on. Clearly, matrices with lower thresholds (typically BLOSUM40 and BLOSUM50) are best fit for distant related proteins, whereas higher thresholds (BLOSUM80 and BLOSUM90) are used to score alignments of higher sequence identity. Nevertheless, this procedure never sets an explicit lower threshold of sequence identity. Therefore, even BLOSUM90 is derived from mismatches between sequences in conserved blocks which may have 30% of sequence identity. As we are going to see in chapter 4 section 4.3, this makes BLOSUM matrices generally inappropriate when aligning very similar sequences,

while they perform much better when searching for distant homologues.

Despite the different methodological approach between PAM and BLOSUM matrices, their scores $s_{i,j}$ are expressed in both cases as *log-odds*:

$$s_{i,j} = k \cdot log\left(\frac{expected\, frequency\, from\, evolution}{expected\, frequency\, from\, null\, model}\right) \tag{2.5}$$

where k is a scaling constant and the expected frequency from null model is, in general, the product of the equilibrium frequencies of state i and state j and accounts for the probability to find i and j aligned by chance. The numerator, on the other hand, measures the expected frequency to find i and j aligned according to the model (f_{ij}^{model}). This gives:

$$s_{i,j} = k \cdot log\left(\frac{f_{ij}^{model}}{f_i \cdot f_j}\right) \tag{2.6}$$

The base of the logarithm is often set to 2 but this has not a big importance as it just scales all the scores by a common prefactor without affecting the ranking of the different alignments. The scores are generally also multiplied by a prefactor and rounded to the nearest integer to give integer scores. This is just a convention and so, when we compute scores in chapter 4, for simplicity we will keep them as they are defined in equation 2.5, with *log* standing for the natural logarithm.

We now quickly review the principal PAM-like scoring matrices on amino acids and on codons that we will use in the following chapters as benchmarks or examples. We also describe the building process of standard PAM matrices that we are going to partially reproduce in chapter 4.

PAM and PAM-like matrices

The original *Point Accepted Mutation* (*PAM*) matrices were introduced in 1968 by Dayhoff and Eck and then recomputed with the same procedure but on a larger dataset in 1978 (Dayhoff *et al.*, 1978). In both cases they collected manually curated alignments with more than 85% of sequence identity and computed the transition probabilities between amino acids from the observed number of interchanges. Given the number of observed interchanges $n_{i,j}$ between the amino acids i and j, they computed the equilibrium frequencies of the amino acids as:

$$f_i = \frac{\sum_j n_{i,j}}{\sum_{j,k} n_{j,k}} \tag{2.7}$$

and the frequency of observing each pair i, j of aligned amino acids as

$$f_{i,j} = \frac{n_{i,j}}{\sum_{k,l} n_{k,l}} \tag{2.8}$$

In this formalism, the transition probability from state i to state j is given by

$$p_{i,j} = \lambda \frac{f_{i,j}}{f_i} \tag{2.9}$$

where λ is a scaling constant. These transition probabilities $p_{i,j}$ were scaled to give the one-step transition probability matrix $\mathcal{P}$, in which on average one mutation occurs per 100 protein sites. Hence, $\mathcal{P}$ is obtain by choosing λ in equation 2.9 so that $\sum_i \sum_{j \neq i} f_i \cdot p_{i,j} = 0.01$. Then, the diagonal entries are computed as $\mathcal{P}_{i,i} = 1 - \sum_{j \neq i} \mathcal{P}_{i,j}$ The log-odd of the transition matrix $\mathcal{P}$ gives the PAM-1:

$$PAM\text{-}1_{i,j} = k \cdot log\left(\frac{f_i \cdot \mathcal{P}_{i,j}}{f_i \cdot f_j}\right) \tag{2.10}$$

while the generic PAM-M matrix can be obtained by raising $\mathcal{P}$ to the power M and similarly doing its log-odd. Here, PAM matrices associated with a large M (typically PAM-250) are used to score distant alignments, whereas small values of M (often PAM-120) are used for shorter evolutionary times.

The rescaling from p to $\mathcal{P}$ implicitly assumes that each mismatch observed in the learning set of alignments at more than 85% of sequence identity is considered the outcome of a single accepted mutation. As we are going to see in chapters 3 and 4, this is a big approximation, because many of those mismatches are instead the result of multiple substitutions. We can easily test this by dividing the amino acids pairs into those that can intermutate by a single nucleotide variation (labeled *single*) and those which cannot (labeled *multiple*) and computing a conservative estimate of the fraction of multiple transitions in $\mathcal{P}$ by counting the transitions necessarily involving multiple substitutions. These are the substitutions between amino acids whose codons always differ from each other by at least two nucleotides. Extra multiple substitutions could also come from interchanges where the amino acids can intermutate by both single and multiple nucleotide substitutions. This procedure gives:

$$f_{multiple} = \sum_{(i,j) \in multiple} f_i \cdot \mathcal{P}_{i,j} = 0.19 \tag{2.11}$$

This estimate is times bigger than all the estimates of the fraction of instantaneous

multiple substitutions found in literature, which give 0.003 (Smith *et al.*, 2003), 0.02 (Averof *et al.*, 2000) or at most 0.03 (Schrider *et al.*, 2011).

In 1992 Jones, Taylor and Thornton recomputed a PAM-like matrix (labeled *JTT* after their names) by using almost the same approach as Dayhoff and Eck but on a much larger dataset, obtaining a more precise statistics, and in the same year also Gonnet *et al.* (1992) published an improved version of the PAM matrices obtained by matching the entire database of sequences available at that time.

In the following years other PAM-like matrices were obtained by analyzing alignments at all sequence divergences and detecting at the same time the evolutionary distances of the alignments and the entries of the instantaneous substitution matrix $\boldsymbol{Q}$ (see section 2.1) by maximum likelihood. The most famous example was published by Whelan and Goldman and named WAG after them (Whelan and Goldman, 2001), but also Müller and Vingron (2000) provided a relevant contribution with their resolvent method. These approaches improve the available statistics by exploiting alignments at all evolutionary divergences and also remove the previously unavoidable bias toward short evolutionary times.

Recently, Le and Gascuel (2008) proposed a modified version of the WAG matrix where, in the maximum likelihood framework, each site is also allowed to have a different overall rate of substitution. We will see in chapter 3 that accounting for this among-site rate variability allows a better estimate of the instantaneous substitution matrix.

Notice that, while the original implementation of PAM matrices (Dayhoff *et al.*, 1978) aimed at the construction of one-step substitution probabilities, $\mathcal{P}$, these last implementations prefer the formulation based on instantaneous rates ($\boldsymbol{Q}$ matrix). In this thesis we will also adopt mostly this last formulation.

Scoring matrices on codons

All the scoring matrices described until now treat evolution at the level of amino acids. However, as shown in section 2.1.1, models on codons should be preferred. Therefore, we now briefly discuss some codon models.

Contrary to the scoring matrices for amino acids analyzed so far, the majority of scoring matrices on codons are mechanistic rather than empirical: they are often based on theoretical assumptions and only some parameters are free to vary. This is probably due to the fact that a substitution rate matrix of 61x61 entries needs much more statistics than that necessary for a 20x20 matrix and so, in the past, empirical

models were out of reach. One of the simplest versions of mechanistic model on codons is the M0 model defined by the following instantaneous rate matrix (Yang *et al.*, 2000):

$$Q^c_{i,j\neq i} \propto \begin{cases} 0 & i \text{ or } j \text{ stop codons} \\ 0 & i \to j \ > 1 \text{ nucl. subst.} \\ \pi^c_j & i \to j \text{ syn. transv.} \\ \pi^c_j \kappa & i \to j \text{ syn. transit.} \\ \pi^c_j \omega & i \to j \text{ nonsyn. transv.} \\ \pi^c_j \kappa\omega & i \to j \text{ nonsyn. transit.} \end{cases} \tag{2.12}$$

where π^c_j is the equilibrium probability for codon j, κ is the transition/transversion[7] rate ratio and ω is the nonsynonymous/synonymous[8] rate ratio.

Recently, also a certain number of empirical or semi-empirical codon matrices have been developed: Kosiol *et al.* (2007) have obtained their Empirical Codon Model (ECM) by a procedure very similar to the one used by Whelan and Golman for amino acids, Schneider *et al.* (2005) have used on codons the same approach that Gonnet *et al.* (1992) had adopted for amino acids, and Doron-Faigenboim and Pupko (2007) have proposed a combined codon model that assimilates empirical amino acid replacement probabilities but still taking into account theoretical assumptions such as the transition-transversion bias.

Scoring insertions and deletions

Besides analyzing substitutions between residues, most of the aligning algorithms allow the insertion of gaps in alignments, which, from an evolutionary point of view, correspond to insertions or deletions (often shortened to *indels*) of residues with respect to the ancestral sequence. Therefore, a score must provide also this possibility. In most cases indels are simply treated as *gaps*, losing their evolutionary meaning, and associated to a penalty in the score. In such models the penalty either grows linearly with the number of neighboring gaps or distinguishes between the opening of a gap (the most penalized) and the extension of an already existing one (affine

[7] Transitions are interchanges of two purines ($A \longleftrightarrow G$) or two pyrimidines ($C \longleftrightarrow T$) and therefore involve DNA bases of similar shape, while transversions are interchanges between one purine and one pirimidine.

[8] A synonymous substitution is a substitution that leaves unchanged the coded amino acid, while a non-synonymous one makes it change.

gap model). Being these gaps introduced independently from an evolutionary model, their optimal penalties are generally empirically optimized on large datasets and differ from one scoring matrix to the other. A more precise approach for the inclusion of gaps is typical of algorithms based on hidden Markov models, where gap-penalties are not uniform along the alignment (Durbin *et al.*, 1998). So, while the number of parameters in scoring matrices only depends on the choice of codon or amino acids, the number of parameters for indels strongly depends on the underlying model.

2.2.2 Algorithms for sequence alignment

Depending on the constraints that they are asked to satisfy, we can classify the algorithms to perform sequence alignment into different categories. The number of sequences that they are asked to treat divides them in *pairwise sequence alignment* and *multiple sequence alignment* (MSA) algorithms, whereas they are defined *global* when they require every residue in every sequence to be aligned with something (gap or residue) or *local* otherwise, where only part of the sequences may be present in the alignment.

Pairwise sequence alignment

The most common tools to perform pairwise alignment are based on *dynamic programming*, which is a method for solving complex problems, often optimizations, by decomposing them into a collection of simpler sub-problems. Each sub-problem is then solved just once and its solution is stored so that, the next time it occurs again, its result can be retrieved without recomputing it, thus saving computational time. Given a scoring scheme, dynamic programming guarantees to find one of the optimal alignments.
Examples of this kind of algorithms are the *Needleman-Wunsch* (Needleman and Wunsch, 1970) for global alignments and the *Smith-Waterman* (Smith and Waterman, 1981) for local ones. They are both based on the compilation of a matrix such that in figure 2.1, where each column corresponds to one letter of the first sequence and each line to one letter of the second sequence. Starting from the cell at the top-left corner, to which the initial score 0 is assigned, the score for each cell in the matrix can be progressively computed from the knowledge of the scores in the three top-left neighboring cells. Indeed, when searching for a *global alignment*, the desired score

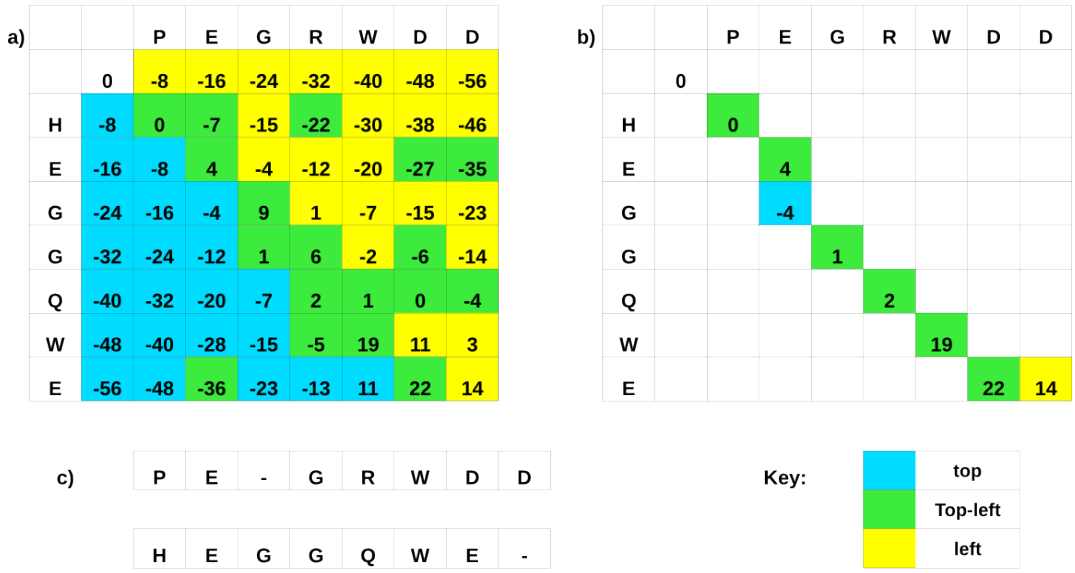

Figure 2.1: *Steps of the aligning procedure with the Needleman-Wunsch algorithm. a) Matrix for the construction of the alignment between the sequences $PEGRWDD$ and $HEGGQWE$ obtained with scoring matrix PAM-250 and gap penalty -8. The color code of each cell tells whether its score comes from the top cell (blue), the left one (yellow) or the top-left (green); b) Matrix containing only the optimal path obtained by backtracking from the bottom-right cell; c) Obtained alignment.*

may derive from either of these three directions according with this criterion:

$$S_{(i,j)} = max \begin{cases} S_{(i-1,j-1)} + M(x_i, y_i), & \text{top-left diagonal} \\ S_{(i-1,j)} + p, & \text{top} \\ S_{(i,j-1)} + p & \text{left} \end{cases} \tag{2.13}$$

where $S_{(i,j)}$ is the score associated to the cell in line i and column j, x_i (resp. y_j) is the amino acid occupying position i (resp j) in the second (resp. first) sequence, $M(x_i, y_j)$ is the value of the scoring matrix for the amino acid pairs (x_i, y_j) and p is the gap penalty.

When the best score $S_{(i,j)}$ is computed starting from the top-left diagonal (green cells in fig. 2.1), this entails the alignment of the two letters corresponding to line i and column j, while when it comes from top or left (respectively blue and yellow cells in fig. 2.1) it represents an insertion/deletion in the alignment. While repeating this procedure iteratively on each cell, it is also necessary to keep track of where each cell's score was computed from, as done in figure 2.1 by means of the colors.

When the full matrix is complete, the optimal global alignment (Needleman-Wunsch algorithm) is obtained by backtracking, from the bottom-right cell, the path of the optimal scores with the help of the colors up to the top-left corner as shown in fig. 2.1 panels b) and c).

A small modification of this procedure gives the best *local alignment*. In particular, with the Smith-Waterman algorithm, the score of each cell $S_{(i,j)}$ is chosen as the maximum among the same three values described above in eq. 2.13 and 0: if 0 is chosen this corresponds to starting a new alignment. The best local alignment is chosen by applying the same traceback procedure described above but now starting from the cell with highest score, instead of the bottom-right corner, and stopping when a score 0 is reached.

Beside dynamic programming, also some *heuristic methods* have become very common in the domain of pairwise sequence alignment because, even if they do not guarantee to find the optimal solution, they are generally much quicker. The most famous algorithm of this family is probably *BLAST* (Altschul *et al.*, 1990), whose search is based on identifying a series of short non-overlapping subsequences - or *words* - and match them to the sequences of the candidate database.

Multiple sequence alignment

MSAs are computationally difficult to manage and most of their formulations lead to NP-complete optimization problems (Wang and Jiang, 1994). In fact, although dynamic programming is in principle extensible to many sequences, it gets extremely slow already for small numbers and is then rarely used for more than three or four sequences.

So, the modelers resorted to approximate methods which generally produce a MSA by first aligning the most similar sequences and successively adding less related sequences (*progressive* methods such as *ClustalW* (Thompson *et al.*, 2002)). Their results depend on the selection of the most related sequences and thus is sensitive to inaccuracies in the initial pairwise alignments. To overcome this problem, some algorithms repeatedly realign the initial group of sequences as well as adding new ones to the growing MSA (*iterative* methods such as Muscle (Edgar, 2004)). Some other aligning algorithms realize ad-hoc scoring schemes for each family by learning the parameters from the sequences of an initial (seed) multiple sequence alignment (*heuristic* methods such as *T-Coffee* (Di Tommaso *et al.*, 2011) and *Clustal Omega* (Sievers *et al.*, 2011)). The information contained in a MSA can be used to detect

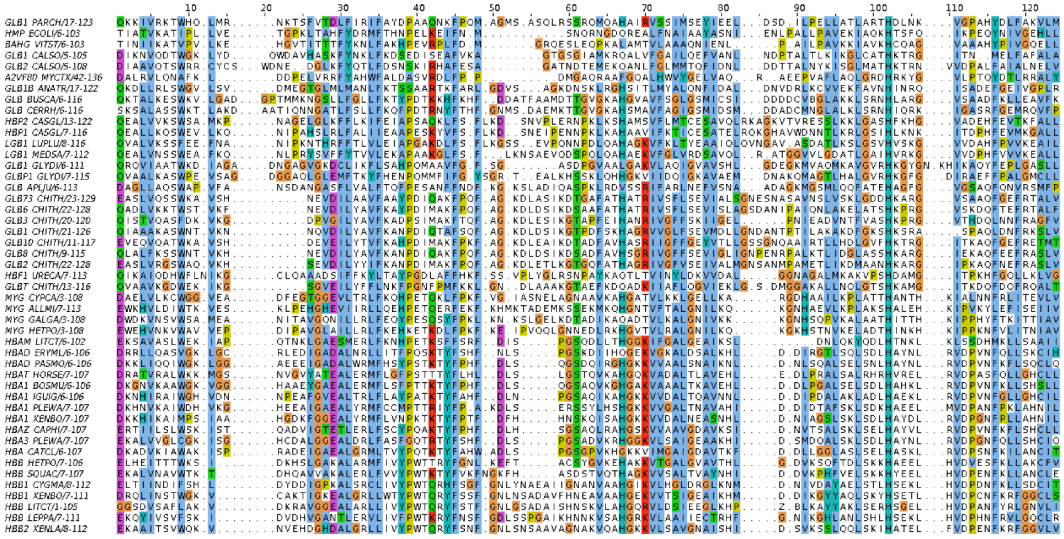

Figure 2.2: *Part of the MSA of the family PF00042 (globins) in Pfam (Finn* et al.*, 2015) viewed by the Java software Jalview (Waterhouse* et al.*, 2009). This alignment has been obtained by HMMer (Finn* et al.*, 2011), an aligning algorithm based on hidden Markov models. One can easily observe that, while some positions strongly vary, others are almost perfectly conserved.*

new members of a protein family by the use of position-specific scoring matrices as in *PSI-BLAST* (Altschul *et al.*, 1997) or hidden Markov models as in *HMMer* (Finn *et al.*, 2011). This last tool is the one used to classify protein domains into families in the renowned database *Pfam* (Finn *et al.*, 2015). See figure 2.2 for an example of MSA obtained by HMMer and figure 2.3 for the corresponding hmm logo.

2.2.3 Phylogenetic tree reconstruction

Besides being used to detect homology and find template structures for homology modeling, MSAs are also fundamental to reconstruct phylogenetic trees, which represent a hypothesis about the evolutionary ancestry of a set of genes.

The algorithms and softwares that perform tree reconstruction can roughly be classified into the following classes:

- In *distance-based* methods the distances between each pair of sequences is taken as a metric for clustering techniques, which are used to recursively divide the dataset. Among them, one of the most used algorithms is the *Neighbor-joining* method (Saitou and Nei, 1987), implemented, for example, in *QuickTree* (Howe *et al.*, 2002).

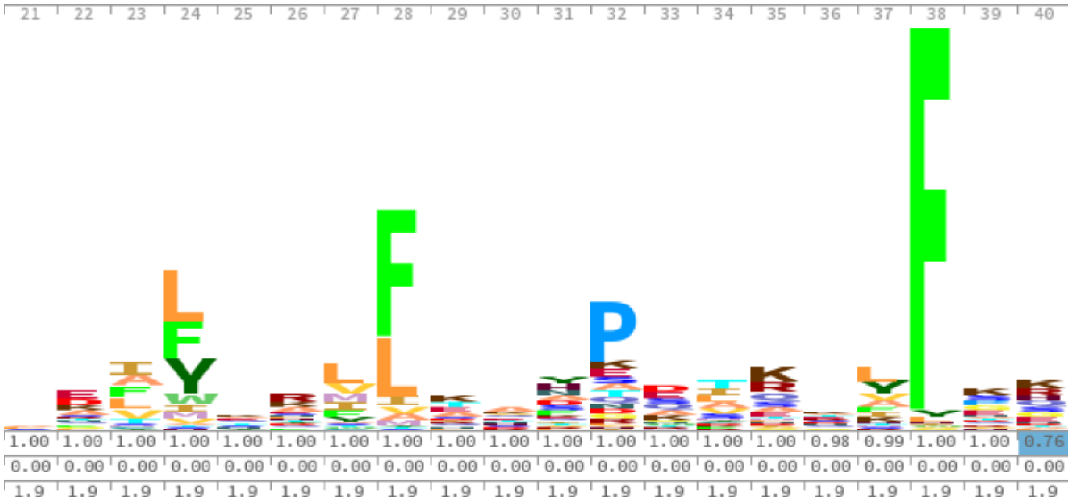

Figure 2.3: *Logo of part of the HMM of the protein family PF00042 (globins) in Pfam (Finn* et al.*, 2015) visualized by Skylign (Wheeler* et al.*, 2014). The total height of each stack corresponds to a measure of the invariance of the column, which corresponds to the information content of that position. The stack's height is spread among the letters based on their probability and letters are sorted such that those with larger probability appear near the top in the stack.*

- *Maximum parsimony* methods, instead, aim at finding the tree that requires the smallest total number of evolutionary events (substitutions, insertions and deletions) to explain the observed sequence data. The first implementation was Fitch's algorithm (Fitch, 1971).

- *Maximum likelihood* techniques use statistical methods to assign probabilities to phylogenetic trees. The "optimal" phylogenetic tree, then, is the one maximizing the likelihood of the data given that tree. In maximum likelihood approaches it is necessary to model sequence evolution, for example by a rate matrix $\boldsymbol{Q}$ such as the WAG matrix (Whelan and Goldman, 2001) or the ECM (Kosiol *et al.*, 2007) described in section 2.1. The *pruning algorithm* (Felsenstein, 1981), a variant of dynamic programming, is often used to reduce the search space by efficiently calculating the likelihood of subtrees. Many famous softwares for tree reconstruction, such as *FastTree*[9] (Price *et al.*, 2010), *PhyML* 3.0 (Guindon *et al.*, 2010), *RAxML* (Stamatakis, 2006) and *PAML* (Yang, 2007), belong to this class.

Besides those presented here, other methods can be used, such as compatibility meth-

[9] *FastTree* rather performs approximately-maximum-likelihood phylogenetics.

ods and Bayesian ones and even algorithms that simultaneously estimate alignments and phylogenies (Suchard and Redelings, 2006). An example of protein family tree is shown in figure 2.4.

2.3 Substitution rate variability

In the simplest formulations of protein sequence alignment as those based on dynamic programming and more in general in all those applications where scoring matrices such as PAMs or BLOSUMs are employed, all protein sites are assumed to evolve identically and independently. This assumption is clearly unrealistic, but it is almost necessary where no a priori information is available. In this section we present algorithms which progressively weaken these assumptions. The easiest part to release is the approximation that all sites have the same identical properties and the easiest way to release it is to allow each site to have a different overall rate of substitution. Indeed, there are overwhelming evidences that rates vary over sites (Uzzell and Corbin, 1971; Fitch, 1971; Echave *et al.*, 2016), with *fast sites*, where evolution act at high rate, and *slow sites* - sometimes even frozen sites - whose evolution takes a much longer time. In particular, it seems that there are few fast sites and many slow sites. In this framework, the transition probabilities on a site characterized by an overall rate r are described by:

$$P_{ij}(r,t) = \left[e^{r \cdot t \cdot \boldsymbol{Q}}\right]_{ij} \tag{2.14}$$

Notice that, even by letting the overall rate vary, we are still assuming that the evolution of each site happens independently from the others and that all sites maintain a common substitution model (i.e. a common matrix $\boldsymbol{Q}$). In the 1990s, Yang (1993, 1994, 1995) investigated deeply this phenomenon and found that the overall substitution rates r seem to follow a Γ distribution:

$$\Gamma(r) = \frac{\beta^{\alpha} \cdot e^{(-\beta \cdot r)} r^{\alpha-1}}{\Gamma^{E}(\alpha)} \tag{2.15}$$

where $\Gamma^{E}(\cdot)$ is the Euler gamma function. This is a flexible distribution characterized by two parameters: α, which determines the shape, and β, which fixes the mean of the distribution. In the present case we are dealing with overall rates that are then multiplied by the rate matrix $\boldsymbol{Q}$, so we can fix $\beta = \alpha$ to obtain a distribution with mean value $\langle r \rangle = 1$. As shown in figure 2.5, the variation of the shape parameter

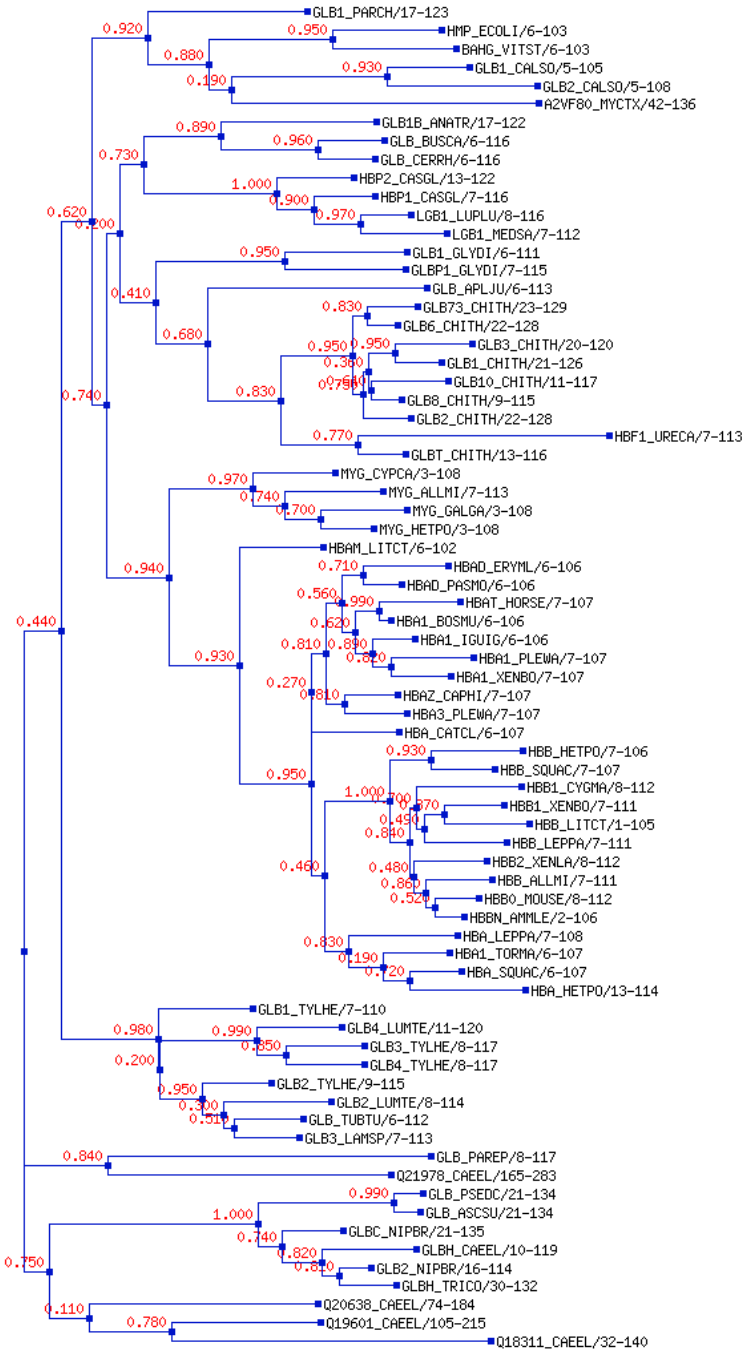

Figure 2.4: *Tree of the protein family PF00042 (globins) in Pfam (Finn* et al.*, 2015) computed by FastTree2 (Price* et al.*, 2010) and visualized by TreeView (Page, 2002). The names on the leaves are the identifiers of the sequences, while the red numbers are the bootstrap values for each node, which measure its reliability: the larger is the bootstrap value the more reliable is the node.*

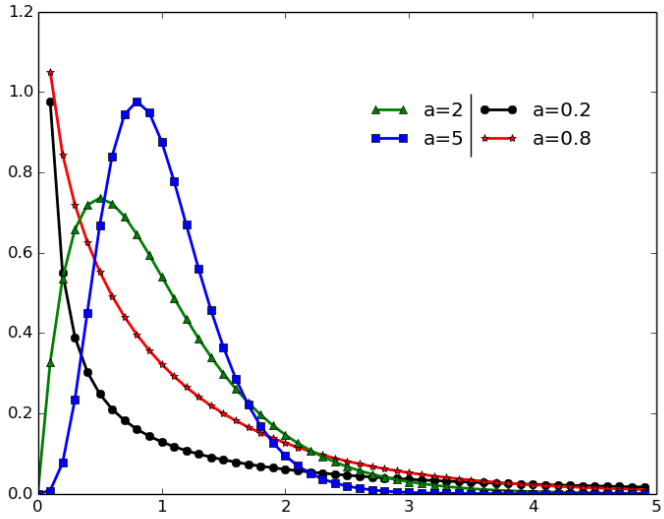

Figure 2.5: *Examples of* Γ *distributions with different values of* α *and* $\beta = \alpha$ *constraining the average value to be 1.*

changes the distribution from an L-shaped one ($\alpha < 1$) to an exponential ($\alpha = 1$), to a distribution that progressively resembles a skewed gaussian distribution when $\alpha > 1$. If the shape parameter is large, then the Γ distribution becomes a spike, with all rates being identical. In conclusion, the larger is α the more the rates resemble each other.

