

UNIVERSITAT ROVIRA I VIRGLI
COMPUTATIONAL INSIGHTS INTO INTERGENIC REGIONS AND OVERLAPPING GENES AMONG PROKARYOTE GENOMES
Albert Pallejà Caro
ISBN:978-84-692-2150-1/DL:T-508-2009



DEPARTAMENT DE BIOQUÍMICA I BIOTECNOLOGIA

FACULTAT DE QUÍMICA

**COMPUTACIONAL INSIGHTS INTO INTERGENIC
REGIONS AND OVERLAPPING GENES AMONG
PROKARYOTE GENOMES**

Memòria presentada per optar al Grau de
Doctor per la Universitat Rovira i Virgili
amb menció europea.

Vist i plau del Director de Tesi:

Vist i plau de l'alumne:

DR ANTONI ROMEU FIGUEROLA

ALBERT PALLEJÀ CARO

Tarragona, 10 de Desembre del 2008

UNIVERSITAT ROVIRA I VIRGILI
COMPUTATIONAL INSIGHTS INTO INTERGENIC REGIONS AND OVERLAPPING GENES AMONG PROKARYOTE GENOMES
Albert Pallejà Caro
ISBN:978-84-692-2150-1/DL:T-508-2009

ALS MEUS PARES

CONTENTS

PREFACE	1
BACKGROUND AND OBJECTIVES	5
<hr/>	
CHAPTER 1: <i>IN SILICO</i> PREDICTION OF THE ORIGIN OF REPLICATION AMONG BACTERIA: A CASE STUDY OF <i>BACTEROIDES THETAIOOTAOMICRON</i>	21
CHAPTER 2: OVERLAPPING GENE STRUCTURES AMONG PROKARYOTE GENOMES	49
CHAPTER 3: LARGE GENE OVERLAPS IN PROKARYOTIC GENOMES: RESULT OF FUNCTIONAL CONSTRAINTS OR MISPREDICTIONS?	80
CHAPTER 4: ADAPTATION OF THE SHORT CO-DIRECTIONAL SPACERS TO THE SHINE-DALGARNO MOTIF IN PROKARYOTE GENOMES	105
CHAPTER 5: PAIRWISE NEIGHBOURS DATABASE: OVERLAPS AND SPACERS AMONG PROKARYOTE GENOMES	136
<hr/>	
CONCLUSIONS	160
LIST OF PUBLICATIONS AND CONGRESS CONTRIBUTIONS	165
AGRAÏMENTS / ACKNOWLEDGEMENTS	169

ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGACAGCTGATAGG
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGGCGCTGGTCAGGCAGCGCAGCATGGGAAAA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
GATCAGATGACAGATAGATAACCACAGAGACATG
CGCGCGGACGCATGATTGATGATCAGATGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGATGATAAAGTGATTAGATAGATGGTGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
GAGAGATAGAGAGATAGGATCGCGCTCGAGCGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AACCGGCCAACCGGTGGCTTAGGATAGATGATGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AAAGTGTGTGTGACAGACAGTTGATGATAGTACA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGACAGCTGATAGC
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CGCGCGCTCGCTCGAGCGCTAGCTCGATCGATCGA
TCGCGCTCAAACGAGCGCTAGCTCGATCGATCC
TCGATAGGTACGCGAAAATGGCAGTAGCTAGCTAG
TCGATAGGTACGCGGATGAATGGCAGTAGCTAGCT
TCGATCGATCGATCGATCGATCGCGCGA

PREFACE

TCAGCTGGGTGGTAGGACG
TCGATCGATCGATCGATCG
TCGACAGACAGTTGA



TCAGCGAAAATGGCA
TCGCTCGAGCGCTAGCTCG,
TCGTACGCGAAAATGGCAGT,
TCGATCGATCGATCGATCGCG,
TCAGCATGACACACACATGAT,
TCGTGCCAGGCAGCATAAAGCAGAC,
TCAGCAGCTGGGTGGTAGGAGTGATG
TCAGTGCCAGGCAGCATAAAGCAGACGA
TCACCAGCAGCTGGGTGGTAGGAGTGATGTA

