Chapter 3

Genetic flux over time

3.1 Introduction

The divergence of *Salmonella* and *E. coli* lineages from their common ancestor has been estimated to have occurred approximately 100-140 million years (Myr) ago (Doolittle *et al.*, 1996; Ochman and Wilson, 1987). Using models of amelioration (i.e. the change of the sequence composition over time) to estimate the time of Horizontal Gene Transfer (HGT) events it has been previously inferred (Lawrence and Ochman, 1997) that the entire *E. coli* chromosome contains more than 600 kilobases (kb) of horizontally transferred, protein-coding DNA and that the two sister lineages (*E. coli* and *S. enterica*) have each gained and lost more than 3 megabases (Mb) of novel DNA since their divergence.

DNA sequences of recent HGT events can deviate strongly from the genome background composition while older insertions have often lost their donor-specific sequence signature (Lawrence and Ochman, 1997). Generally, each genome exhibits a reasonably constant background sequence composition; however some genes, traditionally considered part of the core-gene dataset, such as rRNA and ribosomal protein-coding genes often deviate compositionally from the genome background sequence composition mainly due to specific, well-conserved functional constraints rather than alien origin (although some of them can be horizontally exchanged (Nomura, 1999; Yap *et al.*, 1999)). In those cases the effect of the amelioration over time is expected to be limited since strong selection applies.

Base composition and specifically G+C content is known to be related to phylogeny (Forsdyke, 1996). Consequently closely related organisms tend to have similar G+C content; for example the average G+C content of *E. coli*, *Shigella* and *Salmonella* lineages is approximately 50%, 51% and 52% respectively while for the Gram positive *Staphylococcus* and

Streptococcus lineages the average G+C content is 33% and 38%, respectively.

Usually horizontally acquired genes are introduced into a single lineage, and therefore the acquired DNA sequence will be limited to the descendents of the recipient strain and absent from closely related ones. For example *Salmonella* Pathogenicity Island (SPI) 1, a 40kb island carrying a type-III secretion system (T3SS) that enabled the invasion of epithelial cells (Galan, 1996) is present in both *Salmonella* species, *S. bongori* and *S. enterica* while it is absent from the genome of *E. coli*. Consequently SPI-1 represents an ancient HGT event that took place close to the divergence of the two genera (*E. coli* and *Salmonella*) (Baumler, 1997).

On the other hand SPI-2, which is important for systemic infection, is a mosaic of two independent acquisitions: The tetrathionate reductase (*ttr*) gene cluster (Hensel *et al.*, 1999), a 15kb region present in *S. bongori* and *S. enterica*, and a 25kb region, encoding an additional T3SS (Hensel *et al.*, 1997), present only in *S. enterica*. Consequently, using a reference tree topology, HGT events can be distributed into phylogenetic branches of increasing depth; moreover their relative time of insertion, i.e. the most ancient branch in the tree topology that shares a putative horizontally acquired (PHA) gene present only in descendant lineages, can be inferred. Based on this principle, Daubin and Ochman (Daubin and Ochman, 2004), identified sequences unique to monophyletic groups at increasing phylogenetic depths, and studied the characteristics of sequences with no detectable database match (ORFans) using *E. coli* MG1655 as a reference genome.

A key step in inferring the relative time of insertion of PHA genes is the construction of phylogenetic trees that will capture reliably the evolutionary history of the organisms being studied. rRNA genes have been extensively used as molecular chronometers for inferring phylogeny and building tree topologies (Woese, 1987). However, it has been shown that even these traditionally core components of the cell can be horizontally transferred (Nomura, 1999; Yap *et al.*, 1999). Consequently more reliable phylogenies can be built based on approaches exploiting larger sequence samples, e.g. whole-genome sequence (Doolittle, 1999; Doolittle and Papke, 2006). Moreover homologous recombination might well complicate the inference of the true evolutionary history of the genomes under study (Doolittle and Papke, 2006; Feil *et al.*, 2001; Smith *et al.*, 1993). Many closely related bacteria exchange a significant amount of DNA via homologous recombination through highly similar patches throughout their genome sequence (Didelot *et al.*, 2007). Therefore different regions within those genomes might well have different evolutionary histories that cannot be reliably captured by phylogenies relying on a single tree topology (Doolittle and Papke, 2006).

In this chapter, I describe a comparative analysis (Vernikos *et al.*, 2007) between eleven *Salmonella*, three *E. coli* and one *Shigella* strain in order to infer the relative time of insertion of putative HGT events in three strains of the *S. enterica* lineage, by implementing a whole-genome sequence alignment to construct the phylogenetic tree topology of the organisms under study. The relative time of insertion is inferred taking into account the most parsimonious sequence of events i.e. allowing for deletions or independent acquisitions in some of the descendant or ancestral branches. Moreover I discuss and analyse data suggesting that prophages in the *Salmonella* lineage are shared only between very recently diverged lineages but that their sequence composition is very similar to their host's. Finally I describe the implementation of G+C content, the Codon Adaptation Index (CAI) (Sharp and Li, 1987) and high order compositional vectors (Vernikos and Parkhill, 2006), in order to monitor the amelioration process over time.

3.2 Methods

3.2.1 Whole Genome Alignment

The extent of intra-species diversity of bacterial populations was shown recently in a study focused on whole-genome sequence comparisons of eight *Streptococcus agalactiae* isolates (Tettelin *et al.*, 2005); the results show that the genome of a bacterial species (i.e. pan-genome) (Medini *et al.*, 2005) consisting of core and dispensable (i.e. partially shared) genes, may be many times larger than the genome of a single isolate; consequently single-genome sequences may represent poor samples of the overall complexity and structure of a bacterial species.

In the current analysis the phylogenetic relationship of 15 genomes will be inferred by pursuing a whole-genome based comparative genomics approach; the rationale behind a whole-genome based methodology as opposed to marker-based or core-gene based approaches relies on two important aspects of comparative genomics; gene content information and phylogenetic resolution.

In the first case, the dispensable gene pool of a species often encodes components that drive host-adaptation, antigenic variability and determine pathogenicity and virulence related properties of different isolates; for example, HGT can in a single step transform a normally benign organism into a pathogen, a process often referred to as “evolution in quantum leaps” (Groisman and Ochman, 1996); in *Salmonella* two single-step HGT events enabled the invasion of host cells, evading the host defence system, while its close relative *E. coli* evolved as an opportunistic and commensal pathogen.

In terms of phylogenetic resolution, traditional classification systems geared towards analyzing a handful of genetically distinct, often non-overlapping species representatives are capturing only a tiny fraction (Table 3.1) of the species variation (Medini *et al.*, in press); as such they struggle to cope with the increasingly complex structure, the overlapping (fuzzy) boundaries and the dynamic nature of bacterial populations. Moving from single-gene (e.g. 16s rRNA (Woese, 1987)) phylogenies trying to capture the phylogenetic history of an entire bacterial species exploiting only a tiny sequence sample (~0.07%) of a genome, to approaches using a much larger sequence sample (~0.2%) (e.g. multilocus sequence typing – MLST (Maiden *et al.*, 1998)) and recently to whole-genome (Tettelin *et al.*,

2005) comparative genomics (100% coverage), is definitely a big step closer to understanding and more reliably reconstructing the phylogenetic history of bacterial populations.

Table 3.1: Properties of four methods for the comparative analysis of microbial genomes. Estimates have been calculated based on: [a]*Neisseria meningitidis*: genome size ~2.2 Mb (Bentley *et al.*, 2007), 16S rRNA length ~1.5kb (Sacchi *et al.*, 2002), length of MLST loci ~4kb (Maiden *et al.*, 1998). [b]*Salmonella typhi*: genome size ~4.8 Mb (Deng *et al.*, 2003), SNPs on gene fragments covering ~89 Kb (Roumagnac *et al.*, 2006). Source: (Medini *et al.*, in press).

Method	Genome coverage (%)	Core genes	Dispensable genes
16s rRNA	0.07[a]	Yes	No
MLST	0.2[a]	Yes	No
SNPs	2[b]	Yes	Yes
Whole-genome	100	Yes	Yes

Taking into account the increased genome fluidity and sequence mosaicism of bacterial chromosomes due to genome rearrangements, gene gain, gene loss and recombination events, conventional multiple sequence alignment methods, e.g. ClustalW (Thompson *et al.*, 1994) and MUSCLE (Edgar, 2004) that assume sequence co-linearity are not directly applicable on building whole-genome sequence alignments.

MAUVE (Darling *et al.*, 2004), is a genome comparison tool that merges chromosomal sequence rearrangement analysis with conventional multiple sequence alignment methods, providing an efficient method of building whole-genome multiple sequence alignments taking into account extensive chromosomal reordering. In the case of non collinear chromosomal sequences, MAUVE firstly identifies locally collinear regions (termed locally collinear blocks – LCBs) within the chromosomes that represent regions of sequence similarity shared between two or more genomes; in a second step the identified LCBs are progressively aligned using conventional multiple sequence alignment methods (i.e. ClustalW or MUSCLE). The overall algorithm is summarized in the following pseudocode:

Algorithm: MAUVE.

1. Identify multiple maximal unique matches (multi-MUMs), i.e. local alignments of exactly matching (single-copy) sequences that are shared between 2 or more chromosomes.
2. Calculate a phylogenetic guide tree based on the multi-MUMs sequences.
3. Partition a subset (anchors) of the multi-MUMs into LCBs.
4. Do recursive anchoring to identify new anchors within and outside the LCBs.
5. Align each LCB based on the guide tree.

Source: (Darling *et al.*, 2004).

Note: Formally, an LCB is a sequence of multi-MUMs that satisfies a total ordering property, such that the left end of the *i*th multi-MUM occurs before the left end of the *i*+1 multi-MUM, for all multi-MUMs in the LCB and for all the genomes compared.

In the current analysis, 15 genomic sequences (Table 3.2) were aligned by implementing the MAUVE algorithm with the default parameters. An example of the whole-genome sequence alignment of the 15 chromosomes, visualized through the alignment viewer of MAUVE is shown in Figure 3.1. The whole genome sequence alignment of 122 LCBs shared between the 15 genomes was used to build a whole-genome based phylogenetic tree (discussed in the next section).

Table 3.2: The list of 15 strains used in this comparative analysis.

Organism	Reference	Accession Number
Escherichia coli K-12 MG1655	(Blattner *et al.*, 1997)	U00096
E.coli O157:H7 EDL933	(Perna *et al.*, 2001)	AE005174
E. coli CFT073	(Welch *et al.*, 2002)	AE014075
Shigella flexneri serotype 2a 301	(Jin *et al.*, 2002)	AE005674
Salmonella bongori 12419	http://www.sanger.ac.uk/Projects/Salmonella/	N/A
S. arizonae RSK2980	http://genome.wustl.edu/genome_index.cgi	N/A
S. enterica serovar Typhi CT18	(Parkhill *et al.*, 2001)	AL513382
S. enterica serovar Typhi TY2	(Deng *et al.*, 2003)	AE014613
S. enterica serovar paratyphi A SARB42	(McClelland *et al.*, 2004)	CP000026
S. enterica serovar paratyphi A AKU_12601	http://genome.wustl.edu/genome_index.cgi	N/A
S. enterica serovar Typhimurium SL1344	http://www.sanger.ac.uk/Projects/Salmonella/	N/A
S. enterica serovar Typhimurium LT2	(McClelland *et al.*, 2001)	AE006468
S. enterica serovar Typhimurium DT104	http://www.sanger.ac.uk/Projects/Salmonella/	N/A
S. enterica serovar Enteritidis PT4	http://www.sanger.ac.uk/Projects/Salmonella/	N/A
S. enterica serovar Gallinarum 287/91	http://www.sanger.ac.uk/Projects/Salmonella/	N/A

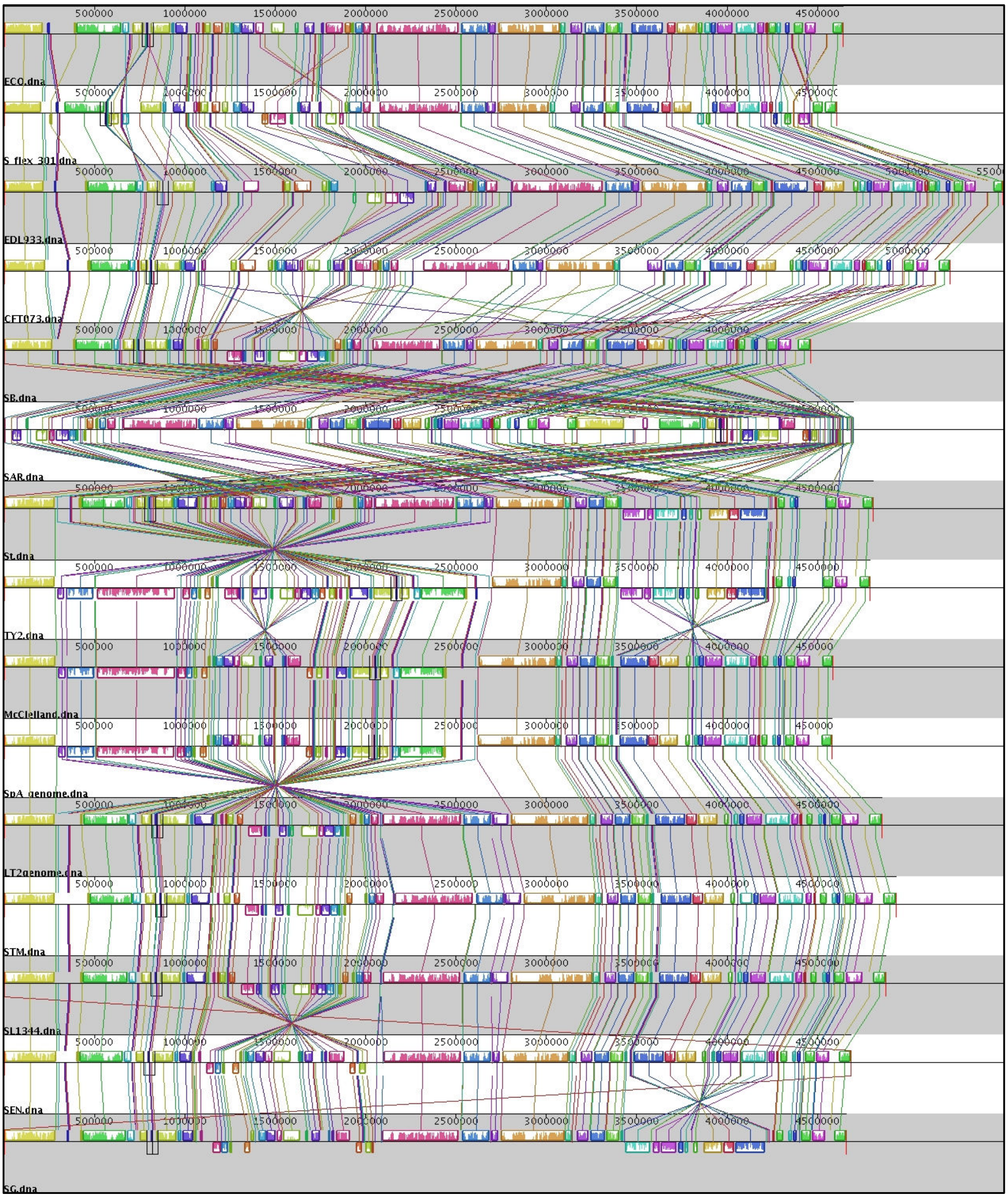


Figure 3.1: MAUVE alignment viewer screenshot: 15 genomes have been aligned (from top to bottom): *E. coli* MG1655, *E. coli* EDL933, *E. coli* CFT073, *S. flexneri* 2a 301, *S. bongori* 12419, *S. arizonae* RSK2980, *S. typhi* CT18, *S. typhi* TY2, *S. paratyphi* A SARB42, *S. paratyphi* A AKU_12601, *S. typhimurium* SL1344, *S. typhimurium* LT2, *S. typhimurium* DT104, *S. enteritidis* PT4, *S. gallinarum* 287/91. Coloured boxes (LCBs) above (forward) and below (reverse orientation) the line represent regions within each chromosome aligned with other regions with sequence similarity, present in the other chromosomes. Inside each aligned box, a similarity profile plot shows the average level of conservation of that sequence. Vertical lines connect the corresponding (same colour) LCBs present in the 15 chromosomes.

