Unione Europea

Ministero dell'Istruzione,
dell'Università e della Ricerca

UNIVERSITÀ DEGLI STUDI DI SALERNO

Dottorato di ricerca in Biologia dei Sistemi
XI ciclo nuova serie
2009-2012

Sequence Analysis in Bioinformatics: methodological and practical aspects

Dottorando: **Vittorio Fortino**

Tutor: **Ch.mo Prof. Roberto Tagliaferri**

Coordinatore: **Ch.mo Prof. Antonietta Leone**

INDEX

CHAPTER 4

CHAPTER 5

ABSTRACT

My PhD research activities has focused on the development of new computational methods for biological sequence analyses.

To overcome an intrinsic problem to protein sequence analysis, whose aim was to infer homologies in large biological protein databases with short queries, I developed a statistical framework BLAST-based to detect distant homologies conserved in transmembrane domains of different bacterial membrane proteins. Using this framework, transmembrane protein domains of all *Salmonella* spp. have been screened and more than five thousands of significant homologies have been identified. My results show that the proposed framework detects distant homologies that, because of their conservation in distinct bacterial membrane proteins, could represent ancient signatures about the existence of primeval genetic elements (or mini-genes) coding for short polypeptides that formed, through a primitive assembly process, more complex genes. Further, my statistical framework lays the foundation for new bioinformatics tools to detect homologies domain-oriented, or in other words, the ability to find statistically significant homologies in specific target-domains.

The second problem that I faced deals with the analysis of transcripts obtained with RNA-Seq data. I developed a novel computational method that combines transcript borders, obtained from mapped RNA-Seq reads, with sequence features based operon predictions to accurately infer operons in prokaryotic genomes. Since the transcriptome of an organism is dynamic and condition dependent, the RNA-Seq mapped reads are used to determine a set of confirmed or predicted operons and from it specific transcriptomic features are extracted and combined with standard genomic features to train and validate three operon classification models (Random Forests - RFs, Neural Networks – NNs, and Support Vector Machines - SVMs). These classifiers have been exploited to refine the operon map annotated by DOOR, one of the most used database of prokaryotic operons. This method proved that the integration of genomic and transcriptomic features improve the accuracy of operon predictions, and that it is possible to predict the existence of potential new operons. An inherent limitation of using RNA-Seq to improve operon structure predictions is that it can be not applied to genes not expressed under the condition studied. I evaluated my approach on different RNA-Seq based transcriptome profiles of *Histophilus somni* and *Porphyromonas gingivalis*. These transcriptome profiles were obtained using the standard RNA-Seq or the strand-specific RNA-Seq method. My experimental results demonstrate that the three classifiers achieved accurate operon maps including reliable predictions of new operons.

RIASSUNTO

L’attività di ricerca svolta in questi tre anni di dottorato è stata incentrata sullo studio e lo sviluppo di nuove metodi computazionali per la risoluzione di problemi derivanti dall'analisi di sequenze biologiche (DNA, RNA). Il primo problema affrontato è stato quello della ricerca di similarità significative (o omologie) tra domini transmembrana (TMs) di diverse proteine integrali di membrana. Per risolvere tale problema ho realizzato un framework statistico, basato sull’utilizzo di BLAST (Altschul et al. 1990), che seleziona e valida sequenze simili trovate in domini TMs. L’obiettivo del framework è quello di valutare, in maniera appropriata, il significato statistico degli allineamenti trovati con sequenze query molto brevi, e quindi verificare se le similarità osservate rappresentino reali casi di omologia. Il framework è stato testato sulle sequenze TMs di proteine integrali di membrana (TMPs) di diverse specie di Salmonella. A partire da 1,760 TMPs sono state identificate 5,216 omologie aventi un p-value minore o uguale a 0.05. Il framework proposto deriva da un concetto più generale che potrebbe essere esteso per la ricerca di sequenze omologhe conservate in specifici domini (homology search domain-oriented). Infatti, nell’ambito di questi studi ho realizzato un framework teorico che può essere utilizzato per implementare algoritmi sia per la ricerca di sequenze omologhe distanti sia per la validazione di allineamenti non ritenuti significativi dalla statistica test usata da BLAST (Altschul and Gish *et al.*, 2006), una statistica molto generale e che soprattutto non si adatta a tutti casi di omologia.

Un secondo problema trattato, derivante dall'analisi di dati RNA-Seq, è stato quello di combinare caratteristiche “sequence-based”, usate dagli algoritmi esistenti per la predizione degli operoni (Dam *et al.*, 2006; Moreno-Hagelsieb and Collado-Videset *et al.*, 2002; Sabatti *et al.*, 2002; Salgado et al., 2000; Taboada *et al.*, 2010), con caratteristiche “transcriptome-based”, per migliorare l’accuratezza dei modelli di predizione degli operoni. L’idea è quella di estrarre caratteristiche “transcriptome-based” da un profilo globale di trascrizione ottenuto tramite esperimenti di sequenziamento del RNA totale (RNA-Seq), che possano essere utilizzate per la classificazione dei geni in operoni. Infatti, profili di trascrizione basati su dati RNA-Seq possono essere impiegati per quantificare il livello di trascrizione delle regioni geniche ed intergeniche, e quindi capire se due geni adiacenti vengono trascritti in blocco o meno. Quindi, ho progettato ed implementato un nuovo sistema di classificazione che usa caratteristiche genomiche (come la distanza intergenica e il bias nell'utilizzo dei codoni) e trascrittomiche (come il livello di espressione della regione intergenica e la differenza di espressione tra due geni adiacenti), per determinare la corretta organizzazione dei geni in operoni. Il sistema proposto si basa su tre modelli di classificazione: Random Forests (RFs), Neural Networks (NNs) e Support Vector Machines (SVMs). Questi modelli sono stati scelti perché, dalla letteratura, risultano quelli più performanti per la classificazione di geni in operoni (Charaniya *et al.*, 2007; Taboada *et al.*, 2010; Tjaden *et al.*, 2002; Tran *et al.*, 2007). I tre classificatori, una volta addestrati su un set di operoni confermati da esperimenti RNA-Seq, vengono impiegati per ridefinire parte della struttura operonica indicata dal database DOOR (Mao *et al.*, 2008) e delineare nuove coppie di geni adiacenti che potrebbero far parte di operoni non conosciuti in letteratura.

Il metodo proposto è stato testato su diversi profili di trascrizione di *Haemophilus somni* e *Porphyromonas gingivalis*. I risultati sperimentali mostrano che i classificatori ottenuti sono capaci di classificare gli operoni con un grado di accuratezza superiore al 96%, e questo prova che i dati trascrittomici estratti da un profilo di trascrizione ottenuto con dati di esperimenti RNA-Seq, possono aiutare a classificare correttamente i geni in operoni e a identificare nuovi potenziali operoni.

CHAPTER 1

Sequence Analysis in Bioinformatics

This is a thesis about computational sequence analysis, particularly applied to characterize both genomic and transcriptomic sequences. The work of several years has been focused on finding solutions to different, biological and computational problems. The research activities have led me to learn the main bioinformatics tools for the analysis of different biological sequences (DNA and RNA) and the study of new high-throughput technologies, such as RNA-Seq, that have opened a new frontier of computational problems in sequence analysis.

This chapter represents an introduction to the first important field of bioinformatics that is the analysis of biological sequences. After giving a historical survey of the first sequence analysis tools in the early days of bioinformatics, I introduce the most common sequence-based applications in this field. Next, I provide a short explanation of the aims and motivations of this thesis, followed by a brief description of the chapters.

1.1 A short history of sequence analysis

The history of sequence analysis began with Frederic Sanger who sequenced in 1956 bovine insulin, a protein consisting of 51 residues. This achievement showed, for the first time, that proteins are composed of linear (poly)peptides formed by aminoacid residues covalently attached in a defined, but apparently random order. Then, eight years elapsed between the first protein sequence reported by F. Sanger and the first nucleotide sequence in 1964 (Holley *et al.*, 1965).

Modern DNA sequencing become available in 1977, with the development of the chemical method by Maxam and Gilbert, the dideoxy method by Sanger, Nicklen *et al.* (1977), and the first complete sequence of a DNA molecule (Sanger *et al.*, 1978), which showed that DNA sequence determination could provide profound insights into genetic organization.

After the first DNA sequencing methods, subsequent improvements allowed the sequencing of increasingly longer molecules, and, consequently, it became possible to accomplish the first sequence analysis tasks. The early sequences were assembled, analyzed, and compared manually, by writing them down in lab notebooks (Figure 1).

When rapid DNA-sequencing technologies became available in the early 80s, the first genome sequencing projects initiated, and the number of generated sequences grew exponentially. With the growing number of DNA sequences it became impractical to analyze and store sequences manually, thus, more effective methods for data collection, storage, and retrieval were needed.

Fortunately, in the same years computers became more widespread and affordable due to the mass production of the microprocessor, and consequently, biologists and computer scientists started to think how to curate, organize, and manipulate the huge amount of information created by modern sequencing technologies. This was the beginning of bioinformatics, and more in general, the use of computers in DNA sequence analysis.

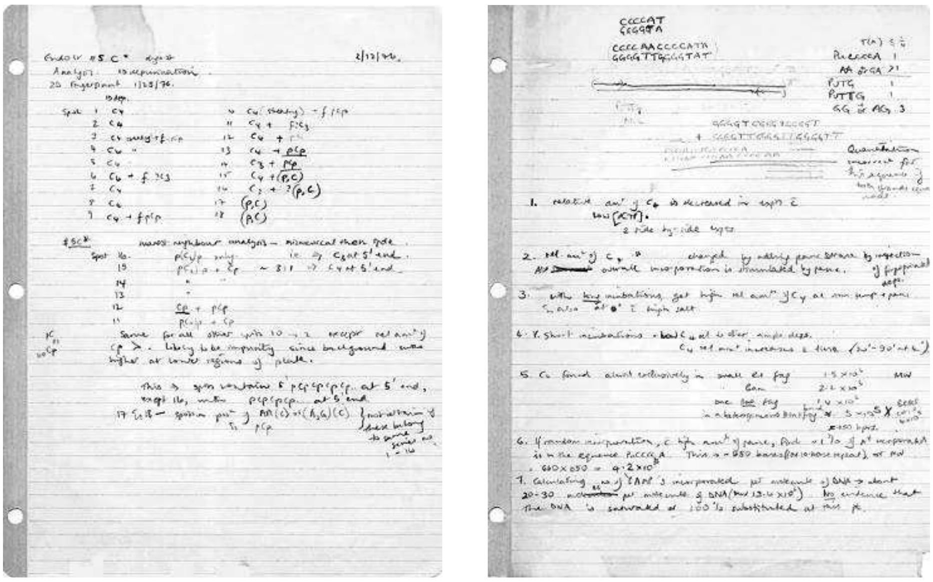


Fig. 1 – Lab Notes (Blackburn *et al.*, 1976).

At that time, the first task to accomplish was to sequence and archive DNA sequences of thousands of organisms in *databanks*, to storage and retrieval information. Before the advent of modern sequencing methods, first databanks were already available, but they were merely collections of sequences of proteins distributed in scientific journals.

In 1965, Dayhoff gathered all the available sequence data to create the first bioinformatics database (Atlas of Protein Sequence and Structure). Then, in 1972 the Protein Data Bank followed (Bernstein *et al.*, 1977) with a collection of ten X-ray crystallographic protein structures. Later, Walter Goad at the Theoretical Biology and Biophysics Group at the Los Alamos National Laboratory (LANL, USA) and others already present at the Los Alamos Sequence Database in 1979, which culminated in 1982 with a public GenBank (Figure 2). In 1980 the European Molecular Biology Laboratory (EMBL, Heidelberg, Germany) started the development of the EMBL Nucleotide Database. In a few years, a large number of several databases (or databank) of different types and sizes became accessible. Mostly of them were organized in flat files, because they are essentially repositories of biological sequences inserted by researchers to make them accessible to the scientific community. Only later, with the growth of data volumes, complexity and diversity, relational database have been made to improve data organization, facilitate the access and information retrieval.

Concomitantly, the first proteins sequences and the development of the earliest databases and, tools for sequence alignment became available. When a new sequence is available, it has to be checked whether the sequence is already present in databanks, or whether it is homologous (sequences which are derived from a common ancestor) to other sequences of different databanks. Usually, if two biological sequences display sufficient similarity, than inference of homology between them is justified. However, before evaluating or recognizing any similarity between two sequences, a plausible alignment must be determined. Therefore, *sequence alignment* tools soon became the basic techniques for sequence analyses, and its importance has increased over the years as more robust and informative techniques were developed.

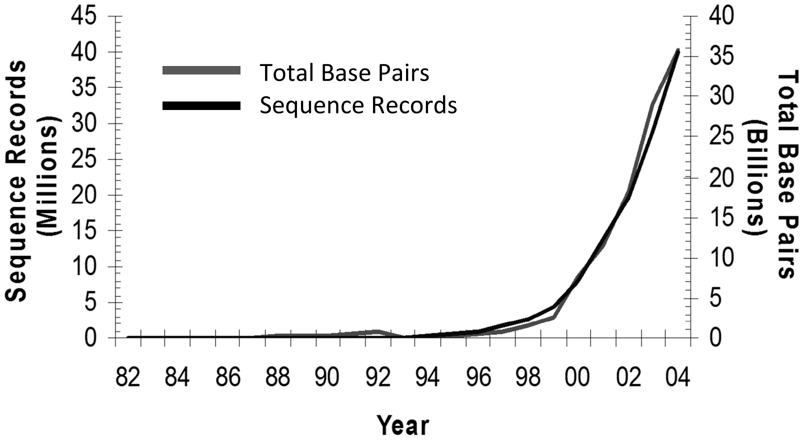


Fig. 2 – Growth of data deposited in GENBANK (1982-2004).

Methods for sequence alignments are now very complex, but the concepts are based on simple rules. Initially, sequence alignments were done manually that required a big effort, since biologists, not acquainted with computer programs for sequence analyses had to understand how many gaps were needed to get better alignments. The first sequence alignment method that was developed, and that could be computerized, was the Needleman-Wunsch method by Needleman and Wunsch (1970). Ten years later, Smith and Waterman (1981) elaborated a new algorithm to identify local sequence alignments. The Needleman-Wunsch and Smith-Waterman algorithms represent optimal solutions to the sequence alignment problem. They find, respectively, the best global or local alignment between two strings by comparing every pairs of characters and using some scoring scheme.

Since these algorithms tend to be slow, when we search in a whole database, new heuristic[1] alignment algorithms were proposed: FASTA (Lipman and Pearson *et al.*, 1985; Pearson and Lipman *et al.*, 1988) and BLAST (Altschul *et al.*, 1990). These programs were developed to speed up the alignment process, discarding optimal solutions. Today, BLAST and FASTA are useful to understand the relatedness of any protein or DNA sequence (query sequence) to other sequences (database), to identify sequences with a common ancestor (orthologs) and paralogs, to discover new genes or proteins, and to explore protein structure and function.

The next important advance in sequence analysis was to design new algorithms capable of performing *multiple sequence alignments*. Aligning two sequences can be very useful to infer biological relationships from the sequence similarity, but aligning multiple sequences can enable the search of interesting patterns in specific protein families, the building of phylogenetic trees and the definition of computational models to predict the secondary and tertiary structures of new sequences. The first computational method for multiple sequence alignment has been CLUSTAL (Higgins and Sharp *et al.*, 1988); it performs alignments based on clustering the sequences.

From the first sequence alignment tools, different and sophisticated programs for any kind of sequence analysis problem were produced over the years. On the other hand, since bioinformatics deals with computational analysis of DNA sequences, this led to solve computational problems and methods that produced a multitude of applications in Bioinformatics (Figure 3).

In this context, the sequencing method developed by Sanger has been the main standard for DNA sequencing for many years, and made possible the first draft of the human genome in 2001, which took more than 10 years and US$2.7 billion to complete (Venter *et al.*, 2010). For this reason, Sanger‘s method is regarded as one of the most important scientific technologies developed in the 20th century. However, since this method is laborious and time consuming, it is not suitable for high-throughput sequencing. This led to a high demand for low-cost sequencing technologies capable of producing thousands or millions of sequences in a short period of time (Church *et al.*, 2006; Hall *et al.*, 2007).

In fact, recent advances in DNA sequencing have resulted in a new generation of DNA sequencing systems followed by a multitude of novel applications in biology and sequence analysis. These systems are based on new massively parallel next-generation sequencing (NGS) platforms, with the main advantage, compared to Sanger’s sequencing, of yelding considerably higher throughput and lower cost per sequenced base (Ansorge *et al.*, 2009).

Sanger’s sequencing is adequate for many applications involving the analysis of single genes, and more general, single stretches of DNA. It generates a single, long, reliable sequence read of one region of DNA at a time, and, under the right conditions, Sanger’s method can sequence, on the average, well over 500 nucleotides. NGS technologies cannot generate long and exact sequences, but they can sequences, simultaneously, millions of different sequences. Thus, this new method can sequence the entire human genome in a week at a cost 200-fold lower than previous methods. For this reason, NSG technologies allow the generation of many kinds of sequence data: for example, they are used to obtain *de novo*

[1] In computer science, the term heuristic is used for algorithms designed for solving a problem more quickly when classic methods are too slow. They find solutions among all possible ones without guarantee that the best will be found.

sequencing, to re-sequence individual genomes when a reference genome already exists, sequence RNA to quantify expression level (RNA-Seq - Marioni *et al.*, 2008; Mortazavi *et al.*, 2008) and study the regulation of genes by sequencing chromatin immunoprecipitation products (ChIP-Seq - Mikkelsen *et al.*, 2007; Robertson *et al.*, 2007).

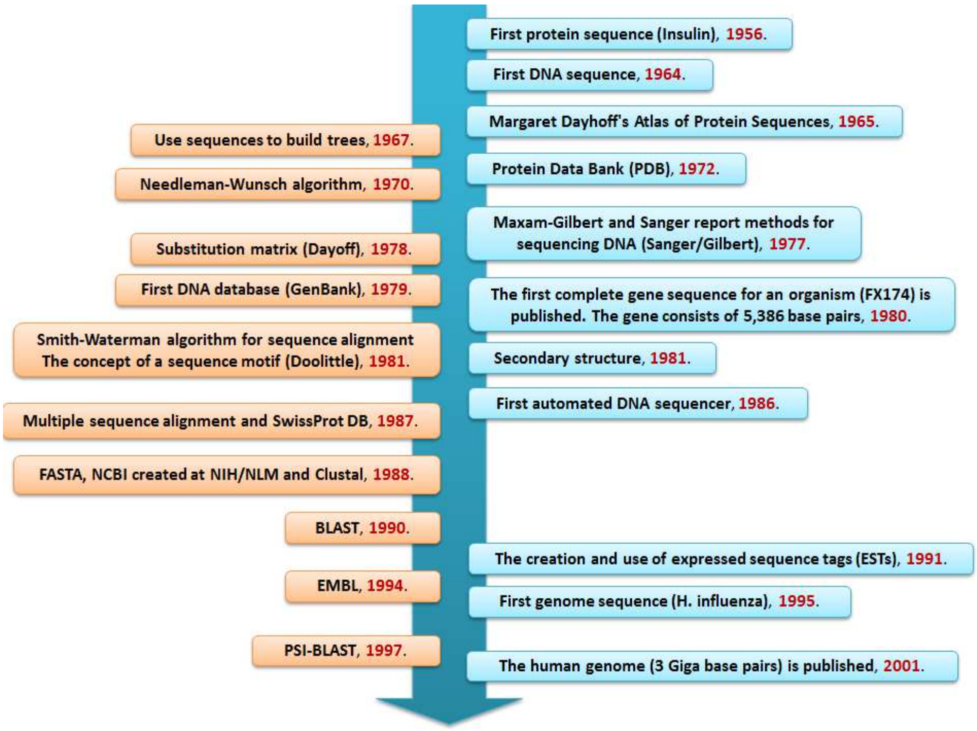

First protein sequence (Insulin), 1956.
First DNA sequence, 1964.
Margaret Dayhoff's Atlas of Protein Sequences, 1965.
Use sequences to build trees, 1967.
Needleman-Wunsch algorithm, 1970.
Protein Data Bank (PDB), 1972.
Maxam-Gilbert and Sanger report methods for sequencing DNA (Sanger/Gilbert), 1977.
Substitution matrix (Dayoff), 1978.
First DNA database (GenBank), 1979.
The first complete gene sequence for an organism (FX174) is published. The gene consists of 5,386 base pairs, 1980.
Smith-Waterman algorithm for sequence alignment The concept of a sequence motif (Doolittle), 1981.
Secondary structure, 1981.
First automated DNA sequencer, 1986.
Multiple sequence alignment and SwissProt DB, 1987.
FASTA, NCBI created at NIH/NLM and Clustal, 1988.
BLAST, 1990.
The creation and use of expressed sequence tags (ESTs), 1991.
EMBL, 1994.
First genome sequence (H. influenza), 1995.
PSI-BLAST, 1997.
The human genome (3 Giga base pairs) is published, 2001.

Fig. 3 – Short history of sequence analysis in the early days of bioinformatics.

The advent of NGS required the development of new statistical methods and bioinformatics tools for the analyses and the management of the huge amounts of data generated. Today, a substantial number of software already exists for managing and analyzing NGS data. These programs can be classified into different categories including alignment of sequence reads to a reference, base-calling and/or polymorphism detection, *de novo* assembly from paired or unpaired reads, structural variant detection and genome browsing. However, since the RNA-Seq technologies allow the construction of single-base resolution expression profile of a cell in un-precedented detail, today, it is important to develop new sequence analysis tools that can improve our knowledge about the RNA-based regulatory mechanisms.

1.2 Overview of bioinformatics applications in sequence analysis

In Bioinformatics the general term “sequence analysis” indicates a process that aims at the discovery of functional and structural similarities and/or differences between multiple biological sequences. The main bioinformatics applications involving sequence analysis tasks are reported in Figure 4.

With the rapid increase in the available biological sequences through a wide variety of databases, *similarity searches* have become essential components of most bioinformatics applications. They form the bases of structural analyses, motif identification, gene identification, and insights into functional associations.

Finding significant similarities between an unknown sequence and a sequence about which something is already known, it is useful to characterize the new sequence in terms of its structure and/or function. It is well known in literature that two similar sequences possess the same functional role, regulatory or biochemical pathway, and protein structure. In addition, if two sequences, belonging to different organisms, are similar enough and this similarity is statistically significant, they are considered *homologous sequences*.

Homology search is important in predicting the nature of a sequence: it helps greatly in the development of new drugs and in performing *phylogenetic analysis*. Evaluation of the similarity of two sequences, is performed by finding a plausible *alignment* between them, and then looking for identical characters or character patterns to quantify the level of similarity of the two sequences.

The alignment of two biological sequences is a fundamental operation that forms part of many bioinformatics applications, including sequence database searching, multiple sequence alignment, genome assembly, and short read mapping. Sequence alignment algorithms works by comparing base-by-base two (*pairwise alignment*) or more (*multiple alignment*) sequences by searching for matches, that are series of individual characters or character patterns that are in the same order in the sequences. With these algorithms we can understand the relatedness of any protein or DNA sequence to other known sequences by *database searching* tasks, identify sequences with a common ancestor (orthologs) and paralogs, discover new genes or proteins, explore protein structure and function.

Database similarity searching allows us to determine which of the hundreds of thousands of sequences present in a biological database are potentially related to a particular sequence of interest (o query sequence). Clearly, optimal alignment algorithms are impractical in database searching, and, therefore, different solutions based on heuristic methods have been proposed. Heuristic strategies make use of approximations to speed up significantly sequence comparisons, but with a minimal risk that true alignments can be missed. BLAST and FASTA are the most used sequence alignment tools based on a heuristic approach.

Further, alignment algorithms can be used for the alignment of multiple sequences. The application of *multiple sequence alignment algorithms* is important to determine conserved regions, and so, to *characterize protein families* and develop mathematical models to identify new member family.

With nearly a thousand genomes partly or fully sequenced, scientists have started to develop sophisticated programs to compare genome sequences and extract relevant information to explore the genetic differences between species, construct evolutionary trees, trace disease susceptibility in populations, and even track down people's ancestry.

Another important class of sequence analysis quests is *gene prediction*. Gene prediction is the process of detecting meaningful signals in uncharacterized DNA sequences. Thanks to the different sequence analysis tools, it is possible to integrate integrate the analysis of many different regions such as promoter regions, translation start and stop codons, reading frame periodicities, polyadenylation (polyA) signals, and, for eukaryotes, intron splicing signals, base compositional bias between codon positions for exons and introns, and various coding statistics.

An active area of research in sequence analysis is comparing whole genomes. Today, with complete genomes and large cDNA sequence collections available for many organisms, we can accomplish complex comparative genomic studies focused on discovering functionally and evolutionarily significant information. For example, the identification of cross-genome protein families may lead to targeting drugs better when the target protein present in bacterial genomes does not correlate with proteins coded by the human genome, thus leading to antibiotic targets. Gene classification can be further enhanced with the availability of evolutionary groupings. This enables tracing structure/functionality back to a single ancestral gene; therefore, classification of the target gene is derived from its ancestral form.

In spite of the many sequence analysis tools that have been developed, different computational problems in sequence analysis are still without an algorithmic solution. For instance, tools capable of inferring homologies using short query sequences do not exist yet, as well as tools to detect homologies in specific target domains (*homology search domain-oriented*).

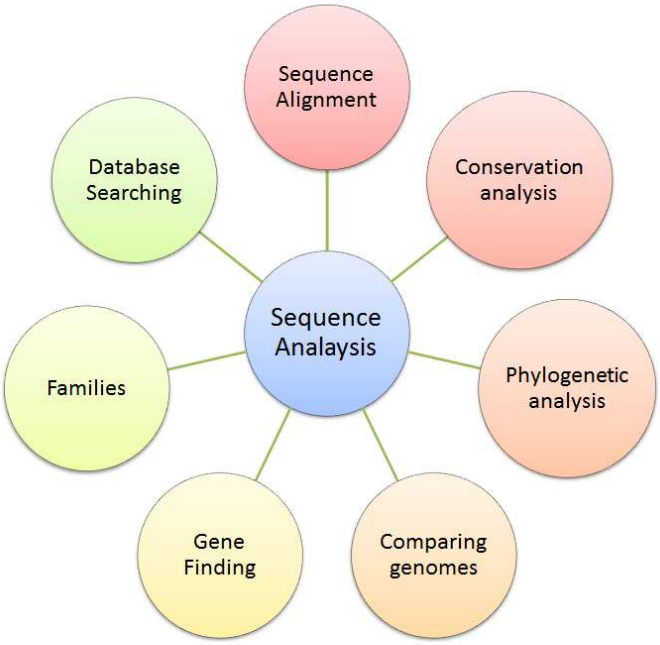


Fig. 4 – Main applications in Sequence Analysis.

Furthermore, new computational problems in sequence analysis arise with the advent of new high-throughput sequencing technologies. For instance, by sequencing transcriptomes with RNA-Seq technologies many new problems related to sequence analyses (ranging from the size and complexity of the datasets, to re-defining the genetic regulatory system found in prokaryotes and eukaryotes) became apparent.

This implies that the demand of new tools in sequence analysis will continue to grow despite the large number of the existing tools. For this reason, the main aim of this thesis is to provide novel bioinformatics methods to resolve computational problems ranging from the analysis of protein sequences, to the analysis of transcripts in RNA-Seq studies.

1.3 Thesis outline

The objective of this thesis is to present two novel bioinformatics method that I designed and implemented to solve two different computational problems in biological sequence analysis:

1. How to identify significant homologies in large databases with short query sequences.
2. How to combine RNA-seq data with DNA sequence features to accurately infer operons.

Therefore, the thesis is organized in two sections, and each section includes two chapters one to discuss the biological questions and the other to present the approach that I developed to solve the corresponding computational problems.

First Section. I found a solution for the first problem by building a statistical framework for homology searching in large biological sequence databases with short queries. The new solution is tightly linked to the problem of inferring homologies between distinct bacterial membrane proteins. The aim of this sequence analysis was to identify conserved and distant domain segments in transmembrane regions of different bacteria. This type of homology analysis uses a domain-oriented search strategy and it can be very useful to explore distantly related homologs or to identify homologs in insignificant Blast hits (Fortino *et al.*, 2010; Fortino *et al.*,in preparation).

Second Section. Regarding the second computational problem I developed a simple and useful prediction program which successfully infers the presence of operons by combining experimental operon identification with RNA-Seq data and standard operon prediction strategies. Essentially, I focused on the development and validation of a new transcriptome-based operon prediction method. This method is based on a multiple classifier system that combines different genomic (or static) and transcriptomic (or dynamic) features to produce condition-dependent operon maps (Fortino *et al.*, 2012; Fortino *et al.*, in preparation).

The results obtained with these different studies are very useful to make sequence analysis, in two important specific, different applications and contexts.

CHAPTER 2

Protein sequence comparison and Protein evolution

Today, with numerous genomes and countless genes sequenced, it is well established that recombination of DNA sequences encoding polypeptide domains has been a key process in the origin and evolution of new proteins (Gilbert *et al.*, 1987; 1997), and in evolution itself. However, there is no evidence to suggest how the first polypeptide domains arise.

Recently, Lupas *et al.* (2001) suggested that protein domains of living organisms originated from ancient short peptide ancestors called antecedent domain segments (ADSs), and these short segments are involved in a self-assembling system to produce single-domain homo- or hetero-multimers.

This model is in accord with recent theories about the Origin of Life (OoL). During the earliest stages of OoL there were no proteins, chaperones or ribosomes and protein-coding genes.

It has been proposed that membrane-spanning peptides constitued the first "information-rich" molecules during the OoL (Bywater *et al.*, 2009). According to this hypothesis, ancient proteins originated by a self-assembling system aggregates of short peptide, likely made from the 7 or 8 amino acids which were synthesized in the pre-biotic Earth (Miller and Urey *et al.*, 1959; Parker *et al.*, 2011).

The self-assembling system, together with the spontaneous formation of liposomes, led to formation of micelles containing short peptides and other components so that were able to transform them and to produce new molecules.

Firstly, this chapter discusses the different mechanisms by which new proteins arose during evolution and the roles of protein domains as independent evolutionary units. Secondly, the general aspects of domain identification and characterization by standard sequence analysis tools are explained. Finally, the new and old theories about the OoL are introduced, focusing on a model proposed by Bywater *et al.*, 2009 called transmembrane-peptide-first. The study reported here lays the foundation for the work of the following chapter where I present a computational BLAST-based method to infer ancient homologies in unrelated bacterial membrane proteins.

2.1 Protein Evolution

Proteins are the building blocks and functional units of all living organisms, involved in virtually every process within cells. Often, proteins with similar functions – which belong to the same or to different organisms - display significant similarity in their amino acid sequences and these similarities can be studied to understand the mechanisms for the emergence of new proteins.

New proteins arose during evolution by several different mechanisms: duplication of a single gene (including their regulatory regions), duplication of genomic regions (Li *et al.*, 2003), in some extreme cases, of the entire genome and divergence (Christoffels *et al.*, 2004; McLysaght *et al.*, 2002; Wolfe and Shields *et al.*, 1997).

Functional fragments of genes, or whole genes, are duplicated producing a family of related genes. They may remain in tandem in the same chromosome or may be inserted in other chromosomal loci. Newly duplicated genes and the original sequence can then undergo mutational events generating genes with new functions through the normal evolutionary processes of chance and selection.

Gene duplication has occurred regularly over evolutionary time and it is a fundamental characteristic of evolution (Ohno *et al.*, 1970). Particularly in multicellular organisms, many proteins derive from basic units (domains) allowing proteins to be grouped into families, super-families, and folds (Patthy *et al.*, 2003). Prokaryotic organisms can acquire new proteins also by more complex genetic events such as lateral (also called horizontal) gene transfer (Beiko *et al.*, 2008). Through this process bacteria acquire in a single event one or several genes from related species by conjugation, transformation, by infection of viruses or by plasmids.

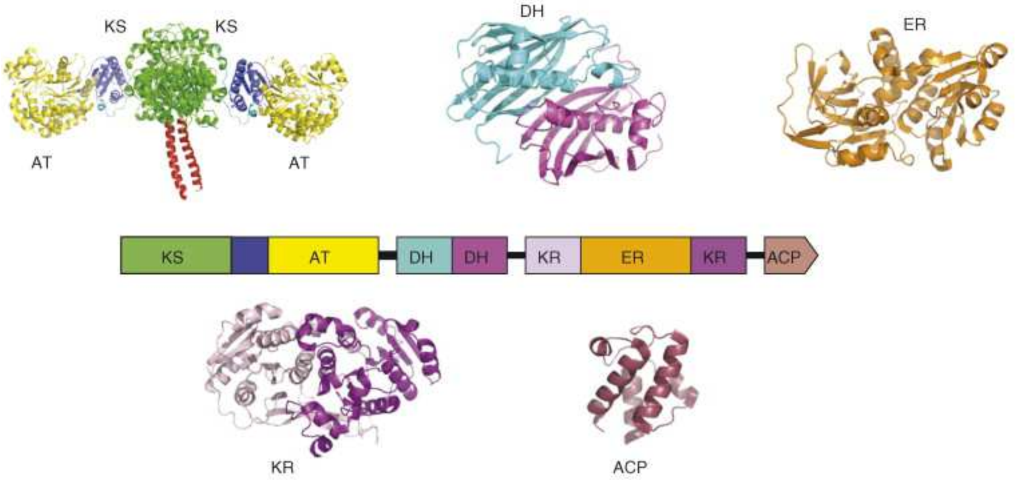


Fig. 5 – PKS domains and their structures (Gokhale *et al.*, 2007). Schematic representation of various domains and linkers are depicted on the primary sequence of a PKS polypeptide chain. The three-dimensional structures available for various catalytic domains and linker regions are also shown.