In the last few decades, bioinformatics has become an important part of research and development in the biomedical sciences [1]. Bioinformatics is the application of tools of computation and analysis to the capture and interpretation of biological data and has become essential for the management of data in modern biology and medicine [2]. In the field of genomics or proteomics, bioinformatics make it possible to connect all the different data formats gathered by new high-throughput techniques such as systematic sequencing, proteomics, expression arrays, yeast two-hybrid, and high throughput screenings [3]. Laboratories are employing local bioinformatics to study fundamental biological questions. The contribution of bioinformatics is related to the development of concepts in theoretical molecular biology, but also to the management and representation of complex biological information.

The exponential growth in molecular sequence data started in the early 1980s when methods for DNA sequencing became widely available. A novel strategy for random sequencing of the whole genome, the “shotgun technique”, was used to sequence the bacterial genome of *Haemophilus influenzae* in 1995 [4]. This was the first genome of any free living organism to be sequenced. Soon after, other bacterial genomes were fully sequenced such as *Mycoplasma genitalium* [5] and *Mycobacterium tuberculosis* [6]. The sequence and annotation of the first eukaryotic genome was in 1996, which was the genome of the yeast *Sacharomyces cerevisiae* [7]. After the initial period of irregular growth of sequenced genomes, the accumulation of fully sequenced genomes of bacteria and archaea showed a remarkably good fit to exponential functions, with a doubling time of 20 months for bacteria and 34 months for archaea [8]. On the other hand, the fully sequenced eukaryotic genomes have grown slower due to their larger extension. However in 2001, the whole human genome was sequenced as a result of the great efforts made by the worldwide human genome project and a private genomic company [9,10]. This may be the biggest scientific news, and a great achievement for the people working in

bioinformatics, since the discovery of the double helix in DNA by Watson and Crick [11]. Apart from all these sequencing projects mentioned above, many other organisms have been completely sequenced. Handling this massive amount of data requires powerful integrated bioinformatic systems. Therefore, methods for the design, management and interpretation of the results are required.

From the 80s the data were accumulated in databases such as GenBank, EMBL (European Molecular Biology Laboratory nucleotide sequence database), DDBJ (DNA Data Bank of Japan), PIR (Protein Information Resource and SwissProt. Computational methods were developed for data retrieval and analysis, including important algorithms for sequence similarity searches such as BLAST [12] (based on mathematical statistics coupled with human intuition), structural predictions and functional predictions. Bioinformatics in the 90s was focused on the understanding of functions and utilities of individual genes or proteins. Later Bioinformatics was dedicated to understanding functions and utilities at the molecular, cellular and organism levels. Nowadays Bioinformatics is trying to understand the basic principles of the higher complexity of biological systems [1].

After finishing my degrees in Chemistry (2003) and Biochemistry (2004), I started my thesis in the Bioinformatic field as a novice. Bioinformatics was born as a tool to manage huge amounts of data and it became a wide research field on its own. Nowadays a large number of Bioinformatics labs are working around the world. People have joined this field from different disciplines such as Biology, Chemistry, Mathematics or Statistics. Even now there are several master degrees available to complete the knowledge of the scientists that are working in this field. A background in computer studies and life sciences is desirable, but also a fair amount of knowledge of statistics and mathematical calculations is important. *"A person needs to be a jack of all trades, and then he or she can become master of bioinformatics,"* (Prof. K Kannan, dean, School of

Biotechnology, Guru Gobind Singh Indraprastha University). I wanted to comment this preface on the complex learning process to adapt my chemistry and biochemistry background to the Bioinformatics field. This process has been really long and I have to admit that I am still learning a lot every day. Learning how to work in a linux system, learning perl programming language, learning mySQL language and dealing with the huge amount of databases that are now available on Internet have been some of my goals in terms of training. In fact, after all these years spent doing my doctorate, I am ready to confess that I have only just started. When I started my thesis I knew a bit, but now I realize how much I still do not know.