3.2.2 Phylogenetic tree building methods

In the current phylogenetic analysis, the aim is to estimate the phylogenetic tree topology that best describes the evolutionary history of the 15 enteric bacteria, taking into account their increased level of genetic fluidity. There are three major "schools" of tree-building methodologies (Table 3.3) widely used to infer the most likely phylogenetic tree: distance-based methods, e.g. unweighted pair-group method with arithmetic mean (UPGMA) (Michener and Sokal, 1957) and Neighbor-Joining (NJ) (Saitou and Nei, 1987); maximum parsimony (MP) methods; and statistical methods, e.g. maximum likelihood (ML) (Felsenstein, 1981) and Bayesian inference (Holder and Lewis, 2003; Huelsenbeck *et al.*, 2001).

Table 3.3: Properties of four widely used tree-building methods.

Method	**Pros**	**Cons**
Neighbor-Joining	Very fast, $O(n^3)$ for n taxa.	Does not necessarily produce the minimum-evolution (optimal) tree.
Maximum-Parsimony	Provides information on the ancestral sequences.	Ambiguous results if homoplasy is common ("long branch attraction"). Underestimates branch lengths.
Maximum-Likelihood	Site-specific likelihoods. Accurate branch lengths.	Computationally intensive.
Bayesian-inference	Faster than ML. Accurate branch lengths.	Relies on the prior distribution over the parameters of the model.

There are numerous previous studies arguing for or against the accuracy and reliability of those methods, exploiting different test datasets and model parameters (Huelsenbeck, 1995; Saitou and Imanishi, 1989; Tateno *et al.*, 1994). Although their evaluation leads to different conclusions, they all converge over the superiority of the NJ and ML methods over the MP method. For those reasons, both of those methods were exploited in the current methodology implementing the DNAML and NEIGHBOR modules of the PHYLIP software (Felsenstein, 1989), discussed in more detail in the following sections.

Generally speaking, the number of all different possible tree topologies grows rapidly with the number of taxa. It can be shown (Felsenstein, 1978) that the number of alternative topologies for an unrooted tree as a function of the number of taxa (T), is:

$$A(T) = \prod_{i=3}^{T} (2i-5) \quad ,$$

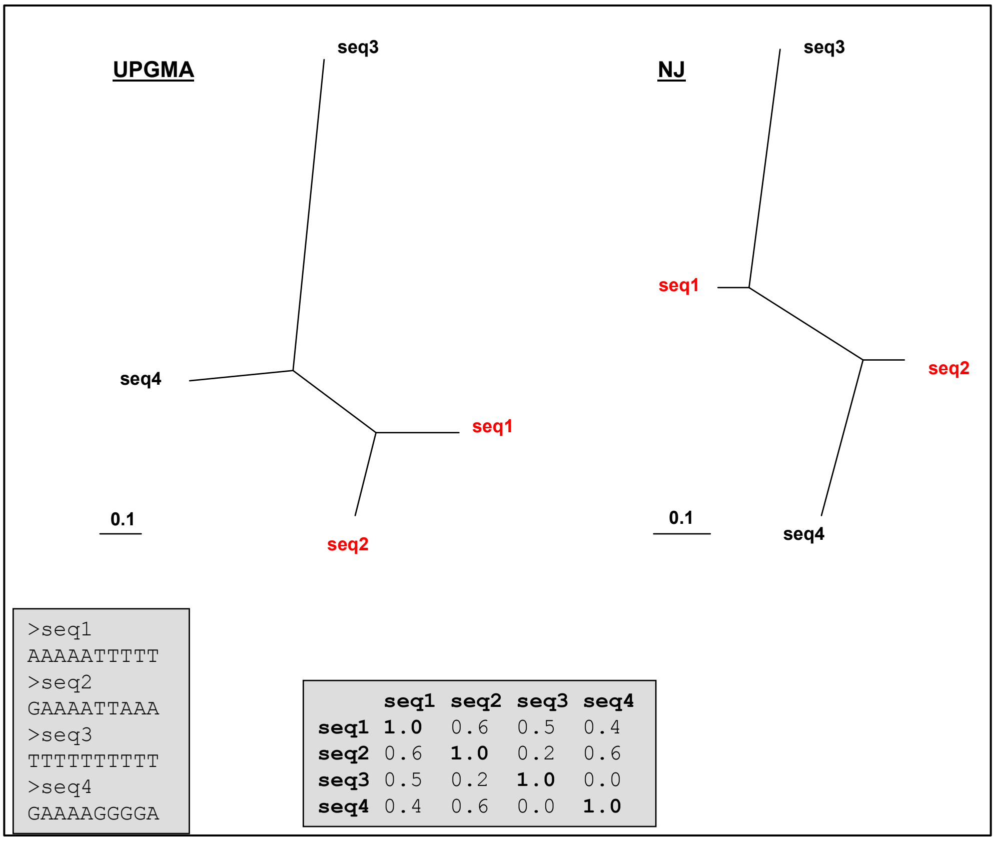
while for a rooted tree, that number is:

$$A(T) = (2T-3)\prod_{i=3}^{T} (2i-5)$$

That means that for 10 and 20 taxa, there are approximately 2 x 10^6 and 2.2 x 10^{20} alternative unrooted tree topologies, respectively.

3.2.2.1 UPGMA

The simplest (and less efficient) tree-building method is UPGMA; this method, exploits a sequential clustering algorithm that starts by identifying the two most similar (given a distance matrix) operational taxonomic units (OTUs) and then builds step-wise the phylogenetic tree topology, evaluating the similarities between the remaining OTUs; the two most similar OTUs of the previous step, are treated as a single OTU in subsequent clustering steps. The main disadvantage of the UPGMA method is that it is based on the assumption that the rate of evolution is constant over time in all the evolutionary lineages (molecular clock hypothesis); in other words, the UPGMA clustering finds the correct tree topology only if the distances between the different taxa are ultrametric, i.e. d(A,B) ≤ *max* [d(A,C), d(B,C)], for all A, B and C; where $d(x,y)$ is the distance metric between OTUs x and y (Figure 3.2).



```
>seq1
AAAAATTTTT
>seq2
GAAAATTAAA
>seq3
TTTTTTTTTT
>seq4
GAAAAGGGGA
```

	seq1	seq2	seq3	seq4
seq1	**1.0**	0.6	0.5	0.4
seq2	0.6	**1.0**	0.2	0.6
seq3	0.5	0.2	**1.0**	0.0
seq4	0.4	0.6	0.0	**1.0**

Figure 3.2: An example of four hypothetical sequences and the inferred true topology, exploiting the UPGMA (left) and the NJ (right) method. The four sequences and the similarity matrix are shown at the bottom of the figure.

3.2.2.2 Maximum Parsimony

MP exploits the concept of parsimony that favours generally simpler over more complicated hypotheses. As such, MP is based on the assumption that the best tree topology is the one that requires the minimum number changes to explain the observed differences between the taxa, and searches for the topology with the minimal cost. If $S_k(a)$ denotes the minimal cost for assignment of character a to node k, such that:

$$S_k(a) = \min_b(S_i(b) + S(a,b)) + \min_b(S_j(b) + S(a,b))$$

the topology with the minimal cost can be found by minimizing the above function for all characters *a* and all nodes *k* of the tree; *i* and *j* denote the daughter nodes of node *k*, and *S*(*a*,*b*) denotes the cost of substituting *a* with *b*.

The MP algorithm consists of two steps: 1) the computation of the cost for a given tree and 2) a search through all trees, to find the overall minimum of this cost; for a small number of taxa e.g. (< 10), an exhaustive search of all the possible tree topologies can be carried out; for a higher number of taxa, however, heuristic methods have to be exploited. Broadly speaking there are two major MP algorithms; weighted parsimony and traditional parsimony (Fitch, 1971). In the first algorithm, each character substitution is assigned a cost while the second algorithm counts simply the number of character substitutions.

3.2.2.3 Bayesian inference

A Bayesian approach produces the tree (or a set of equally optimal trees) that is most likely to be explained by the data (i.e. sequences); in other words it estimates the posterior probability P(H/D) of the hypothesis given the data. This is different from ML that finds the tree that is most likely to have produced the data, evaluating the probability of seeing the data given the hypothesis, i.e. P(D/H). The posterior probability, in a Bayesian implementation, is calculated exploiting Bayes' theorem:

$$P(\vartheta / D) = \frac{P(\vartheta) \cdot P(D / \vartheta)}{P(D)}$$

where *P(θ/D)* is the posterior probability of the tree, *P(θ)* is the prior probability of the tree, *P(D/θ)* is the likelihood of the data given the tree and *P(D)* is the probability of the data (can be calculated as a marginal probability and serves as a normalizing constant, i.e. the sum of the posterior probabilities is 1). The posterior probabilities can be approximated by a Markov Chain Monte Carlo (MCMC) approach (Hastings, 1970; Metropolis *et al.*, 1953) that performs a random walk through the parameter space, randomly modifying the parameters (e.g. the tree topology, a branch length or a substitution model parameter) accepting or rejecting proposed moves based on their posterior probability. If the new posterior computed is larger than the current one, the proposed move is taken, otherwise depending on the level of decrease the move is rejected or accepted; therefore, the Markov chain visits the different regions in the parameter space proportionally to their posterior probability.

3.2.2.4 Neighbor – Joining

NJ (Saitou and Nei, 1987) exploits the concept of minimum evolution (Rzhetsky and Nei, 1993), i.e. at each step the topology with the minimum total branch length is preferred. The NJ algorithm is a star-decomposition algorithm, i.e. the initial tree is a star-like topology that does not however guarantee that the optimal tree topology will be found (greedy algorithm) given that it is prone to converge over a local rather than a global maxima.

NJ is a distance-based, tree building algorithm like UPGMA that nonetheless overcomes the limitation of assuming a constant evolutionary rate for all lineages. This property is very important, and can efficiently avoid converging over the wrong tree topology in case of different evolutionary rates (i.e. the ultrametric condition does not apply); instead of selecting simply the taxa with the minimum distance $d(x,y)$ (that might well not be true neighbouring taxa, see Figure 3.2), NJ builds a new distance matrix (that corrects for different rates) by subtracting from $d(x,y)$ distance the average distances of the two taxa x and y to all the other taxa. The pseudo-code describing the NJ algorithm is given below:

Algorithm: Neighbor-Joining.

Define:

$D_{ij} = d_{ij} - (r_i + r_j)$, where

$$r_i = \frac{1}{|L|-2}\sum_{k \in L} d_{ik}$$

$|L|$ denotes the size of the set L of leaves, and d_{ij} is the distance between taxa i and j.

Initialization:

Define T to be the set of leaf nodes, one per sequence, and set $L = T$.

Iteration:

Pick a pair i, j in L for which D_{ij} is minimal.

Define a new node k, and set $d_{km} = ½\ (d_{im} + d_{jm} - d_{ij})$ for all $m \in L$.

Add k to T, with edges of lengths $d_{ik} = ½\ (d_{ij} + r_i - r_j)$, $d_{jk} = d_{ij} - d_{ik}$, joining k to i and j respectively. Remove i, j from L and add k.

Termination:

When L consists of two leaves, i and j, add the remaining edge between them, with length d_{ij}.

Source: (Durbin *et al.*, 1998).

3.2.2.5 Maximum Likelihood

As mentioned earlier, the aim in a maximum likelihood approach is to maximize the likelihood of a tree P(data | tree), i.e. the probability of the data given a tree topology and a model of evolution (see next section). For a set x of n sequences x_i, for $i = 1 \ldots n$, given a model of evolution, the aim is two-fold: (1) to search through all the possible tree topologies T with the n sequences assigned at the corresponding leaves of the tree and (2) to search over all possible branch lengths t, with the objective of finding the maximum likelihood tree, i.e. the tree with topology T and branch lengths t that maximizes $P(x \mid T, t)$.

In the case of two sequences x_1 and x_2, there is only one possible rooted tree topology T, therefore the likelihood of the tree will vary relative to the branch lengths t_1 and t_2. In this example, let $x_{1,\ m}$ and $x_{2,\ m}$ denote the residues at the mth site of the two sequences. Assigning a residue a to the root of the tree, we can calculate the probability (q_a) of having a at the root of T and of having substitutions of a by $x_{1,\ m}$ and $x_{2,\ m}$, as follows:

$$P(x_{1,m},x_{2,m},a\mid T,t_1,t_2)=q_aP(x_{1,m}\mid a,t_1)P(x_{2,m}\mid a,t_2)$$

In a second step, in order to calculate the probability of generating $x_{1,\ m}$ and $x_{2,\ m}$ residues at the two leaves of T, we have to sum over all different possible values of a, since we do not have any prior knowledge of what the residue at the root of the tree is:

$$P(x_{1,m},x_{2,m}\mid T,t_1,t_2)=\sum_a q_aP(x_{1,m}\mid a,t_1)P(x_{2,m}\mid a,t_2)$$

The final step is to calculate the full likelihood over the entire length (M) of the two sequences x_1 and x_2:

$$P(x_1,x_2\mid T,t_1,t_2)=\prod_{m=1}^{M}P(x_{1,m},x_{2,m}\mid T,t_1,t_2)$$

In order to calculate the probability P(z|y,t) of a sequence z arising from an ancestral sequence y over the branch length t, we need a model of evolution that describes how residues are substituted by others. Details of such evolutionary models will be discussed in the next section. It can be shown that given a *transition-probability* matrix $P(t) = e^{Qt}$ that determines the probability that a given residue a will become b after time t (Q denotes the *substitution-rate* matrix that determines the rate of change between pairs of nucleotides in an infinitely small time interval dt), we can compute the maximum likelihood estimate (MLE) of a given branch length t, i.e. the value of t that maximizes the likelihood of the tree.

