Genome Evolution and Regulatory Network Structure in Bacteria

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Nacho Molina
aus Madird (Spain)

Basel 2010

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

- Eugene V. Koonin
- Mihaela Zavolan
- Erik van Nimwegen (Supervisor)

Basel, den 9-12-2008

Dekan:
Prof. E. Parlow

To Blanca (my almost wife).
To my Mother. To my Family
All of them my most important support.

...the present was almost intolerable in its richness and sharpness, as were his most distant and trivial memories... He knew by heart the forms of the southern clouds at dawn on the 30th of April, 1882, and could compare them in his memory with mottled streaks on a book in Spanish binding he had only seen once... Two or three times he had reconstructed a whole day; he never hesitated, but each reconstruction had required a whole day... He was, not forget, almost incapable of ideas of a general, Platonic sort. Not only was it difficult for him to comprehend that the generic symbol dog embraces so many unlike individuals of diverse size and form; it bothered him that the dog at three fourteen (seen form the side) should have the same name as the dog at three fifteen (seen from the front)... He was not very capable of thought. To think is to forget differences, generalize, make abstractions...

"Funes, the memorious", Jorge Luis Borges.

Contents

1 Introduction

Funes, in spite of his infallible memory, was not capable of thought since, as J.L. Borges writes, "to think is to forget differences, generalize, make abstractions." Due to the latest technological advances, biology seems to be entering in a Funes-like state: biologists can amass more experimental data about the organisms they study than ever before; and, store these "memories" in huge databases. A fundamental question rises: can the scientific community synthesize this information and turn it into powerful abstract theories? Is abstraction possible or even desirable in such a complex discipline as biology? From the point of view of a physicist I believe that a theoretical biology is both possible and desirable.

Several quantitative laws have recently come to light in biology, particularly in the evolution and regulatory architecture of genomes. This thesis explores the implications on genome evolution and regulatory network structure of one such law: the scaling of functional content of genomes with their size [1, 2]. This was the starting point of this thesis which hopefully represents a tiny little step towards a general theory of genome evolution and regulatory network structure in bacteria.

1.1 Genome evolution

Darwin's original work established the basis of the theory of evolution postulating that traits spread in populations by natural selection [3]. This fundamental understanding was partially changed by the discovery that DNA carries heritable genetic information leading to the began of the new era of molecular evolution. Comparing orthologous mammalian DNA sequences to the fossil record indicated that the rate of amino acid substitutions was roughly constant in time [4]. However, these substitutions fixed in populations too often to have been the result of selection [5]. The high rate of fixation led Kimura to formulate his neutral theory of molecular evolution [6]. Since then, neutral evolution became the null model of sequence evolution which permitted the rigorous reconstruction of phylogenies [7] and detection of selection on gene sequences [8, 9].

Today the sequences available have grown from a few genetic loci to hundreds of whole annotated genomes [1]. This wealth of data permits us to look beyond amino acid substitutions and study the variation in gene content and structure of genomes at a whole. In fact, several studies have shown that even closely related genomes with few substitutions often have enormous differences in gene content [10]. These results highlight that changes at higher level of organization have an essential role in the evolutionary process and therefore in life diversity. The main forces causing these changes, i.e. shaping the gene-content of genomes, are gene duplication, gene deletion and horizontal gene transfer leading to the acquisition of genes with new functions, subfunctionalizing existing functions, or deleting genes whose functions are no longer required.

Studies of gene content have uncovered several striking quantitative laws that are directly related to genome evolution. First of all, it was noticed [11, 12, 13] that a number of key genomic quantities show power-law distributions. In particular, the distribution of gene families is a power-law in each genome, whose exponent appears to depend mostly on the size of the genome. Several theoretical models have been put forth for explaining these power-law distributions which all include gene duplications, gene deletions and gene innovation as key ingredients [14, 15, 16]. Another striking observation [1] is that the numbers of genes in different functional categories scale as power-laws

[1] At the moment of writing there were 770 completed prokaryotic genomes and 1287 in-progress in the NCBI database

in the total number of genes in the genome. For example, whereas the numbers of genes involved in different types of metabolism scale approximately linear with genome size, the number of genes involved with regulatory processes such as transcription regulation and signal transduction scales almost quadratically with genome size, and the number of genes involved with basic processes such as DNA replication or cell division scales with an exponent less than 1. Such scaling laws are observed for the large majority of high-level functional categories. As argued before [1, 2], these scaling laws have important implications for the evolutionary dynamics of gene duplications and deletions.

This thesis focuses on how the functional content of genomes scales with genome size. We show that these scaling laws hold across bacterial clades, and formulate the simplest null model which accounts for these scaling laws. The scaling exponents emerge as universal constants of genome evolution. We test the model's predictions against the protein domain content of closely related genomes by estimating the number of domain additions and deletions in each pair of genomes since they diverged from their last common ancestor. The available data support nearly all of the model's predictions. Finally, we discuss the implications of our work on the role of horizontal gene transfer in genome evolution.

1.2 Regulatory networks

We can view a bacterial cell as an entity made up of many molecular components that is capable of sensing many internal and external physico-chemical signals, and executing specific cellular programs in response. The realization of each program produces certain concentrations of specific proteins that act in some fashion beneficial to the cell. Thus, to understand the cell's dynamics, we must know how the protein concentrations change in response to the environment.

Transcription of genes into mRNA molecules is one of the most important stages of protein biosynthesis. Transcription is regulated by specific proteins which are collectively called transcription factors. In response to stimuli, transcription factors bind specifically to DNA by recognizing short DNA sequences upstream of genes. Upon binding, they activate or repress transcription of genes into mRNA, i.e. transcription factor activate or repress gene expression. The set of all interactions between transcription factors and their regulated target genes form the so-called transcriptional regulatory network. Therefore, understanding this network is essential to understand the cell's response to its environment.

The topological features of the transcriptional regulatory networks of E. coli and S. cerevisiae have been intensely studied and some of their global and local properties have been uncovered in recent years. For instance, some studies have shown that the distribution of the number of genes that are regulated by a particular transcription factor (or out-degree) follows a power law, while the number of transcription factors regulating a particular gene (or in-degree) follows an exponential distribution [17].

Globally, these network are organized into subnetworks which show a hierarchical internal structure with very few feedback interactions except for self-regulation. Interestingly, it has experimentally been demonstrated that these subnetworks process specific environmental signals [18]. Locally, certain motifs formed by few nodes appear more often than in random networks with the same degree distributions [19]. The information-processing properties of these motifs has been studied individually [20, 21, 22] as well as how they aggregate to form higher structures [23]. However, it is not clear whether these motifs have been positively selected by evolution due to their particular functions, or they are a side effect of the evolution of the regulatory network [24, 25]. Some of these results are still controversial and it is important to recall that they were obtained on incomplete networks. They may not hold once the full networks are known [26].

All the results above come from a small number of model organisms. Therefore, little is known about how the global structure of transcription regulatory networks varies across bacteria. Strikingly, the number of transcription factors grows roughly quadratically with the size of the genome

[27, 1]. For example, according to the DBD database [28], the number of transcription factors per genome in bacteria varies from only 3 (of a total of 504 genes) in Buchnera aphidicola, to 801 (of a total of 7717 genes) in Burkholderia sp. 383. To put the latter number in perspective, the vastly bigger genomes of *C. elegans* and *D. melanogaster* have a lower estimated total number of transcription factors according to the same database. The enormous range in the number of transcription factors across bacteria reflects a corresponding range in complexity of gene regulation. For example, Buchnera lives in a very stable environment as an endosymbiont of aphids, and shows little transcriptional regulation [29]. In contrast, Burkholderia can live under extremely diverse ecological conditions including soil, water, as a plant pathogen, and as a human pathogen, which most likely require complex regulatory mechanisms.

This scaling property of the number of transcription factors has important implications for the structure of transcription regulatory networks. The total number of interactions between transcription factors and regulated genes is given by the number of transcription factors r times the average number of interactions per transcription factor $\langle o \rangle$, but also by the total number of genes g times the average number of transcription factor that regulate a gene $\langle i \rangle$, we have: $r \langle o \rangle = g \langle i \rangle$. Since the number of transcription factors *per gene* grows linearly with the total number of genes we cannot have that both the average number of interactions per transcription factor and the average number transcription factors that regulate a gene are the same in bacteria of different genome size. In particular, we must have $\langle i \rangle / \langle o \rangle \propto g$. That is, either genes are regulated by more transcription factors in larger genomes or the regulon size decreases with genome size. Which of these scenarios is the one that occurs in nature? This thesis addresses this question.

However, answering this question directly requires knowing a large number of transcriptional regulatory networks, but very few such networks are available. Instead, we use an indirect procedure based on the assumption that regulatory sites on the genome evolved under purifying selected. We develop a novel method to measure purifying selection in intergenic regions. Our procedure starts from a set of related bacterial genomes (a *clade*) as provided by the NCBI microbial genome database [30], of which one is denoted as the *reference species*. For each gene and each intergenic region of the reference species we extract orthologous genes and intergenic regions from the other species and produce multiple alignments. We determine cliques of orthologous proteins (sets of genes that are all mutual orthologs between all species in the clade) and infer the topology of the phylogenetic tree from the concatenated alignment of all cliques. Then, we evaluate the amount of selection for each alignment column by the likelihood ratio of two evolutionary models: the background model that assumes a simple F81 substitution rate model [7] which is parameterized by an overall mutation rate and a vector of equilibrium base frequencies. And, the foreground model that assumes the same substation rate model but with a unknown specific set of base frequencies that account for the selection action on that site that are integrate out of the likelihood. Some of these techniques were integrated into MotEvo, a novel tool for detecting binding sites in intergenic alignments given known weight matrices.

We applied our method to 22 different bacterial clades which span widely the whole phylogenetic tree. We identified segments in the intergenic regions of the analyzed bacteria that show evidence of purifying selection. To evaluate the performance of our method for detecting real binding sites we studied the overlap between the identified segments and experimental verified binding sites of *E. coli*. The results show that we are available to detect real binding sites based on conservation. We obtained purifying selection profiles respect to gene start and stop sites revealing universal patterns across species. One of the most remarkable pattern is the selection that takes place around the start codon which is shown to be connected to translational efficiency. We observed, almost in all clades, a relatively higher frequency of adenine around the start codon which we showed is related to the avoidance of RNA secondary structure in that region.

Coming back to our starting question: how the number of binding sites scales with genome size? To answer this, we studied the amount of purifying selection from intergenic regions across the 22 bacterial clades. Strikingly, the amount of purifying selection in intergenic regions does not vary

with genome size. Moreover, the most conserved DNA words in intergenic regions showed higher diversity in large genomes than in small ones. These results strongly indicate that the structure of transcription regulatory networks changes dramatically with genome size: small genomes have few transcription factors each binding to many sites, while large genomes have many transcription factors each binding to a few sites. In other words, gene regulatory complexity is limited across bacteria while transcription factors become specialized in large genomes.

1.3 Outline of the thesis

The content of the thesis is organized as follow: in chapter 2 we show that measuring protein domains using Pfam annotations reproduces the known scaling laws in the functional contents of the genomes. Then, we check whether the scaling laws established for all genomes hold within clades. This is an essential question since universal and clade-independent scaling laws indicate that fundamental constraints, which are independent of bacterial lifestyle, shape genome functional organization. We focus on the scaling laws of transcription factors due to its singular relevance in regulatory networks, and we study how the exponents of the scaling laws vary for different annotation procedures and bacterial clades.

In chapter 3 we present the simplest evolutionary model that can account for the observed scaling laws. We show that a time-invariance hypothesis, i.e. assuming that the scaling laws held at any time in evolutionary history, uniquely determines the relative rates of addition and deletion of protein domains. In particular, our model predicts that the relative rates of addition and deletion of domains in a given functional category is proportional to the current number of domains in the category multiply by a category-dependent constant which is the *same* for all evolutionary lineages. These category-dependent constants, that we called *evolutionary potentials*, represent the relative probabilities of an addition or deletion of a domain in a functional category to be fixed in the population. Our model, also, predicts that these constants equal the exponents of the scaling laws. These results established a direct quantitative connection between the scaling laws in the functional content of genomes and the rate of duplications and deletions during short evolutionary time intervals. We analyze the domain content of several pairs of closely-related genomes from all over the bacterial phylogenetic tree demonstrating that the predictions are supported by available genome-sequence data. Finally, we discuss the implications that our results have on horizontal gene transfer.

Next, we turn to the structure of transcriptional regulatory network, and the topological constraints that our scaling laws imply. In particular, we investigate how the average number of transcription factors regulating each gene and the average number of genes regulated by each transcription factor scale with genome size. Very few regulatory networks are known, so we rely on an indirect measurement: the amount of selection that take place in intergenic regions. In chapter 4 we present an integrated set of algorithms to detect purifying selection across sites. Our methodology includes new algorithms for mapping of orthologs, inferring phylogenetic trees, and aligning orthologous intergenic regions. We describe in detail the underlying evolutionary model used to measure selection and identify conserved segments in intergenic regions.

In chapter5 we apply these algorithms to a comprehensive set of bacterial genomes. We find several patterns of purifying selection shared by all bacteria, and show that some of these patterns are directly related to translation efficiently and the avoidance of RNA secondary structure.

Finally, in chapter 6 we investigate how the average number of regulatory sites per intergenic region and the average number of sites regulated by a particular transcription factor vary with genome size. We measure how the average length of intergenic regions, the number of operons and the degree of selection scale with genome size. We study the clustering of transcription factors across all genomes and the diversity of the most and least conserved DNA words across clades. We conclude that the structure of transcriptional regulatory networks changes dramatically with

genome size. Small genomes have few transcription factors, each binding to a large number of sites. Large genomes have more transcription factors, each binding to fewer sites.

2 Scaling laws in the functional content of genomes

It has been established that, for many high-level functional categories, the number of genes in the category scales as a power-law in the total number of genes in the genome. With the large number of bacterial genomes now available it has become possible to compare these scaling laws across individual clades of bacteria. Recently it has been reported that, for the category of transcription regulators, there are substantial differences in the scaling across clades. Here we present an comprehensive analysis of the scaling in functional gene content across different clades for a large number of functional categories. Strikingly, we find that for almost all functional categories, including transcription regulators, the available data suggest that all bacterial clades follow a common universal scaling law. This result strongly suggests that these universal scaling laws reflect fundamental physical and biological design principles of bacterial genomes that are independent of life-style and lineage. A small number of categories, including amino acid metabolism and oxidoreductase activity, suggesting the clade-specific functional organization affects mostly amino acid metabolism and energy pathways.

2.1 Introduction

A few years ago, we studied the gene content of the fully-sequenced genomes that were then available and found that, for many high-level functional categories, the number of genes n_c in each category c scales as a power-law in the total number of genes n in the genome, i.e.

$$n_c = e^{\beta_c} n^{\alpha_c} \tag{2.1}$$

with the exponent α_c the constant β_c depending on the functional category c [31]. At the time the number of available genomes precluded studying the gene-content scaling for individual bacterial clades but with currently more than 600 bacterial genomes available such analysis is now possible. Indeed, in a recent work [32], Cordero and Hogeweg studied the scaling in the number of transcription factors with genome size across different bacterial clades and found significant variation in the scaling exponents between clades, including exponents as low as 1 (i.e. linear scaling). Here we infer scaling laws for a large number of high-level functional categories separately for 24 different bacterial clades. Strikingly, our results show that, for most categories, there is no significant variation in the offsets and exponents of the scaling laws across bacterial clades. That is, for almost all functional categories that we study, all bacterial clades obey the *same* scaling laws.

2.2 Reproducing the scaling laws at the protein domain level

Although genes are natural units in genome analysis there are some disadvantages to using genes as the central units in the analysis of the evolution of genome content. For example, apart from being able to mutate, duplicate, and be deleted, it is well-known that, not unfrequently, two genes can fuse into one, single genes can split into two [33], and genes can evolve *de novo* from non-coding sequence. Such events significantly complicate the analysis of the evolution of gene content.

Protein domains form more natural units for the study of the evolution of gene-content for several reasons. It can be argued that protein domains act like 'evolutionary atoms' to a certain extent

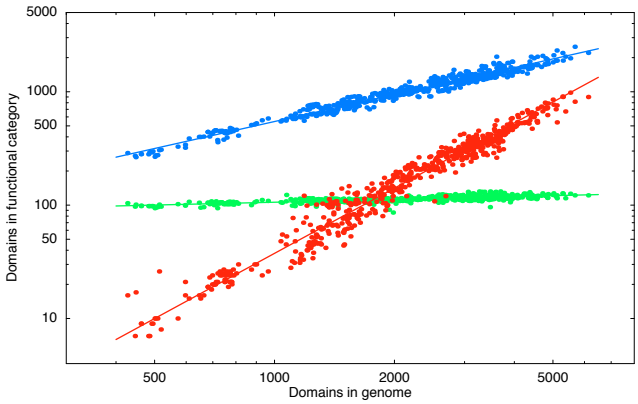

Figure 2.1: The number of protein-domains associated with functional categories 'translation' (green), 'metabolic process' (blue), and 'regulation of transcription' (red) as a function of the total number of domains in the genome for which a functional annotation is available. Each dot corresponds to a fully-sequenced microbial genome, with the total number of domains on the horizontal axis and the number of domains in a particular functional category on the vertical axis. Both axes are shown on a logarithmic scale. The straight lines show power-law fits.

[34]; Protein domains form functional units [35] that cannot be split into smaller units, and a single protein domain can, in general, not be constructed by fusing multiple occurrences of other protein domains. Therefore, we can safely assume that almost all changes in the number of occurrences in the genome of a given protein domain are due to deletions, duplications, or the horizontal transfer of a domain from another organism's genome. We thus decided to study the evolution of functional gene content in terms of the number of occurrences of different protein domains. Among databases of protein domains Pfam [36] is attractive because the Pfam domain families are disjoint, i.e. at the default settings it is guaranteed that any given DNA sequence segment will be classified to belong to at most one domain family. We thus used Pfam domains as our evolutionary 'atoms'.

We functionally annotated 630 bacterial genomes available (at the time of the study) at the NCBI database [30]. To do that, first we ran HMMer [37] using all Pfam models. A hit was considered a valid domain if its score was equal or bigger than the so-called *gathering score* of the model provided by the Pfam web site, and it did not overlap with any other hit of lower E-value. To count the number of domain occurrences per functional category we used a mapping from Pfam domains to Gene Ontology terms [38] which is available at http://www.geneontology.org/. If a domain-family f maps to category c it will be associated with c and all parent categories of c in the Gene Ontology hierarchy.

We counted the number of occurrences of each Pfam domain in each fully sequenced bacterial genome. Using a mapping from Pfam to Gene Ontology categories [38] we determined, for each genome g, the total number of domains $n(g)$ that can be associated with *any* GO category and, for each GO category c, the number of domains $n_c(g)$ occurring in the genome.

Figure 2.1 shows, for 3 example categories, the number of domains in that category as a function of the total number of domains in the genome (that can be mapped to a GO category).

As the figure shows, for all three categories the number of genes in the category n_c scales as a power-law in the total number of domains in the genome n, i.e.

$$n_c = e^{\beta_c} n^{\alpha_c}, \tag{2.2}$$

with both the prefactors β_c and the exponents α_c varying between categories. These power-laws are observed for the large majority of high-level functional categories. For each GO category we fitted a power-law of the form (2.2) using a Bayesian procedure which in particular provides a posterior probability distribution for the exponent α_c (see appendix). We selected 156 GO

categories that occur in at least 95% of all genomes and that show good power-law fits. The inferred exponents match what we found previously based on the gene-number analysis of a much smaller number of genomes [1, 2], i.e. for basic processes such as translation and DNA repair exponents are low, whereas exponents for regulatory functions such a regulation of transcription and signal transduction are largest. The inferred exponents for all 156 selected categories are listed in the appendix.

2.3 Same scaling laws across all bacterial lineages

To group the bacterial genomes into clades we used the taxonomy provided for each genome by NCBI. To select categories that can be meaningfully fitted we collected, for each clade, all categories c for which the domain-count $n_c(g)$ varies by a factor of at least 2 across the genomes in the clade, and fitted a power-law using a Bayesian model (see appendix). We denote by $\alpha_{i,c}$ and $\beta_{i,c}$ the fitted exponent and offset for category c in clade i. We denote by $\overline{\alpha}_c$ and $\overline{\beta}_c$ the exponent and offset obtained from fitting all genomes. To measure how the clade-specific exponents $\alpha_{i,c}$ deviate from the overall exponent $\overline{\alpha}_c$ we introduce the following Z-scores:

$$Z_{i,c} = \frac{(\alpha_{i,c} - \overline{\alpha}_c)}{\sqrt{\sigma_{i,c}^2 + \overline{\sigma}_c^2}} \tag{2.3}$$

where the $\sigma_{i,c}$ and $\overline{\sigma}_c$ are the error-bars on the clade-dependent and overall exponent, respectively, which were obtained form the 99% posterior probability intervals on $\alpha_{i,c}$ (see appendix). We calculated analogous Z-scores for the deviations of the clade-specific offsets $\beta_{i,c}$ from the overall offset $\overline{\beta}_c$.

To quantify the overall amount of variation in fitted exponent for each category we averaged the clade-dependent scores $Z_{i,c}$ to obtain an overall Z-score for each category:

$$Z_c = \sqrt{\frac{1}{N_c}\sum_i Z_{i,c}^2}, \tag{2.4}$$

where N_c is the number of clades (24). We calculated analogous Z-scores for the variation in fitted offsets.

In figure 2.4a (top) we show the scores Z_c for the variation in fitted exponents across functional categories. In the other three top panels we show the fitted exponents for selected functional categories that have a high, medium and low Z-score. The selected categories are indicated in colored font in Fig. 2.5a (top) and the corresponding overall exponents $\overline{\alpha}_c$ are shown as dashed lines with corresponding colors in the other panels. The results show that, for the large majority of categories including important categories such as transcription factor activity, translation, transport, and metabolic process, the fits in all clades are consistent with a single universal power-law. Moreover, even for the cases with the highest Z-scores, such as 'amino acid metabolic process' shown in the figure 2.5c (top), the variation of the exponents across clades is very moderate, with most clades still consistent with a single common exponent.

The four panels in the bottom of the Figure 2.5 show analogous results for the offsets $\beta_{i,c}$. The distribution of Z-scores again shows that for the majority of categories the data are consistent with a single underlying offset across all clades. Also, for important categories such a transcription factor activity, translation, transport, and metabolic process the variation is not larger than would be expected by chance. Finally, even for functional categories with the largest Z-score the variation of fitted offset is limited.

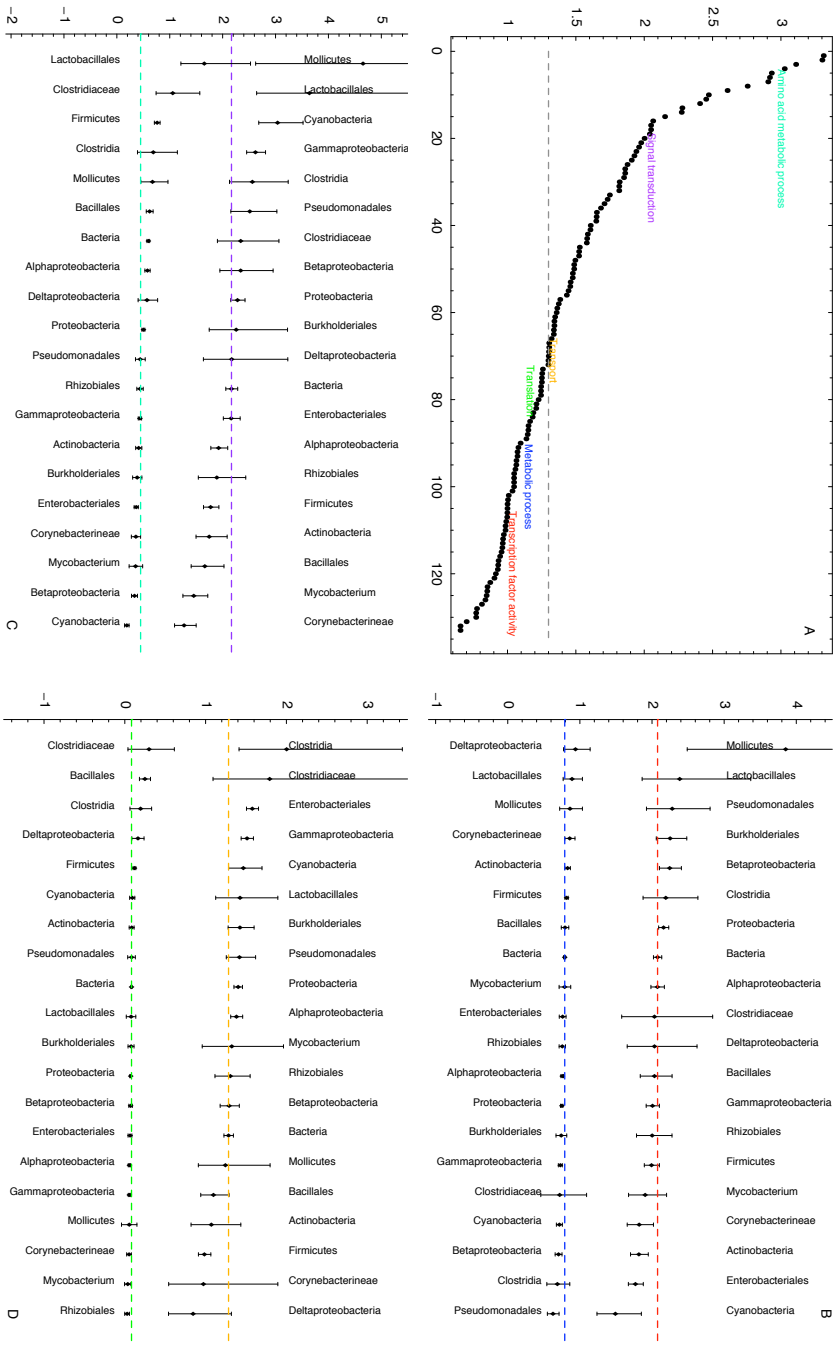

Figure 2.2: a) Z-score of the selected functional categories. Exponents for different lineages and the their 99% posterior intervals of, b) transcription factor activity and metabolic process, c) signal transduction and amino acid metabolic process. d) transport and translation.

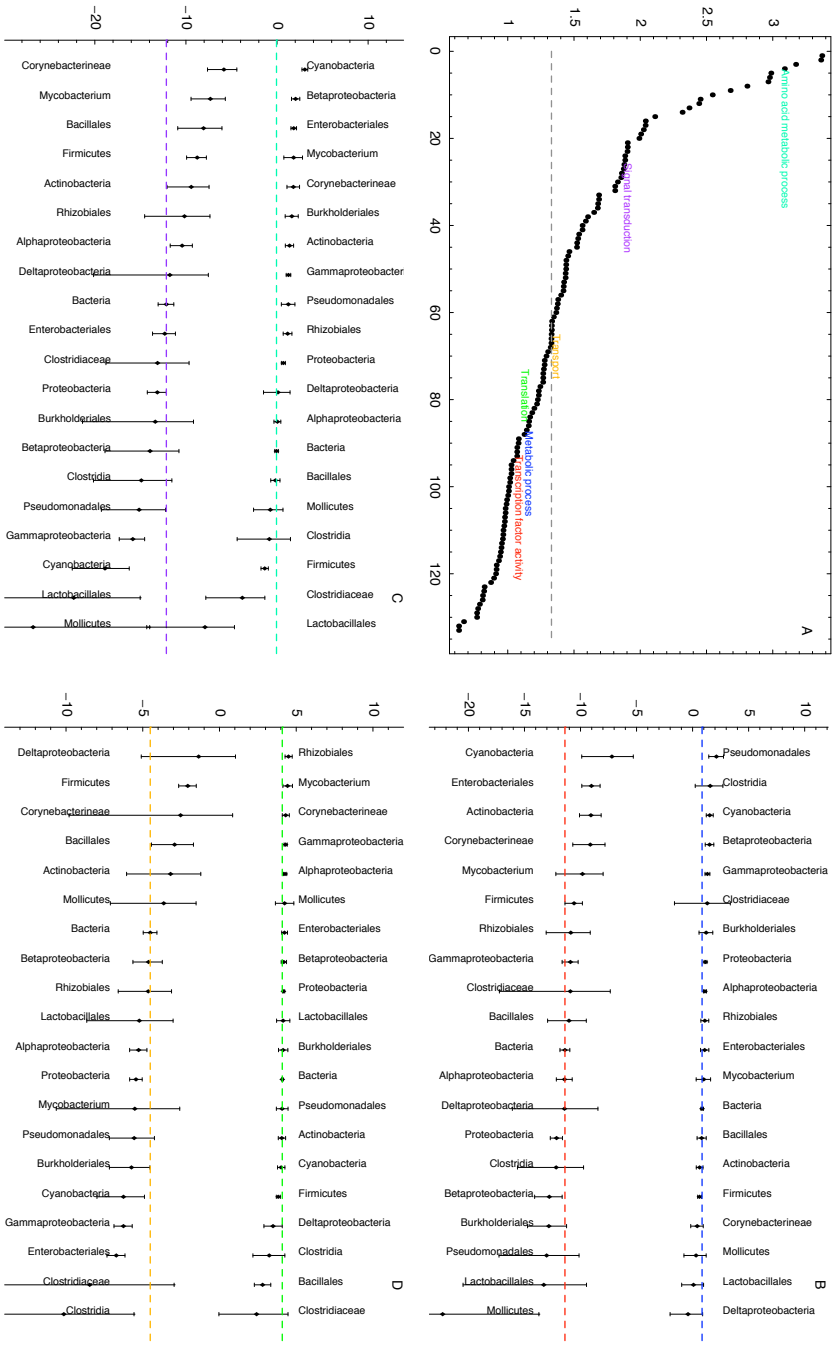

Figure 2.3: a) Z-score of the selected functional categories. Offsets for different lineages and the their 99% posterior intervals of, b) transcription factor activity and metabolic process, c) signal transduction and amino acid metabolic process. d) transport and translation.

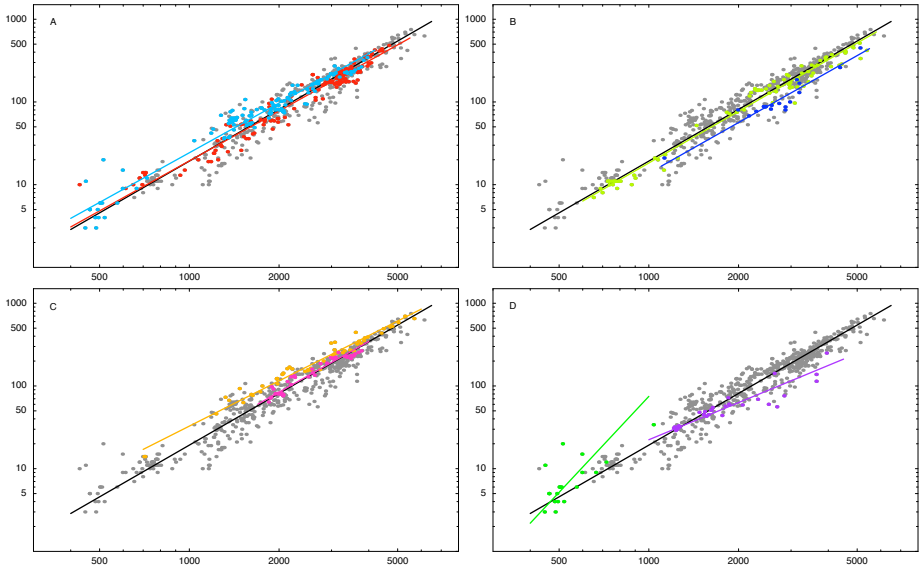

Figure 2.4: Number of domains in the functional category transcription factor activity as a function of the the total number of domains in the genome. The gray dots show all bacteria and the colored dots correspond to different clades. Light green: mollicutes. Purple: cyanobacteria. Light blue: firmicutes. Red: γ-proteobacteria. Orange: actinobacteria. Pink: Bacillales. Green: α-proteobacteria. Blue: δ-proteobacteria.

Discrepancy with the results of Cordero and Hogeweg

Our results for transcription factors contradict findings by Cordero and Hogeweg [32] (CH from now on) which found significant differences in the scaling of the number of transcription factors across clades. To illustrate our findings, Fig. 2.4 shows the scaling of the number of domains that map to the category 'transcription factor activity' against the total number of domains in the genome for all genomes (grey dots), as well as for 8 different clades of bacteria (colored dots). The figure clearly illustrates that essentially the same scaling in the number of transcription factors is found in all clades, including γ-proteobacteria, α-proteobacteria, δ-proteobacteria, firmicutes, and bacillales. Note that the clades whose exponents deviate most from the common one (Mollicutes and Cyanobacteria, Fig. 2.4d) correspond to scatters that are very noisy and that have a relatively small range in the total number of domains across genomes.

There are a number of possible explanations for the discrepancy of our results with those of CH. First, we perform our analysis at the level of domains whereas CH's analysis is at the level of proteins. To check the effect of this difference, we recalculated the scaling laws across different clades, as well as the Z-scores, using protein rather than domain counts and the results are essentially unchanged (see section 2.4). Using domains versus proteins is clearly not the main source of the discrepancy.

Second, we have used the number of domains that map to at least one GO term, i.e domains that have known function, as the quantity on the horizontal axis of the scatter (as opposed to the total number of domains). As shown in section 2.6, the number of domains with functional annotation n_{annot} scales as a power-law in the total number of domains n with exponent about 0.94, i.e. $n_{\text{annot}} \propto n^0.94$. Note that this implies that the quality of annotation cannot be uniform across all genomes. That is, the fraction of unannotated domains is somewhat larger in large genomes. This effect occurs as well at the level of proteins, i.e. the fraction of unannotated proteins grows with genome size. This clearly affects the fitted exponents. That is, if we fit a slope of $\alpha_c = 2$

for the number of transcription factors as a function of $n_{\rm annot}$, then this corresponds to a slope $2/0.94 = 1.88$ in terms of the total number of domains n.

Our rational for fitting the scaling laws in terms of $n_{\rm annot}$ instead of in terms of n is that, if the fraction of annotated domains decreases with genome size, then we expect this to also apply to the fraction of annotated domains in a functional category c. That is, if n_c is the true number of domains of category c, and $n_{c,annot}$ is the number that are captured by the annotation, we expect that these also obey a relation $n_{c,annot} \propto n_c^{\gamma_c}$, with some exponent γ_c. Note that if the true number of domains n_c in category c scales as n^{α_c} then we find

$$n_{c,annot} \propto n^{\gamma_c \alpha_c} \propto (n_{\rm annot})^{\alpha_c \gamma_c/\gamma}. \tag{2.5}$$

Since the total number of annotated proteins obeys this law with exponent $\gamma = 0.94$, we in general expect $\gamma_c < 1$ for more specific categories as well. In particular, one cannot have $\gamma_c = 1$ for all categories, because this would imply $\gamma = 1$ as well. Therefore, γ_c must be less than one for many categories, and fitting $n_{c,annot}$ in terms of n would lead to consistent underestimation of the exponents α_c for all those categories.

In our opinion the simplest assumption is to assume that all γ_c are equal, i.e. $\gamma_c = \gamma$ for all categories. As equation (2.5) demonstrates, under this assumption the correct exponents are inferred when fitting in terms of $n_{\rm annot}$. One source of discrepancy between the results of CH and ours is that CH fits results in terms of the total number of proteins, not the number of proteins with annotation, leading to systematically lower exponents.

Another source of discrepancy is the fitting procedure itself. We use a Bayesian procedure which essentially finds the first principal component whereas CH use standard linear regression. Note that our Bayesian procedure is symmetric with respect to the axes. That is, if we fit a slope α for y as a function of x, we fit a slope $1/\alpha$ for x as a function of y. Since standard regression assumes that all deviations from the power-law are only in the vertical direction it does not obey this symmetry and will typically infer exponents closer to $\alpha = 1$. In particular, standard regression will fit lower slopes for categories that scale superlinearly, especially when the data is noisy.

Finally, the discrepancy could result from the functional annotation procedure: we use Pfam domains and gene ontology whereas CH use COGs. To investigate this effect we analyzed the scaling of the number of transcription regulators (according to COG) as a function of the total number of proteins that map to at least 1 COG (see section 2.5). Somewhat surprisingly, at least qualitatively the results are very similar to those we obtained based on Pfam and GO annotation. The Z-statistic ($Z = 1.08$) indicates that almost all clades are consistent with a universal scaling law. Moreover, the exponent most significantly less than 2 is 1.71 ± 0.20 (Actinobacteria). In contrast, CH report an exponent 1.34 ± 0.11 for the clade Actinobacteria.

We decided to track in detail the discrepancy for the clade Actinobacteria. Using our Pfam annotation, and using Bayesian fitting in terms of the number of annotated domains we find a slope of 1.73 ± 0.13 for Actinobacteria, which compares with 1.34 ± 0.11 reported by CH. First, there are currently significantly more genomes available than at the time of CH's study. With the current set of genomes, applying CH's procedure (using COG annotation, fitting using standard regression as a function of the total number of proteins), we find a exponent of 1.54. That is, with the larger number of genomes available the slope has already increased significantly. If we use Bayesian fitting instead of standard regression we find a slope of 1.61 ± 0.17. If we fit in terms of the number of proteins that have a COG annotation we recover our result 1.71 ± 0.16. This is almost indistinguishable from the result obtained with Pfams. That is, we find that the low slope estimated by CH is a result of a combination of: fewer genomes, using standard regression, and fitting in terms of the total number of proteins as opposed to the number of annotated proteins.

In section 2.7 we compare in detail the estimated slopes and quality of the fits that are obtained for the category 'transcription regulation' when using Pfam or COG annotation and using the total number of proteins/domains or the number of annotated proteins/domains. The results show that the highest quality fits and lowest Z-statistic (variance of fitted exponents across clades) are

obtained using Pfam annotation and fitting in terms of the number of annotated proteins, followed by Pfam annotation fitting in terms of all domains, then COGs fitting in terms of number of COG-annotated proteins, and finally COG fitting in terms of the total number of proteins. That is, using Pfam annotation and fitting in terms of the number of annotated domains both increases the quality of the fits, and decreases the variance in fitted exponents. In addition, we find that using COG annotation there is a significant correlation between the quality of the fit and the fitted exponent, i.e. the low exponents tend to correspond to clades who have poor fits. This correlation is absent when using Pfam annotation. Finally, we find that there is no significant correlation between the exponents fitted using Pfams and using COGs. That is, those clades for which exponents come out small according to COG tend not to be the same clades for which exponents come out small according to Pfam.

Together these results strongly suggest that more reliable fitting is obtained when using Pfam annotation and fitting in terms of the number of annotated domains, and that the significant variation of exponents across clades that CH find is an artifact of the annotation procedures used by CH.

Categories with non-universal scaling laws

In figure 2.5, for both the exponents and the offsets, there are a little under 20 categories that have a Z-statistics larger than 2 which are separated from the other > 110 categories by a little gap. These GO categories show the most evidence of variation in their scaling laws across clades. Interestingly, we find that the high-variance categories are essentially the *same* for both exponents and offsets, i.e. those categories with significantly varying exponents also have significantly varying offsets. Manual inspection shows that these 17 categories mainly consist of five groups of related categories around the categories: 'amino acid metabolic process' ($Z = 3.03$), 'vitamin binding' ($Z = 2.84$), 'oxidoreductase activity' ($Z = 2.75$), 'lyase activity' ($Z = 2.37$), and 'GTP binding' ($Z = 2.34$). For all these categories we find that some clades show high exponent α_c and low offset β_c, whereas others show low exponent α_c and high offset β_c. Interestingly, the clade cyanobacteria is always at one of the extremes. Cyanobacteria show a high exponent in GTP binding and a low exponent in all other 4 categories. In contrast, the clades firmicutes and lactobacillales show high exponents in 'amino acid metabolic process', 'oxidoreductase', and 'vitamin binding'. The category 'lyase activity' is interesting in that it separates the sister clades Bacillales (low exponent) and Lactobacillales (high exponent).

Although it is hard to extract a single essential feature of these 5 categories it is clear that broad themes are amino acid metabolism, enzymes that need to bind cofactors, and energy pathways. It is tempting to suggest that these broad themes define the different 'life styles' of the bacteria in the different clades.

2.4 Scaling laws across different bacterial clades using number of proteins

We have recalculated the exponents and the offsets, as well as, the Z-scores of the scaling laws across different bacterial clades at the level of proteins. To do that each protein is mapped to a GO terms (and all its parents in the GO hierarchy) if it contains a Pfam domain that maps to that GO term. Then, we fit a power-law in each clade independently and we compute Z-scores for the exponents and the offsets as we did for the case of the scaling laws at the level of domains. In figure 2.5a we show the Z-scores of all functional categories. In figure 2.5b, 2.5c and 2.5d we show the clade-dependent exponents for some relevant functional categories. In figure 2.6 we show similar results for the fitted offsets. As it can be seen the results are consistent with the ones we obtain performing the analysis at the level of domains.

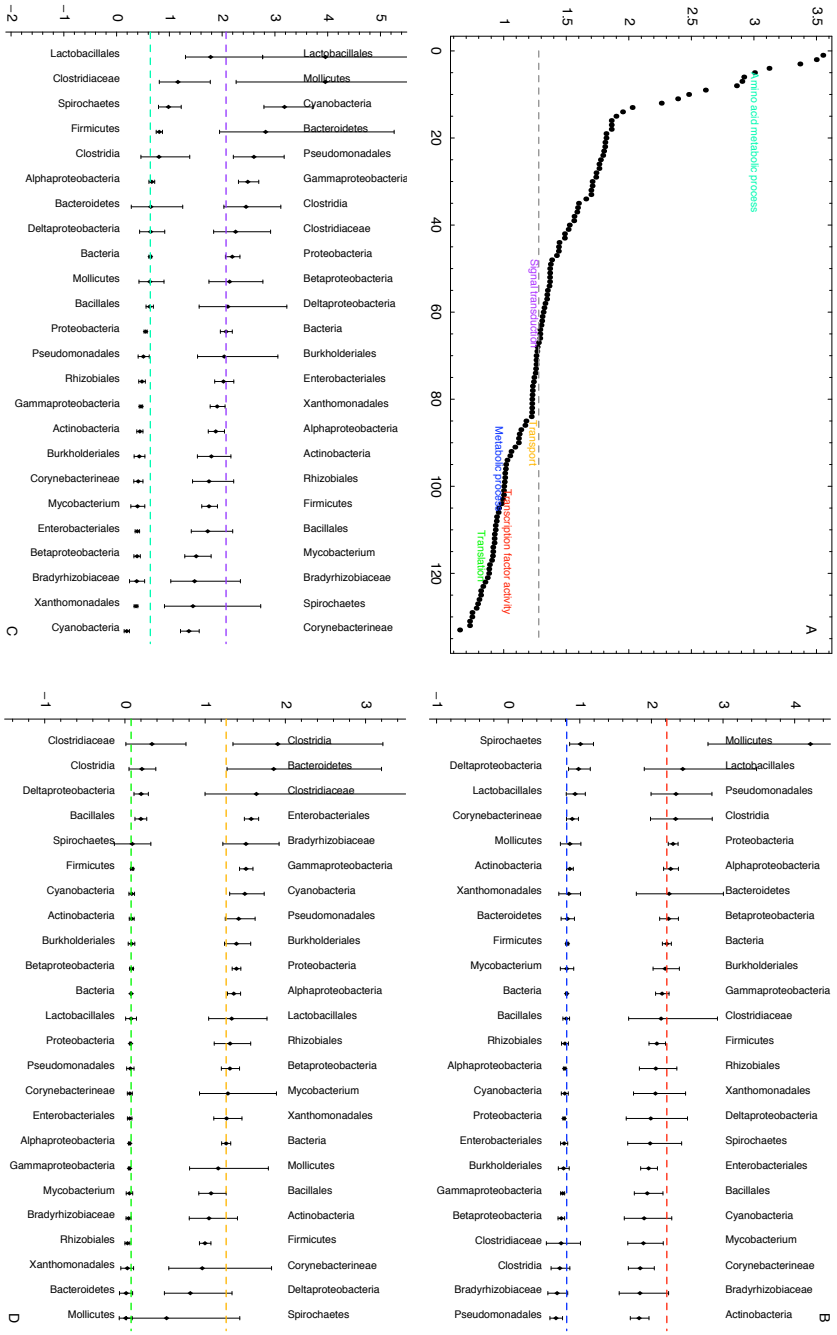

Figure 2.5: a) Z-score for the variation of the exponents across functional categories. Clade-dependent exponents and the their 99% posterior intervals for the categories b) transcription factor activity and metabolic process, c) signal transduction and amino acid metabolic process, and d) transport and translation.

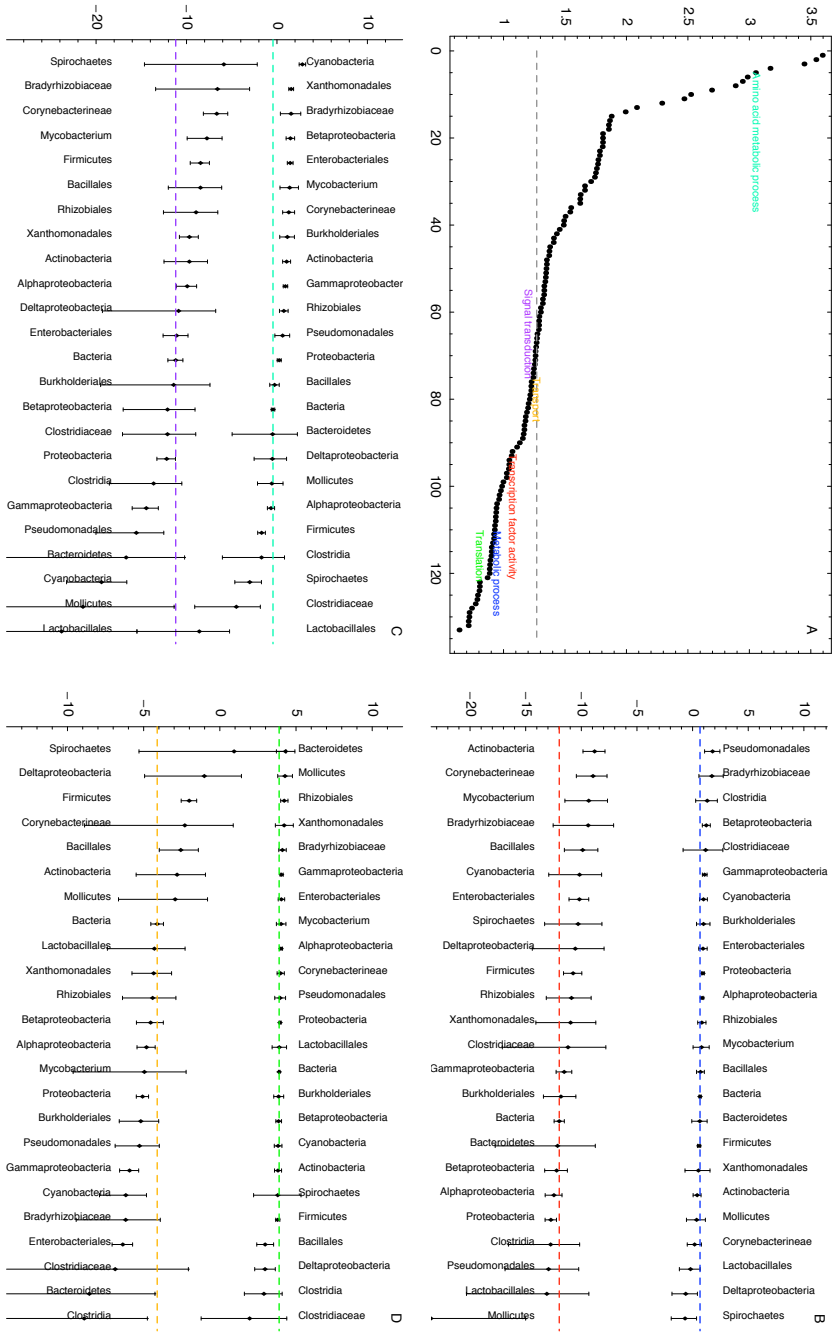

Figure 2.6: a) Z-scores for the variation of the offsets across functional categories. Clade-dependent offsets and the their 99% posterior intervals for the categories b) transcription factor activity and metabolic process, c) signal transduction and amino acid metabolic process, and d) transport and translation.