Eukaryotic genes (particularly those of Metazoa) typically contain coding exons interrupted by noncoding introns. The exons often correspond to the functional domains of the coded proteins (Liu and Grigoriev *et al.*, 2004). In addition, exons with similar sequence can be found as parts of very different genes. Therefore, exons are believed to be functional units in their own right and eukaryotic genes have arisen by shuffling individual exons. Gilbert *et al.* (1978) proposed that in Metazoa and plants new proteins are generated by shuffling genomic sequences encoding specific polypeptide domains. Recombination of DNA sequences encoding polypeptide domains is believed to be a key process in the origin and evolution of proteins, (Gilbert *et al.*, 1997) and in evolution itself. The intron–exon organization of eukaryotic genes implies that new arrangement of exons is obtained by recombination within the intervening intron sequences, yielding rearranged genes with different functions (Schmidt and Davies *et al.*, 2007).

Analyses of three-dimensional structures of proteins have shown that proteins, particularly those of Metazoa, are generally organized in distinct domains (Figure 5), that are compact and spatially distinct units with specific functions. These units could represent the basic evolutionary components that formed proteins (Rossmann *et al.*, 1974; Murzin *et al.*, 1995). From an evolutionary perspective, protein domains can be described as significantly sequence-similar homologs that are often present in different molecular contexts.

2.2 Domain duplication and recombination

The duplication and shuffling by recombination of functional domains is an important force driving protein evolution (Vogel *et al.*, 2005). The fact that many extant proteins contain duplicated domains suggests that complex proteins have evolved from simple proteins mainly via domain duplication. Domain duplication is an important prerequisite of evolution on organism complexity and speciation (Ranea *et al.*, 2004; Vogel *et al.*, 2003; Muller *et al.*, 2002).

A crucial characteristic of a protein domain is that once synthesized it folds independently of the rest of the protein. Thus, domain shuffling is considered the main mechanism for the rapid generation of novel domain combinations (Doolittle *et al.*, 1995). It is reasonable that the diversification of proteins by recombination of domain has contributed significantly to the observed accelerated evolution of Metazoa, since this mechanism facilitates the rapid construction of multi-domain extracellular and cell surface proteins that are indispensable for multicellularity (Patthy *et al.*, 2003; Lundin *et al.*, 1999). Besides, since the same domains are reused in different combinations, domain duplication is an important prerequisite for novel domain by shuffling rearrangements. Observed domain combinations are only a small fraction of all possible combinations (Chothia *et al.*, 2003). This shares a similarity with the evolution of protein folds and suggests that protein evolution could be affected by functional and structural constraints at all levels.

2.3 Domain origin

Sequence analyses of proteins have revealed that various domains are of ancient origins because they are widespread in each of the three forms of cellular life, archaea, bacteria, and eukarya, whose common ancestor may existed over three billion years ago. The persistence of such domains implies that they are either hyper adaptable, suited to many beneficial functional niches, or that they are essential to fundamental cellular processes (Doolittle *et al.*, 1994; Gerstein *et al.*, 1997; Wolf *et al.*, 1999).

The formation of a completely new DNA sequence coding for a novel domain is very difficult to achieve due to the enormous combinatorial possibilities that this would require and to time needed to select for a given function. Lupas *et al.* (2001)

proposed that modern protein domains evolved by joining together short polypeptide segments (or sub-domains) that were capable of folding and conveying valuable functions. It has been suggested that the ancestors of these sub-domains arose by spontaneous association of peptides (or fragments) and that these assemblies led to functional advantage. Therefore, fusion of primitive genes encoding these fragments was preferentially selected by evolution

The importance of short polypeptides for domain origin is evident in modern proteins that contain either homologous repeats or very short sequence-similar motifs embedded in unrelated or non-homologous structures. Single amino acids or short peptide motifs may be repeated in proteins, thus they likely derived from a short polypeptide ancestor with only a single repeat. Furthermore, short highly sequence-similar motifs, such as Asp-box and helix hairpin- helix (HhH) motifs, have been identified in non-homologous structures (Copley *et al.*, 2001; Doherty *et al.*, 1996).

Currently, there are no valid models that may explain why unusual short motif sequences (or short gene segments) have been successfully duplicated and incorporated into different proteins. Of course, whether these motifs are ancient or more modern, their presence suggests that domains might be divisible and have arisen by recombination of shorter sequences.

The past occurrence of short polypeptides lead us to hypothesize think that complex single-domain structures might have arisen by the fusion of simpler substructures (or first-genes) and that complex multi-domain proteins arose by domain shuffling.

2.4 Domains in sequence analysis

In sequence analysis, domains are viewed as segments of aminoacid of minimal length (i) with a specific biological function (ii), that are characterized by significant sequence-similar homologs occurring in different molecular contexts (iii). This is a complementary domain definition based on the sequence homology and it is widely used in domain annotation.

Homology between protein regions can be identified using sophisticated tools such as BLAST (Altschul *et al.*, 1990). However, not all residues in a protein domain/family are equally well conserved. Domain detection may employ pairwise algorithms or, more effectively, generalized profile methods that determine multiple alignments of known members of a domain protein family in order to estimate the frequency of site-specific residues (Gribskov *et al.*, 1987). These methods provide “domain descriptor” can then be used to identify other homologs that might not have been previously thought to be members of the corresponding domain family.

A widely used profile method is based on Hidden Markov models (HMMs by Eddy *et al.*, 1998). These models have several advantages over standard profiles. HMM profiles have a formal probabilistic basis and robust theory behind gap and insertion scores, in contrast to standard profile methods which use heuristic methods. HMMs apply a statistical method to estimate the true frequency of a residue at a given position in the alignment from its observed frequency while standard profiles use the observed frequency itself to assign the score for that residue. Application of profile HMMs for domain detection has been shown to be very successful and has had a high impact on the understanding of the structure of newly sequenced genes and genomes (Bateman *et al.*, 2002).

The most important collection of domain models, based on HMM profiles, is Pfam database (Finn *et al.*, 2010). The Pfam database has two components: Pfam-A and Pfam-B. Pfam-A contains, for each well-characterized protein domain families, a hidden Markov model, full alignments, associated annotation, literature references, and database links. Pfam-B is an automatically generated set of protein families derived from sequence clusters taken over from the ADDA database (Heger *et al.*, 2005). Pfam-B families have no associated functional annotation and no profile-HMMs.

2.5 Infer ancient domain homologies

In the previous paragraphs, I formulated the hypothesis that complex single-domain structures might have arisen by the fusion of smaller sequences such as short peptides. In addition, short highly sequence-similar motifs have been identified in non-homologous structures. For the evolution and spread of these short peptides and structure motifs, Ponting and Russell *et al.* (2002) suggested that these segments represent ancient conserved domain cores that have been preserved because of their importance for some function and structure, while their flanking structures have been exposed to greater changes. As results, these short peptides have been duplicated and incorporated into unrelated protein structures.

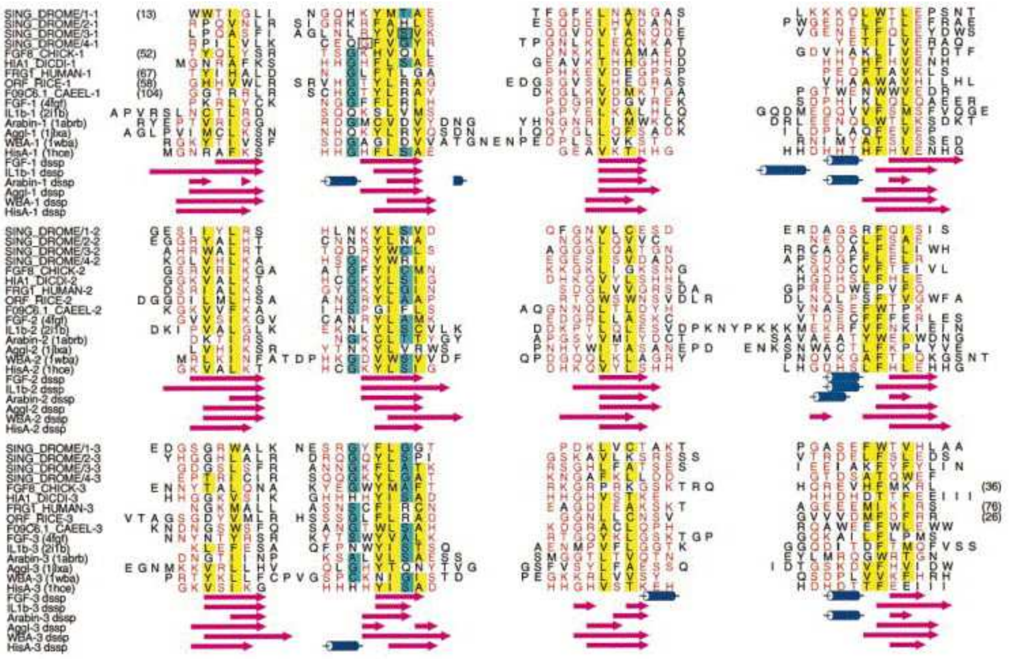


Fig. 6 – Identification of distant homologues in fibroblast growth factors (Ponting and Russell *et al.*, 2000).

The mentioned authors have performed a homology analysis on all β-trefoil proteins to identify distant homologues of fibroblast growth factors. The fibroblast growth factor families have distinct functions and occur in different subcellular localizations. Significant sequence identities identified by a multiple alignment of all proteins, are partially reported in Figure 6. These data have provided statistical evidence that b-trefoil proteins are all homologues (having arisen from a common ancestor), despite their differences.

This means that the search of small homologous sequences in unrelated protein domains may help us in finding ancient domain homologies and that a homology analysis domain-centric can assist this invistigation. However, short homologous segments in domains are hard to identify via currently available methods in sequence analysis. Comparison of any short aminoacid sequence, with fewer than 50 aa, in protein databases generates a plethora of insignificant alignments, a few with proteins that possess the same arrangement of domains, many more with proteins with different domain arrangements, and some with multiple hits within the same sequence. From such results it is difficult to identify true relationships among the different regions of multi-domain protein sequences.

2.6 OoL: information storage and propagation

Many scientists have studied the origin of life. Different schools of thought have developed different theories for the pre-biotic origin of life. These theories proposed various initial mechanisms and thus, each theory proposes a different molecule as the initial one. These theories are: RNA-based, metabolism-based, lipid-based and peptide-based.

Fundamental requirements in a hypothetical pre-life system are the storage and the propagation of information. Indeed, the main aim of research of the origin of life has been the understanding of how the first macromoleculse allowed storing of information in the sequence of its monomers (Joyce and Orgel *et al.*, 1993). Since, both genes and catalysts play a key role in supporting theories about the origin of life, a widespread view is that life began with the formation of a polymer having both catalytic activity and the ability to contain and propagate its chemical identity over generations.

Therefore, most scientists believe that life went through an early RNA-dominated phase (called “RNA World”). According to this hypothesis, RNA stored genetic information and catalyzed the chemical reactions of primitive cells (RNA-first model). Only later in evolutionary time DNA took over as the genetic material and proteins become the major catalysts and structural components of cells. But, it is difficult to understand how RNA molecule as a whole emerged from small molecules (Joyce *et al.*, 1991) and with the support of simple metabolic processes. In addition, the RNA-first model needs mechanisms that allowed the concentration of molecules to avoid that activated nucleotides capable of self-polymerization into RNA in solution could be diluted (dilution problem), compromising the subsequent reactions. Therefore, in such system both information storage and propagation were jeopardized. Many hypothesis have been proposed for the “dilution problem” (Robertson and Miller *et al.*, 1995; Stribling and Miller *et al.*, 1991; Wächtershäuser et al., 1988; Ferris *et al.*, 1996; Sowerby *et al.*, 2002).

On the other hand, these physical mechanisms seem to be inconsistent with the first stages of generating monomers (Deamer *et al.*, 2002). Several models reject the idea of a RNA-first model and suggest that a simple metabolism emerged first followed by the capacity to duplicate polymers. A metabolism first model would have emerged either spontaneously or by a process of random drift, and once established, this have also exhibited a simple, non-genomic replicating capability (Dyson *et al.*, 1985; Kauffman *et al.*, 2000, 1993; Segré *et al.*, 2000; New and Pohorille *et al.*, 2000).

The metabolism-first hypothesis consists of several different hypotheses proposed by different researchers about how life first emerged. These hypotheses suggest that ordered chemical reactions, and not information replication, was the property of the initial molecules. They propose that a primitive type of metabolic life characterized by a series of self-sustaining reactions based on monomeric organic compounds made directly from simple constituents (CO_2, CO) arose in the vicinity of mineral-rich hydrothermal systems. According to this theory, at first, life did not have any molecule~~s~~ with capacity to storage information. Subsequently, interconnected networks of self-sustaining reactions evolved more complex over time. At some point during these stages of evolution, genetic molecules were somehow incorporated into the system and life as we know it took form.

The different hypotheses differ in the nature of the self-sustaining chemical reactions that characterized early life. One of these hypotheses was proposed by Wächtershäuser *et al.* (1992). He proved that iron pyrite can catalyze biochemical reactions and that the existence of iron pyrite on the early Earth would have rapidly led to a proliferation of chemical reactions that take place in living organisms today. Indeed, many enzymes present in modern cells contain iron pyrite clusters within their structure, attesting to the importance of its role. The “metabolism first” hypothesis basically suggests that in the early stages of prebiotic era nucleic acids that replicated themselves did not exist, but a structure that metabolized inorganic compounds were present on the surface of iron pyrite, most likely around small deep-sea hydrothermal vents.

Both models, “RNA-first” and “metabolism-first”, offer possibilities for storage and transmission of information, although in very different ways (Bywater *et al.*, 2009). However, RNA is highly water soluble and very susceptible to hydrolysis. Nucleic acids of extant species have a high fidelity of DNA replication or RNA transcription that is based on the activity of complex proteins (DNA and RNA polymerase) and on proof reading activity that ensure a very low level of mutations (ca. 10^{-7} per genome per replication in extant prokaryotes, Drake *et al.*,1998). It has been hypothesized that ancient RNA molecules may have had only a frequency of mutation rate of ca. 10^{-2}, too high to maintain a randomly generated sequence with self-replicative activity.

On the other hand, this hypothesis is based on the suggestive argument that catalytic cycles selected a subset of molecules from the overwhelmingly diverse repertoire of organic molecules that could arise under prebiotic conditions (Morowitz *et al.*, 2000; Smith *et al.*, 2004; Schuster *et al.*, 2000). Unfortunately, almost all proposed models are not supported by experimental data and the role of larger cycles at the origin of life are rather questionable (Orgel *et al.*, 2008).

An alternative model is based on a self-organization system that arranged different molecules without guidance or management from an outside source (Koch *et al.*, 1985; Morowitz *et al.*, 1992; Deamer *et al.*, 1997). Such system would have used the liquid water as a source of free energy, and a minimal set of organic compounds (or prebiotic chemicals) capable of self-assembly to assemble more (Monnard and Deamer *et al.*, 2002).

2.6.1 Necessity of a self-assembling system

Self-assembly in liquid water occurs when small amphiphilic molecules spontaneously associate by hydrophobic interactions into more complex structures with defined compositions and organization. Examples include the assembly of amphiphilic molecules into micelles, monolayers, and bilayers in the form of vesicles (Figure 7).

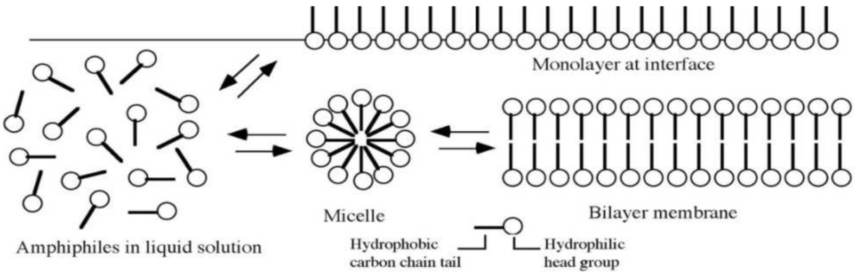


Fig. 7 – Self-assembled structures of amphiphiles (Deamer *et al.*, 2002). Lipids into water form spontaneously aggregates, which is because of their amphiphilic structure. The hydrophilic part tries to yield the hydrophobic one from the water. Common aggregates are micelles and lipid bilayer. A very important aggregate is the lipid bilayer, where two monolayers are arranged with the fatty acid chains being exposed to each other and lipid vesicles are formed.

Therefore, the amphiphilic compounds allow self-assembly into membrane-bounded structures that encapsulate subsets of the mixed components. Solutions inside these structures are highly concentrated allowing to overcome the "dilution problem". In addition, since the membrane-bounded structures can maintain specific groups of macromolecules within, they allow the selection of variations, leading to the "speciation". An important function carry out by the membrane boundaries is that to provide permeability barriers which control the flow of information between the building structures and the external environment (or between the cytoplasm and the external environment). The membranous vesicles capture and maintain light energy and redox energy by using pigment systems and electron transport to generate electrochemical proton gradients as a source of free energy. This free energy can drive transport processes across the membrane boundary. The amphiphilic compounds, with their selective permeability, have significantly facilitated exchange of solutes between primitive cells and their environments, leading to accumulation of nutrients from the environment into the protocell interior.

On the early Earth, it seems likely that spontaneous self-assembly processes led to formation of enormous numbers of microscopic structures that were able to use the energy and nutrients available in the external environment to reproduce their structures. This self-assembling system led to the first forms of cellular life.

2.6.2 Minimal sets of complex and structurally diverse chemicals

In a hypothetical self-assembling system, a minimal set of prebiotic chemicals must has been available to provide building blocks for polymer synthesis and formation of membrane-bounded compartments (Deamer *et al.*, 2006). This minimal set would have to incorporate some lipids that, through their assembling, were able to form primitive “cells”, and peptides, inside the first membranes, that were able to realize some of the basic life functions that are necessary for a living system to exist (Bywater *et al.*, 2009). This led to a new theory about the origin of life: “lipid-first’’ model or “Lipid World” scenario (Luisi *et al.*, 1996; Deamer *et al.*, 2002; Segré *et al.*, 2001).

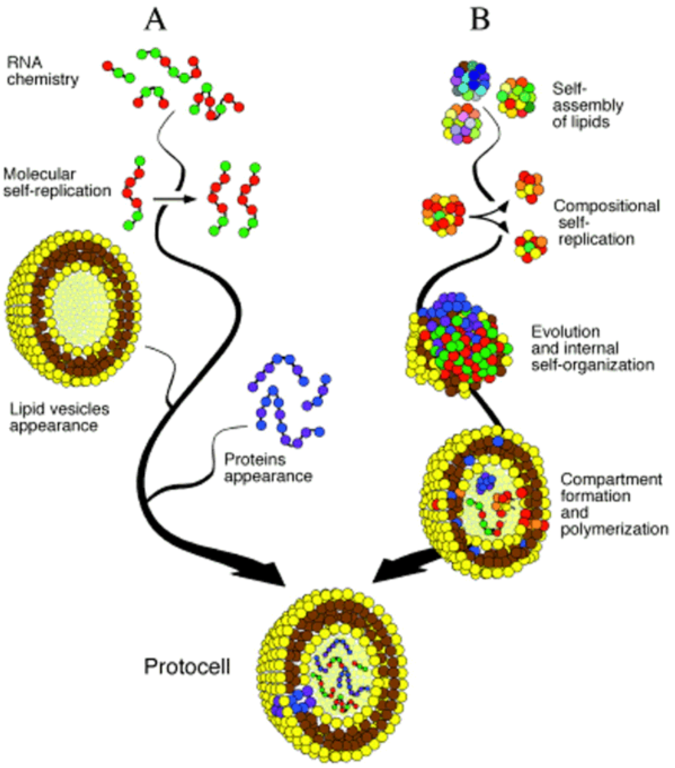

A
B
RNA chemistry
Molecular self-replication
Lipid vesicles appearance
Proteins appearance
Self-assembly of lipids
Compositional self-replication
Evolution and internal self-organization
Compartment formation and polymerization
Protocell

Fig. 8 – Two possible scenarios leading from a 'primordial soup' to a rudimentary protocellular structure (Segré and Lancet et al., 2000). (A) The 'biopolymer first' scenario, according to which the emergence of self-replicating informational strings such as RNA and proteins are assumed to have had an independent origin from that of lipid encapsulation. (B) The 'Lipid World' scenario, which maintains that the roots of life could have been aggregates of spontaneously assembling lipid-like molecules endowed with capabilities for dynamic self-organization and compositional inheritance.

This model combines the potential chemical activities of lipids and other amphiphiles, with their capacity to go through spontaneous self-organization into complex molecular structures, such as micelles and bilayers (Figure 8). These structures can contain the products of other prebiotically occurring chemical processes. The “Lipid World” scenario may represent an intermediate “mesobiotic” phase that linked a set of abiotic random collection of organic molecules with a biotic protocell that contains long biopolymers, and so allowed information storage, catalysis and replication (Shenhav *et al.*, 2003).

An interesting model based on the lipid world scenario was proposed by Bywater and Conde-Frieboes *et al.* (2005). The authors introduced new features to the previous models: (1) rapid cycles of catalysis and transport of material, (2) desegregation and segregation (3) cross-catalysis and (4) compartmentalization. All these features were essential to build up a framework capable of developing information-rich molecules like peptides and RNA. However, the lipid-first model seems to be critical for OoL, because lipids are not able to storage and propagate information. Bywater *et al.* (2009) suggested that, in addition to the lipids, the minimal set would have to incorporate also some peptides, inside the first membranes, that allowed the first complex molecular structures to communicate with the external environment, and to replicate/propagate their structures.

During the earliest stages of OoL there were no proteins, chaperones or ribosomes. It has been proposed that membrane-spanning peptides must have been the first "information-rich" molecules during the OoL (Bywater *et al.*, 2009). According to this model, ancient proteins originated by self-assembling aggregates of short polypeptide chains, likely made from the simplest and oldest aminoacids which were synthesized in the pre-biotic Earth (Miller and Urey *et al.*, 1959; Parker *et al.*, 2011). Membrane-spanning peptides are stabilized in membrane since lipids have molecular chaperone properties that assist (poly) peptides in their folding similarly to canonical protein-based chaperones (Dowhan and Bogdanov *et al.*, 2011; Bywater and Conde-Frieboes *et al.*, 2005; Bogdanov et al., 1996; Bogdanov *et al.*, 1999). It has been suggested that during the early stages of prebiotic syntheses, the physical and chemical properties of proteins and lipids may have co-evolved within the lipid environment of membrane. More complex peptides with catalytic properties associated with the membrane could have developed later as independently folding functional units formed by extension of the protruding ends of the transmembrane peptides within an aqueous environment.

However, though lipids have the intrinsic characteristic of self-assembly in an aqueous environment forming various micelles and vesicles (Luisi *et al.*, 2004), they do not permit the exchange of molecules with the environment. Thus, the emergence of short trans-membrane peptides not only have stabilized membrane but generated the segregation and compartmentalization of molecules inside the membrane vesicles in which life eventually emerged.

CHAPTER 3

Homology analysis of membrane-spanning regions

According to "the theory of mini genes" by Gilbert *et al.* (1997), in the early stages of evolution ca. 4 billion years ago, the first genes were originally constituted by small stretches of DNA coding for short polypeptides of 15-20 aminoacids.

In spite of the extensive theoretical and experimental studies dedicated to the above theory, so far the presence of any "signature" about the existence of such first genes has been elusive. Lupas *et al.* (2001) has suggested that modern protein domains evolve by combinations of ancient short polypeptides that together are able to fold and convey valuable functions. Therefore complex single-domain structures might be arisen by the fusion of these substructures as well as we think that complex multi-domain proteins arise by domain shuffling or exon shuffling.

In addition, since it has been proposed that membrane-spanning peptides must have been the first "information-rich" molecules during the OoL (Bywater *et al.*, 2009), it is possible that primitive forms of life may have had origin in their membranes short peptides and that during the expansion of prokaryotes those peptides may have been used in unrelated membrane-bound proteins.

Therefore, TM domains of integral membrane proteins (TMPs) can support the existence of ancient conserved TM domain cores that have been preserved because of their importance during evolution, while their flanking structures have been exposed to greater changes (Ponting and Russell *et al.*, 2002). As a result, these short peptides may have been duplicated and incorporated into distinct TMPs.

In order to support these hypotheses, I have started an investigation to identify significant and distant homologs in TM domains of different classes of TMPs (in collaboration with Prof. Bruno Maresca). This investigation has been accomplished with a statistical framework BLAST-based that finds and statistically validates homologies in large biological sequence databases using short queries.

This chapter presents a novel statistical framework BLAST-based to detect distant homologies conserved in transmembrane domains of different bacterial membrane proteins. Using this framework, 5,216 distant homologies in TM domains have been identified by screening 11,771 transmembrane sequences of *Salmonella* spp. The results show that the proposed framework detects distant homologies which could represent ancient signatures about the existence of primeval genetic elements (or mini-genes) coding for short polypeptides that formed, through a primitive assembly process, more complex genes.

3.1 The biological question

Recombination of DNA sequences encoding polypeptide domains is believed to be a key process in the origin and evolution of proteins (Gilbert *et al.*, 1997), and in evolution itself. However, there is less evidence to suggest how early genes and protein domains, themselves, arose.

In agreement with the ”Exon Theory of Genes” by Gilbert, in the primeval stages of evolution the earliest genetic elements (or first genes) coded for short polypeptides of 15–30 aminoacids. The formation of a completely new DNA sequence coding for a novel domain is very difficult to achieve due to the enormous combinatorial possibilities that this would require and to time needed to select for a given function. Therefore, it has been suggested that modern protein domains evolved by fusion and recombination from a more ancient peptide world in which short peptides were involved in a self-assembling system to produce the first single-domain structures. Then, fusion of primitive genes encoding these fragments was preferentially selected by evolution (Lupas *et al.*, 2001).

The theories about the mechanisms of protein and domain evolution involve recent theories about the OoL. Bywater *et al.* (2009) proposed that membrane-spanning peptides must have been the first “information-rich” molecules for the 'Lipid World' scenario. The 'Lipid World' scenario (Segré *et al.*, 2001) maintains that the roots of life could have been aggregates of spontaneously assembling lipid-like molecules endowed with capabilities for dynamic self-organization and compositional inheritance. However, though lipids have the intrinsic characteristic of self-assembly in an aqueous environment, they do not allow the exchange of molecules with the environment, and so the self-replication (storage) and transmission of information (propagation). Therefore, other prebiotic components such as short transmembrane peptides were necessary to stabilize the membrane, generate the segregation and compartmentalization of molecules inside the membrane vesicles in which life eventually emerged.

Hence, we suppose that ancient short transmembrane peptides supported the first primitive life forms and that during the expansion of prokaryotes these peptides may have been used in unrelated membrane-bound proteins. In support of this hypothesis I developed a homology analysis of bacterial TMPs can show the presence of significant homologue short trans-membrane fragments (trans-membrane fragment - TMF) in different bacteria and in distinct TMPs.

3.2 Homology search for genomic sequences database

The advent of high-throughput sequencing technologies and initiatives such as the Human Genome Project and very recently the ENCODE project (Bernstein B.E., Birney E., Dunham I., Green E.D., Gunter C., Snyder M. and 594 collaborators. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489:57-74, 2012) led to an explosion of genomic data in recent years. As a result, available data banks such as GenBank (Benson *et al.*, 2005), EMBL (Kanz *et al.*, 2005) and DDBJ (Miyazaki *et al.*, 2004) contain over 60 gigabytes of uncompressed sequence data and continue to exhibit exponential growth. Exploring such data collections provide valuable insights into the characteristics of proteins and DNA that will lead to important discoveries in biology or to confirm important not yet proved theories.

In sequence analysis one of the most important tasks is homology search. Homology search tools compare a query sequence to every sequence present in a collection to identify highly-similar and possibly related entries or matches. A poorly-annotated or unknown sequence can be used to query a large database of biological sequences to identify a related and well-annotated sequence in the collection. Similar sequences often share a common three-dimensional structure, have the same function in an organism and share a common evolutionary origin. Therefore, homology search can provide an insight into the structure, function and evolutionary origin of a newly sequenced or poorly annotated protein or strand of DNA.

To detect homologies significant similarities into the sequences under study must be recognized. The degree of similarity between biological sequences is measured using a sequence alignment algorithm, a comparison method that models the process through which proteins or DNA evolve and measures the number of elementary changes required to transform one sequence into another. Comparison by sequence alignment is a very powerful analysis tool. Dynamic programming is one of the most popular and efficient approach to compute the optimal alignment between two sequences (Smith-Waterman and Needleman-Wunsch). However, the exhaustive alignment algorithms are too slow for searching large genomic databases such as GenBank (a search of GenBank can take hours or even days on a modern workstation). Therefore, several heuristic approaches have been developed that renounce accuracy for substantially faster search. The most successful heuristic approach is BLAST (Altschul *et al.*, 1997).

BLAST provides fast yet sensitive search of large genomic databases and it is the most popular homology search tool. The blast software is installed in almost every medium- to large-scale molecular biology facility, and it has been widely-adapted to different hardware, operating systems, and tasks. The online interface at the popular National Center for Biotechnology Information (NCBI[2]) website is used to evaluate over 120,000 queries each day (McGinnis and Madden *et al.*, 2004), and the 1997 article describing the algorithm has been cited more than 10,000 times. BLAST remains the most successful approach to homology search despite a huge number of more recent methods such as index-based approaches (Kent *et al.*, 2002; Williams and Zobel *et al.*, 2002) and discontinuous seeds (Ma *et al.*, 2002; Li *et al.*, 2004).

3.3 Working with BLAST

The blast algorithm is a four-stage process that is both efficient and effective for searching genomic databases. The steps progressively reduce the search space, but each one is more fine-grain and takes longer to process each sequence than the previous; blast algorithm works with three parameters (the word size *W*, the word similarity threshold *T* and the minimum match score *S)* and two inputs (a query sequence and a database sequence).

In the first stage, a list of all words of length *W* that have similarity *T* to some word in the query sequence *q* is generated. Then, the database sequence is scanned for all alignments, called *hits*, of words *s* in the list. In the second stage, each such hit is extended in both directions without using *gaps* until its similarity score reaches the

[2] http://www.ncbi.nlm.nih.gov/

threshold *S*. Such extended ungapped local alignments are called high-scoring segment pairs (HSPs). In the third stage, gapped alignments are performed between sequences using a similar approach to the Smith-Waterman algorithm. In the final stage the alignments themselves are recorded using the traceback process and displayed to the user.

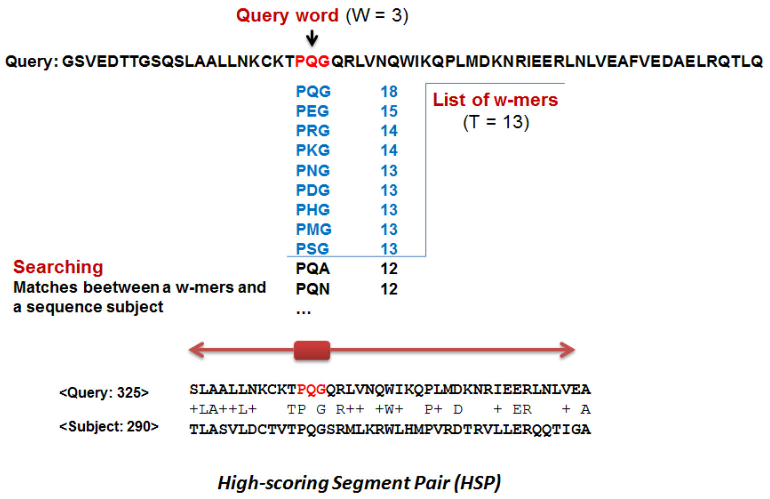

Query word (W = 3)
Query: GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQ
PQG 18
PEG 15
PRG 14
PKG 14
PNG 13
PDG 13
PHG 13
PMG 13
PSG 13
PQA 12
PQN 12
...
List of w-mers
(T = 13)
Searching
Matches beetween a w-mers and a sequence subject
<Query: 325>
<Subject: 290>
SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA
+LA++L+   TP G R++ +W+   P+ D   + ER   + A
TLASVLDCTVTPQGSRMLKRWLHMPVRDTRVLLERQQTIGA
High-scoring Segment Pair (HSP)

Fig. 9 – BLAST procedure.

```
Score = 248 bits (129), Expect = 1e-63
Identities = 213/263 (80%), Gaps = 34/263 (12%)
Strand = Plus / Plus
Query: 161 atatcaccacgtcaaaggtgactccaactcca---ccactccattttgttcagataatgc 217
           ||||||||||||||||||||||||||||| ||     | |   || ||||||||||||||
Sbjct: 481 atatcaccacgtcaaaggtgactccaact-tattgatagtgttttatgttcagataatgc 539
Query: 218 ccgatgatcatgtcatgcagctccaccgattgtgagaacgacagcgacttccgtcccagc 277
           |||||||   ||||||||||||||||||||| || |            ||||||||||||
Sbjct: 540 ccgatgactttgtcatgcagctccaccgatttg-g------------ttccgtcccagc 586
Query: 278 c-gtgcc--aggtgctgcctcagattcaggttatgccgctcaattcgctgcgtatatcgc 334
           |  || |  | ||||||||||||||||||||||||||||||||||||||| |||||||||
Sbjct: 587 caatgacgta-gtgctgcctcagattcaggttatgccgctcaattcgctgggtatatcgc 645
Query: 335 ttgctgattacgtgcagctttcccttcaggcggga------------ccagccatccgtc 382
           |||||||||||||||||||||||||||||||||||            |||||||||||||
Sbjct: 646 ttgctgattacgtgcagctttcccttcaggcgggattcatacagcggccagccatccgtc 705
Query: 383 ctccatatc-accacgtcaaagg 404
            |||||||| |||||||||||||
Sbjct: 706 atccatatcaaccacgtcaaagg 728
```

Fig. 10 – Example of a query match reported by BLAST. In this example the default scoring system has been changed to highlight the occurrence of gaps and mismatches with a match scoring 1 point, a mismatch scoring -2 points, a gap opening scoring -2 points and gap extension scoring -1 points. Default gap penalties are -5 and -2 respectively. The word size used was 11 nt (default).