"Life is a long lesson of humility" James Mathew Barrie (Scottish Dramatist and Novelist best known as the creator of Peter Pan, 1860-1937).

In my opinion, a research thesis must not be reduced to the number of papers that you have been able to write. For me the thesis is the sum of competences that you have acquired during your thesis period. In the beginning you do not know what to ask and which questions are scientifically important? In the end you become an autonomous scientist that is able to do experiments on his own and with a critical scientific point of view. In these five years I have grown scientifically and personally. Otherwise, I would not be satisfied right now. During these years I frequently got lost, but I was strong enough and I had the willpower enough to carry on my research. This has helped me learn not only about scientific concepts, but also about myself and life in general.

"Learn from yesterday, live for today, hope for tomorrow. The important thing is not to stop questioning." Albert Einstein (German Physic, 1879- 1955).

REFERENCES

1. Kanehisa M, Bork P (2003) Bioinformatics in the post-sequence era. *Nat Genet* 33 Suppl: 305-310.
2. Bayat A (2002) Science, medicine, and the future: Bioinformatics. *BMJ* 324: 1018-1022.
3. Valencia A (2002) Bioinformatics: biology by other means. *Bioinformatics* 18: 1551-1552.
4. Fleischmann R, Adams M, White O, Clayton R, Kirkness E, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496-512.
5. Fraser C, Gocayne J, White O, Adams M, Clayton R, et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270: 397-403.
6. Cole S, Brosch R, Parkhill J, Garnier T, Churcher C, et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393: 537-544.
7. Goffeau A, Barrell B, Bussey H, Davis R, Dujon B, et al. (1996) Life with 6000 genes. *Science* 274: 546, 563-547.
8. Koonin E, Wolf Y (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.*
9. Lander E, Linton L, Birren B, Nusbaum C, Zody M, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
10. Venter J, Adams M, Myers E, Li P, Mural R, et al. (2001) The sequence of the human genome. *Science* 291: 1304-1351.
11. Watson J, Crick F (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171: 737-738.
12. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403-410.

ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGACAGCTGATAGG
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGGCGCTGGTCAGGCAGCGCAGCATGGGAAAA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
GATCAGATGACAGATAGATAACCACAGAGACATG
CGCGCGGACGCATGATTGATGATCAGATGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGATGATAAAGTGATTAGATAGATGGTGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTTT
GAGAGATAGAGAGATAGGATCGCGCTCGAGCGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AACCGCCAACCGGTGGCTTAGGATAGATGATGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AAAGTGTGTGTGACAGACAGTTGATGATAGTACA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGACAGCTGATAGC
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CGCGCGCTCGCTCGAGCGCTAGCTCGATCGATCGA
CGCGCGCTCAAACGAGCGCTAGCTCGATCGATCC
CGATAGGTACGCGAAAATGGCAGTAGCTAGCTAG
CGATAGGTACGCGGATGAATGGCAGTAGCTAGCT
GATCGATCGATCGATCGATCGCGCGGAT
TCAGCATGACACACACACATGAT
TGCCAGGCAGCATAAAGCAGAC
TCAGCTGGCTGGTAGGACG
GATCGATCGATCGATCG
TCAGCTGGCTGGTAGGATGTA



BACKGROUND AND OBJECTIVES

ACGCGAAAATGGCA
JCTCGAGCGCTAGCTCG,
GTACGCGAAAATGGCAGT,
CGATCGATCGATCGATCGCGC
JAGCATGACACACACACATGAT,
JTGCCAGGCAGCATAAAGCAGACC
JGAGCAGCTGGGTGGTAGGAGTGATG,
JAGTGCCAGGCAGCATAAAGCAGACGAC
JACCAGCAGCTGGGTGGTAGGAGTGATGTA,