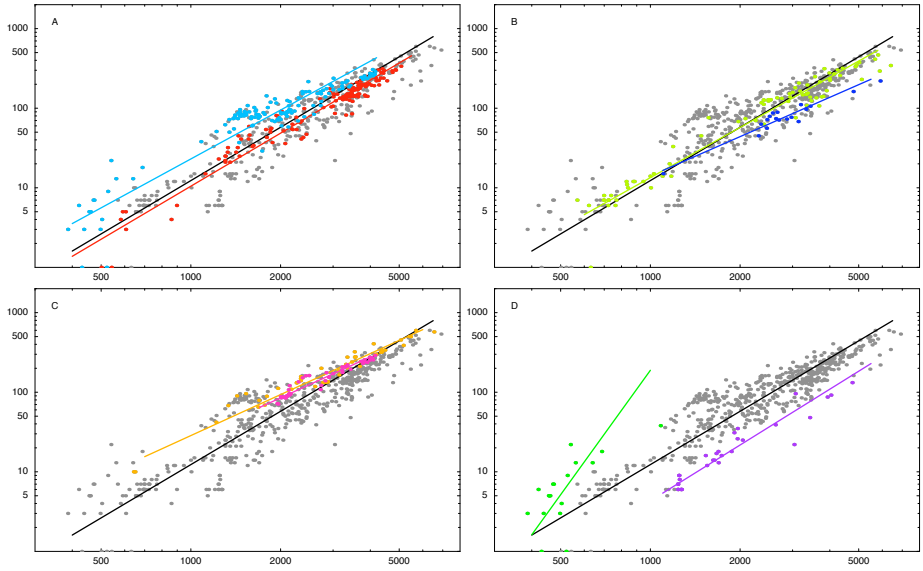

Figure 2.7: Number of proteins in the COG functional category 'transcription regulators' as a function of the the total number of domains in the genome. The gray dots show all bacteria and the colored dots correspond to different clades. Light green: mollicutes. Purple: cyanobacteria. Light blue: firmicutes. Red: γ-proteobacteria. Orange: actinobacteria. Pink: Bacillales. Green: α-proteobacteria. Blue: δ-proteobacteria.

2.5 Scaling law of transcription regulators using COGs

We use the COG annotation of each bacterial genome that is available from the NCBI ftp site. To determine the number of transcription factors we count, for each genome, how many proteins belong to any of the COGs that are functionally classify as transcription regulators'. This functional category is part of the more general category 'transcription' (letter code: K). Then, for each clade, we used a Bayesian model to fit a power-law of the form $n_R = e^{\beta_c} n^{\alpha_c}$ where n_R is the number of regulators and n the number of proteins that belong to at least one COG.In figure 2.8 we show the clade-dependent exponents (left) and the offsets (right) for 23 different clades and the overall bacterial exponent and offset. As it can be seen, the variation of the exponents across clades is very moderate, with most clades still consistent with a single common exponent and none of them below 1.55. In figure 2.7 we plot the number of transcription regulators against the total number of proteins that belong to at least one COG. In gray we show all bacteria and in colored dots 8 different clades. Even though the scatters are more noisy compared with the ones we obtain with Pfam domains we still see that there is a general trend which is obey by the different clades.

2.6 Functional annotation coverage

To calculate the scaling laws we have used the number of Pfam domains that map to, at least, one GO term. Here we want to study if the functional annotation coverage depends on the clade, i.e. if the amount of domains that are functionally annotated in a genome is different depending on which clade the genome belong to. In the left panel of the figure 2.9 we show in a log log plot how the number of domains with a known function scale with the total number of domains. The exponents is almost one (0.94 ± 0.01) and as it can see in the right panel almost all clades have the same exponent. Interestingly, the clade cyanobacteria, the only outlier is the one that shows

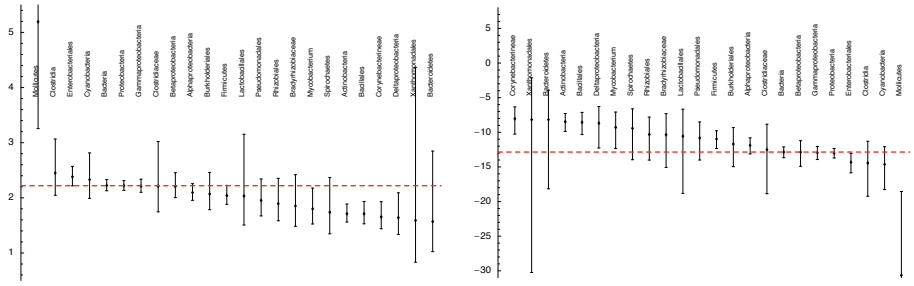

Figure 2.8: Clade-depend exponents (left) and offsets (right) with their 99% posterior probability interval of the COG functional category 'transcription regulator'

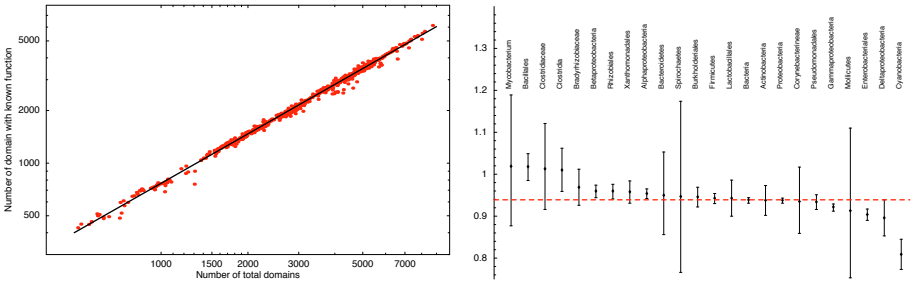

Figure 2.9: *Left*: number of domains that map to, at least, one GO term against total number of domains. *Right*: fitted exponents and their 99% posterior probability interval for different across different bacterial clades.

the most different exponent in the scaling law for the transcription factors.

2.7 Scaling of transcription regulators using different annotation procedures

We investigated the variation in fitted scaling laws for the category 'transcription regulation' using 4 different annotation procedures

1. Pfamfunc: Using Pfam domains annotations and fitting the number of transcription regulation domains in terms of the total number of domains that are annotated (to at least one category in the GO hierarchy).

2. Pfamall: Using Pfam annotations and fitting in terms of the total number of domains (including those without GO annotation).

3. COGfunc: Using COGs and fitting in terms of the total number of proteins that are mapped to at least one COG.

4. COGall: Using COGs and fitting in terms of the total number of proteins (including those not in COGs).

First we determined the distribution of the fitted scaling exponents α across the 24 different clades, using the 4 different annotation procedure. For each annotation procedure we determined the fitted exponent α_i and error-bar σ_i for each clade i and we approximated the total distribution of α as

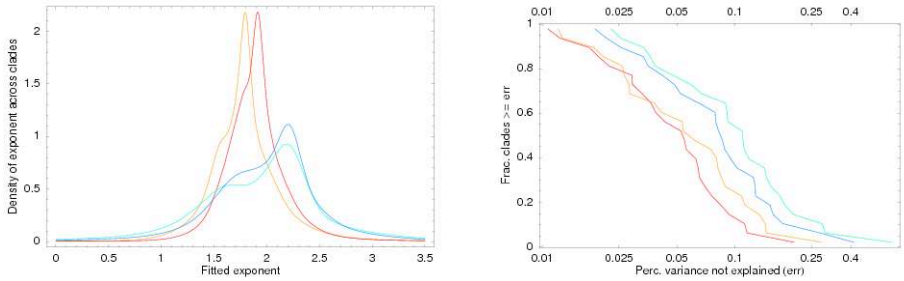

Figure 2.10: Left panel: The distribution of fitted exponents α across the 24 clades for annotation procedures Pfamfunc (red), Pfamall (orange), COGfunc (dark blue), and COGall (light blue). Right panel: Reverse-cumulative distribution of the error (one minus fraction of variance explained by the fit) of the fit for the four annotation procedures. The horizontal axis is shown on a logarithmic scale.

a mixture of Gaussians with means α_i and variances $(\sigma_i)^2$. The left panel of Fig. 2.10 shows the four distributions of α that are obtained in this way.

The left panel of Fig. 2.10 shows that annotation procedure Pfamfunc leads to the smallest variation in fitted exponent α, with a strong peak around $\alpha = 1.9$ and low probability to find an exponent less than 1.5 or larger than 2.5. Annotation procedure Pfamall has a similar distribution, only slightly shifted toward smaller exponents. The distributions for procedures COGfunc and COGall show much wider distributions of α, with a long tail stretching all the way to $\alpha = 1$ for COGall. That is, using COG annotation leads to a significantly higher variance in fitted exponents across clades.

For each power-law fit we can define an 'error' as the fraction of the variance *not* explained by the fit, i.e. the ratio between the average squared-distance of the data points from the line, and the average squared-distance of the data points to the centroid of the data. The right panel of Fig. 2.10 shows the reverse-cumulative distribution of the 'error' across the 24 fits (one for each clade) for each of the 4 annotation procedures. We see that clearly the highest quality fits are obtained with Pfamfunc followed by Pfamall. Significantly lower quality of fits are obtained with COGfunc, and the worst fits are obtained with COGall. The combination of lower variance of inferred exponents and higher fit quality suggests quite strongly that the Pfam annotations are more reliable for the purpose of fitting scaling laws than the COG annotations. This might be the result of the completeness of the annotation varying more across genomes for the COG than for Pfam annotations.

We next investigated to what extend the different annotation procedures infer consistent exponents. We see that the exponents inferred by Pfamfunc correlate reasonably well with those inferred by Pfamall ($r^2 = 0.74$, p-value $1.5 * 10^{-7}$, and similarly those inferred by COGfunc correlate well with those inferred by COGall ($r^2 = 0.87$, p-value $8.4 * 10^{-11}$. In contrast, the exponents inferred by Pfamfunc and COGfunc do not correlate significantly ($r^2 = 0.11$, p-value 0.12) and neither do the exponents inferred by Pfamall and COGall ($r^2 = 0.015$, p-value 0.58). This result shows that deviations from the overall scaling law observed for different clades are typically *not* robust to the annotation procedure that is used. That is, a clade that shows an significantly low exponent using one annotation procedure will typically not show a low exponent when another annotation procedure is used.

Finally, we investigated if there is a systematic bias in the inferred exponent as a function of the quality of the fit. Figure 2.12 shows the correlation between 'error' of the fit and fitted exponent for each annotation procedure.

The figure shows that for the Pfam annotations there is no correlation between the quality of the fit and the fitted exponent (Pfamfunc $r^2 = 0.004$, p-value 0.766, and Pfamall $r^2 = 0.007$, p-

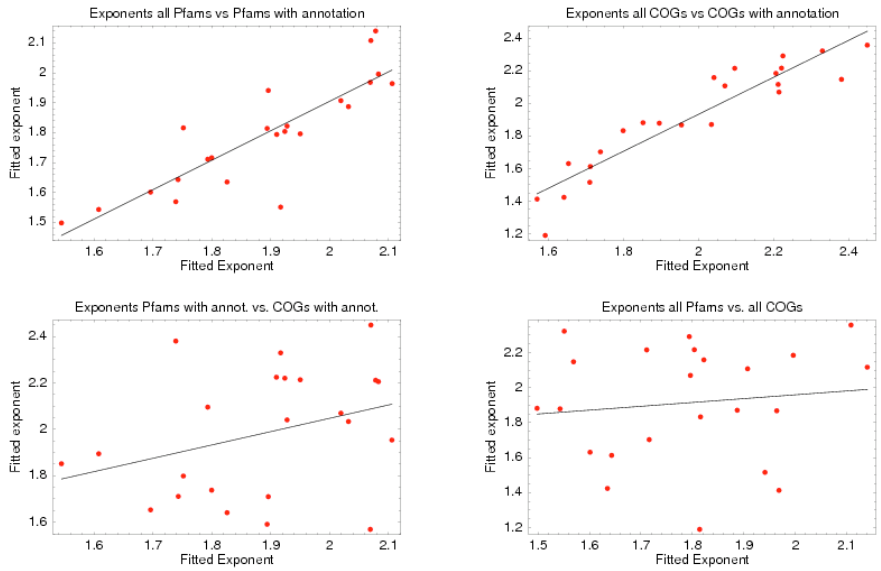

Figure 2.11: Correlation between fitted exponents using the four different annotation procedures. In each panel the red dots show the fitted exponents using the first and second annotation procedure. The straight lines are linear regression fits. Upper left: Pfamfunc against Pfamall. Upper right: COGfuc against COGall. Bottom left: Pfamfunc against COGfunc. Bottom right: Pfamall against COGall.

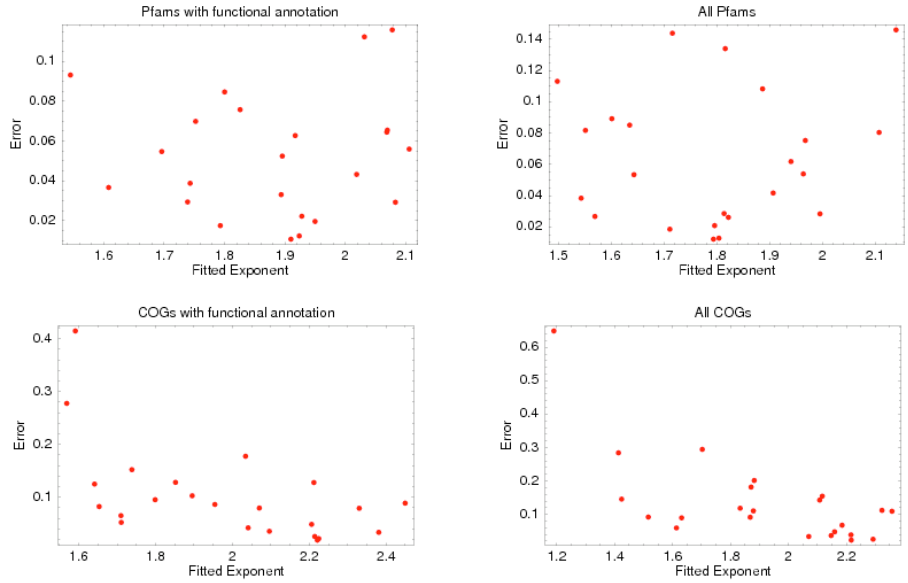

Figure 2.12: Correlation between fitted exponents and the 'error' of the fit for each of the four different annotation procedures. Upper left: Pfamfunc. Upper right: Pfamall. Bottom left: COGfunc. Bottom right: COGall.

value 0.71). In contrast, significant correlations are observed for the COG annotations (COGfunc $r^2 = 0.31$, p-value 0.006, and COGall $r^2 = 0.42$, p-value 0.0008). That is, for the COG annotations there is a systematic bias that leads to lower exponents when the quality of the fit is less. This is a final piece of evidence that, for the purpose of fitting scaling laws, the Pfam annotations are more reliable than the COG annotations.

2.8 Discussion

We investigated to what extent different clades of bacteria show different scaling laws of the number of domains/genes in particular functional categories as a function of the total number of domains/genes in the genome. Somewhat surprisingly, we find that for the large majority of high-level functional categories (116/133) the data are consistent with a common universal scaling law across all clades. The small set of categories that show significant variations across clades are mainly associated with amino acid metabolism, and energy pathways. It is tempting to suggest that the functional categories whose scaling laws vary across clades are characteristic for differences in 'life-style' across clades. Our results would then suggest that differences in 'life-style' among bacteria affects mainly the functional organization of amino acid metabolism and energy pathways. To elucidate this further it might be worthwhile to group genomes not by phylogenetic similarity but rather by life-style, i.e. either their habitat or the type of metabolism they employ, and investigate if there are clear variations observed for bacteria grouped by life-style.

In spite of the considerable interest in elucidating further clade-specific scaling laws, the main striking result presented in this chapter is that the scaling laws are *universal*, i.e. do not vary across clades, for the large majority of high-level functional categories. Since bacteria in different clades have quite different life-styles this shows that the observed scaling laws are independent of bacterial life-style and that the origins of these scaling laws must lie in fundamental physical, biological, and/or evolutionary principles. At this point it is still unclear if the scaling laws originate mainly from physico-chemical constraints that apply to all bacteria, or that they are an inherent result of the evolutionary dynamics. In any case, the fact that most categories show universal scaling supports the simple model of genome evolution that we put forward recently [39], and will be discussed in section 3.3, which assumes relative duplication and deletion rates of domains depend only on the functional category of the domain, i.e. are invariant in time and across different lineages.

It is important to note that we selected our 133 GO categories based on their abundance across bacterial genomes (present in at least 95% of all genomes and with at least 5 domains occurring on average) as well as on the quality of the fit to a scaling law when fitting the data from all genomes (explaining at least 0.95 of the variance). In fact, there are 403 GO categories that pass the abundance criterium, and 270 were discarded because of a poor fit. However, most of these have quite low counts. If we demand that a category accounts for at least 1% of the domains in the genome on average (at least 25 domains on average) only 200 domains are left, of which 125 pass the fit quality threshold. However, it cannot be excluded that some important categories do not show a good fit when fitting to all clades but *do* show good fits to scaling laws for individual clades and this is something we wish to study further in the future.

Finally, one of the most intriguing scaling laws is the approximately quadratic scaling of transcription factors. In a recent work by Cordero and Hogeweg (CH) [32] it was claimed that only some clades obey this quadratic scaling law and that others show much lower exponents, including close to linear scaling for specific clades. We believe that the results presented in this chapter demonstrate that, upon more in-depth analysis of the data, this claim cannot be maintained. Among all categories transcription regulators show in fact a rather low variance in their fitted exponent, with no exponent lower estimated to be lower than 1.7. The discrepancy between our results and those of CH are due to a combination of: 1. a smaller number of genomes with reduced range in size, 2. using standard linear regression rather than Bayesian fitting, 3. using COG annotations rather

than Pfams, and 4. fitting the number of transcription regulators in terms of the total number of proteins rather than the total number of proteins with functional annotation. Effects 1. and 2. lead to systematic underestimation of the exponent and this effect is significantly enhanced by 1. (small range) and 3. (more noisy scatters than with Pfam annotations). We showed that our annotation procedures consistently lead to better quality fits to power-laws and a significantly smaller variation in fitted exponents. The latter is also a strong indication that the underlying scaling law is probably universal: whereas it is easy to see how imperfect and non-homogeneous annotation would lead to increased variations in the fitted exponents across clades, it is hard to imagine how an imperfect annotation could systematically produce scaling laws with very similar slope if the exponents were truly different for different clades. Finally, if the variation in fitted exponents across clades were meaningful we would expect that the variations observed using different annotation methods would correlate. But we observe no such correlation, providing further evidence that these variations are likely an artifact of imperfect annotation and limited data. Finally, as we have discussed in [40] and will be presented in chapter 6 of this thesis, the quadratic scaling of transcription factors has interesting consequences for the topology of transcription regulatory networks across bacteria of different size. The fact that the quadratic scaling is observed for essentially all clades implies that the same regulatory design principle shape the transcription regulatory networks across all bacteria.

3 The evolution of domain-content in bacterial genomes

As we have seen in the previous chapter, across all sequenced bacterial genomes, the number of domains n_c in different functional categories c scales as a power-law in the total number of domains n, i.e. $n_c \propto n^{\alpha_c}$, with exponents α_c that vary across functional categories. Here we investigate the implications of these scaling laws for the evolution of domain-content in bacterial genomes. We show that, using only an assumption of time invariance, uniquely determines the relative rates of domain additions and deletions across all functional categories and evolutionary lineages. In particular, the model predicts that the rate of additions and deletions of domains of category c is proportional to the number of domains n_c currently in the genome and we discuss the implications of this observation for the role of horizontal transfer in genome evolution. Second, in addition to be proportional to n_c, the rate of additions and deletions of domains of category c is proportional to a category-dependent constant ρ_c, which is the same for all evolutionary lineages. This 'evolutionary potential' ρ_c represents the relative probability for additions/deletions of domains of category c to be fixed in the population by selection and is predicted to equal the scaling exponents α_c. By comparing the domain content of 93 pairs of closely-related genomes from all over the phylogenetic tree of bacteria, we demonstrate that the predictions are supported by available genome-sequence data. Our results establish a direct quantitative connection between the scaling of gene numbers with genome size, and the rate of duplications and deletions during short evolutionary time intervals.

3.1 Introduction

When the first gene sequences became available in the 1960s some striking and unexpected patterns were observed. For example, comparison of the fossil record with the number of amino acid substitutions separating orthologous proteins in mammals [4] suggested a constant rate of amino acid substitutions. In addition, the inferred rate of amino acid substitutions was so high that it was hard to imagine how all of these substitutions could have been fixed by the action of natural selection [5]. This famously lead Kimura to propose the neutral theory of molecular evolution [6]. Neutral evolution became the *de facto* null model of sequence evolution and the availability of such a null model in was crucial to the development of rigorous methods for reconstructing evolutionary phylogenies (e.g. [7]) and methods for detecting selection acting on gene sequences (e.g. [8, 9]).

Evolution of course also takes place at higher levels of organization than substitutions within protein-coding genes. In particular, large genomic segments containing one or more genes can be duplicated or deleted, and segments can be 'horizontally transfered', i.e. taken from one organism's genome and inserted into another organism's genome. Through such events organisms can vary the gene content of their genomes, acquiring genes with new functions, subfunctionalizing existing functions, or deleting genes whose functions are no longer required. Now that the sequences of several hundred of whole microbial genomes have become available over the last decade it has become possible to investigate variation in gene-content across genomes in a quantitative manner.

Studies of gene content have uncovered several striking quantitative 'laws'. First of all, it was noticed [11, 12, 13] that a number of key genomic quantities show power-law distributions. In particular, the distribution of gene families is a power-law in each genome, whose exponent appears to depend mostly on the size of the genome. Several theoretical models have been put forth for

explaining these power-law distributions which all include gene duplications and deletions as key ingredients. Another striking observation [1] is that the numbers of genes in different functional categories scale as power-laws in the total number of genes in the genome. For example, whereas the numbers of genes involved in different types of metabolism scale approximately linear with genome size, the number of genes involved with regulatory processes such as transcription regulation and signal transduction scales almost quadratically with genome size, and the number of genes involved with basic processes such as DNA replication or cell division scales with an exponent less than 1. Such scaling laws are observed for the large majority of high-level functional categories of genes and appear to apply to all bacterial genomes.

As we have argued previously [1, 2], these scaling laws have important implications for the evolutionary dynamics of gene duplications and deletions and we will here investigate these implications in detail. The organization of the chapter is as follows. We study genome evolution at the level of protein domains and we start by demonstrating that scaling laws are also observed at the level of the number of protein-domains. We re-estimate the scaling exponents α_c using all 630 currently available genomes. Next, using the assumption that the scaling laws are time invariant, we derive a 'null model' for genome evolution that accounts for the observed scaling laws. In this model the exponents of the scaling laws are identified as universal constants of the evolutionary process.

We collected 93 pairs of closely-related bacterial genomes and tested the model's predictions by analyzing the protein-domain content of these genomes and estimating, for each pair, the number of domain additions/deletions that have occurred since their common ancestor. We show that essentially all of the model's predictions are supported by the available genome data. Finally, we also discuss the important implications of our results for the role of horizontal gene transfer in genome evolution.

3.2 Evolutionary model

We want to investigate the implications of the scaling laws (2.2) for evolutionary dynamics. That is, we want to infer what the scaling laws imply for the behavior of the domain number counts $n_c(t)$ as a function of time t. It is important to define precisely what we mean by $n_c(t)$. A sequenced genome g represents a particular bacterial strain and can idealistically be thought of as representing the genome of a single bacterial organism living today with domain counts $n_c(g)$. Since bacteria reproduce clonally we can imagine tracing this individual back through time, back to the its mother cell, its grandmother, and eventually all the way back until the common ancestor of all currently sequenced genomes. We denote by $n_c(g,t)$ the number of domains of category c that were present in the ancestor organism of genome g that was living at time t.

Let t_{now} denote today and let $x_c(g,t)$ denote the logarithm of the domain-number, i.e. $x_c(g,t) = \log[n_c(g,t)]$, and similarly $x(g,t) = \log[n(g,t)]$. In these variables the scaling laws are just straight lines, i.e all genomes g (approximately) obey the linear relation

$$x_c(g, t_{\text{today}}) = \alpha_c x(g, t_{\text{today}}) + \beta_c \; \forall g. \tag{3.1}$$

We will now derive how these scaling laws constrain how the domain-numbers have *changed* throughout time. Let $t = 0$ denote the time at which the last common ancestor of all sequenced bacterial genomes was alive. Note that, since the GO categories that we consider occur in almost all genomes, it is reasonable to assume that they all had nonzero count in the last common ancestor. We let $x_c(0)$ denote the log-domain counts in this common ancestor and $x(0)$ the logarithm of the total domain count. Further, we denote by $dx_c(g,t)$ the change in the log domain-count for category c, that occurred in a small interval of time centered around time t in the evolutionary history of genome g. The log domain-counts $x_c(g,t)$ and $x(g,t)$ are then by definition given by the integrals

$$x_c(g, t_{\text{now}}) = x_c(0) + \int_0^{t_{\text{now}}} dx_c(g,t), \tag{3.2}$$

and

$$x(g, t_{\text{now}}) = x(0) + \int_0^{t_{\text{now}}} dx(g, t). \tag{3.3}$$

Comparing equations (3.2) and (3.3) with equation (3.1) the scaling laws thus imply that we have

$$x_c(0) + \int_0^{t_{\text{now}}} dx_c(g, t) = \beta_c + \alpha_c \left[x(0) + \int_0^{t_{\text{now}}} dx(g, t) \right] \ \forall g. \tag{3.4}$$

Since (3.4) must hold for *all* genomes g, this equation first of all implies a relation between the offsets β_c and the domain counts in the last common ancestor:

$$\beta_c = x_c(0) - \alpha_c x(0). \tag{3.5}$$

More importantly, we find that all genomes must obey

$$\int_0^{t_{\text{now}}} dx_c(g, t) = \alpha_c \int_0^{t_{\text{now}}} dx(g, t) \ \forall g. \tag{3.6}$$

For short time intervals in which the changes in n_c are small relative to n_c itself, the changes in x_c are related to the changes in n_c through

$$dx_c(g, t) = \frac{dn_c(g, t)}{n_c(g, t)}, \tag{3.7}$$

and similarly

$$dx(g, t) = \frac{dn(g, t)}{n(g, t)}. \tag{3.8}$$

Substituting these in (3.6) we obtain

$$\alpha_c = \frac{\int_0^{t_{\text{now}}} \frac{dn_c(g,t)}{n_c(t)}}{\int_0^{t_{\text{now}}} \frac{dn(g,t)}{n(g,t)}} \ \forall g. \tag{3.9}$$

Equation (3.9) summarizes the implications for domain-count dynamics implied by the scaling laws. It states that, *independent* of which evolutionary history we take, the ratio of the integrals of dn_c/n_c and dn/n over all evolutionary time must match the scaling exponent α_c. This is illustrated on the left-hand side of figure 3.1, i.e. equation (3.9) implies that the ratio of integrals is the same for each of the evolutionary histories indicated as colored lines.

3.3 Time invariance

The equations (3.9) reflect the constraints on domain-count dynamics implied by the scaling laws but they don't uniquely determine an evolutionary model. To derive a unique evolutionary *null model* we will assume *time invariance* of the scaling laws. We assume that, if we had collected genomes of bacteria living several tens or even hundreds of million years ago, as opposed to the bacteria living today, we would have observed the *same* scaling laws as we observe today. That is, we assume that there is nothing particularly special about our current time, and that the same scaling laws have held since the last common ancestor, or at least since the origin of the clades from which our current genome sequences derive. We feel that this is by far the simplest assumption that can be made about the evolutionary dynamics and will here analyze its implications.

Given that the scaling laws are invariant in time, we immediately obtain that (3.9) should hold for each *short* time interval, i.e. we have that

$$\frac{dn_c(g, t)}{n_c(g, t)} = \alpha_c \frac{dn(g, t)}{n(g, t)} \ \forall g, t, \tag{3.10}$$

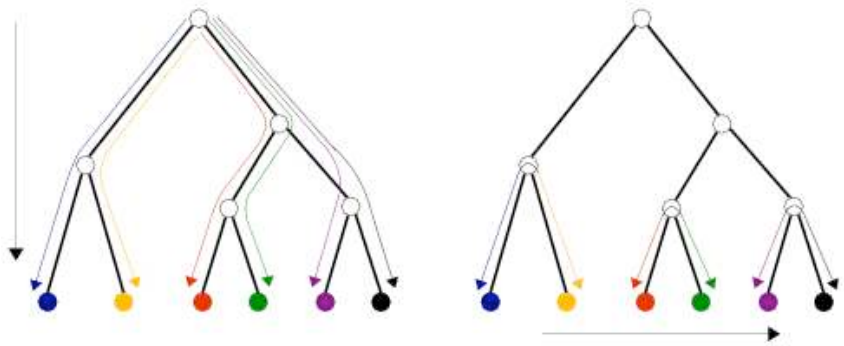

Figure 3.1: Evolutionary histories of different organisms. The scaling laws constrain integrals of domain-count changes over long evolutionary times, i.e. from the common ancestor up to the present (left panel). Our assumption of time invariance now implies relations between the domain-count changes during short time intervals which can be tested by comparing domain-counts in closely-related genomes (right panel).

or

$$\frac{dn_c(g,t)}{dn(g,t)} = \alpha_c \frac{n_c(g,t)}{n(g,t)} \; \forall g,t. \tag{3.11}$$

That is, the assumption of time invariance implies that, for each genome g, and for each short time interval in its evolution, the ratio between the change $dn_c(g,t)$ in the domain-count of category c and the total change $dn(g,t)$ in domain-count is given by the product of the exponent α_c and the fraction $n_c(g,t)/n(g,t)$ of all domains that are of category c. In particular, equation (3.11) will apply to the domain-count changes that occurred since the common ancestors of pairs of closely-related species, as illustrated on the right-hand side of Fig. 3.1. Therefore, we can test the validity of the null model by comparing the domain-counts in the genomes of closely-related bacteria.

3.4 Implications for closely-related pairs of genomes

We now discuss how the prediction (3.11) can be tested with data from closely-related genomes. Note that, strictly speaking, (3.11) holds only in the limit of infinitesimally small $dn(g,t)$ and that we have so far implicitly assumed that the $n_c(g,t)$ are continuous variables, whereas in reality the smallest possible change is $dn(g,t) = 1$. For the integer-valued quantities $n_c(g,t)$ equation (3.11) can be interpreted as follows: whenever a single domain is added to the genome, i.e. $dn = 1$, then the *probability* that this domain is of category c is given by $\alpha_c n_c/n$. Similarly, whenever a single domain is removed, i.e. $dn = -1$, then the probability that this domain is of category c is also given by $\alpha_c n_c/n$. Equivalently, if r denotes the overall rate at which additions or deletions occur, and r_c the rate at which additions/deletions of domains of category c occur, then the model predicts

$$\frac{r_c}{r} = \alpha_c \frac{n_c}{n}. \tag{3.12}$$

For pairs of closely-related genomes the number of domain-count changes that occurred since they diverged from a common ancestor is generally very small compared to the total number of domains. Therefore, the fractions n_c/n have generally changed little during the time since the two genomes diverged from their ancestor and we will make the assumption that the fraction n_c/n can be considered constant. Under this approximation equation (3.12) predicts that, if during the time interval since the pair's common ancestor, a total of Δn domain-count changes occurred, i.e. counting both additions and deletions, then the expected number of domain-count changes Δn_c in category c should equal $\alpha_c \frac{n_c}{n} \Delta n$.

We collected 93 pairs of fully-sequenced genomes that are evolutionary relatively closely related, using the tree of life that was inferred by Bork et al. [41] as a guide. For each pair of genomes i we counted the numbers of domain occurrences for each Pfam family and used these to estimate the number of domain-count changes Δn_c^i for each category c and the total number of domain-count changes Δn^i. Similarly we estimated, for each genome pair i, the fractions fraction n_c^i/n^i by averaging the domain counts over the two genomes in the pair. Our model thus predicts that, for each pair i, the ratio $\Delta n_c^i/\Delta n^i$ should be proportional both to the fraction n_c^i/n^i and to scaling law exponent α_c.

3.5 Estimating domain-count changes Δn_c

We extracted the phylogenetic tree of bacteria from the tree of life that was produced by Bork et al. [41] based on the concatenated protein sequences of 31 protein families. From this tree we considered all pairs of species for which the average identity at the amino acid level of orthologous proteins was at least 0.75, i.e. distance less than 0.25. To avoid having redundant pairs we clustered all species whose distances were 0.01 or less and took a single representative genome from each cluster. With these cutoffs we obtained 93 pairs of bacterial genomes which are listed in the appendix.

We estimate the number of domain-count changes Δn and Δn_c by comparing domain counts for each Pfam family separately. Let n_f^1 and n_f^2 denote the number of occurrences of domains from family f in the first and second genome of the pair. We will assume that, during the time from the common ancestor of the two genomes, the rates at which domains were added and deleted for each family f is an unknown constant. In principle there are 4 unknown rates for each domain family f: the rate λ_f^1 at which domains of family f are added to genome 1, the rate λ_f^2 at which domains of family f are added to genome 2, the rate μ_f^1 at which domains of family f are removed from genome 1, and the rate μ_f^2 at which domains of family f are removed from genome 2. Since we cannot distinguish between additions to genome 1 and removals from genome 2 (and similarly for removals from genome 1 and additions to genome 2) we define the following rate sums

$$\lambda_f = \lambda_f^1 + \mu_f^2, \tag{3.13}$$

and

$$\mu_f = \lambda_f^2 + \mu_f^1. \tag{3.14}$$

We denote by a_f the number of additions in genome 1 plus deletions in genome 2, and by d_f the number of additions in genome 2 plus deletions in genome 1. Since the rates of additions and deletions are assumed constant, both a_f and d_f are Poisson distributed

$$P(a_f, d_f|\lambda_f, \mu_f, t) = \frac{(\lambda_f t)^{a_f}(\mu_f t)^{d_f}}{a_f! d_f!} e^{-(\lambda_f+\mu_f)t} \tag{3.15}$$

The expected total number of additions is

$$\lambda = \sum_f \lambda_f t, \tag{3.16}$$

and the expected total number of deletions is given by

$$\mu = \sum_f \mu_f t. \tag{3.17}$$

The fractions of changes (additions or deletions) involving domain family f is

$$x_f = \frac{(\lambda_f + \mu_f)t}{\lambda + \mu}. \tag{3.18}$$

In terms of these variables the probability of obtaining the set of additions and deletions $\{a_f, d_f\}$ is

$$P(\{a_f, d_f\}|\lambda, \mu, \{x_f\}) = \prod_f \frac{(\lambda x_f)^{a_f}}{a_f!} \frac{(\mu x_f)^{d_f}}{d_f!} e^{-(\lambda+\mu)}. \tag{3.19}$$

Assume that the number n_f^1 of domains of family f in genome 1 is bigger than the number n_f^2 of domains of family f in genome 2 and denote by δn_f the difference, i.e. $\delta n_f = n_f^1 - n_f^2$. We know that the number of additions a_f must be at least δn_f. Let e_f the number of "extra" additions. Note that the number of deletions d_f is necessarily equal to e_f. Similarly, if $n_f^2 > n_f^1$ we define, $\delta n_f = n_f^2 - n_f^1$ and we write $d_f = \delta n_f + e_f$, and $a_f = e_f$. In terms of the δn_f and the extra moves e_f the probability is given by

$$P(\{\delta n_f, e_f\}|\lambda, \mu, \{x_f\}) = e^{-(\lambda+\mu)} \lambda^{A+E} \mu^{D+E} \prod_f \frac{(x_f)^{\delta n_f + 2e_f}}{e_f!(\delta n_f + e_f)!}, \tag{3.20}$$

where we have defined

$$A = \sum_{f|n_f^1 > n_f^2} \delta n_f, \tag{3.21}$$

$$D = \sum_{f|n_f^2 > n_f^1} \delta n_f, \tag{3.22}$$

and

$$E = \sum_f e_f. \tag{3.23}$$

To estimate the number of additions and deletions for each family f we maximize the probability (3.20) with respect to λ, μ, the fractions x_f, and the number of extra moves e_f. To do this we use an iterative procedure. Note that for given extra moves e_f the optimal λ, μ, and x_f are given by

$$\lambda = A + E, \tag{3.24}$$

$$\mu = D + E, \tag{3.25}$$

and

$$x_f = \frac{\delta n_f + e_f}{\sum_{\tilde{f}} \delta n_{\tilde{f}} + e_{\tilde{f}}}. \tag{3.26}$$

Similarly, when x_f is given, the probability of e_f conditioned on these variables is given by

$$P(e_f|\lambda, \mu, x_f, \delta n_f) \propto \frac{(x_f)^{\delta n_f + 2e_f}}{e_f!(\delta n_f + e_f)!}, \tag{3.27}$$

and we can numerically solve for the e_f that maximizes this likelihood. We start by setting all $e_f = 0$ and use the above equations to, iteratively, solve for λ, μ and the x_f given the e_f, and then the e_f given the x_f. This is repeated until a fixed point is reached. Finally, the estimated total number of events Δn_f for family f equals $\delta n_f + 2e_f$. In this way we estimate the number of events Δn_f^i separately for each of the genome pairs i we analyze.

The estimated total number of changes in category c is given by $\Delta n_c^i = \sum_{f \in c} \Delta n_f^i$, where the sum is over all Pfam domain families f associated with category c. The estimated total number of changes is given by $\Delta n^i = \sum_f \Delta n_f^i$, where the sum is over all Pfam domain families. To calculate the fractions n_c^i/n^i for a given closely-related pair i we calculate the average number of domains associated with category c as $n_c^i = \sum_{f \in c}(n_f^1 + n_f^2)/2$ and the average total number of domains $n^i = \sum_f (n_f^1 + n_f^2)/2$.

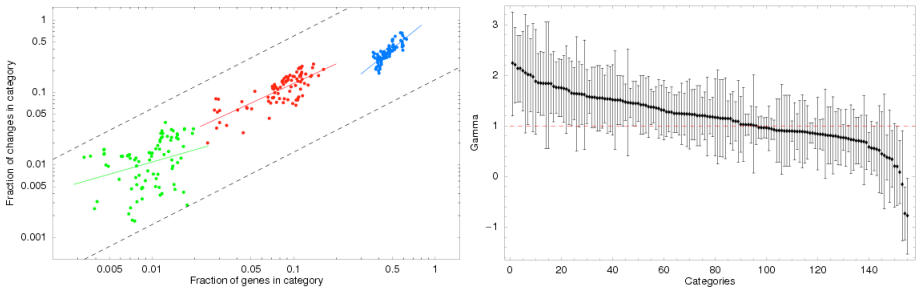

Figure 3.2: Linear dependency of the fraction of domain-count changes on the domain-count itself. **Left panel**: For each genome pair i the fraction $\Delta n_c^i/\Delta n^i$ of domain-count changes that involve domains of category c is shown (vertical axis) as a function of the fraction n_c^i/n^i of all domains in the genome that are associated with category c (horizontal axis) for the categories 'metabolic process' (green) , 'regulation of transcription' (red), and 'two-component sensor activity' (blue). Each dot corresponds to the data for one pair i of closely-related genomes. Both axes are shown on a logarithmic scale. The straight-lines show least-squares fits of the form $\log[\Delta n_c^i/\Delta n^i] = \gamma_c \log[n_c^i/n^i] + \delta_c$. The fitted slopes for the three categories are $\gamma_{\text{twocomp.sensor}} = 0.71 \pm 0.46$, $\gamma_{\text{reg.transcr.}} = 0.9 \pm 0.2$, and $a_{\text{met.proc.}} = 1.58 \pm 0.32$. For comparison the dotted lines show linear scaling. **Right panel**: A 99% posterior probability interval for the slope γ_c was estimated for all selected GO categories. The fitted slopes were ordered from high to low and are shown in the right panel from left to right with the vertical bars corresponding to the 99% posterior probability intervals for each slope γ_c. The slope $\gamma = 1$, corresponding to a linear dependency, is shown as a horizontal dotted line.

3.6 Scaling of the fraction of domain-count changes

Equation (3.12) puts very strong constraints on the dynamics of domain-counts which we will check in three steps. First, we check that, for each category c, the estimated fractions $\Delta n_c/\Delta n$ of domain-count changes grow linearly with the fractions n_c/n. The left panel of figure 3.2 shows scatter plots of $\Delta n_c^i/\Delta n^i$ as a function of n_c^i/n^i for three selected categories. The axes are shown on logarithmic scales and the straight lines show least-squares linear fits of the form $\log[\Delta n_c^i/\Delta n^i] = \gamma_c \log[n_c^i/n^i] + \delta_c$.

The left panel of Fig. 3.2 demonstrates two points. First, comparing the three categories with each other, we see that most domain-count changes occur in the most abundant category and least domain-count changes occur in the least abundant category, with the fraction of domain-count changes $\Delta n_c^i/\Delta n^i$ indeed scaling roughly linearly with n_c^i/n^i (compare with the dotted guide lines showing linear scaling). Beyond that, if we compare the numbers of domain-count changes across the different genomes *within* each category we see that, in those genomes where the domains of the category are most abundant domain-count changes in that category are also most abundant. That is, although the data is quite noisy, it is clear that all three clouds of points show a close to linear increase of $\Delta n_c^i/\Delta n^i$ with n_c^i/n^i.

The estimated slopes γ_c for all selected GO categories are shown in the right panel of Fig. 3.2 (and listed in the appendix). The estimated γ_c are very roughly symmetrically distributed around 1 with a median γ_c of 1.18 and a mean γ_c of 1.16. For almost 75% of the categories a slope of $\gamma_c = 1$ is within the 99% posterior probability interval. This thus supports the prediction of our evolutionary null model that the fraction of all domain-count changes that involve domains of category c is proportional to the fraction n_c/n of all domains in the genome that belong to category c.

For about 25% of the categories we infer slopes significantly deviating from 1. It should be noted, however, that the least-squares fitting assumes simple Gaussian noise in $\log[\Delta n_c/\Delta n]$, whereas in reality the size of the noise in $\log[\Delta n_c/\Delta n]$ increases as Δn decreases. Moreover, whereas the

fitting assumes that the numbers of domain-count changes are given, in reality these are estimated and thus themselves subject to uncertainty. We therefore are significantly underestimating the uncertainty in the fitted slope for many categories, and it is reasonable to conclude that for most if not all categories the data is consistent with the predicted linear dependence of $\Delta n_c/\Delta n$ on n_c/n.

3.7 Evolutionary Potentials

The results of the previous section strongly suggest that the rate r_c of domain-count changes involving domains of category c is proportional to the number of domains n_c currently present in the genome. Let r_c^i denote the rate of addition/deletion of domains of category c for genome pair i and let r^i denote the overall rate of addition/deletion of domains for genome pair i. Assuming only that r_c^i is proportional to n_c^i we can generally write for the relative rates

$$\frac{r_c^i}{r^i} = \rho_c^i \frac{n_c^i}{n^i}, \tag{3.28}$$

which is the generalization of equation (3.12). The proportionality constants ρ_c^i defined by this equation quantify the extent to which domain-count changes of category c are more or less frequent in the lineages of pair i than expected based on their frequency n_c^i/n^i. For this reason we will refer to these proportionality constants as *evolutionary potentials*. That is, when ρ_c^i is high it indicates that, apparently, domain additions and deletions involving domains of category c are fixed in evolution at a higher rate in the evolutionary lineages of pair i.

Our evolutionary null model predicts that the evolutionary potentials ρ_c^i are the same for all evolutionary lineages, and in addition that the evolutionary potentials ρ_c^i are equal to the scaling law exponents α_c. We will check these two predictions in turn.

3.8 The evolutionary potentials ρ_c^i are constant across lineages

Given the estimated numbers of domain-count changes Δn_c^i, and the total number of domain-count changes Δn^i we can estimate the lineage-specific evolutionary potentials ρ_c^i as follows. For every domain-count change that occurs, the probability that it will involve a domain of category c is simply given by the relative rate r_c^i/r^i. Therefore, if Δn^i domain-count changes occur in total, the probability that Δn_c^i involve domains of category c is simply given by

$$P(\Delta n_c^i|\Delta n^i, \rho_c^i) = \binom{\Delta n^i}{\Delta n_c^i} \left(\rho_c^i \frac{n_c^i}{n^i}\right)^{\Delta n_c^i} \left(1 - \rho_c^i \frac{n_c^i}{n^i}\right)^{\Delta n^i - \Delta n_c^i}, \tag{3.29}$$

where we used the definition (3.28). Using a uniform prior over ρ_c^i we find for the posterior probability of ρ_c^i given the estimated domain-count changes

$$P(\rho_c^i|\Delta n^i, \Delta n_c^i)d\rho_c^i = \frac{n_c^i}{n^i} \frac{(\Delta n^i + 1)!}{\Delta n_c^i!(\Delta n^i - \Delta n_c^i)!} \left(\rho_c^i \frac{n_c^i}{n^i}\right)^{\Delta n_c^i} \left(1 - \rho_c^i \frac{n_c^i}{n^i}\right)^{\Delta n^i - \Delta n_c^i} d\rho_c^i \tag{3.30}$$

Using (3.30) we determined posterior probability intervals $[l_c^i, h_c^i]$ defined by

$$\int_0^{l_c^i} P(\rho|\Delta n^i, \Delta n_c^i)d\rho = 0.01, \tag{3.31}$$

and

$$\int_0^{h_c^i} P(\rho|\Delta n^i, \Delta n_c^i)d\rho = 0.99, \tag{3.32}$$

for each category c and each genome pair i. Figure 3.3 shows these posterior probability intervals, for all genome pairs i, for the categories 'translation', 'metabolic process', and 'regulation of transcription'.

