UNIVERSITY OF QUEBEC AT MONTREAL

FUNCTIONAL GENOMIC REGION AND HORIZONTAL GENE TRANSFER DETECTION USING SEQUENCE VARIABILITY CLUSTERING: APPLICATIONS TO VIRAL AND PROKARYOTIC EVOLUTION

DISSERTATION
PRESENTED
AS A PARTIAL REQUIREMENT
OF THE DOCTORATE IN COMPUTER SCIENCE

BY
DUNAREL BADESCU

NOVEMBER 2015

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

DÉTECTION DES RÉGIONS GÉNOMIQUES FONCTIONNELLES ET DES TRANSFERTS HORIZONTAUX DE GÈNES PAR LA VARIABILITÉ DES SÉQUENCES ET L'ANALYSE DE REGROUPEMENTS : APPLICATIONS À L'ÉVOLUTION DES VIRUS ET DES PROCARYOTES

THÈSE
PRÉSENTÉE
COMME EXIGENCE PARTIELLE
DU DOCTORAT EN INFORMATIQUE

PAR
DUNAREL BADESCU

NOVEMBRE 2015

To my beloved son Elian and wife Mihaela

ACKNOWLEDGMENTS

I am heartily thankful to my two supervisors, Professors Vladimir Makarenkov and Abdoulaye Baniré Diallo, for their guidance and support as well as for the gradual uncovering of the three great pillars of our research, which are the strength of Computer Science, the beauty of Biology and the wisdom of Bioinformatics.

I also thank my colleagues from the UQAM bioinformatics lab, for providing me with such a dynamic and vibrant research environment.

I would like to express my gratitude to all professors and administrative staff at UQAM who served as good examples, helping me to forge a personalized set of technological skills, all along my studies.

I am grateful to the *Natural Sciences and Engineering Research Council of Canada* (NSERC) and the *Fonds Québécois de la Recherche sur la Nature et les Technologies* (FQRNT) for their state of the art financial support.

I also benefited from the moral support of my wife, especially on the difficult moments, when dealing with the inherent uncertainties of research.

A special thought goes to my parents who cultivated the great seed of Science in me. May they dwell in my memories!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ADENO	Adenocarcinome
ADN	Acide DésoxyriboNucléique
ARN	Acide RiboNucléique
GenBank	NIH genetic sequence database
HGT	Horizontal Gene Transfer
HIV	Human Immunodeficiency Virus
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
NJ	Neighbor-Joining
NNI	Nearest Neighbour Interchange
PARS	Parsimony Program
PHYLIP	PHYLogeny Inference Package
SPR	Subtree Pruning and Regrafting
SQUAM	Carcinome aux cellules squameuses
TBR	Tree Bisection and Reconnection
VIH	Virus de l'Imunodéficience Humaine
VPH	Virus du Papillome Humain

RÉSUMÉ

La biologie évolutive est régie par des forces écologiques correspondant à des échelles géographiques et temporelles différentes. L'interrelation hôte-pathogène constitue une des principales forces évolutives, menant à la croissance de la variabilité génétique. Dans cette thèse, nous présentons d'abord un nouveau modèle permettant de retrouver des régions génomiques fonctionnelles en se basant sur la variabilité des séquences ainsi que sur une analyse de regroupement d'espèces faite selon des critères booléens de pathogénicité. Les méthodes et les fonctions de regroupement qui en découlent ont été appliquées à des jeux de données réelles impliquant la carcinogénicité et l'invasivité des espèces. Ces méthodes et fonctions doivent varier dépendamment de la combinaison des mécanismes évolutionnaires (sélection positive et lignée spécifique) de même que des types de regroupement variés (monophylétique et polyphylétique). Nous utilisons l'index de *Rand* ajusté pour valider les résultats. Par la suite, nous étudions sur une plus grande échelle le phénomène du transfert horizontal de gènes, complet et partiel, chez les procaryotes. Cette analyse détaillée est effectuée sur plusieurs niveaux taxonomiques, génétiques et écologiques pour permettre d'estimer statistiquement l'ampleur de l'acquisition du matériel génétique tout au long de l'histoire évolutive des procaryotes. Finalement, nous décrivons une nouvelle méthode rapide de détection des transferts horizontaux de gènes complets qui est basée sur des fonctions de regroupement, existantes et nouvelles, accompagnée de la procédure de validation utilisant les *p-values*.

Mots clés : analyse de regroupement; arbre phylogénétique; bipartition; carcinogénicité; détection de régions fonctionnelles; invasivité; recombinaison; transfert horizontal de gènes; variabilité génétique.

ABSTRACT

Evolutionary biology is driven by different ecological forces, acting on the geographical and temporal scales. Host-pathogen interaction is one such major evolutionary force, leading to higher genetic variability. In this thesis, we first present a new model allowing for recovering functional genomic regions responsible for a given disease. The new model relies on sequence variability cluster analysis and Boolean pathogenicity criteria. The proposed clustering functions and methods have been applied to real datasets characterized by carcinogenicity and invasivity of certain species. The considered clustering functions vary according to the involved evolutionary mechanisms (positive selection or lineage specific selection) and phylogenetic relationships between species (monophyletic or polyphyletic). Our results were validated by using the adjusted Rand index. Then, we carried out a comprehensive study to measure the impact of horizontal gene transfer on the evolution of prokaryotes. Complete and partial forms of horizontal gene transfer were studied. This detailed analysis was performed on taxonomic, genetic and ecological levels in order to assess statistically the rate of horizontal acquisition of genetic material along the evolutionary history of prokaryotic species. Moreover, in the final chapter, we introduced a new fast method for detecting complete horizontal gene transfer events. The proposed method is based on the above-mentioned clustering functions and accompanied by a validation procedure using *p-values*.

Keywords : bipartition; carcinogenicity; cluster analysis; functional region detection; genetic variability; horizontal gene transfer; invasivity; phylogenetic tree; recombination

INTRODUCTION

Evolutionary biology is best explained by ecological forces acting over different geographic and temporal timescales. Biotic factors, such as competition and predation, shape ecosystems locally and over short time spans, as assumed by the *Red Queen* hypothesis. On the contrary, abiotic factors, explained by the *Court Jester* hypothesis, such as climate, oceanographic and tectonic events, shape larger-scale patterns regionally and globally over millions of years.

According to this view, *host-parasite* relationships stand as the main evolutionary force on the microscale, increasing genetic variability. Many parasites, usually prokaryotes and viruses, have the advantage of shorter generation time. Hosts, usually eukaryotes, developed sexual reproduction, which by the means of recombination speeds up evolution, while also developing an immune system capable of generating hypervariable genetic regions. Sequence conservation, a measure of negative selection, has been used extensively to detect functional regions, but other forces are responsible of driving change in this host-parasite setting.

In this thesis, we present models and algorithms, using variability clustering to detect the forces, active on the parasite side, such as positive selection, lineage specific selection and horizontal gene transfer (HGT), followed by recombination. In respect to one such change driving force, namely HGT, we cluster prokaryotes in phylogenetic and ecological groups, quantifying its presence and extent at gene and subgene detailing level, and also time its distribution at the genomic scale.

Chapter I discusses the basic notions and models used in bioinformatics, which are necessary for understanding the application context of our algorithms.

Chapter II presents the state of the art of functional sequence detection. It describes efforts made to uncover genomic regions under evolutionary forces. We can classify these forces according to the degree of change they imply. Most studied of all is *negative selection*, which is based on conservation measures, and uncovers fundamental structures needed by a majority of organisms in order to function independently. On the contrary, species need variation among individuals in order

to escape pathogenic attack, as has been described previously. We focus on *positive selection, site specific and lineage specific selection* methods at first. An even higher degree of change is brought by HGT and *recombination*. Finally, we describe *genetic association studies* and efforts made to use this natural clustering around virulence factors defined as a Boolean criterion, such as carcinogenicity or invasivity of microorganisms.

Chapter III describes a novel prediction method for discovering genomic regions associated with a disease. This method relies on transfers between groups to optimize bipartitions in order to maximize various variability metrics. Using simulations, we showed relations and limits of our detection method for each proposed metric, under a combination of evolutionary mechanisms (positive selection or lineage specific selection) and clustering types (monophyletic or polyphyletic). We then used *Adjusted Rand Index* to validate the obtained results.

Chapter IV presents a large study of the extent of HGT in the prokaryotic world. Here, we consider clustering around phylogenetic and ecological groups. Applying the efficient and highly accurate HGT-Detection algorithm, which is backed by a bootstrap-based statistical validation, we first quantified the global extent of this phenomenon at the gene level (complete transfers) and then detailed this extent at the nucleotide level, accounting for subgenetic regions (partial transfers). Interrelations between groups and important transfer statistics were inferred and discussed in detail. The existence of phylogenetic and ecological HGT-related clusters was also revealed. Finally, dating of HGT events was inferred and presented for complete and partial horizontal gene transfers.

Chapter V describes a new fast method intended to detect complete horizontal gene transfers. We show that this method is able to recover a majority of low-confidence and almost all high-confidence transfers found by the well known HGT-Detection algorithm, at the price of a higher, but still acceptable, false positive rate. The presented method is based on new aggregation functions, similar to those presented in chapter III, but using different clustering criteria. *P-values* were estimated in the proposed algorithm using a stochastic Monte Carlo procedure.

We finally, present a synthesis of our work and some perspective ways for improving our methods.

CHAPTER I

BASIC GENETIC AND EVOLUTIONARY NOTIONS AND MODELS OF MICROORGANISMAL EVOLUTION

Bioinformatics is a vast and multidisciplinary field. According to the National Health Institute (NH) of the United States of America (Huerta et al. 2000) it includes: "Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data." Bioinformatics is useful in many life sciences, but its primary applications concern genetic data, first of all DNA and RNA sequences, stored in large public databases, such as *GenBank* (Benson et al. 2009) and *Entrez Gene* (Maglott et al. 2005). Analysis of such data is done by a combination of computer science, biology, statistics and mathematics. In this chapter, we present basic biological and evolutionary notions and definitions, as well as elementary models used in computer science, in order to represent biological data and species evolution. They are all essential to the understanding of this thesis. We also present the basic information of the microorganisms studied by other chapters including *Neisseria Meningitidis* and *Human Papilloma Virus.*

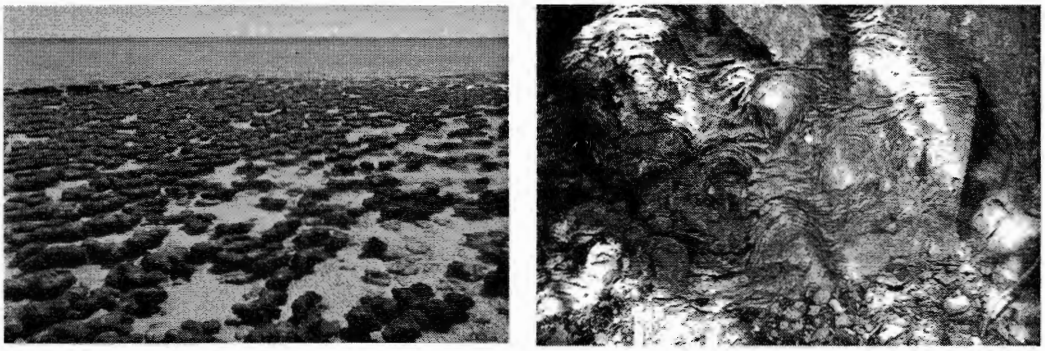
1.1 Microorganisms

Microorganism is consists of organisms that are observable only under microscope. Antonie Van Leeuwenhoek (1632–1723) was amongst the first to build such a microscope. Another contemporary of him, Robert Hooke, at 1665 wrote a book describing his observations. He also introduced the term "cell", as the first organisms seen with such microscope were indeed single

celled. With the development of the electron microscope, the internal structure of these microorganisms has become visible. The existence of a distinct nucleus separated eukaryotes (organisms with nucleus) from prokaryotes (organisms without nucleus). Present day classification of all living organisms is based on the work pioneered by Carl Woese and others, who used genetic material and cell membrane structure. There are three main lineages, called domains, Bacteria, Archaea and Eukarya. First two of them are prokaryotic lineages. Eukarya includes all eukaryotes, from the kingdoms Animalia, Plantae, Fungi and Protista. Bacteria, archaea and a set of Eukarya (almost all the protozoa, some fungi, algae, and animals) represent the set of microorganisms. Our algorithms were developed and tested on Prokaryotes and Viruses. For the scope of this thesis we will limit our description to the latter two types of organisms. Viruses are organism-dependent entities that are generally not considered as micro-organisms. For practical purposes, they are studied in Virology, a subfield of microbiology.

1.1.1 Prokaryotes

Prokaryotes are ubiquitous organisms, living in all environments, including the most extreme, like boiling springs, permanently frozen or extremely salty waters, at the depths of the ocean, or environments without oxygen or radioactively contaminated. Prokaryotes also normally reside in the human digestive system and skin. They are sometimes responsible for several kind of illnesses, but also serve an important role in the preparation of many foods, like yoghourt, vinegar or chocolate. Prokaryotes are probably the first inhabitants of Earth, able to withstand harsh conditions of very high temperatures, volcanic eruption, mutagenic radiation from the sun and no oxygen conditions. There is evidence of a fossilized microbial mat, in Australian sandstone, estimated to date over 3 billion years (OpenStax College 2014). These sedimentary rocks are called stromatolites (Figure 1.1).



(a) (b)

Figure 1.1 Stromatolites

(a) These living stromatolites are located in Shark Bay, Australia.

(b) These fossilized stromatolites, found in Glacier National Park, Montana, are nearly 1.5 billion years old. (credit a: Robert Young; credit b: P. Carrara, NPS), OpenStax College, Biology. OpenStax CNX. 16 Apr 2014 http://cnx.org/contents/185cbf87-c72e-48f5-b51e-f14f21b5eabd@9.43.

Only a small percentage of all prokaryotes are pathogens. Their role in ecological processes is very important. Life would not be possible without them. Human life is also dependent on symbiosis with microbes, which help us digest food, produce nutrients, protect from pathogenic microbes and train our immune system.

1.1.2 Viruses

Viruses constitute another important group of microorganisms. There are non-cellular, lacking most of the components of cells, such as organelles, ribosomes, and the plasma membrane. Single virus particles are called virions. Each virion consists of a nucleic acid core, an outer protein capsid, and sometimes an outer envelope. The envelope is not of its own structures, but rather made of proteins and phospholipids derived from the host cell membranes. Sometimes viruses contain enzymes. It is important to notice that the complexities of the viruses with their hosts are not related. They are obligate intracellular parasites, incapable of replicating themselves outside of a host. This property, but also the maintaining of activity after having been crystallized (Stanley et al. 1935), makes their consideration as living organisms debatable. They are only visible in the electron microscope, about 20–250 nanometers in diameter, with some exceptions for some large virions of the poxvirus family. The tobacco mosaic virus was the first to be discovered (Figure 1.2). Host and parasite

complexity usually are not related. Bacteriophages are amongst the most complex virion structures, but they infect only bacteria (OpenStax College 2013).

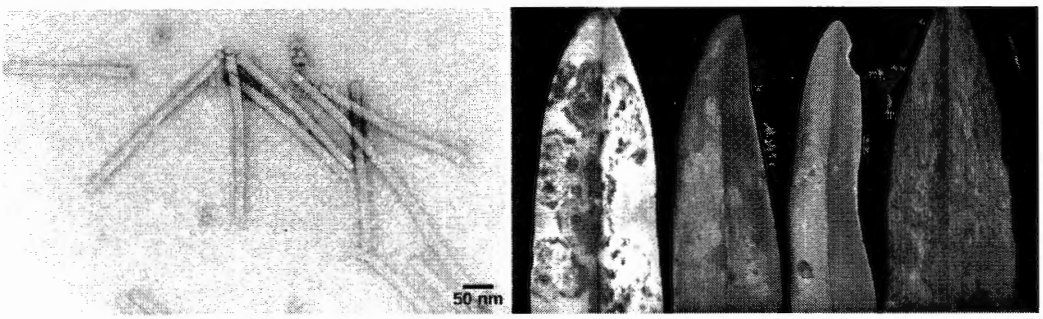


Figure 1.2 Tobacco mosaic virus

Transmission electron microscopy (left).

Sample of orchids affected by disease caused by viruses (right).

(credit a: USDA ARS; credit b: modification of work by USDA Forest Service, Department of Plant Pathology Archive North Carolina State University; scale-bar data from Matt Russell).

OpenStax College, Introduction. OpenStax CNX. 5 Mar 2013 http://cnx.org/contents/ed0fb5c2-ce30-4a76-8d58-77fb7cf9c7ec@2.

1.2 Elementary genetic notions

1.2.1 DNA

Deoxyribonucleic acid (DNA) is a high weight macromolecule, a polymer of smaller weight nucleotides (Saenger 1984). It stores information necessary to normal functioning and development of the whole organism. DNA is the basic element that constitutes genes corresponding to the support of heredity. It is linear, unbranched polymer in which monomeric subunits are four chemically distinct nucleotides that can be linked together. Each nucleotide is constituted of three elements: a phosphate group and a monosaccharide (deoxyribose), which belongs to the backbone and a nitrogenous base which gives its name to the corresponding nucleotide. There are four types of nucleotides composing DNA (Lodish et al. 2000). Adenine (A) and guanine (G) are double-ringed purines, while cytosine (C) and thymine (T) are single-ringed pyrimidines. DNA's form is a complementary double helix, as shown in Figure 1.3 (Watson and Crick 1953). DNA is present in the nucleus of eukaryotic cells, in prokaryotic cytoplasm, mitochondrial matrix and also in chloroplasts. There are also some viruses containing DNA, placed inside a protein protecting structure which is called the capsid.

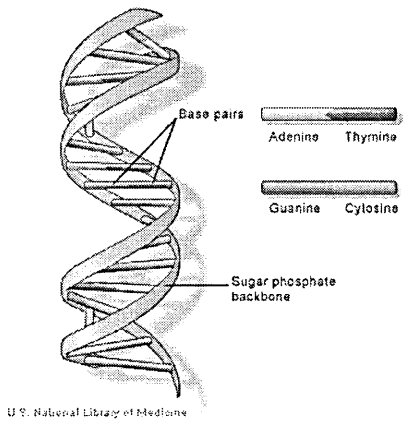


Figure 1.3 DNA strands complementarity
Possible interactions are: A-T and T-A, G-C and C-G

(credit: (National Library of Medicine (US). Genetics Home Reference [Internet]. Bethesda (MD): The Library; 2013).

1.2.2 Central dogma of molecular biology

General protocol of information flow in biological systems is as follows: DNA is copied (**replication**) to DNA, which is transformed (**transcription**) into messenger RNA (mRNA), which

serves as a model for protein synthesis (**translation**) (Crick et al. 1970). Translation involves the use of codons, triplets of nucleotides, each associated with one amino-acid in the proteic primary chain (Figure 1.4). There are some exceptions to this general rule, known to date as reverse transcription and replication of RNA, which will be described subsequently.

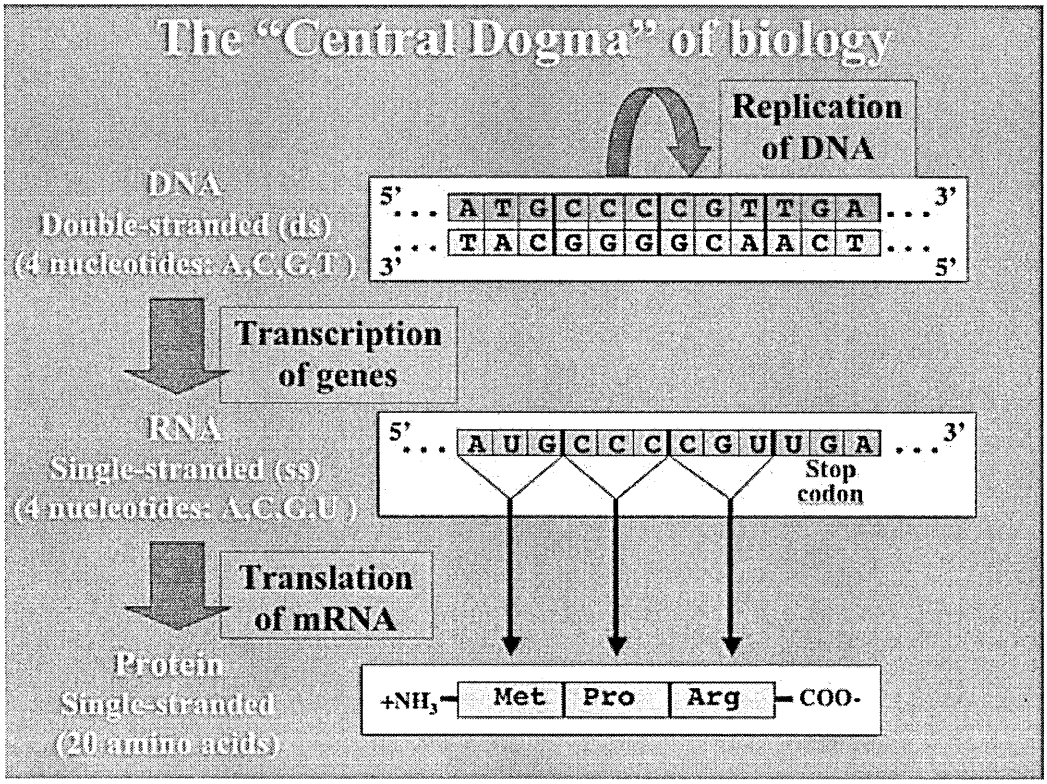


Figure 1.4 Central dogma of molecular biology.
(Bruce Fouke, 2006).

1.2.3 Gene

Gene is a DNA segment containing biological information that corresponds to either coding sequences (which gives a protein) or a non-coding sequences corresponding to non-coding RNA (see 1.2.4). It is located usually on a chromosome and it is the functional unit of inheritance (Johannsen 1911). Gene is mainly composed of two parts: the exons that contain the DNA that will bé transformed into proteins and introns that are regions containing regulatory elements and untralslated DNA. For a given coding gene, the coding sequence is a region of the gene coding for a

protein. The coding region of a gene, also known as the coding sequence or CDS (from Coding DNA Sequence), is the portion of the gene's DNA or RNA, composed of exons, that codes for a protein.

1.2.4 RNA

Ribonucleic acid (RNA) is a macromolecule, a polymer of nucleotides, similar to DNA. There are however some differences. RNA is single stranded, and ribose here replaces desoxyribose, while Uracil (U) replaces thymine (T). There exist several families of RNA. These are grouped according to their function or secondary (or tertiary) structure.

Table 1.1 Main types and functions of RNA

Name	Acronym	Function
messenger RNA	mRNA	Represents the template for protein assembly.
transfer RNA	tRNA	Transports an amino-acid corresponding to a specific codon.
ribosomal RNA	rRNA	Constitutes the ribosome after maturation and association to proteins.
micro RNA	miRNA	Blocks translation of certain mRNA by ribosomes. They can regulate gene expression.

RNA can have a secondary structure showing base pairing interactions. It results in alpha-helix structures and beta-sheets (Doty et al. 1959). RNA could either encode for protein (coding RNA) or not (Non coding RNA). Non-coding RNA principally regulates gene expression (Birney et al. 2007). Non-coding RNA genes include highly abundant and functionally important RNAs such as transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), microRNAs, the long non coding RNAs and several other classes.

1.2.5 Codon

A codon is a triplet of mRNA nucleotides A, C, U or G. It will be transcribed into one of the 20 natural amino-acids. Certain codons are synonyms, several of them coding for the same amino-acid (Figure 1.5).

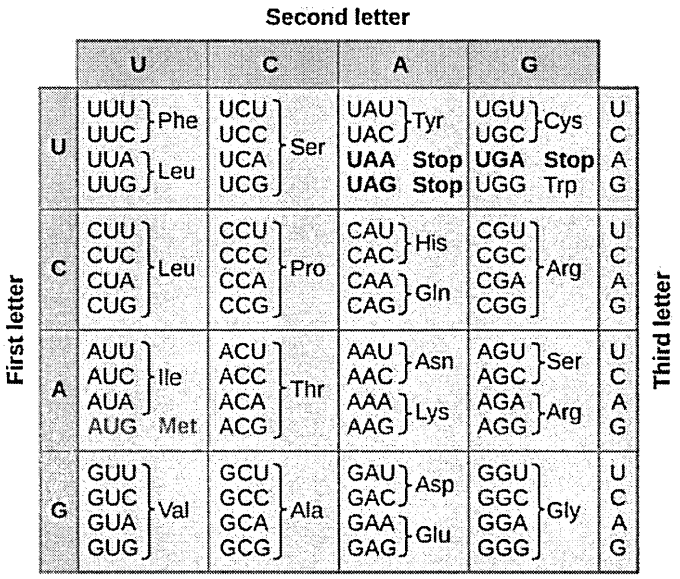


Figure 1.5 Genetic code

Genetic code is used for translating nucleotide triplets, found in mRNA, into amino acids or a termination signal in a nascent protein (credit: modification of work by NIH). OpenStax College, The Genetic Code. OpenStax CNX. 24 Feb 2014 http://cnx.org/contents/40489b84-9322-47be-96dc-4f80079cb868@7.

Three of the 64 codons are called stop codons. They terminate protein synthesis. Another codon, AUG, in addition to specifying the amino acid methionine, also serves as the start codon. The reading frame for translation is set by the AUG start codon. The genetic code is almost universal. Purified mRNA from one species can be used by another species for protein synthesis. This serves of evidence of common origin of life on Earth. Even viruses share the same genetic code. The genetic code is also degenerate, or redundant, which makes it fault tolerant. Codons specifying same amino acids typically differ only by a single nucleotide. Also, similar codons encode chemically similar side chains.

1.2.6 Protein

A protein is a macromolecule composed of one or many amino acid chains, bounded by peptidic bonds (Branden and Tooze 1996). When their molecular weight is under 10kDa, they are called

peptides (Oliva et al. 2004). Proteins are at the foundation of cellular functions. They are responsible for the catalysis of chemical reactions, transport, communication, signaling and signal recognition. Numerous proteins also have a structural role, for instance, those belonging to the viral capsid (Lodish et al. 2000). Following translation, amino acid order constitutes protein's primary structure. Then, the molecule folds into itself, with help from hydrogen bonds, to form secondary structures, the most important being alpha-helix and beta-sheets. Different secondary structures arrange themselves into tertiary structures, governed by hydrophobic interactions and disulfide bonds (Figure 1.6). There is even a quaternary structure with the association of multiple peptidic units (Lodish et al. 2000).

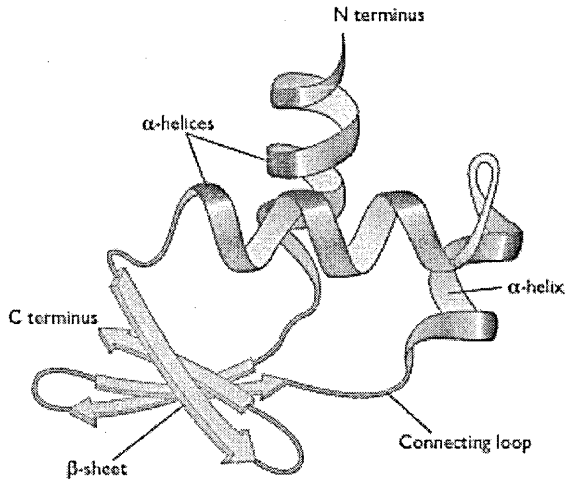


Figure 1.6 Example of a protein structure
Hypothetical protein comprised of 3 α-helices and 4 β-sheets.
Reproduced from (Turner et al. 1997), redesigned in (Brown 2006).

1.2.7 Chromosome

In a genome, the DNA is mainly packaged into a complex macromolecule with several genes, regulatory elements and non coding DNA. This structure called chromosome, codes most of the genetic information. Specific proteins help package and control its functions. Some prokaryotes also store DNA in plasmids. Eukaryotic cells have large linear chromosomes and prokaryotic cells have smaller circular ones.

1.3 Human papilloma virus (HPV)

HPV is a DNA virus with a dimension of 8kpb . Its genome is composed of eight genes, coding for the same number of proteins, and one regulatory region. Genes are designated by letter E for early and L for late, according to their epithelial differentiation. E1, E2, E5, E6, and E7 are expressed early in the differentiation processes, E4 is expressed all along, while L1 and L2 are expressed during final stages (Figure 1.7). Early proteins are expressed at low levels that could explain long latency. L1 is a major capsid protein; L2 serves as intermediary with plasmidic DNA (Doorbar 2006, Schiffman et al. 2007). E1 and E2 are regulatory proteins that modulate transcription and replication, while E5, E6, and E7 modulate transformation. The role of E4 is not completely elucidated; several studies indicate the possibility of facilitating genome replication and activation of late functions (Wilson et al. 2007) as well as virus assembly (Prétet et al. 2007).

First discovered HPV was classified together with polyomaviruses of the Papovaviridae family, due to their similar non-enveloped capsid characteristic and analogous double-strained DNA genome. The unique common element between these two families is, in fact, a proteic domain of gene E1 (De Villiers et al. 2004). It codes for a helicase, very similar to simian 40 (SV40) T antigen in a polyomavirus, to the NSI protein of parvoviruses and even an extra-chromosomal element in a flat worm - Girardia tigrina (Rebrikov et al. 2002). T antigen of SV40 is ligating the tumoral suppressor p53 and inhibit its transcription (Dobbelstein and Roth 1998). Over 200 papilloma virus genotypes exist, and more than 100 have been classified (Büchen-Osmond 2006).

Traditionally, based on tissue tropism, classification of HPV was made in three groups – cutaneous, muquous and mixed (Segondy 2008). Contrary to many viruses the modern HPV classification is not based on morphological criteria but rather on genetic similarity (De Villiers et al. 2004). A classification given by genomic similarity, pathogenicity and potential to determine cancer, divide papillomavirus in Genera such as Alpha, Beta and Gamma-papillomavirus (De Villiers et al. 2004). HPV is responsible for frequently sexually transmitted diseases. Certain strains infect genital mucosa, some other infect skin. Most known clinical manifestation is *Condyloma Acuminata.*

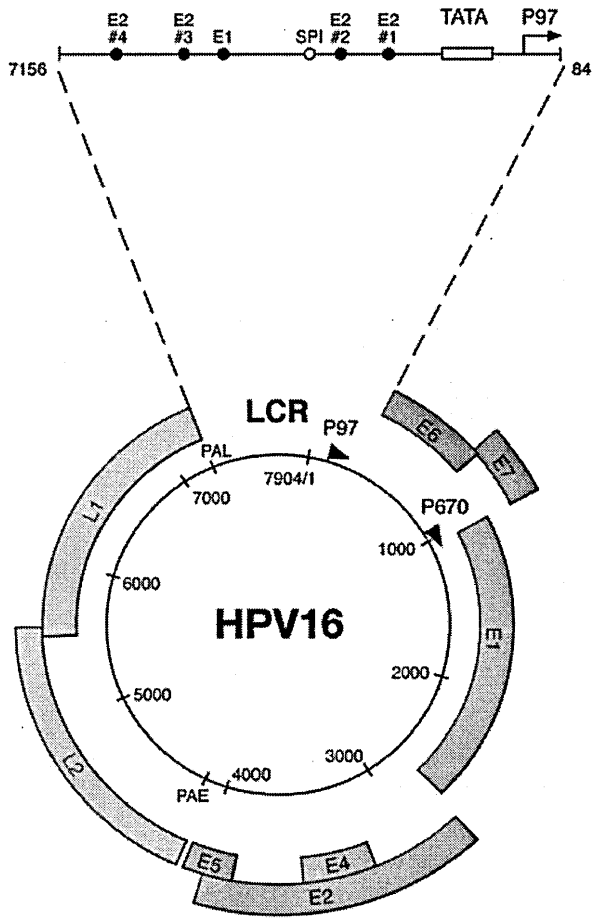


Figure 1.7 HPV16 genome structure

The genome has length 7904 bp. It is represented as black circle with early promoters (p97) and late (p670) which are depicted by black arrows. Early ORFs (E1, E2, E4 and E5) E6 and E7, are expressed starting at p97 or p670 at different stages of cellular epithelial differentiation. ORFs L1 and L2 are also expressed starting at p670, following changes in splicing models and polyadenilation sites. Viral genes are encoded on the same strand. Long control region (LCR) spanning from 7156 to 7184 is in large, for visualization of binding site E2 and TATA box of promoter p97. Binding sites E1 and SP1are also shown. Reproduced from (Doorbar).

1.3.1 Carcinogenicity of HPV

Some strains of HPV are involved in cervical cancer (Schiffman et al. 2007). Non-carcinogenic strains or the absence of infection by HPV do not correlate, or negatively correlate with initial modifications, seen in cervical cancer (Castle et al. 2007). However, most of HPV are not carcinogenic, especially strains causing common and plantar warts. Over 40 genotypes infect mycosis and among them 13 to 18 types belong to the high-risk category. This category is considered a precondition to cervical cancer development. It is involved in genesis of part of ano-genital and aero-digestive cancers as well.

Even low risk strains are still responsible for high morbidity and are source of genital warts (Trottier and Franco 2006). A study involving 11 countries and 15 613 women aged 15 to 74 years showed a variable prevalence, ranging from 1-4% in Spain to 20 times higher - 25-26% in Nigeria (Clifford et al. 2005). This form of cancer is the second most frequent in female populations and seventh amongst all. This is a globally public health concern with an estimated 493,000 new cases and 274,000 deaths for year 2002, everywhere in the world.

1.4 *Neisseria Meningitidis*

Meningococcus is a Gram-negative bacterium, known for its role in the development of meningitis in humans. It has aerial transmission, by inhalation. Because of its invasivity the contact with patients infected with disease heighten the risk of transmission by 500 to 2000 times (Peltola 1983). The website «neisseria.org» is one of the best resources serving research community for centralizing public available information about *N.meningitidis*. Until now, 4 invasive strains and 3 asymptomatic ones have been sequenced. Very subtle differences between them could be responsible for their virulence.

Genome of *N.meningitidis* is made up of one circular chromosome, with a medium size of 2.2 Mpb and a G+C content of around 51%. There are at average 1971 CDS, with an average 885 bp (Schoen et al. 2009).

Comparative genomics studies showed a broad range of mechanisms that support genomic flexibility. *N.meningitidis* would be a paradigm for organisms using variability to adapt to a hostile and changing environment (Schoen et al. 2009). Its genome abounds of 20% repetitive DNA, being ranked among the most repetitive in a study conducted by (Achaz et al. 2002). In another study of bacterial families using gene order as a measure of stability, *N. meningitidis* is ranked among the

less stable genomes (Rocha 2006). Intra-genomic recombination is the main mechanism used to generate phenotypic diversity (Schoen et al. 2007, Schoen et al. 2009).

Many horizontal gene transfers, originating in the same or related *N.meningitidis* species, have been identified (Maiden et al. 1996). Their biology is complex, being comprised of minimal mobile elements (Saunders and Snyder 2002), DNA islands horizontally transferred (Tettelin et al. 2000), canonic genomic islands (Hotopp et al. 2006) and defective phages (Schoen et al. 2009). For instance, the only factor proved to be associated to any one pathological type of *N.meningitidis* is the polysacharidic capsule, which has been obtained by horizontal transfer (Elias et al. 2006).

1.5 Molecular evolution

The major results of the Darwinian Theory is that species evolve through changes occurring over time (Darwin, 1859). The independent changes in the genomic patrimonies of living species lead to the rise of new organisms. Evolution happens because of processes that affect individual organisms, as primary source of change, and their fate at the population level. Evolution includes *genetic variability* (i.e. genetic code modifications appearing at an individual level), and *changing allele frequency* (i.e. frequency of different versions of same corresponding sequence) in the population during time (i.e. impact of individual modifications over the entire population) (Duret 2008).

1.5.1 Evolutionary hypotheses

Two important hypotheses explain at least partially the need for change, namely the *Red Queen* and *Court Jester* hypotheses.

1.5.1.1 Red Queen

It states that survival is equilibrium between co-evolving opposing organisms in a continuously changing environment. Under host-parasite coevolutionary relationships (Penn 2001), parasites have the advantage of shorter generations, while hosts developed sex, which uses recombination, in order to achieve greater genetic variation. Many microorganisms studied in this thesis are parasites, expressing coevolutionary relationships, among them being Human Papilloma Virus (HPV) (Lace et al. 2009, Schwarz and Leo 2008, Tindle 2002) and *Neisseria Meningitidis* (Jolley et al. 2005).

The name of the concept comes from Lewis Carroll's novel, *Through the Looking-Glass*, in which Red Queen states: "Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!" (Pearson

2001). Competition, predation, and other biotic factors seem to explain ecosystems on the short temporal time scale (Benton 2009).

1.5.1.2 Court Jester

Court Jester model, explains the evolution of ecosystems on a global scales as adaptation to geological events, like climate, landscape, or food supply changes. This term is used in opposition to the *Red Queen* philosophical concept, by using a Tarot card, *the Fool,* or *the Joker*, as a suggestion of the non-correlation between individual efforts and global results (Barnosky 2001, Benton 2009).

1.5.2 Genetic variability

Genetic variability consists of the difference between individuals genetic patrimonies due to several types of mechanisms such as *mutations, recombination and horizontal gene transfer (HGT)*. These evolutionary mechanisms can also lead to translocations, duplications, insertions and deletions that modify long genomic regions, generally chromosome-wide.

1.5.2.1 Mutations

Mutation constitutes a mechanism in which DNA is altered to give different sequence during evolution. Genome variability is principally due to different types of mutations. They represent the key force of evolution, as they create permanent change to the genetic material. They are originating as errors of cellular division, particularly DNA replication, but could also arise as a result of radiation, chemical substances or viruses. Sometimes, mutations are generated by controlled mechanisms during the course of cellular reproductive lines division (i.e. meiosis), or the hypermutation needed for antibodies production. According to the produced effect, mutations can be disfavoring (e.g. interruption of an important cellular function), favoring or neutral. Neutral ones do not modify organism's survival or reproductive capacity, in its particular environment and, therefore, can accumulate over time. Mutations affect DNA composition of a short region of a genome. Point mutations replace one nucleotide with another, and are called substitutions, while insertions or deletions can affect several nucleotides.

1.5.2.2 Recombination

Contrary to mutations, recombination is carried out and regulated by enzymes and other proteins. It can be homologous or heterologous, in respect to the relative position, value, or structure of DNA. This concept is flexible enough to include alleles at the molecular level or species at the conceptual one. In each case, genetic material is exchanged between one or more parts of the corresponding sequences.

1.5.2.3 Horizontal Gene Transfer - HGT

The principle way of acquiring genetical patrimonies is through parent-child inheritance. However, genetical material could also be transferred within distant organisms. Often, microorganisms transfer DNA between individuals that results in strains with beneficial uptakes from more than one parent. This is sometimes achieved by transformation (first mode of HGT), when release of DNA to the environment is followed by its uptake and recombination. Homologous recombination is limited to similar organisms, but recently "homology-facilitated illegitimate recombination" (HFIR) is being able to extend into areas with little similarity (De Vries et al. 2004, Meier and Wackernagel 2003). Plasmids and conjugation can spread genetic material even beyond species barrier (Figure 1.8). Heterologous recombination, also known as "illegitimate recombination", is one of the sources of horizontal gene transfers (Vetsigian and Goldenfeld 2005). Integrative conjugative elements usually use "site-specific recombination". Transduction, consisting of DNA transfer by phage, is yet another mode of HGT. It is able to use "non-homologous recombination" between short repeats of length 5 to 12bp. Barriers to these mechanisms exist but are limited and thus cannot prevent gene acquisition in most cases (Thomas and Nielsen 2005). HGT can pose several risks to humans including antibiotic-resistant genes spreading to pathogenic bacteria, transgenic DNA insertion into human cells and possible cancer triggering, as well as disease-associated genes spreading and recombining to create new viruses and bacteria (Boc et al. 2010, David and Alm 2011).

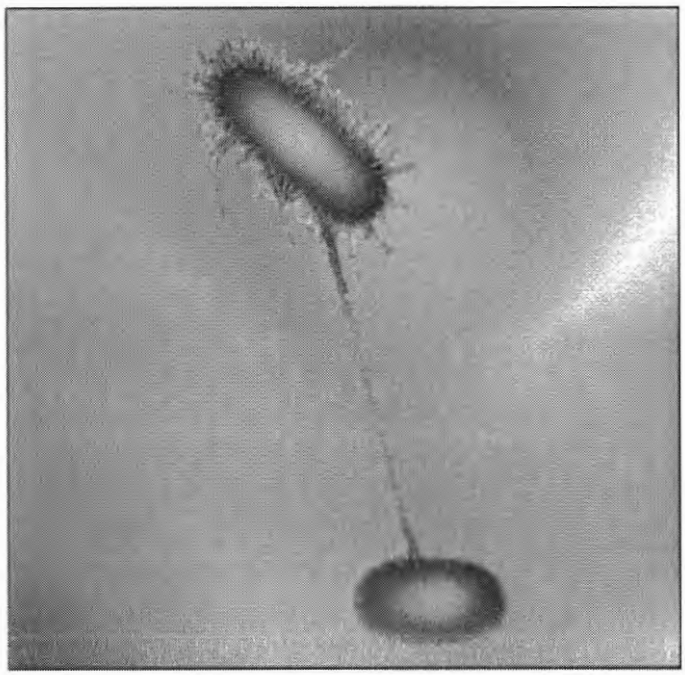


Figure 1.8 Bacterial transfer using a conjugation plasmid
Reproduced from Systematic Biology (59)-2, March 2010. Photo credit AJC1 Flickr.

1.5.2.4 Mosaic genes and intragenic recombination

Following HGT, intragenic recombination occurs, leading to mosaic gene formation (Hollingshead et al. 2000). This phenomenon is significant, but sometimes misinterpreted (Zhaxybayeva et al. 2004). Bacteria and Archaea are adapting to changing environments by creation of mosaic genes. This term comes from alternating blocs of sequences that despite having different histories are combined in an allele following a recombination event (see Figure 1.9). These recombined segments can come from similar strains or even from very distant species (Gogarten et al. 2002, Hollingshead et al. 2000). A mosaic gene is composed of segments identical to the original allele, and others derived from recently integrated DNA. When entrant DNA is very different from the host DNA, mosaic genes can express new phenotypes. There are biological proofs, for constant generation of mosaic genes in transformable populations, and probably in all genes. Multilocus sequence typing (MLST) is useful to evaluate the extent of recombination in a bacterial population (Maiden et al. 1998). Non-transformable bacteria exhibit mosaic genes too, but at a lower rate (e.g. some *Neisseria* species). Mosaic alleles have been reported for many genes, including those

encoding surface antigens, IgA protease, and antibiotic targets (Hollingshead et al. 2000, Maiden et al. 1998). A good mosaic gene example resulting from HGT between two different species is the Penicillin resistance in *Streptococcus pneumoniae,* mediated by penicillin-binding proteins (PBPs) (Claverys et al. 2000).

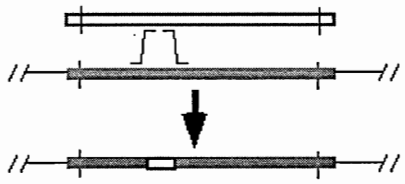


Figure 1.9 Diagram depicting formation of a mosaic gene
White sub-sequence of the new mosaic gene originating from another strain and the blue one from the original strain (credit: Stanley Maloy, 2002).

1.5.3 Natural selection

Natural selection is a result of differential mortality and fertility. As such, it is responsible of the fate of such mutations that modify the *adaptive value* of organisms. Alleles conferring more value have the tendency to heighten their proportion, until fixation occurs, by means of positive selection (i.e. their proportion goes up to 100%, when they become fixed, by completely eliminating all other alleles). On the contrary, alleles that reduce adaptive value are under negative selection, also known as *purifying selection.* Between these extremes there is *balancing selection*, when a group of selective population processes, actively maintaining frequencies of a pool of genes, above that of gene mutation. In a host-parasite relationship, for eukaryotic hosts, this selection occurs in the immune system, where the *Major Histocompatibility Complex* (MHC) loci are known to be highly polymorphic (Hughes and Nei 1988). Also it occurs when some alleles give advantages to a heterozygosis state (i.e. only part of alleles is different for the same genetic locus). They are maintained in equal proportions in their population by means of balancing selection (Duret 2008). In a predator-prey relationship, *frequency-dependent* selection is yet another form of balancing selection (Endler and Greenwood 1988). Mutations that do not affect adaptive value are not affected by natural selection but they are left to genetic drift (Duret 2008).

According to selectionist theory, natural selection is a primordial force of evolution, the influence of non-adaptive processes being reduced to minor contributions. It explains differences between species as the effect of positive selection, as a consequence of adaptation to environment, and polymorphism as the work of balancing selection (Duret 2008). Motoo Kimura proposed a different view of evolution, molecular neutral evolution (Kimura et al. 1968), sustained by King and Jukes, who introduced their own non-Darwinian type of evolution (King et al. 1969). The latter theory affirms that the majority of molecular changes are caused by *random fixation* of mutants (due to genetic drift of populations of finite sizes) under neutral adaptive value, and continuous flow of mutations. It also affirms that polymorphism of DNA and proteins, forming variability inside the same species boundaries are selectively neutral, and are maintained in the species by balance between mutational entries and random extinction (Kimura 1985). Natural selection may favor HGT, as a more rapid way of adaptation than the accumulation of numerous point mutations, leading to alteration of gene functions. Prokaryotes have sophisticated mechanisms for the acquisition of new genes via HGT, which is considered rampant among various groups of genes in bacteria (Boc et al. 2010).

1.6 Phylogenetic tree as support of evolution

We described above genetic and evolutionary notions, as they are understood today. The idea of evolution was first applied to biology, and models have been developed, long before molecular biology arose as a standard basis for classification. The most popular representation of evolutionary history is that of a phylogenetic tree.

1.6.1 Phylogenetic trees

1.6.1.1 Definition

Phylogenetic trees are acyclic and connected graphs in which contemporary species are associated with tree leaves.

Four main components of a phylogenetic tree are as follows (Figure 1.10):

- *Root* indicating common ancestry of species or strains represented;
- *External nodes – leaves* – representing contemporary species, which are also called taxa;
- *Internal nodes*, representing putative inferred ancestors;

- *Branches*, showing the ancestry relations between nodes. They can have length (representing mutation rate, genomic distance, etc...).

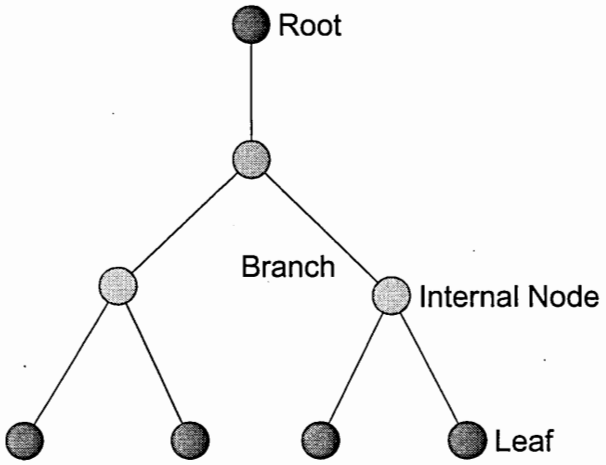


Figure 1.10 Example of a phylogenetic tree

1.6.1.2 Characteristics

A node's degree is defined as the number of branches adjacent to it. Nodes having degree higher than three are called *unresolved*, otherwise they are *resolved*. An unrooted phylogenetic tree - having n leaves and all internal nodes resolved - is composed of:

- $2n$-2 nodes (n-2 internal nodes and n leaves) and
- $2n$-3 branches.

Whenever the common ancestor of all species is determined, the tree is rooted. It is oriented following species evolution. A rooted tree allows for defining an ancestry relation among two successive nodes. It is impossible to objectively identify the origin (i.e. root) of species diversification based on the analyzed species alone. Usually, the root is inferred either using the midpoint, or using an outgroup. The outgroup technique consist of including in the study a species known having distant relationships to all of the present taxa. The obtained bifurcation between such a distant taxon and all other taxa will define the tree root. The midpoint technique consists of putting the root at the middle of the two distantly related taxa.

1.6.2 Evolutionnary biology and the introduction of phylogeny

Charles Darwin initiated this scientific discipline in his seminal work «On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life», which was published in 1859. He introduced a theory of population evolution through natural selection. He illustrated life diversity and presented arguments for branching evolution, based on common ancestry. His work reflected the observations of his famous voyage around the world on board of the ship *Beagle* in 1830, but also his own subsequent work and experimentations. The only figure of his book is specifically a phylogenetic tree used to classify species. It is reproduced in Figure 1.12.

Ernst Haeckel also relied on phylogenetic trees to describe species evolution, see Figure 1.13. He represented the organismal evolution using an almost linear progressive model in contrast to Darwin's widely branching one.

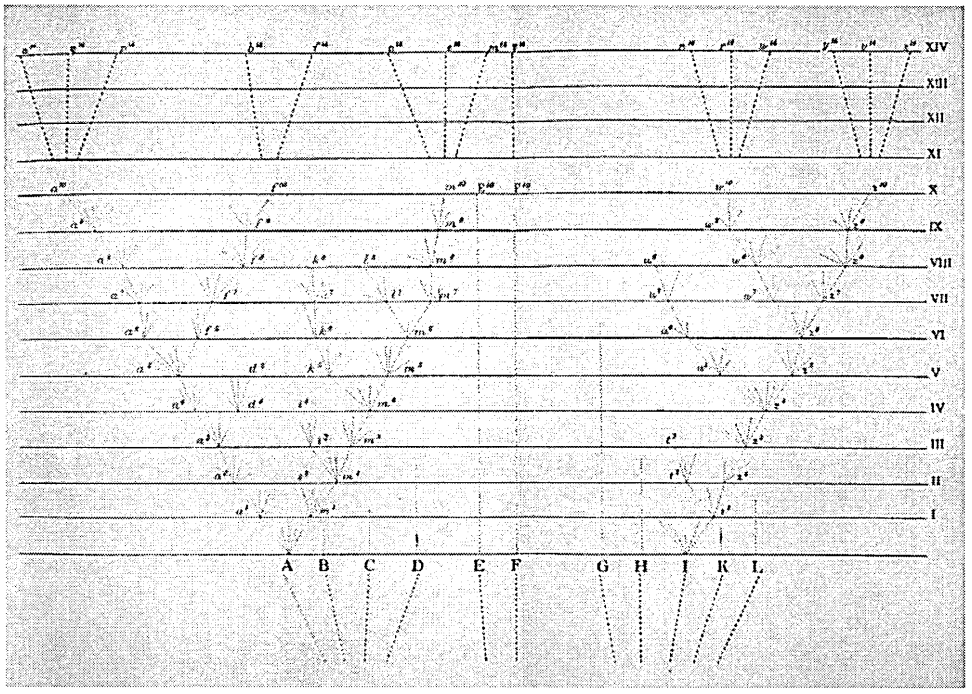


Figure 1.11 A species tree
Tree of life image from Darwin's "On the Origin of Species by Natural Selection" (Darwin 1859).