For example, in the case of two hypothetical nucleotide sequences x_1 and x_2, each 95 nucleotides long with 9 different nucleotides, exploiting the simplest evolutionary model of Jukes and Cantor (Jukes and Cantor, 1969) (see next section for details) the MLE of the branch length between x_1 and x_2 can be estimated (Figure 3.3) applying an expectation maximization (EM) algorithm. Generally in the case of n sequences x_1, ..., x_n with m residues, the probability of generating those residues at the n leaves of T with branch lengths t can be calculated by taking the product of the probabilities of substitutions on all branches of the tree:

$$P(x_{1,m} \dots x_{n,m} \mid T, t) = \\ \sum_{a_{n+1}, a_{n+2}, \dots a_{2n-1}} q_{a_{2n-1}} \prod_{i=n+1}^{2n-2} P(a_i \mid a_{a(i)}, t_i) \prod_{i=1}^{n} P(x_{i,m} \mid a_{a(i)}, t_i)$$

where $a(i)$ denotes the parent node of node i. Note that the sum is over all possible assignments of a_k to non-leaf nodes k, i.e. nodes $n+1$ … $2n-1$. The above probability can be calculated pursuing a post-order traversal (i.e. leaves → root direction) of the tree, exploiting the *pruning algorithm* introduced by Felsenstein (Felsenstein, 1981). If the residue at node k is a then the probability of all the leaves below k is $P(L_k \mid a)$. Having computed the probabilities $P(L_i \mid b)$ and $P(L_j \mid c)$ of all b and c, at the daughter nodes i and j of k, the probability $P(L_k \mid a)$ can be calculated as follows:

```
Algorithm: Maximum-Likelihood (Felsenstein).
Initialise:
Set: $k = 2n - 1$.
Recursion: Compute $P(L_k \mid a)$ for all $a$ as follows:
If $k$ is leaf node:
        Set $P(L_k \mid a) = 1$ if $a = x_{k,m}$, $(L_k \mid a) = 0$ if $a \neq x_{k,m}$.
If $k$ is an internal node:
        Compute $P(L_i \mid a)$ and $P(L_j \mid a)$ for all $a$ at the daughter nodes $i$ and $j$, and set:
$P(L_k \mid a) = \left[ \sum_b P(b \mid a, t_i) P(L_i \mid b) \right] \times \left[ \sum_c P(c \mid a, t_j) P(L_j \mid c) \right]$   (3.1)
Termination:
Likelihood at site $m$:
$P(x_m \mid T, t) = \sum_a q_a P(L_{2n-1} \mid a)$
```

Source: (Durbin *et al.*, 1998).

Assuming that all *M* sites are independent, the full likelihood is:

$$P(x \mid T,t) = \prod_{m=1}^{M} P(x_m \mid T,t)$$

Note that the pruning algorithm of Felsenstein's calculates successively the probabilities of the data on each subtree of the tree topology *T*. Therefore it is crucial to sum over all the ancestral states of a node only after having done so for all of its child nodes. In equation 3.1, the two terms represent the probability that residue *a* will become *b* (or *c*) over the branch length t_i (or t_j) times the probability of observing the tips of node *i* (or *j*) given the state *b* (or c), summed over all possible states *b* (or c).

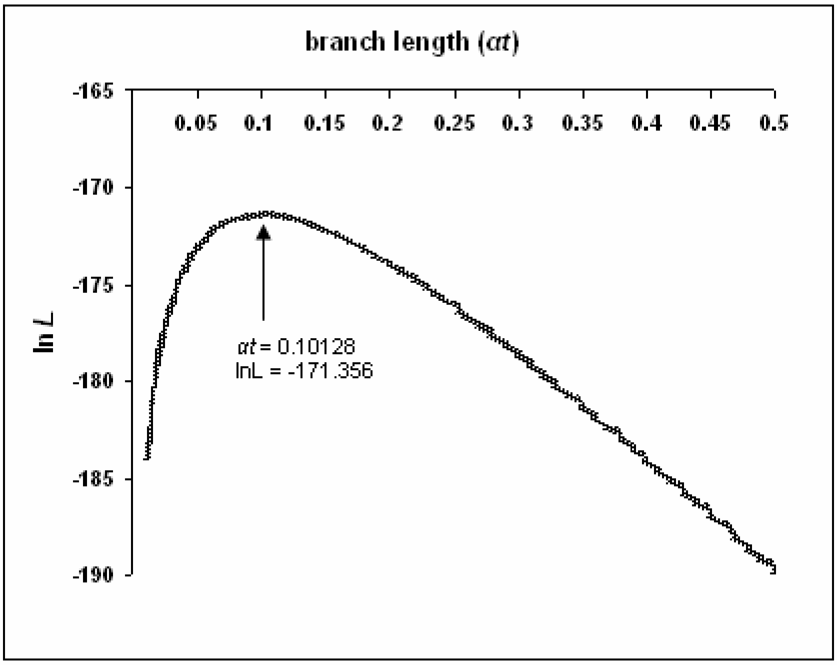


Figure 3.3: The *log* likelihood P(x_1,x_2 | *T*,*t*) for two sequences x_1, x_2 with 9 different nucleotides (nt) and a total length of 95nt, exploiting the Jukes and Cantor model. The MLE (0.10128) of the branch length is shown.

3.2.3 Nucleotide substitution models

Generally, DNA sequences derived from a common ancestor will, over time, gradually diverge due to substitution of their nucleotides. The distance between two sequences reflects the expected number of nucleotide substitutions per site, and assuming a constant over time evolutionary rate, the distance is a linear function of the time of divergence. The simplest estimate of the distance between two sequences is the proportion (p) of sites at which the two sequences differ. For example for two sequences, each 100nt long with 20 different sites, $p = 20\% = 0.2$. However because over time, the two sequences will accumulate more and more substitutions and some sites will have changed multiple times, the observed differences do not necessarily represent the true number of substitutions that have occurred since the divergence of the two sequences.

Therefore, for sequences diverged long time ago, p underestimates the number of substitutions, since it does not take into account multiple substitutions (Figure 3.4). For that reason, more sophisticated and realistic evolutionary models have to be exploited in order to estimate more reliably the true evolutionary time elapsed since the divergence of two sequences, taking into account the various aspects of the dynamics dictating the substitutions of nucleotide residues.

3.2.3.1 Jukes-Cantor model

The simplest evolutionary model (Figure 3.5), introduced by Jukes and Cantor (Jukes and Cantor, 1969), assumes that every nucleotide changes into any other nucleotide with exactly the same rate a. For two nucleotide residues i and j (where i, j = T, C, A or G), let q_{ij} denote the instantaneous rate of substitution of i by j. Those substitution rates for all 16 different combinations of nucleotide pairs can be represented in the form of a *substitution-rate* matrix $\boldsymbol{Q}$:

$$Q = \{q_{ij}\} = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}$$

Note that for any nucleotide *i* the total rate of substitution is 3α, and the order of nucleotides in the matrix is: T, C, A, G.

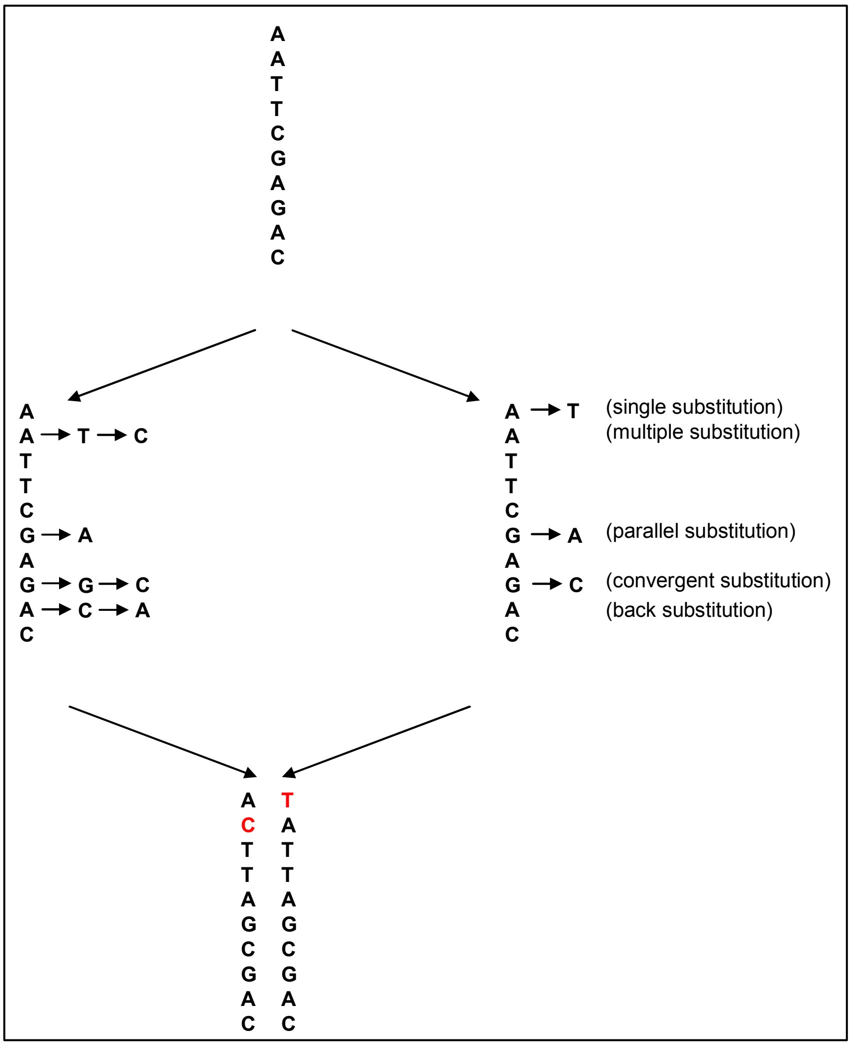


Figure 3.4: An example of multiple substitutions at the same site for a set of two hypothetical sequences diverged from a common ancestral sequence (top). Only two observed substitutions ($p = 0.2$) are inferred, while the true number of substitutions is 10, i.e. 1.0 substitutions per site.

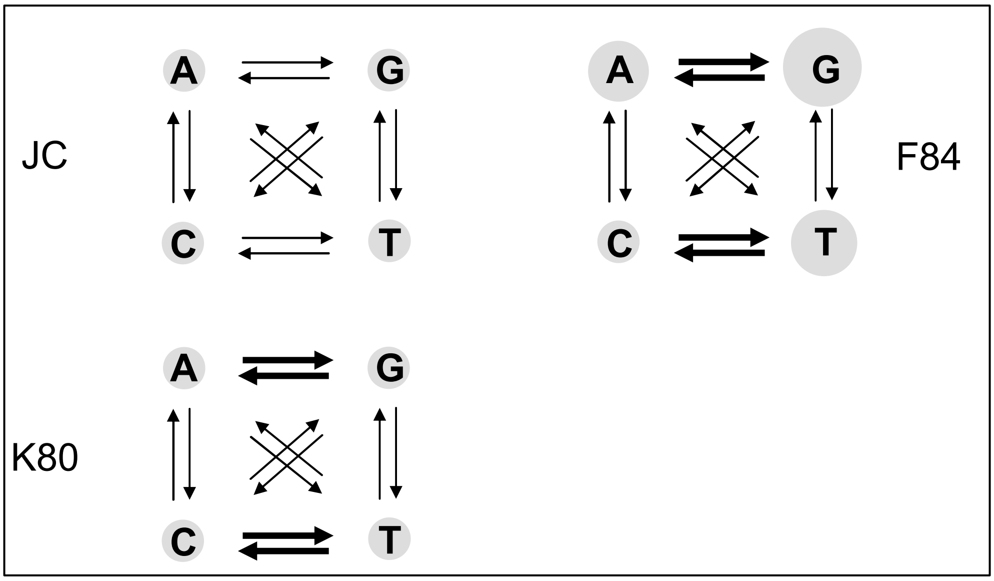


Figure 3.5: Three models of nucleotide substitution; JC (Jukes and Cantor, 1969), K80 (Kimura, 1980) and F84 (Kishino and Hasegawa, 1989).Arrows of different thickness represent different substitution rates and circles of different size the different nucleotide equilibrium frequencies.

q_{ij} *dt* represents the probability of $i \to j$ change over an infinitely small time interval *dt*. However in the case of biological sequences, we are more interested in longer time t ($t > 0$) periods, over which residue substitutions occur. In other words we want to estimate the transition probability p_{ij} (t) of i being substituted by j after time t. The 16 different transition probabilities $p_{ij}(t)$ can be represented in the form of a *transition-probability* matrix:

$$P(t) = e^{Qt} = \begin{bmatrix} p_r(t) & p_s(t) & p_s(t) & p_s(t) \\ p_s(t) & p_r(t) & p_s(t) & p_s(t) \\ p_s(t) & p_s(t) & p_r(t) & p_s(t) \\ p_s(t) & p_s(t) & p_s(t) & p_r(t) \end{bmatrix} ,$$

where:

$$p_r(t) = \frac{1}{4}\left(1 + 3e^{-4at}\right)$$

$$p_s(t) = \frac{1}{4}\left(1 - e^{-4at}\right) \quad .$$

Using the *transition-probability* matrix $P(t)$ we can calculate over the time period t, the probability of nucleotide i having being substituted by j (Figure 3.6). Note that for $t\to\infty$, $p_r(t) = p_s(t) = ¼$, suggesting that the nucleotide equilibrium frequencies according to the JC model are $q_T = q_C = q_A = q_G = ¼$. In other words, after time $t\to\infty$, at every site of the sequence so many substitutions have occurred that the target nucleotide is random (i.e. with equal probability of observing any of the four nucleotides).

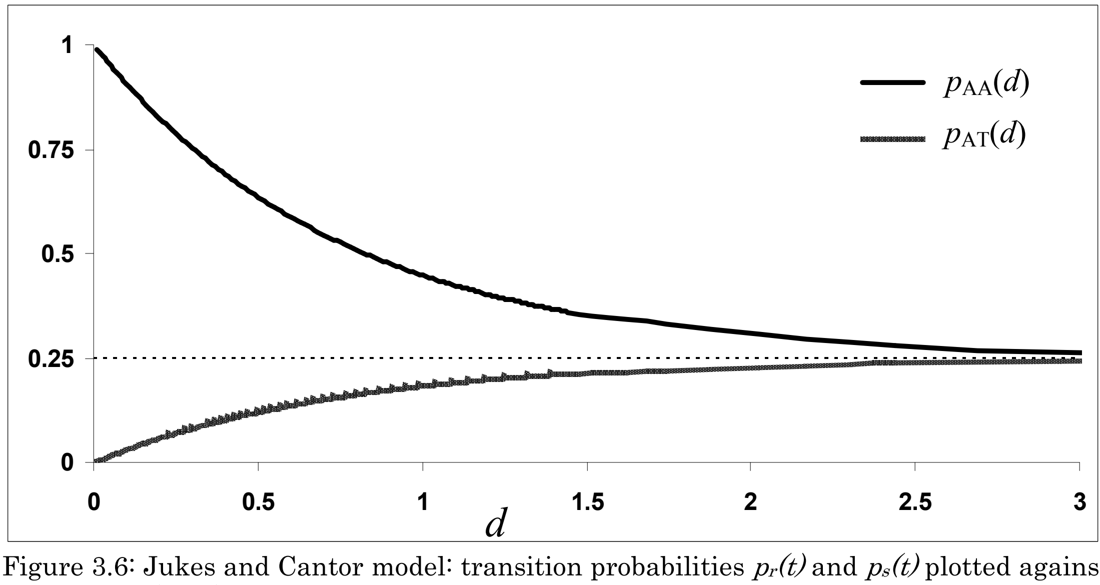


Figure 3.6: Jukes and Cantor model: transition probabilities $p_r(t)$ and $p_s(t)$ plotted against distance d (=$3at$); d is expressed as the expected number of substitutions per site. Assuming that for any nucleotide, the total substitution rate is $3a$ (see *substitution-rate* matrix Q), if two hypothetical sequences are separated by time t (i.e. diverged from their common ancestor $t/2$ ago) the distance d between them is $3at$.

Under the JC model, for any nucleotide the total substitution rate is 3*a*, while the probability *p* of a nucleotide being different from the nucleotide of the ancestral sequence is:

$$p = 3p_s(t) = \frac{3}{4}\left(1 - e^{-4at}\right) = \frac{3}{4}\left(1 - e^{-\frac{4}{3}d}\right)$$

Consequently, if we know the proportion $\hat{p}$ of different sites between two sequences, we can estimate their distance:

$$\hat{d} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\hat{p}\right)$$

The above equation represents the MLE (Figure 3.3) of the distance between the two sequences. Note that if two sequences are different in over 75% of their sites, the above estimate is not applicable, since their estimated distance becomes infinite.

