Chapter 6

Discussion

6.1 Conclusions

In this thesis, I have introduced and discussed a multi-factorial methodology for modelling, predicting and analyzing mobile genetic elements, present in microbial genomes, termed genomic islands (GIs). The reason that this project was chosen, is the observation that GIs drive accelerated rates of evolution (Groisman and Ochman, 1996) in microbial populations that in return shape host-pathogen interactions, adaptations to specific niches and the overall population structure, in a way fundamentally different from the biological processes and dynamics shaping eukaryotic genomes.

In order to predict and study GIs I firstly introduced a novel compositional-based algorithm (chapter 2), exploiting the principle that at the time of insertion, horizontally acquired genomic DNA carries the sequence signature of its donor and often deviates compositionally from the sequence signature of its new host. Although this assumption might not hold in many cases (e.g. in the case of compositionally similar donor and host genomes, host genes under functional constraints and horizontally acquired genes that have converged to the host composition due to the time-dependent process of amelioration), it can be tolerated for the sake of developing unsupervised algorithms that can be directly applied on raw genomic datasets, with a minimal (if any) level of annotation (Table 6.1).

The novelty of this methodology relies on the fact that it exploits a new compositional algorithm, i.e. the Interpolated Variable Order Motifs (IVOMs) that overcomes the limitations of pre-existing, fixed (low or high) order compositional based methodologies, by introducing an interpolated variable order approach in analyzing local compositional biases. Under this principle, no *a priori* assumption is made about the order of the compositional distribution that best captures departures from the genome backbone compositional distribution.

Obviously relying more on a higher level of annotation (e.g. gene prediction and functional/structural annotation) more accurate algorithms, that capture more reliably the true origin of putative horizontally acquired genomic regions, can be devised. Exploiting this principle, I introduced in chapter 4 a machine learning approach that quantifies our posterior belief that a genomic structure is likely to be a true GI.

This methodology did not make any *a priori* assumptions about the structure of GIs, but instead implemented a bottom-up search, sampling both putative GIs and non-GI genomic regions from Gram positive and Gram negative bacteria, rather than relying on a previous GI structural definition (Hacker *et al.*, 1997). The data showed that GIs represent a superfamily of mobile elements with core and variable structural features, characterized by increased structural variation, approaching probably a structural continuum, under which families and subfamilies are distinguishable but also conditional on the assumptions made and the arbitrarily chosen criteria used.

The novelty of this methodology relies on the fact that traditional machine learning approaches were exploited under a "forward-reverse" concept; a training dataset was used to train structural GI models, and those models were exploited not only to make predictions ("forward" implementation) on unseen examples, but most importantly to use their estimated parameters (weights) in "reverse" to draw conclusions about the structural variation of GIs.

Although the benchmarking analysis showed that structural-based predictions of GIs can be more reliable than methodologies exploiting purely compositional based information, they form supervised solutions that require a higher level of annotation (Table 6.1).

A feature of GIs, independent of any *a priori* compositional or structural assumption, is their horizontal origin, i.e. GIs are horizontally acquired mobile elements of limited phylogenetic distribution. Exploiting this principle in chapter 3 I discussed a comparative-based approach for the prediction of GIs, the modelling of their compositional amelioration over time, and the mapping of their inferred relative time of acquisition on the phylogenetic history of the reference genomic dataset.

Comparative based methodologies, applied in the prediction of GIs, can be more accurate and reliable than structural and compositional based approaches, purely due to the fact that they make no *a priori* assumptions about how GIs should "look"; instead they utilize information about a more fundamental property, i.e. their origin. However comparative-based methods, require a very wide (sequenced) species sample, a prerequisite that might well prohibit the application of such approaches in the case of species with very few sequenced representatives (Table 6.1).

It becomes obvious that in the case of predicting genetic elements characterized by increased levels of mobility, exploiting information (e.g. composition) derived from a single genome sequence provides only a very narrow and static "snapshot" of their mobile life and history. On the other hand, capturing a dynamic rather than a static picture of a bacterial population, allowing inter- and intra-species genetic-flux (i.e. gene loss, gene gain, duplication, recombination and chromosomal-rearrangements), key evolutionary steps and host adaptations to be explicitly modelled, provides a more reliable description of those highly mobile genetic elements. Under this "genetic-flux" framework, a more comprehensive picture of bacterial populations can be built taking into account both static (e.g. sequence information) and dynamic (i.e. genetic-flux) parameters (see future work section below).

In chapter 5, I carried out a blind-test, applying, in an integrative fashion, the compositional and the structural-based techniques described in the previous chapters on a newly sequenced, un-annotated genome with the specific aim of performing a "real-life" implementation of this prediction pipeline utilizing only the minimum level of information, i.e.

raw genomic sequence. Exploiting an experimentally validated test dataset, I discussed results showing that such methodologies can be directly applicable on genomic datasets even at the very early stages of the annotation pipelines, acting as complementary tools to the currently existing annotation methods.

Table 6.1: Properties of three different *in silico* methods developed and discussed in this thesis, for the analysis and study of Genomic Islands.