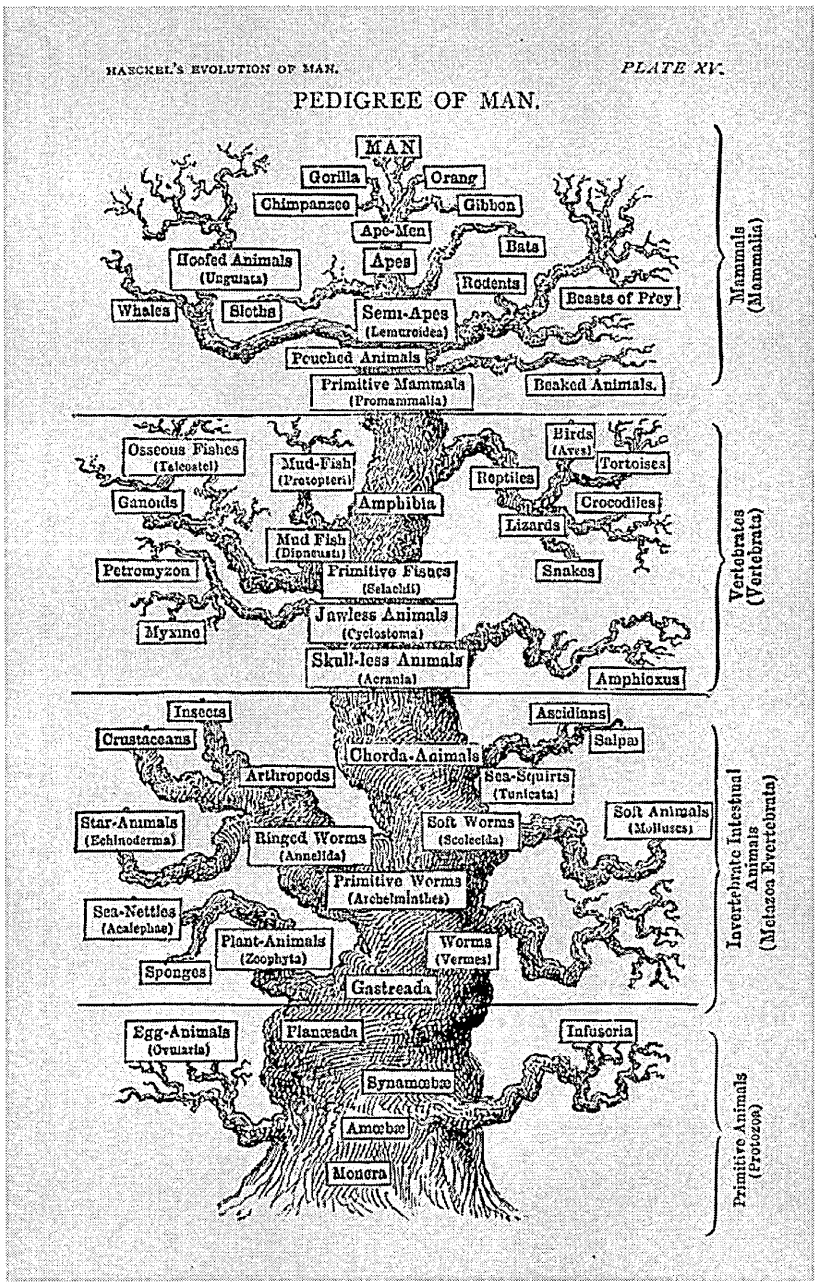


Figure 1.12 A species tree (including human genealogy) (Haeckel 1879).

1.6.3 Lineage in a phylogenetic tree

A lineage in a phylogenetic tree consists of a path including the given species and all its ancestors up to the tree root. When phylogenies are represented using trees, this becomes a polyline, joining all ancestors of the existing organism, up to its root (Figure 1.14).

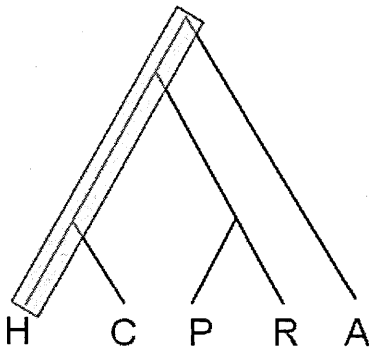


Figure 1.13 Lineage representation inside a phylogenetic tree
This figure depicts a phylogenetic tree, each node standing for a species. Letters represent existent species. For the species H, the gray rectangle represents its lineage.

1.6.4 Tree of life

Tree of Life is an old and complex metaphor to indicate the evolutionary history of all living species. It has been the subject of many reviews (Mindell 2013). Charles Darwin presented the first scientific phylogenetic tree (Darwin 1859). Then, Ernst Haeckel built one universal tree for all species and groups known at that time (Haeckel 1879). For a long time classification was carried out based on observable traits such as anatomical, physiological or, later, biochemical features. Woese first produced a classification of living organisms based on 16S ribosomal RNA (Woese and Fox 1977) prior to discovering the Archaea kingdom (Woese et al. 1978). He then proposed a widely-accepted phylogeny of the three domains of life including Eukarya, Bacteria and Archaea (Woese et al. 1990). Today, tendency is more towards phylogenomics, using concatenated genes or genomes (Pierce 2007). This type of information is useful for long range reconstructions, spanning geological ages. The phylogenetic reconstruction is still a complex and debatable subject (Lecointre and Le Guyader 2006). One of the debated subjects is tree rooting (Becerra et al. 2007). To avoid the influences of HGT or recombination events, ubiquitary genes, core genes, or alternatively 16S RNA are used to build phylogenies. Several collaborative projects to reconstruct the tree exist. One

of the most famous is called *Tree of Life Web Project* (ToL). The latter is the main international project intended to infer a phylogeny of all currently leaving species (Maddison et al. 2007).

1.7 Approaches for phylogenetic reconstruction

1.7.1 Phylogenetic classification – Cladistics

This classification is based on observable characteristics, pertaining to species, as a testimony of ancient history (Lecointre and Le Guyader 2001). It is based on evolutionary proximity relations between species and is, therefore, tied to the modern vision of evolution. Its schematic representation is a cladogram, which is a phylogenetic, unrooted tree, containing nodes and leaves (i.e. species or taxa). Groups including a common ancestor and all its descendents are called monophyletic. They represent clades (see figure 1.15) (Hennig 1975). A taxon is a classification entity, grouping together organisms having in common certain well-defined characteristics (Wiley et al. 1991). Also, subtle differences between mono-, holo-, and paraphyletic groups make this subject a very debatable one (Envall 2008).

Sauropsids group is constituted of reptiles and birds. It is considered monophyletic because all its descendents are present inside this group. Without birds we obtain reptiles. This forms a paraphyletic group. Mammals and birds form the "warm blood animals" group. This one is polyphyletic because its members have different ancestors (see Figure 1.1.5).

Modern taxonomy is under control of several International Committees (e.g. International Committee on Taxonomy of Viruses - ICTV, International Commission on Zoological Nomenclature - ICZN). A fully phylogenetic classification, called PhyloCode (Cantino and De Queiroz 2004), failed to become a gold standard in the field.

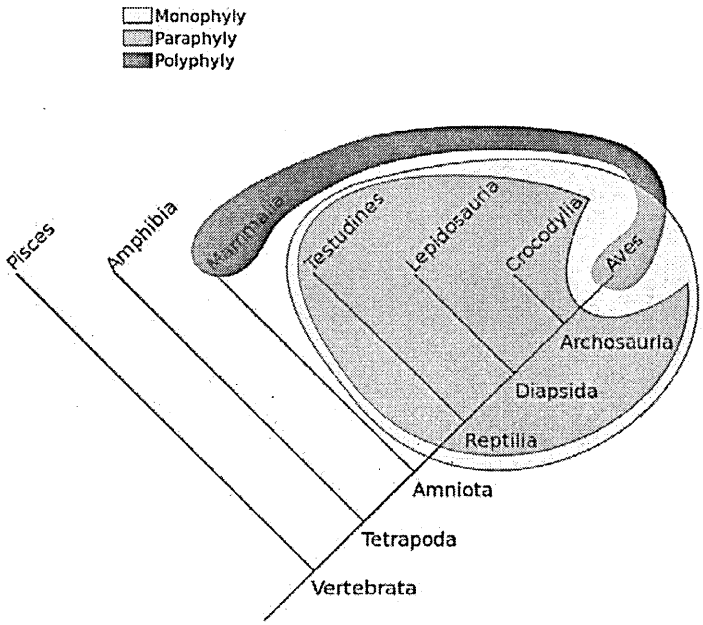


Figure 1.14 Differences between monophily, paraphily and polyphily
Image credit TotoBaggins.

1.7.2 Cladistic phylogenetic tree reconstruction

Cladistic reconstruction is based on an evolutionary model, inferring an optimal tree, which is evaluated at each tree node for optimality. Computational difficulties often arise when a huge number of trees have to be evaluated. The three main cladistic approaches are *maximum likelihood*, *Bayesian reconstruction* and *maximum parsimony* (Felsenstein 1981). The maximum likelihood approach is the most widely used nowadays. Homology is defined as common resemblance between taxa that can be attributed to common ascendance. Many characters are homologous. Modern tree reconstructions are almost always based on homologous DNA, RNA or protein sequences (Bear and Rintoul 2014). Therefore, methods determining which sequences are homologous strongly influence the quality of phylogenetic inference (e.g. quality of multiple sequence alignments).

1.7.2.1 Phylogenetic tree inference using maximum parsimony

Maximum parsimony is a principle known as Occam's razor, or in Latin as *lex parsimoniae*. It was devised and used for problem-solving by William of Ockham (1287–1347) and states that we should always select the hypotheses involving the fewest number of assumptions. Stephen Hawking writes in *A Brief History of Time*: "*We could still imagine that there is a set of laws that determines events completely for some supernatural being, who could observe the present state of the universe without disturbing it. However, such models of the universe are not of much interest to us mortals. It seems better to employ the principle known as Occam's razor and cut out all the features of the theory that cannot be observed*" (Hawking and Jackson 1993). This idea has been used in phylogenetic reconstruction to recover the tree of the smallest total number of mutations (i.e. the most parsimonious tree in terms of the number of mutations) (Edwards and Sforza 1963). Fitch introduced the most known parsimony algorithm (Fitch 1971). Some comparative reviews show that nonparametric methods, including maximum parsimony, can be resistance to biases in real datasets (Kolaczkowski and Thornton 2004).

1.7.2.2 Maximum likelihood principle

Given a variable sample, maximum likelihood is a general statistical method allowing for the inference of the parameters for a probability distribution, which maximize the probability of the sample. It was first employed in bioinformatics by Edwards and Cavalli-Sforza (1963) in a study of gene frequencies. First application on molecular sequences was that by statistician Jerzy Neyman (1971). Given a family of parameterized density functions, where θ is the parameter and x is the experiment's result:

$x \mapsto f(x \mid \theta)$, the likelihood function is: $L(\theta \mid x) = f(x \mid \theta)$, where $f(x \mid \theta)$ is a function of probability density, x is the variable, θ is the model's parameter and $L(\theta \mid x)$ is the likelihood function.

1.7.2.3 Phylogenetic tree inference using maximum likelihood

Application to phylogenetic trees requires an evolutionary model allowing for the computation of transition probabilities depending on the evolutionary time. Several evolutionary models for genomic sequences have been developed. They regularly differ on the way of fixing parameters such type of mutations, rate of mutations and frequency of nucleotide. The models JC (Jukes and Cantor 1969), HKY (Hasegawa et al. 1985) and GTR (Tavaré 1986) are the most widely used

models of DNA evolution. They offer a good compromise between precision and computability. We suppose that evolution is independent for different sites and lineages. In order to find the most likely tree, nucleotides of all sequences are separately compared, final probability being the product of individual probabilities. Theoretically, we have continuous variables, but practically, for computational reasons, several discrete classes of rate variations are used. If the evolutionary model is reversible, we obtain an unrooted tree.

There are three main steps in a maximum likelihood inference procedure:

1st step: Generate tree topologies to be tested

Normally, we should already have a set of tree topologies to be tested. We can start with a Neighbor-Joining tree topology (Saitou and Nei 1987), and then let a *branch-and-bound* or *greedy* search procedure recover more optimal topologies. Several tree transformation methods have been developed to search for optimal tree topologies:

- Nearest Neighbor Interchange (NNI);
- Subtree Pruning and Regrafting (SPR);
- Tree Bisection and reconnection (TBR).

A maximization-optimization algorithm that guarantees non-diminishing scores has been proposed (Friedman et al. 2002). *Simulated annealing* is another heuristic used frequently.

2nd step: Branch length optimization

The most widely used method is Newton-Raphson numerical optimization, included in RAxML (Stamatakis et al. 2005) and PhyML (Guindon and Gascuel 2003). We will use both of these methods in our experiments with real data described in Chapter IV.

3rd step: Calculate the total maximum likelihood of the given tree

ML programs use the Felsenstein pruning algorithm (Felsenstein 1981) for calculating the Phylogenetic Likelihood Function (PLF). It assumes data interdependence and updates already calculated values, inferring ancestral states at each tree node.

Many computer programs using ML approach exist, the most known of them being: DNAML (DNA maximum likelihood program) from Felsenstein (1981), PHYML (Guindon and Gascuel 2003) and RAxML (Stamatakis et al. 2005). The last algorithm has efficiently vectorized and parallelized implementations.

1.7.3 Phenetic tree reconstruction: the distance methods

The phenetic category of tree reconstruction methods estimates first evolutionary distances between each pair of species and then infers the phylogenetic tree that fits best these distances. We obtain the estimate of the evolutionary distance between two species by summing up branch lengths of the unique path relating these species in the inferred phylogenetic tree. The obtained distance is a tree metric (Barthélemy and Guénoche 1991). When the differences between the evolutionary distances and the obtained tree distances are small, the correct tree is usually inferred (Kim and Warnow 1999). The main advantage of the distance methods is their low algorithmic complexity. This makes them useful for the analysis of large datasets. One of the first proposed distance-based algorithms is UPGMA (Unweighted Pair-Group Method using arithmetic Averages) (Sneath et al. 1973). The most popular distance-based algorithm is certainly Neighbor-Joining (NJ) (Saitou and Nei 1987). Its output tree often constitutes the starting tree for more advanced tree reconstruction methods.

1.8 Multiple sequence alignment - MSA

Alignment is an operation leading to the identification of homologous elements. Homology is the concept of character present in extant species to share a common ancestry. Hence alignment problem can be defined as finding the alignment necessary for all sequence comparisons (see Figure 1.16). Diverse MSA building heuristics have been proposed. Most of them are based on Hidden Markov Models (HMM) (e.g., see the *hmmalign* program from the HMMER package (Eddy 1998, Finn et al. 2011) or simulated annealing (Kim et al. 1994). One of the popular approaches is progressive alignment. It usually starts by aligning the most similar sequences, then by using a guiding phylogenetic tree. It treats all the sequences one by one. The most known of these progressive methods is *ClustalW* (Thompson et al. 1994). However the *T-Coffee* approach is slower than ClustalW but yields better results for more distant relations (Notredame et al. 2000). *DIALIGN* is based on local alignments (Brudno et al. 2003) and *MUSCLE* is a more precise MSA

implementation (Edgar 2004). Another fast MSA method is *MAFFT* which is based on a fast Fourier transform technique (Katoh et al. 2002).

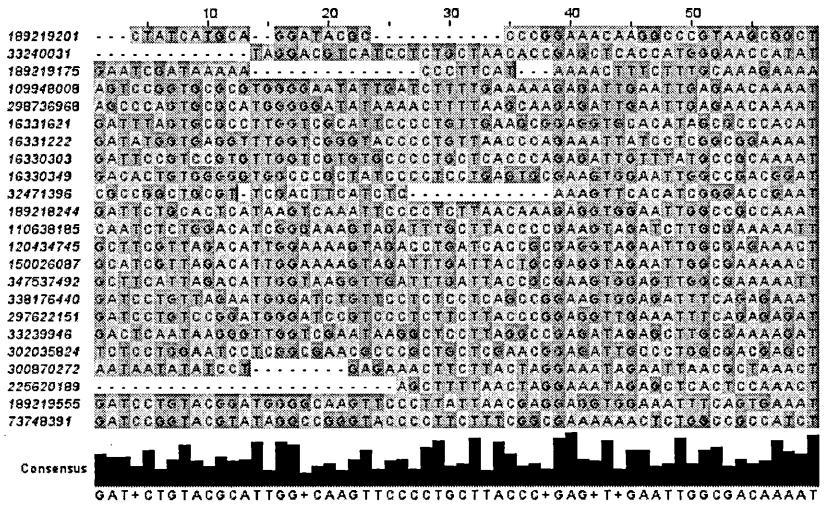


Figure 1.15 An example of a multiple sequence alignment – MSA

MSA displayed in Jalview (Clamp et al. 2004). Lines represent sequences. Columns represent homologous nucleotides. Gaps are represented by «-» characters and stand for inferred indels (insertions or deletions of nucleotides).

1.9 Reticulated evolution and networks

Phylogenetic trees are appropriate models for Darwinian evolution but they lack support for phenomena such as horizontal gene transfer (see Chapter II), hybridization or genetic recombination (Sonea 2000). Reticulated networks, instead, can represent relations where an individual inherits genetic material from multiple ancestors (Legendre and Makarenkov 2002).

In the reticulogram reconstruction, for example, we start with a phylogenetic tree and then add reticulations (supplementary branches) to the supporting tree structure (Legendre and Makarenkov 2002). The T-Rex server (Boc et al. 2012), (Makarenkov 2001) is one of the most comprehensive web servers allowing for inferring phylogenetic trees and networks (see figure 1.17).

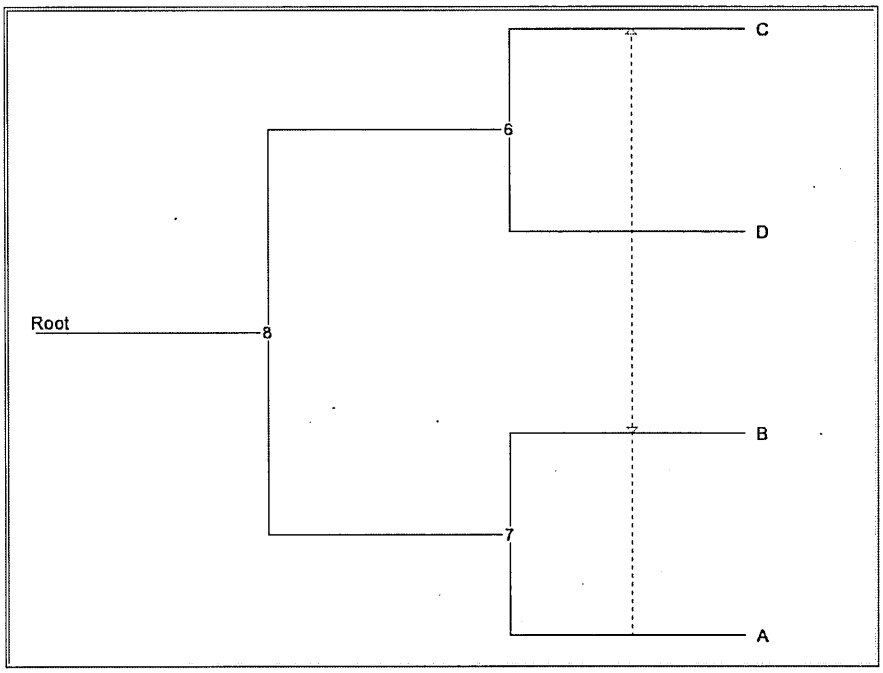


Figure 1.16 A phylogenetic network modelling a scenario of horizontal gene transfers inferred using the T-Rex web site

(Boc et al. 2012, Makarenkov 2001).

1.10 Methods for detection of Recombination

Despite more than 20 years of research and a high number of detection methods available, recombination analysis is still considered imperfect (Maydt and Lengauer 2006). The problem consisting of determining if sequences belonging to a multiple sequence alignment include elements originating in recombination is hard. Locating recombination breakpoints is even harder (Posada and Crandall 2001, Wiuf et al. 2001). The problem complexity is producing an impressive number of approaches, each standing for a different aspect of recombination. Such studies, evaluating performance, conclude that we should choose different methods, based on "a priori" knowledge of our data, especially depending on divergence rate. Scoring methods are faster and have higher sensitivity, but phylogenetic ones are more precise and do not generate excessive numbers of false positives. Some methods are described in greater detail in a review chapter (Husmeier and Wright 2005).

First proposed methods for detecting recombination were based on statistical tests verifying non-uniformity of substitution distribution, such as the χ^2 test. They were not based on an explicit evolutionary model, but usually yielded good results (Posada and Crandall 2001). One of the most widely-used methods remains *GENECONV* (Sawyer 1989). It searches for the longest conserved fragment between two sequences and determines whether it is significant. Extensions of this method allow for including mutations in the fragments. Some methods can detect signal differences between adjacent regions of a multiple sequence alignment, including PLATO (Grassly and Holmes 1997), TOPAL (McGuire et al. 1997), (McGuire and Wright 2000), PhyPro and SimPlot. Finally, there are methods based on coalescence (Brown et al. 2001), minimizing cost of substitution and topology change following tree-like history. *RecPars* is probably the most known of them (Hein 1993). It defines optimality in terms of parsimony and is based on a recombination versus nucleotide substitution cost ratio.

A more accurate statistical framework, including Bayesian Hidden Markov Models and Markov Chain Monte Carlo (Husmeier and McGuire 2003) was proposed, but the considered tree inference at each sequence position implied super-exponential complexities. The *Recco* method (Maydt and Lengauer 2006) is generally comparable with older methods in terms of results. It is able to improve detection in certain scenarios, while suffering from some limitations (e.g. mutual masking of similar recombinant sequences). It uses cost optimization and dynamic programming. Another method based on sliding window procedure, comparable to *RecPars* and inspired by *DSS* (McGuire and

Wright 2000) and *PDM* (probabilistic divergence measure) (Husmeier and Wright 2001), uses a clustering pruning scheme. It needs not an a priori recombination to substitution ratio, but an estimated maximum number of recombination events, in order to pre-establish the number of clusters (Husmeier et al. 2005). The latter authors used it to detect interspecific recombination. For all fixed size sliding window methods, there is a compromise between the power to detect the recombination signal and the method's time complexity.

A study of mitochondrial DNA shows widespread evidence of recombination. An aggregate score of the overall evidence has been proposed using a total of nine local and global scoring methods based on *p*-values. All local methods were present in *RDP2* (Tsaousis et al. 2005). The algorithms such as *GARD* (Pond et al. 2006) and *3SEQ* (Boni et al. 2007) have been used in studies on homologous recombination in the avian viruses (He et al. 2009) and (Boni et al. 2008), with mixed results.

1.11 Methods for detecting Horizontal Gene Transfer

There are two main approaches to detect horizontally transferred genes. First, sequence analysis of the host genome may suggest fragments with different GC content or codon atypical usage patterns (Lawrence and Ochman 1997). Finding sequences not likely to arise from a selective process means that they might have been acquired horizontally. An original method to detect such sequences has been proposed by (Tsirigos and Rigoutsos 2005). The main limitation of this method is the need of codon boundaries knowledge. The second approach is based on phylogeny reconciliation between the given *species tree*, or molecular tree based on a molecule that is assumed not to undergo HGT (e.g. 16S rRNA or 23S rRNA), and the given *gene tree* defined for the same set of organisms. Ribosomal genes can be also affected by HGT, but at a seemingly lower rate. Thus, a ribosome tree can serve as a better approximation to a species phylogeny in the absence of more reliable data (Acinas et al. 2004). The main limitation of this approach is that its accuracy is strongly dependent on the reliability of the gene phylogeny (i.e. bootstrap support of the gene tree branches).

Several proposed methods model tree reconciliation by minimizing the subtree prune and regraft (SPR) distance. Computing the SPR distance for rooted binary trees was shown to be NP-hard (Bordewich and Semple 2005) as well as for unrooted trees (Hickey et al. 2008). An exact algorithm, called *LatTrans,* computing all shortest SPR scenarios is available. However it is exponential on the number of transfers (Hallett and Lagergren 2001).

Several distance methods have been developed to detect HGT. They rely on heuristics running in polynomial time. One of the most known is *RIATA-HGT* based on the divide-and-conquer approach (Nakhleh et al. 2005). The latest version of *RIATA-HGT* is considerably faster than *LatTrans* providing the almost equally accurate (Than and Nakhleh 2008). An even faster and more accurate algorithm is *HGT-Detection* (Boc et al. 2010). It uses an improved distance measure called *bipartition dissimilarity*. It is implemented as a package running on the T-Rex web site (Boc et al. 2012). Our study described in chapter IV relies on the results provided by *HGT-Detection.* Another well-known HGT detection method is Efficient Evaluation of Edit Paths (EEEP) that uses tree comparisons and evolutionarily reasonable constraints (Beiko and Hamilton 2006). It achieves faster speeds than *LatTrans* but is less accurate. A probabilistic model has also been developed, but applied only to gene family size problems (Csűrös and Miklós 2006). A combinatorial model incorporating HGT and duplication events has been proposed as well (Hallett et al. 2004). It consists of the improvement of *LatTrans* algorithm (Hallett and Lagergren 2001). Hallett, Lagergren and Tofigh presented the proof of the NP-completeness of the problem, and gave tractable and polynomial solutions when cycles are disregarded and restrictive parameterization is performed. Unfortunately, their algorithms are not publicly available.

An interesting development in handling time constraints was the introduction of dated species trees (Merkle and Middendorf 2005) in the context of host-parasite coevolution. Its implementation in CoRe-PA (Merkle et al. 2010) is based on a dynamic programming parameter-adaptive approach. This approach helps keep polynomial time complexities for all algorithms of this category. Other methods and implementations using the same idea are *AnGST* (David and Alm 2011), *Mowgli* (Doyon, Scornavacca, Gorbunov, Szöllősi,) and *Jane* (Conow et al. 2010, Libeskind-Hadas and Charleston 2009). A general comparison of HGT detection methods is available (Doyon, Ranwez, Daubin and Berry 2011). It presents some important discrepancies between theoretical results of *AnGST* and *CoRe-Pa*, expected from methods descriptions and their implementations. New optimization algorithms that treat distance-dependent transfer costs are implemented in the *RANGER-DTL* package (Bansal et al. 2012). Comparisons with other algorithms are shown only for the time cost, but not for the quality of individual transfer scenarios.

A precision improvement of parsimony methods has been achieved by including elements of detection based on population genetics and the coalescent model. This allows for modeling incomplete lineage sorting (ILS) phenomenon. A new software package, called *Notung,* has been

applied to phylogenetic datasets in which ILS, HGT and hybridization may be present (Stolzer et al. 2012). Issues with reliability of inferred evolutionary events over multiple transfer scenarios have been outlined as well as the need of defining support values. As a majority-rule consensus of 50% support cannot be guaranteed, a median reconciliation has been proposed (Nguyen et al. 2013). When several HGTs occur between two given species, then these species are considered as being linked by a highway of gene sharing. A polynomial time algorithm, using parsimony principle and quartet trees has been designed (Bansal et al. 2011) and, later, a software package called *HiDe* (Highway Detection) has been developed (Bansal et al. 2013). HGT modeling has also been used recently to show that a set of extant species carry information about extinct lineages, and specifically about the size and dynamics of ancient biodiversity (Szöllösi et al. 2013). Another interesting development is the application of *partial gene transfer models* to the problem of mosaic genes detection (Zhaxybayeva et al. 2004) and the development of efficient algorithms and implementations for inferring partial HGTs (Boc and Makarenkov 2011). The latter method brings statistical bootstrap confidence to the problem of detecting genetic regions first horizontally transferred and then affected by intragenic recombination (Boc and Makarenkov 2011, Makarenkov et al. 2006).

CHAPTER II

METHODS FOR DETECTION OF FUNCTIONAL SEQUENCES AND RETICULATED EVOLUTIONARY EVENTS

The algorithms we will present in this thesis are applied to detect functional sequences (i.e. those parts of the genome whose existence, composition or structure is related to a known molecular function). There are many functional levels. The gene is certainly the most important one; it has been studied for many years. Translation into proteins is considered as a proof of biological function (Brown 2006). Later, the role of non-coding RNA has been emphasized and regulatory regions became the focus of detection (Macdonald and Long 2005). Interactions between RNA, DNA and proteins have been included into the functional domain, and have become the scope of bioinformatics development.

Although functional characterization remains an ultimate molecular biology task, the search for such candidates remains an important bioinformatics challenge (Huerta et al. 2000), (Wooley et al. 2005). One way to proceed is to find regions that evolve at different speed or follow different sequence patterns than natural random evolution (Vitti et al. 2013). Many statistical models are in place for different aspects of evolution, and computational methods have been developed to evaluate statistical scores, or likelihood related to them (Wooley et al. 2005). Most studied approaches are related to *sequence conservation* as a mean to detect negative selection (Siepel et al. 2005). Many tree-like evolution models have being studied in terms of positive selection (Yang 2007) and lineage specific selection (Hubisz et al. 2011), as well as reticulate evolution models (Dagan et al. 2008), which include HGT and recombination.

Functional sequence detection "*in silico*" has also a cost saving benefit, as genome-wide analyses in molecular "wet lab" are usually very costly. Computational methods usually offer a limited number of high probability candidates, which are further analyzed by molecular biologists. There are also following aspects of scope and filtering of the obtained results:

- Identification of all functional regions;
- Identification of regions responsible for disease or specific pathogenic characteristics, such as invasivity or carcinogenicity.

2.1 Negative (purifying) selection

Methods belonging to this category have the purpose of finding conserved genomic regions, evolving under negative selective pressure. This is the case of essential regions for cell functioning. Once lost, these functions lead to cellular death, or its incapacity to reproduce. *BLAST* (Altschul et al. 1990) is one of the first and still widely used tools to rapidly, locally align and identify sequence similarity between a relatively short query sequence and a large sequence database. The improvements of the reference implementation at NCBI include better alignment statistics (Schäffer et al. 2001), usability (Ye et al. 2006), indexing (Morgulis et al. 2008), specialized searches (Johnson et al. 2008) and programmability (Camacho et al. 2009). The alternative implementation at EBI has been also improved recently (Flicek et al. 2014). Cross-species comparisons require alignment of many sequences, and thus global alignment algorithms are needed. There is evidence that not only coding, but also regulatory sequences, can be also conserved (Pennacchio and Rubin 2001).

More advanced methods to detect "Multi-species Conserved Sequences" have been developed (Margulies et al. 2003). Two basic methods are available to validate functional sequences, "Conserved RNA Secondary Structures" obtained with QRNA program (Rivas and Eddy 2001) and "Transcription Factor–Binding Sites" obtained with TRANSFAC (Matys et al. 2003).

A very popular algorithm to detect conservation patterns is *PhastCons*, which is based on a two-state phylogenetic hidden Markov model (phylo-HMM). *PhastCons* fits a phylo-HMM to the data by maximum likelihood, adding constraints designed to calibrate across species groups (Siepel et al. 2005). Implementation and programmability techniques have recently been improved (Hubisz et al. 2011). Another algorithm is *Sequence CONservation Evaluation* (*SCONE*). It uses the Bayes or

maximum-likelihood estimates of the evolutionary rate. It also defines a probability (*p*-value) of neutrality for each site in a MSA. Its applications have been limited to the mammalian genomes and, specifically, to the human genome (Asthana et al. 2007).

2.2 Positive selection

Methods belonging to this category detect regions under positive selection, which bring new functions to the cell, giving an advantage of survival or reproduction. The widely used criterion is to highlight abnormal ratios of d_N/d_S (i.e. between non-synonymous and synonymous codon sites) (Hubisz et al. 2011, Yang et al. 2000). Non-synonymous sites induce changes in proteins, and are thus proven functional. Synonymous sites are expected to be distributed along neutral selection. There are some measures of neutrality such as *z-test* – based on normal distribution, or the LRT (*likelihood ratio test*), or based on χ^2 distribution.

These methods need gene annotations to operate in order to establish scope and reading frame. They also lack power when positive selection affects a reduced number of sites or when long branches are saturated with mutations. Also, the most important limitation is that they are not able to detect non-coding control regions. One of the most popular software packages for detecting positive selection is PAML (Yang 2007). It includes various models of codon substitution, evaluated using maximum likelihood. The site-wise log-likelihood score is yet another good predictor of genes under positive selection (Wang et al. 2013).

Selectome is a database of positive selection, based on variation in selective pressure *dN/dS* ratio) over branches and over sites (Moretti et al. 2014). Limits of detection power of such *branch-site tests* of positive selection have been investigated, showing robustness but lack of power under synonymous substitution saturation and high GC content variation (Gharib and Robinson-Rechavi 2013). *Bayesian* estimates of this ratio appear to have better statistical properties than the ML estimates. A new computationally efficient method has become available and may be useful for genome-scale comparisons of protein-coding gene sequences (Angelis et al. 2014). The implementation optimizations, such as detection of frame-shift mutations and premature stop-codons, have become of a significant interest for genome-wide applications (Zhang et al. 2013).

2.3 Site specific, lineage specific or signature selection methods

Contrast between conserved and non-conserved regions is not always very sharp. DNA conservation among diverged species is able to successfully identify noncoding regulatory regions. At the same time, rapidly evolving regulatory regions will not generally be conserved across species and fall out of purely conservation-based methods detection resolution. Finer grain methods have been developed, able to discriminate inside lineages, or gene families. They generally use previously available models and methods to detect negative or positive selection, but combine them with phylogenetic analysis and statistical tests in an automated or semi-automated manner.

Automated algorithms are usually time consuming, but are precise. The most commonly known algorithms are *phyloP,* which is based on phylogenetic *p*-values, and *DLESS,* which is based on a phylo-HMM model (Siepel et al. 2006). While *DLESS* determines itself the clade, it is able to find only "gain" or "lost" events. On the contrary, *phyloP* is able to cope with "accelerated" events but needs the phylogenetic subtree of interest be provided. It implements four statistical phylogenetic tests: a likelihood ratio test, a score test, a test based on exact distributions of numbers of substitutions, and the genomic evolutionary rate profiling (GERP) test (Pollard et al. 2010). The method *DivE* addresses the latter limitations and allows for the detection of "accelerated" events in non-coding regions (Pertea et al. 2011).

A genetic algorithm has also been developed, but it has a high computational cost and is largely tied to prior model assumptions (Pond and Frost 2005). Positive selection detection that allows for lineage-specific rate variation has been also integrated into the maximum likelihood framework (Guindon et al. 2004).

A method involving "signatures of nonneutral evolution", i.e. an examination of the pattern of polymorphism both within and between populations as well as divergence with sibling species, detects several nonneutrally evolving regions not identified by conservation (Macdonald and Long 2005). A pattern referred to ''selective signature'' of a gene is defined by its evolutionary speed and is associated with gene function and ecology inside specific phylogenetic groups. Such methods, able to study signatures by combining previously described methods, have been recently designed (Shapiro and Alm 2008). Signature methods can recover site-specific selective pressure using machine learning classifiers, including Naïve Bayes, *k*-nearest neighbors and support vector

machine (SVM). They are usually able to outperform common measures of sequence conservation (Sadri et al. 2011).

2.4 Scope of this thesis

Detection of functional sequences is a major goal for biology as a whole and bioinfomatics in particular. It can be achieved on multiple levels, from proteins to genes and genomes. Historically, the uncovering of functional patterns has switched from direct observation of highly conserved proteins to DNA and non-coding RNAs, for an ever more abstract and systemic model based discovery. The advent of genomics era has brought an increasing size for studied datasets, rapidly extending their scope. This brings the opportunities for developing methods that explore the newly available genomic regions and discover global patterns of evolution, at high speed and intensive computation to maintain the high level of detail.

Genomic sequences are under selection of various evolutionary forces. Conservation is one major and most studied one, accounting for essential molecular structures preservation. We found that others, like positive selection and lineage specific selection, are particularly active on the host-pathogen interaction. The main focus of this thesis is the development of methods able to detect such regions, evolving at different speed or following different patterns than natural random evolution. Due to the host-pathogen relationship, these regions have a high probability of being associated for disease. The main purpose of this thesis is to provide algorithms, methods and procedures for variability clustering and studies, as a general framework able to take into account, in addition to the DNA sequence, set of global group classifications, like partial and complete horizontal gene transfers, variability, epidemiologic categories, phylogenetic families or habitats. These considerations are explore in three different articles representing chapters III, IV and V.

Chapter III addresses current limitations of the existing algorithms to effectively use pathogenicity information, on the pathogen side of the host-pathogen relationship. Genome wide association studies have addressed this problem, but only on the host side, where algorithms do not address positive selection or lineage specific selection specifically. We also account for both monophyletic and polyphyletic studies. We developed an algorithm able to work with, or without previous external knowledge such as species carcinogenicity or invasivity.

Chapter IV provides a framework to study horizontal gene transfer patterns of evolution, of different clusters of habitats, occupied by multiple phylogenetic families. We tried to shed more

light on local and global rate of this phenomenon, adding partial horizontal gene transfer to this study, as previous studies were limited to complete gene transfers on the genomic level. The current state of the art attributes low values for this phenomenon when the authors favor advanced phylogenetic analysis, and follow tree like evolution. On the contrary, when they use pairwise distance measures, and reticulate evolution, they find much higher levels of interaction. Usually the first group of researchers use a core set of genes, that exhibit stable behaviour across evolutionary time, called "core genes", in order to preserve tree likelihood. The growing sequencing effort has shrinked this "core" set of genes. We here reconcile both views, stating that horizontal gene transfer is a continuous phenomenon, which rarely affects alleles, but accumulates at higher clustering levels affecting "core" genes many times during their history. Partial gene transfer is more frequent than complete transfer, showing a gradual integration of complete transfer, by intraspecific recombination. We also tried to uncover the gradient of values that is linked to different confidence intervals.

In Chapter V we used sequence variability clustering for horizontal gene transfer or recombination detection. By introducing new asymmetric operators and variability functions, we developed a fast algorithm, having same quadratic asymptotic complexity as the Hamming distance measure. It has a higher constant cost, used to maintain *p-values.* It is meant to rapidly detect candidates on the genomic level, used as an alternative to conservation measures, leaving the precise, statistically proven methods for subsequent validation.

CHAPTER III

DETECTING GENOMIC REGIONS ASSOCIATED WITH A DISEASE USING AGGREGATION FUNCTIONS AND ADJUSTED RAND INDEX

Published in:

BMC Bioinformatics 2011, 12:S9 doi:10.1186/1471-2105-12-S9-S9

3.1 Abstract

3.1.1 Background

The identification of functional regions contained in a given multiple sequence alignment constitutes one of the major challenges of comparative genomics. Several studies have focused on the identification of conserved regions and motifs. However, most of existing methods ignore the relationship between the functional genomic regions and the external evidence associated with the considered group of species (e.g., carcinogenicity of Human Papilloma Virus). In the past, we have proposed a method that takes into account the prior knowledge on external evidence (e.g., carcinogenicity or invasivity of the considered organisms) and identifies genomic regions related to a specific disease.

3.1.2 Results and conclusion

We present a new algorithm for detecting genomic regions that may be associated with a disease. Two new aggregation functions and a bipartition optimization procedure are described. We validate and weigh our results using the Adjusted Rand Index (ARI), and thus assess to what extent the selected regions are related to carcinogenicity, invasivity, or any other species classification, given as input. The predictive power of different hit region detection functions was assessed on synthetic and real data. Our simulation results suggest that there is no a single function that provides the best results in all practical situations (e.g., monophyletic or polyphyletic evolution, and positive or negative selection), and that at least three different functions might be useful. The proposed hit region identification functions that do not benefit from the prior knowledge (i.e., carcinogenicity or invasivity of the involved organisms) can provide equivalent results than the existing functions that take advantage of such a prior knowledge. Using the new algorithm, we examined the *Neisseria meningitidis* FrpB gene product for invasivity and immunologic activity, and human papilloma virus (HPV) E6 oncoprotein for carcinogenicity, and confirmed some well-known molecular features, including surface exposed loops for *N. meningitidis* and PDZ domain for HPV.

3.2 Background

Many bacteria and viruses adapt to changing environmental conditions through several evolutionary mechanisms such as homologous recombination (Posada and Crandall 2001), nucleotide substitutions, insertions-deletions (Kimura 1985), horizontal gene transfer (Boc et al. 2010), etc. These mechanisms lead to the formation of different polymorphic strands of the same group of organisms, in which the variation on the DNA composition is spread randomly throughout the genomes. The survival of these strands depends on their ability to overcome the environmental changes (Moran 1962). One of the goals of comparative genomics consists of finding the variation among aligned genomic sequences in order to identify functional regions. Several comparative genomic tools allow for the identification of genomic regions in an alignment that have evolutionary patterns different from the neutral evolution. For instance, PhastCons (Siepel et al. 2005) predicts, from a given alignment and the related phylogenetic tree, the genomic regions under negative selection. PAML (Yang 1997, Yang 2007, Yang et al. 2000) allows for the comparison of synonymous versus non-synonymous mutations in an alignment in order to predict regions under selective pressure. RDP3 (Martin et al. 2010) and TOPAL (Milne et al. 2004) are software packages including several methods for detecting recombination. Most of these methods and software do not

take into consideration external epidemiological evidence associated with many bacterial and virus strands. Such evidence can allow for the clustering of organisms based not only on the similarity of their genomic sequences, but also, on their association to different diseases. Hence, intra-specific and inter-specific variation among carcinogenic and non-carcinogenic human papilloma viruses can lead to the identification of regions related to carcinogenicity. In our previous works, we introduced a hit region identification function using prior knowledge information (Badescu et al. 2008) and described the related validation framework based on Monte-Carlo simulations (Diallo et al. 2009). Then, we extended the latter study by presenting and testing four variants of the hit region identification function, still using the available prior knowledge (Badescu et al. 2010). In this chapter, we present a new algorithm for the identification of specific genomic regions associated with an external disease. The introduced algorithm uses a bipartition optimization procedure to maximize a specific clustering function Q, based on inter- and intragroup variability, for each window position, over the given sequence alignment. It can be applied *with or without prior knowledge information* characterizing species in hand. Hit regions (i.e., putative regions related to a disease) can be validated using ARI (Hubert and Arabie 1985) - a corrected-for-chance version of the Rand index (Rand 1971) - and organismal bipartitions are constructed using the available epidemiological data. The new algorithm has been applied to two independent datasets: The human papilloma viruses and the *Neisseria meningitidis* data. The obtained results suggest that genomic regions with important biological features in both datasets can be associated with either carcinogenicity or invasivity.

3.3 Dataset description

3.3.1 Neisseria meningitidis dataset

Neisseria meningitidis is a Gram negative bacterium responsible for meningitis and septicemia. It has a relatively small genome size of 2.2 Mbp. In March 2011, the PubMLST database listed a total of 8,793 genetically distinct members of Neisseria organisms (Jolley et al. 2004). All these facts make *N. meningitidis* well suited for testing comparative genomics methods (Maiden 2008). Proteins expressed under iron limitation (e.g. FrpB(FetA)) are considered as potential vaccine components (Pettersson et al. 1997). Bacteria grown under iron starvation express several proteins, the most abundant of them being FrpB, a 70kDa outer membrane protein (OMP). It is expressed in large amounts in all strains, and antibodies against this protein appear to be bactericidal. A putative FrpB topology was first proposed with a 26-stranded β-barrel (Pettersson et al. 1995), and later

reassessed to a plug domain and a 22-stranded β-barrel with 11 surface-exposed loops (Kortekaas et al. 2007). These loops are accessible to the host immune system, which produces natural antibodies against these regions. In general, bacteria express genetic sequence variability in order to evade this defense mechanism.

The data we considered were classified on the invasivity basis using a list of identified hyperinvasive meningococci (Urwin et al. 2004). We then built a list of unique FetA sequence tags carried by the alleles of these organisms. Using local BLAST operations (Altschul et al. 1990), we searched for the presence of these tags in the distinct sequences belonging to the selected multiple sequence alignment (MSA), first examined in (Badescu et al. 2010). We classified as belonging to the invasive category (subset X) any allele that contained at least one of the selected invasive tags. All the other alleles were put in the non-invasive category (subset Y). We annotated the MSA with the information regarding surface-exposed loops, beta-sheets and periplasmic loops (Kortekaas et al. 2007). Translating indexes from the amino-acid sequences to DNA sequences were also computed. Each single value of the hit region identification function Q (the Q-type functions will be used to identify genetic regions that may be related to a disease) corresponds to an interval of a certain length (i.e., 9 or 20 nucleotides in this study) and depends on the starting position of the sliding window used in our algorithm.

3.3.2 Human papilloma virus dataset

Human papilloma viruses (HPV) have a causal role in cervical cancer with almost half a million new cases occurring each year (Angulo and Carvajal-Rodrguez 2007, Bosch et al. 1995, Munoz 2000). About a hundred of HPV types have been identified, and the whole genomes of more than eighty of them have been sequenced (see the latest Universal Virus Database report by International Committee on Taxonomy of Viruses (ICTV)). A typical HPV genome is a double-stranded, circular DNA genome of size close to 8 Kbp, with a small set of genes (L1, L2, E1, E2, E4, E5, E6 and E7). In this study, we focused on the gene E6, which is predominantly linked to cancer due to the binding of its product to the p53 tumor suppressor protein. It contains a PDZ domain-binding motif (-X-T-X-V) at its carboxy terminus, which is essential for targeting the PDZ proteins for proteasomal degradation. Such proteins include hDlg, hScrib, MAGI-1, MAGI-2, MAGI-3 and MUPP1 (Lee and Laimins 2004). The interaction between E6 and hDlg, or the other PDZ domain-containing proteins, may be an underlying mechanism in the development of HPV-associated cancers (Kiyono et al. 1997). The gene E6 was also shown to contain two stable folded domains,

E6N and E6C (Lipari et al. 2001, Nominé et al. 2003). Models of these domains have been built in the absence of complete crystallographic data (Nominé et al. 2006).

To define carcinogenic types, we used the epidemiological data from a large international survey on HPV in cervical cancer and from a multicenter case-control study conducted on 3,607 women with histologically confirmed cervical cancer (Munoz et al. 2004, Muñoz et al. 2003). More than 89% of them had squamous cell carcinoma (i.e., Squam cancer) and about 5% had adenosquamous carcinoma (i.e., Adeno cancer). More than a half of the infection cases were due to the types 16 and 18 of HPV, which are later referred to as High-Risk HPV (Chan et al. 1995). In this study, we examined the content of the gene E6 for 83 different HPV types.

We fixed the window size to 20 nucleotides for HPV datasets in order to be consistent with our previous works (Badescu et al. 2008, Diallo et al. 2009), where we conducted simulations with windows of different sizes and used the size of 20 bp to present the results. In the same way, we considered the window size of 9 nucleotides for the *N. meningitidis* dataset to be consistent with another our study (Badescu et al. 2010).

3.4 Methods

3.4.1 Description of the algorithm

The new algorithm takes as input a MSA established for a set of organisms. Assume that this set of organisms is partitioned into two different subsets according to a Boolean criterion (e.g., invasivity vs. non-invasivity or carcinogenicity vs. non-carcinogenicity). The corresponding subsets are denoted X (invasive/carcinogenic) and Y (non-invasive/non-carcinogenic), respectively. The region of interest is scanned using a non-overlapping sliding window, as shown in Figure 3.1, of a fixed width (20 sites for HPV and 9 sites for *N. meningitidis*). For each window position, we carry out a bipartition optimization algorithm in order to search for maximum values of the hit region identification function. A specific version of the Q-type function (see below) can be taken as the algorithm parameter. We denote by Q' a specific version of the Q-type function computed under condition that the subsets bipartition is unknown (i.e., prior knowledge). The complete algorithmic scheme is presented in Algorithm A.1 in Additional file 1 (see Appendix A).

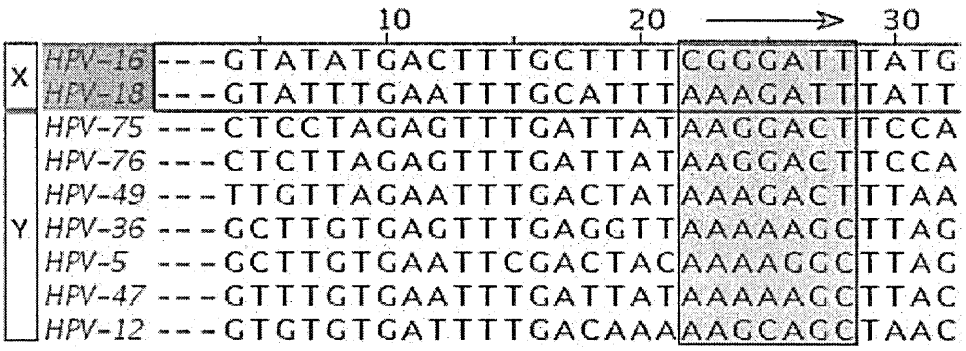


Figure 3.1. Sliding window procedure
Sliding window of a fixed width was used to scan the HPV gene E6. The sequences in black belong to the set X (carcinogenic HPV; in this example HPV 16 and 18), all the other sequences belong to the set Y (non-carcinogenic HPV). The HPV type is indicated in the left column.