3.2.3.2 Kimura – 2 parameter model

The JC model fails to capture a very important parameter driving the dynamics behind nucleotide substitutions; purine to purine (A ↔ G) or pyrimidine to pyrimidine (T ↔ C) substitutions (i.e. transitions) occur more frequently than substitutions between purines and pyrimidines (A,G ↔ G,C), i.e. transversions. A slightly more complex model of nucleotide substitutions that accounts for different transition and transversion rates, was introduced by Kimura (Kimura, 1980). However this model is still far from realistic, since it assumes (as the JC model does) that the nucleotide equilibrium frequencies are equal. The *substitution-rate* matrix for the Kimura 2-parameter model (K80) is:

$$Q = \begin{bmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{bmatrix},$$

where α denotes the transition and β the transversion substitution rates, respectively. Note that the distance d between two sequences is now $(\alpha + 2\beta)t$, and the total substitution rate for each nucleotide is $\alpha + 2\beta$. In a similar principle to the one used for the JC model, it can be shown that the estimate of the distance between two sequences is:

$$\hat{d} = -\frac{1}{2}\ln(1-2S-V) - \frac{1}{4}\ln(1-2V) \quad ,$$

where S and V are the fractions of transitions and transversions in the alignment of two sequences, respectively. Exploiting the K80 model with transition/transversion rate (k = 0.75), for the same example of the two sequences (each 95nt long with 9 different nucleotides) used in Figure 3.3, the MLE of their distance is 0.10136, (JC distance = 0.10128); note that the K80 model with k = 0.5 reduces to the JC model, giving the same distance estimate.

3.2.3.3 F84 model

A more sophisticated model (F84) of substitution with five free parameters, allowing different transition and transversion substitution rates ($\alpha \neq \beta$), as well as different nucleotide equilibrium frequencies ($q_T \neq q_C \neq q_A \neq q_G$) was proposed by Felsenstein; this model is the one exploited by the DNAML module of the PHYLIP package (Felsenstein, 1989) and the transition probabilities for this model were firstly described by Kishino and Hasegawa (Kishino and Hasegawa, 1989). The F84 model reduces to the K80 model for $q_T = q_C = q_A = q_G$, and the JC model for $2\alpha = \beta$ and $q_T = q_C = q_A = q_G$.

3.2.3.4 Substitution rate variation

So far all the evolutionary models discussed rely on a very simplifying assumption; each site in the sequence is evolving with the same rate, i.e. a single substitution matrix describes all the different nucleotide sites. However in biological sequences, this assumption rarely holds; for example, in the case of protein coding genes for each codon there are three different nucleotide positions, i.e. position 1, 2 and 3, and because of the genetic code degeneracy each position is under different mutational pressure. In the case of RNA coding genes, secondary loop and stem structures evolve with different substitutions rates. Therefore, assuming a single evolutionary rate across all the nucleotide sites underestimates the true distance between two sequences.

The rate variation among sites can be approximated by a statistical distribution, in which case the rate *r* for any site is a random variable drawn from that distribution. It has been shown that the rate variation among sites approximates the gamma distribution (Yang, 1994; Yang, 1996):

$$g(r,\alpha,\beta)=\frac{e^{-\beta r}r^{\alpha-1}\beta^{a}}{\Gamma(\alpha)}$$

for $0 < r, \alpha, \beta < \infty$, where α and β are the shape and the scale parameters, respectively. The mean of the distribution is $E(r)=\alpha / \beta$ and the variance $\mathrm{var}(r) = \alpha / \beta^2$. The rate variation among sites is inversely correlated with the α parameter (Figure 3.7):

- If $\alpha \leq 1$, then most sites have very low substitution rates, and very few have very high rates,
- if $\alpha \to \infty$, then all sites have the same rate,
- if $\alpha > 1$, then most sites have intermediate rates and few sites have either very high or very low rates.

I will give an example showing that ignoring the rate variation among sites, leads to underestimation of the true distance between two sequences. Considering again the hypothetical sequences (length: 95nt, mismatches: 9nt) discussed in the Maximum Likelihood section above, the JC distance with the α parameter set to 0.5 (i.e. most sites have very low substitution rate), is 0.11627, much higher than the JC distance (= 0.10128) ignoring the rate variation among sites.

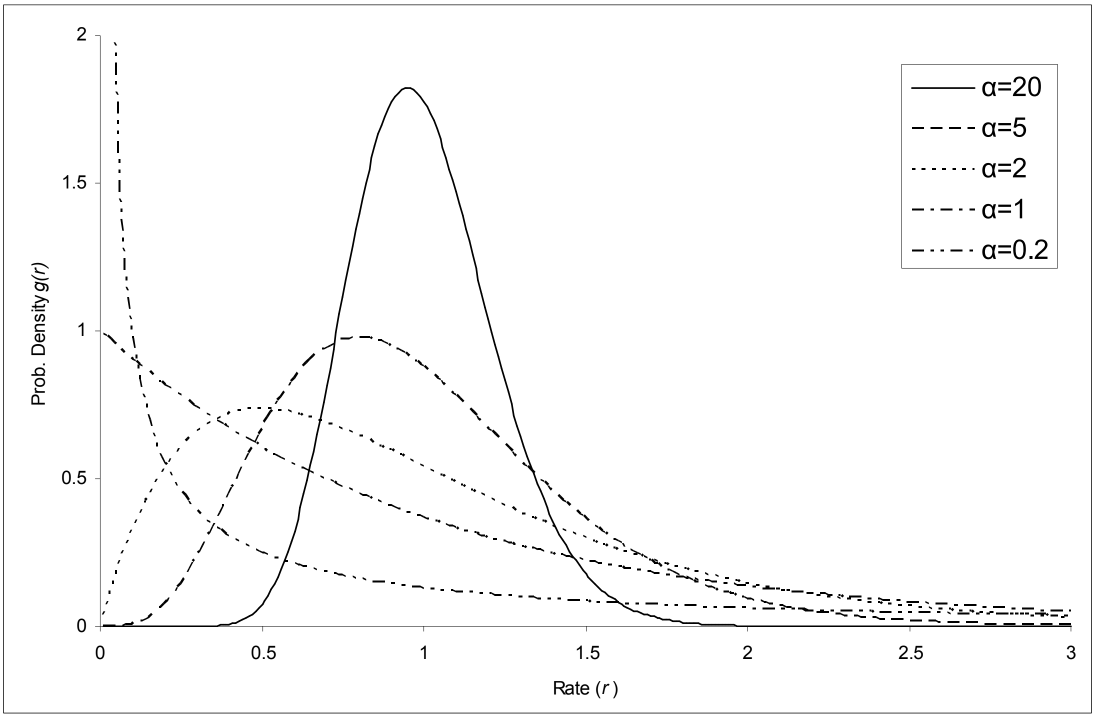


Figure 3.7: *Gamma* distribution $g(r, \alpha, \beta)$; probability densities for different values of the α parameter. In this example, $\alpha = \beta$. The mean of the distribution is $E(r)=\alpha / \beta = 1$ and the variance $\mathrm{var}(r) = \alpha / \beta^2 = 1/\alpha$.

One way of estimating the different substitution rates of different sites in a multiple-alignment of sequences, is to treat the unknown r_i rate of each site i as the hidden state and the residues of each column in the alignment as the observed state in a Hidden Markov Model (HMM). With a HMM implementation, we can estimate the most probable state (i.e. rate) path that best describes the data. Defining the number of expected number k of different rates r_i and a prior probability distribution that determines the probabilities of occurrence of each rate, we can infer for each site i the most probable rate r_i. An EM technique, e.g. the Baum-Welch algorithm (Baum, 1972) can be used to estimate the parameters (i.e. emission and transition probabilities) of the HMM and a dynamic programming approach, e.g. the Viterbi algorithm (Viterbi, 1967) can be used to estimate the most probable rate path (Figure 3.8). For details about the Viterbi and the Baum-Welch algorithm refer to chapter 2. A HMM-based implementation for inferring different rates of evolution at different sites, was introduced by Felsenstein and Churchill (Felsenstein and Churchill, 1996) and implemented in the DNAML module of the PHYLIP package (Felsenstein, 1989).

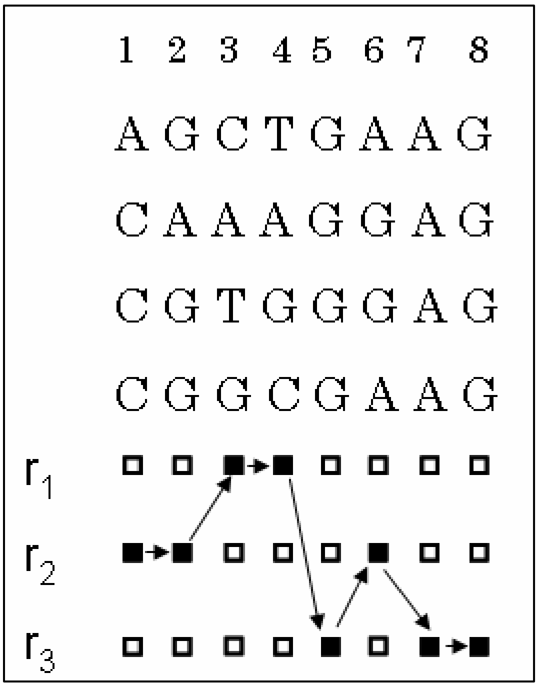


Figure 3.8: An example of four hypothetical sequences, each 8nt long. Each nucleotide site evolves under a different substitution rate (r_1 > r_2 > r_3). Assuming that there are k (=3) different substitution rates, implementing a Hidden Markov Model (HMM) approach, we can infer the most likely rate r_i for each site.

3.2.3.5 Parameter estimation

Although the Maximum Likelihood method can produce a very reliable tree topology with all the parameters (e.g. node/branch order and branch length) optimized, in the case of a large number of sequences it can be very computationally intensive. The overall aim is two-fold; search through all the possible tree topologies and then for each topology compute the maximum likelihood estimate of its branch lengths. Although the ML method is not applicable in the case of a large number of sequences, searching for the ML tree for a set of four (nucleotide or protein) sequences is a very straight forward computation (15 different rooted tree topologies).

This concept is exploited by the quartet puzzling algorithm (Strimmer and von Haeseler, 1996) and implemented by the TREE-PUZZLE software (Schmidt *et al.*, 2002). The quartet puzzling algorithm consists of three steps: 1. All possible quartet ML trees are reconstructed (ML step), 2. The quartet trees are repeatedly combined to an overall intermediate tree (puzzling step) adding sequences step-wise (with multiple input orders), 3. In the consensus step, a majority rule consensus of all intermediate trees is constructed. Because the quartet puzzling algorithm is efficiently fast, the parameters e.g. the α shape-parameter of the gamma distribution for among site rate variation, the transition/transversion rate and the nucleotide frequencies can be accurately estimated from the data, prior to the tree building (e.g. NJ or ML) method.

Using the whole-genome sequence alignment of the 15 (11 *Salmonella* and four outgroup strains) reference genomes, built by the MAUVE method, and running the TREE-PUZZLE algorithm the parameters of the evolutionary model were estimated from the data (Table 3.4). The multiple sequence alignment and the estimated model parameters were fed into the NEIGHBOR and the DNAML modules of PHYLIP (Felsenstein, 1989) to build the Neighbor-Joining and the Maximum Likelihood tree topology of the dataset, respectively.

Table 3.4: Evolutionary model parameters, estimated from the data, by the TREE-PUZZLE method, exploiting a whole-genome based multiple sequence alignment of 11 *Salmonella* and four outgroup strains.

Model of substitution	HKY85 (Hasegawa *et al.*, 1985)	
Expected transition/transversion ratio	2.22	
Expected pyrimidine transition/purine transition ratio	1.01	
Rate matrix R	A-C rate	1.00000
	A-G rate	4.38068
	A-T rate	1.00000
	C-G rate	1.00000
	C-T rate	4.38068
	G-T rate	1.00000
Nucleotide frequencies	pi(A)	23.9%
	pi(C)	26.2%
	pi(G)	26.0%
	pi(T)	23.9%
Gamma distribution – alpha parameter	*a* = 0.26, S.E. 0.00	
	Number of Gamma rate categories: 4	
	Category	Relative rate
	1	0.0008
	2	0.0696
	3	0.5975
	4	3.3321
	Categories 1-4 approximate a continuous Gamma-distribution with expectation 1 and variance 3.87.	
Quartet Puzzling	Number of puzzling steps	1000
	Analysed quartets	1365
	Fully resolved quartets	1365
	Partly resolved quartets	0
	Unresolved quartets	0

3.2.4 Relative time of HGT events

In order to differentiate more reliably gene loss from gene gain (HGT) in the *Salmonella* lineage, a genomic dataset of three *E. coli* (MG1655, EDL933, CFT073) and one *S. flexneri* strain was used; those four genomes form the outgroup lineage in the reference tree topology. For example a gene that is present in the *Salmonella* lineage and absent from *E. coli*

MG1655 might well be either a true HGT in the former or deletion in the latter. However, if for example, the same gene is also present in *E. coli* EDL933 and *E. coli* CFT073 then we can infer more reliably that this event probably represents a deletion (in *E. coli* MG1655) rather a true HGT in the *Salmonella* lineage. Conversely, a sequence that is confined to one lineage is more likely to have been horizontally acquired than to have been deleted independently from multiple lineages (Lawrence and Ochman, 1998).

In a parsimony model the least complex (i.e. with the lowest cost) interpretation of an observation is always favoured; in our case this is the minimum number of events or changes within a phylogenetic tree that can explain the current state of phylogenetic relationships between the taxa compared.

In the current analysis, the tree topology was used as the reference phylogenetic history of the 15 genomes compared, and genes present in three representative *S. enterica* strains (i.e. Typhi CT18, Paratyphi A SARB42 and Typhimurium LT2) were distributed on increasing depth branches of the phylogenetic tree. For example, in the case of CT18, a gene X in CT18 that has orthologs only in TY2 and the four outgroup genomes, is more likely to represent an independent HGT event in the parent node of CT18 and TY2, rather than the result of multiple deletions in the other nine genomes (Figure 3.9 A). Similarly, a gene X in CT18 that has no ortholog in the four outgroups and the *S. bongori* genome but has orthologs in the other nine genomes is more likely to have been acquired on the branch predating the divergence of *S. enterica* from *S. arizonae*. (Figure 3.9 B).

The algorithm for inferring the most likely relative time of acquisition of a PHA gene in the *Salmonella* lineage, taking into account the most parsimonious sequence of events, is summarized in the following pseudocode.

Algorithm: Maximum Parsimony for inferring the relative time of HGT events.
Define: k is the number of the node. a is the state of k (“0” or “1” for gene absence or presence, respectively).

A. Ancestral state reconstruction:
Iteration (post-order tree traversal, i.e. leaves → root direction):

```
If $k$ is a leaf node:
        Set $S_k = a$.
If $k$ is an internal node:
        Compute $S_i$ and $S_j$ for all $a$ at the daughter nodes $i$ and $j$, of $k$.
        $S_k$:
        For $a = 0$, compute:
                $A = |a - S_i| + |a - S_j|$    (1)
        For $a = 1$, compute:
                $B = |a - S_i| + |a - S_j|$    (2)
        if (A<B) then set $S_k = 0$
        elsif (A>B) then set $S_k = 1$
        else set $S_k = [0,1]$
```

Note: In case of equally parsimonious ancestral states, i.e. $S_x = [0,1]$ then compute (1) and (2) for both states of S_x.
Termination:
If $k = 2n - 1$, where n is the number of taxa.