Using a maximum likelihood approach to infer phylogenetic trees, Yang proved that the inference benefits from the inclusion in the algorithm of the among-site variability of substitution rates, in particular when their distribution is modeled by eq. 2.15. Even if a continuous distribution as the Γ seems biologically reasonable, dealing with such distributions involves intensive computation which makes it often impractical. Therefore, Yang (1994) provided also an approximation of the Γ distribution performed by using a discrete number of categories. This *discrete-*Γ*-correction*, which often goes under the name of *Rates-Across-Sites (RAS)*, is nowadays implemented in most algorithms for phylogenetic tree reconstruction based on maximum likelihood (Stamatakis, 2006; Price *et al.*, 2010; Guindon *et al.*, 2010; Yang, 2007).

Some more complex algorithms for phylogeny reconstruction labeled as *mixture models* (Halpern and Bruno, 1998; Pagel and Meade, 2004; Le *et al.*, 2008) allow not only for different overall rates r but also for different substitution models (i.e. a different rate matrix $\boldsymbol{Q}$) at different sites. This has been done to account for the

across-sites heterogeneity in the pattern of evolution due to factors such as different equilibrium distribution of amino acids, different solvent exposure or involvement in proteins structure or function. When properly applied, this approach gives of course better predictions, but unfortunately needs a large number of aligned sequences to correctly estimate the larger number of parameters that it includes.

2.3.1 Estimating the shape parameter of the rate distribution from the number of substitutions per site

Given a protein family characterized by a MSA and a reconstructed phylogenetic tree, if we assume that the among-site distribution of the rates is well approximated by a Γ distribution, we are left with the problem of estimating its parameter α. From the phylogenetic tree we can compute an estimate of the number of substitutions per site (Fitch, 1971; Gu and Zhang, 1997). Then, if we assume that substitutions at each site are described by a Poisson process whose substitution rate r is taken from a Γ distribution, the number of substitutions per site (k) is described by a negative binomial distribution:

$$P(k|\alpha, \langle k \rangle) = \frac{\Gamma^E(\alpha + k)}{\Gamma^E(\alpha) \cdot k!} \left(\frac{\langle k \rangle}{\langle k \rangle + \alpha} \right)^k \cdot \left(\frac{\alpha}{\langle k \rangle + \alpha} \right)^\alpha \tag{2.16}$$

where again Γ^E is the Euler gamma function, α is the parameter of the Γ distribution and $\langle k \rangle$ is the average number of substitutions per site computed from the data. From eq. 2.16 the value of α can be estimated by maximum likelihood.

2.3.2 Heterotachy: the time variability of the substitution rates

With the introduction of the Γ-correction, we have released one of the unrealistic constraints of the simplistic description of protein sequence evolution as a set of identical and independent Markov processes: now the protein sites may differ in their overall rate of substitutions. But still, when deducing the shape of the rate distribution from eq. 2.16, we are implicitly assuming that each site maintains its own rate for the entire evolutionary process. In principle, there is no reason why this should be true, especially for long evolutionary times.

To check the validity of this assumption we can make the following test: we take a large MSA and build its phylogenetic tree (T_0) by FastTree2 (Price *et al.*, 2010) with

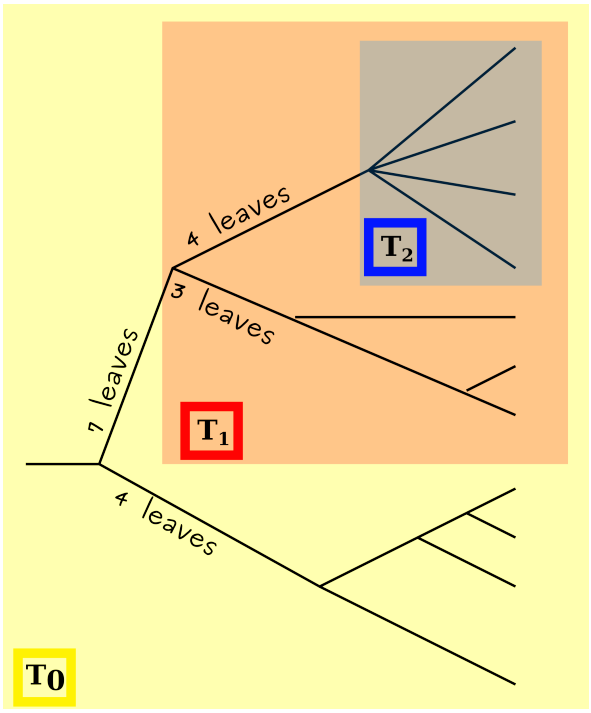

Figure 2.6: *Cartoon describing the procedure used in section 2.3.2 to progressively select the most populated subtrees from the root to the leaves of a given tree.*

the inclusion of the Γ-correction. Then, starting from the root of T_0, we label T_1 its 1-level subtree[10] containing the largest number of leaves. From the common ancestor of subtree T_1, we label T_2 its most populated 1-level subtree, and so on until the leaves are reached (see figure 2.6 for a visual description of this procedure). For each subtree we can compute the average sequence identity between the leaves, $\langle seqID \rangle$, and a new estimate of α recomputed from the sequences of that subtree only (again by FastTree2). If the rates remain constant in time, the estimates of α obtained from all subtrees should be compatible, whereas, if the rates change with time, we expect to find different estimations of α. In fact, if rates changes in time, the procedure described in section 2.3.1 would estimate time-averaged rates, and averages would be more and more uniform as the time lag over which they are averaged grows. Such an underestimation of the heterogeneity of the substitution rates (visually explained in figure 2.7) would produce an overestimation of the parameter α that get worse for larger evolutionary times. To assess this, we performed the described analysis on five Pfam families (Finn *et al.*, 2015) characterized by large MSAs and in all

[10] Given a tree, the level of each of its subtree is given by the number of edges between the root of the tree and the root of the subtree.

cases we observed an increase of α at low average sequence identity, namely at larger evolutionary times (figure 2.8).

A similar phenomenon was already observed by Gaucher *et al.* (2001) when analyzing the phylogenies of elongation factors: they observed that the value of α estimated on bacteria alone and on eukaryotes alone were almost one half of the α obtained when analyzing the two groups together. Also in that case the fake increase of α when computed on the full set of sequences seemed to be due to the mechanism described in figure 2.7.

These examples seem to suggest that in real sequence evolution there is a non-negligible amount of time-variation of the rate, especially for medium and low sequence identity. Already in the 1970s, indeed, Fitch and Markowitz (1970) proposed that at any given time only a restricted number of sites in the protein can fix mutations, but that these sites are not always the same. He suggested to label this idea as the *covarion* model, meaning that there are groups of *COncomitanly VARIable codONS*. In the following years many evidences emerged in favour of the idea that substitution rates vary in time and this phenomenon was given the name of *heterotachy* (Lopez *et al.*, 2002). If the first appearance of the covarion model was pretty qualitative and did not provide a recipe for its application, successive implementations were proposed by Penny *et al.* (2001) and by Galtier (2001) respectively in the framework of HMM and maximum likelihood. Anyway, to make this model easier to manage, they had to give up one of the basic properties of Fitch's original covarions: the covariation of protein sites. We will discuss the concept of protein site covariation from another perspective in the next section and we give our interpretation of this phenomenon in chapter 5.

2.3.3 Spatial covariation of substitution rates

Although the main implementations of the covarion model (Penny *et al.*, 2001; Galtier, 2001) neglected the among-site correlation of rates originally proposed by Fitch and Markowitz (1970), this feature was successively integrated into other models, but, in turn, these neglected the time variability of the rates, thus developing an orthogonal approach with respect to the one proposed in the previous section.

Yang (1995), after theorizing the Γ distribution for the substitution rates discussed above, proposed the *auto-gamma* model of evolution, where the among-site rate variation, still globally described by the discretization of a Γ distribution, is correlated by assuming a Markov dependence of rates at adjacent sites. Felsenstein and Churchill

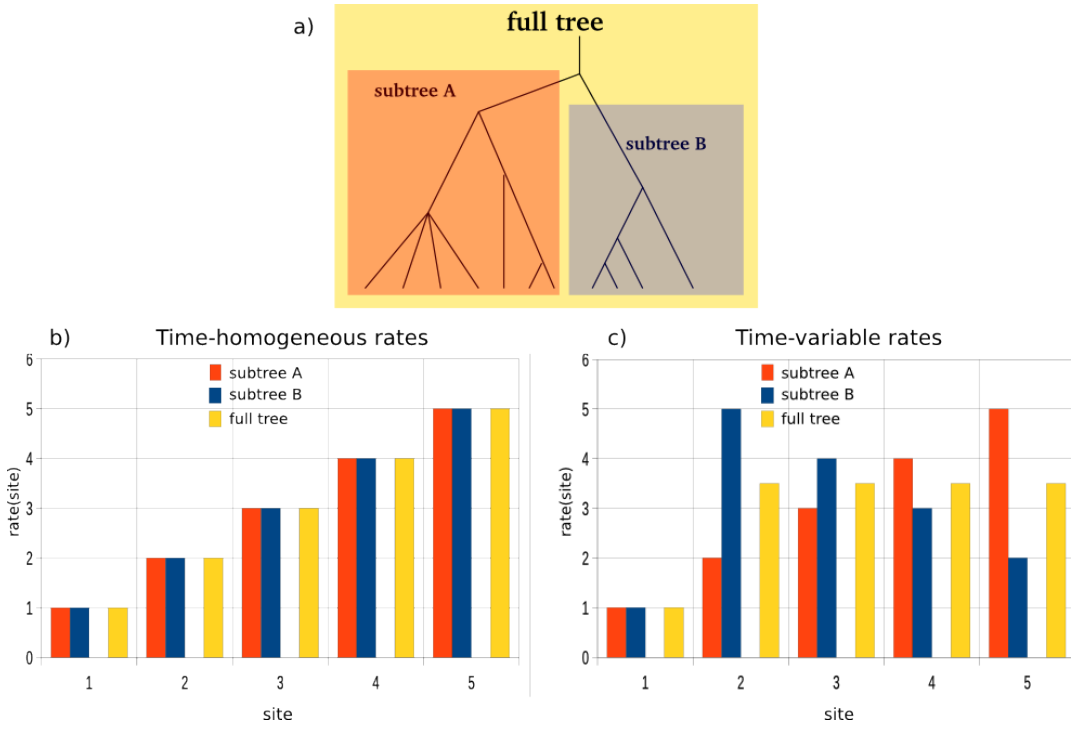

Figure 2.7: *Description of one of the possible mechanisms leading to the underestimation of the rate heterogeneity. Given a tree composed by two subtrees as in panel a), in panels b) and c) we characterize two possible scenarios: time-homogeneous rates and time-variable ones. In both cases 5 protein sites are analyzed and for each of them we show in the histogram the rate in subtree A (red), the rate in subtree B (blue) and the rate averaged on the two lineages (full tree, yellow). Notice that the distribution of the rates (and so their α) is the same in both lineages and in both panels. However, when the rates are homogeneous (case b) the estimates on the full tree coincide with those on subtrees A and B. When the rates differ in the two lineages (case c) the estimates on the full tree correspond to average rates, and this leads to apparently more uniform rates. The apparent uniformity induces a fake increase in the estimation of α. This mechanism was described by Gaucher* et al. *(2002) and used to explain their estimations of α for elongation factors in bacteria and eukaryotes (Gaucher* et al.*, 2001).*

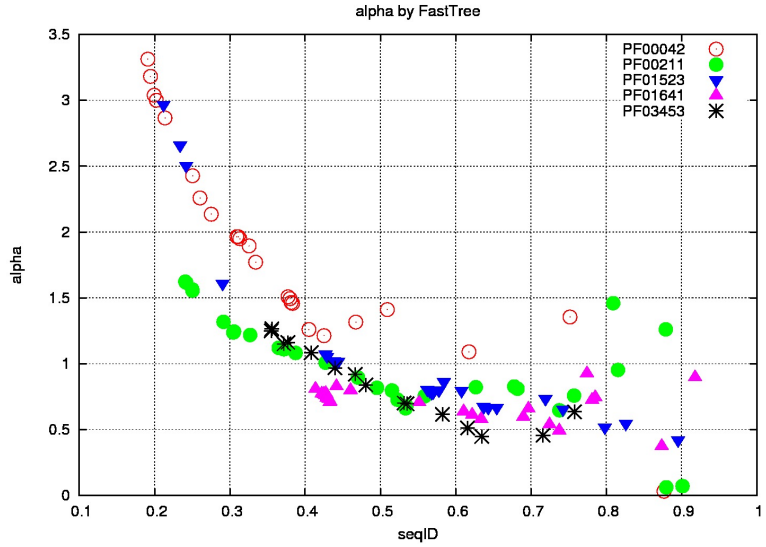

Figure 2.8: *Analysis on five Pfam families of the estimates of α obtained by FastTree-2 (Price* et al.*, 2010) on subtrees characterized by different $\langle seqID \rangle$. The procedure of selection of the subtrees is described in figure 2.6.*

(1996) also proposed a similar model, where the discrete-Γ distribution is embedded in the framework of a HMM along the protein sites. Still, both these approaches only allow a nearest-neighbor coupling among sites and, as already mentioned before, had to sacrifice the time-variability of rates to applicability.

We are going to see in the next section an approach that looks even more deeply into rate covariation, by describing the coevolution of spatially coupled sites by means of the covariation of the amino acids in those positions.

2.4 Coevolution of residues in a protein sequence

In the last 20 years much effort has been dedicated to investigate amino acid covariation, seeing in it the signature of the protein structure and function. Indeed, covariation[11] suggests the presence of *compensatory mutations* occurring between entangled residues (see figure 2.9). The back-of-the-envelope idea is that, when one site accidentally mutates, its contacts with the neighboring residues may get lost, modifying the original structure of the protein. To avoid this dangerous outcome, the residues in contact with the mutated site become prone to accept forthcoming mutations that re-establish the original balance. Because of this phenomenon sites i

[11]Here covariation is intended as the presence of correlated amino acid substitutions at different protein sites.

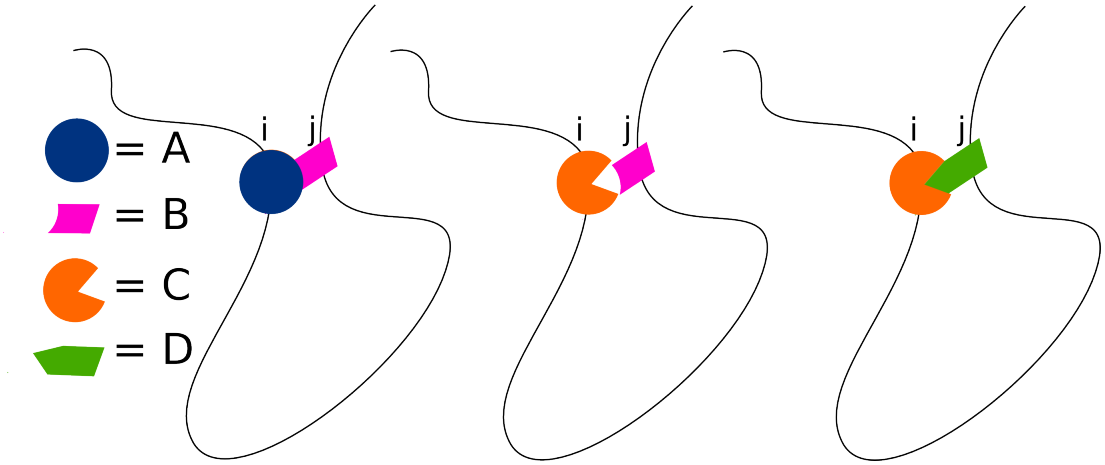

Figure 2.9: *Mechanism of compensatory mutations. Sites i and j are in contact. Initially site i is occupied by amino acid A and site j by amino acid B, whose physico-chemical properties are such that they form a stabilizing contact. At a certain point amino acid A is mutated into amino acid C, which can not bind B. To overcome this vulnerability, when B gets accidentally mutated into D, the mutation is easily fixed by selection if the contact between C and D is about as stable as the initial one between A and B.*

and j get entangled and, in presence of large ($\sim$ 1000 sequences) multiple sequence alignments, where the statistics on the amino acid occupancy of each site is copious, the covariation of amino acids at different sites can signal the presence of a three-dimensional contact between them. To detect these covariations (often labeled *residue covariations*) many computational tools have been developed. In parallel, other methods have been developed to search for larger groups of residues (*specificity-determining postitions*, or *SDPs*) co-conserved within protein subfamilies, which define functional properties specific for each subfamily, such as substrate binding specificity. We are here going to focus mainly on the first of these two mechanisms of coevolution, i.e. residue covariation, because in chapter 5 we will employ some related ideas.

Three main approaches have been developed for evaluating the coevolution between two residues from a MSA (de Juan *et al.*, 2013):

- by means of substitution correlation, where substitution patterns are built from an amino acids substitution matrix and their similarity are assessed by linear correlation (Gobel *et al.*, 1994; Fares and Travers, 2006);

- by analyzing the mutual information of amino acids frequencies (Fodor and

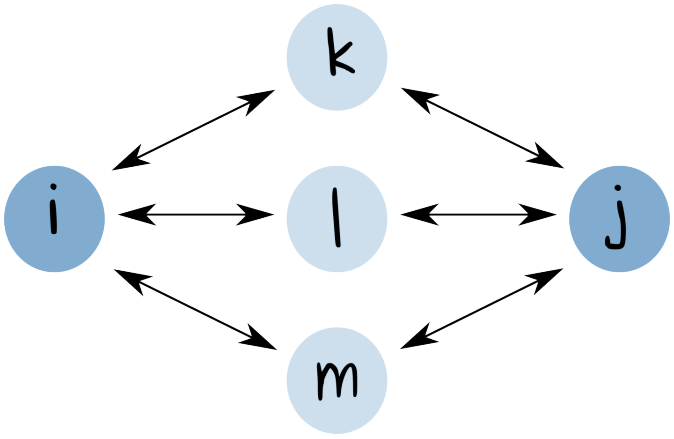

Figure 2.10: *Difference between direct and indirect couplings. The residues (labeled circles) connected by arrows are directly coupled so will tend to covary. These covariations will likely lead also to the covariation between site i and site j, even if they are not directly coupled (indirect covariation). Algorithms such as DCA (Weigt* et al.*, 2009; Morcos* et al.*, 2011) aim at disentangling direct coupling from this more complex pattern of covariation.*

Aldrich, 2004; Dunn *et al.*, 2008);

- by a global statistical model of the MSA such as in *Direct Coupling Analysis* (*DCA*) (Weigt *et al.*, 2009; Morcos *et al.*, 2011; Ekeberg *et al.*, 2013; Cocco *et al.*, 2013) or in *PSICOV* (Jones *et al.*, 2012) or in the Bayesian framework described by Burger and Van Nimwegen (2010).

These last ones, in particular, focus on disentangling directly coupled residues from the network of indirectly correlated positions. The observed covariation of two residues may indeed be due to a direct interaction between them, or by the mediation of residues at other positions that, being coupled with both of them, induce an indirect coupling (see the visual description of this mechanism in figure 2.10).

Even if recently Talavera *et al.* (2015) have questioned the assumption that the covariation observed in MSAs is a consequence of coevolution and have suggested that most of them are rather due to independent changes at highly conserved sites, these last methods seem to accurately predict the contact maps of protein families (de Juan *et al.*, 2013).

Chapter 3

Non-Markovian effects on protein sequence evolution due to site dependent substitution rates

3.1 Introduction

Since the publication of the work by Dayhoff and Eck (1968) introducing for the first time the concept of PAM matrices, protein sequence evolution has been typically modeled as a time-homogeneous Markov process and each protein site is assumed to be ruled by the same dynamic laws and to evolve independently from the others and from its own past history. This concept, which is examined in detail in chapter 2, is a milestone in the modeling of protein evolution and is, for example, at the basis of several successful approaches for structure prediction. After Dayhoff's first paper, PAM matrices have been further developed and specialized by using larger datasets (Dayhoff *et al.*, 1978; Jones *et al.*, 1992) and different methods to infer the instantaneous substitution rate matrix (Gonnet *et al.*, 1992; Whelan and Goldman, 2001; Mueller *et al.*, 2002). However, the availability of larger and larger substitution datasets has started challenging this theoretical framework. For example, Benner *et al.* (1994) and Mitchison and Durbin (1995) observed qualitative differences in protein evolution at different sequence divergence, raising concerns on treating the substitution process as Markovian. More recently, Kosiol and Goldman (2011) proved that, if the substitution process is Markovian at the codon level, it is not Markovian at the amino acid level. With that paper it became evident that scoring matrices on codons (Kosiol *et al.*, 2007; Schneider *et al.*, 2005; Doron-Faigenboim and Pupko, 2007) should be preferred to those on amino acids, but it is still unclear up to which

point evolution at the codon level can be considered Markovian. In particular, scoring matrices both on amino acids and on codons tend to present high rates for double and triple instantaneous substitutions, i.e. substitutions between codons differing by more than one nucleotide or between amino acids whose codons differ all for more the one nucleotide. This phenomenon, according to the biochemical wisdom, does not seem realistic and may hint to further violation of the Markov assumption not kept into account even when describing the evolution at the codon level.

Another important result in the description of protein sequence evolution was obtained in 1993 by Yang (Yang, 1993; Yang *et al.*, 1994), who proved that the estimations of evolutionary distances and evolutionary trees improve if the variability of substitution rates among sites is accounted for. This rate variability, which is typically modeled by a gamma distribution (Yang, 1993; Yang *et al.*, 1994), is due to many effects, including different structural and functional constraints (Echave *et al.*, 2016) and coevolution inducing a coupling between substitutions at close-by sites (Yang, 1995; Felsenstein and Churchill, 1996; Fitch and Markowitz, 1970).

The importance of taking rate variability into account is widely recognized in phylogenetics and many methods have been developed to include it when dealing with large multiple sequence alignments (Halpern and Bruno, 1998; Pagel and Meade, 2004; Lartillot and Philippe, 2004; Yang, 1994). However, these findings are generally neglected when building scoring matrices or applying them to alignments where no further information on the rate distribution is available. One noteworthy exception is due to Le and Gascuel (2008), who improved the amino acid replacement matrix by Whelan and Goldman (2001) by incorporating the variability of evolutionary rates across sites, but still proposing a model on amino acids rather than on codons.

In this chapter we describe how the among-site variability of substitution rates, by allowing each site to evolve at a different rate, makes the evolution of full protein sequences effectively non-Markovian even if the single protein site still evolve by a Markov process. The observed non-Markovian behavior at the full-sequence scale can be seen as the consequence of a reduction in the state space: the full state space, consisting in the 64 codons on sites characterized by different rates, is implicitly reduced to the simple set of codons, independently of the specific rate of that site. This gives rise to ensemble average transition probabilities on the reduced state space which are not Markovian. The main consequence is that simple Markov models of protein evolution that neglect rate variability (such as PAM and PAM-like matrices), no matter if they are empirical or mechanistic and if they are developed at the codon or at the amino acid level, are affected by systematic errors that, for example, may

lead to underestimating the evolutionary times. We will also show that one of the effects of treating protein evolution as a Markov process is a general overestimation of instantaneous double and triple substitutions, which might explain the corresponding high values found in the most common instantaneous substitution matrices such as the Jones-Taylor-Thornton (JTT) (Jones *et al.*, 1992), the Whelan and Goldman (WAG) (Whelan and Goldman, 2001) and the Empirical Codon Model (ECM) (Kosiol *et al.*, 2007).