This is a computational thesis which has been developed in the Biochemistry and Biotechnology Department at the *Rovira i Virgili* University. The director of this thesis has been Dr Antoni Romeu who is Professor of the University and Head of research in the Biochemistry and Biotechnology Department of the *Rovira i Virgili* University. Regarding me, I am a PhD student in the same Department of the same University. I have degrees in Chemistry (2004) and in Biochemistry (2005). During my Doctorate period I finished The Nutrition and Metabolism Doctorate program (from 2003 to 2005). In addition, to complete my formation in the Bioinformatics field I did some courses about Perl programming language, Phylogenetics, MySQL language, Shell Scripting and Linux Systems.

Our research group has been involved in the following research projects:

- ***“Developing Bioinformatic Tools for the Characterization of Prokaryote Genomes (Desarrollo de herramientas bioinformáticas para la caracterización de genomas de procariontes)”*** funded by Ministerio de Ciencia y tecnología (BIO2003-07672)
- ***“Bioinformatic characterization of bacterial origin and terminus of replication and its impact in genome evolution”*** funded by Fundación BBVA (Ref.: BIO 04)
- ***“Estudio de la posible regulación epigenética e impronta de genes humanos sometidos a un regulación por la dieta”*** funded by Ministerio de Educación y Ciencia (AGL2007-65678/ALI).

From all these projects I have mainly been working on the characterization of the prokaryote genomes. Concretely, I focused in the characterization of the Origin of replication, in the study of the intergenic regions

between genes and in the analysis of the overlapping genes among prokaryote genomes.

SIZE AND OVERALL ORGANIZATION OF THE PROKARYOTE GENOMES

The first bacterial genome was sequenced in 1995. Concretely it was the pathogenic bacterium *Haemophilus influenzae* [1]. After the initial period of irregular growth, the accumulation of sequenced genomes of bacteria and archaea showed a remarkably good fit to exponential functions [2,3]. In the middle of the next year (2009), probably, we will be reaching the 1,000 fully sequenced genomes. Comparative analysis of the hundreds of sequenced bacterial and dozens of archaeal genomes leads to several generalizations on the principles of genome organization and evolution. Although the huge variety of life styles, as well as metabolic and genomic complexity, bacterial and archaeal genomes show common architectural principles [2]. In terms of genome sizes and overall genome organization, bacteria do not qualitatively differ from archaea, although the currently characterized archaea have smaller and compact genomes. Bacteria show a clear-cut bimodal distribution of genome sizes, with the highest peak at ~2 Mb and the second, smaller one at ~5 Mb [2]. Archaea are less diverse in genome size, from ~0.5 Mb in the parasite *Nanoarchaeum equitans* to ~5.8 Mb in the acetate-utilizing methanogen *Methanosarcina acetivorans*, and show a sharp peak at ~2 Mb that coincides with the highest bacterial peak, even though there are larger archaeal genomes [2]. Comparing with eukaryotes, prokaryotes accommodate a rather narrow range of variation in genome size [4]. Whereas eukaryote genomes vary in size about four orders of magnitude, there is only one order of magnitude difference across prokaryote genome sizes [5]. However, the difference in the ranges of genome size in eukaryotes and prokaryotes is not reflected in corresponding differences in gene number. The genome size variation in prokaryotes is almost directly proportional to the biochemical, physiological and organismal complexity [4]. For instance, *Mycoplasma genitalium* has 525 genes along its 580,076

nucleotides [6], while *Bacillus subtilis* has 4225 genes along its 4,214,630 nucleotides [7]. In contrast, yeast and humans have genomes that differ almost 300-fold in size, although they have only a six-fold difference in gene content [8,9,10].