Method	**Annotation level**	**Information**	**Chapter**	**Pros**	**Cons**
Alien Hunter	Low	Composition	2	▪ Automated ▪ Fast ▪ Unsupervised ▪ Applicable on newly sampled and sequenced, un-annotated genomic datasets	▪ Composition might "lie" (compositionally similar donors-hosts, genes under functional constrains, amelioration)
RVM	Medium	Structure	4	▪ Very fast ▪ Reliable ▪ Good generalization properties	▪ Supervised ▪ Requires known examples to form the training dataset ▪ Requires structural annotation
Phylogenetic tree	High	Gene content, phylogenetic distribution	3	▪ Very reliable predictions if the correct model of evolution is applied ▪ Gives estimates about the relative time of acquisition ▪ Allows mapping of key evolutionary events on the phylogenetic history of the genomes of interest	▪ Time consuming (phylogenies) ▪ Manual curation ▪ Requires pre-existing sequenced closely and distantly related genomes

6.2 Future work

Although methodologies exploiting the dynamic properties of bacterial populations have just started to emerge (Daubin and Ochman, 2004; Didelot *et al.*, 2007; Fuxelius *et al.*, 2008; Vernikos *et al.*, 2007) providing a step-wise decomposition of the evolutionary history of species over time, and revealing key evolutionary events that drive host-adaptation and pathogenicity, they are still far from being complete, efficiently automated, standardized and high-throughput.

This challenge could well form the focus of a future project; to perform a bioinformatic whole-genome based comparative study of bacterial genomes in order to quantify explicitly inter- and intra-species differences and interactions. The results could be used to implement a high-throughput *in silico* platform for fast and reliable step-wise decomposition of the evolutionary history of bacterial populations, focused on identifying virulence genes and potential vaccine candidates (Vernikos, 2008). In the following sections I provide a brief outline of how this methodology could be implemented.

6.2.1 High-throughput modelling of genetic flux

6.2.1.1 Selection of bacterial genomes

For the purposes of studying inter- and intra-species genetic-flux, a set of query as well as outgroup genomes is needed. Two options can be exploited:

A. Manual: The user can select manually a set of species and outgroup representative genomes, based on prior knowledge.

B. Composition-based: Variable-order compositional distributions can be used, implementing the Interpolated Variable Order Motifs (IVOMs) theory (Vernikos and Parkhill, 2006). IVOMs is a very powerful and sensitive method that can reliably estimate the relatedness, by means of compositional analysis, of different closely or distantly related bacterial chromosomes, overcoming the limitations of fixed-order compositional indices (e.g. % G+C content). Its increased resolution can discriminate even very similar genomes e.g. of the same serovar, while the fact that it is alignment-free makes it efficiently fast and automated. This method can automatically select appropriate closely and distantly related (i.e. outgroup) genomes for a reliable study of genetic-flux.

6.2.1.2 Whole-genome, all-against-all comparative analysis

A. Orthologous genes: In order to identify orthologous genes, each genome in the dataset can be compared against all the other genomes, by means of a best reciprocal FASTA (Pearson, 1990) approach. Although this methodology has been optimized and fine-tuned to predict reliably orthologous genes (Bentley *et al.*, 2007; Thomson *et al.*, 2006; Vernikos *et al.*, 2007), the best matches between genes of the different genomes may well be paralogs rather than true orthologs. However this limitation is a desired property in the current methodology; gene duplication is part of the genetic-flux concept and such prediction ambiguities can be analyzed in a second step taking into account their syntenic relationship to differentiate true orthologs from paralogs.

B. Phyletic profile: From the above all-against-all comparison the different patterns of presence or absence (i.e. phyletic profile) of all the genes in the pan-genome (i.e. the genome of a bacterial species consisting of core and dispensable genes, (Medini *et al.*, 2005)), can be grouped and coded in a binary fashion, i.e. [1,0] to denote [presence, absence] respectively. The phyletic profile can be analyzed for the purpose of a three-fold strategy:

1. The patterns of gene presence or absence can be grouped into core (shared among all genomes) and dispensable (partially shared and strain-specific) gene sets; modelling the number of strain-specific genes in the pan-genome as a function of adding step-wise new genomes, could enable us to draw conclusions about the pan-genome properties (i.e. open or closed pan-genome) and its rate of growth (Tettelin *et al.*, 2005).

2. The phyletic profile can be used to build the phylogenetic tree of the dataset relying on an alignment-free, distance-based approach (Fitz-Gibbon and House, 1999; Snel *et al.*, 1999). The phyletic profile can be converted into a distance matrix, in which the distance will reflect the fraction of genes that two genomes have in common. This alignment-free methodology is key for the development of a high-throughput approach since it is very fast compared to sequence-based techniques, exploits the

entire pan-genome and takes into account the various aspects of genetic-flux.

3. The phylogenetic tree of the dataset can be exploited as the reference tree topology, in order to infer putative gene gain and gene loss events, analyzing the phyletic profile by means of a maximum parsimony model (Mirkin *et al.*, 2003; Vernikos *et al.*, 2007). This methodology will enable us to estimate the relative time (Daubin and Ochman, 2004; Vernikos *et al.*, 2007) and rate of gene-transfer events on branches of increasing depth within the tree, revealing potential key host-adaptation strategies, e.g. genome-degradation (Gomez-Valero *et al.*, 2007; Parkhill *et al.*, 2001).