3.2 Methods

We describe the dynamics in the framework of codons by the M0 model introduced by Yang *et al.* (2000), which is characterized by the following instantaneous rate matrix:

$$Q_{i,j\neq i} \propto \begin{cases} 0 & i \text{ or } j \text{ stop codons} \\ 0 & i \to j \ > 1 \text{ nucl. subst.} \\ \pi_j & i \to j \text{ syn. transv.} \\ \pi_j\kappa & i \to j \text{ syn. transit.} \\ \pi_j\omega & i \to j \text{ nonsyn. transv.} \\ \pi_j\kappa\omega & i \to j \text{ nonsyn. transit.} \end{cases} \tag{3.1}$$

where π_j is the equilibrium probability for codon j, κ is the transition/transversion rate ratio and ω is the nonsynonymous/synonymous rate ratio. The parameters are set to their typical values for protein-coding DNA: $\omega = 0.2$, $\kappa = 2.5$ and the codon distribution π_i is chosen as in Kosiol and Goldman (2011).
Starting from this $\boldsymbol{Q}$ matrix, the transition probability matrix $\boldsymbol{P}(t)$ between all the codon pairs at all evolutionary times can be computed as

$$P_{ij}(t) = \left[e^{t\cdot\boldsymbol{Q}}\right]_{ij} \tag{3.2}$$

3.2.1 Including rate variability by ensemble average transition probabilities

We now consider the effect on protein sequence evolution of a site-dependent substitution rate. Consistently with what proposed by Yang (1993), we assume that the

overall rate[1] of substitution, r, is distributed among sites according to a Γ-shaped probability density:

$$\rho(r) = \frac{\beta^{\alpha} \cdot e^{(-\beta \cdot r)} r^{\alpha-1}}{\Gamma^{E}(\alpha)} \tag{3.3}$$

where $\Gamma^{E}(\cdot)$ is the Euler gamma function. Unless otherwise indicated, in this chapter the shape parameter of this distribution is set to $\alpha = 0.286$. This specific choice of α is here irrelevant, being the scope of this work to provide a demonstration of the consequences of a plausible rate distribution on protein sequence evolution. Anyway, in section 3.3.4 we will analyze some results for different choices of α.

The transition probability from state i to state j in a time interval of t for a site characterized by an overall rate r is given by:

$$P_{ij}(r,t) = \left[e^{r \cdot t \cdot \boldsymbol{Q}}\right]_{ij} \tag{3.4}$$

When no information is available on the specific overall rate of each site, which is the typical premise when using general scoring matrices, we can score an alignment by comparing it to average transition probabilities. So we are interested in estimating the average probability for a site being in state i at time zero to be in state j at time t, considering that the rate distribution is given by eq. 3.3 and that each site evolves according to the Markovian dynamics described by eq. 3.4:

$$\begin{aligned} \widetilde{P}_{ij}(t) &= \int_0^{\infty} P_{ij}(r,t)\,\rho(r)\,dr \\ &= \left[\int_0^{\infty} e^{r \cdot t \cdot \boldsymbol{Q}} \rho(r)\,dr\right]_{ij} \end{aligned} \tag{3.5}$$

We will call $\widetilde{\boldsymbol{P}}(t)$ the *ensemble average transition probability matrix at time t*. Here the term *ensemble average* should be intended as an average over the ensemble of sites subject to the distribution of the substitution rate described by equation 3.3. We want to highlight that the definition of eq. 3.5 implicitly entails that each site is characterized by a substitution rate that remains constant over time. This is, of course, an approximation, because during evolution the propensity of a site to accept mutations may change (Lopez *et al.*, 2002), but, for short evolutionary times and in the range of sequence identity considered in this study ($\sim 80\%$), this approximation should hold. In fact, this is the same approximation implicitly used in the vast

[1] In this chapter we will consider separately the overall rate of substitution of a site (r) and the relative instantaneous rates of substitutions, which are grouped in the matrix $\boldsymbol{Q}$.

majority of phylogenetic algorithms for tree reconstruction, where each protein site is assumed to maintain the same rate along the branches of the full tree.

3.2.2 Non-Markovian behavior of ensemble average transition probabilities

According to eq. 3.5, the ensemble average transition probability matrix is a combination of many Markovian transition probabilities and, in general, combinations of non-identical Markov processes are not Markovian. In other words, even if here the single-site dynamics is assumed to be Markovian, when the full protein sequence evolution is approximated by neglecting site specificity, as done in general scoring matrices, the state space is implicitly reduced. But only some special reductions, with respect to which that process is *lumpable* (Kemeny and Snell, 1960) still give rise to Markov dynamics. With this in mind, the non-Markovian behavior of the full protein sequence dynamics can be simply proved either by checking that $\widetilde{\boldsymbol{P}}(t) \neq \left[\widetilde{\boldsymbol{P}}(\tau)\right]^{t/\tau}$, namely that $\widetilde{\boldsymbol{P}}$ violates the Chapman-Kolmogorov equation, or by exploiting the properties of lumpable Markov processes.

Proof of the violation of the Markov assumption by the properties of lumpable processes

Given a set of states $s = \{s_1, s_2, ...s_N\}$ and a partition on it $A = \{A_1, A_2, ...A_r\}$, a necessary and sufficient condition for a Markov chain on s to be lumpable with respect to A is that, for every pair of sets A_i and A_j, the sum $\sum_{s_l \in Aj} P_{s_k,s_l}$ of the transition probabilities from state s_k to states $s_l \in A_j$ has the same value for every $s_k \in A_i$ (Kemeny and Snell, 1960).

We exploit this property to prove the non-Markovian behavior of sequence evolution in presence of rate heterogeneity. Here the full state space is given by all the possible pairs $\{r, c\}$ with r a real number in $[0 : \infty[$ corresponding to a rate value and c one of the 64 codons. We consider the following transition probability from state s_1 to state s_2:

$$P_{\{r1,c1\},\{r2,c2\}} = \delta(r_1, r_2) \cdot \left[e^{r1 \cdot \Delta t \cdot \boldsymbol{Q}}\right]_{c1,c2} \tag{3.6}$$

where Δt is an arbitrarily small time.

We partition the state space into a 64-dimensional reduced space given by the set of possible codons: $A = \{c_1, c_2, ...c_{64}\}$, thus each set in A contains all the states characterized by a same codon and different rates. The dynamics described by

$P_{\{r1,c1\},\{r2,c2\}}$ is lumpable with respect to A only if

$$\int P_{\{r,c1\},\{r2,c2\}} dr_2 = \int P_{\{s,c1\},\{r2,c2\}} dr_2 \tag{3.7}$$

for all the possible r and s. But the first term gives $\int \delta(r, r_2) \cdot \left[e^{rt\boldsymbol{Q}}\right]_{c1,c2} dr_2 = \left[e^{rt\boldsymbol{Q}}\right]_{c1,c2}$ while the second term gives $\left[e^{st\boldsymbol{Q}}\right]_{c1,c2}$ which are equal only if $r = s$.
So, the dynamics of the reduced process, in presence of rate variability, is not Markovian.

3.3 Results

3.3.1 A simple example

In order to understand qualitatively the effects of the variation of the rate over sites, let us first consider a simplified world with only three codons, A, B and C, characterized all by the same frequency $\pi_A = \pi_B = \pi_C = \frac{1}{3}$. We assume that the instantaneous substitution matrix for this model is:

$$\boldsymbol{Q}^E = \begin{pmatrix} -1 & 0.9 & 0.1 \\ 0.9 & -1.1 & 0.2 \\ 0.1 & 0.2 & -0.3 \end{pmatrix}$$

If the rate of substitution is constant over sites, the transition probability matrix at time t is

$$P^E(t) = e^{tQ^E} \tag{3.8}$$

and describes a Markovian dynamics. On the other hand, we can imagine a sequence where, still keeping the same average rate, half of the sites has a reduced substitution rate of 0.5 and the other half has a faster substitution rate of 1.5. For this second system the ensemble average transition probability matrix at time t will be

$$\widetilde{\boldsymbol{P}}^E(t) = \frac{e^{0.5 \cdot t\boldsymbol{Q}^E} + e^{1.5 \cdot t\boldsymbol{Q}^E}}{2} \tag{3.9}$$

It may be of interest to monitor the value of these two sets of transition probabilities as functions of time and compare them. Being the three codons equiprobable, $\boldsymbol{Q}^E$, $\boldsymbol{P}^E$ and $\widetilde{\boldsymbol{P}}^E$ are symmetric and so we can limit ourselves to control only 3 different transition probabilities: $P^E_{AB} = P^E_{BA}$, $P^E_{AC} = P^E_{CA}$ and $P^E_{BC} = P^E_{CB}$. In figure 3.1 we

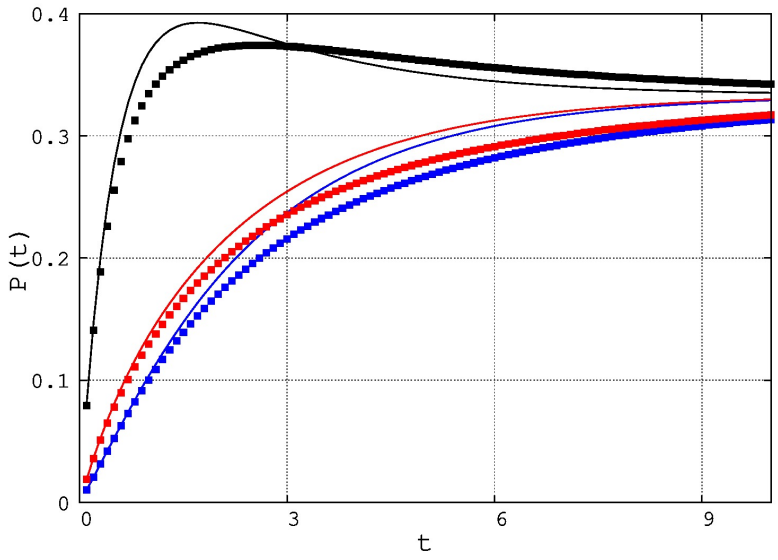

Figure 3.1: *Transition probabilities in a simplified world.*
Comparison between the transition probabilities in a sequence with constant substitution rate over sites and in a sequence with two equiprobable classes of rates for a simplified system described in Results. Black: $P^E_{AB}(t)$ (solid line) and $\tilde{P}^E_{AB}(t)$ (points); Blue: $P^E_{AC}(t)$ (solid line) and $\tilde{P}^E_{AC}(t)$ (points); Red: $P^E_{BC}(t)$ (solid line) and $\tilde{P}_{BC}(t)$ (points).

compare the time evolution of these three quantities (respectively in black, blue and red) in the two systems (in solid line for eq. 3.8 and points for eq. 3.9): clearly $\boldsymbol{P}^E(t) \neq \widetilde{\boldsymbol{P}}^E(t)$. So, it is evident that the variation of the rate induces a change in the average transition probabilities even if both $\boldsymbol{Q}^E$ and the average substitution rate do not change.

3.3.2 Non-Markovian behavior in the framework of codons

We now estimate the quantitative importance of the violation of the Markov assumption on the more realistic model described by the instantaneous substitution matrix defined in eq. 3.1 and the rate distribution of equation 3.3 (see sec. 3.2). Since now the rate distribution is continuous, the sum in equation 3.9 is replaced by an integral and gives equation 3.5. To quantify the variation of the ensemble average transition probability matrix with respect to the Markovian transition probability matrix, we compare $\boldsymbol{P}(t)$ to $\widetilde{\boldsymbol{P}}(t)$ at time $t = 0.235$, which corresponds, for $\boldsymbol{P}$, to the 80% of sequence identity. In figure 3.2(a) we show the entry-by-entry comparison between them in log-log scale: each point corresponds to a pair i, j of codons and its x-value is given by the Markovian evolution $P_{ij}(t)$, while its y-value is its non-Markovian

counterpart $\widetilde{P}_{ij}(t)$. If the two dynamics gave the same results, the points would lie on the line $y = x$, but this is not the case. In particular, one can see four separate subsets: the black squares (zoomed in the inset) are the entries corresponding to $j = i$ (the diagonal terms in the matrix), while the red, green and blue points correspond to $j \neq i$ (the off-diagonal ones), where i and j differ respectively by one, two or three nucleotides. It is evident that, with respect to the Markovian dynamics, $\widetilde{\boldsymbol{P}}(t)$ gives rise to higher entries for $j = i$, enhances double and triple substitutions, and discourages single ones.

A first consequence is that the expected sequence identity between two sequences separated by an evolutionary time t is lower for the Markovian dynamics than for the non-Markovian one. This happens because, even if the average rate of substitutions is the same, in the non-Markovian case it is much more likely that substitutions cumulate on the few sites with rate larger than 1. In this way, a much larger fraction of substitutions takes place on a site that has already mutated, without further modifying the global sequence identity. The Markovian assumption produces therefore a systematic underestimation of evolutionary times. This result may be considered the theoretical explanation of the observation by Yang *et al.* (1994) that, when taking substitution rate variability into account, one gets larger estimates of branch lengths in phylogenetic trees. The difference of sequence identity between two sequences separated by a given evolutionary time in the two processes can be found in fig. 3.2 (b). In particular, at the time $t = 0.235$ the non-Markovian dynamics presents the 85.7% of sequence identity, while the Markovian one only the 80%.

It is, then, more appropriate to compare the two processes at fixed sequence identity: in fig. 3.2(c), one can find the same comparison of fig. 3.2(a) with the time $\tilde{t}$ of the Non-Markovian process chosen to produce a sequence identity of the 80%, which gives $\tilde{t} = 0.4$. Even if this choice balances the entries corresponding to $i = j$, the non-Markovian dynamics still enhances double and triple substitutions with respect to its Markovian analogue. For example, at sequence identity of 80%, the estimated probability of finding a substitution from codon ATC to codon TGG (3 different nucleotide, so one of the blue points in figures 3.2(a)-(c)) is $5.21 \cdot 10^{-8}$ when the Markovian approximation is adopted, while is more than one hundred times bigger if the rate is Γ-distributed. In table 3.1 one can find some other examples of how transition probabilities change in the two frameworks.

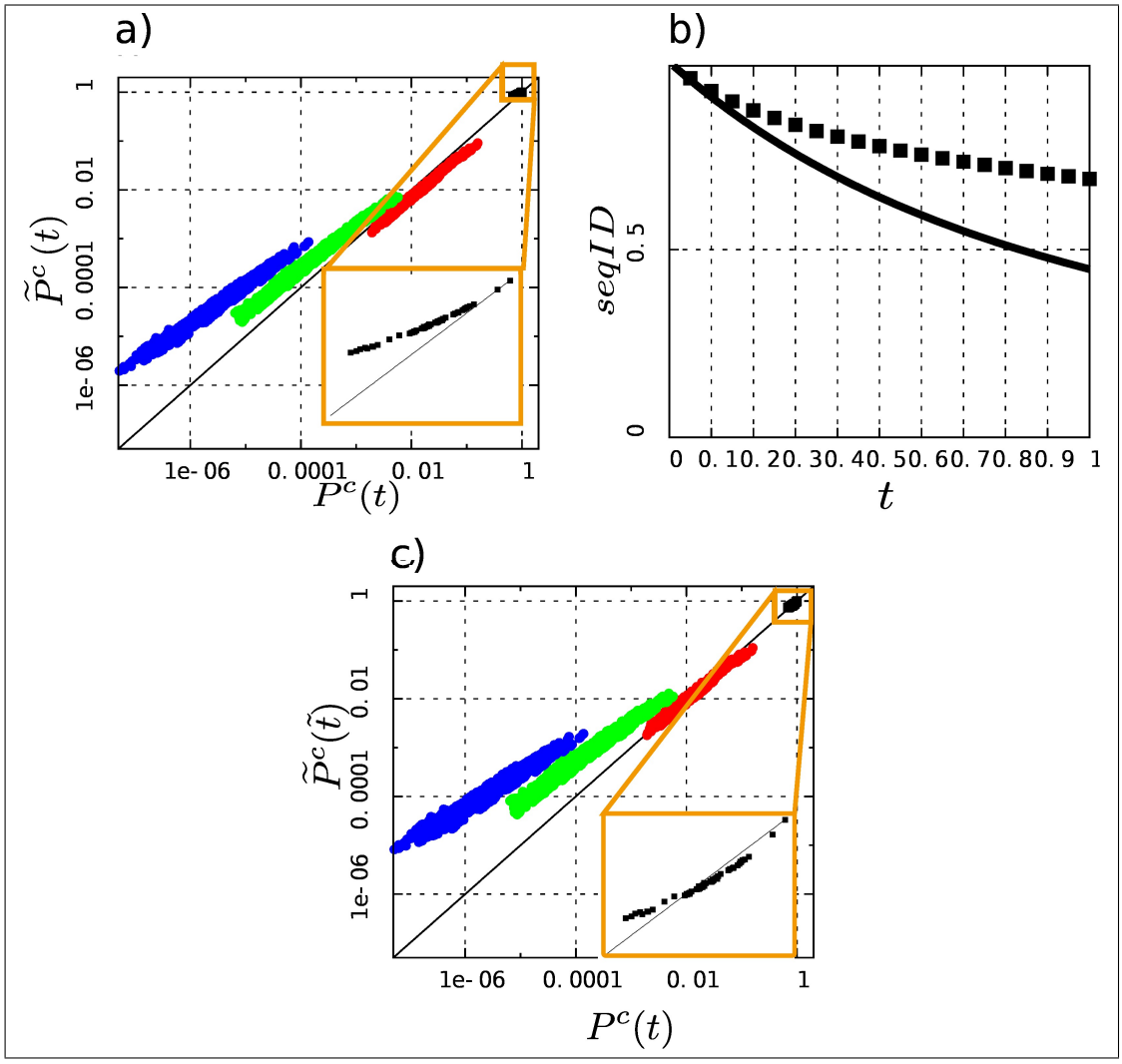

Figure 3.2: *Comparison between Markovian and non-Markovian substitution probabilities. (a) Points: entry-by-entry comparison of* $\boldsymbol{P}(t)$ *and* $\widetilde{\boldsymbol{P}}(t)$ *in log-log scale, with* $t = 0.235$*. Each point corresponds to a pair* i, j *of codons with* $x = P_{ij}(t)$ *and* $y = \widetilde{P}_{ij}(t)$*. The black squares (zoomed in the yellow inset) are the entries with* $i = j$*, while red, green and blue points are respectively the entries where codon* i *and codon* j *differ by one, two or three nucleotides. Solid line:* $y = x$*. (b) Comparison of the sequence identity as a function of* t *for the Markovian (*$I_M(t)$ *in solid line) and the non-Markovian (*$I_{NM}(t)$ *with points) dynamics. (c) Comparison of* $\boldsymbol{P}(t)$ *and* $\widetilde{\boldsymbol{P}}\left(\tilde{t}\right)$ *with* $t = 0.235$ *and* $\tilde{t} = 0.4$*, so that* $I_M(t) = I_{NM}\left(\tilde{t}\right) = 0.8$*. Coordinates, lines and colors have the same meaning as in panel (a).*

Table 3.1: *Examples of the variation of the transition probabilities at the sequence identity of 80% between Markovian ($\boldsymbol{P}$) and non-Markovian ($\widetilde{\boldsymbol{P}}$) dynamics.*

Initial state	Final state	$\boldsymbol{P}(t)$	$\widetilde{\boldsymbol{P}}(\tilde{t})$	$\boldsymbol{P}(t)/\widetilde{\boldsymbol{P}}(\tilde{t})$
ATC	TGG	$5.21 \cdot 10^{-8}$	$8.23 \cdot 10^{-6}$	0.006
TTC	ATG	$3.39 \cdot 10^{-5}$	$3.09 \cdot 10^{-4}$	0.110
GTC	GTT	0.1507	0.0951	1.58

3.3.3 Impact on the estimation of $\boldsymbol{Q}$ of the Non-Markovian behavior due to the rate variability

We now show that treating full protein sequence evolution as Markovian, neglecting substitution rate variability, determines also a wrong estimation of $\boldsymbol{Q}$, the instantaneous substitution matrix. In particular, we will see that, when learning $\boldsymbol{Q}$ from pairwise alignments, substitution rates between codons differing by more than one nucleotide are systematically magnified. This is somehow intuitive: rate variability allows substitutions to accumulate on the few sites with high substitution rate and so, when learning substitution frequencies from alignments, we find a larger number of double and triple substitutions than expected if the rates were constant. Then, when inferring $\boldsymbol{Q}$ from these data without taking rate variability into account, the only way to encompass the extra number of double substitution is to enhance instantaneous double and triple transition probabilities. For simplicity we are going to show this for a particular case, where $\boldsymbol{Q}$ is estimated from alignments all at the same sequence identity, but the reasoning can be generalized for alignments at various sequence identity and for multiple sequence alignments.

To evaluate the order of magnitude of this overestimation of instantaneous double and triple substitutions, we recover a measure of $\boldsymbol{Q}$, $\widetilde{\boldsymbol{Q}}(t)$, from the ensemble average transition probability matrix at time $\tilde{t} = 0.4$, $\widetilde{\boldsymbol{P}}\left(\tilde{t} = 0.4\right)$. If, when estimating $\widetilde{\boldsymbol{Q}}\left(\tilde{t}\right)$, we are considering the process as Markovian, for a sequence identity of 80% we would infer the evolutionary time being not $\tilde{t} = 0.4$ but rather $t = 0.235$ (see previous calculations and figure 3.2(b)). So we can calculate $\widetilde{\boldsymbol{Q}}\left(\tilde{t}\right)$ by inverting eq. 3.2:

$$\widetilde{\boldsymbol{Q}}\left(\tilde{t}\right) = \frac{log\left(\widetilde{\boldsymbol{P}}\left(\tilde{t}\right)\right)}{t} \tag{3.10}$$

with $\tilde{t} = 0.4$ and $t = 0.235$.

Figure 3.3 shows the entry-by-entry comparison between the original $\boldsymbol{Q}$ and $\widetilde{\boldsymbol{Q}}\left(\tilde{t}\right)$.

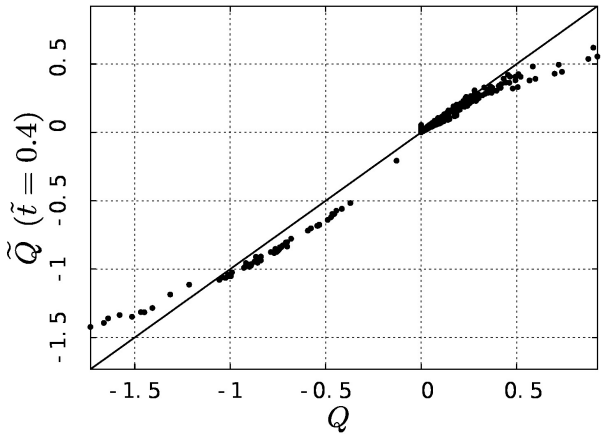

Figure 3.3: *Impact of the Markovian assumption on the estimation of* $\boldsymbol{Q}$*. Points: entry-by-entry comparison between* $\boldsymbol{Q}$ *and* $\widetilde{\boldsymbol{Q}}(\tilde{t}=0.4)$ *estimated by equation 3.10. Solid line:* $y=x$*.*

The two matrices clearly do not correspond, as the points do not lie on the line $y = x$, so the estimate of $\boldsymbol{Q}$ from alignments when neglecting rate variability is affected by systematic errors. In particular, we can calculate the fraction of instantaneous double substitutions in the original $\boldsymbol{Q}$, f_2^{true}, and in the estimated $\widetilde{\boldsymbol{Q}}(\tilde{t})$, f_2^{est}, by:

$$f_2^{true} = \frac{\sum_{i,j|\,2\neq nucl}\left[\pi_i \cdot Q_{ij}\right]}{\sum_{i,j\neq i}\left[\pi_i \cdot Q_{ij}\right]} \tag{3.11}$$

$$f_2^{est} = \frac{\sum_{i,j|\,2\neq nucl}\left[\pi_i \cdot \widetilde{Q}_{ij}\left(\tilde{t}=0.4\right)\right]}{\sum_{i,j\neq i}\left[\pi_i \cdot \widetilde{Q}_{ij}\left(\tilde{t}=0.4\right)\right]} \tag{3.12}$$

where π_i is the equilibrium probability of codon i and the sum at the numerator is the restricted sum over the entries involving a pair of codons i, j differing by two nucleotides. The fractions of triple substitutions for the original $\boldsymbol{Q}$, f_3, and for the estimated $\widetilde{\boldsymbol{Q}}$, f_3^{est}, are computed in a similar way.

In the original instantaneous rate matrix $\boldsymbol{Q}$ (equation 3.1) double and triple substitutions are not allowed, so $f_2^{true} = f_3^{true} = 0$ by construction, while, in the estimated matrix $\widetilde{\boldsymbol{Q}}\left(\tilde{t}\right)$, we get $f_2^{est} = 0.153$ and $f_3^{est} = 0.017$. So, the sum of the fractions of instantaneous double and triple substitution estimated from alignments at the 80% of sequence identity would make up the 17% of all the instantaneous substitutions, while in the original Markovian model they are the 0%.

This result might cast some light on the anomalous high entries for double and triple substitutions in the $\boldsymbol{Q}$ matrix of many models: the sum of the fractions of

instantaneous double and triple substitutions is 25% in the ECM, 22% in the WAG and 14% in the JTT matrix (for the last two, double substitutions are defined as the substitutions between amino acids whose most similar pair of codons differ by two nucleotides). Considering that mutations take place by chance, one would rarely expect double or triple substitutions to happen in an infinitesimal time on the same codon, which is also the underlying hypothesis in the definition of the mechanistic $\boldsymbol{Q}$ of equation 3.1. In the literature we found that the estimates of the ratio of double and triple instantaneous substitutions vary between 0.3% (Smith *et al.*, 2003) and 3% (Averof *et al.*, 2000) of the total. A possible explanation of the high value of double and triple substitution rates in standard scoring matrices is that the Markov assumption may have induced a fictitious increase for double and triple substitutions. A full proof of this idea would require recalculating $\boldsymbol{Q}$ from the same alignments used to build each matrix by including the rate variability. However, this explanation is consistent with two previous results: when the WAG matrix was re-examined by Le and Gascuel by including the Γ correction, they found smaller values for the triple substitutions (Le and Gascuel, 2008) and De Maio *et al.* (2012) observed that accounting for rate variability by hidden Markov Models reduces the estimate of instantaneous multiple substitutions in the ECM matrix.

3.3.4 Results for various values of α in the rate distribution

To further generalize our tests, we performed similar analysis with other values of α in the range $[0.2 : 1]$. These results are shown in fig. 3.4. It is evident that, despite the speficif choice of α, the rate variability always induces non-Markovian effects on the evolution and that this principally affects the estimation of double and triple substitution probabilities. On top of this, we see that smaller values of α, which correspond to less homogeneous rates, correspond, as may be expected, to stronger differences between the Markovian and non-Markovian dynamics.