B. Relative time of acquisition inference:

```
For all $k$ in the node path leading from the root of the tree to the node of the reference genome:
If $S_k = 1$ then set $t^* = k$, (break loop).
else $k--$;
```

where t^* denotes the relative time of HGT in the *Salmonella* lineage (relative to the reference genome).

This algorithm consists of two parts; in the first part (A), the ancestral states of gene presence/absence are reconstructed in a post-order tree traversal, starting from the leaves moving towards the root of the tree. If the presence (or absence) of a gene X on ancestral branches can be unambiguously inferred, a state character 1 (or 0) is assigned on the node following the corresponding branch; alternatively both state characters (1, 0) are assigned. In the second part (B) of this algorithm, for each reference genome the relative time of acquisition of the gene in question is inferred following the node path leading from the root of the tree to the node of the reference genome; the relative time of acquisition is assigned to be the first node (more specifically the parental branch of this node) of the reference path, for which a character state 1 has been assigned.

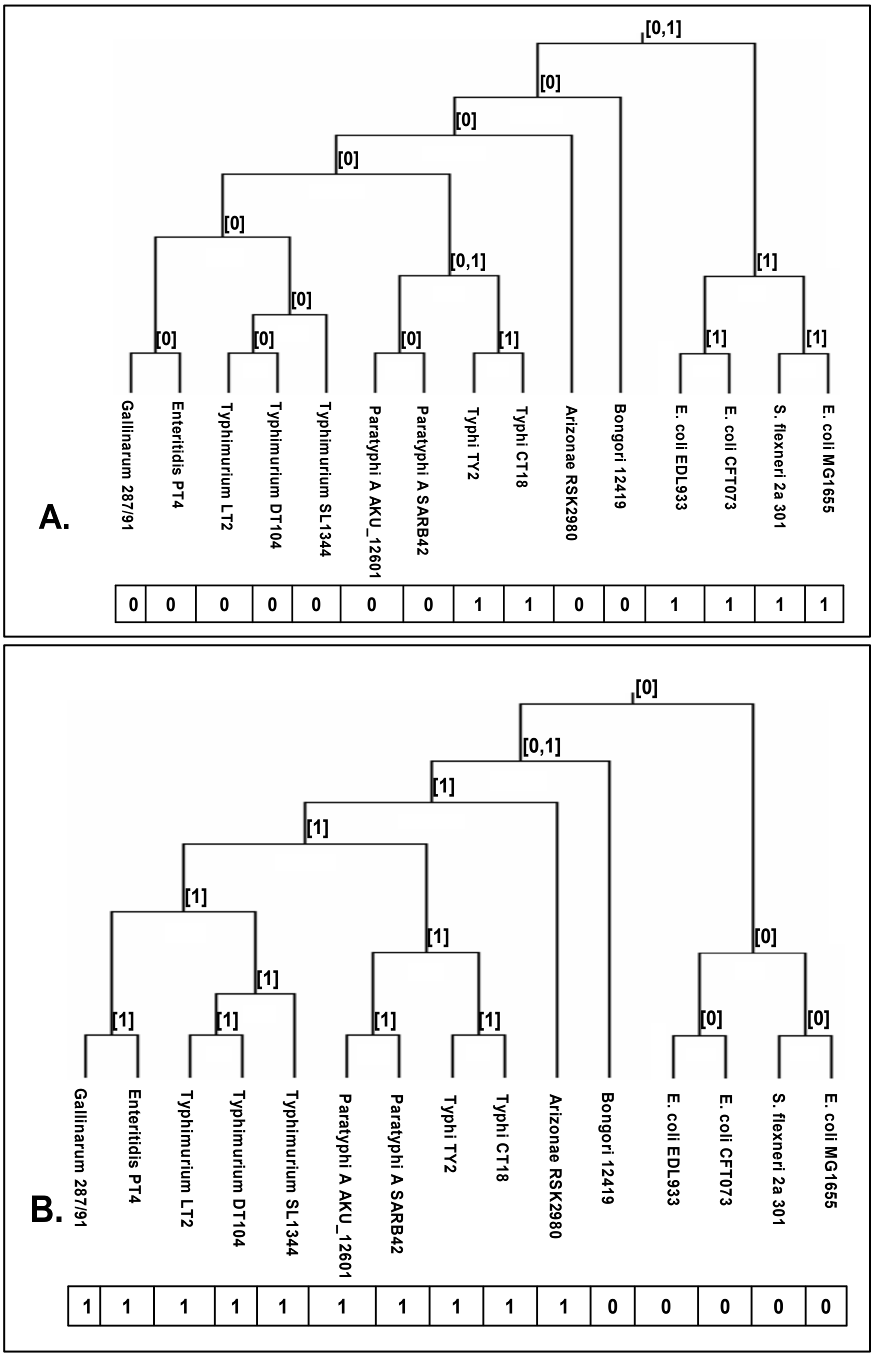


Figure 3.9: The phylogenetic distribution of a hypothetical gene X, present only in the four outgroups and the two Typhi genomes (A). The phylogenetic distribution of a hypothetical gene X, absent from the four outgroups and the Bongori genome (B). In the columns below the tree topology the presence or absence of the gene X is shown as “1” and “0” respectively. On each node, the inferred ancestral state that gives the minimum cost is shown in a binary fashion, i.e. [1] or [0]. In case of equally parsimonious ancestral states, both possible states are assigned to each node. In the first case, the inferred relative time of insertion is the branch predating the Typhi node, while in the second case it is the branch predating the divergence of *S. enterica* from *S. arizonae*.

3.2.5 Compositional analysis

In order to monitor the level of amelioration with respect to the inferred relative time of insertion for each gene in each of the three query genomes, the overall as well as the codon-position specific G+C content was calculated. The G+C content of the second codon position is generally very constrained to similar values across species (Lawrence and Ochman, 1997), given that most possible nucleotide substitutions would result in a change in the encoded aminoacid residue (non-synonymous substitutions); therefore calculating the codon-position specific G+C content increases the compositional resolution.

Furthermore in order to increase the sensitivity of capturing compositionally deviating genes (genes that do not deviate in terms of G+C content but show higher order compositional bias), I implemented the Interpolated Variable Order Motifs (IVOMs) method (Vernikos and Parkhill, 2006). In order to differentiate horizontally acquired from highly expressed genes that can also deviate compositionally, I also performed a CAI analysis (for details refer to section 1.3 of the introduction), measuring the adaptation of each gene to the codon usage of a reference set of highly expressed genes, proposed by Sharp and Li (Sharp and Li, 1986).

3.2.6 Orthologous genes

In order to identify orthologous genes, each genome in the reference dataset was compared against all the other genomes, by means of best reciprocal FASTA (Pearson, 1990) approach; overall an all-against-all comparison of 67,553 genes was performed (details of the best reciprocal FASTA algorithm are given in section 2.2.5 of chapter 2). The results were manually curated taking into account the syntenic relationship among the putative orthologs by visualizing the comparison using ACT (Carver *et al.*, 2005).

3.3 Results

3.3.1 Time distribution of PHA genes

In order to construct the tree topology that best describes the phylogenetic history of the strains studied in this analysis, I implemented the Neighbor Joining (Saitou and Nei, 1987) and the Maximum Likelihood (Felsenstein and Churchill, 1996) method, exploiting three different models of nucleotide substitution, namely JC, K80 and F84. Both (NJ and ML) methods resulted in identical tree topology illustrated in Figure 3.10. These data suggest that using whole-genome sequence information the true phylogeny of the organisms at hand can be captured reliably (see discussion for more details).

For each of the three query genomes the total number of PHA genes, as well as their relative time of insertion was inferred (Appendix B, C and D). The results are summarized in Table 3.5 and Figure 3.10. Using each of the three query genomes, on the branches prior to nodes 1, 2 and 3, I inferred similar numbers of PHA genes for the corresponding relative time of insertion (for the sake of simplicity, from this point on I will refer to the branch prior to node X as branch X). The different number of PHA genes is principally due to small differences in the number of genes in each genome (insertions, deletions, gene-remnants) as well as differences in the genome annotation.

From this point on I assign on branches 1, 2 and 3 the intersection of the respective number of genes determined on each branch using each one of the three query genomes. Overall, this reciprocal FASTA analysis suggests that approximately 2,500 orthologous genes form a core gene set shared by all the 11 *Salmonella* strains; this number reduces to approximately 2,000 orthologous genes shared by the *E. coli*, *S. flexneri* and *Salmonella* strains, used in this study (Figure 3.11). Interestingly this figure is very close to the 2,049 native genes in the γ-Proteobacteria, proposed by Daubin and Ochman (Daubin and Ochman, 2004).

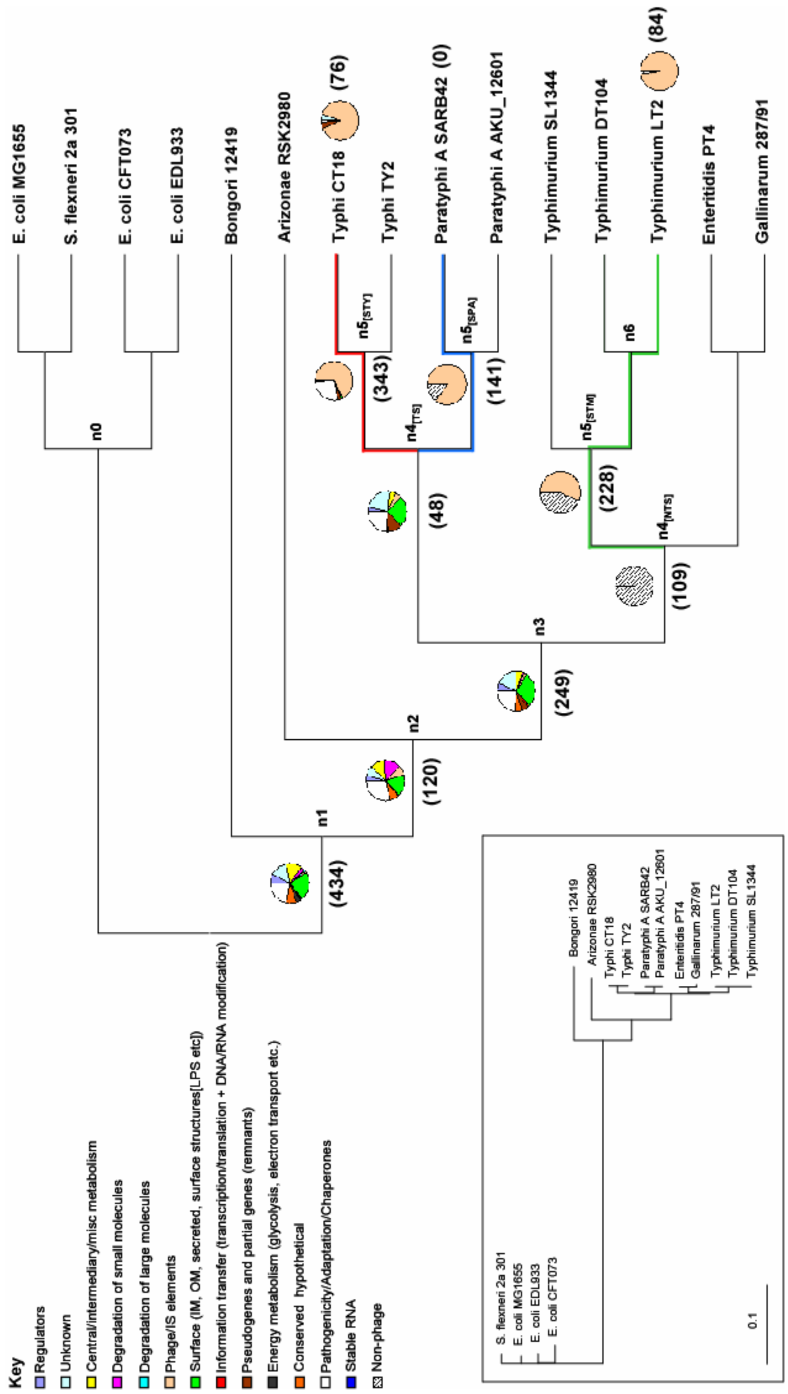


Figure 3.10: Numerical and functional distribution of PHA genes. The cladogram (main) shows the phylogenetic relationship between the 15 genomes used in this study, ignoring branch length. The topology of the tree is based on whole-genome sequence alignment. For the true phylogenetic distance with the respective branch lengths drawn to scale refer to the phylogram detailed in the inset of this figure; the phylogram is built using the ML method exploiting the F84 model. Numbers within parenthesis (main) reflect the number of PHA genes. Pie charts on each branch represent the functional classification of genes based on the colour-class detailed in the key. The non-phage functional class (black and white diagonal hatching) was introduced to classify CDSs without colour-coded functional classification in their annotation; those CDSs assigned into the "non-phage" pseudo-class represent CDSs that belong to any of the thirteen functional classes apart from the phage class. Numbers of genes on branches 1, 2 and 3 reflect the intersection of the respective number of genes determined on each branch using one of the three query genomes; the same applies for genes assigned to branch $4_{[TS]}$.

Table 3.5: A list of PHA genes, and their inferred relative time of insertion.

S. typhi CT18		*S. paratyphi* A SARB42		*S. typhimurium* LT2	
Relative time of insertion	**PHA genes**	**Relative time of insertion**	**PHA genes**	**Relative time of insertion**	**PHA genes**
Branch 1	493	Branch 1	434	Branch 1	473
Branch 2	124	Branch 2	120	Branch 2	128
Branch 3	316	Branch 3	268	Branch 3	249
Branch 4 [TS]	62	Branch 4 [TS]	48	Branch 4 [NTS]	109
Branch 5 [STY]	343	Branch 5 [SPA]	141	Branch 5 [STM]	228
Branch CT18	76	Branch SARB42	0	Branch LT2	84
Total	1,414	Total	1,011	Total	1,271

This analysis revealed a surprisingly high number of 434 PHA genes inserted at the base of the *Salmonella* lineage (branch 1). Based on two independent previous studies (Doolittle *et al.*, 1996; Ochman and Wilson, 1987) the divergence of the *E. coli* and *Salmonella* lineage occurred approximately 100-140 Myr ago. Consequently putative HGT events on branch 1 represent ancient insertions, close to the divergence of these two lineages and include 76 coding sequences (CDSs) of “ancient” SPIs such as SPI-5, SPI-4, a part of SPI-2 (ttr-region), SPI-9, SPI-1 and a part of SPI-3 (magnesium transport ATPase – mgt region).

The *cob* operon of *S. enterica*, which encodes vitamin B12 biosynthesis, has been previously shown to be horizontally acquired in the *Salmonella* lineage following its divergence from the *E. coli* lineage (Lawrence and Roth, 1995; Lawrence and Roth, 1996). In a later study, Lawrence and Ochman (Lawrence and Ochman, 1997) showed, using a model of reverse amelioration, that the *cob* operon was probably introduced into the *Salmonella* lineage 71 Myr ago. The current analysis assigned the *cob* operon to branch 2 which predates the divergence of *S. arizonae* from the *S. enterica* lineage. Based on the data available, we can infer that the divergence of *S. arizonae* from the *S. enterica* lineage occurred approximately 100-71 Myr ago, and further suggest that the 120 inferred PHA genes assigned to branch 2 have an absolute time of insertion of the same order of magnitude.