The extension of alignments requires a scoring system and a procedure for locally maximizing the score. This scoring system assigns points and penalties for matches, mismatches, gap formation, and gap extension – all these parameters being provided either as defaults or by the user. Besides, the scoring system takes in account the positive matches, that are residue pairs can commonly substitute for one another in proteins in a given substitution matrix.

An example of a blast match is shown in Figure 10. For each match, BLAST reports the raw score (calculated from a scoring system based on a selected substitution matrix) the corresponding normalized score (indicated as bit score), the expect value *E* (that is the number of alignments with scores greater than or equal to score *S* that are expected to occur by chance in a database search), the number of identities, the number of gaps and the information about the strand.

3.4.1 Statistical significance of alignments in BLAST

BLAST aligns the query to sequences in a collection and report high-scoring alignments to the user. In such context it is useful, particularly when searching large databases, to calculate the statistical significance of alignments, and the probability that an optimal alignment score is due to a chance similarity rather than a homologous relationship.

Tools like BLAST determine the statistical significance of each HSP computing the Expect-value (E-value). The E-value is used to normalize results and to determine the number of sequences that would match ours, if we were searching a database of random sequences.

In more details, the significance of alignments relies upon the random independence model that considers genomic sequences as string of letters randomly drawn from the relevant alphabet (Karlin and Altschul *et al.*, 1990; Altschul and Gish *et al.*, 1996).

Although each aminoacid and nucleotide base has a different frequency of occurrence (Robinson and Robinson *et al.*, 1991), this model assumes that there is no relationship between adjacent aminoacids and that genomic sequences can be represented by a zero-order Markov model.

This assumption is supported by the poor compressibility of protein and DNA sequence data (Nevill-Manning and Witten *et al.*, 1999). Under this assumption, Karlin and Altschul *et al.*, (1990) show analytically that optimal alignment scores between random or unrelated sequences follow an extreme value or Gumbel distribution when gaps are not permitted. That is, the probability of a pair of unrelated sequences having an optimal alignment score S that is greater than a specific alignment score s is defined as:

$$P(S > s) = 1 - \exp\left(-Kmne^{-\lambda}\right) \qquad (1)$$

Where *m* and *n* are the lengths of the sequences being compared and the constants *K* and *λ a*re synonymous to the location and scale of the distribution, and are dependent on the alignment scoring scheme.

Therefore, the E-value reflects the random background noise that exists for matches between sequences, and depends on the length of a query sequence, the sequences stored into database used for search and the dimension of this database.

Thus, shorter sequences have a high probability to be found by chance in the database with higher similarity scores thus, the E-value for these sequences is very high. In addition, when a database become greater in size, there is a higher chance that it will contain matching sequences, and this determines a lower the E-value.

However, in many contexts it would be very useful to establish the significance of a pair of sequences aligned by BLAST independently from the selected database and the length of the query sequences that we are using.

3.4.1 Filtering low-complexity regions

In the previous section, I described the method used by BLAST for measuring the significance of an alignment. This measure is based on the random independence model, in which it is assumed that sequences have a standard composition and each amino acid or base in the sequence has an independent probability of occurrence. Unfortunately, many sequences contain low-complexity or repeat regions, containing a bias towards certain amino acids, strongly violating this model (Kreil and Ouzounis *et al.*, 2003).

Common types of low complexity regions include short period repeats, aperiodic mosaics and homopolymers (Wootton and Federhen *et al.*, 1993). One solution to this problem is the use a low-complexity filter. The filter removes these regions by replacing them with the ambiguous residue character X for protein, or N for nucleotide data.

The default filtering method used by BLAST for protein query sequences is the popular SEG algorithm (Wootton and Federhen *et al.*, 1993; 1996). This filter function masks part of the query sequence so not enough of the query is left for use as alignment seed; this means that for short queries we need to set off any kind of filter for low complexity regions.

3.5 A new statistical framework to search homologies in TM domains of different bacterial membrane proteins

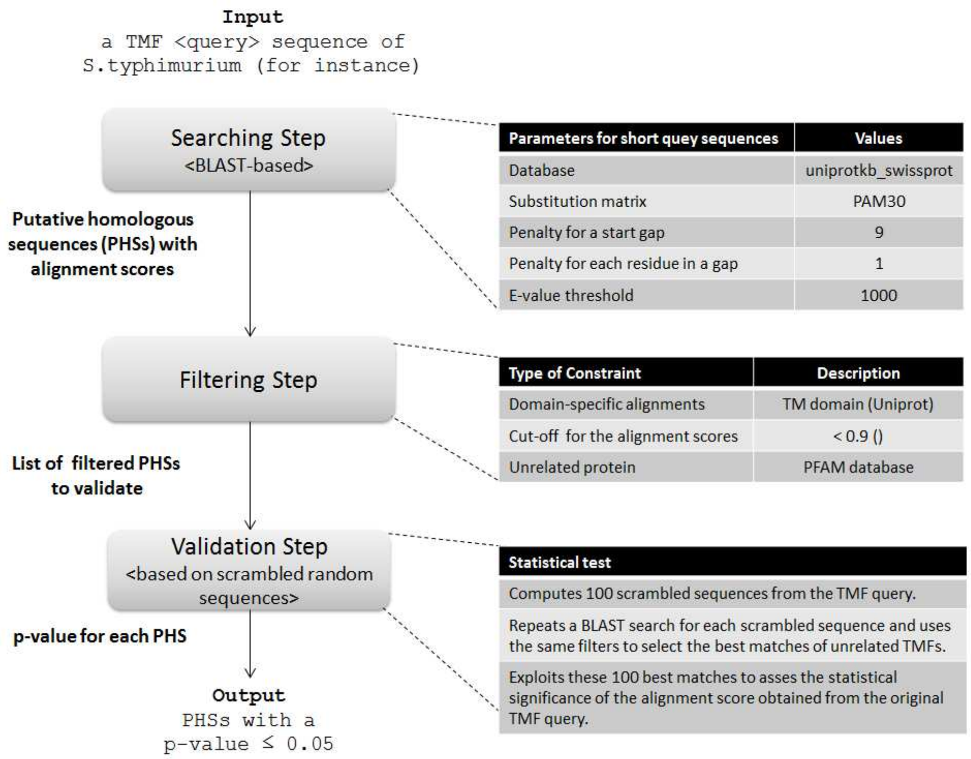
In this section I present a new, highly efficient homology search method to look for significant homologies among short TMFs of different integral membrane proteins and distinct bacteria.

Homologous TMFs can be identified using tools like BLAST. The heuristic BLAST uses a four-stage approach to effectively align a given query to sequences in a collection (or database) and report high-scoring alignments (or HSPs) to the user. Since similarity does not imply homology, BLAST uses alignment E-values (Karlin *et al.*, 1990) to distinguish alignments between homologous sequences from alignments between unrelated sequences. The E-value provides a measure of the statistical significance of the alignment. It approximates the probability that two random sequences, one the length of the query sequence and the other the entire length of the database (which is approximately equal to the sum of the lengths of all of the database sequences), could produce the calculated HSP score. Therefore, with short query sequences (eg. TMFs), the E-value increases because they have a high probability of occurring in the database purely by chance. Besides, the E-value becomes higher when larger databases are used.

Since I was interested in finding homologous TMFs among unrelated integral membrane proteins and since these sequences usually are very short, I had to develop a statistical framework that screens the insignificant hits obtained by BLAST and re-evaluates the statistical significance of those HSPs representing alignments between two TMFs of different bacterial membrane proteins (TMPs).

This framework relies on two external tools: a transmembrane prediction program (the ensemble method used by UniProt[3]) and a protein family classification system (the classifier used by PFAM). The first is used to determine the TM domains in bacterial membrane proteins and the second is employed to detect functional relatedness of two proteins.

The general schema is reported in Figure 11. The framework uses a three step approach. Firstly, it performs the BLASTp algorithm on a given TMF (used as query sequence) to achieve an initial set of alignments with the parameter setting to search with short query sequence. Secondly, it filters the aligned sequences to obtain a set of putative homologous TMFs: high-scoring alignments in TM domains of different bacterial TMPs. Finally, it estimates the statistical significance of each high-scoring alignment evaluating the corresponding score by a permutation test.


Input
a TMF <query> sequence of
S.typhimurium (for instance)
Searching Step
<BLAST-based>
Putative homologous sequences (PHSs) with alignment scores
Filtering Step
List of filtered PHSs to validate
Validation Step
<based on scrambled random sequences>
p-value for each PHS
Output
PHSs with a
p-value ≤ 0.05

Parameters for short quey sequences	Values
Database	uniprotkb_swissprot
Substitution matrix	PAM30
Penalty for a start gap	9
Penalty for each residue in a gap	1
E-value threshold	1000

Type of Constraint	Description
Domain-specific alignments	TM domain (Uniprot)
Cut-off for the alignment scores	< 0.9 ()
Unrelated protein	PFAM database

Statistical test
Computes 100 scrambled sequences from the TMF query.
Repeats a BLAST search for each scrambled sequence and uses the same filters to select the best matches of unrelated TMFs.
Exploits these 100 best matches to asses the statistical significance of the alignment score obtained from the original TMF query.

Fig. 11 – A graphical representation of the proposed statistical framework.

The permutation test scrambles 100 times the starting TMF query and repeat the first two steps, searching for homologs in all TM domains of the genome in which BLAST has found the matching sequence. By doing so, we can check whether the

[3] http://www.uniprot.org/manual/transmem.

alignments achieved with scrambled sequences score better than the alignment with the starting TMF query.

3.5.1 Searching step

The searching step is accomplished by BLASTp: the version of BLAST program that compares an aminoacid query sequence against a database of protein sequences. The TMFs are very short sequences, therefore, I adjusted BLASTp parameters in order to identify high-scoring alignments representing short and nearly exact matches. Here is a summary of the parameter set that I used to search high-scoring alignments with TMFs:

- The penalty for the initiation of a gap is equal to 9.
- The penalty for each residue in a gap is equal to 1.
- The low complexity filtering option is removed since this filters out larger percentage of a short sequence, resulting in little or no query sequence remaining.
- PAM-30 must be used as scoring matrix which is better suited to finding short regions of high similarity.
- The E-value must be equal to 1000 because in spite of all these expedients the E-value will be high when searching from shorter sequences.

With this paramter setting an initial list of putative homologous sequences (PHSs) is achieved for each TMF. These alignments can represent ancient conserved domain cores conserved in unrelated TMPs.

3.5.2 Filtering step

The proposed framework uses five different filters to achieve, from the list of PHSs, the matches (or alignments) having a high similarity score and belong to different bacteria and to distinct TMPs.

The first two filters are useful to massively reduce the number of PHSs, acting directly upon the percentage identity and the number of gaps. Matches having a percentage identity greater than or equal to 90% were considered as an indication of homology between two TMFs of proteins with the same biological function, and so they were eliminated. Besides, matches including more than three gaps were discarded in order to search for short, nearly exact matches.

The third filter is conceived to remove from PHSs all the matches that are associated to the same protein family of the starting TMF query, that has been used to generate the list of PHSs. This filter uses the Pfam program to classify the TMPs in protein families and to link each matched protein to the corresponding Pfam protein family. The Pfam database is queried using the accession number that BLAST returns for each matched TMP. Therefore, I exploited the Pfam entries, retrieved by using the accession number, to check whether the Pfam protein family of a matched TMP is equal to the Pfam protein family of the starting TMF query.

The fourth filter is developed to verify whether the matching sequences overlap a transmembrane domain. More precisely, the matches with a percentage of overlapping transmembrane residues greater than or equal to 80% are removed from the list of PHSs.

The last filter selects the best matches in terms of alignment score. The proposed framework defines a simplified version of the alignment scoring function used by BLAST:

$$AS(TMF_m) = \frac{\#identities + \#positives}{length(TMF_q)} \qquad if\ \#gaps \leq nag \tag{2}$$

Where *TMF_m* and *TMF_q* indicate the starting TMF query and the aligned TMF, respectively. While, *nag* indicates the maximum number of accepted gaps inserted into either query or subject sequence (e.g., *nag*=3). If the number of total gaps (*#gaps)* is greater than the constant *nag*, the framework assigns a default value equal to 0.1. The final computed alignment score is compared with a minimum scoring value (e.g., 0.5) in order to select only high-scoring alignments.

The proposed scoring function considers the ratio between the sum of identity and positive matches on the transmembrane residues. Therefore, for those alignments involving partially aligned TM domains (with an overlapping TM region ≥ 80% and < 100%), the identities and positive matches not correspondent to transmembrane segments were discarded. Furthermore, the scoring function takes in account also the number of positives in a match using PAM30 as scoring matrix. A positive match indicates a pair of residues that replace each other more frequently than expected by chance into a specified scoring matrix.

3.5.3 Validation step

After identifying the best PHS involving high-scoring alignments in TM domains, a statistical test is performed to evaluate whether the scores of filtered PHSs are significant or not. For a filtered PHS, the statistical test will concern only the TM domains of the genome having the matching sequence (indicated as the target genome for the validation of a match).

The statistical test performs the following steps: (i) scrambles the starting TMF query, by itself, for 100 times in order to obtain a set of random TMFs with same aminoacid composition; (ii) runs BLAST to define a list of similar sequences (or matches) for each scrambled sequence; (iii) selects the matches involving transmembrane regions; (iv) selects the match with the best alignment score from each list; and (v) uses the list of the best 100 alignment scores to assess the statistical significance of the score achieved with the starting TMF query.

$$p_{val}\big(AS(TMF_m)\big) = (1 - X)/100 \tag{3}$$

Where

$$X = count(AS(TMF_m) > AS(TMF_{sm})) \tag{4}$$

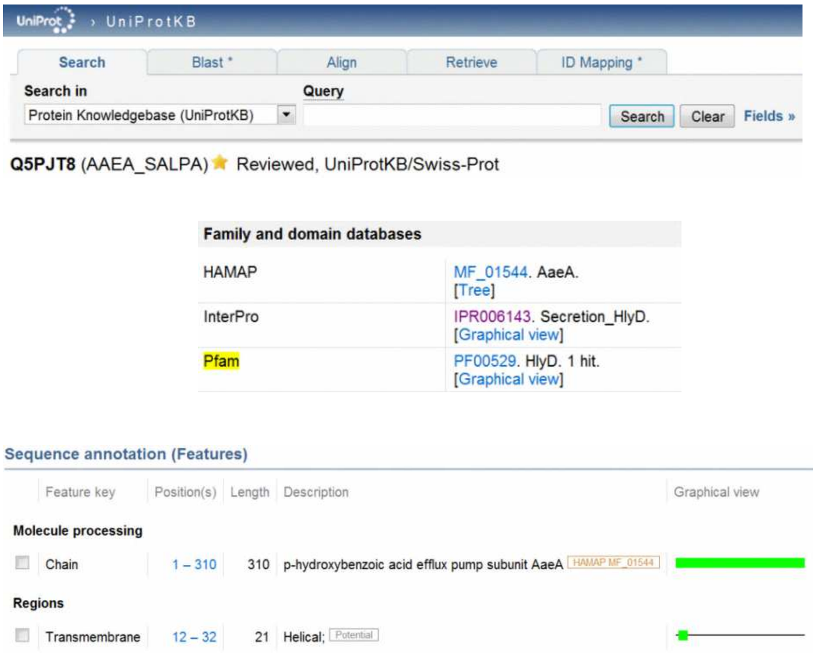
The TMF_{sm} indicates the best high-scoring alignments in a TM of the target genome, that the framework achieved with a scrambled sequence of the TMF query.

3.5.2 Implementation issues and annotations

For each selected PHS, a statistical test is performed to evaluate whether the corresponding alignment score is significant or not. The framework accomplishes this step using some external programs. Firstly, it queries the UniProtKB/Swiss-Prot (Wu *et al.*, 2006) database in order to organize an internal archive of ".dat" files that contain sequence and functional annotations about each TM protein record of a specific bacterial genome. Secondly, it uses the cross-references of each TM protein record to determine the exact location of all annotated TM domains and to identify the Pfam protein families associetd with it (Figure 12). This information is important to identify matches in TM domains and between distinct TMPs.

The most relevant homologies are organized and stored in a XML file. A short example of this XML file is shown in Figure 13. For each analyzed TMP, the XML file stores (i) the protein sequence, (ii) the first associated Pfam protein family and (iii) the list of the corresponding TMFs. Next to each TMF, the XML file records the homologies found in TMF related to different bacteria and distinct TMPs.

Furthermore, the XML file describes the homologies showing generic information about the TMP having the matching sequence, the part of the query sequence that has been aligned (*querySeq*), the matching sequence (*matchSeq*), the pattern with identity and positive matches (*pattern*), the TMF involved in the match (*bestTMFragment*) and the information about the p-value (*statisticalSignificance*).



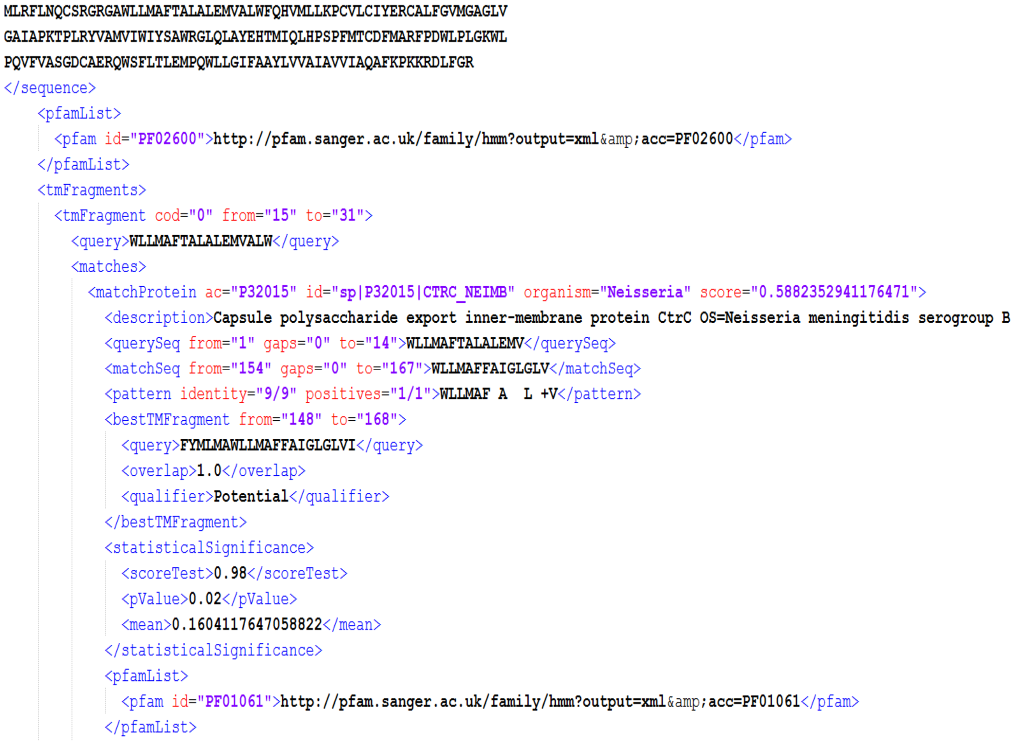
Q5PJT8 (AAEA_SALPA) ★ Reviewed, UniProtKB/Swiss-Prot

Family and domain databases	
HAMAP	MF_01544. AaeA. [Tree]
InterPro	IPR006143. Secretion_HlyD. [Graphical view]
Pfam	PF00529. HlyD. 1 hit. [Graphical view]

Sequence annotation (Features)

	Feature key	Position(s)	Length	Description	Graphical view
Molecule processing					
□	Chain	1 – 310	310	p-hydroxybenzoic acid efflux pump subunit AaeA HAMAP MF_01544	
Regions					
□	Transmembrane	12 – 32	21	Helical; Potential	

Fig. 12 – Screenshots of UniProt record information.



```
MLRFLNQCSRGRGAWLLMAFTALALEMVALWFQHVMLLKPCVLCIYERCALFGVMGAGLV
GAIAPKTPLRYVAMVIWIYSAWRGLQLAYEHTMIQLHPSPFMTCDFMARFPDWLPLGKWL
PQVFVASGDCAERQWSFLTLEMPQWLLGIFAAYLVVAIAVVIAQAFKPKKRDLFGR
</sequence>
    <pfamList>
      <pfam id="PF02600">http://pfam.sanger.ac.uk/family/hmm?output=xml&amp;acc=PF02600</pfam>
    </pfamList>
    <tmFragments>
      <tmFragment cod="0" from="15" to="31">
        <query>WLLMAFTALALEMVALW</query>
        <matches>
          <matchProtein ac="P32015" id="sp|P32015|CTRC_NEIMB" organism="Neisseria" score="0.5882352941176471">
            <description>Capsule polysaccharide export inner-membrane protein CtrC OS=Neisseria meningitidis serogroup B
            <querySeq from="1" gaps="0" to="14">WLLMAFTALALEMV</querySeq>
            <matchSeq from="154" gaps="0" to="167">WLLMAFFAIGLGLV</matchSeq>
            <pattern identity="9/9" positives="1/1">WLLMAF A  L +V</pattern>
            <bestTMFragment from="148" to="168">
              <query>FYMLMAWLLMAFFAIGLGLVI</query>
              <overlap>1.0</overlap>
              <qualifier>Potential</qualifier>
            </bestTMFragment>
            <statisticalSignificance>
              <scoreTest>0.98</scoreTest>
              <pValue>0.02</pValue>
              <mean>0.1604117647058822</mean>
            </statisticalSignificance>
            <pfamList>
              <pfam id="PF01061">http://pfam.sanger.ac.uk/family/hmm?output=xml&amp;acc=PF01061</pfam>
            </pfamList>
```

Fig. 13 – An overview of typical information reported in the XML file.

UniprotKB

The database UniprotKB[4] represents a comprehensive, high-quality and freely-accessible resource of protein sequences and it is provided with rich functional information. The UniProt-Knowledgebase[5] consists of two sections: Swiss-Prot and TrEMBL.

UniProtKB/Swiss-Prot contains records combining full manual annotation with computer-assisted, manually verified annotation performed by biologists and biochemists and based on published literature and sequence analysis. While, UniProtKB/TrEMBL contains records with computationally generated annotation and large scale functional characterization.

UniProtKB/Swiss-Prot records provide an integrated presentation of annotations such as protein name and function, taxonomy, enzyme-specific information (catalytic activity, cofactors, metabolic pathway, and regulatory mechanisms), domains and sites, post-translational modifications, subcellular locations, tissue-specific or developmentally specific expression, interactions, and diseases. Literature citations provide evidence from experimental data, which, along with feedback information from contacted authors, are regarded as information of the

[4] Uniprot website: http://www.uniprot.org/.

[5] Uniprot-knowledgebase website: http://web.expasy.org/docs/userman.html.

highest value, and are constantly being added to each record as they become available.

Furthermore, Swiss-Prot is currently cross-referenced to more than 50 different databases. Cross-references are provided in the form of pointers to information related to Swiss-Prot entries and found in data collections other than Swiss-Prot. For instance, we can look for the protein families related to a specific protein.

Transmembrane domains

UniprotKB provides information about the membrane-spanning regions of proteins. It stores information about the presence of both alpha-helical transmembrane regions and the membrane spanning regions of beta-barrel transmembrane proteins. The annotated transmembrane domains can be in three states of "reliability": (i) the transmembrane domains have been experimentally determined; (ii) the corresponding protein is related to a well-characterized family known to contain transmembranes; (iii) the TM domains have been predicted/detected by different prediction tools.

For the prediction of transmembrane domains, UniprotKB uses a combination of different prediction programs:

1) ESKW (Eisenberg *et al.*, 1984) – it gives the good length for the transmembrane domains.
2) TMHMM (Krog *et al.*, 2001) – it generally gives the good number of domains, but not the correct ranges;
3) MEMSAT (Jones *et al.*, 1994) – its predictions can be retained when confirmed by Phobius;
4) Phobius (Kall *et al.*, 2007) – it is used primarily to discriminate between signal sequences and transmembranes located at the N-terminus.

These tools predict only alpha-helical membrane spanning regions the positions of membrane spanning beta-sheet regions are annotated strictly according to experimental information. In Figure 14 is displayed an example of TMFs predicted by TMHMM (Krog *et al.*, 2001).

For predicted alpha-helical transmembrane regions at least two methods must return a positive prediction in order for a region to be annotated as transmembrane in UniProtKB/Swiss-Prot. When predicted N-terminal signal peptides and transmembrane domains overlap, then the Phobius prediction is used to discriminate between the two.

Pfam annotations

For each record in UniprotKB information can be accessed about similarities with other proteins. For instance, we can use the PFAM[6] protein families associated with a membrane protein to have information about the functions of the protein and understand if two homologous TMFs belong to two TMPs with the same functions.

The PFAM database is a large collection of protein families, based on multiple sequence alignments and hidden Markov model (HMMs) profiles (Eddy *et al.* 1990). My statistical framework uses Pfam annotations to select homologous trans-membrane sequences of different bacterial membrane proteins. Moreover, since a protein may belong to different families, during the comparison it considers the most

[6] PFAM Genome Distribution website: http://www.sanger.ac.uk/.

representative protein family of each protein, thus if two aligned TMFs have the same first PFAM protein family then the framework discards the coresponding match.

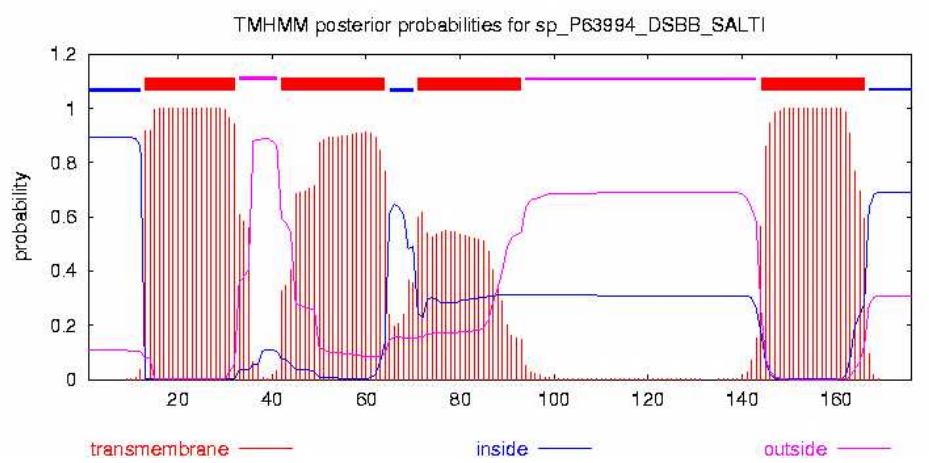
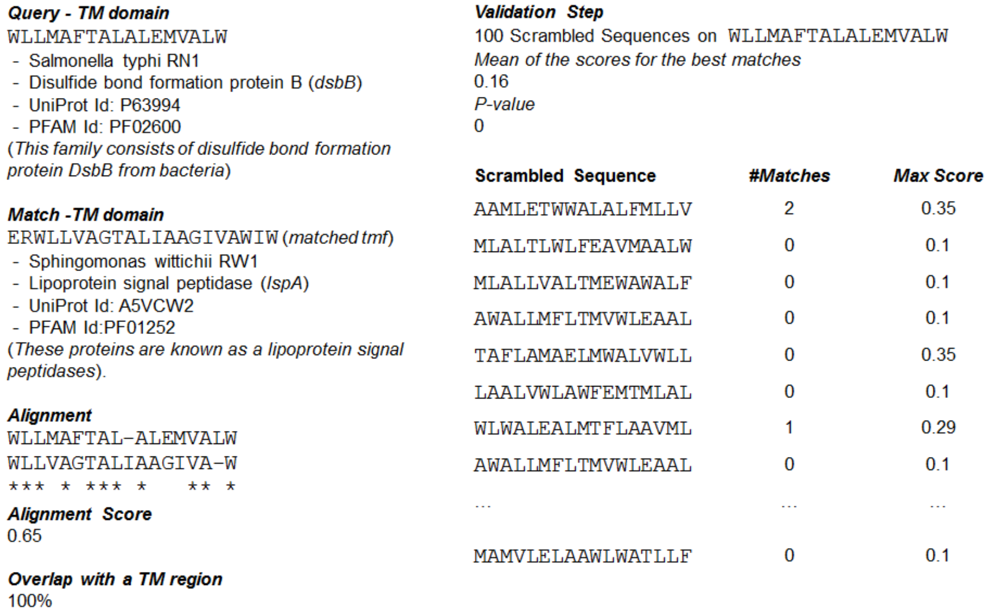


Fig. 14 – Example of TMFs predicted by TMHMM. Results using Disulfide bond formation protein B (Disulfide oxidoreductase) *Salmonella typhi*. TMHMM allows you to predict the location of transmembrane alpha helices and the location of intervening loop regions. This program will also predict which loops between the helicies will be on the inside or outside of the cell or organelle. This program will not detect beta sheet transmembrane domains. It takes about 20 aminoacids to span a lipid bilayer in an alpha helix.

3.6 Evaluation results

The proposed statistical framework has been tested on the TM domains of *Salmonella* spp. For this analysis 1,760 TMPs entries have been retrieved from the UniProt database, including the information about their membrane-spanning regions and their Pfam protein families. The framework, using the 11,771 TMFs of these proteins, identified 5,216 high-scoring alignments having a p-value less than 0.05 (see Fig. 17). Furthermore, to reduce the number of false positive matches, a cut-off value of 0.30 has been used. This cut-off value removes matches that achieved high-scoring alignments with scrambled sequences. Besides, the cut-off value was applied on the average alignment score obtained for aligning scrambled sequences; for instance, on the column “Max Score” in Figure 15.

Most of the homologies inferred with TMFs of *Salmonella* have an optimal p-value (which is very close to zero), because it did not identify a scrambled sequence that scored better than the TMF query. This can be observed from the histograms reported in Figure 16. The histogram of p-values shows a very high peak right at the point 0.0. Figure 15 describes an example of the validation step performed to assess the statistical significance of an alignment between two TMFs of different bacterial membrane proteins. In this regard, the framework tries to validate an alignment between the first TMF of the *dsbB* - disulfide bond formation protein B, and a TM domain of IspA – lipoprotein signal peptidase protein that belongs to *Sphingomonas wittichii*. The alignment overlaps a TM domain with a score equals to 0.65. During the scrambling process, the framework generates 100 random permutations of the query sequence (“WLLMAFTALALEMVALW”), and, for each one, it selects the best matching TM domain. Most of the scrambled sequences do not give matches belonging to a TM domain and, consequently, a default score value (equal to 0.1) is provided. Finally, the validation step provides a p-value; if this p-value is close to 0, then none of the 100 permutations generates a match with an alignment score greater than 0.65.



Query - TM domain
WLLMAFTALALEMVALW
- Salmonella typhi RN1
- Disulfide bond formation protein B (*dsbB*)
- UniProt Id: P63994
- PFAM Id: PF02600

(*This family consists of disulfide bond formation protein DsbB from bacteria*)

Match -TM domain
ERWLLVAGTALIAAGIVAWIW (*matched tmf*)
- Sphingomonas wittichii RW1
- Lipoprotein signal peptidase (*lspA*)
- UniProt Id: A5VCW2
- PFAM Id:PF01252

(*These proteins are known as a lipoprotein signal peptidases*).