3.4 Discussion

We have discussed the effects of the among-site variability of substitution rates in the process of protein sequence evolution. The relative difference of the rates mixes Markov processes with different speed, which makes the process of full sequence evolution effectively not Markovian. The first consequence of the violation of the Markov assumption is a systematic underestimation of evolutionary distances. We

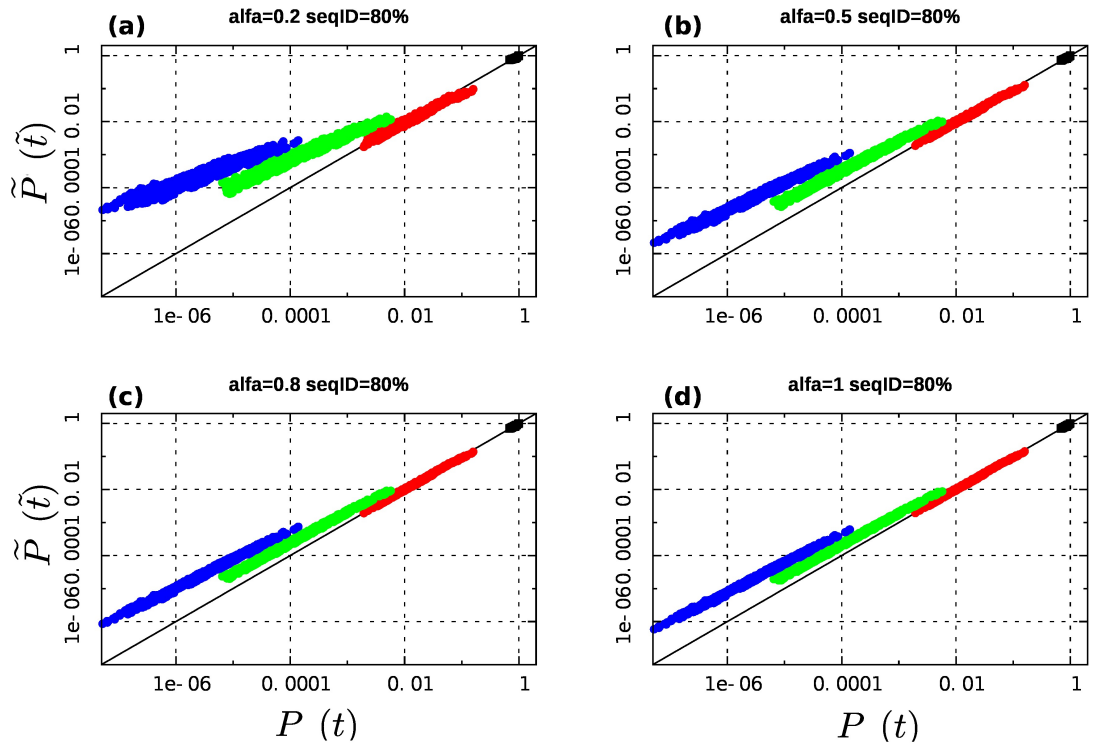

Figure 3.4: *(a) Entry-by-entry comparison of* $\boldsymbol{P}(t)$ *and* $\widetilde{\boldsymbol{P}}\left(\tilde{t}\right)$ *in log-log scale, for* t *and* $\tilde{t}$ *chosen so that* $I_M(t) = I_{NM}\left(\tilde{t}\right) = 0.8$*. Each point corresponds to a pair* i, j *of codons and its* x*-value is given by* $P_{ij}(t)$*, while its* y*-value is* $\widetilde{P}_{ij}(t)$ *where parameter* α *in the* Γ*-distribution is set to* $\alpha = 0.2$*. The black squares are the entries with* $i = j$*, while red, green and blue points are respectively the entries where codon* i *and codon* j *differ by one, two or three nucleotides. Solid line: line* $y = x$*. (b) Same comparison as in panel (a), but for* $\alpha = 0.5$*. (c) Same comparison as in panel (a), but for* $\alpha = 0.8$*. (d) Same comparison as in panel (a), but for* $\alpha = 1$*.*

have quantified the violation of the Markov assumption for a realistic model of evolution on codons, demonstrating that neglecting the rate variability may cause two orders of magnitude of difference in the relative probability for triple substitutions and one order of magnitude for double substitutions. We have also shown that this approach modifies in a radical way the estimate of $\boldsymbol{Q}$ itself by especially magnifying double and triple substitutions, which might explain the correspondent high transition probabilities in the main instantaneous substitution matrices (e.g. JTT, WAG, ECM).

Statistical inference of phylogenies under Markov models including Γ-distributed rate variation (Yang, 1993; Yang *et al.*, 1994) as well as CAT models (Lartillot and Philippe, 2004) can effectively deal with this problem. Mixture models (Halpern and Bruno, 1998; Pagel and Meade, 2004; Le *et al.*, 2008) and hidden Markov models

(Eddy, 1998; Krogh *et al.*, 1994; De Maio *et al.*, 2012), which allow not only site-dependent substitution rates but also site-dependent substitution probabilities, can go even beyond. However, the scoring matrices for codons and amino acids are most of the times derived without taking into account the among-site rate variability and these matrices enter necessarily even in the construction of the seed multiple sequence alignment at the basis of any hidden Markov Model. According to our findings, Markovian models for protein evolution based on most of the available scoring matrices are affected by errors that get worse when inferring information far from the learning set. This is valid both for models at the codon level and at the amino acid level, for which Kosiol and Goldman (2011) have already showed that a further source of memory is present.

The results shown here are robust with respect to the specific choice of the rate matrix and rate distribution: as can be guessed by the first simple example in Results, any non-trivial rate distribution combined in equation 3.5 with whatever $\boldsymbol{Q}$ gives rise to ensemble average transition probabilities $\widetilde{\boldsymbol{P}}$ which differ from the simple $\boldsymbol{P} = e^{t\boldsymbol{Q}}$. The results presented in fig. 3.2 should then be intended as a "proof-of-principle" that variable substitution rates cause a non-Markovian full protein sequence evolution and as a plausible estimate of the entity of the systematic errors arising when using standard substitution models in a naive way.

Even if further and more specific analysis would be necessary to quantify the impact of the effect described in this work on specific applications, the present results seem to discourage the use of simple Markovian models that neglect among-site rate variability for both amino acid and codon sequence alignments, especially when the scoring matrices are learned on alignments in a range of sequence identity very different from the test set. On the other hand, they encourage the use of models that account for among-site rate variability, for example mechanistic codon models with the Γ correction, Hidden Markov Models (Eddy, 1998; Krogh *et al.*, 1994; De Maio *et al.*, 2012), CAT models (Lartillot and Philippe, 2004) or other mixture models (Halpern and Bruno, 1998; Pagel and Meade, 2004; Le *et al.*, 2008) that allow it by construction, or substitution models of the type of Le and Gascuel (LG) (Le and Gascuel, 2008) that account for it explicitly.

In the next chapter we are going to apply the evolutionary model described by equation 3.5, which accounts for among-site rate variability and is based on the concept of ensemble average transition probabilities, to the special case of an instantaneous substitution matrix learned from a database of Single Nucleotide Polimorphisms.

Chapter 4

Using Single Nucleotide Polymorphisms to predict substitution probabilities between amino acids

4.1 Introduction

As we described in section 2.2, the quality of a protein sequence alignment is often measured by scores learned by assuming protein sequence evolution to be a succession of *point accepted mutations*. In this domain it is also generally assumed that these substitutions take place on independent protein sites obeying all to the same dynamic laws (Dayhoff *et al.*, 1978; Jones *et al.*, 1992; Schneider *et al.*, 2005; Gonnet *et al.*, 1992; Whelan and Goldman, 2001; Kosiol *et al.*, 2007). The recent improvements in the methods for DNA sequencing (Ronaghi *et al.*, 1996; Wheeler *et al.*, 2008; Bentley *et al.*, 2008) provide an increasing amount of sequenced genomes (1000 Genomes Project Consortium *et al.*, 2010, 2015), hence larger datasets to estimate scoring matrices. At the same time, these datasets also allow the identification of a large number of Single Nucleotide Polymorphisms (SNPs) (Sherry *et al.*, 2001), which are isolated single nucleotide variations in the DNA sequence that are present among the individuals of a species in a significant frequency.

In this chapter we are going to prove that frequent SNPs can be used to infer with high accuracy substitution probabilities among protein variants present in different species in the high sequence identity range. Describing DNA or protein sequence evolution as a succession of point mutations is an approximation which

neglects complicate evolutionary mechanisms such as gene duplication, exon shuffling and horizontal gene transfer and focuses on the two most common processes: the random appearance of a mutation in the DNA sequence, due to failures of the mechanisms of DNA replication or repair, and its consequent fixation or rejection by natural selection. The final destiny of a mutation depends on many factors and in particular on its positive or negative impact on the *fitness*[1] of its carrier. The principle on which scoring matrices and simple Markov models of evolution are rooted is that knowing which amino acids are involved in a mutation can give some information on its probability of fixation and rejection. But, especially for non-detrimental mutations, other factors are relevant in their probability of fixation, which pertain more to population genetics than to biochemistry. In fact, the destiny of a mutation is strongly linked to that of the population hosting it, which may be largely independent of the mutation itself, its specific amino acid composition and its fitness. Polymorphisms lie in the no-man's-land after the occurrence of a random mutation and before its definitive fixation or rejection. So, it is not clear a priori if their nature and their statistics are similar to those of random mutations, to those of fixed substitutions or to neither of them. Polymorphisms have been used for inference about population phenomena, such as migration and selection (Kingman, 1982; Rosenberg and Nordborg, 2002) and the interplay between polymorphisms within and divergence between species has also been recently analyzed (Wilson *et al.*, 2011; De Maio *et al.*, 2015) to asses the distribution of selection coefficients and the signature of adaptation in different zones of the genome. However, we are not aware of any use of polymorphisms in the domain of scoring matrices.

We selected from the SNP database the polymorphisms characterized by no known clinical significance and present in at least 1% of the population, having high probability to be nearly neutral. The results presented in this chapter provide a hint that the component of natural selection determined by the fitness of a mutation acts principally on short time scales and has already completed its purifying action on the polymorphisms of our selection. According to this view, these frequent polymorphisms are separated from fixation mainly by factors dictated by chance, which act on longer time scales but are probably largely independent from the specific mutation and its fitness.

We are here going to prove that not only that genetic variations present frequently in a single species and substitutions accumulated between species at short evolutionary times have similar statistical properties, but also that they can both be embedded in

[1]The fitness of a phenotype describes its average reproductive success.

a common evolutionary model. This does not mean that we are going to model at the same time the temporal evolution of both polymorphisms and substitutions, but rather that we are going to exploit frequent polymorphisms to build an evolutionary model for substitutions. This model, obtained by combining the data from SNPs with the procedure for treating rate variability described in chapter 3, predicts the observed transition probabilities between amino acids in the range of 90-100% of sequence identity better than any other scoring matrix we are aware of, including LG (Le and Gascuel, 2008), WAG (Whelan and Goldman, 2001) and JTT (Jones *et al.*, 1992) for amino acid models or ECM (Kosiol *et al.*, 2007) for codon models.

This result is particularly significant because the SNP-model presented here, in contrast to all the other models against which it is benchmarked, is not learned from alignments and so, in principle, should be disadvantaged with respect to the others in the prediction of substitution frequencies from alignments. Indeed, tests at medium and low sequence identities show that our matrices derived from the SNPs perform worse than the standard ones. This indicates that the statistics on the SNPs alone is not precise enough to properly extrapolate the behaviour of sequences also at long evolutionary distances and that, in the medium-low sequence identity range, using the information embedded in the alignments becomes important for building a reliable model.

Even if our model of protein sequence evolution based only on the SNPs becomes less accurate as the sequences diverge, these results provide a hint that two traditionally separated domains, SNPs and sequence alignments, may provide compatible information and encourage to exploit SNP-based matrices for bioinformatic analysis of highly similar protein sequences. In particular, the scoring matrices presented here might provide a more accurate fine-grain resolution of phylogenies in human evolutionary genetics and population analysis.

4.2 Methods

4.2.1 Download and selection of SNP data

From the dbSNP database (Sherry *et al.*, 2001) we selected the SNPs in the coding part of the human genome whose *Global Minor Allele Frequency* (GMAF)[2] is at least 1%. We chose to set this lower threshold in order to ensure that a certain amount of

[2]The Global Minor Allele Frequency, or GMAF, is the fraction of individuals in a species presenting the less frequent between the two alleles (variants) of a polymorphism.

time has passed since the appearance of the mutation, so that, if the stability of the protein was seriously affected, the purifying natural selection had the time to reject it. The robustnress of our results with respect to the selected threshold is discussed in section 4.3. To remove also those SNPs that are likely to entail a positive or negative structural modification, we also excluded the SNPs with a known clinical significance ("benign", "likely benign", "pathogenic" and "likely pathogenic")[3]. With the described procedure we gathered 109082 polymorphisms, 53288 of which preserving the original amino acid (synonymous) and 55794 changing it (non-synonymous, or missense).

4.2.2 From the SNP database to codon substitutions

The information on the codons involved in a single nucleotide polymorphism has been derived as follows:

- We downloaded the desired set of SNPs both in *flatfile* format and in *brief* format, because none of the two contains the full necessary information.
- From the flatfile format we extracted the following data:
  - the label of that SNP;
  - its Global Minor Allele Frequency (GMAF);
  - the orientation (+ or -) in which the sequence must be read to be correctly translated to amino acids.
  - The two (or more) alleles, the corresponding amino acids and the frame of the codon in which the different nucleotides were found[4].
- If more than one orientation is present or GMAF<0.01 that entry was discarded[5].
- For each of the remaining entries we selected from the brief format the entry with the correct label and read the parts of the nucleotide sequence before and

[3]the SNPs used in this chapter were downloaded on the 17/07/2015 at `http://www.ncbi.nlm.nih.gov/snp/advanced`
Query: *(((homo sapiens[Organism] AND snp[SNP Class]) AND missense) NOT ((((benign[Clinical Significance]) OR likely benign[Clinical Significance]) OR likely pathogenic[Clinical Significance]) OR pathogenic[Clinical Significance]))) NOT no info[Validation Status]* for the non-synonymous polymorphisms and same query with "synonymous codon" in place of "missense" for the synonymous polymorphisms.

[4]If the frame is equal to 1, then the different alleles of the polymorphism are on the first letter of the codon. Same for frames 2 and 3

[5]We also prepared two subsets respectively containing those entries with $0.01 < GMAF < 0.2$ and $GMAF > 0.2$ for the consistency tests described in section 4.3.1

after the polymorphism.

- We combined the information on orientation and frame with the sequence surrounding the polymorphism and reconstructed the two (or more) involved codons.

- As usually done for alignments, the substitution process was treated as stationary and the substitutions from codon c_1 to codon c_2 were assumed to be as frequent as those from c_2 to c_1. So, the 64 x 64 matrix of occurrences $\boldsymbol{n}^{SNP-c}$ was built by setting both $n^{SNP}_{c1,c2}$ and $n^{SNP}_{c2,c1}$ to one half of the observed interchanges between the unordered pair of codons $\{c_1, c_2\}$. The entries of matrix $\boldsymbol{n}^{SNP-c}$ for substitutions involving codon pairs that differ by more than one nucleotide were not observed by construction. The least populated entry is $n^{SNP-c}_{ACG,CCG} = n^{SNP-c}_{CCG,ACG} = 22.5$.

- From the matrix of occurrences we computed the frequency of codon interchanges:

$$f^{SNP-c}_{c1c2} = \frac{n^{SNP-c}_{c1c2}}{\sum_{c3}\sum_{c4\neq c3} n^{SNP-c}_{c3c4}} \tag{4.1}$$

  This codon matrix can be summed to a 20 x 20 matrix of interchanges between amino acids by: $f^{SNP}_{AB} = \sum_{\{c1\in A\}}\sum_{\{c2\in B\}} f^{SNP-c}_{c1c2}$ where $\{c1 \in A\}$ stands for the set of codons coding for amino acid A.

4.2.3 Computing the transition probabilities for codons and amino acids

We here use the frequencies of interchanges between codons derived from SNPs, $\boldsymbol{f}^{SNP-c}$, as a direct measure of instantaneous substitution rates and we deduced the non-diagonal entries of the substitution rate matrix $\boldsymbol{Q}$ (see chapter 2):

$$Q_{c1,c2\neq c1} = \frac{f^{SNP-c}_{c1c2}}{f_{c1}} \tag{4.2}$$

where f_{c1} is the equilibrium frequency of codon $c1$ in the dataset, which we assume to be equal to those tabulated for the human codon usage bias (Nakamura *et al.*, 2000). The diagonal entries of $\boldsymbol{Q}$ are, instead, defined as $Q_{ii} = -\sum_{j\neq i} Q_{ij}$. Due to the normalization of $\boldsymbol{Q}$ such that $\sum_{c1}\sum_{c2\neq c1}(f_{c1}Q_{c1,c2}) = 1$, the time t is measured in units of expected substitutions per site.

The matrix of the average transition probabilities in a time interval t for a sequence characterized by a Γ-distributed substitution rate ($SNP + \Gamma$ model) was here estimated as proposed in the previous chapter:

$$P_{c1,c2}^{SNP+\Gamma}(t) = \left[\int_0^\infty e^{r \cdot t \cdot \boldsymbol{Q}} \Gamma_\alpha(r)\, dr\right]_{c1c2} \tag{4.3}$$

where r is the overall rate associated to a certain site and $\Gamma_\alpha(r)$ is its among-sites Γ-shaped probability distribution. This distribution depends on the shape parameter α, which is known to vary from protein family to protein family. Since we aim at describing an average situation, unless otherwise specified, we here use the effective value $\alpha = 0.25$, which is a plausible value for high sequence identity evolution (see Table 1 of ref. (Zhang and Gu, 1998)).

We also computed the substitution probabilities in the more traditional framework of identical rates on all sites (SNP-*no*-Γ model). In this case, the transition probabilities at time t are given by:

$$P_{c1,c2}^{SNP-no-\Gamma}(t) = \left[e^{t \cdot \boldsymbol{Q}}\right]_{c1c2} \tag{4.4}$$

From eq. 4.3 or eq. 4.4 the transition probabilities between amino acids, which are those in which we are interested for our comparisons, are computed by:

$$P_{A,B}(t) = \frac{\sum_{\{c1 \in A\}} \sum_{\{c2 \in B\}} f_{c1} P_{c1,c2}(t)}{f_A} \tag{4.5}$$

where $\{c1 \in A\}$ stands for the set of codons coding for amino acid A.
The expected sequence identity between an initial sequence and its evolution after a time t can be deduced from equations 4.3 or 4.4 by:

$$seqID(t) = \sum_A f_A P_{A,A}(t) \tag{4.6}$$

This equation implicitly specifies the time at which an evolutionary model should be compared with real data characterized by a certain sequence identity.

4.2.4 Selection of alignments and computation of experimental substitution frequencies

To test how well the evolutionary models described above predict the transition probabilities in the alignments, we computed the frequencies of amino acid interchanges,

$f_{AB}^{align}(seqID)$, from alignments at various sequence identities, grouping them in windows of 3% of sequence identity from 50% to 99%. These frequencies were learned from UniRef (Suzek *et al.*, 2007), an arrangement of the UniProt database (The UniProt Consortium, 2015) that clusters sequences above a certain sequence identity threshold. The following procedure was followed:

- We downloaded from UniRef the clusters at 50% of sequence identity with at least one human sequence[6].
- From each cluster we detected the human sequence and aligned it with each of the others. Sequences were aligned locally by the algorithm *water* (Smith and Waterman, 1981) in the *emboss* software package (Rice *et al.*, 2000) and only ungapped parts of at least 80 residues were considered.
- We then collected at most one ungapped alignment per cluster per window of sequence identity, to avoid weighting bigger clusters more.
- For each sequence identity window, the average sequence identity ($seqID$) was computed and the mismatches in the alignments for every amino acid pair (A, B) were counted. The substitution process was treated as an equilibrium one: we set both $n_{AB}^{align}(seqID)$ and $n_{BA}^{align}(seqID)$ equal to one half of the number of mismatches found between the unordered pair of amino acids A and B in that window of sequence identity.
- The entries where $n_{AB}^{align}(seqID) < 5$ were neglected in all further calculations, being affected by too large statistical errors.

4.2.5 Download and implementation of benchmark models

We are going to compare our evolutionary model with JTT (Jones *et al.*, 1992), LG (Le and Gascuel, 2008), ECM unrestricted (Kosiol *et al.*, 2007), WAG (Whelan and Goldman, 2001) and BLOSUM90 (Henikoff and Henikoff, 1992), so we now describe how we implemented them.

- The JTT matrix was deduced from the counts (Table 1) in the original paper (Jones *et al.*, 1992) by the same procedure described in the paper and in section 2.2.1. The $\boldsymbol{Q}$ matrix was obtained by inverting equation 4.4.

[6]The alignments were downloaded on the 23/07/2015 from UniRef at `http://www.uniprot.org/help/uniref` with query: [query:count:[2 TO $*$] length:[50 TO *] taxonomy:Homo sapiens (Human) [9606] AND identity:0.5]

- The LG $\boldsymbol{Q}$ matrix was reconstructed from the website of one of its authors (`http://www.atgc-montpellier.fr/download/datasets/models/lg_LG.PAML.txt`).
- The ECM-unrestricted[7] $\boldsymbol{Q}$ matrix was reconstructed from the Supplementary material of the original paper (Kosiol *et al.*, 2007) and was chosen instead of its restricted version because it was proved by the authors to give better results.
- The WAG $\boldsymbol{Q}$ matrix was reconstructed from the website of one of its authors (`http://www.ebi.ac.uk/goldman/WAG/wag.dat`).
- The BLOSUM90 matrix was downloaded from the site of NCBI (`http://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_DOC/lxr/source/data/BLOSUM90`). To adapt those scores to our definition (see eq. 4.9), which is deprived of any prefactor and computed by a natural logarithm, we multiplied them by $ln(2)/2$.

In all cases except BLOSUM we derived the substitution probabilities at the desired times from the $\boldsymbol{Q}$ matrix either by equation 4.4 (no-Γ version) or by equation 4.3 (+Γ version).

4.2.6 Maximum Likelihood Tests

We also benchmarked the quality of our model by computing the likelihood of phylogenetic trees with respect to other substitution models. Maximum likelihood tests were performed using the PAML software package (Yang, 2007) on reference alignments and phylogenetic trees downloaded from the Phylome Database[8] (Huerta-Cepas *et al.*, 2014). PhylomeDB provides a collection of highly accurate gene phylogenies for a wide variety of gene families and species. PhylomeDB phylogenies are derived from maximum likelihood tree inference, alignment trimming and evolutionary model testing.

The reference Primates and Model Species Metaphylomes[9], seeded on Homo sapiens,

[7] Here *unrestricted* means that, in the process of maximum likelihood used to determine the entries of matrix $\boldsymbol{Q}$, all the entries corresponding to multiple nucleotide substitutions were optimized. On the contrary, in the *restricted* version, multiple nucleotide substitutions were constrained to 0.

[8] PhylomeDB v4, data accessed on February 1st, 2016

[9] A *phylome* is the set of all gene phylogenies reconstructed starting from one proteome, which is called the seed proteome. This seed proteome will be the only proteome fully represented in the phylome, while the other analyzed ones will only appear in those trees where they have homologous sequences to the seed. *Metaphylomes* are groups of phylomes that use the same set of proteomes but start using different seed proteomes.

were downloaded from this database (Phylome IDs: 098 and 0500 respectively). We only analysed the collections made of those multiple sequence alignments (and their corresponding phylogenetic trees) characterized by 500-900 pairs of protein sequences, with average sequence identity in the range 45% to 99%. This filter produces two reduced collections of respectively 111 and 251 alignments for the two phylomes.

The PAML software package was used to compute maximum likelihoods keeping tree topologies intact and optimizing branch lengths. The Γ-correction was included and the free parameter α was optimized during the maximum likelihood estimations for all the tested models.

We computed maximum likelihood values for each of these reference phylogenies using the SNP, the JTT (Jones *et al.*, 1992), the WAG (Whelan and Goldman, 2001) and the LG (Le and Gascuel, 2008) evolutionary models.

The $\boldsymbol{Q}^{SNP}$ matrix used here is the reduction on amino acids of the one in eq. 4.2:

$$Q_{A,B} = \frac{\sum_{\{c1 \in A\}} \sum_{\{c2 \in B\}} f_{c1} Q_{c1,c2}}{f_A} \tag{4.7}$$

4.3 Results

4.3.1 Consistency tests

In this work we use the SNPs as a source of point accepted mutations to build an evolutionary model. The choice of SNPs as a starting point comes from the desire of a source of genuine point mutations: while observed mismatches in alignments may be the result of more than one substitution in a row, SNPs guarantee by construction to actually derive the scores from one-step interchanges. This is an important advantage, because it does not require to decouple overlapping substitutions. However, in order to use the SNPs to estimate a substitution rate matrix, it is necessary to verify that the purifying natural selection, which depends on the fitness of a mutation, has already accomplished its action on the SNPs selected by our procedure, making their statistical properties resemble those of substitutions cumulated along the lineages in protein variants of similar species.

With this in mind we performed the following consistency test. We divided our collection of SNPs into two subsets corresponding to different GMAF values (see section 4.2.1). In the subset at low GMAF (label L) we selected the entries with $0.01 < GMAF \leq 0.2$ and in the one at high GMAF (label H) we included those with $GMAF > 0.2$. We computed the frequency of substitutions $f_{c1,c2}$ between codons

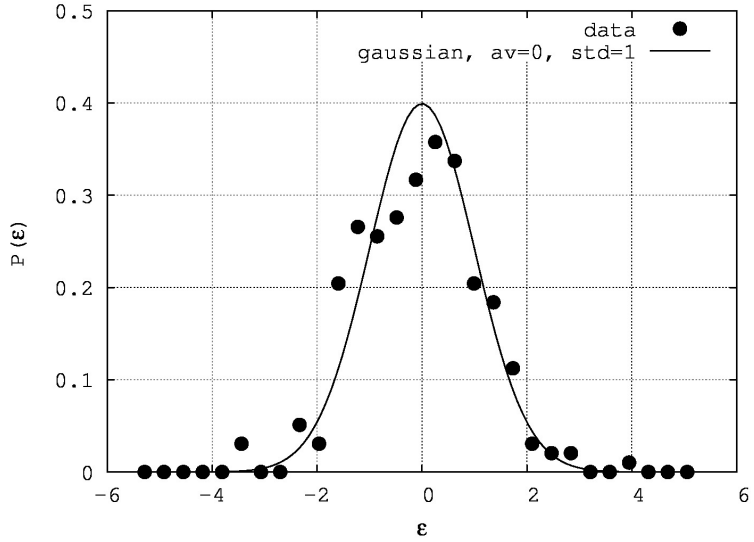

Figure 4.1: *Analysis of the statistical consistency between SNP frequencies at small* $(0.01 < GMAF \leq 0.2)$ *and high* $(GMAF > 0.2)$ *GMAF.*
The histogram of the relative difference between the two datasets (ϵ)*, defined as in eq. 4.8, (points) is compared with a zero-mean unit-variance gaussian distribution (line).*

$c1$ and $c2$ in both datasets and the corresponding error according to the Poisson statistics. For each pair of different codons we computed:

$$\epsilon_{c1,c2} = \frac{f^L_{c1,c2} - f^H_{c1,c2}}{\sqrt{\sigma^2(f^L_{c1,c2})^2 + \sigma^2(f^H_{c1,c2})}} \tag{4.8}$$

where $\sqrt{\sigma^2(f^H_{c1,c2}) + \sigma^2(f^L_{c1,c2})}$ corresponds to the statistical error on the difference at the numerator. Figure 4.1 shows the normalized histogram of $\epsilon_{c1,c2}$ and a gaussian distribution with expected value 0 and standard deviation 1. The similarity between the two curves shows that the difference (ϵ) between the H and L datasets can be ascribed to statistical errors. We may then conclude that, for the subset of the SNPs chosen by our procedure, there is no evident bias in the frequency of substitutions due to different ranges of GMAF. This, according to our interpretation, means that the fitness-related natural selection has already completed its work even for the SNPs at low GMAF ($0.01 < GMAF < 0.2$). Another possible interpretation would be that the purifying selection needs much longer times even with respect to polymorphisms at large GMAF, but, in the light of the results that we are going to present, this last interpretation seems to make little sense. Indeed, we will show that the frequencies of interchange observed in the SNPs are compatible with the frequency of substitutions

observed in alignments.