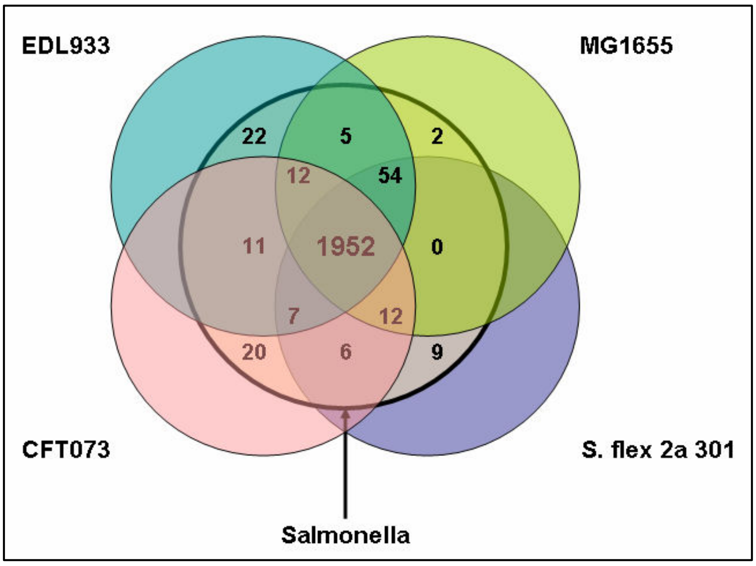


Figure 3.11: Venn diagram illustrating the orthologous genes shared between all the 11 *Salmonella* strains (bold circle in the middle) and the genomes of *E. coli* MG1655, *E. coli* EDL933, *E. coli* CFT073 and *S. flexneri* 2a 301. The number highlighted in bold, represents the total number of orthologous genes (core genes) shared between the 15 genomes used in this study.

On branch 3 (*S. enterica* lineage), there are 249 inferred PHA genes. On this branch are found SPIs that are restricted to the *S. enterica* lineage, such as part of SPI-3 (3' end), part of SPI-10 (fimbrial-sef operon), SPI-6, SPI-16 and SPI-17. Finally on more recent branches, i.e. branch $5_{[\mathrm{STY}]}$ (STY: *S. typhi*), branch $5_{[\mathrm{SPA}]}$ (SPA: *S. paratyphi* A), branch $5_{[\mathrm{STM}]}$ (STM: *S. typhimurium*) and strain-specific genes, (relative to each of the three query genomes), I have inferred a significant number of putative HGT events which are mainly dominated by CDSs that belong to annotated prophage structures (discussed in more detail below).

3.3.2 Functional analysis of PHA genes

Implementing a classification of 14 functional classes, listed in Figure 3.10, each of the PHA genes, with a given relative time of insertion, was assigned into one of the 14 colour-coded functional classes. The results are summarized, via pie charts assigned to each branch, in Figure 3.10. Overall, from this functional classification, it is clear that PHA genes on branches 1-3, branch 4$_{[TS]}$ (Typhoidal *Salmonella*) and branch 4$_{[NTS]}$ (Non-Typhoidal *Salmonella*) show a wide distribution over almost all the 13 functional classes (e.g. cell-surface, regulation, central metabolism, pathogenicity), while gene-remnants/pseudogenes are mainly restricted to recently diverged lineages, i.e. the *S. enterica* species. Moreover CDSs that belong to annotated structures of prophages (light pink-coloured functional class) are predominant in very recent lineages (i.e. on branches 5$_{[STY], [SPA], [STM]}$, or strain-specific CDSs).

On branch 4$_{[TS]}$, which predates the Typhi-Paratyphi A divergence, overall 24% of PHA genes have unknown function, 26% encode cell surface-related components, 11% are remnants/pseudogenes and 24% are related to pathogenicity or adaptation. Also on this branch are the CDSs of a 8.5kb, previously uncharacterized, Genomic Island (GI) at position 2187521-2195992bp, of very low G+C (36.29%) content that encodes 16 CDSs (STY2349-STY2364 in CT18) of unknown function, without significant similarity with previously annotated CDSs. Furthermore, this novel GI does not have any of the “classical” GI-related features e.g. direct/inverted repeats, integrase gene or insertion adjacent to RNA locus. Details about the composition of this putative GI and other genes assigned to branch 4$_{[TS]}$ will be discussed in the following section.

The functional analysis of the PHA genes assigned to recent branches (branches 5$_{[STY], [SPA], [STM]}$ and strain-specific) is in line with a previous study focused on *E. coli* MG1655 showing that IS elements and prophage remnants represent mostly very recent insertion events in MG1655 (Lawrence and Ochman, 1998); the same study suggests that very few acquired DNA sequences are maintained for more than 10 Myr in the genome of *E. coli* MG1655. In the current study, there is no complete-intact prophage structure, inserted at the base of *Salmonella* lineage that is present in all the 11 *Salmonella* strains or even prophages inserted in the *S. enterica* lineage that are shared between the Typhi, Paratyphi A and the Typhimurium strains. Using Typhi CT18 as a query genome, on branch $5_{[STY]}$, 67% (231) of PHA CDSs belong to prophage structures, while 93% (71) of CT18-restricted PHA CDSs are of phage origin. Similarly, in the case of Typhimurium LT2, 57% and 98% of PHA genes that are on branch $5_{[STM]}$ and LT2-restricted, respectively, belong to annotated prophage structure. In the lineage of Paratyphi A, 85% of PHA CDSs acquired on branch $5_{[SPA]}$ are of phage origin; interestingly there are no SARB42-specific CDSs relative to Paratyphi A AKU_12601.

In a previous study, Thomson *et al.* (Thomson *et al.*, 2004) provided data showing that many prophage structures present in Typhi CT18 are predicted to be Typhi-specific, further suggesting that these bacteriophages have a level of specialization for their host and play a key role in generating genetic diversity in the *S. enterica* lineage. Moreover the same authors suggested that Typhi has indeed a unique pool of prophage elements that distinguish it from other serovars, in contrast with the *Salmonella* specific SPIs which show a wider distribution within the *Salmonella* lineage (Ochman and Groisman, 1996).

Generally in microbial genomes, some PHA genes are retained over long evolutionary distances and therefore contribute to species diversification (Lawrence, 1999; Lawrence, 2001), while PHA genes that might be detrimental, or not advantageous for the host are rapidly removed (Lawrence *et al.*, 2001; Lawrence and Ochman, 1998). Horizontally acquired DNA is more likely to be deleted than are native, core genes; for example, prophage structures often harbor direct repeats forming their endpoints i.e. phage attachment sites attL (left) and attR (right) that can, via homologous recombination, efficiently remove those “parasitic” elements. Furthermore, some prophage genes can be detrimental (e.g. the *N* gene of bacteriophage λ), neutral (e.g. integrases) or advantageous (e.g. immunity repressors) (Lawrence *et al.*, 2001). Based on this model, parasitic-detrimental DNA sequence (e.g. prophage elements) is removed by sequential deletion over time. This bias of deletion over insertion (Andersson and Andersson, 1999) can equilibrate HGT events, and this is further supported by the comparable genome size of closely related genomes (Bergthorsson and Ochman, 1998). Overall the current study suggests that indeed prophage structures are not retained for a long time in the *Salmonella* lineage, while complete, intact prophage structures represent very recent insertions in the Typhi, Paratyphi A and the Typhimurium lineage, which based on their impact (detrimental, neutral or advantageous) on the host, will eventually be retained or removed from those genomes.

3.3.3 Compositional analysis

The aim of the compositional analysis in this study was to determine if there is any clear trend for genes assigned to relatively old branches in the reference tree topology to show sequence composition closer (compared to more recent insertions) to the average composition of the host genome, thus supporting the effect of amelioration as a time-dependent process. It should be noted that because this analysis is focused on the effects of the amelioration in the *Salmonella* lineage which diverged fairly recently from *E. coli* and the rest of the enteric bacteria, we expect to identify, if any, mild effects of the amelioration on the sequence composition of the gene datasets under study. For example, Daubin and Ochman (Daubin and Ochman, 2004) applying a similar approach on a much broader phylogenetic sample (the γ-Proteobacteria), showed a strong correlation between the G+C content and different phylogenetic depths in their reference tree topology.

As a starting point for the compositional analysis of PHA genes, I applied the Alien_Hunter algorithm, which implements the Interpolated Variable Order Motifs (IVOMs) method (Vernikos and Parkhill, 2006), to the three query genomes, and performed a benchmarking analysis of its sensitivity versus the inferred relative time of insertion of PHA genes; the results are shown in Figure 3.12. Overall it can be concluded that the sensitivity of this HGT prediction method correlates strongly with the relative time of insertion. Indeed, in all the three query genomes regression analysis showed a correlation ($0.45 \leq R^2 \leq 0.74$) between the sensitivity and the relative time of insertion. For example, PHA genes inserted at the base of *Salmonella* lineage, e.g. on branch 1, can be identified with a False Negative (FN) rate of 0.55 while more recent insertions with a much lower FN rate of 0-0.2. It is worth noting that the high sensitivity of Alien_Hunter on very recent branches is in contrast with the drop in the IVOMs score distribution (Figure 3.13); the majority of the PHA genes assigned to these branches belong to prophage structures, consequently their clustering and not their composition should mainly explain the high sensitivity of this algorithm on these branches. It is important to note that the analysis of the sensitivity of this algorithm relies on the assumption that all the PHA genes identified in the current analysis are true horizontally acquired genes and the conclusions drawn about its performance are specific for this set of PHA genes.

Calculating the G+C content, both overall and codon position specific, as well as higher order compositional biases, implementing the IVOMs method, the amelioration process versus the relative time of insertion of PHA genes was monitored (Figure 3.13, Figure 3.14). Using Typhi CT18 and Paratyphi A SARB42 as query genomes this analysis revealed that there is a clear correlation (R^2 = 0.98 for branches 1-3, R^2 = 0.65 for branches 1-$4_{[TS]}$) between the G+C content or the IVOMs score of PHA genes and the relative time of their insertion on the earlier branches; however, this strong correlation seems to "break down" in the case of very recent putative HGT events, i.e. insertions that took place after the divergence of Typhi and Paratyphi A lineages (Figure 3.13, Figure 3.14).

For example genes assigned to branches 1 and 2 show an average G+C content of 51.4% and 50.6% respectively, close to the average gene G+C content of 53.2% and 53.3% (CT18 and SARB42 respectively). The same observation becomes much clearer when calculating higher order compositional biases (Figure 3.13). Based on the IVOMs score, genes on branches 1 and 2 have an average score of 0.06 and 0.063 respectively while more recently acquired genes, i.e. on branches 3 and 4[TS] have a score of 0.072 and 0.093 respectively; the average, genome-wide IVOMs score in Typhi CT18 is 0.059.

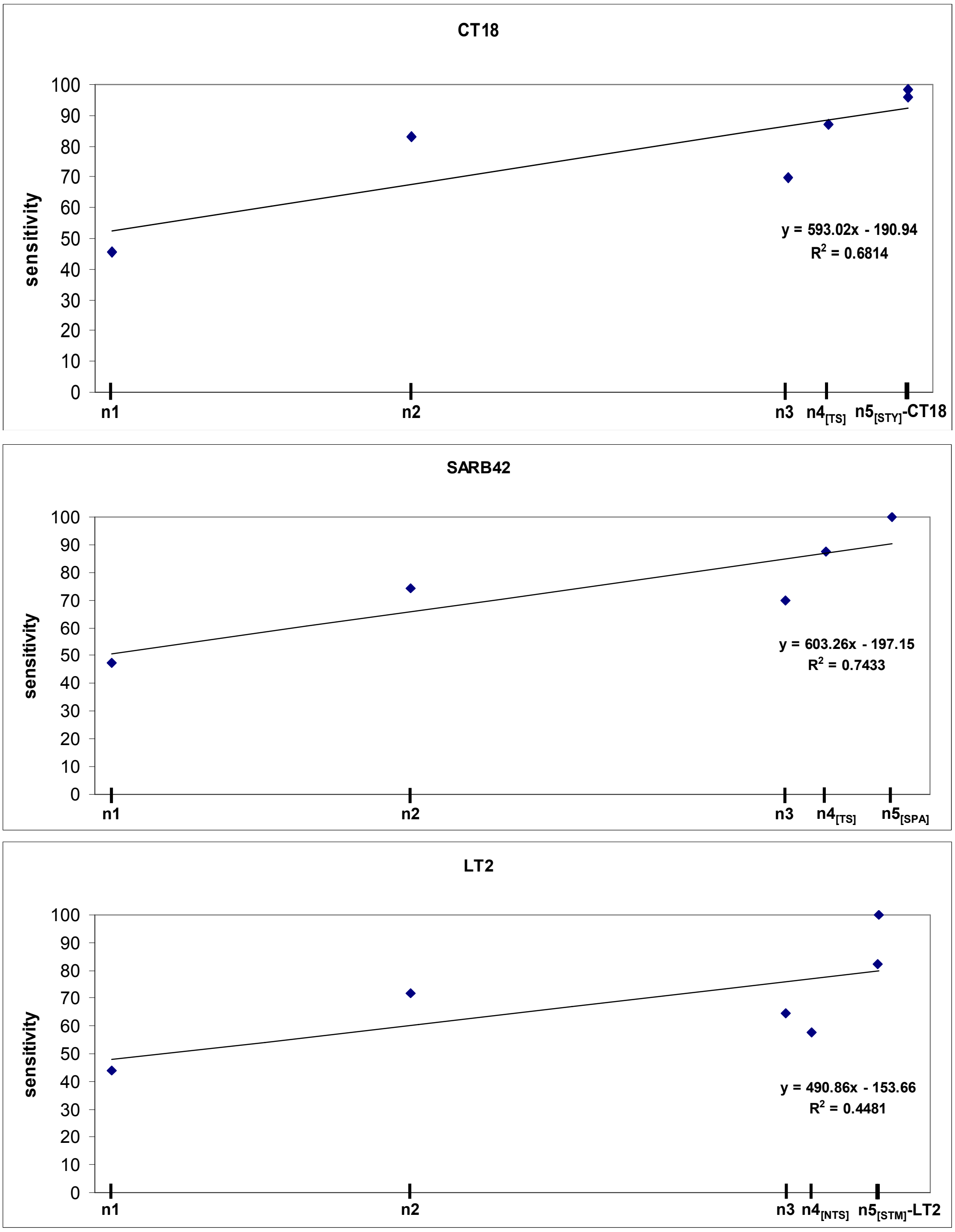


Figure 3.12: Sensitivity of the Alien_Hunter algorithm, which implements the IVOMs method, versus the inferred relative time of insertion of PHA genes for the three query genomes: *S. typhi* CT18 (top), *S. paratyphi* A SARB42 (middle), *S. typhimurium* LT2 (bottom). The nodes on the X axis are scaled according to the respective branch lengths of the tree topology shown in the inset of Figure 3.10. Regression analysis is provided embedded within the three graphs; *p*-values: 0.04, 0.05 and 0.14 respectively.