Since the total number of domain-count changes Δn^i is often small, it is not surprising that the posterior probability intervals are often rather wide. In spite of this, it can be clearly seen that, consistent with the scaling exponents α_c, ρ_c^i is largest for the category 'regulation of transcription', and smallest for the category 'translation'. Moreover, Fig. 3.3 shows that data by and large support the prediction that the potentials ρ_c^i are *the same* for all evolutionary lineages. That is, for each of the three categories the probability intervals ρ_c^i of the majority of genome pairs i are consistent with a common underlying potential ρ_c. This is a further piece of support for the evolutionary null model.

3.9 Evolutionary potentials ρ_c correlate with scaling exponents α_c

The previous section has shown that the data are consistent with constant evolutionary potentials across the genome pairs and we will now that the evolutionary potentials ρ_c^i are equal to a common potential ρ_c and estimate it by combining data from all genome pairs. We find for the probability of the observed domain-count changes $\{\Delta n_c^i\}$ and $\{\Delta n^i\}$

$$P(\rho_c|\{\Delta n_c^i\},\{\Delta n^i\}) \propto \prod_i \left(\rho_c \frac{n_c^i}{n^i}\right)^{\Delta n_c^i} \left(1-\rho_c \frac{n_c^i}{n^i}\right)^{\Delta n^i - \Delta n_c^i}. \tag{3.33}$$

Using this equation we estimate ρ_c for each selected category c. Equation (3.12) predicts that the evolutionary potentials ρ_c equal the scaling exponents α_c. Figure 3.4 shows a scatter plot of α_c against the estimated ρ_c.

Note that, since the evolutionary potential ρ_c is a measure of frequency in domain-count changes between closely-related species, and α_c is a measure of the scaling of the number of domains with genome size, there is *a priori* no reason why these two quantities should be strongly correlated. However, as predicted by our evolutionary null model, there is a clear evidence of a linear dependency between the exponents α_c and the evolutionary potentials ρ_c.

Rather than a simple relation $\rho_c = \alpha_c$ we find that ρ_c varies over a somewhat smaller range, i.e. the 99% posterior probability interval for the slope of the correlation runs from 0.7 to 0.83. One possible explanation is that, because the estimation of the numbers of domain-count changes Δn_c is the same for all categories, we might underestimate the numbers of domain-count changes more for categories with large ρ_c than for categories with low ρ_c.

3.10 Implications for the rates of horizontal transfer

In general, the rate at which additions/deletions occur is the product of two independent factors. First, the rate at which domain additions and deletions are *introduced* into individuals of the population, and second the fraction of the time that such mutations are being fixed into the population. There are likely three main mechanisms through which domain additions or deletions are introduced: duplications, deletions, and horizontal transfers. To a first approximation, the rates at which duplications, deletions, and horizontal transfers are being introduced into individuals will be determined by the biases inherent in the mechanisms underlying these processes and not by selection. In contrast, the fraction of the time that such mutations are fixed in evolution will strongly depend on selection.

It is clear that, for duplications and deletions, the rate at which such mutations are introduced is naturally proportional to the number of existing domains n_c. That is, when the number of domains

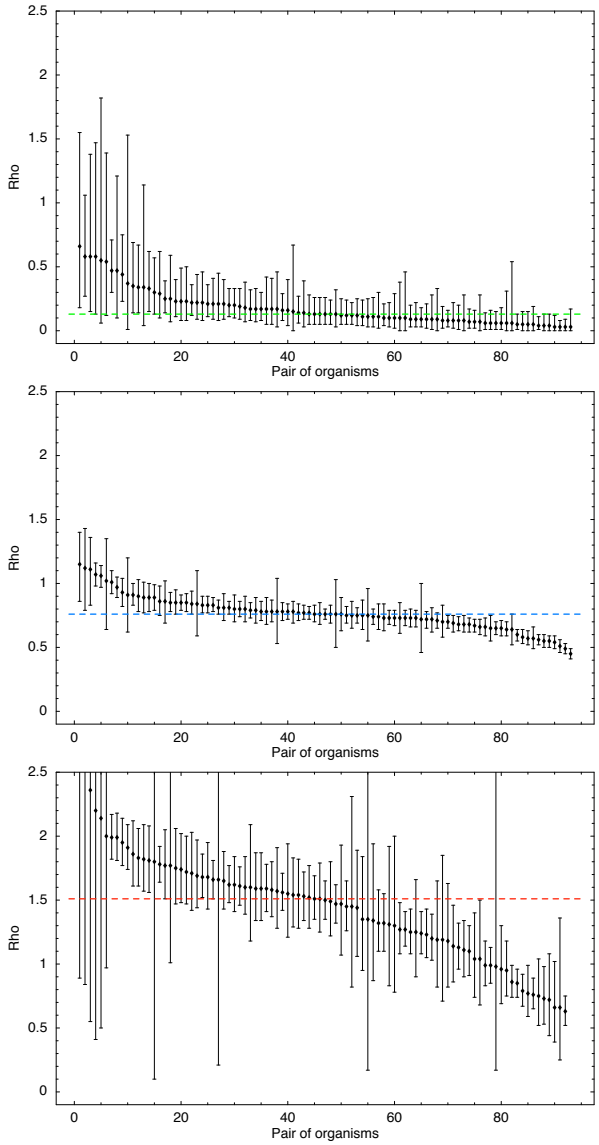

Figure 3.3: Distribution of inferred evolutionary potentials ρ_c^i for the categories 'translation' (left panel), 'metabolic process' (middle panel), and 'regulation of transcription' (right panel) across all genome pairs i. Each panel shows the 98% posterior probability intervals $[l_c^i, h_c^i]$ for the potentials ρ_c^i as vertical bars (sorted from left to right by their means). The dotted horizontal lines show the average ρ_c^i, averaged over all pairs i.

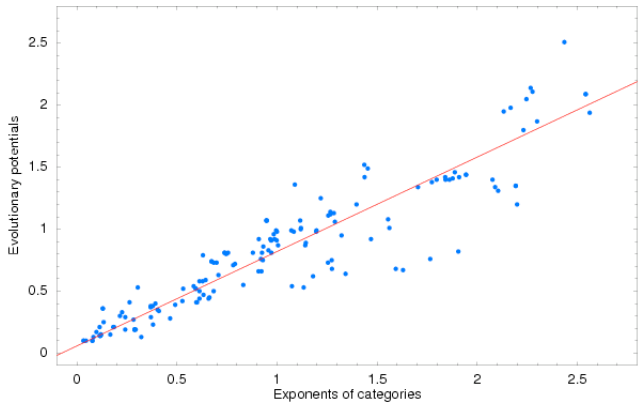

Figure 3.4: Correlation between the inferred evolutionary potentials ρ_c (vertical axis) and the exponents α_c (horizontal axis) of the scaling laws. Each dot corresponds to one of the 156 selected GO categories. The line shows the linear fit $\rho_c = 0.76\alpha_c + 0.06$ with correlation coefficient $r^2 = 0.85$.

n_c doubles, the total rate of duplication and deletion within this category also doubles. Moreover, since selection is not involved, the rate of introduction of duplications and deletions will be the *same* for all functional categories c (except of course for transposable elements which are duplicated through a separate mechanism). Therefore, as the rate of introduction is proportional to n_c, with the same proportionality constant for each category, and the total rate must be proportional to $\rho_c n_c$, this implies that the relative rate of *fixation* through selection must be proportional to the evolutionary potential ρ_c. Thus, the evolutionary potentials ρ_c (and the scaling exponents α_c) have a particularly simple interpretation: they give the average relative rate with which additions and deletions of domains in category c are fixed by selection.

As we have mentioned already, evidence has accumulated over recent years that horizontal transfers are common (e.g. [42, 43, 44, 10]) and that they account for a non-negligible fraction of changes in gene content, at least among closely-related genomes. Although we have no direct evidence, it is attractive to assume that the probability that a domain addition will be fixed in the population does *not* depend on the mechanism by which it was introduced. That is, the relative rate of fixation of domain additions in category c should be proportional to ρ_c for both duplicated domains as well as horizontally transfered domains. If this is indeed the case, it follows immediately from the fact that the overall rate should be proportional to $\rho_c n_c$, that the rate at which horizontal transfers are *introduced* must be proportional to the number of domains n_c present in the genome. However, whereas this is naturally the case for gene duplications, it is not clear at all why this should also hold for horizontal transfers. Therefore, our results put rather strong constraints on the rate of horizontal transfer.

One possibility is that horizontal transfer is negligible and that domain additions are dominated by duplications. This assumption, which we have made in previous work [1, 2] is at odds, however, with recent studies that establish a significant role for horizontal transfer [43, 42]. One possibility is that most horizontal transfers are only transient and that the domain-count changes that are maintained across long evolutionary times are mostly due to duplications and deletions, although this a priori does not seem plausible.

Alternatively, there are several hypotheses that could explain why the rate at which horizontal transfers of domains of category c are introduced is proportional to the number of domains n_c already in the genome. First, it is possible that horizontal transfer is highly biased to occur predominantly between genomes that are closely-related phylogenetically. Since closely-related species are likely to have highly correlated domain counts, it is likely that the fraction n_c/n of category c domains in the donor genome is close to the fraction of domains of category c in the

receiver genome. However, many of the horizontal transfers detected through sequence analysis involve transfers between distally related species.

Another possible explanation is that bacterial habitats naturally separate into different genome-size classes. That is, it is conceivable that bacteria tend to be surrounded by other bacteria of the same genome size. Because the scaling laws apply to all genomes, the fractions n_c/n are similar for similarly sized genomes and one would naturally have that the rate at which horizontal transfers of domains of category c occur is proportional to n_c. However, as far as these authors are aware, there seems to be no evidence suggesting bacteria cluster with other bacteria of similar genome size.

Finally, it is possible that, even though a given bacterium would generally be surrounded by other bacteria of very different sizes, that horizontal transfer is highly biased to occur predominantly between organisms that have genomes with similar sizes. In fact, there is some evidence in the literature that bacteria can recognize and silence horizontally transfer ed genes that have an AT-content which is significantly higher than the AT-content of the genome itself [45]. In addition, there is generally a good correlation between genome size and GC-content [46]. It is therefore conceivable that horizontal transfer between genomes of similar size are much more common than horizontal transfers between genomes of significantly different sizes.

In any case, whatever the underlying mechanism, if horizontal transfers account for a significant fraction of domain additions through evolution, then something must ensure that the rate of such horizontal transfers is proportional to the number of existing domains n_c in the receiving genome.

3.11 Discussion

We have shown that, across all bacteria and for most high-level GO categories c, the number of domain occurrences n_c scales as a power-law in the total number of domains n, with scaling exponents α_c varying from close to zero to a bit larger than 2. We have derived what we believe is the simplest evolutionary model that can account for the observed scaling laws. This 'null model assumes that, across all evolutionary lineages and all evolutionary times, the relative rate r_c/r at which additions and deletions of domains of category c are fixed in evolution is proportional to the current fraction n_c/n of domains in category c and a characteristic *evolutionary potential* ρ_c which equals the scaling exponent α_c.

By comparing genome-wide domain-counts n_f for each Pfam family f across 93 pairs of closely-related species we have estimated the rates at which domain additions and deletions occur across GO categories and across different evolutionary lineages. The results of this analysis support the predictions made by the evolutionary null model. First, we have shown that, for most categories c, the relative rate r_c/r of domain additions and deletions is proportional to the fraction of domains n_c/n already occurring in the genome. Second, we estimated the relative rates r_c^i/r^i of domain additions and deletions independently for different evolutionary lineages i and used these to estimate lineage-dependent evolutionary potentials ρ_c^i. We found that, whereas the evolutionary potentials ρ_c^i clearly vary between categories c, the data support the null model's prediction that for a given category c the potentials ρ_c^i are the same across all evolutionary lineages i. Finally, by combining data from all lineages we estimated average evolutionary potentials ρ_c and found that, as predicted by the model, there is a good correlation between these evolutionary potentials and the scaling law exponents α_c. Importantly, this result establishes that, there is a direct relation between the scaling of domain-counts with genome size and the rates with which domains are added and removed during short evolutionary time intervals. This reinforces our proposal that the evolutionary potentials ρ_c are fundamental constants of the evolutionary process.

If, as recent work suggests, horizontal transfer is an important force in shaping the gene-content of genomes, then our results put strong constraints on the rates r_c at which horizontal transfers of domains of different functional categories c can occur. In particular, we find that the rate at which domains of category c are horizontal transfer ed into a genome must be proportional to the

number of domains n_c already existing in the receiving genome. An important avenue for future research is to clarify the underlying mechanism that is responsible for this surprising fact.

As our results have made plausible that the evolutionary potentials ρ_c (and the corresponding scaling exponents α_c) are fundamental constants of the evolutionary process that apply across all time and all evolutionary lineages, the major challenge is now to elucidate what determines these numbers. In this respect it is important to note that the functional categories c that we consider are taken directly from the human-defined Gene Ontology and are thus rather subjective. A first challenge for future work is therefore to identify a procedure that divides domain families into functional groups in a more objective manner. Although difficult with the current amount of available data, one possible approach is to estimate evolutionary potentials ρ_f for individual domain families and to investigate if these fall into a small number of natural classes. That is, it is conceivable that on some more fundamental level there are only a small number of distinct exponents, for example $\alpha = 0$, $\alpha = 1$, and $\alpha = 2$, and that the observed scaling laws with more complex exponents are different mixtures of these more fundamental scaling laws. Finally, we believe that the exponents α_c reflect fundamental design principles of bacterial life, maybe similar to the way geometry and architectural design principles demand that the number of windows in a building scales as the 2/3 power of the building's volume. Seen from this point of view the exponents α_c encode crucial information about the basic design that is shared by all bacterial life.

4 A novel method to detect purifying selection

We have developed an integrated set of algorithms for comprehensive footprinting of bacterial genomes starting from the genomes of a reference species and a set of related species. Our methodology includes new algorithms for comprehensive mapping of orthologs, inferring the phylogenetic tree relating the species, and aligning orthologous intergenic regions. Finally, we developed a Bayesian probabilistic model for identifying sequence segments that are under selection, and using this model we identify conserved segments in all intergenic regions of the reference species. Comparison of our predictions in E. coli with known transcription factor binding sites shows a high overlap between predicted segments and known binding sites. We have upload to SwissRegulon [47], a database with genome-wide annotations of regulatory sites, all highly conserved segments of 22 reference genomes that span widely the whole bacterial phylogenetic tree.

4.1 Introduction

At the time of writing, there are 742 fully-sequenced bacterial and archael genomes available in the NCBI database [48]. Reliable annotations of the positions of predicted protein-coding and RNA genes are easily available for all these genomes. In addition, by comparing the protein sequences with models of protein families and protein domains [49] it is possible to obtain rough functional annotations for the large majority of all predicted genes [50]. In contrast, very little is known about the occurrence of regulatory sites and other functional sites in the intergenic regions between the genes. For the highly studied model organisms E. coli and B. subtilis there are databases available [51, 52] that collect a significant number of known binding sites from the extensive experimental literature, but for other organisms there is currently almost nothing known. A significant amount of work over the last years has shown that through a combination of phylogenetic footprinting, i.e. the identification of conserved sequence segments in orthologous intergenic regions, and motif finding methods, a substantial fraction of all regulatory sites can be recovered genome-wide. For example, by comparing the orthologous intergenic regions of proteo-gamma bacteria thousands of putative regulatory sites can be recovered in E. coli [53, 54, 55]. By clustering these according to similarity of their sequence motifs a large number of known and newly predicted regulons can be reconstructed [56]. In another example, a number of Saccharomyces species were sequenced and through phylogenetic footprinting conserved regulatory motifs were identified genome-wide [57, 58]. More recently ChIP-on-chip experiments were performed that identified which intergenic regions are bound by each of a large number of transcription factors from yeast [59] and using motif finding methods a first draft genome-wide annotation of regulatory sites was made for Saccharomyces cerevisiae [59]. Besides algorithms for finding regulatory sites in promoter regions of co-regulated genes, e.g. [60, 61], over the last years a number of computational methods have been developed for phylogenetic footprinting, e.g. [53, 54, 55, 62, 63], and more recently algorithms have been developed which combine general motif finding approaches with phylogenetic footprinting into an integrated frame work [64, 65, 66, 67]. Using these methods it was recently shown that one can dramatically improve the genome-wide annotation of regulatory sites in yeast [67, 68, 69]. With the development of more sophisticated tools for regulatory site identification, and with the large increase in the number of available fully-sequence bacterial genomes, it seems that the time is ripe

for computational approaches to comprehensively identify regulatory sites in bacterial genomes, including the large number of genomes for which currently virtually no regulatory sites are known.

Here we develop a new method that, starting from the genbank genome sequence files of a set of related organisms, automatically performs all the necessary steps for comprehensively identifying significantly conserved sequence segments in intergenic regions genome-wide. Our method maps orthologs between the genomes, reconstructs the phylogenetic tree relating the species, aligns the orthologous intergenic regions, and finally identifies all sequence segments that are significantly more conserved than could be expected given the phylogenetic relations of the species. We validate our method by comparing the predictions in a group of species related to E. coli with the locations of known binding sites [70]. We also apply our algorithm to 17 other groups of related bacteria and upload the results to the web-based database SwissRegulon [47].

4.2 Algorithm outline

Our algorithm for identifying conserved sequence segments genome-wide in outline consists of the following steps.

1. The input consists of the genome of a reference species, plus the genomes of a number of related species.
2. We first identify all orthologous pairs of genes between all pairs of species. We use an iterated procedure that identifies best reciprocal pairs based on their estimated evolutionary distance, re- constructs syntenic sets of such pairs, and iteratively "fills in" missing orthologs within the syntenic regions.
3. Using alignments of cliques of orthologous genes we determine the topology of the phylogenetic tree relating the species.
4. Using third positions of fourfold degenerate codons, and taking into account the different codon biases of different species, we determine the phylogenetic distances between all pairs of species.
5. We determine the branch lengths in the phylogenetic tree by fitting the pairwise distances to the tree topology.
6. For each intergenic region in the reference species we collect the orthologous intergenic regions from the other species and construct multiple alignments.
7. We scan all alignments for sequence segments that are significantly conserved using a probabilistic model that compares the probability of the alignment under a neutral background model with the probability of the alignment assuming that selection is constraining the evolution of the bases in the column.t compares the probability of the alignment under a neutral background model with the probability of the alignment assuming that selection is constraining the evolution of the bases in the column.

Before describing in details each of the steps above listed we want first to introduce the mathematical model it was used to formalize the evolution of DNA sequences. This evolutionary model is the central pillar on which the whole method relay on.

4.3 Evolutionary model

The molecular evolution of natural populations is an extraordinarily complex process, involving so many different confounding influences (e.g. mutational biases, epistatic interactions, heterogeneous recombination rates, population mixing patterns, temporal variations in population size,

time-dependent selection, frequency-dependent selection, and so on), that essentially all models of molecular evolution are not more than simple cartoons that focus on a few processes which are judged to be the most relevant. Consequently, there is a large variety of models and approaches to detecting natural selection from sequence data, see e.g. [71] for a review. Detecting sequence substitutions that are the result of adaptive evolution, i.e. that were positively selected, is especially challenging and typically requires the comparison of polymorphism data within one species with substitution data between closely-related species, see [72] for a recent review.

Here we are concerned with using conservation statistics of multiple alignments of orthologous DNA from related species to infer sites that are under purifying selection. A simple and robust approach to this problem is to compare conservation statistics of *pairwise* alignments of presumed 'neutral segments with the statistics of conservation in nearby segments that may contain constrained sites. This approach has for instance been applied to estimate the fraction of sites that are under purifying selection in intergenic DNA of Drosophila [73, 74]. In the context of bacterial genomes, a very similar approach has been used to extract putative regulatory sites in *E. coli* using pairwise alignments of orthologous intergenic regions from related species [55]. Such approaches can be generalized to the analysis of alignments of multiple species. Here the most commonly used approach is to introduce an explicit model of the substitution rates along the branches of the phylogenetic tree relating the species. Such models assign probabilities to multiple alignment columns in terms of the substitution rates and lengths of the branches [7]. Hidden Markov models are then used to segment multiple alignments into two (or a small number of) classes of sites [75, 76], i.e. those that evolve at slower overall rates and those that evolve at higher overall rates. Maximum likelihood is used to estimate the substitution rates in the different classes of sites. This approach has for example been used to estimate the fraction of DNA that is evolving slowly, presumably because of purifying selection, in the genomes of a substantial number of eukaryotes [77].

Here we are interested in estimating the density of conserved transcription factor binding sites (TFBSs) from multiple alignments of orthologous intergenic regions in bacteria. To do this we introduce two types of evolutionary models, a 'background model that describes the overall evolution of a category of sites (such as all sites in intergenic regions or all sites at third positions of a particular fourfold degenerate codon), and and a a 'foreground model describing the evolution of positions in regulatory sites, or more generally positions that evolve under substitution rates that are significantly different from those of the background model. We then use the likelihood-ratio of the 'foreground and 'background model for each alignment column to quantify the evidence that the position is part of a regulatory site.

Binding sites for a given TF are generally represented through position specific weight matrices $\mathbf{w}$ where w_α^i denotes the fraction of regulatory sites (for the TF in question) having nucleotide α at position i. Biophysical models of TFs binding to their target sites show [78, 79, 80] that, to a good approximation, the total binding free energy of a TF to a binding site is the sum of independent binding energies from each nucleotide in the site. In addition, the binding energy E_α^i of nucleotide α at position i is, to a reasonable approximation, proportional to the logarithm $\log(w_\alpha^i)$ of the frequency w_α^i of α at position i. Because the binding energies E_α^i vary significantly, with both the identity of the preferred nucleotides and the strength of the preference varying from position to position, one generally cannot assume uniform substitution rates across different positions in TFBSs. Indeed, studies of the evolution of known regulatory sites show that substitution rates vary significantly from position to position and in correspondence with the equilibrium frequencies w_α^i [81, 82, 80].

We thus felt it to be essential that our model for the evolution of TFBSs takes into account that both the preferred nucleotides and the strength of the preference vary from position to position. Our model assumes that different positions in regulatory sites evolve independently from each other. Since selection most likely acts on the binding energy of the entire site to the TF, this assumption is only an approximation, as stressed in [80]. However, the fact that different positions in known TFBSs show only marginal correlation indicates that this approximation is fairly accurate,

and indeed this approximation is followed by virtually all currently used models of regulatory site evolution. For each position i in a regulatory site we assume there is a (generally unknown) set of 4 selection coefficients for the possible nucleotides at this position, which are constant through time and, in the limit of large time, lead to equilibrium frequencies w^i_α. Following Golding and Felsenstein [83] Halpern and Bruno [84] have shown that, in the weak mutation limit of the standard Kimura-Ohta theory, one can uniquely determine substitution rates in terms of the mutation rates and the equilibrium frequencies w^i_α. In particular, if $r^i_{\alpha\beta}$ is the rate of substitution from β to α at position i, $\mu_{\alpha\beta}$ the rate of mutation from β to α, and w^i_α the equilibrium frequency of α at this position, we have [84]

$$r^i_{\alpha\beta} = \mu_{\alpha\beta} \frac{\log\left[\frac{\mu_{\beta\alpha} w^i_\alpha}{\mu_{\alpha\beta} w^i_\beta}\right]}{1 - \frac{\mu_{\alpha\beta} w^i_\beta}{\mu_{\beta\alpha} w^i_\alpha}}. \tag{4.1}$$

Under the Halpern-Bruno (HB) model, the probability to evolve from nucleotide β in the ancestor to nucleotide α in the descendant over the course of a time t is then given by

$$P_{\mathrm{HB}}(\alpha|\beta, \mu, w^i, t) = \left(e^{\mathbf{r}^i t}\right)_{\alpha\beta}, \tag{4.2}$$

where μ denotes the matrix of mutation rates, w^i denotes the vector of equilibrium frequencies at position i, and $\mathbf{r}^i$ the matrix of substitution rates at this position. The matrix exponential $e^{\mathbf{r}^i t}$ is generally calculated by (numerically) diagonalizing the matrix $\mathbf{r}^i$.

Given the transition probabilities (4.2) and given a phylogenetic tree T, one can then calculate the likelihood $L_{\mathrm{HB}}(C|w, \mu, T)$ for an alignment column C. Formally the likelihood is the product over transition probabilities $P_{\mathrm{HB}}(\alpha|\beta, \mu, w^i, t)$ for each branch of the tree, summed over all possible nucleotides for the internal nodes, and can be calculated efficiently using the recursive algorithm introduced by Felsenstein [7]. This calculation requires, however, that we know the mutation matrix μ and the equilibrium frequencies w^i_α. In some situations, these quantities may indeed be known. For example, for a given TF one can determine the equilibrium frequencies w^i_α from collections of known binding sites and one can then use the model with substitution rates (4.1) to identify conserved binding sites for the TF in multiple alignments of intergenic regions. This approach has been implemented by the MONKEY algorithm [81]. In our situation, however, the equilibrium frequencies w^i_α are intrinsically unknown. The rigorous Bayesian solution in this situation is to treat the equilibrium frequencies as nuisance parameters that need to be integrated out of the likelihood. That is, given a prior probability distribution $P(w)$ over possible equilibrium frequencies, we would calculate

$$L_{\mathrm{HB}}(C|\mu, T) = \int L_{\mathrm{HB}}(C|w, \mu, T) P(w) dw, \tag{4.3}$$

where the integral is over all vectors w such that $w_\alpha \geq 0$ for all α, and $\sum_\alpha w_\alpha = 1$. Unfortunately, because of the complicated dependence of the rates $r_{\alpha\beta}$ on the equilibrium frequencies w, these integrals are generally intractable. If the likelihood were sharply peaked as a function of w we could approximate the integral by the value at its peak and a correction factor such as the Bayesian Information Criterion [85]. However, since in our case the 'data consists of only a single alignment column C with nucleotides from typically a handful of species, the likelihood function is typically not sharply peaked so that such approximations are not suitable.

We thus sought to approximate the Halpern-Bruno model with a simpler model for which the integral (4.3) *can* be performed and that maintains the feature that selection coefficients (and correspondingly the limit frequencies w^i_α) can vary from position to position in regulatory sites. This can be achieved by using the following substitution rate model introduced by Felsenstein [7]

$$r^i_{\alpha\beta} = \mu w^i_\alpha, \tag{4.4}$$

also known as the F81 model. The F81 model makes the simplification that the substitution rate is dependent only on the identity of the target base. In addition, whereas the HB model explicitly separates the effects of mutation rate biases and selection on the equilibrium frequencies, the F81 model parametrizes the overall mutation rate by a single parameter μ and subsumes the effect of mutational biases and position-dependent selection into the position-dependent equilibrium frequencies w_α^i. Alternatively, one can think of the F81 model as assuming equal rates of all mutations and assuming that, at position i, the probability of a mutation to base α has a probability w_α^i to be fixed in the population. Under the F81 model the probability $P(\alpha|\beta, t, w)$ to evolve from ancestral base β to offspring base α over a time t is

$$P_{\mathrm{F81}}(\alpha|\beta, q, w) = e^{-\mu t}\delta_{\alpha\beta} + (1 - e^{-\mu t})w_\alpha. \tag{4.5}$$

In spite of the conceptual differences between the HB and F81 models in practice the transition probabilities of the HB and F81 models are typically not very different numerically. The F81 model we use here has been successfully applied in a number of algorithms [67, 66, 86] for regulatory motif finding in alignments of orthologous intergenic DNA.

To calculate the likelihood $L_{\mathrm{F81}}(C|\mu, T)$ of an alignment column C we now need to calculate the integral

$$L_{\mathrm{F81}}(C|\mu, T) = \int L_{\mathrm{F81}}(C|w, \mu, T)P(w)dw. \tag{4.6}$$

For the prior we use standard Dirichlet priors of the form

$$P(w) \propto \prod_\alpha (w_\alpha)^{\lambda_\alpha - 1}, \tag{4.7}$$

with the λ_α being the so-called pseudocounts. Since the likelihood $L_{\mathrm{F81}}(C|w, \mu, T)$ is simply a polynomial in the equilibrium frequencies w_α, we can perform the integral term by term using the general identity

$$\int \prod_\alpha (w_\alpha)^{n_\alpha - 1} dw = \frac{\prod_\alpha \Gamma(n_\alpha)}{\Gamma(\sum_\alpha n_\alpha)}. \tag{4.8}$$

In summary, in order to incorporate the fact that in regulatory sites the selection coefficients vary significantly from position to position we used a simplified version of the general Halpern-Bruno model, i.e. the F81 model, to calculate the likelihood $L_{\mathrm{F81}}(C|\mu, T)$ of any alignment column C as a function of the mutation rate μ and phylogenetic tree T.

4.4 Mapping Orthologs

Our procedure for mapping orthologs modifies the standard "best-reciprocal hit" procedure to be both conservative and take advantage of the significant amount of gene-order conservation between the closely related species. For each pair of organisms in a clade we estimate the evolutionary distances between each pair of genes using PAML [87], i.e. as in [88]. An initial set of "trusted orthologous pairs" is constructed by taking only those best-reciprocal hits that align over 50% of both proteins and for which the evolutionary distance of the second best hit is at least twice the evolutionary distance of the best hit. We then resolve additional ortholog relations by making use of gene-order information. We first construct diagonals of trusted pairs that are consecutive in both genomes and search for additional orthologous pairs that lie within the gaps or at the edges of the diagonals of already identified orthologs.

The detailed process is as follow: first we collect the list of all (predicted) protein sequences for each genome from the corresponding genbank file. A list of putative orthologs for each pair of genomes is obtained by running WU-BLAST [89]. As shown in [88], ortholog identification becomes more accurate if evolutionary distances, estimated by maximum likelihood, are used

instead of BLAST scores. Thus, for each reported hit, we globally align the corresponding pair of proteins using CLUSTALW [90]. To avoid mistaking single domain matches for orthologs we only retain alignments that cover at least 50% of both proteins. We estimate the evolutionary distance d of the pair using PAML [87] and assign a score $H = -\log(d)$ to the pair. Then, we number all the proteins in both genomes according to their position on the chromosome and identify orthologs by the following iterative procedure:

1. A pair of genes (α, β) are considered orthologs, which we will denote as $\alpha \backsim \beta$, if they are best reciprocal hits, and there is no other hit with a score larger than a fraction f of the score of the pair. That is $\alpha \backsim \beta$ if $H_{\alpha j} < f H_{\alpha\beta}$ for all $j \neq \beta$ and $H_{i\beta} < f H_{\alpha\beta}$ for all $i \neq \alpha$. We search for all pairs satisfying these conditions. After that, for each identified orthologous pair $\alpha \backsim \beta$ we set all scores $H_{\alpha j}$ and $H_{i\beta}$, i.e. hits to other proteins, to zero. We then repeat the search for orthologs until no more new orthologs are found.

2. We construct diagonals of consecutive or "anti-consecutive" pairs, i.e. runs of syntenic orthologous pairs of the form $\{\alpha \backsim \beta, (\alpha+1) \backsim (\beta+1), \ldots, (\alpha+n) \backsim (\beta+n)\}$ or $\{\alpha \backsim \beta, (\alpha+1) \backsim (\beta-1), ..., (\alpha+n) \backsim (\beta-n)\}$.

3. We now collect the set of pairs of proteins (i, j) that lie at the start or end of any of the syntenic runs of orthologs. Note that this includes all pairs of genes that lie in "gaps" between consecutive syntenic runs. We then perform the ortholog search (step 1) on only this subset of pairs.

4. When no more orthologs are found we identify the remaining set of best reciprocal pairs, i.e. no longer demanding that all other scores are less than a fraction f of the best reciprocal pair score.

We used $f = 0.5$, i.e. the score of the best pair should be twice as high as the next best pair. We find that, even for the sets of genomes of relatively closely related species that we work with, this procedure increases the number of orthologous pairs found by 10% or more over just using best reciprocal hits.

4.5 Reconstructing the phylogenetic tree

Estimating the tree topology

Having determined all the pairwise orthologous relations for all pairs of genomes in a clade, it is straightforward to find cliques of orthologs. A 'clique of orthologs is a set of genes, one from each species in the clade, that all are mutually orthologous. First, we assign to each gene an n-dimensional vector (with n the number of genomes in the clade) where the ith entry in the vector is the identity of the ortholog in the ith genome (if i is the genome from which the gene itself stems, then the entry is the identity of the gene itself). For each genome we produce a list of such vectors. Cliques are identified as those vectors that occur in the list of vectors of all genomes.

We align the DNA sequences of all orthologous cliques using T-coffee [91]. For each alignment we identify all third positions in codons of the sequence of the reference species and check what fraction of the aligned bases from the other species in the clade is conserved. In this way a conservation statistic is assigned to each multiple alignment. For each clade we sort all multiple alignments by this conservation statistic and remove the top 10% and bottom 10% of the multiple alignments. These 'outliers' will not be used for our parameter estimation. This is done to avoid that outliers, such as the ribosomal genes that are significantly more conserved at silent positions than other genes, or genes whose orthologs have been misidentified, would skew the parameters of the background models. For each clade we concatenated the remaining 80% of the multiple alignments and let the TREE-PUZZLE algorithm [92] determine a phylogenetic tree from this concatenated alignment.

Estimating pairwise species distances

The likelihoods of foreground and background models still depend on the product μt of overall mutation rate μ and branch length t, i.e. equation (5.1), for each branch of the tree. Note that since the likelihood depends only on the product μt we can set $\mu = 1$ without loss of generality. To estimate the branch lengths t for each branch of the tree we use third positions in fourfold degenerate codons to estimate distances between every pair of species in the clade. For each clade, and each pair of species in this clade, we start by collecting the data-set D of all pairwise aligned third positions in fourfold degenerate codons from the filtered set of cliques (excluding the first and last 20 amino acids in each protein). We use only those third positions for which the amino acid is conserved and count the number of times $n^c_{\alpha\beta}$ that base α occurs in the first species and base β in the other in codons of type c. Further let w^c_α and $\tilde{w}^c_\alpha$ denote the frequency of base α in codons of type c in the first species and second species respectively. We will approximate the probability to observe the pair of bases $\alpha\beta$ at a codon of type c by the average of the probabilities (under the F81 model) to start with base α in the first species and evolve to base β in the second and the probability to start with base β in the second species and evolve base α in the first. That is, the probability $P(\alpha|\beta, t, w^c)$ that the third position of a codon of type c will evolve from base β in the second genome to base α in the first genome assuming distance t between the genomes, is given by

$$P(\alpha|\beta, t, w^c) = \delta_{\alpha\beta} e^{-t} + (1 - e^{-t}) w^c_\alpha \tag{4.9}$$

where w^c_α is the fraction of all codons of type c in the first genome that have base α at the third position. Analogously, the probability to evolve from α in the first genome to β in the second genome is given by

$$P(\beta|\alpha, t, \tilde{w}^c) = \delta_{\alpha\beta} e^{-t} + (1 - e^{-t}) \tilde{w}^c_\beta. \tag{4.10}$$

We now approximate the probability to find the pair of bases $\alpha\beta$ at the third positions of a codon of type c in the alignment of two orthologous proteins from the two genomes as the average of $P(\alpha|\beta, t, w^c)\tilde{w}^c_\beta$ and $P(\beta|\alpha, t, \tilde{w}^c)w^c_\alpha$:

$$P(\alpha\beta|t, w^c, \tilde{w}^c) = \frac{(w^c_\alpha + \tilde{w}^c_\beta)}{2} \delta_{\alpha\beta} e^{-t} + (1 - e^{-t}) w^c_\alpha \tilde{w}^c_\beta. \tag{4.11}$$

Using this expression, the probability $P(D|t, w, \tilde{w})$ of the observed dataset D of counts $n^c_{\alpha\beta}$ is then given by

$$P(D|t, w, \tilde{w}) = \prod_{c,\alpha,\beta} \left[\frac{(w^c_\alpha + \tilde{w}^c_\alpha)}{2} \delta_{\alpha\beta} e^{-t} + (1 - e^{-t}) w^c_\alpha \tilde{w}^c_\beta \right]^{n^c_{\alpha\beta}}, \tag{4.12}$$

where the product is over all 8 fourfold degenerate codons c and 16 base combinations $\alpha\beta$. We determine the distance t of the pair of species by maximizing this expression with respect to t. We take the derivative of the logarithm of the expression (4.12) and set the result equal to zero. This leads to the following algebraic equation

$$N^{\text{diff}} = \sum_c \sum_\alpha \frac{N^{\text{cons}}_{c,\alpha} (W^c_\alpha - w^c_\alpha \tilde{w}^c_\alpha) e^{-t}}{e^{-t} W^c_\alpha + (1 - e^{-t}) w^c_\alpha \tilde{w}^c_\alpha} \tag{4.13}$$

where $N^{\text{cons}}_{c,\alpha}$ is the number of occurrences of a conserved pair $\alpha\alpha$ among codons of type c. N^{diff} is the total number of pairs with different bases in the two species, and

$$W^c_\alpha = \frac{w^c_\alpha + \tilde{w}^c_\alpha}{2}. \tag{4.14}$$

The equation above can be solved by standard numerical techniques since the expression on the right is a monotonically increasing function of t on the positive real axis.

Fitting the tree from the pairwise distances

As described above, we have already determined the topology of the phylogenetic tree for each clade. In addition, we have determined all pairwise distances t_{ij} between each pair of species (ij) for each clade. To determine the distance t_b on each of the branches b in each phylogenetic tree we use the standard least-square fitting with a fixed tree topology [93]. For completeness we describe the procedure here.

We find the set of distances t_b such that, for all pairs ij, the distances t_{ij} are best approximated by the total distance along the branches connecting i and j. That is, we minimize

$$F = \sum_{ij} \left(t_{ij} - \sum_{b \in \Pi_{ij}} t_b \right)^2 \tag{4.15}$$

where Π_{ij} is defined as the set of branches connecting nodes i and j of the tree. Taking the derivative with respect to t_b and setting it zero we obtain

$$0 = \sum_{ij} \delta(b \in \Pi_{ij}) \left[t_{ij} - \sum_{b' \in \Pi_{ij}} t_{b'} \right], \tag{4.16}$$

where $\delta(b \in \Pi_{ij})$ is 1 if branch b is an element of Π_{ij} and 0 otherwise. If we define the vector $\sum_{ij} \delta(b \in \Pi_{ij}) t_{ij} = A_b$, and the matrix $V_{bb'} = \sum_{ij} \delta(b \in \Pi_{ij}) \delta(b' \in \Pi_{ij})$, then equation (4.15) becomes

$$0 = A_b - \sum_{b'} V_{bb'} d_{b'} \Leftrightarrow d_b = \sum_{b'} \left(V^{-1} \right)_{bb'} A_{b'}. \tag{4.17}$$

Thus, the optimal set of branch lengths can be determined by a simple matrix inversion. Notice also that A_b is the sum of all pairwise distances between species that are connected through b and $V_{bb'}$ is the number of pairs of species ij for which the path that connects them passes through both b and b'.

4.6 Identification of segments under selection

For each gene in the reference species we extract the upstream region up to the previous gene. Then, we collect all the upstream regions of the orthologous genes from the other species of the clade. Finally, we construct multiple alignments of the sets of intergenic regions using T-COFFEE [91]. Having these alignments we identify segments that likely have evolved under purifying selection as follow:

Let C denote an alignment column for the species of the clade. The probability $P(C|\mathrm{bg})$ to observe alignment column C under the background evolution model is given by taking the product of (5.1) over all branches in the phylogenetic tree, and summing over the bases at all internal nodes:

$$P(C|\mathrm{bg}) = \sum_{\beta_i | i \in I} b_{\beta_r} \prod_{n \neq r} P(\beta_n | \beta_{a(n)}, t_b, b), \tag{4.18}$$

where β_i is the base at node i, I is the set of internal nodes of the tree, $a(n)$ is the ancestral node of node n, r is the root, and the product is over all nodes except for the root. Notice that we have replaced the vector of equilibrium frequencies w by the background base frequencies b. The sum over the bases at the internal nodes is calculated using the standard recursive method introduced by Felsenstein [7]. That is, let C^n_α denote the probability of the subtree rooted at node n, assuming that the base at node n was α. We then have the recursion relation

$$C^n_\alpha = \prod_{m \in c(n)} \left[\sum_\beta P(\beta|\alpha, t_m, w) C^m_\beta \right], \tag{4.19}$$

where the product is over all nodes m that are in the set of children $c(n)$ of node n, and t_m is the length of the branch leading from n to child m. Note that for leafs n we have

$$C_\alpha^n = \delta_{\alpha\alpha_n}, \tag{4.20}$$

with α_n the base at leaf n. Starting from the leafs we can determine the C_α^n at all nodes recursively. Once we have determined C_α^r of the root r we finally have

$$P(C|\text{bg}) = \sum_\alpha b_\alpha C_\alpha^r. \tag{4.21}$$

The foreground model is calculated by assuming that the nucleotide frequencies w are not given but *unknown*. That is, we integrate over all possible vectors of nucleotide frequencies:

$$P(C|\text{fg}) = \int P(C|w)P(w)dw, \tag{4.22}$$

where $P(C|w)$ is the exact same expression as (4.18) but with the specific vector of frequencies b replaced by the unknown vector of frequencies column w, and the integral is over the simplex $w_A + w_C + w_G + w_T = 1$. Finally, $P(w)$ gives the prior probability distribution that a foreground column will have frequency vectors w. We choose for $P(w)$ a Dirichlet prior, and we set the parameters of this prior to match the base composition in this class:

$$P(w) = \prod_\alpha \frac{(w_\alpha)^{w_\alpha - 1}}{\Gamma(w_\alpha)}. \tag{4.23}$$

Note that to calculate this integral we again have to sum over the bases at all internal nodes of the tree. Whereas each term in this sum can be integrated analytically using the general expression

$$\int \prod_\alpha (w_\alpha)^{n_\alpha - 1} dw = \frac{\prod_\alpha \Gamma(n_\alpha)}{\Gamma(\sum_\alpha n_\alpha)}, \tag{4.24}$$

there is no simple recursive way to calculate the sum and we are forced to sum all terms explicitly. However, we only need to do this once for each clade.

For each possible alignment column C we calculate the ratio $R(C)$ between the foreground and background model,

$$R(C) = \frac{P(C|\text{fg})}{P(C|\text{bg})}, \tag{4.25}$$

which quantifies the amount of evidence that column C is evolving according to a selection pressure different from the background model.

For a segment s of l contiguous columns the posterior probability $P(sel|s)$ that the segment have evolved under selection pressure is simply given in terms of the product of the scores $R(C_i)$ of the columns C_i in the segment:

$$P(sel|s) = \frac{\pi_{sel}\Pi_{i=1}^l R(C_i)}{\pi_{sel}\Pi_{i=1}^l R(C_i) + (1 - \pi_{sel})} \tag{4.26}$$

where π_{sel} is the prior probability that a randomly chosen segment of length l is under selection.

Finally, for a given segment length l our algorithm calculates the score of all segments in all intergenic regions. It then places windows of length l in the intergenic regions consecutively by at each step placing the window at the segment with highest score that does not overlap any of the windows placed previously. This process stops either when the highest scoring segments falls below a certain cut-off, or if no more windows can be placed without overlapping previously placed windows. At the end we have a genome-wide list of non-overlapping segments of length l ordered by their conservation score.

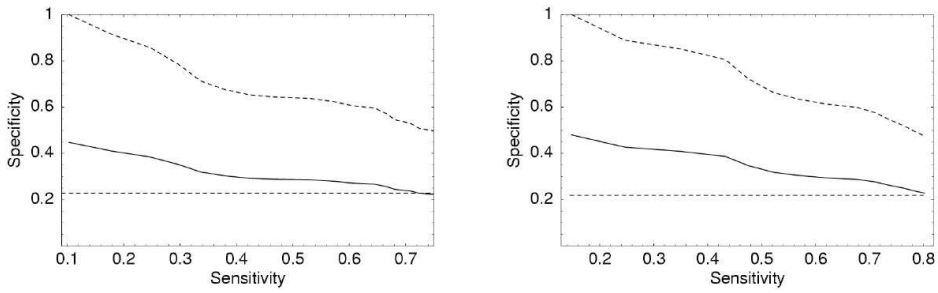

Figure 4.1: Comparison of predicted conserved segments with known binding sites in E. coli. For different cutoffs in segment score the horizontal axis shows the fraction of known sites that match predicted segments (sensitivity) and the vertical axis the fraction of predicted segments that match known sites (specificity). The left panel shows the results for all intergenic regions with at least one known site, and the right panel the results for only those regions for which an orthologous intergenic region exists in at least 3 other species.

4.7 Validation of E. coli predictions

As a first test of our algorithm we used it to predict conserved segments in E. coli and compared its predictions with known binding sites from the regulonDB database [70]. We took the *E. coli K12* genome as reference species and added the genomes of the related species *Salmonella typhi*, *Yersinia pestis KIM*, *Photorhabdus luminescens*, and *Photobacterium profundum SS9*. The phylogenetic tree that our algorithm estimated is shown in the appendix and the numbers of orthologous intergenic regions that were found in each of the species are 1935, 1401, 827 and 827 respectively. We predicted conserved sequence segments of length $L = 18$ genome-wide using a range of different cutoffs. At each cutoff we compared the predicted segments (those with score above the cutoff) with the known sites. We focused on the 481 intergenic regions for which at least one site is annotated in regulonDB and for which there is an ortholog in at least one other species. At our highest cutoff there were a total of 250 predicted segments in all 481 regions, and at the lowest cutoff (zero) there were 4263 predicted segments, i.e. almost nine per intergenic region. At each cutoff we calculated sensitivity and specificity as follows. Each predicted site that overlapped a known site by at least half of its length was considered a true positive. Similarly, each known site that overlapped a predicted site by half of its length was considered "predicted". We then calculated the fraction of all known sites that was predicted (sensitivity) and the fraction of all predictions that were true positives (specificity). The left panel of Fig. 4.1 shows the specificity and sensitivity for different cutoffs as the solid line. The horizontal dashed line corresponds to the fraction of all bases in these intergenic regions that correspond to known sites (about 0.22). Thus, at cutoff zero the specificity matches roughly what would be expected under random placement of segments. However, as the cutoff is increased the specificity of the predictions rises and when only the top 250 scoring windows are considered the specificity matches about 0.45, i.e. almost half of the 250 predictions match known sites. Since the set of known sites is almost certainly only a fraction of the total set of true sites the true specificity of our predictions is much higher. For example, if only half of all true sites are known, then the true specificity of the top 250 windows is not 0.45 but $0.45/0.5 = 0.9$. In fact, given the very strong evidence of conservation in the top segments, we believe it is not unreasonable to assume that essentially all of the top windows correspond to functional segments, i.e. that the known sites correspond to only 45% of all true functional segments. This is equivalent to assuming that the true density of functional positions is $0.22/0.45 = 0.49$, i.e. that half of the intergenic regions are under selection. The dashed diagonal line in the left panel shows the specificity of our predictions under this assumption. In summary,

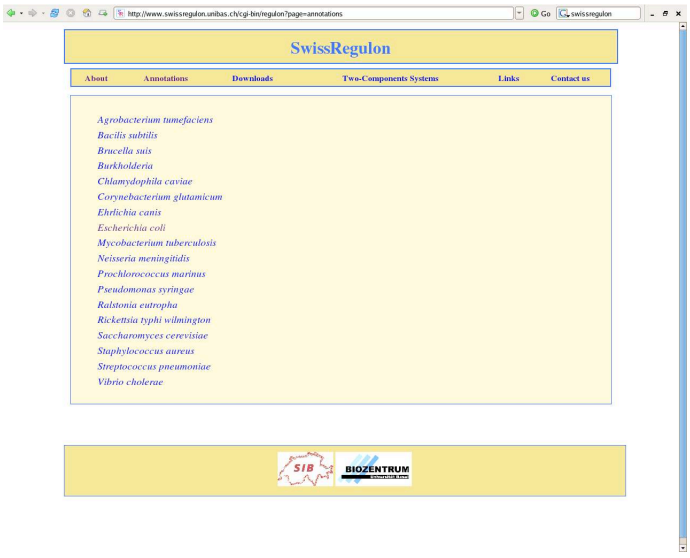

Figure 4.2: Screen shot of SwissRegulon web site

from the comparison with known sites we estimate that our method predicts almost 75% of all true sites at a specificity of about 50%, and 25% of all true sites at a specificity of over 80%. In the right hand panel of Fig. 4.1 we show the same results but now focusing on only the 231 intergenic regions that had at least one known site and for which there were orthologous intergenic regions in at least 3 of the 4 related species. We see that, as might be expected, for the regions for which more orthologs are available both the specificity and the sensitivity are increased, although the difference is fairly small.