INTERGENIC LENGTHS IN PROKARYOTE GENOMES

According to the genomic compactness that suffer the prokaryote genomes, they have intergenic distances that are much shorter than the gene lengths and are relatively short compared to those in eukaryote genomes [2]. The eukaryote genomes show a much larger range of genome sizes and contain protein-coding genes that are typically, interrupted by introns, and have longer intergenic regions. In contrast, prokaryote genomes are considered wall-to-wall genomes, which consist largely of genes for proteins and structural RNAs, with only a small fraction of the genomic DNA containing intergenic regions, which are thought to typically contain regulatory signals [11]. There are variations in percentage of non-coding DNA among the prokaryote genomes. These variations do not depend on the genome size or the gene content, whereas the latter variables strongly correlate [4]. The spacers between a pair of genes were classified into three types according to their transcriptional direction: i) unidirectional, ii) convergent and iii) divergent [11]. Here we decided to use co-directional instead of unidirectional. These three classes of spacers differ in the type of regulatory signals that they contain. The co-directional spacers may contain an upstream gene terminator, a promoter and an operator for a downstream gene. The convergent spacers may contain terminators for both genes while the divergent ones have only promoters and other upstream transcriptional signals. The different types of intergenic regions in prokaryotes, including the convergent and divergent ones (all of them inter-operonic) and the co-directional (largely intra-operonic), evolve under the same evolutionary pressures. The principal evolutionary force is the selective pressure to minimize the amount of non-functional DNA [11]. However, in prokaryotes, these

- BACKGROUND AND OBJECTIVES -

intergenic regions must maintain a minimal extension to accommodate essential regulatory signals [11] and the DNA replication sequences [12]. Therefore, in general, short spacers between prokaryote genes are expected. However, there are long intergenic distances between the transcription units due to special regulation requirements, extensive movements of mobile elements or active pseudogene formation. The pseudogenes in prokaryote genomes undergo processes such as niche change, host specialization or weak selection strength [13]. This process is extremely clear in certain intracellular parasitic bacteria, such as *Mycobacterium leprae* and *Rickettsia*, which appear to be in the process of extensive genome degradation via pseudogenization [14].

ORIGIN OF REPLICATION IN PROKARYOTES

As has been commented above, the origin and terminus of replication usually is in the intergenic regions between the genes. The initiation of chromosomal replication occurs only once during the prokaryote cell cycle. Most of the prokaryote genomes contain a single, bidirectional replication origin site [15]. This origin can affect the global genome architecture [16]. The bidirectional origin is the switch point between the leading and lagging strand that in prokaryotes are replicated in different modes, continuous and discontinuous, respectively. The leading and lagging strands show substantial asymmetries in nucleotide composition, gene orientation and gene content [17,18]. E. Chargaff experimentally determined the approximated equimolarities $A \sim T$ and $C \sim G$ for long, single stranded DNA molecules [19]. These equalities are observed in the lack of bias between the two DNA strands for mutation and selection [20,21]. However, in prokaryotic genomes there are local and systematic deviations for many reasons, even though the main reason appears to be the different mutational pressure associated with the different mechanisms for replication between leading and lagging strands [22,23,24,25]. This asymmetry generally divides the chromosome into two regions with opposite signs for base composition skews [26]. The skew parameter is the difference in base

- BACKGROUND AND OBJECTIVES -

composition between leading and lagging strands. The switch in skew direction generally occurs at the origin and terminus of replication [17,26,27,28,29]. Usually the GC or AT skew patterns in prokaryotes chromosomes are significant enough to make a good prediction of the origin of replication. Actually the origins of the uncharacterized genomes are often detected using the DNA compositional asymmetry [27]. The leading and lagging strands also show asymmetric distributions of genes, with greater density of genes in the leading strand [30]. In addition, it is known that highly expressed genes such as ribosomal proteins-coding genes or essential ones, are overrepresented in the leading strand while alien genes are encoded mainly in the lagging strand [31,32]. The initiation and end of the chromosomal replication seem to be really important factor that determine the genome architecture. Therefore the predictions of these sites must be very accurate.