Alignment

```
WLLMAFTAL-ALEMVALW
WLLVAGTALIAAGIVA-W
*** * *** *    ** *
```

Alignment Score
0.65

Overlap with a TM region
100%

Validation Step
100 Scrambled Sequences on WLLMAFTALALEMVALW
Mean of the scores for the best matches
0.16
P-value
0

Scrambled Sequence	***#Matches***	***Max Score***
AAMLETWWALALFMLLV	2	0.35
MLALTLWLFEAVMAALW	0	0.1
MLALLVALTMEWAWALF	0	0.1
AWALLMFLTMVWLEAAL	0	0.1
TAFLAMAELMWALVWLL	0	0.35
LAALVWLAWFEMTMLAL	0	0.1
WLWALEALMTFLAAVML	1	0.29
AWALLMFLTMVWLEAAL	0	0.1
…	…	…
MAMVLELAAWLWATLLF	0	0.1

Fig. 15 – Validation process example. In this example, the validation process identifies a statistically significant similarity between two transmembrane regions of different bacterial membrane proteins.

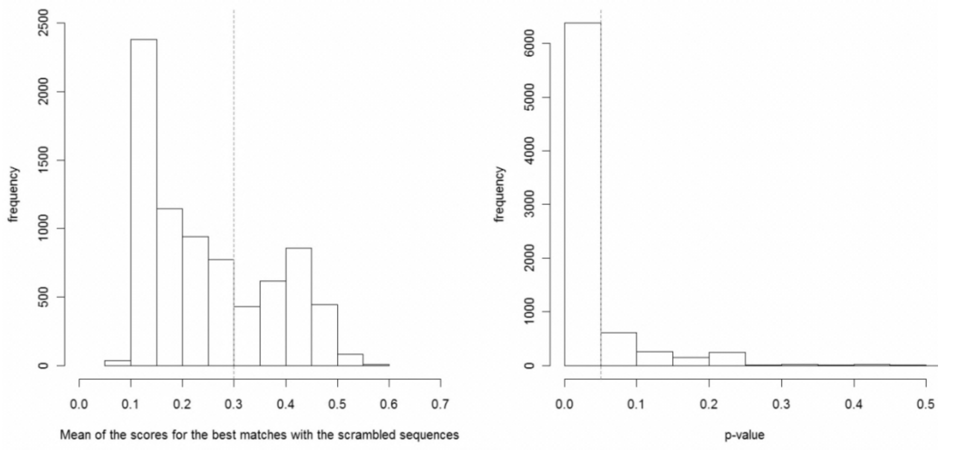


Fig. 16 – (Left) Histogram of observed similarity scores of randomly scrambled sequences. (Right) Histogram of corresponding p-values.

Moreover, it is not necessary to apply a procedure for multiple testing corrections. This procedure, usually, is used to prevent the problem of multiple comparisons when one tests *n* dependent or independent hypotheses on the same set of data. On the other hand, each statistical hypothesis was tested on a different set of data: all bacterial TM domains that have matching TM sequence. Therefore, the application of a multiple test correction was useless and self-defeating.

An important result obtained with my statistical framework is shown in Table 1. The table provides the number of significant matches found among the TM domains of *Salmonella* and TM domains of other bacterial genomes. It is possible to notice that significant homologous TM sequences are found in both gram-negative and gram-positive bacteria. This implies that bacterial genomes share significant similarities in their TM domains and that these similarities exist regardless of their class.

Species	#matches	Type
Bacillus	644	Gram-positive
Escherichia	472	Gram-negative
Staphylococcus	445	Gram-positive
Pseudomonas	286	Gram-negative
Mycobacterium	254	Gram-positive
Mycoplasma	242	Gram-positive
Shigella	213	Gram-negative
Yersinia	203	Gram-negative
Haemophilus	156	Gram-negative
Edwardsiella	151	Gram-negative
Methanocaldococcus	142	Gram-negative
Lactobacillus	136	Gram-positive
Rhizobium	115	Gram-negative
Klebsiella	112	Gram-negative
Buchnera	101	Gram-negative
Streptomyces	90	Gram-positive
Campylobacter	82	Gram-negative
Archaeoglobus	70	Gram-positive

Tab. 1 - List of bacteria sharing a high number of significant TM homologies with the TM domains of *Salmonella* spp.

As a consequence, we can hypothesize that these homologous TM domains emerged before the appearance of true bacteria, and that they may represent important signatures about the existence of the first trans-membrane-peptides that contributed to the formation of the earliest primitive life forms and organisms. All analytical results have been stored in a XML file and a corresponding excel file.

In Table 2 it is possible to observe the typical information reported in the excel file containing all the aligned TMFs. For each significant match that the framework found with a TMF of *Salmonella* spp., the table reports (i) the starting TMF query, (ii) general information about the TMP having the matching TM sequence (the accession number, the genome name and a synthetic description of this TMP), (iii) the alignment score and the corresponding p-value. Furthermore, XML file records also those matches with an irrelevant p-value, to allow other searchers to be applied with different statistical tests for the validation of high-scoring alignments that the framework found in different bacteria and distinct TMPs.

TMP	**Description**				
P63993	OS=Salmonella typhimurium RN1=Disulfide bond formation protein B				
TMFs	**TMPs**	**Organism**	**Description**	**Score**	**P-val**
WLLMAF TALALEM VALW	P32015	*Neisseria*	Capsule polysaccharide export inner-membrane protein CtrC. OS=Neisseria meningitidis serogroup BGN=ctr PE=3 SV=2	0,59	0.02
	A9BAR6	*Prochlorococcus*	Cytochrome c biogenesis protein CcsA. OS=Prochlorococcus marinus (strain MIT 9211) GN=ccsA PE=3 SV=1	0,59	~0
	A5VCW2	*Sphingomonas*	Lipoprotein signal peptidase. OS=Sphingomonas wittichii (strain RW1 / DSM 6014 / JCM 10273) GN=lspA PE=3 SV=1	0,65	~0
ALFGVM GAGLVG AIAP	Q9V2N1	*Pyrococcus*	Cobalamin synthase. OS=Pyrococcus abyssi GN=cobS PE=3 SV=1	0,56	0.02
	B9L178	*Thermomicrobium*	NADH - quinone oxidoreductase subunit N. OS= Thermomicrobium roseum (strain ATCC 27502 / DSM 5159 / P-2) GN= nuoN PE=3 SV=1	0,56	0.01

C0R5U2	*Wolbachia*	ATP synthase subunit c. OS=Wolbachia sp. subsp. Drosophila simulans (strain wRi) GN=atpE PE=3 SV=1	0,63	0.02
A9BZN0	*Delftia*	Probable intracellular septation protein. OS=Delftia acidovorans (strain DSM 14801 / SPH-1) GN=ispZ PE=3 SV=1	0,63	0.01

Tab. 2 - The best results achieved with the first two TM domains of the membrane protein DsbB of *Salmonella typhimurium*.

3.4 A theoretical approach to infer domain-centric homologies

The idea behind the proposed method can be understood in a more general context. Sometimes we need to find conserved short sequences or conserved amino acid positions in specific protein domains and/or under some specified constraints.

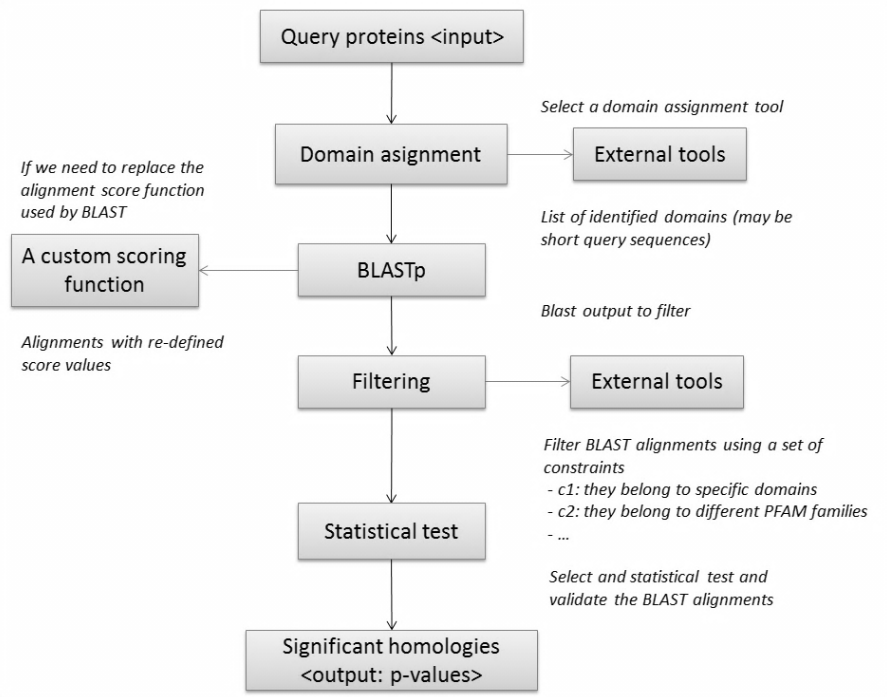

Query proteins <input>
Select a domain assignment tool
Domain asignment
External tools
If we need to replace the alignment score function used by BLAST
List of identified domains (may be short query sequences)
A custom scoring function
BLASTp
Blast output to filter
Alignments with re-defined score values
Filtering
External tools
Filter BLAST alignments using a set of constraints
- c1: they belong to specific domains
- c2: they belong to different PFAM families
- …
Statistical test
Select and statistical test and validate the BLAST alignments
Significant homologies <output: p-values>

Fig. 17 – A general similarity search scheme to identify distant homologies in short sequence domains.

This is a type of homology analysis uses a domain-centric search strategy, in which the aim is the identification of homologies in specific target-domains. Such homology searching can be obtained exploring distantly related homologs (Hollich *et al.*, 2007) or to identify homologs in insignificant blast hits.

Figure 17 shows a general similarity-based approach to search for significant and distant homologies under some specified constraints, such as domain-specific constraints or label information (a Pfam family ID). This schema is based on BLAST, the most versatile and popular algorithm to detect sequence similarity. Blast algorithm, however, can result in many insignificant (by e-value) hits that are nevertheless homologs. This occurs when, for instance, short query sequences are used. Increased sensitivity can be obtained by setting accurately the BLAST parameters or using more sophisticated methods, such as algorithms like PSI-BLAST (Altschul *et al.*, 1997) and hidden Markov models (HMMs).

However, these methods need to include some basic information to improve the identification of homologous proteins compared to BLAST. The scheme reported in the Figure 17 represents an alternative pipeline to implement more complicated similarity searching methods. It should be used for the identification of domain-centric homologs in insignificant blast hits by exploiting different external sources of information. The pipeline that we proposed can be adapted for homology searches with different domain-specific targets.

The proposed pipeline (i) assigns domains to a given query protein and extracts the corresponding list of short fragments using annotations or prediction tools. Subsequently, (ii) it runs BLAST and filters its output to determine a first set of putative homologous domains without considering if they belong to significant blast hits. The filtering step is necessary to remove matches that are biologically uninteresting or that are not meaningful for some specified constraints and/or domain-specific targets. Finally, the alignments are statistically validated considering the probability that the alignment occurred by chance with specific protein domains and/or under some specified constraints. For example, if we measure the statistical significance of alignments in TM domains, we must calculate the probability that this match has occurred by chance considering only TM matches.

An important aspect of this pipeline is the possibility to improve the detection of homology by using information not derived from the primary sequence of proteins, such as alignment information (e.g., identities, positives, etc). For instance, we can use external tools to have specific information of the aligned sequences (e.g., TM domains, protein family information, etc.) and use them to assess the statistical significance of alignments that are statistically irrelevant with BLAST.

3.6 Conclusion

I developed a new statistical BLAST-based framework to perform sequence similarity searching with short amino acid queries representing TM domains. The proposed program is able to identify ancient homologies in TM domains of proteins of different bacteria and distinct TMPs.

I used this framework to screen the TMPs of *Salmonella* spp., and constructed a XML data base containing 11,771 trans-membrane fragments derived from 1,760 TMPs. This framework identified 5,216 significant homologies between TMFs of different bacteria. These homology sequences were found in both gram-negative and gram-positive bacteria. This implies that extant bacterial genomes share significant similarities in their TM domains and that these similarities exist regardless of their classes. Therefore, my approach provides an insight on the theory of minigenes because the high-scoring alignments that the framework identified are evolutionary distant and are characterized by short and nearly exact matches.

All results have been stored in a XML file that can be parsed/transformed for further analysis or validation steps, or can be converted into web pages.

Finally, I defined a theoretical framework to perform homology searches domain-oriented. This framework can represent an optimal guideline for effective searching statistically significant homologies in specific target-domains.

CHAPTER 4

RNA-Seq: a new frontier in whole-transcriptome analysis

The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition.

Transcriptome studies are essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and for understanding RNA-based regulatory mechanisms.

Several technologies have been developed to study the transcriptome, but, recently a new advanced technique called RNA Sequencing (RNA-Seq) uses massively parallel sequencing to allow transcriptome analyses of genomes at a high resolution. This technology aims to the sequencing of the entire transcriptome in order to have transcript discovery and sequence-based quantitative analysis of gene expression levels.

Currently, the most used sequencing platforms for RNA-Seq are: Illumina (with paired-end reads up to 100 bp), SOLiD (up to 50-35 bp) and Roche 454 (up to 400bp). Apart from Roche 454, the other sequencing technologies give very short reads compared to the average length of transcripts (~2000 bp). This provides computational challenges for analyzing and interpreting the data.

RNA-Seq methods can generate thousands of millions of short sequences or reads that need to be aligned to a reference genome or assemblied in transcripts. Then, several programs can be used for data handling and visualisation, transcript expression inference, differential expression analysis, data quality assessment, integrative bioinformatic and statistical analysis.

However, the popularity of transcriptome sequencing RNA-Seq based projects demands new data mining tools and sophisticated programs with a comprehensive set of features to support a wide range of new analysis tasks.

For instance, once the reads have been aligned, we can carry out complex tasks to elicit fascinating information about the transcriptome dynamics including all the species of transcript units, non-coding RNAs, small RNAs and transcriptional gene~~s~~ structures. Moreover, studies based on RNA-Seq can resolve interesting questions about biological processes in bacterial cells; they can also contribute to genomics, allowing better quality annotation.

This chapter presents a brief description of the RNA-Seq techonology and its uses to improve the understanding of the regulatory mechanisms that control the transcriptome. It provides basic information about the RNA-Seq method and its experimental process. There is also a brief introduction to the main bioinformatics challenges for the analysis of RNA-Seq results and the commonly used algorithms to align reads. Finally, some RNA-Seq studies are presented to show their contributes to the discovery of new knowledge about regulatory mechanisms in eukaryotic and prokaryotic cells.

4.1 RNA-Seq: introduction

RNA-Seq refers to the method of using Next-Generation Sequencing (NGS)[7] technology to measure RNA levels; this methodology was developed and initially utilized for identifying the transcriptional map of yeasts (Nagalakshmi *et al.*, 2008).

RNA-Seq technology enables experimenters to capture the RNA expression profile of a cell in un-precedented detail. The process starts by shearing up the RNA (actually cDNA) into small fragments, typically around 30-400 bp in length depending on the DNA sequencing technology used. A short sequence or "read" is then determined by a NSG method from one end (single-end reads) or both ends (paired-end reads) of each of these fragments.

Today, there are three widely accepted commercially available NGS devices for RNA-Seq: 454 GS FLX (Roche), Genome Analyzer II (Illumina) and SOLiD (Applied Biosystems). These systems have already been applied for RNA-Seq studies (Foster City, Cloonan *et al.*, 2008; Eveland *et al.*, 2008; Marioni *et al.*, 2008; Nagalakshmi *et al.*, 2008).

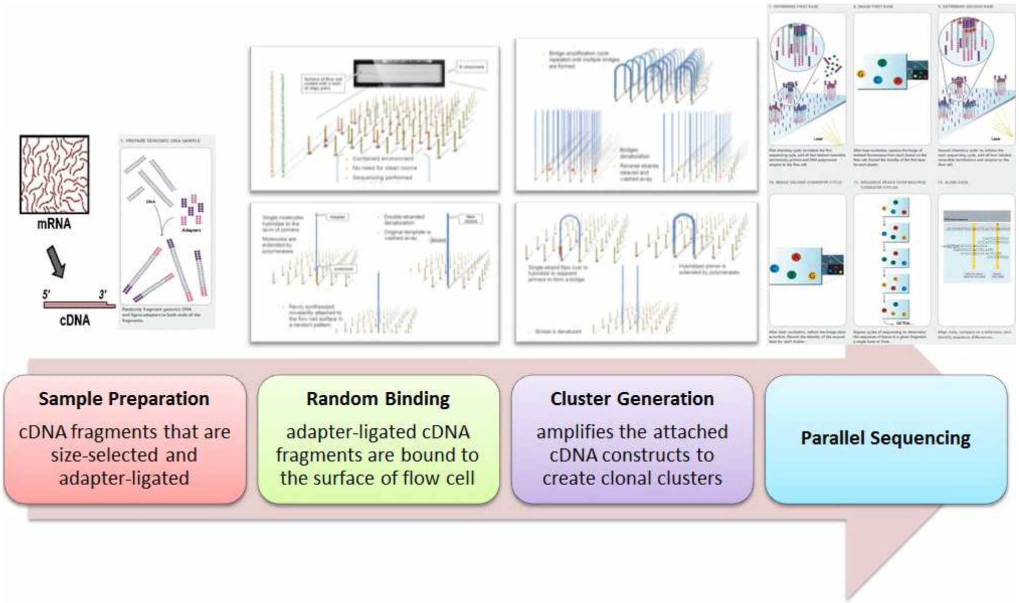

mRNA
5'
3'
cDNA
Sample Preparation
cDNA fragments that are size-selected and adapter-ligated
Random Binding
adapter-ligated cDNA fragments are bound to the surface of flow cell
Cluster Generation
amplifies the attached cDNA constructs to create clonal clusters
Parallel Sequencing

Fig. 18 – Example of RNA-Seq protocol. The basic RNA-Seq protocol of Illumina consists of the following steps (i) polyadenylated RNAs in the biological sample are extracted, (ii) these RNAs are converted into more stable cDNA molecules which are randomly sheared, (iii) a size selection on the sheared fragments is done for optimization of later steps or paired-end sequencing, (iv) the fragments are amplified and adapters are ligated to the fragments, and finally (v) sequencing of the fragments is carried out using an NGS approach.

[7] NGS technology is an ultra-high-throughput technology to measure DNA sequences.

Each system works differently, but they are all based on similar principals: shear target DNA into small pieces, bind individual DNA molecules to a solid surface, amplify each molecule into a cluster, copy one base at a time and detect different signals for A, C, T, and G bases. Figure 18 reports the main steps involved in an Illumina RNA-Seq analysis.

After sequencing, the resulting short reads are either aligned to a reference genome or reference transcripts, or assembled *de novo* without the genomic sequence to produce a genome-scale transcription map that consists of both the transcriptional structure and/or level of expression for each gene.

RNA-Seq offers several advantages compared with other transcriptomics methods (Wang Z., *et al.*, 2009). They provide high-throughput solutions for the construction of single-base resolution expression profiles with a low background noise and a low required amount of RNA. Besides, RNA-Seq is not limited to detecting transcripts that correspond to existing genomic sequences. They can be very useful for discovering new transcripts, identifying mutations, deletions and insertions, and alternatives splicing; it provides excellent coverage and can generate millions of reads in a single run.

4.1.1 Bioinformatics challenges

The main bioinformatics challenges in RNA-Seq data analysis can be divided into four categories: (i) store/treat a huge amount of data, (ii) align the short reads, (iii) re-build the transcriptome and (iiii) quantify the expression level for each transcribed region. The NSG methods used for RNA-Seq experiments (including Roche/454, Illumina and SOLiD^{TM}) are able to produce data of the order of giga base-pairs (Gbp) per machine day (Metzker *et al.*, 2010). This implies that several computational problems must be faced, including the implementation of proficient methods to store, retrieve and process large amounts of data.

Usually, the mentioned RNA-Seq machines output the results in “fastq" format. The fastq format stores four lines of information for each read: (i) an identifier beginning with an "@" character, (ii) the sequence of bases called by the machine (an N means a call could not be made), (iii) a separator line containing only the "+" character and (iv) a line containing the corresponding quality scores for each base. The quality score (Q score) is widely used to measure "reliability" of base-calling, it can differ depending on the sequencer and even the version number of the base calling software.

```
$ head -n 8 ../../Sample_reads.fastq

@SRR037945.1 HWUSI-EAS627_1:2:1:0:1629
NNNANNNNNNNATCTCTTTAGATTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGAAAGAAGAAA
+
!!!#!!!!!!!#################################################################
@SRR037945.2 HWUSI-EAS627_1:2:1:0:681
NNNANNNNNNNCAGAAGAGGGCATCAGATCTCATTACAGATGGTTTTTAGCCACCATTTTGTTTCTTGGGTTTTAA
+
!!!#!!!!!!!#################################################################
```

Fig. 19 – Example of information reported in a fastq file containing single reads.

In general: *Q = -10*log10(p)*, where *p* is the probability that the corresponding base call is incorrect. For example, quality score of 20 means there is 1/100 chance that the base-calling is wrong, quality score of 30 means there is 1/1000 chance that the base-calling is wrong. In order to keep the size of the fastq files down, this Q score is rounded to the nearest integer and then converted to the ascii character corresponding to that integer.

The Q score is defined for each character and the corresponding ascii characters are reported in the fourth line. Therefore, the fastq files are often very large because they should contain information about hundreds of millions of reads, and so, good strategies necessity to be implemented to manipulate simultaneously all those reads.

Once high-quality reads have been obtained, the first task to accomplish is to align the short reads from RNA-Seq to the reference genome, or to assemble them into contigs before aligning them to the genomic sequence to reveal transcription structures. There are several programs for mapping reads to the genome, including SOAP (Li R., *et al.*, 2008), MAQ (Li H., *et al.*, 2008) and BOWTIE (Langmead *et al.*, 2009). For large transcriptomes, alignment can be very complicated by the fact that a significant portion of sequence reads match multiple locations in the genome. Another problem is due to the sequencing errors and polymorphisms that can negatively influence the mapping. Generally, single base differences are not problematic, because most mapping algorithms accommodate one or two base differences. However, resolving larger differences will require a better reference genome annotation for polymorphisms and a deeper sequencing coverage. Obtaining longer sequence reads should help in alleviating the multi-matching problem.

After the alignment, a single-nucleotide transcription profile is obtained, and it can be used to elucidate different types of transcriptomic features. For instance, we can define computational models to identify microRNAs, non-coding RNAs, small RNAs or determine transcription units that are transcribed in monocistronic or polycistronic mRNAs. A standard format to represent the single-nucleotide transcription profile comes from a RNA-Seq experiment is the “pile-up” format, a file showing local coverage, mismatches and consensus calls and indels. The typical information stored in a pile-up file is shown in Figure 20.

```
I 25514 G G 42 0  25 5  ....^:.                    CCCCC
I 25515 T T 42 0  25 5  .....                      CC?CC
I 25516 A G 48 48 25 7  GGGGG^:G^:g                CCCCCC5
I 25517 G G 51 0  25 8  ......,^:,                 CCCCCC1?
I 25518 T T 60 0  25 11 ......,,^:.^:,^:,          CCCCCC3A<:;
I 25519 T T 60 0  25 11 ......,,.,,                CCCCCC>A@AA
I 25520 G G 60 0  25 11 ......,,.,,                CCCACC>A@<A
I 25521 T T 60 0  25 11 ......,,.,,                CCCCCC?ACAA
I 25522 A A 60 0  25 11 ......,,.,,                CCCCCC>ACAA
I 25523 A A 72 0  25 15 ......,,.,,^:.^:,^:,^:.    CCCCCC;ACAAC??C
I 25524 C C 72 0  25 15 ......,,.,,.,,.            CCCCCC6<<A?C=9C
I 25525 C C 56 0  24 18 ......,,.,,.,,.^:,^!.^:T   CCCCCC>ACA?C=AC<CC
```

Fig. 20 – Example of information found in a pile-up file.

The pile-up files can be used for the next gene expression analysis. But, before this step, the pile-up files need to be converted into a table of counts. The aim is to count the number of reads overlapping some annotation object: single base, coding-sequences, well-known small RNAs, intergenic regions, operon structures. After obtaining the table of counts (Table 3), a test can be performed to look for statistically significant differences in expression level.

Therefore, RNA-Seq technologies can be used to quantify gene expression by measuring the number of reads counting the number of reads overlapping a gene (Mortazavi *et al.*, 2008; Nagalakshmi *et al.*, 2008). The challenge, however, is that read-counts need to be normalized to extract meaningful expression estimates. Usually, two are the main sources of systematic variability that require normalization.

Gene	Sample A		Sample B		… etc.
	Rep 1	Rep 2	Rep 1	Rep 2	
ENSG00000209432	4	6	35	45	
ENSG00000209432	0	0	2	1	
ENSG00000209432	110	96	177	203	
ENSG00000209432	12685	10897	9246	9873	
ENSG00000212678	148	201	112	93	
…etc.					

Tab. 3 – Example of table of counts for annotated genes.

First, RNA fragmentation during library construction causes longer transcripts to generate more reads compared to shorter transcripts present at the same abundance in the sample (Marioni *et al.*, 2008). Second, the variability in the number of reads produced for each run causes fluctuations in the number of fragments aligned across samples. In order to resolve these problems we can use the RPKM (Reads Per Kilobase of transcript per Million mapped reads) normalization method that normalizes read counts by both the length of reads and the total number of mapped reads in the sample.

After expression quantification and normalization, an important question to address is to understand how these expression levels differ across conditions and how to identify genes that are statistically and differentially expressed among different conditions. The power to detect differential expression depends on the sequencing depth of the sample, the expression level of the gene, and even the length of the gene. To accommodate the count-based nature of RNA-Seq data, initial methods modeled the observed reads using count-based distributions such as the Poisson distribution (Trapnell *et al.*, 2009).

However, several studies have reported that these distributions, exemplified by Poisson, do not account for biological variability across samples (Langmead *et al.* 2010; Robinson *et al.*, 2007). Ideally, if we had enough replicates the variability across replicates could be empirically estimated using a permutation-derived approach, similar to that used by the Myrna method (Langmead *et al.* 2010). However, few RNA-Seq expression studies have generated a sufficient number of replicates to achieve this goal. To overcome this problem, many methods-attempt to model biological variability and to provide a measure of significance in the absence of a large number of biological replicates. These methods, such as EdgeR (Robinson *et al.*, 2010), DESeq (Anders and Huber *et al.*, 2010), and DEGseq (Wang K., *et al.*, 2010) model the count variance across replicates as a nonlinear function of the mean counts and use various different parametric approaches (such as the Normal and Negative Binomial distributions).

4.1.2 Coverage, sequencing depth and costs

Number of expressed genes (coverage) and costs increase with sequence depth (Wang K., *et al.*, 2010). This implies an important question on sequence coverage. If we want to know more and more, a greater coverage is necessary and this leads to require more sequencing depth. In addition, with larger genome and, in general, with organisms having a complex transcriptome, the request of sequencing depth increases. RNA-Seq can reveal transcript structure and splicing and can even identify novel isoforms, gene fusions and allele-specific variants. Therefore, the ability to observe any object involved in the RNA-based regulatory mechanismsin is dependent upon the coverage, and so the sequence depth.

4.2 Read alignment algorithms

The generation of reliable RNA-Seq data relies heavily on proper alignment of sequencing reads to corresponding reference genomes or on their efficient de novo assembly. A wide variety of alignment algorithms and software have been developed an these can be grouped in two categories. The first category is represented by aligners that map reads to reference without allowing any large gaps between exons, this are also called unspliced read aligners. This category of aligners fall into two main sub-categories: Seed methods and Burrows-Wheeler (BW) transform methods.

Unspliced Aligners	Reference	Methods
SeqMap	Jiang and Wong *et al.*, 2008	Seed
MAQ	Li, H. *et al.*, 2008	Seed
SOAP	Li, R. *et al.*, 2008	Seed
RMAP	Smith, A.D. *et al.*, 2008	Seed
BFAST	Homer *et al.*, 2009	Seed
BOWTIE	Langmead *et al.*, 2009	BW transform
BWA	Li and Durbin *et al.*, 2009	BW transform
SHRiMP	Rumble *et al.*, 2009	Seed
GASSST	Rizk and Lavenier *et al.*, 2010	Seed
Stampy	Lunter and Goodson *et al.*, 2011	Seed

Tab. 4 – Main unspliced read aligners.

Seed methods such us MAQ (Li H., *et al.*, 2008) and Stampy (Lunter and Goodson *et al.*, 2011) find matches for short sub-sequences, termed “seeds”, supposing that at least one seed within a read will be perfectly aligned to the reference genome. These sub-sequences within a read are then used to initiate the alignment process. Finally, the alignments for each read are computed and ranked and the best scoring mapping is then reported.

On the other hand, Burrows-Wheeler transform methods like BWA (Li and Durbin *et al.*, 2009) and Bowtie (Langmead *et al.*, 2009), compact the genome into a search-efficient data structure, called index, which allows for faster alignment, as long as the number of mismatches is small.

These type of aligners are not optimal for direct alignment of RNA-Seq reads to the genome since many reads that originate from exon-exon junctions map discontinuously to the genome. However, short read aligners are ideal for mapping reads against a known cDNA databases for quantification purposes (Mortazavi *et al.*, 2008; Berger *et al.*, 2010). Usually, BW methods are more commonly used because they are significantly faster than seed based methods and because they require less memory. Besides, the BW methods are more straightforward to program than BW methods that seem more complicated.

The unspliced read aligners are inadequate to identifying known exons and junctions, and, generally, they do not allow the identification of splicing events involving novel exons.

Therefore, a second category of aligners has been defined: spliced aligners. These aligner can align the entire genome, including intron-spanning reads that require large gaps for proper placement. They can be split in two main sub-categories: exon-first and seed-and-extend.

Spliced Aligners	Referefnce	Methods
QPALMA	De Bona *et al.*, 2008	Seed-extend
TopHat	Trapnell *et al.*, 2009	Exon-first
MapSlice	Wang K. *et al.*, 2010	Exon-first
SpliceMap	Au *et al.*, 2010	Exon-first
GSNAP	Wu and Nacu *et al.*, 2010	Seed-extend

Tab. 5 – Main spliced read aligners.

Exon-first methods like MapSplice (Wang *et al.*, 2010), SpliceMap (Au *et al.*, 2010) and TopHat (Trapnell *et al.*,2009) are based on a two-step process. First, they map reads to the genome, forbidding large gaps. Second, unmapped reads are split into shorter segments and aligned independently. The genomic regions surrounding the mapped read segments are then searched for possible spliced connections. Exon-first aligners are very efficient when only a small portion of the reads require the more computationally intensive second step. Alternatively, seed-extend methods like GSNAP (Wu and Nacu *et al.*, 2010) and QPALMA (De Bona *et al.*, 2008) break reads into short seeds, which are placed onto the genome to localize the alignment. Candidate regions are then examined with more sensitive methods, such as the Smith-Waterman algorithm, or iterative extension and merging of initial seeds (Kent *et al.*, 2002) to determine the exact spliced alignment for the read.

Exon-first approaches are faster and require fewer computational resources compared to seed-extend methods. Exon-first approaches can miss spliced alignments for reads that also map to the genome contiguously, as can occur for genes that have retrotransposed pseudogenes. In contrast, seed-extend methods can produce a good alignment for both gapped and ungapped locations, yielding the best placement of each read. Seed-extend methods are better than exon-first approaches when mapping reads from polymorphic species (Mikkelsen *et al.*, 2007) and, in addition, most of them allow also paired-end read mapping. Paired-end read mapping can increase the specificity of the alignment and the corresponding alignments can be used for both reconstruction and quantification.

All software for the alignment process output the results in the SAM format (Li and Handsaker *et al.*, 2009), the emerging standard alignment format which is widely supported by alignment viewers such as BamView (Carver *et al.*, 2010). The SAMtools[8] , online available, offer the possibility to convert between SAM and BAM format (a binary version of a SAM file, suitable for fast processing), a text-file human-readable, sort and merge different SAM files, index SAM and FASTA files for fast access, produce pile-up files.

4.3 Study of RNA-based regulatory mechanisms

RNA-Seq methods generate an unprecedented global view of the transcriptome and its organization giving the possibility to realize interesting studies of RNA-based regulatory mechanisms. In eukaryotes, RNA-Seq has been used for the discovery of splice variants, RNA editing sites, and new microRNAs (Glazov *et al.*, 2008; Sultan *et al.*, 2008; Li *et al.*, 2009). The identification of exon boundaries with RNA-Seq data is a task of special interest, since it can lead to discovery new splicing isoform of known genes (David *et al.*, 2006).

Besides, other RNA-Seq based studies have proved the existence of a large number of new transcribed regions in different genomes, including the *A. thaliana* (Lister *et al.*, 2008), mouse (Cloonan *et al.*, 2008), human (Morin *et al.*, 2008), *S. cerevisiae* (Nagalakshmi *et al.*, 2008) and *S. pombe* (Wilhelm *et al.*, 2008). These novel transcribed regions, combined with many undiscovered novel splicing variants, suggest that there is considerably more transcript complexity than previously appreciated.