4.8 Segments under selection are available in SwissRegulon

Note that for many bacterial genomes little or nothing is known about their regulatory elements. Here we have presented a novel method that allows us to focus in those concrete sequences within intergenic regions that have evolved under purifying selection pressure. This conserved regions are likely to contain regulatory elements as we have shown in the case of *E. coli* where known binding sites are available. Therefore, we have apply our method to detect segments that have evolved under selection on 17 different bacterial clades. The genomes analyzed span widely the whole bacterial phylogenetic tree. The results are available in SwissRegulon, a web-based database which contains genome-wide annotations of regulatory sites in the intergenic regions of genomes. Figures 4.2 and 4.3 show screen shots of the web site.

4.9 Discussion

Bacterial genomes contain on average between 200 and 300 bps of intergenic DNA per gene which are thought to contain the bulk of regulatory signals that encode the organism's regulatory networks, including transcription factor binding sites, translation control sites, and small regulatory RNAs. While the hundreds of available fully-sequenced genomes are well annotated for their protein coding genes, the regulatory signals in intergenic DNA remain largely unexplored. With the large number of fully-sequenced bacterial available, including many related species, and with recent advances in computational methods for phylogenetic footprinting both the data and the methods are available for identifying putative regulatory sites genome-wide using sets of related genomes.

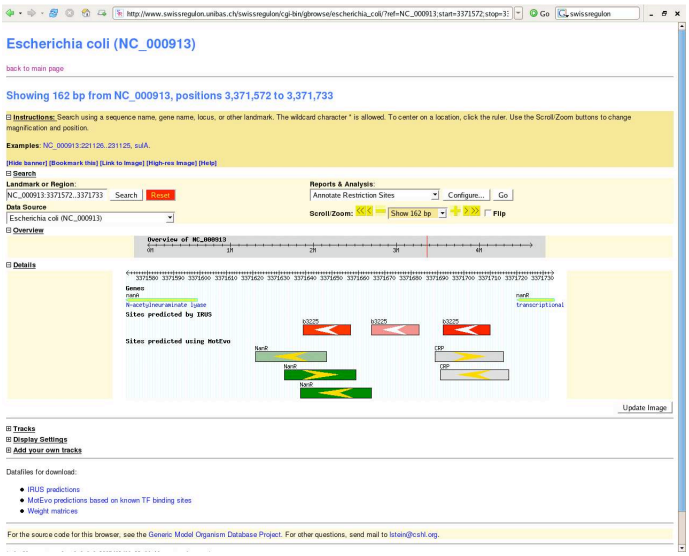

Figure 4.3: Screen shot of the SwissRegulon web site

To do this one needs to map all orthologs between the genomes in the set, reconstruct the phylogenetic tree relating the species, identify and align all orthologous intergenic regions, and search these alignments for segments that show evidence of selection. Here we have presented an integrated set of algorithms that accomplishes all these tasks in an automated fashion. Our methodology includes several novel features including an iterative mapping of orthologs in which at each iteration new orthologs are identified using synteny information of already mapped orthologs, and a novel method for inferring the phylogenetic tree. In the latter method we first identify the topology of the tree using standard methods, then identify all pairwise distances using third positions of fourfold degenerate codons while taking codon bias into account, and finally infer the branch lengths of the phylogenetic tree from the set of pairwise distances.

Finally, we also developed a novel procedure for identifying selected positions that takes the phylogenetic tree of the species rigorously into account. Together our methods allows us, starting simply with the genbank genome files of a set of related species, to comprehensively identify significantly conserved sequence segments in their intergenic regions. We believe that this set of methods can now form the basis for more sophisticated genome-wide annotation of regulatory sites in bacteria. For example, sets of significantly conserved regions could now be further searched using motif finding algorithms or one may search these conserved segments for motifs resembling known regulatory motifs. We also made an initial analysis of the large scale statistical features of our mappings of conserved sequence segments. We found that in all genomes analyzed a substantial fraction of the intergenic regions seems to be under selection. Finally, all the significantly segments of 22 bacterial clades that show a high degree of purifying selection where upload to the public-available database SwissRegulon.

5 Universal patterns of purifying selection at non-coding positions

We used the method previously described to comprehensively quantified evidence of purifying selection at non-coding positions across bacteria and found several striking patterns that are shared by all bacteria. Whereas most silent positions in genes show no deviation from the background evolution model, all intergenic regions show evidence of purifying selection. Consistent with selection acting at transcriptional regulatory elements we find most evidence of selection in upstream regions and a universal positional profile with respect to gene starts and ends, showing most selection immediately upstream and weakest immediately downstream of genes. Further universal features are a peak in purifying selection at ribosomal binding sites, and a pattern of high adenine frequency, significant selection at silent positions, and avoidance of RNA secondary structure concentrated in the areas immediately around translation starts. These features indicate that selection for translation initiation efficiency is the major determinant of the sequence composition around translation start in all clades.

5.1 Introduction

We briefly outline our procedure for quantifying evidence of purifying selection across non-coding positions genome-wide in bacterial genomes. The whole method is extensively described in chapter 4. Our procedure takes as input a set of related bacterial genomes (a *clade*) as provided by the NCBI microbial genome database [30], of which one is denoted as the *reference species*. For each gene and each intergenic region of the reference species we extract orthologous genes and intergenic regions from the other species and produce multiple alignments. We determine cliques of orthologous proteins (sets of genes that are all mutual orthologs between all species in the clade) and infer the topology of the phylogenetic tree from the concatenated alignment of all cliques.

For each alignment column we calculate the likelihood under two evolutionary models: a 'foreground' and a 'background' model. The background model assumes a simple F81 substitution rate model [7] which is parametrized by the branch lengths of the phylogenetic tree and a vector w of nucleotide frequencies, with w_α being the frequency of nucleotide α. In the F81 model the rate of substitution $r_{\alpha\beta}$ from base β to base α is simply proportional to w_α and independent of β. As nucleotide frequencies vary significantly between intergenic positions, coding positions, and third positions of four-fold degenerate codons, we separate positions into 12 different categories and construct a background model for each. The categories we distinguish are first, second, and third codon positions in genes, intergenic positions, and third positions in each of the 8 fourfold degenerate codons (silent positions). To estimate the parameters of the background models we determine the overall nucleotide frequencies w_α in each of the 12 categories of positions and fit the branch lengths of the phylogenetic tree from the alignments of silent positions using maximum likelihood. Our background models thus explicitly incorporates the overall nucleotide and codon biases of different classes of sites.

For each of the 12 background evolution models we have a corresponding foreground model. The only difference between the foreground and background model is that, whereas the background model assumes that all positions undergo substitutions from base β to base α at the same rate $r_{\alpha\beta} \propto w_\alpha$, in the foreground model we assume that at a given position i, the substitution rates $r^i_{\alpha\beta} \propto w^i_\alpha$ are altered due to specific selection preferences for certain bases at this position, which

are parametrized by the 'target' nucleotide frequencies w^i_α. Since the w^i_α at each position are unknown we treat them as nuisance parameters that are integrated out of the likelihood. Such evolutionary models have been used by several groups to model the evolution of positions in regulatory sites [67, 66, 86, 81]. The reason we use the simpler F81 substitution rate model rather than the related, but more general, Halpern-Bruno model [84] is that the necessary integrals over the unknown position-dependent frequencies w^i_α can only be performed for the F81 model.

For each alignment column of the reference species, both in genes and in intergenic regions, we calculate the ratio R of likelihoods of foreground and background evolutionary models. This statistic quantifies the evidence that the alignment column evolves under a different set of substitution rates than the background model. In addition, we estimate the effective substitution rate reduction Q relative to the substitution rate of the background model at each alignment column (see chapter 4). In practice we find that columns of high R (clear deviation from the background model) correspond to columns of high Q (low substitution rate). We thus interpret R and Q as quantifying the amount of purifying selection at each alignment column relative to the background model.

5.2 Quantifying evidence of purifying selection at non-coding positions

General derivation

Let c generally denote a class of positions. The background evolutionary model for positions within class c is given in terms of the base frequencies w^c_α for this class, and the branch lengths t_b for each branch b of the phylogenetic tree of the clade. Along a single branch of the tree, the probability to evolve from ancestral base β to descendant base α is given by

$$P(\alpha|\beta, t_b, w^c) = \delta_{\alpha\beta} e^{-t_b} + (1 - e^{-t_b}) w^c_\alpha. \tag{5.1}$$

Let C denote an alignment column for the species of the clade. The probability $P(C|\mathrm{bg}, c)$ to observe alignment column C under the background evolution model of class c is given by taking the product of (5.1) over all branches in the phylogenetic tree, and summing over the bases at all internal nodes:

$$P(C|\mathrm{bg}, c) = \sum_{\beta_i | i \in I} w_{\beta_r} \prod_{n \neq r} P(\beta_n|\beta_{a(n)}, t_b, w^c), \tag{5.2}$$

where β_i is the base at node i, I is the set of internal nodes of the tree, $a(n)$ is the ancestral node of node n, r is the root, and the product is over all nodes except for the root. The sum over the bases at the internal nodes is calculated using the standard recursive method introduced by Felsenstein [7]. That is, let C^n_α denote the probability of the subtree rooted at node n, assuming that the base at node n was α. We then have the recursion relation

$$C^n_\alpha = \prod_{m \in c(n)} \left[\sum_\beta P(\beta|\alpha, t_m, w^c) C^m_\beta \right], \tag{5.3}$$

where the product is over all nodes m that are in the set of children $c(n)$ of node n, and t_m is the length of the branch leading from n to child m. Note that for leafs n we have

$$C^n_\alpha = \delta_{\alpha\alpha_n}, \tag{5.4}$$

with α_n the base at leaf n. Starting from the leafs we can determine the C^n_α at all nodes recursively. Once we have determined C^r_α of the root r we finally have

$$P(C|\mathrm{bg}, c) = \sum_\alpha w^c_\alpha C^r_\alpha. \tag{5.5}$$

For each class c the foreground model is calculated by assuming that the nucleotide frequencies w are not given but *unknown*. That is, we integrate over all possible vectors of nucleotide frequencies:

$$P(C|\text{fg}, c) = \int P(C|w)P(w|c)dw, \tag{5.6}$$

where $P(C|w)$ is the exact same expression as (5.2) but with the class specific vector of frequencies w^c replaced by the unknown vector of frequencies column w, and the integral is over the simplex $w_A + w_C + w_G + w_T = 1$. Finally, $P(w|c)$ gives the prior probability distribution that a foreground column in class c will have frequency vectors w. We choose for $P(w|c)$ a Dirichlet prior, and we set the parameters of this prior to match the base composition in this class:

$$P(w|c) = \prod_\alpha \frac{(w_\alpha)^{w_\alpha^c - 1}}{\Gamma(w_\alpha^c)}. \tag{5.7}$$

Note that to calculate this integral we again have to sum over the bases at all internal nodes of the tree. Whereas each term in this sum can be integrated analytically using the general expression

$$\int \prod_\alpha (w_\alpha)^{n_\alpha - 1} dw = \frac{\prod_\alpha \Gamma(n_\alpha)}{\Gamma(\sum_\alpha n_\alpha)}, \tag{5.8}$$

there is no simple recursive way to calculate the sum and we are forced to sum all terms explicitly. However, we only need to do this once for each clade.

For each class c and each possible alignment column C we calculate the ratio $R(C|c)$ between the foreground and background model for this class,

$$R(C|c) = \frac{P(C|\text{fg}, c)}{P(C|\text{bg}, c)}, \tag{5.9}$$

which quantifies the amount of evidence that column C is evolving according to a selection pressure different from the background model for this class. Finally, we analyze the evidence of selection in different groups of non-coding positions by calculating the average value of $R(C|c)$ for different groups of positions. In particular, we determine the average value of R in different types of intergenic regions, the average value of R within different classes of positions within genes, and the average value of R at a given locations relative to the start and stop codons of genes.

Background evolution models

Our evolutionary model for regulatory sites thus assumes an F81 substitution rate model with independent equilibrium frequencies w_α at each position, which are treated as unknown nuisance parameters that are integrated out of the likelihood. We contrast this 'foreground' model with a 'background' model which is exactly the same, except that the equilibrium frequencies w_α are not assumed unknown and varying from position to position, but rather they are assumed the same at each position and are estimated from the overall nucleotide frequencies. It is clear, however, that using a single background model for all non-coding positions is not appropriate. One generally finds significantly higher AT-content in intergenic regions than in genes and, moreover, different fourfold-degenerate codons show significantly different frequencies of the nucleotide in their third position. We thus introduce separate background models for intergenic positions and for each of the 8 fourfold degenerate codons. To compare the likelihood-ratios between foreground and background models at non-coding positions with those at coding positions we also introduce background models for first, second, and third positions in codons in general.

For each background model c we need to determine the vector of equilibrium frequencies w^c, with w_α^c the frequency of base α in class c. To estimate the equilibrium frequencies we average

over all organisms in the clade. Base frequencies in intergenic regions are determined from all intergenic regions in all genomes in the clade. For the coding positions and the silent positions for the 8 fourfold degenerate codons we used all the remaining orthologous cliques, but have excluded the first and last 20 amino acids in each clique. The latter is done because, as our results show, there are significant deviations in base composition at the starts and ends of genes.

Finally, for each of the 12 classes of sites we set the pseudo-counts in the prior (4.7) for the foreground model equal to the estimated nucleotide frequencies of the background model in the corresponding class, i.e. $\lambda_\alpha = w_\alpha$. As shown in the next subsection, this guarantees that in the limit of very short branch length $t \to 0$, the foreground and background model will obtain the same likelihood.

R values in the limit of $t \to 0$

Imagine an alignment column for only two species that are so closely-related that their phylogenetic distance is essentially zero, i.e. $t \approx 0$, and that all nucleotides are conserved. Obviously there is no useful conservation information whatsoever in this alignment and as a consequence our R statistic should be equal at all alignment columns and not indicate any evidence of the foreground over the background model, i.e. we should have $R = 1$ for all alignment columns.

Let's assume a given alignment column of class c has α in both sequences. Under the background evolution model the probability of this data is just given by the frequency w_α^c of nucleotide α in this class. Assume that for the foreground evolution model we integrate over w with Dirichlet prior

$$P(w) = \Gamma(\lambda) \prod_\alpha \frac{(w_\alpha)^{\lambda_\alpha - 1}}{\Gamma(\lambda_\alpha)}, \tag{5.10}$$

where the λ_α are the pseudocounts of the prior and $\lambda = \sum_\alpha \lambda_\alpha$. A simple calculation shows that the probability of an alignment column with nucleotide α in both species (at distance $t = 0$) under this foreground model is given by λ_α/λ. Therefore we find that $R = 1$ if and only if $\lambda_\alpha \propto w_\alpha^c$. That is, the pseudocounts of the prior should be proportional to the overall frequencies w_α^c of the background model. This leaves the overall scale λ free to determine. The overall scale λ sets the expected bias for columns in the foreground model with $\lambda = 4$ corresponding roughly to a uniform prior. We set $\lambda = 1$ which corresponds roughly to the bias observed in known regulatory sites in *E. coli*.

R values for positions evolving according to the background model

For the bulk of silent positions in proteins we observe an average R value of $R = 1$. Here we show that this suggests that these positions evolve according to the background model. Assume that, for a certain class of positions, a fraction $f(C)$ show alignment column C. The average R value in these positions is then given by

$$\langle R \rangle = \sum_C f(C) \frac{P(C|\text{fg})}{P(C|\text{bg})}. \tag{5.11}$$

If the positions in this set are evolving according to the background model we have

$$f(C) = P(C|\text{bg}). \tag{5.12}$$

Therefore, we have

$$\langle R \rangle = \sum_C f(C) \frac{P(C|\text{fg})}{P(C|\text{bg})} = \sum_C P(C|\text{bg}) \frac{P(C|\text{fg})}{P(C|\text{bg})} = 1, \tag{5.13}$$

where the last equality follows because the foreground distribution $P(C|\text{fg})$ is of course also normalized. In summary, the fact that $R = 1$ on average at silent positions suggests that the fractions $f(C)$ at these positions are close to the background model frequencies $P(C|\text{bg})$.

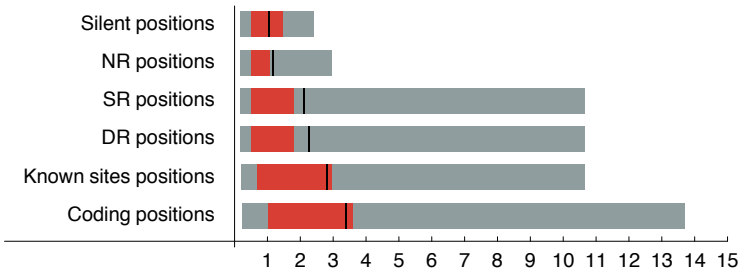

Figure 5.1: Distributions of R values in different classes of positions in *E. coli*. For each category of positions the black line denotes the average R value, the red bar the 25 to 75 percentile, and the grey bar the 5 to 95 percentile.

5.3 Multiple alignments of syntenic regions

When calculating average R values across intergenic regions and when calculating R values as a function of their position with respect to the starts and ends of genes, we want to make sure not to erroneously align intergenic regions that have undergone rearrangement since the species diverged from their common ancestor. To do this we consider an intergenic region in a given species orthologous to an intergenic region in the reference species only if the genes at both ends are orthologous and their orientation is conserved.

Let X denote an intergenic region in the reference species, let g_l and g_r denote the genes in the reference species at the left and right end of the intergenic region X, and let o_l and o_r be the orientations of the genes g_l and g_r, i.e. $o_l = 1$ means gene g_l is on the plus strand and $o_l = -1$ means it is on the negative strand. Similarly, let $\tilde{X}$ denote an intergenic region in another species of the clade with $\tilde{g}_l$ and $\tilde{g}_r$ the genes on the right and left of $\tilde{X}$, and $\tilde{o}_l$ and $\tilde{o}_r$ their orientations. The regions X and $\tilde{X}$ are considered orthologous if one of the following two sets of conditions holds

1. g_l is orthologous to $\tilde{g}_l$, g_r is orthologous to $\tilde{g}_r$, $o_l = \tilde{o}_l$, and $o_r = \tilde{o}_r$.

2. g_l is orthologous to $\tilde{g}_r$, g_r is orthologous to $\tilde{g}_l$, $o_l = -\tilde{o}_r$, and $o_r = -\tilde{o}_l$.

For each intergenic region X in the reference species we collect all orthologous intergenic regions in the other species of the clade. We then extracted, from each species, the DNA sequences of the intergenic region *plus* the two flanking genes. This set of sequences was then aligned with the T-Coffee algorithm [91] using default parameters.

Finally, we classified the intergenic regions into 3 different types: non-regulatory (NR) regions that are downstream of two convergently transcribed genes, single-regulatory (SR) regions upstream of the first gene in an operon and downstream of another gene, and double-regulatory (DR) regions that lie between two divergently transcribed genes. For the calculation of average R values across intergenic regions of different type we only considered intergenic regions that were at least 50 bp wide.

5.4 Distribution of R values in different regions of E. coli

To investigate the ability of our R statistic to detect positions in TFBSs we focused on *E. coli* for which a large collection of experimentally determined TFBSs is available [70]. Figure 5.1 summarizes the distribution of R values at silent sites, sites in NR regions, SR regions, DR regions, positions in known TFBSs, and sites in coding regions. Silent positions in *E. coli* have an average R close to 1 which suggests that most silent positions evolve according to their background model (see section 5.2). Sites downstream of genes (in NR regions) also typically have small R values. On

the contrary, sites upstream of genes (SR and DR regions) show significantly higher R values. The 25 percentile occurs at similarly low values of R for silent, NR, SR and DR positions, indicating that there is a significant fraction of positions in upstream regions that are not under purifying selection. In contrast, the 75 and 95 percentiles are shifted significantly upwards for SR and DR regions, indicating that a substantial number of positions in SR and DR regions are under purifying selection. This is also evident from the fact that the average for SR and DR is above the 75 percentile. Known TFBSs show even larger average R values than upstream positions in general, and both the 25 and 75 percentile are shifted upwards with respect to SR and DR regions. Nonetheless, not all positions in known sites show large R values, which is to be expected, since many TFBSs have internal spacers that are presumably not under purifying selection. Finally, positions in coding regions show the largest R values with about 75% of all positions having an R larger than 1. In summary, the results in Fig. 5.1 show that the R statistic clearly detects purifying selection at coding positions, that upstream regions show increased purifying selection compared to downstream and silent positions, and that known binding sites are characterized by elevated R values whose average nears the average R at coding positions.

5.5 Purifying selection at different types of non-coding positions

Comparing R values across clades

We next turned to comparing R values between silent positions, intergenic positions, and coding positions across all 22 clades. For each clade we averaged the R values of sites at silent positions, at positions in NR regions, in SR regions, in DR regions, and at coding positions (Figure 5.2). We see that, in all clades, silent positions appear to evolve according to the background model, i.e. R is close to 1. Note that the fact that $R = 1$ at silent positions does not necessarily mean that there is no purifying selection at third positions, but it does imply that the selection which may exist at silent positions is accurately captured by the overall codon bias which is incorporated into the background model. In contrast, all intergenic regions show evidence for purifying selection deviating from the background model (which incorporates the overall nucleotide bias in intergenic regions). Even for NR regions downstream of genes there is some evidence for purifying selection deviating from the background model, i.e. most dots in the top-left panel occur to the right of $R = 1$. The same panel also shows that SR regions always show more evidence of purifying selection than NR regions, i.e. all red dots are above the diagonal. The top-right panel shows that DR regions generally exhibit more evidence of purifying selection than SR regions, i.e. most green dots are above the diagonal. The bottom-left panel demonstrates that, for all clades, the purifying selection at coding positions is still significantly larger than that in DR regions, i.e. all blue dots are above the diagonal. In summary these three panels show that our observations from *E. coli* generalize to all clades. This universal order in average R values (largest in DR, followed by SR, then NR, and $R = 1$ at silent positions) strongly suggests that conserved regulatory elements occur in the upstream regions of all clades and are responsible for the observed increase in average R.

Estimating branch lengths with PAML

Our model makes various simplifying assumptions that might affect our results, e.g. it ignores transition-transversion bias. To check the robustness of our results we performed an analogous analysis using a completely different method. For each region type (NR, SR, DR, coding, silent) we extracted all alignment columns. Each set of alignment columns was then concatenated into a pseudo-alignment of all positions in regions of that type. These pseudo-alignments were then given as input to the PAML program [87], which performed a maximum likelihood inference of

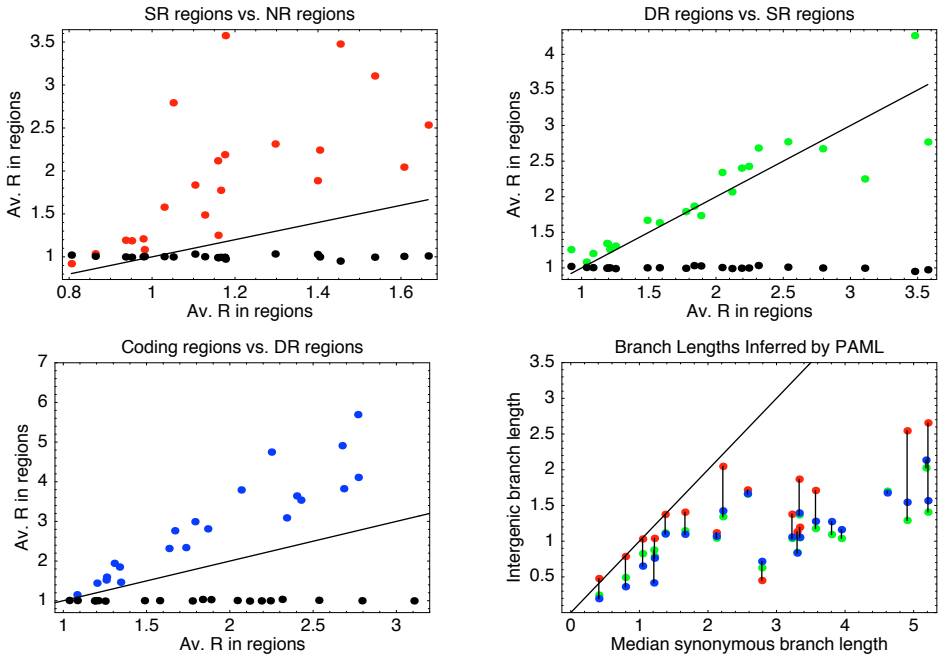

Figure 5.2: Comparison of average R values in different regions for 22 clades of bacteria. The red dots in the top-left panel show the average R in SR regions (vertical axis) against the average R in NR regions (horizontal axis). The green dots in the top-right panel show the average R in DR regions (vertical) against the average R in SR regions (horizontal). The blue dots in the bottom-left panel show the average R in coding positions (vertical) against the average R in DR regions (horizontal). The black dots in all panels show the average R in silent positions (vertical). The line $y = x$ is also shown in all panels. The bottom right panel shows, for each clade, the total branch lengths in the phylogenetic trees as inferred by PAML on alignment columns from NR (red), SR (green), and DR (blue) regions, as a function of the total branch length in the phylogenetic tree inferred from the silent positions (horizontal). Dots corresponding to the same clade are connected by vertical lines.

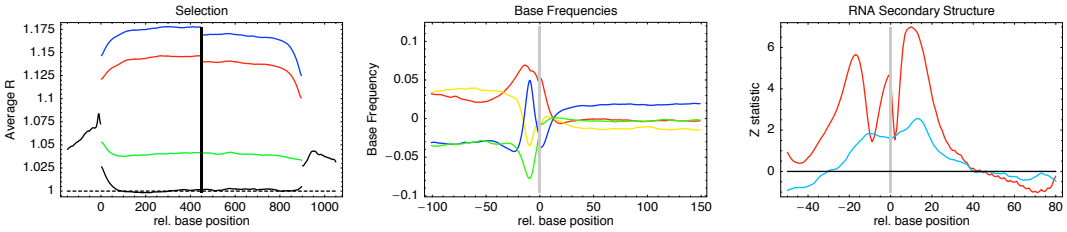

Figure 5.3: Universal position-dependent profiles in purifying selection, base frequencies, and secondary structure as a function of position relative to translation start (position 0) and, in the left panel, stop (position 900). Statistics are averaged over all 22 clades in each panel. Left panel: Average R value profile for first (red), second (blue), and third (green) positions in codons as well as intergenic/silent positions (black). Middle panel: Relative base frequencies around translation start, i.e. position-dependent frequencies of A (red), C (green), G (blue),and T (yellow) nucleotides relative to their genome-wide frequencies. Right panel: z-statistics for the probability of a given position to be unpaired relative to the average over the regions (-50,-31) and (31,80) (red) and relative to synthetic sequences with the same base composition (blue).

the branch lengths of the phylogenetic tree of each pseudo-alignment using a HKY85 evolutionary model [94]. We then run PAML with the following set of options:

1. Evolutionary model: HKY85 which treats transition and transversion events separately.
2. Kappa: we allow the program to estimate the ratio between transitions and transversions kappa.
3. Clock: we do not assume a molecular clock.
4. Alpha: we set parameter alpha to zero, which means that the rate of mutation is assumed constant across sites.

We then compared the branch lengths of the phylogenetic trees that PAML inferred for each region type. The bottom-right panel shows the total branch length of the tree inferred by PAML from the pseudo-alignments of positions in NR (red), SR (green), and DR (blue) regions, as a function of the the total branch length of the tree inferred from the silent positions, together with the diagonal $y = x$. The more purifying selection acts to conserve positions in regions of a given type, the shorter the inferred branch lengths will be. The PAML results agree with the results in the three other panels of Fig. 5.2: in essentially all clades the inferred distance in all types of intergenic regions is lower than that in silent positions, i.e. there is evidence of purifying selection acting in all three types of intergenic regions. Also, SR and DR regions have always more evidence of purifying selection than NR regions. In contrast to the results we obtained with our R statistic, the PAML results do not show a consistent ordering of the inferred branch lengths for the DR and SR regions.

5.6 Purifying selection profiles relative to gene starts and ends

To gain further insight in the selection patterns across bacteria we calculated the average value of R as a function of the relative position of the alignment column with respect to the start and stop codons of genes. The left panel of Fig. 5.3 shows this position-dependent selection profile averaged over all 22 clades.

Strikingly, the main characteristics of this profile are shared across all 22 clades (see figures 5.4 and 5.5). As shown in figure 5.2, the highest R values are observed for those positions that most often affect the amino acid sequence, i.e. in order: second (blue), first (red), and third (green) codon positions. Interestingly, whereas there is a clear drop in R at first and second positions near the starts and ends of the genes, at the third positions there is an *increase* in R near the starts of genes. Intergenic regions show clear evidence of purifying selection ($R > 1$) with R significantly higher upstream of genes than downstream, though selection is lower than at coding positions. Consistent with a pattern in which regulatory elements are most common near the starts of genes we find that R values are highest near the translation start and fall off progressively further upstream. In contrast to the coding positions and intergenic regions, the bulk of the silent positions seems to evolve according to the background model, i.e. $R = 1$.

The R value profiles in addition show a number of universal features that, as we will argue in sections 5.10 and 5.11, relate to efficiency and regulation of translation initiation. First, we find a sharp peak in R just upstream of translation start which is accompanied by a sharp peak in the frequency of guanines (middle panel of Fig. 5.3). Closer inspection shows that this peak corresponds to highly-conserved Shine-Dalgarno sequences [95] to which the ribosome binds. As shown in the section 5.10, although varying significantly in strength between clades, this Shine-Dalgarno peak is found in essentially all clades. In addition, 20 of the 22 clades show a sharp peak in G nucleotide frequency at this position matching the known Shine-Dalgarno consensus. Interestingly, this peak in G nucleotides is absent in the two clades of Cyanobacteria where instead a peak in C nucleotides is observed. The R statistic thus detects universally occurring purifying selection at ribosome binding sites.

In addition, in essentially all bacterial clades (see figures 5.4 and 5.5), R rises sharply at silent positions immediately downstream of the ATG and this heightened selection is accompanied by an increase in the frequency of adenines which extends into the upstream region. This rise in R is not caused by an increase of codon bias at gene starts, nor is it caused by misannotation of start codon as we will see later. In fact, an increase in adenine frequency around the start codon, accompanied by elevated conservation at silent positions immediately downstream, has been observed previously, i.e. [96] observed this pattern in *E. coli* and suggested that it is the result of selection for the avoidance of RNA secondary structure in this area of the mRNA, which in turn is the result of selection for translation initiation efficiency. In *B. subtilis* the same pattern was observed, accompanied by reduced secondary structure in this area [97]. Moreover, experimental studies showed that increasing the frequency of A nucleotides immediately following translation start increases translation efficiency [98, 99, 100].

Profiles of R values for all clades

The R value profiles, as shown in figures 5.4 and 5.5 were calculated from the gene-intergenic-gene multiple alignments just described. To obtain the profiles we need to calculate the average R value of alignment columns at a given positions relative to the start codon, and alignment columns at a given position relative to the stop codon. To do this we calculate, for each alignment column in each multiple alignment, the relative position r_l of the nucleotide in the reference species to the start/stop codon of the gene on the left and relative position r_r to the start/stop codon of the gene on the right (whether these are start or stop codons depends on the orientations of the flanking genes for the intergenic region under study). The R value of the column in question is then added to both averages at positions r_l and r_r. In this way two average profiles were created, average R values around the start codons of genes, and average R values around stop codons of genes. We concatenated these profiles into a single profile by taking from each profile 150 bps in the intergenic region and 300 bps in the coding region.

Figures 5.4 and 5.5 show the R value profiles for all 22 clades we analyzed. Each figure shows 12 panels with 11 panels corresponding to the R value profiles in different clades and one panel

corresponding to the profile averaged over all clades.

Note that although individual clades show differences in the details of the R value profiles, there are a number of features shared by essentially all clades. Selection is strongest at coding positions, in the order: second positions in codons, first positions in codons, and third positions in codons. That is, in order of the frequency with which substitutions at these positions effect the amino acid. Selection at coding positions drops at the starts and ends of genes. Silent positions away from the starts and ends of genes evolve according to the background model ($R = 1$). Selection in intergenic regions is almost always higher than at silent positions and is higher in upstream than in downstream regions. Generally selection in intergenic regions is highest immediately upstream of genes and lowest immediately downstream. There is almost always a sharp peak in selection a few bases upstream of selection start. This peak corresponds to conserved Shine-Dalgarno sequences. Finally, in all clades there is heightened selection at silent positions immediately downstream of translation start.

5.7 Total branch length in the phylogenetic tree versus R values

We study in this section how our R score, which measure evidence of selection, behaves with the total branch length of the phylogenetic tree of the clade. For each clade we calculate the total branch length T in its phylogenetic tree by summing the branch lengths t_b over all branches in the tree, i.e.

$$T = -\sum_b t_b. \tag{5.14}$$

In figure 5.6 we show how the average values of R in intergenic and coding regions depend on this total branch length T.

As the figure shows, there is a clear correlation between the average value of R and the total tree length, both in intergenic (red dots) and in coding regions (purple dots). For intergenic regions there seems to be an approximately linear relationship whereas for coding regions R seems to increase even faster than linearly. The reason there is this general correlation between R and the length of the branches in the tree is that for longer branches the evidence of selection is easier to detect than for short branches, i.e. for very close species most bases are already conserved due to evolutionary proximity.

5.8 Profiles of effective substitution rate

General derivation

The $R(C)$ statistic of a column C calculates the likelihood ratio of the column under the foreground and background evolutionary model. As we have seen before, when the branch lengths in the phylogenetic tree grow it generally becomes easier to distinguish if a column is evolving under the foreground or the background model and R values thus typically grow with the total branch length of the phylogenetic tree (see figure 5.6). As the scaling of R with the branch lengths of the phylogenetic tree may complicate the comparison of the results across different clades we calculated an alternative measure of the amount of selection on an alignment column which does not scale with branch length. Instead of assuming that a column evolves either according to a background model, or according to some unknown WM column, we will instead assume that each alignment column evolves according to a WM column and infer the effective substitution rate at this position.

Note that, if a given position evolves according to WM column w, then the overall rate of substitution at this position depends on the WM column w. That is, given a total mutation rate

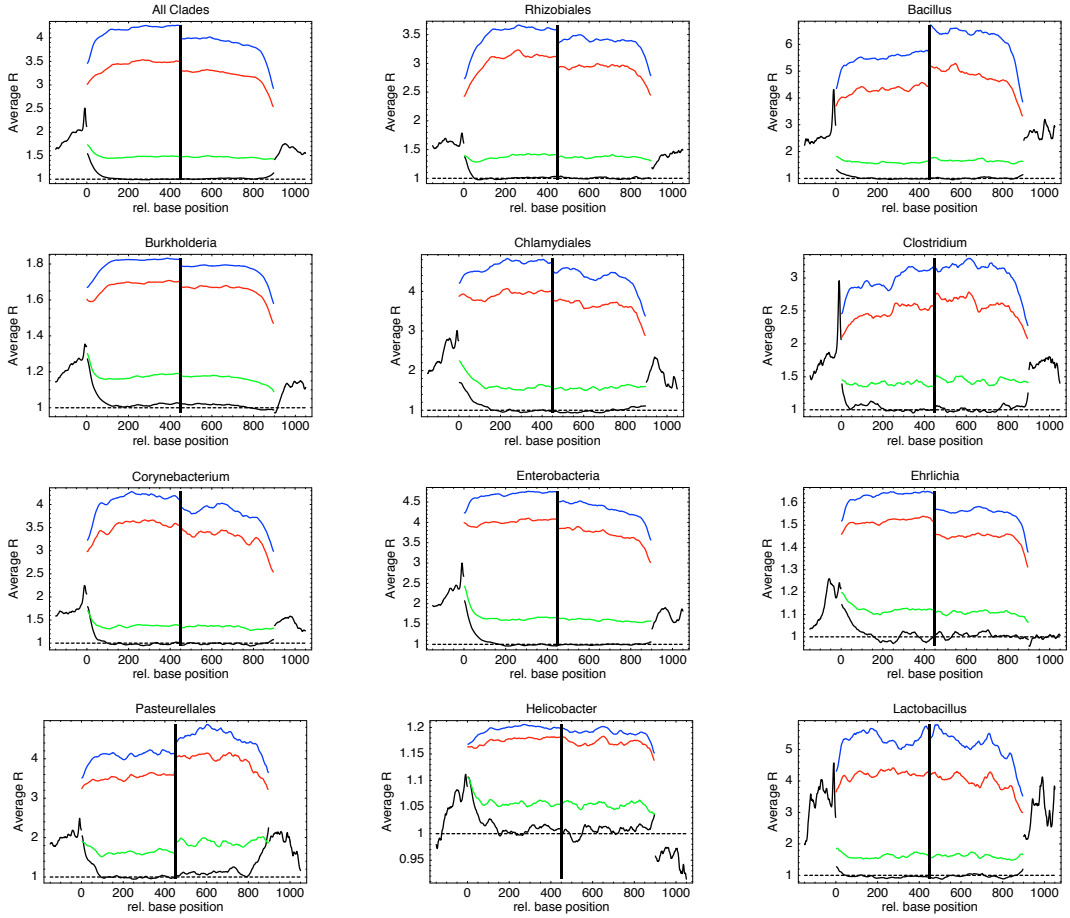

Figure 5.4: Evidence of selection (average R value) as a function of position with respect to translation start (position 0) and end (position 900) in 11 different clades of species, and averaged over all clades. The left half of each panel shows R values in 150 bps upstream regions and the initial 300 bps of genes. The right half shows the last 300 bps of genes plus 150 bps downstream. Average R values at first (red), second (blue), and third positions of codons within genes are shown, as well as average R values within intergenic regions and at silent positions (black). The dotted horizontal line shows $R = 1$ in each panel, which corresponds to evolution according to the background model.

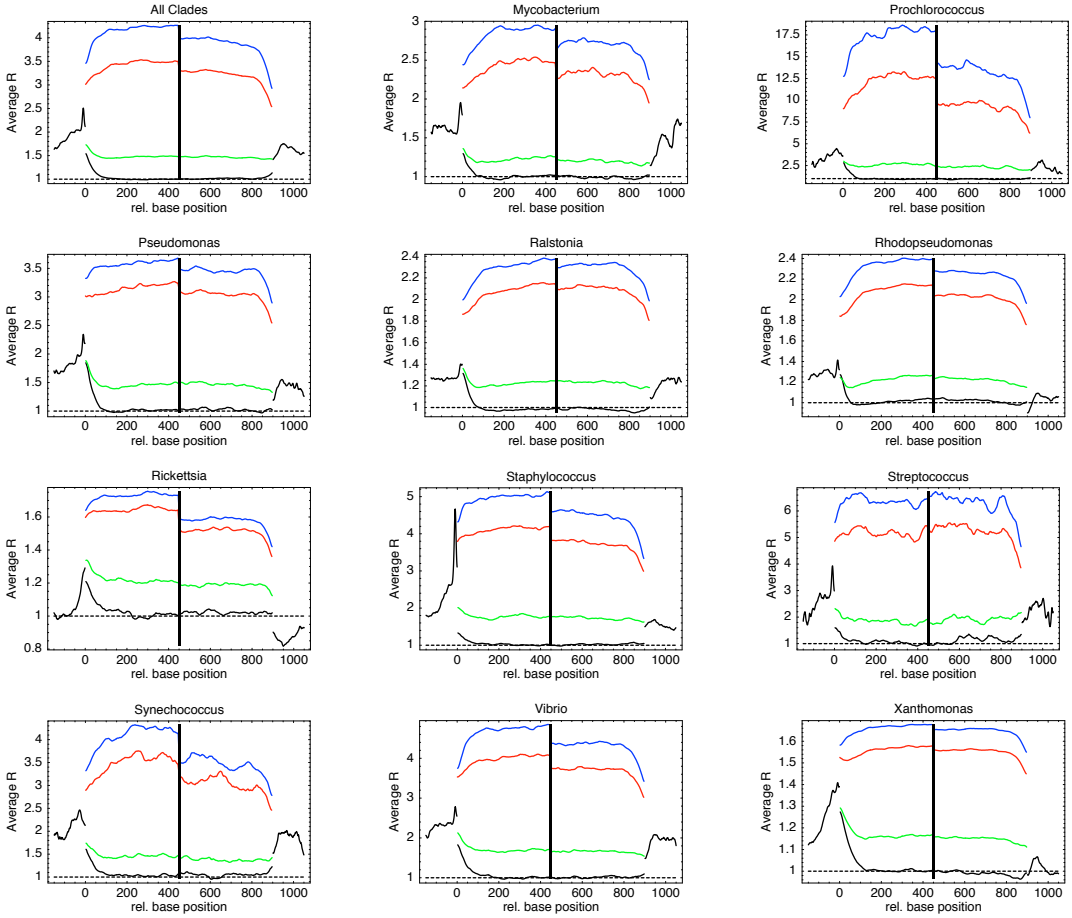

Figure 5.5: Evidence of selection (average R value) as a function of position with respect to translation start (position 0) and end (position 900) in 11 different clades of species, and averaged over all clades. The left half of each panel shows R values in 150 bps upstream regions and the initial 300 bps of genes. The right half shows the last 300 bps of genes plus 150 bps downstream. Average R values at first (red), second (blue), and third positions of codons within genes are shown, as well as average R values within intergenic regions and at silent positions (black). The dotted horizontal line shows $R = 1$ in each panel, which corresponds to evolution according to the background model.

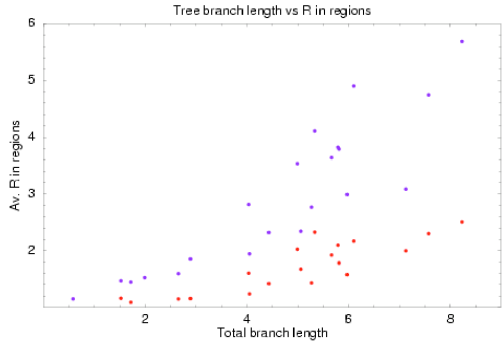

Figure 5.6: Average R values (vertical axis) as a function of total branch length T of the phylogenetic tree of the clade (horizontal axis). Each dot represents one of the 22 clades. Red dots show the average R value in intergenic regions (averaged over DR, SR, and NR regions). The purple dots shows the average R values in coding regions.

of μ, the rate $s_{\alpha\beta}$ of substitution from base α to base β is

$$s_{\alpha\beta} = \mu w_\beta, \tag{5.15}$$

and total rate of substitution away from α is given by

$$s_\alpha = \sum_{\beta \neq \alpha} s_{\alpha\beta} = \mu(1 - w_\alpha). \tag{5.16}$$

Since w_α also gives the equilibrium frequency of occurrence of base α at this position, the probability to find nucleotide α at this position at a given point in time is w_α. Therefore, the average rate of substitution in this column is given by

$$s(w) = \sum_\alpha \mu w_\alpha (1 - w_\alpha) = \mu \left(1 - \sum_\alpha (w_\alpha)^2 \right). \tag{5.17}$$

The prefactor μ just gives the overall rate at which mutations are introduced, independent of w, and the second factor encodes the effect on substitution rate by selection (and mutational bias) as encoded by WM column w. The stronger this bias in the WM column w, the lower the mutation rate. We can thus quantify the strength of selection and mutational bias at this column by a *substitution rate reduction* $\mathrm{SSR}(w)$, which we define as 1 minus the relative substitution rate $s(w)/\mu$:

$$\mathrm{SRR}(w) = 1 - \frac{s(w)}{\mu} = \sum_\alpha (w_\alpha)^2. \tag{5.18}$$

Given an alignment column C, we can thus calculate an expected overall substitution rate reduction $\mathrm{SRR}(C)$ at this position:

$$\mathrm{SRR}(C) = \int \sum_\alpha (w_\alpha)^2 P(w|C) dw = \frac{\int \sum_\alpha (w_\alpha)^2 P(C|w) P(w) dw}{\int P(C|w) P(w) dw}. \tag{5.19}$$

That is, just as we calculated an R value for each alignment column C, we now calculate an expected overall substitution rate at this column $\mathrm{SRR}(C)$. Finally, to quantify the evidence of selection in a column C we defined the Q-statistic $Q(C)$ by normalizing $\mathrm{SRR}(C)$ to the expected $\mathrm{SRR}(C)$ given the background model:

$$Q(C) = \frac{\mathrm{SRR}(C)}{\sum_C \mathrm{SRR}(C) P(C|bg)}. \tag{5.20}$$

and we again calculate average Q values over classes of positions.

Q value profiles

Apart from the R values we also estimated, for each alignment column, the effective substitution rate statistic Q, i.e. the observed reduction in substitution rate reduction Q at this column relative to the substitution rate reduction expected from the background model (see previous subsection). These profiles are shown in figures 5.7 and 5.8.