OVERLAPPING GENES AMONG PROKARYOTE GENOMES

Another consistent phenomenon that is found in prokaryote genomes is the gene overlaps [33]. Overlapping genes were originally discovered in viruses, mitochondria and other extra chromosomal nuclear elements [34,35]. Nowadays, there are thousands of examples of overlapping genes in bacteriophages, animal viruses and mitochondria genomes, as well as in all bacterial and archaeal genomes sequenced to date [33,34,35,36,37]. Overlapping genes have been classified into three types according to their transcriptional direction equally to the spacers between genes [36,38,39,40]. These are: i) co-directional (genes in the same strand overlapping an upstream gene 3'-end and a downstream gene 5'-end), ii) convergent (genes in opposite strands overlapping the 3'-ends) and iii) divergent (genes in opposite strands overlapping the 5'-ends) In prokaryotes, these overlaps have been hypothesized to be involved (i) in compressing the maximum amount of genetic information as a result of the evolutionary pressure to minimize genome size and increase the density of genetic information [36,41,42,43,44] and (ii) to be a mechanism for

- BACKGROUND AND OBJECTIVES -

regulating gene expression through translational coupling of functionally related polypeptides [33,36,44,45,46]. The co-directional and convergent overlaps can arise because of the loss of a stop codon in either gene, resulting in the elongation of the 3'-end of the gene's coding region. More specifically, the loss of a stop codon may result of one from the following events: i) deletion of the stop codon, ii) point mutation at the stop codon or iii) frameshift at the end of the coding region [38]. The co-directional and divergent overlaps can simply arise when the downstream gene adopts a new start codon within the upstream coding sequence [47]. Genes follow the rules imposed by the genetic code to overlap. Overlaps of one and four nucleotides are extremely common [33,39,44,47], especially the 4 bps co-directional overlap which includes, in the overlapping region, the start and the stop codon of both genes favoring translational coupling [48]. The overlapping lengths tend to be short due to the selective pressure against long overlaps. As longer is the overlap as higher is the risk that a deleterious mutation can affect both genes. Because of such mutation the cell could lose two proteins at the same time.

SHINE-DALGARNO SEQUENCE IN PROKARYOTES

Under this scenario of short spacers between genes and genes overlapping, the regulatory signals may be compromised. One of such regulatory signals is the Shine-Dalgarno (SD) sequence, which was discovered by Shine and Dalgarno on 1974 [49], and plays a key role in the translation initiation. Usually this SD sequence (5'-GGAGGU-3') is found at the 5' UTR regions and binds a motif sequence (3'-CCUCCA-5') at the 16S rRNA tail [49]. The complementarity between the 3' tail of 16S rRNA and the region 5' of the start codon on the mRNA is enough to create a stable, double-stranded structure that could position the ribosome correctly on the mRNA during translation initiation. The motif 5'-GGAGGU-3' and variations on it, which are also complementary to parts of the 3' 16S rRNA tail, have been referred as SD sequence. This sequence was experimentally verified on 1987 by Hui and de