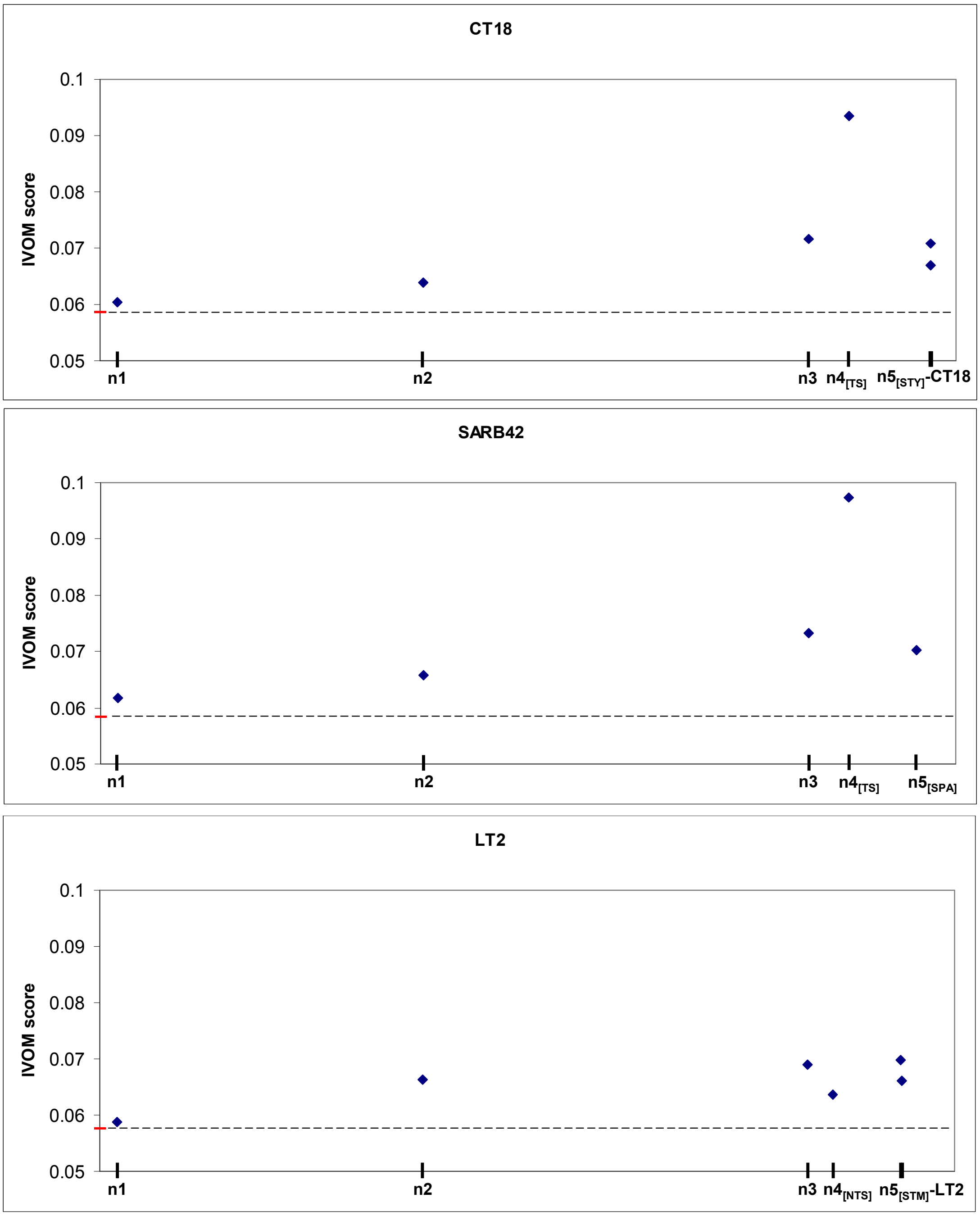


Figure 3.13: Average score, taking into account higher order compositional biases, of putative horizontally acquired genes, versus the inferred relative time of insertion, in the three query genomes: *S. typhi* CT18 (top), *S. paratyphi* A SARB42 (middle), *S. typhimurium* LT2 (bottom). The score is calculated implementing the IVOMs method. The average score for the three query genomes is highlighted in red (the embedded dashed line is provided for ease of comparison). The nodes on the X axis are scaled according to the respective branch lengths of the tree topology shown in the inset of Figure 3.10.

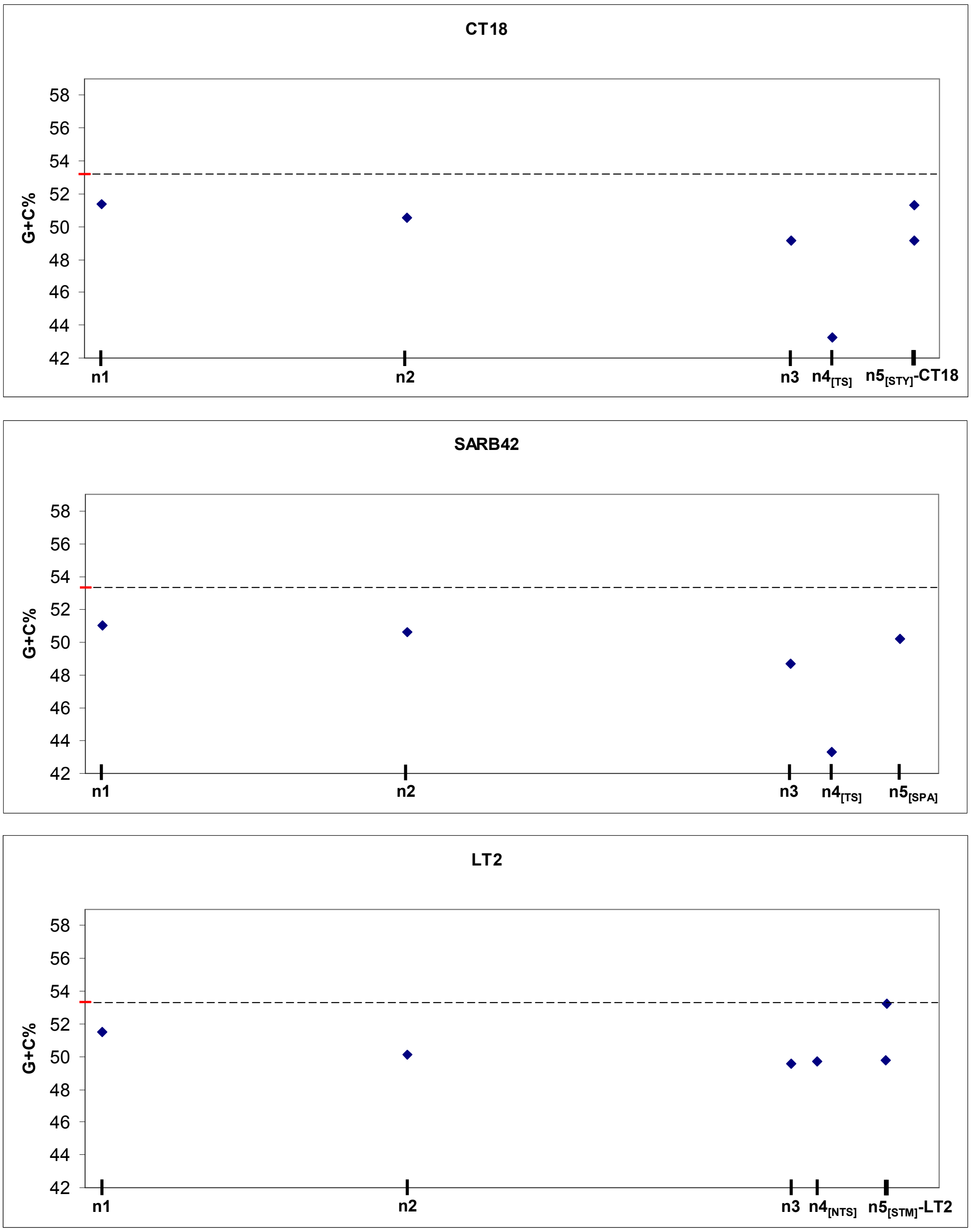


Figure 3.14 Average G+C content of putative horizontally acquired genes, versus the inferred relative time of insertion, in the three query genomes: *S. typhi* CT18 (top), *S. paratyphi* A SARB42 (middle), *S. typhimurium* LT2 (bottom).The average G+C content for the three query genomes is highlighted in red (the embedded dashed line is provided for ease of comparison). Error bars could not be visualized (the standard deviation is in the range of 0.05-0.08). The nodes on the X axis are scaled according to the respective branch lengths of the tree topology shown in the inset of Figure 3.10.

A similar observation can be made for Typhimurium LT2. More specifically, there is a very strong correlation ($R^2 = 0.89$) between G+C content or IVOMs score and the relative time of insertion which breaks-down on branches descendent of node 3 (Figure 3.13, Figure 3.14). More specifically, the average G+C content of genes assigned to branches 1, 2 and 3, is 51.5%, 50% and 49.6% respectively while for genes on the branch $4_{[NTS]}$, the average G+C content is 49.7%. Similarly, using the IVOMs method, the corresponding scores for the four branches are: 0.059, 0.066, 0.069 and 0.064 respectively.

PHA genes assigned to branch $4_{[TS]}$ on the Typhi-Paratyphi A lineage show a very strong compositional deviation, indicated both by their very low G+C content of 43.3% (gene average: 53.2%) and the IVOMs score of 0.093 (genome average: 0.059). Furthermore, the codon-position specific G+C content of genes assigned to branch $4_{[TS]}$, deviates strongly (GC_1 = 49%, GC_2 = 37%, GC_3 = 43%) (Figure 3.15) from the expected values (GC_1 = 59%, GC_2 = 41%, GC_3 = 56%, respectively) based on the three linear equations (13, 14, 15) provided by Lawrence and Ochman (Lawrence and Ochman, 1997). Those three linear equations are based on the observation that the G+C% content at the three codon positions shows a linear positive correlation with the genomic (i.e. genome average) G+C% content, although with different rates of correlation (Muto and Osawa, 1987), Figure 3.15:

$$GC_1 = 0.615 \times GC_{Genome} + 26.9$$

$$GC_2 = 0.270 \times GC_{Genome} + 26.7$$

$$GC_3 = 1.692 \times GC_{Genome} + 32.3$$

Source: (Lawrence and Ochman, 1997).

Therefore, the above linear equations can be used to infer the level of departure from those expected values of codon-position specific G+C% content and evaluate the level of amelioration of PHA genes; the codon-position specific G+C% content of PHA genes that have been recently horizontally acquired, follow the expected GC_1, GC_2 and GC_3 values based on the GC_{Genome} of their donor; on other hand PHA genes that have been acquired a long time ago, follow the expected GC_1, GC_2 and GC_3 values based on the GC_{Genome} of the their “new” host; PHA genes still undergoing the amelioration process are expected to fall somewhere in between.

The G+C content of the second codon position is generally very constrained to similar values across species (Lawrence and Ochman, 1997). Interestingly, genes assigned to branch $4_{[TS]}$ in Typhi-Paratyphi A lineage show a significant deviation also in this compositionally well-conserved codon position possibly suggesting a distantly related donor genome (Figure 3.15).

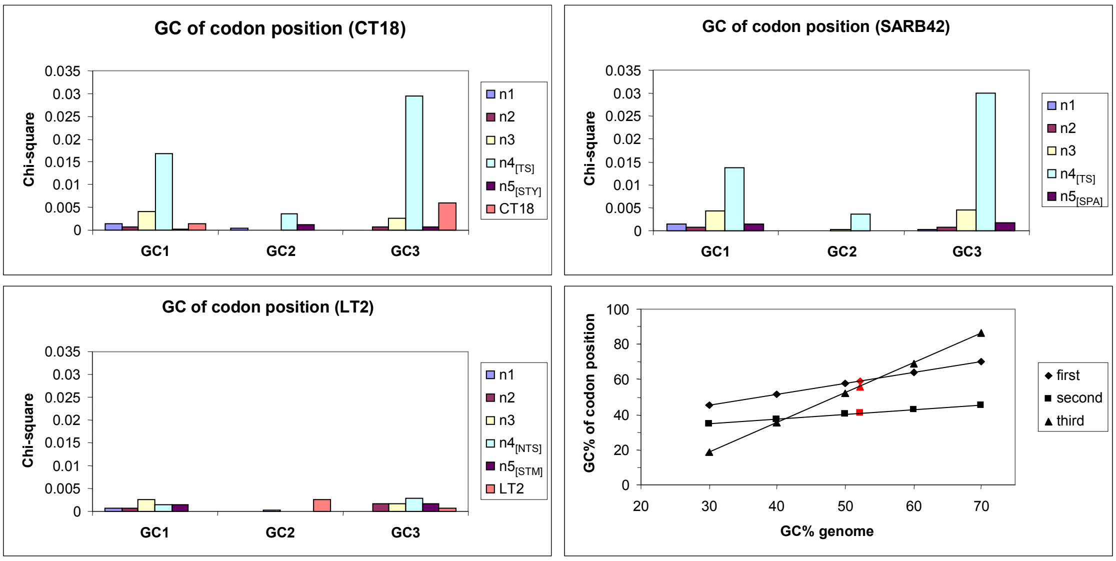


Figure 3.15 Chi-square values of G+C content over the three codon positions, for genes assigned to lineages of increasing depth in the reference tree topology. Chi-square values are calculated using the expected G+C codon-position values derived from the three linear equations (13, 14, 15) provided by Lawrence and Ochman (Lawrence and Ochman, 1997). At the right-bottom side of the figure, the correlation between genomic G+C content and G+C content at the three codon positions based on the data provided by Muto and Osawa (Muto and Osawa, 1987), is provided. Genes that are still under the amelioration process are expected to deviate from those expected values. The expected G+C content for each codon position in the *Salmonella* lineage is highlighted in red.

Codon usage analysis revealed that genes on branch $4_{[TS]}$ show a bias towards A+T rich codons (Figure 3.16). For example, the ‘AAA’ codon is overrepresented in CDSs of this branch, compared to its average frequency in the genome; the AAA codon (encoding lysine) has been previously shown to be overrepresented in highly expressed genes (Sharp and Li, 1987). To test further whether genes on this branch deviate compositionally due to their highly expressed pattern, rather than their alien origin, I performed a CAI analysis (summarized in Table 3.6). It can be clearly seen that genes on branch $4_{[TS]}$ deviate compositionally from the genome background composition, more likely due to their alien origin, rather than their high rate of expression, representing the "left ear" in the "rabbit-like" codon bias vs CAI plot described in (Karlin *et al.*, 1998).

Indeed, genes on branch $4_{[TS]}$ show an average CAI value of 0.221, significantly lower (*p*-value = 4.95 10^{-13}) than the average gene CAI value (= 0.31) and much lower than the CAI values of highly expressed genes, e.g. ribosomal protein coding (CT18: 0.554, SARB42: 0.560, LT2: 0.561) and aminoacyl-tRNA synthetase genes (CT18: 0.437, SARB42: 0.453, LT2: 0.434). Furthermore, the CAI analysis revealed that genes inferred in this study of being PHA do not show CAI values of highly expressed genes, and overall their CAI values are significantly lower (*p*-value = 3.75 10^{-74}) than the average gene CAI values.

Overall, using any of the three query genomes (CT18, SARB42, LT2) this analysis indicates that very recent acquisitions e.g. on branches $5_{[STY], [SPA], [STM]}$ seem to have been equally "ameliorated" with acquisitions on older branches e.g. branches 1, 2; moreover, in the case of LT2 genome, strain specific acquisitions (see LT2 branch) show sequence composition very close to the genome composition. Very recent acquisitions are expected to deviate strongly from the host backbone composition, unless the donor is very close compositionally to the host. Amelioration, a time-dependent process, can not have significantly affected their sequence composition, which should still reflect mostly the donor rather than the host specific compositional signature. However, recent acquisitions identified in this study either show very close composition to the host backbone composition, for example PHA genes on LT2 branch have an average G+C content of 53.26% very close to the gene average G+C of

53.33%, or deviate compositionally equally to PHA genes acquired on older branches; for example the G+C content of PHA genes in CT18 on branches 1 and $5_{[STY]}$ is 51.4 and 51.3 respectively. Similarly the G+C content of PHA genes in SARB42 on branches 2 and $5_{[SPA]}$ is 50.6% and 50.2% respectively.