Comparing figures 5.7 and 5.8 with the R value profiles Figs. 5.4 and 5.5 we see that all main characteristics of the R value profiles are reproduced in the Q value profiles. In fact, the two pairs of figures look very similar. The evidence of selection is highest at coding positions in the order second positions, first positions, and than third positions in codons. In most clades substitution rate reduction is lowest at silent positions in the middle of genes. As in the R profiles substitution rate reduction is higher in upstream regions than in downstream regions, generally is highest immediately upstream of translation start, and lowest immediately downstream of the stop codon. We also again see the sharp peak a few bases upstream of translation start, corresponding to the Shine-Dalgarno sequences, in most clades. Finally, the increase in selection at silent positions immediately downstream of translation start is again observed in essentially all clades.

5.9 Nucleotide composition profiles

We determined the average base composition at positions from 150 bps upstream of translation start to 100 bps downstream of translation start in all 22 clades of bacteria. These nucleotide composition profiles are shown in Figures 5.9 and 5.10.

Although there are significant differences in the base composition profiles between different clades, there are again several features that are universal. For example, in all clades (except for the two cyanobacteria clades Prochlorococcus and Synechococcus) there is a peak in the frequency of A nucleotides around the translation start. In particular, within genes the frequency of A nucleotides is maximal at the start of the gene and decreases over the first 20 nucleotides. G nucleotides have a minimum at the start of the gene and increase over the first 20 nucleotides. The peak in A nucleotide frequency extends into the upstream region. A few bps upstream of translation start a sharp peak in G nucleotides is observed which corresponds to the Shine-Dalgarno sequence. As mentioned, cyanobacteria are the only clades that do not show these patterns. Instead of a peak in the frequency of A nucleotides around translation start the cyanobacteria show a peak in C nucleotides. The cyanobacteria also do not show the peak in G nucleotide frequency immediately upstream of start. These observations suggest cyanobacteria use another mechanism for translation initiation than all other clades. Note that in many clades there seems to be a small but significant minimum in the frequency of A nucleotides between 10 and 20 codons downstream of translation start. We currently have no idea what the meaning or the role of this minimum might be but it seems plausible that it is also related to translation initiation.

In [101] it was shown that, in almost all bacteria DR regions have the highest AT content followed by SR regions, and then NR regions. As demonstrated in Figs. 5.9 and 5.10, we addition find that in all clades the AT content upstream of translation start is higher than the AT content downstream of translation start.

5.10 Selection at silent sites immediately downstream of the start codon

We performed a number of controls to check if the observed elevated selection at silent sites immediately downstream of translation start can be an artefact of another bias. Some of these controls are presented below.

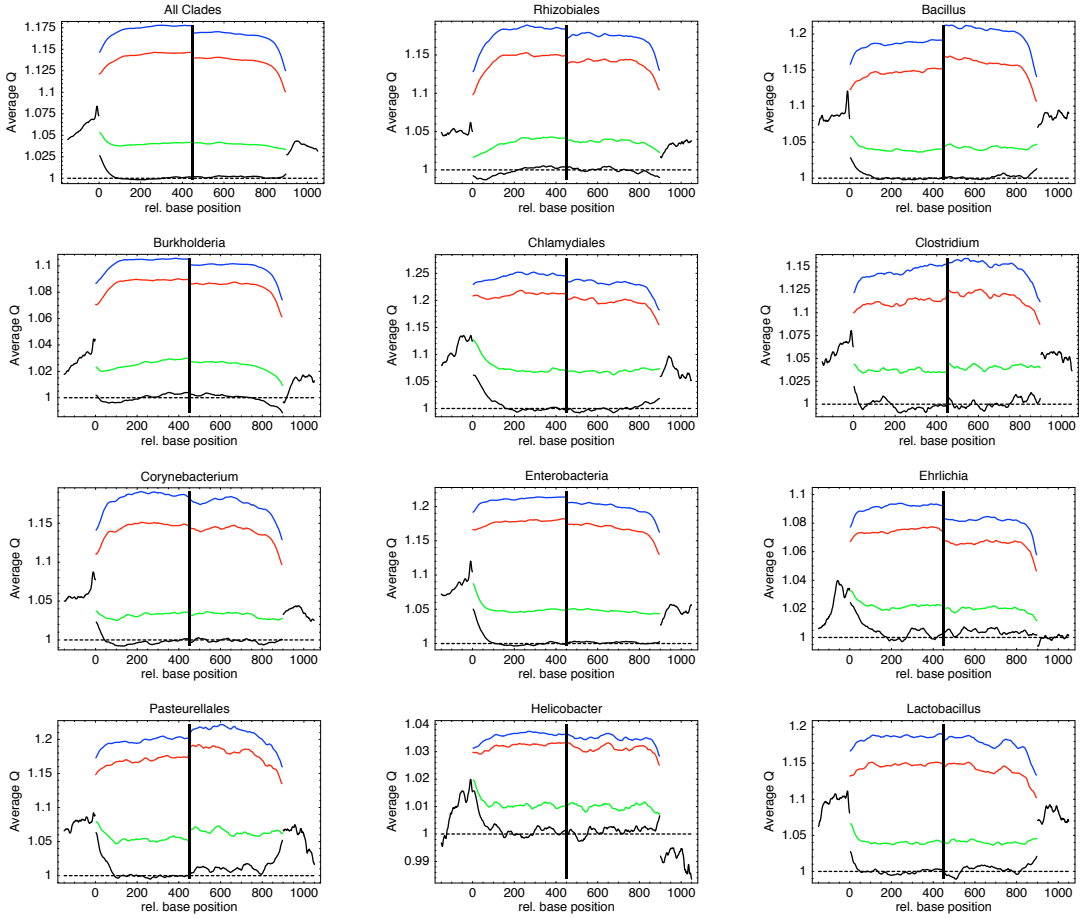

Figure 5.7: Evidence of selection as measured by reduction in effective substitution rate (Q values, see supporting methods) as a function of position with respect to translation start (position 0) and end (position 900) in 11 different clades of species, and averaged over all clades. The left half of each panel shows Q values in 150 bps upstream regions and the initial 300 bps of genes. The right half shows the last 300 bps of genes plus 150 bps downstream. Average values at first (red), second (blue), and third positions of codons within genes are shown, as well as average values within intergenic regions and at silent positions (black). The dashed lines show $Q = 1$, corresponding to the substitution rate expected from the background model.

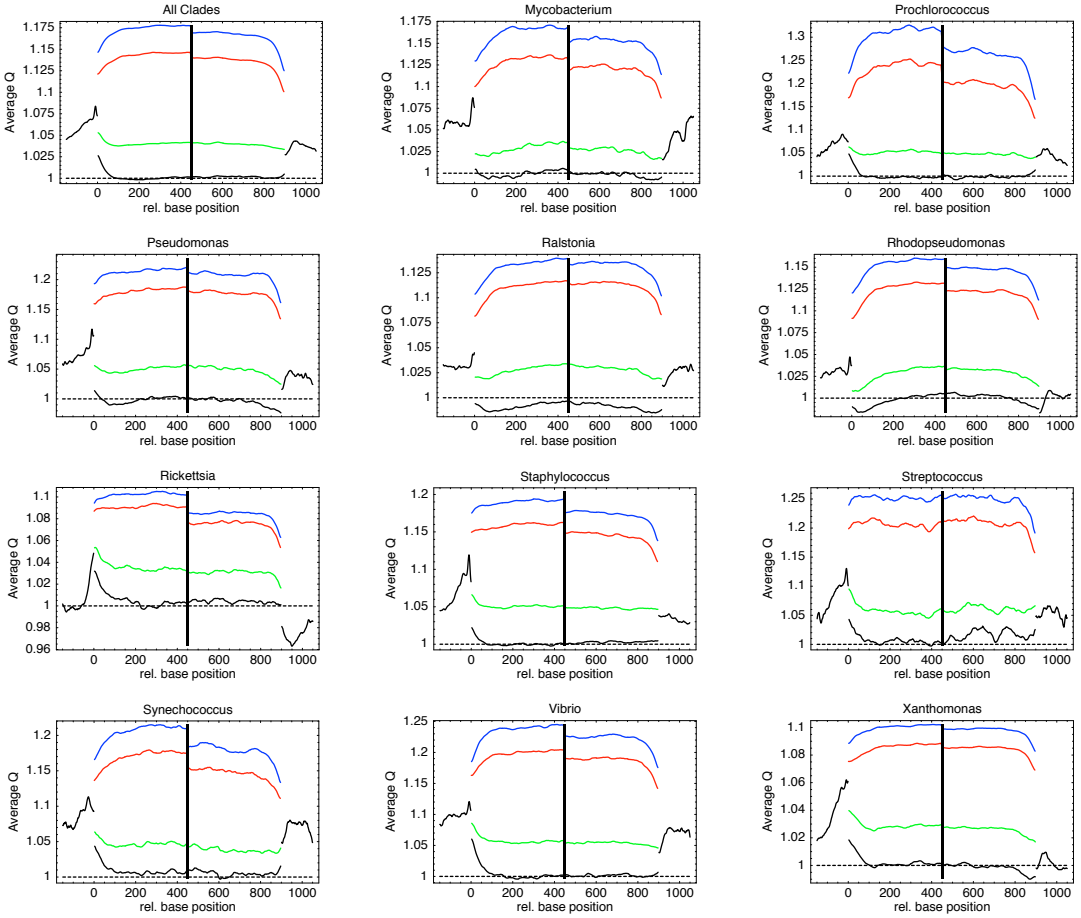

Figure 5.8: Evidence of selection as measured by reduction in effective substitution rate (Q values, see supporting methods) as a function of position with respect to translation start (position 0) and end (position 900) in 11 different clades of species, and averaged over all clades. The left half of each panel shows Q values in 150 bps upstream regions and the initial 300 bps of genes. The right half shows the last 300 bps of genes plus 150 bps downstream. Average values at first (red), second (blue), and third positions of codons within genes are shown, as well as average values within intergenic regions and at silent positions (black). The dashed lines show $Q = 1$, corresponding to the substitution rate expected from the background model.

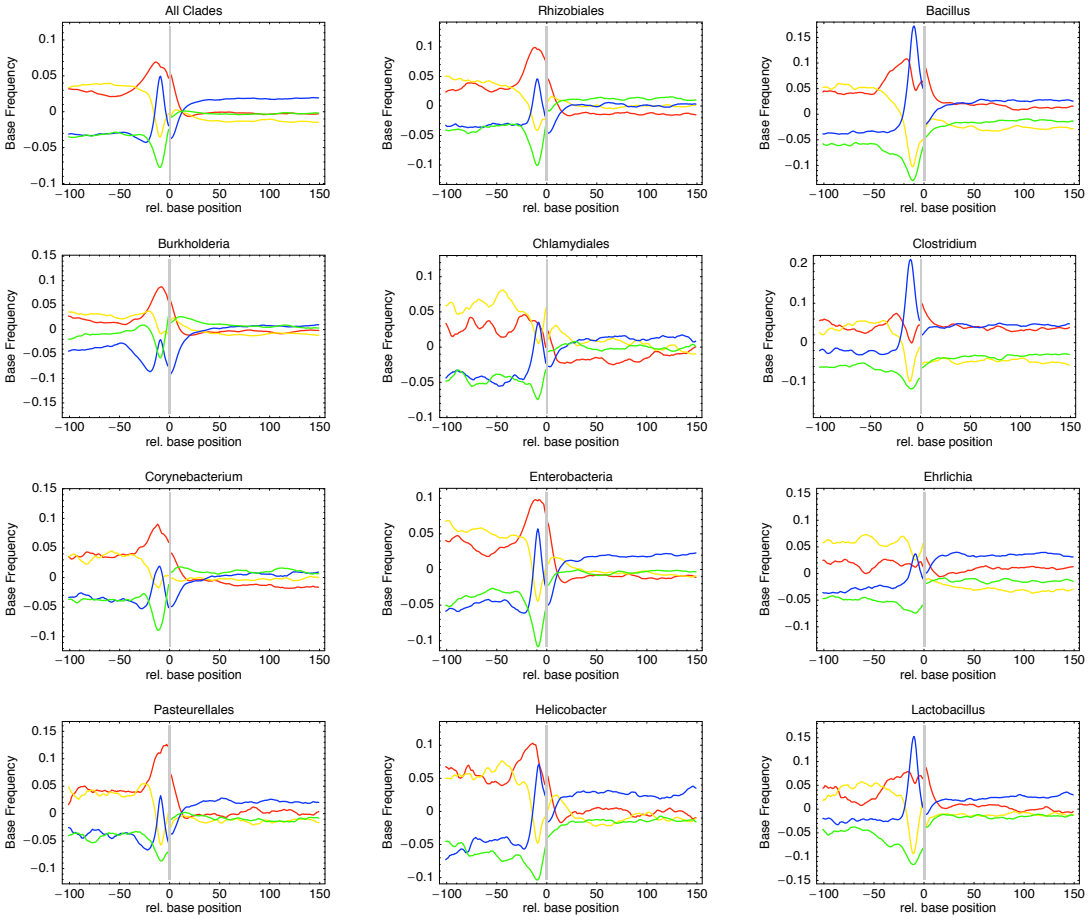

Figure 5.9: Nucleotide composition profiles from 100 bps upstream of translation start to 150 bps downstream of translation start for 11 different clades and the average profiles over all clades. The vertical axis shows the difference between the frequency of A (red), C (green), G (blue), and T (yellow) nucleotides at each position and the average frequency of the corresponding nucleotides in the entire genome.

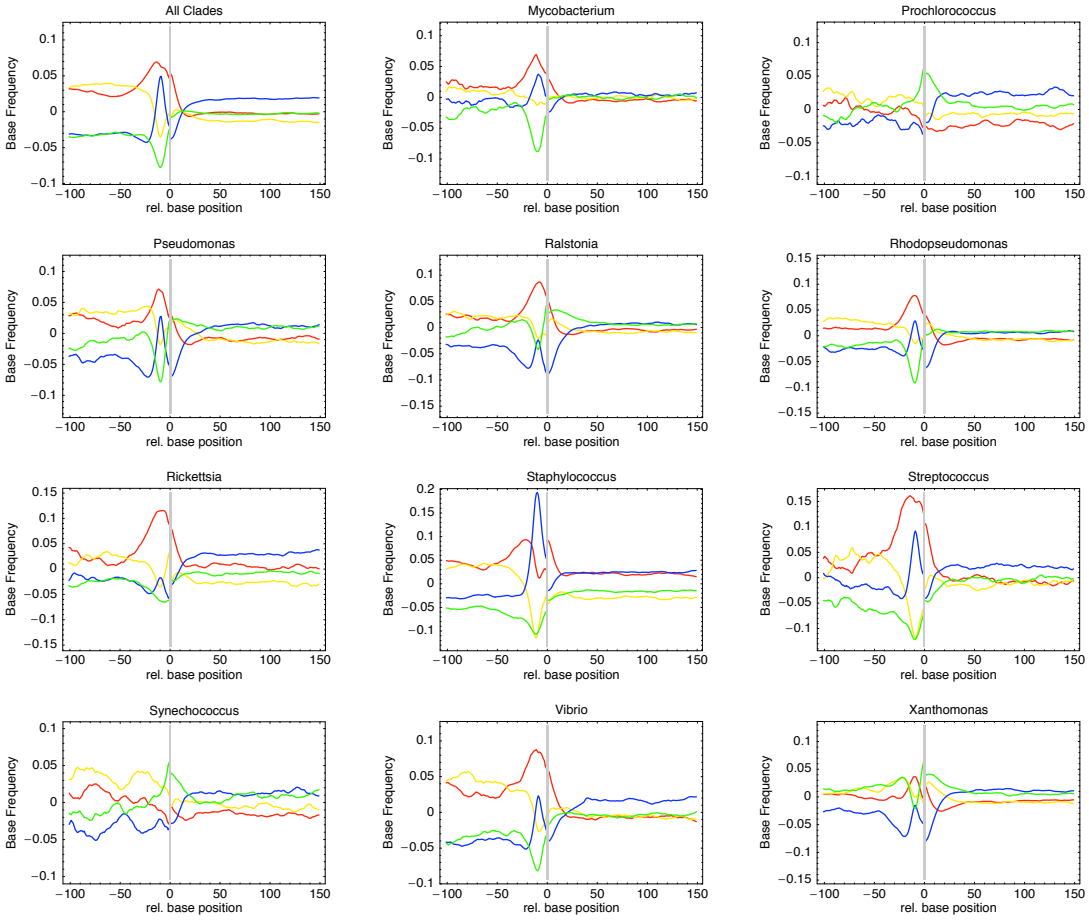

Figure 5.10: Nucleotide composition profiles from 100 bps upstream of translation start to 150 bps downstream of translation start for 11 different clades and the average profiles over all clades. The vertical axis shows the difference between the frequency of A (red), C (green), G (blue), and T (yellow) nucleotides at each position and the average frequency of the corresponding nucleotides in the entire genome.

Reannotation of gene starts

Another hypothesis is that the apparent increase in selection immediately downstream of translation start, and the corresponding lower selection at first and second positions, is an artefact of the incorrect annotation of gene starts in a subset of the genes. That is, if the 'true starts' of a significant fraction of the genes were downstream of the annotated ones, then what we consider to be the initial coding positions of these genes are in fact intergenic positions. Given that the amount of selection at intergenic positions is higher than at silent positions and lower than at coding positions this would produce the pattern of a lowered selection at coding positions and an increase in selection at silent positions, i.e. similar to what we observe.

We implemented a simple procedure, using conservation information, in order to identify gene starts that have potentially been placed too far upstream. First, we search, in the multiple alignment, for an alternative start codon (ATG, GTG or TTG) which is conserved across all species. If such an alternative start exists, we compute the fraction of conserved amino acids $f(i)$ for all columns i of the alignment, the average $\bar{f}_s$ over the positions from the first to the second start codon, and the average $\bar{f}_r$ over the reset of the protein. In this context the 'conservation' fraction $f(i)$ at a position i is the fraction of amino acids in the other species that match the amino acid of the reference species. We then calculate the z-statistic

$$Z = \frac{\bar{f}_r - \bar{f}_s}{\sqrt{\sigma_r^2 + \sigma_s^2}} \tag{5.21}$$

where σ_r and σ_s are the standard errors of the fraction of conserved amino acids in the region between the first and second codon, and in the rest of the protein respectively. Whenever $Z \geq 5$ and the length of the region between the first and second start is less than half of the protein, we reannotate the start of the gene, i.e. move it to the downstream start position.

Table 5.1 shows a number of general statistics on our clades such as the total number of genes in the reference species, the number of regions that were used for constructing the R value profiles and the average number of orthologs per region. It also shows the number of gene starts that were reannotated in our reannotation control. The fraction of reannotated gene starts varies from 5% in Staphylococcus to 35% in Xanthomonas.

Figures 5.11 and 5.12 show the original R value profiles together with the R value profiles using the reannotated gene starts.

As the figure shows, although the reannotation decreases the amount of selection immediately downstream of translation start, and increases the conservation at second and first positions in this region, the changes are small and significant evidence of selection immediately downstream of translation start remains. We also observed that, using the reannotated gene starts, intergenic regions now exhibit more evidence of selection at second and first positions then at third positions (relative to the start codon), which was not the case with the original annotations (data not shown). This suggests that our reannotation has already misclassified a significant number of coding regions as intergenic.

An alternative way of refuting that the selection immediately downstream of translation start is a result of misannotated gene starts is to calculate R values for a set of proteins with well-known amino acid sequences, i.e. with known starts. We built such a set by collecting all E. coli K12 proteins for which the function has been experimentally determined. Again we observed that the R profiles of this set are very similar to the R profiles of all genes. In summary, we believe we can exclude the hypothesis that the observed selection downstream of translation start is an artefact of misannotated gene starts.

Position-dependent codon adaptation index

One hypothesis for the apparent increase of selection immediately downstream of translation start is that it is caused by an increase in codon bias in this region. For example, highly expressed genes

Clade Name	Genes	Starts	Stops	O. start	O. stop	Rean.	Frac.
Rhizobiales	5469	849	522	2.26	2.16	171	20%
Bacillus	4224	818	492	1.93	2.12	112	14%
Burkholderia	7805	1916	1551	2.04	1.96	429	22%
Chlamydiales	1046	325	224	2.90	2.80	27	8%
Clostridium	3954	196	152	2.88	2.89	14	7%
Corynebacterium	3072	589	432	1.89	1.79	123	21%
Enerobacteria	4400	1127	706	2.19	2.08	169	15%
Ehrlichia	967	368	301	2.54	2.27	64	17%
Pasteurealles	1735	236	100	1.83	1.96	38	16%
Helicobacter	1660	238	136	2.74	2.60	69	29%
Lactobacillus	1938	224	146	1.89	1.88	34	15%
Mycobacterium	4237	479	302	2.29	2.08	134	28%
Prochlorococcus	2324	294	236	2.81	2.52	57	19%
Pseudomonas	5684	454	341	2.90	2.77	70	15%
Ralstonia	6532	1172	735	1.80	1.81	346	30%
Rhodopseudo.	4958	1230	893	3.02	2.85	359	29%
Rickettsia	877	308	301	3.12	2.78	44	14%
Staphylococcus	2698	1076	1034	2.63	2.50	55	5%
Streptcoccus	2164	126	72	2.62	2.62	13	10%
Synechococcus	2697	289	166	2.08	1.84	66	23%
Vibrio	3958	713	539	3.09	2.90	213	30%
Xanthomonas	4242	1333	1201	2.45	2.37	460	35%

Table 5.1: Number of regions used in R value profiles, and number of reannotated regions. For each clade the columns show (from left to right): the total number of genes in the reference species, the number of regions around gene starts used for building the R value profiles, the number of regions around gene ends used for building the R value profiles, the average number of orthologs per gene start region, the average number of orthologs per gene end region, the number of reannotated gene starts, and the fraction of gene starts that were reannotated.

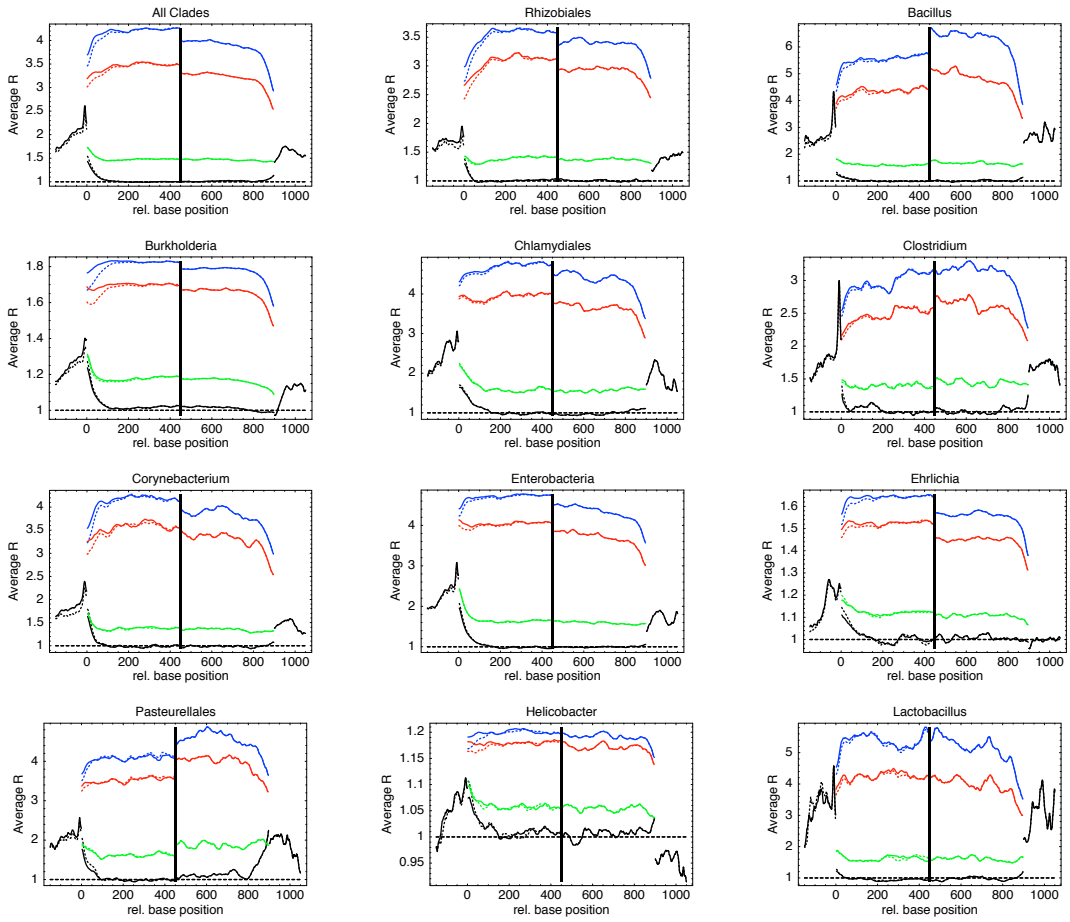

Figure 5.11: Comparison of R value profile with the original and reannotated gene starts. The R profiles with the original gene starts are shown as dotted lines whereas the R profiles with the reannotated gene starts are shown as solid lines. See the caption of figure 5.4 for a description of the data shown.

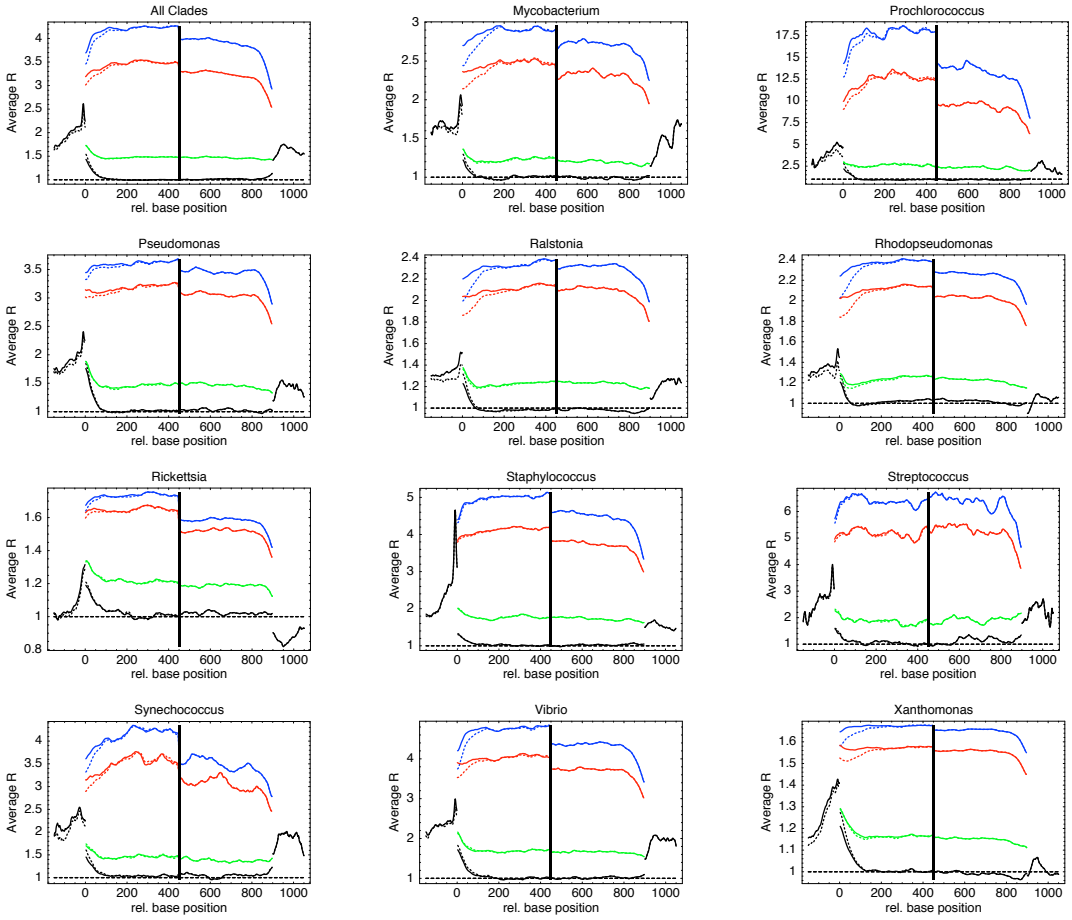

Figure 5.12: Comparison of R value profile with the original and reannotated gene starts. The R profiles with the original gene starts are shown as dotted lines whereas the R profiles with the reannotated gene starts are shown as solid lines. See the caption of figure 5.5 for a description of the data shown.

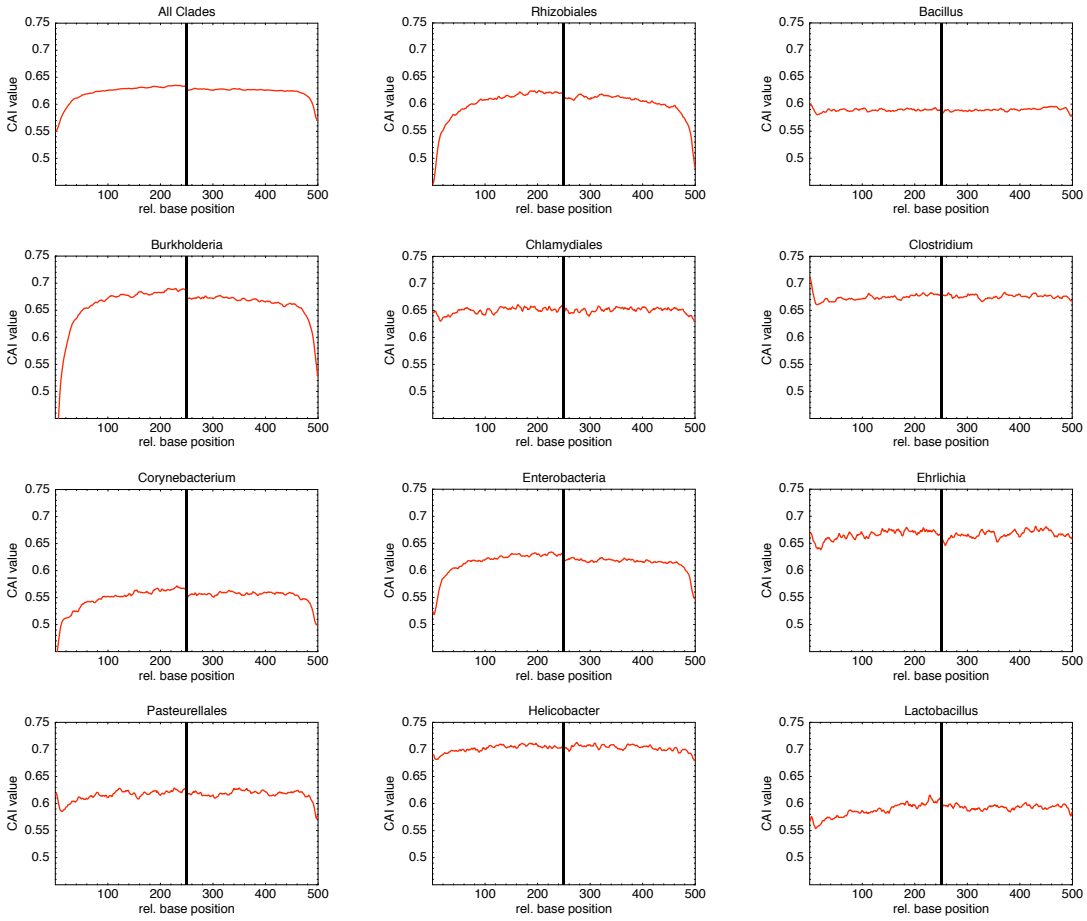

Figure 5.13: Codon adaptation index (CAI) profiles as a function of position relative to translation start (position 0) and translation end (position 500) for the reference species of 11 clades and averaged over all reference species. Each panel corresponds to one reference species. The left half of each panel corresponds to the first 250 codons downstream of translation start, and the right half to the last 250 codons before the stop codon.

such as ribosomal genes generally show elevated codon bias which is likely the result of selection for translation efficiency. It is conceivable that the initial positions of genes are generally under a stronger selection for efficient translation than positions further downstream in the genes, which would lead to higher codon bias and the elevated selection would be the result of an elevated codon bias only.

To test this hypothesis we have computed position-dependent codon adaptation index CAI(d) [102] profiles as a function of the position relative to the start and to the end of the gene. These profiles are shown in figures 5.13 and 5.14.

The profiles show that, for almost all clades, the CAI values go *down* rather than up near the start and end of the genes. For the remaining clades an approximately flat CAI profile is observed. These clades have a codon bias that prefers A nucleotides at the third positions of codons such that the elevated frequency of A nucleotides immediately downstream of translation start matches the overall codon bias. In summary, it is clear that the selection immediately downstream of translation start generally reflects a selection for A nucleotides at these positions and not a selection to match

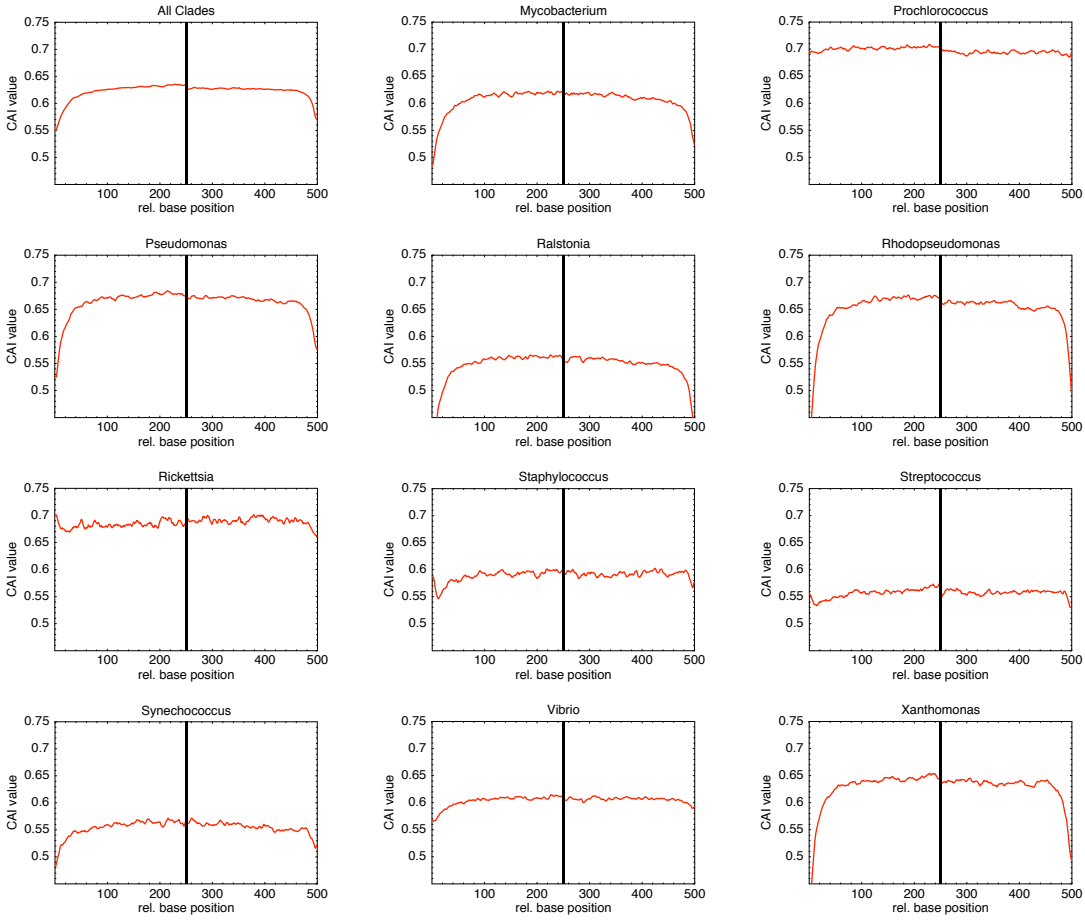

Figure 5.14: Codon adaptation index (CAI) profiles as a function of position relative to translation start (position 0) and translation end (position 500) for the reference species of 11 clades and averaged over all reference species. Each panel corresponds to one reference species. The left half of each panel corresponds to the first 250 codons downstream of translation start, and the right half to the last 250 codons before the stop codon.

the codon bias in the species.

R value profiles at intra-operonic regions

If the observed selection acting at silent sites immediately downstream of the start codon was not related to translation and instead, for instance, was related to transcription we would see this effect only in genes that are located as heads of their operons. To check that we have calculated the average R value profiles upstream and downstream from genes that are not the first in their operon. The profiles in 5.15 and 5.16 show the R values in intra-operonic regions at the starts of genes that are not the first in the operon. Each figure shows 12 panels with 11 panels corresponding to the R value profiles in different clades and one panel corresponding to the profile averaged over all clades.

We see again the sharp peak a few bases upstream of translation start, corresponding to the Shine-Dalgarno sequences, in most clades. And more important, the increase in selection at silent positions immediately downstream of translation start. This suggests that the force responsible for the selection on these sites is related to translation.

Shine-Dalgarno peak and downstream selection signal

If the avoidance of secondary structure around gene starts were related to transcription initiation we would expect to observe this pattern only in genes that are the first in their operon. In the left panel of Fig. 5.17 we compare the selection at the first 20 silent positions immediately downstream of ATG (the 'downstream signal') in genes with small and large upstream regions. Although the downstream signal is often largest in genes with large upstream regions there is clear evidence of downstream signal in genes with small upstream regions, which in some cases is even larger than in genes with large upstream regions.

If the downstream signal is associated with translation initiation we might expect a correlation of this signal with the strength of selection at the Shine-Dalgarno sequences. As shown in the right panel of Fig. 5.17, there is in general a linear correlation between the height of the Shine-Dalgarno peak and the downstream signal. Interestingly, the firmicutes clades (green dots) deviate from this pattern and show relatively little downstream signal and very strongly conserved Shine-Dalgarno sequences. The results in Fig. 5.17 strongly suggest that the avoidance of secondary structure around translation start is the result of a selection pressure for ensuring efficient translation initiation.

5.11 Avoidance of RNA secondary structure around start codons

To provide further evidence that both the increased selection immediately downstream of the start codon, as well the increased frequency of adenines are the result of selection for avoiding RNA secondary structure at the start of the open reading frame, we extracted for each gene the RNA sequence from 60 bp upstream (which is the typical length of 5' UTRs in *E. coli*, see below) to 90 bp downstream and used the Vienna RNA package [103] to determine the probability, for each nucleotide, to be paired with another nucleotide in the RNA secondary structure. By averaging over all genes in the genome we then obtained an average 'open-ness' profile around the translation starts of genes for each clade (see figures 5.18 and 5.19). The red curves show a z-statistic profile for the average openness at a given position compared to the average openness in the flanking regions (-50,-31) and (+31,80), averaged over all clades. There is a clear preference for the region immediately upstream and downstream of translation start to be more free of secondary structure than regions further away. Again this pattern is observed in all clades. Second, for each clade we determined the position-dependent nucleotide frequencies in the regions $(-60, +90)$ around translation starts. We

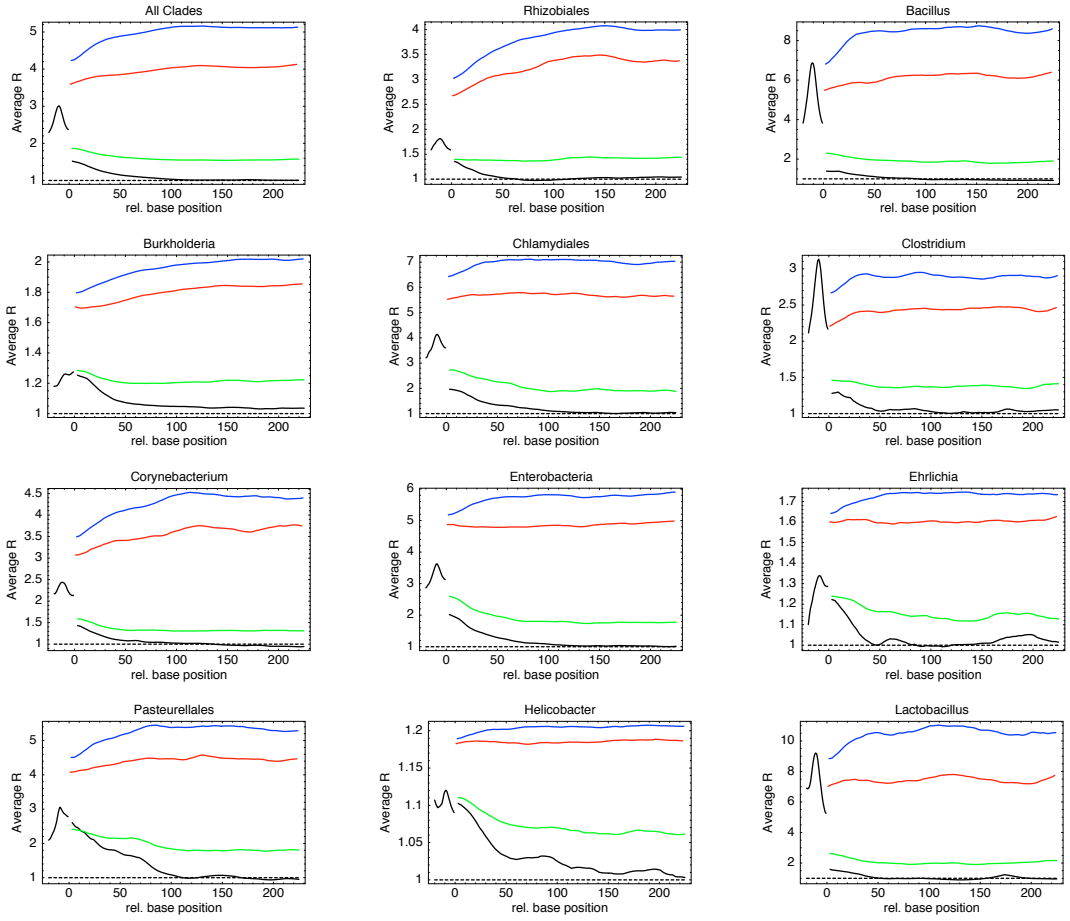

Figure 5.15: Evidence of selection (average R value) as a function of position with respect to translation start (position 0) of genes that are not the first in their operon in 11 different clades of species, and averaged over all clades. Each panel shows R values in 50 bps upstream regions and the initial 250 bps of genes. Average R values at first (red), second (blue), and third positions of codons within genes are shown, as well as average R values within intergenic regions and at silent positions (black). The dotted horizontal line shows $R = 1$ in each panel, which corresponds to evolution according to the background model.

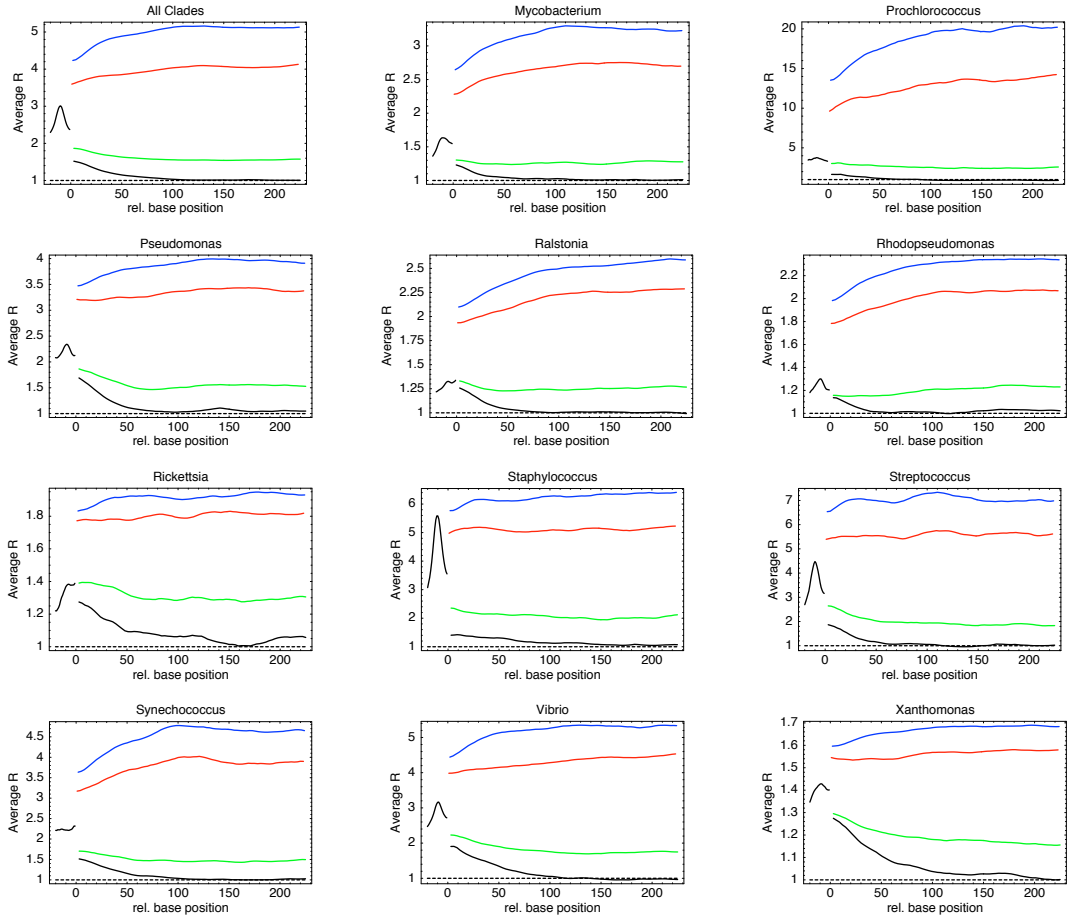

Figure 5.16: Evidence of selection (average R value) as a function of position with respect to translation start (position 0) of genes that are not the first in their operon in 11 different clades of species, and averaged over all clades. Each panel shows R values in 50 bps upstream regions and the initial 250 bps of genes. Average R values at first (red), second (blue), and third positions of codons within genes are shown, as well as average R values within intergenic regions and at silent positions (black). The dotted horizontal line shows $R = 1$ in each panel, which corresponds to evolution according to the background model.