Let us denote f_{AB}^{SNP} the frequency of substitutions between amino acids A and B in the full[10] SNP dataset and $f_{AB}^{align}\,(seqID)$ the analogous frequency of substitutions from pairwise alignments at sequence identity $seqID \geq 98\%$ (see *Methods*), namely at very high sequence identity. In figure 4.2 we plot the entry-by-entry comparison in log-log scale of $\boldsymbol{f}^{SNP}$ with $\boldsymbol{f}^{align}\,(seqID \geq 98\%)$ both with error-bars estimated according to Poisson statistics. The datasets are clearly correlated and lay along the line $y = x$, as they should if they were harvested from the same distribution. The only important exception is the set of blue points without error bars: these points correspond to the amino acid pairs whose codons have at least two different nucleotides. For example histidine is coded by CAU and CAC and phenylalanine by UUU and UUC, so no single nucleotide substitution can transform any codon of histidine into a codon of phenylalanine. These points have $f_{AB}^{SNP} = 0$ by construction, but one can easily notice that they occur with non-zero probability in the alignments even at such a high sequence identity. This is the major difference between the two sets of frequencies and the impossibility to deduce instantaneous multiple substitutions represents a possible drawback of using SNPs alone to predict protein sequence evolution. Contrasting estimates have been made to quantify instantaneous multiple substitutions, but in all cases they seem to cover between the 0.3% (Smith *et al.*, 2003) and the 3% (Schrider *et al.*, 2011) of the total number of instantaneous substitutions. This can be taken as a measure of the error introduced by our approximation of "no-multiple-substitutions". However, we will see in the following that allowing for instantaneous multiple substitutions is not strictly necessary in order to achieve a high consistency with substitutions observed in alignments if one models sequence evolution accounting for among-site rate variability.

Going back to the points corresponding to single nucleotide substitutions, it is evident that, even if the correlation is rather good, the statistical error is not large enough to explain entirely the deviations from the diagonal (statistical errors cover about the 30% of the real difference). Extra errors are likely to be mostly due to other hidden double substitutions in the alignments for amino acid pairs where both single and double nucleotide substitutions can lead to the same amino acid exchange.

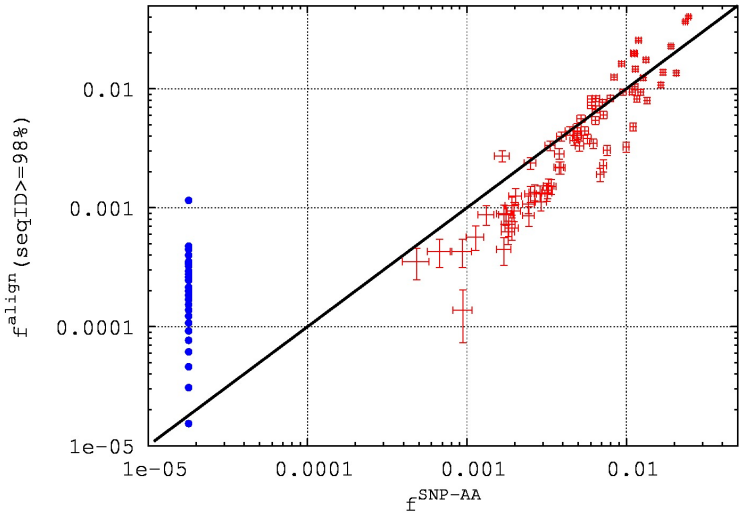

Figure 4.2: *Red symbols with error bars: entry-by-entry comparison in log-log scale of the substitution frequencies between amino acids in SNPs from Homo sapiens (f_{AB}^{SNP}) and the equivalent substitution frequencies in ungapped pairwise alignments at $seqID \geq 98\%$ with one of the sequences labeled as human ($f_{AB}^{align}\,(seqID \geq 98\%)$). Each point corresponds to a pair (A, B) of amino acids and its x-value is given by f_{AB}^{SNP}, while its y-value is f_{AB}^{align}. The error bars show the statistical errors expected on each point estimated according to Poisson statistics. Blue points: same as for red symbols, but for the amino acid pairs that can not mutate into one another by a single nucleotide substitution. For these points $f_{AB}^{SNP} = 0$ is modified to the tiniest appreciable value ($f_{AB}^{SNP} = \frac{1}{N\ total\ mut}$) to make them visible in the log-log scale. Dashed line: line $y = x$.*

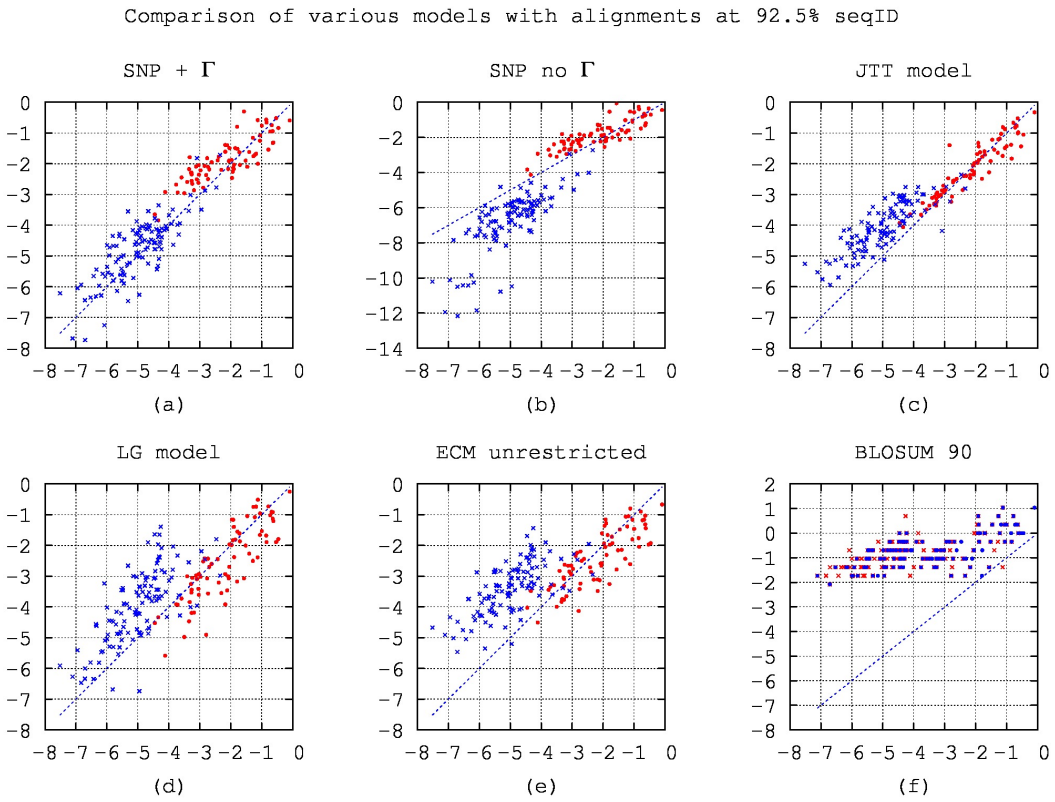

Figure 4.3: *Entry-by-entry comparison of the scores* s_{AB}^{align} $(seqID = 92.5\%)$ *from ungapped alignments in UniRef (procedure described in detail in section 4.2) on the x-axis and* s_{AB}^{model} $(seqID = 92.5\%)$ *computed for different models of evolution on the y-axis. Each point corresponds to a pair of amino acids* (A, B)*. A different point style is adopted to distinguish amino acid pairs whose interchange can be determined by a single nucleotide change (red circles) from the pairs where at least two nucleotides must mutate (blues crosses). The scores between amino acid pairs where the first and the second amino acids coincide are not shown for the sake of visibility. Panels: (a)* $SNP + \Gamma$*; (b)* SNP*-no-*Γ *(notice that here the y axis is not the same as in the other panels); (c): JTT model (Jones* et al.*, 1992); (d): LG model (Le and Gascuel, 2008); (e): ECM-unrestricted model (Kosiol* et al.*, 2007); (f): BLOSUM90 (Henikoff and Henikoff, 1992). Dashed line:* $y = x$

4.3.2 Prediction of transition probabilities in alignments

We now consider the transition probabilities between amino acids predicted by our $SNP+\Gamma$ model (see section 4.2.3), obtained by an instantaneous rate matrix derived from SNPs and evolved by accounting for rate variability as described in chapter 3. To evaluate the quality of a model of protein sequence evolution we analyze its scores (see section 2.2.1):

$$s_{AB}^{model}\left(seqID\right) = \log\left[\frac{P_{A,B}^{model}\left[t\left(seqID\right)\right]}{f_B^{model}}\right] \tag{4.9}$$

$P_{A,B}(t)$ is the transition probability from amino acid A to amino acid B in an evolutionary time t corresponding to the desired sequence identity $seqID$, whereas f_B is the equilibrium frequency of amino acid B. To check if the dynamics predicted by our $SNP+\Gamma$ model (eq. 4.3) is a good descriptor of the real one we compare $s_{AB}^{SNP+\Gamma}$ with the equivalent scores extracted from pairwise sequence alignments around the same sequence identity (see section 4.2):

$$s_{AB}^{align}\left(seqID\right) = \log\left(\frac{f_{AB}^{align}\left(seqID\right)}{f_A^{align} f_B^{align}}\right) \tag{4.10}$$

where f_A is the equilibrium frequency of amino acid A in the considered set of alignments.

In figure 4.3 panel (a) we show the entry-by-entry comparison of s_{AB}^{align} in alignments around the 92.5% of sequence identity and $s_{AB}^{SNP+\Gamma}$ derived from our model at the same sequence identity. In such plots a good model is characterized by points lying along the line $y = x$ with the smallest deviations. Amino acid pairs whose interchange may be determined by a single nucleotide change are identified by red circles, while those pairs for which this is not possible are labeled by blues crosses. For example histidine is coded by CAT and CAC, glutamine by CAA and CAG and phenylalanine by TTT and TTC. So a single nucleotide substitution can transform one of histidine's codons to glutamine, while at least two substitutions are necessary to transform any of them to phenylalanine. The amino acid pair histidine-glutamine is thus labeled by a red circle, while the pair histidine-phenylalanine is labeled by a blue cross. The points lie near the line $y = x$ in both subsets, proving that the $SNP+\Gamma$ model correctly predicts the substitution probabilities in the alignments at 92.5% of sequence identity for what concerns both single and multiple substitutions.

[10] We now consider all entries with $GMAF > 0.01$.

In figure 4.3 panel (b) we compare s_{AB}^{align} (x-axis) and $s_{AB}^{SNP-no-\Gamma}$ (y axis), with the same color code as in panel (a), at the same sequence identity. Here the scores are computed from a model still based on SNPs, but neglecting the rate variability as described in Methods. Contrary to panel (a), here the comparison is not very good: even if the points lie in the proximity of the line $y = x$ and the two datasets are certainly correlated, a significant shift is clearly visible for those amino acid interchanges where double mutations are necessary. These scores are systematically underestimated by the *SNP-no-*Γ model. It is evident that the inclusion of the variability of substitution rates is necessary for making the SNP model accurate.

In the other panels of figure 4.3 we compare $s^{align}(seqID = 92.5\%)$ with the scores from some popular models for protein sequence evolution: JTT (Jones *et al.*, 1992) in panel (c), LG (Le and Gascuel, 2008) in panel (d), the codon matrix ECM-unrestricted (Kosiol *et al.*, 2007) in panel (e) and BLOSUM90 (Henikoff and Henikoff, 1992) in panel (f). For each model time has been chosen in order to attain a sequence identity of 92.5%. For JTT, LG, ECM, and also for the WAG matrix (Whelan and Goldman, 2001) that we will use below, we evolved the respective $\boldsymbol{Q}$ matrices both with (eq. 4.4) and without (eq. 4.3) considering the variation of rates across sites. We have noticed that in all those cases computing transition probabilities by equation 4.3, namely averaging over the distribution of the rates, worsens the performances rather than improving them (see next paragraph for more specifications), so we used for all of them the $no - \Gamma$ version in the comparison in figure 4.3 and in all the following ones. Although the dispersion of the points could be influenced by the choice of alignments with a human sequence, it is clear that none of the analyzed models correctly estimates the amount and distribution of multiple substitutions (the subset labeled by the blue crosses): while the *SNP-no-*Γ underestimates them, the standard models (JTT, LG, ECM and BLOSUM) overestimate them. BLOSUM90, in particular, dramatically fails to reproduce experimental data. This is somehow expected by the construction method of BLOSUM matrices. In fact, while all the other models considered here are based on an evolutionary model of substitutions, the scores of BLOSUM90 were learned from conserved blocks of multiple sequence alignments whose maximum sequence identity is 90% and without any explicit lower bound. The main consequence is that, while BLOSUM matrices are perfectly suited to score alignments at medium and low sequence identity, they are not optimal in scoring those at high sequence identity (see chapter 2).

To assess the quality of the score comparisons in a more quantitative way we compute the average distance from the diagonal of the points in an entry-by-entry

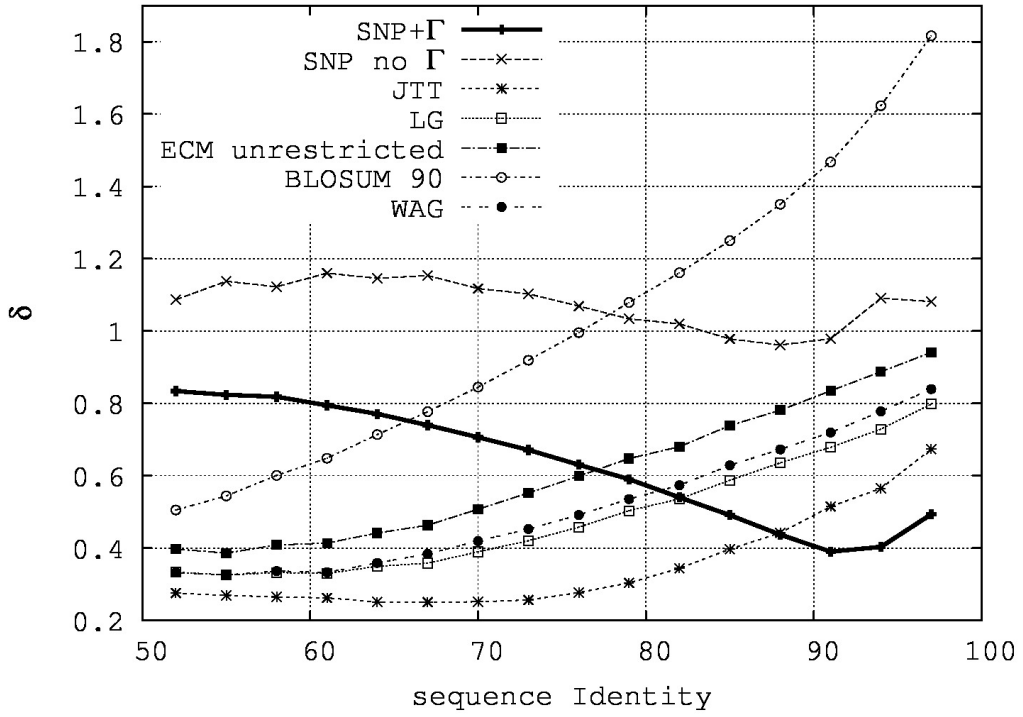

Figure 4.4: *δ^{model} (defined in eq. 4.11) as a function of the sequence identity for different models (key on the figure)*

plot (as those in the panels of figure 4.3) in units of the variance of the data[11]:

$$\delta^{model} = \sqrt{\frac{\left\langle \left(s_{AB}^{model} - s_{AB}^{align} \right)^2 \right\rangle_{A \neq B}}{\left\langle \left(s_{AB}^{align} - \left\langle s_{AB}^{align} \right\rangle_{A \neq B} \right)^2 \right\rangle_{A \neq B}}} \tag{4.11}$$

This is a sort of relative error, so lower values of δ mean better predictions.

Figure 4.4 shows the behavior of δ in the window of sequence identity 50-100% for the following models: $SNP + \Gamma$, SNP-*no*-Γ, JTT, LG, ECM-unrestricted, WAG and BLOSUM. At very high sequence identity (90-100% seqID) the model $SNP + \Gamma$ performs better than any other method against which it is benchmarked and it keeps very high performances also for 80-90% of sequence identity. There only JTT, which is learned from alignments in that interval of sequence identity, performs better. The poor performances of the SNP-*no*-Γ model indicates again that taking into account the among-site rate variability is essential when dealing with SNPs. The

[11] With the aim of obtaining a quantification of the quality independent of the sequence identity, we chose to divide the mean squared error by the variance of the data.

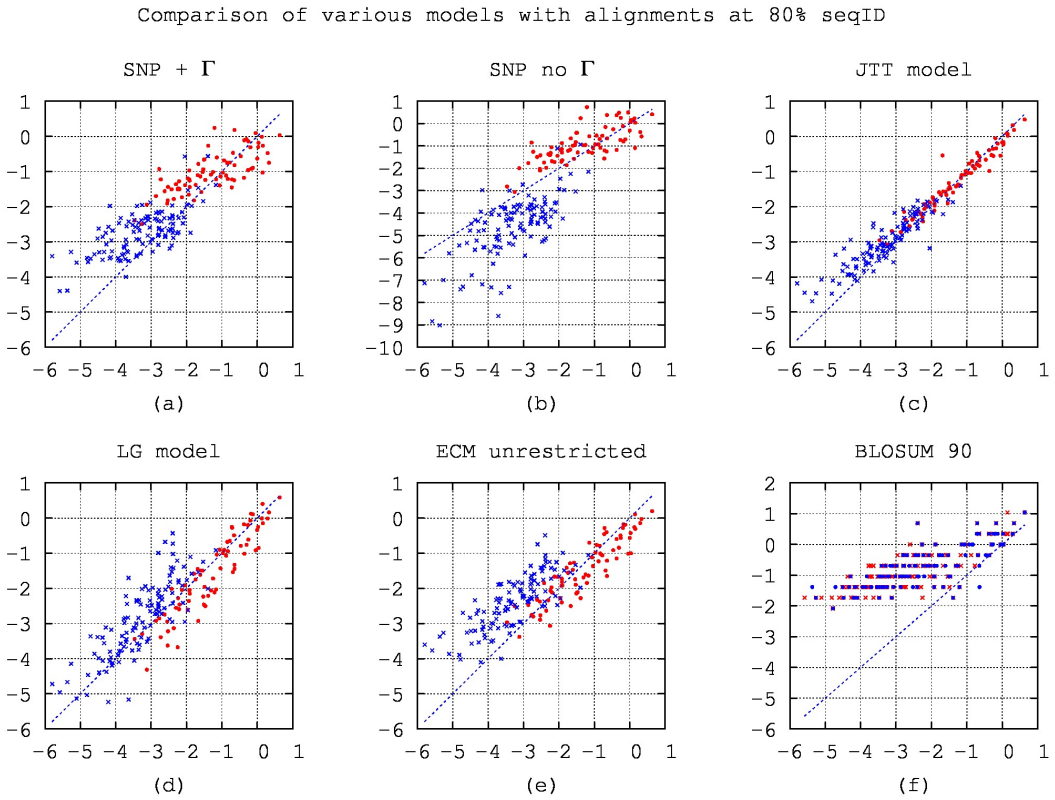

Figure 4.5: *Entry-by-entry comparison of the scores* $s_{AB}^{align}\,(seqID = 80\%)$ *on the x-axis and* $s_{AB}^{model}\,(seqID = 80\%)$ *on the y-axis. Point style, lines and panels are analogous to those in figure 4.3.*

analogous of figure 4.3 calculated at the 80% of sequence identity is shown in figure 4.5, confirming the worsening of the scores of the $SNP + \Gamma$ model with respect to JTT, LG and ECM.

Figure 4.6 shows the relative error (δ) at 92.5% of seqID for JTT, LG, ECM-unrestricted and WAG models evolved by equation 4.3, i.e. averaging transition probabilities over Γ-distributed substitution rates. For all models we tried several shape parameters α obtaining better performances for larger values of α. Anyway, the best performances in each of these cases were obtained when evolving these matrices without averaging over different rates, thus by equation 4.4 (shown in figure as $\alpha = \infty$). This may seem counterintuitive at first sight, because it is well known that the Γ-correction tends to improve phylogenetic estimations with all these models, but one must consider that here we have no a priori information on the rates of specific sites and in equation 4.3 we employ the Γ distribution just to perform ensemble average transition probabilities. In our opinion the worsening in the performances obtained in this case may be due to the fact that these matrices are

learned at medium and low sequence identity, where this phenomenon has already been averaged out, and thus they estimate an effective $\boldsymbol{Q}$ matrix.

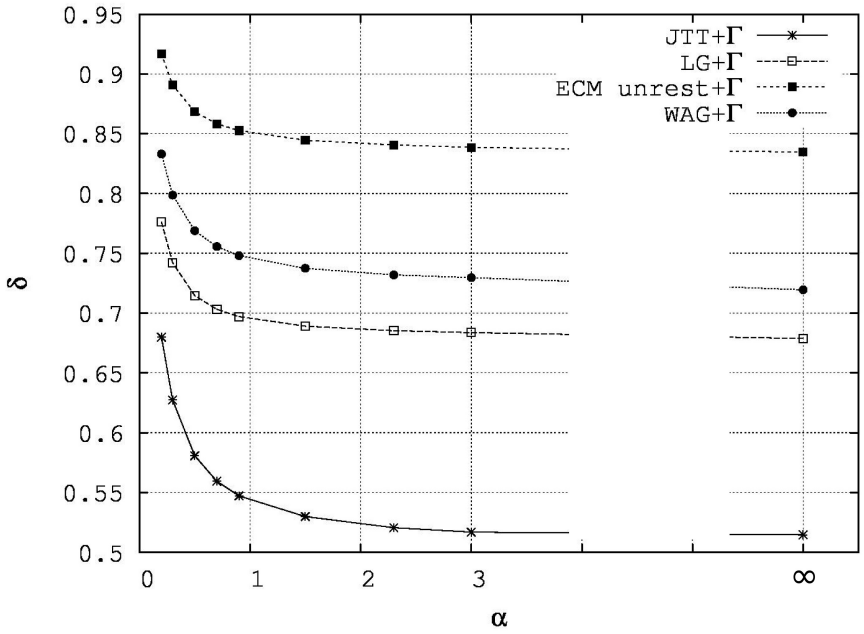

Figure 4.6: *$\delta^{model}(seqID = 92.5\%)$ for JTT, LG, ECM unrestricted and WAG models evolved by eq. 4.3, as functions of the shape parameter (α) of the Γ distribution of substitution rates over which transition probabilities are averaged. When $\alpha = \infty$ the Γ distribution becomes a δ function, thus representing the case where all rates are equal and transition probabilities are simply obtained by equation 4.4.*

As explained in Methods, the $SNP + \Gamma$ model uses an effective value of α, while the true one changes from protein family to protein family. In the previous tests and figures α was set to 0.25, which seems the best choice for high sequence identity, but we now test various values in the interval $[0 : 2]$. Figure 4.7 shows that the $SNP + \Gamma$ model, at the sequence identity of 92.5% (the one used in fig. 4.3), performs better than the other models for all $\alpha \in [0.1 : 0.7]$ and in the range $\alpha \in [0.7 : 1.5]$ is second only to the JTT. The considered intervals include the values of α of most protein families (see Table 1 of (Zhang and Gu, 1998)), so we can conclude that the performance of the $SNP + \Gamma$ model is relatively robust with respect to the choice of the effective value of α. The important point is allowing for a certain variability among rates.

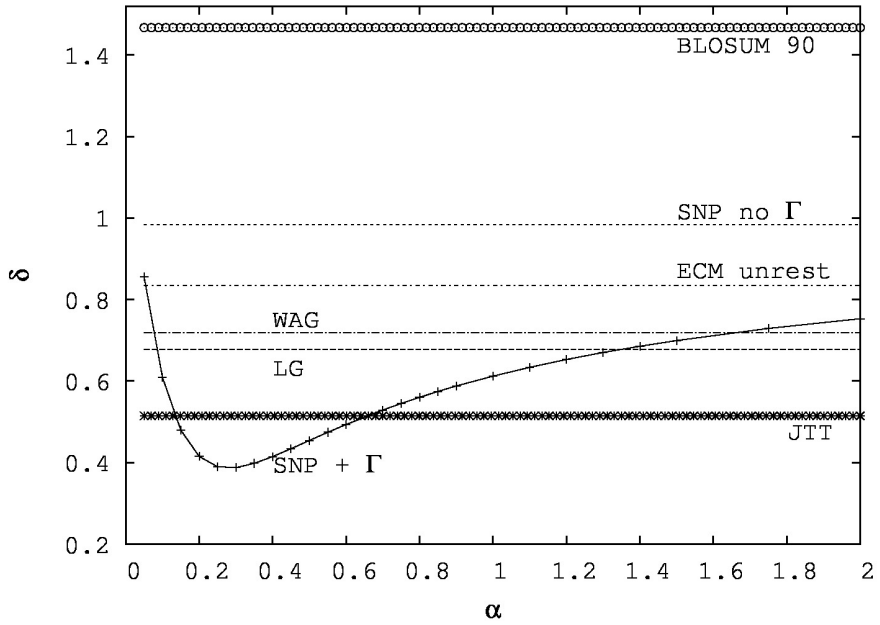

Figure 4.7: *$\delta^{SNP+\Gamma}$ from eq 4.11 at 92.5% of sequence identity as a function of the shape parameter α of the substitution rate distribution $\Gamma_\alpha(r)$. Horizontal lines: reference values of the other analyzed models according to the figure key.*

Direct minimization of δ

We discussed in section 2.3.2 how substitution rates are not only variable along the sites but also in time and how, when this feature is not accounted for, the estimated substitution rates are time averages of the true values. We observed that this phenomenon leads to larger estimates of the shape parameter α for large evolutionary times. Given the impossibility to include time-variability in our model in a viable way we decided at least to allow α to vary at the different sequence identity, choosing each time the value of α giving the best agreement with the data (dark blue line in figure 4.8). As we expected, for lower sequence identities the optimal α is no more $\alpha = 0.25$ but tends to grow ($\alpha = 0.6$ at 55% of sequence identity) and, with this inclusion, the comparison with the scores from alignments show considerable improvements.

Moreover, we analyzed the influence of the statistical errors coming from the SNPs on the performances and the limitation implied by the assumption of $Q_{ij} = 0$ for the codon pairs (i, j) differing by more than 1 nucleotide by optimizing the entries of $\boldsymbol{Q}$.

We thus minimized the quantity:

$$\delta^{SNP\,mod+\Gamma} + \sum_i \sum_{j \neq i} \left(\frac{Q_{ij}^{SNP} - Q_{ij}^{SNP\,mod}}{\sigma_{ij}} \right)^2$$

at 72% of the sequence identity with $\alpha = 0.5$. Here $\boldsymbol{Q}^{SNP\,mod}$ is the optimized $\boldsymbol{Q}$ matrix, $\delta^{SNP\,mod+\Gamma}$ has the same definition of δ in the main text (eq. 3) but using $\boldsymbol{Q}^{SNP\,mod}$ rather than $\boldsymbol{Q}^{SNP}$ and σ_{ij} is the statistical error on Q_{ij}^{SNP} calculated with propagation of errors by considering the observed counts n_{ij}^{SNP} as Poisson-distributed. We artificially set $\sigma_{ij} = 0.005$ for the pairs of codons i, j differing for two nucleotides and $\sigma_{ij} = 0.0001$ for those pairs differing for all three nucleotides because the corresponding counts where 0 in the SNP database. The minimization was constrained to have $Q_{i,j\neq i} \geq 0$ and the entries on the diagonal were recalculated at every step by $Q_{ii} = -\sum_{j\neq i} Q_{ij}$.

The light blue curve in figure 4.8 shows $\delta^{SNP\,opt+\Gamma}$ obtained with the result of the described optimization $\boldsymbol{Q}^{SNP\,opt}$. Here α was optimized separately at each sequence identity.