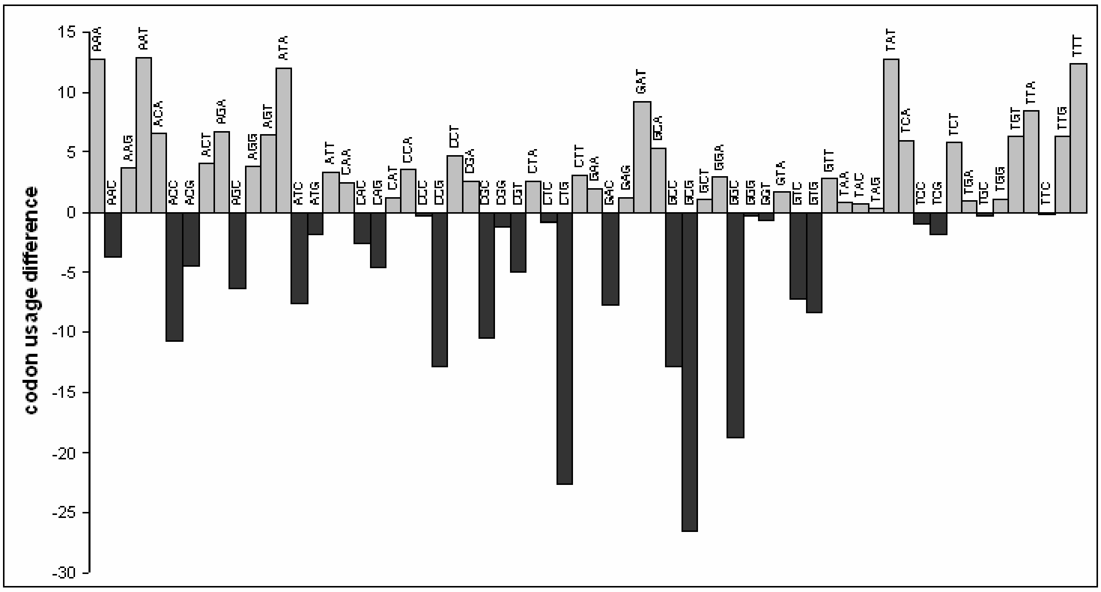


Figure 3.16: Codon usage difference of CDSs assigned on branch $4_{[TS]}$ relative to the average codon usage in Typhi CT18. Positive values in the Y axis indicate overrepresentation (light grey bars) of certain codons in CDSs of this branch relative to the average codon usage and vice versa.

Interestingly, branches descendant of nodes $4_{[TS]}$ and $4_{[NTS]}$ are dominated by genes of phage origin (57-98% of genes at the given relative time of insertion), Figure 3.10. For example on branch $5_{[STY]}$, 67% of Typhi CT18 genes assigned to this branch belong to one of the six prophage structures present both in Typhi CT18 and TY2. On branch $5_{[STY]}$, SPI-7 and the phage-related gene G+C content is 50.87% and 51.98% respectively. In a previous study, it has been shown that the last common ancestor of Typhi existed 15,000-150,000 years ago, during the human hunter-gatherer period (Kidgell *et al.*, 2002); consequently PHA genes assigned to branch $5_{[STY]}$, have a time of insertion of the same order of magnitude. Similarly, in Typhimurium LT2, there are two prophage (Fels-1, Fels-2) structures that represent very recent acquisitions (LT2-specific), and are absent from the other two Typhimurium strains. CDSs of these prophage elements have an average G+C content of 53.57% and 52.94% respectively, while their CAI value is 0.307, very close to the LT2 genome average CAI of 0.313.

Table 3.6: Average CAI values for genes of different inferred relative time of insertion for the three query genomes. Average CAI values for all genes in the genome, ribosomal protein coding and aminoacyl-tRNA synthetase genes, are also provided as a reference. Genes ≤ 300bp were excluded. The reference gene set of highly expressed genes was the one proposed by Sharp and Li (Sharp and Li, 1986) using the genome of *E. coli*.

S. typhi CT18		*S. paratyphi* A SARB42		*S. typhimurium* LT2	
Genes	**CAI**	**Genes**	**CAI**	**Genes**	**CAI**
PHA on branch 1	0.264	PHA on branch 1	0.264	PHA on branch 1	0.264
PHA on branch 2	0.258	PHA on branch 2	0.258	PHA on branch 2	0.258
PHA on branch 3	0.256	PHA on branch 3	0.256	PHA on branch 3	0.256
PHA on branch 4 [TS]	0.221	PHA on branch 4 [TS]	0.221	PHA on branch 4 [NTS]	0.275
PHA on branch 5 [STY]	0.283	PHA on branch 5 [SPA]	0.297	PHA on branch 5 [STM]	0.269
PHA on branch CT18	0.282	PHA on branch SARB42	NA	PHA on branch LT2	0.307
All genes	0.310	All genes	0.315	All genes	0.313
Ribosomal	0.554	Ribosomal	0.560	Ribosomal	0.561
tRNA synthetase	0.437	tRNA synthetase	0.453	tRNA synthetase	0.434

3.4 Discussion

The aim of this analysis was to study the distribution of PHA genes in a time-dependent manner i.e. to infer the relative time of insertion based on the reference tree topology, throughout the *Salmonella* lineage, applying an extensive comparative analysis between 11 *Salmonella*, three *E. coli* and one *Shigella* strain. The selection of four genome sequences that form an outgroup of the *Salmonella* lineage was made in order to differentiate gene loss from gene gain more reliably, two mechanisms that could explain the presence of a gene in one lineage and its absence from a sister, closely related lineage.

However, because the *E. coli* and *Salmonella* lineage represent very closely related, sister lineages the 434 PHA genes inferred to have been acquired at the base of the of the *Salmonella* lineage might equally represent deletion events in the *E. coli* lineage subsequent to the common ancestor with *Salmonella*. To investigate further this alternative scenario, I used a set of three more distantly related enteric outgroup genomes: *Erwinia carotovora* SCRI1043 (accession number: BX950851), *Yersinia enterocolitica* 8081 (accession number: AM286415) and *Y. pseudotuberculosis* IP32953 (accession number: BX936398). Less than 5% of the 434 PHA genes inferred to have been acquired on branch 1 have orthologous genes present in this distant outgroup. These data suggest that the majority (>95%) of 434 PHA genes most likely represent true HGT events that occurred quite early in the evolution of the *Salmonella* lineage, rather than deletion events in the *E. coli* lineage.

In the current study I exploited a much larger sequence sample, i.e. whole genome sequence, rather than selected gene/protein sequences to serve as “molecular chronometers”, thus the phylogenetic signature seems to be strong enough for the NJ and ML method to result in identical tree topologies, inferring the same phylogenetic history of the query genomes at hand. However, care should be taken when interpreting whole-genome sequence based phylogenies, since extensive HGT events, homologous recombination or other homoplastic events might well obscure the true phylogenetic history of the genomes under study (Doolittle and Papke, 2006) whose phylogeny may therefore be more efficiently described using phylogenetic nets rather than single tree topologies (Doolittle, 1999; Hilario and Gogarten, 1993; Martin, 1999).

For example, Didelot *et al.* (Didelot *et al.*, 2007) showed that Paratyphi A and Typhi genomes are, over 75% of their sequence, distantly related *S. enterica* members (both in terms of nucleotide divergence and gene content), while the remaining 25% of their sequence is much more similar (average nucleotide divergence 0.18%, instead of 1.2%); the authors suggested that the two genomes have recently exchanged, via homologous recombination, a significant amount of DNA, now representing seemingly similar lineages (convergence evolution). In the current analysis, the two lineages, have been mapped on very close branches of the *Salmonella* phylogenetic tree (Figure 3.10), representing perhaps a limitation of the applied, strictly bifurcating tree topology; using instead a reticulate phylogenetic network, would have probably (correctly) mapped the two seemingly similar lineages on more distant branches, taking also into account (by means of multi-furcating branches connecting the two lineages) the extensive amount of exchanged homologous DNA.

It is worth noting that whole-genome based phylogenetic approaches are capturing the “overall” phylogenetic signal based on whole chromosome sequences. In the case of very closely related organisms, e.g. strains of the same serovar, minor differences in terms of gene content (e.g. prophages, GIs) cannot be reliably represented in the “overall” phylogenetic signal. In other words, whole genome-based phylogenies focusing on a wide range of strains may suffer from low resolution in the case of very closely related genomes. Moreover mobile elements may show similarity on the sequence level (e.g. prophages) but differ on the structural level (i.e. different phage types). Relying on sequence information only, these seemingly similar mobile elements will bias the relatedness of closely related strains (e.g. the three Typhimurium strains used in this study).

The reason why I pursued a comparative, rather than a compositional based approach (i.e. defining PHA genes based simply on their compositional deviation, but ignoring their distribution throughout the lineage of interest), was the fact that compositional based approaches frequently underestimate the true number of HGT events (Lawrence and Ochman, 1997), either due to the amelioration process, in the case of ancient insertions, or due to compositionally similar donor genomes, in the case of new insertions. The current comparative analysis suggests that approximately 30, 25 and 28% of protein-coding sequences in Typhi CT18, Paratyphi A SARB42 and Typhimurium LT2 respectively, represent putative HGT events. The distribution of those PHA genes on different branches of the reference tree topology reveals that approximately 35-40%

of them were acquired at the base of the *Salmonella* lineage (branch 1), very close to its divergence from *E. coli*, reflecting perhaps the acquisition of genes that enabled the exploration of new niches e.g. the acquisition of SPI-1 which enabled *Salmonella* to invade epithelial cells (Galan, 1996). Moreover, 20% of those genes were acquired at the base of the *S. enterica* lineage (branch 3); overall 60-70% were inserted after the divergence of the *Salmonella* from the *E. coli* lineage and prior to the divergence of the *S. enterica* subspecies. This suggests that approximately 60-70% of the putative HGT events are probably shared between most of the subspecies of the *S. enterica* lineage.

Based on the functional classification of genes assigned to branches 1, 2 and 3 that predate the *S. enterica* lineage, it becomes evident that generally, genes within almost all functional classes, e.g. regulation, energy metabolism, cell surface, virulence-related, have been horizontally acquired. Moreover the functional distribution of genes assigned to branches 1 and 3 correlates strongly with the functional distribution of all the genes in the genome (R: 0.71 and 0.63 – p-value: 0.01 and 0.03 respectively) whereas this correlation is weaker (and not significant) for genes on branches 2 and 4 (R: 0.54 and 0.45 – p-value: 0.07 and 0.1 respectively) and disappears (R<0.01, p-value: 0.98) completely in the case of very recent branches (branch 5 or genome-specific).

On branches 1-4 there is a fairly constant percentage of genes encoding cell-surface structures (18-28%), genes related to pathogenicity and adaptation (22-29%) and regulatory elements (4-8%). Furthermore, the percentage of genes with unknown function ranges from 8-18%, while fragmented gene-remnants (pseudogenes) account for 6% and 11% on branches 3 and 4[TS] respectively with almost no pseudogenes (< 0.1%) on branches 1 and 2. The increased number of genes acquired at the base of *S. enterica* lineage that have been inactivated suggests that some of these early-acquired functions are no longer necessary, and are being lost in these serovars. The increased number of pseudogenes (11%) in the Typhi-Paratyphi A lineage that are absent from the Typhimurium lineage supports a genome degradation process via pseudogene formation suggested to be due to the recent change in niche of these serovars (Parkhill *et al.*, 2001).

The compositional analysis of the inferred PHA genes indicates that there is indeed a strong correlation between the time of insertion and amelioration towards the host-specific genomic signature. In other words, anciently horizontally acquired genes have ameliorated more towards the host composition, compared to more recent acquisitions. However, even HGT events inferred to have inserted at the base of the *Salmonella* lineage still preserve some of their donor genome sequence signature, as indicated by their overall and codon-position specific G+C content, suggesting that these genes are still undergoing the amelioration process.

On the other hand, in the case of very recent acquisitions that represent mostly insertion of prophage elements, it seems that their sequence composition is already much closer to the host background composition, presumably not due to the amelioration process, since they have been acquired fairly recently, but rather due to an adaptation to the specific sequence signature of the their host. Perhaps the compositional adaptation of those prophages is pivotal for the masking of their alien sequence identity in order to successfully integrate into the bacterial chromosome, without being detected by the histone-like nucleoid structuring (H-NS) protein that selectively silences horizontally acquired DNA of lower G+C content than the host genome (Navarre *et al.*, 2006).

If we take into account both the absence of complete-intact prophage structures from old branches (1-3, $4_{[\mathrm{TS}]}$ and $4_{[\mathrm{NTS}]}$), and the significant compositional similarity of those prophage-related genes to the host sequence composition, when the effects of the amelioration process are expected to be mild, it would be tempting to speculate that prophage elements in the *Salmonella* lineage have undergone an adaptation to specific serotypes. However this hypothesis does not explain why anciently inserted prophages e.g. those inserted at the base of *Salmonella* lineage, prior to the divergence of *S. bongori* and *S. arizonae* from the *S. enterica*, have not been retained in descendent lineages e.g. the Typhi, Paratyphi A and Typhimurium strains.

Perhaps anciently inserted bacteriophages at the base of the *Salmonella* lineage carried genes that were either neutral or detrimental, providing no profound advantage to the host, and over time the host has lost those parasitic elements via a deletion process which has left behind molecular fossils of those elements. This observation is further supported by the absence of pseudogenes on very old branches, i.e. branches 1 and 2; perhaps the ongoing time-dependent process of deleting redundant or detrimental DNA sequence has already removed a much higher proportion of pseudogenes on very old branches, compared to recent ones further suggesting that genome degradation is still a continuous process in the *Salmonella* lineage (Lawrence *et al.*, 2001).

3.5 Conclusions

Overall the current analysis has shown that the impact of amelioration, a time-dependent process, is still detectable even in fairly recent HGT events, e.g. that occurred 100-140Myr ago; moreover it sheds more light on the relative time of insertion of HGT events in the *Salmonella* lineage, and presents data that show that prophage structures are not retained for long periods in the *Salmonella* lineage.

Whether this last observation is related to an ongoing genome degradation process that over time removes redundant or detrimental DNA sequences, equilibrating the horizontal influx of genes, maintaining a fairly constant genome sequence size, still remains to be clarified. Perhaps the study of the very recently acquired prophage elements which seem to account for the majority of the strain or serovar specific genes (McClelland *et al.*, 2004; Thomson *et al.*, 2004), and their impact (detrimental, neutral, advantageous) on the evolution, life-style and host adaptation of the *Salmonella* strains, might shed more light on the underlying principles of the observed genome degradation process.

The prophage elements present in the *Salmonella* lineage show a very close sequence composition to the host-specific background composition, strongly suggesting that those parasitic elements have specialized and adapted to their hosts, playing a key role in driving bacterial evolution (Thomson *et al.*, 2004), or even speciation itself supporting the notion of "evolution in quantum leaps", introduced by Groisman and Ochman (Groisman and Ochman, 1996). Overall, the distribution of PHA genes in the *Salmonella* lineage coincides strongly with the divergence of the major *Salmonella* species, underlining the major impact of horizontal transfer in the evolution of the salmonellae.