3.4.2 Clustering using the Q-type functions

To perform the clustering of our data into two groups A and B, we first calculate the intragroup variability of the sequences from the group A, denoted by $V(A)$, the group B, denoted by $V(B)$, and, finally, the intergroup variability $D(A, B)$, as described in Equations 3.1, 3.2 and 3.3. These measures are defined as the means of the squared Hamming distances, *dist*, among the sequence fragments (bounded by the sliding window position) of the taxa from the group A only, from the group B only, and between the sequence fragments from the distinct groups A and B:

$$V(A) = \frac{\sum_{\{a_1, a_2 \in A | a_1 \neq a_2\}} dist_h^2(a_1, a_2)}{N(A) \times (N(A) - 1)/2}, \tag{3.1}$$

$$V(B) = \frac{\sum_{\{b_1, b_2 \in B | b_1 \neq b_2\}} dist_h^2(b_1, b_2)}{N(B) \times (N(B) - 1)/2}, \tag{3.2}$$

$$D(A, B) = \frac{\sum_{\{a \in A, b \in B\}} dist_h^2(a, b)}{N(A) \times N(B)}. \tag{3.3}$$

In (Badescu et al. 2008, Badescu et al. 2010, Diallo et al. 2009) four different hit region identification functions, Q_1, Q_2, Q_3 and Q_4, which could be summarized by the following equation, were defined:

$$Q = D(A,B) - k \times V(A) - l \times V(B), \tag{3.4}$$

where the $[k, l]$ combinations are as follows:

$$Q_1 \rightarrow (1,0), Q_2 \rightarrow (0,1), Q_3 \rightarrow (\frac{1}{2}, \frac{1}{2}), \text{and } Q_4 \rightarrow (0,0).$$

The function Q_4 (Equation 3.5), along with new versions of the hit region identification function, denoted by Q_5 (Equation 3.6) and Q_6 (Equation 3.7), will be tested and discussed in this study:

$$Q_4 = D(A,B), \tag{3.5}$$

$$Q_5 = |V(A)\text{-}V(B)|, \tag{3.6}$$

$$Q_6 = |V(A)/V(B)|. \tag{3.7}$$

Measuring the agreement between the reference and the optimal calculated bipartitions using the Adjusted Rand Index (ARI)

The Adjusted Rand Index (Hubert and Arabie 1985) has become a criterion of choice for measuring agreement between two partitions in clustering analysis (Milligan and Cooper 1986). Having a calculated bipartition $U'' = A \mid B$ and a reference bipartition $U' = X \mid Y$, for all $\binom{n}{2}$ pairs of elements, one can compute how many of them fall into the same group and how many in different groups. One can then calculate ARI (Santos and Embrechts 2009) according to Equation 3.8. ARI is the corrected-for-chance version of the Rand index (Rand 1971). It ranges between -1 and 1, and expresses the level of concordance between two bipartitions (Hubert and Arabie 1985). The values of ARI close to 1 indicate an almost perfect concordance between the two compared bipartitions, whereas the values close to -1 indicate a complete discordance between them:

$$ARI = \frac{\binom{n}{2}(a+d)\text{-}[(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2\text{-}[(a+b)(a+c)+(c+d)(b+d)]}, \tag{3.8}$$

where $\binom{n}{2} = a+b+c+d$, a is the number of pairs that are in the same group in the bipartitions U'' and U', b is the number of pairs that are in the same group in the bipartition U'' and in different groups in the bipartition U', c is the number of pairs that are in different groups in the bipartition U'' and in the same group in U', and d is the number of pairs that are in different groups in the bipartitions U'' and U'.

Validation of the obtained hit regions using the Adjusted Rand Index

We define a new function Q'' reflecting the quality of the reference bipartition, as follows:

$$Q'' = ARI \times Q'. \tag{3.9}$$

The difference between Q' and Q'' indicates the level of concordance of the reference bipartition U' with the selected function Q. Throughout this study, Q will denote the hit identification function using prior knowledge information, Q' – not using any prior knowledge information and Q'' – using prior knowledge information and based on ARI.

3.4.3 Bipartition optimization

For each window position, we generated a fixed number of random initial bipartitions. For each such a bipartition, we moved elements from one subset to the other and back again in cycles, each time accepting the move that maximized the objective function Q, until no further improvement was possible. Once a local maximum was reached, we compared it to the best current value obtained for all starting random bipartitions tested so far. ARI was used to compare the level of concordance of the obtained bipartition (i.e., the one that was maximizing the given function Q) with the reference bipartition (carcinogenic vs. non-carcinogenic taxa for HPV and invasive vs. non-invasive taxa for *N. meningitidis*) given as a parameter to the algorithm.

3.4.4 Time complexity

The time complexity of the new algorithm carried out with an overlapping sliding window of a fixed width, and advancing one alignment site by step, is $O(l \times n^2 \times w \times r)$, where l is the length of the MSA, n the number of considered species, r the number of random initial partition generations and w the window width. In order to ensure this complexity, we have to limit the optimization cycle to a constant number of iterations.

3.4.5 Simulation study

In order to validate the hit region identification functions Q_4', Q_5' and Q_6', we conducted a Monte-Carlo simulation study involving two major evolutionary mechanisms: Positive selection (PS) and Lineage specific selection (LSS). Two cases of group selection were also tested: The cases of the monophyletic and polyphyletic clustering. An approach involving the computation of p-values was implemented to asses the predictive ability of each of the three functions for each combination of evolutionary parameters. The following procedure was carried out. A phylogenetic tree T with 16 leaves was first generated using the algorithm described by (Kuhner and Felsenstein 1994). The edge lengths of T were generated using an exponential distribution. Following the approach of (Guindon and Gascuel 2002), we added some noise to the tree edges in order to provide a deviation

from the molecular clock hypothesis. The random trees yielded by this procedure had depth of $O(log(16))$. The tree was then rooted by midpoint. For the monophyletic test, the left and right sub-trees, denoted by T_1 and T_2, were determined, depending on the position of the root. For the polyphyletic test, two sets of leaves were randomly chosen and the corresponding sub-trees, denoted by T_3 and T_4, were extracted.

In the PS simulations, we used the original lengths of the edges of the subtrees T_1 and T_2 (i.e., monophyletic case), and T_3 and T_4 (i.e., polyphyletic case), while all edge lengths of T were gradually multiplied by the scaling factor a, varying from 0.05 to 1 (with the step of 0.05).

In the LSS simulations, all edge lengths of T were multiplied by 0.5 (thus simulating neutral evolution), while all edge lengths of T_1 and T_3 were multiplied by the scaling factor $a_1 = 0.5 + 0.025x$, and all edge lengths of T_2 and T_4 by $a_2 = 0.5 - 0.025x$, where x was varying from 1 to 19.

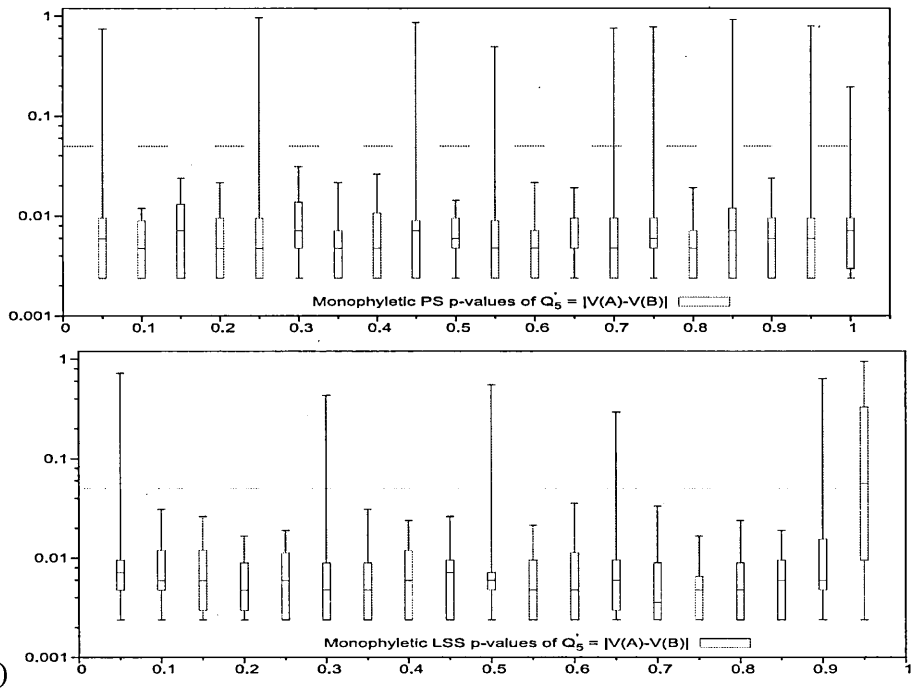
Second, we executed the SeqGen program (Rambaut and Grass 1997) to generate random MSAs of nucleotide sequences along the edges of the phylogenetic trees constructed at the first step. The SeqGen program was used with the Jukes-Cantor model of sequence evolution. DNA sequences with 440 bp were generated for each tree T. In addition, MSAs of the length 20 bp were generated for each of the trees T_1, T_2, T_3 and T_4. Two different variants of MSA were produced to simulate monophyletic and polyphyletic evolution. In the sequence alignment generated for the original tree T, we inserted those generated for the trees T_1 and T_2 in the monophyletic case, and those generated for the trees T_3 and T_4 in the polyphyletic case. The location of the inserted sequence blocks was known.

Thus, depending on the scaling factor parameters, for the PS case we simulated a variable homogeneous region inside a conserved context, and for the LSS case a more divergent region inside a neutral context. Third, we scanned the resulting sequence alignment using a sliding window of size 20 bp with the step of 1. We calculated the value of the hit region identification functions Q_4', Q_5' and Q_6' for each fixed position of the window and assessed the proportion of their values that were higher than the reference value corresponding to the inserted region.

These steps were repeated over 100 different replicates and the distributions of the best (in each case) function over each combination of testing parameters were represented using quartiles.

3.5 Results and discussion

We proposed a new algorithm for finding genomic regions that may be related to a disease along with two new hit region identification functions Q_5 and Q_6. Both new functions along with the best existing function Q_4 were tested in simulations. The functions yielding the best results for each case were illustrated in Figure 3.2: Monophyletic evolution (case a: PS, case b: LSS) and Figure 3.3: Polyphyletic evolution (case a: PS, case b: LSS). The remaining results for the Q_4', Q_5' and Q_6' functions are presented in Additional file 1 (see Appendix A). Figures 3.2 and 3.3 clearly show that the hit zone identification in the monophyletic case is much easier than in polyphyletic case. We can suggest that in order to be recognized, the hit region has to have a different evolutionary speed than the context in which it resides. The polyphyletic lineage specific case represents the hardest evolutionary situation. Also, one can notice that different Q-type functions, Q_4', Q_5' or Q_6', should be used in different practical situations.



(a)

(b)

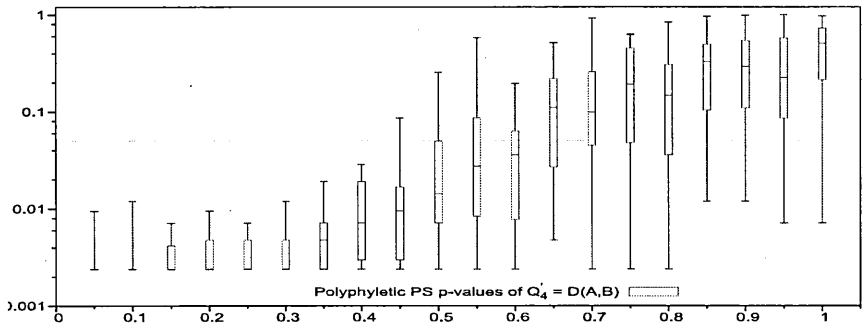
Figure 3.2. *P*-values obtained for monophyletic evolution hit region detection

(a) Positive selection - Variable hit region inside conserved context.

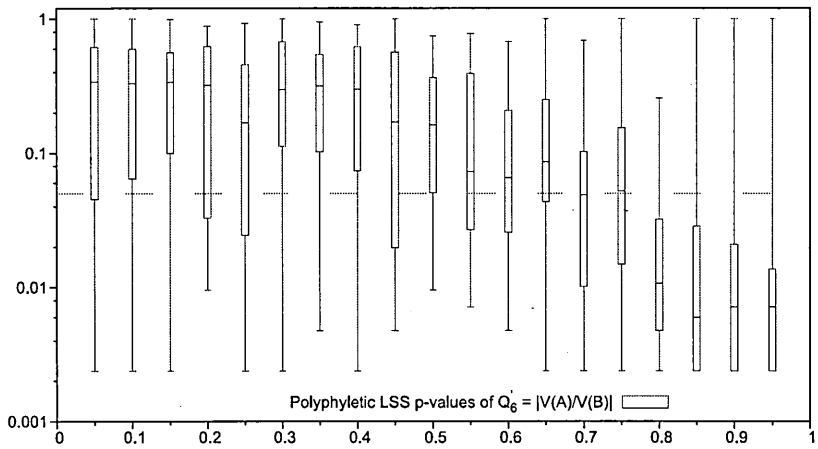
Quartile distribution of p-values obtained for the function Q_5'. Abscissa represents scaling factor of the conserved context in which the variable hit region resides. Values close to 0 represent conservation (maximum discrimination), while values close to 1 represent variability (identical to context). Variable hit region is always maintained at a scaling factor of 1. Ordinate represents p-values in log-scale. Horizontal dashed line represents the significance threshold of 0.05.

(b) Lineage specific selection - Heterogeneous hit region inside neutral context.

Quartile distribution of p-values obtained for the function Q_5'. Abscissa represents the difference in scaling factors among the two lineages present in the hit region. Values close to 0 represent homogeneous evolutionary speed (similar to the neutral context in which it resides), while values close to 1 represent divergence among these lineages. Context is always maintained at a scaling factor of 0.5, simulating neutral evolution. Horizontal dashed line represents the significance threshold of 0.05. In the case of lineage specific selection, the value of the Q'-type functions corresponding to 1 on the abscissa scale cannot be computed because it involves a sub-tree with 0 edge lengths.



(a)



(b)

Figure 3.3. P-values obtained for polyphyletic evolution hit region detection
(a) Positive selection - Variable hit region inside conserved context.

Quartile distribution of p-values obtained for the function Q_4'. Variable hit region is always maintained at a scaling factor of 1. Abscissa represents scaling factor of the conserved context in which the variable hit region resides. Values close to 0 represent conservation (maximum discrimination), while values close to 1 represent variability (identical to context). Ordinate represents p-values in log-scale. Horizontal dashed line represents the significance threshold of 0.05.

(b) Lineage specific selection - Heterogeneous hit region inside neutral context. Quartile distribution of p-values obtained for the function Q_6'. Context is always maintained at a scaling factor of 0.5, simulating neutral evolution. Abscissa represents difference in scaling factors among the two lineages present in the hit region. Values close to 0 represent homogeneous evolutionary speed (similar to the neutral context in which it resides), while values close to 1 represent divergence among these lineages, and from the neutral context. Horizontal dashed line represents significance threshold of 0.05.

The procedure for the identification of hit regions was carried out to detect the variability zones in the FrpB gene of *N. meningitidis* as well as the regions potentially responsible for cancer in the gene E6 of HPV. In both cases, we also carried out the ARI validation.

3.5.1 Neisseria meningitidis analysis

We scanned the MSA of the FrpB gene using the new algorithm with a sliding window of size 9 nucleotides. We compared the obtained results to the putative topology model of the FrpB protein described in (Kortekaas et al. 2007) (see Figure 3.4a). The results are presented in Figure 3.4b and c. Remarkably, all surface exposed loops confirmed by enzyme-linked immunosorbent assay (i.e., L2, L3, L4, L5 and L10) (Kortekaas et al. 2007) were properly detected using the functions Q_4' and Q_5'. It is worth noting that our algorithm was able to find the loop L4, which is hidden between the loops L5 and L3. The model loops L1, L8 and L9 were found at their predicted positions. The loops L2 and L11 were found at different positions, while the loops L6 and L7 were missed regardless of the availability of the prior knowledge information (see Figure 3.4b and c). As protein models gradually improve and more crystallographic data become available, it will be interesting to reassess these results in the future. Both presented Q'-type functions (Equations 3.5-3.6) overlap along the alignment, with the exception of the largest loop (L5) and the second largest loop (L3), where the amino acid variability is largely confined. The function Q_4' correlates best with surface exposed loops structure. This suggests that the divergences in shape between the functions Q_4' and Q_5' might be used to detect immunologic activity. It is known that bactericidal antibodies are directed against variable regions situated in the largest loops of proteins (Van Der Ley et al. 1991). Note that the organisms compared here were strains of the same bacterium; their genetic variant being alleles and evolutionary distances between them being very small. On such a small timescale, underlining evolutionary processes are usually not very diverse. It would be also interesting to verify whether similar conclusions could be made for other outer membrane proteins.

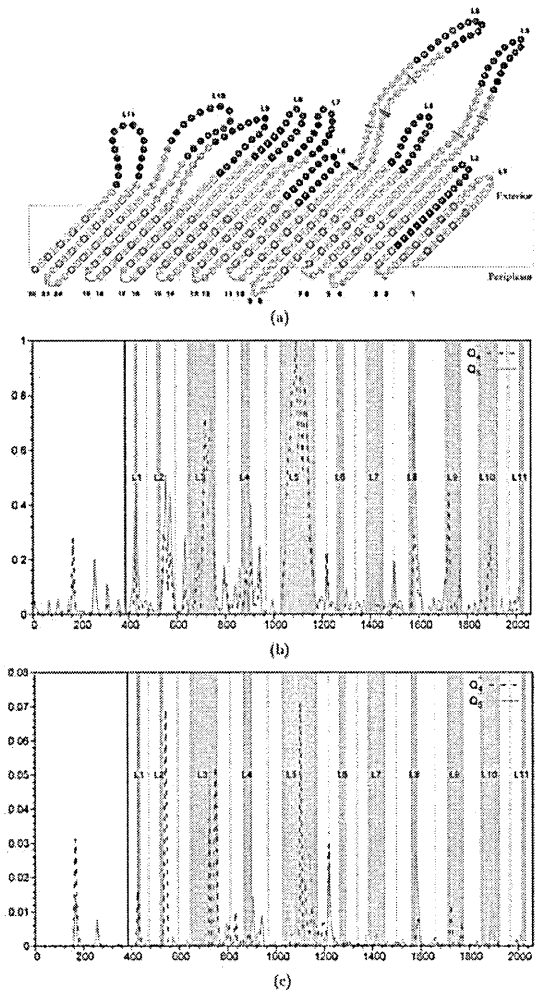


Figure 3.4. N. meningitidis FrpB protein variability zone detection

(a) Topology model of the FrpB protein of N. meningitidis strain H44/76. Topology of the β-barrel. Surface-exposed loops (L) and β-strands are numbered. Residues are framed according to their predicted secondary structure: Amino acid residues in β-strands are depicted by diamonds. Amino acid residues present in exposed loops and periplasmic turns are depicted by circles - reproduced from (Kortekaas et al. 2007).

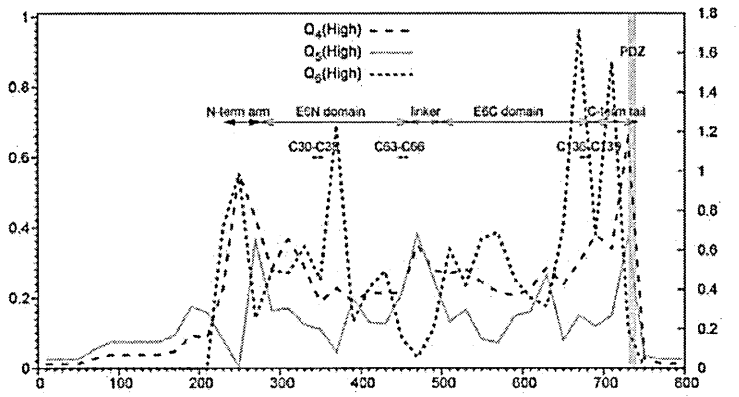
(b)-(c) Variability zone detection by the hit region identification Q′-type functions, achieved without prior knowledge of invasive taxa (case b), and Q″-type functions, using this prior knowledge along with the ARI coefficient (case c). Functions Q_4' and Q_4'' are depicted by a dashed line and functions Q_5' and Q_5'' are depicted by a continuous line. A non-overlapping sliding window of size 9 nucleotides was used during the scan of the gene FrpB MSA. The abscissa axis represents the window position along the nucleotide MSA. 11 gray zones correspond to extracellular loops. Annotations start at the solid vertical line (near the 400 abscissa mark).

3.5.2 Human Papilloma Virus analysis

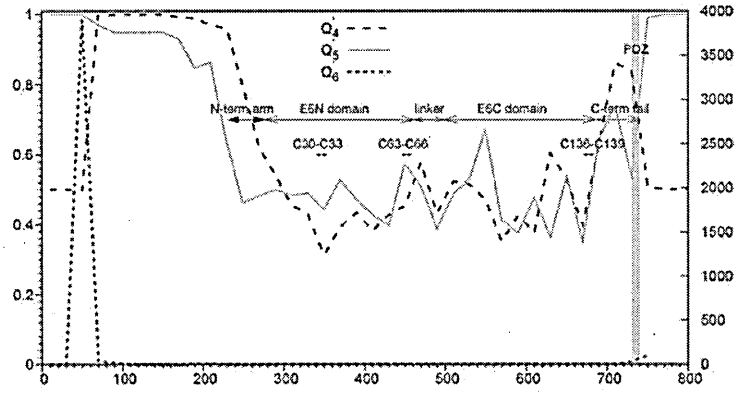
We performed a scan of the MSA of the gene E6 for 83 HPV organisms (using non-overlapping windows of size 20 nucleotides). Each time the species bipartition was known, High-Risk HPV against all other HPV types in Figure 3.5a, Squam-Risk HPV against all other HPV types in Figure 3.6a, and Adeno-Risk HPV against all other HPV types in Figure 3.6b, it was incorporated in the computational procedure as shown in Algorithm A.1. The comparative results for the High Risk HPV subset provided by the new algorithm *without prior knowledge of carcinogenic taxa* and those yielded by the former one (Badescu et al. 2008), are presented in Figure 3.5 using annotations for HPV-16. Figure 3.5a illustrates the results obtained using the functions Q_4 and Q_5 *using a prior knowledge* on the species carcinogenicity.

According to the new algorithm, see Figure 3.5b, the PDZ domain is ranked first in the annotated part of the alignment. A detailed view of the terminal aligned region, within the index interval 680-740, shows a small left shift in the peak positions of the function Q_4' (3.5b), but inside the same C-terminal tail domain. On the left side, flanking the PDZ domain, one can find the E6C domain which is related to the DNA binding (Nominé et al. 2006). One can notice that the function peaks (see Figure 3.5a and 3.5b) of Q_4' are almost in the same positions than those found using Q_4, exception being a region at the beginning of the alignment (i.e., at the beginning of the E6N domain). As for *N. meningitidis* loops, it would be interesting to study in greater details the regions recognized by both tested functions, Q_4' and Q_5'.

We can conclude, by comparing Figures 3.5a and 3.5b, that the new functions, Q_4' and Q_5', provide almost identical hit region recovery than the existing functions Q_4 and Q_5, which take advantage of a prior knowledge on the species carcinogenicity.



(a)



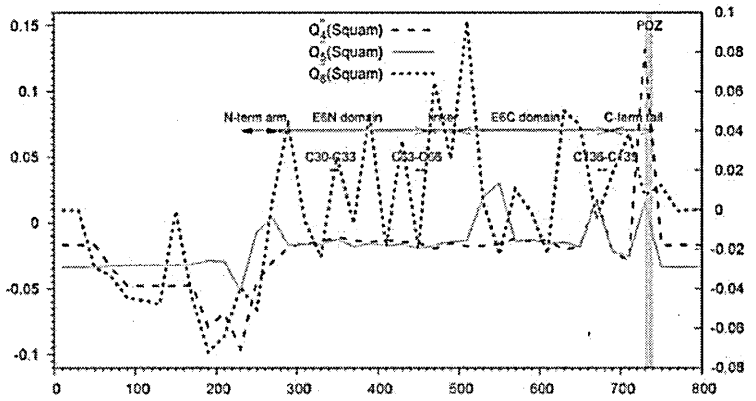
(b)

Figure 3.5 Hit region identification functions for High-Risk HPV

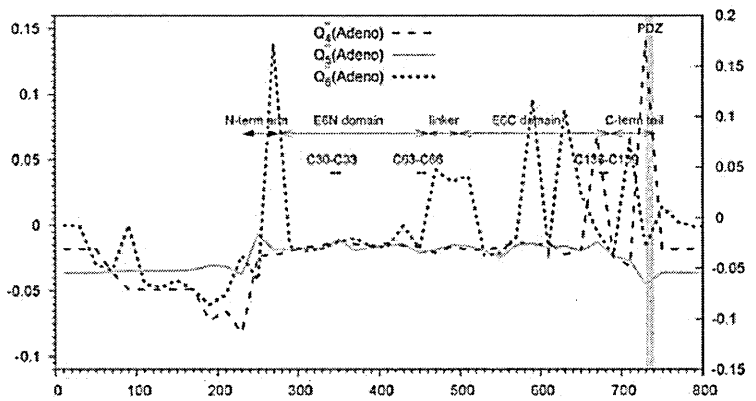
(a) Functions obtained *using prior knowledge* on the taxa carcinogenicity.

The hit region identification functions Q_4, depicted by a dashed line, Q_5, depicted by a continuous line, and Q_6, depicted by a dotted line, for the High-Risk HPV (HPV-16 and 18) (Badescu et al. 2008),(Diallo et al. 2009), during the scan of the gene E6.

(b) Functions computed *without prior knowledge* on the taxa carcinogenicity. The hit region identification functions Q_4', depicted by a dashed line, Q_5', depicted by a continuous line, and Q_6', depicted by a dotted line, during the scan of the gene E6. The abscissa axis represents the window position along the nucleotide multiple sequence alignment. The PDZ-domain is highlighted in gray. Annotations for the N and C-terminal arms, E6N and E6C domains are represented for HPV16 coordinates, from (Nominé et al., 2006) (Nominé et al. 2006). Zn^{2+}-ligating Cys residues annotations reproduced from (Lipari et al. 2001).



(a)



(b)

Figure 3.6. Q″-type functions, depending on ARI

(a) Squam HPV dataset. (b) Adeno HPV dataset.

Variation of the function Q_4", depicted by a dashed line, Q_5", depicted by a continuous line, and Q_6", depicted by a dotted line, obtained with the non-overlapping sliding window of width 20 nucleotides during the scan of the gene E6. The abscissa axis represents the window position along the nucleotide MSA. The *PDZ*-domain is highlighted in gray. Annotations for the N and C-terminal arms, E6N and E6C domains are represented for HPV16 coordinates, from (Nominé et al., 2006) [30]. Zn^{2+}-ligating Cys residues annotations reproduced from (Lipari et al. 2001).

The Q'' function validation was also carried out for HPV data. The results are presented in Figure 3.6. Here, the PDZ domain ranks first for both tested datasets, related to the Squam and Adeno cancers (Figures 3.6a and 3.6b). The peaks were found at almost the same positions as in Figure 3.5, with the exception that only some of the peaks shown in Figure 3.5 are present here. The function Q_4'' seems to be less variable than the function Q_5''. For the Squam dataset, there is one peak in the E6C domain, absent in the Adeno dataset, with a high monophyletic signal and unknown annotation.

On the other hand, the peak located at the index 660, and corresponding to the window positions 660-680, includes two putative Zn^{2+}-ligating Cys residues whose absence in mutants results in a dramatic loss in the p53 degradation activity (Lipari et al. 2001).

By analyzing Figures 3.4, 3.5 and 3.6, one can notice that in some situations prior knowledge information brings an important advantage to the method (see the case of Figure 3.4c when the use of the prior knowledge along with the ARI coefficient allows for getting rid of some false positive hits; for instance, the false positive picks found using Q'*-type* functions around the indices 1225 and 1500 presented in Figure 3.4b were not found by the Q''-type functions presented in Figure 3.4c as well as the case of an almost perfect PDZ domain recovery provided by the Q''-type functions as shown in Figures 3.6a and 3.6b), but in the other cases, the new algorithm is capable of correct recovering hit regions without any prior knowledge (e.g., see the cases of the loops L1, L3, L5, L8, L9 and L10 for the *N. meningitidis* dataset).

3.6 Conclusion

We described a new algorithm for finding genomic regions that may be associated with a disease. It is capable of detecting hit regions without prior knowledge on the carcinogenicity or invasivity of related organisms. This is an important improvement over previous works in the field (Badescu et al. 2010, Diallo et al. 2009). We also showed as the Adjusted Rand Index (Hubert and Arabie 1985, Milligan and Cooper 1986, Santos and Embrechts 2009) can be incorporated in the hit detection procedure. The discussed algorithm can be directly used to study organisms that have an ambivalent behavior and are, thus, more difficult to classify. For instance, some strains of *Neisseria Meningitidis* show a hyperinvasive behavior during epidemics, but are non-invasive, otherwise. The behavior of some other organisms, like human papilloma viruses (HPV), is more consistent. Such organisms can be clearly classified with respect to their level of carcinogenicity. Species

bipartitions, established according to carcinogenicity or invasivity criterion, suggested in the literature are important for the identification of genomic regions responsible for a related disease. We showed, however, that a successful identification of these regions can be accomplished without any prior knowledge of the species classification (Figure 3.5). Considering, in parallel, several hit region identification functions can provide more insight into the structure of genomic regions (Figures 3.4, 3.5 and 3.6). Simulation results suggest that there is no a unique function that provides the best overall results in all practical situations (e.g., the case of monophyletic or polyphyletic evolution and positive or negative selection), and that at least three different functions might be useful (Figures 3.2 and 3.3). It is worth noting that the monophyletic scenarios are easier to detect than the polyphyletic ones. The function Q_5 allows for a better detection of monophyletic scenarios, while in the polyphyletic case, the functions Q_4 and Q_6 provide the best results in the positive selection context and in the lineage specific selection context, respectively. The application of the described functions to the HPV gene E6 allows one to retrace the hit regions that are well-known to be related to carcinogenicity (Lee and Laimins 2004),(Kiyono et al. 1997),(Lipari et al. 2001),(Nominé et al. 2006).

Furthermore, the results given by these functions while analyzing the FetA sequences of *Neisseria meningitidis* suggest a large overlap between the regions with surface-exposed loops and those detected by the hit region identification functions (Figure 3.4). All these results indicate the ability of the proposed algorithm to identify regions with bipartite evolutionary signatures according to different patterns of evolution. Each time the species bipartition was known, High-Risk HPV against all other HPV types in Figure 3.5a, Squam-Risk HPV against all other HPV types in Figure 3.6a, and Adeno-Risk HPV against all other HPV types in Figure 3.6b, it was incorporated in the computational procedure as shown in Algorithm A.1. In the future, it will be important to assess the correlation between different non-overlapping detected hit regions present in the given alignment. It would be also interesting to compare the performance of the introduced bi-clustering algorithm with the existing bi-clustering methods currently used in bioinformatics, including SAMBA (Tanay et al. 2004), Crossing Minimization (Abdullah and Hussain 2006) and cMonkey (Reiss et al. 2006). Another possibility consists of using a k-means (MacQueen et al. 1967) type of algorithms that can suggest partitioning of the given dataset in several, and not necessarily in two, classes when the exact number of classes is unknown. For instance, in the case of HPV data, one could consider the three following HPV classes: High-Risk HPV (types 16 and 18), Low-Risk HPV (types 6, 11, 26,

31, 33, 35, 39, 45, 51, 52, 53, 55, 56, 58, 59, 66, 73, 81, 82 and 83) and No-Risk HPV (all other HPV types).

It is worth noting that the presented algorithm, like most of the comparative genomics methods, relies on the assumption of the alignment correctness. Thus, it will be also important to analyze the impact of alignment errors on the results of the proposed hit detection procedure.

We have provided the complete source code of our application allowing one to carry out the methods presented in this chapter; the application's name is QFUNC v.0.5. A Makefile along with the examples of the input and output data have been also made available. The ReadMe documentation file provides an explanation of the main steps to follow for executing the application. The source code and the accompanying files have been uploaded to the GitHub public repository (with the BSD licence). It is freely available at the following URL address: https://github.com/dunarel/dunphd-thesis/tree/master/Chapter3/Main/q_funcb.

3.7 Acknowledgements

This study was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC). This article has been published as part of *BMC Bioinformatics* Volume 12, 2011: Proceedings of the Ninth Annual Research in Computational Molecular Biology (RECOMB) Workshop on Comparative Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/12?issue=S9.

CHAPTER IV

COMPLETE AND PARTIAL HORIZONTAL GENE TRANSFERS AT THE CORE OF PROKARYOTIC ECOLOGY AND EVOLUTION

4.1 Abstract

Horizontal Gene Transfer (HGT) is one of the major evolutionary processes affecting prokaryotic species. Two known types of horizontal gene transfer are complete and partial HGT. Identifying the origins and the rates of horizontal gene transfers in the context of complete and partial HGT models, and this for different phylogenetic families and ecological habitats, is a very relevant and challenging problem. In this chapter, we describe a novel bioinformatics framework designed to estimate and compare the rates of complete and partial HGT at different phylogenetic and ecological levels. Well-known methods of phylogenetic tree inference (e.g. RAxML) and horizontal gene transfer detection (e.g. HGT-Detection) will be used in our experiments. We support a "genome space" view of prokaryotic evolution, in which individual strains interact based on ecological habitat and phylogenetic similarity. Our results suggest that partial HGTs are almost twice more frequent than their complete counterparts. Moreover, we show that partial HGTs, detected by the contemporary HGT detection algorithms, seem to be more recent than complete HGTs.

At the allele level, HGT seems to be rather a rare event. We estimated, using a 75% confidence HGT detection threshold, that the average HGT rate is 2.94 x 10^{-2} for complete transfers and 8.07 x 10^{-2} in overall (complete + partial transfers). This HGT rate is the probability that a contemporary prokaryotic allele or one of its direct ancestors (i.e. species located on the allele's lineage) have been ever affected by HGT coming from another prokaryotic organism during its evolutionary history. Thus, the majority of the existing prokaryotic alleles have not been affected by HGT. On the contrary, the majority of prokaryotic genes (i.e. a gene here is represented by a multiple alignment of the corresponding alleles) have been affected multiple times by gene transfers during its evolutionary history: 82.7% of the considered prokaryotic genes have been affected by at least one complete HGT and 96.3% - by at least one HGT in overall (these results are indicated for the HGT confidence threshold of 75%). We determined that the accuracy of the HGT age inference, which is another problem we addressed in this study, is the highest within the most recent 1000 Mya time period. It decreases progressively according to the time of HGT occurrence. The comparison between complete and partial HGTs also highlights the fact that the ages of partial HGTs, which are more recent than complete transfers, can be detected with a better confidence.

4.2 Introduction

Horizontal gene transfer is an important and widespread phenomenon in prokaryotic evolution (Koonin et al. 2001, Sjöstrand et al. 2014, Wolf et al. 2012). HGT has an important impact on microbial cooperation and bacterial virulence (Nogueira et al. 2009, Takeuchi et al. 2014). There exist three well-known HGT mechanisms, including transformation, transduction and conjugation, which allow DNA sequence acquisition either from the environment or directly from the donor species (Boc et al. 2010). The facility with which some bacteria develop antibiotic resistance is clearly an evidence of traits being transferred among species (Ochman et al. 2000), rather than *de novo* multiple mutations in each lineage (Davies and Davies 2010). High prevalence of HGT in prokaryotes has been demonstrated by the discovery of pathogenicity islands and virulence attributes (Koonin and Wolf 2008, Ochman et al. 2000). The latter events are relatively recent, and have a clear ecological component associated with maintenance, expansion or change of microorganism's ecological habitat (Smillie et al. 2011). Furthermore, bacteriophages, as gene transfer agents, stand as another compelling evidence of recent HGT (Koonin and Wolf 2008). Recently, bacterial sequences in cancer samples were found to integrate into the human somatic as well as into mitochondrial genomes (Riley et al. 2013).

Thus, gene transfer can be considered a well-established phenomenon on the "microscale": numerous biological experiments with bacteria and viruses, direct genome comparisons using simple heuristics (i.e. BLAST) and detection of anomalous characteristics of certain genomic sequences provide a compelling evidence of HGT (Smillie et al. 2011).

According to Koonin and Wolf (2009), in the 6th edition of the Origin of Species Darwin explicitly introduced the notion of the Tree of Life (TOL), (Darwin, 1872). Since then, phylogenetic tree thinking in biology became standard. We should mention, however, that the work of Darwin does not contradict the notion of reticulate evolution which is based on the use of phylogenetic networks for representing reticulate evolutionary mechanisms. Unfortunately, a phylogenetic tree accounts only for vertical (i.e. direct) transfer of genetic material and cannot be used for representing horizontal gene transfer events (Legendre and Makarenkov 2002). For instance, the traditional tree model is not convenient for studying the evolution of prokaryotic species.

The use of the term 'prokaryote' has been recently disapproved by some researchers, because Archaea and Bacteria do not form a monophyletic clade (Pace 2006). Similar arguments have been shown to exist for the eukaryotes, underlining the rigidness of the present nomenclature rules and their ability to deal with HGT (Syvanen 2012). The modern implications of reticulate evolution on the Tree of Life, and on Prokaryotes in general, have been extensively analyzed in a recent book (Doolittle and Zhaxybayeva 2013). The latter authors discuss the historic debate opposing Woese (Woese and Fox 1977) and Mayr (Mayr 1998), involving 'three' versus 'two' domains of life.

Many researchers have consequently developed a more appropriate concept, known as the "genome space", which is supported by phylogenetic networks (Huson and Bryant 2006). Rather than eliminating HGTs from the tree reconstruction, some authors used them as a support for the tree of life (Abby et al. 2012). In order to preserve the notion of phylogeny at the genomic level, the concept of "core of genes", has been proposed (Charlebois and Doolittle 2004). The core genes include a set of genes which are relatively "immune" to HGT and show a slow rate of evolution. However, the extent of HGT has been intensely debated, as the discussed HGT rates vary among different studies and clearly depend on the applied statistical models and HGT detection methods (Boc et al. 2010). In this vein, the debate opposing "genome space" and "core of genes" models of evolution has been going on for a long time, each party having its arguments, the former explaining

for the rapid adaptation of populations to the changing environmental conditions, the latter being more relevant the traditional view of species evolution (Koonin and Wolf 2008).

Continuing genome sequencing projects constantly contribute to the decrease of the set of core genes. Confronted to growing evidence that even the essential function of photosynthesis can be spread by HGT (Mulkidjanian et al. 2006), researchers struggle to identify core genes that potentially allow the separation of true (i.e., tree-like) phylogenetic signals from ''noise''(Shi and Falkowski 2008). Thus, relaxed measures should be introduced to account for sequencing and annotation artifacts and some genes' tendency to form multidomain proteins when establishing a set of "core genes" (Charlebois and Doolittle 2004).

Moreover, the relation between ecology and phylogenetics has been further refined by the observation that there is a relatively narrow variation in the prokaryotic genome sizes, which leads to an emerging view of bacterial genomes as samplers and not accumulators of genes. Thus HGT, which greatly contributes to the diversification of bacterial genomes, redefines the ecological niches of the microorganisms and promotes bacterial speciation (Ochman et al. 2000, Smillie et al. 2011). The existence of habitat-specific gene pools and their relationship with the core genome can explain how prokaryotic populations exhibit both ecological cohesion and high genomic diversity (Polz et al. 2013). Arguments in favor of an even larger pan-domain gene pool emphasized the role of HGT in ecologically important processes, ranging from heavy-metal detoxification to glycerol uptake and metabolism (Schönknecht et al. 2013).

The new emerging view of prokaryotic world is that of a single connected and compartmentalized gene pool, allowing for a gene exchange at variable rates, with fuzzy boundaries between species (Gogarten and Townsend 2005, Koonin and Wolf 2008, Smillie et al. 2011). An important study, based on the assumptions of relative constancy of ancestral prokaryotic genome sizes, estimates a minimum lower bound of the average rate of complete HGT at 1.1 event per gene family and family lifespan; the maximum rate can reach 2.1 events per gene family (Dagan and Martin 2007). Using the median tree method of inferring a species phylogeny (Kim and Salisbury 2001), the average rate of complete HGT among prokaryotes, estimated at the allele level, was found to be around 2% (Ge et al. 2005).

In this study, we propose a computational framework for estimating the rates of complete and partial types of HGT among prokaryotes. Extending the work of Smillie et al. (2011), who carried out their experiments for complete HGT only, we explore the impact of habitat and phylogenetic family affiliation on the exchange of genetic material in the context of both complete and partial HGT models (the HGT-Detection algorithms of Boc et al. (2010) and Boc and Makarenkov (2011) were used to detect and statistically validate HGT events).

Smillie and colleagues estimated the rate of the complete recent HGT for bacteria colonizing human environments at the maximum level of 20%. However, for non-human environments this rate was much lower, i.e. around 2% (see Fig. 1 in Smillie et al. 2011). On average, for all bacteria, the HGT rate was around 10%. In the pathogenic genes, this rate was much higher, with a maximum bound of 40% (see Fig. 4 in Smillie et al. 2011). Another recent study showed that the gut inflammation can boost pathogenic horizontal gene transfers (Stecher et al. 2012).

Using a directed network analysis Popa and colleagues discovered reliable donor–recipient relationships leading to a general HGT rate of 7% (Popa et al. 2011). The latter study also found that Proteobacteria form a highly connected cluster in the inferred HGT network. On the other hand, Crenarchaeota was found to be one of the groups exhibiting genetic mosaicism due to partial HGT (Ching et al. 2014).

We also identified the exact ages of the obtained complete and partial HGT events by using the B.E.A.S.T. v.1.7.5 (Drummond and Rambaut 2007) and TreePL (Sanderson 2002) programs. Szollosi and colleagues showed that the HGT phylogenetic modeling can contribute to the reconstruction of the relative speciation timing (Szollosi et al. 2012).

4.3 Materials and Methods

4.3.1 Data acquisition and classification

In our study, prokaryotic species were selected from the database of completely sequenced species genomes available at the NCBI Genomes ftp site. All of the completely sequenced prokaryotic genomes available at this site (1465, as of November 2011) were considered. Among them, we first selected 100 of the most complete prokaryotic genomes (belonging to the 23 available prokaryotic families) in terms of the number of genes.

The species selection was made proportionally to the percentage of the family representatives in the whole set of 1465 prokaryotic species. Then, we added to them 11 additional species to ensure that our dataset includes at least one representative from each of the 23 prokaryotic families (some families had less representatives than 1% of the total number of species). This yielded us a total number of 111 species, denoted throughout this chapter as *Species set*. Detailed information on the considered species can be found in Supplementary Table 1. We also identified 110 of the most complete genes (see Supplementary Table 3) belonging to the selected set of 111 species. The latter genes were labeled as *Ubiquitous gene* set. A *Core gene* set was then defined as its subset including 36 genes previously identified by Charlebois and Doolittle (Charlebois and Doolittle 2004).

The limits that we imposed on the number of considered species and genes was dictated by the high time complexities of the HGT detection and validation algorithms used in our study (e.g. Partial HGT detection program with the HGT bootstrap validation of the scanned sequence fragments (Boc and Makarenkov 2011) and has an exponential time complexity on the number gene transfers). At the same time, the Bayesian inference of the ages of partial HGTs, carried out over multiple sliding windows, took more than 3 months for a complete execution. We also tried to run this analysis with the datasets including 250 and 500 of the most complete prokaryotic genomes, but these computations were stopped as they should have taken over 12 months with the computational resources allowed to us at Compute Canada High Performance Cluster. We opted for an experimental design in which we invested the available computational resources in performing statistical validation of the inferred transfers by means of bootstrapping. Another valid approach would be to abandon the idea of statistical validation of transfers in favor of sampling a higher number of organism/genes. Yet, another approach would be to sample repeatedly different sets of organisms/genes and then to average the obtained results in order to assess their robustness.

Afterward, we constituted 110 multiple sequence alignments (MSAs) of amino-acid sequences (one MSA per selected gene) from which we excluded misclassified paralogs using TribeMCL (Enright et al. 2002). When multiple alleles of the same species were available, all of them were included in the corresponding MSA. The TribeMCL program, which implements a Markov Chain Clustering (MCL) algorithm (van Dongen 2000) on all-to-all BLASTP hits, is known to be conservative in terms of the number of groups (Li et al. 2012). We carried out the TribeMCL version of the program, bundled with "mcl" v11.294, with default parameters (I=2.0). In order to obtain more accurate results of BLASTP, we selected a Smith-Waterman backend and an E-value threshold of 10^{-4}. About 1% of the original alleles were identified as potential paralogs, using this procedure, and thus excluded from the original MSAs.

The nucleotide sequences corresponding to the selected amino-acid sequences were retrieved from the associated chromosomes available at GenBank. The retrieved nucleotide sequences were aligned using the MUSCLE tool (v3.8.31, Edgar 2004) and then corrected using the GBlocks program (v0.91b, Castresana 2000) which eliminates misaligned sequence fragments. In our analysis, we were less restrictive than the default option of GBlocks, allowing 50% of the sequences for flank positions (-b2 parameter), a maximum of 10 contiguous nonconserved positions (-b3 parameter), minimum block length of 5 (-b4 parameter) and half gap positions (-b5 parameter).

The 110 multiple sequence alignments analyzed in our study are available at the following URL address: http://www.info2.uqam.ca/~makarenkov_v/alignments.zip. We also provide Gene IDs (still available with the new NCBI prokaryotic genome annotation pipeline, as of November 2015) - for the considered Ubiquitous dataset (presented in Supplementary table 3), of the amino-acid sequences considered in this study at the same URL address (Angiuoli et al., 2008), (Tatusova et al., 2013). The corrected nucleotide MSAs were then used as basis for building gene trees, given as input to the HGT detection algorithms (Boc et al. 2012). We constructed the gene trees by means of the RAxML method (Stamatakis 2006). Species taxonomy (i.e. species tree in the HGT context) was retrieved from the NCBI Taxonomy website (Benson et al. 2009). Taxonomic groups (i.e. families) were those assigned by the NCBI Genome Project. Note that each species was assigned to one established prokaryotic family.

In this study, we explored the patterns of HGT by considering two different classifications of prokaryotic species. The first way follows the taxonomic species classification provided by the

NCBI Genome Project. It is based on the established prokaryotic family classification. The second way takes into account possible ecological localizations, or habitats, of the selected species. The set of the available habitats, described by MIGS Field (Field et al. 2008), was extracted from the Genomes OnLine Database – GOLD (Pagani et al. 2012). It is worth noting that the Extreme habitat is a heterogeneous collection of habitats corresponding to extreme environmental conditions, (e.g., superheated waters, acid-laden streams around old mines, frigid Antarctic ice, super-salty waters of the Dead Sea). This classification complies with the annotation of the GOLD database. Mention that many of the organisms belonging to the Extreme habitat also belong to some other habitats. In the GOLD classification (Pagani et al. 2012) each species could belong either to a unique or to multiple habitats.

4.3.2. Phylogenetic reconstruction and HGT detection

In order to detect and validate complete and partial horizontal gene transfers using the HGT-Detection algorithms (Boc et al. 2010; Boc and Makarenkov 2011), we need to have a species tree and a gene tree (or a gene MSA for the partial HGT inference). These algorithms proceed by reconciliation of the trees, gradually transforming the species tree into the gene tree in order to infer horizontal gene transfers. An important advantage of these algorithms is that they allow for validating the obtained HGTs statistically by estimating their bootstrap support (Boc et al. 2010).

To reconstruct the species tree, representing the traditional taxonomic species pattern, for the selected set of 111 prokaryotic species, we considered the available NCBI species Taxonomy (Benson et al. 2009). Then, for each considered multiple sequence alignment, we computed a gene tree representing the evolutionary history of the given gene. This history may be different from the classical taxonomic pattern due, for example, to HGT, recombination or hybridization phenomena (Legendre and Makarenkov 2002). In the case of prokaryotes, complete and partial HGTs (i.e. partial HGT is a complete HGT followed by intragenic recombination) are the most plausible explanations for topological discrepancy between the species and gene trees. To infer the gene trees, the RAxML reconstruction method (the RAxML program v7.2.8 with multithreads; see (Stamatakis 2006) was used with the following parameters: GTR Gamma model, 20 starting random trees, and 100 bootstrap replicates. Then, we reconciled each gene tree with the species tree to identify statistically plausible HGT scenarios. To this end, we inferred complete gene transfers using the HGT-Detection program (v.3.4), (Boc et al. 2010) available on the T-Rex web site (Boc et al.

2012). Partial gene transfer detection was based on a sliding window procedure described in Boc and Makarenkov (2011). We implemented this procedure using the KSH, JRuby and Java scripts and a multilayer approach. Thus, the 110 multiple sequence alignments we considered were scanned with a sliding window algorithm. Sliding window sizes were equal to 10%, 25% and 50% of the total alignment length. Partial HGTs were recovered from the overlapping MSA fragments (when the same HGT was found for multiple consecutive positions of the sliding window) based on a Jaccard similarity of at least 75%. This allowed us to account for the tree inference instability from short MSA fragments (Boc and Makarenkov 2011). To gain in computing time, we parallelized the partial HGT-Detection program and ran it on a parallel Mammoth cluster (Compute Canada High Performance Cluster), in addition to using the parallel version of the RAxML program. Partial HGTs spanning to multiple MSA fragments could have several bootstrap support scores in each of them. We assigned a single value to these multiple-fragment partial HGTs that corresponded to the maximum bootstrap support of the components.

For the two methods of HGT prediction (complete and partial), we used, in turn, bootstrap thresholds of 50%, 75% and 90%, respectively, to assess the robustness of the obtained HGT.

Once complete and partial HGTs were predicted for the whole dataset, we computed HGT rates between the species of the same phylogenetic family (or of the same habitat), and then between the species of different phylogenetic families (or of different habitats). Formulas (4.1-4.6) below were used to compute the presented intragroup and intergroup HGT statistics. These formulas normalize the obtained HGT rates with respect to the number of considered alleles. We also provide a more general result regarding prokaryotic alleles, which is the average probability estimation that a prokaryotic allele has been affected by HGT during its evolution (see Formula 4.7). Moreover, we highlight ten of the most frequent HGT events for each level of confidence we considered (i.e. 50%, 75% and 90% HGT bootstrap thresholds).

As the applied partial HGT-Detection algorithm (Boc and Makarenkov 2011) was inferring both partial and complete HGTs, the obtained absolute rates were denoted as overall gene transfer rates (i.e. complete HGT was a particular case of partial HGT in this algorithm; thus, the indicated overall HGT rates account for both complete and partial transfers; see Tables 4.1-4.3).

4.3.3. Computation of HGT statistics

In this section, we present the main formulas used to calculate HGT statistics and the corresponding explanations regarding the computation of the HGT weights (Figure 4.1). The presented formulas were used to generate: (a) heat maps of horizontal gene transfer events between 23 prokaryotic families (Figures 4.2 and 4.3), (b) heat maps of intragroup HGT rates (Figures 4.2 and 4.3, on the main diagonal), (c) histograms of outgoing HGT rates (Figure 4.4), (d) histograms of incoming HGT rates (Figure 4.5), (e) histograms of intragroup HGT rates (Figure 4.6), and, finally, (f) the overall probability that a prokaryotic allele has been affected by HGT during its evolution (Table 4.1a). We detailed the probability results by indicating the rates for the three selected HGT bootstrap thresholds for the entire set of considered genes and for the genes explicitly classified as core genes. These formulas were then adopted to calculate the HGT rates for the habitat study (Figures 4.9-4.13).

If several groups were involved in an HGT, the obtained transfers were weighted taking into account all involved alleles. Figure 4.1 illustrates a possible case: the transfer between the cluster including alleles belonging to species of families F1 and F4 and the cluster including alleles belonging to species of families F2 and F3 is accounted for as follows:

W(F1 -> F3) = 0.5, W(F1 -> F2) = 1.5,

W(F4 -> F3) = 0.5, W(F4 -> F2) = 1.5.

Thus, the horizontal gene transfer event depicted in Figure 4.1 is decomposed into four weighted HGTs. The resulting weights depend on the number of affected alleles of each family.