In addition, RNA-Seq methods allow the identification of transcript boundaries that can be used to correct the exist annotations (Yoder and Himes *et al.*, 2009) and reveal several novel features of eukaryotic gene organization. For instance, Nagalakshmi *et al.* (2008) found several yeast genes that were overlapping at their 3′ ends, while Cloonan *et al.* (2008) found that antisense expression is enriched in the 3′ exons of mouse transcripts. For multicellular organisms, antisense transcription could modulate gene expression through the production of siRNAs or through dsRNa editing. For yeast, which seems to lack siRNA and dsRNA-editing functions, the transcription from one gene might interfere with that from an overlapping gene, or coordinate gene expression through other mechanisms.

RNA-Seq studies have also been used to refine our understanding of bacterial gene expression. For instance, a first insight was that the expression level of coding-sequences is constantly distributed without having a background transcription level (Passalacqua *et al.*, 2009). Instead, in a case study based on marine metatranscriptome it has been found that gene sequences, highly representing in cDNA samples, are often occasional from the corresponding genomic DNA samples, suggesting some bacteria may be transcribing a set of uncharacterized genes at an unusually high level (Gilbert J.A., *et al.*, 2008).

Furthermore, using RNA-Seq data the annotation of bacterial coding sequences has been improved and novel transcribed regions have been identified in most studies (Filiatrault *et al.*, 2010; Perkins *et al.*, 2009), including that of *M. pneumoniae* which has a genome just 816 kb in size. In addition, existing gene models have been remodeled, often correcting the choice of the start codon, and associated with one another into operons, which can include the identification of untranslated regions.

[8] http://samtools.sourceforge.net/.

However, in both *M.pneumoniae* and *H.pylori*, the annotation of transcriptional units was complicated by an unexpectedly high level of flexibility in the structure of operons (Güell *et al.*, 2009). Evidence from both tiling microarray and RNA-Seq data indicated different promoters appeared to be driving expression of the same genes under different conditions, leading to an alternative transcript system that works similar to the well-known alternative splicing system of some eukaryotes.

Bacterial whole transcriptome studies based on RNA-Seq experiment have led to a high success rate of non-coding RNA discovery. Such transcripts have even been identified and mapped to genomes from marine metatranscriptome data, where certain putative ncRNA showed distinct spatial distributions throughout the water column (Shi *et al.*, 2009). In *H. pylori*, both in silico analysis and mutational inactivation suggested that one novel ncRNA uncovered by RNA-Seq regulated a chemotaxis receptor as an antisense RNA (Sharma *et al.*, 2010), and a similar mechanism was posited for a novel ncRNA in *V. cholerae*, which was found to down regulate mannitol metabolism (Liu *et al.*, 2009).

4.3.1 Transcriptome analysis of bacteria using RNA-Seq

Recent studies have employed RNA-Seq for bacterial transcriptome research and demonstrated its effectiveness in accurate operon definition, discovery of non-coding RNAs, and correction of gene annotation (Perkins *et al.*, 2009).

Though the pioneering study was done with eukaryotic organisms, because their mRNAs with poly-A tails are easier to isolate, RNA sequencing technology has also been applied to prokaryotes (van Vliet *et al.*, 2010), as shown in Table 6.

With RNA-Seq method all transcription is studied in an unbiased manner, because it provides direct access to the sequence without a reference genome or predesigned probes. This allows the discovery of novel genetic features, as well as the delineation of operons and untranslated regions, allowing the improvement and extension of sequence annotation.

Furthermore, RNA-Seq technology has been shown to be highly precise in the quantification of transcription levels, giving results similar to those provided by quantitative PCR (qPCR) (Wang Z., *et al.*, 2009). Until now, quantitative RT-PCR has been the reference and it is the most precise means to measure expression; according to Roberts *et al.* (2011), although it is not a perfect assay, it is the best option except for RNA-Seq. RNA sequencing data are highly reproducible, with few differences between technical replicates, according to the research carried out with data from Illumina, provided that it is sequenced from the same library (Marioni *et al.*, 2008); thus, it is necessary to sequence only once.

Studies based on RNA-Seq can help to improve the quality annotation of microbial genomes, refine our understanding about the regulatory systems in prokaryotes and address interesting questions on biological processes in bacterial cell. For instance, there are transcriptome RNA-Seq studies defining new genes, re-defining the structure of the annotated gene, determining the true start codon, detecting untranslated regions (UTRs), including riboswitches and binding sites of regulatory small RNAs (sRNAs), and identifying new operons (Sorek and Cossart *et al.*, 2010).

Despite the many advantages RNA-Seq offers, it is still a relatively new methodology with developments continuing for both experimental procedures and subsequent data analyses. For instance, new protocols for strand-specific RNA-Seq library preparation have been developed (Levin *et al.*, 2010). These protocols are used by the community for the detection of large amounts of cis-antisense ncRNA:

regions of CDSs that are bi-directionally transcribed, and suggested to act to block expression the encoded protein (Filiatrault *et al.*, 2010; Sharma *et al.*, 2010; Wurtzel *et al.*, 2010).

An interesting study based on directional RNA-Seq data was realized by Güell *et al.* (2009) using two technologies, tiling arrays and direct strand-specific sequencing (DSSS) by means of Illumina, for analyzing bacteria.

Using a combination of these techniques, the authors were able to observe the expression of all the genes. The analysis revealed the versatility of operons in response to different conditions (173 different conditions were tested); that is, one gene coded as polycistronic under one condition, can be transcribed as monocistronic in another.

Species	Phylum	Sequecing platform	Reference
Bacillus anthracis	Firmicutes	SOLiD$^{\mathrm{TM}}$	Passalaqua *et al.*, 2009
Burkholderia cenocepacia	Betaproteobacteria	Illumina	Yoder - Himes *et al.*, 2009
Listeria monocytogenes	Firmicutes	Illumina	Oliver *et al.*, 2009
Mycoplasma penumoniae	Firmicutes	Illumina	Güell *et al.*, 2099
Salmonella typhi	Gammaproteobacteria	Illumina	Perkins *et al.*, 2009
Acinetobacter baumannii	Gammaproteobacteria	Illumina	Camarena *et al.*, 2010
Chlamydia trachomatis	Verrucomicrobia	Roche FLX	Albrecht et al., 2010
Helicobacter pylori	Epsilonproteobacteria	Roche FLX	Sharma *et al.*, 2010
Pesudomonas syringae	Gammaproteobacteria	Illumina	Filiatrault *et al.*, 2010
Staphylococcus aureus	Firmicutes	Illumina	Beaume *et al.*, 2011
Neisseria gonorrhoeae	Betaproteobacteria	SOLiD$^{\mathrm{TM}}$	Isabela - Clark *et al.*, 2011
Streptococcus pneumoniae	Firmicutes	Illumina	Croucher *et al.*, 2011
Porphyromonas gingivalis	Bacteroidetes	Illumina	Hövik *et al.*, 2012
Haemophilus somnus	Gammaproteobacteria	Illumina	Kumar *et al.*, 2012

Tab. 6 – Representatives of the domain bacteria that have had their transcriptomes studied by RNA-Seq to date.

This versatility of the operon was observed in more than 40% of the transcriptions of *M. pneumoniae* and has already been documented in another transcriptome study in Archaea (Koide *et al.*, 2009). These reports reinforce the notion of the operons as non-static structures, which increase the regulatory capacity of bacterial transcriptomes, so that they are functionally analogous to alternative promoters or alternative splicing in eukaryote transcriptomes.

Even today, after many years of studies of bacteria, new discoveries continue to surprise us. Through RNA-Seq, it can be seen how the microbial transcriptome is more complex and dynamic than initially thought and how it approximates that of eukaryotes in various aspects. Perhaps, it is due to this transcriptomic versatility and complexity that bacteria are able to adapt to different environments with such agility.

4.4 Request of new computational tools in sequence analysis

NSG technologies are revolutionizing genomic/trascriptomic analysis by improving existing high throughput methods and making other sequence analyses feasible for the first time. In such context, RNA-Seq has started to change the way we analize and study the complexity and dynamics of transcriptomes and genome regulation. In a few years, this technology brought to light more extensively expressed genomes, more complex transcriptomes and unknown regulatory mechanisms.

The potential applications of the new sequencing technologies are unlimited, while the extraction of novel information from such data is severely limited by a lack of new computational methods. In sequence analysis, for instance, the existing tools need to evolve to meet new requirements, and new tools must be realized to enable new analysis tasks such as the identification of condition-dependent operon maps.

CHAPTER 5

Transcriptome dynamics-based operon prediction

The identification of genes that are grouped together into operons is a key step toward the reconstruction of complex regulatory networks.

However, the mechanisms of operon formation are poorly understood and experimental methods to identify genome-wide operon structures genome-wide are laborious (Walters *et al.*, 2001) and time consuming. Therefore, many computational approaches have been proposed for predicting operons based on inherent DNA sequence properties.

Current methods predict operons using a model trained on a set of experimentally-defined transcription units in prokaryotic organisms (Edwards *et al.*, 2005). Unfortunately, prokaryotic genomes with experimentally verified operon data sets exist for only a few model organisms. In fact, operon prediction methods show a high-prediction accuracy only because they focus on the prediction (and verification) of operons in *Escherichia coli* K12 and/or *Bacillus subtilis*, model organisms that are well-studied and have well defined operon maps. Therefore, these methods tend to generalize well only to the genomes of species closely related to these model organisms.

In addition, recent whole-transcriptome RNA-Seq analysis have identified new operon pairs that were not predicted by standard operon prediction algorithms and genes within predicted or experimentally verified operons that were not co-transcribed (Hövik *et al.*, 2012; Kumar *et al.*, 2012). In particular, an interesting discovery by Güell *et al.* (2009) was obtained on the genome of *Mycoplasma pneumoniae*. Güell, using tiling arrays and RNA-Seq by Illumina, observed changes in operon structures in response to different experimental conditions.

These findings indicate that bacterial operon structure can be more complex and dynamical than previously appreciated, and that transcriptomic data deriving from RNA-Seq experiments can help us to better understand these dynamics and improve the accuracy of operon predictions.

This chapter contains the description of a novel computational method to improve whole-genome prokaryotic operon map inference by using RNA-Seq data analysis. This method allows to refine predictions generated with standard sequence feature-based strategies and to produce condition-dependent operon maps. The proposed solution is evaluated on different RNA-Seq based transcriptome profiles of *Histophilus somni and Porphyromonas gingivalis*. The results show that classification algorithms, based on features dependent on transcriptome dynamics and DNA sequence feature, can accurately classify >96% of gene pairs in a set of operons collected by DOOR and verified with

5.1 Biological background and problem domain

Transcriptional regulation is perhaps the most fundamental control in gene expression. Many functionally related genes are often co-regulated, meaning that their expression is coordinated temporally or even spatially in response to the need of the organism in a given environmental condition.

In prokaryotes these co-regulated genes are often organized in their genomes into physical clusters called operons. An operon thus consists of more than one adjacent gene expressed as a transcription unit, often identified by the presence of promoters and terminators.

Operons allow an organism to simultaneously express the genes that are needed for cell survival under the same condition, providing a control circuit that is both simple and economical. It has been reported that genes transcribed in a single operon are functionally related and make up a part of a metabolic pathway. Therefore, understanding the operon organization of a genome is important also to better understand the functions of genes and the genome.

Due to the laborious experimental methods to determine operons on an individual basis, several computational approaches have been proposed for predicting them. The main methods to predict operons are reported in Table 6.

Most methods rely on features based on the genome structure or the functional similarity of genes of interest. Since adjacent genes in an operon often are physically closer to each other than those not in the same operon, intergenic distance provides information about the likelihood that two adjacent genes may be on the same operon (Salgado *et al.*, 2000). The conservation of gene order in multiple organisms and similarity of codon usage are also taken into account (Ermolaeva *et al.*, 2001; Price *et al.*, 2005). Additionally, information about co-regulation derived from microarray gene expression data is also used for operon predictions because identifying adjacent genes whose transcription levels are well correlated provides much information on the likelihood of their being in the same operon (Sabatti *et al.*, 2002).

Using these features, a number of computational methods have been developed including hidden Markov models, machine learning, simple statistical and Bayesian methods, neural networks, support vector machine and a few, as shown in Table 6.

Most of these methodologies show high prediction accuracy because they were mainly validated using information from *Escherichia coli* and *Bacillus subtilis* that are well-studied model organisms and have well defined operon maps. Therefore, they tend to generalize well only to the genomes of species closely related to these model organisms (Romero and Karp *et al.*, 2004). Among factors that could have affected their generalization ability there is the (possibly unintentional) use of genome-specific features, leading to performance decrease of these methods when applied to a new genome.

Though several studies have been carried out to combine different features in various ways for operon prediction, very little has been done to examine the contribution of transcriptomic features that can be extracted, for instance, from whole-transcriptome RNA-Seq data in order to improve the classification performance and/or to develop methods that identify operons in a prokaryotic genome without using information about other genomes.

Furthermore, recent application of whole-transcriptome RNA-Seq analysis to prokaryotic organisms reveals that the operon structure of bacteria is dynamic and condition dependent (Güell *et al.*, 2009; Hövik *et al.*, 2012; Kumar *et al.*, 2012). Indeed, these studies have shown the ability of operon structures to change in response to different environmental conditions.

Such findings indicate that bacterial operon structure can be more dynamic than previously appreciated, suggesting that transcriptomic data deriving from RNA-Seq experiments can help us to define condition-dependent operon map.

In this chapter, I present a novel method that combines standard DNA sequence features with transcriptome data of a prokaryotic organism, studied under a specific environmental or genetic perturbation, in order to train and validate a classification system that accurately infers changes in the operonic organization of this micro-organism.

5.2.1 The central dogma

A deoxyribonucleic acid (DNA) contains all the required information to build and maintain an organism. More precisely, an organism's DNA is a molecule that encodes all the ribonucleic acid (RNA) and the protein molecules that are needed to make its cells. All the cells of an organism, except blood and reproductive cells contain DNA. The entire DNA of an organism is called a genome. Prokaryotes are unicellular or multicellular organisms, such as bacteria, whose genomes are contained in a single double-stranded circular DNA molecule. Some prokaryotic organisms also have smaller DNA molecules called plasmids. Today, more than 300 bacterial genomes have been fully sequenced and these genomes can range in size from approximately 0.49 million base pairs (Mbps) in hyperthermophilic archaeal parasite *Nanoarchaeum equitans* Kin4-M (Waters *et al.*, 2003) to 9.12 Mbps in gram-positive bacterium *Streptomyces avermitilis* MA-4680 (Ikeda *et al.*, 2003).

All cells in an organism have exactly the same DNA and approximately the same DNA is also found in cells in different stages of development (Albert et al., 1994). However, different portions of the DNA are transcribed and translated under different conditions or in different cells of an organism. More precisely, when a cell needs new proteins a transcription process is activated. The DNA is copied (transcribed) into ribonucleic acid (RNA), a less stable nucleic acid that can be rapidly degraded. The segment of the DNA that is transcribed into RNA is called a gene. The RNA that codes for a protein is called messenger RNA (mRNA) and the DNA segment that provides that code is known as open reading frame (ORF). When read in the 5' to 3' direction, the region of the DNA before an ORF is called upstream, and the portion following an ORF is called downstream. Although about 90% of all genes in a prokaryotic organism are protein coding, only about 4% of cell's total RNA codes for proteins (Kohane *et al.*, 2003). Overall, RNA makes up a few percent of a cell's dry weight. Most of the RNA in cells is ribosomal RNA (rRNA) and transfer RNA (tRNA). They are important for the translation process. In addition, cells have many types of small RNAs whose function is under rigorous investigation in major laboratories today.

Figure 21 shows the central dogma of molecular biology in the form of a process in which the genetic information flows from DNA via RNA to proteins. In this process a gene produces its product and the product that carries out its function is called gene expression. Gene regulation refers to a mechanism that controls the synthesis of a particular gene product. Gene expression in prokaryotes is mainly regulated through transcription. At any given time, only a fraction of genes in an organism is expressed, and cells are capable of modulating the expression of their genes in response to external signals. Many of the genes are always expressed, while some others become active only when the cell needs their products. Even though gene expression is said to occur when gene products are needed, cells always maintain the minimum amount of RNA from every gene in the genome.

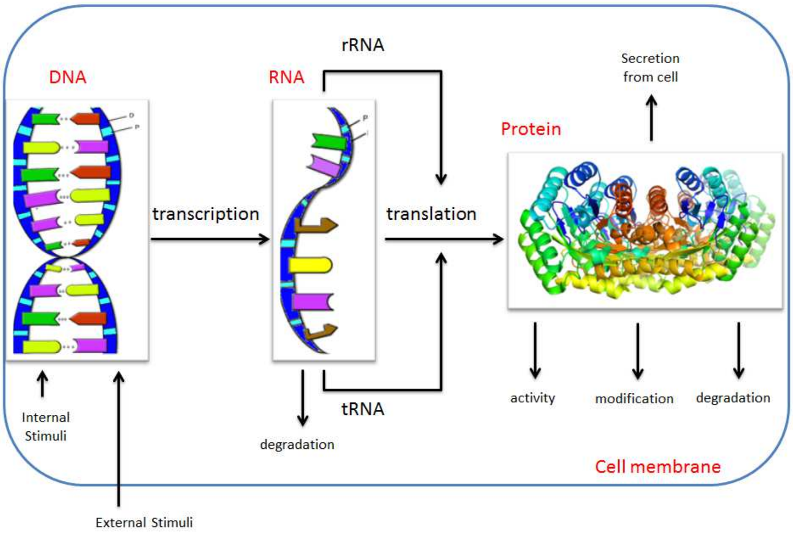


Fig. 21 – Central Dogma of Gene expression. The DNA segment of a gene is transcribed into mRNA. The mRNAs are translated into proteins, which have different roles within and outside of cells. Three types of RNA molecules, mRNA, rRNA and tRNA participate in the translation process. The translation products are folded, modified, and sent to their final destinations. In prokaryotes, mRNA is degraded within a few minutes after translation.

5.2.1 Operon definition

Experimental studies of the *E. coli* bacterium by F. Jacob and J. Monod in the 1950s revealed a special type of genes that are consecutive on the same transcriptional strand and co-expressed under the same condition (Jacob *et al.*, 1961). These genes are grouped into multi-gene clusters, called operons. Operon genes often have the same cellular function and their products form complex molecules (Salgado *et al.*, 2000). According to the standard definition for prokaryotic organisms, an operon is defined as a transcription unit (TU) consisting of a promoter followed by two or more genes and a transcription terminator. In Figure 22 a typical genomic structure of operon is reported.

The transcription process starts when RNA polymerase binds to a promoter before the first gene in an operon. The RNA polymerase then moves along the DNA using it as a template to produce an RNA molecule. When the RNA polymerase gets past the last gene in the operon, it encounters a special sequence called a terminator that signals it to release the DNA and finish transcription.

A promoter is a DNA sequence located upstream of a gene And it serves as a recognition site for the transcriptional machinery of the RNA polymerase complex.

The genes in an operon are usually co-transcribed to become part of a single primary transcript (mRNA molecule).

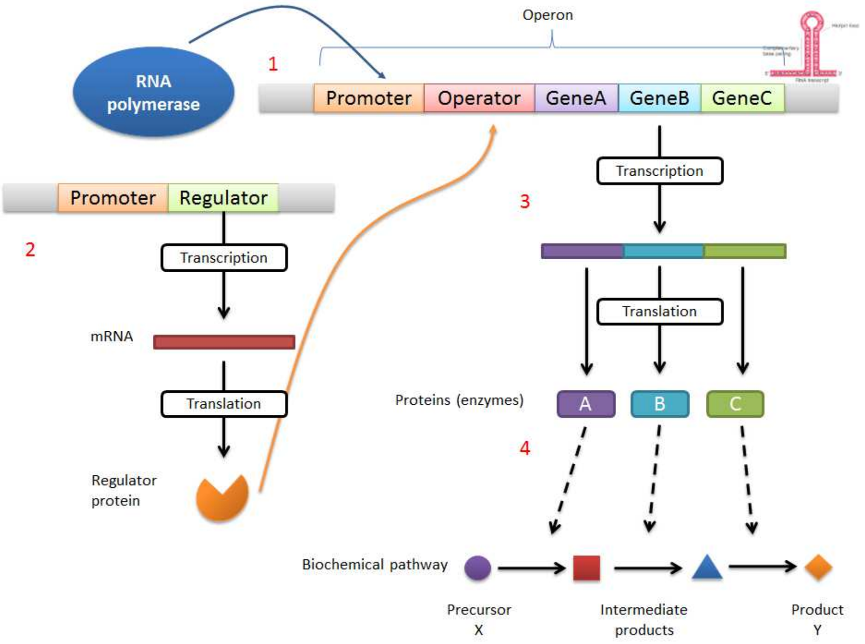


Fig. 22 – The definition of an operon. (1) An operon is a single transcriptional unit that includes a series of structural genes, a promoter, and an operator. (2) A separate regulator gene – with its own promoter – encodes a regulatori protein. (3) The regulator protein may bind to the operator site to regulate the transcription of mRNA. (4) The product of mRNA catalyze reactions in a biochemical pathway.

The mRNA of this type is called a polycistronic mRNA; mRNAs coding for a single protein are called monocistronic. The same ribosome translates all the proteins coded by the polycistronic mRNA. The actual quantity of each protein synthesized from a polycistronic mRNA can differ. These differences are partly due to the failure of ribosome to reinitiate with the mRNA when translating downstream genes. There are operons with several promoters, some of which are found inside the operon structures. These alternative promoters are used by RNA polymerases in certain conditions. Thus, sometimes all genes in an operon are transcribed and other times, only a subset. Several other regulatory sequences are involved in the mechanisms of operon formation. These include the operator and the terminator. The operator is a short region of a regulatory DNA, located between the promoter sequence and the structural genes, used for binding of a special protein, called regulator that can either repress or induce transcription of an operon. This regulator does not have to be adjacent to the genes in the operon.

Operons can be induced or repressed under different conditions or by different regulators. The terminator, on the other hand, indicates to the RNA polymerase the termination of the transcription process. As such, the promoter serves as a transcriptional start site, the terminator serves as a stop site, and the operator helps determine whether transcription will occur.

5.2.2 Operon formation

Operons are prevalent in all microbial genomes; about half of all protein-coding genes are in operon structures. The genes in an operon often (but not always) code for gene products in the same functional pathway; they are frequently conserved across species by vertical inheritance (Itoh *et al.*, 1999) and tend to be very compact.

The operon structure leads to several advantages in regulation. Genes that are placed in the same operon are strongly coordinated and functional related; transferring an entire operon allows an organism to acquire a complete, new capability. Moreover, there is just one promoter region for the activation or repression of a set of consecutive genes, this makes the system regulation efficient and faster.

Besides, operons tend to have more complex conserved regulatory sequences than individually transcribed genes (Price *et al.*, 2005). This should explain why certain operons, and in particularly new operons, contain genes with no apparent functional relationship (Price *et al.*, 2005; Regozin *et al.*, 2002)—the genes may be necessary to respond to a specific environmental condition despite they are involved in different pathways. For example, some conserved operons contain genes for ribosomal proteins and enzymes of central metabolism, perhaps because both are required in proportion to growth rates (Regozin *et al.*, 2002).

Two main theories have been advanced to elucidate the mechanisms of operon formation: the co-regulation and the selfish operon model.

The co-regulation model of operon formation claims that the co-regulation of genes that are co-adaptive provides a selective advantage to species. This model provides an important advantage: all the genes in an operon are under the control of a single operator and, consequently, are active and repressed at the same time. For instance, it only makes sense to express all the genes that are needed for tryptophan formation at the same time, and this will provide a selective advantage to any species that can co-regulate co-adaptive genes. This is a very simple model that does a better job at explaining why operons are kept than how they are formed. Essentially, it relies on operon randomly through deletions and juxtaposition of genes. However, the co-regulation model has a number of substantial problems associated with it. Since the model provides no selective benefit for clustering until co-transcription, rare and precise chromosomal rearrangements would be required for every gene added to the operon (Lawrence and Roth *et al.*, 1996). In addition, the co-regulation model provides no real explanation for why genes can also be co-regulated without being in the same operon and why genes are often clustered but not in a single operon (Ferrandez *et al.*, 1998).

Another popular model is the selfish operon model developed by Lawrence *et al.* (2006). This model tries to explain why non-essential genes primarily form operons and how these genes become juxtaposed. According to this theory, a large amount of genes from one species is horizontally transferred to the next one in a single horizontal transfer step, and therefore all of these newly acquired genes are positioned together in the receiving genome. Most of these genes do not bring any benefits to the species which inherit them except for several genes which gives the species a new ability. This theory is consistent with the compactness of operons. However, it has been proved that essential and other non–horizontally transferred (non-HGT) genes are likely to be in operons (Pal and Hurst *et al.*, 2004), and non-HGT genes are forming new operons at significant rates. Also, the selfish theory cannot explain the many operons that contain functionally unrelated genes. Thus, it seems that the selfish operon model may increase the prevalence of some operons, but it cannot explain any case of operon formation.

5.2.3 Operon identifcation problem

Operons are the basic functional units in a prokaryotic cell to make up the more complex functional units such as regulons, and pathways. Therefore, characterization of operon structures of a genome is crucial for understanding biological functions of the genome. However, the mechanisms of operon formation are poorly understood and experimental characterization of operons is expensive and time consuming.

Thus, developing computational methods to effectively predict operons has become a very important challenge in computational biology.

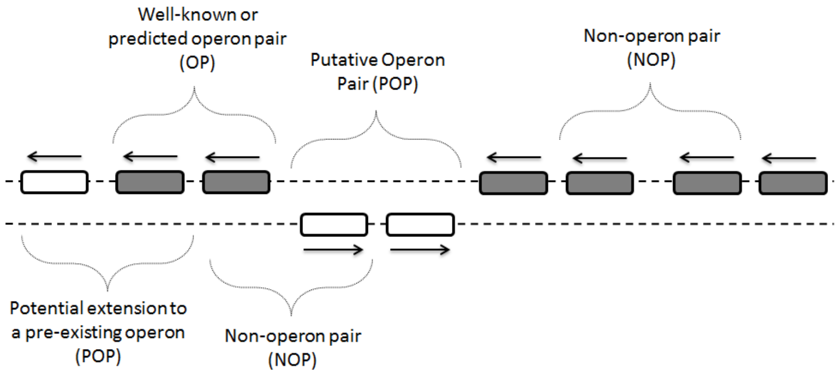

Well-known or predicted operon pair (OP)
Putative Operon Pair (POP)
Non-operon pair (NOP)
Potential extension to a pre-existing operon (POP)
Non-operon pair (NOP)

Fig. 23 – Illustration of the operon pair definitions. Grey boxes indicate genes that are part of a known operon, white boxes indicate genes of unknown status. Arrows show the direction of transcription of each gene.

The operon prediction problem can be simply considered as the partitioning of a genome into gene clusters, where all genes within the cluster share a promoter and terminator. Alternatively, it can be treated as a classification problem, i.e., determining whether an adjacent gene pair belongs to the same operon or not. Therefore, any pair of genes which are adjacent on a single DNA strand can be divided into two groups, OP (operon pair) or NOP (non-operon pair) using appropriated features.

The prediction is to divide all pairs of adjacent genes into predicted OPs and predicted NOPs. The prediction is evaluated using a set of gene pairs for which it is known whether they belong to the same operon (class OP) or to different operons (class NOP).

5.2.4 Machine Learning and Supervised Classification

The Machine Learning field evolved from the broad field of Artificial Intelligence, which aims to mimic intelligent abilities of humans by machines. The aim of machine learning research is to address the following question: how to make machines able to "learn" from experience?

Learning in this context is understood as inductive inference, where some examples are observed in order to collect incomplete information about some "statistical phenomenon". In unsupervised learning one typically tries to discover hidden regularities (e.g. clusters) or to detect anomalies in the data (for instance some unusual machine function or a network intrusion). In supervised learning, there is a class label associated with each example and it is supposed to be the answer to a question about the example. If the label is discrete, then the task is called classification problem – otherwise, for real valued labels we speak of a regression problem.

Using these examples (including the class labels), one is particularly interested in predicting the answer for other cases not yet explicitly observed. Therefore, the process of learning is not mainly focused on remembering the observed examples, but on generalizing the recognition of new examples.

An important task in Machine Learning is classification: to realize algorithms capable of automatically distinguish between different types of data from examples, based on their input features. More formally, the (supervised) classification goal is to find a functional mapping between the input data X, describing the input feature set, to a class label Y (e.g. −1 or +1), such that Y = f(X). This function should be able to associate specific feature-values to the different class labels. Besides, the construction of the mapping is based on so-called training data supplied to the classification algorithm. The final goal is to accurately predict the correct label of unseen data.

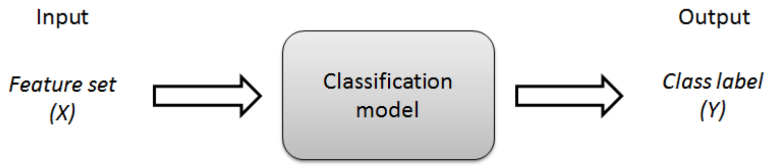


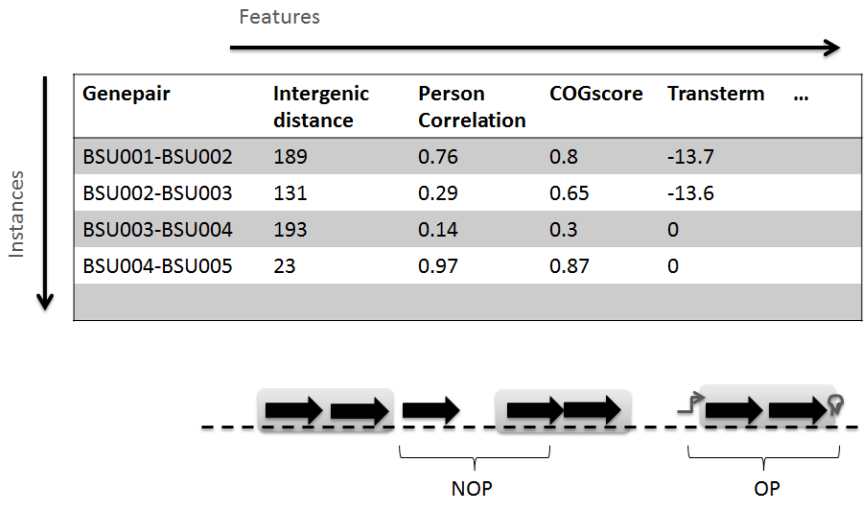
Fig. 24 – Classification problem. Classification is the task of mapping an input attribute set x into its class labels y.

A classification system takes into account the features of the observed examples in order to learn how to distinguish among the class labels. For instance, in a face recognition task some features could be the color of the eyes or the distance between the eyes. Thus, the input to a pattern recognition task can be viewed as a two-dimensional matrix, whose rows are the examples and columns are the different features.

Generally, the realization of a supervised classifier involves several sub-tasks: (i) data collection and representation, (ii) feature selection and/or feature reduction, and (iii) classification.

Data collection and representation are mostly problem-specific. Therefore it is difficult to give general statements about this step of the process. The general objective is to find invariant features, that describe the differences in classes as best as possible. On the other hand, feature selection and feature reduction attempt to reduce the dimensionality (i.e. the number of features) for the remaining steps of the task. Finally, the classification phase of the process learns the actual mapping between examples and labels (or targets). In many applications the second step is not essential or is implicitly performed in the third step.

For the operon classification problem the aim is to develop a machine that, using a features set of known OPs and NOPs, can learn how to determine the right operon status of an unlabeled gene pair.



Genepair	Intergenic distance	Person Correlation	COGscore	Transterm	...
BSU001-BSU002	189	0.76	0.8	-13.7	
BSU002-BSU003	131	0.29	0.65	-13.6	
BSU003-BSU004	193	0.14	0.3	0	
BSU004-BSU005	23	0.97	0.87	0	

Fig. 25 – Example of selected features for the classification of OPs and NOPs.

5.2.5 Prediction step

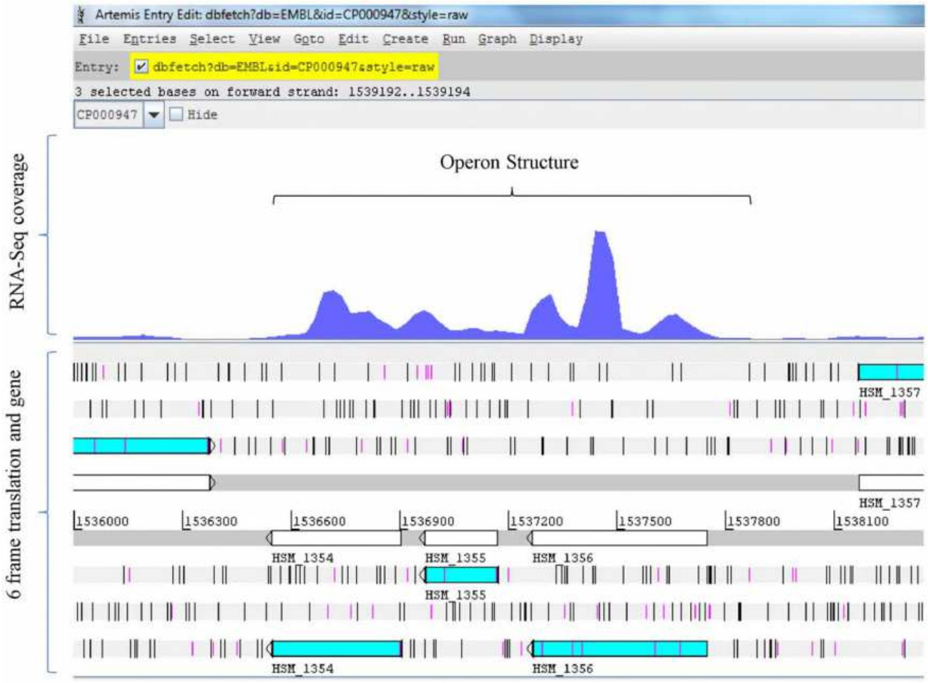
Supervised learning generates a function that maps inputs to desired outputs, which are also called labels, because they are often provided by human experts labeling the training examples. For example, in our classification problem, the learner approximates a function mapping a vector of genomic/transcriptomic feature values (in Figure 25 these vectors correspond to the rows of the table) into OP and NOP classes by looking at input-output examples of the function. Then, follow the inference step that tries to predict new outputs on specific and fixed cases. The inference step is useful to determine the changes in the operon map predicted with standard methods (e.g, DOOR).