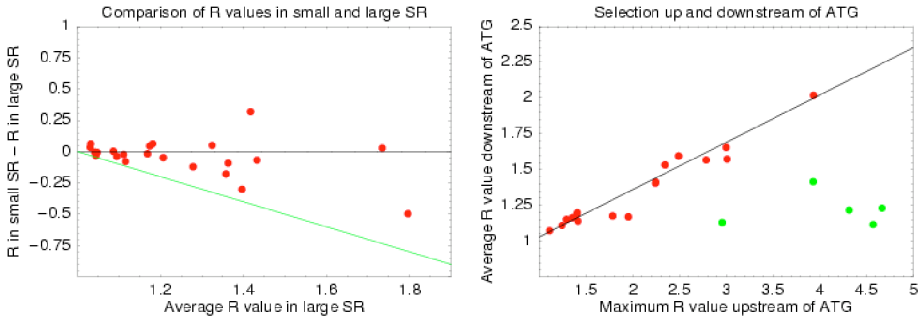

Figure 5.17: Left Panel: Difference of the 'downstream signal' (average value of R in the first 20 silent positions downstream of translation start) between genes with small (< 50 bp) upstream regions and genes with large (> 150 bp) upstream regions as a function of the downstream signal in genes with large upstream regions. The green line corresponds to a value of $R = 1$ in genes with small upstream regions. Right panel: The downstream signal (vertical axis) as a function of the height (in R value) of the peak corresponding to the Shine-Dalgarno signal. The green dots correspond to firmicutes clades. Each dot corresponds to one of the 22 clades in both panels.

then created synthetic sequences that have the exact same position-dependent base composition as the true sequences in that clade, and folded them. The blue curves in figures 5.18 and 5.19 show the z-statistic of the openness of the true sequences compared to these synthetic sequences. Again we see a clearly positive z-statistic in the region immediately around translation start. In summary, show that the base composition around translation start significantly reduces the amount of secondary structure in this area (red curve) and that, beyond this, correlations between bases at different positions further reduce the amount of secondary structure compared to sequences with the same base composition (blue curve).

Two further tests indicate that the avoidance of RNA secondary structure around translation start is associated with selection for translation initiation efficiency. If the avoidance of secondary structure around gene starts were related to transcription rather than translation initiation we would expect to observe this pattern only in genes that are the first in their operon. However, we observe elevated R values immediately downstream of ATGs of both genes with large and genes with small intergenic regions (see figures 5.15 and 5.16). Second, there is an approximately linear correlation between R at the Shine-Dalgarno peak and the average R in the first 20 amino acids downstream of ATG (see figure 5.17) suggesting a link between these two signals. Interestingly, the 5 firmicutes clades deviate from this pattern: they have very strong Shine-Dalgarno sequences but only moderately increased R immediately downstream of ATG. This suggests that in firmicutes translation initiation is dependent mainly on the ribosome binding site. In summary, a pattern of increased conservation and increased frequency of A nucleotides was observed in *E. coli* [96] and *B. subtilis* [97] and was hypothesized to be the result of selection for translation initiation efficiency which leads to avoidance of RNA secondary structure around translation start. Here we provided additional evidence which supports that selection for translation initiation efficiency is indeed the cause of this pattern, and showed that this pattern extends to all bacteria.

RNA secondary structure profiles

Figures 5.18 and 5.19 show position dependent z-statistics for the average probability of bases at that position to be unpaired in the RNA secondary structure of the mRNA around translation start, both compared to the average probability of of being unpaired in the flanking regions $(-50, -31)$ and $(31, 80)$, and compared to the average probability of being unpaired in random sequences with the same position-dependent base composition (see supplementary methods).

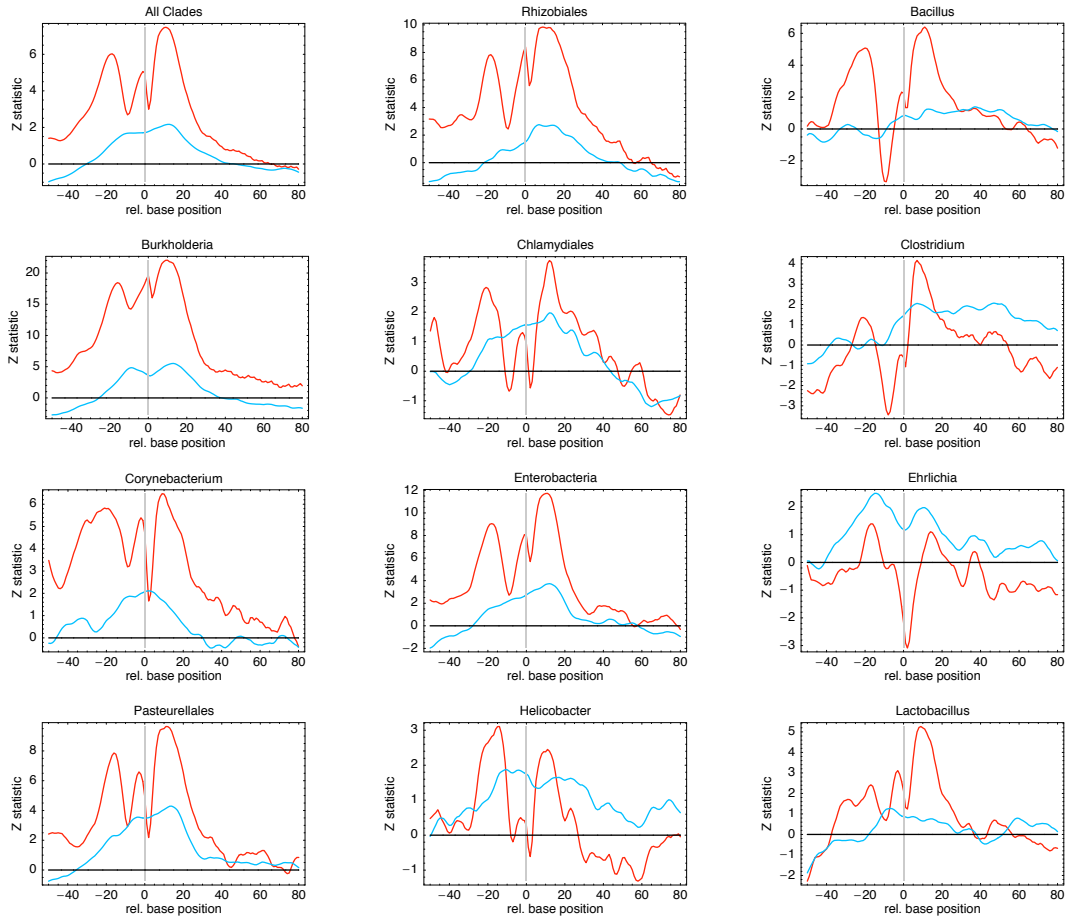

Figure 5.18: RNA secondary structure profiles for 11 clades and averaged over all clades. The horizontal axis in each panel shows the position relative to translation start, from 50 bp upstream to 80 bp downstream. The vertical axes show two z-statistics for the probability of the nucleotide at that position to be *unpaired*. The red lines show the z-statistic of the probability for the position to be unpaired relative to the average probability over the flanking segments $(-50, -31)$ and $(31, 80)$. The blue lines show the z-statistics for the position to be unpaired relative to the average probability of the same position being unpaired in random sequences with the same position-dependent base composition as observed in the clade.

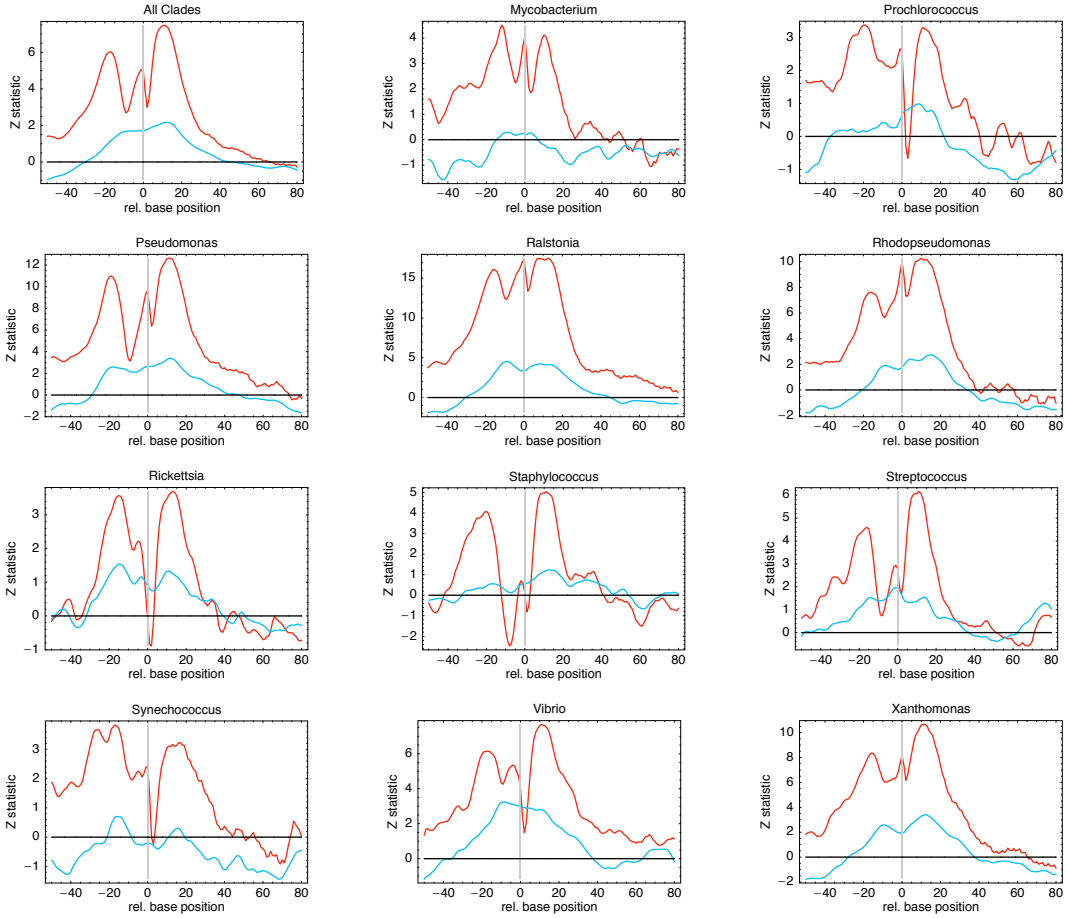

Figure 5.19: RNA secondary structure profiles for 11 clades and averaged over all clades. The horizontal axis in each panel shows the position relative to translation start, from 50 bp upstream to 80 bp downstream. The vertical axes show two z-statistics for the probability of the nucleotide at that position to be *unpaired*. The red lines show the z-statistic of the probability for the position to be unpaired relative to the average probability over the flanking segments $(-50, -31)$ and $(31, 80)$. The blue lines show the z-statistics for the position to be unpaired relative to the average probability of the same position being unpaired in random sequences with the same position-dependent base composition as observed in the clade.

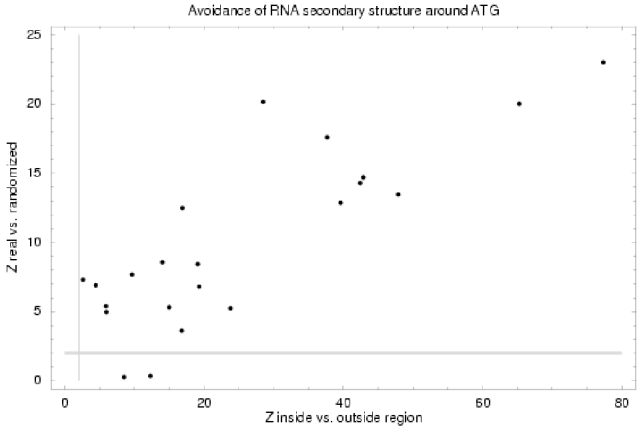

Figure 5.20: Z-statistics of the RNA secondary structure for the region $(-20, 20)$ immediately around translation start. Each dot in the plot corresponds to one clade. On the horizontal axis is the z-statistic of the average openess in the region $(-20, 20)$ compared to the average openess in the flanking regions $(-50, -31)$ and $(31, 80)$. On the vertical axis is the z-statistic of the average openess in the region $(-20, 20)$ compared to the average openess in random sequences with the same position-dependent base composition. The grey lines show the value $z = 2$.

We see that for all clades there are peaks in 'openess' immediately upstream and downstream of translation start compared to the flanking regions more to the left and right (red curves). Note that the G nucleotides of the Shine-Dalgarno sequence and the start codon itself tend to lead to a minimum in openess at these positions. In addition, the blue curves show that the region immediately around translation start shows even more 'openess' then random sequences with the exact same base composition. This strongly suggests that base composition in these regions is the result of a selection for avoiding secondary structure in essentially all clades.

Z-values for the region immediately around translation start

We calculated z-values for the average openess in the region $(-20, 20)$ immediately around translation start, compared with the average openess in the flanking regions (regions $(-50, -31)$ and $(31, 80)$) and z-values for the average openess in the region $(-20, 20)$ compared with the average openess of the same region in random sequences with the same position-dependent base composition. The results are shown in figure 5.20.

The figure shows that in all clades there is significantly more openess, i.e. $z > 2$, in the region immediately around translation start than in the flanking regions. In addition, for all but two clades (Mycobacterium and Synechococcus) there is significantly more openess in the region $(-20, 20)$ then in random sequences with the same position-dependent base composition.

5' UTR lengths in E. coli

For folding the region around translation start we assumed that transcription start occurs 60 bp upstream of translation start, i.e. we include 60 bp upstream of translation start in the sequence to be folded. The estimate of 60 bp is based on analysis of the distribution of 5' UTR lengths in E. coli.

RegulonDB [51] contains a collection of experimentally determined transcription start sites in E. coli. For each TSS we calculated the distance to the start of the downstream ORF and determined the distribution of 5' UTR lengths. Fig 5.21 shows the distribution of 5' UTR lengths that we observed. Note that the majority of 5' UTRs is less than 50 bp long but that there is a fairly long

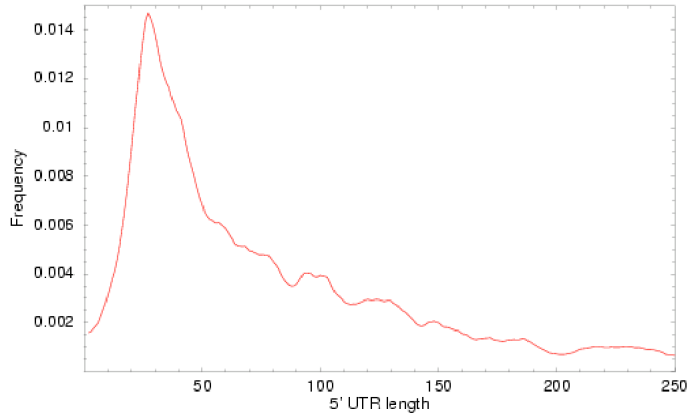

Figure 5.21: Distribution of 5' UTR lengths in E. coli as estimated from the collection of transcription start sites in RegulonDB [51]. The horizontal axis shows the length of the 5' UTR and the vertical axis shows the the frequency of 5' UTRs of the corresponding length. The distribution was smoothed with an exponential kernel (see supporting methods).

tail including apparent very long 5' UTRs. Some of these very long 5' UTRs are possibly due to misidentification of the downstream gene or other sources of error. If one excludes all 5' UTRs longer than 150 bps the average length is 60 bp. If one excludes 5' UTRs longer than 250 bp the average length is 77 bps.

RNA secondary structure statistics

For each gene in each clade we wanted to determine the secondary structure in the mRNA immediately around the start codon. It is hard to do this accurately for two main reasons: First, the secondary structure will depend on the precise transcription start site, i.e. where the mRNA starts, and this is generally unknown. Second, folding algorithms can reasonably accurately determine RNA secondary structure in thermodynamic *equilibrium* but it is likely that the true RNA secondary structure at the start of the mRNA is determined by an essentially kinetic process in which the RNA starts folding as the nascent transcript emerges from the RNA polymerase. That is, we are likely to get a more accurate approximation of the true RNA secondary structure by folding only an initial piece of the mRNA. As an approximation we chose to position the hypothesized 'start' of each transcript 60 bps upstream of its translation start and to fold a region of 150 bps long, i.e. up to 90 bps downstream of translation start.

We thus extracted, for each gene in each clade, the region from 60 bps upstream of the start codon to 90 bps downstream of the start codon and folded it using the Vienna RNA package [103]. Among the statistics that the Vienna package provides is the probability p_i (or fraction of time in equilibrium) for each nucleotide i to *not* be paired to another nucleotide. For each position i, with i running from -60 to $+90$, we calculated the average probability $\langle p_i \rangle$ by averaging p_i over all genes,

$$\langle p_i \rangle = \frac{1}{G} \sum_{g=1}^{G} p_i(g), \tag{5.22}$$

with $p_i(g)$ the probability that position i is open in gene g and G is the total number of genes in the genome. We also calculated the variance v_i as

$$v_i = \frac{1}{G} \sum_{g=1}^{G} p_i(g) \left(1 - p_i(g)\right), \tag{5.23}$$

The standard error e_i in the estimated $\langle p_i \rangle$ is then given by

$$e_i = \sqrt{\frac{v_i}{G}}. \tag{5.24}$$

We now want to compare the probabilities $\langle p_i \rangle$ at different locations relative to translation start. To do this we compare $\langle p_i \rangle$ at each position with the average value of $\langle p_i \rangle$ in the regions away from translation start. That is, we define 'flanking' regions running from $[-50, -31]$ and $[31, 80]$ and calculate the average openess in these areas as,

$$\langle p_{\text{flanking}} \rangle = \frac{1}{70} \left[\sum_{i=-50}^{-31} \langle p_i \rangle + \sum_{i=31}^{80} \langle p_i \rangle \right]. \tag{5.25}$$

Note that we have excluded the regions $[-60, -51]$ and $[81, 90]$ at the ends of the sequence that we fold to avoid boundary effects, i.e. the regions at the ends of the sequence show less base pairing in general.

The standard error e_{flanking} in the estimate of $\langle p_{\text{flanking}} \rangle$ is

$$e_{\text{flanming}} = \frac{1}{72} \sqrt{\left[\sum_{i=-50}^{-31} (e_i)^2 + \sum_{i=31}^{80} (e_i)^2 \right]}. \tag{5.26}$$

Finally, we calculate the Z-statistic at each position i as

$$z_i = \frac{\langle p_i \rangle - \langle p_{\text{flanking}} \rangle}{\sqrt{(e_i)^2 + (e_{\text{flanking}})^2}}. \tag{5.27}$$

Positive z_i values indicate that position i tends to be more open than the flanking regions, and negative values indicate that the position tends to be more closed than the flanking regions.

The openess value p_i at different positions are to a large extent driven by base composition, e.g. the elevated frequency of A nucleotides immediately upstream and immediately downstream of translation start leads to less secondary structure in these areas. It is a priori not clear if the base composition is driving the RNA secondary structure in this area or that a selection for avoiding RNA secondary structure in this area is driving base composition. To test this we compared the observed openess values $\langle p_i \rangle$ with those observed at this position in G randomly generated sequences with the exact same base composition. That is, if the selection is on the RNA secondary structure then one might expect that the openess in the regions around translation start is larger even than the openess in random sequences with the same base composition.

For each clade we created G random sequences where, at each position i, the probability to put A, C, G, or T match the observed base frequencies at that position. We then fold all G sequences and calculate $\langle p_i(\text{rand}) \rangle$ for this random data-set as well as the standard errors $e_i(\text{rand})$. Finally, we calculate the Z-statistics

$$z'_i = \frac{\langle p_i \rangle - \langle p_i(\text{rand}) \rangle}{\sqrt{(e_i)^2 + (e_i(\text{rand}))^2}}. \tag{5.28}$$

Note that positive z'_i values indicate positions at which the openess is larger than in random sequences with the same base composition.

5.12 Discussion

We comprehensive quantified the evidence for purifying selection acting at non-coding positions genome-wide for all 22 clades and found a number of remarkably universal features. First, we found

that the bulk of the silent positions within genes evolve according to the estimated background model, whereas essentially all intergenic regions show evidence of purifying selection. Experimental studies suggest [104] that transcription itself can increase mutation rates (although comparative genomic studies suggest precisely the opposite, see e.g. [105]) and one may wonder if the apparent increase in purifying selection can be explained by a lower mutation rate in intergenic regions. Several of our observations strongly argue against this possibility. An overall lower rate of mutation in intergenic regions would affect all intergenic regions equally, whereas we clearly find most evidence of purifying selection in DR regions, followed by SR regions, and much lower evidence of purifying selection in NR regions. Furthermore, the universal pattern of high R immediately upstream of starts and low R immediately downstream, the universal Shine-Dalgarno peak, and the elevated R at known *E. coli* regulatory sites all demonstrate that R is capturing conserved regulatory elements and not a decrease in mutation rate.

Another universal pattern that we uncovered is a sharp increase of R values at silent positions immediately *downstream* of translation start, which is accompanied by a peak in the frequency of adenines around translation start. Previously this pattern has been observed in *E. coli* [96, 99] and *B. subtilis* [97], and was suggested to result from selection for avoiding secondary structure in the region around translation start. In addition, several experimental studies [98, 100] have shown that increase of adenines immediately downstream of the start codon lead to high translation efficiency. Here we showed that this pattern characterizes all bacterial clades, and we provide evidence that indeed, avoidance of RNA secondary structure around the start codon is important for translation initiation efficiency. We believe that it should be possible to use the biased base composition around gene starts, and the even stronger bias for avoiding RNA secondary structure, to significantly improve *ab initio* gene finding and gene start annotation in bacterial genomes, especially since the pattern seems to apply universally.

The main global statistic of genome organization that we have left largely unexplored is the role of base composition and codon bias. There are a number of intriguing observations that suggest that there may be intimate connections between genomic GC content, codon bias, genome size and regulatory complexity, and selection acting at intergenic and silent positions. First, highly expressed genes tend to show more codon bias [102] and, as tRNA abundances generally correlate with codon bias, this is interpreted as a result of selection at silent positions to ensure translation efficiency of highly expressed genes. Second, more recently evidence has been presented that codon bias is largely driven by an underlying bias in GC content of the genome [106, 107]. Traditionally it has been assumed that GC contents of genomes simply reflect the underlying mutational biases and [105] and [107] provide some evidence in support of this hypothesis. If this is indeed the case then compositional bias, codon bias, the relative abundances of different tRNAs, and the selection at silent sites in highly expressed genes would all derive from an underlying mutational bias. Moreover, our background models would accurately reflect mutational biases, so that the deviations from these background models measure selection directly. There are several observations, however, that suggest that reality may be more complicated. First, experimental studies of mutational biases as well as comparative studies on pseudogenes all suggest a general bias of GC to AT mutations [105]. Second, from a metabolic perspective AT nucleotides are energetically less costly than GC nucleotides, and it has been suggested [108] that this leads to selection for AT over GC nucleotides in situations where energy resources are limiting. Both of these observations beg the question as to why there are genomes with very high GC content at all. Third, there is a clear correlation between GC content and genome size, with very small genomes being almost all AT rich and large genomes being almost all GC rich [46]. It is hard to imagine why genome size and mutational biases would be directly correlated, suggesting again that GC content may be the result of a more complex interplay of effects including selection. Finally, GC content differs in a consistent way between different intergenic regions [101] and genes, suggesting a link between GC content and the regulatory organization of a genome. In essentially all species NR regions have the lowest GC content, followed by SR regions, followed by DR regions, and it was suggested in [101] that this is

a result of the preference of regulatory sites for AT rich sequences. We in addition find that GC content is higher in genes than in intergenic region in all clades. Together all these observations form pieces of a puzzle that relates GC content, codon bias, genome size, and selection in intergenic and silent positions. Working out how these pieces fit together is one of the main issues regarding bacterial genome evolution that remain to be solved.

6 Limited complexity in bacterial regulatory networks

To investigate the dependence of the number of regulatory sites per intergenic region on genome size we have developed a new method, described in the chapter 4, for detecting purifying selection at non-coding positions in clades of related bacterial genomes. Surprisingly, although the number of transcription factors increases quadratically with genome size, we present several lines of evidence that small and large genomes have the same average number of regulatory sites per intergenic region. By comparing the sequence diversity of the most and least conserved DNA words in intergenic regions across clades we provide evidence that the structure of transcription regulatory networks changes dramatically with genome size: small genomes have a small number of TFs with a large number of target sites, whereas large genomes have a large number of TFs with a small number of target sites each.

6.1 Introduction

What is the global structure of transcription regulatory networks in bacteria of disparate genome size? In this chapter we address this question through a comprehensive and quantitative analysis of conservation statistics in intergenic regions across sequenced bacterial genomes. Our main motivation stems from the observation [27, 1], that the number of transcription regulators grows approximately quadratically as a function of the total number of genes in the genome. For example, according to the DBD database [28], the number of transcription factors per genome in bacteria varies from only 3 (of a total of 504 genes) in Buchnera aphidicola, to 801 (of a total of 7717 genes) in Burkholderia sp. 383. To put the latter number in perspective, the vastly bigger genomes of *C. elegans* and *D. melanogaster* have a lower estimated total number of transcription factors according to the same database.

The simplest interpretation for the large range in the number of transcription factors (TFs) across bacteria is that it reflects a large range in complexity of gene regulation across bacteria. For example, as an endosymbiont of aphids, Buchnera lives in a very stable environment and some evidence suggests it shows little transcriptional regulation [29]. In contrast, Burkholderia can live under extremely diverse ecological conditions including soil, water, as a plant pathogen, and as a human pathogen, which most likely require complex regulatory mechanisms.

Quantitatively, the approximately quadratic scaling of the number of TFs thus means that the largest bacterial genomes have about a 20 times higher fraction of genes involved in transcriptional regulation than the smallest, i.e. increasing from about 0.5% in Buchnera to about 10% in Burkholderia. Put differently, the number of TFs *per gene* increases from 1 per 200 genes to 1 per 10 genes. This has important implications for the structure of transcription regulatory networks. One can think of the transcription regulatory network as a graph, with nodes corresponding to genes, and directed edges going from TFs to their target genes. The total number of edges in this network is given by the number of TFs times the average number of outgoing edges per TF, but also by the total number of genes times the average number of incoming edges per gene. That is, if r is the number of TFs, g the number of genes, $\langle i \rangle$ the average number of incoming edges per gene and $\langle o \rangle$ the number of outgoing edges per TF we have $r\langle o \rangle = \langle i \rangle g$. Since the number of TFs *per gene* grows linearly with the total number of genes, i.e. $r/g \propto g$, we cannot have that both the average number of outgoing edges per TF and the number of incoming edges per gene are the

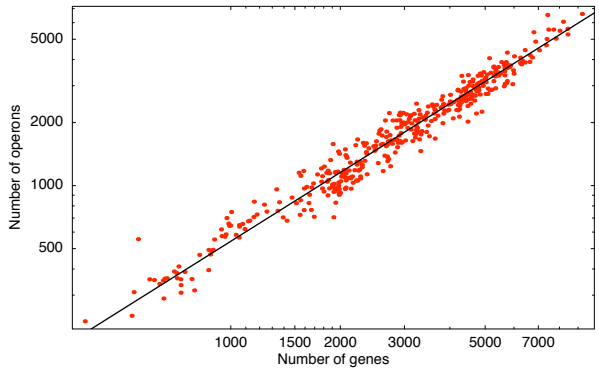

Figure 6.1: The estimated number of operons (vertical axis) as a function of the total number of genes (horizontal axis) for all 416 currently fully-sequenced genes in the NCBI database. Each red dot corresponds to one genome. The black line shows a power-law fit.

same in bacteria of different genome size. In particular, we must have $\langle i \rangle / \langle o \rangle \propto g$. That is, either the number of incoming edges per gene must increase with genome size, i.e. genes are regulated by more TFs in larger genomes, or the number of outgoing edges per TF must decrease with genome size, i.e. the regulon size decreases with genome size (or of course a combination of these two). The main aim of this study was to investigate how the number of incoming edges per gene and the number of outgoing edges per TF depends on the genome size across bacteria.

Transcription regulation is generally implemented through the sequence-specific binding of transcription factors (TFs) to transcription factor binding sites (TFBSs) located mostly in intergenic regions upstream of genes [109]. Therefore, the average number of incoming regulatory edges per gene is directly related to the average number of TFBSs per intergenic region. Here we will assume that the average numbers of regulatory sites can be estimated by comparing conservation statistics of non-coding positions in alignments of orthologous sequences from clades of related bacterial genomes. For a large number of sequenced bacterial genomes one can find other sequenced genomes that are closely related, meaning that orthologous genes and intergenic regions can be identified for a large number of genes, and the intergenic regions show enough conservation to be aligned, yet are sufficiently diverged such that a substantial fraction of nucleotides has undergone substitution since they diverged from their common ancestor. Under the assumption that much of the regulation of orthologous genes is conserved across closely-related species within a clade, we below infer the presence of regulatory sequences from the conservation statistics at non-coding positions in genes and intergenic regions, i.e. measuring the amount of selection that is detected in intergenic regions. In particular, by calculating the likelihood of alignment columns under 'foreground' and 'background' evolutionary models, we quantify the evidence for purifying selection in intergenic regions of genomes form 22 clades of bacteria.

6.2 Operon number and intergenic region sizes

Operon number as function of genome size

Before turning to the analysis of conservation patterns, one might ask to what extent the large range in the number of TFs is reflected in the overall organization of intergenic regions across bacteria. In prokaryotes, genes are organized in operons, i.e. sets of genes which are transcribed together and are under the control of common regulatory elements that occur in the intergenic region upstream of the first gene in the operon. Thus, as TFBSs likely occur predominantly upstream of the first gene in each operon it is relevant ask how the total number of operons grows as a function of

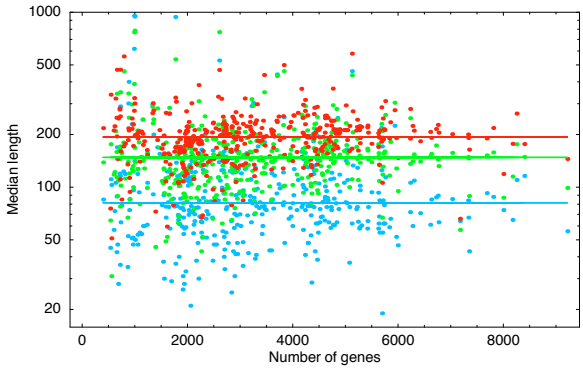

Figure 6.2: Median lengths of intergenic regions (vertical axis) as a function of the total number of genes (horizontal axis) for NR regions (blue), SR regions (green), and DR regions (red) across all sequenced bacteria. Each dot corresponds to one genome. Both axes are shown on logarithmic scale. The horizontal lines correspond to the medians of median region lengths over all genomes.

the total number of genes. Previous studies have shown that the number of operons increases only slightly faster than linear with the total number of genes [110, 2]. We redo this analysis for 416 currently sequenced bacteria, using operon predictions from a recent Bayesian method [111], and find that the number of operons grows approximately as the number of genes to the power 1.09 ± 0.03 as can be seen in figure 6.1. This implies that the number of TFs *per operon* still grows almost quadratically with the total number of genes.

Average intergenic length as function of genome size

Another relevant question is how the lengths of intergenic regions depend on genome size. In eukaryotes, there is a trend for more complex organisms to possess larger amounts of intergenic DNA per gene, and one might expect that large bacterial genomes, with their much larger number of TFs, may also have longer intergenic regions as a result of containing more regulatory sites. This question too has been investigated previously [112, 2] and, somewhat surprisingly, no correlation was found between the average size of intergenic regions and overall genome size. Here we want to farther investigate this observation. To determine the median intergenic region lengths in we used the predictions for 416 currently fully-sequenced bacterial genomes of a recent Bayesian operon-prediction algorithm [111] which we downloaded from `http://www.microbesonline.org/operons/`. Figure 6.2 shows the median size of intergenic regions across currently sequenced bacteria as a function of the total number of genes in the genome. We classified the intergenic regions into 3 different types: non-regulatory (NR) regions that are downstream of two convergently transcribed genes (blue dots), single-regulatory (SR) regions upstream of the first gene in an operon and downstream of another gene (green dots), and double-regulatory (DR) regions that lie between two divergently transcribed genes (red dots).

We found no evidence of correlation between the number of genes and intergenic region size in any of the 3 classes. In [112] it was suggested that intergenic regions in bacteria are under selection pressure to minimize their size while maintaining the necessary regulatory sites. This view is supported by our observation (Fig. 6.2) that DR regions, which contain regulatory signals for two genes, are largest, followed by SR regions, and that NR regions which presumably contain few (if any) regulatory sites are clearly smallest. Interestingly, if the length of an intergenic region indeed reflects the number of regulatory sites that occurs in it, then the absence of a correlation between intergenic region length and genome size would imply that the average number of regulatory sites per intergenic region is the same in small and large genomes. We now investigate this in more detail by analyzing the evidence for purifying selection across non-coding positions in 22 clades of

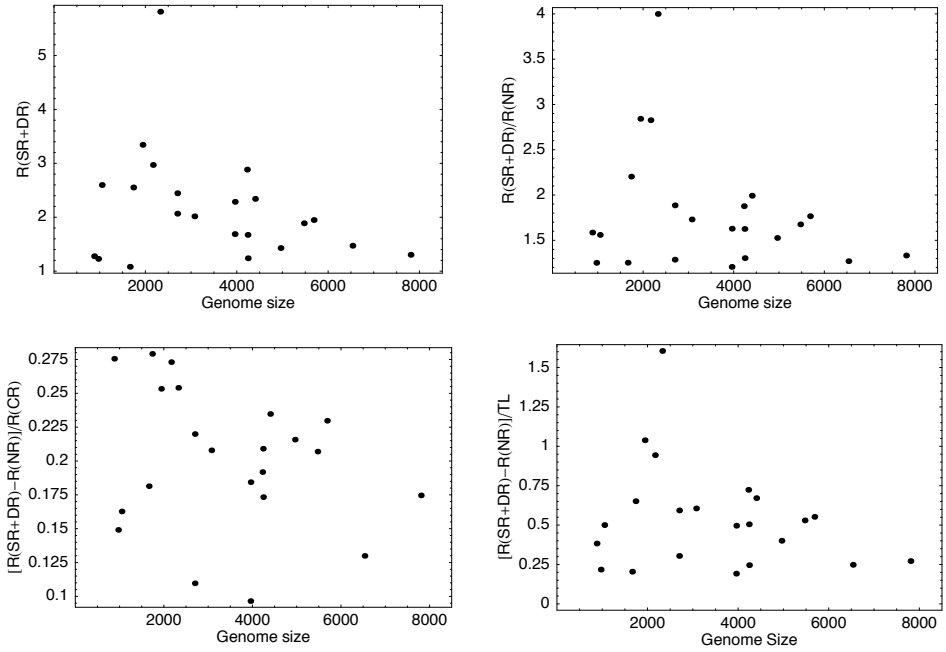

Figure 6.3: Various average R statistics as a function of genome size. In each panel each dot represents one clade. The horizontal axis in each panel shows the total number of genes in the reference species of the clade. The vertical axes show respectively **Top left:** The average value of R in SR and DR regions. **Top right:** The ratio between the average value of R in SR and DR regions and the average value of R in NR regions. **Bottom-left:** The difference between R in SR and DR regions and the average R in NR regions, relative to the average R value in coding positions. **Bottom-right:** The difference between the average R in SR and DR regions and the average R in NR regions relative to the total branch length in the tree.

closely-related bacterial genomes.

6.3 Density of regulatory sites as a function of genome size

Having shown that our R statistic accurately describes sites under purifying selection including known regulatory elements such as the TFBSs in *E. coli* and the Shine-Dalgarno sequences in all clades (see chapter 5), we now return to the main motivation of our study: investigating how the density of regulatory sites in intergenic regions varies with genome size. Since, as mentioned in the introduction, organisms with large genomes appear to have complex life styles that require much greater regulatory complexity, and the number of TFs per gene is much larger in larger genomes, we *a priori* expected that either R itself or a suitably normalized version would correlate with genome size. However, no such correlation exists. We analyzed 12 different statistics to investigate if a correlation between genome size and the amount of evidence for selection in intergenic regions can be detected. In figure 6.3 we show 4 different R value statistics as a function of the number of genes in the genome. the absolute values of R, as well as different combinations of relative differences or ratios of R values in different regions, but none showed any correlation with genome size.

In the top left panel of figure 6.3 we show the average value of R, averaged over all SR and DR regions, directly against the number of genes in the genome. There is no significant correlation (p-value 0.23). As we showed in chapter 5 the R values in DR and SR (upstream) regions are

substantially higher than those in NR (downstream) regions, which is most likely the result of regulatory elements being more abundant upstream of genes than in regions downstream of genes. Therefore, one might argue that a more 'accurate' assessment of the density of regulatory sites can be made by comparing the R values in SR and DR regions with those in NR regions. In the upper-right panel we show the ratio of the average R values in SR and DR regions and the average R in NR regions as a function of the number of genes in the genome. Again there is no significant correlation (p-value 0.21). The difference between R values in SR and DR regions and R values in NR regions also shows no correlation with genome size (data not shown). Another issue that might complicate observation of a correlation with genome size is that the rate of turnover of regulatory sites may be significantly different in different clades. Of course, given that we do not know what the TFs in almost all of these genomes bind, it is hard to estimate the rate of regulatory site turnover directly. However, we would generally expect the rate of turnover to be smallest if the organisms in the clade occupy very similar niches. To some extent we can estimate this from the rate of protein evolution. That is, the amount of conservation at the amino acid level will be higher for organisms living in a similar niche, compared to those that occupy different niches. In the lower-left panel of figure 6.3 we show the relative *difference* $[R(\mathrm{SR}+\mathrm{DR}) - R(\mathrm{NR})]/R(\mathrm{CR})$ between R in SR and DR regions and R in NR regions, *relative* to the average R in coding positions $R(\mathrm{CR})$. That is, we have normalized the difference between R in upstream and downstream regions to the R values at coding positions. We again see that there is no significant correlation (p-value 0.24). Finally, we also showed (see section 5.7) that R values generally correlate positively with the sum of the branch lengths in the phylogenetic tree of the clade. Therefore, one might argue that to obtain properly 'normalized' R values we should divide the R values by the total branch length in the tree. In the bottom-right panel of figure 6.3 we show the relative difference $[R(\mathrm{SR}+\mathrm{DR}) - R(\mathrm{NR})]/\mathrm{TL}$ relative to the tree length TL. Here too there is no correlation (p-value 0.29). We also tried other combinations such as non-normalized differences, or normalized versions of $R(\mathrm{SR}+\mathrm{DR})$ but none gave significant correlations (data not shown).

Finally, note that the R values are calculated compared to what would be expected based on the phylogenetic tree of the species, which was calculated from the silent positions in genes. If intergenic regions are subject to different mutational mechanisms than coding regions than the tree inferred from silent positions may not be appropriate for intergenic regions. To control for this possibility we also build phylogenetic trees from the NR regions in the clade and then calculated R values in intergenic regions using this phylogenetic tree. The results again showed no signs of correlations between the R values in upstream regions and the genome size (data not shown).

Substitution rate reduction

To verify the robustness of this result we performed an analogous analysis using the Q statistic which measures the substitution rate reduction at each alignment column relative to the background model. This Q statistic is thus an alternative measure for the strength of selection in an alignment column that doesn't intrinsically scale with the length of the branches in the phylogenetic tree. As detailed in section 5.8, the Q statistic recovers all the results we found using the R statistic, e.g. substitution rates are lower upstream than downstream of genes, the silent positions evolve according to the background model, substitution rates are lowest upstream of ATG and increase with distance from ATG, and the pattern of lower substitution rates at the Shine-Dalgarno sequence and immediately downstream of ATG. However, as with the R statistic, we found that neither the substitution rates themselves, nor differences of substitution rates between different regions show any correlation with genome size.

The results are shown in figure 6.4. In the top-left panel we show the average Q in SR and DR (upstream) regions as a function of the number of genes in the reference species of the clade. Although by eye there may appear to be some negative correlation, this correlation is not significant (p-value 0.38). Note though that even if the correlation was significant it would go in the *wrong*

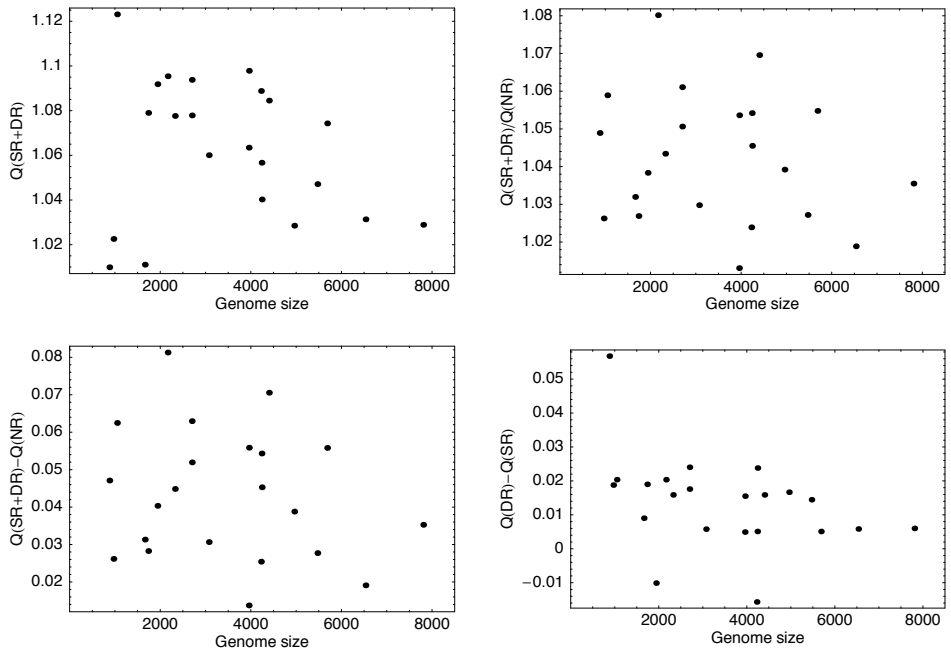

Figure 6.4: Estimated average substitution rate statistics as a function of the number of genes in the genome. In each panel each dot represents one clade. The horizontal axis in each panel shows the total number of genes in the reference species of the clade. The vertical axis in each panel shows: **Top-left:** The average Q in SR and DR regions **Top-right:** The ratio between the average Q in SR and DR regions and the average Q in NR regions **Bottom-left:** The difference between the average Q in SR and DR regions and the average Q in NR regions. **Bottom-right:** The difference between the average Q in DR regions and the average Q in SR regions.

direction, i.e. larger genomes would show less evidence of selection. In the top-right panel we show the average Q in SR and DR (upstream) regions relative to the average Q in NR (downstream) regions. As discussed before, we generally find more evidence of selection in upstream (SR and DR) regions than in downstream (NR) regions and we interpret this as the result of a higher density of regulatory sites in upstream than in downstream regions. Thus, it can be argued that the abundance of regulatory sites should be reflected in the relative sizes of Q in upstream and downstream regions. However, we see in the upper-right panel that there is no correlation whatsoever between this relative substitution rate and genome size (p-value 0.40). Instead of the *ratio* of Q values in upstream and downstream regions we can also consider their *difference* and this is shown in the bottom-left panel of figure 6.4. Again there is no evidence of correlation with genome size (p-value 0.38). Finally, we have also observed that DR regions often show more evidence of selection than SR regions. In the bottom-right panel we show the difference between the average Q value in DR regions and the average Q value in DR regions as a function of genome size. Here there is a marginally significant correlation (p-value 0.07), but again this correlation goes in the wrong direction, i.e. the difference in Q value between DR and SR regions is less in larger genomes.

Branch lengths inferred by PAML

Instead of using our methods to estimate the strength of selection we performed an analogous analysis using PAML. In particular, for each clade, we let PAML infer 11 different phylogenetic trees and calculated the total branch length in each of the trees. One tree was inferred from all alignment columns in NR regions and we denote its branch length by $BL(NR)$. The second tree was inferred from all alignment columns in SR regions, and we denote its total branch length by $BL(SR)$. The third tree was inferred from all alignment columns in DR regions and we denote its total branch length by $BL(DR)$. We denote by $BL(SR + DR)$ the average of $BL(SR)$ and $BL(DR)$. Finally, 8 different trees were inferred from the silent positions of each of the 8 fourfold degenerate codons. We denote by $BL(syn)$ the median of the total branch lengths of these 8 trees. We, then, looked for correlations between the number of genes in the genome and the branch lengths inferred by the PAML algorithm for alignment columns from different regions

The results are shown in figure 6.5. In the top-left panel we show the measure $BL(SR + DR)/BL(syn)$ for each clade as a function of the total number of genes in the reference species of that clade. That is, we compare the branch lengths in upstream regions with those at silent positions. Selection conserving regulatory elements in upstream regions would lead to lowered branch lengths in upstream regions relative to silent positions. As the density of regulatory sites increase the ratio $BL(SR + DR)/BL(syn)$ should thus decrease. However, as the figure shows, even though the ratio is less than 1 in all clades, there is no observable correlation between these branch lengths and genome size (p-value 0.20). In the top-right panel we show the ratio $BL(SR + DR)/BL(NR)$, that is the total branch length in upstream regions relative to the total branch length in downstream regions. Upstream regions are expected to contain much more regulatory elements than downstream regions so that it can be argued that the ratio $BL(SR+DR)/BL(NR)$ quantifies the density of regulatory sites in upstream regions. However, we again observe no correlation with genome size (p-value 0.41). In the bottom-left panel we look at the relative difference $[BL(NR) - BL(SR + DR)]/BL(syn)$. Here we look at the difference in branch lengths between upstream and downstream regions and normalize this using the branch lengths of the silent positions. Again, no correlation with genome size is observed (p-value 0.34). Finally, we look at the difference in total branch length for SR and DR regions (normalized again by $BL(syn)$). DR regions generally should have more regulatory sites than SR regions and their difference can again be argued to reflect the average density of regulatory elements per gene, but again no correlation with genome size is observed (p-value 0.40).

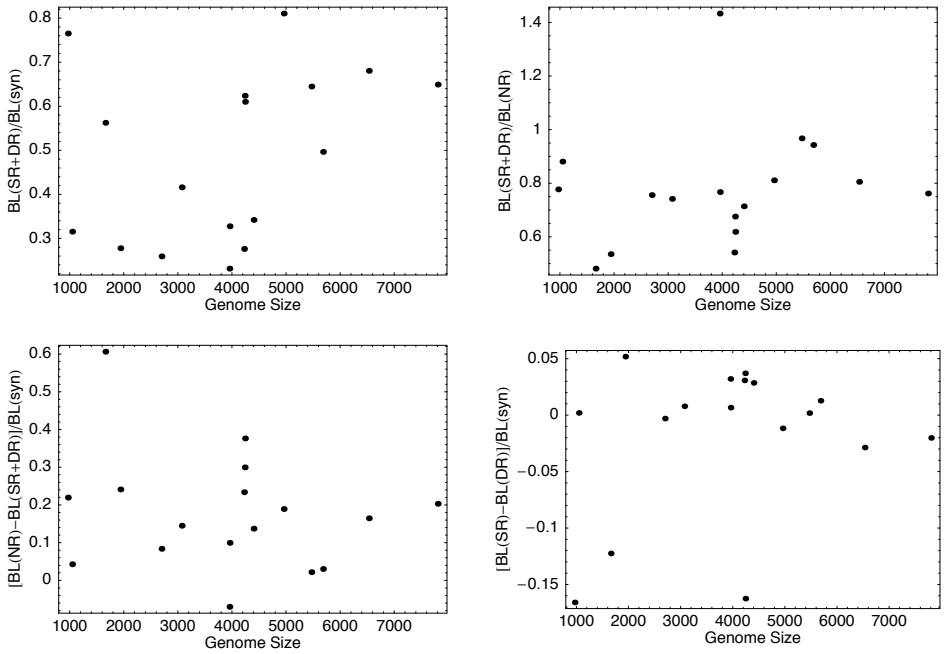

Figure 6.5: Statistics of total branch lengths in the phylogenetic trees of different clades, as inferred by PAML from alignment columns of different regions. Each dot in each panel represents one clade. The horizontal axis in each panel shows the total number of genes in the reference species of the clade. The vertical axes in the four panels show: **Top-left:** The average total branch lengths $BL(SR + DR)$ in the phylogenetic trees inferred from the alignment columns of SR and DR regions relative to the total branch length BL(syn) inferred from alignment columns of silent positions. **Top-right:** The average total branch length $BL(SR + DR)$ inferred from SR and DR regions relative to the total branch length $BL(NR)$ inferred from alignment columns in NR regions. **Bottom-left:** The difference of the total branch length for NR regions and the total branch length for SR and DR regions relative to the total branch length for silent positions, i.e. $[BL(NR) - BL(SR + DR)]/BL(syn)$. **Bottom-right:** The difference between the total branch length for SR regions and for DR regions relative to the total branch length for silent positions, i.e. $[BL(SR) - BL(DR)]/BL(syn)$.