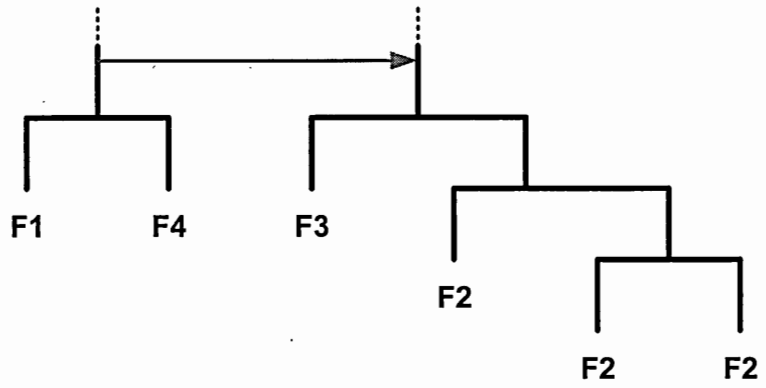


Figure 4.1 Example of a horizontal gene transfer event involving alleles belonging to species of four different families (F1, F2, F3 and F4).

Given the HGT weights calculated for individual transfers, we computed the HGT-related statistics. First, the HGT rates between families (see Figures 4.2, 4.3 and Tables 4.2, 4.3) were calculated as follows:

$$HGT(F2 \rightarrow F1) = \frac{1}{G(F1,F2)} \sum_{g=1}^{G(F1,F2)} \left(\frac{1}{N_{F1}(g) \times N_{F2}(g)} \sum_{i=1}^{N_{F2 \rightarrow F1}(g)} W_{F2 \rightarrow F1}(g,i) \right), \quad (4.1)$$

where $HGT(F2 \rightarrow F1)$ is the HGT rate for the alleles belonging to species of family $F1$ that were affected by gene transfers from alleles of family $F2$, $W_{F2 \rightarrow F1}(g,i)$ is the weight of the i^{th} HGT from $F2$ to $F1$ found for gene (i.e., multiple sequence alignment) g, $N_{F2 \rightarrow F1}(g)$ is the total number of detected HGTs for gene g that stemmed from alleles belonging to species of family $F2$ and affected species of family $F1$, $N_{F1}(g)$ and $N_{F2}(g)$ are the total numbers of alleles belonging to species of family $F1$ and $F2$, respectively, and $G(F1,F2)$ is the number of genes (i.e., multiple sequence alignments) containing at least one allele of family $F1$ and one allele of family $F2$.

Second, the non-normalized HGT rate between the alleles belonging to species of families $F2$ and $F1$ (from $F2$ to $F1$) was calculated as follows:

$$HGT_{nn}(F2 \rightarrow F1) = \sum_{g=1}^{G(F1,F2)} \left(\sum_{i=1}^{N_{F2 \rightarrow F1}(g)} W_{F2 \rightarrow F1}(g,i) \right), \quad (4.2)$$

where $HGT_{nn}(F2 \rightarrow F1)$ is the non-normalized HGT rate from $F2$ to $F1$.

The *local* intragroup HGT rates (Figures 4.2 and 4.3 on the diagonal) were computed as follows:

$$Intra\_HGT\_gene(F1) = \frac{1}{G(F1)} \sum_{g=1}^{G(F1)} \left(\frac{2}{N_{F1}(g) \times (N_{F1}(g)-1)} \sum_{i=1}^{N_{F1 \rightarrow F1}(g)} W_{F1 \rightarrow F1}(g,i) \right), \quad (4.3)$$

where $Intra\_HGT\_gene(F1)$ is the internal HGT rate for family $F1$ (i.e., this rate accounts for alleles belonging to species of family $F1$ that were affected by gene transfers from another alleles of this family), $W_{F1 \rightarrow F1}(g,i)$ is the weight of the i^{th} HGT from $F1$ to $F1$ found for gene (i.e., multiple sequence alignment) g, $N_{F1 \rightarrow F1}(g)$ is the total number of detected HGTs for gene g from alleles belonging to species of family $F1$ and affecting species of the same family, and $G(F1)$ is the number of genes (i.e., multiple sequence alignments) containing at least one allele of family $F1$.

The outgoing HGT rates (Figure 4.4) were computed as follows:

$$Outg\_HGT(F1) = \frac{1}{N_{F1}} \sum_{j=1,(F1 \neq Fj)}^{P} HGT_{nn}(F1 \rightarrow Fj), \quad (4.4)$$

where $Outg\_HGT(F1)$ is the outgoing HGT rate for family $F1$, representing the probability that an allele (or its part for the case of partial HGT) of a species of family $F1$ was transferred to a species from another prokaryotic family, $HGT_{nn}(F1 \rightarrow Fj)$ is the non-normalized HGT rate calculated according to Equation 4.2, N_{F1} is the total number of considered alleles belonging to species of family $F1$ (counted over all 110 MSAs), and P is the total number of considered prokaryotic families ($P = 23$ in our study).

The incoming HGT rates (Figure 4.5) were computed as follows:

$$Incom\_HGT(\mathrm{F1}) = \frac{1}{N_{F1}} \sum_{j=1,(Fj \neq F1)}^{P} HGT_{nn}(Fj \rightarrow F1), \tag{4.5}$$

where $Incom\_HGT(F1)$ is the proportion of alleles belonging to species of family $F1$ affected by HGT stemming from alleles belonging to species of the other prokaryotic families, $HGT_{nn}(Fj \rightarrow F1)$ is the non-normalized HGT rate calculated according to Equation 4.2.

The *global* intragroup HGT rates (Figure 4.6) were computed as follows:

$$Intra\_HGT(F1) = \frac{1}{N_{F1}} HGT_{nn}(F1 \rightarrow F1), \tag{4.6}$$

where $Intra\_HGT(F1)$ is the proportion of alleles belonging to species of family $F1$ that were affected by HGT stemming from the same prokaryotic family.

Finally, the average probability that a prokaryotic allele has been affected by HGT was computed as follows:

$$HGT_{average} = \frac{\sum_{i=1}^{P} (Incom\_HGT\ (Fi) + Intra\_HGT\ (Fi)) \times N_{Fi}}{\sum_{i=1}^{P} N_{Fi}} = \frac{\sum_{i=1}^{P} (Outg\_HGT\ (Fi) + Intra\_HGT\ (Fi)) \times N_{Fi}}{\sum_{i=1}^{P} N_{Fi}} \tag{4.7}$$

The average HGT rates for the set of all considered genes, as well as for its subset of core genes, are reported in Table 4.1a.

We studied ecological habitats using the same statistical measures and formulas as for the 23 phylogenetic families. The number of considered prokaryotic habitats, P, was equal to 8 in the habitat study.

The decomposed transfers, shown in Figure 4.1, were further weighted according to the species habitat membership (note that a species can live in more than one habitat; see Supplementary Table 2 for more details). Let us consider a possible case: an HGT from Allele A1, belonging to species X which is present in habitats H1 and H2, to Allele A2, belonging to species Y which is present only in habitat H1.

The weighted transfer from Allele A1 to Allele A2, with the weight $W(A1 \rightarrow A2)$, will be decomposed as follows:

$$W(A1 \rightarrow A2) = W(H1 \rightarrow H1) \oplus W(H2 \rightarrow H1);$$

$$W(H1 \rightarrow H1) = 0.5 \times W(A1 \rightarrow A2);$$

$$W(H2 \rightarrow H1) = 0.5 \times W(A1 \rightarrow A2).$$

The implementation details are available in Appendix B (see Formulas B.3 and B.4). These formulas and implementation ensures that the average HGT rates computed using Formula 4.7 are the same regardless the selected species classification (per phylogenetic family or per habitat; see Table 4.1a).

4.3.4 HGT time estimation

An important part of our study addresses the problem of estimating the age distribution of the identified complete and partial HGTs. We dated the inferred maximum likelihood gene trees using a Bayesian method implemented in B.E.A.S.T. v.1.7.5 with "beagle" library v.1.1.0 (Drummond and Rambaut 2007), and then compared the obtained results with those given by a semi-parametric method based on penalized likelihood implemented in TreePL v.1.0 (Sanderson 2002). Secondary constraints were applied to the tree nodes using genomic timescale from a well-known geological and phylogenetic study of prokaryotic evolution (Battistuzzi et al. 2004). Based on the available species tree, we established a list of 26 constraints representing prokaryotic groups and their most recent common ancestor (MRCA) nodes up to the roots of Bacteria and Archaea (see Supplementary Table 4). For each node of the species tree multiple values were available, each corresponding to the different root calibrations (3 root calibrations for Archaea and 4 for Bacteria; see (Battistuzzi et al. 2004). We aggregated the provided constraints in order to infer the mean age

and the mean standard deviation for each tree node, besides using the already available confidence intervals. For each gene tree, the corresponding nodes were found using MRCAs of the present strains. When multiple constraints existed for the same gene tree node, we sorted them according to the mean time and chose the most recent one. This generally corresponds to narrower and hence more precise group classifications, given that dating methods are less precise for remote geological events. Due to the mechanisms of reticulate evolution, such as HGT and recombination, and to the absence of representatives of certain families in the gene trees, some incompatibilities could exist between the applied constraints. We verified the compatibility of each constraint using the time analysis. The constraints incompatibilities were treated differently in the TreePL and B.E.A.S.T. analyses. As TreePL uses only discrete constraints in the form of a time interval, we used the confidence intervals for this purpose. Using a greedy approach, we sorted these constraints in the ascending order based on the mean age and enabled them progressively, starting by the most recent ones. We eliminated the constraints that led to the execution errors in TreePL and reran the program. In the case of B.E.A.S.T., we used the available normal distribution information, defined by the mean time and the standard deviation. Due to the continuous nature of these probability functions, there was no incompatibility between the applied constraints.

As B.E.A.S.T. was only needed for the time estimation, and not for phylogenetic inference, we used similar parameters of nucleotide substitution as in the RAxML gene tree inference (GTR + Gamma) and disabled the tree operators. Gene trees were rooted using the HGT-Detection program (Boc et al. 2010), which selects the gene tree root in order to minimize the Robinson and Foulds topological distance between the species and gene trees. We scaled gene trees using a value of 4290 Mya for the last common ancestor (LCA) (Sheridan et al. 2003). Besides using this unique value as an initial starting point, we also bounded LCA between the origin of life on Earth and the origin of aerobic methanotrophy. This gave us a uniform prior of 2500-4500 Mya, which we used for the root calibration in the gene trees (we did not consider any other organisms apart from prokaryotes).

We used the inferred RAxML gene trees as input trees. The branch length distributions were obtained by fitting normal and lognormal distributions to the branch lengths of the gene trees by means of the R statistical language, v.2.15.1. (R Core Team 2014). We used the birth-death tree model (Gernhard 2008) and an uncorrelated relaxed clock (Drummond et al. 2006) with a lognormal distribution model with the parameters "ucld.mean" and "ucld.stdev" set to the mean and the standard deviation inferred previously from branch lengths. We also defined a lognormal prior

for "ucld.mean", based on the location and scale values inferred previously, and an exponential prior for "ucld.stdev". Markov Chain Monte Carlo (MCMC) algorithm with 20 million generations, burn-in of 5%, sampled each 10,000 iterations, was carried out. Tracer v.1.2 was used to evaluate the method's convergence and marginal density. Treeannotator with a posterior probability limit of 0.5 was used to transfer estimated tree node ages back to the original tree, with uncertainty in parameter estimates corresponding to the 95% highest probability density (HPD). We chose to apply a relaxed molecular clock model (Drummond et al. 2006) in order to address the difficulties in inferring a strict molecular clock when studying genes affected by HGT (Novichkov et al., 2004).

4.4 Results

4.4.1 Gene transfer rates in complete and ovreall HGT scenarios

The obtained average HGT rates are quite similar for the two considered sets of genes: ubiquitous genes that represent the entire set of 110 genes examined in this study and core genes that include 36 genes identified by (Charlebois and Doolittle 2004); also see Supplementary Table 3 which reports both sets of genes. For the 75% bootstrap threshold, our benchmark threshold in this study, the obtained mean rate of complete HGT was about 3% per allele, whereas the overall (complete + partial) mean HGT rate was about 8% per allele (see Table 4.1a). Note that most of the existing studies focus on complete and recent HGTs only (Smillie et al. 2011). The mean HGT rates in Table 4.1a are indicated for the three following HGT bootstrap thresholds: 90%, 75% and 50%. Obviously, the mean HGT rates increase as the value of the bootstrap threshold decreases (low threshold values can lead to the inclusion of more conflicting or erroneous transfers) for both complete and overall HGT scenarios. Our findings suggest that core genes are slightly less prone to complete HGT than ubiquitous genes. However, somewhat surprisingly, core genes show more partial HGTs. Although complete HGTs have already been found at the heart of the ribosome (Brochier et al. 2000), partial gene transfers would be more likely to overcome the constraints imposed by the complexity theory (Jain et al. 1999). Mention that the obtained complete HGT rates are compatible with those found by (Dagan and Martin 2007).

Table 4.1a. Mean HGT rates, indicated for 100 comparisons, for complete and overall (complete + partial) HGT scenarios and three different bootstrap thresholds 90%, 75% and 50%.

Ubiquitous genes represent the entire set of 110 genes considered in this study. Core genes (assumed to be more resistant to HGT) include 36 genes identified by (Charlebois and Doolittle 2004).

Gene set	Complete HGT [90%, 75%, 50%]	Overall HGT [90%, 75%, 50%]
Ubiquitous (110)	[1.545, 2.944, 6.216]	[3.585, 8.066, 25.949]
Core (36)	[1.471, 2.653, 5.969]	[3.728, 8.215, 28.349]

For each considered gene, represented by the corresponding multiple sequence alignment, and for the three selected bootstrap thresholds of 90%, 75% and 50%, we also identified the exact number of genes that have been affected at least once during their evolutionary history by the complete and overall HGTs (see Table 4.1b). The results presented in this table confirm that there is no major difference in the number of the HGT-affected genes between the core genes and all the genes, whereas the core genes seem to be more prone to partial HGTs than the ubiquitous genes. The presented statistics also suggest that a large majority of genes have undergone multiple HGT events during their evolutionary history. Thus, our results show that although HGT events are rather rare at the allele level, their impact at the gene level is very significant.

Table 4.1b. Percentages of genes affected by at least one HGT during their evolutionary history, indicated for complete and overall (complete + partial) HGT scenarios and three different bootstrap thresholds 90%, 75% and 50%.

Ubiquitous genes represent the entire set of 110 genes considered in this study. Core genes (assumed to be more resistant to HGT) include 36 genes identified by (Charlebois and Doolittle 2004).

Gene set	Complete HGT [90%, 75%, 50%]	Overall HGT [90%, 75%, 50%]
Ubiquitous (110)	[64.50, 82.70, 96.30]	[85.40, 96.30, 100]
Core (36)	[66.66, 80.55, 97.22]	[94.44, 100, 100]

4.4.2 General overview of patterns of complete and overall HGT scenarios for the phylogenetic family study

Figures 4.2 and 4.3 present the intensity of transfers between the source and destination families for complete and overall HGTs, respectively. Here, only the results for the 75% HGT bootstrap threshold are described in detail. Mention that similar trends were observed for the two other HGT bootstrap thresholds we considered (i.e. 50% and 90%). Even though the intensity of overall, and partial, HGTs is higher than that of complete HGTs, the corresponding hit maps share most of the displayed intensity patterns (see Figures 4.2-4.3 and Tables 4.2-4.3).

First, we can notice that the intragroup HGT intensities (see the main diagonal in Tables 4.2-4.3) are usually higher than intergroup intensities for both complete and overall transfers. Although HGT-related clusters of prokaryotic families are not very clearly defined, we can observe two meta-clustering with the Archaea and Proteobacteria groups, including more transfers within each of these groups than between them. These patterns are noticeable for both complete and overall HGTs (Figures 4.2-4.3). They are more perceptible for higher bootstrap confidence levels (i.e. 75% bootstrap level - presented results and 90% bootstrap level - results not presented here). Surprisingly, transfers among phylogenetically close prokaryotic families are not necessarily well supported. On the contrary, several evolutionary remote prokaryotes show transfer affinities, as for example Spirochaetes and Thermotogae, or Crenarchaeota and Aquificae. The other closely interacting prokaryotic families are Thermotogae / Epsilonproteobacteria and Planctomycetes / Verrucomicrobia. Interacting families generally show reciprocal, but rather asymmetrical transfer intensity. This trend can be observed for both complete and overall HGTs (Figures 4.2-4.3).

Cyanobacteria, for example, exhibit much higher intra vs. intergroup HGT rate, what confirms the results of the previous studies (Zhaxybayeva et al. 2006). Some others families exhibiting similar behavior are Alphaproteobacteria, Betaproteobacteria, Bacteroidetes/Chlorobi and Actinobacteria (see Figures 4.3c and 4.2c). We also found that the Firmicutes family is the top groups in terms of the intragroup global HGT rate.

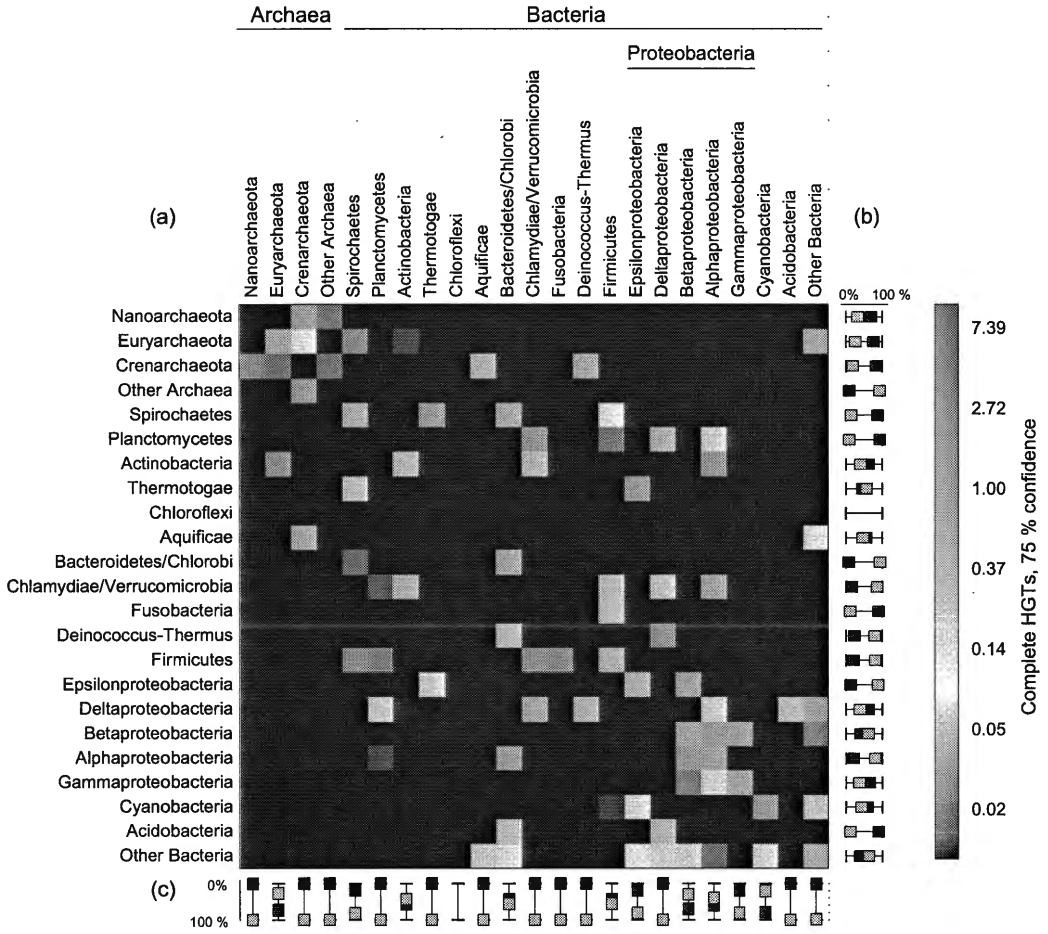


Figure 4.2 Complete HGT rates among prokaryotic phylogenetic groups, indicated for 100 comparisons, obtained for 75% bootstrap confidence level. This hit map corresponds to the results from Table 4.2.

a) HGT source group is represented by row (left) and HGT destination group is represented by column (top). Color scale on the right is natural log scale; b) Ratio of source (black squares) vs. destination (gray squares) HGT rate for a phylogenetic group indicated on the left; c) Intra- (black squares) vs. intergroup (gray squares) ratio for a phylogenetic group indicated on top.

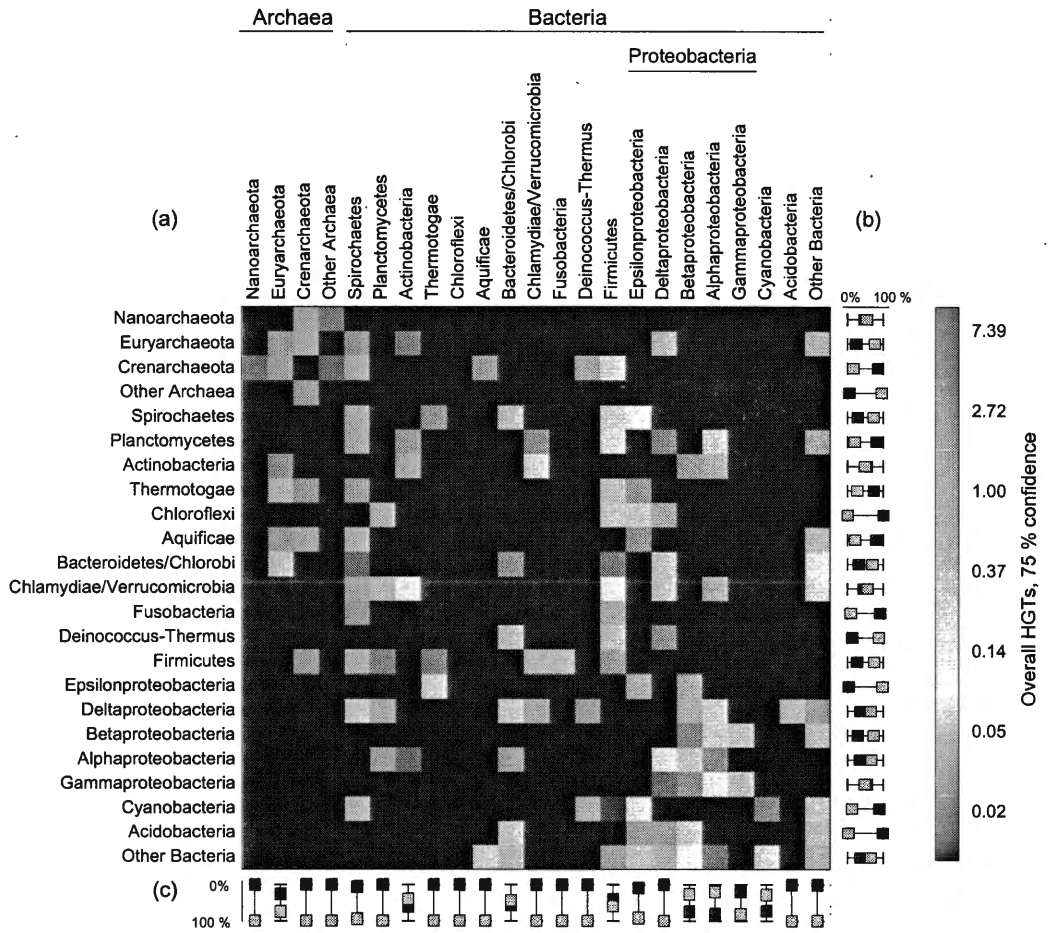


Figure 4.3 Overall (complete + partial) HGT rates among prokaryotic phylogenetic groups, indicated for 100 comparisons, obtained for 75% bootstrap confidence level. This hit map corresponds to the results from Table 4.3.

a) HGT source group is represented by row (left) and HGT destination group is represented by column (top). Color scale on the right is natural log scale; b) Ratio of source (black squares) vs. destination (gray squares) HGT rate for a phylogenetic group indicated on the left; c) Intra- (black squares) vs. inter-group (gray squares) ratio for a phylogenetic group indicated on top.

Table 4.2 *Complete HGT* rates among prokaryotic phylogenetic groups for 75% bootstrap confidence level, indicated for 100 comparisons.

a) Source group is represented by row (left) and destination by column (top). Group cardinality in terms of the number of species is indicated between parentheses and the number of alleles (counted over all considered MSAs) in square brackets. Values with low statistical support (when the alleles of two involved phylogenetic groups were present together in less than or equal to 5% of MSAs) are highlighted in gray. 10 highest values for groups with high statistical support (when the alleles of two involved phylogenetic groups were present together in more than 5% of MSAs) are highlighted in red. Intragroup HGT rates are underlined. Incoming (Inc) and outgoing (Out) HGT rates are highlighted in dark green and green, respectively. Grand total, highlighted in violet, represents the average complete HGT rate among prokaryotes.

Group Name	\	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Out
Nanoarchaeota(1),[6]	1	0	0	0.83	7.42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4.95
Euryarchaeota(6),[410]	2	0	0.5	0.07	0	0.02	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.04	0.58
Crenarchaeota(3),[167]	3	3.91	0.02	0	8.32	0	0	0	0	0	0.34	0	0	0	0.59	0	0	0	0	0	0	0	0	0	2.41
Other Archaea(1),[6]	4	0	0	1.41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4.3
Spirochaetes(3),[303]	5	0	0	0	0	0.39	0	0	1.42	0	0	0.04	0	0	0	0.09	0	0.01	0	0	0	0	0	0	2.95
Planctomycetes(1),[129]	6	0	0	0	0	0	0	0	0	0	0	1.69	0	0	0	0.02	0	0.33	0	0.13	0	0	0	0	3.62
Actinobacteria(10),[1077]	7	0	0.02	0	0	0	0	0.24	0	0	0	0	0.05	0	0	0	0	0	0	0.03	0	0	0	0	0.56
Thermotogae(1),[84]	8	0	0	0	0	0.2	0	0	0	0	0	0	0	0	0	0	1.1	0.01	0	0	0	0	0	0	1.73
Chloroflexi(2),[192]	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Aquificae(1),[98]	10	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0	0	0	0	0.09	1
Bacteroidetes/Chlorobi(5),[508]	11	0	0	0	0	0.01	0	0	0	0	0	0.53	0	0	0	0.01	0	0	0	0.01	0	0	0.01	0.01	0.35
Chlamydiae/Verrucomicrobia(2),[148]	12	0	0	0	0	0	0.01	0.05	0	0	0	0	0	0	0	0.06	0	0.17	0	0.04	0	0	0	0	2.97
Fusobacteria(1),[19]	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.18	0	0	0	0	0	0	0	0	4.71
Deinococcus-Thermus(1),[89]	14	0	0	0.01	0	0	0	0	0	0	0	0.23	0	0	0	0	0	0.02	0	0	0	0	0	0	1.3
Firmicutes(21),[2506]	15	0	0	0	0	0.02	0.02	0	0	0	0	0	0.02	0.02	0	0.37	0	0	0	0	0	0	0	0	0.12
Epsilonproteobacteria(3),[286]	16	0	0	0	0	0	0	0	0.14	0	0	0	0	0	0	0	0.36	0.01	0.03	0	0	0	0	0	0.37
Deltaproteobacteria(3),[257]	17	0	0	0	0	0.01	0.15	0	0.01	0	0.01	0	0.04	0	0.36	0.01	0.01	0	0	0.07	0.01	0.01	0.06	0.97	2.56
Betaproteobacteria(7),[617]	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0.59	0.03	0.03	0	0	0.02	1.17
Alphaproteobacteria(11),[1126]	19	0	0	0	0	0	0.01	0.01	0	0	0	0.03	0	0	0	0	0	0	0.03	0.71	0.01	0	0	0	0.54
Gammaproteobacteria(21),[2131]	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0.02	0.07	0.04	0	0	0	0.27
Cyanobacteria(3),[348]	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0.11	0	0	0	0	1.12	0	0.06	0.75
Acidobacteria(1),[94]	22	0	0	0	0	0	0	0	0	0	0	0.21	0	0	0	0	0	0.3	0	0	0	0	0	0	1.94
Other Bacteria(3),[339]	23	0	0.01	0	0	0	0	0	0	0	0.16	0.12	0	0	0	0	0.1	0.14	0.06	0.01	0	0.12	0	0.05	2.1
	Inc	11.7	0.4	1.1	31.5	0.79	1.04	0.28	4.27	0	1.08	1.23	1.98	0.55	2.2	0.42	1.07	1.72	1.09	1.28	0.26	0.41	0.19	1.06	2.94

Table 4.3 Overall (complete + partial) HGT rates among prokaryotic phylogenetic groups for 75% bootstrap confidence level, indicated for 100 comparisons.

a) Source group is represented by row (left) and destination by column (top). Group cardinality in terms of the number of species is indicated between parentheses and the number of alleles (counted over all considered MSAs) in square brackets. Values with low statistical support (when the alleles of two involved phylogenetic groups were present together in less than or equal to 5% of MSAs) are highlighted in gray. 10 highest values for groups with high statistical support (when the alleles of two involved phylogenetic groups were present together in more than 5% of MSAs) are highlighted in red. Intragroup HGT rates are underlined. Incoming (Inc) and outgoing (Out) HGT rates are highlighted in dark green and green, respectively. Grand total, highlighted in violet, represents the average overall HGT rate among prokaryotes.

Group Name	\	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Out
Nanoarchaeota(1),[6]	1	0	0	0.83	7.42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4.95
Euryarchaeota(6),[410]	2	0	1.58	0.65	0	0.02	0	0.02	0	0	0	0	0	0	0	0	0	0.33	0	0	0	0	0	0.05	2.46
Crenarchaeota(3),[167]	3	9.46	1.35	0	18.1	0.52	0	0	0	0	1.17	0	0	0	0.59	0.07	0	0	0	0	0	0	0	0	9.15
Other Archaea(1),[6]	4	0	0	1.41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4.3
Spirochaetes(3),[303]	5	0	0	0	0	0.49	0	0	1.9	0	0	0.22	0	0	0	0.14	0.08	0.01	0	0	0	0	0	0	5.02
Planctomycetes(1),[129]	6	0	0	0	0	0.54	0	0.03	0	0	0	0	1.92	0	0	0.11	0	3.31	0	0.15	0	0	0	0.47	12.6
Actinobacteria(10),[1077]	7	0	0.02	0	0	0	0	0.53	0	0	0	0	0.07	0	0	0	0	0	0.03	0.04	0.01	0	0	0.01	0.94
Thermotogae(1),[84]	8	0	0.35	2.17	0	1.23	0	0	0	0	0	0	0	0	0	0.06	1.85	0.01	0	0	0	0	0	0	11.4
Chloroflexi(2),[192]	9	0	0	0	0	0	0.38	0	0	0	0	0	0	0	0	0.07	0.17	0.85	0	0	0	0	0	0	3.36
Aquificae(1),[98]	10	0	1.41	0.83	0	0.35	0	0	0	0	0	0	0	0	0	0	2.06	0.01	0	0	0	0	0	0.26	6.2
Bacteroidetes/Chlorobi(5),[508]	11	0	0.21	0	0	0.01	0	0	0	0	0	2.15	0	0	0	0.02	0	0.19	0	0.01	0	0	0.01	0.07	1.35
Chlamydiae/Verrucomicrobia(2),[148]	12	0	0	0	0	0.81	0.33	0.09	0	0	0	0	0	0	0	0.08	0	0.2	0	0.04	0	0	0	0.13	6.02
Fusobacteria(1),[19]	13	0	0	0	0	1.85	0	0	0	0	0	0	0	0	0	0.64	0	0	0	0	0	0	0	0	21.1
Deinococcus-Thermus(1),[89]	14	0	0	0.01	0	0	0	0	0	0	0	0.23	0	0	0	0.06	0	0.02	0	0	0	0	0	0	2.45
Firmicutes(21),[2506]	15	0	0	0.03	0	0.04	0.02	0	0.02	0	0	0.01	0.05	0.26	0	1.18	0	0	0	0	0	0	0	0.01	0.42
Epsilonproteobacteria(3),[286]	16	0	0	0	0	0	0	0	0.14	0	0	0	0	0	0	0	0.54	0.01	0.04	0	0	0	0	0	0.46
Deltaproteobacteria(3),[257]	17	0	0	0	0	0.17	0.83	0	0.01	0	0.01	0.07	0.04	0	1.14	0.01	0.01	0	0.05	0.07	0.01	0.01	0.06	1.17	5.43
Betaproteobacteria(7),[617]	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	1.9	0.07	0.06	0	0	0.06	2.17
Alphaproteobacteria(11),[1126]	19	0	0	0	0	0	0.03	0.01	0	0	0	0.03	0	0	0	0	0	0.1	0.06	2.88	0.01	0	0	0	1.24
Gammaproteobacteria(21),[2131]	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0.03	0.08	0.05	0	0	0	0.34
Cyanobacteria(3),[348]	21	0	0.01	0	0	0.32	0.01	0.01	0	0	0	0	0	0	0.38	0.01	0.11	0.01	0.01	0.01	0.01	3.06	0.01	0.17	1.92
Acidobacteria(1),[94]	22	0	0	0	0	0	0	0	0	0	0	0.22	0	0	0	0	1.09	1.08	0.19	0	0	0	0	0.27	7.46
Other Bacteria(3),[339]	23	0	0.01	0	0	0	0	0.01	0	0	0.16	0.4	0	0	0	0.03	0.21	0.43	0.08	0.02	0	0.14	0	0.05	5.65
	Inc	28.4	2.37	5.64	52.5	5.09	4.06	0.51	6.03	0	2.1	2.72	2.98	5.82	5.66	1.21	3.57	9.52	2.36	1.69	0.52	0.48	0.21	3.42	8.07

4.4.3. Source and destination species most commonly affected by HGT

Common sources and destinations of the obtained complete and overall HGTs are not uniformly distributed among the species of the 23 prokaryotic families considered in our study (see Figures 4.4-4.6). Figures 4.4a (destination HGT) and 4.5a (source HGT) present the percentage of transfers that originated from the prokaryotic families whose representatives appeared in at least 5% of the multiple sequence alignments (i.e. genes) examined here. The obtained results show some discrepancies between the complete and overall HGT scenarios, as well as between the intragroup and intergroup relationships. For example, Firmicutes were found to be among the lowest ranking groups in terms of both outgoing and incoming HGT rates (Figures 4.4a and 4.5a), in a strong contrast to the top position they occupy in the intragroup ranking (Figure 4.6a).

Other important trends which can be observed in Figures 4.4-4.6 are the following. Fusobacteria are by far the top HGT donors in both partial and complete scenarios, Deltaproteobacteria being the top HGT receivers for partial HGT. These families exhibit a very asymmetric behavior, as they rank much lower in the reverse direction (i.e. receiver vs. donor). Some other prokaryotic families sharing asymmetric behavior are Crenarchaeota and Planctomycetes. Families more symmetric in respect to the direction of transfers are Betaproteobacteria and Euryarchaeota.

Figures 4.4b-4.6b present the results for the two families, Nanoarchaeota and Other Archaea, whose representatives are rarely present in the examined multiple sequence alignments (they are present in less than 5% of MSAs). Thus, the average HGT rates obtained for these two families are less significant from the statistical point of view, even though some important incoming HGT rates were obtained for both of them (see Figure 4.5b). Note that the alleles of the species belonging to the Nanoarchaeota family are the origin of transfer in only 4% of the considered MSAs.

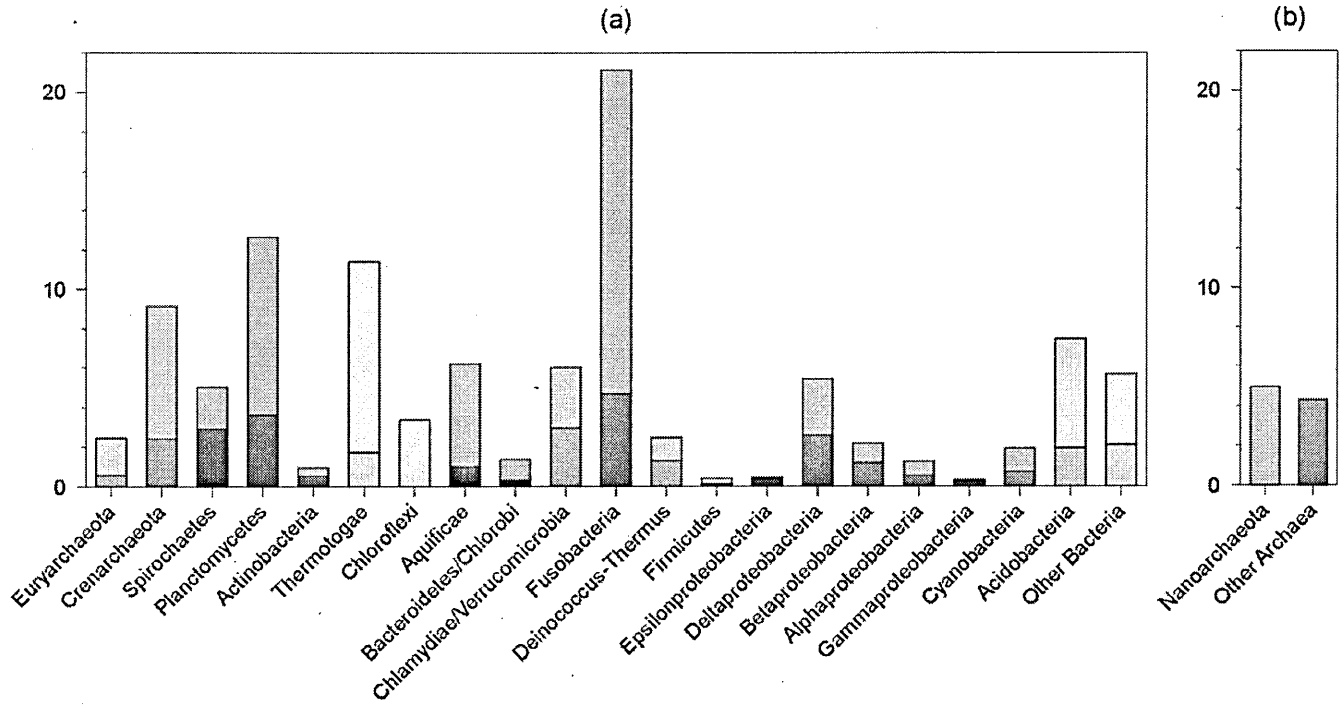


Figure 4.4 Overall outgoing HGT rates obtained for prokaryotic phylogenetic groups for 75% bootstrap confidence level, indicated for 100 comparisons, including complete and partial HGTs.

Lower parts of represented bars – in darker colors – depict *complete HGT*. Total height bars - including both darker and lighter colors - depict *partial HGT*. a) Groups with high statistical support (when the group alleles were present in more than 5% of MSAs); b) Groups with low statistical support (when the group alleles were present in less than or equal to 5% of MSAs).

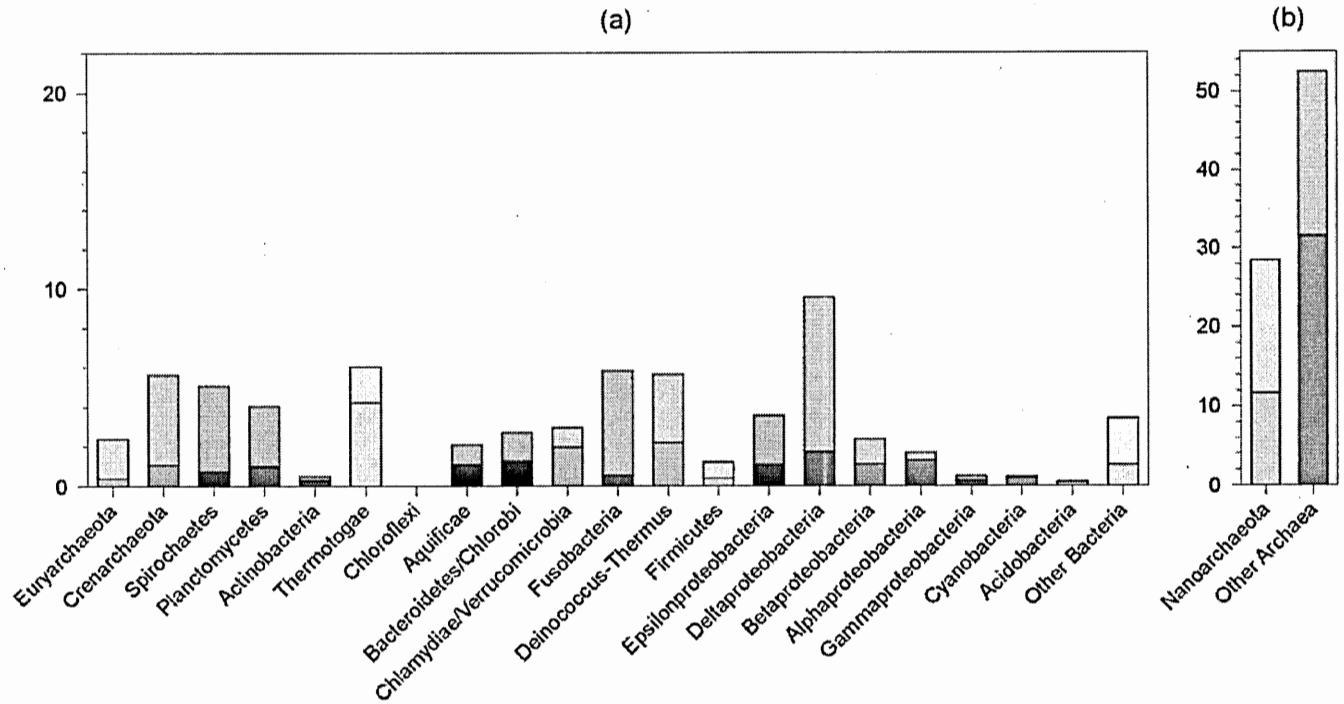


Figure 4.5 Overall incoming HGT rates obtained for prokaryotic phylogenetic groups for 75% bootstrap confidence level, indicated for 100 comparisons, including complete and partial HGTs.

Lower parts of represented bars – in darker colors – depict *complete HGT*. Total height bars - including both darker and lighter colors - depict *partial HGT*. a) Groups with high statistical support (when the group alleles were present in more than 5% of MSAs); b) Groups with low statistical support (when the group alleles were present in less than or equal to 5% of MSAs).

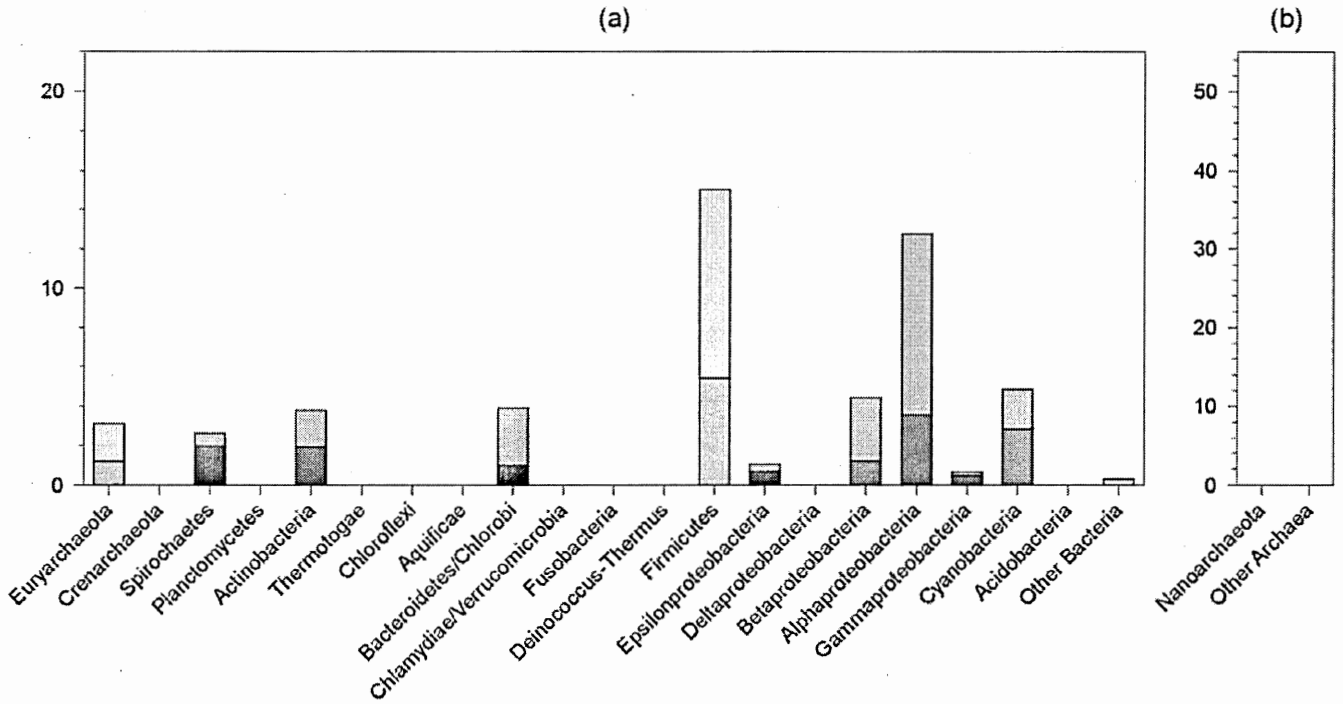


Figure 4.6 Overall global intragroup HGT rated obtained for prokaryotic phylogenetic groups for 75% bootstrap confidence level, indicated for 100 comparisons, including complete and partial HGTs.

Lower parts of represented bars – in darker colors – depict *complete HGT*. Total height bars - including both darker and lighter colors - depict *partial HGT*. a) Groups with high statistical support (when the group alleles were present in more than 5% of MSAs); b) Groups with low statistical support (when the group alleles were present in less than or equal to 5% of MSAs).

4.4.4 Ten most frequent horizontal gene transfer patterns among prokaryotes

Here, we also present the ten most frequent horizontal gene transfers among phylogenetic families for each selected bootstrap level. This has been done separately for complete and overall HGTs. The most significant transfers are mapped into the phylogenetic tree of 111 prokaryotic species (see Figures 4.7-4.8). Circular tree views were selected for this presentation. We put together all of the 10 most significant transfers obtained for the 50%, 75% and 90% bootstrap thresholds. This resulted in 18 distinct transfers for the complete HGT (Figure 4.7) and 16 distinct transfers for the overall HGT (Figure 4.8). The large majority of them (i.e. 13) were shared between both HGT scenarios, but obviously at different HGT rates. One of them was found in the reverse direction (i.e. from Crenarchaeota to Aquificae). Two of them shared the same source (from Crenarchaeota to Deinococcus-Thermus and to Euryarchaeota). Two others shared the same destination (from Acidobacteria and Chloroflexi to Deltaproteobacteria). Only two of them were completely different: the local intragroup complete transfers for Spirochaetes and Firmicutes.

The obtained results confirm that the intragroup HGTs are very important for the process of the prokaryotic diversification. A majority of the highly-ranked intragroup HGTs (i.e. top 7 out of 11 intragroup HGTs illustrated in Figure 4.6) are also ranked among the ten most frequent HGTs in general (see Figures 4.7-4.8). The only notable exception is Actinobacteria. Further analysis of its intragroup rates reveals lower local interactions (see Formula 4.3) and higher global ones (see Formula 4.6).