5.2.6 Transcriptome features from high-throughput technologies

An important feature characterizing operon structures is that their structural genes are co-transcribed; therefore the transcriptome analysis at single cell resolution becomes crucial to support the prediction of operon pairs.

Nowadays, gene expression microarrays and RNA-Seq are the most popular methods for single-cell transcriptome profiling. Both methods enable high throughput analysis of many cells and gene targets.

Developed in the 1990s, high density microarrays are still a preferable choice for projects that involve large numbers of samples for profiling transcripts in model organisms with well-annotated genomes. Several studies use microarrays for their robust sample processing and analysis pipelines that quickly turn raw data into spreadsheets of gene expression values and significant differentially expressed genes. However, microarray suffers from background hybridization, limited accuracy of expression for transcripts in low abundance, and cannot be used to detect splice variants or unknown genes (Tang *et al.*, 2011), and to estimate the expression level for different transcriptional features such as exons, introns and intergenic regions.


Artemis Entry Edit: dbfetch?db=EMBL&id=CP000947&style=raw

Entry: ☑ dbfetch?db=EMBL&id=CP000947&style=raw

3 selected bases on forward strand: 1539192..1539194

CP000947 ▾ □ Hide

Fig. 26 – Example of gene pairs expressed in tandem (Kumar *et al.*, 2012). The RNA-Seq coverage shows three genes annotated as ribosomal proteins (IF3, L35, and L20), that seem to be expressed as a transcription unit.

More recently, direct sequencing of transcripts by high-throughput sequencing technologies (RNA-Seq) has become an additional alternative to microarrays. This technology provides direct access to the sequence without a reference genome or predesigned probes, and so, bias and variation due to hybridization and labeling efficiencies are avoided. Besides, RNA-Seq experiments generate a huge amount of reads that cover coding and non-coding regions (see CHAPTER 4 for more details). Therefore, we can use RNA-Seq transcriptome profiles to determine whether two adjacent genes, on the same strand, are transcribed in “lock-step” as shown in Figure 26.

5.3 Current bioinformatics methods for operon prediction

The prediction of operon structures is essential not only because it provides the information about which genes are co-regulated, but also because the characterization of other regulatory elements, such as transcription binding sites, promoters, etc., often relies on the delineation of operon structures. Besides, the understanding of the operon organization is important to improve computer annotation of genomes and to infer the function of uncharacterized proteins.

However, experimental methods to identify operon structures are very difficult to implement, and so many computational approaches have been proposed for predicting operons. Several methods rely on features based on inherent DNA sequence properties. On the other hand, it has been shown that a number of genomic features relevant to adjacent gene pairs (on the same strand) are useful for predicting whether the pairs belong to the same operons. These features include (i) the intergenic distance (Salgado *et al.*, 2000; Moreno-Hagelsieb and Collado-Videset *et al.*, 2002), (ii) the conservation of gene pairs (gene neighbourhood) across multiple genomes (Overbeek *et al.*, 1999; Tamames *et al.*, 1997), (iii) commonality of function (Taboada *et al.*, 2010), (iv) similarity of codon usage (the frequency with which synonymous codons encode aminoacids in neighboring genes, Bockhorst *et al.*, 2003) and (v) the correlation of their gene expression patterns (Sabatti *et al.*, 2002).

Furthermore, as shown in Table 6, several computational methods have been tested including (i) hidden Markov model-based methods (Yada *et al.*, 2009), (ii) support vector machines (Zhang *et al.*, 2006), (iii) simple statistical methods (Chen *et al.*, 2004), (iv) Bayesian methods (Westover *et al.*, 2005), (v) graph-theoretic approaches (Bockhorst *et al.*, 2003, Tjaden *et al.*, 2002), (vi) neural networks (Tran *et al.*, 2007), and (vii) classifiers from PRTools Matlab toolbox (Dam *et al.*, 2007).

Authors	Feature					Model
	IGR	***CL***	***Functional relations***	***Genome properties***	***Experimental evidence***	
Yada *et al.*, 1999	X			Promoters, terminators,		HMM
Craven *et al.*, 2000	X		functional classificati	Promoters, terminators, operon size	39 DNA microarray datasets	Naive Bayes
Salgado *et al.*, 2000	X		functional classification			Log-likelihood scores
Moreno-Hagelsieb and Collado-Vides *et al.*, 2002	X	X	functional classification			Log-likelihood scores
Sabatti *et al.*, 2002	X				72 DNA microarray datasets	Bayesian classifier
Tjaden *et al.*, 2002					Tilling DNA microarrays	
Bockhorst *et al.*,	X			Codon usage,	39 DNA	Bayesian

2003				promoters, terminators, operon length	microarray datasets	network
Chen *et al.*, *2004*	X	X	COG	Transcriptional terminators, conserved promoters		Log-likelihood scores
Paredes *et al.*, 2004	X			Promoters, transcriptional terminators,		Empirical scoring scheme
Jacob *et al.*, 2005	X	X	Metabolic pathways, protein function			Fuzzy guided genetic algorithm
Price *et al.*, *2005*	X	X	COG	Codon adaptation index		Naive Bayes approach
Westover *et al.*, *2005*	X	X	Functional relatedness			Naïve Bayes approach
Janga *et al.*, *2006*				Oligo-nucleotide signatures		Log-likelihood scores
Zhang *et la.*, *2006*	X	X	Metabolic pathways, interacting protein domains			Support vector machine
Charaniya *et al.*, *2007*	X			Transcriptional terminators	67 DNA microarray datasets	Support vector machine
Dam *et al.*, *2007*	X	X	GO	TTTTT motif, gene length ratio		11 classifiers from PRTools
Tran *et al.*, *2007*	X		Metabolic pathways, GO			Neural network
Laing *et al.*, *2008*				Transcription factor binding sites		

Tab. 6 – Main methods to predict operons.

5.3.1 Limitation of current approaches

Several methods predict operons using a model trained on a set of experimentally-defined transcription units in prokaryotic organisms (Edwards *et al.*, 2005). Unfortunately, prokaryotic genomes with experimentally verified operon data sets exist only for a few model organisms. In fact, the above-mentioned methods show a high-prediction accuracy only because they focus on the prediction (and verification) of operons in *E. coli* and *B. subtilis*, model organisms that have well defined operon maps. Therefore, these methods tend to generalize well only to the genomes of species closely related to these model organisms, and so we cannot completely rely on these prediction models for distantly related species.

For instance, an operon-prediction program, trained on *E.coli* data, could have 91% prediction accuracy on (other) *E.coli* operonic gene pairs but have its accuracy dropped to 64% when tested on *B.subtilis* operonic gene pairs (Romero and Karp *et al.*, 2004).

In addition, recent whole-transcriptome RNA-Seq analysis have identified new operon pairs that were not predicted by standard operon prediction algorithms and genes within predicted or experimentally verified operons that were not co-transcribed (Kumar *et al.*, 2012; Hövik *et al.*, 2012). In particular, an interesting discovery by (Güell et al., 2009) was obtained on the genome of *Mycoplasma pneumoniae*. Güell *et al.* (2009) applied two technologies, tiling arrays and direct strand-specific sequencing (DSSS) by means of Illumina, for analyzing bacteria. Using a combination of these techniques, they were able to observe the expression of all the genes. The analysis revealed the ability of the operonic structure to change in response to different experimental conditions (173 different conditions were tested); that is, one gene, coded on operons under one condition, can be transcribed as a single transcription unit in another one. Furthermore, other studies (Brinza *et al.*, 2010; Oliver *et al.*,2009) have revealed that often within operons transcription start and end sites occur that indicate operon formation could be due to the use of alternative promoters and terminators instead of rearrangement or deletion events.

These findings indicate that bacterial operon structures can be more complex and dynamic than previously appreciated, and that transcriptomic data deriving from RNA-Seq experiments can help us to better understand these dynamics and improve the accuracy of operon predictions.

5.4 A novel method to improve the accuracy of operon predictions

Here I propose a novel method to improve the whole-genome prokaryotic operon map inference, by using RNA-Seq data analysis, and to refine predictions generated with standard sequence feature based strategies. Since the transcriptome of an organism is dynamic and condition dependent, we use RNA-Seq mapped reads to determine a set of confirmed or predicted operons and extract from it transcriptomic features that are combined with standard genomic features in order to train and validate three classifiers: Random Forests (RFs), Neural Networks (NNs) and Support Vector Machines (SVMs). These models are then exploited to refine the operon map predicted by DOOR (Mao *et al.*, 2008), including the prediction of potential new operons. An inherent limitation of using RNA-Seq to improve operon structure predictions is that it does not apply to genes not expressed under the studied condition.

The proposed method has been evaluated on whole-transcriptome profiles related to two different RNA-Seq studies. In the first study (Kumar *et al.*, 2012), a single nucleotide resolution transcriptome map of the pathogen *Haemophilus somni* (NC010519) was reported using a standard RNA-Seq method. In the second study (Hövik *et al.*, 2012), the researchers used a strand-specific RNA-Seq method to determine three different transcriptional profiles from the periodontal pathogen *Porphyromonas gingivalis* (NC002950) grown under three different experimental conditions. For more details see the information reported in Table 7.

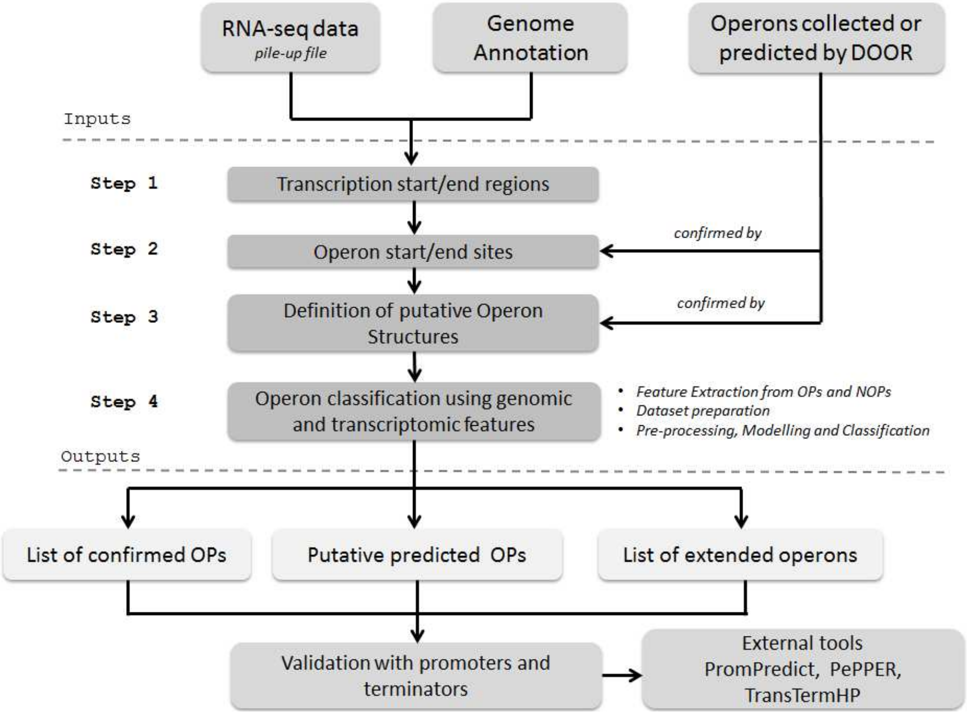

RNA-seq data
pile-up file
Genome Annotation
Operons collected or predicted by DOOR
Inputs
Step 1
Transcription start/end regions
Step 2
Operon start/end sites
confirmed by
Step 3
Definition of putative Operon Structures
confirmed by
Step 4
Operon classification using genomic and transcriptomic features
Feature Extraction from OPs and NOPs
Dataset preparation
Pre-processing, Modelling and Classification
Outputs
List of confirmed OPs
Putative predicted OPs
List of extended operons
Validation with promoters and terminators
External tools PromPredict, PePPER, TransTermHP

Fig. 27 – Workflow of the proposed operon prediction method. The inputs are: a whole-transcriptome RNA-Seq profile (pile-up file) of a prokaryotic organism and the corresponding map of operons collected by DOOR. The core process is represented by four steps. The first three steps determine a partial operon structure from experimental data. While, in the last step the system trains and validates the NN-, a RF- and a SVM-based classifiers on a list of confirmed OPs and NOPs. In output, these classifiers are used to reassess the operon structure annotated in DOOR. Furthermore, a validation process is accomplished to verify that there are not regulatory signals between adjacent genes predicted as operon pairs.

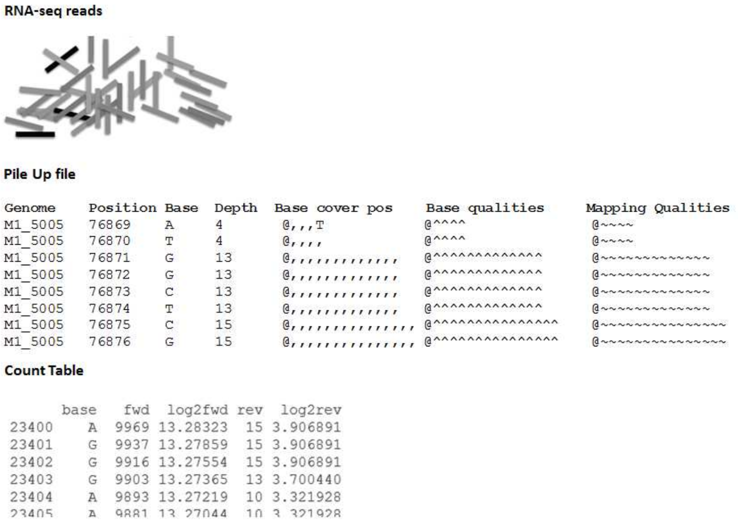
The key elements of my computational approach are (i) to identify the start/end transcription sites and the expression levels of annotated genes and intergenic regions, (ii) to determine a set of confirmed operons using DOOR annotations (iii), to use both genomic and transcriptomic features to train and validate models for the classification of known operon pairs (positive class) and non-operon pairs (negative class). Finally, (iv) the classifiers are exploited to refine the whole operon structures, comprising the prediction of potential new operons. Since genes in an operon are transcribed as a group, no regulatory signals should be present between the genes representing an operon pairs. Therefore, (v) promoters and terminators are predicted across the genome, using prediction programs such as PromPredict (Rangannan and Bansal *et al.*, 2010; 2011) and TransTermHP (Kingsford *et al.*, 2007), to validate the gene pairs classified as operon pairs. Figure 27 illustrates the synthetic scheme of the proposed operon prediction method.

5.4.1 Identification of transcript boundaries with RNA-Seq data

The proposed method finds the boundaries of transcriptionally active regions using a pile-up file and a sliding window correlation procedure.

Pile-up file

A pile-up file represents the signal map for the whole genome in which alignment results are represented in per-base format. This file is used to generate a count table (or table of counts) that reports the number of reads mapped for each genomic position (Figure 28). This table is then employed to determine putative transcription start/end points and to obtain the expression level for the annotations.



RNA-seq reads

Pile Up file

```
Genome     Position Base  Depth  Base cover pos     Base qualities      Mapping Qualities
M1_5005    76869    A     4       @,,,T             @^^^^                @~~~~
M1_5005    76870    T     4       @,,,,             @^^^^                @~~~~
M1_5005    76871    G     13      @,,,,,,,,,,,,,    @^^^^^^^^^^^^^       @~~~~~~~~~~~~~
M1_5005    76872    G     13      @,,,,,,,,,,,,,    @^^^^^^^^^^^^^       @~~~~~~~~~~~~~
M1_5005    76873    C     13      @,,,,,,,,,,,,,    @^^^^^^^^^^^^^       @~~~~~~~~~~~~~
M1_5005    76874    T     13      @,,,,,,,,,,,,,    @^^^^^^^^^^^^^       @~~~~~~~~~~~~~
M1_5005    76875    C     15      @,,,,,,,,,,,,,,, @^^^^^^^^^^^^^^^     @~~~~~~~~~~~~~~~
M1_5005    76876    G     15      @,,,,,,,,,,,,,,, @^^^^^^^^^^^^^^^     @~~~~~~~~~~~~~~~
```

Count Table

```
      base   fwd  log2fwd rev  log2rev
23400    A  9969 13.28323  15 3.906891
23401    G  9937 13.27859  15 3.906891
23402    G  9916 13.27554  15 3.906891
23403    G  9903 13.27365  13 3.700440
23404    A  9893 13.27219  10 3.321928
23405    A  9881 13.27044  10 3.321928
```

Fig. 28 – Example of a table of counts. Mapped reads are assembled into expression summaries representing by tables.

Tables of counts can be displayed in Artemis (Carver *et al.*, 2012). Artemis is a free genome browser and annotation tool that allows visualisation of sequence features, next generation data and the results of analyses within the context of the sequence. An example of mapped RNA-Seq reads displayed by Artemis is reported in Figure 26.

Determination of transcript boundaries

The procedure for the identification of transcript borders measures the correlation value between a vector of 100 integers (for instance x=[0_{50},1_{50}]) modeling a simply shape of sharp increases in transcription, and a window that, sliding on the genome, represents segments of coverage depth of one hundred consecutives bases. Therefore, the procedure selects only those windows having a positive correlation coefficient (exceeding 0.7) with the adopted sharp increases model, and with a significant test of correlation p-value (< 10^{-7}). Clearly, the points with a negative correlation coefficient (<-0.7) represent sharp decreases in transcription, and so they are seletcted to determine transcription end points.

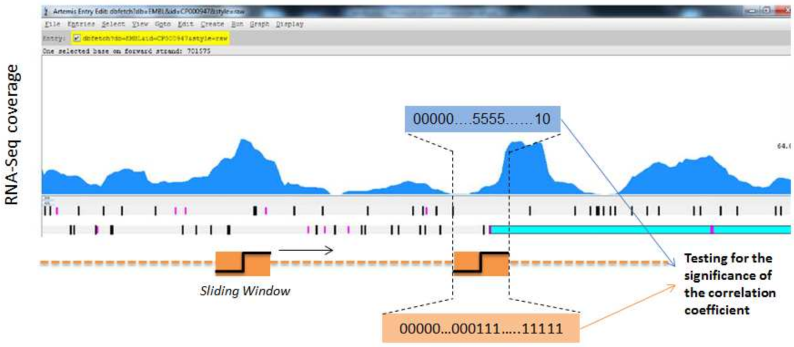
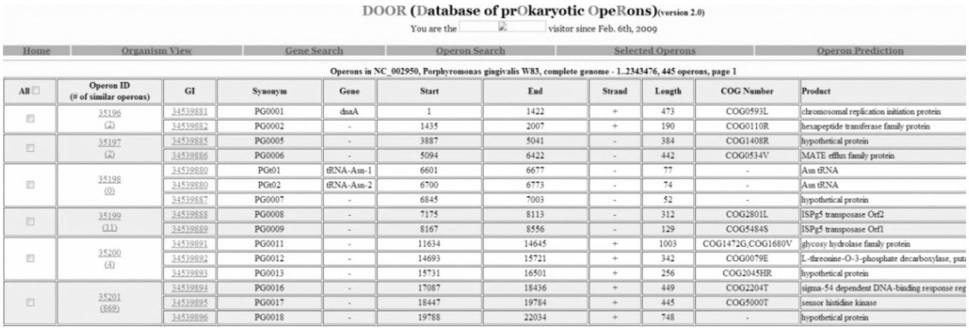

Artemis Entry Edit: dbfetch?db=EMBL&id=CP000947&style=raw
File Entries Select View Goto Edit Create Run Graph Display
Entry: dbfetch?db=EMBL&id=CP000947&style=raw
One selected base on forward strand: 701575
RNA-Seq coverage
00000….5555……10
Sliding Window
00000...000111.....11111
Testing for the significance of the correlation coefficient

Fig. 29 – Sliding window correlation procedure to identify putative transcription start regions.

Finally, these points of sharp increases or decreases in transcription are validated to determine reliable transcription start/end regions (TSR/TER), that refer to regions rather than specific start/end sites (TSS/TES). For each start point my method estimates the average of coverage depth per nucleotide on the left side of the sliding window (called Start Transcription Level - STL) and filters out those points with an STL value greater than 1.5. In the same way, using the average depth on the right side of the sliding window (called End Transcription Level - ETL), the putative transcription end points are validated.

5.4.2 Annotations

The genome sequences and the annotation files were retrieved from the NCBI[9] database. The NCBI Genome database is a collection of complete large scale sequencing, assembly, annotation, and mapping projects for cellular organisms.


DOOR (Database of prOkaryotic OpeRons)(version 2.0)

You are the visitor since Feb. 6th, 2009

Operons in NC_002950, Porphyromonas gingivalis W83, complete genome - 1..2343476, 445 operons, page 1

All □	Operon ID (# of similar operons)	GI	Synonym	Gene	Start	End	Strand	Length	COG Number	Product
□	35196 (2)	34539881	PG0001	dnaA	1	1422	+	473	COG0593L	chromosomal replication initiation protein
		34539882	PG0002	-	1435	2007	+	190	COG0110R	hexapeptide transferase family protein
□	35197 (2)	34539885	PG0005	-	3887	5041	-	384	COG1408R	hypothetical protein
		34539886	PG0006	-	5094	6422	-	442	COG0534V	MATE efflux family protein
□	35198 (0)	34539880	PGt01	tRNA-Asn-1	6601	6677	-	77	-	Asn tRNA
		34539880	PGt02	tRNA-Asn-2	6700	6773	-	74	-	Asn tRNA
		34539887	PG0007	-	6845	7003	-	52	-	hypothetical protein
□	35199 (11)	34539888	PG0008	-	7175	8113	-	312	COG2801L	ISPg5 transposase Orf2
		34539889	PG0009	-	8167	8556	-	129	COG5484S	ISPg5 transposase Orf1
□	35200 (4)	34539891	PG0011	-	11634	14645	+	1003	COG1472G,COG1680V	glycosy hydrolase family protein
		34539892	PG0012	-	14693	15721	+	342	COG0079E	L-threonine-O-3-phosphate decarboxylase, puta
		34539893	PG0013	-	15731	16501	+	256	COG2045HR	hypothetical protein
□	35201 (869)	34539894	PG0016	-	17087	18436	+	449	COG2204T	sigma-54 dependent DNA-binding response reg
		34539895	PG0017	-	18447	19784	+	445	COG5000T	sensor histidine kinase
		34539896	PG0018	-	19788	22034	+	748	-	hypothetical protein

Fig. 30 – Screenshot from DOOR database.

Annotation files of known (or predicted) operons were recovered from DOOR[10]. DOOR is a database containing computationally predicted operons of all the

[9] http://www.ncbi.nlm.nih.gov/genome/.

sequenced prokaryotic genomes. It is based on a data-mining classifier. The features include intergenic distance, neighborhood conservation, phylogenetic distance, information from short DNA motifs, similarity score between GO terms of gene pairs and Length ratio between a pair of genes. The classifier is a trained decision tree based one the training data from *E. coli* and *B. subtilus*. Currently, DOOR contains predicted operons for 675 organisms with 736 chromosomes and 489 plasmids, with a total of 450,986 operons.

5.4.3 Comparison with annotations

As showed by Oliver *et al*. (2009), RNA-Seq coverage-depth is correlated with qRT-PCR transcript levels indicating that RNA-Seq data is quantitative. Therefore, after the identification of transcript borders, the whole coverage depth is compared with annotations in order to estimate the expression level of annotated coding sequences (CDS regions) and intergenic non-coding sequences (IGR regions). Then, the expression values are normalized with the RPKM method (Mortazavi *et al.*, 2008) to allow a comparison in terms of expression levels between different genes within the same RNA-Seq experiment. The number of reads per kb of transcript per million mapped reads or RPKM has been proposed as a useful metric that normalizes for variation in transcript length and sequence yield. Since RPKM values are log-normally distributed, they ar expressed as *log₂(RPKM)*. RPKM is computed as follows:

$$RPKM = \frac{10^6\ \mathrm{C}}{\mathrm{NL}/10^3} \tag{5}$$

Given RPKM to be the expression of gene *x*, *C* is the number of reads uniquely aligned to gene *x*, *N* is the total number of reads that uniquely aligned to all the genes, and *L* is the number of bases on gene x. The RPKM method is able to eliminate the influence of different gene length and sequencing discrepancy when calculating the gene expression. The 50th percentile of the two distributions of *log_2(RPKM)* values, respectively for CDS and IGR regions, has been used as a minimum expression threshold. These minimum expression thresholds are employed in the next step in which operon structures are explained.

5.4.4 Explanation of operon structures

Before the identification of operons, the proposed method uses the transcription start and end regions to determine a list of operon start- and end-points (OSPs and OEPs). The OSPs are selected using the following filters: (i) the downstream gene is transcribed and the expression level is higher than a selected minimum expression threshold for the CDS regions (for instance, the 50th percentiles of the two distributions of log2(RPKM)), (ii) the corresponding gene overlaps a structural gene of an operon collected by DOOR and (iii) there is enough space for the 5'UTR, that is an untranslated region between the transcriptional start regions and the start codon of an mRNA. In the same way we select OEPs.

Each operon start site is linked to a confirmed operon, therefore I defined a linkage process that adds the next structural genes of an operon until one of the following rules is not verified: (i) the expression level of the intergenic region is higher

[10] http://csbl1.bmb.uga.edu/OperonDB_10142009/DOOR.php.

than the minimum expression threshold for the IGR regions, (ii) the expression level of the next gene is higher than the minimum expression threshold for the CDS regions, (iii) the intergenic region neither contains transcription start points nor contains transcription end points. At the end of this linkage process a collection of operons annotated in DOOR and confirmed by experimental data is determined.

From this collection I select operon pairs (OPs), that are two genes located on the same DNA strand, adjacent to one another, transcribed together and confirmed by DOOR. Moreover, I use the transcription start and end points to also determine non-operon pairs (NOPs), and new potential operon pairs (POPs).

NOPs are two genes that are adjacent, transcribed in the same direction and with a point of start/end transcription into the intergenic region. On the other hand, the POPs are adjacent genes transcribed in the same direction with an unknown operon status or an operon status to refine.

For the POP class we have a special case: gene pairs formed by the last structural gene of an operon collected by DOOR, and the following gene that DOOR indicated as a single gene. Practically, the POP class contains pair of adjacent genes that are transcribed in the same direction and can be part of unidentified, new operons or can represent the extensions to known operons.

The POPs are selected without verifying if the corresponding adjacent genes are co-expressed and if they have a short intergenic distance. This task is assigned to the classifiers, that learn how to distinguish confirmed OPs/NOPs in order to verify if a selected POP can be classified as an operon pair or not. Figure 23 shows the three types of pairs of adjacent genes that have been considered.

5.4.5 Selected features

Several computational methods predict operons based on the properties of adjacent genes, which they try to identity as either OP or NOP. Often, distances between genes and, generally, genomic comparative features have been used for predicting operons. The aim of this work is to provide a transcriptome dynamics-based operon prediction method, that use features extracted from RNA-Seq transcriptome profiles with standard DNA sequence features to correctly classify OPs and NOPs. Therfore, I selected two features based on genome sequence, intergenic distance and codon usage similarity, and two features that depend on the transcriptome, difference in expression levels of adjacent genes and the expression level within the intergenic region.

Intergenic distance

The Intergenic distance represents the number of base pairs separating two consecutives genes. A distance-based operon prediction technique was first described by Salgado *et al.* (2000). The authors, using the genomic sequence of *E. coli* K-12, found that the distribution of distances between adjacent genes in operons differs from the distribution of distances between adjacent genes at the boundaries of transcription units. Also Moreno-Hagelsieb and Collado-Videset *et al.* (2002) provided evidence that the distance-based method can be used to predict operons in any prokaryotic genome. Therefore, I decided to adopt the distance between two consecutive genes as the first genomic feature for our classifiers. The intergenic distance was calculated as follows:

$$igrLength(g_i, g_{i+1}) = g_i(start) - \ g_{i+1}(end + 1) \qquad (6)$$

Codon usage features

It is well-known that the genes in the same operon often exhibit similar codon-usage patterns while genes in different operons exhibit different codon bias (Harayama *et al.*, 1994). Consequently, to decide whether or not two consecutive genes constitute an operon pair, it may be helpful to consider the codon usage of the genes (Bockhorst *et al.*, 2003). We associate with each gene g_i a vector of Relative Synonymous Codon Usage bias (RSCU value), that contains a RSCU value for each aminoacid *a*:

$$RSCU_a(g_i) \tag{7}$$

Using the RSCU values, we define the codon usage similarity:

$$RSCU_a(g_i, g_{i+1}) = \sum_a RSCU_a(g_i) * \ RSCU_a(g_{i+1}) \tag{8}$$

This measure is symmetric and indicates the consistency and degree to which the bias vectors are characterized by a similar codon bias.

RSCU is a simple measure of non-uniform usage of synonymous codons in a coding sequence (Sharp et al., 1986). An RSCU value for a codon is the observed frequency of that codon divided by the frequency expected under the assumption of equal usage of the synonymous codons for an aminoacid. In the absence of any codon usage bias, the RSCU value would be 1.00. A codon that is used less frequently than expected will have a value less than 1.00 and vice versa for a codon that is used more frequently than expected.

Difference in expression levels

The difference of expression level between two consecutive genes represents the first selected feature extracted from an RNA-Seq transcriptome profile.

$$diffExpr(g_i, g_{i+1}) = |log_2(RPKM(g_i)) - \ log_2(RPKM(g_{i+1})| \tag{9}$$

Genes that are transcribed as a part of an operon should exhibit a similar transcription level, because these genes were in the same transcript. Therefore, two consecutive genes that constitute an operon pair should have a difference close to zero. While for non-operon pair this feature should exhibit a high value of mean and variance. My results confirmed that this feature is a good discriminator for operon prediction.

Expression level of intergenic regions

The expression level of intergenic regions represents the second transcriptomic feature that I selected. If the intergenic region between two consecutive genes is highly expressed and these genes are transcribed in the same direction with a similar expression value, than they can be part of an operon structure. We compare the RNA-Seq based transcriptome profile with the available genome annotation to identify expression level for the intergenic regions.

$$igrExpr(g_i, g_{i+1}) = log_2(RPKM(igr(g_i, g_{i+1})) \tag{10}$$

The *igr()* function takes the intergenic region between g_i and g_{i+1}. It is excepted that the intergenic expression levels of consecutive genes in the same operon to be higher than the intergenic expression levels of genes that are not in the same operon. Indeed, in our experimental results, the intergenic expression level of confirmed OPs is almost always greater than the intergenic expression level of NOPs. Moreover, in order to obtain a value of this feature for overlapping genes, we use a pessimistic model to compute an approximated transcription level of the overlapping section.

$$igrExpr(g_i, g_{i+1}) = minExpr(g_i, g_{i+1}) + \mathrm{w}(g_i, g_{i+1}) \tag{11}$$

$$minExpr(g_i, g_{i+1}) = \min\big(log_2(rpkm(g_i), log_2(RPKM(g_{i+1})\big) \tag{12}$$

$$\mathrm{w}(g_i, g_{i+1}) = diffExpr(g_i, g_{i+1}) * \frac{overlapping\_base\_pairs}{\min(length(g_i), length(g_{i+1}))} \tag{13}$$

Essentially, this model determines an increment of the expression level resulting from the overlapping section and adds this increment to the expression level of the gene with the minimum expression value, as shown in equation 11. In this way, the larger is the difference in gene level expression and/or the overlapping section, the greater is the increment value *w*, and this influences the expression values estimated for the overlapping region. The experimental results obtained in this thesis evidenced that this feature is a good discriminator for operon prediction.

5.4.6 Supervised classification

From the collection of operons confirmed by RNA-Seq transcriptome profiles (Section 5.4.4), a dataset of operon and non-operon pairs were identified and the values for the genomic and transcriptomic features were selected using the functions introduced in Section 5.4.5. This dataset was used to train and validate three classification models: RFs, NNs and SVMs. These three models represent the most recently machine learning methods used for operon prediction (Charaniya *et al.*, 2007; Taboada *et al.*, 2010; Tran *et al.*, 2007; Tjaden *et al.*, 2002). Before the validation process, a heuristic grid search step was used to select the best parameters for the models. This selection was performed searching for the combination of parameters that gives the best cross-validation accuracy. We used the R package rminer (Cortez *et al.*, 2010) to accomplish this step. For the RF model a different procedure was used to determine the best parameters by taking into account its characteristic learning process.