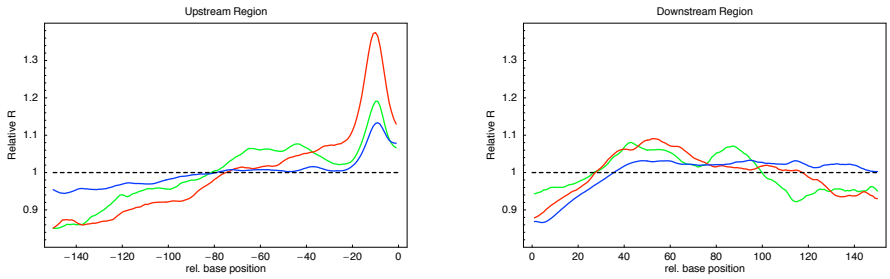

Figure 6.6: Relative average R values upstream and downstream of genes as in the left panel of Fig. 5.3 but now averaged separately over genomes with less than 2000 genes (green), genomes with between 2000 and 4500 genes (red), and genomes with more than 4500 genes (blue). In order to compare the shapes of the R value profiles the values on the vertical axis are scaled to have a mean of 1 when averaged over the 150 bps upstream and when averaged over the 150 bps downstream.

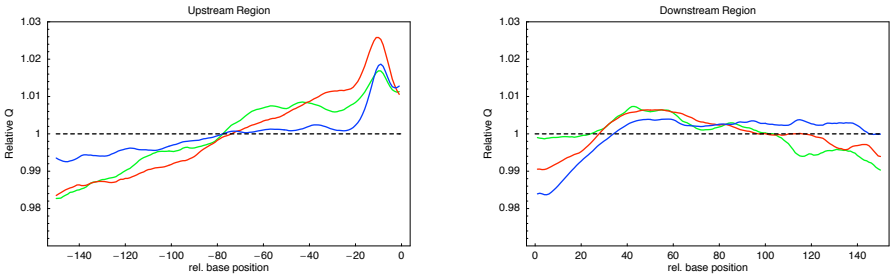

Figure 6.7: Relative average Q values upstream and downstream of genes averaged separately over genomes with less than 2000 genes (green), genomes with between 2000 and 4500 genes (red), and genomes with more than 4500 genes. In order to compare the shapes of the Q value profiles the values on the vertical axis are scaled to have a mean of 1 when averaged over the 150 bps upstream and when averaged over the 150 bps downstream.

Selection profiles of small, medium and large genomes

To verify this further we investigated if there are clear differences in the *shape* of the R statistic profile upstream and downstream of genes for genomes of different size. Figure 6.6 shows the shapes of the R profiles upstream and downstream of genes, i.e. as in figures 5.4 and 5.5, but now separately for small, medium-sized, and large genomes. The shapes of the profiles are very similar for the three classes of genome sizes. In Shine-Dalgarno peak is most pronounced in medium-sized genomes and least pronounced in large genomes. Similarly, the R profile appears to drop fastest with distance from ATG for medium-sized genomes and slowest for large genomes. The shape of the small genome profile falls somewhere in between the shapes of the profiles for large and medium-sized genomes. Thus, although there are some small differences in the shapes of the profiles, these differences do not show a consistent trend with genome size. As shown in figure 6.7 we find essentially the same result with the substitution rate statistic Q. Overall, the similarity of the profile shapes for small, medium, and large genomes supports that there is a common architecture of regulatory sites which is independent of genome size. Note that, as mentioned in the discussion of figure 6.2, this result is also supported by the absence of a correlation between intergenic region size and genome size.

Conclusion

In summary, in spite of of using three different methods (R values, reduction in substitution rates, and branch lengths inferred by PAML), using both silent positions and NR regions to infer the phylogenetic trees, and using a number of different statistics for each of these methods, we did not find any indication that the density of regulatory sites increases with genome size (the total number of genes in the genome). Although it could be argued that more sophisticated models than the ones we employed might be able to uncover a subtle correlation it seems highly unlikely that the density of regulatory sites in intergenic regions changes substantially between the smallest and largest genomes. For example, the fraction of genes in the genome that are regulatory genes increases by about a factor of 20 between the smallest and largest genomes. If the density of regulatory sites would have increased by a similar factor then our methods would have detected such a increase. Note that our methods do infer more evidence of regulatory sites upstream then downstream of genes, they detect the elevated selection at silent sites immediately downstream of translation start, and they correctly infer the strong selection on the Shine-Dalgarno sequence immediately upstream of translation start. It thus seems highly unlikely that a significant increase in the density of regulatory sites would have gone undetected.

The combination of results just presented provides compelling evidence that the average number of regulatory sites per upstream region is independent of genome size. This implies that, whereas the number of TFs increases quadratically with genome size, the total number of regulatory sites increases only linearly with genome size. There are now two possibilities. The first possibility is that in small genomes there are significantly more TFBSs per TF than in large genomes, i.e. regulon size decreases with genome size. The second possibility is that TFs in large genomes more often *share* TFBSs, i.e. that each TFBS is bound by multiple TFs. In eukaryotes one often finds families of TFs with highly similar DNA binding domains that have essentially identical sequence specificities, such that a given binding site can be bound by all members of the family [113]. In prokaryotes, however, such potential sharing of binding sites by families of related TFs has so far not been investigated in detail.

6.4 Clustering of TFs with similar DNA binding domains

If sharing of TFBSs by multiple TFs is more common in large genomes we would expect more clusters of TFs with highly similar DNA binding domains in large genomes than in small genomes. In particular, we would expect that, whereas the total number of TFs grows approximately quadratically with genome size, the number of distinct families of TFs would grow more slowly with genome size. Fig 6.8 shows that this is not case. We collected the DNA binding domains of all TFs in each genome using Pfam [36]. For different similarity cut-offs p we then used single-linkage clustering to cluster all domains with at least p percent identity. We find that, at various cut-offs p, the number of clusters grows roughly as a power-law of the total number of genes (Fig 6.9). Fitting the exponents of the power-laws that are obtained for different cut-offs p (Fig. 6.8), we found essentially the same exponent when we clustered DNA binding domains, as when we fitted the power-law of the total number of TFs as a function of the total number of genes (1.85). That is, even if we cluster all TFs whose DNA binding domains are 50% identical (at the amino acid level) we still find that the number of clusters grows with almost the same exponent as when we each TF is counted independently. For comparison, we compared the DNA binding domains of all *E. coli* TFs for which the binding specificity is known [70] and found 10 pairs of TFs with at least 50% similarity in their DNA binding domains. Of these 10 pairs only 4 show similarity in their binding specificity (data not shown). In summary, there is little evidence for families of TFs with high similarity in their DNA binding domains, and no evidence that such families are more common in large than in small genomes.

Figure 6.9 shows the number of different clusters of paralogous transcription factors as a function

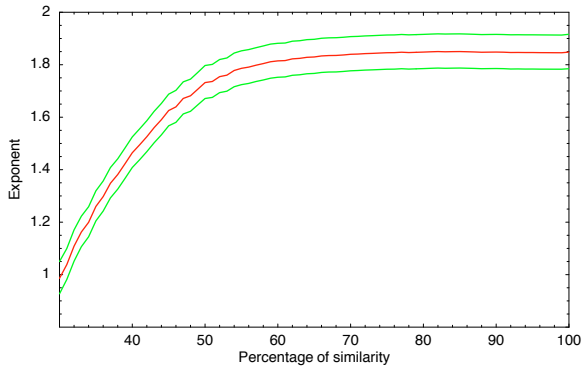

Figure 6.8: Fitted exponent (vertical axis) for the number of DNA binding domain clusters as a function of genome size for different similarity cut-offs (horizontal axis). For a given similarity p we clustered all TFs in whose DNA binding domains had a similarity of at least p percent were clustered using single-linkage (separately for each genome). We then fitted the number of clusters as a function of the total number of genes in the genome to a power-law. The fitted exponent is show as the red line, with the green lines indicating the 95% posterior probability interval.

of the total number of genes in the genome at different similarity cut-offs.

The figure shows that, at all similarity cut-offs the function can be reasonably well-fitted by a power law. The fitted intersects and exponents are

- Intercept -9.656, Exponent 1.845 at a cut-off of 100% similarity.
- Intercept -9.699, Exponent 1.847 at a cut-off of 85% similarity.
- Intercept -9.568, Exponent 1.827 at a cut-off of 65% similarity.
- Intercept -8.209, Exponent 1.625 at a cut-off of 45% similarity.

We thus find that at all three higher cut-offs there is very little evidence of TF clustering, and the amount of clustering does not increase with genome size. At 45% identity there is some clustering but the exponent is still as high as 1.6 (i.e. far from linear) and at this low similarity there is little guarantee that the TFs will bind similar motifs.

6.5 Sequence diversity of DNA 7-mers under purifying selection

As the 'sharing' of TFBSs does not seem to increase with genome size, and the number of regulatory sites per intergenic region appears constant, the necessary consequence is that the number of TFBSs *per TF* must decrease with genome size. That is, our results suggest that small genomes have a small number of large regulons, while large genomes have a large number of small regulons. To test this directly, we compared the sequence diversity of the most conserved sequence segments with the diversity of the least conserved sequence segments in the intergenic regions of each genome. For each clade, we enumerated all 4^7 7-mers, counted their number of occurrences in intergenic regions and ranked them by the amount of evidence they show of being under purifying selection (see next subsection). Then, starting from the most significantly selected 7-mer, we counted how many distinct 7-mers are necessary to account for 5% of all intergenic sequence segments of length 7. We denote this number by n_t. Similarly, starting from the bottom of the list, we counted how many distinct 'un-selected' 7-mers n_b are necessary to account for 5% of all intergenic sequence segments.

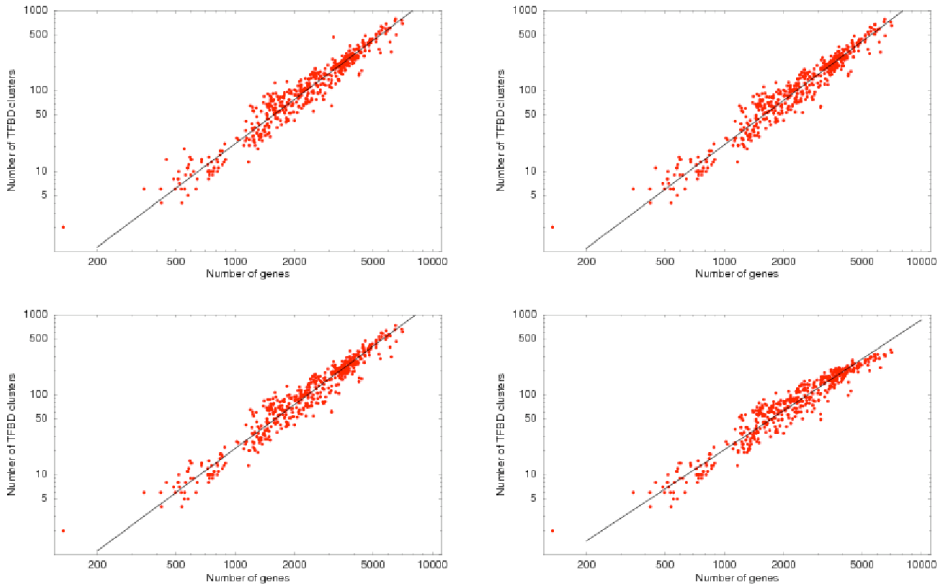

Figure 6.9: Number of clusters of transcription factors with similar DNA binding domains at cut-offs of 100% amino acid identity (top-left), 85% amino acid identity (top-right), 65% amino acid identify (bottom-left) and 45% amino acid identity (bottom-right) as a function of the total number of genes in the genome. Both axes are shown on logarithmic scales. The black lines are power-law fits.

Figure 6.10 (top) shows the ratio n_t/n_b for each genome as a function of the number of TFs in the genome. We find that in small genomes one needs only a small number of highly-selected 7-mers to account for 5% of all intergenic sequence segments, whereas in large genomes a large number of highly-selected 7-mers is needed to account for 5% of all sequence segments (this also holds when taking 10% or 20% instead of 5%, see figure 6.10 middle and bottom). To put it differently, in small genomes the most selected 7-mers are much more frequent than poorly selected 7-mers whereas in large genomes the most selected segments are much less frequent than poorly selected segments. This observation provides a strong piece of independent evidence that, indeed, the regulon sizes of small genomes are significantly bigger than regulon sizes in large genomes. Note that the changes in the ratio n_t/n_b are large: n_t/n_b increases over almost two orders of magnitude, i.e. roughly by the same factor as the number of TFs (straight line fit in Fig. 6.10). In fact, besides the number of TFs and the number of signal transduction genes [2] we are not aware of any other genome statistic that increases by such a large factor between small and large genomes as the ratio n_t/n_b.

Sequence diversity of most and least conserved 7-mers

The probability that a sequence segment evolves under the foreground rather than the background model is quantified by the sum of the $\log(R)$ values of the alignment columns in the segment. Moving with a sliding window of length 7 over all intergenic region alignments we assigned a score X, equal to the sum of $\log(R)$ values, to each window. For each of the 4^7 possibles 7-mers s we collected all $n(s)$ occurrences of s in intergenic regions and calculated the average score $\langle X(s)\rangle$ and its variance $\mathrm{var}(X(s))$. We also calculated the overall average $\langle X\rangle$ over all n windows of length 7 and the overall variance $\mathrm{var}(X)$. Assuming that the scores of the $n(s)$ windows with 7-mer s were drawn from a Gaussian distribution with unknown mean and variance, the probability that the

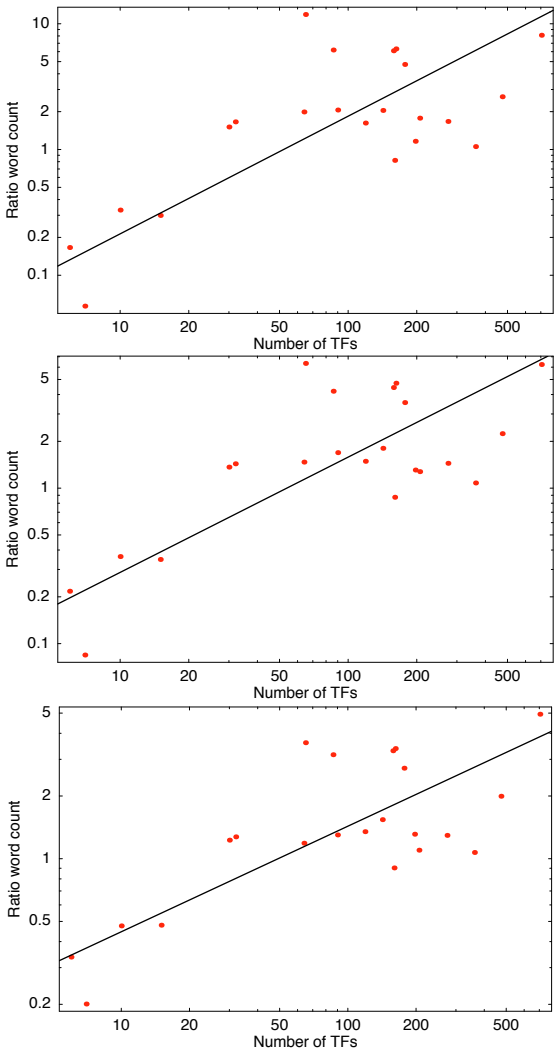

Figure 6.10: Relative sequence diversity of the 7-mers under most and least purifying selection as a function of the number of TFs in the genome. For each genome we ordered all 7-mers by their evidence for being under purifying selection and collected the most and least conserved unique 7-mers such that the 7-mers of both sets each account for 5%(top), 10% (middle) and 20% (bottom) of all sequence segments in the genome. The vertical axis shows the ratio between the number of most selected and least selected 7-mers in the corresponding set as a function of the total number of TFs in the genome (horizontal axis). Both axes are shown on logarithmic scale. The black line shows a linear fit.

mean differs from the overall mean $\langle X \rangle$ is quantified by the z-statistic

$$z(s) = (\langle X(s) \rangle - \langle X \rangle) \sqrt{\frac{n(s)}{\mathrm{var}(X(s)) + \mathrm{var}(X)/n(s)}}. \tag{6.1}$$

For each clade we calculate the z-statistics $z(s)$ for each 7-mer s and produced an ordered list of 7-mers, with the most selected at the top and least selected at the bottom. We then collected the top n_t 7-mers such that the sum of their occurrence counts $n(s)$ equals $0.05n$, i.e. 5% of all windows. Similarly we collected the bottom n_b 7-mers such that the sum of their occurrence counts $n(s)$ equals $0.05n$. Finally we calculated the ratio n_t/n_b for each clade.

6.6 Discussion

The intriguing observation that the number of TFs increases almost quadratically with the total number of genes in bacteria implies that there must be important structural differences between the transcription regulatory networks in small and large bacterial genomes. As the number of TFs per gene increases linearly with genome size, either large genomes have on average more regulatory inputs per gene, i.e. more regulatory sites per upstream region, or TFs in large genomes have on average less regulatory outputs per TF, i.e. smaller regulons (or a combination of the two). In order two investigate these possibilities we set to estimate the density of conserved sites in intergenic regions of 22 'clades', comprising a total of 105 bacterial species.

We produced multiple alignments of orthologous genes and intergenic regions, and estimated the phylogenetic tree of each clade from third positions in fourfold degenerate codons using a 'background' evolution model that takes codon bias into account. We defined an R statistic that measures, at each alignment column, the likelihood that the position evolves under substitution rates significantly different from the substitution rates of the background model. We showed that our statistic accurately captures known selection pressures and reveals known regulatory elements. For example, in all clades we find a sharp peak in R at Shine-Dalgarno sequences a few bases upstream of the start codon. In addition, we showed that the average R values at known regulatory sites in *E. coli* are almost as high as at coding positions, and significantly higher than the average R in upstream regions overall. We comprehensive quantified the evidence for purifying selection acting at non-coding positions genome-wide for all 22 clades and found a number of remarkably universal features (see chapter 5). Results that finally show that R is capturing conserved regulatory elements and not a decrease in mutation rate. We then applied this measure to evaluate the correlation between genome size and the amount of purifying selection in intergenic regions.

Previous work has shown that operon sizes decrease only slightly with genome size [110, 2] and that the sizes of intergenic regions are independent of genome size [112, 2]. This implies that every time the size of a bacterial genome doubles, the total amount of intergenic DNA upstream of operons roughly doubles as well. Yet the number of TFs roughly quadruples, implying that large genomes have a larger number of TFs per upstream region. One may therefore expect that large genomes have a larger number of regulatory sites per upstream region, especially considering that bacteria with large genomes are generally thought to exhibit much more complex transcription regulation than small parasitic bacteria. In spite of attempts to identify such a correlation using three different methods for measuring purifying selection, and using a large number of different statistics, we found no correlation whatsoever between genome size and the amount of purifying selection in intergenic regions, suggesting that large and small genomes have on average the same density of regulatory sites per gene.

Given that our conservation statistics can only measure the density of *conserved* regulatory sites, an alternative possibility is that large genomes have a higher density of regulatory sites but that these sites tend to be less conserved. Although in principle possible, this scenario would require a general correlation between genome size and the rate of regulatory site turnover and,

moreover, it would require that as the density of sites increases the turnover rate increases so as to precisely counterbalance the increased site density, leaving no correlation between the number of *conserved* binding sites with genome size. Assuming that site densities simply do not correlate with genome size seems to us a much more parsimonious assumption. In addition, the profiles of R and Q upstream of gene starts have similar shapes for small, medium-sized and large genomes which further supports that promoter architectures and regulatory site distributions are similar for large and small genomes. Finally, it is thought that bacteria are generally under selection to minimize the size of their genomes and pseudogenes are typically removed from the genomes relatively quickly. It has therefore been argued [112] that the *sizes* of intergenic regions reflect the amount of regulatory sites within them. Consistent with this hypothesis, we find that DR regions are longer than SR regions and that NR regions are by far the shortest. Yet the sizes of different types of intergenic regions also do not show any correlation with genome size.

All these observations are consistent with the simple conjecture that the number of regulatory sites per intergenic region is constant for small and large genomes, leading us to hypothesize that the basic molecular mechanisms of transcription regulation in bacteria strongly constrain the number of different TFs that can co-regulate a given bacterial gene. That is, we hypothesize that bacteria do not have the molecular mechanisms that allow them to place a gene under the control of many different regulatory elements. As a consequence, bacterial genes have on average the same (small) number of regulatory elements per gene, independent of the genome size and the total number of TFs in the genome. This is in stark contrast to what is observed in eukaryotes. Especially in higher eukaryotes genes can receive regulatory inputs from many different regulatory modules that can be located many tens of kilobases from the transcription start site and it is generally assumed that the number of inputs per gene increases with the complexity of the organism. Correspondingly, the sizes of intergenic regions increase dramatically as one moves from simple to more complex eukaryotes. We thus propose that a key difference between the transcription regulatory networks of prokaryotes and eukaryotes is that prokaryotes are constrained to only a small number of regulatory inputs per gene.

The quadratic growth of TFs with genome size together with an on average constant number of regulatory sites per gene now imply that the number of unique regulatory sites per TF decreases significantly with genome size, i.e. by a factor of 20 between the smallest and largest genomes. Given that we find that clusters of TFs with highly similar DNA binding domains are typically small and the size of these clusters does not grow with genome size, we conclude that there is little evidence of 'site sharing' in bacteria, which in turn implies that TFs have on average much fewer TFBSs per TF in large compared to small genomes. This conclusion is further supported by our observation that there is a highly significant correlation between genome size and the sequence diversity of the most conserved sequence segments: whereas in small genomes the most conserved 7-mers tend to also be the most common 7-mers in intergenic regions, in large genomes the most conserved 7-mers are the least common 7-mers. This provides a strong independent piece of evidence that regulon sizes are large in small genomes and small in large genomes.

7 Discussion

7.1 Summary of results

Scaling laws across bacterial clades

In chapter 2 we showed that the scaling laws in the functional content of genomes are reproducible using Pfam annotations of protein domains. More importantly, we found that for the large majority of high-level functional categories the data is consistent with a common scaling law across all bacterial clades, i.e. the scaling laws are *universal*. This observation, supports the assumption that the rate of addition and deletion of domains depends only on the functional category of the domain, not on time or evolutionary history. This assumption is the basis of the model of genome evolution studied in chapter 3. Unfortunately, how the scaling laws emerge from fundamental principles remains unknown.

Implications for the rates of additions and deletion across evolutionary lineages

In chapter 3 we presented a simple evolutionary model previously formulated in [31]. We showed that a time-invariant hypothesis, i.e. assuming that the observed scaling laws have held since the last common ancestor, led us to derive the following constrains on the evolutionary dynamics:

1. The relative rates of fixing additions and deletions of protein domains in each functional category are proportional to the fraction of all domains in the genome currently in that category.

2. The relative rates are proportional to a constant which depends on the functional category, but not on time or evolutionary lineage. We called this constant the *evolutionary potential*.

3. The *evolutionary potential* of each functional category are equal to the scaling exponent of that category.

We tested these predictions by comparing the number of protein domains in several pairs of closely related species. With this data, we estimated the rates of addition and deletion for different functional categories and evolutionary lineages, and showed that prediction 1 above holds. Then, we estimated functional-specific evolutionary potentials for each pair of closely-related bacteria. We found that while the *evolutionary potential* clearly vary between categories, they are similar across evolutionary lineages. Finally, we estimated overall *evolutionary potentials* for each functional category and we showed that there is a good correlation between these estimates and the exponents of the scaling laws. In summary, these results support the predictions made by our evolutionary model and reinforce the idea that *evolutionary potentials* are fundamental constants of the evolutionary process.

Implications for horizontal gene transfer

Our finding that the rate of addition of protein domains is proportional to the number of preexisting domains in the genome places an important constraint on horizontal gene transfer. The rates of domain duplication and domain deletion are naturally proportional to the number of preexisting domains, but it is not at all obvious that this should be the case for horizontal gene transfer.

If horizontal gene transfer is an essential force in shaping the gene-content of genomes as recent studies have suggested, then the rates at which it occurs must be proportional to the number of existing domains. We propose a few hypotheses for why this might be so:

Horizontal gene transfer may be much more common between closely related organisms, as would be the case for DNA conjugation and DNA insertion by bacteriophages. Closely related organisms are likely have similar numbers of domains in each functional category, so the rate of horizontal gene transfer would again be roughly proportional to the number of domains in the receiver genome. However, many horizontal gene transfer has been observed between distantly related species.

It is also possible that the genome size of a bacterium is approximately dictated by its habitat. One could think that bacteria living in a similar environment needs similar genomic tools to optimally survive, and therefore they will show similar genome sizes. Since genomes with similar sizes have similar number of genes in different functional categories the rates of horizontal gene transfer would be naturally proportional to these numbers. Although this is an attractive idea, we are unaware of any evidence suggesting that bacteria from the same habitat have similar genome sizes.

Finally, horizontal gene transfer might occur preferentially between organisms with similar genome sizes. In fact, there is some evidence that bacteria silence foreign DNA with a GC-content significantly different than their GC-content of its own genome. The GC-content is correlated with genome size, so this could make successful horizontal transfer more likely between organisms with similar genome sizes.

A novel method to detect selection

We developed a novel method to comprehensively detect DNA sites and segments that evolve under selection pressure. Starting from genome sequences of a set of organisms we extract and align both genes and intergenic regions. We infer the phylogenetic tree topology using cliques of orthologous genes and the distances of the branches using four fold degenerate positions incorporating rigorously the codon usage bias of each species. We then use an evolutionary model that quantifies the strength of selection acting on different regions of the genome.

With this method we fund highly conserved sequences in the intergenic regions of 22 bacterial clades. For most of these species, nothing is known about the regulatory sites in their genomes, and we believe our method can provide, at least, partial predictions for binding sites in these cases. The sequences we found are publicly available in the SwissRegulon database. We also measured the amount of selection at each site of our alignments, and found several interesting and general patterns. Some of our methods have been integrated in MotEvo, a novel tool for detecting binding sites in alignments of intergenic regions given known weight matrices.

Profiles of selection and translational efficiency

For all 22 bacterial clades, we measured purifying selection on noncoding sites. Several selection patterns appear universally across clades. The bulk of silent sites evolve according to our background model, while essentially all intergenic regions endure purifying selection. At silent sites just downstream of translation start sites there is a sharp increase of the strength of selection and an increase in codons encoding adenine. This pattern appears in all bacterial clades and correlates with reduced RNA secondary structure around the translational start site, which is known to be important for the efficiency of initiating translation. We believe that these patterns could be used to significantly improve ab initio gene prediction in bacteria.

Limited complexity of regulatory networks

We studied how the scaling law governing the number of transcription factors in an organism affects the structure of transcriptional regulatory networks. The fraction of all genes which are

transcription factors grows linearly with the total number of genes, which suggests that the regulation of a gene in a large genome involves more transcription factors than in a small genome. Given this, we would expect that the number of regulatory sites upstream of each gene, and therefore its regulatory complexity, increases with genome size. Our findings show that this is not the case.

First, we showed that the operon size and average size of intergenic regions are independent of genome size, thus the fraction of the genome reserved for the regulation of each gene is roughly independent of genome size. Moreover, three measures of purifying selection in intergenic regions across 22 clades find no correlation with genome size, indicating that large and small genomes have the same density of regulatory sites per gene. Finally, the selection profiles upstream of gene starts are similar for all genome sizes, which suggests that architecture of promoter regions do not depend on genome size.

Both the length and the amount selection on intergenic regions are independent of genome size which strongly indicate that the number of binding sites per intergenic region is also independent of genome size. Yet it could be that in large genomes, many transcription factors bind the same sites. However, we found that clusters of transcription factors with highly similar DNA binding domains are typically small and the size of these clusters does not grow with the total number of genes in the genome. This gives us a piece of evidence to support the hypothesis that in prokaryotes transcription factors generally do not bind similar sites. Therefore, the number of transcription factor that bind upstream of a gene is independent of genome size since the number of binding sites does not correlate with genome size. This conclusion is further supported by the diversity and frequency of the most conserved DNA words. That is, there is a highly significant correlation between genome size and the sequence diversity of the most conserved sequence segments: whereas in small genomes the most conserved 7-mers in intergenic regions tend to also be the most common, in large genomes the most conserved 7-mers are the least common. This indicates that sequence diversity increases with genome size, thus the number of sites bound by a particular transcription factor decreases with genome size. This provides a strong independent piece of evidence that regulon sizes are large in small genomes and small in large genomes.

All these observations show that the number of regulatory sites per intergenic region is independent of genome size, which leads to the hypothesis that the molecular mechanisms of transcription regulation in bacteria does not permit a particular gene to be under the control of arbitrarily many transcription factors. Instead, bacteria have the same, small number of regulatory elements per gene, independent of genome size and the total number of transcription factors. This is completely different from eukaryotes, especially in higher eukaryotes, where genes may be controlled by many regulatory modules that can be located tens of kilobases distant from the transcription start site. In general, it is assumed that the number of regulatory inputs per gene increases with the complexity of the organism. Correspondingly, the size of intergenic regions increase dramatically with the complexity of eukaryotes. The fundamental limits on regulatory complexity we have found for bacteria may be one of their key differences from eukaryotes.

7.2 Open questions and future work

What are the fundamental principles?

We have found that the *evolutionary potentials* are constants of genome evolution, independent of time and lineage, but what determines the value of these numbers? At this stage it is yet unclear if the scaling laws are determined by physico-chemical constraints or the evolutionary dynamics. Elucidate the origin of these scaling laws will be, in the coming years, a major challenge in theoretical biology. Moreover, we believe that fundamental principles of biology and evolution will be uncovered in order to answer this question.

The scaling law of transcription factors is particularly interesting, due to its implications on the structure of transcriptional regulatory networks and its evolution. It is tempting to imagine that

regulatory network have been optimized to *efficiently* process environmental stimuli under certain structural constraints which are due to the molecular character of the system. In this case the scaling laws will emerge naturally as an optimum solution of a constrained optimization problem.

Towards a general theory of genome evolution

I believe this is a special moment in the history of biology. Experimental techniques have recently changed dramatically, and we are inundated with data. We are probably in a similar situation as physicist in times of Tycho Brahe when they started to have enormous amount of astronomical data. Moreover, as we have seen first phenomenological laws in genome evolution begin to be uncovered as Kepler's ones regarding planetary motion[1]. The distribution of gene family sizes and the scaling laws of functional content of genomes are examples of such evolutionary laws. In my opinion a possible interesting step forward would be to unify both laws under a unique evolutionary model. It is probably time for a general theory of genome evolution.

[1]This analogy was borrowed from professor Erik van Nimwegen

Appendix

Power-law fitting

In several places we fit a power-law of the form $n = e^{\beta} g^{\alpha}$ to a scatter of points, i.e. the number of operons as a function of the number of genes, the number of TFs as a function of the number of genes, the number of TF clusters as a function of the number of genes, etc. To perform these fits we have used a Bayesian model. First, we log-transform all the data points, i.e. $(x_i, y_i) = (\log[n_i], \log[g_i])$ discarding all data points with zero counts, i.e. $n_i = 0$. We assume that the pairs (x_i, y_i) derive from a line $y_i = \alpha x_i + \beta$ plus noise of unknown size in both x- and y-direction. In addition we assume a rotationally invariant prior for the slope α. Under these assumptions the posterior probability density for the slope α given the data D after integrating out the size of the noise and the offset is given by,

$$P(\alpha|D)d\alpha \propto \frac{\left(\alpha^2+1\right)^{(N-3)/2}}{\left(\sigma_{yy} + \sigma_{xx}\alpha^2 - 2\alpha\sigma_{xy}\right)^{(N-1)/2}} d\alpha, \tag{7.1}$$

where N is the number of data points, σ_{xx} is the variance of x values, σ_{yy} the variance of y values, and σ_{xy} the covariance of x and y values. Then, we obtain the slope α^*that maximizes [7.1] and the 99% posterior probability interval $[\alpha_{min}, \alpha_{max}]$ numerically.

The posterior distribution of the offset β can be easily obtained applying the following relationship between this parameter and the slope α given by,

$$\langle y \rangle = \alpha \langle x \rangle + \beta \tag{7.2}$$

where $\langle y \rangle$ and $\langle x \rangle$ are the averages values of the variables y and x respectively. Therefore, the offset that maximize the posterior distribution is $\beta^* = \langle y \rangle - \alpha^* \langle x \rangle$ and similarly for the 99%posterior probability interval $[\beta_{min}, \beta_{max}]$.

Note that the optimal line, given by the slope α^* and the offset β^*, obtained by this procedure corresponds to the first principal component of the data.

Smoothed profiles

All the position dependent profiles that are shown in figures of this thesis were smoothed to reduce fluctuations on short distance scales. We produced the smoothed profiles $\overline{S}(x)$ of a statistic S using a double-exponential kernel:

$$\overline{S}(x) = \frac{1}{N} \sum_y S(x-y) e^{-\frac{|x-y|}{\alpha}} \tag{7.3}$$

where N is a normalization factor,

$$N = \sum_y e^{-\frac{|x-y|}{\alpha}} \tag{7.4}$$

and α is a length-scale which, for this study, was set to 3. In order to avoid the mixture of the statistics in intergenic or coding regions, special boundaries were take into account for summing in (7.3) and (7.4). That is, for calculating the smoothed statistic $\overline{S}(x)$ at a position x that lies within intergenic, the sum on the right runs only over positions y that are in intergenic as well. Similarly, for calculating the smoothed statistic $\overline{S}(x)$ at a position x within the coding region, the sum on the right runs only over positions y that are in the coding region as well.

List of results for selected functional categories

Fitted overall exponents α_c and offsets β_c for different functional categories c and the Z-scores which measure the deviation of the clade-dependent exponents (offsets) from the overall ones:

GO term	α_c	β_c	Z_α	Z_β
ribosome	0.03 ± 0.01	3.8 ± 0.1	0.77	0.77
structural constituent of ribosome	0.03 ± 0.01	3.8 ± 0.1	0.77	0.77
ribonucleoprotein complex	0.04 ± 0.01	3.7 ± 0.1	0.81	0.82
intracellular non membrane bounded organelle	0.08 ± 0.01	3.7 ± 0.1	1.24	1.23
non membrane bounded organelle	0.08 ± 0.01	3.7 ± 0.1	1.24	1.23
translation	0.08 ± 0.01	4.1 ± 0.1	1.16	1.12
cytoplasmic part	0.10 ± 0.02	3.5 ± 0.1	1.00	0.96
aminoacyl tRNA ligase activity	0.11 ± 0.01	2.7 ± 0.1	1.82	1.83
ribonucleoprotein complex biogenesis and assembly	0.11 ± 0.02	1.3 ± 0.2	1.00	0.96
tRNA aminoacylation for protein translation	0.12 ± 0.01	2.6 ± 0.1	1.85	1.90
ligase activity forming aminoacyl tRNA and related compounds	0.12 ± 0.01	2.7 ± 0.1	1.81	1.81
ligase activity forming carbon oxygen bonds	0.12 ± 0.01	2.7 ± 0.1	1.81	1.81
amino acid activation	0.12 ± 0.01	2.6 ± 0.1	1.86	1.88
tRNA aminoacylation	0.12 ± 0.01	2.6 ± 0.1	1.86	1.88
RNA polymerase activity	0.13 ± 0.02	1.6 ± 0.2	0.65	0.63
DNA directed RNA polymerase activity	0.13 ± 0.02	1.6 ± 0.2	0.65	0.63
structural molecule activity	0.13 ± 0.02	3.1 ± 0.1	0.70	0.67
tRNA metabolic process	0.17 ± 0.02	2.6 ± 0.1	1.71	1.69
intracellular organelle	0.18 ± 0.02	3.1 ± 0.2	0.95	0.97
organelle	0.18 ± 0.02	3.1 ± 0.2	0.96	0.98
gene expression	0.24 ± 0.01	3.3 ± 0.1	2.04	1.99
macromolecular complex	0.24 ± 0.02	2.8 ± 0.2	0.83	0.77
cytoplasm	0.26 ± 0.02	2.9 ± 0.1	1.06	1.01
macromolecule biosynthetic process	0.28 ± 0.02	2.9 ± 0.1	1.87	1.85
guanyl ribonucleotide binding	0.29 ± 0.02	1.5 ± 0.2	1.04	1.02
guanyl nucleotide binding	0.29 ± 0.02	1.5 ± 0.2	1.04	1.02
GTP binding	0.29 ± 0.02	1.5 ± 0.2	1.04	1.02
RNA binding	0.29 ± 0.02	2.0 ± 0.2	2.15	2.04
intracellular part	0.30 ± 0.02	2.8 ± 0.1	1.24	1.18
nucleotidyltransferase activity	0.37 ± 0.03	0.7 ± 0.2	0.87	0.81
cellular protein metabolic process	0.37 ± 0.02	2.5 ± 0.1	1.73	1.69
ligase activity	0.37 ± 0.02	1.5 ± 0.1	1.91	1.88
protein metabolic process	0.38 ± 0.02	2.5 ± 0.1	1.61	1.56
GTPase activity	0.38 ± 0.03	-0.7 ± 0.2	1.36	1.34
cellular macromolecule metabolic process	0.39 ± 0.02	2.4 ± 0.1	1.68	1.65
cellular biosynthetic process	0.40 ± 0.01	2.4 ± 0.1	2.28	2.37
nuclease activity	0.41 ± 0.03	0.5 ± 0.2	1.47	1.42
RNA metabolic process	0.47 ± 0.03	1.0 ± 0.2	2.47	2.45
biosynthetic process	0.53 ± 0.02	1.8 ± 0.1	2.06	2.11
macromolecule metabolic process	0.53 ± 0.02	2.0 ± 0.2	0.97	0.94
nucleoside phosphate metabolic process	0.59 ± 0.03	-1.1 ± 0.2	0.98	0.98
nucleotide metabolic process	0.59 ± 0.03	-1.1 ± 0.2	0.98	0.98
amino acid metabolic process	0.60 ± 0.03	-0.1 ± 0.2	3.02	3.09
amino acid and derivative metabolic process	0.60 ± 0.03	-0.1 ± 0.2	3.11	3.17
primary metabolic process	0.61 ± 0.02	1.8 ± 0.1	0.84	0.79
cellular metabolic process	0.63 ± 0.02	1.7 ± 0.1	0.93	0.91
intracellular	0.63 ± 0.02	0.7 ± 0.2	1.65	1.68
amine metabolic process	0.63 ± 0.03	-0.3 ± 0.2	2.90	2.96
carboxylic acid metabolic process	0.66 ± 0.03	-0.3 ± 0.2	3.30	3.36
organic acid metabolic process	0.66 ± 0.03	-0.4 ± 0.2	3.31	3.37
ribonucleotide binding	0.67 ± 0.02	0.6 ± 0.2	1.25	1.26
purine ribonucleotide binding	0.67 ± 0.02	0.6 ± 0.2	1.25	1.26
nitrogen compound metabolic process	0.68 ± 0.03	-0.6 ± 0.2	2.61	2.68
purine nucleotide binding	0.68 ± 0.02	0.6 ± 0.2	1.25	1.26
nucleotide binding	0.70 ± 0.02	0.5 ± 0.2	1.21	1.22

hydrolase activity acting on ester bonds	0.71 ± 0.03	-1.4 ± 0.2	1.08	1.01
ATP binding	0.74 ± 0.02	-0.0 ± 0.2	1.30	1.33
adenyl ribonucleotide binding	0.74 ± 0.02	-0.0 ± 0.2	1.30	1.33
adenyl nucleotide binding	0.75 ± 0.02	-0.0 ± 0.2	1.30	1.32
cellular process	0.76 ± 0.02	1.0 ± 0.1	1.15	1.07
transferase activity transferring phosphorus containing groups	0.78 ± 0.03	-1.4 ± 0.3	1.46	1.42
metabolic process	0.79 ± 0.01	0.8 ± 0.1	1.15	1.16
response to stress	0.83 ± 0.04	-2.5 ± 0.3	1.60	1.60
transferase activity	0.88 ± 0.02	-1.3 ± 0.1	1.19	1.14
transferase activity transferring one carbon groups	0.91 ± 0.04	-3.1 ± 0.3	1.26	1.24
binding	0.91 ± 0.01	-0.1 ± 0.1	0.85	0.87
methyltransferase activity	0.92 ± 0.04	-3.3 ± 0.3	1.18	1.17
cellular catabolic process	0.93 ± 0.04	-3.5 ± 0.3	1.34	1.37
hydrolase activity	0.93 ± 0.03	-1.4 ± 0.2	1.14	1.20
cellular component	0.95 ± 0.03	-0.8 ± 0.2	1.43	1.38
cell	0.95 ± 0.03	-0.8 ± 0.2	1.48	1.43
cell part	0.95 ± 0.03	-0.8 ± 0.2	1.48	1.43
catabolic process	0.96 ± 0.04	-3.7 ± 0.3	1.30	1.33
nucleic acid binding	0.97 ± 0.03	-1.5 ± 0.2	1.06	1.08
catalytic activity	0.97 ± 0.01	-0.4 ± 0.1	1.65	1.59
biological process	0.98 ± 0.01	-0.2 ± 0.0	0.96	0.93
molecular function	1.00 ± 0.00	-0.1 ± 0.0	1.65	1.54
membrane part	1.00 ± 0.05	-3.1 ± 0.4	1.34	1.29
transition metal ion binding	1.07 ± 0.04	-4.2 ± 0.3	0.85	0.82
response to stimulus	1.09 ± 0.04	-4.1 ± 0.3	1.13	1.16
cation binding	1.11 ± 0.04	-4.4 ± 0.3	1.01	0.95
ion binding	1.12 ± 0.04	-4.1 ± 0.3	0.93	0.89
metal ion binding	1.12 ± 0.04	-4.2 ± 0.3	0.99	0.96
lyase activity	1.14 ± 0.05	-5.1 ± 0.4	2.45	2.54
kinase activity	1.14 ± 0.05	-4.9 ± 0.4	1.46	1.44
integral to membrane	1.20 ± 0.05	-5.0 ± 0.4	1.94	1.87
intrinsic to membrane	1.20 ± 0.05	-5.0 ± 0.4	1.92	1.86
membrane	1.22 ± 0.03	-3.6 ± 0.3	1.09	1.01
oxidoreductase activity acting on CH OH group of donors	1.25 ± 0.06	-6.3 ± 0.4	2.27	2.32
establishment of localization	1.26 ± 0.05	-4.2 ± 0.4	1.37	1.37
localization	1.27 ± 0.05	-4.3 ± 0.4	1.30	1.30
DNA binding	1.27 ± 0.04	-4.1 ± 0.3	1.22	1.27
coenzyme binding	1.27 ± 0.04	-5.7 ± 0.3	1.97	2.04
oxidoreductase activity acting on the CH OH group of donors NAD or NADP as acceptor	1.27 ± 0.06	-6.6 ± 0.5	2.75	2.81
transport	1.28 ± 0.06	-4.5 ± 0.4	1.34	1.36
transporter activity	1.29 ± 0.06	-4.7 ± 0.4	1.38	1.40
hydrolase activity acting on carbon nitrogen but not peptide bonds	1.32 ± 0.06	-7.0 ± 0.4	1.30	1.27
cofactor binding	1.40 ± 0.04	-6.3 ± 0.3	1.96	2.00
acyltransferase activity	1.44 ± 0.07	-7.7 ± 0.5	1.32	1.28
transferase activity transferring acyl groups	1.44 ± 0.06	-7.5 ± 0.5	1.44	1.42
transferase activity transferring groups other than amino acyl groups	1.45 ± 0.07	-7.7 ± 0.5	1.36	1.33
phosphotransferase activity alcohol group as acceptor	1.55 ± 0.08	-8.3 ± 0.6	1.52	1.53
oxidoreductase activity	1.56 ± 0.06	-6.5 ± 0.5	2.92	2.99
pyridoxal phosphate binding	1.59 ± 0.08	-9.2 ± 0.6	2.41	2.44
transferase activity transferring nitrogenous groups	1.63 ± 0.09	-10.0 ± 0.7	2.00	2.02
biological regulation	1.70 ± 0.03	-7.8 ± 0.2	0.90	0.91
vitamin binding	1.77 ± 0.08	-10.1 ± 0.6	2.93	2.98
regulation of biological process	1.77 ± 0.03	-8.4 ± 0.3	0.93	0.95
regulation of cellular process	1.80 ± 0.04	-8.6 ± 0.3	0.96	0.97
regulation of gene expression	1.84 ± 0.04	-9.0 ± 0.3	1.00	1.01
regulation of metabolic process	1.84 ± 0.04	-9.0 ± 0.3	0.91	0.91
regulation of cellular metabolic process	1.86 ± 0.04	-9.2 ± 0.3	0.99	0.99

regulation of nucleobase nucleoside nucleotide and nucleic acid metabolic process	1.88 ± 0.04	-9.3 ± 0.3	1.00	1.00
transcription regulator activity	1.89 ± 0.04	-9.5 ± 0.3	0.94	0.94
regulation of transcription	1.91 ± 0.04	-9.6 ± 0.3	1.05	1.04
regulation of transcription DNA dependent	1.95 ± 0.04	-9.9 ± 0.3	1.07	1.07
regulation of RNA metabolic process	1.95 ± 0.04	-9.9 ± 0.3	1.07	1.07
transcription factor activity	2.08 ± 0.06	-11.4 ± 0.4	1.03	1.08
protein kinase activity	2.09 ± 0.11	-13.0 ± 0.8	1.58	1.56
phosphotransferase activity nitrogenous group as acceptor	2.10 ± 0.10	-13.2 ± 0.8	1.21	1.22
cell communication	2.13 ± 0.11	-11.9 ± 0.8	1.49	1.44
signal transduction	2.17 ± 0.11	-12.1 ± 0.9	1.52	1.46
protein histidine kinase activity	2.19 ± 0.12	-13.9 ± 0.9	1.34	1.33
two component sensor activity	2.19 ± 0.12	-13.9 ± 0.9	1.34	1.33
two component signal transduction system phosphorelay	2.23 ± 0.10	-13.5 ± 0.8	1.47	1.44
acetyltransferase activity	2.25 ± 0.14	-14.6 ± 1.0	1.53	1.45
N acetyltransferase activity	2.27 ± 0.15	-14.8 ± 1.1	1.58	1.52
N acyltransferase activity	2.28 ± 0.15	-14.9 ± 1.1	1.58	1.52
two component response regulator activity	2.30 ± 0.12	-14.2 ± 0.9	1.75	1.68
response to chemical stimulus	2.44 ± 0.16	-15.7 ± 1.2	1.05	0.99
molecular transducer activity	2.54 ± 0.13	-15.2 ± 1.0	2.05	1.90
signal transducer activity	2.54 ± 0.13	-15.2 ± 1.0	2.05	1.90
sequence specific DNA binding	2.56 ± 0.12	-16.5 ± 0.9	0.77	0.81