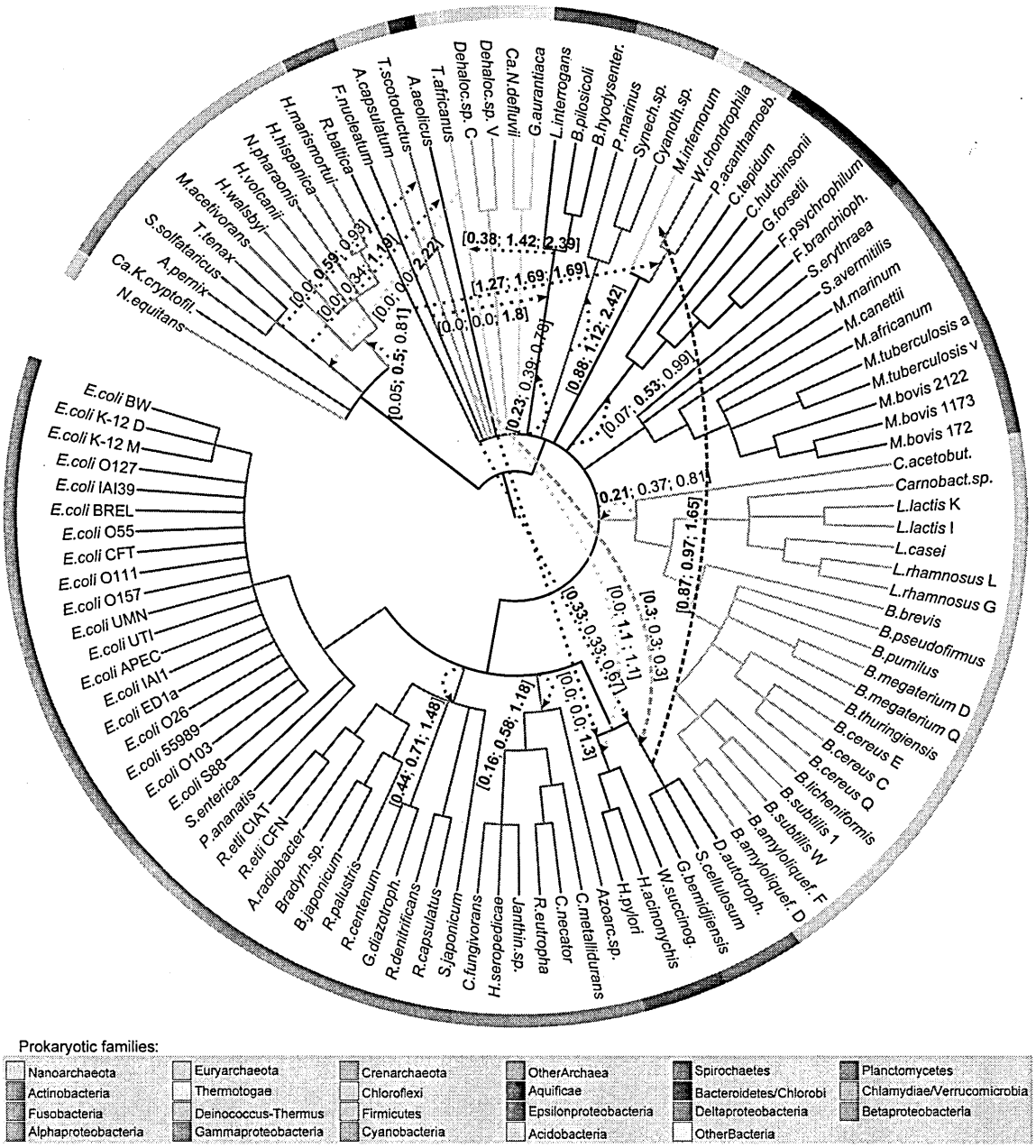
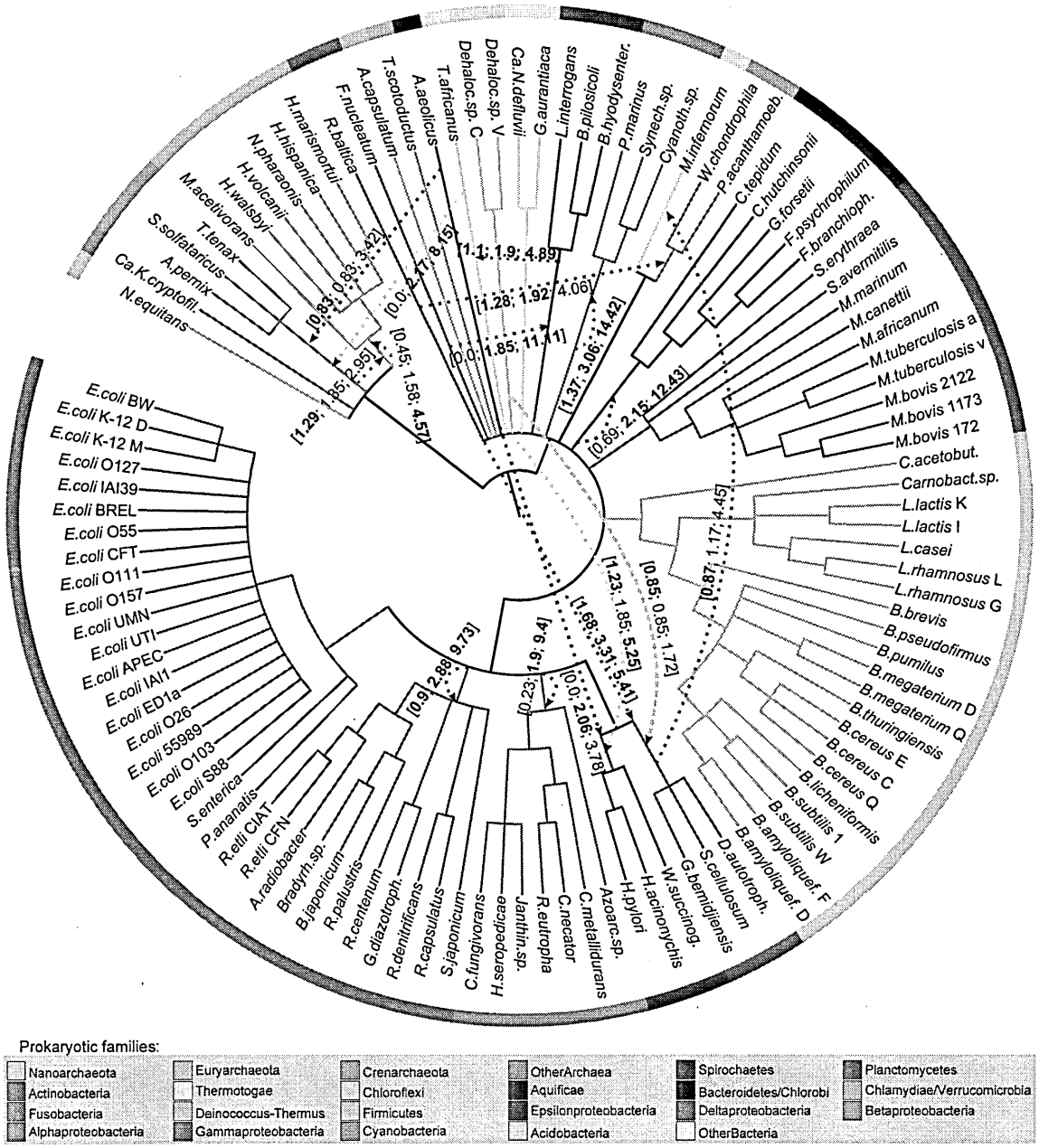


Figure 4.7 Phylogenetic network inferred for 111 prokaryotic species belonging to 23 different prokaryotic families, including 18 most significant *complete HGTs*.

Here, the HGT rate is given for each of the three following HGT bootstrap confidence levels: 90%, 75% and 50%. Interval format is: [90%, 75%, 50%]. Arrows are colored according to the HGT source group. Values are boldfaced when they belong to the top 10 list of the corresponding bootstrap confidence level.



Prokaryotic families:

Nanoarchaeota	Euryarchaeota	Crenarchaeota	OtherArchaea	Spirochaetes	Planctomycetes
Actinobacteria	Thermotogae	Chloroflexi	Aquificae	Bacteroidetes/Chlorobi	Chlamydiae/Verrucomicrobia
Fusobacteria	Deinococcus-Thermus	Firmicutes	Epsilonproteobacteria	Deltaproteobacteria	Betaproteobacteria
Alphaproteobacteria	Gammaproteobacteria	Cyanobacteria	Acidobacteria	OtherBacteria	

Figure 4.8 Phylogenetic network inferred for 111 prokaryotic species belonging to 23 different prokaryotic families, including 16 most significant *overall HGTs*.

Here, the HGT rate is given for each of the three following HGT bootstrap confidence levels: 90%, 75% and 50%. Interval format is: [90%, 75%, 50%]. Arrows are colored according to the HGT source group. Values are boldfaced when they belong to the top 10 list of the corresponding bootstrap confidence level.

4.4.5 General overview of patterns of complete and overall HGT scenarios for the habitat study

In this work, we also extend the study of Smillie et al. (2011) on defining the clusters of habitats associated to complete and overall HGTs (see Figures 4.9-4.12 and Tables 4.4 and 4.5). We can observe a wide range of habitats involved in HGT events for the overall HGT scenarios showing the presence of interaction between all the habitats, except the Human respiratory habitat (see Figures 4.10). For instance, Marine shows an exchange of genetic material with the Animal and Soil habitats only for the overall HGT scenarios. The symmetrical aspect of the presented hit maps in both scenarios can be observed. However, the evidence of mutual exchange of genetic material within the cluster of Human others, Plant, Animal and Soil habitats, first, as well as within the cluster of Marine, Fresh water and Extreme habitats, second, is also clearly visible. This finding is coherent with two classes of habitat relationships of the water and non-water-related habitats. Mention that these two habitat HGT interaction clusters are much more clearly defined compared to the phylogenetic family interaction clusters, underlining the paramount role played by ecological habitats in shaping HGT patterns.

Detailed analysis of Figures 4.9 and 4.10 suggests that the Human respiratory habitat involves a group of species that mutually exchange genetic material at a much higher rate than they do it with the species from the other habitats. This could be related to the fact that the only mechanism of acquisition of genetic material for this habitat is through the air, which have less probability to happen in the exchange with the solid-based habitats. This trend is opposite to the Human others habitat, which entails intestinal and skin host species having the propensity to exchange genetic material with the other habitats through direct contact and food vectors. In fact, the species from the Human others habitat constitute the most frequent source as well as the most frequent destination of HGT events (see Figures 4.11 and 4.12).

A previous HGT-based study showed the existence of such a network connecting the human microbiome (Smillie et al. 2011), and underlined the role of ecology in its definition. Here, we extend those findings, revealing an even larger HGT-related cluster of habitats, comprising Humans (including digestive system), Plants, Animals and Soil. Prokaryotes colonizing human respiratory system apparently have their own HGT-related habitat. This habitat has the lowest global HGT rate, totaling the outgoing, incoming and intragroup HGTs, thus suggesting that the immediately acquired advantages (i.e. new genes allowing species to survive in the changing environment) could

be also rapidly lost. Moreover, we can observe the existence of another HGT-related cluster of habitats, constituted by the three water-based environments, Marin, Fresh water and Extreme. This cluster is particularly well separated from the other environments in the case of the complete HGTs (Figure 4.9). For overall gene transfer scenarios, the water-based habitat tends to merge with the other prokaryotic habitats (Figure 4.10).

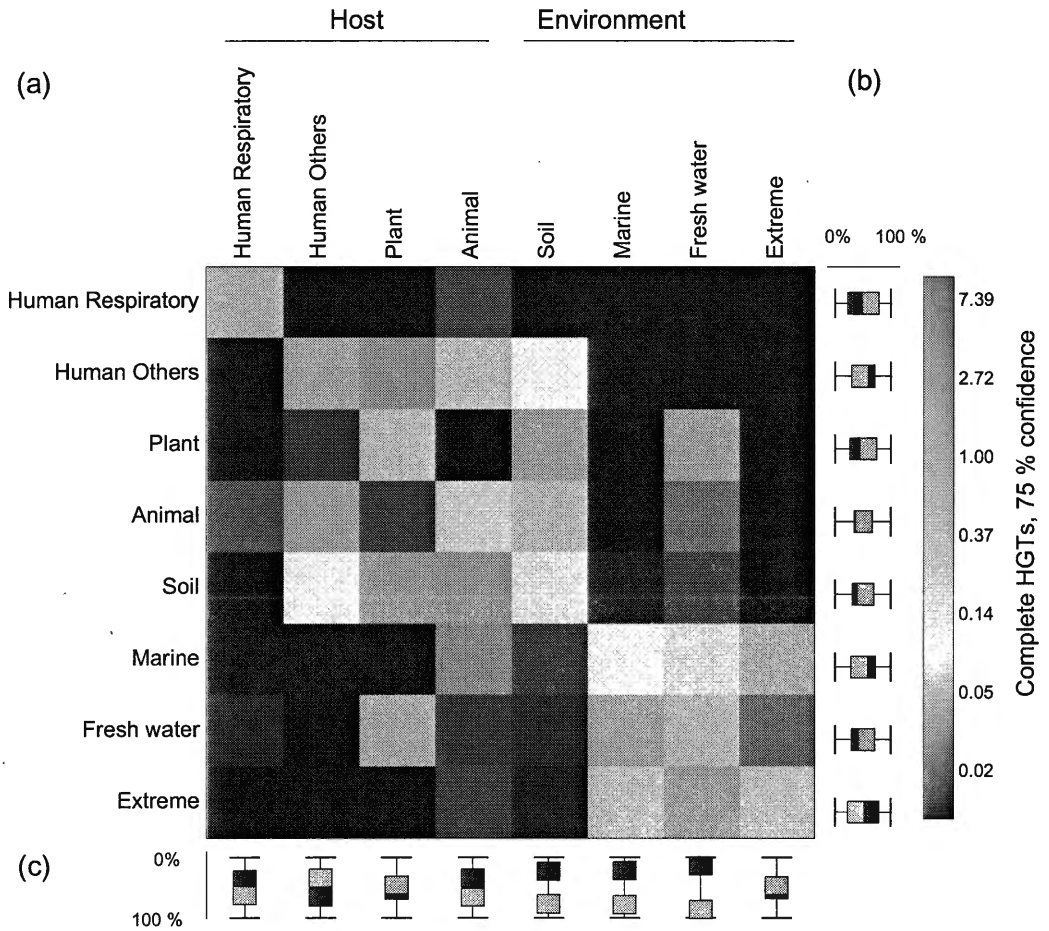


Figure 4.9 Complete HGT rates among prokaryotic habitats (indicated for 100 comparisons) obtained for 75% bootstrap confidence level. This hit map corresponds to the results from Table 4.4.

a) HGT source group is represented by row (left) and HGT destination group is represented by column (top). Color scale on the right is a natural log scale; b) Ratio of source (black squares) vs. destination (gray squares) HGT rate for a habitat indicated on the left; c) Intra- (black squares) vs. inter-group (gray squares) ratio for a habitat indicated on top; Habitat called Human Others includes Digestive and Urogenital habitats.

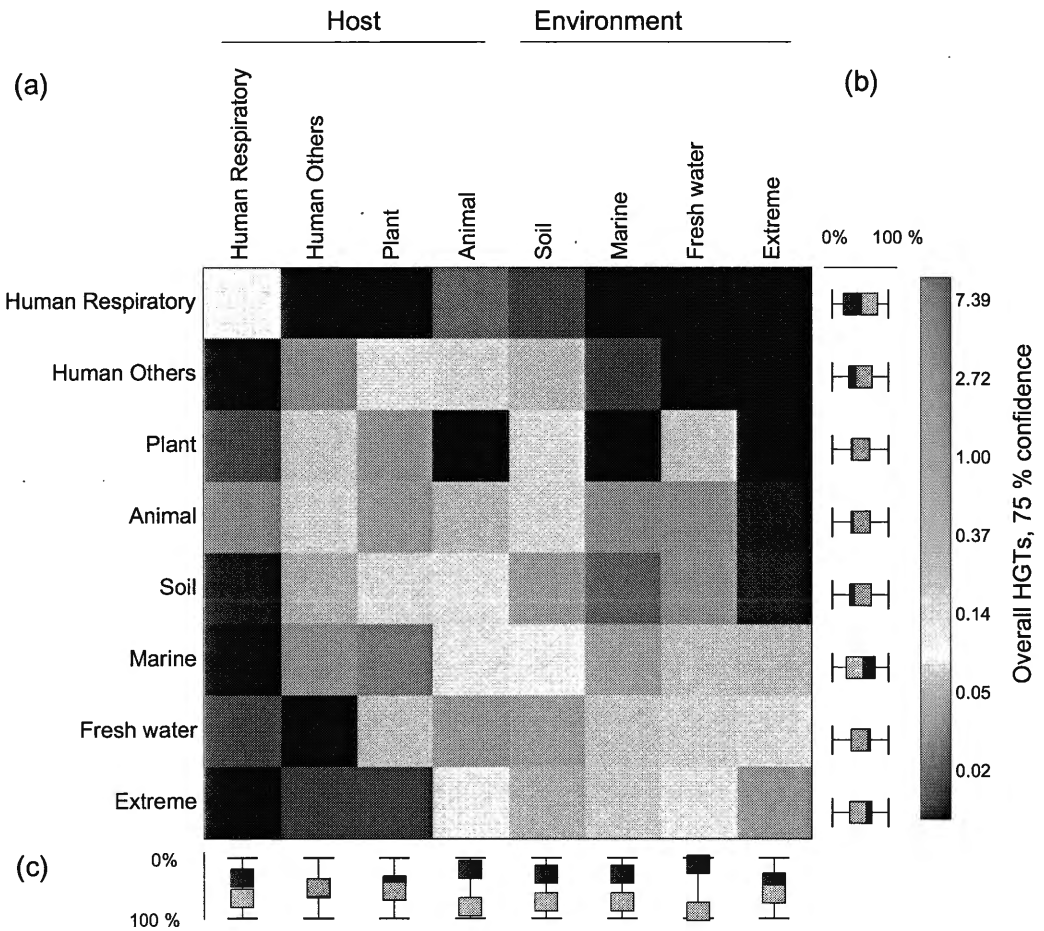


Figure 4.10 Overall (complete + partial) HGT rates among prokaryotic habitats (indicated for 100 comparisons) obtained for 75% bootstrap confidence level. This hit map corresponds to the results from Table 4.5.

a) HGT source group is represented by row (left) and HGT destination group is represented by column (top). Color scale on the right is a natural log scale; b) Ratio of source (black squares) vs. destination (gray squares) HGT rate for a habitat indicated on the left; c) Intra- (black squares) vs. inter-group (gray squares) ratio for a habitat indicated on top; Habitat called Human Others includes Digestive and Urogenital habitats.

Table 4.4 Complete HGT rates among prokaryotic habitats for 75% bootstrap confidence level, indicated for 100 comparisons.

a) Source group is represented by row (left) and destination by column (top). Group cardinality in terms of the number of species is indicated between parentheses and the number of alleles (counted over all considered MSAs) in square brackets; 10 highest values are highlighted in red. Intragroup HGT rates are underlined. Incoming (Inc) and outgoing (Out) HGT rates are highlighted in dark green and green, respectively. Grand total, highlighted in violet, represents the average complete HGT rate among prokaryotes.

Group Name	\	1	2	3	4	5	6	7	8	Out
Human Respiratory (25),[2365]	1	0.05	0	0	0.02	0.01	0	0	0	0.48
Human Others (15),[1082]	2	0.01	0.68	0.03	0.06	0.09	0.01	0.01	0.01	3.13
Plant (11),[992]	3	0.01	0.01	0.3	0	0.04	0	0.04	0	2.09
Animal (11),[907]	4	0.02	0.04	0.01	0.17	0.05	0.01	0.02	0	3.07
Soil (34),[2592]	5	0.01	0.09	0.04	0.04	0.12	0.01	0.02	0	2.08
Marine (13),[866]	6	0.01	0	0.01	0.03	0.01	0.09	0.07	0.05	2.19
Fresh water (29),[1767]	7	0.01	0	0.05	0.01	0.01	0.05	0.06	0.02	1.63
Extreme (12),[370]	8	0	0	0.01	0.02	0.01	0.06	0.04	0.25	2
	Inc	0.75	2.78	2.69	2.68	2.15	1.56	1.74	1.12	2.94

Table 4.5 Overall (complete + partial) HGT rates among prokaryotic habitats for 75% bootstrap confidence level, indicated for 100 comparisons.

a) Source group is represented by row (left) and destination by column (top). Group cardinality in terms of the number of species is indicated between parentheses and the number of alleles (counted over all considered MSAs) in square brackets; 10 highest values are highlighted in red. Intragroup HGT rates are underlined. Incoming (Inc) and outgoing (Out) HGT rates are highlighted in dark green and green, respectively. Grand total, highlighted in violet, represents the average complete HGT rate among prokaryotes.

Group Name	\	1	2	3	4	5	6	7	8	Out
Human Respiratory (25),[2365]	1	0.08	0	0.01	0.03	0.02	0	0	0	0.9
Human Others (15),[1082]	2	0.01	1.19	0.11	0.13	0.22	0.02	0.01	0.01	7.9
Plant (11),[992]	3	0.02	0.17	0.8	0.01	0.12	0.01	0.16	0.01	6.8
Animal (11),[907]	4	0.04	0.13	0.05	0.21	0.12	0.04	0.04	0.01	6.97
Soil (34),[2592]	5	0.01	0.27	0.12	0.1	0.47	0.03	0.04	0.01	5.98
Marine (13),[866]	6	0.01	0.04	0.03	0.08	0.08	0.39	0.18	0.2	6.8
Fresh water (29),[1767]	7	0.02	0.01	0.18	0.05	0.05	0.16	0.13	0.12	5.61
Extreme (12),[370]	8	0.01	0.02	0.02	0.08	0.06	0.15	0.11	0.52	6.03
	Inc	1.43	9.3	8.4	6.31	6.27	4.44	4.59	4.18	8.07

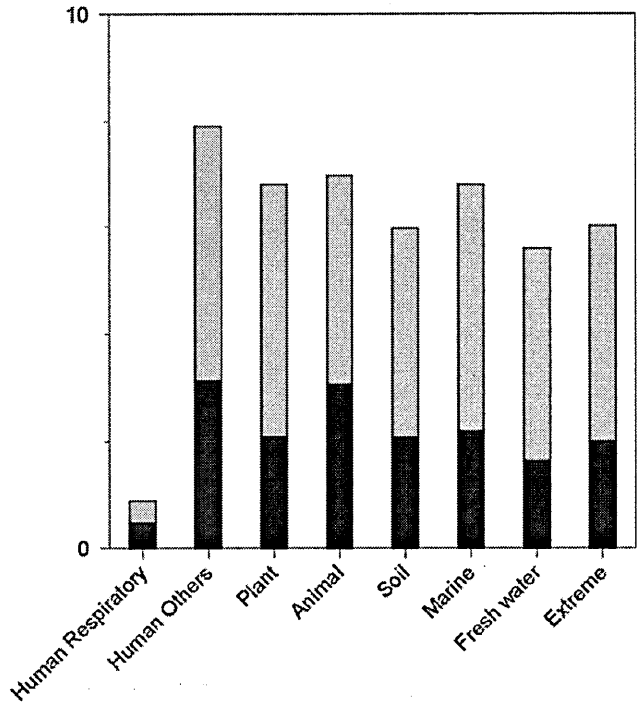


Figure 4.11 Overall outgoing HGT rates obtained for prokaryotic habitats for 75% bootstrap confidence level, indicated for 100 comparisons, including complete and partial HGTs.

Lower parts of represented bars – in darker colors – depict *complete HGT*. Total height bars - including both darker and lighter colors - depict *partial HGT*.

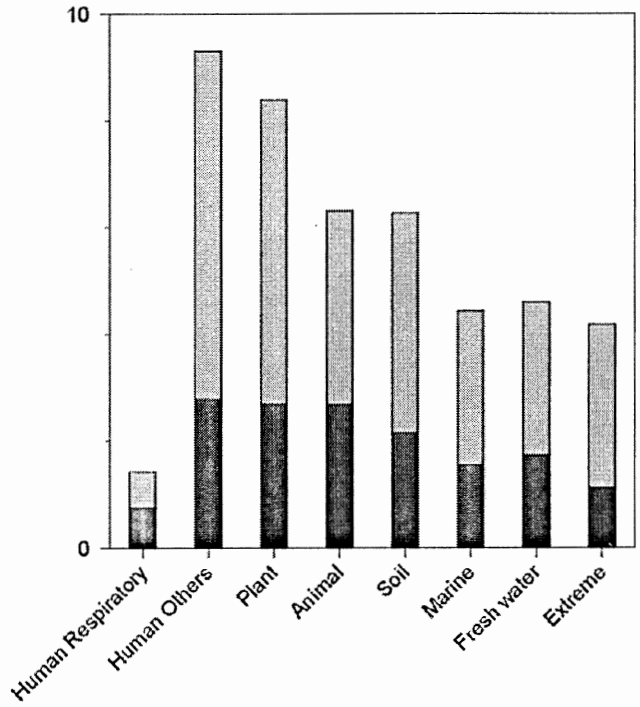


Figure 4.12 Overall incoming HGT rates obtained for prokaryotic habitats for 75% bootstrap confidence level, indicated for 100 comparisons, including complete and partial HGTs.

Lower parts of represented bars – in darker colors – depict *complete HGT*. Total height bars - including both darker and lighter colors - depict *partial HGT*.

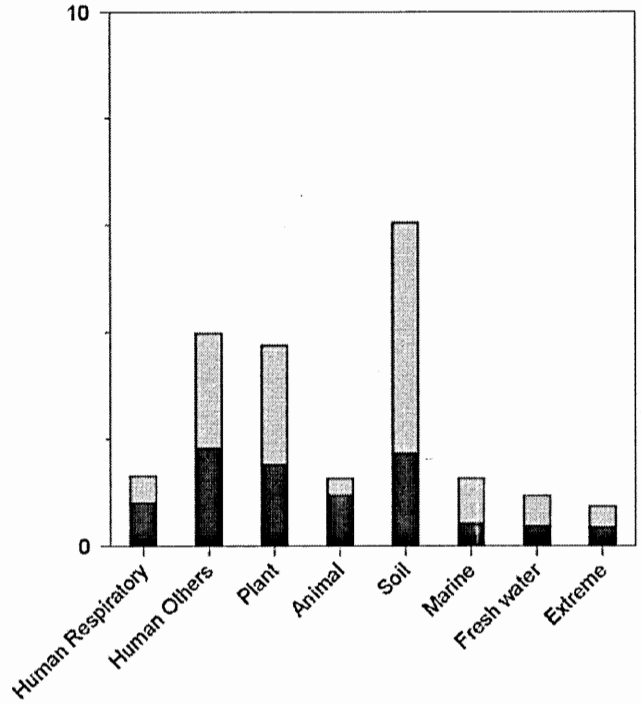


Figure 4.13 Overall global intragroup HGT rates obtained for prokaryotic habitats for 75% bootstrap confidence level, indicated for 100 comparisons, including complete and partial HGTs.

Lower parts of represented bars – in darker colors – depict *complete HGT*. Total height bars - including both darker and lighter colors - depict *partial HGT*.

4.4.6 Prediction of the HGT ages

Here, we discuss the results of our study aimed at the prediction of ages of the identified complete and overall HGT events (Figures 4.14-4.17). The experimental setup and the methods used in our analysis are described in Section 4.3.4 above. We compare the HGT age predictions obtained by two different maximum likelihood prediction methods (TreePL and B.E.A.S.T). The predictions were made for both complete and overall HGT events. We can observe a multimodal curve, corresponding to the general division of geological time (Figure 4.14). It shows a very low HGT rate during the Archaean period (before 2500 Mya), then a progressively higher HGT rate during the Proterozoic period (500 Mya – 2500 Mya), and finally a very high HGT rate during the most recent Phanerozoic period (under 500 Mya). These results suggest that partial transfers, at least as they can be detected by the modern HGT-Detection methods (Boc et al. 2012), are generally more recent than the complete ones. Our findings contrasts with another recent study, addressing the age recovery of complete gene transfers only, that argue that the HGT rate might be constant across the time scale (David and Alm 2011).

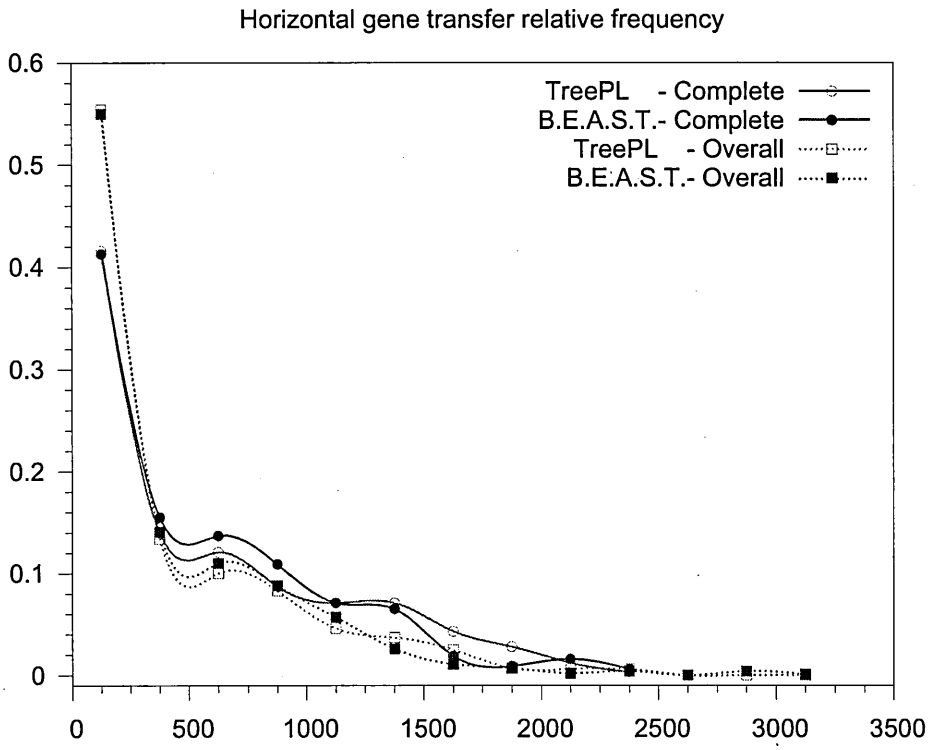


Figure 4.14 Frequency of complete (red and blue circles) and overall HGTs (red and blue squares) according to the time period.

Each represented value is drawn in the middle of the corresponding 250 Mya (million of years) time interval. The neighbor points are connected using natural smoothing *Gnuplot* splines. TreePL and B.E.A.S.T. software were used to infer both complete and overall HGT ages. The sum of all represented values for each of the four curves is 1.0.

Figures 4.15 and 4.16 compare the results obtained by using TreePL and B.E.A.S.T. Figure 4.15 illustrates the general traits of the time distribution of the detected complete (case a) and overall (case b) HGT events by using a boxplot representation. It shows that relative differences in the results provided by TreePL and B.E.A.S.T. are more important for the complete than for the partial HGTs. Mention that the central value of the presented distribution (i.e. its median value) is almost identical for the two time inference methods.

Figure 4.16 estimates the distributions obtained with TreePL and B.E.A.S.T by using Gaussian kernels. It also depicts the limits of the 95% High Probability Density interval of B.E.A.S.T (i.e. for the 5% and the 95% boundary). The TreePL curve is almost completely bound by this interval. Not being the optimal one (i.e. from the Bayesian point of view), the

TreePL estimation can be viewed as an acceptable outcome of B.E.A.S.T. Thus, the results provided by the two methods can be seen as compatible.

The curves in Figure 4.17 confirm the previous trends, while representing the correlation via a Quantile-Quantile plot. These results show very little divergence of the prediction of the transfer ages between the partial and overall HGT scenarios for the most recent period of 1000 Mya. Then, the two predictors diverge slightly until 2000 Mya. Older than the latter period, the results provided by TreePL and B.E.A.S.T. tend to be different. It is worth noting that more than three-quarters of the transfers fall into the most recent time interval (i.e. less than 1000 Mya). Mention than the latter result is consistent with the findings of (David and Alm 2011).

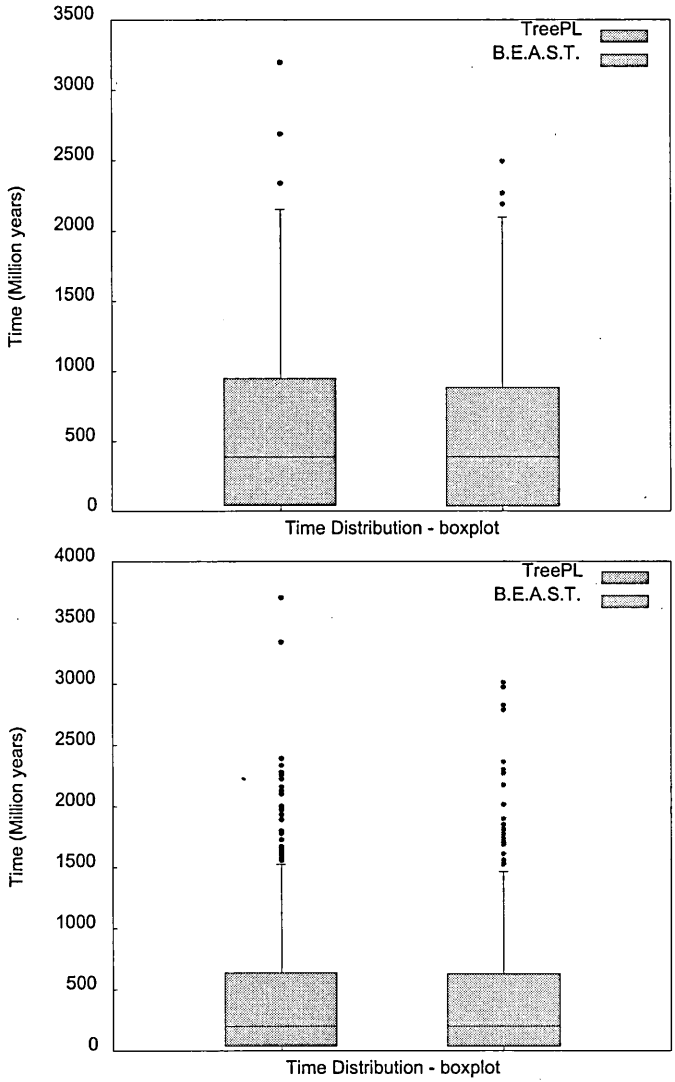


Figure 4.15 Boxplot of time distribution of the detected HGT events.
Time scale represents HGT ages in Mya (million of years).

a) Boxplot for complete HGT events; b) Boxplot for overall (complete + partial) HGT events.

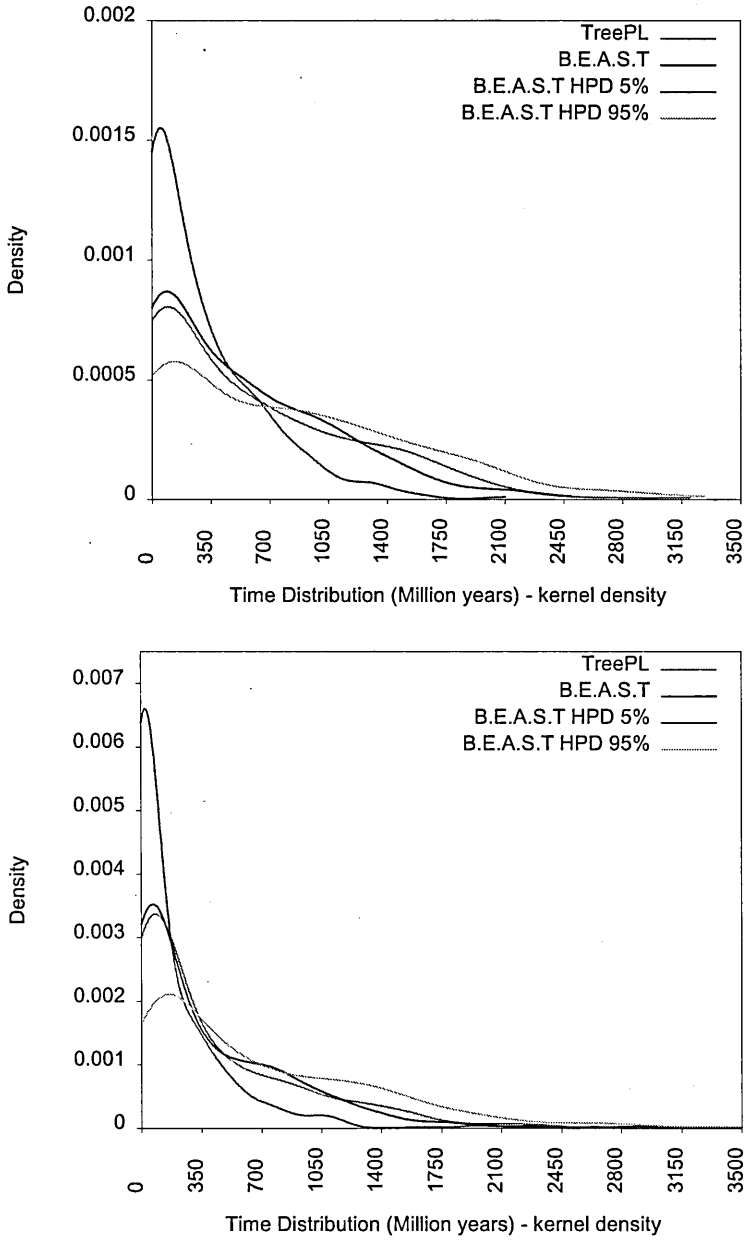


Figure 4.16 Gaussian kernel graphs of time distribution of the detected HGT events.
The curves represent: TreePL mean value, B.E.A.S.T. median value, B.E.A.S.T high probability density 5% and B.E.A.S.T high probability density 95%. a) Gaussian kernels for complete HGT events; b) Gaussian kernels for overall (complete + partial) HGT events.

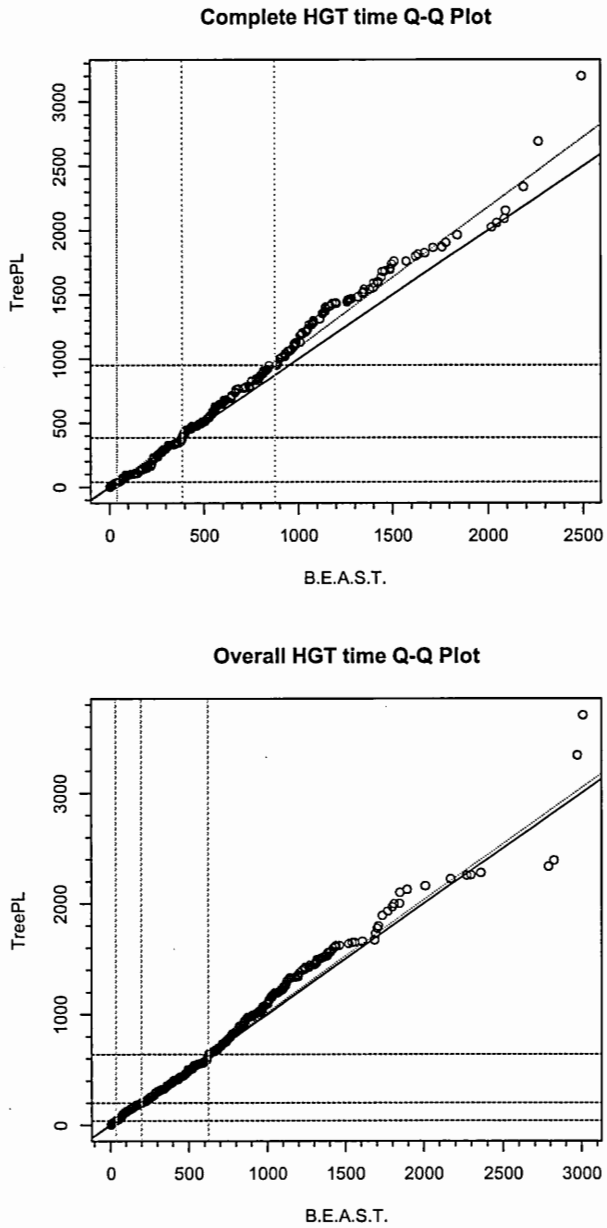


Figure 4.17 Q-Q (Quantile-Quantile) Plot of TreePL mean values vs. B.E.A.S.T. median values. Q-Q plot in which 25%, 50% and 75% percentiles are represented by dashed lines; interquantile lines are dotted and 45 degree lines are solid. a) Q-Q plot for complete HGT events; b) Q-Q plot for overall (complete + partial) HGT events.

4.5 Conclusion

In this chapter, we presented a comprehensive comparative study of complete and partial HGTs affecting the evolution of prokaryotic species. All the methods we applied in the framework of our analysis included a statistical validation step. We started by showing that the rate of HGT among core genes (Charlebois and Doolittle 2004) is generally comparable to that among ubiquitous genes (i.e. all the genes). Precisely, core genes undergo slightly more partial and slightly less complete HGTs than ubiquitous genes. Our results are generally compatible with previous analyses conducted for complete HGTs only (Ge et al. 2005); Smillie et al. 2011), showing that most of the prokaryotic genes have been affected by HGT multiple times during their evolutionary history, but for a reduced number of alleles. According to our estimation, the percentage of prokaryotic genes affected by at least one complete HGT during their evolutionary history varies between 64.5% (for the 90% HGT bootstrap threshold) and 96.3% (for the 50% HGT bootstrap threshold), while the percentage of genes affected by at least one overall (complete + partial) HGT varies between 85.4% (for the 90% HGT bootstrap threshold) and 96.3% (for the 100% HGT bootstrap threshold); see Table 4.1b.

Moreover, our findings suggest that depending on the selected bootstrap confidence level, the ratio between the overall and complete HGT rates is between 2.3 (for the 90% HGT bootstrap threshold) and 4.7 (for the 50% HGT bootstrap threshold) - see Table 4.1a. We highlighted the main differences in the HGT rates between the two scenarios for different groups of taxa. Thus, we showed that Archaea and Proteobacteria are the highest-level phylogenetic clusters regarding HGT. On the individual group level, the Firmicutes family is exhibiting a high intragroup and a very low intergroup HGT interactions. Most of the prokaryotic families show an asymmetric behavior in regards to the incoming and outgoing HGTs.

Furthermore, we depicted and compared the ten most common complete and partial HGTs (see Figures 4.7-4.8) characterizing the evolution of the selected set of the most frequent prokaryotic genes. The presented comparisons emphasize the fact that the complete HGT patterns are very similar to the partial ones, especially in the case of the family analysis (i.e. phylogenetic classification). In addition, we compared the patterns of complete and overall HGTs within different ecological habitats (i.e. ecological classification). The obtained results show some disagreement between the two HGT scenarios. For instance, unlike complete HGTs, overall (and respectively partial) HGTs favor a more reciprocal exchange of genetic material between prokaryotes. Except the Extreme habitat, all the other considered habitats include the species that share at least one HGT

with the species of any other considered ecological habitat. Two major ecological clusters of habitats, regarding HGT exchange, can be easily identified from the presented hit maps (see Figures 4.9-4.10): the cluster including Human others, Plant, Animal and Soil habitats and that including the Marine, Fresh water and Extreme habitats. Mention that our habitat findings are coherent with the results of Smillie et al. (2011).

Finally, the comparison of the ages of the inferred complete and partial HGTs underlines a high correlation between the results provided by the B.E.A.S.T. and TreePL methods for the HGTs falling within the most recent period of 1000 Mya; the two predictions diverge slightly until 2000 Mya; after this date, the results provided by TreePL and B.E.A.S.T. tend to be different. Our analysis also indicates that the ages of the recent HGTs can be predicted with a much higher confidence than those of the ancient ones.

Our results emphasize the importance of considering partial HGTs in the process of phylogenetic or ecological classification of prokaryotic species. A detailed study of overall HGT scenarios in prokaryotes, including both complete and partial HGTs, was one of the main original contributions of our work. We also showed that both complete and overall HGT rates are not the unique, and well-established, values but should be rather estimated by means of intervals of possible values. The boundaries of such intervals depend on the selected minimum and maximum HGT bootstrap acceptance thresholds (Boc et al. 2010). Thus, in the future, it would be important to design further techniques for statistical validation of the obtained complete and partial horizontal gene transfer events. For instance, it would be interesting to verify whether the flow of genetic material across habitats is not a reflection of the distribution of strains and their probability of exchanging genetic material. A hypergeometric test (Rice and John 2007), based on the p-value calculation, could be carried out for addressing this issue. However, such a computation should be very lengthy as the identification of complete and partial HGTs should be done for various data samples.

We have provided the complete source code of our application allowing one to carry out the methods for detecting and validating horizontal gene transfer events discussed in this chapter; the application's name is HGT-QFCLUST v.0.2. The related scripts written in the Python programming language have been also made available. The ReadMe documentation file provides an explanation of the main steps to follow for executing the application. The source code and the accompanying files have been uploaded to the GitHub public repository (with the BSD licence).

It is freely available at the following URL address:
https://github.com/dunarel/dunphd-thesis/tree/master/Chapter4/Main/linalgebra_impl.

We also supplied the original scripts allowing one to carry out the whole computational pipeline of the project presented in this chapter; these scripts are freely available in different directories at the following URL address: https://github.com/dunarel/dunphd-thesis.

Supplementary Table 1. Species sampled: Taxon ID from the NCBI Taxonomy database, scientific and abbreviated species names used in tree representation.

Id	Taxon-id	Scientific name	Tree name
1	228908	*Nanoarchaeum equitans* Kin4-M	*N.equitans*
2	374847	*Candidatus Korarchaeum cryptofilum* OPF8	*Ca.K.cryptofil.*
3	272557	*Aeropyrum pernix* K1	*A.pernix*
4	273057	*Sulfolobus solfataricus* P2	*S.solfataricus*
5	768679	*Thermoproteus tenax* Kra 1	*T.tenax*
6	188937	*Methanosarcina acetivorans* C2A	*M.acetivorans*
7	362976	*Haloquadratum walsbyi* DSM 16790	*H.walsbyi*
8	309800	*Haloferax volcanii* DS2	*H.volcanii*
9	348780	*Natronomonas pharaonis* DSM 2160	*N.pharaonis*
10	634497	*Haloarcula hispanica* ATCC 33960	*H.hispanica*
11	272569	*Haloarcula marismortui* ATCC 43049	*H.marismortui*
12	243090	*Rhodopirellula baltica* SH 1	*R.baltica*
13	190304	*Fusobacterium nucleatum* subsp. nucleatum ATCC 25586	*F.nucleatum*
14	240015	*Acidobacterium capsulatum* ATCC 51196	*A.capsulatum*
15	743525	*Thermus scotoductus* SA-01	*T.scotoductus*
16	224324	*Aquifex aeolicus* VF5	*A.aeolicus*
17	484019	*Thermosipho africanus* TCF52B	*T.africanus*
18	255470	*Dehalococcoides sp.* CBDB1	*Dehaloc.sp.* C
19	311424	*Dehalococcoides sp.* VS	*Dehaloc.sp.* V
20	330214	*Candidatus Nitrospira defluvii*	*Ca.N.defluvii*
21	379066	*Gemmatimonas aurantiaca* T-27	*G.aurantiaca*
22	267671	*Leptospira interrogans* serovar Copenhageni str. Fiocruz L1-130	*L.interrogans*
23	759914	*Brachyspira pilosicoli* 95/1000	*B.pilosicoli*
24	565034	*Brachyspira hyodysenteriae* WA1	*B.hyodysenter.*
25	167539	*Prochlorococcus marinus* subsp. marinus str. CCMP1375	*P.marinus*
26	1148	*Synechocystis sp.* PCC 6803	*Synech.sp.*
27	43989	*Cyanothece sp.* ATCC 51142	*Cyanoth.sp.*
28	481448	*Methylacidiphilum infernorum* V4	*M.infernorum*
29	716544	*Waddlia chondrophila* WSU 86-1044	*W.chondrophila*
30	765952	*Parachlamydia acanthamoebae* UV7	*P.acanthamoeb.*
31	194439	*Chlorobium tepidum* TLS	*C.tepidum*
32	269798	*Cytophaga hutchinsonii* ATCC 33406	*C.hutchinsonii*
33	411154	*Gramella forsetii* KT0803	*G.forsetii*
34	402612	*Flavobacterium psychrophilum* JIP02/86	*F.psychrophilum*
35	1034807	*Flavobacterium branchiophilum* FL-15	*F.branchioph.*
36	405948	*Saccharopolyspora erythraea* NRRL 2338	*S.erythraea*
37	227882	*Streptomyces avermitilis* MA-4680	*S.avermitilis*
38	216594	*Mycobacterium marinum* M	*M.marinum*
39	1048245	*Mycobacterium canettii* CIPT 140010059	*M.canettii*
40	572418	*Mycobacterium africanum* GM041182	*M.africanum*
41	419947	*Mycobacterium tuberculosis* H37Ra	*M.tuberculosis* a
42	83332	*Mycobacterium tuberculosis* H37Rv	*M.tuberculosis* v
43	233413	*Mycobacterium bovis* AF2122/97	*M.bovis* 2122
44	410289	*Mycobacterium bovis* BCG str. Pasteur 1173P2	*M.bovis* 1173
45	561275	*Mycobacterium bovis* BCG str. Tokyo 172	*M.bovis* 172
46	991791	*Clostridium acetobutylicum* DSM 1731	*C.acetobut.*
47	208596	*Carnobacterium sp.* 17-4	*Carnobact.sp.*
48	684738	*Lactococcus lactis* subsp. lactis KF147	*L.lactis* K
49	272623	*Lactococcus lactis* subsp. lactis Il1403	*L.lactis* I
50	543734	*Lactobacillus casei* BL23	*L.casei*
51	568704	*Lactobacillus rhamnosus* Lc 705	*L.rhamnosus* L
52	568703	*Lactobacillus rhamnosus* GG	*L.rhamnosus* G
53	358681	*Brevibacillus brevis* NBRC 100599	*B.brevis*
54	398511	*Bacillus pseudofirmus* OF4	*B.pseudofirmus*
55	315750	*Bacillus pumilus* SAFR-032	*B.pumilus*
56	592022	*Bacillus megaterium* DSM 319	*B.megaterium* D
57	545693	*Bacillus megaterium* QM B1551	*B.megaterium* Q
58	281309	*Bacillus thuringiensis* serovar konkukian str. 97-27	*B.thuringiensis*
59	288681	*Bacillus cereus* E33L	*B.cereus* E
60	637380	*Bacillus cereus* biovar anthracis str. CI	*B.cereus* C
61	361100	*Bacillus cereus* Q1	*B.cereus* Q
62	279010	*Bacillus licheniformis* ATCC 14580	*B.licheniformis*
63	224308	*Bacillus subtilis* subsp. subtilis str. 168	*B.subtilis* 1
64	655816	*Bacillus subtilis* subsp. spizizenii str. W23	*B.subtilis* W
65	326423	*Bacillus amyloliquefaciens* FZB42	*B.amyloliquef.* F
66	692420	*Bacillus amyloliquefaciens* DSM 7	*B.amyloliquef.* D
67	177437	*Desulfobacterium autotrophicum* HRM2	*D.autotroph.*
68	448385	*Sorangium cellulosum* 'So ce 56'	*S.cellulosum*
69	404380	*Geobacter bemidjiensis* Bem	*G.bemidjiensis*
70	273121	*Wolinella succinogenes* DSM 1740	*W.succinog.*
71	382638	*Helicobacter acinonychis* str. Sheeba	*H.acinonychis*
72	693745	*Helicobacter pylori* B8	*H.pylori*
73	62928	*Azoarcus sp.* BH72	*Azoarc.sp.*
74	266264	*Cupriavidus metallidurans* CH34	*C.metallidurans*
75	1042878	*Cupriavidus necator* N-1	*C.necator*
76	381666	*Ralstonia eutropha* H16	*R.eutropha*
77	375286	*Janthinobacterium sp.* Marseille	*Janthin.sp.*
78	757424	*Herbaspirillum seropedicae* SmR1	*H.seropedicae*
79	1005048	*Collimonas fungivorans* Ter331	*C.fungivorans*
80	452662	*Sphingobium japonicum* UT26S	*S.japonicum*
81	272942	*Rhodobacter capsulatus* SB 1003	*R.capsulatus*
82	375451	*Roseobacter denitrificans* OCh 114	*R.denitrificans*
83	272568	*Gluconacetobacter diazotrophicus* PAl 5	*G.diazotroph.*
84	414684	*Rhodospirillum centenum* SW	*R.centenum*
85	258594	*Rhodopseudomonas palustris* CGA009	*R.palustris*
86	224911	*Bradyrhizobium japonicum* USDA 110	*B.japonicum*
87	288000	*Bradyrhizobium sp.* BTAi1	*Bradyrh.sp.*
88	311403	*Agrobacterium radiobacter* K84	*A.radiobacter*
89	347834	*Rhizobium etli* CFN 42	*R.etli* CFN
90	491916	*Rhizobium etli* CIAT 652	*R.etli* CIAT
91	706191	*Pantoea ananatis* LMG 20103	*P.ananatis*
92	321314	*Salmonella enterica* subsp. enterica serovar Choleraesuis str. SC-B67	*S.enterica*
93	585035	*Escherichia coli* S88	*E.coli* S88
94	585395	*Escherichia coli* O103:H2 str. 12009	*E.coli* O103
95	585055	*Escherichia coli* 55989	*E.coli* 55989
96	573235	*Escherichia coli* O26:H11 str. 11368	*E.coli* O26
97	585397	*Escherichia coli* ED1a	*E.coli* ED1a
98	585034	*Escherichia coli* IAI1	*E.coli* IAI1
99	405955	*Escherichia coli* APEC O1	*E.coli* APEC
100	364106	*Escherichia coli* UTI89	*E.coli* UTI
101	585056	*Escherichia coli* UMN026	*E.coli* UMN
102	544404	*Escherichia coli* O157:H7 str. TW14359	*E.coli* O157
103	585396	*Escherichia coli* O111:H- str. 11128	*E.coli* O111
104	199310	*Escherichia coli* CFT073	*E.coli* CFT
105	701177	*Escherichia coli* O55:H7 str. CB9615	*E.coli* O55
106	413997	*Escherichia coli* B str. REL606	*E.coli* BREL
107	585057	*Escherichia coli* IAI39	*E.coli* IAI39
108	574521	*Escherichia coli* O127:H6 str. E2348/69	*E.coli* O127
109	511145	*Escherichia coli* str. K-12 substr. MG1655	*E.coli* K-12 M
110	316385	*Escherichia coli* str. K-12 substr. DH10B	*E.coli* K-12 D
111	595496	*Escherichia coli* BW2952	*E.coli* BW

Supplementary Table 2. Habitat membership of sampled species: species-family presence-absence matrix.
Columns represent: Taxon ID from the NCBI Taxonomy database, abbreviated species names used in tree representation; it is followed by 1 if species is present in the corresponding habitat or by 0 if it is absent in it -according to the GOLD database (also see Appendix B, Formula B.3).