Random Forests

Random Forests (RFs) are an ensemble learning method proposed by Breiman *et al.*, (2001). They classify data generating many decision trees and aggregating their results to obtain a more accurate classifier. In random forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run. Each generated tree is constructed using a different bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample (out-of-bag, oob) and not used in the construction of the k_{th} tree. In this way, a test set classification is obtained for each case in about one-third of the trees. The procedure has proven to be unbiased in many tests. RFs have only one tuning parameter, *mtry*, which is the number of the descriptors randomly sampled as candidates for the splitting at each node during the tree induction. The *mtry* parameter was estimated using the *tuneRF* function in the R package *randomForest* (Liaw and Wiener *et al.*, 2002). This function automatically selects the optimal value of mtry with respect to the out-of-bag correct classification rates.

Neural Networks

Neural networks (NNs) are one of the most commonly used approaches in data classification. NN is used when we want to combine multiple sources of information, without assuming the underlying relationships among the individual data sources. For the classification of OPs/NOPs, a multilayer perceptron neural network (MLP) has been used. MLP is a modification of the standard linear perceptron and can distinguish non-linearly separable data. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a non-linear activation function. MLP utilizes a supervised learning technique called back-propagation for training the network. The classification performance of a NN model is mainly affected by one hyperparameter, that is the number of hidden nodes: *H*. Therefore, the NNs model was optimized using a grid search for the best *H* parameter.

Support Vector Machines

SVMs are a class of kernel-based machine learning methods that use the principle of structural risk minimization to identify a decision function that separates objects from two classes with maximum margin. More formally, a support vector machine constructs a hyperplane or a set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger is the margin the lower is the generalization error of the classifier. For any kernel function the soft margin parameter *C* must be determined. For our scope we employed a binary SVM classifier with a Gaussian kernel which presents less parameters than other kernels.

$$K(x, x') = \exp(-\gamma \|x - x'\|) \, , \gamma > 0 \tag{14}$$

With a Gaussian kernel, there is a single parameter *γ* to optimize. Therefore, the classification performance of the SVMs was affected by two hyperparameters: the kernel parameter *γ*, and *C*. The best combination is selected using a grid search with exponentially growing sequences of *C* and *γ*, for example:

$$C \in \{2^{-5}, 2^{-3}, \ldots, 2^{-13}, 2^{-15}\}, \gamma \in \{2^{-15}, 2^{-13}, \ldots, 2^{1}, 2^{3}\} \qquad (15)$$

Typically, each parameter combination choice is checked using cross validation, and the parameters with best cross-validation accuracy are picked.

5.4.7 Model validation

After searching the best parameters, from a dataset of confirmed OPs and NOPs, I randomly selected and held out the 30% of the dataset as test set and used the rest of the 70% for a 5-fold cross validation procedure. The 5-fold cross validation procedure was repeated ten times in order to estimate several evaluation metrics and thus compare the performance of the three classifiers.

The classification system did not run cross validation in RFs. With this model, there is no need for cross-validation to get an unbiased estimate of the test set error, because it is internally estimated, during the run. However, in order to compare the different classifiers, the RF model run ten times the training/validation process.

Evaluation metrics

The following metrics were used to compare the performance of the three different classifiers:

- True Positive Rate (TPR): TP/(TP + FN).
- Positive predictive value or Precision (PPV): TP/(TP + FP).
- False Positive Rate or Recall (FPR): FP/(FP + TN).
- Error rate (ER): (FP + FN)/(TP + FN + TN + FP).
- Accuracy (ACC): (TP + TN)/(TP + TN + FP + FN).

Where: TP (True Positives)=Number of OPs accurately classified as op-eron pairs by the model; FN (False Negatives)=Number of OPs falsely classified as non-operon pairs by the model; FP (False Positives)=Number of NOPs falsely classified as operon pairs; TN (True Negatives)=Number of NOPs accurately classified as non-operon pairs. Recall quantifies the sensitivity of the model, i.e. how many OPs can be predicted as operon pairs by the model, and precision quantifies the specificity of the model, i.e. how many of the operon pairs predicted from the training set (OPs and NOPs) are OPs. Then, error rate is the percentage of errors made over the whole set of instances (records) used for testing. Finally, the accuracy is the percentage of well classified data in the testing set.

K-fold cross validation and ROC curves

The 5-fold cross-validation randomly splits data into five subsets (called 5-fold) of the same size. The first four fold are used for training and the remaining fold, instead, is used for testing the classifier. The 5-fold cross-validation is performed 10 times (10 x 5) and the true class of the gene pairs in each of the 50 test subsets are then used to generate receiver operating characteristics (ROC) graphs, or ROC curves (Egan *et* *al.*, 1975), and get the main evaluation metrics. A single ROC curve for each classifier were calculated in order to measure the overall accuracy of our three classification models. An ROC curve plots FP versus TP fractions. The percentage of false positives is shown on the x-axis and the percentage of true positives is shown on the y-axis. Every ROC plot has a diagonal line indicating the performance of a predictor that randomly assigns genes to the OP and NOP classes. Consequently, an ROC curve that is well above the diagonal random line represents a significant predictive power and a curve below the diagonal suggests that the predictor consistently gives wrong results. The area under ROC curve (AUC) can be used as a further evaluation metric. The AUC score for a random classifier is 0.5 and that of an ideal classifier is 1.

5.4.8 Promoters and Terminators

Transcription of a unit encoding a single gene or an operon is controlled by a promoter and a terminator. Therefore, I validated the gene pairs classified as an operon pair (OPs or POPs) verifying the absence of any promoter or terminator in the corresponding intergenic regions. The promoters and terminators were predicted across the genome, using prediction programs such as PromPredict (Rangannan-Bansal *et al.*, 2010; 2011), Pepper (Prediction of Prokaryote Promoter Elements and Regulons) tool (de Jong *et al.*, 2012), and TransTermHP (Kingsford *et al.*, 2007), to add confidence to the identified novel operon pairs.

PromPredict can identify putative promoters using the whole-genome percentage GC of selected bacterial genomes (Rangannan and Bansal *et al.*, 2009). The average free energy (E) over known promoter sequences and the difference (D) between E and the average free energy over downstream random sequences (REav) are used to search for promoters in the genomic sequences. This classification system has been used to predict promoters in 913 microbial genomes that have been accumulated in a database called PromBase. Pepper Toolbox provides an improved promoter prediction tool for prokaryotes based on curated PWM and HMMs models. The training of HMMs is based on DBTBS, RegulonDB and MolGen. While, TransTermHP finds rho-independent transcription terminators in bacterial genomes. The algorithm searches for mRNA motifs that potentially form a hairpin structure and are followed by a short uracil-rich region both within and between the genes.

5.5 Result and Discussion

In order to evaluate the performance of the presented method, I tested it on two different RNA-Seq studies: Kumar *et al.* (2012) and Høvik *et al.* (2011). In both the studies the compiled transcriptome profiles were based on total RNA samples isolated from different laboratory culturing conditions, and the mapped RNA-Seq reads covered by both coding and non-coding sequence regions. In the first study the authors used a standard RNA-Seq method for the experimental annotation of the *H. somni* (strain 2336) genome (here indicated as HS2336) and to construct a single nucleotide resolution transcriptome profile. In the second study, the authors applied a strand-specific RNA-Seq protocol to analyze the transcriptome of the periodontal pathogen *P. gingivalis* (strain W83). From this second study three different strand-specific transcriptome profiles were obtained using three different experimental growth conditions, here called PG1, PG2 and PG3 (Table 7).

GEO-Accession	GSE30452	GSE29578
Organism	*P. gingivalis* (W38)	*H. somni* (2336)
#Conditions	3	1
Platform	Illumina Genome Analyzer II	Illumina Genome Analyzer II
Library	Strand-specific cDNA	Illumina protocol
Sample	PG1 (MIN - a chemically defined minimal medium), PG2 (TSB - trypticase soy broth) PG3 (BAPHK - blood agar plates)	HS2336 (a single nucleotide resolution transcriptome map)

Tab. 7 – Information about the RNA-Seq transcriptome profiles used for testing.

I focused these RNA-Seq studies for two reasons. In these studies, the comparison of co-expressed gene pairs identified by RNA-Seq and operons predicted by DOOR reveals the presence of potential new operons that were not predicted by DOOR. They hypothesized that some co-expressed gene pairs should have been part of new operon structures and, that, the existing computational approaches for operon prediction fail because they do not use transcriptome RNA-Seq data. Moreover, the transcription profiles were obtained using different RNA-Seq methods: strand- and not strand-specific. The strand-specific RNA-Seq protocol compared to the standard RNA-Seq methods provides additional valuable information to improve the accuracy of annotations. Therefore, the test aims to prove that the proposed technique is valuable also when the RNA-Seq transcriptome profile is based on strandness reads.

General Information	*P. gingivalis* (W38)	*H. somni* (2336)
Annotated genes	2,053	2,065
Annotated operons (DOOR)	445	464
Platform	67.73%	69.98%

Tab. 8 – Information about the number of genes and operons annotated in *P. gingivalis* (W38) and *H. somni* (2336).

5.5.1 Empirical Evaluation

This section shows the predictive accuracy of the proposed method. Before the classification process, my method defines a set of confirmed OPs, NOPs and POPs for each considered RNA-Seq transcriptome profile, as described in Section 5.4.4. OPs are co-expressed gene pairs collected by DOOR and confirmed by experimental data; NOPs are gene pairs having a transcription start or point in the corresponding intergenic region; POPs are pairs of adjacent genes that are transcribed in the same direction and can be part of unidentified, new operons or can represent the extensions to known operons. The set of OPs and NOPs is then used to build a dataset containing the values of selected genomic and transcriptomic features, as described in Section 5.4.6. This dataset is then exploited for the training and validation of three classifiers (NN, RF and SVM), as described in Section 5.47.

Table 9 reports the number of confirmed and annotated OPs and NOPs that have been used to generate the training and testing datasets. While, Table 10 presents the number of co-expressed gene pairs that have an unknown operon status (POPs) and expressed gene pairs that are annotated in DOOR and are not linked to an identified transcription operon start-point. For these last gene pairs we need to re-define the operon status reported in DOOR.

Confirmed OPs and NOPs	PG1	PG2	PG3	HS2336
Number of OPs	121	124	124	101
Number of NOPs	63	53	45	168

Tab. 9 – List of adjacent gene pairs with a known operon status. For each transcriptome RNA-Seq profile (columns) this table reports the number of confirmed and annotated OPs and NOPs.

Gene pairs with questionable operon status	PG1	PG2	PG3	HS2336
Number of POPs	80	83	81	53
Number of adjacent and expressed genes	268	271	286	164

Tab. 10 – List of adjacent gene pairs with an operon status to correct. For each RNA-Seq profile this table reports the number of gene pairs with an unknown operon status and selected as putative operon pairs. In addition, we show the number of operon pairs that are expressed, annotated in DOOR, and not linked to a TSR (Section 5.4.1).

From each list of OPs and NOPs, the proposed method randomly selects the 30% as test set and uses the rest of the 70% for a 5-fold cross validation. Then, several evaluation metrics are calculated to assess the performance of selected classifiers. A comparison of all the accuracy values, for each transcriptome RNA-Seq profile, is reported in Table 11. Results show good performances for all the classifiers, with accuracy values greater than 96%. Other two important metrics are the precision and the recall. The precision quantifies the specificity of the model, that is how many operon pairs predicted from the training set are annotated in DOOR and confirmed by RNA-Seq data. On the other hand, the recall quantifies the sensitivity of the model, that is how many annotated and confirmed OPs can be predicted as operon pairs by the model. All the tests indicate that the average precision values range in [0.91-0.98] for NN-based classifiers, in [0.95-0.99] for RF-based classifiers and in [0.94-0.99] for SVM-based classifiers; for what concerns the recall metric, the values range in [0.97-1] for NNs, in [0.92-0.99] for RFs and in [0.92-0.99] for SVMs [0.94-1]. Figure 31 shows the ROC curves generated to evaluate the overall accuracy of the three methods. The ROC curves display the full picture of the trade-off between sensitivity (TPR) and “1-specificity” (FPR) across a series of cut-off points. From these plots, we observe that the RF-based classifiers showed better result than NN or SVM classifier.

Confirmed OPs and NOPs	Dataset	NNs	RFs	SVMs
HS2336 (70%)	Training	0.9643	0.9578	0.9739
HS2336 (30%)	Testing	0.9638	0.9759	0.9759
PG1 (70%)	Training	0.9223	0.9636	0.9607
PG1 (30%)	Testing	1	0.9824	0.9824
PG2 (70%)	Training	0.9650	0.9690	0.9853
PG2 (30%)	Testing	0.9814	0.9814	0.9814
PG3 (70%)	Training	0.9623	0.97	0.9880
PG3 (30%)	Testing	0.9807	0.9615	0.9807
HS2336 (70%)	Training	0.9643	0.9578	0.9739

Tab. 11 – Accuracy values from the 5-fold cross validation process. For each transcriptome RNA-Seq profile this table reportes the accuracy values obtained with training and testing datasets. The results we report are aggregates from all five folds.

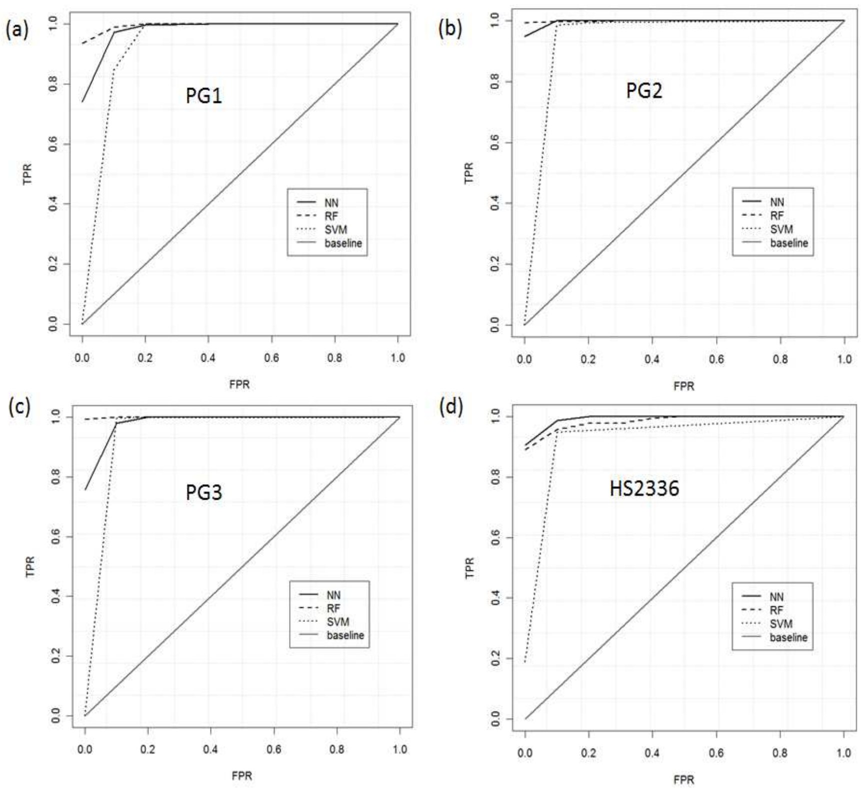


Fig. 31 – Comparison of different classifiers in each RNA-Seq based transcriptome profiles by ROC curves. The goal of an ROC curve analysis is to compare the accuracy of the three classification models trained, every time, on a set of different, confirmed OPs/NOPs in PG1 (a), in PG2 (b), in PG3 (c) and in HS2336 (d). ROC curves foun False positive rate (FPR) is the percentage of non-operon pairs (NOPs) misclassified as operon pairs and the true positive rate (TPR) is the percentage of known operon pairs (OPs) correctly classified as operon pairs. The ROC curves have been generated for each classifier by a 5-fold cross-validation, as described in the Section 5.4.7.

Furthermore, it is possible observe that also with a transcriptome profile based on strandness reads, the accuracy is greater than 95%. This evidences that the proposed technique is valuable for strand- and not strand-specific RNA-Seq experiments. In addition, both the evaluation metrics and the ROC curves demonstrates that the classification models (NNs, RFs and SVMs) are enough robust to yield a reasonable predictive performance about new potential operon pairs in the same genome and with the same transcriptome data.

5.5.2 Classification performance with strandness reads

Strand-specific RNA-Seq improves on standard RNA-Seq in three ways: accurately identifying antisense transcripts, determining the transcribed strand of non-coding RNAs (e.g. lincRNAs), and demarcating the boundaries of transcribed genes. Consequently, it is reasonable to expect more accurate predictions when “stranded” information are available. However, as shown in Table 11, the performance of classifiers based on stranded reads is slightly better than the classifiers based on strandness reads. Therefore, the obtained results prove that the proposed operon classification system is valuable also when the RNA-Seq transcriptome profile is based on strandness reads.

5.5.3 Evaluation of selected genomic and transcriptomic features

In all the transcriptome profiles, the confirmed OPs have shorter intergenic distance and higher similarity of codon usage bias. In addition, the expression differences between adjacent genes constituting an operon pair are close to zero. On the other hand, NOPs have the expression differences that have higher mean values. Furthermore, genes within operons are likely to have a higher expression level in their intergenic regions compared to gene pairs that are labeled as NOP. Clearly, for the presence of particular transcription regulatory elements in the intergenic regions, such as small non-coding RNAs, there is always a minimum level of transcription.

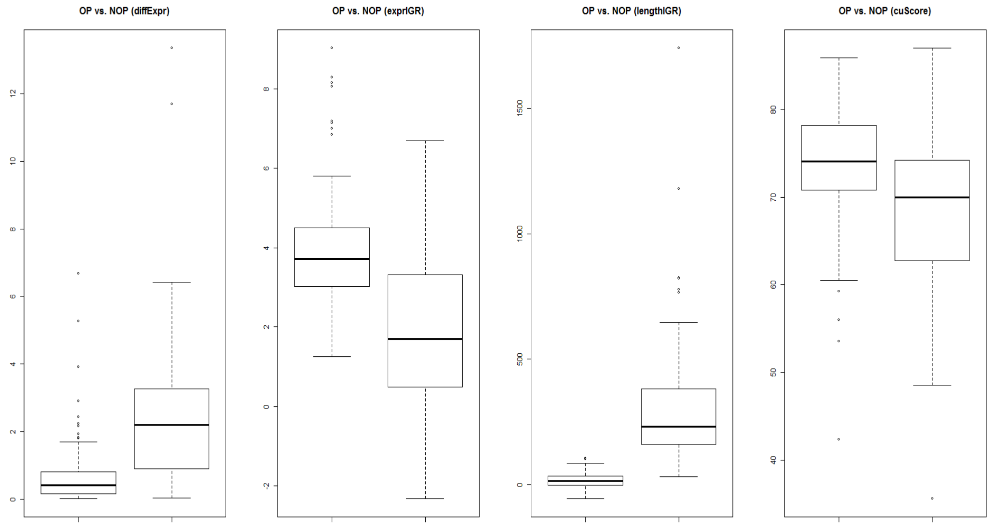


Fig. 32 – Box plots showing the distribution of feature values for OPs and NOPs in PG1.

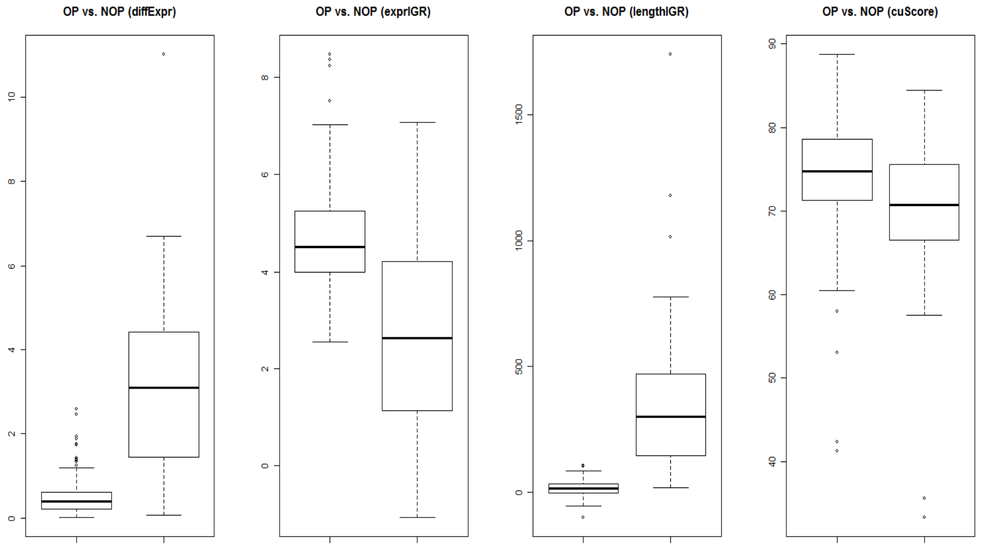


Fig. 33 – Box plots showing the distribution of feature values for OPs and NOPs in PG2.

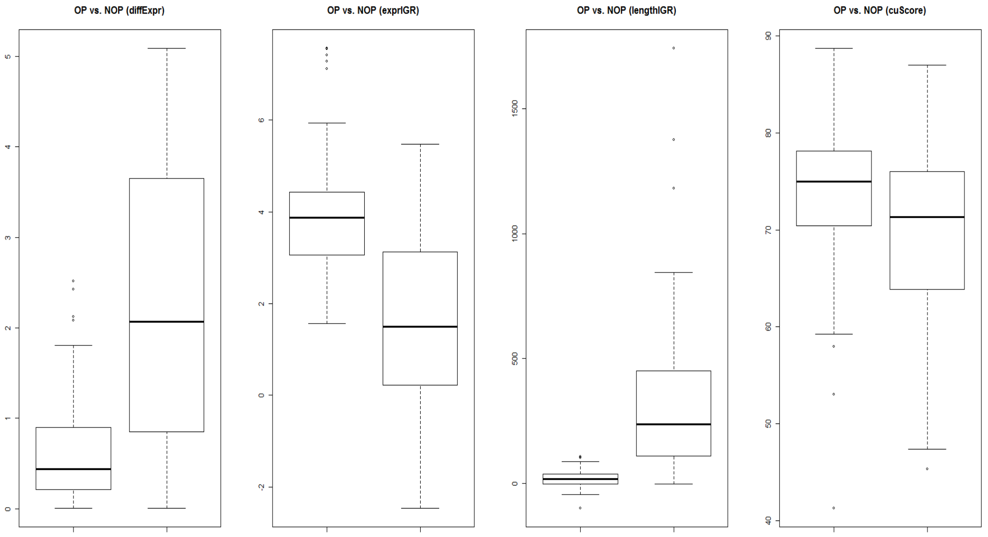


Fig. 34 – Box plots showing the distribution of feature values for OPs and NOPs in PG3.

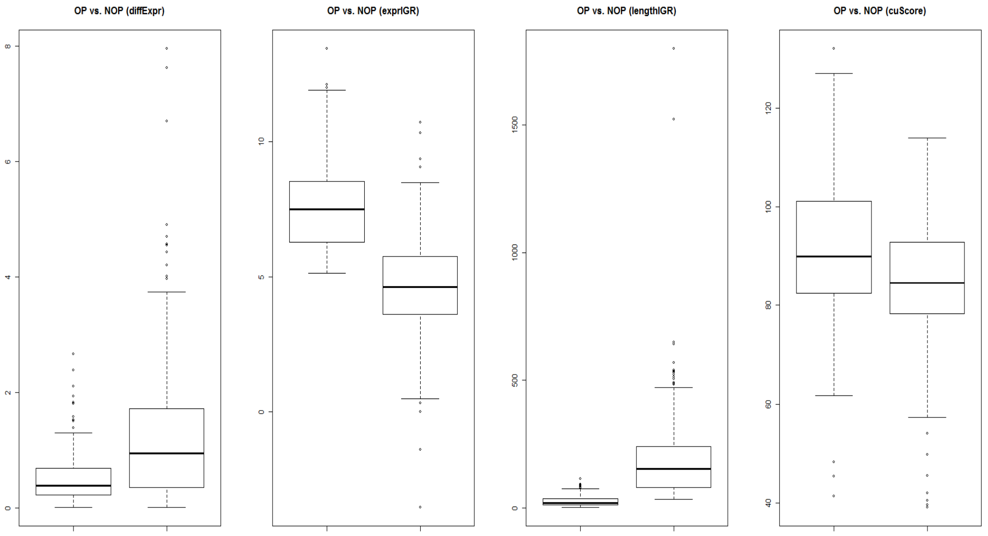


Fig. 35 – Box plots showing the distribution of feature values for OPs and NOPs in HS2336.

5.5.4 Classification performance of different groups of features

The purpose of this section is to examine the predictive power of groups of features. Groups of features are defined in order to evaluate the performance of classifiers based on genomic features with those based on transcriptomic features, and verify the contribution of transcriptomic features in improving the classification accuracy.

As shown in Table 12, when comparing the performance, we found that effectively the combination of all selected features gives the best classification performance in all the datasets. If only transcriptome data is used for classification, the average accuracy ranges in [0.82-0.89]. While, if only genomic properties are used for classification, the average accuracy value ranges in [0.88-0.91]. Therefore, the models trained with standard DNA sequence features performs marginally better than the models trained with only transcriptomic features. However, it is clear that when using genomic and transcriptomic features together we can achieve higher levels of accuracy.

Therefore, we can say that the transcriptomic features extracted from RNA-Seq data are an important factor for determining whether two adjacent genes represent an operon pair or not. Clearly, the discrimination power of these transcriptomic features depends on the quality of RNA-Seq data in terms of sequencing depth, strand specificity, coverage uniformity, and read distribution over the genome structure.

Dataset	Genomic Features			Transcriptomic Features			IGR, CuScore and IGR-Expr			IGR, CuScore and Diff-Expr		
	NN	RF	SVM	NN	RF	SVM	NN	RF	SVM	NN	RF	SVM
PG1	0.84	0.9	0.9	0.84	0.84	0.8	0.84	0.97	0.95	0.76	0.91	0.91
PG2	0.88	0.96	0.9	0.9	0.9	0.89	0.93	0.92	0.92	0.9	0.95	0.97
PG3	0.89	0.94	0.86	0.92	0.84	0.88	0.92	0.95	0.95	0.94	0.96	0.95
HS2336	0.87	0.88	0.9	0.87	0.84	0.87	0.94	0.95	0.93	0.85	0.89	0.90

Tab. 12 – The contribution of transcriptomic features in improving the classification accuracy. The first column reports the accuracy results with all the features. The next two columns show the accuracy values achieved, respectively, with genomic and transcriptomic features. Finally, the last two columns display the improvement, in classification accuracy, obtained combining each transcriptomic feature to the two genomic features.

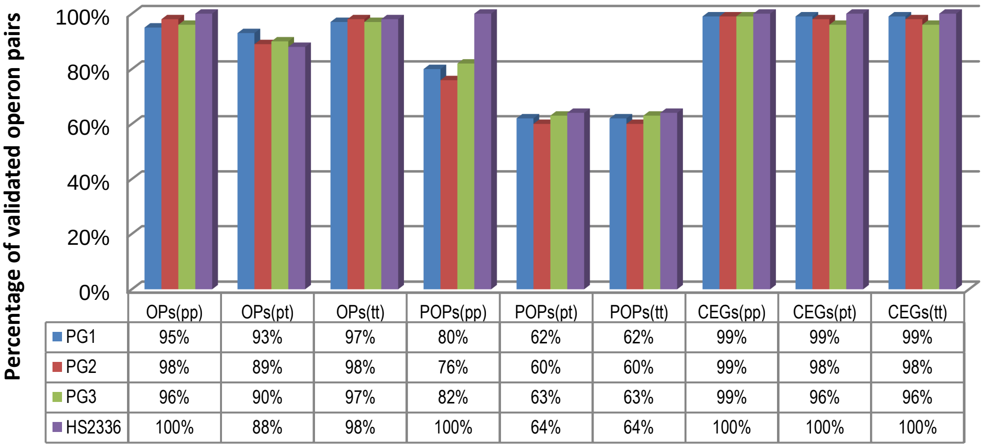
5.5.5 Operon predictions

After the training and validation steps, I used the classification models to predict the class of gene pairs with an unknown operon status or an operon status to redefine. In addition, a simple majority voting schema (SMVS) was adopted to combine the classifier predictions and so to improve prediction accuracies. The voting system tags a gene pair as an OP whether at least two classifiers have predicted that gene pair as an OP. For what concerns the POPs, the proposed voting system identifies 45/80 (in PG1), 52/83 (in PG2), 55/81 (in PG3) and 11/53 new OPs (in HS2336). While, for all gene pairs, annotated as operon pairs in DOOR, the voting system indicates that only 124/268 (in PG1), 143/271 (in PG2), 164/286 (in PG3) and 55/164 (in HS2336) are true OPs. In addition, the voting scheme gives the prediction accuracy at 99%, 99%, 100% and 98%, respectively, in PG1, PG2, PG3 and HS2336.

5.5.6 Validation of operon predictions

Figure 4, we show the percentage of validated gene pairs that have been predicted as operon pairs by our classification system. This percentage is determined verifying the absence of control signals (promoter and terminator) in the intergenic regions of gene pair predicted as operon pairs.

We used three prediction programs to identify standard promoter regions (Pepper toolbox), putative promoter regions (PromPredict) and rho-independent transcription terminators (TransTermHP). Validation results showed that the percentage of operon pairs without promoters/terminators is very high in all the datasets and that at least 60% of putative operon pairs, predicted as OPs, do not have a promoter or terminator in the corresponding intergenic regions. Besides, we avoided to verify the presence of control signals in the flanking regions of an operon pair because we do not know if it is part of an operon with more than three genes or not. In addition, in the past, different authors have proved that the identification of signals that occur on the boundaries of an operon, such as promoter and terminator, does not improve the accuracy of operon prediction methods (Yada et al., 1999). Therefore, we did not verify the presence of control signals into the flanking regions of gene pairs predicted as operon pairs.



	OPs(pp)	OPs(pt)	OPs(tt)	POPs(pp)	POPs(pt)	POPs(tt)	CEGs(pp)	CEGs(pt)	CEGs(tt)
🟦 PG1	95%	93%	97%	80%	62%	62%	99%	99%	99%
🟥 PG2	98%	89%	98%	76%	60%	60%	99%	98%	98%
🟩 PG3	96%	90%	97%	82%	63%	63%	99%	96%	96%
🟪 HS2336	100%	88%	98%	100%	64%	64%	100%	100%	100%

Fig. 36 – Summary of validated gene pairs that were predicted as OP. OPs – percentage of annotated operon pairs which are correctly classified and without internal control signals. POPs - percentage of operon pairs which are classified as operon pairs and without internal control signals. CEGs – percentage of co-expressed gene pairs predicted as operon pairs and with no internal control signals. The percentage of validated operon pairs are reported for each used prediction program: <pp> PromPredict, <pt> Pepper Toolbox and <tt> TransTerm.

5.5.7 Identification of condition-dependent operon maps

The proposed method, finally, determines condition-dependent operon maps through a linkage process that finds adjacent genes predicted as OPs and groups them into operons. These operons are split into three categories: (1) confirmed (2) modified and (3) putative operons. The category “confirmed” represents those operons annotated in DOOR and confirmed by the RNA-seq data analysis. The category of modified operons (or modifications to pre-existing operons) includes the identification of new structural genes for known operons and genes that are transcribed as a single transcription unit and are annotated in operons. Note that here an ‘operon’ is defined as a set of consecutive genes that are transcribed as a unit under some condition.

5.7 Conclusion

The proposed method proves that using features dependent on transcriptome dynamics and genome sequence, a NN-, a RF- or a SVM-based classification algorithm can accurately classify >96% of gene pairs in a set of operons collected by DOOR and verified with RNA-Seq experiments.

Furthermore, my results indicate the combination of DNA sequence data and expression data results in more accurate predictions than either alone. Furthermore, the trained classifiers can be used to identify new potential operons, extensions to pre-existing operons or re-define the operon status of gene pairs collected by DOOR.

An inherent limitation of using RNA-Seq to improve operon structure predictions is that it does not apply to genes not expressed under the condition studied. On the other hand, the proposed method using an RNA-Seq transcriptome profile of a prokaryotic organism (studied under an experimental growth condition), can identify condition-specific changes in the operon organization of that microorganism.

CONCLUSION

This thesis makes two contributions in the area of computational sequence analysis. My first contribution is the implementation of a new tool BLAST-based to identify distant homologies in TM domains of different membrane proteins. This computational method allows to discover a large number of short and higlly conserved TM sequences that, because of their conservation in distinct bacterial membrane proteins, could represent ancient signatures about the existence of primeval genetic elements (or mini-genes) coding for short polypeptides that formed, through a primitive assembly process, more complex genes.