Even if the optimization of $\boldsymbol{Q}^{SNP}$ was performed at a single sequence identity (72%, in fig. 4.8 the square is the starting point and the dot the end point) the improvements brought by using $\boldsymbol{Q}^{SNP\,opt}$ are general and stable in the whole sequence identity range.

4.3.3 Likelihood ratio tests for phylogenetic trees

In order to further benchmark the robustness of our $SNP + \Gamma$ model of protein sequence evolution, we performed likelihood ratio tests of our model against other popular models using reference datasets of phylogenetic trees and multiple sequence alignments retrieved from the Phylome Database. Since this database does not include codon information, we here used a reduced rate matrix on amino acids instead of the one on codons, as described in 4.2. Two phylomes were considered: one containing homologous sequences of closely related species (only primates) and another covering a wide diversity of species (mammals, birds, insects, plants, fungi, bacteria...). On a collection of 111 multiple sequence alignments of primates and their corresponding phylogenetic trees, our model provides the best likelihood values over the four tested models (SNP, JTT, WAG and LG, all employed together with the Γ-correction) for 50% of the phylogenetic trees being assessed. In this optimization tree topologies were kept fixed, while branch lengths were optimized. Surprisingly,

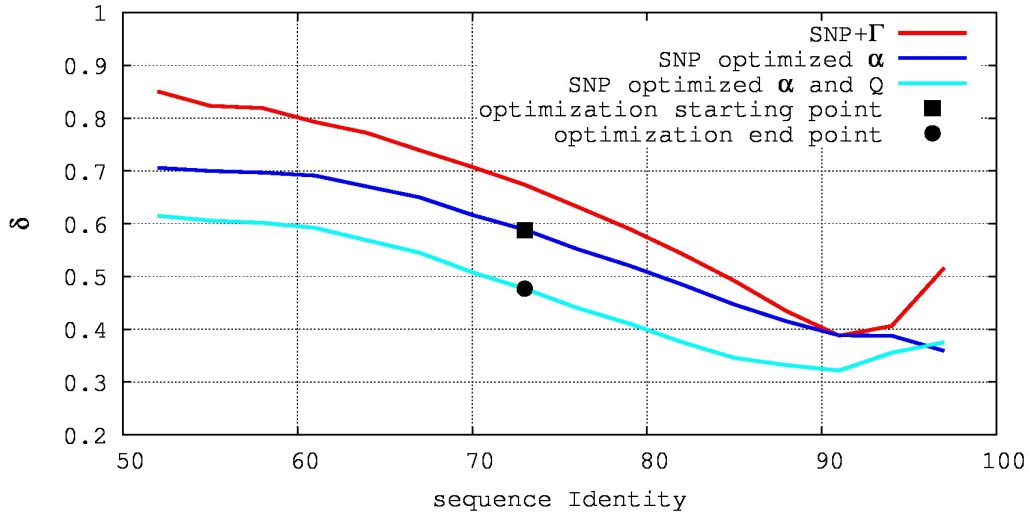

Figure 4.8: *Comparison of the relative error given by equation 4.11 ($\delta^{SNP+\Gamma}$) with the shape parameter of the Γ distribution fixed at $\alpha = 0.25$ (red line), $\delta^{SNP+\Gamma}$ with α optimized separately at each sequence identity (blue line) and $\delta^{SNP\,opt+\Gamma}$ (light blue line). The square is the value of $\delta^{SNP+\Gamma}$ at the beginning of the optimization of $\boldsymbol{Q}$ ($\alpha = 0.5$) and the dot is its final value, corresponding to $\boldsymbol{Q}^{SNP\,opt}$ and $\alpha = 0.5$.*

these results were obtained at practically any level of average sequence identity (40%-99%) in this first dataset (see Table 4.1). This indicates that our model can also be considered for phylogenetic inference from multiple alignments of sequences of evolutionary close species, such as the primates phylome tested here. We are confident that the quality of these results could be further improved if using our substitution rate matrix on codons instead of its reduction on amino acids, since they lead to different dynamics (Kosiol and Goldman, 2011).

When the same test is performed on a phylome that covers a wide variety of species, the likelihood improves only for alignments at very high average sequence identity. Indeed our model is derived only from data at extremely high sequence identity and from the same species (Homo sapiens SNPs). However, it is still evident that also in the multiple species phylome half of the phylogeny reconstructions in the range of high sequence identity can be improved with our SNP model.

4.4 Discussion

Our results imply that frequent SNPs and alignments at high sequence identity provide consistent information on amino acid substitution probabilities. In particular, we verified that frequent SNPs, when combined with an adequate treatment of the

Avrg Seq. Id.	Primates	Multiple species
40%-99%	57 (51.3%)	23 (9.2%)
90%-99%	12 (42.9%)	11 (47.8%)
80%-89%	14 (53.8%)	10 (30.3%)
70%-79%	15 (60.0%)	2 (10.5%)
50%-69%	14 (50.0%)	0 (0.0%)
40%-49%	2 (50.0%)	0 (0.0%)

Table 4.1: *Number and fraction of phylogenies with improved likelihood values for the $SNP + \Gamma$ model compared to other three models (JTT, WAG and LG) for two different phylomes: primates phylome and multiple species phylome (see section 4.2 for further details).*

among-site substitution rate variability, can be successfully used to learn scoring matrices and score alignments at very high sequence identity (80-100%) with performances comparable to the best scoring matrices used nowadays. The capability of the $SNP + \Gamma$ model of reproducing substitution probabilities is remarkable if we consider that it is the only one that does not learn the transition probabilities from alignments. Moreover, these results are obtained without considering the ratio between synonymous and non-synonymous substitutions as a free parameter: we measure it direclty from the SNP dataset.

The observed worsening of the performances at lower sequence identities (see fig. 4.4) is somehow expected: one significant source of error is the approximation that the rate of each site is assumed to be constant in time. This approximation is less adequate at long evolutionary times (Lopez *et al.*, 2002) because, for instance, of co-evolution. Indeed, we verified that optimizing the value of α in the Γ distribution separately at each sequence identity can encompass this variation and partially improve the performances giving a relative error reduced by $\sim$15%. A second relevant source of error depends on the impossibility to produce non-zero instantaneous multiple substitutions from a database of SNPs, which by definition contains only single nucleotide polymorphisms. One last identified source of error comes from the extrapolation to low sequence identity of information collected from the SNPs, making the statistical errors propagate. To estimate the importance of these two last sources of error, we optimized the entries of $\boldsymbol{Q}$, included those concerning multiple substitutions, constraining them around their initial value. We obtained a further reduction of the relative error of an additional 15% with respect to the optimization of α described above (fig. 4.8).

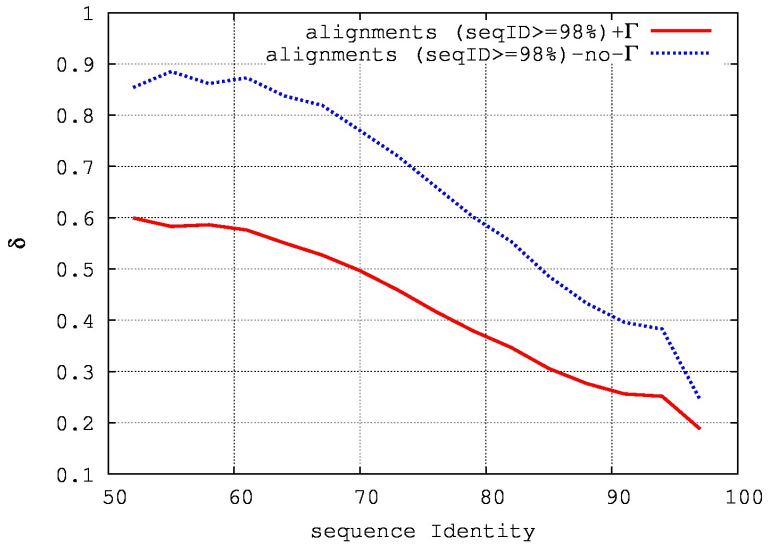

Figure 4.9: *Relative error δ from equation 4.11 in the range 50-100% of sequence identity for the $\boldsymbol{Q}$ matrix obtained from the alignments at more than 98% of sequence identity in the database UniRef. The two curves compare the performances when this $\boldsymbol{Q}$ is evolved by eq. 4.3 (red curve), thus averaging over Γ-distributed rates choosing the best α at each sequence identity ($\alpha \in [0.35 : 0.8]$), and when evolved by eq. 4.4, so with the assumption of constant rates.*

The poor performances obtained by the *SNP-no-*Γ model may give important hints for the developement of new evolutionary models. Indeed, in the previous chapter we proved that the rate variability induces non-Markovian effects on the full protein sequence evolution. Remarkably, this effect seems much stronger for the model learned from the SNPs than for the other tested models. Indeed, as shown in figure 4.6, computing transition probabilities as averages over Γ-distributed rates for JTT, LG, WAG and ECM worsens their performances rather then improving them. A possible interpretation is that the matrices learned from alignments at medium sequence identity estimate an effective Markovian dynamics[12], thus needing a Markovian propagator and performing better when comparing sequences separated by large enough evolutionary times. Instead, the $\boldsymbol{Q}$ matrix of the SNP model describes the original non-Markovian process and consequently needs a non-Markovian propagator to give meaningful results. This would also explain why evolving the SNP-model without a proper inclusion of rate variability leads to wrong results both at small and at large evolutionary time.

[12]effective with respect to the violation of the Markov assumption due both to the genetic code (section 2.1.1) and to the rate variability (chapter 3)

To verify this guess we also computed a substitution rate matrix from the alignments downloaded from UniRef having sequence identity larger than 98% (with the same procedure used for SNP). In figure 4.9 we analyze the performances obtained by evolving it both by eq. 4.4 and by eq. 4.3, thus averaging over Γ-distributed rates. As for the $\boldsymbol{Q}$ matrix learned from SNPs, we found a clear improvement when accounting for the rate variability. This phenomenon further strengthens our belief that matrices learned at very short evolutionary times describe the original non-Markovian process and thus need a non-Markovian propagator. According to our interpretation, two regimes exist: a pseudo-Markovian regime at medium and low sequence identity and a strictly non-Markovian regime at high sequence identity. The threshold seems to lie around 85% of sequence identity, above which the $SNP + \Gamma$ model, or other models learned from alignments at very high sequence identity, starts to perform better than standard models, probably thanks to their adequate inclusion of the non-Markovian factors (rate variability and genetic code).

Even if at high sequence identity homology can be detected also by generic models, the use of more specific models for highly similar sequences can correct small local misalignments and errors in the alignment scores and in the calculation of pairwise distances. In phylogenetic trees, a precise estimation of the evolutionary distances between very similar sequences is extremely important. In particular, this feature arises clearly when using distance-based approaches such as the neighbor-joining algorithm (Saitou and Nei, 1987), in which the first step of the tree construction consists in joining each sequence with its nearest neighbor. If more then two sequences are comparably similar, the exact ranking of their similarity becomes sensitive on the exact measure of similarity.

The SNP model presented here provides marginally better estimations of phylogenetic relationships in species closely related to Homo sapiens, even when codon information is not available. The different performances between the two analyzed phylomes can be ascribed to the fact that both the primate phylome and the SNPs are specific for Homo sapiens and probably share features such as the equilibrium probability of amino acids which improve the predictive power.

Finally, scoring the alignments with the approach introduced in this work may become relevant in the framework of massive human genome sequencing projects aimed at deciphering human genetic variations among populations (1000 Genomes Project Consortium *et al.*, 2015).

Chapter 5

A model for substitution rate variability based on finite memory coevolution

5.1 Introduction

In chapter 2 we mentioned that evolution conserves structure and function more than the sequence and we have sketched there the mechanism of *compensatory mutations* by which they induce evolutionary constraints between entangled residues. For example, after a mildly destabilizing mutation at a site, the residues in contact with it may more easily fix mutations in order to re-establish the original structural and functional balance. In the last twenty years, many investigations on coevolving residues have been performed showing that one can infer structural information from residue covariation in large multiple sequence alignments (de Juan *et al.*, 2013; Gobel *et al.*, 1994; Weigt *et al.*, 2009; Morcos *et al.*, 2011; Ekeberg *et al.*, 2013; Burger and Van Nimwegen, 2010; Jones *et al.*, 2012). This suggests an important contribution of structural coevolution in the process of sequence evolution.

We also mentioned that the rate at which each site fixes random mutations may vary both among sites (Yang, 1993, 1994, 1995; Halpern and Bruno, 1998) and in time (Fitch and Markowitz, 1970; Gaucher *et al.*, 2001; Lopez *et al.*, 2002) making the process of sequence evolution more complicated than a simple set of identical and independent Markov processes on the protein sites. In particular Fitch and Markowitz (1970) theorized that only a limited number of protein sites at a time can fix mutations, and once this happens others residues coupled with them will gain mutational freedom. This phenomenon would produce groups of *COncomitantly*

VARIable codONS (for simplicity *covarions*) which change over time in a correlated way. Unfortunately, despite this initial qualitative intuition, most of the quantitative implementations of the covarion model (Penny *et al.*, 2001; Galtier, 2001) had to sacrifice for the sake of computability the inclusion of any spatial pattern of covariation.

In the line of thought introduced by the covarion model, we here propose a simple model to describe the time evolution of the substitution rates of protein sites by explicitly including spatial and temporal interdependence. The core idea is that the probability of fixing a substitution at a site is enhanced if this site is in spatial proximity with other recently mutated sites, mimicking the mechanism of compensatory mutations. At variance with standard coevolution models, we assume that this mechanism acts only for a finite time: if compensatory mutations do not take place in a few generations the initial substitution is efficaciously forgotten. This, as we will show, allows reproducing by a simple model several non trivial features observed in protein sequence evolution.

With our model we accurately reproduce the experimental patterns of along-chain conditional probability of substitutions in the sequence identity range 50-100%. Moreover, the number of substitutions per site produced by our model is well described by a negative binomial distribution, as generally found in phylogenetic analysis. The shape parameter of this negative binomial distribution falls in a realistic range and increases for growing evolutionary time in qualitative accordance with experimental data (see chapter 2). This indicates that the model reproduces, at least qualitatively, the distribution of the rates observed in real families.

Our model predicts that substitutions take place in *avalanches* localized not only in three-dimensional space, as commonly predicted by coevolution, but also in time. We found qualitative evidence of these avalanches in the sequence evolution of Influenza Hemagglutinin.

The simplicity of this model lies in the presence of a single free parameter and on an implementation which neglects both specific sequence and structure focusing only on the set of times of the last mutation[1] of each site.

These results seem to foster the hypothesis that the variability of substitution rates, both along-chain and in time, is strongly connected with coevolution and that a mutation triggers indeed the acceptance of other mutations in nearby sites for

[1] In this chapter we are going to confound the time in which a mutation arises with the time at which it is fixated by evolution. We will call them indifferently *time of mutation* or *time of substitution*. This approximation becomes acceptable when dealing, as in our case, with time scales much longer than the interval between the appearance and fixation of a mutation.

a limited time. The minimal model described here, even if simple, may provide a handy implementation of combined space and time variation of substitution rates that might be included in more complex models. For example, it could be combined with a model of codon or amino acid substitutions. It will also be interesting for the future to encompass the idea of substitution avalanches into existing algorithms for pairwise and multiple protein sequence alignment.

5.2 Methods

5.2.1 Experimental along-chain correlation of substitutions

We first describe how we selected the data used to test our model. We retrieved these data from UniRef (Suzek *et al.*, 2007), an arrangement of the UniProt database (The UniProt Consortium, 2015) that clusters sequences above a certain sequence identity threshold:

- We downloaded from UniRef the clusters at 50% of sequence identity with at least one human sequence[2].
- From each of these clusters we detected the human sequence and aligned it with each of the others. Sequences were aligned locally by the algorithm *water* (Smith and Waterman, 1981) in the *emboss* software package (Rice *et al.*, 2000) and only ungapped parts of at least 80 residues were considered.
- We splitted the full range of 50-100% of sequence identity into windows of 2% of sequence identity each (50-52%, 52-54% ...98-100%) and then collected at most one ungapped alignment per cluster per window of sequence identity, to avoid weighting bigger clusters more.
- Each alignment was translated to a binary sequence, with 0 corresponding to two identical paired amino acids (a persistence) and 1 to different paired residues (a substitution).
- For each window of sequence identity s we computed the average sequence identity ($seqID$) between the observed alignments.

[2]The alignments were downloaded on the 23/07/2015 from UniRef at `http://www.uniprot.org/help/uniref` with query: [query:count:[2 TO $*$] length:[50 TO *] taxonomy:Homo sapiens (Human) [9606] AND identity:0.5].

- For each window of sequence identity s, we computed the overall conditional probability of having a substitution d sites away from another substitution by

$$P_s^{data}(d) = \frac{\sum_{a \in alignments} N_s^{a,1}(d)}{\sum_{a \in alignments} \left[N_s^{a,0}(d) + N_s^{a,1}(d) \right]} \tag{5.1}$$

  where $N_s^{a,1}(d)$ and $N_s^{a,0}(d)$ are respectively the number of 1 and 0 found at a distance d from another substitution in alignment a, where the first and last 5 residues were neglected to reduce boundary effects.

5.2.2 Contact probability between protein sites

A fundamental quantity for our approach is the contact probability, $G(|i-k|)$, between two sites (i,k) separated on the primary sequence by $|i-k|$ sites. We computed it from a set of proteins belonging to the top500H database (Lovell *et al.*, 2003) by considering two residues to be *in contact* when the distance between their α carbon was shorter than 6.5 Å. We observed that at short distance this contact probability can be effectively approximated by the function

$$G(|i-k|) \equiv \exp\left(|i-k|/\xi\right) \tag{5.2}$$

where the exponential fit on real data with $|i-k| < 6$ gave $\xi = 5.6$.
We also tested other functional forms and observed that, notably, it is necessary for the function $G(\cdot)$ to decrease exponentially in order to reproduce the experimental data.

5.2.3 Mean field model

We developed a minimal dynamic model for the substitution rates in protein sequence evolution in which mutations in the neighborhood of recently mutated sites have increased chances to be fixed by natural selection. We first model the temporal evolution of substitution rates by treating protein contacts in a mean field (MF) approximation, while in the next section we present a modified version which stochastically accounts for specific contact maps.

In the mean field approximation, we model the substitution rate of site i at time t by:

$$r_{MF}^i(t) = r_0 + J \sum_k G(|i-k|) \cdot \exp\left[-(t-t_k)\right] \tag{5.3}$$

where the sum runs over all the protein sites and t_k corresponds to the time of the last mutation at site k. All times are measured in units of an implicit memory time and are thus dimensionless. There are two different terms involved in equation (5.3), whose relative weight is fixed by parameter J that is the only free parameter of this model. The first term consists in a constant rate r_0 describing a uniform mutational background and the second term is a coupling term characterized by the spatial kernel $G(\cdot)$ and the memory kernel $\exp[-(t-t_k)]$. We will show that the results are insensitive to the value of r_0 in the limit of small enough r_0, which is implicit in our assumption of coevolutionary-driven rate variation. In the coupling term, the function $G(|i-k|)$ approximates the short-range contact probability as described above in equation 5.2, while the memory kernel reduces the impact of a substitution on the rest of the chain as time passes. If no other substitution appears in the neighborhood, after a sufficient amount of time, the substitution rates of that zone recovers their unperturbed value r_0. From a biological point of view this mimics the case in which a segregating mutation keeps being transmitted from generation to generation without the early emergence of any compensatory mutations, being probably a mutation with no significant detrimental effects on protein functions. An example of the evolution of the rate profile obtained with this model is shown in figure 5.1, where the rate of a sequence of 500 residues is plotted at different times.

Notice that, since we do not include any information on the precise sequence of amino acids, each sequence S in our model is only characterized by a length N_S and by the vector of the times of latest mutation for each site $\{t_k\}_{k=1,\cdots,N_S}$.

5.2.4 Stochastic contact model

We also introduce a modified version of the previous model which accounts for both short-range and long-range contacts by means of stochastic contact (SC) maps. We then model the substitution rate of site i at time t by

$$r^i_{SC}(t) = r_0 + J\sum_k C_{(i,k)} \exp[-(t-t_k)] \tag{5.4}$$

where J, r_0 and t_k have the same meaning as in eq. 5.3 and $\boldsymbol{C}$ is a contact map, namely $C_{(i,k)} = 1$ if sites i and k are in contact and $C_{(i,k)} = 0$ otherwise.

With this model the desired observable quantities are obtained by averaging on

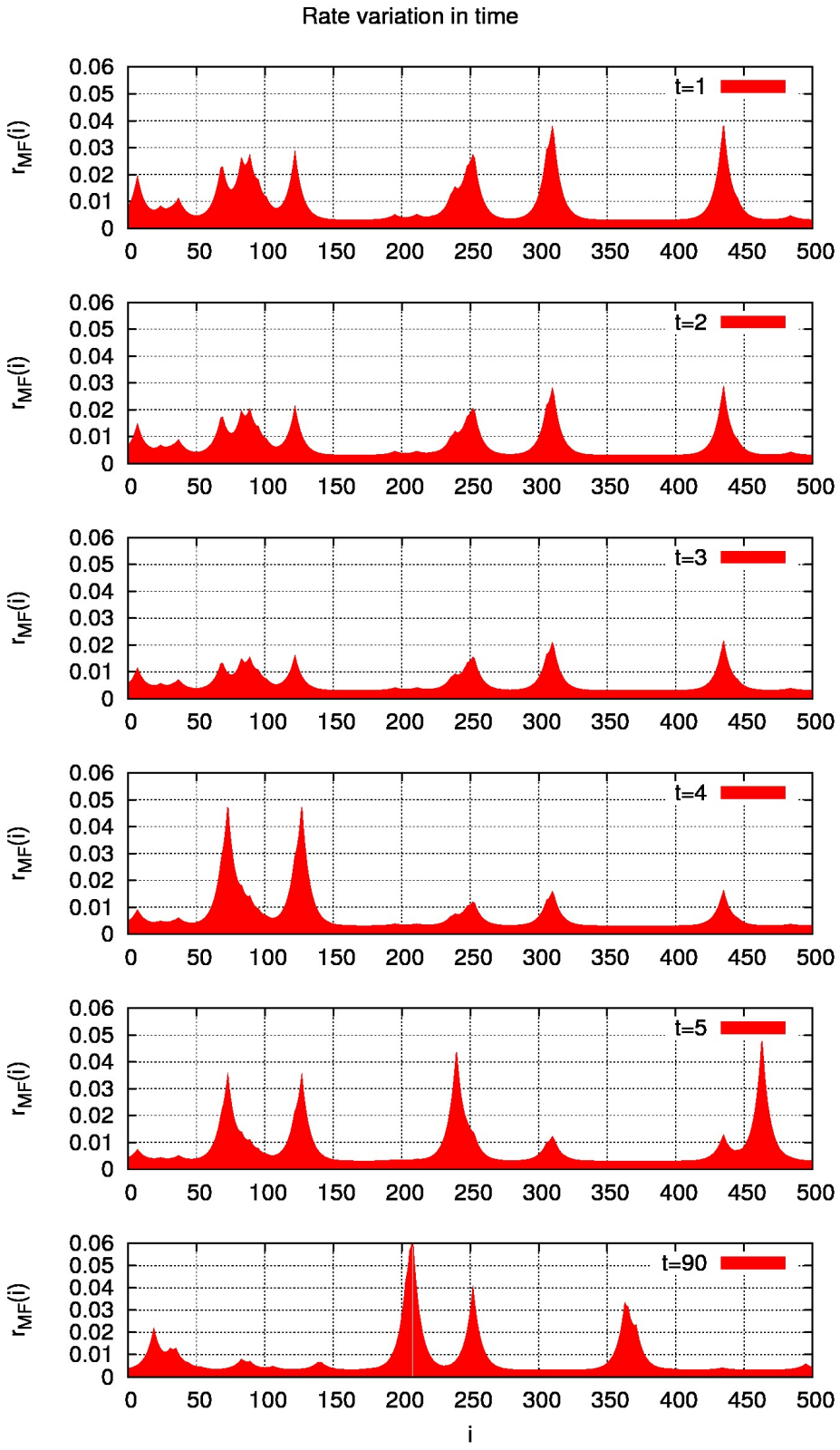

Figure 5.1: *Profile of the rates on a 500 residue sequence plotted at six different times of a same simulation. The simulation was performed using the mean field model with* $r_0 = 0.003$ *and* $J = 0.045$*. Each screenshot is separated from the previous and the following ones by a time interval of 1, corresponding to the memory time, except for the last screenshot which is separated from the previous one by a much longer time.*

different realizations of contact maps $\boldsymbol{C}$ in consistency with the contact probability:

$$P^c_{i,k} = a(N_S) \cdot G(|i-k|) + b(N_S)$$

where $G(|i-k|)$ is the average short-range contact probability defined in eq. 5.2 and $a(N_S)$ and $b(N_S)$, functions of the length of the considered sequence, are determined in compliance with a realistic partition of contacts between short-range ($|i-k| <= 6$) and long-range ones ($|i-k| > 6$). In particular we observed that in the proteins belonging to the top500H database (Lovell *et al.*, 2003) the following conditions hold:

$$\begin{aligned} \sum_{j=1, j\neq N/2}^{N} P^c_{N/2,j} &= 13.58 \\ \sum_{j=N/2-6, j\neq N/2}^{N/2+6} P^c_{N/2,j} &= 6.58 \end{aligned}$$

This means that, on average, a residue makes 13.58 contacts, 6.58 of which with nearby residues. The quantities $a(N_S)$ and $b(N_S)$ are thus computed from these equations.

5.2.5 Simulations and parameter optimization

For both the mean field and the stochastic contact models, rate and sequence evolutions were simulated by discretizing the time into small time steps dt and computing the probability of substitution at each site i in such time intervals by:

$$p_i(t) = r_i(t) \cdot dt$$

Each site i mutates during that time step if a random number drawn from a uniform distribution in $[0, 1]$ is smaller than p_i. We verified that, with our choice of dt, two or more substitutions along the chain occurred at the same time step in less than 1% of the cases.

During each simulation we kept track of the number of mutated sites as well as of the number of substitutions per site (k_i). In this way we could stop the simulation when the desired number of sites had mutated and therefore collect data for different sequence identities. The sequence identity was measured between the original sequence and the final one. We also made the simplifying approximation

that consecutive substitutions at the same site can not bring the site back to its initial amino acid type: the number of mutated sites is simply the number of site that mutate at least once. By simulating sequences of different lengths we observed that in the limit of large protein sizes (larger than 4000) the results do not depend on the size. However, the same is not true for the typical lengths in the test set of observed alignments. So, for each range of sequence identity we simulated many sequences whose distribution of lengths was compatible with the experimental one and obtained the desired quantities as averages over these different realizations. For the stochastic contact version of the model we also simulated many contact maps for each length and averaged on both lengths and contact maps. The optimization of parameter J has been accomplished by minimizing the root mean square displacement (RMSD) between the experimental $P_s^{data}(d)$ (eq. 5.1) and the corresponding quantity computed from the simulations. Only data corresponding to distances d along the chain shorter than 30 amino acids have been used during this optimization.

5.2.6 Data from Influenza Hemagglutinin

We also qualitatively validate our model on sequences of Influenza Hemagglutinin, a viral protein exposed to strong evolutionary pressure which has been largely studied and systematically sequenced in the last thirty years.