TAXON_ID	Species name	Human Respiratory	Human Others	Plant	Animal	Soil	Marine	Fresh water	Extreme
228908	*N.equitans*	0	0	0	0	0	1	0	1
374847	*Ca.K.cryptofil.*	0	0	0	0	0	0	1	1
272557	*A.pernix*	0	0	0	0	0	1	0	1
273057	*S.solfataricus*	0	0	0	0	0	0	1	1
768679	*T.tenax*	0	0	0	0	0	0	1	0
188937	*M.acetivorans*	0	0	0	0	0	1	0	1
362976	*H.walsbyi*	0	0	0	0	0	0	1	1
309800	*H.volcanii*	0	0	0	0	0	0	1	0
348780	*N.pharaonis*	0	0	0	0	0	0	1	0
634497	*H.hispanica*	0	0	0	0	0	1	0	1
272569	*H.marismortui*	0	0	0	0	0	1	0	0
243090	*R.baltica*	0	0	0	0	0	1	0	0
190304	*F.nucleatum*	1	0	0	0	0	0	0	0
240015	*A.capsulatum*	0	0	0	0	1	0	1	1
743525	*T.scotoductus*	0	0	1	0	1	0	1	0
224324	*A.aeolicus*	0	0	0	0	0	1	0	1
484019	*T.africanus*	0	0	0	0	0	0	1	0
255470	*Dehaloc.sp. C*	0	0	0	0	1	0	1	0
311424	*Dehaloc.sp. V*	0	0	0	0	0	0	1	0
330214	*Ca.N.defluvii*	0	0	0	0	0	0	1	0
379066	*G.aurantiaca*	0	0	0	0	0	0	1	1
267671	*L.interrogans*	0	1	0	0	1	0	1	0
759914	*B.pilosicoli*	0	0	0	1	0	0	0	0
565034	*B.hyodysenter.*	0	0	0	1	0	0	0	0
167539	*P.marinus*	0	0	0	0	0	1	0	0
1148	*Synech.sp.*	0	0	0	0	0	0	1	0
43989	*Cyanoth.sp.*	0	0	0	0	0	1	0	0
481448	*M.infernorum*	0	0	0	0	0	0	1	1
716544	*W.chondrophila*	0	0	0	1	0	0	0	0
765952	*P.acanthamoeb.*	0	1	0	0	0	0	0	0
194439	*C.tepidum*	0	0	0	0	0	0	1	0
269798	*C.hutchinsonii*	0	0	0	0	1	1	1	0
411154	*G.forsetii*	0	0	0	0	0	1	0	0
402612	*F.psychrophilum*	0	0	0	1	0	0	1	0
1034807	*F.branchioph.*	0	0	0	0	0	0	1	0
405948	*S.erythraea*	0	0	0	0	1	0	0	0
227882	*S.avermitilis*	0	0	0	0	1	0	0	0
216594	*M.marinum*	0	1	0	0	0	0	0	0
1048245	*M.canettii*	0	1	0	0	0	0	0	0

TAXON_ID	Species name	Human Respiratory	Human Others	Plant	Animal	Soil	Marine	Fresh water	Extreme
572418	*M.africanum*	0	1	0	0	0	0	0	0
419947	*M.tuberculosis a*	0	1	0	0	0	0	0	0
83332	*M.tuberculosis v*	0	1	0	0	0	0	0	0
233413	*M.bovis 2122*	0	1	0	1	0	0	0	0
410289	*M.bovis 1173*	0	0	0	1	0	0	0	0
561275	*M.bovis 172*	0	1	0	0	0	0	0	0
991791	*C.acetobut.*	0	0	0	0	1	0	0	0
208596	*Carnobact.sp.*	0	0	0	0	0	1	0	0
684738	*L.lactis K*	0	0	1	0	0	0	0	0
272623	*L.lactis I*	0	0	0	1	0	0	0	0
543734	*L.casei*	1	0	0	0	0	0	0	0
568704	*L.rhamnosus L*	1	0	0	1	0	0	0	0
568703	*L.rhamnosus G*	1	0	0	0	0	0	0	0
358681	*B.brevis*	0	0	0	0	1	0	0	0
398511	*B.pseudofirmus*	0	0	0	0	1	0	0	0
315750	*B.pumilus*	0	0	0	0	1	0	0	0
592022	*B.megaterium D*	0	0	0	0	1	0	0	0
545693	*B.megaterium Q*	0	0	0	0	1	0	0	0
281309	*B.thuringiensis*	0	1	0	0	1	0	0	0
288681	*B.cereus E*	0	1	0	1	1	0	0	0
637380	*B.cereus C*	0	1	0	0	1	0	0	0
361100	*B.cereus Q*	0	0	0	0	1	0	0	0
279010	*B.licheniformis*	1	0	0	0	1	0	0	0
224308	*B.subtilis 1*	0	0	0	0	1	0	0	0
655816	*B.subtilis W*	0	0	0	0	1	0	0	0
326423	*B.amyloliquef. F*	0	0	0	0	1	0	0	0
692420	*B.amyloliquef. D*	0	0	0	0	1	0	0	0
177437	*D.autotroph.*	0	0	0	0	0	1	0	0
448385	*S.cellulosum*	0	0	0	0	1	0	0	0
404380	*G.bemidjiensis*	0	0	0	0	1	0	1	0
273121	*W.succinog.*	0	0	0	1	0	0	0	0
382638	*H.acinonychis*	0	0	0	1	0	0	0	0
693745	*H.pylori*	1	0	0	0	0	0	0	0
62928	*Azoarc.sp.*	0	0	1	0	0	0	0	0
266264	*C.metallidurans*	0	0	0	0	1	0	1	0
1042878	*C.necator*	0	0	0	0	1	0	1	0
381666	*R.eutropha*	0	0	0	0	1	0	1	0
375286	*Janthin.sp.*	0	0	0	0	0	0	1	0
757424	*H.seropedicae*	0	0	1	0	1	0	0	0
1005048	*C.fungivorans*	0	0	0	0	1	0	0	0
452662	*S.japonicum*	0	0	0	0	1	0	0	1
272942	*R.capsulatus*	0	0	0	0	1	0	0	0
375451	*R.denitrificans*	0	0	0	0	0	0	1	0
272568	*G.diazotroph.*	0	0	1	0	0	0	0	0
414684	*R.centenum*	0	0	0	0	0	0	1	0

TAXON_ID	Species name	Human Respiratory	Human Others	Plant	Animal	Soil	Marine	Fresh water	Extreme
258594	*R.palustris*	0	0	0	0	1	0	1	0
224911	*B.japonicum*	0	0	1	0	0	0	0	0
288000	*Bradyrh.sp.*	0	0	1	0	1	0	0	0
311403	*A.radiobacter*	0	1	1	0	0	0	0	0
347834	*R.etli CFN*	0	0	1	0	0	0	0	0
491916	*R.etli CIAT*	0	0	1	0	0	0	0	0
706191	*P.ananatis*	0	0	1	0	0	0	0	0
321314	*S.enterica*	0	1	0	0	1	0	1	0
585035	*E.coli S88*	1	0	0	0	0	0	0	0
585395	*E.coli O103*	1	0	0	0	0	0	0	0
585055	*E.coli 55989*	1	0	0	0	0	0	0	0
573235	*E.coli O26*	1	0	0	0	0	0	0	0
585397	*E.coli ED1a*	1	0	0	0	0	0	0	0
585034	*E.coli IAI1*	1	0	0	0	0	0	0	0
405955	*E.coli APEC*	1	0	0	0	0	0	0	0
364106	*E.coli UTI*	1	0	0	0	0	0	0	0
585056	*E.coli UMN*	1	0	0	0	0	0	0	0
544404	*E.coli O157*	1	0	0	0	0	0	0	0
585396	*E.coli O111*	1	0	0	0	0	0	0	0
199310	*E.coli CFT*	1	0	0	0	0	0	0	0
701177	*E.coli O55*	1	0	0	0	0	0	0	0
413997	*E.coli BREL*	1	0	0	0	0	0	0	0
585057	*E.coli IAI39*	1	1	0	0	0	0	0	0
574521	*E.coli O127*	1	0	0	0	0	0	0	0
511145	*E.coli K-12 M*	1	0	0	0	0	0	0	0
316385	*E.coli K-12 D*	1	0	0	0	0	0	0	0
595496	*E.coli BW*	1	0	0	0	0	0	0	0

Supplementary Table 3. Genes sampled.

a) *Core genes* (36 in total) according to (Charlebois and Doolittle 2004).

alaS	*argS*	*atpD*	*cysS*	*dnaG*	*eno*	*gltX*	*glyA*	*groEL*
guaA	*hisS*	*ileS*	*infB*	*ksgA*	*leuS*	*lysS*	*map*	*nusA*
nusG	*pheS*	*pheT*	*proS*	*pyrG*	*pyrH*	*recA*	*rplB*	*rplC*
rplX	*rpoB*	*rpsN*	*secY*	*serS*	*thrS*	*trpS*	*trxB*	*valS*

b) *Rest of the genes* (74 in total) considered in this study.

adk	*argD*	*aroA*	*aroE*	*asd*	*aspS*	*atpA*	*atpB*	*atpC*
carB	*clpP*	*clpX*	*dapA*	*def*	*dnaJ*	*dnaK*	*fabG*	*galE*
gatA	*gatB*	*glmS*	*glyS*	*grpE*	*gyrA*	*gyrB*	*hemA*	*hisB*
hisD	*hisH*	*ilvC*	*ilvD*	*metK*	*nadD*	*nadE*	*oppA*	*pgk*
proA	*proC*	*purA*	*purB*	*purD*	*purE*	*purF*	*purL*	*purM*
pyrB	*pyrC*	*pyrD*	*pyrE*	*pyrF*	*ribH*	*rnhB*	*rplF*	*rplP*
rplW	*rpmB*	*rpmG*	*rpoA*	*rpoD*	*ruvB*	*sdhA*	*secE*	*serA*
thrC	*trpA*	*trpB*	*trpC*	*trpD*	*truA*	*trxA*	*tyrS*	*uvrA*
uvrB	*uvrC*							

Supplementary Table 4. Time constraints applied to the gene tree nodes, corresponding to the considered phylogenetic families and some of their Most Recent Common Ancestors (MRCA), up to their Last Common Ancestor (LCA).

ID	Node name	Confidence interval Minimum time	Confidence interval Maximum time	Mean time	Standard deviation	Node name (or MRCA)
0	34	3417	4482	4149	289	*Archaea*
1	106	3437	4391	3917	272	*Bacteria*
2	105	3181	4038	3590	244	*76 Thermotogae*
3	104	2815	3530	3139	207	*76 103*
4	103	2738	3434	3051	201	*102 Bacteroidetes/Chlorobi*
5	102	2658	3339	2963	196	*101 Chlamydiae Spirochaetes*
6	101	2460	3128	2761	189	*100 Epsilonproteobacteria*
7	100	2156	2844	2476	187	*a99 Alphaproteobacteria*
8	99	1569	2310	1924	192	*Gamaproteobacteria Betaproteobacteria*
9	80	2438	3115	2744	191	*Chlamydiae Spirochaetes*
10	76	2713	3382	3009	196	*75 Fusobacteria Firmicutes*
11	75	2512	3076	2743	173	*71 Actinobacteria*
12	71	2342	2814	2519	159	*Deinococcus-Thermus Cyanobacteria*
13	68	2602	3242	2880	210	*Fusobacteria Firmicutes*
14	33	3154	4168	3826	270	*Euryarchaeota*
15	20	3046	3919	3617	229	*Crenarchaeota*
16	98	1382	2113	1732	189	*Gammaproteobacteria*
17	87	1107	1837	1455	187	*Betaproteobacteria*
18	86	1650	2390	2007	192	*Alphaproteobacteria*
19	81	858	1666	1236	207	*Epsilonproteobacteria*
20	79	1423	2254	1839	213	*Spirochaetes*
21	78	320	1042	592	187	*Chlamidia*
22	74	1032	1727	1357	179	*Actinobacteria*
23	70	706	1355	1020	169	*Cyanobacteria*
24	67	2367	3013	2650	181	*Firmicutes*
25	lca	2500	4500	4290	0	*LCA*

Node names and numbers, confidence intervals and standard deviations were those provided by (Battistuzzi et al. 2004).

CHAPTER V

A NEW FAST ALGORITHM FOR DETECTING AND VALIDATING HORIZONTAL GENE TRANSFER EVENTS USING PHYLOGENETIC TREES AND AGGREGATION FUNCTIONS

5.1 Abstract

5.1.1 Background

Until recently, the traditional view of prokaryotic evolution has been based on divergence and periodic selection. Mutation has been assumed to be the main diversifying force and selection was the unifying one, until accumulation of mutations led to a speciation event. A new evolutionary model has slowly emerged, in which Horizontal Gene Transfer (HGT) is the main diversifying force and recombination is the main unifying one, speciation being an ecological adaptation. Negative selection has been the most studied evolutionary force, for which simple and efficient detection methods, based on sequence conservation (see chapters I and II), exist. In chapter 3 we described an efficient algorithm, applied to the strains based on distinct pathogenic populations, for detecting functional genomic regions associated with positive and lineage specific selection (in both variants, monophyletic or polyphyletic).

5.1.2 Results and conclusion

In this chapter, we present a new algorithm, called HGT-QFUNC, for detecting genomic regions that can be associated with complete HGT events. The aggregation functions described in chapter III, which yielded good results in detecting selection, will be tested in the context of HGT identification. New clustering functions which perform better in presence of HGT and recombination will be also introduced.

We will validate our results using *p*-values estimated by means of a Monte Carlo approach. To estimate the rates of complete HGT among prokaryotes, we will compare our results to the highly accurate but slower HGT-Detection algorithm based on the calculation of bootstrap support of considered gene trees (see chapter IV). We will also compare the results provided by HGT-QFUNC and HGT-Detection using simulated data, which will be representative of the prokaryotic landscape. We will show that the proposed new functions and algorithm are capable of providing good detection rates for most of the highly probable HGT events. The main advantage of the proposed algorithm is its quadratic time complexity on the number of considered species. This makes it applicable to the study of large genomic datasets. Note that the proposed HGT-QFUNC algorithm yields better performances than a simple conservation approach, running at the same quadratic asymptotic time. The obtained results confirm the prime importance of HGT in the light of prokaryotic evolution.

5.2 Background

The mechanisms by which bacteria and viruses adapt to changing environmental conditions are well known. These mechanisms include homologous recombination (Posada and Crandall 2001), nucleotide substitutions, insertions-deletions (Kimura 1985) and horizontal gene transfer (Boc et al. 2010). The variation of the DNA composition is spread throughout prokaryotic genomes leading to the formation of different polymorphic strands of the same group of organisms. The survival of these strands depends on their ability to overcome environmental changes (Moran 1962). Multiple mechanisms can overlap, and limits between groups are sometimes "fuzzy" (Hanage et al. 2005). The classical Linnaean paradigm of biological classification is that of a hierarchical organization of species into increasingly narrower groups based on their shared characteristics. It is the most used framework for interpreting Darwinian evolution. According to it, the most narrowly defined biological group is the species, and the formation of a new lineage is a speciation, which entails the diversification of one species into two different species. Inside the species, a free exchange of

genetic information is allowed, but outside the species boundaries, genetic information is passed solely to descendant individuals.

This model of evolution is challenged on the prokaryotic level, where there exists experimental evidence of massive transfer of genetic material between different organisms (Fraser et al. 2007). Such a transfer can occur by two distinct routes: homologous recombination and HGT (Thomas and Nielsen 2005). Homologous recombination is often limited to closely related organisms, having sufficient sequence similarly to allow for efficient integration of genetic material (Ochman et al. 2000). HGT can occur between both closely related and genetically distinct organisms.

5.3 Data description

5.3.1 Real prokaryotic (genomic) data

We assembled a real-world dataset, representative of the prokaryotic genomic landscape, to serve as a basis for testing our algorithm against a well-known HGT-Detection (v.3.4) algorithm (Boc et al. 2010) available on the T-Rex web site (Boc et al. 2012). A complete description of this dataset is available in chapter IV. Here we outline the most important features of the dataset that we examined.

All of the completely sequenced prokaryotic genomes available at the NCBI Genomes ftp site (1465 as of November 2011) were considered. Among them, we first selected 100 of the most complete genomes in terms of the number of genes. Then, we added to them 11 additional species to ensure that our dataset includes at least one representative from each of the 23 available prokaryotic families. This yielded us a total number of 111 species. Detailed information on the considered species can be found in Supplementary Table 1. We also identified 110 of the most complete genes (Supplementary Table 3) from the selected set of 111 species (see also chapter IV for more details). Multiple sequence alignments could contain multiple alleles of the same species.

Afterward, we constructed 110 multiple sequence alignments (one MSA per selected gene) from which we excluded misclassified paralogs using TribeMCL (Enright et al. 2002). The latter tool, which uses a Markov Chain Clustering (MCL) algorithm (van Dongen 2000) on all-to-all BLASTP hits, is known to be conservative in the number of groups (Li et al. 2012). We carried out the TribeMCL version bundled with "mcl" v11.294, with default parameters (I=2.0). In order to obtain more accurate results of BLASTP, we set a Smith-Waterman backend and an E-value threshold of

10^{-4}. Using this procedure, 1% of initial alleles were identified as paralogs and excluded from the original MSA.

Nucleotide sequences were retrieved from protein sequences identified above. They were aligned using MUSCLE v3.8.31 (Edgar 2004), with default parameters, and trimmed with GBlocks v0.91b (Castresana 2000). In our analysis, we were less restrictive than the default option of GBlocks, allowing 50% of the sequences for flank positions (-b2 parameter), a maximum of 10 contiguous nonconserved positions (-b3 parameter), minimum block length of 5 (-b4 parameter) and half gap positions (-b5 parameter).

The obtained MSAs were then used as a basis for the detection of complete HGT. Species taxonomy (i.e. species tree in the HGT context) was retrieved from the NCBI Taxonomy website (Benson et al. 2009). Taxonomic groups were those assigned by the NCBI Genomes Project. Each species was then assigned to one established prokaryotic family.

We constructed the gene trees using the RAxML method (Stamatakis 2006). We used the RAxML v.7.2.8 - multithreaded implementation, and GTR Gamma model, 20 starting random trees and 100 bootstrap trees as RAxML input options.

5.3.2 Synthetic data

For a simulation study conducted with synthetic data, we used the real prokaryotic dataset as a basis, in order to maintain the same real-world relationships between sequences, and the same limitations for our detection algorithm as it would be in a real-world situation. To simulate our data, we chose as benchmark the gene *hisH*, which is the gene with the highest number of different prokaryotic strains (i.e. 99) in which the HGT-Detection algorithm did not find any HGT at the bootstrap level of 50%. This threshold was considered as a minimum quality requirement in our study. The detailed description of the simulated synthetic data and the corresponding simulation study can be found in section 5.4.7.

5.4 Methods

5.4.1 Clustering using aggregation functions

Considering a collection of different prokaryotes, classified as belonging to different taxonomic groups, we can model the simplest case of HGT as the transfer of one single gene sequence between two different species (e.g. x_0 and y_0) belonging to two different monophyletic groups (e.g. X and Y).

If there was a genetic transfer from source strain y_0 to destination species x_0, then species x_0 would have the same or very similar genetic sequence as the source species y_0. This would lead to an inverse direction shift in phylogenetic classification. HGT can involve either heterologous (e.g. site-specific) or homologous recombination, or direct host gene replacement followed by intragenic recombination. The end result at the phylogenetic level is the integration of species x_0 into the group Y closely to source species y_0. This situation is depicted in Figure 5.1. Genetic transfer direction and resulting phylogenetic neighborhood are dependent on the relative fractions of species x_0 and y_0 involved in the process of recombination. In this chapter, we consider the case of complete HGT when source species is integrated into the host genome without intragenic recombination (i.e. without formation of a mosaic gene).

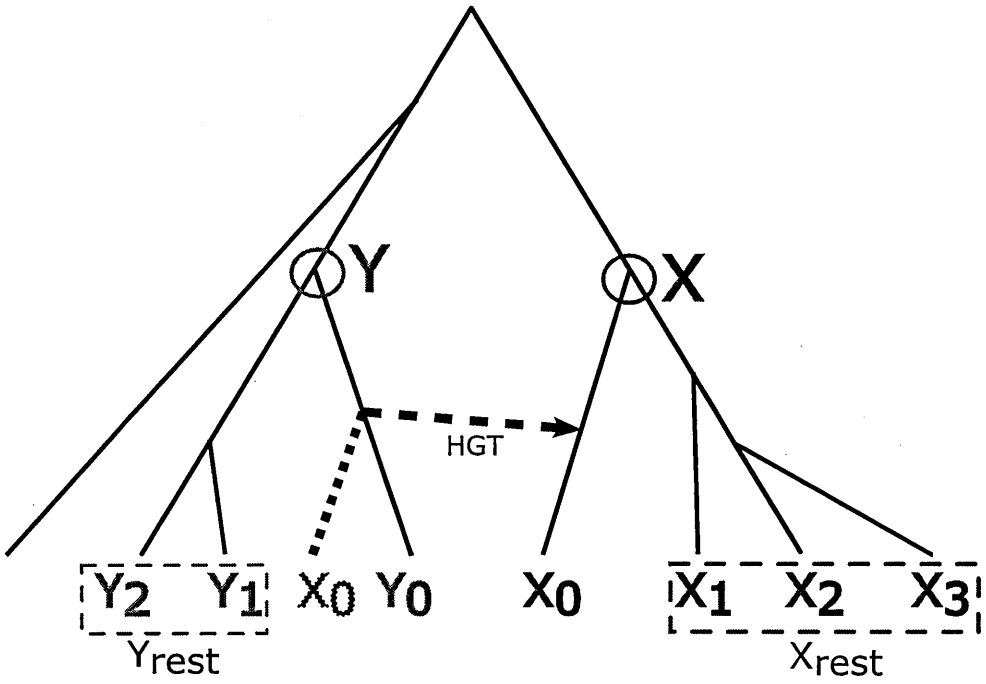


Figure 5.1. Intragroup and intergroup phylogenetic relationships following an HGT
A horizontal gene transfer from species y_0 of the group Y to species x_0 of the group X is shown by an arrow; dotted line shows the position of species x_0 in the tree after the transfer; X_{rest} denotes the rest of the species of the X group, and Y_{rest} denotes the rest of the species of the group Y. Each species corresponds to a unique nucleotide sequence in this example.

Here we describe the HGT detection problem mathematically. To perform the clustering of our data, we first define the following sets, involving the HGT-related species x_0 and y_0:

$$R = \{x_0 \cup y_0\}, \tag{5.1}$$

$$X_{rest} = X \setminus x_0, \tag{5.2}$$

$$Y_{rest} = Y \setminus y_0. \tag{5.3}$$

Note that in a general case, x_0 and y_0 can be clusters (i.e. sub-trees) including several species. We define the intergroup and intragroup variability. Consider two groups of species A and B not having common members. The measures in question are calculated as the means of the Hamming distances (any other evolutionary distance can be used instead), $dist_h$, among the sequences of the same group A (or B) only, and among the sequences from the distinct groups A and B.

First, the intragroup variability of the group A, denoted by $V(A)$, is defined by equation 5.4:

$$V(A) = \sum_{\{a_1, a_2 \in A | a_1 \neq a_2\}} dist_h(a_1, a_2). \tag{5.4}$$

We then normalize $V(A)$ by the number of possible different pairs of elements in A (equation 5.5):

$$V_{norm}(A) = \frac{V(A)}{N(A) \times (N(A)\text{-}1)/2}, \tag{5.5}$$

where $N(A)$ is the number of elements in the group A.

The intergroup variability of the groups A and B, denoted by $D(A,B)$, is defined as follows:

$$D(A,B) = \sum_{\{a \in A, b \in B\}} dist_h(a,b). \tag{5.6}$$

We then normalize $D(A,B)$ by the number of possible pairs of species:

$$D_{norm}(A,B) = \frac{D(A,B)}{N(A) \times N(B)}. \tag{5.7}$$

Using previously described groups and functions, we define a new function Q_7 as follows:

$$Q_7(R) = Max(D_{norm}(R,X_{rest}); D_{norm}(R,Y_{rest})) - V_{norm}(R), \quad (5.8)$$

where R is defined by equation 5.1.

When a complete HGT happens (Figure 5.1), the transferred gene is assumed to replace a homologous gene in the host genomes. As a result of this event, destination species x_0 migrates close to source species y_0 into the phylogenetic network representing the evolution of the given gene (Figure 5.1). Thus, in the obtained gene tree destination species x_0 will be a part of the group $\boldsymbol{Y}$ to which belongs source species y_0. Formula 5.8 reflects such a principle. Also $V_{norm}(R)$, in this particular case, defines the distance between species x_0 and y_0.

We also introduce the aggregation function, Q_8, similar to the function that provided good results in detecting lineage specific selection in (Badescu et al. 2010), (i.e. $Q_6 = |V(A)/V(B)|$):

$$Q_8(R) = \frac{D_{norm}(R,XY_{rest})}{V_{norm}(R)}. \quad (5.9)$$

Because this function uses the division instead of the summation, such a function underlines the asymmetry between the two groups. Note that both HGT and lineage specific selection exhibit asymmetrical properties.

Finally, we define the function $Q_9(R)$ as follows:

$$Q_9(R) = -V_{norm}(R). \quad (5.10)$$

Another clustering option would be to consider both interacting species as a destination group and merge the rest of the sequences into the source group:

$$XY_{rest} = \{X \cup Y \setminus R\}. \quad (5.11)$$

5.4.2 Other variants of clustering functions as implemented in the algorithm

Here we describe particular cases of formulas used in our implementation when the HGT occurred between the tree leaves (i.e. individual species of X and Y; see formulas 5.12-5.18). Let us define:

$$D(x_0,y_0) = dist_h(x_0,y_0), \quad (5.12)$$

$$D(x_0,X_{rest}) = \sum_{\{x_i \in X; x_i \neq x_0\}} dist_h(x_0,x_i) = D(x_0,X), \tag{5.13}$$

$$D(y_0,Y_{rest}) = \sum_{\{y_i \in Y; y_i \neq y_0\}} dist_h(y_0,y_i) = D(y_0,Y), \tag{5.14}$$

$$D(x_0,Y) = \sum_{\{y_i \in Y\}} dist_h(x_0,y_i), \tag{5.15}$$

$$D(y_0,X) = \sum_{\{x_i \in X\}} dist_h(y_0,x_i), \tag{5.16}$$

$$D(x_0,Y_{rest}) = \sum_{\{y_i \in Y; y_i \neq y_0\}} dist_h(x_0,y_i) = D(x_0,Y) - D(x_0,y_0), \tag{5.17}$$

$$D(y_0,X_{rest}) = \sum_{\{x_i \in X; x_i \neq x_0\}} dist_h(y_0,x_i) = D(y_0,X) - D(x_0,y_0), \tag{5.18}$$

$$n = N(X), \tag{5.19}$$

$$m = N(Y). \tag{5.20}$$

We also introduced an epsilon (ε) value (i.e. in our implementation we set the value of ε equal to 0.00001) to avoid the division by zero. Some other constants were also added to formulas (5.9) – (5.22) in order to normalize the results and to obtain the same minimum value.

$$Q_{7a}(x_0,y_0) = Max\left(\frac{D(x_0,X_{rest}) + D(y_0,X_{rest})}{2(n\text{-}1)}, \frac{D(x_0,Y_{rest}) + D(y_0,Y_{rest})}{2(m\text{-}1)}\right) - D(x_0,y_0) + 2, \text{ and} \tag{5.21}$$

$$Q_{8a}(x_0,y_0) = \frac{D(x_0,X_{rest}) + D(y_0,X_{rest}) + D(x_0,Y_{rest}) + D(y_0,Y_{rest}) + \varepsilon}{2(xy+\varepsilon)(n+m\text{-}2)}. \tag{5.22}$$

In our implementation, we used the following variants of the functions Q_8 and Q_9 as well:

$$Q_{8b}(x_0,y_0) = \frac{\frac{D(x_0,X) + D(y_0,X)}{n\text{-}1} + \frac{D(x_0,Y) + D(y_0,Y)}{m\text{-}1} + \varepsilon}{2(D(x_0,y_0)+\varepsilon)}, \text{ and} \tag{5.23}$$

$$Q_{9a}(x_0,y_0) = -D(x_0,y_0) + 2. \tag{5.24}$$

For each pair of species (x,y) belonging to two different groups, we maximize Q_{7a}, Q_{8a} and Q_{8b}, over all possible sets R in order to identify the best HGT candidates. At the same time, maximizing Q_{9a} is equivalent to minimizing the distance between x_0 and y_0.

5.4.3 Description of the algorithm

Here we present a new algorithm allowing one to estimate the values of functions Q_{7a} (called later Q_7), Q_{8a}, Q_{8b} and Q_{9a} (called later Q_9) and to validate the results using the p-value estimation procedure. Our algorithm takes as input a multiple sequence alignment (MSA) of n species, a set of groups (e.g. species families) and a unique association of each species to one of these groups. The algorithm's output consists of pairs of clusters that could be involved in horizontal gene transfers. The detailed algorithmic scheme is presented in Algorithm 5.1.

The p-value estimation is done by carrying out a Monte Carlo procedure with a fixed p-value threshold. For a constant number of steps, this procedure simulates permuted MSAs. A constant number of nucleotides are permuted within each of the original sequences. Then, we compare the obtained values of the selected function Q to the reference value obtained with the original data. The detailed p-value estimation scheme is presented in Algorithm 5.2.

The main algorithm consists of the three major steps. First, it calculates the pairwise distance matrix between all given species. Second, it calculates the distance between each species and all other species belonging to the other groups. Third, it estimates the intergroup and intragroup distances and aggregation solutions by using the formulas (5.12)-(5.24).

There is one more step needed to complete the detection of HGT. The obtained potential HGTs are ranked according to the value of the corresponding Q-function, first, then p-value, second. Those HGTs whose Q-function values were greater than a fixed threshold are considered as valid. This threshold can be set based on the p-values or a fixed percentage of the total number of species (called here *percentage of positive values*). We also considered an alternative results ranking: by p-value, first, and by the Q-function value, second. Such a ranking strategy allowed us to better emphasize the strength of statistical signal. All the tests of the new HGT detection algorithm were carried out in parallel with both ranking strategies.

5.4.4 Implementation

The presented algorithm can be parallelized to improve its performances. At least three different parallelization schemes exist. The first one uses fine grained parallelism with global atomic reductions that would be better suited for graphic cards. The second one involves the parallelization of higher granularity, implying the p-value estimation steps. It would be better suited to multicore processors. The third one, which we implemented in our program, proceeds by mapping each group into a CPU core. Although this is not the most efficient scheme, it has the advantage to accelerate calculations even in the absence of the p-value estimation step. We developed a C++ code for this algorithm for multicore CPUs, parallelizing using OpenMP, and SIMD vectorizing using SSE3 instructions.

Our implementation is available at the following URL address:

http://www.info2.uqam.ca/~makarenkov_v/fastHGT.zip.

Algorithm 5.1.

HGT-QFUNC algorithm for detecting species related to each other by the way of complete horizontal gene transfer (HGT)

Require:
MSA : Multiple sequence alignment,
FI: Aggregation function to be optimized Q_7, Q_{8a}, Q_{8b} or Q_9,
GR : Groups,
SG : Unique association of each sequence in the MSA to one group (G),
Ensure:
QVAL: Matrix of Q_{FI} values for each pair of sequences ∈ MSA

1: MSA_N ← Number of sequences in MSA
2: GR_N ← Number of groups
3: N_SEQS [GR_N] ← Number of sequences (i.e. species) in each group
4: D_SEQ_SEQ ← Matrix [MSA_N] [MSA_N] // sequence to sequence distance matrix
5: **for all** $seq\_i \in$ MSA **do**
6: **for all** $seq\_j \in$ MSA **do**
7: D_SEQ_SEQ $[seq\_i]\,[seq\_j] = D(seq\_i, seq\_j) = dist_h(seq\_i, seq\_j)$
8: **end for**
9: **end for**
10: D_SEQ_GR ← Matrix [MSA_N] [GR_N] // sequence to group distance matrix
11: **for all** $seq\_i \in$ MSA **do**
12: **for all** $seq\_j \in$ MSA **do**
13: gr_j ← SG $[seq\_j]$
14: D_SEQ_GR $[seq\_i]\,[gr\_j]$ += D_SEQ_SEQ $[seq\_i]\,[seq\_j]$
15: **end for**
16: **end for**

...(continued on next page)...

...(continued from previous page)...

17: QVAL ← Matrix [MSA_N] [MSA_N]
18: **for all** $seq\_i \in$ MSA **do**
19: **for all** $seq\_j \in$ MSA **do**
20: $gr\_i \leftarrow$ SG [seq_i]
21: $gr\_j \leftarrow$ SG [seq_j]
22: $n \leftarrow$ N_SEQS [gr_i]
23: $m \leftarrow$ N_SEQS [gr_j]
24: **go to 37 if** ($seq\_i >= seq\_j$) **or** ($n < 2$) **or** ($m < 2$)
25: $D(x_0,y_0) =$ D_SEQ_SEQ[seq_i][seq_j]
26: $D(x_0,X_{rest}) = D(x_0,X) =$ D_SEQ_GR [seq_i][gr_i]
27: $D(y_0,Y_{rest}) = D(y_0,Y) =$ D_SEQ_GR [seq_j][gr_j]
28: $D(x_0,Y) =$ D_SEQ_GR [seq_i][gr_j]
29: $D(x_0,Y_{rest}) = D(x_0,Y) - D(x_0,y_0)$
30: $D(y_0,X) =$ D_SEQ_GR [seq_j][gr_i]
31: $D(y_0,X_{rest}) = D(y_0,X) - D(x_0,y_0)$
32: $$Q_7(x_0,y_0) = Max\left(\frac{D(x_0,X_{rest}) + D(y_0,X_{rest})}{2(n\text{-}1)}, \frac{D(x_0,Y_{rest}) + D(y_0,Y_{rest})}{2(m\text{-}1)}\right) - D(x_0,y_0) + 2$$
33: $$Q_{8a}(x_0,y_0) = \frac{D(x_0,X_{rest}) + D(y_0,X_{rest}) + D(x_0,Y_{rest}) + D(y_0,Y_{rest}) + \varepsilon}{2(xy+\varepsilon)(n+m\text{-}2)}$$
34: $$Q_{8b}(x_0,y_0) = \frac{\frac{D(x_0,X) + D(y_0,X)}{n\text{-}1} + \frac{D(x_0,Y) + D(y_0,Y)}{m\text{-}1} + \varepsilon}{2(D(x_0,y_0)+\varepsilon)}$$
35: $Q_9(x_0,y_0) = -D(x_0,y_0) + 2$
36: QVAL [seq_i] [seq_j] ← $Q_{FI}(x_0,y_0)$
37: **end for**
38: **end for**
39: **return** QVAL

Algorithm 5.2

P-value validation for HGT-QFUNC algorithm (see Algorithm 5.1.) using Monte Carlo estimation

---

Require:
MSA : Multiple sequence alignment,
FI: Aggregation function to be optimized Q_7, Q_{8a}, Q_{8b} or Q_9,
GR : Groups,
SG : Unique association of each sequence in the MSA, to one group (G),
PVST : Constant number of *p*-value steps,
PERM: Nucleotide permutation percentage.
Ensure:
PQVAL: Matrix of Q_{FI} *p*-values for each pair of sequences ∈ MSA

1: PQVAL ← Matrix [MSA_N] [MSA_N]
2: QVAL ← **call Algorithm 5.1**(MSA and *FI,GR,SG*)
3: **for** *i* ∈ (1...PVST) **do**
4: MSA_PERM ← MSA
5: //introduce a level of uncertainty
6: **for all** *seq_i* ∈ MSA **do**
7: permute PERM nucleotides
8: **end for**
9: //calculate regular values with permuted MSA
10: QVAL_PERM ← **call Algorithm 5.1**(MSA_PERM and *FI,GR,SG*)
11: //test if the obtained values are at least as good as those obtained without permutation
12: **for all** *seq_i* ∈ MSA **do**
13: **for all** *seq_j* ∈ MSA **do**
14: **if** QVAL_PERM [*seq_i*] [*seq_j*] >= QVAL [*seq_i*] [*seq_j*] **then**
15: PQVAL [*seq_i*] [*seq_j*] ++
16: **end if**
17: **end for**
18: **end for**
19: **end for** //end permutations
20: //update *p*-value
21: **for all** *seq_i* ∈ MSA **do**
22: **for all** *seq_j* ∈ MSA **do**
23: PQVAL [*seq_i*] [*seq_j*] ++
24: PQVAL [*seq_i*] [*seq_j*] /= (PVST +1)
25: **end for**
26: **end for**
27: **return** QVAL, PQVAL

5.4.5 Time complexity

The time complexity of the new algorithm in a general case, carried out over a MSA of n species and DNA or amino acid sequences of size l, is $O(ln^2+n^4)$. When we consider only HGT between individual species (i.e. leaves) – the most common case in HGT analysis – the time complexity is $O(ln^2)$ only (see Algorithm 5.1). The p-value estimation procedure (see Algorithm 5.2) adds a constant overhead to the algorithm's running time in order to maintain the desired p-value precision. This constant is usually 100, 1000 or 10000, for a precision of 0.01, 0.001 or 0.0001, respectively.

5.4.6 Simulation with the real prokaryotic dataset and comparison to HGT-Detection

We tested the ability of the described HGT-QFUNC algorithm to detect complete HGT events by comparing it to a highly accurate but much slower HGT-Detection algorithm (Boc et al. 2010). We used the presented functions Q_7, Q_{8a}, Q_{8b} and Q_9, side by side, in order to identify their strengths and limitations in general use-case scenarios on the real prokaryotic data described in detail in chapter IV. We used the *Sensitivity* measure to compare the performances of HGT-QFUNC and HGT-Detection. *Sensitivity*, which reflects the ability to detect true positives is defined as follows:

$$Sensitivity = \frac{number\_of\_true\_positives}{number\_of\_true\_positives + number\_of\_false\_negatives}. \qquad (5.25)$$

The true positive and false negative HGTs were determined by comparing the obtained results to those provided by the HGT-Detection algorithm (i.e. the transfers found by HGT-Detection were considered as true positives). We excluded from the analysis the alignments where HGT-Detection did not return any HGT to avoid the division by zero in the *Sensitivity* formula.

We assured comparability of the output formats between the two compared algorithms. HGT-Detection provides its results as a list of pairs of HGT-related source and destination branches defined by the corresponding nodes of the species tree. We decomposed HGT-Detection transfer scenarios into a list of all affected leave pairs between respective source and destination subtrees. As the species tree was not always completely resolved in our case, some trivial transfers (i.e. transfers amongst branches of the same multifurcating node) could occur (see chapter IV for more details). For quality reasons, we discarded such trivial transfers in this simulation study.

HGT-Detection first applies sophisticated phylogenetic tree-based manipulations and then filters results by HGT bootstrap values. The output of the HGT-Detection program usually contains a very

small number of transfers due to the applied evolutionary constraints and imposed bootstrap threshold. We tried to mimic this behavior by limiting the HGT-QFUNC algorithm to a restrictive p-value threshold of 0.001, but still had too many (compared to HGT-Detection) HGT events identified in the end. Therefore, we limited the number of detected HGT events by imposing a fixed threshold (as described below).

We carried out the HGT-Detection algorithm over our prokaryotic dataset with minimum bootstrap supports of 50%, 75% and 90%, respectively. The obtained results are shown in Figure 5.2, while the corresponding results of HGT-QFUNC, based on the p-value ordering, are shown in Figure 5.6. Corresponding runs of HGT-QFUNC had a maximum allowed number of HGTs per alignment of 300, 200 and 100 events, respectively. How restrictive these thresholds are, in terms of possible number of detected HGT events for each considered gene, is shown in figure 5.3.

5.4.7 Simulation with artificial data and comparison to HGT-Detection

After we have explored the real-life detection performances of the HGT-QFUNC algorithm, we tested its ability to recover correct HGT events by simulating different HGT rates in artificially generated multiple sequence alignments. We performed a series of tests involving random nonreciprocal sequence transfers between the species of different prokaryotic groups. Both complete (involving only gene replacement) and partial (involving intragenic recombination and creation of mosaic genes) horizontal gene transfer cases were considered in this simulation.

All simulated transfers were supposed to occur between single species of the considered MSA (a single species always corresponds to a tree leave in a phylogenetic tree). From an evolutionary standpoint such transfers are the most recent, and also the most recoverable ones (Boc et al. 2010). Therefore, they are also the most reported ones.

We considered the cases with 1 to 128 simulated transfers, following the logarithmic scale (i.e. 1, 2, 4, 8, 16, 32, 64 and 128 transfers). One of our goals in this simulation was to discriminate between the functions Q_7, Q_{8a}, Q_{8b} and Q_9 when detecting different numbers of complete HGT. We set a maximum allowed number of positive values as the double of the number of transfers (i.e. 2, 4 ,8, 16, 32, 64, 128 and 256 transfers, respectively). Note that in the above-described real-life experiments, we allowed 100, 200 and 300 transfers, depending on the bootstrap support fixed for the HGT-Detection algorithm. Mention that for the artificial data the algorithm was carried out

under more restrictive conditions (lower number of positive values) than those imposed in the experiments with real-life prokaryotic data.

We first simulated gene transfers without recombination (i.e. complete HGT), as a simple replacement of the source sequence by the destination sequence. Second, we added an average percent of recombination of 25% to the data (i.e. partial HGT). This process was simulated as a random recombination between the source and destination sequences. The new resulting sequence (i.e. mosaic gene) contained 75% of the source sequence and 25% of the destination sequence. We also considered the case of a maximum recombination rate of 50%, where the resulting mosaic sequence was a hybrid of 50% source and 50% destination sequence.

Every combination of the simulation parameters was tested with 50 replicates. The distribution of the obtained average results based on the Q functions ordering is shown in Figures 5.4 and 5.5. The additional results based on the p-value ordering, with a maximum threshold of 0.05, are shown in Figures 5.7 and 5.8.

5.5 Results and discussion

In this chapter we described a new algorithm for determining genomic regions that may be related to HGT and recombination, and introduced three new clustering functions Q_7, Q_{8a}, Q_{8b}. We compared the performances of these functions to those yielded by a simple distance measure Q_9. All of the considered aggregation functions were tested on the real-life genomic data (see figures 5.2 and 5.3) as well as on the synthetic data (see figures 5.4 and 5.5).

5.5.1 Analysis of prokaryotic data

For all the functions we introduced in this study, i.e. Q_7, Q_{8a}, Q_{8b} and Q_9, we can observe the following trend: better detection sensitivity corresponds to higher HGT bootstrap confidence thresholds. The function Q_7 always provided better results than Q_9, while Q_{8a} and Q_{8b} were better than Q_9 only for 75% and 90% bootstrap thresholds (based on the median values shown by a vertical black line on each of the boxes in Figure 5.2).

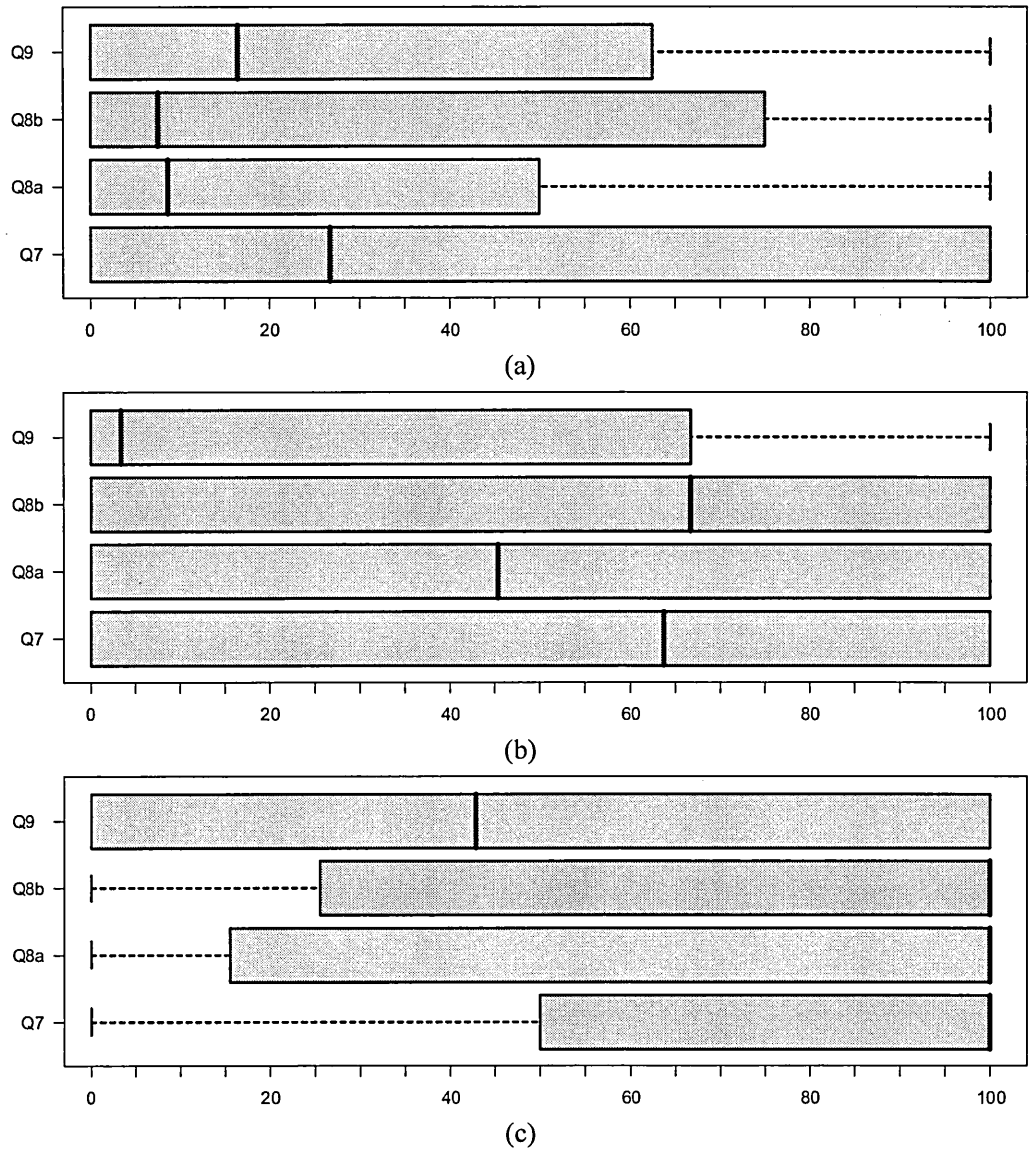


Figure 5.2 HGT-QFUNC sensitivity results for functions Q_7, Q_{8a}, Q_{8b} and Q_9 when detecting complete HGT in prokaryotic dataset based on Q-value ordering – boxplot representation

Abscissa represents the sensitivity percentage and ordinate represents the tested function. The median value is shown by a vertical black line within each box. The HGT-QFUNC algorithm was limited to the following maximum numbers of positive values:

(a) 300 HGTs (corresponds to 50% bootstrap support in the HGT-Detection algorithm);
(b) 200 HGTs (corresponds to 75% bootstrap support in the HGT-Detection algorithm);
(c) 100 HGTs (corresponds to 90% bootstrap support in the HGT-Detection algorithm).

The p-value based ordering, established with the threshold of 0.05, yields very good detection results for all of the tested functions Q_7, Q_{8a}, Q_{8b} and Q_9 (see Figure 5.6). The functions Q_{8a} and Q_{8b} provided better results than Q_9 for the HGT detection threshold of 50% bootstrap support. Moreover, the presented results suggest that the function Q_{8a} is able to detect almost all of high confidence HGT (90% bootstrap support). The main differences can be observed in the tail of the distribution, for the lower 25% quartile, as the median and high quartile are already at the same maximum value (of 100%). It should be noticed that for the 75% HGT detection threshold, which was our benchmark threshold throughout this thesis, the best average results were provided by the function Q_7.

One of the limitations of the HGT-QFUNC algorithm, compared to the HGT-Detection algorithm (Boc et al. 2010), is that our new algorithm imposes a fixed number of positive values (100, 200 or 300 in our case) regardless the number of species in the given multiple sequence alignment. These constant values were selected in order to find on average less than 2%, 4% and 6% of the maximum possible number of transfers between individual species for, respectively, 90%, 75% and 50% bootstrap support levels adopted by the HGT-Detection algorithm (see Figure 5.3).

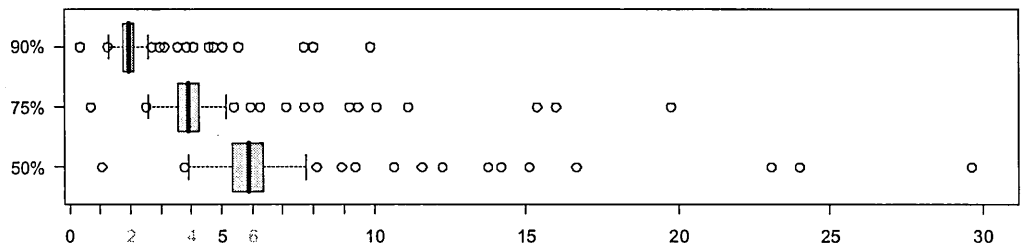


Figure 5.3 Distribution of the HGT-QFUNC maximum percentages of positive values chosen for prokaryotic data

Abscissa represents the percentage of the maximum possible number of HGTs between individual species. Ordinate represents the corresponding HGT-Detection bootstrap confidence level. Average values correspond to less than 6%, 4% and 2% of the maximum possible number of HGTs for the 50%, 75% and 90% bootstrap confidence levels, respectively.