The second contribution is a new operon prediction method which combines different sequence-based features to improve whole-genome prokaryotic operon map inference by using RNA-Seq data analysis and DNA sequence feature based methods. This method allows refining the operon structures annotated in DOOR, identifying new potential operons, and determining genome-wide transcription units

Those results are presented against the background of the changing technological landscape affecting life sciences and bioinformatics research and the resulting need for new computing solutions.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my my major advisor, Prof. Roberto Tagliaferri for his guidance, advice and constant support throughout my thesis work. I would like to thank him for this wonderful experience.

I am also immensely grateful to Prof. Bruno Maresca for his valuable suggestions and discussions, for his guidance and support, especially during the early learning years.

I would like to thank to Prof. Leif Peterson for giving me the opportunity to visit The Mehodist Hospital Research Institut (TMHRI), and to Prof. Stephen B. Beres for a very rewarding and enjoyable collaboration.

A special thank also to my colleague, Francesco Napolitano, for his generosity, continuous support and interest in my work and for countless illuminating discussions.

I would also like to thank all the faculty members of the Department of Pharmaceutical and Biomedical Sciences (FARMABIOMED, University of Salerno), especially Prof.ssa Antonella Leone, coordinator of this PhD course, for her guidance and valuable support. I would also like to thank all the members of the NeuroneLab and WebLab research group.

Finally, and most importantly, I would like to thank Giuliana. Her support, encouragement, quiet patience and love were undeniably. My parents, Filomena and Giovanni, receive my deepest gratitude and love for their dedication and the many years of support during my studies.

Alberts B. (1994). Molecular biology of the cell. 3rd ed. Carland Publishing.

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

Altschul S.F., Gish W. (1996). Local alignment statistics. Methods in Enzymology, 266:460-480.

Altschul S.F., Gish W., Myers E.W., Lipman D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215**:403-10.

Ansorge W.J. (2009). Next-generation DNA sequencing techniques. New Biotechnology. **25**:195-203.

Au K.F., Jiang H., Lin L., Xing Y., Wong W.H. (2010). Detection of splice junctions from paired-end RNA-Seq data by SpliceMap. Nucleic Acids Res. **38**:4570-4578.

Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S.R., Griffiths-Jones S., Howe K.L., Marshall M., Sonnhammer E.L. (2002). The Pfam protein families database. Nucleic Acids Res. **30**:276-280.

Beiko RG, Ragan MA. (2008). Detecting lateral genetic transfer: a phylogenetic approach. Methods Mol Biol. **452**:457-469.

Bemstein F.C., Koetzle T.F., Williams G.J.B., Meyer E.F., Brice M.D., Rodgers J.R., Kennard 0., Shimanouchi T., Tasumi M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. J Mol Biol **112**:535-542.

Benson D., Karsch-Mizrachi I., Lipman D., Ostell J., Wheeler D. (2005). Genbank. Nucleic Acids Res. **33**:D34-D38.

Berger M.F., Levin J.Z., Vijayendran K., Sivachenko A. (2010). Integrative analysis of the melanoma transcriptome. Genome Res. **20**(**4**):413-427.

Bockhorst J., Craven M., Page D., Shavlik J., Glasner J. (2003). A Bayesian network approach to operon prediction. Bioinformatics **19**:1227-1235.

Bogdanov M., Sun J., Kaback, H.R., Dowhan W.A. (1996). A phospholipid acts as a chaperone in assembly of a membrane transport protein. J Biol Chem. **271**:11615-11618.

Bogdanov M., Umeda M., Dowhan W. (1999). Phospholipid-assisted refolding of an integral membrane protein. Minimum structural features for phosphatidy-lethanolamine to act as a molecular chaperone. J Biol Chem. **274**:12339-12345.

Brinza L., Calevro F., Duport G., Gaget K., Gautier C., CharlesH. (2010). Structure and dynamics of the operon map of *Buchnera aphidicola* sp. strain APS.

BMC Genomics **11**:666-682.

Bernstein F.C., Koetzle T.F., Williams G.J., Meyer E.F., Jr, Brice M.D., Rodgers J.R., Kennard O., Shimanouchi T., Tasumi M. (1997). The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol. **112**(**3**):535–542.

Bywater R.P. (2009). Membrane-spanning peptides and the origin of life. J Theor Biol. **261**:407-413.

Bywater R.P., Denny-Gouldson P.R. (2006). Molecular Mechanisms of GPCR Activation, in Ligand Design for G Protein-coupled Receptors.

Bywater R.P., Conde-Frieboes K. (2005). Did life begin on the beach? Astrobiology **5**:568–574.

Carter R.J., Dubchak I., Holbrook S.R. (2001). A computational approach to identify genes for functional RNAs in genomic sequences. Nucleic Acids Res. **29**:3928-3938.

Carver T., Harris S.R., Berriman M., Parkhill J., McQuillan J.A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. **28**(**4**):464–469.

Carver T., Bohme U., Otto T. (2010). BamView: viewing mapped read alignment data in the context of the reference sequence. Bioinformatics **26**(**5**):676-677.

Charaniya S. (2007). Transcriptome dynamics-based operon prediction and verification in Streptomyces coelicolor. Nucleic Acids Res. **35**:7222-7236.

Church G.M. (2006). Genomes for all. Sci. Am. 294(1):46-54.

Chothia C., Gough J., Vogel C., and Teichmann S.A. (2003). Evolution of the protein repertoire. Science **300**:1701-1703.

Christoffels A., Koh E.G., Chia J.M., Brenner S., Aparicio S., Venkatesh B. (2004). Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol Biol Evol. **21**:1146–1151.

Cloonan N. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature Methods **5**:613-619.

Copley R.R., Russell R.B., Ponting C.P. (2001). Sialidase like Asp-boxes: sequence similar structures within different protein folds. Protein Sci. **10**:285–92.

Cortez P. (2010). Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool. Advances in Data Mining - Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining, (Berlin, Germany, July), 572-583.

Craven M., Page D., Shavlik J., Bockhorst J., Glasner J. (2000). A probabilistic learning approach to whole-genome operon prediction. Proc. Int. Conf. Intell. Syst. Mol. Biol. **8**:116-127.

de Jong A., Pietersma H., Cordes M., Kuipers O.P., Kok J. (2012). PePPER: a webserver for prediction of prokaryote promoter elements and regulons. BMC Genomics. **13**:299.

Dayhoff M.O., Eck R.V., Chang M.A., Sochard M.R. (1965). Atlas of Protein Sequence and Structure Vol. 1. National Biomedical Research Foundation, Silver Spring, MD.

David L. (2006). A high-resolution map of transcription in the yeast genome. Proc. Natl Acad. Sci. USA **103**:5320-5325.

Dam P., Olman V., Harris K., Su Z., Xu Y. (2006). Operon prediction using both genome-specific and general genomic information. Nucleic Acids Res. **35**:288-298.

De Bona F., Ossowski S., Schneeberger K., Ratsch G. (2008). Optimal spliced alignments of short sequence reads. Bioinformatics **24**:i174-180.

Deamer D.W., Dworking J.P., Sanford S.A., Bernstein M.P., Allamandola L.J. (2002). The first cell membranes. Astrobiology. **2**: 371-381.

Deamer D.W. (1997). The first living systems: a bioenergetics perspective. Microbiol. Mol. Biol. Rev. **61**:230-261.

Doherty A.J., Serpell L.C., Ponting C.P. (1996). The helix-hairpin-helix DNA-binding motif: a structural basis for non sequence-specific recognition of DNA. Nucleic Acids Res. **24**:2488-97

Doolittle R.F. (1995). The multiplicity of domains in proteins. Annu Rev Biochem. 64:287-314.

Doolittle W.F., Brown J.R. (1994). Tempo, mode, the progenote, and the universal root. Proc. Natl. Acad. Sci. USA **91**:6721-28.

Dowhan W., Bogdanov M. (2012). Molecular genetic and biochemical approaches for defining lipid-dependent membrane protein folding. Biochimica et biophysica acta. **1818**:1097-1107.

Drake J.W., Charlesworth B., Charlesworth D., Crow J.F. (1998). Rates of spontaneous mutation. Genetics **48**:1667-1686.

Dyson F. (1985). Origins of Life, Cambridge, Cambridge U.

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

Edwards M.T., Rison S.C., Stoker N.G. and Wernisch L. (2005). Universally applicable method of operon map prediction on minimally annotated genomes using con-served genomic context. Nucleic Acids Res. **33**:3253-3262.

Eisenberg D., Schwarz E., Komaromy M., Wall R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J. Mol. Biol. **179**:125-142.

Eveland A.L., McCarty D.R., KOCH K.E. (2008). Transcript profiling by 3′-untranslated region sequencing resolves expression of gene families. Plant Physiol. **146**:32–44.

Ermolaeva M.D., White O., Salzberg S.L (2001). Prediction of operons in microbial genomes. Nucleic Acids Res. **29**:1216-1221.

Ferrandez A., Mininambres B., Garcia B., Oliveira E.R., Luengo J. M., Garcia J.L., Diaz E. (1998) Catabolism of phenylacetic acid in Escherichia coli. Characterization of a new aerobic hybrid pathway. J Biol Chem. **273**:25974-25986.

Ferris J.P, Hill A.R., Liu R., Orgel L.E (1996). Synthesis of long prebiotic oligomers on mineral surfaces. Nature **381**:59-62.

Filiatrault M.J., Stodghill P.V., Bronstein P.A., Moll S., Lindeberg M., Grills G., Schweitzer P., Wang W., Schroth G.P., Luo S. (2010). Transcriptome analysis of Pseudomonas syringae identifies new genes, noncoding RNAs, and antisense activity. J Bacteriol. **192**:2359-2372.

Finn R.D., Mistry J., Tate J., Coggill P., Heger A., Pollington J.E., Gavin O.L., Gunasekaran P., Ceric G., Forslund K. (2010). The Pfam protein families database. Nucleic Acids Res. **38**:D211-222.

Fortino V., Tagliaferri R. (2012). Computational methods for predicting transcriptional units in bacteria using different data sources. BBCC2012 - Bioinformatica e Biologia Computazionale in Campania. (http://bioinformatica.isa.cnr.it/BBCC/BBCC2012/programma.html)

Fortino V., Tagliaferri R. [in preparation - 2013]. A novel method for accurate operon predictions in whole transcriptome sequencing. Bioinformatics.

Fortino V., Porta A., Maresca B. [In preparation - 2013]. Conservation analysis of membrane spanning regions in prokaryotic proteins - Are these relics of an ancestral set of peptides coded by mini genes? Mol. Biol. Evol.

Fortino V., Porta A., Tagliaferri R. (2012). Computational methods for predicting transcriptional units in bacteria using different data sources. BBCC2012 - Bioinformatica e Biologia Computazionale in Campania.

Fortino V., Porta A., Tagliaferri R., Maresca B. (2010). Transmembrane regions: a bioinformatics-approach to infer homologies. BioDN@ work10 - Computational Biology & Bioinformatics Workshop for PhD students (http://www1.na.infn.it/PhDBioinformatica/posterws2011.pdf).

Gerstein M. (1997). A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. J. Mol. Biol. **274**:562–76.

Gilbert J.A., Field D., Huang Y., Edwards R., Li W., Gilna P., Joint I. (2008). Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities PLoS One **3**(**8**): e3042.

Gilbert W. (1978). Why genes in pieces? Nature **271**:501.

Gilbert W. (1987). The exon theory of genes. Cold Spring Harb Symp Quant Biol. **52**:901-905

Gilbert W., de Souza S.J., Long M. (1997). Origin of genes. Proc Natl Acad Sci USA. **94**:7698-703.

Gokhale R.S., Sankaranarayanan R., Mohanty D. (2007). Versatility of polyketide synthases in generating metabolic diversity. Curr Opin Struct Biol. **17**:736-743.

Gribskov M., McLachlan A.D., Eisenberg D. (1987). Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA **84**:4355-4358.

Güell M., Van Noort V., Yus E., Chen W.H., Leigh-Bell J., Michalodimitrakis K., Yamada T. (2009). Transcriptome complexity in a genome-reduced bacterium. Science (New York, N.Y.) **326**:1268-1271.

Hall N. (2007). Advanced sequencing technologies and their wider impact in microbiology. J. Exp. Biol. **210**(**9**):1518-25.

Higgins D.G., Sharp P.M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene **73**: 237-244.

Holley R.W., Apoar J., Evebett G.A., Madison J.T., Maequisse M., MERRILL S.H., Pexs~icic J.R., Zamir A. (1965). Structure of a ribonucleic acid. Science. **147**:1462-1465.

Hollich V., Sonnhammer E.L.L. (2007). PfamAlyzer: Domain-centric homology search. Bioinformatics. **23**:3382-3383.

Homer N., Merriman B., Nelson S.F. (2009). BFAST: An alignment tool for large scale genome resequencing. PLoS ONE. **4**(**11**):e7767.

Hövik H., Yu W.H., Olsen I., Chen T. (2012). Comprehensive transcriptome analysis of the periodontopathogenic bacterium Porphyromonas gingivalis W83. J Bacteriol. **194**:100-114.

Hui J. and Wing H.W. (2008). SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics **24**(**20**):2395-2396.

Ikeda H. (2003). Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. Nat Biotechnol **21**(**5**):p.526-531.

Itoh T., Takemoto K., Mori H., Gojobori T. (1999). Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Mol Biol Evol. **16**:332-346.

Janga S.C., Moreno-Hagelsieb G. (2004). Conservation of adjacency as evidence of paralogous operons. Nucleic Acids Res. **32**(**18**):5392-5397.

Jacob F., J. Monod (1961). Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology **3**:p.318-356.

Jacob E., Sasikumar R., Nair K.N.R. (2005). A fuzzy guided genetic algorithm for operon prediction. Bioinformatics **21**:1403-1407.

Joyce G.F., Orgel L.E. (2006). Prospects for understanding the origin of the RNA world. In: Gesteland R.F., Cech T.R., Atkins J.F. (eds). The RNA World 3rd edn. Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY, pp 23–56.

Jones D.T., Taylor W.R., Thornton J.M. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry. **133**:3038-3049.

Kall L., Krogh A., Sonnhammer E.L. (2007). Advantages of combined transmembrane topology and signal peptide prediction: the Phobius web server. Nucleic Acids Res. **35**:W429-W432.

Kanz C., Aldebert P., Althorpe N., Bakerc W., Baldwin A., Bates, P. Browne, van den Broek A., Castro M., Cochrane G., Duggan K., Eberhardt R., Faruque N., Gamble J., Diez F.G., Harte N., Kulikova T., Lin Q., Lombard V., Lopez R., Mancuso R., McHale M., Nardone F., Silventoinen V., Sobhany S., Stoehr P., Tuli M. A., Tzouvara K., Vaughan R., Wu D., Zhu W., and Apweiler R. (2005). The EMBL nucleotide sequence database. Nucleic Acids Res. **33**:D29-D33.

Karlin S., Altschul S. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proceedings of the National Academy of Sciences USA, **87**(**6**):2264-2268.

Kauffman S.A. (1993). The Origins of Order. Self-Organization and Selection in Evolution, Oxford, Oxford U.P.

Kent W. (2002). BLAT - the BLAST-like alignment tool. Genome Res. **12**(**4**):656-664.

Kingsford C.L., Ayanbule K., Salzberg S.L. (2007). Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol. **8**:R22.

Koch A.L. (1985). Primeval cells: possible energy-generating and cell division mechanisms. J. Mol. Evol. **21**:270-277.

Kohane I.S. (2003). Microarrays for an Integrative Genomics. 2003: The MIT Press.

Koide T., Reiss D.J., Bare J.C., Pang W.L. (2009). Prevalence of transcription promoters within archaeal operons and coding sequences. Mol. Syst. Biol. **5**:285.

Kreil D., C. Ouzounis (2003). Comparison of sequence masking algorithms and the detection of biased protein sequence regions. Bioinformatics. **19**(**13**):1672-1681.

Krogh A., Larsson B., von Heijne G., Sonnhammer E. (2001). Predicting transmembrane protein topology with a hidden Markov model. Application to complete genomes. J. Mol. Biol. **305**:567-580.

Kumar R., Lawrence M.L., Watt J., Cooksey A.M., Burgess S.C. (2012). RNA-Seq Based Transcriptional Map of Bovine Respiratory Disease Pathogen Histophilus somni 2336. PLoS ONE **7**:e29435.

Langmead B., Trapnell C., Pop M., Salzberg S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology **10**:R25.

Lawrence J. G., Roth J.R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics **143**:1843-60.

Levin J.Z., Yassour M., Adiconis X., Nusbaum C., Thompson D.A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods **7**:709-715.

Laing E., Mersinias V., Smith C.P. and Hubbard S.J. (2006). Analysis of gene expression in operons of Streptomyces coelicolor. Genome biology **7**:R46.

Li H., Handsaker B., Wysoker A. (2009). The sequence alignment/map format and SAMtools. Bioinformatics **25**:2078-9.

Li H., Ruan J., Durbin R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. **18**:1851-8.

Li R., Li Y., Kristiansen K., Wang J. (2008). SOAP: short oligonucleotide alignment program. Bioinformatics **24**:713-714.

Liu J.M. Livny J., Lawrence M.S., Kimball M.D., Waldor M.K., Camilli A. (2009).Experimental discovery of sRNAs in Vibrio cholerae by direct cloning, 5S/tRNA depletion and parallel sequencing. Nucleic Acids Res. **37**:223-322.

Lister R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell **133**:523-536.

Lunter G., Goodson M., (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. **21**:936-939.

Li W., Jaroszewski L., Godzik A. (2002). Sequence clustering strategies improve remote homology recognitions while reducing search times. Protein Engineering. **15**(**8**):643-649.

Li W.H., Gu Z., Cavalcanti A.R., Nekrutenko A. (2003). Detection of gene duplications and block duplications in eukaryotic genomes. J Struct Funct Genomics **3**:27-34.

Lipman D.J., Pearson W.R. (1985). Rapid and sensitive protein similarity searches. Science **227**:1435–1441.

Liu M., Grigoriev A. (2004). Protein domains correlate strongly with exons in multiple eukaryotic genomes-evidence of exon shuffling? Trends Genet. **20**:399-403.

Luisi P.L., Rasi P.S., Mavelli F. (2004). A possible route to prebiotic vesicle reproduction. Artif Life. **10**:297-308.

Luisi P.L.The Emergence of Life: From Chemical Origins to Synthetic Biology; Cambridge University Press: Cambridge, UK, 1996.

Lundin L.G. (1999). Semin Gene duplications in early metazoan evolution. Cell Dev Biol.**10**:523-30.

Lupas A.N., Ponting C.P., Russell R.B. (2001). On the evolution of protein folds. Are similar motifs in different protein folds the result of convergence, insertion or relics of an ancient peptide world? J. Struct. Biol. **134**:191-203.

Ma B., Tromp J., Li M. (2002). PatternHunter: faster and more sensitive homology search. Bioinformatics. **18**(**3**):440-445.

Maxam A.M., Gilbert W. (1997). A new method for sequencing DNA. Proc. Natl Acad. Sci. USA. **74**:560–564.

Marioni J.C., Mason C. E., Mane S. M., Stephens M., Gilad Y. (2008). RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. **18**:1509–1517.

Mao F., Dam P., Chou J., Olman V., Xu Y. (2008). DOOR: a database for prokaryotic operons. Nucleic Acids Res. **37**:459-463.

McLysaght A., Hokamp K., Wolfe K.H. (2002). Extensive genomic duplication during early chordate evolution. Nature Genet **31**:200-204.

McGinnis S., Madden T. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. **32**:W20-W25.

Metzker ML. (2010). Sequencing technologies – the next generation. Nat RevGenet. **11**:31-46.

Mikkelsen T.S., Ku M., Jaffe D.B., Issac B., Lieberman E., Giannoukos G., Alvarez P., Brockman W., Kim T.K., Koche R.P. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature **448**: 553-560.

Miller S.L., Urey H.C. (1959). Organic compound synthesis on the primitive Earth. Science **130**:245-251.

Miyazaki S., Sugawara H., Ikeo K., Gojobori T., Tateno Y. (2004). DDBJ in the stream of various biological data. Nucleic Acids Res. **32**:D31-D34.

Monnard P.A., Deamer D. (2002). Membrane self-assembly processes: steps toward the first cellular life. Anatomical Record. **268**:196-207.

Moreno-Hagelsieb G., Collado-Vides J. (2002). A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics **18**:329-36.

Morowitz H.J., Kostelnik J.D., Yang J., Cody G.D. (2000). The origin of intermediary metabolism. Proc Natl Acad Sci USA. **97**:7704-7704.

Morowitz H.J. (1992). Beginnings of Cellular Life, Yale University Press, New Haven, CT.

Mortazavi A., Williams B. A., McCue K., Schaeffer L., Wold B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods **5**: 621-628.

Muller A., MacCallum R. M., Sternberg M. J. (2002). Structural characterization of the human proteome. Genome Res. **12**:1625-1641.

Murzin A.G., Brenner S.E. Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247**:536-540.

Murzin A.G. (1992). Structural principles for the propeller assembly of β-sheets: the preference for seven-fold symmetry. Proteins **14**:191-201.

Nagalakshmi U., Wang Z., Waern K., Shou C., Raha D., Gerstein M., Snyder M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science **320**:1344-1349.

Needleman S.B., Wunsch C.D. (1970). A general method applicable to the search for similarities in the aminoacid sequence of two proteins. J. Mol. Biol. **48**: 443–453.

New M.H., Pohorille A. (2000). An Inherited Efficiencies Model of Non-Genomic Evolution, Simulation Practice Theory. **8**:99-108.

Nirenberg M.W. (1966). The RNA code and protein synthesis. Cold Spring Harbor Symposia on Quantitative Biology. **31**:11-24.

Ohno S. (1970). Evolution by gene duplication. New York, Springer-Verlag, Berlino 1970.

Oliver H.F., Orsi R.H., Ponnala L., Keich U., Wang W., Sun Q., Cartinhour S.W., Filiatrault M.J., Wiedmann M., Boor K.J. (2009). Deep RNA sequencing of L. monocytogenes reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs. Biomed Central (BMC) Genomics **10**:641.

Orgel L.E. (2000). Self-organizing biochemical cycles. Proc Natl Acad Sci USA. **97**:12503-12507.

Orgel L.E. (2008). The implausibility of metabolic cycles on the prebiotic Earth. PLoS Biol **6**, e18.

Pal C., Hurst L.D. (2004). Evidence against the selfish operon theory. Trends Genet **20**:232-234.

Paredes C.J., Rigoutsos I., Papoutsaki, E.T. (2004). Transcriptional organization of the Clostridium acetobutylicum genome. Nucleic Acids Res. **32**:1973-1981.

Parker E.T., Cleaves H.J., Dworkin J.P., Glavin D.P., Callahan M., Aubrey A., Lazcano A., Bada J.L. (2011). Primordial synthesis of amines and aminoacids in a 1958 Miller H2S-rich spark discharge experiment. Proc Natl Acad Sci USA. **108**:5526-5531.

Passalacqua K.D., Varadarajan A., Ondov B.D., Okou D.T., Zwick M.E., Bergman N.H. (2009). The structure and complexity of a bacterial transcriptome. J Bacteriol. **191**:3203-3211.

Patthy L. (2003). Modular assembly of genes and the evolution of new functions. Genetica. **118**:217-31.

Perkins T.T., Kingsley R.A., Fookes M.C., Gardner P.P., James K.D., Yu L., Assefa S.A., He M., Croucher N.J., Pickard D.J. (2009). A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhiPLoS Genet. **5**:e1000569.

Pearson W.R., Lipman D.J. (1988). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. **85**:2444-2448.

Ponting C.P., Russell R. (2002). The natural history of protein domains. Annu Rev Biophys Biomol Struct. **31**:45-71.

Ponting C.P., Russell R. (2000). Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all β-trefoil proteins. J. Mol. Biol. **302**:1041-47.

Price M.N., Huang K.H., Alm E.J., Arkin, A.P. (2005). A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res. **33**:880-

892.

Pross A. (2004). Causation and the origin of life. Metabolism or replication first? Orig Life Evol Biosph **34**:307-321.

Ranea J.A., Buchan D.W., Thornton J.M., Orengo C.A. (2004). Evolution of protein superfamilies and bacterial genome size. J. Mol. Biol. **336**:871-887.

Rangannan V, Bansal M. (2011). PromBase: a web resource for various genomic features and predicted promoters in prokaryotic genomes. BMC Res Notes. **4**:257-268.

Rangannan V., Bansal M. (2010). High-quality annotation of promoter regions for 913 bacterial genomes. Bioinformatics **26**:3043-3050.

Rizk G., Lavenier D. (2010). GASSST: Global Alignment Short Sequence Search Tool, Bioinformatics Advance Access published online on August 24, 2010.

Robertson G., Hirst M., Bainbridge M., Bilenky M., Zhao Y., Zeng T., Euskirchen G., Bernier B., Varhol R., Delaney A. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Methods **4**:651-657.

Roberts A., Trapnell C., Donaghey J., Rinn J.L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. **12**:R22.

Robertson M.P., Miller S.L. (1995). An efficient prebiotic synthesis of cytosine and uracil. Nature. **375**:772-774.

Rogozin I.B., Makarova K.S., Murvai J., Czabarka E., Wolf Y.I. (2002). Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res. **30**:2212-2223.

Romero P.R., Karp P.D. (2004). Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases. Bioinformatics **20**:709-717.

Rossmann M.G., Moras D., Olsen K.W. (1974). Chemical and biological evolution of nucleotidebinding protein. Nature. **250**:194-199.

Rumble S.M. (2009). SHRiMP: accurate mapping of short color-space reads. PLoS Comput. Biol. 2009. **5**:e1000386.

Sabatti C. (2002). Co-expression pattern from DNA microarray experiments as a tool for operon prediction. Nucleic Acids Res. **30**:2886-2893.

Salgado H. (2000). Operons in Escherichia coli: genomic analyses and predictions. Proc Natl Acad Sci USA **97**:6652-6657.

Salgado H., Moreno-Hagelsieb G., Smith T.F., Collado-Vides J. (2000).

Operons in Escherichia coli: genomic analyses and predictions. Proc Natl Acad Sci USA **97**:6652-6657.

Sanger F., Nicklen S., Coulson A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl Acad. Sci. USA. **74**:5463–5467.

Sanger F., Air G.M., Barrell B.G., Brown N.L., Coulson A.R., Fiddes C.A., Hutchison C.A., Slocombe P.M., Smith M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. Nature. **265**:687–695.

Sanger F., Coulson A.R., Friedmann T., Air G.M., Barrell B.G., Brown N.L., Fiddes J.C., Hutchison C.A., III, Slocombe P.M. (1978). The nucleotide sequence of bacteriophage phiX174. J. Mol. Biol. **125**:225–246.

Segré D., Lancet D. (2000). Composing Life, EMBO Reports. **1(3)**:217-222.

Segré D., Ben-Eli D., Deamer D.W., Lancet D. (2001). The lipid world. Origins of Life and Evolution of the Biosphere. **31(1-2)**:119-45.

Segré D. (2000). Compositional Genomes: Prebiotic Information Transfer in Mutually Catalytic Non covalent Assemblies, Proc. Nat. Acad. Sci. USA. **97**:4112–7.

Sharma C.M., Hoffmann S., Darfeuille F., Reignier J., Findeiss S., Sittka A., Chabas S., Reiche K., Hackermuller J., Reinhardt R. (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. Nature. **464**:250-255.

Shenhav B., Segre D., Lancet D. (2003). Mesobiotic emergence: Molecular and ensemble complexity in early evolution. Adv. Complex Syst. **6**:15-35.

Schmidt E.E., Davies C.J. (2007). The origin of polypeptide domains BioEssays **29**:262-270.

Schulze A., Downward J. (2001). Navigating gene expression using microarrays—a technology review. Nature cell biology **3**(**8**):E190-195.

Schuster P. (2000). Taming combinatorial explosion. Proc Natl Acad Sci USA. **97**:7678-7680.

Shi Y., Tyson G.W., De Long E.F. (2009).Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. Nature. **459**(7244):266-269.

Sinsheimer R.L. (1959). A single-stranded DNA from bacteriophage phi X174. J. Mol. Biol. **1**:43.

Smith A.D., Xuan Z., Zhang M.Q.(2008). Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics. **9**:128.

Smith E., Morowitz H.J. (2004). Universality in intermediary metabolism. Proc Natl Acad Sci USA. **101**:13168-13173.

Smith T.F., Waterman M.S (1981). Identification of common molecular subsequencees. J. Mol. Biol. **147**:195-197.

Sorek R., Cossart P. (2010). Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nat. Rev. Genet. **11**:9-16.

Sowerby S.J., Petersen G.B., Holm N.G. (2002). Primordial coding of aminoacids by adsorbed purine bases. Orig. Life Evol. Biosph. **32**:35-46.

Stribling R., Miller S.L. (1991). Template-directed synthesis of oligonucleotides under eutectic conditions. J.Mol. Evol. 32, 289–295.

Taboada B., Verde C., Merino E. (2010). High accuracy operon prediction method based on STRING database scores. Nucleic Acids Res. **38**:e130.

Tang F., Lao K., Surani M.A. (2011). Development and applications of single-cell transcriptome analysis. Nature methods **8**(**4**):S6-11.

Tjaden B., Haynor D.R., Stolyar S., Rosenow C., Kolker E. (2002). Identifying operons and untranslated regions of transcripts using Escherichia coli RNA expression analysis. Nat Methods **18**:S337-S344.

Tran T.T., Dam P., Su Z., Poole F. L., Adams M.W., Zhou G.T., Xu Y. (2007).Operon prediction in Pyrococcus furiosus. Nucleic Acids Res. **35**:11-20.

Trapnell C., Pachter L., Salzberg S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. **25**:1105-1111.

van Vliet A.H. (2010). Next generation sequencing of microbial transcriptomes: challenges and opportunities. FEMS Microbiol. Lett. **302**:1-7.

Vapnik V.N., (1995). The Nature of Statistical Learning Theory Springer, New York.

Vapnik V.N. (1998). Statistical Learning Theory Wiley, New York.

Vogel C. (2005). The relationship between domain duplication and recombination. J Mol Biol. **346(1)**:355–365.

Vogel C., Teichmann S.A., Chothia C. (2003). The immunoglobulin superfamily in Drosophila melanogaster and Caenorhabditis elegans and the evolution of complexity. Development. **130**:6317-6328.

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R..A, Gocayne J.D., Amanatides P., Ballew R.M., Huson D.H., Wortman J.R., Zhang Q., Kodira C.D., Zheng X.H., Chen L., Skupski M., Subramanian G., Thomas P.D., Zhang J., Gabor Miklos G.L., Nelson C., Broder S., Clark A.G., Nadeau J., McKusick V.A., Zinder N., et al: The sequence of the human genome. Science **291**:1304-1351.

Wang K., Singh D., Zeng Z., Coleman S.J.,Huang Y. (2010). MapSplice: Accurate mapping of RNA-Seq reads for splice junction discovery. Nucleic Acids Res. **38**(**18**):e178.

Wang Z., Gerstein M. Snyder M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. **10**:57-63.

Wächtershäuser G. (1988). Before enzymes and templates: theory of surface metabolism. Microbiol Rev. **52**:452–484.

Wächtershäuser G. (1990). Evolution of the first metabolic cycles. Proc Natl Acad Sci. **87**:200–204.

Wächtershäuser G. (1992). Groundworks for an Evolutionary Biochemistry: The Iron Sulfur World, Prog. Biophys. Mol. Biol. **58**:85-201.

Waters E. (2003). The genome of Nanoarchaeum equitans: Insights into early archaeal evolution and derived parasitism. PNAS **100**(**22**):12984-12988.

Westover B.P. (2005) .Operon prediction without a training set. Bioinformatics, **21**(**7**):880-888.

Wilhelm B.T. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature, **453**:1239-1243.

Wolf Y.I., Brenner S.E., Bash P.A., Koonin E.V. (1999). Distribution of protein folds in the three super kingdoms of life. Genome Res. **9**:17-26.

Wolfe K.H., Shields D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. Nature. **87**:708-713.

Wu C.H., Apweiler R., Bairoch A. (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res., **34**:D187-191.

Wu T.D., Nacu S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. **26**:873-881.

Wurtzel O., Sapra R., Chen F., Zhu Y., Simmons B.A., Sorek R. (2010). A single-base resolution map of an archaeal transcriptome. Genome Res. **20**:133-141.

Yada T. (1999). Modeling and predicting transcriptional units of Escherichia coli genes using hidden Markov models. Bioinformatics **15**:987-993.

Yan B., Methe B.A., Lovley D.R., Krushkal J. (2004). Computational prediction of conserved operons and phylogenetic footprinting of transcription regulatory elements in the metal-reducing bacterial family Geobacteraceae. J. Theor. Biol. **230**:133-144.

Yoder-Himes D.R., Chain P.S., Zhu Y., Wurtzel O., Rubin E.M., Tiedje J.M., Sorek R. (2009). Mapping the Burkholderia cenocepacia niche response via highthroughput sequencing. Proc Natl Acad Sci. **106**:3976-3981.

Zhang G.Q., Cao Z.W., Luo Q.M., Cai Y.D., and Li Y.X. (2006) .Operon prediction based on SVM. Comput. Biol Chem. **30**:233-240.