These sequences were downloaded (April 27^{th} 2016) from the NIAID Influenza Research Database (Squires *et al.*, 2012) by selecting protein data of virus type A and subtype H3N2 for the period 1981-2015 in Homo sapiens (complete segments only). Each sequence is characterized by a year, so its temporal evolution can be easily reconstructed. On average, the sequences of the same year are much more similar among themselves than to those of other years (Łuksza and Lässig, 2014). So, for each year, we computed a consensus sequence which has at each position the most common amino acid.

In order to search for avalanches of substitutions on this dataset, we followed the following procedure:

- From the PDB database (Berman *et al.*, 2000) we retrieved the entry 2WR0, containing one x-ray structure of the homotrimer of Influenza Hemagglutinin.

- The sequences downloaded from the Influenza Research Database have all of the same length and so they can be considered as a multiple sequence alignment. Using the program *PdbTool* (`https://github.com/christophfeinauer/PdbTool`.

`jl`) we were able to map each position in the sequence dataset to the corresponding residue on the PDB file.

- From the PDB entry we built a contact map between the protein residues by considering in contact two residues whose α-carbons are nearer than 8.5Å. We mapped this contact map on the indexes of our MSA obtaining matrix $C_{(i,j)}$, with i and j in $[1, 566]$ and $C_{(i,j)} = 1$ if the residues corresponding to sites i and j on the PDB are in contact and $C_{(i,j)} = 0$ otherwise. Of course not all the 566 residues in the MSA were mapped on a residue of the PDB, so we also added a contact between these sites i not mapped on the the PDB file and their along-chain neighbors j ($C_{(i,j)} = 1$ if i or j not mapped on the PDB and $|i - j| < 5$).

- Then we compared the consensus sequences of consecutive years and translated this information into a binary matrix $m_{i,t}$, where $m_{i,t} = 0$ corresponds to a match between the amino acids at site i in the consensus sequences at times t and $t - 1$ and a $m_{i,t} = 1$ corresponds to a mismatch (implying that a mutation got fixed during that year). This operation is particularly easy on this dataset because the sequences maintained the same number of amino acids and the multiple sequence alignment is consequently ungapped. Matrix m is, then, a 566 (sites) x 33 (years) matrix.

- We built a network whose nodes are labeled (i, t) by a site i and a year t and whose links $l_{(i,t_1)(j,t_2)}$ obey the following conditions:

$$\begin{aligned} &l_{(i,t_1)(j,t_2)} = 1 \quad if\ C_{(i,j)} = 1 \ \wedge\ |t_1 - t_2| \leq 2 \ \wedge\ m_{i,t_1} = m_{j,t_2} = 1 \\ &l_{(i,t_1)(j,t_2)} = 0 \qquad\qquad\qquad\qquad otherwise \end{aligned}$$

  This means that nodes (i, t_1) and (j, t_2) are connected if they both correspond to a substitution, if the two examined protein sites are in contact and if the two substitutions took place at similar times (precisely if their associated years differ by no more than 2 years).

- We found the connected subgraphs of this network and computed their size. These sizes are a measure of how correlated in space and time the substitutions are.

- To verify whether the observed distribution of these sizes can be observed by chance, we separately shuffled the temporal axis and the label of the

protein sites. When shuffling the temporal axis we considered 100 independent permutations of time labels $\tau(t)$ with the constraint that, for each pair of t_1 and t_2 if $|t_1 - t_2| \leq 2$ then $|\tau(t_1) - \tau(t_2)|$ must be larger than 1. When shuffling the indexes of the protein sites we considered again 100 independent permutations of the 566 indexes. For each of these permutations we computed again the network described above and its connected subgraphs and we compared the ones observed in the original dataset.

5.3 Results

5.3.1 Along-chain conditional probability of substitutions

We first investigate the along-chain conditional probability of having a substitution d sites away from another one, $P_s^{data}(d)$, in real alignments (red curves in fig. 5.2). At all sequence identity ranges this quantity exhibits a strong correlation decreasing with the distance. In the long-distance regime the conditional probability of observing two substitutions is approximately equal to the single-point substitution probability.

In the curves at lower sequence identity, the experimental pattern is perturbed by a short-range periodic modulation which is made evident by peaks in the conditional probability at distances of 3-4,7,10-11 and 15 amino acids. As can be noticed from figure 5.3, this modulation almost completely disappears after filtering out the sequences that JPred4 (Drozdetskiy *et al.*, 2015) predicts to have a fraction of α-helical residues larger than 38%. This is a strong hint that the spatial correlation is related with structural contacts between residues.

For each analyzed sequence identity ranges s, we than compare $P_s^{data}(d)$ with the corresponding predictions of our models. In this first part all the simulations have been performed by using the mean field model described by eq. 5.3. We first set $r_0 = 0.003$ and later on we will discuss the effects of changing this parameter. The blue dashed lines in figure 5.2 show the conditional probability of substitution for four different sequence identities as predicted by our model. It is evident that the model accurately reproduces the experimental probabilities in each investigated case. Sizable deviations from experimental data are visible only for $10 < d < 50$ at the 90% of sequence identity. We verified that this could be cured by employing a different models for the contact probability[3] but the modification would introduce a similar deviation from experimental values at lower sequence identities. This deviation may

[3]we successfully tried with the contact probability of the gaussian chain (Doi and Edwards, 1988)

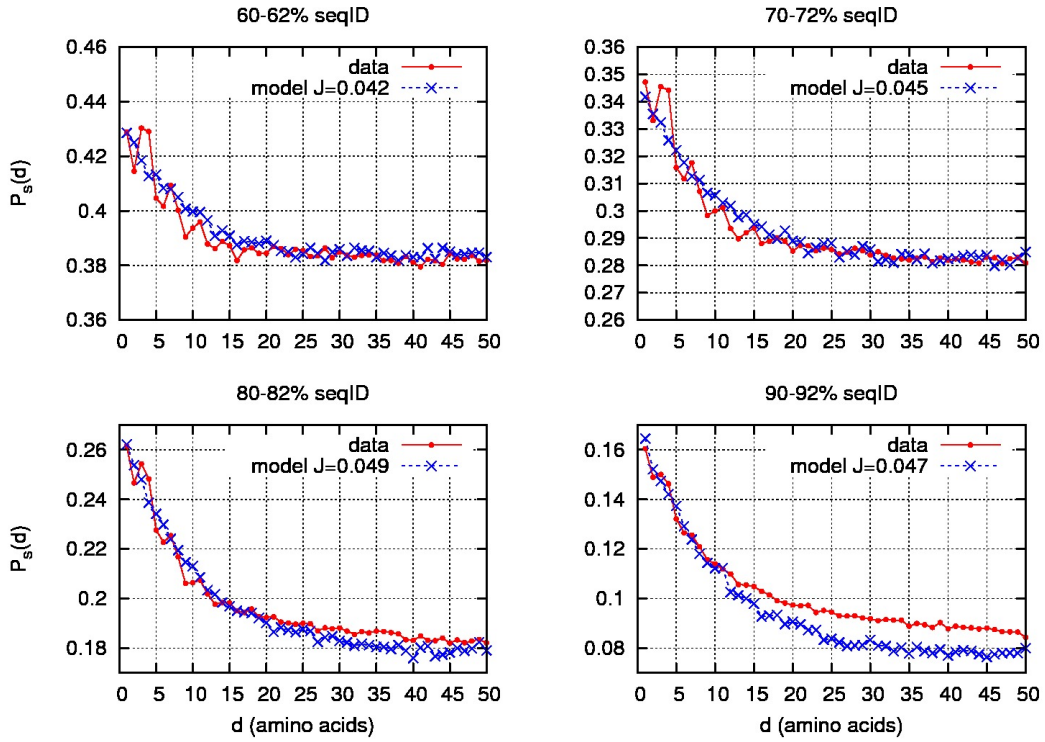

Figure 5.2: *Conditional probability $P_s(d)$ of observing a substitution d sites away along the chain from another substitution at various sequence identities s, respectively (from top left to bottom right) 60-62%,70-72%, 80-82% and 90-92%. Mean field model in blue and data in red. Parameter J is optimized at each sequence identity and $r_0 = 0.003$.*

be due to our requirement of ungapped alignments, which at low sequence identity seems to select more structured regions with respect to higher sequence identity. With this in mind, it is natural, then, to expect slightly different average contact probabilities, slightly different values of J and a different fraction of α-helices in the pool of structures. In short, these effects can be attributed to the non uniformity in the set of data at various sequence identity rather than to the model.

The optimal values of J does not vary strongly with respect to the sequence identity (see the keys in figure 5.2). We also computed $P_s^{model}(d)$ with a common value of J at all sequence identities (figure 5.4). We found only marginal degradation with respect to the case in which J is free to vary. This indicates that, at this level of accuracy, we can assume that J is basically constant.

To prove that the results do not depend on the choice of r_0 in the limit of small r_0 we plot in figure 5.5 the correlations obtained by the mean field model at 70-72% of sequence identity by using different values for r_0. From this comparison it is evident

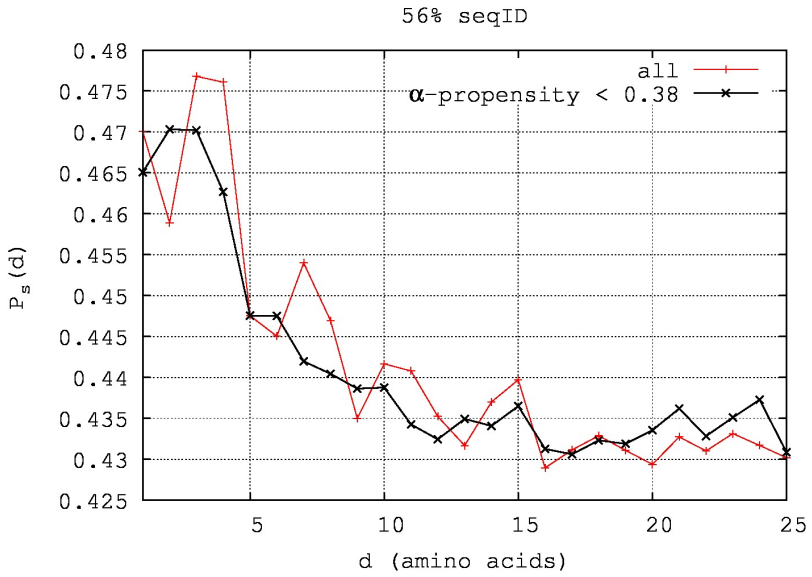

Figure 5.3: *Experimental conditional probability $P_s(d)$ of observing a substitution d sites away along the chain from another substitution at sequence identity 56% computed on all the available alignments (red) and on a subset characterized by α-helical propensity smaller than 0.38 according to the JPred4 predictor (Drozdetskiy* et al.*, 2015).*

that $P_s(d)$ does not significantly vary, so r_0 is substantially irrelevant as long as small and different from zero.

With the stochastic contact model we obtain an agreement similar to the case of the mean field model. These comparisons are shown in figure 5.6.

5.3.2 Distribution of the number of substitutions per site

In most algorithms for phylogenetic reconstruction based on likelihood maximization, the among-site rate variability is modeled by a Γ-distribution whose shape parameter α is estimated from the distribution of the number of substitutions per site. Indeed, when dealing with a mixture of Poisson processes characterized by Γ-distributed rates, the number of substitutions per site is necessarily a negative binomial whose shape parameter is the same α (see chapter 2). The probability to have k substitutions at a site is then:

$$P(k|\alpha, \langle k \rangle) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha) \cdot k!} \left(\frac{\langle k \rangle}{\langle k \rangle + \alpha} \right)^k \cdot \left(\frac{\alpha}{\langle k \rangle + \alpha} \right)^\alpha \tag{5.5}$$

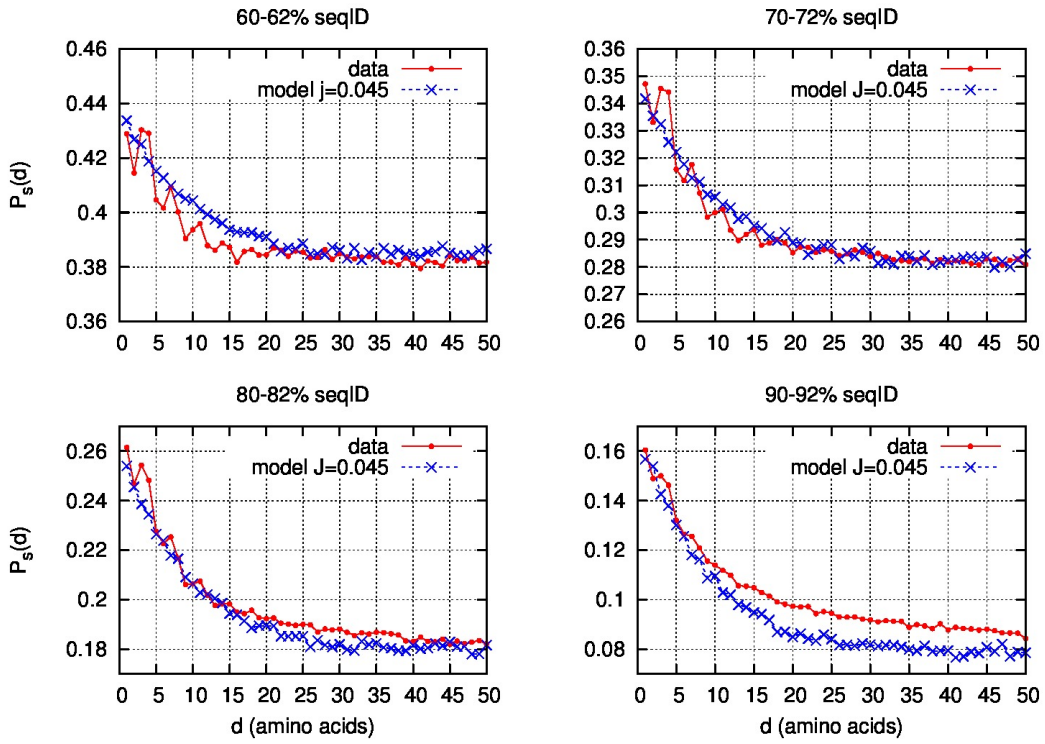

Figure 5.4: *Conditional probability $P_s(d)$ of observing a substitution d sites away from another substitution at various sequence identities s, respectively 60-62%,70-72%, 80-82% and 90-92% obtained by using a single value of J at all sequence identities and $r_0 = 0.003$. Simulations were done in the mean field version of the model.*

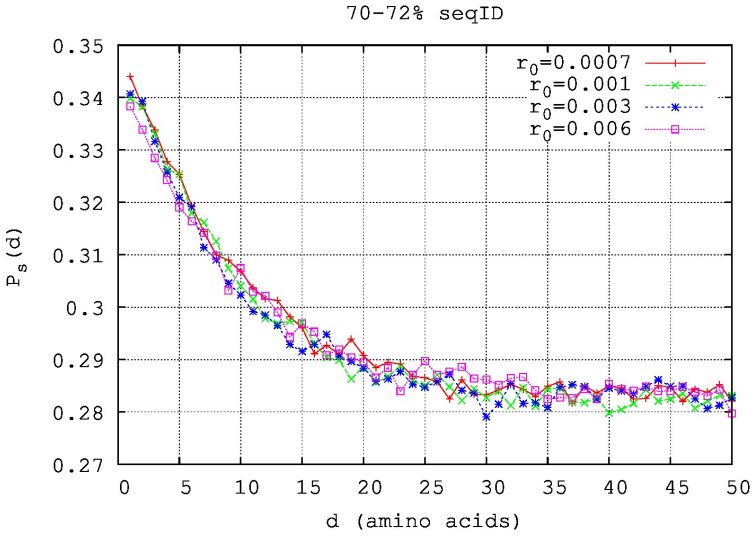

Figure 5.5: *Conditional probability $P_s(d)$ of observing a substitution d sites away from another substitution in the sequence identity range 70-72% obtained for different r_0 with our mean field model. In each simulation J was fixed to 0.045.*

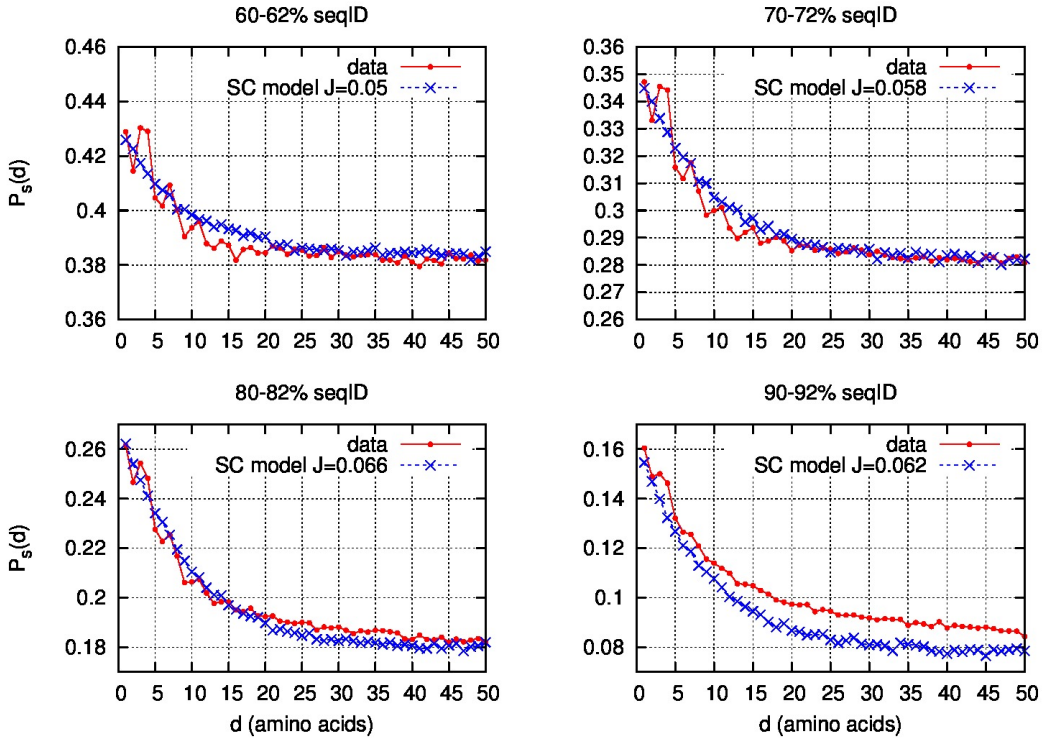

Figure 5.6: *Conditional probability $P_s(d)$ of observing a mutation d sites from another mutation at various sequence identities s, respectively 60-62%,70-72%, 80-82% and 90-92%. Stochastic contact model in blue and data in red. Parameter J is optimized at each sequence identity and $r_0 = 0.003$.*

where $\langle k \rangle$ is the average number of substitutions per site computed from the data and α is the parameter to be estimated.

We here show that also our model predicts a negative binomial for $P(k|\alpha, \langle k \rangle)$, consistently with what inferred for real substitutions from phylogenetic trees (Gu and Zhang, 1997). We kept track of the number of substitutions per site and, for each sequence identity range s, we analyzed its distribution in our simulations with the value of J previously optimized on the experimental spatial correlations $P_s(d)$.

In figure 5.7 we show the normalized histogram of the number n_k of substitutions per site k predicted by our model at four different sequence identities. On top of each histogram we show its fit with a negative binomial which was performed by weighting each bin according to its error ($\sqrt{n_k}$). It is clear that the negative binomial in all four cases reproduces well the statistics of the substitutions in our model.

Moreover, the estimates of the shape parameter α, especially when computed by the stochastic contact model, fall in a range compatible with those observed in real protein families when dealing with Γ-distributed substitution rates. Contrary to the models with Γ-distributed rates across sites, in our model the rate variability is totally ascribed to the action of compensatory mutations. This is a severe approximation, since in real proteins the structure is well known to affect the rates. However, the model presented in this work, thanks to its schematic formulation, suggests the possible importance of coevolution in determining rate variability. Indeed, according to our model, sites will have a rate that changes during their history. Then, the value of α describes the distribution of time-averaged rates rather than real rates. As we reduce the sequence identity, we are more and more dealing with averaged rates, which are by definition more uniform than instantaneous rates. This corresponds to larger values of parameter α. We have discussed in section 2.3.1 of chapter 2 the relationship between parameter α and the degree of uniformity of the underlying rates. In figure 5.8 we quantify this phenomenon by showing the value of α obtained by fitting to equation 5.5 the number of substitutions per site obtained by our model at different sequence identities. The estimated values of α span about an order of magnitude for the MF model, while are sensibly smaller for the more realistic stochastic contact implementation. We have seen in chapter 2 that similar results are found in real data, as proved, among others, by Gaucher *et al.* (2001, 2002). In chapter 2 we have also shown the progressive growth in the estimation of α for five Pfam families (see figure 2.8). Our model, especially in the stochastic contact version, predicts a similar trend but with a larger variation of α with respect of the examples discussed above. This might depend on our simplifying approximation of equal dynamics on

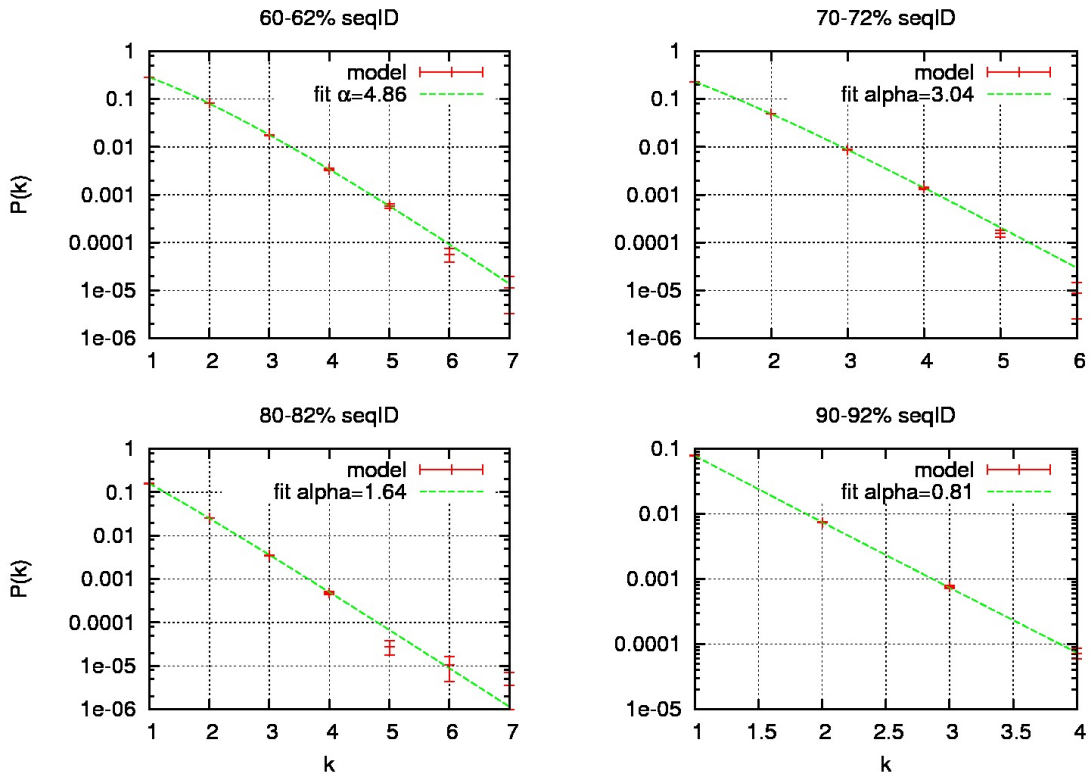

Figure 5.7: *Weighted fit of the normalized histogram of the number of substitutions k per site to a negative binomial distribution at various sequence identities. The fit defines the value of α written in the key. The reduced χ^2 of these fit are respectively, from top−left to bottom−right: 0.827, 1.129, 0.948, 0.060.*

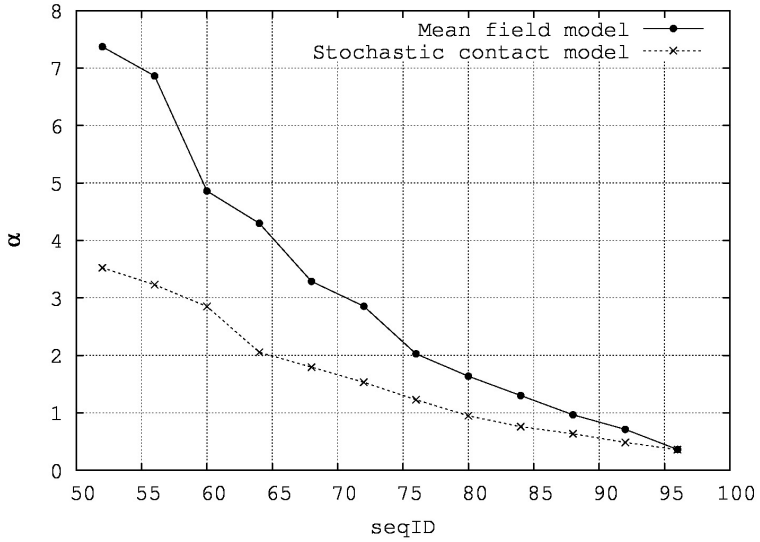

Figure 5.8: *Estimated α by our models as a function of the sequence identity. The value of α is estimated by fitting the number of substitutions per site to a negative binomial (eq. 5.5).*

all sites. This approximation could be straightforwartly weakened by increasing the number of parameters. For example it would be possible to stochastically divide sites among two or more classes each characterized by a different J. This would allow the simultaneous presence of slow-varying-sites and fast-varying-sites and keep a more reasonable difference among the rates also for longer evolutionary times.