Fitted overall exponents α_c, gamma exponenents γ_c and evolutionary potentials ρ_c for different functional categories c:

GO term	α_c	γ_c	ρ_c
ribosome	0.03 ± 0.01	1.4 ± 0.6	0.11 ± 0.02
structural constituent of ribosome	0.03 ± 0.01	1.4 ± 0.6	0.11 ± 0.02
ribonucleoprotein complex	0.04 ± 0.01	1.4 ± 0.6	0.11 ± 0.02
non membrane bounded organelle	0.08 ± 0.01	1.5 ± 0.5	0.10 ± 0.01
intracellular non membrane bounded organelle	0.08 ± 0.01	1.5 ± 0.5	0.10 ± 0.01
translation	0.08 ± 0.01	1.2 ± 0.4	0.13 ± 0.01
cytoplasmic part	0.10 ± 0.02	1.3 ± 0.5	0.18 ± 0.02
aminoacyl tRNA ligase activity	0.11 ± 0.01	1.6 ± 0.6	0.15 ± 0.02
ribonucleoprotein complex biogenesis and assembly	0.11 ± 0.02	0.5 ± 0.9	0.21 ± 0.09
tRNA aminoacylation for protein translation	0.12 ± 0.01	1.6 ± 0.6	0.15 ± 0.03
ligase activity forming carbon oxygen bonds	0.12 ± 0.01	1.6 ± 0.6	0.15 ± 0.02
ligase activity forming aminoacyl tRNA and related compounds	0.12 ± 0.01	1.6 ± 0.6	0.15 ± 0.02
tRNA aminoacylation	0.12 ± 0.01	1.6 ± 0.6	0.16 ± 0.02
amino acid activation	0.12 ± 0.01	1.6 ± 0.6	0.16 ± 0.02
RNA polymerase activity	0.13 ± 0.02	1.8 ± 1.2	0.37 ± 0.07
DNA directed RNA polymerase activity	0.13 ± 0.02	1.8 ± 1.2	0.37 ± 0.07
structural molecule activity	0.13 ± 0.02	0.9 ± 0.6	0.26 ± 0.02
tRNA metabolic process	0.17 ± 0.02	1.7 ± 0.7	0.16 ± 0.02
intracellular organelle	0.18 ± 0.02	1.4 ± 0.4	0.22 ± 0.02
organelle	0.18 ± 0.02	1.4 ± 0.4	0.22 ± 0.02
translation factor activity nucleic acid binding	0.21 ± 0.03	1.7 ± 0.7	0.31 ± 0.07
translation regulator activity	0.23 ± 0.03	1.6 ± 0.6	0.34 ± 0.07
gene expression	0.24 ± 0.01	1.2 ± 0.4	0.19 ± 0.01
macromolecular complex	0.24 ± 0.02	1.4 ± 0.4	0.30 ± 0.02
cytoplasm	0.26 ± 0.02	1.0 ± 0.4	0.42 ± 0.02
macromolecule biosynthetic process	0.28 ± 0.02	0.9 ± 0.4	0.28 ± 0.01
guanyl ribonucleotide binding	0.29 ± 0.02	0.9 ± 0.7	0.20 ± 0.03
guanyl nucleotide binding	0.29 ± 0.02	0.9 ± 0.7	0.20 ± 0.03
GTP binding	0.29 ± 0.02	0.9 ± 0.7	0.20 ± 0.03
RNA binding	0.29 ± 0.02	1.8 ± 0.6	0.20 ± 0.02

intracellular part	0.30 ± 0.02	1.2 ± 0.5	0.54 ± 0.02
RNA processing	0.32 ± 0.03	2.2 ± 0.7	0.14 ± 0.03
nucleotidyltransferase activity	0.37 ± 0.03	1.6 ± 0.5	0.39 ± 0.04
cellular protein metabolic process	0.37 ± 0.02	0.9 ± 0.3	0.38 ± 0.01
ligase activity	0.37 ± 0.02	1.4 ± 0.4	0.30 ± 0.02
protein metabolic process	0.38 ± 0.02	0.8 ± 0.3	0.39 ± 0.01
GTPase activity	0.38 ± 0.03	0.5 ± 0.8	0.23 ± 0.06
cellular macromolecule metabolic process	0.39 ± 0.02	0.8 ± 0.3	0.41 ± 0.01
cellular biosynthetic process	0.40 ± 0.01	1.5 ± 0.4	0.36 ± 0.01
nuclease activity	0.41 ± 0.03	1.5 ± 0.5	0.35 ± 0.03
RNA metabolic process	0.47 ± 0.03	1.8 ± 0.7	0.28 ± 0.02
isomerase activity	0.49 ± 0.03	1.8 ± 0.5	0.40 ± 0.03
biosynthetic process	0.53 ± 0.02	1.4 ± 0.3	0.43 ± 0.01
macromolecule metabolic process	0.53 ± 0.02	1.2 ± 0.3	0.53 ± 0.01
nucleobase nucleoside nucleotide and nucleic acid metabolic process	0.58 ± 0.03	1.6 ± 0.5	0.55 ± 0.01
nucleotide metabolic process	0.59 ± 0.03	2.1 ± 0.7	0.53 ± 0.04
nucleoside phosphate metabolic process	0.59 ± 0.03	2.1 ± 0.7	0.53 ± 0.04
amino acid metabolic process	0.60 ± 0.03	1.9 ± 0.6	0.42 ± 0.02
amino acid and derivative metabolic process	0.60 ± 0.03	1.8 ± 0.6	0.42 ± 0.02
nucleobase nucleoside and nucleotide metabolic process	0.61 ± 0.04	1.5 ± 0.7	0.45 ± 0.03
primary metabolic process	0.61 ± 0.02	1.4 ± 0.3	0.59 ± 0.01
nucleobase nucleoside and nucleotide biosynthetic process	0.61 ± 0.03	2.0 ± 0.7	0.50 ± 0.04
cellular metabolic process	0.63 ± 0.02	1.5 ± 0.3	0.59 ± 0.01
intracellular	0.63 ± 0.02	0.3 ± 0.6	0.80 ± 0.02
amine metabolic process	0.63 ± 0.03	1.8 ± 0.6	0.48 ± 0.02
biopolymer metabolic process	0.64 ± 0.04	1.5 ± 0.5	0.60 ± 0.01
carboxylic acid metabolic process	0.66 ± 0.03	1.8 ± 0.6	0.45 ± 0.02
organic acid metabolic process	0.66 ± 0.03	1.8 ± 0.6	0.46 ± 0.02
ribonucleotide binding	0.67 ± 0.02	0.8 ± 0.6	0.75 ± 0.01
purine ribonucleotide binding	0.67 ± 0.02	0.8 ± 0.6	0.75 ± 0.01
purine nucleotide binding	0.68 ± 0.02	0.8 ± 0.6	0.74 ± 0.01
nitrogen compound metabolic process	0.68 ± 0.03	1.5 ± 0.6	0.51 ± 0.02
nucleotide binding	0.70 ± 0.02	0.8 ± 0.6	0.74 ± 0.01
hydrolase activity acting on ester bonds	0.71 ± 0.03	1.0 ± 0.5	0.63 ± 0.03
ATP binding	0.74 ± 0.03	0.9 ± 0.6	0.82 ± 0.02
adenyl ribonucleotide binding	0.74 ± 0.03	0.9 ± 0.6	0.82 ± 0.02
adenyl nucleotide binding	0.75 ± 0.02	0.9 ± 0.6	0.81 ± 0.02
cellular process	0.76 ± 0.02	1.1 ± 0.5	0.82 ± 0.01
transferase activity transferring phosphorus containing groups	0.78 ± 0.03	1.3 ± 0.6	0.72 ± 0.03
metabolic process	0.79 ± 0.01	1.6 ± 0.3	0.73 ± 0.01
response to stress	0.83 ± 0.04	1.5 ± 0.7	0.56 ± 0.04
transferase activity	0.88 ± 0.02	1.6 ± 1.1	0.82 ± 0.02
transferase activity transferring one carbon groups	0.91 ± 0.04	0.7 ± 0.9	0.67 ± 0.04
binding	0.91 ± 0.01	1.1 ± 0.8	0.93 ± 0.01
peptidase activity	0.92 ± 0.05	1.3 ± 0.5	0.77 ± 0.04
methyltransferase activity	0.92 ± 0.04	1.0 ± 1.0	0.67 ± 0.04
cellular catabolic process	0.93 ± 0.04	1.2 ± 0.8	0.82 ± 0.04
proteolysis	0.93 ± 0.05	1.3 ± 0.5	0.76 ± 0.04
hydrolase activity	0.93 ± 0.03	1.3 ± 0.6	0.87 ± 0.02
cell part	0.95 ± 0.03	1.2 ± 0.6	1.08 ± 0.01
cellular component	0.95 ± 0.03	1.2 ± 0.6	1.08 ± 0.01
cell	0.95 ± 0.03	1.2 ± 0.6	1.08 ± 0.01
catabolic process	0.96 ± 0.04	1.1 ± 0.8	0.84 ± 0.04
nucleic acid binding	0.97 ± 0.03	1.5 ± 1.0	0.93 ± 0.01
pyrophosphatase activity	0.97 ± 0.05	1.2 ± 0.8	0.92 ± 0.03
hydrolase activity acting on acid anhydrides in phosphorus containing anhydrides	0.97 ± 0.05	1.2 ± 0.8	0.92 ± 0.03
catalytic activity	0.97 ± 0.01	2.1 ± 0.8	0.82 ± 0.01
biological process	0.98 ± 0.01	0.4 ± 1.1	0.97 ± 0.01
nucleoside triphosphatase activity	0.99 ± 0.05	1.2 ± 0.8	0.93 ± 0.03

molecular function	1.00 ± 0.00	1.5 ± 1.1	0.99 ± 0.00
hydrolase activity acting on acid anhydrides	1.00 ± 0.05	1.2 ± 0.7	0.92 ± 0.03
known function	1.00 ± 0.00	0.0 ± 0.0	1.00 ± 0.00
membrane part	1.00 ± 0.05	1.2 ± 0.6	0.88 ± 0.03
transition metal ion binding	1.07 ± 0.04	2.0 ± 1.0	1.00 ± 0.04
monocarboxylic acid metabolic process	1.07 ± 0.06	2.2 ± 1.0	0.55 ± 0.06
ATPase activity	1.08 ± 0.06	1.0 ± 0.7	0.99 ± 0.03
response to stimulus	1.09 ± 0.04	2.1 ± 1.2	1.37 ± 0.04
cation binding	1.11 ± 0.04	1.9 ± 1.0	1.07 ± 0.04
ion binding	1.12 ± 0.04	1.6 ± 1.1	1.01 ± 0.03
metal ion binding	1.12 ± 0.04	2.0 ± 1.1	1.01 ± 0.04
coenzyme metabolic process	1.13 ± 0.06	0.8 ± 0.9	0.54 ± 0.05
lyase activity	1.14 ± 0.05	0.8 ± 0.8	0.88 ± 0.04
kinase activity	1.14 ± 0.05	1.2 ± 0.9	0.90 ± 0.04
FAD binding	1.18 ± 0.06	1.7 ± 0.9	0.63 ± 0.07
integral to membrane	1.20 ± 0.05	1.2 ± 0.5	0.99 ± 0.04
intrinsic to membrane	1.20 ± 0.05	1.2 ± 0.5	1.00 ± 0.04
membrane	1.22 ± 0.03	1.4 ± 0.5	1.26 ± 0.02
oxidoreductase activity acting on CH OH group of donors	1.25 ± 0.06	0.4 ± 0.9	0.73 ± 0.05
establishment of localization	1.26 ± 0.05	1.2 ± 0.4	1.12 ± 0.02
localization	1.27 ± 0.05	1.1 ± 0.4	1.12 ± 0.02
DNA binding	1.27 ± 0.04	1.6 ± 0.6	1.15 ± 0.02
coenzyme binding	1.27 ± 0.04	0.2 ± 0.8	0.76 ± 0.04
oxidoreductase activity acting on the CH OH group of donors NAD or NADP as acceptor	1.27 ± 0.06	-0.2 ± 1.1	0.69 ± 0.05
transport	1.28 ± 0.06	1.2 ± 0.4	1.14 ± 0.02
transporter activity	1.29 ± 0.06	1.3 ± 0.4	1.07 ± 0.02
hydrolase activity acting on carbon nitrogen but not peptide bonds	1.32 ± 0.06	0.1 ± 0.6	0.96 ± 0.06
aromatic compound metabolic process	1.34 ± 0.07	0.9 ± 0.9	0.65 ± 0.04
cofactor binding	1.40 ± 0.04	0.7 ± 0.6	1.21 ± 0.04
acyltransferase activity	1.44 ± 0.07	0.9 ± 0.7	1.53 ± 0.07
transferase activity transferring acyl groups	1.44 ± 0.06	0.7 ± 0.7	1.43 ± 0.06
transferase activity transferring groups other than amino acyl groups	1.45 ± 0.07	0.8 ± 0.7	1.50 ± 0.06
phosphoric ester hydrolase activity	1.47 ± 0.08	0.7 ± 0.5	0.93 ± 0.10
phosphotransferase activity alcohol group as acceptor	1.55 ± 0.08	1.1 ± 0.6	1.08 ± 0.05
oxidoreductase activity	1.56 ± 0.06	1.1 ± 0.4	1.02 ± 0.02
pyridoxal phosphate binding	1.59 ± 0.08	-0.8 ± 0.7	0.69 ± 0.06
transferase activity transferring nitrogenous groups	1.63 ± 0.09	0.2 ± 0.8	0.68 ± 0.07
biological regulation	1.70 ± 0.03	1.0 ± 0.2	1.35 ± 0.02
vitamin binding	1.77 ± 0.08	-0.7 ± 0.5	0.77 ± 0.05
regulation of biological process	1.77 ± 0.03	1.0 ± 0.2	1.39 ± 0.02
regulation of cellular process	1.80 ± 0.04	1.0 ± 0.2	1.40 ± 0.02
regulation of gene expression	1.84 ± 0.04	0.9 ± 0.2	1.43 ± 0.02
regulation of metabolic process	1.84 ± 0.04	1.0 ± 0.2	1.41 ± 0.02
regulation of cellular metabolic process	1.86 ± 0.04	1.0 ± 0.2	1.41 ± 0.02
regulation of nucleobase nucleoside nucleotide and nucleic acid metabolic process	1.88 ± 0.04	0.9 ± 0.2	1.42 ± 0.02
transcription regulator activity	1.89 ± 0.04	1.0 ± 0.2	1.47 ± 0.03
FMN binding	1.91 ± 0.12	0.4 ± 0.4	0.83 ± 0.10
regulation of transcription	1.91 ± 0.04	0.9 ± 0.2	1.43 ± 0.02
regulation of RNA metabolic process	1.95 ± 0.04	0.9 ± 0.2	1.45 ± 0.03
regulation of transcription DNA dependent	1.95 ± 0.04	0.9 ± 0.2	1.45 ± 0.03
transcription factor activity	2.08 ± 0.06	0.7 ± 0.3	1.41 ± 0.03
protein kinase activity	2.09 ± 0.11	0.6 ± 0.4	1.35 ± 0.07
phosphotransferase activity nitrogenous group as acceptor	2.10 ± 0.10	0.8 ± 0.5	1.32 ± 0.07
cell communication	2.13 ± 0.11	1.0 ± 0.4	1.96 ± 0.04
signal transduction	2.17 ± 0.11	1.0 ± 0.4	1.98 ± 0.04
protein histidine kinase activity	2.19 ± 0.12	0.7 ± 0.5	1.36 ± 0.08

two component sensor activity	2.19 ± 0.12	0.7 ± 0.5	1.36 ± 0.08
electron carrier activity	2.20 ± 0.16	0.8 ± 0.3	1.21 ± 0.04
two component signal transduction system phosphorelay	2.23 ± 0.10	0.9 ± 0.5	1.81 ± 0.06
acetyltransferase activity	2.25 ± 0.14	0.8 ± 0.5	2.06 ± 0.10
N acetyltransferase activity	2.27 ± 0.15	0.6 ± 0.5	2.15 ± 0.10
N acyltransferase activity	2.28 ± 0.14	0.5 ± 0.5	2.12 ± 0.10
two component response regulator activity	2.30 ± 0.12	0.9 ± 0.6	1.88 ± 0.07
response to chemical stimulus	2.44 ± 0.16	0.6 ± 0.3	2.52 ± 0.09
molecular transducer activity	2.54 ± 0.12	1.2 ± 0.4	2.10 ± 0.04
signal transducer activity	2.54 ± 0.12	1.2 ± 0.4	2.10 ± 0.04
sequence specific DNA binding	2.56 ± 0.12	0.7 ± 0.5	1.95 ± 0.07

List of genome pairs

Pairs of closely-related spacies with their phylogenetic distances that were used to estimate addition/deletion rates:

Genome 1	Genome 2	Distance
Bacillus anthracis Ames	Bacillus cereus ATCC14579	0.012
Helicobacter pylori 26695	Helicobacter pylori J99	0.015
Salmonella enterica Choleraesuis	Escherichia coli 536	0.017
Xanthomonas campestris	Xanthomonas citri	0.020
Xylella fastidiosa M12	Xylella fastidiosa	0.021
Vibrio vulnificus YJ016	Vibrio parahaemolyticus	0.032
Streptococcus agalactiae 2603	Streptococcus pyogenes MGAS315	0.042
Chlamydia muridarum	Chlamydia trachomatis	0.046
Streptomyces coelicolor	Streptomyces avermitilis	0.054
Vibrio cholerae	Vibrio vulnificus YJ016	0.058
Staphylococcus aureus N315	Staphylococcus epidermidis RP62A	0.058
Pseudomonas putida F1	Pseudomonas syringae	0.059
Corynebacterium efficiens YS-314	Corynebacterium glutamicum	0.060
Vibrio cholerae	Vibrio parahaemolyticus	0.065
Mycobacterium avium paratuberculosis	Mycobacterium bovis	0.067
Rickettsia conorii	Rickettsia prowazekii	0.067
Pasteurella multocida	Haemophilus influenzae	0.072
Streptococcus agalactiae 2603	Streptococcus mutans	0.074
Streptococcus pyogenes MGAS315	Streptococcus mutans	0.081
Mycobacterium bovis	Mycobacterium leprae	0.082
Yersinia pestis biovar Mediaevails	Photorhabdus luminescens	0.085
Escherichia coli 536	Yersinia pestis biovar Mediaevails	0.085
Salmonella enterica Choleraesuis	Yersinia pestis biovar Mediaevails	0.086
Mycobacterium avium paratuberculosis	Mycobacterium leprae	0.087
Streptococcus pneumoniae R6	Streptococcus agalactiae 2603	0.088
Pseudomonas aeruginosa	Pseudomonas putida F1	0.088
Pseudomonas aeruginosa	Pseudomonas syringae	0.093
Streptococcus pneumoniae R6	Streptococcus pyogenes MGAS315	0.094
Streptococcus pneumoniae R6	Streptococcus mutans	0.100
Escherichia coli 536	Photorhabdus luminescens	0.102
Salmonella enterica Choleraesuis	Photorhabdus luminescens	0.102
Pasteurella multocida	Haemophilus ducreyi 35000HP	0.103
Rhodopseudomonas palustris BisA53	Bradyrhizobium japonicum	0.105
Haemophilus influenzae	Haemophilus ducreyi 35000HP	0.107
Rhizobium leguminosarum bv viciae 3841	Agrobacterium tumefaciens C58 UWash	0.108
Photobacterium profundum SS9	Vibrio vulnificus YJ016	0.116
Photobacterium profundum SS9	Vibrio cholerae	0.118
Prochlorococcus marinus MIT9313	Synechococcus sp WH8102	0.122
Photobacterium profundum SS9	Vibrio parahaemolyticus	0.123
Corynebacterium diphtheriae	Corynebacterium efficiens YS-314	0.127

Corynebacterium diphtheriae	Corynebacterium glutamicum	0.136
Chlamydophila caviae	Chlamydophila pneumoniae J138	0.164
Xanthomonas campestris	Xylella fastidiosa M12	0.170
Xanthomonas citri	Xylella fastidiosa M12	0.171
Bacillus subtilis	Bacillus anthracis Ames	0.171
Xanthomonas campestris	Xylella fastidiosa	0.175
Xanthomonas citri	Xylella fastidiosa	0.176
Bacillus subtilis	Bacillus cereus ATCC14579	0.177
Brucella melitensis	Rhizobium etli CFN 42	0.182
Chlamydia muridarum	Chlamydophila caviae	0.188
Chlamydia trachomatis	Chlamydophila caviae	0.195
Buchnera aphidicola	Buchnera aphidicola Sg	0.198
Lactococcus lactis	Streptococcus pneumoniae R6	0.199
Prochlorococcus marinus MED4	Prochlorococcus marinus MIT9313	0.205
Lactococcus lactis	Streptococcus agalactiae 2603	0.212
Pasteurella multocida	Escherichia coli 536	0.213
Pasteurella multocida	Salmonella enterica Choleraesuis	0.213
Haemophilus influenzae	Escherichia coli 536	0.216
Haemophilus ducreyi 35000HP	Escherichia coli 536	0.217
Haemophilus influenzae	Salmonella enterica Choleraesuis	0.217
Haemophilus ducreyi 35000HP	Salmonella enterica Choleraesuis	0.217
Bacillus subtilis	Bacillus halodurans	0.218
Prochlorococcus marinus MED4	Synechococcus sp WH8102	0.219
Lactococcus lactis	Streptococcus pyogenes MGAS315	0.219
Rhizobium leguminosarum bv viciae 3841	Brucella melitensis	0.219
Pasteurella multocida	Yersinia pestis biovar Mediaevails	0.219
Mycoplasma pneumoniae	Mycoplasma genitalium	0.222
Haemophilus influenzae	Yersinia pestis biovar Mediaevails	0.223
Haemophilus ducreyi 35000HP	Yersinia pestis biovar Mediaevails	0.223
Chromobacterium violaceum	Neisseria meningitidis FAM18	0.224
Rhizobium leguminosarum bv viciae 3841	Rhizobium etli CFN 42	0.224
Lactococcus lactis	Streptococcus mutans	0.224
Clostridium tetani E88	Clostridium perfringens	0.225
Agrobacterium tumefaciens C58 UWash	Brucella melitensis	0.225
Agrobacterium tumefaciens C58 UWash	Rhizobium etli CFN 42	0.230
Bacillus halodurans	Oceanobacillus iheyensis	0.231
Bacillus anthracis Ames	Bacillus halodurans	0.232
Chlamydia muridarum	Chlamydophila pneumoniae J138	0.234
Pasteurella multocida	Photorhabdus luminescens	0.236
Vibrio vulnificus YJ016	Escherichia coli 536	0.237
Vibrio vulnificus YJ016	Salmonella enterica Choleraesuis	0.238
Bacillus cereus ATCC14579	Bacillus halodurans	0.238
Haemophilus influenzae	Photorhabdus luminescens	0.239
Haemophilus ducreyi 35000HP	Photorhabdus luminescens	0.239
Vibrio cholerae	Escherichia coli 536	0.240
Vibrio cholerae	Salmonella enterica Choleraesuis	0.240
Chlamydia trachomatis	Chlamydophila pneumoniae J138	0.241
Vibrio vulnificus YJ016	Yersinia pestis biovar Mediaevails	0.244
Photobacterium profundum SS9	Escherichia coli 536	0.244
Photobacterium profundum SS9	Salmonella enterica Choleraesuis	0.245
Vibrio parahaemolyticus	Escherichia coli 536	0.245
Vibrio parahaemolyticus	Salmonella enterica Choleraesuis	0.245
Vibrio cholerae	Yersinia pestis biovar Mediaevails	0.246

Publications

This thesis was based on the following papers:

- Nacho Molina and Erik van Nimwegen (prepared for submission). *Scaling laws in the functional content of genomes across bacterial clades*.
- Nacho Molina and Erik van Nimwegen (2008). *The evolution of domain-content in bacterial genomes.* Biology Direct. 2008.
- Nacho Molina and Erik van Nimwegen (2007). *Universal patterns of purifying selection at non-coding positions in bacteria.* Genome Research.
- Michail Pachkov, Ionas Erb, Nacho Molina and Erik van Nimwegen (2006). "S*wissRegulon: a database of genome-wide annotations of regulatory sites*". Nucleic Acids Research.
- Phil Arnold, Ionas Erb, Nacho Molina and Erik van Nimwegen (in preparation). "*MotEvo*: *A Motif Scanner Combining Phylogeny with a New Background Model*".

Curriculum Vitae

Personal Data

Nationality:	Spanish
Birth:	24/04/1979
Address:	77, Avenue de Bâle. 68300 St-Louis (France)
Telephone:	0033 389 697 920
Mobil:	0041 677 177 089
Email:	j.molina@unibas.ch

Research Interests

- **Evolutionary genomics:** I am interesting in the characterization of the basic forces and constrains that take place in the evolution of genomes and that could explain several phenomenological laws that were discovered recently. For example: the scaling laws of the functional content of genomes or the protein family size distributions.
- **Regulatory networks:** I am also interesting in how the maintained scaling laws, in particular how the number of transcription factors grows with genome size, constrain the topology of the regulatory networks.
- **General interests:** statistical mechanics, stochastic processes, probability theory, information theory, graph theory, game theory and econophysics.

Education

2004-2008 **PhD in computational biology.** University of Basel. Switzerland.
Title: *Genome Evolution and Regulatory Networks Structure.*
Supervisor: Erik van Nimwegen.

2003-2004 **Master thesis in theoretical physics**. NIKHEF. Amsterdam. Netherlands.
Title: *Quantum Scalar Field in a Classical Background.*
Supervisor: Jan-Willem van Holten.

2001-2003 **Master in theoretical physics**, Universidad Complutense de Madrid. Spain.

1998-2001 **Bachelor in physics**, Universidad Complutense de Madrid. Spain.

Publications

- Nacho Molina and Erik van Nimwegen (prepared for submission). *Scaling laws in the functional content of genome across bacterial clades.*

- Nacho Molina and Erik van Nimwegen (2008). *The evolution of domain-content in bacterial genomes*. Biology Direct.
- Nacho Molina and Erik van Nimwegen (2007). *Universal patterns of selection at non-coding positions in bacterial*. Genome Research.
- Michail Pachkov, Ionas Erb, Nacho Molina and Erik van Nimwegen (2006). S*wissRegulon: a database of genome-wide annotations of regulatory sites*". Nucleic Acids Research.
- Phil Arnold, Ionas Erb, Nacho Molina and Erik van Nimwegen (in preparation). *MotEvo*: *A Motif Scanner Combining Phylogeny with a New Background Model*.

Conferences and talks

- Contributed talk in the Otto Warburg International Summer School (2006). Berlin: *Comprehensive phylogenetic foot-printing in bacterial genomes*.
- Contributed talk in the conference Biology without borders. Microsoft Research-Center of Computational and System Biology (2007). Trento: *Universal patterns of selection at non-coding positions in bacteria*.
- Contributed talk in the meeting Advances in molecular biology by junior researches abroad (2007). Centro Nacional de Biotecnologia (CNB). Madrid: *Limited complexity in the regulatory networks of bacteria*.
- Contributed talk in the conference Computational and experimental molecular biology. (2008). Max-Delbruck-Centrum for Molecular Medicine. Berlin: *Universal patterns of selection at non-coding positions in bacteria*.
- Invited speaker at:
  - Center for Genomic Regulation. Barcelona. (2007).
  - Seminars of the soft condensed matter group. University of Munich. (2008).
  - Zurich Interaction Seminar on Evolution and Ecology. ETH Zurich. (2008).

Summer schools and special courses

- Otto Warburg International Summer School and Workshop. Max Plank Institute of Molecular Genetics, Berlin.
  - *Networks and Regulation* (2005).
  - *Evolutionary Genomics* (2006).
- Advanced Statistics Summer-school. Escuela Politecnica de la Universidad San Pablo CEU, Madrid
  - *Bayesian Networks* (2008).
  - *Times Series* (2008).
- PhD courses at the University of Basel:
  - Transcription, regulation and gene expression in eukaryotes (2008).
  - Machine learning (2008).
  - Quantitative reasoning with biological data (2007).
  - Computational modeling and simulation (2007).

Computer and Languages skills

- Programming: Perl, C++, Matlab, Mathematica and Python.
- Operating Systems: Linux/Unix and Windows XP.
- Language: Spanish, native. English, fluent. French, basic.

Bibliography

[1] E. van Nimwegen. Scaling laws in the functional content of genomes. *Trends in Genet.*, 19(9):479–484, 2003.

[2] E. van Nimwegen. Scaling laws in the functional content of genomes: Fundamental constants of evolution? In E. Koonin, G. Karev, and Y. Wolf, editors, *Power Laws, Scale-free Networks and Genome Biology*, pages 236–253. Landes Bioscience, 2004.

[3] C. R. Darwin. *On the origin of species by means of natural selection*. J. Murray, London, 1859.

[4] E. Zuckerkandl and L. B. Pauling. Molecular disease, evolution, and genetic heterogeneity. In M. Kasha and B. Pulman, editors, *Horizons in Biochemistry*, pages 189–225. Academic Press, New York, 1962.

[5] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–626, 1968.

[6] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1983.

[7] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[8] M. Nei and T. Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3(5):418–426, 1986.

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

[10] Mark W. J. van Passel, Pradeep Reddy Marri, and Howard Ochman. The emergence and fate of horizontally acquired genes in escherichia coli. *PLoS Comput Biol*, 4(4):e1000059, Apr 2008.

[11] M. Huynen and E. van Nimwegen. The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.*, 15(5):583–589, 1998.

[12] N. M. Luscombe, J. Qian, Z. Zhang, T. Johnson, and M. Gerstein. The dominance of the population by a selected few: power-law behavior applies to a wide variety of genomic properties. *Genome biology*, 3(8), 2002.

[13] E. V. Koonin, Y. I. Wolf, and G. P. Karev. The structure of the protein universe and genome evolution. *Nature*, 420:218–222, 2002.

[14] MA Huynen and Erik van Nimwegen. The frequency distribution of gene family sizes in complete genomes. *Molecular Biology and Evolution*, 15:583–589, 1998.

[15] G. Karev, Y. Wolf, A. Rzhetsky, F. Berezovskaya, and E. Koonin. Birth and death of protein domains: A simple model of evolution explains power law behavior. *BMC Evolutionary Biology*, 2(1):18, 2002.

[16] G. Karev, Y. Wolf, F. Berezovskaya, and E. Koonin. Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. *BMC Evolutionary Biology*, 4(1):32, 2004.

[17] N. Guelzim, S. Bottani, and F. Kepes P. Bourgine. Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics*, 31:60–63, 2002.

[18] G. BalÃ¡zsi, A.-L. BarabÃ¡si, and Z. N. Oltvai. Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 102(22):7841–7846, 2005.

[19] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature genet.*, 31:64–68, 2002.

[20] F.M. Camas, J. BlÃ¡zquez, and J.F. Poyatos. Autogenous and nonautogenous control of response in a genetic network. *Proceedings of the National Academy of Sciences*, 103(34):12718–12723, 2006.

[21] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21):11980–11985, 2003.

[22] U. Alon. Network motifs: theory and experimental approaches. *Nature Review Genetics*, 8(6):450–461, 2007.

[23] Radu Dobrin, Qasim Beg, Albert-Laszlo Barabasi, and Zoltan Oltvai. Aggregation of topological motifs in the escherichia coli transcriptional regulatory network. *BMC Bioinformatics*, 5(1):10, 2004.

[24] J. J. Ward and J. M. Thornton. Evolutionary Models for Formation of Network Motifs and Modularity in the Saccharomyces Transcription Factor Network. *PLOS Computational Biology*, 2007.

[25] M. Lynch. The evolution of genetic networks by non-adaptive processes. *Nature Review Genetics*, 8(10):803–813, 2007.

[26] Hong-Wu Ma, Bharani Kumar, Uta Ditges, Florian Gunzer, Jan Buer, and An-Ping Zeng. An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. *Nucl. Acids Res.*, 32(22):6643–6649, 2004.

[27] C. K. Stover, X. Q. T Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrener, M. J. Hickey, F. S. L. Brinkman, W. O. Hufnagle, D. J. Kowalik, M. Lagrou, R. L. Garber, L. Goltry, E. Tolentino, S. Westbrook-Wadman, Y. Yuan, L. L. Brody, S. N. Coulter, K. R. Folger, A. Kas, K. Larbig, R. M. Lim, K. A. Smith, D.H. D. H. Spencer, G. K.-S. Wong, Z. Wu, I. T. Paulsen, J. Reizer, M. H. Saier, R. E. W. Hancock, S. Lor, and M. V. Olson. Complete genome sequence of pseudomonas aeruginosa pa01, an opportunistic pathogen. *Nature*, 406:959–964, 2000.

[28] S. K. Kummerfeld and S. A. Teichmann. DBD: a transcription factor prediction database. *Nucl. Acids Res.*, 34:D74–D81, 2006.

[29] N. A. Moran, H. E. Dunbar, and J. L. Wilcox. Regulation of transcription in a reduced bacterial genome: nutrient-provisioning genes of the obligate symbiont Buchnera aphidicola. *J. Bacteriol.*, 187(12), 2005.

[30] http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi.

[31] Erik van Nimwegen. Scaling laws in the functional content of genomes. *Trends in Genetics*, 19(9):479–484, 2003.

[32] O. X. Cordero and P. Hogeweg. Large changes in regulome size herald the main prokaryotic lineages. *Trends in genetics*, 23, 2007.

[33] B. Snel, P. Bork, and M. Huynen. Genome evolution. gene fusion versus gene fission. *Trends in genetics*, 16:9–11, 2000.

[34] E. V. Koonin, Y. I. Wolf, and Georgy P. Karev. The structure of the protein universe and genome evolution. *Nature*, 420:218–223, 2002.

[35] Branden C. and Tooze J. *Introduction to Protein Structrue*. Garland Publishing, 1999.

[36] A. Bateman, L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L.L. Sonnhammer, D.J. Studholme, C. Yeats, and S.R. Eddy. The Pfam protein families database. *Nucl. Acids Res.*, 32:D138–D141, 2004.

[37] SR Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–763, 1998.

[38] Michael Ashburner. Gene ontology: tool for the unification of bioloy. *Nature Genetics*, 25:25–29, 2000.

[39] N. Molina and E. van Nimwegen. The evolution of domain-content in bacterial genomes. *Biolgy Direct*, -, 2008.

[40] Nacho Molina and Erik van Nimwegen. Universal patterns of purifying selection and non-coding positions in bacteria. *Genome Research*, 18:148–160, 2007.

[41] F.D. Ciccarelli, T. Doerks, C. von Mering, C.J., B. Snel, and P. Bork. Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science*, 311(5765):1283–1287, 2006.

[42] Jeffrey G. Lawrence Howard Ochman and Eduardo A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, 2000.

[43] J. Peter Gogarten and Jeffrey P. Townsend. Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Micro.*, 3(9):679–687, 2005.

[44] Simone Linz, Achim Radtke, and Arndt von Haeseler. A Likelihood Framework to Measure Horizontal Gene Transfer. *Mol Biol Evol*, 24(6):1312–1319, 2007.

[45] W. W. Navarre, McClelland M, S. J. Libby, and F. C. Fang. Silencing of xenogeneic dna by H-NS—facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. *Genes and Dev.*, 21:1456–1471, 2007.

[46] S. D. Bentley and J . Parkhill. Comparative genomic structure of prokaryotes. *Annual Review of Genetics*, 38(1):771–791, 2004.

[47] M. Pachkov, I. Erb, N. Molina, and E. van Nimwegen. SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucl. Acids Res.*, 35:D127–131, 2007. http://www.swissregulon.unibas.ch.

[48] http://www.ncbi.nlm.nih.gov/genomes/index.htm.

[49] R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, M. Oinn, M. Pagni, F. Servant, C. J. Sigrist, and E. M. Zdobno. The interpro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, 29(1):37–40, 2001.

[50] Paul Kersey, Lawrence Bower, Lorna Morris, Alan Horne, Robert Petryszak, Carola Kanz, Alexander Kanapin, Ujjwal Das, Karine Michoud, Isabelle Phan, Alexandre Gattiker, Tamara Kulikova, Nadeem Faruque, Karyn Duggan, Peter Mclaren, Britt Reimholz, Laurent Duret, Simon Penel, Ingmar Reuter, and Rolf Apweiler. Integr8 and genome reviews: integrated views of complete genomes and proteomes. *Nucl. Acids Res.*, 33:D297–D302, 2005.

[51] H. Salgado, A. Santos-Zavaleta, S. Gama-Castro, D. Millan-Zarate, F.R. Blattner, and J. Collado-Vides. RegulonDB (version 3.0): transcriptional regulation and operon organization in Escherichia coli K-12. *Nucl. Acids Res.*, 28:65–7, 2000.

[52] Yuko Makita, Mitsuteru Nakao, Naotake Ogasawara, and Kenta Nakai. DBTBS: database of transcriptional regulation in bacillus subtilis and its contribution to comparative genomics. *Nucl. Acids Res.*, 32:D75–D77, 2004.

[53] Lee Ann McCue, William Thompson, C. Steven Carmack, Michael P. Ryan, Jun S. Liu, Victoria Derbyshire, and Charles E. Lawrence. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucl. Acids Res.*, 29(3):774–782, 2001.

[54] Lee Ann McCue, William Thompson, C. Steven Carmack, and Charles E. Lawrence. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.*, 12:1523–1532, 2002.

[55] Nikolaus Rajewsky, Nicholas D. Socci, Martin Zapotocky, and Eric D. Siggia. The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Res.*, 12:298–308, 2002.

[56] E. van Nimwegen, M. Zavolan, N. Rajewsky, and E. D. Siggia. Probabilistic clustering of sequences: Inferring new bacterial regulons by comparative genomics. *Proc. Natl. Acad. Sci. USA*, 99:7323–7328, 2002.

[57] Paul Cliften, Priya Sudarsanam, Ashwin Desikan, Lucinda Fulton, Bob Fulton, John Majors, Robert Waterston, Barak A. Cohen, and Mark Johnston. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, 301:71–76, 2003.

[58] Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423:241–254, 2003.

[59] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok DK, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004.

[60] W. Thompson, E. C. Rouchka, and C. E. Lawrence. Gibbs recursive sampler: finding transcription factor binding sites. *Nucl. Acids res.*, 31(13):3580–3585, 2003.

[61] T. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 2:28–36, 1994.

[62] M. Blanchette and M. Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, 12(5):739–748, 2002.

[63] M. Blanchette, B. Schwikowski, and M. Tompa. Algorithms for phylogenetic footprinting. *J. Comput. Biol.*, 9(2):211–223, 2002.

[64] T. Wang and G. Stormo. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, 19(18):2369–2380, 2003.

[65] A. M. Moses, D. Y. Chiang, and M. B. Eisen. Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac. Symp. Biocomput.*, pages 324–335, 2004.

[66] S. Sinha, M. Blanchette, and M. Tompa. PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, 5:170, 2004.

[67] R. Siddharthan, E. D. Siggia, and E. van Nimwegen. Phylogibbs: A gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol*, 1(7):e67, 2005.

[68] K. D. Macisaac, T. Wang, B. D. Gordon, D. K. Gifford, G. D. Stormo, and E. Fraenkel. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, 7:113, 2006.

[69] Ionas Erb and Erik van Nimwegen. Statistical features of yeast's transcriptional regulatory code. In *IEE Proceedings Systems Biology, ICCSB*, 2006.

[70] H. Salgado, S. Gama-Castro, M. Peralta-Gil, E. Diaz-Peredo, F. Sanchez-Solano, A. Santos-Zavaleta, I. Martinez-Flores, V. Jimenez-Jacinto, C. Bonavides-Martinez, J. Segura-Salazar, A. Martinez-Antonio, and J. Collado-Vides. RegulonDB (version 5.0): Escherichia coli k-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucl. Acids Res.*, 34:D394–D397, 2006.

[71] R. Nielsen. Molecular signatures of natural selection. *Annu Rev Genet*, 39:197–218, 2005.

[72] A. Eyre-Walker. The genomic rate of adaptive evolution. *Trends Ecol Evol*, 21(10):569–575, 2006.

[73] D. L. Halligan, A. Eyre-Walker, P. Andolfatto, and P. D Keightley. Patterns of evolutionary constraints in intronic and intergenic DNA of Drosophila. *Genome Res.*, 14:273–279, 2004.

[74] D. L. Halligan and P. D Keightley. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Res.*, 16:875–884, 2006.

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

[76] J. Felsenstein and G. A. Churchill. A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, 13(1):93–104, 1996.

[77] A. Siepel, G. Bejerano, J. S. Pedersen, A. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15:1034–1050, 2005.

[78] O. G. Berg and P. H. von Hippel. Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, 193:723–750, 1987.

[79] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev.*, 15(2):116–124, 2005.

[80] V. Mustonen and M. Lässig. Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. *PNAS*, 102(44):15936–15941, 2005.

[81] A. M Moses, D. Y Chiang, D. A. Pollard, V. N. Iyer, and M. B. Eisen. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.*, 5:R98, 2004.

[82] C. T. Brown and C. G. Callan Jr. Evolutionary comparisons suggest many novel cAMP response protein binding sites in Escherichia coli. *PNAS*, 101(8):2404–2409, 2004.

[83] B. Golding and J. Felsenstein. A maximum likelihood approach to the detection of selection from a phylogeny. *J. Mol. Evol.*, 31:511–523, 1990.

[84] A. L. Halpern and W. J. Bruno. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.*, 5(7):910–917, 1998.

[85] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[86] S. Sinha, E. van Nimwegen, and E. D. Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19 suppl. 1:i292–i301, 2003.

[87] Z. Yang. Paml: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in BioSciences*, 13:555–556, 1997.

[88] D. P. Wall, H. B. Fraser, and A. E. Hirsh. Detecting putative orthologs. *Bioinformatics*, 19(13):1710–1711, 2003.

[89] Wu-blast, 1996-2004. http://blast.wustl.edu.

[90] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.

[91] C. Notredame, D. Higgins, and J. Heringa. T-Coffee: A novel method for multiple sequence alignments. *J. Mol. Biol.*, 302:205–217, 2000.

[92] H. A. Schmidt, K. Strimmer, M. Vingron, and A. von Haeseler. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18:502–504, 2002.

[93] L. Cavalli-Sforza and A. W. F. Edwards. Phylogenetic analysis: models and estimation procedures. *Am. J. Hum. Genet.*, 19:233–257, 1967.

[94] M. Hasegawa, H. Kishino, and T. Yano. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22:160–174, 1985.

[95] J. Shine and L. Dalgarno. The 3'-terminal sequence of Escherichia coli 16s ribosomal RNA: Complementarity to nonsense triplets and ribosome binding sites. *PNAS*, 71(4):1342–1346, 1974.

[96] A. Eyre-Walker and M. Bulmer. Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucl. Acids Res.*, 21(19):4599–4603, 1993.

[97] E. P. Rocha, A. Danchin, and A. Viari. Translation in Bacillus subtilis: roles and trends of initiation and termination, insights from a genome analysis. *Nucl. acids res.*, 27(17):3567–3576, 1999.

[98] C. M. Stenstrom, H. Jin, L. L. Major, W. P. Tate, and L. A. Isaksson. Codon bias at the 3'-side of the initiation codon is correlated with translation initiation efficiency in Escherichia coli. *Gene*, 263:273–284, 2001.

[99] T. Sato, M. Terabe, H. Watanabe, T. Gojobori, C. Hori-Takemoto, and K. Miura. Codon and base biases after the initiation codon of the open reading frames in the Escherichia coli genome and their influence on the translation efficiency. *J. Biochem (Tokyo)*, 129:851–860, 2001.

[100] D. Voges, M. Watzele, C. Nemetz, S. Wizemann, and B. Buchberger. Analyzing and enhancing mrna translational efficiency in an Escherichia coli in vitro expression system. *Biochem Biophys Res Commun*, 318:601–614, 2004.

[101] G. Mitchison. The regional rule for bacterial base composition. *Trends Genet.*, 21(8):440–443, 2005.

[102] P. M. Sharp and W. H. Li. The Codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucl. Acids Res.*, 15(3):1281–1295, 1987.

[103] IL Hofacker, W. Fontana, PF Stadler, LS Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188, 1994.

[104] A. Beletskii and Ashok S. Bhagwat. Transcription-induced mutations: Increase in c to t mutations in the nontranscribed strand during transcription in escherichia coli. *PNAS*, 93:13919–13924, 1996.

[105] H. Ochman. Neutral mutations and neutral substitutions in bacterial genomes. *Mol. Biol. Evol.*, 20(12):2091 – 2096, 2003.

[106] R. D. Knight, S. J. Freeland, and L. F. Landweber. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology*, 2(4):research0010.1–0010.13, 2001.

[107] S. L. Chen, W. Lee, A. K. Hottes, L. Shapiro, and H. H. McAdams. Codon usage between genomes is constrained by genome-wide mutational processes. *PNAS*, 101(10):3480–3485, 2004.

[108] E. P. C. Rocha and A. Danchin. Base composition bias might result from competition for metabolic resources. *Trends Genet.*, 18:291–294, 2002.

[109] Rolf Wagner. *Transcription regulation in prokaryotes*. Oxford University Press, 2000.

[110] J. L. Cherry. Genome size and operon content. *J. theor. Biol.*, 221:401–410, 2003.

[111] M. N. Price, K. H. Huang, E. J. Alm, and A. P. Arkin. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucl Acids Res*, 33:880–892, 2005.

[112] I. B. Rogozin, K. S. Makarova, D. A. Natale, A. N. Spiridonov, R. L. Tatusov, Y. I. Wolf, and E. V. Koonin. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucl. Acids Res.*, 30(19):4264–4271, 2002.

[113] A. Sandelin and W. W. Wasserman. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol.*, 338(2):207–215, 2004.