5.5.2 Analysis of synthetic data

We also tested the detection sensitivity of our method for randomly generated HGTs between terminal tree branches using synthetically generated data and different levels of recombination. In the case of artificial data, the functions Q_7, Q_{8a} and Q_{8b} provided better performances than the function Q_9 only when recombination was considered. The results obtained for the Q-function ordering are shown in Figure 5.4 (for the 25% recombination level in the left column and for the 50% recombination level in the right one). Figure 5.5 presents in the left panel the results for HGT with no recombination (i.e. 0%) and in the right panel, the limits of our simulation, where there is no difference between the functions involved.

The p-value-based ordering, with the threshold of 0.05, shows only minor improvements, especially when the number of simulated transfers is low (i.e. 2, 4 and 8) (see Figures 5.7 and 5.8). Specifically, for 1 simulated transfer we obtained an almost perfect detection rate, while for 128 transfers we obtained the worst HGT recovery rates that were around 40%.

Thus, the following general trend can be observed: the higher number of transfers we have, the lower detection rates are. Higher degrees of recombination also lead to a lower detection rate for all the functions, but favor the functions Q_{8a} and Q_{8b}, as their performance degrades less, especially in the middle range. The function Q_7, which showed very good performances for the real-life prokaryotic data, does not outperform the function Q_9 in this particular testing framework. It is important to notice that even without recombination the functions Q_{8a} and Q_{8b} can be also used as they yield almost the same detection rates as the function Q_9.

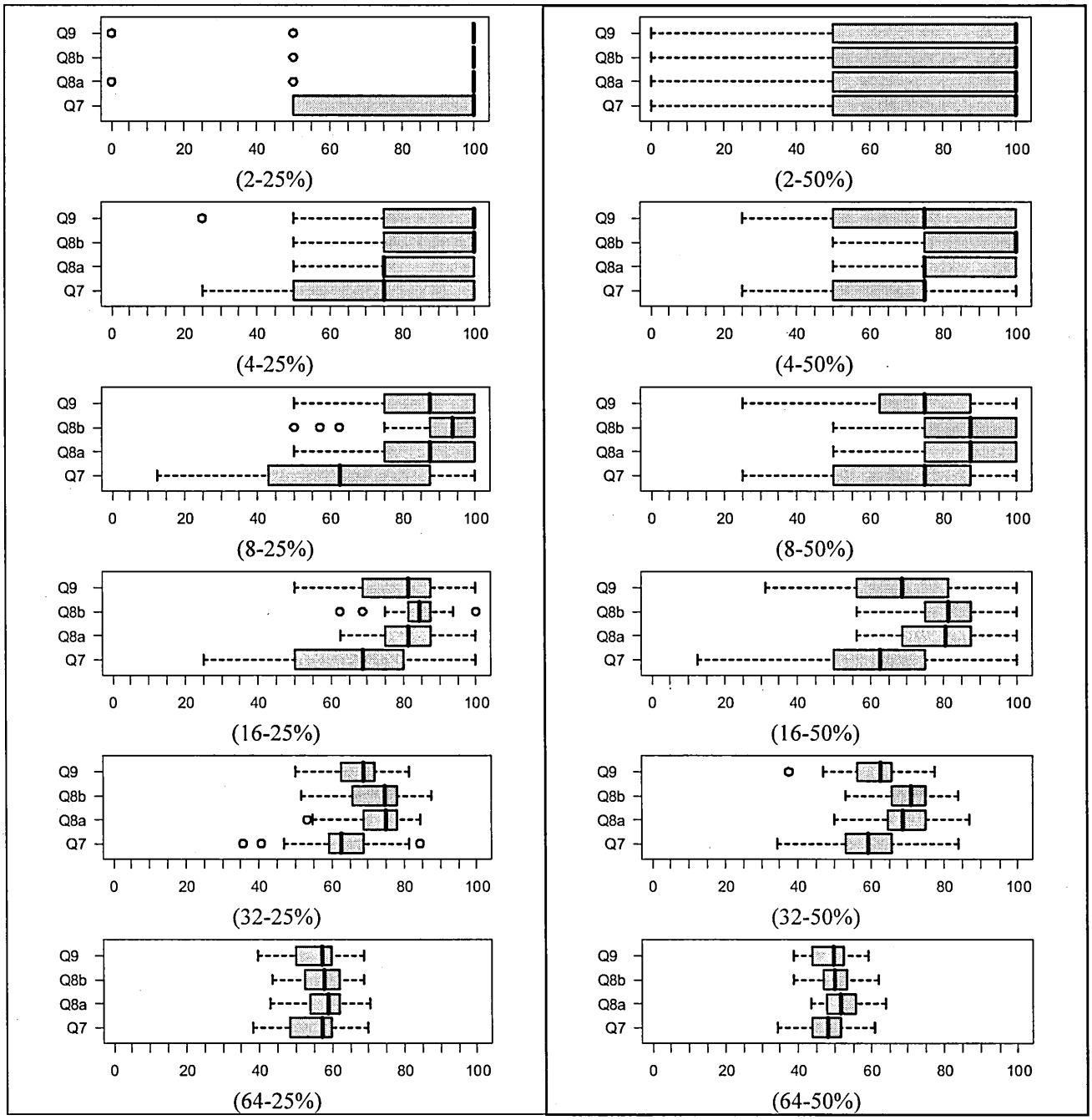


Figure 5.4 HGT-QFUNC sensitivity results for functions Q_7, Q_{8a}, Q_{8b} and Q_9 when detecting partial HGT in synthetic dataset based on Q-value ordering – boxplot representation

Abscissa represents the sensitivity percentage and ordinate represents the tested function. The median value is shown by a vertical black line within each box. Simulations for 2, 4, 8, 16, 32 and 64 random nonreciprocal sequence transfers between prokaryotic species (first value between parentheses) were carried out. Average simulation results under the medium degree of recombination (when 25% of the resulting sequence belong to one of the parent sequences) are depicted in the left panel. Average simulation results under the highest level of recombination (when 50% of the resulting sequence belong to the source sequence and 50% to the destination sequence) is depicted in the right panel. For each dataset, the maximum allowed number of positive values was the double of the number of transfers (i.e. 4, 8, 16, 32, 64 and 128, respectively). Calculations were done over 50 replicates for each parameters combination.

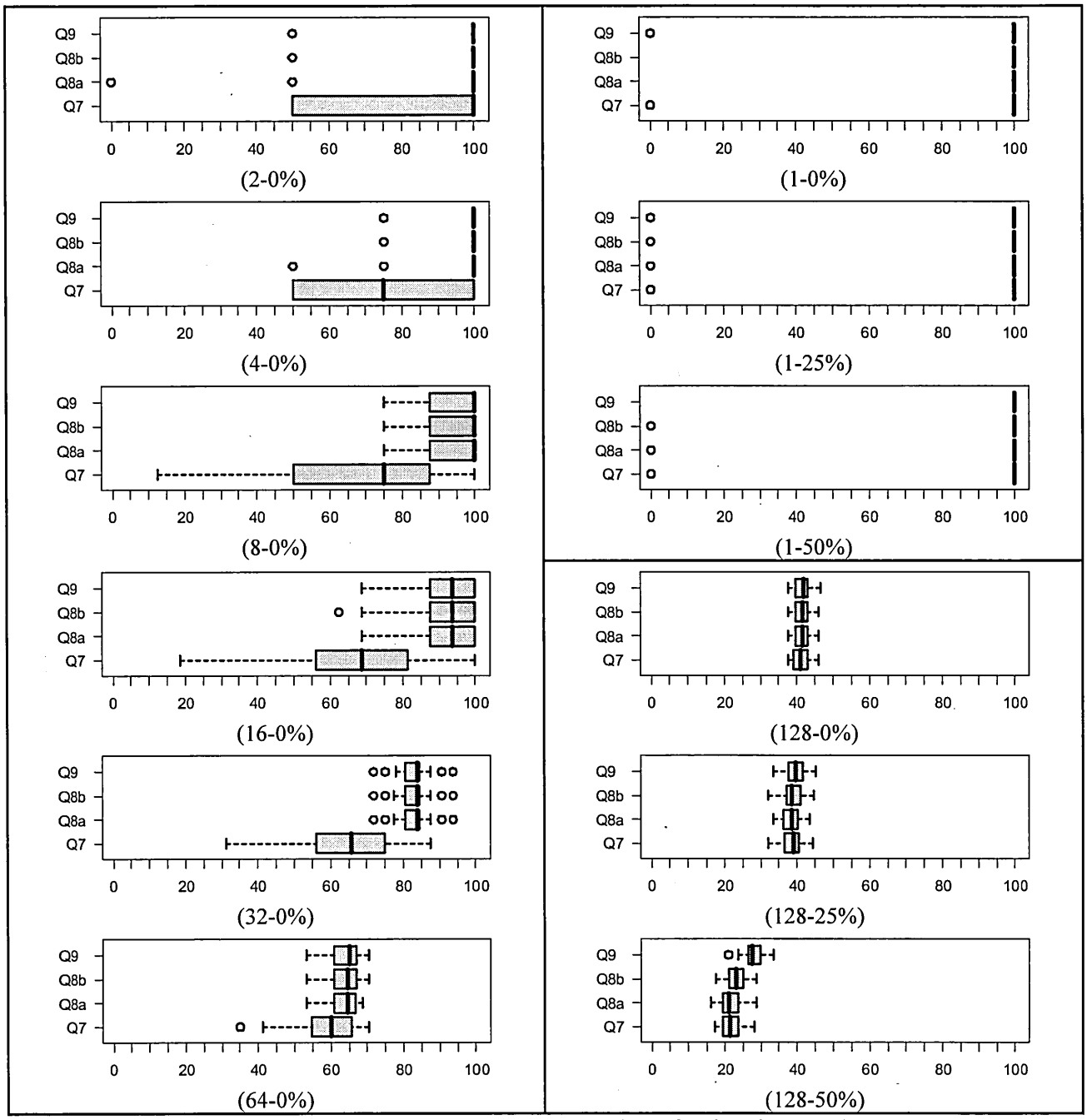


Figure 5.5 Remaining HGT-QFUNC sensitivity results for functions Q_7, Q_{8a}, Q_{8b} and Q_9 when detecting complete and partial HGT in synthetic dataset based on Q-value ordering – boxplot representation

Abscissa represents the sensitivity percentage and ordinate represents the tested function. The median value is shown by a vertical black line within each box. Simulations for 2, 4, 8, 16, 32 and 64 random nonreciprocal sequence transfers between prokaryotic species (first value between parentheses) were carried out. Average simulation results for data without recombination are depicted in the left panel. Right panel depicts the results of the same simulations, for the cases of 1 and 128 transfers, with recombination levels of 0% (no recombination), 25% and 50%. Average simulation results under the highest level of recombination (when 50% of the resulting sequence belong to the source sequence and 50% to the destination sequence) is depicted in the right panel. For each dataset, the maximum allowed number of positive values was the double of the number of transfers (i.e. 4, 8, 16, 32, 64 and 128, respectively). Calculations were done over 50 replicates for each parameters combination.

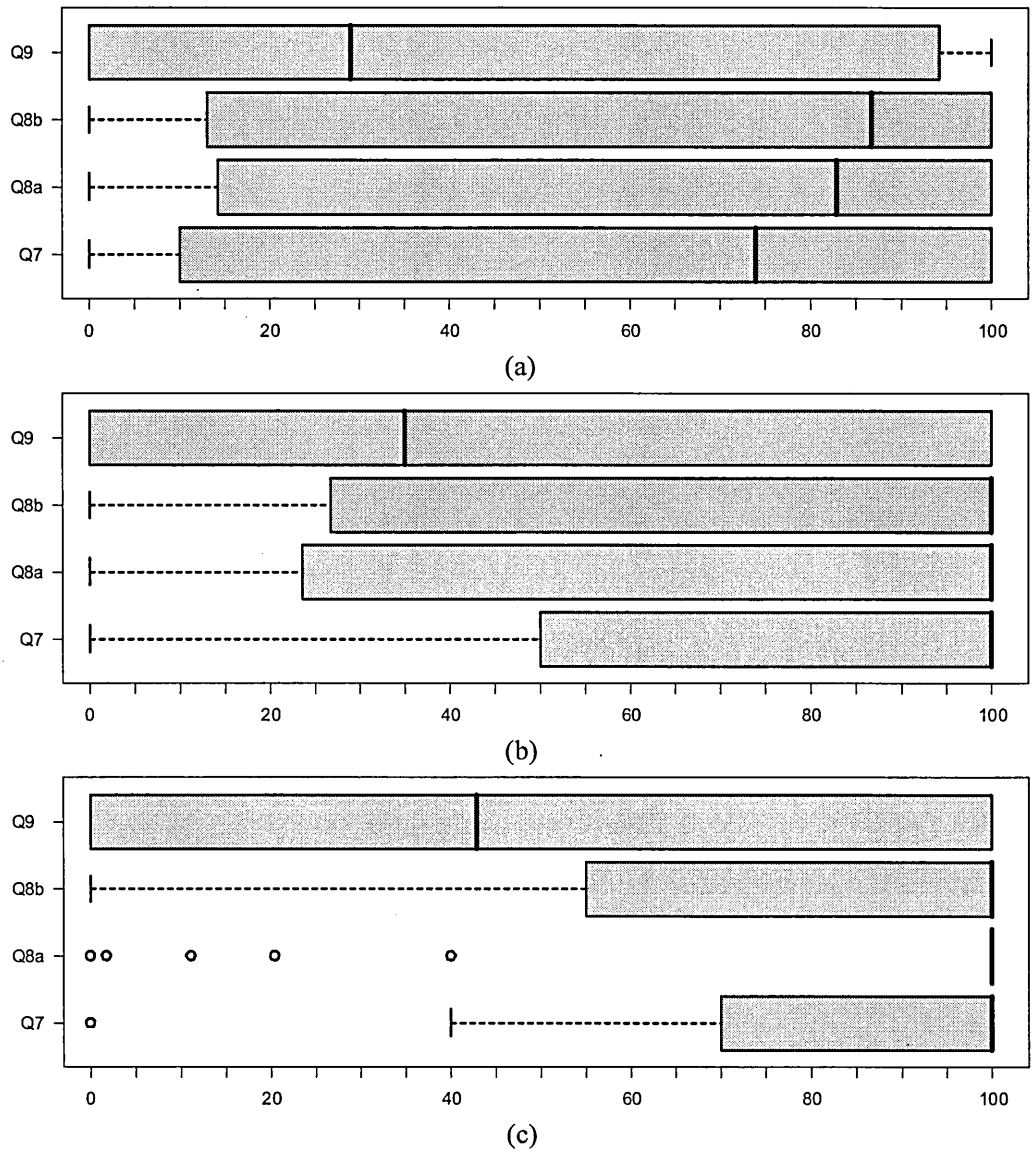


Figure 5.6 HGT-QFUNC sensitivity results for functions Q_7, Q_{8a}, Q_{8b} and Q_9 when detecting complete HGT in prokaryotic dataset based on *p*-value ordering (maximum *p*-value of 0.05) – boxplot representation

Abscissa represents the sensitivity percentage and ordinate represents the tested function. The median value is shown by a vertical black line within each box. The HGT-QFUNC algorithm was limited to the following maximum numbers of positive values:

(a) 300 HGTs (corresponds to 50% bootstrap support in the HGT-Detection algorithm);
(b) 200 HGTs (corresponds to 75% bootstrap support in the HGT-Detection algorithm);
(c) 100 HGTs (corresponds to 90% bootstrap support in the HGT-Detection algorithm).

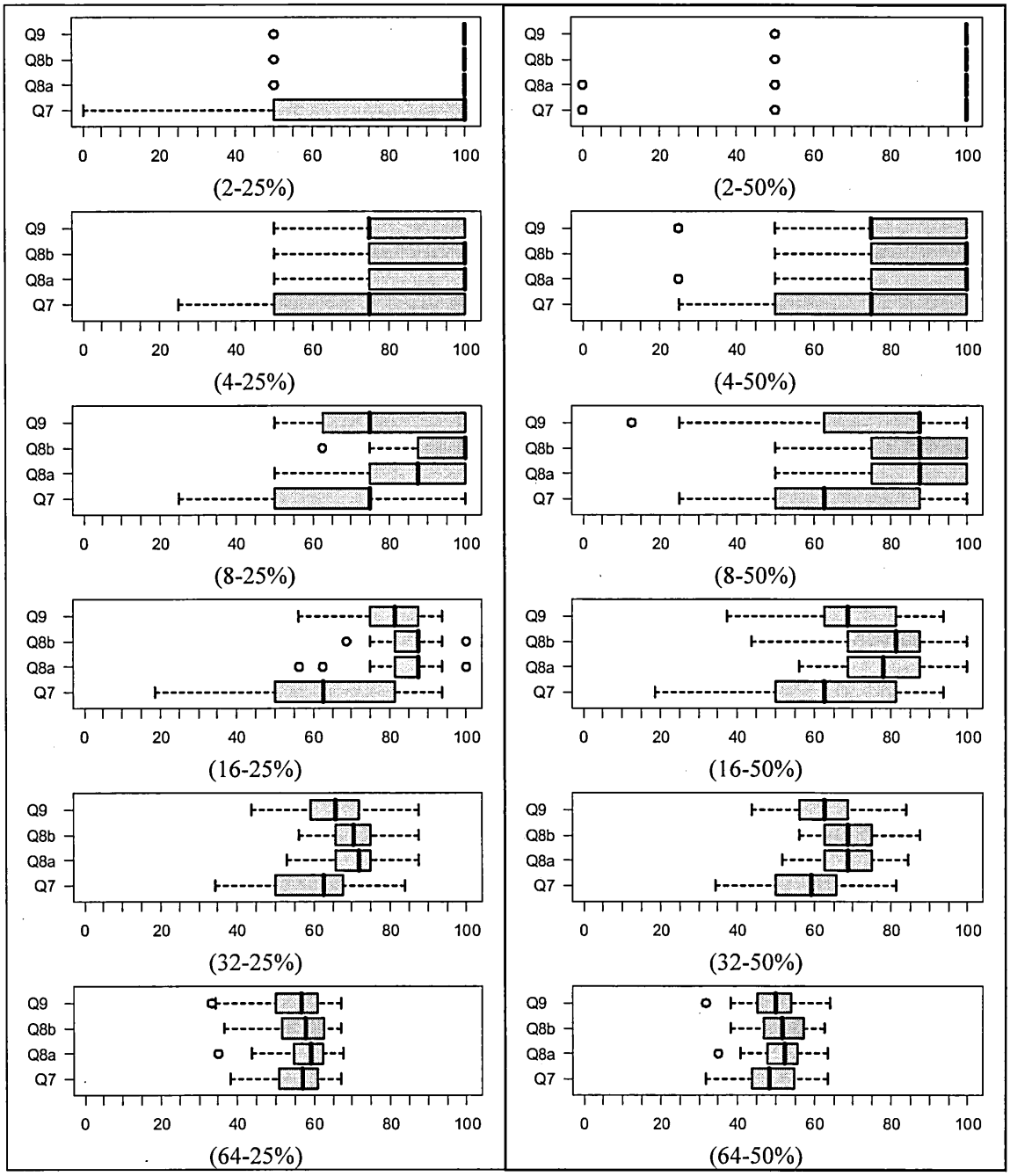


Figure 5.7 HGT-QFUNC sensitivity results for functions Q_7, Q_{8a}, Q_{8b} and Q_9 when detecting partial HGT in synthetic dataset based on *p*-value ordering (maximum *p*-value of 0.05) – boxplot representation

Abscissa represents the sensitivity percentage and ordinate represents the tested function. The median value is shown by a vertical black line within each box. Simulations for 2, 4, 8, 16, 32 and 64 random nonreciprocal sequence transfers between prokaryotic species (first value between parentheses) were carried out. Average simulation results under the medium degree of recombination (when 25% of the resulting sequence belong to one of the parent sequences) are depicted in the left panel. Average simulation results under the highest level of recombination (when 50% of the resulting sequence belong to the source sequence and 50% to the destination sequence) is depicted in the right panel. For each dataset, the maximum allowed number of positive values was the double of the number of transfers (i.e. 4, 8, 16, 32, 64 and 128, respectively). Calculations were done over 50 replicates for each parameters combination.

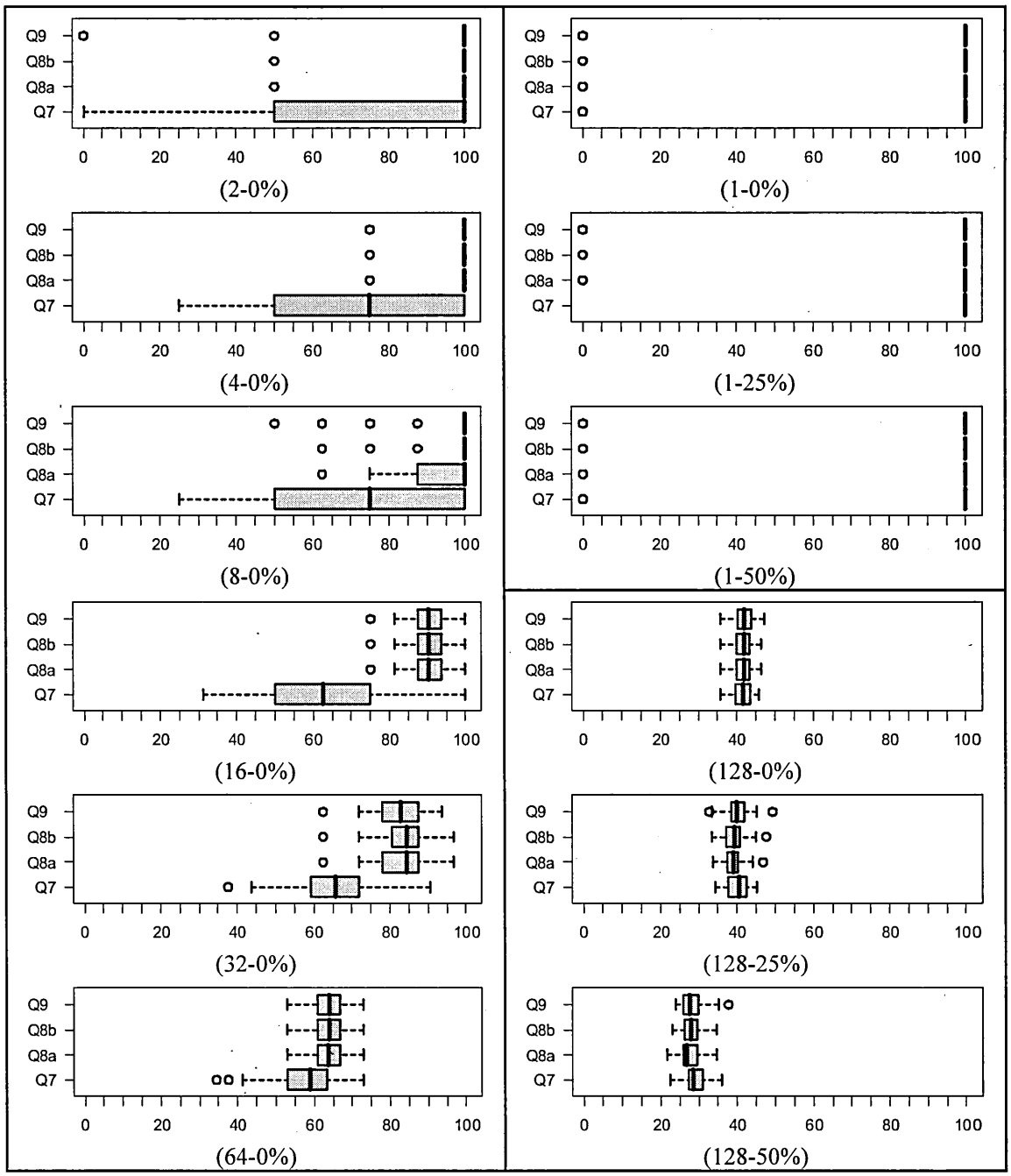


Figure 5.8 Remaining HGT-QFUNC sensitivity results for functions Q_7, Q_{8a}, Q_{8b} and Q_9 when detecting complete and partial HGT in synthetic dataset based on *p*-value ordering (maximum *p*-value of 0.05) – boxplot representation

Abscissa represents the sensitivity percentage and ordinate represents the tested function. The median value is shown by a vertical black line within each box. Simulations for 2, 4, 8, 16, 32 and 64 random nonreciprocal sequence transfers between prokaryotic species (first value between parentheses) were carried out. Average simulation results for data without recombination are depicted in the left panel. Right panel depicts the results of the same simulations, for the cases of 1 and 128 transfers, with recombination levels of 0% (no recombination), 25% and 50%. Average simulation results under the highest level of recombination (when 50% of the resulting sequence belong to the source sequence and 50% to the destination sequence) is depicted in the right panel. For each dataset, the maximum allowed number of positive values was the double of the number of transfers (i.e. 4, 8, 16, 32, 64 and 128, respectively). Calculations were done over 50 replicates for each parameters combination.

5.6 Conclusion

Horizontal gene transfer is a well-structured evolutionary paradigm as the recent studies show higher levels of transfers between certain prokaryotic groups (Beiko et al. 2005) or certain ecological habitats (Smillie et al. 2011). The impact of horizontal transfers on the creation of many prokaryotes and viruses, as well as the cumulative effect of recombination over multiple generations, remains to be investigated in greater detail.

Despite the general availability of quality controlling HGT detection methods based on complex phylogenetic analyses, simple distance measures can still be useful for recovering HGT events. The computational complexity of more precise HGT detection methods as well as the high volume of considered genomic data are the main motivations behind the development of fast and effective HGT detection algorithms.

In this chapter we described a new fast HGT detection algorithm which runs in quadratic time when HGTs between terminal branches are considered. It allows for an efficient parallel implementation. The discussed method also benefits from a Monte Carlo p-value validation procedure, obviously at the cost of the associated validation constant needed for maintaining precision. Because of its low time complexity, the new algorithm can be used in complex phylogenetic and genomic studies involving thousands of species. Mention that the Hanssen-Kuipers Skill Score (Hanssen and Kuipers 1965), allowing for decomposing different sources of error, could be used instead of sensitivity for measuring the performances of our algorithm.

Even though the presented method is designed to identify complete HGT, we investigated how it copes with partial HGT (i.e. HGT followed by the intragenic sequence recombination) and showed that in many cases it can be used to identify both complete and partial HGT.

The new variability clustering functions Q_7, Q_{8a}, Q_{8b} and Q_9 were introduced and tested in our simulations. We also tested the function in the context of complete and partial HGT recovery. In overall, the functions Q_7 and Q_{8a} provided the best HGT detection performances.

We have provided the complete source code of our application allowing one to carry out the new method for detecting complete horizontal gene transfer events discussed in this chapter; the application's name is HGT-QFUNC.v.0.5.2. A Makefile along with the examples of the input and output data have been also made available. The ReadMe documentation file provides an explanation

of the main steps to follow for executing the application. The source code and the accompanying files have been uploaded to the GitHub public repository (with the BSD licence). It is freely available at the following URL address:

https://github.com/dunarel/dunphd-thesis/tree/master/Chapter5/Main/hgt-qfunc.v.0.5.2.

CONCLUSION AND PERSPECTIVES

The detection of functional genomic regions is a fundamental goal of genetic research. Our understanding of relations between genetic structure and biological function is instrumental for developing new drug targets, genetic treatments as well as for improving biotechnological engineering (e.g. for better livestock management or better productivity of food microorganisms).

Neutral theory of DNA evolution has provided a theoretical basis for the development of a wealth of methods aimed at detecting selection and its multiples modes, by using statistical significance tests against neutrality. Negative selection, which is associated with essential biological functions common to all individuals of the same family, is very well studied and can be detected by studying sequence conservation patterns.

On the other hand, organisms engaging in host-parasite relationships are subject to accelerated evolution, where they are forced to develop variability of their populations in order to survive. Eukaryotes usually play the role of hosts, while prokaryotes and viruses play the role of parasites in this relationship.

Many statistical methods exist for the detection of regions evolving under a different pattern than that of neutral evolution (see chapters I-II). On the host side, many methods have been designed to take advantage of the particular diploid nature of human DNA. On the parasite side, phylogenetic models have been extensively used. Mathematical modeling of these concepts usually results in formulating NP-complete problems. Moreover, the heuristics developed for solving these problems involve high computational costs.

We dedicated our thesis to the development of new computational methods taking advantage of several well-known classification criteria, such as pathogenic factors (carcinogenicity, invasivity), phylogenetic families or ecological habitats (i.e. working at both taxonomic and ecological levels).

As we could see in chapter III, the combination of sequence analysis with aggregation functions can allow for fast detection of major differences across known groups as well as for discovering general data patterns. The modern comparative genomics has introduced huge real-life datasets. This explosion of data means that only a limited number of computational biology methods can be used at a large scale, namely those with low computational complexity.

In chapter III, we described a novel algorithm intended for optimizing species clustering into two groups. This methodology could be further improved to account for three or more species groups (i.e. considering three or more clustering criteria). Our algorithm uses a *k-means*-like principle to move elements between groups. A possible future development could include the use of a *fuzzy c-means* clustering criterion to allow for elements belonging to several distinct groups. Another possible improvement of the presented algorithm could consist of an alternate bipartition optimization using those bipartitions that are already present in reference trees.

In chapter IV, we presented a comprehensive genomic study of prokaryotes in order to detect both complete and partial horizontal gene transfer events between the contemporary prokaryotic organisms as well as between their ancestors. To the best of our knowledge, we first applied the HGT bootstrap validation methods on such a large genomic scale. We put in place a weighted statistical scheme to account for individual HGT events as well as for the selected species clusters (i.e. species classifications). In our study, we considered species classifications according to phylogenetic families and ecological habitats, but this methodology could be easily expanded in the future to other readily available classifications, such as molecular functions, protein structures or cellular locations. It is worth noting that when the habitat classification was considered, one individual prokaryotic strain could belong to one or multiple habitats. Thus, in the future, we could consider *fuzzy classifiers*, which are able to take into account such combined classifications.

We also carried out local and global HGT interaction rate analysis using different normalization schemes. In the future, we plan to develop these ideas in the context of a comprehensive probabilistic framework.

Mention that we not only estimated a global HGT rate characterizing the evolution of prokaryotes, but also provided a good level of detail, using taxonomic clustering by strain and by species. Another level of detail was provided by detecting the size of the transferred genetic material: a whole gene (i.e. complete HGT) or a part of the gene (i.e. partial HGT, which an HGT followed by

intragenic recombination and leading to formation of mosaic genes). Our results confirm the arguments in favor of the continuous nature of the HGT phenomenon and its ubiquity even at the core genes, the most conserved and restrictive to horizontal transfers according to some works. A similar type of study could be conducted in the future on viruses and bacteriophages in order to compare their respective HGT rates with those obtained in the case of prokaryotes.

We also presented for the first time at a large scale the intergroup relationships and dating results for partial horizontal transfers between prokaryotes and compared the overall (complete + partial) HGT rates to the complete gene transfer rates (a common case of HGT analysis). In the future, it would be interesting to compare the partial HGT results with those obtained by the methods used to detect recombination.

We estimated the precise timing of the detected complete and overall HGTs by using dated gene trees. In the future, it would be interesting to integrate our fast distance-based HGT detection method (see chapter V) into the HGT-timing framework in order to reduce the algorithmic complexity of the HGT age estimation. The HGT detection algorithm introduced in chapter V can be also used in complex phylogenetic and genomic studies involving thousands of species because of its quadratic time complexity, on the number of species, in most of the practical situations.

In two chapters of this thesis (III and V), we considered new sequence aggregation functions. Chapter III corresponds to the classic view of mutation as the main diversifying force, paired with selection as the unifying one. There, we focused on the study of positive selection and lineage specific selection processes. On the contrary, in chapter V, we brought arguments in favor of the new emerging view of evolution in the prokaryotic world, in which HGT is considered as the main diversifying force. In chapter IV, we also explored the relation of HGT to speciation through clusters of HGT-related habitats.

In the future, it would be interesting to investigate other existing evolutionary events such as small insertions and deletions, for example. They are usually overlooked and deleted from multiple sequence alignments, as a part of quality assurance steps, because they can be easily confounded with alignment errors.

We showed that the introduced aggregation functions (see chapters III and V) have different sensitivities in presence of different modes of evolution but also different species clustering types

(monophyletic or polyphyletic). We suggested that several aggregation functions should be used in combination for detecting genetic regions responsible for pathogenicity. They could be also used in order to detect signatures of immunological features. This conclusions comes together with the observation that the new algorithm presented in chapter III could be used to detect linear epitopes, and thus be useful to vaccine design.

APPENDIX A

FULL CONTENTS OF THE CHAPTER III SUPPLEMENT (ADDITIONAL FILE 1)

Algorithm A.1.

Algorithm for computing genomic regions responsible for carcinogenicity or invasivity

---

Require: *FI*: Hit region identification function to be optimized Q_4, Q_5 or Q_6,

MSA : Multiple sequence alignment,

X : Subset of carcinogenic or invasive taxa,

Y : Subset of non-carcinogenic or non-invasive taxa,

WIN_MIN : Minimum sliding window width,

WIN_MAX : Maximum sliding window width,

S : Sliding window step,

RPG : Constant number of random bipartition generations.

Ensure: Set of Hit Regions: (*win_width,idx,Q',ARI,Q''*), where

win_width : Current sliding window width,

idx : Hit Index (i.e., its genomic position),

Q' : Hit region identification function without knowledge of X and Y,

ARI : Adjusted Rand index,

Q'' : Validation function depending on ARI.

1: MSA_L ← Length of MSA
2: **for** win_width **from** WIN_MIN **to** WIN_MAX **do**
3: **for** idx **from** 0 **to** *MSA_L−win_width* **with step** *S* **do**
4: **for all** *r* such that $1 \leq r \leq RPG$ **do**
5: Randomly select a bipartition A|B
6: $MSA_A \leftarrow MSA[A][idx..idx + win\_width]$
7: $MSA_B \leftarrow MSA[B][idx..idx + win\_width]$
8: $Q_{Partition}$ = Calculate $Q(FI, A, B, MSA_A, MSA_B)$
9: Update $Q' = \text{Max}(Q_{Partition}, r)$
10: **repeat**
11: **for all** $i \in A$ **do** // i is randomly chosen
12: $A \leftarrow A \setminus i$, $B \leftarrow B \cup i$
13: Update $Q_{Partition}$
14: $\text{MaxQ}_{\text{Partition}} = \text{Max}(\text{Max}Q_{Partition}, Q_{Partition}, r)$
15: keep old A and B if $\text{Max}Q_{Partition}$ is unchanged
16: **end for**
17: Swap(A,B)
18: **until** No improvement of $MaxQ_{Partition}$ is possible
19: $Q'[win\_width,idx] = \text{Max}(Q', \text{Max}Q_{Partition})$
20: $ARI[win\_width,idx]$ = Calculate ARI(A|B, X|Y)
21: $Q''[win\_width, idx] = ARI \times Q'$
22: **end for**
23: **end for**
24: **end for**
25: **return** Q', ARI, Q''

Figure A.1. p-values obtained for hit region detection using the remaining (i.e., not presented in Figures 3.2 and 3.3) Q'-type functions

(a),(b),(c),(d) Monophyletic evolution - (e),(f),(g),(h) Polyphyletic evolution

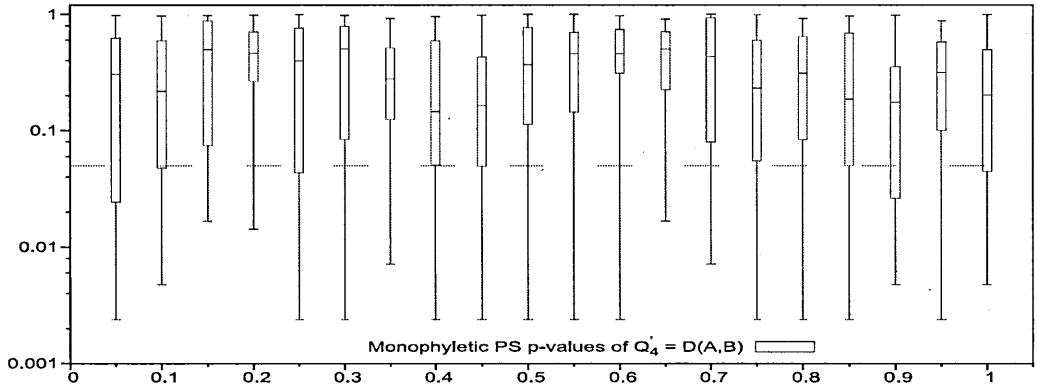
(a),(c),(e),(g) Positive selection - Variable hit region inside conserved context.

Quartile distribution of p-values obtained for the functions Q_4'(a), Q_6'(c), Q_5'(e), and Q_6'(g). Abscissa represents scaling factor of the conserved context in which the variable hit region resides. Values close to 0 represent conservation (maximum discrimination), while values close to 1 represent variability (identical to context). Variable hit region is always maintained at a scaling factor of 1. Ordinate represents p-values in log-scale. Horizontal dashed line represents the significance threshold of 0.05.

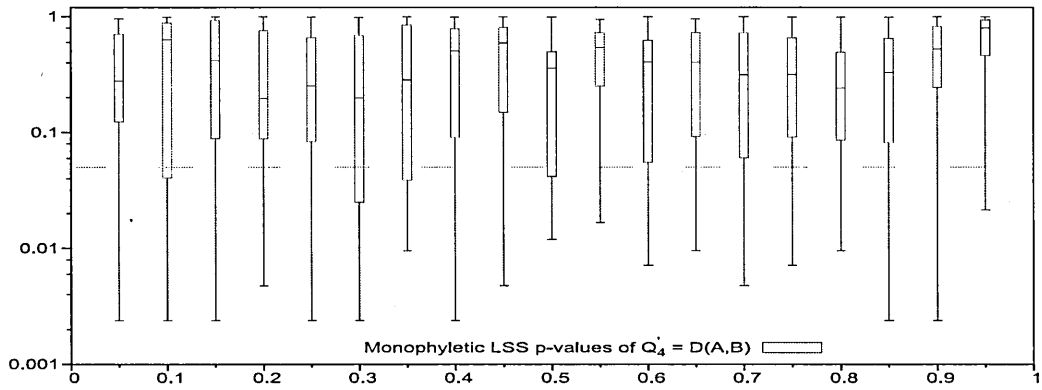
(b),(d),(f),(h) Lineage specific selection - Heterogeneous hit region inside neutral context.

Quartile distribution of p-values obtained for the functions Q_4'(b), Q_6'(d), Q_5'(f), and Q_4'(h). Abscissa represents the difference in scaling factors among the two lineages present in the hit region. Values close to 0 represent homogeneous evolutionary speed (similar to the neutral context in which it resides), while values close to 1 represent divergence among these lineages. Context is always maintained at a scaling factor of 0.5, simulating neutral evolution. Horizontal dashed line represents the significance threshold of 0.05. In the case of lineage specific selection, the value of the Q'-type functions corresponding to 1 on the abscissa scale cannot be computed because it involves a sub-tree with 0 edge lengths.

Figure A.1 (a), (b) – Remaining monophyletic evolution hit detection p-values

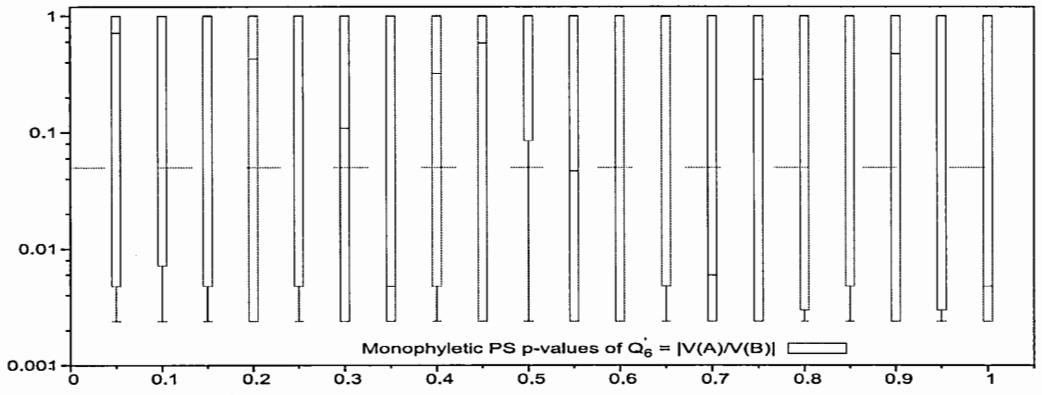


(a)

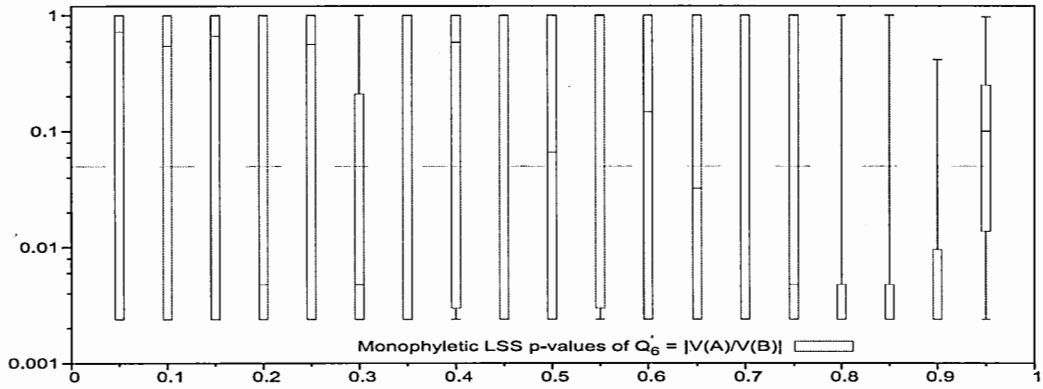


(b)

Figure A.1 (c), (d) – Remaining monophyletic evolution hit detection p-values

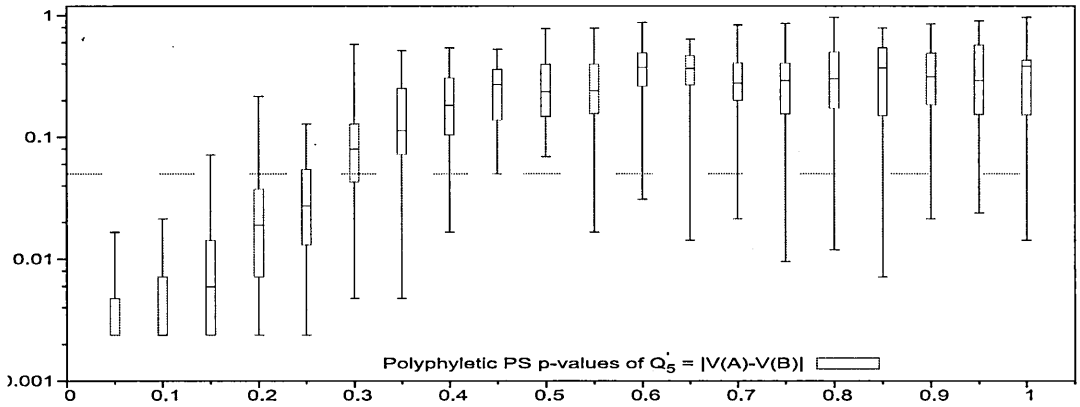


(c)

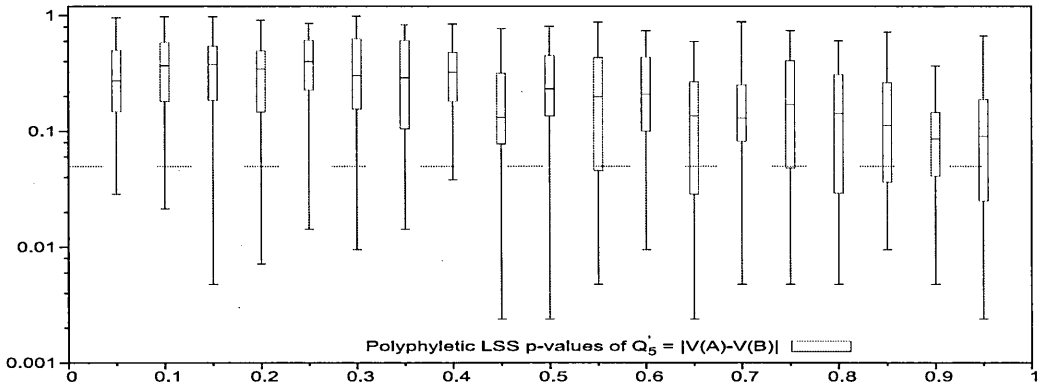


(d)

Figure A.1 (e), (f) – Remaining polyphyletic evolution hit detection p-values

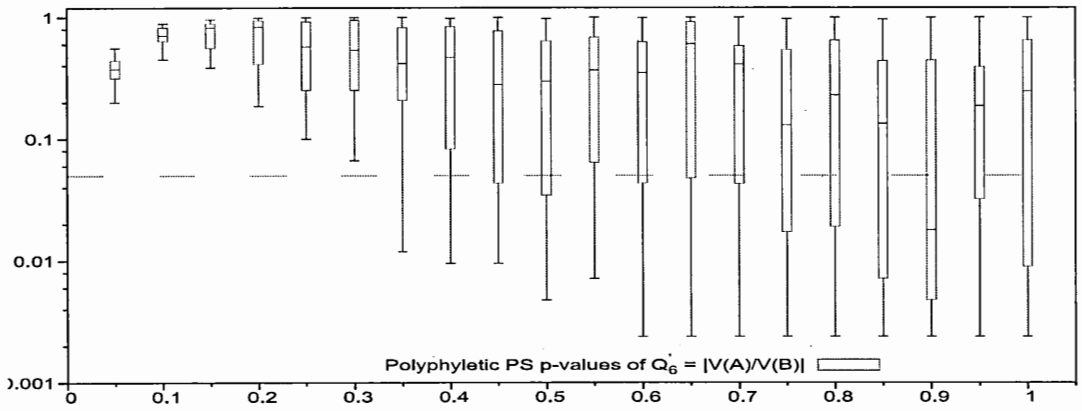


(e)

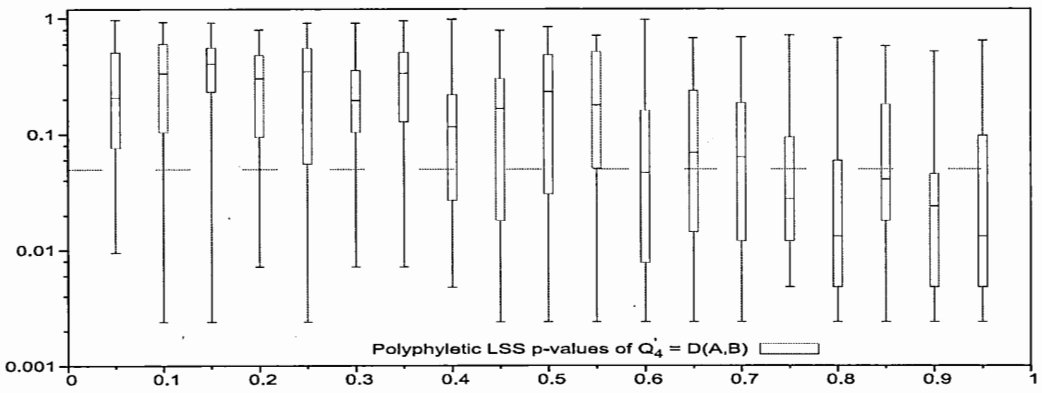


(f)

Figure A.1 (g), (h) – Remaining polyphyletic evolution hit detection p-values



(g)



(h)

APPENDIX B

IMPLEMENTATION DETAILS OF CHAPTER IV CLUSTERING

Let T(g) be the total number of HGT detected for the gene g. Each transfer is occurring between two branches of the associated Gene Tree. We define U and V, as the set of alleles of the associated subtrees. Accordingly, the transfer is considered between the most recent common ancestors (MRCA) of the associated node of the source and respectively destination branches.

We represent the i^{th} HGT of gene g, as a transfer matrix called K, defined over the Cartesian product of the source and destination alleles. A is the vector of alleles present in all multiple sequence alignments. G is the vector of genes.

$K(g,i) = [|A| \times |A|]$, where $g \in G$, and $a,b \in A$. Its values are defined with following formula:

$$K_{a \to b}(g,i) = \begin{cases} 1/|U| \cdot |V|, if\ a \in U \wedge b \in V, \\ 0, \qquad\qquad otherwise, \end{cases} \tag{B.1}$$

For the same HGT, we define an associated matrix W, between prokaryotic families. $W = [|P| \times |P|]$, where P is the vector of prokaryotic families. We can calculate its values using following formula:

$$W_{F2 \to F1}(g,i) = R \times Z \times K(g,i) \times Z^T \times R^T \tag{B.2}$$

This matrix is obtained by weighting the previous transfer matrix by two other matrices, corresponding to classifications. These are the allele membership to the species (Z) and the species membership to the prokaryotic families (R). These matrices are defined as follow:

Let S be the vector of species.

Let $V[|P| \times |S|]$ be the family-species presence matrix.

$$\forall p \in P \wedge s \in S, V_{p,s} = \begin{cases} 1, if\ s\ is\ classified\ as\ belonging\ to\ p \\ 0, otherwise \end{cases} . \tag{B.3}$$

We obtain the weighted matrix $R = [|P| \times |S|]$, by dividing each element by the number of groups to which each corresponding species belong:

$$R_{p,s} = V_{p,s} / \sum_{i \in 1}^{|P|} V_{i,s} \tag{B.4}$$

Let $M = [1...|P|]$ be the row vector of number of species belonging to each family p of P. See equations (10)-(15).

$$M_p = \sum_{s=1}^{|S|} V_{p,s} \tag{B.5}$$

Let $Z[|S| \times |A|]$ be the species-alleles association matrix:

$$\forall s \in S \wedge a \in A, Z_{s,a} = \begin{cases} 1, if\ a\ is\ an\ allele\ of\ s \\ 0, otherwise \end{cases} \tag{B.6}$$

Let $L[|A| \times |G|]$ be the alleles-genes association matrix:

$$\forall a \in A \wedge \forall g \in G\ and\ L_{a,g} = \begin{cases} 1, if\ a \in MSA(g) \\ 0, otherwise \end{cases} \tag{B.7}$$

Let $H = [|P| \times |G|]$ be the family-gene weight-count matrix: $H = R \times Z \times L$. Using a different notation, $N_{F1}(g) = H_{F1,g}$, that is used in equations (4.1) and (4.3).

Let $N = [1...|P|]$ be the row vector of number of alleles belonging to group p. See equations (4),(6),(8) and (9).

$$N_p = \sum_{g=1}^{|G|} H_{p,g} \tag{B.8}$$

REFERENCES

Abby, S. S., Tannier, E., Gouy, M. and Daubin, V. (2012), 'Lateral gene transfer as a support for the tree of life', *Proceedings of the National Academy of Sciences* **109**(13), 4962–4967.