5.3.3 Avalanches of substitutions on Influenza Hemagglutinin

We have described a model that enhances the probability of correlated substitutions taking place at similar times in sites which are in contact. This process can be resumed by saying that our model produces avalanches of substitutions confined not only in space, as commonly expected by coevolution, but also in time. We searched for qualitative footprints of such avalanches in the sequence evolution of Influenza Hemagglutinin, a viral protein exposed to strong evolutionary pressure which has been largely studied and systematically sequenced in the last thirty years. The procedure used to identify these avalanches is described in detail in section 5.2.6. Figure 5.9 shows each substitution found in the time evolution of Hemagglutinin by a square placed in correspondence of its site (x axis) and its year of appearance (y axis). Black squares correspond to isolated substitutions, which, according to our

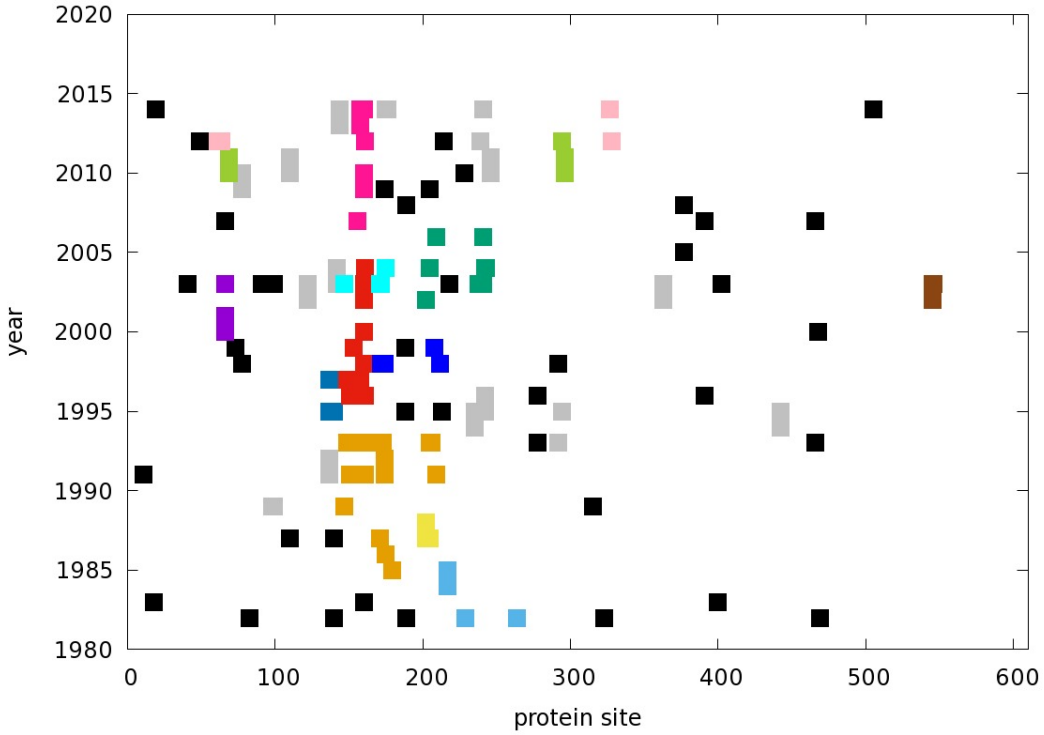

Figure 5.9: *Avalanches of substitutions on sequences of Influenza Hemagglutinin. Each square represents one substitution which took place on the site corresponding to the x value and in the year corresponding to the y value. Black squares are isolated substitutions, grey ones are substitutions in avalanches made up by two substitutions. The other substitutions are colored according to the avalanche (or subgraph) to which they belong according to the procedure described in section 5.2.6*

procedure and thresholds, do not belong to any avalanche. Grey squares label those substitutions which are part of small avalanches, composed by only two substitutions. The remaining squares are colored according to the avalanche to which they belong. Notice that there may be substitutions with very different x value still belonging to the same avalanche: this is because the spatial proximity is here computed three-dimensionally from the PDB structure of the protein, thus presenting contacts also between some sites which are far away along the sequence.

The exact partitioning in avalanches depicted in figure 5.9 is a consequence of the threshold chosen for the memory when analyzing the data (here set at three years). Similar results have been obtained by choosing two or four years as alternative temporal memory. This choice is somehow arbitrary, but a significant degree of time correlation is qualitatively visible in the rough data and this choice helps highlighting this correlation. Indeed, avalanches estimated with the same procedure but on a dataset in which either the years or the protein sites have been reshuffled according to the procedure described in section 5.2.6 are sizably smaller. In the real case the fraction of residues which are part of an avalanche is 72%, while they are only the 63% when the temporal axis is shuffled and the 57% when the indexes of the sites are shuffled.

Unfortunately no quantitative comparison can be performed between our models and such data, because Influenza Hemagglutinin is one rare case in which the temporal evolution of the sequence is known thus does not need to be inferred. We are not aware of any dataset giving similar information on scales large enough to allow quantitative comparisons with the avalanches predicted by our model.

5.4 Discussion

It is well known that substitutions exhibit a strong spatial correlation along protein sequences which vary with the sequence identity between the aligned sequences. It is also well known that the rate of substitution varies significantly among sites (Yang, 1993, 1994, 1995; Halpern and Bruno, 1998) and in time (Fitch and Markowitz, 1970; Gaucher *et al.*, 2001; Lopez *et al.*, 2002). In this chapter we show how these two phenomena can be explained together by a simple model based on the idea that mutations may perturb the stability and functionality of a protein by introducing (or reducing) frustration. As a consequence, all sites that are in contact with a mutated one are themselves stimulated to accept mutations in order, for instance, to reduce frustration again. This idea has alredy emerged in the domain of protein sequence

coevolution (de Juan *et al.*, 2013; Gobel *et al.*, 1994; Weigt *et al.*, 2009; Morcos *et al.*, 2011; Ekeberg *et al.*, 2013; Burger and Van Nimwegen, 2010; Jones *et al.*, 2012): the novel ingredient that we introduce in this work is that this perturbation acts only for a finite time. In fact, if an isolated mutation has survived for many generations, this means that it has no more effects on the contacting residues. We presented two models based on this idea, which revisit the covarion model proposed for the first time by Fitch and Markowitz (1970). The first one is characterized by a mean field approximation of short-range contact probability, while the second one, slightly more complex, deals stochastically both with short-range and long-range contacts. Both these models accurately reproduce the observable along-chain correlation of substitutions in a large range of sequence identity (50-100%). This suggests that the observed correlation may be due to structural contacts between residue pairs, as also confirmed by the peaks at the α-helical contact periodicity that vanish once the predicted α-helices are removed from the dataset.

Remarkably, the same models also reproduce a distribution of the number of substitutions per site largely consistent with a negative binomial, a distribution also observed in phylogenetics. If the negative binomial distribution produced by our model is interpreted by the traditional Γ-distributed rates, it shows the typical overestimation of the Γ shape parameter for lower sequence identities due to the time variation of rates. The amount of overestimation is much larger in our mean field model with respect to real data, while the stochastic contact model gives more meaningful predictions, even if still significantly larger than data. We suppose that any inclusion of time-independent differences among site will reconcile the stocastic contact's estimation and data at the expense of a larger number of parameters (for example different coupling terms for different structural motifs).

The models presented here predict that substitutions, far from being isolated, are part of a complex spatio-temporal pattern and gather in avalanches confined in space and in time.

The simplicity of these models is showed by the presence of a single free parameter, which tunes the strength of coupling between the protein sites. This parameter seems to be similar at all the analyzed sequence identities, suggesting that sequence evolution can be fairly approximated by a stationary process. Our model employs average contact probabilities, but could be easily modified to include given contact maps for more specific applications.

Our results seem to underline the importance of accounting for coevolution in the modeling of substitution rates. The minimal models described here may provide a

handy implementation of combined space and time variation of substitution rates that might be included in a more complex framework and, for example, combined with a model of codon or amino acid substitutions. Moreover, in view of the observed pattern of along-chain correlation of substitutions it would be interesting to account for an increased probability of nearby substitutions into the existing models and algorithms for protein sequence alignment, especially for pairwise alignments where no a priori information is available on the substitution rate or on the structural and functional constraints.

Concluding remarks and future perspectives

In this thesis we have presented three original contributions to basic-level modeling of protein sequence evolution. In chapter 3 we introduced a simple procedure that allows including, in a mean-field-like framework, the among-site rate variability in the time evolution of PAM-like substitution matrices. We showed that including rate variability leads to evolutions violating the Markov assumption when considering the whole protein sequence even when using codon matrices. In chapter 4 we proposed a procedure for deriving a substitution rate matrix from Single Nucleotide Polymorphism (SNP) data. We showed that this matrix faithfully describes substitution rates for short evolutionary times, if one takes into account the rate variability as described in Chapter 3. Finally, in chapter 5, we presented a simple model, inspired by coevolution, capable of predicting at the same time the along-chain correlation of substitutions and the time variability of substitution rates. The model is based on the idea that a mutation at a site enhances the probability to fix mutations in the other protein sites in its spatial proximity, but only for a certain amount of time: after this time, the occurrence of this mutation is effectively forgotten.

We showed that the approach introduced in chapter 3 is well suited to be applied on matrices learnt at very short evolutionary times such as our matrix obtained from SNPs or a matrix obtained from alignments with sequence identity higher than 98%. We also compared the performance of the substitution matrix learned from the SNPs with more standard ones such as JTT (Jones *et al.*, 1992) or WAG (Whelan and Goldman, 2001). This analysis convinced us that our substitution rate matrix obtained from SNPs is appropriate to describe sequence evolution in the limit of short evolutionary time. In this regime, it is important to take into account the non-Markovianity due to the combined effects of the among-site overall rate variability and of the degeneracy of the genetic code. Standard substitution rate matrices are, on the other hand, derived in the framework of a Markovian approximation, which

is correct and appropriate when evolutionary times are longer. We estimated the crossover between these two regimes to be around 85% of sequence identity. These two dynamics are described by different propagators: at high sequence identity one should use a propagator accounting for the among-site rate variability and the codons, while at medium-low sequence identity a simple Markovian propagator is more appropriate. In our opinion, only the first class of models may hope to properly reproduce the dynamics in the short evolutionary time regime. The attempt described in chapter 4 to evolve our substitution rate matrix on codons derived from SNPs by ensemble averages on the among-site distribution of substitution rates attains this result only partially: the obtained dynamics describes better than Markovian models the evolution at short timescales, but its performance progressively degrades at lower sequence identities. A natural development of the results presented in this thesis would be developing a model capable to describe both regimes with a precision at least comparable with that obtained by our SNP model at high sequence identity. For example, we are considering the possibility to insert the information derivable from SNPs in a mutation-selection model like the one described in ref. (Tamuri *et al.*, 2012). More in general, we hope that, now that we have shown that SNPs can be effectively employed to describe the short-time regime of protein sequence evolution, this information will be embedded in more precise and complex frameworks and appropriately integrated with the information available in proteomics.

In perspective one could also combine our substitution rate matrix obtained from the SNPs with the model of rate evolution described in chapter 5 and assess whether the inclusion of effective time variability of substitution rates can improve the performances presented in chapter 4. In particular one could apply this combined model on protein families where the three-dimensional structure is known and then the real contact map can be exploited to infer a precise model of the rate evolution.

Another advantage of deriving a substitution rate matrix out of SNP data is that scoring matrices, which are used to align sequences in the first steps of most bioinformatic analysis, are themselves derived from alignments and, a priori, it is not clear whether this circular process could induce artifacts in the quality of the results. Deriving a substitution matrix from data that are totally independent from alignments is, according to us, an important result: if no artifact is found by applying a matrix derived from the SNPs, this gives further credit and larger theoretical validity to the current procedures.

We also believe that it would be interesting to investigate the relationship between the instantaneous substitution rate matrix and the chemical similarity of amino acids.

If the rate matrix is expressed in terms of amino acid interchanges, the redundancy of the genetic code may partially hide the signal of chemical similarity. Problems may arise also due to the Markov approximation on the propagator. Instead, we believe that a rate matrix derived as described in chapter 4 could be a good starting point for such an analysis. For example, by comparing instantaneous codon interchanges, the similarity between amino acids could be quantified in terms of effective free energy barriers between their codons.

Finally, since when we started analyzing the experimental along-chain correlation of substitutions described in chapter 5, we got convinced that finding a way to include this feature in aligning algorithms could improve their performances. This could be particularly significant, for example, in pairwise alignments, where very little information is available on the specificities of the studied case: with pairwise alignments only scoring matrices and their average information can be exploited and we believe that including the spatial correlation of substitutions, even in an approximate manner, could help to achieve a better sorting of near-optimal alignments. Indeed it has been shown (Sierk *et al.*, 2010) that among the near-optimal alignments found, for instance, by dynamic programming there is big chance to find the optimal structural alignment, which is generally considered much more precise, but that is not very often the one with the best score. In some preliminary analysis we have tested that the optimal structural alignments are often among the near-optimal alignments with a higher fraction of spatially correlated substitutions. Therefore, including this information may help in better selecting which near-optimal alignment to choose. Unfortunately, it is still not evident to us how to include this feature: a spatial coupling breaks the independence of sites at the basis of dynamic programming algorithms and so new solutions should be found. The easiest way is probably to select not only optimal alignments but also near-optimal ones, for example by the forward-backward algorithm in aligning tools based on hidden Markov models (Durbin *et al.*, 1998), and to rescore them a posteriori by accounting also for the spatial correlation of substitutions. Clearly taking correlation of mutations into account would be interesting also in multiple sequence alignments, but with a further increase in the complexity of the problem.

List of abbreviations

CAT: mixture model that classifies sites into CATegories Lartillot and Philippe (2004).
COVARIONS: COncomitantly VARIable codONS (Fitch and Markowitz, 1970).
DCA: Direct coupling analysis (Weigt *et al.*, 2009; Morcos *et al.*, 2011).
DNA: DeoxyriboNucleic acid.
ECM: Empirical Codon Model Kosiol *et al.* (2007).
JTT: Jones Taylor Thornton Jones *et al.* (1992).
GMAF: Global Minor Allele Frequency. It is the fraction of individuals in a species presenting the less common between the two alleles (variants) of a polymorphism.
HMM: Hidden Markov Model.
LG: Le Gascuel Le and Gascuel (2008).
MF: Mean Field.
MSA: Multiple Sequence Alignment.
PAM: Point Accepted Mutation Dayhoff *et al.* (1978).
SC: Stochastic Contact. Referred to one of the models proposed in chapter 5.
SNP: Single Nucleotide Polymorphism.
WAG: Whelan And Goldman Whelan and Goldman (2001);

Acknowledgements

... with gratitude
to those who instilled the best doubts in me,
always *scenting my wardrobe*
with their intelligence and curiosity.

Incoraggiate coloro che cercano
di tener vigile la vostra mente
tenete in serbo i loro pensieri
metteteli dentro una cassapanca
insieme ad alcune mele cotogne
così i vostri panni avranno profumo
d'intelligenza per un anno intero

ARISTOFANE, *Le vespe*
P. RUMIZ, *La cotogna di Istanbul*

My most grateful thanks go to my supervisor, Alessandro Laio, and co-supervisor, Alex Rodriguez, for being patiently present in each single moment of this adventure, never letting me despair and always searching for new ideas and solutions at my side. I wonder if I will ever manage to get their amazing research style: working passionately, but without taking themselves too seriously, staying humble, and keeping in mind that work is far from being the only side of life.

Then, I want to thank Prof. Carolin Kosiol and Prof. Martin Weigt for accepting to read and evaluate this manuscript.

I also want to acknowledge the people who actively contributed in the work presented here: Stefano Zamuner and Andrea Pagnani for the project in chapter 5 and Xevi Biarnés for collaborating in the project in chapter 4. Their contribution, besides being important for the project, enriched me with meaningful competences and interesting perspectives. Many other people have directly or indirectly contributed to reach this final goal and it is impossible to mention them all here. I have very much enjoyed and benefited from the discussions with Edoardo Sarti, Daniele Granata, Antonio Trovato, Flavio Seno, Giovanni Pinamonti, Elena Facco, Rute da Fonseca, Marco Punta and with all the "Laio group" present and past. Special thanks go also to Michele Allegra for the precise and patient proofreading.

Another special mention goes to the SISSA Medialab, and in particular to Simona Cerrato, for believing in me and giving me the moral and material support to amuse myself by communicating science. My gratitude goes also to the friends with whom I shared these science communication adventures, from FameLab, to the Scratch lab and, in particular, to Lucia T. for being at my side in most of these crazy but demanding moments.

And then, I regret I have to thank this (literally) crazy, "surly but graceful" town[4], Trieste, which I have hated since the very first moment, and which I will keep loving much after my departure. I have spent four happy years here, and my gratitude goes to those many people, in SISSA and outside, who gave their small or large contribution for this to happen with their presence, support and friendship, starting from my (I have to say...very brave!) flatmates and to the SBP group. I cannot mention all of them here without doubling the size of this thesis, but their names are all well secured in my heart.

Finally, my gratitude goes to my parents, who grew my curiosity since its very, very beginning as the most precious creature. They, together with my expanded family in the neighborhood of Camin, are the persons to whom I owe the most of who I am and the most of what I believe in.

Thank you.

Francesca

[4]U. SABA, *Trieste*

Bibliography

1000 Genomes Project Consortium *et al.* 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319): 1061–1073.

1000 Genomes Project Consortium *et al.* 2015. A global reference for human genetic variation. *Nature*, 526(7571): 68–74.

Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. 2013. *Essential Cell Biology*. Garland Science.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3): 403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17): 3389–3402.

Anfinsen, C. B., Haber, E., Sela, M., and White, F. H. 1961. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences*, 47(9): 1309–1314.

Averof, M., Rokas, A., Wolfe, K. H., and Sharp, P. M. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science*, 287(5456): 1283–1286.

Benner, S. A., Cohen, M. A., and Gonnet, G. H. 1994. Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Engineering*, 7(11): 1323–1332.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., *et al.* 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218): 53–59.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. N., and Bourne, P. E. 2000. The protein data bank. *Nucleic Acids Research*, 28(1): 235–242.

Burger, L. and Van Nimwegen, E. 2010. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.*, 6(1): e1000633.

Cocco, S., Monasson, R., and Weigt, M. 2013. From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction. *PLoS Comput. Biol.*, 9(8): 1–17.

Cossio, P., Granata, D., Laio, A., Seno, F., and Trovato, A. 2012. A simple and efficient statistical potential for scoring ensembles of protein structures. *Scientific Reports*, 2.

Cox, D. R. and Miller, H. D. 1977. *The theory of stochastic processes*, volume 134. CRC Press.

Dayhoff, M. and Eck, R. 1968. A model of evolutionary change in proteins. *In Atlas of Protein Sequences and Structure*, pages 33–41.

Dayhoff, M., Schwartz, R., and Orcutt, B. 1978. A model of evolutionary change in proteins. *In Atlas of Protein Sequences and Structure*, 5: 345–352.

de Juan, D., Pazos, F., and Valencia, A. 2013. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4): 249–261.

De Maio, N., Holmes, I., Schlötterer, C., and Kosiol, C. 2012. Estimating empirical codon hidden Markov models. *Molecular Biology and Evolution*.

De Maio, N., Schrempf, D., and Kosiol, C. 2015. Pomo: An allele frequency-based approach for species tree estimation. *Systematic Biology*, 64(6): 1018–1031.

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.-M., Taly, J.-F., and Notredame, C. 2011. T-coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Research*, 39(suppl 2): W13–W17.

Doi, M. and Edwards, S. F. 1988. *The theory of polymer dynamics*, volume 73. Oxford University Press.

Doron-Faigenboim, A. and Pupko, T. 2007. A combined empirical and mechanistic codon model. *Molecular Biology and Evolution*, 24(2): 388–397.

Drozdetskiy, A., Cole, C., Procter, J., and Barton, G. J. 2015. Jpred4: a protein secondary structure prediction server. *Nucleic Acids Research*.

Dunn, S. D., Wahl, L. M., and Gloor, G. B. 2008. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3): 333–340.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. 1998. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press.

Echave, J., Spielman, S. J., and Wilke, C. O. 2016. Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics*, 17: 109–121.

Eddy, S. R. 1998. Profile hidden Markov models. *Bioinformatics*, 14(9): 755–763.

Edgar, R. C. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5): 1792—1797.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. 2013. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Physical Review E*, 87(1): 012707.

Fares, M. A. and Travers, S. A. A. 2006. A novel method for detecting intramolecular coevolution: Adding a further dimension to selective constraints analyses. *Genetics*, 173(1): 9–23.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6): 368–376.

Felsenstein, J. and Churchill, G. A. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13(1): 93–104.

Finn, R. D., Clements, J., and Eddy, S. R. 2011. Hmmer web server: interactive sequence similarity searching. *Nucleic Acids Research*.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. 2015. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1): D279–D285.

Fitch, W. M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20(4): 406–416.

Fitch, W. M. and Markowitz, E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, 4(5): 579–593.

Fodor, A. A. and Aldrich, R. W. 2004. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Structure, Function, and Bioinformatics*, 56(2): 211–221.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution*, 18(5): 866–873.

Gaucher, E. A., Miyamoto, M. M., and Benner, S. A. 2001. Function–structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. *Proceedings of the National Academy of Sciences*, 98(2): 548–552.

Gaucher, E. A., Gu, X., Miyamoto, M. M., and Benner, S. A. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends in Biochemical Sciences*, 27(6): 315 – 321.

Gobel, U., Sander, C., Schneider, R., and Valencia, A. 1994. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18(4): 309–317.

Gonnet, G. H., Cohen, M. A., and Benner, S. A. 1992. Exhaustive matching of the entire protein sequence database. *Science*, 256(5062): 1443–1445.

Gu, X. and Zhang, J. 1997. A simple method for estimating the parameter of substitution rate variation among sites. *Molecular Biology and Evolution*, 14(11): 1106–1113.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3): 307–321.

Halpern, A. L. and Bruno, W. J. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular Biology and Evolution*, 15(7): 910–917.

Henikoff, S. and Henikoff, J. G. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Research*, 19(23): 6565–6572.

Henikoff, S. and Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22): 10915–10919.

Howe, K., Bateman, A., and Durbin, R. 2002. Quicktree: building huge neighbour-joining trees of protein sequences. *Bioinformatics*, 18(11): 1546–1547.

Huang, E. S., Subbiah, S., and Levitt, M. 1995. Recognizing Native Folds by the Arrangement of Hydrophobic and Polar Residues. *Journal of Molecular Biology*, 252(5): 709–720.

Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L. P., Marcet-Houben, M., and Gabaldón, T. 2014. Phylomedb v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Research*, 42(D1): D897–D902.

Jones, D. T., Taylor, W. R., and Thornton, J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences: CABIOS*, 8(3): 275–282.

Jones, D. T., Buchan, D. W. A., Cozzetto, D., and Pontil, M. 2012. Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2): 184–190.

Kemeny, J. G. and Snell, J. L. 1960. *Finite Markov chains*, volume 356. van Nostrand Princeton, NJ.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610): 662–666.

Kingman, J. F. 1982. On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43.

Kosiol, C. and Goldman, N. 2011. Markovian and non-Markovian protein sequence evolution: Aggregated Markov process models. *Journal of Molecular Biology*, 411.4-6: 910–923.

Kosiol, C., Holmes, I., and Goldman, N. 2007. An empirical codon model for protein sequence evolution. *Molecular Biology and Evolution*, 24(7): 1464–1479.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235(5): 1501–1531.

Lartillot, N. and Philippe, H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6): 1095–1109.

Le, S. Q. and Gascuel, O. 2008. An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, 25(7): 1307–1320.

Le, S. Q., Lartillot, N., and Gascuel, O. 2008. Phylogenetic mixture models for proteins. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1512): 3965–3976.

Lopez, P., Casane, D., and Philippe, H. 2002. Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution*, 19(1): 1–7.

Lovell, S. C., Davis, I. W., Arendall, W. B., de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S., and Richardson, D. C. 2003. Structure validation by C alpha geometry: phi, psi and C beta deviation. *Proteins: Structure, Function, and Bioinformatics*, 50(3): 437–450.

Łuksza, M. and Lässig, M. 2014. A predictive fitness model for influenza. *Nature*, 507(7490): 57–61.

Mitchison, G. and Durbin, R. 1995. Tree-based maximal likelihood substitution matrices and hidden Markov models. *Journal of Molecular Evolution*, 41(6): 1139–1151.

Miyazawa, S. and Jernigan, R. L. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3): 534–552.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49): E1293–E1301.

Mueller, T., Spang, R., and Vingron, M. 2002. Estimating amino acid substitution models: A comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Molecular Biology and Evolution*, 19(1): 8–13.

Müller, T. and Vingron, M. 2000. Modeling amino acid replacement. *Journal of Computational Biology*, 7(6): 761–776.

Nakamura, Y., Gojobori, T., and Ikemura, T. 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Research*, 28(1): 292–292.

Needleman, S. B. and Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3): 443 – 453.

Page, R. D. 2002. Visualizing phylogenetic trees using TreeView. *Current Protocols in Bioinformatics*, pages 6–2.

Pagel, M. and Meade, A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology*, 53(4): 571–581.

Papoulis, A. and Pillai, S. U. 2002. *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education.

Penny, D., McComish, B. J., Charleston, M. A., and Hendy, M. D. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *Journal of Molecular Evolution*, 53(6): 711–723.

Price, M. N., Dehal, P. S., and Arkin, A. P. 2010. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3): e9490.

Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics*, 16(6): 276–277.

Roberts, E., Eargle, J., Wright, D., and Luthey-Schulten, Z. 2006. Multiseq: unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics*, 7(1): 1–11.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P. 1996. Real-time {DNA} sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, 242(1): 84 – 89.

Rosenberg, N. A. and Nordborg, M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, 3(5): 380–390.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4): 406–425.

Schneider, A., Cannarozzi, G., and Gonnet, G. 2005. Empirical codon substitution matrix. *BMC Bioinformatics*, 6(1): 134.

Schrider, D. R., Hourmozdi, J. N., and Hahn, M. W. 2011. Pervasive multinucleotide mutational events in eukaryotes. *Current Biology*, 21(12): 1051 – 1054.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1): 308–311.

Sierk, M. L., Smoot, M. E., Bass, E. J., and Pearson, W. R. 2010. Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments. *BMC Bioinformatics*, 11(1): 1–15.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7: 539.

Smith, N. G. C., Webster, M. T., and Ellegren, H. 2003. A low rate of simultaneous double-nucleotide mutations in primates. *Molecular Biology and Evolution*, 20(1): 47–53.

Smith, T. F. and Waterman, M. S. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1): 195–197.

Squires, R. B., Noronha, J., Hunt, V., García-Sastre, A., Macken, C., Baumgarth, N., Suarez, D., Pickett, B. E., Zhang, Y., Larsen, C. N., *et al.* 2012. Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza and other respiratory viruses*, 6(6): 404–416.

Srinivasan, R. and Rose, G. D. 1995. Linus: a hierarchic procedure to predict the fold of a protein. *Proteins: Structure, Function, and Bioinformatics*, 22(2): 81–99.

Stamatakis, A. 2006. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21): 2688–2690.

Suchard, M. A. and Redelings, B. D. 2006. Bali-phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, 22(16): 2047–2048.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10): 1282–1288.

Talavera, D., Lovell, S. C., and Whelan, S. 2015. Covariation is a poor measure of molecular coevolution. *Molecular Biology and Evolution*, 32(9): 2456–2468.

Tamuri, A. U., dos Reis, M., and Goldstein, R. A. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics*, 190(3): 1101–1115.

The UniProt Consortium 2015. Uniprot: a hub for protein information. *Nucleic Acids Research*, 43: D204–D212.

Thompson, J. D., Gibson, T., Higgins, D. G., *et al.* 2002. Multiple sequence alignment using ClustalW and ClustalX. *Current protocols in bioinformatics*, pages 2–3.

Uzzell, T. and Corbin, K. W. 1971. Fitting discrete probability distributions to evolutionary events. *Science*, 172(3988): 1089–1096.

Wang, L. and Jiang, T. 1994. On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4): 337–348.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. 2009. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9): 1189–1191.

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. 2009. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1): 67–72.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T., *et al.* 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189): 872–876.

Wheeler, T. J., Clements, J., and Finn, R. D. 2014. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics*, 15(1): 1–9.

Whelan, S. and Goldman, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, 18(5): 691–699.

Wilson, D. J., Hernandez, R. D., Andolfatto, P., and Przeworski, M. 2011. A population genetics-phylogenetics approach to inferring natural selection in coding sequences. *PLoS Genet*, 7(12): e1002395.

Wuthrich, K. 2001. The way to NMR structures of proteins. *Nat Struct Mol Biol*, 8(11): 923–925.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10(6): 1396–1401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39(3): 306–314.

Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics*, 139(2): 993–1005.

Yang, Z. 2007. Paml 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8): 1586–1591.

Yang, Z., Goldman, N., and Friday, A. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular Biology and Evolution*, 11(2): 316–324.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.-M. K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1): 431–449.

Zhang, J. and Gu, X. 1998. Correlation between the substitution rate and rate variation among sites in protein evolution. *Genetics*, 149(3): 1615–1625.