C. Recombination events: The first step for the detection of putative recombination events can be based on the following assumption: if the topology of individual gene trees is statistically different from the reference tree topology of the entire dataset, those genes can be considered candidates for inter or intra-species recombination (Dykhuizen and Green, 1991; Feil *et al.*, 2001).

In a second step, a sliding window can be exploited to analyze local discrepancies in the sequence similarity of consecutive genes with their corresponding orthologs in the other genomes. Significantly different (higher or lower) sequence similarity not expected by chance after evaluating the gene neighbourhood of the query and the target genomes can be combined with violations of the reference tree topology (previous step) in order to determine the possible direction of recombination (i.e. inter- or intra-species).

D. Chromosomal rearrangements: Analyzing the co-linearity of the orthologous gene sets between two genomes will enable us to detect “breaks” in the syntenic relationship between the two chromosomes and infer possible large-scale rearrangements (e.g. inversions) (Eisen *et al.*, 2000; Liu and Sanderson, 1995; Tillier and Collins, 2000); their location relative to the terminus and the origin of replication could reveal the level of selective pressure for maintaining the genome order.

6.2.1.3 Quantification of genetic-flux

Generalized Linear Models (GLMs) (McCullagh and Nelder, 1989) can be used to build species and cross-species specific models of genetic-flux, quantifying the genome fluidity of bacterial populations. In the current framework, each GLM will be the weighted sum of K basis functions, where K denotes the different parameters of genetic-flux (e.g. gene gain, gene loss, duplication, chromosomal rearrangements, and recombination) used to describe a bacterial population exploiting a generalized genetic-flux alphabet; a similar approach to that described in chapter 4.

In the current genetic-flux framework, GLMs can be trained using species and cross-species genomes, quantifying explicitly under a probabilistic framework the contribution of each of the genetic-flux parameters in shaping the dynamic structure of specific bacterial populations. Consequently each GLM will provide in a single linear equation a step-wise decomposition of the evolutionary history of those bacterial populations. The gene-flux GLMs can be used in a machine learning method in order to evaluate how reliably genomic datasets can be classified into different bacterial species, based on their genetic-flux profile. Misclassifications, due to overlapping genetic-flux properties of seemingly distinct bacterial species can be further analyzed to re-evaluate the relatedness of the latter.

6.2.1.4 Biological significance

The results of this study could be directly applicable to: 1. The identification and classification of different or similar adaptation mechanisms to the same or different hosts, respectively.

2. The study and characterization of the genetic boundaries between free-living and host-adapted bacteria, as well as between pathogenic and commensal bacteria.

3. Defining the minimum number of species isolates to be sequenced in order to have a reliable sample of the diversity of a given bacterial population (open or closed pan-genome).

4. Guiding the identification of new vaccine candidates, using the concept of “reverse vaccinology" (Rappuoli, 2000; Rappuoli and Covacci, 2003); whereby comparative genomics has enabled the successful development of novel vaccines against major pathogens (Behr *et al.*, 1999; Maione *et al.*, 2005; Pizza *et al.*, 2000).

5. Quantifying explicitly the genetic-fluidity of bacterial species using a single, linear equation. Utilising a generalized gene-flux alphabet, new whole-genome based classification systems can be devised.

6. Mapping the relative time of gene transfer events from the evolutionary history of bacteria to the evolutionary history of their host, enabling us to begin to understand how the interactions between key gene-transfer events in the evolution of pathogenic bacteria (Parkhill *et al.*, 2001) and behavioural or demographic changes in their host population (Thomson *et al.*, 2008), lead to the emergence of novel pathogens.

6.3 Final remarks

To end, I would like to make a comment on the application of quantitative or qualitative models in modern biology. Initially, when the very first steps towards understanding the rules and principles that govern biological systems were made, simplistic assumptions had to be introduced, to keep the complexity of the hypotheses low enough for biologists to be able to draw valuable and, most importantly, interpretable conclusions. During the last ten years, or so, the transition from single-isolate genomics to comparative genomics of entire biological populations, has introduced new (previously unknown) parameters that in some cases threaten to question or even to reject our initial assumptions about fundamental biological concepts and definitions. For example, in the current context of increased microbial genome fluidity, the fundamental definition of the biological species (Mayr, 1942), does not provide a realistic and representative description of the dynamic relationships that shape microbial evolution. Moving from intuition-driven or even macroscopic observation-driven hypotheses to data-driven hypotheses represents a more realistic approach in the study of biological systems, even when this requires revisiting and perhaps rejecting our initial, intuitively correct but biologically erroneous assumptions and definitions; a recent example, derived from microbial populations that extensively exchange genetic material, involves the rejection of the strictly bifurcating tree of life (Darwin, 1859) by a more realistic model-structure, that of the reticulate phylogenetic network (Huson and Bryant, 2006).