Abdullah, A. and Hussain, A. (2006), 'A new biclustering technique based on crossing minimization', *Neurocomputing* **69**, 1882–1896.

Achaz, G., Rocha, E. P., Netter, P. and Coissac, E. (2002), 'Origin and fate of repeats in bacteria', *Nucleic Acids Research* **30**(13), 2987–2994.

Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V. and Polz, M. F. (2004), 'Divergence and redundancy of 16s rrna sequences in genomes with multiple rrn operons', *Journal of bacteriology* **186**(9), 2629–2635.

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

Angelis, K., dos Reis, M. and Yang, Z. (2014), 'Bayesian estimation of nonsynonymous/synonymous rate ratios for pairwise sequence comparisons', *Molecular Biology and Evolution* .

Angiuoli, S. V., Gussman, A., Klimke, W., Cochrane, G., Field, D., Garrity, G. M., and White, O. (2008). 'Toward an online repository of Standard Operating Procedures (SOPs) for (meta) genomic annotation'. *OMICS A Journal of Integrative Biology*, 12(2), 137-141.

Ángulo, M. and Carvajal-Rodrguez, A. (2007), 'Evidence of recombination within human alpha-papillomavirus', *Virology Journal* **4**, 33.

Asthana, S., Roytberg, M., Stamatoyannopoulos, J. and Sunyaev, S. (2007), 'Analysis of sequence conservation at nucleotide resolution', *PLoS Computational Biology* **3**(12), e254.

Aulchenko, Y. S., De Koning, D.-J. and Haley, C. (2007), 'Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis', *Genetics* **177**(1), 577–585.

Badescu, D., Diallo, A., Blanchette, M. and Makarenkov, V. (2008), An evolutionary study of the human papillomavirus genomes, *in* 'Proceedings of RECOMB Comparative Genomics 2008', Vol. 5267 of *LNCS*, Springer, pp. 128–142.

Badescu, D., Diallo, A. and Makarenkov, V. (2010), 'Identification of specific genomic regions responsible for the invasivity of Neisseria Meningitidis', *Classification as a Tool for Research* pp. 491–499.

Baltimore, D. (2003), 'Rna-dependent dna polymerase in virions of rna tumour viruses', *A century of Nature: twenty-one discoveries that changed science and the world* p. 173.

Bansal, M. S., Alm, E. J. and Kellis, M. (2012), 'Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss', *Bioinformatics* **28**(12), i283–i291.

Bansal, M. S., Banay, G., Gogarten, J. P. and Shamir, R. (2011), 'Detecting highways of horizontal gene transfer', *Journal of Computational Biology* **18**(9), 1087–1114.

Bansal, M. S., Banay, G., Harlow, T. J., Gogarten, J. P. and Shamir, R. (2013), 'Systematic inference of highways of horizontal gene transfer in prokaryotes', *Bioinformatics* **29**(5), 571–579.

Barnosky, A. D. (2001), 'Distinguishing the effects of the red queen and court jester on miocene mammal evolution in the northern rocky mountains', *Journal of Vertebrate Paleontology* **21**(1), 172–185.

Barthélemy, J.-P. and Guénoche, A. (1991), *Trees and proximity representations*, John Wiley & Sons.

Battistuzzi, F. U., Feijao, A. and Hedges, S. B. (2004), 'A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land', *BMC evolutionary biology* **4**(1), 44.

Bear, R. and Rintoul, D. (2014), 'Taxonomy and phylogeny.'. http://cnx.org/contents/12696f5e-80cb-4399-9449-b753db45280b@5.

Becerra, A., Delaye, L., Islas, S. and Lazcano, A. (2007), 'The very early stages of biological evolution and the nature of the last common ancestor of the three major cell domains', *Annual Review of Ecology, Evolution, and Systematics* pp. 361–379.

Beiko, R. G. and Hamilton, N. (2006), 'Phylogenetic identification of lateral genetic transfer events', *BMC evolutionary biology* **6**(1), 15.

Beiko, R. G., Harlow, T. J. and Ragan, M. A. (2005), 'Highways of gene sharing in prokaryotes', *Proceedings of the National Academy of Sciences of the United States of America* **102**(40), 14332–14337.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Sayers, E. W. (2009), 'Genbank', *Nucleic Acids Research* **37**(suppl 1), D26–D31.

Benton, M. J. (2009), 'The red queen and the court jester: species diversity and the role of biotic and abiotic factors through time', *Science* **323**(5915), 728–732.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E. et al. (2007), 'Identification and analysis of functional elements in 1% of the human genome by the encode pilot project', *Nature* **447**(7146), 799–816.

Boc, A. and Makarenkov, V. (2011), 'Towards an accurate identification of mosaic genes and partial horizontal gene transfers', *Nucleic acids research* **39**(21), e144–e144.

Boc, A., Makarenkov, V. et al. (2012), 'T-rex: a web server for inferring, validating and visualizing phylogenetic trees and networks', *Nucleic acids research* **40**(W1), W573–W579.

Boc, A., Philippe, H. and Makarenkov, V. (2010), 'Inferring and validating horizontal gene transfer events using bipartition dissimilarity', *Systematic biology* **59**(2), 195–211.

Boni, M. F., Posada, D. and Feldman, M. W. (2007), 'An exact nonparametric method for inferring mosaic structure in sequence triplets', *Genetics* **176**(2), 1035–1047.

Boni, M. F., Zhou, Y., Taubenberger, J. K. and Holmes, E. C. (2008), 'Homologous recombination is very rare or absent in human influenza a virus', *Journal of virology* **82**(10), 4807–4811.

Bordewich, M. and Semple, C. (2005), 'On the computational complexity of the rooted subtree prune and regraft distance', *Annals of Combinatorics* **8**(4), 409–423.

Bosch, F., Manos, M., Muñoz, N., Sherman, M., Jansen, A., Peto, J., Schiffman, M., Moreno, V., Kurman, R. and Shan, K. (1995), 'Prevalence of human papillomavirus in cervical cancer: a worldwide perspective', *Journal of the National Cancer Institute* **87**(11), 796.

Branden, C. and Tooze, J. (1996), *Introduction à la structure des protéines*, De Boeck Supérieur.

Brochier, C., Philippe, H. and Moreira, D. (2000), 'The evolutionary history of ribosomal protein rps14: horizontal gene transfer at the heart of the ribosome.', *Trends in genetics: TIG* **16**(12), 529.

Brown, C. J., Garner, E. C., Dunker, A. K. and Joyce, P. (2001), 'The power to detect recombination using the coalescent', *Molecular Biology and Evolution* **18**(7), 1421–1424.

Brown, T. A. (2006), *Genomes 3*, Garland Science; 3 edition (May 3, 2006).

Brudno, M., Chapman, M., Göttgens, B., Batzoglou, S. and Morgenstern, B. (2003), 'Fast and sensitive multiple alignment of large genomic sequences', *BMC bioinformatics* **4**(1), 66.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. (2009), 'Blast+: architecture and applications', *BMC bioinformatics* **10**(1), 421.

Cantino, P. D. and De Queiroz, K. (2004), 'Phylocode: A phylogenetic code of biological nomenclature. version 2b', *Published electronically at http://www. ohiou. edu/phylocode* .

Castle, P. E., Stoler, M. H., Solomon, D., Schiffmanx, M. et al. (2007), 'The relationship of community biopsy-diagnosed cervical intraepithelial neoplasia grade 2 to the quality control pathology-reviewed diagnoses an alts report', *American journal of clinical pathology* **127**(5), 805–815.

Castresana, J. (2000), 'Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis', *Molecular biology and evolution* **17**(4), 540–552.

Chan, S., Delius, H., Halpern, A. and Bernard, H. (1995), 'Analysis of genomic sequences of 95 papillomavirus types: uniting typing, phylogeny, and taxonomy', *Journal of virology* **69**(5), 3074.

Charlebois, R. L. and Doolittle, W. F. (2004), 'Computing prokaryotic gene ubiquity: rescuing the core from extinction', *Genome research* **14**(12), 2469–2477.

Charleston, M. A. and Perkins, S. L. (2006), 'Traversing the tangle: algorithms and applications for cophylogenetic studies', *Journal of biomedical informatics* **39**(1), 62–71.

Ching, T. H., Yoza, B. A. and Li, Q. X. (2014), 'Quartet analysis of putative horizontal gene transfer in crenarchaeota', *Journal of molecular evolution* **78**(2), 163–170.

Clamp, M., Cuff, J., Searle, S. M. and Barton, G. J. (2004), 'The jalview java alignment editor', *Bioinformatics* **20**(3), 426–427.

Claverys, J.-P., Prudhomme, M., Mortier-Barrière, I. and Martin, B. (2000), 'Adaptation to the environment: Streptococcus pneumoniae, a paradigm for recombination-mediated genetic plasticity?', *Molecular microbiology* **35**(2), 251–259.

Clifford, G., Gallus, S., Herrero, R., Munoz, N., Snijders, P., Vaccarella, S., Anh, P., Ferreccio, C., Hieu, N., Matos, E. et al. (2005), 'Worldwide distribution of human papillomavirus types in cytologically normal women in the international agency for research on cancer hpv prevalence surveys: a pooled analysis', *The Lancet* **366**(9490), 991–998.

Conow, C., Fielder, D., Ovadia, Y. and Libeskind-Hadas, R. (2010), 'Jane: a new tool for the cophylogeny reconstruction problem', *Algorithms for Molecular Biology* **5**(1), 16.

Crick, F. et al. (1970), 'Central dogma of molecular biology', *Nature* **227**(5258), 561–563.

Csűrös, M. and Miklós, I. (2006), A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer, *in* 'Research in computational molecular biology', Springer, pp. 206–220.

Dagan, T., Artzy-Randrup, Y. and Martin, W. (2008), 'Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution', *Proceedings of the National Academy of Sciences* **105**(29), 10039–10044.

Dagan, T. and Martin, W. (2007), 'Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution', *Proceedings of the National Academy of Sciences* **104**(3), 870–875.

Darwin, C. (1859), 'On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life: Murray'.

David, L. A. and Alm, E. J. (2011), 'Rapid evolutionary innovation during an archaean genetic expansion', *Nature* **469**(7328), 93–96.

Davies, J. and Davies, D. (2010), 'Origins and evolution of antibiotic resistance', *Microbiology and Molecular Biology Reviews* **74**(3), 417–433.

De Villiers, E.-M., Fauquet, C., Broker, T. R., Bernard, H.-U. and zur Hausen, H. (2004), 'Classification of papillomaviruses', *Virology* **324**(1), 17–27.

De Vries, J., Herzfeld, T. and Wackernagel, W. (2004), 'Transfer of plastid dna from tobacco to the soil bacterium acinetobacter sp. by natural transformation', *Molecular Microbiology* **53**(1), 323–334.

Diallo, A., Badescu, D., Blanchette, M. and Makarenkov, V. (2009), 'A whole genome study and identification of specific carcinogenic regions of the human papilloma viruses', *Journal* of Computational Biology 16(10), 1461–1473.

Doolittle, W., Ford and Zhaxybayeva, O. (2013), 'What Is a Prokaryote ?', in: The Prokaryotes, Prokaryotic Biology and Symbiotic Associations, ISBN: 978-3-642-30193-3 (Print) 978-3-642-30194-0 (Online), Springer, 21-37

Doorbar Doorbar, J. (2006), 'Molecular biology of human papillomavirus infection and cervical cancer', *Clinical science* **110**, 525–541.

Doty, P., Boedtker, H., Fresco, J., Haselkorn, R. and Litt, M. (1959), 'Secondary structure in ribonucleic acids', *Proceedings of the National Academy of Sciences of the United States of America* **45**(4), 482.

Doyon, J.-P., Scornavacca, C., Gorbunov, K. Y., Szöllősi, G. J., Ranwez, V. and Berry, V. (2011), An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers, *in* 'Comparative genomics', Springer, pp. 93–108.

Doyon, J., Ranwez, V., Daubin, V. and Berry, V. (2011), 'Models, algorithms and programs for phylogeny reconciliation', *Briefings in bioinformatics* **12**(5), 392–400.

Drummond, A. J., Ho, S. Y., Phillips, M. J. and Rambaut, A. (2006), 'Relaxed phylogenetics and dating with confidence', *PLoS biology* **4**(5), e88.

Drummond, A. J. and Rambaut, A. (2007), 'Beast: Bayesian evolutionary analysis by sampling trees', *BMC evolutionary biology* **7**(1), 214.

Duret, L. (2008), 'Neutral theory: The null hypothesis of molecular evolution', *Nature Education* **1**, 803–806.

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

Edgar, R. C. (2004), 'Muscle: multiple sequence alignment with high accuracy and high throughput', *Nucleic acids research* **32**(5), 1792–1797.

Edwards, A. W. and Sforza, C. L. (1963), 'The reconstruction of evolution', *Heredity* **18**.

Elias, J., Harmsen, D., Claus, H., Hellenbrand, W., Frosch, M. and Vogel, U. (2006), 'Spatiotemporal analysis of invasive meningococcal disease, Germany', *Emerging infectious diseases* **12**(11).

Endler, J. and Greenwood, J. (1988), 'Frequency-dependent predation, crypsis and aposematic coloration [and discussion]', *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **319**(1196), 505–523.

Enright, A., Van Dongen, S. and Ouzounis, C. (2002), 'An efficient algorithm for large-scale detection of protein families', *Nucleic acids research* **30**(7), 1575–1584.

Envall, M. (2008), 'On the difference between mono-, holo-, and paraphyletic groups: a consistent distinction of process and pattern', *Biological journal of the Linnean Society* **94**(1), 217–220.

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M. and Servin, B. (2013), 'Detecting signatures of selection through haplotype differentiation among hierarchically structured populations', *Genetics* **193**(3), 929–941.

Felsenstein, J. (1981), 'Evolutionary trees from dna sequences: a maximum likelihood approach', *Journal of molecular evolution* **17**(6), 368–376.

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M., Angiuoli, S. et al. (2008), 'The minimum information about a genome sequence (migs) specification', *Nature biotechnology* **26**(5), 541–547.

Finn, R. D., Clements, J. and Eddy, S. R. (2011), 'Hmmer web server: interactive sequence similarity searching', *Nucleic acids research* **39**(suppl 2), W29–W37.

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

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S. et al. (2014), 'Ensembl 2014', *Nucleic acids research* **42**(D1), D749–D755.

Fraser, C., Hanage, W. P. and Spratt, B. G. (2007), 'Recombination and the nature of bacterial speciation', *Science* **315**(5811), 476–480.

Friedman, N., Ninio, M., Pe'er, I. and Pupko, T. (2002), 'A structural em algorithm for phylogenetic inference', *Journal of Computational Biology* **9**(2), 331–353.

Ge, F., Wang, L.-S. and Kim, J. (2005), 'The cobweb of life revealed by genome-scale estimates of horizontal gene transfer', *PLoS biology* **3**(10), e316.

Gernhard, T. (2008), 'The conditioned reconstructed process', *Journal of theoretical biology* **253**(4), 769–778.

Gharib, W. H. and Robinson-Rechavi, M. (2013), 'The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in gc', *Molecular Biology and Evolution* .

Gilbert, W. (1986), 'Origin of life: The rna world', *Nature* **319**(6055).

Gogarten, J. P., Doolittle, W. F. and Lawrence, J. G. (2002), 'Prokaryotic evolution in light of gene transfer', *Molecular biology and evolution* **19**(12), 2226–2238.

Gogarten, J. P. and Townsend, J. P. (2005), 'Horizontal gene transfer, genome innovation and evolution', *Nature Reviews Microbiology* **3**(9), 679–687.

Grassly, N. and Holmes, E. (1997), 'A likelihood method for the detection of selection and recombination using nucleotide sequences.', *Molecular Biology and Evolution* **14**(3), 239.

Guindon, S. and Gascuel, O. (2002), 'Efficient biased estimation of evolutionary distances when substitution rates vary across sites', *Molecular biology and evolution* **19**(4), 534.

Guindon, S. and Gascuel, O. (2003), 'A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood', *Systematic biology* **52**(5), 696–704.

Guindon, S., Rodrigo, A. G., Dyer, K. A. and Huelsenbeck, J. P. (2004), 'Modeling the site-specific variation of selection patterns along lineages', *Proceedings of the National Academy of Sciences of the United States of America* **101**(35), 12957–12962.

Haeckel, E. (1879), *The evolution of man*, Vol. 1, CK Paul & Company.

Hallett, M., Lagergren, J. and Tofigh, A. (2004), Simultaneous identification of duplications and lateral transfers, *in* 'Proceedings of the eighth annual international conference on Resaerch in computational molecular biology', ACM, pp. 347–356.

Hallett, M. T. and Lagergren, J. (2001), Efficient algorithms for lateral gene transfer problems, *in* 'Proceedings of the fifth annual international conference on Computational biology', ACM, pp. 149–156.

Hanage, W. P., Fraser, C. and Spratt, B. G. (2005), 'Fuzzy species among recombinogenic bacteria', *BMC Biology* **3**(1), 6.

Hasegawa, M., Kishino, H. and Yano, T.-a. (1985), 'Dating of the human-ape splitting by a molecular clock of mitochondrial dna', *Journal of molecular evolution* **22**(2), 160–174.

Hawking, S. and Jackson, M. (1993), *A brief history of time*, Dove Audio.

He, C.-Q., Xie, Z.-X., Han, G.-Z., Dong, J.-B., Wang, D., Liu, J.-B., Ma, L.-Y., Tang, X.-F., Liu, X.-P., Pang, Y.-S. et al. (2009), 'Homologous recombination as an evolutionary force in the avian influenza a virus', *Molecular biology and evolution* **26**(1), 177–187.

Hein, J. (1993), 'A heuristic method to reconstruct the history of sequences subject to recombination', *Journal of Molecular Evolution* **36**(4), 396–405.

Hennig, W. (1975), 'Cladistic analysis or cladistic classification a reply to ernst mayr', *Systematic Biology* **24**(2), 244–256.

Hickey, G., Dehne, F., Rau-Chaplin, A. and Blouin, C. (2008), 'Spr distance computation for unrooted trees', *Evolutionary bioinformatics online* **4**, 17.

Hollingshead, S. K., Becker, R. and Briles, D. E. (2000), 'Diversity of pspa: mosaic genes and evidence for past recombination in streptococcus pneumoniae', *Infection and immunity* **68**(10), 5889–5900.

Holsinger, K. E. and Weir, B. S. (2009), 'Genetics in geographically structured populations: defining, estimating and interpreting fst', *Nature Reviews Genetics* **10**(9), 639–650.

Hotopp, J. C. D., Grifantini, R., Kumar, N., Tzeng, Y. L., Fouts, D., Frigimelica, E., Draghi, M., Giuliani, M. M., Rappuoli, R., Stephens, D. S. et al. (2006), 'Comparative genomics of neisseria meningitidis: core genome, islands of horizontal transfer and pathogen-specific genes', *Microbiology* **152**(12), 3733–3749.

Hubert, L. and Arabie, P. (1985), 'Comparing partitions', *Journal of classification* **2**(1), 193–218.

Hubisz, M. J., Pollard, K. S. and Siepel, A. (2011), 'Phast and rphast: phylogenetic analysis with space/time models', *Briefings in bioinformatics* **12**(1), 41–51.

Huerta, M., Downing, G., Haseltine, F., Seto, B. and Liu, Y. (2000), 'Nih working definition of bioinformatics and computational biology', *US National Institute of Health* .

Hughes, A. L. and Nei, M. (1988), 'Pattern of nucleotide substitution at major histocompatibility complex class i loci reveals overdominant selection'.

Husmeier, D. and McGuire, G. (2003), 'Detecting recombination in 4-taxa dna sequence alignments with bayesian hidden markov models and markov chain monte carlo', *Molecular Biology and Evolution* **20**(3), 315–337.

Husmeier, D. and Wright, F. (2001), 'Probabilistic divergence measures for detecting interspecies recombination', *Bioinformatics* **17**(suppl 1), S123–S131.

Husmeier, D. and Wright, F. (2005), Detecting recombination in dna sequence alignments, *in* 'Probabilistic modeling in bioinformatics and medical informatics', Springer, pp. 147–190.

Husmeier, D., Wright, F. and Milne, I. (2005), 'Detecting interspecific recombination with a pruned probabilistic divergence measure', *Bioinformatics* **21**(9), 1797–1806.

Huson, D. H. and Bryant, D. (2006), 'Application of phylogenetic networks in evolutionary studies', *Molecular biology and evolution* **23**(2), 254–267.

Jacob, F. and Monod, J. (1961), 'Genetic regulatory mechanisms in the synthesis of proteins', *Journal of molecular biology* **3**(3), 318–356.

Jain, R., Rivera, M. C. and Lake, J. A. (1999), 'Horizontal gene transfer among genomes: the complexity hypothesis', *Proceedings of the National Academy of Sciences* **96**(7), 3801–3806.

Johannsen, W. (1911), 'The genotype conception of heredity', *American Naturalist* pp. 129–159.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. and Madden, T. L. (2008), 'Ncbi blast: a better web interface', *Nucleic acids research* **36**(suppl 2), W5–W9.

Jolley, K., Chan, M.-S. and Maiden, M. (2004), 'mlstdbnet - distributed multi-locus sequence typing (mlst) databases', *BMC Bioinformatics* **5**(1), 86.

Jolley, K., Wilson, D., Kriz, P., McVean, G. and Maiden, M. (2005), 'The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in neisseria meningitidis', *Molecular biology and evolution* **22**(3), 562–569.

Jukes, T. and Cantor, C. (1969), 'Evolution of protein molecules', *Mammalian Protein Metabolism. Academic Press, New York* pp. 21–132.

Katoh, K., Misawa, K., Kuma, K.-i. and Miyata, T. (2002), 'Mafft: a novel method for rapid multiple sequence alignment based on fast Fourier transform', *Nucleic acids research* **30**(14), 3059–3066.

Kim, J., Pramanik, S. and Chung, M. J. (1994), 'Multiple sequence alignment using simulated annealing', *Computer applications in the biosciences: CABIOS* **10**(4), 419–426.

Kim, J. and Salisbury, B. A. (2001), A tree obscured by vines: Horizontal gene transfer and the median tree method of estimating species phylogeny., *in* 'Pacific Symposium on Biocomputing', Vol. 6, pp. 571–582.

Kim, J. and Warnow, T. (1999), 'Tutorial on phylogenetic tree estimation', *Intelligent Systems for Molecular Biology, Heidelberg* .

Kimura, M. (1985), *The neutral theory of molecular evolution*, Cambridge Univ Pr.

Kimura, M. et al. (1968), 'Evolutionary rate at the molecular level', *Nature* **217**(5129), 624–626.

King, J. L., Jukes, T. H. and Clarke, B. (1969), *Non-darwinian evolution*, Bobbs-Merrill.

Kiyono, T., Hiraiwa, A., Fujita, M., Hayashi, Y., Akiyama, T. and Ishibashi, M. (1997), 'Binding of high-risk human papillomavirus e6 oncoproteins to the human homologue of the drosophila discs large tumor suppressor protein', *Proceedings of the National Academy of Sciences of the United States of America* **94**(21), 11612.

Kolaczkowski, B. and Thornton, J. W. (2004), 'Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous', *Nature* **431**(7011), 980–984.

Koonin, E. V., Makarova, K. S. and Aravind, L. (2001), 'Horizontal gene transfer in prokaryotes: quantification and classification 1', *Annual Reviews in Microbiology* **55**(1), 709–742.

Koonin, E. V. and Wolf, Y. I. (2008), 'Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world', *Nucleic Acids Research* **36**(21), 6688–6719.

Kortekaas, J., Pettersson, A., Van der Biezen, J., Weynants, V., Van der Ley, P., Poolman, J., Bos, M. and Tommassen, J. (2007), 'Shielding of immunogenic domains in neisseria meningitidis frpb (feta) by the major variable region', *Vaccine* **25**(1), 72–84.

Kuhner, M. and Felsenstein, J. (1994), 'A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates.', *Molecular Biology and Evolution* **11**(3), 459.

Lace, M. J., Isacson, C., Anson, J. R., Lörincz, A. T., Wilczynski, S. P., Haugen, T. H. and Turek, L. P. (2009), 'Upstream regulatory region alterations found in human papillomavirus type 16 (hpv-16) isolates from cervical carcinomas increase transcription, ori function, and hpv immortalization capacity in culture', *Journal of virology* **83**(15), 7457–7466.

Lampson, B., Inouye, M. and Inouye, S. (2005), 'Retrons, msdna, and the bacterial genome', *Cytogenetic and genome research* **110**(1-4), 491–499.

Lawrence, J. G. and Ochman, H. (1997), 'Amelioration of bacterial genomes: rates of change and exchange', *Journal of molecular evolution* **44**(4), 383–397.

Lecointre, G. and Le Guyader, H. (2001), *Classification phylogénétique du vivant*, Belin.

Lecointre, G. and Le Guyader, H. (2006), *The tree of life: a phylogenetic classification*, Vol. 20, Harvard University Press.

Lee, C. and Laimins, L. (2004), 'Role of the pdz domain-binding motif of the oncoprotein e6 in the pathogenesis of human papillomavirus type 31', *Journal of virology* **78**(22), 12366.

Legendre, P. and Makarenkov, V. (2002), 'Reconstruction of biogeographic and evolutionary networks using reticulograms', *Systematic Biology* **51**(2), 199–216.

Lewis, C. M. and Knight, J. (2012), 'Introduction to genetic association studies', *Cold Spring Harbor Protocols* **2012**(3), pdb–top068163.

Li, J., Dai, X., Liu, T. and Zhao, P. (2012), 'Legumeip: an integrative database for comparative genomics and transcriptomics of model legumes', *Nucleic Acids Research* **40**(D1), D1221–D1229.

Libeskind-Hadas, R. and Charleston, M. A. (2009), 'On the computational complexity of the reticulate cophylogeny reconstruction problem', *Journal of Computational Biology* **16**(1), 105–117.

Lipari, F., McGibbon, G., Wardrop, E. and Cordingley, M. (2001), 'Purification and biophysical characterization of a minimal functional domain and of an n-terminal zn2+-binding fragment from the human papillomavirus type 16 e6 protein', *Biochemistry* **40**(5), 1196–1204.

Lodish, H., Berk, A. and Zipursky, S. (2000), 'Molecular cell biology 4th edition', *New York: W. H. Freeman* .

Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. and Hirschhorn, J. N. (2003), 'Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease', *Nature genetics* **33**(2), 177–182.

Macdonald, S. J. and Long, A. D. (2005), 'Prospects for identifying functional variation across the genome', *Proceedings of the National Academy of Sciences* **102**(suppl 1), 6614–6621.

MacQueen, J. et al. (1967), Some methods for classification and analysis of multivariate observations, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', Vol. 1, California, USA, p. 14.

Maddison, D. R., Schulz, K.-S. and Maddison, W. P. (2007), 'The tree of life web project', *Zootaxa* **1668**(Linnaeus Tercentenary: Progress in Invertebrate Taxonomy), 19–40.

Maglott, D., Ostell, J., Pruitt, K. D. and Tatusova, T. (2005), 'Entrez gene: gene-centered information at ncbi', *Nucleic acids research* **33**(suppl 1), D54–D58.

Maiden, M. (2008), 'Population genomics: diversity and virulence in the neisseria', *Current opinion in microbiology* **11**(5), 467–471.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A. et al. (1998), 'Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms', *Proceedings of the National Academy of Sciences* **95**(6), 3140–3145.

Maiden, M., Malorny, B. and Achtman, M. (1996), 'A global gene pool in the neisseriae.', *Molecular microbiology* **21**(6), 1297.

Makarenkov, V. (2001), 'T-rex: reconstructing and visualizing phylogenetic trees and reticulation networks', *Bioinformatics* **17**(7), 664–668.

Makarenkov, V., Kevorkov, D. and Legendre, P. (2006), 'Phylogenetic network construction approaches', *Applied mycology and biotechnology* **6**, 61–97.

Margulies, E. H., Blanchette, M., Haussler, D., Green, E. D. et al. (2003), 'Identification and characterization of multi-species conserved sequences', *Genome research* **13**(12), 2507–2518.

Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D. and Lefeuvre, P. (2010), 'Rdp3: a flexible and fast computer program for analyzing recombination', *Bioinformatics* **26**(19), 2462–2463.

Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V. et al. (2003), 'Transfac: transcriptional regulation, from patterns to profiles', *Nucleic acids research* **31**(1), 374–378.

Maydt, J. and Lengauer, T. (2006), 'Recco: recombination analysis using cost optimization', *Bioinformatics* **22**(9), 1064–1071.

Mayr, E. (1998). 'Two empires or three?',*Proceedings of the National Academy of Sciences*, **95**(17), 9720-9723.

McGuire, G. and Wright, F. (2000), 'Topal 2.0: improved detection of mosaic sequences within multiple alignments.', *Bioinformatics (Oxford, England)* **16**(2), 130–4.

McGuire, G., Wright, F. and Prentice, M. (1997), 'A graphical method for detecting recombination in phylogenetic data sets.', *Molecular Biology and Evolution* **14**(11), 1125.

Meier, P. and Wackernagel, W. (2003), 'Mechanisms of homology-facilitated illegitimate recombination for foreign dna acquisition in transformable pseudomonas stutzeri', *Molecular microbiology* **48**(4), 1107–1118.

Merkle, D. and Middendorf, M. (2005), 'Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information', *Theory in Biosciences* **123**(4), 277–299.

Merkle, D., Middendorf, M. and Wieseke, N. (2010), 'A parameter-adaptive dynamic programming approach for inferring cophylogenies', *BMC bioinformatics* **11**(Suppl 1), S60.

Milligan, G. and Cooper, M. (1986), 'A study of the comparability of external criteria for hierarchical cluster analysis.', *Multivariate Behavioral Research* .

Milne, I., Wright, F., Rowe, G., Marshall, D. F., Husmeier, D. and McGuire, G. (2004), 'Topali: software for automatic identification of recombinant sequences within dna multiple alignments.', *Bioinformatics (Oxford, England)* **20**(11), 1806–7.

Mindell, D. P. (2013), 'The tree of life: Metaphor, model, and heuristic device', *Systematic Biology* **62**(3), 479–489.

Moran, P. (1962), 'The statistical processes of evolutionary theory.', *The statistical processes of evolutionary theory.* .

Moretti, S., Laurenczy, B., Gharib, W. H., Castella, B., Kuzniar, A., Schabauer, H., Studer, R. A., Valle, M., Salamin, N., Stockinger, H. and Robinson-Rechavi, M. (2014), 'Selectome update: quality control and computational improvements to a database of positive selection', *Nucleic Acids Research* **42**(D1), D917–D921.

Morgan, T. H., Sturtevant, A. H., Muller, H. J. and Bridges, C. B. (1922), *The mechanism of Mendelian heredity*, Holt.

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R. and Schäffer, A. A. (2008), 'Database indexing for production megablast searches', *Bioinformatics* **24**(16), 1757–1764.

Mulkidjanian, A. Y., Koonin, E. V., Makarova, K. S., Mekhedov, S. L., Sorokin, A., Wolf, Y. I., Dufresne, A., Partensky, F., Burd, H., Kaznadzey, D. et al. (2006), 'The cyanobacterial genome core and the origin of photosynthesis', *Proceedings of the National Academy of Sciences* **103**(35), 13126–13131.

Munoz, N. (2000), 'Human papillomavirus and cancer: the epidemiological evidence', *Journal of clinical virology* **19**(1-2), 1–5.

Munoz, N., Bosch, F., Castellsague, X., Díaz, M., de Sanjose, S., Hammouda, D., Shah, K. and Meijer, C. (2004), 'Against which human papillomavirus types shall we vaccinate and screen? the international perspective', *International Journal of Cancer* **111**(2), 278–285.

Muñoz, N., Bosch, F., de Sanjose, S., Herrero, R., Castellsagué, X., Shah, K., Snijders, P. and Meijer, C. (2003), 'Epidemiologic classification of human papillomavirus types associated with cervical cancer', *New England Journal of Medicine* **348**(6), 518–527.

Nakhleh, L., Ruths, D. and Wang, L.-S. (2005), Riata-hgt: a fast and accurate heuristic for reconstructing horizontal gene transfer, *in* 'Computing and Combinatorics', Springer, pp. 84–93.

National Library of Medicine (US). Genetics Home Reference [Internet]. Bethesda (MD): The Library; (2013). http://ghr.nlm.nih.gov

Nguyen, T.-H., Ranwez, V., Berry, V. and Scornavacca, C. (2013), 'Support measures to estimate the reliability of evolutionary events predicted by reconciliation methods', *PloS ONE* **8**(10), e73667.

Nogueira, T., Rankin, D. J., Touchon, M., Taddei, F., Brown, S. P. and Rocha, E. P. (2009), 'Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence', *Current Biology* **19**(20), 1683–1691.

Nominé, Y., Charbonnier, S., Ristriani, T., Stier, G., Masson, M., Cavusoglu, N., Van Dorsselaer, A., Weiss, É., Kieffer, B. and Travé, G. (2003), 'Domain substructure of hpv e6 oncoprotein: biophysical characterization of the e6 c-terminal dna-binding domain', *Biochemistry* **42**(17), 4909–4917.

Nominé, Y., Masson, M., Charbonnier, S., Zanier, K., Ristriani, T., Deryckère, F., Sibler, A., Desplancq, D., Atkinson, R., Weiss, E. et al. (2006), 'Structural and functional analysis of e6 oncoprotein: insights in the molecular pathways of human papillomavirus-mediated pathogenesis', *Molecular cell* **21**(5), 665–678.

Notredame, C., Higgins, D. G. and Heringa, J. (2000), 'T-coffee: A novel method for fast and accurate multiple sequence alignment', *Journal of molecular biology* **302**(1), 205–217.

Novichkov, P. S., Omelchenko, M. V., Gelfand, M. S., Mironov, A. A., Wolf, Y. I., and Koonin, E. V. (2004). 'Genome-wide molecular clock and horizontal gene transfer in bacterial evolution'. *Journal of bacteriology*, 186(19), 6575-6585.

Ochman, H., Lawrence, J. G. and Groisman, E. A. (2000), 'Lateral gene transfer and the nature of bacterial innovation', *Nature* **405**(6784), 299–304.

Oliva, A., Fariña, J. B. and Llabrés, M. (2004), 'Measurement of uncertainty in peptide molecular weight determination using size-exclusion chromatography with multi-angle laser light-scattering

detection and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry', *Analytica chimica acta* **512**(1), 103–110.

OpenStax College, O. C. (2013), 'Viruses'. http://cnx.org/contents/e5ec0a23-64ce-4c45-93a0-d1e24faf7be5@3.

OpenStax College, O. C. (2014), 'Prokaryotic diversity'. http://cnx.org/contents/e5edf495-b4e0-4e41-94d5-6dddee616da9@5.

Pace N. R. Time for a change, Nature, Nature Publishing Group, 2006, 441, 289-289

Pagani, I., Liolios, K., Jansson, J., Chen, I., Smirnova, T., Nosrat, B., Markowitz, V. and Kyrpides, N. (2012), 'The genomes online database (gold) v. 4: status of genomic and metagenomic projects and their associated metadata', *Nucleic Acids Research* **40**(D1), D571–D579.

Pearson, P. N. (2001), *Red Queen Hypothesis*, John Wiley & Sons, Ltd. http://dx.doi.org/10.1038/npg.els.0001667

Peltola, H. (1983), 'Meningococcal disease: still with us', *Review of Infectious Diseases* **5**(1), 71–91.

Penn, D. J. (2001), *Coevolution: Host-Parasite*, John Wiley & Sons, Ltd. http://dx.doi.org/10.1038/npg.els.0001765

Pennacchio, L. A. and Rubin, E. M. (2001), 'Genomic strategies to identify mammalian regulatory sequences', *Nature Reviews Genetics* **2**(2), 100–109.

Pertea, M., Pertea, G. M. and Salzberg, S. L. (2011), 'Detection of lineage-specific evolutionary changes among primate species', *BMC bioinformatics* **12**(1), 274.

Pettersson, A., Maas, A., Van Wassenaar, D., Van der Ley, P. and Tommassen, J. (1995), 'Molecular characterization of frpb, the 70-kilodalton iron-regulated outer membrane protein of neisseria meningitidis.', *Infection and immunity* **63**(10), 4181.

Pettersson, A., Poolman, J., van der Ley, P. and Tommassen, J. (1997), 'Response of neisseria meningitidis to iron limitation', *Antonie van Leeuwenhoek* **71**(1), 129–136.

Pierce, B. A. (2007), 'Genetics: A conceptual approach'.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. and Siepel, A. (2010), 'Detection of nonneutral substitution rates on mammalian phylogenies', *Genome research* **20**(1), 110–121.

Polz, M. F., Alm, E. J. and Hanage, W. P. (2013), 'Horizontal gene transfer and the evolution of bacterial and archaeal population structure', *Trends in Genetics* **29**(3), 170–175.

Pond, S. L. K. and Frost, S. D. (2005), 'A genetic algorithm approach to detecting lineage-specific variation in selection pressure', *Molecular biology and evolution* **22**(3), 478–485.

Pond, S. L. K., Posada, D., Gravenor, M. B., Woelk, C. H. and Frost, S. D. (2006), 'Automated phylogenetic detection of recombination using a genetic algorithm', *Molecular biology and evolution* **23**(10), 1891–1901.

Popa, O., Hazkani-Covo, E., Landan, G., Martin, W. and Dagan, T. (2011), 'Directed networks reveal genomic barriers and dna repair bypasses to lateral gene transfer among prokaryotes', *Genome research* **21**(4), 599–609.

Posada, D. and Crandall, K. A. (2001), 'Evaluation of methods for detecting recombination from dna sequences: computer simulations', *Proceedings of the National Academy of Sciences* **98**(24), 13757–13762.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006), 'Principal components analysis corrects for stratification in genome-wide association studies', *Nature genetics* **38**(8), 904–909.

Pritchard, J. K., Stephens, M., Rosenberg, N. A. and Donnelly, P. (2000), 'Association mapping in structured populations', *The American Journal of Human Genetics* **67**(1), 170–181.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Rambaut, A. and Grass, N. (1997), 'Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees', *Computer applications in the biosciences: CABIOS* **13**(3), 235.

Rand, W. (1971), 'Objective criteria for the evaluation of clustering methods', *Journal of the American Statistical association* **66**(336), 846–850.

Reiss, D., Baliga, N. and Bonneau, R. (2006), 'Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks', *BMC Bioinformatics* **2**, 280–302.

Riley, D. R., Sieber, K. B., Robinson, K. M., White, J. R., Ganesan, A., Nourbakhsh, S. and Hotopp, J. C. D. (2013), 'Bacteria-human somatic cell lateral gene transfer is enriched in cancer samples', *PLoS computational biology* **9**(6), e1003107.

Rivas, E. and Eddy, S. R. (2001), 'Noncoding rna gene detection using comparative sequence analysis', *BMC bioinformatics* **2**(1), 8.

Rocha, E. P. (2006), 'Inference and analysis of the relative stability of bacterial chromosomes', *Molecular biology and evolution* **23**(3), 513–522.

Sadri, J., Diallo, A. B. and Blanchette, M. (2011), 'Predicting site-specific human selective pressure using evolutionary signatures', *Bioinformatics* **27**(13), i266–i274.

Saenger, W. (1984), *Principles of nucleic acid structure*, Springer-Verlag.

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

Sanderson, M. J. (2002), 'Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach', *Molecular biology and evolution* **19**(1), 101–109.

Santos, J. and Embrechts, M. (2009), 'On the use of the adjusted rand index as a metric for evaluating supervised classification', *Artificial Neural Networks–ICANN* pp. 175–184.

Saunders, N. J. and Snyder, L. A. (2002), 'The minimal mobile element', *Microbiology* **148**(12), 3756–3760.

Sawyer, S. (1989), 'Statistical tests for detecting gene conversion.', *Molecular biology and evolution* **6**(5), 526–538.

Schäffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V. and Altschul, S. F. (2001), 'Improving the accuracy of psi-blast protein database searches with composition-based statistics and other refinements', *Nucleic acids research* **29**(14), 2994–3005.

Schiffman, M., Castle, P. E., Jeronimo, J., Rodriguez, A. C. and Wacholder, S. (2007), 'Human papillomavirus and cervical cancer', *The Lancet* **370**(9590), 890–907.

Schoen, C., Joseph, B., Claus, H., Vogel, U. and Frosch, M. (2007), 'Living in a changing environment: Insights into host adaptation in *Neisseria meningitidis* from comparative genomics', *International Journal of Medical Microbiology* **297**(7), 601–613.

Schoen, C., Tettelin, H., Parkhill, J. and Frosch, M. (2009), 'Genome flexibility in *Neisseria meningitidis*', *Vaccine* **27**, B103–B111.

Schönknecht, G., Chen, W.-H., Ternes, C. M., Barbier, G. G., Shrestha, R. P., Stanke, M., Bräutigam, A., Baker, B. J., Banfield, J. F., Garavito, R. M. et al. (2013), 'Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote', *Science* **339**(6124), 1207–1210.

Schwarz, T. F. and Leo, O. (2008), 'Immune response to human papillomavirus after prophylactic vaccination with as04-adjuvanted hpv-16/18 vaccine: improving upon nature', *Gynecologic oncology* **110**(3), S1–S10.

Segondy, M. (2008), 'Classification des papillomavirus (hpv)', *Revue Francophone des Laboratoires* **2008**(405), 23–25.

Shapiro, B. J. and Alm, E. J. (2008), 'Comparing patterns of natural selection across species using selective signatures', *PLoS genetics* **4**(2), e23.

Sheridan, P. P., Freeman, K. H. and Brenchley, J. E. (2003), 'Estimated minimal divergence times of the major bacterial and archaeal phyla', *Geomicrobiology Journal* **20**(1), 1–14.

Shi, T. and Falkowski, P. G. (2008), 'Genome evolution in cyanobacteria: the stable core and the variable shell', *Proceedings of the National Academy of Sciences* **105**(7), 2510–2515.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W. and Richards, S. (2005), 'Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes', *Genome research* **15**(8), 1034–1050.

Siepel, A., Pollard, K. S. and Haussler, D. (2006), New methods for detecting lineage-specific selection, *in* 'Research in Computational Molecular Biology', Springer, pp. 190–205.

Sjöstrand, J., Tofigh, A., Daubin, V., Arvestad, L., Sennblad, B. and Lagergren, J. (2014), 'A bayesian method for analyzing lateral gene transfer', *Systematic biology* p. syu007.

Smillie, C. S., Smith, M. B., Friedman, J., Cordero, O. X., David, L. A. and Alm, E. J. (2011), 'Ecology drives a global network of gene exchange connecting the human microbiome', *Nature* **480**(7376), 241–244.

Sneath, P. H., Sokal, R. R. et al. (1973), *Numerical taxonomy. The principles and practice of numerical classification.*

Sonea, S. (2000), *Prokaryotology: A coherent view*, PUM.

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

Stamatakis, A., Ludwig, T. and Meier, H. (2005), 'Raxml-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees', *Bioinformatics* **21**(4), 456–463.

Stanley, W. M. et al. (1935), 'Isolation of a crystalline protein possessing the properties of tobacco-mosaic virus', *Science* **81**(2113), 644–645.

Stecher, B., Denzler, R., Maier, L., Bernet, F., Sanders, M. J., Pickard, D. J., Barthel, M., Westendorf, A. M., Krogfelt, K. A., Walker, A. W. et al. (2012), 'Gut inflammation can boost horizontal gene transfer between pathogenic and commensal enterobacteriaceae', *Proceedings of the National Academy of Sciences* **109**(4), 1269–1274.

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B. and Durand, D. (2012), 'Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees', *Bioinformatics* **28**(18), i409–i415.

Stranger, B. E., Stahl, E. A. and Raj, T. (2011), 'Progress and promise of genome-wide association studies for human complex trait genetics', *Genetics* **187**(2), 367–383.

Syvanen, M. (2012). 'Evolutionary implications of horizontal gene transfer'.*Annual review of genetics*, **46**, 341-358.

Szollosi, G. J., Boussau, B., Abby, S. S., Tannier, E. and Daubin, V. (2012), 'Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations', *Proceedings of the National Academy of Sciences* **109**(43), 17513–17518.

Szöllösi, G. J., Tannier, E., Lartillot, N. and Daubin, V. (2013), 'Lateral gene transfer from the dead', *Systematic Biology* **62**(3), 386–397.

Takeuchi, N., Kaneko, K. and Koonin, E. V. (2014), 'Horizontal gene transfer can rescue prokaryotes from Muller's ratchet: Benefit of dna from dead cells and population subdivision', *G3: Genes| Genomes| Genetics* **4**(2), 325–339.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Ciufo, S., and Li, W. (2013). 'Prokaryotic genome annotation pipeline'. In '*The NCBI Handbook*', 2nd edition., Bethesda (MD): National Center for Biotechnology Information (US)

Tanay, A., Sharan, R., Kupiec, M. and Shamir, R. (2004), 'Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data', *Proceedings of the National Academy of Sciences* **9**, 2981–2986.

Tavaré, S. (1986), 'Some probabilistic and statistical problems in the analysis of dna sequences', *Lectures on mathematics in the life sciences* **17**, 57–86.

Temin, H. M. and Mizutani, S. (2010), 'Rna-dependent dna polymerase in virions of rous sarcoma virus', *A Century of Nature: Twenty-One Discoveries that Changed Science and the World* p. 181.

Tettelin, H., Saunders, N. J., Heidelberg, J., Jeffries, A. C., Nelson, K. E., Eisen, J. A., Ketchum, K. A., Hood, D. W., Peden, J. F., Dodson, R. J. et al. (2000), 'Complete genome sequence of neisseria meningitidis serogroup b strain mc58', *Science* **287**(5459), 1809–1815.

Than, C. and Nakhleh, L. (2008), Spr-based tree reconciliation: Non-binary trees and multiple solutions., *in* 'APBC', Citeseer, pp. 251–260.

Thomas, C. M. and Nielsen, K. M. (2005), 'Mechanisms of, and barriers to, horizontal gene transfer between bacteria', *Nature reviews microbiology* **3**(9), 711–721.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994), 'Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice', *Nucleic acids research* **22**(22), 4673–4680.

Tindle, R. W. (2002), 'Immune evasion in human papillomavirus-associated cervical cancer', *Nature Reviews Cancer* **2**(1), 59–64.

Trottier, H. and Franco, E. L. (2006), 'The epidemiology of genital human papillomavirus infection', *Vaccine* **24**, S4–S15.

Tsaousis, A. D., Martin, D., Ladoukakis, E., Posada, D. and Zouros, E. (2005), 'Widespread recombination in published animal mtdna sequences', *Molecular Biology and Evolution* **22**(4), 925–933.

Tsirigos, A. and Rigoutsos, I. (2005), 'A new computational method for the detection of horizontal gene transfer events', *Nucleic acids research* **33**(3), 922–933.

Turner, P., Mclennan, A., Bates, A. et al. (1997), 'Instant notes in molecular biology, school of biological sciences, university of Liverpool, UK', *BIOS Scientific Publishers Limited* pp. 239–240.

Urwin, R., Russell, J., Thompson, E., Holmes, E., Feavers, I. and Maiden, M. (2004), 'Distribution of surface protein variants among hyperinvasive meningococci: implications for vaccine design', *Infection and immunity* **72**(10), 5955.

Van Der Ley, P., Heckels, J., Virji, M., Hoogerhout, P. and Poolman, J. (1991), 'Topology of outer membrane porins in pathogenic neisseria spp', *Infection and immunity* **59**(9), 2963.

van Dongen, S. (2000), 'Graph clustering by flow simulation', *PhD thesis, Univ. Utrecht.* .

Vetsigian, K. and Goldenfeld, N. (2005), 'Global divergence of microbial genome sequences mediated by propagating fronts', *Proceedings of the National Academy of Sciences of the United States of America* **102**(20), 7332–7337.

Vitti, J. J., Grossman, S. R. and Sabeti, P. C. (2013), 'Detecting natural selection in genomic data', *Annual review of genetics* **47**, 97–120.

Wang, H.-C., Susko, E. and Roger, A. J. (2013), 'The site-wise log-likelihood score is a good predictor of genes under positive selection', *Journal of molecular evolution* **76**(5), 280–294.

Watson, J. and Crick, F. (1953), 'A structure for deoxyribose nucleic acid', *Nature* **171**.

Wiley, E. O., Siegel-Causey, D., Brooks, D. R., Funk, V. et al. (1991), *The compleat cladist: A primer of phylogenetic procedures*, Vol. 19, Museum of Natural History, University of Kansas Lawrence, Kansas.

Wiuf, C., Christensen, T. and Hein, J. (2001), 'A simulation study of the reliability of recombination detection methods', *Molecular Biology and Evolution* **18**(10), 1929–1939.

Woese, C. R. and Fox, G. E. (1977), 'Phylogenetic structure of the prokaryotic domain: the primary kingdoms', *Proceedings of the National Academy of Sciences* **74**(11), 5088–5090.

Woese, C. R., Kandler, O. and Wheelis, M. L. (1990), 'Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya.', *Proceedings of the National Academy of Sciences* **87**(12), 4576–4579.

Woese, C. R., Magrum, L. J. and Fox, G. E. (1978), 'Archaebacteria', *Journal of Molecular Evolution* **11**(3), 245–252.

Wolf, Y. I., Makarova, K. S., Yutin, N. and Koonin, E. V. (2012), 'Updated clusters of orthologous genes for archaea: a complex ancestor of the archaea and the byways of horizontal gene transfer', *Biol Direct* **7**, 46.

Wooley, J. C., Lin, H. S. et al. (2005), *Catalyzing inquiry at the interface of computing and biology*, National Academies Press.

Wright, S. (1949), 'The genetical structure of populations', *Annals of eugenics* **15**(1), 323–354.

Yang, Z. (1997), 'Paml: a program package for phylogenetic analysis by maximum likelihood', *Comput. Appl. Biosci* **13**, 555–556.

Yang, Z. (2007), 'Paml 4: phylogenetic analysis by maximum likelihood', *Molecular biology and evolution* **24**(8), 1586.

Yang, Z. and Nielsen, R. (2002), 'Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages', *Molecular biology and evolution* **19**(6), 908–917.

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, a. M. (2000), 'Codon-substitution models for heterogeneous selection pressure at amino acid sites.', *Genetics* **155**(1), 431–49.

Ye, J., McGinnis, S. and Madden, T. L. (2006), 'Blast: improvements for better sequence analysis', *Nucleic acids research* **34**(suppl 2), W6–W9.

Zhang, C., Wang, J., Long, M. and Fan, C. (2013), 'Gkaks: The pipeline for genome level ka/ks calculation', *Bioinformatics* .

Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F. and Papke, R. T. (2006), 'Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events', *Genome Research* **16**(9), 1099–1108.

Zhaxybayeva, O., Lapierre, P. and Gogarten, J. P. (2004), 'Genome mosaicism and organismal lineages', *TRENDS in Genetics* **20**(5), 254–260.