- BACKGROUND AND OBJECTIVES -

Boer [50] and Jacob and co-workers [51]. Since Shine Dalgarn's publication two methods have been used to identify and locate the SD sequence in prokaryotes: sequence similarity and free energy calculations. Methods based on sequence similarity include searching upstream from start codons for sub-strings of the SD sequence of at least three nucleotides long [52]. However, there is not a threshold of similarity that can clearly separate actual SD sequences from spurious sites with a significant, but low, degree of similarity to the SD sequence. The lack of certainty leads to separate the genes into two categories: those with obvious SD sequence and those without. The inability of sequence techniques to pinpoint the exact location of the SD sequence is a problem because the SD location is believed to affect translation initiation [53,54,55,56]. The SD motif is mostly found between 7th and 12th base upstream of the start codons [55,57,58]. The free energy calculations method is based on thermodynamic considerations of the proposed mechanism of 30S binding to the mRNA and overcomes the limitations of sequence analysis. Here we used the Starmer and co-workers method that to identify SD sequences calculating the ΔG° values for progressive alignments of the rRNA tail with the mRNA in the region around the translation initiation (upstream and downstream of the start codon) [59]. The free energy calculations approaches can both identify the SD and pinpoint its exact location as that having the minimal ΔG° value. However, recently, more leaderless genes (short leader genes) or genes without SD sequence have been detected among prokaryote genomes [60]. Three different structures can be identified based on the existence of SD and leading sequence: the genes led by SD, genes not led by SD and leaderless genes [61]. Several studies have examined SD sequence dispersion in various prokaryotes including bacteria and archaea genomes [55,58,59,61]. Despite SD sequence is widely found upstream of the bacteria and archaea genomes, these studies have provided evidences that lead to question the common belief that most bacteria and archaea genes have SD sequence [57]. For instance, the genomes of *Sulfolobus sulfataricus* and *Mycoplasma pneumoniae* appear to do not have

- BACKGROUND AND OBJECTIVES -

genes led by SD sequences [62,63]. The population of non SD led genes is considerably larger than previously thought and the SD content varies widely in different prokaryotes [61]. This suggests that alternative unknown mechanisms may be involved in the translation initiation [62,64,65].

In prokaryotes, many co-directional genes are sufficiently close together that the end of one gene may overlap with the SD or even the coding sequence of the next gene. This constrains the end of the upstream gene and the stop codon usage. The changes in composition at the end of genes are consistent with selection against the formation of mRNA secondary structure around the start codon of the next gene in the chromosome. A is likely to be the most favored base in such regions since it only binds U weakly, whereas G is likely the least favored base because it binds U weakly and C strongly [66]. The three stop codons are used unequally in prokaryotes. TAA is used in preference to TGA, which itself is used in preference to TAG [67]. TAA is the preferred stop codon because of faster termination or lower levels of translational read-through. It seems that the use of TGA and TAG increases when the stop codon has other coding functions [68]. For instance the 4bps overlaps, which are extremely common in prokaryotes, require TGA as stop codon. Also the three stop codons may be part of the SD sequence [68]. Recently, in the fusellovirus SSV4 a significantly part of the TGA stop codons analysed were part of a SD of the next gene (GGTGA). As the prokaryote genomes are also highly compacted, we could expect such SD sequence adopting the stop codon within its sequence. Here we assess how the stop codon usage and the short intergenic spacers adapt themselves to the SD location in prokaryote genes.

The analysis of intergenic regions and overlapping genes among prokaryotes can be hampered by annotation errors [69]. These errors can be incorrectly predicted genes, mispredicted start codons or loss of stop codons due to sequencing errors. Therefore is worth studying the overlapping genes, the origin and terminus of replication, the intergenic regions and the regulatory

signals compromised within these regions in order to minimize the annotation errors and find simple rules for improving the automatic annotation algorithms.

OBJECTIVES OVERVIEW

Because of the research lines of the research group, where I am involved, and the available background of this topic, the objectives of this thesis have been the following:

- *In silico* characterization of the Origin of replication of *Bacteroides thetaiotaomicron*.

Some origins of replication have been experimentally determined and have led to the development of *in silico* approaches to find the origin of replication among other prokaryotes. DNA base composition asymmetry is the basis of numerous *in silico* methods used to detect the origin and terminus of replication in prokaryotes. However the composition asymmetry does not allow us to locate precisely the positions of the origin and terminus. Some genome projects directly annotate the origin of replication around the region where the skews switch polarity, close to the *dnaA* gene, without going beyond in the study. Since DNA replication is a key step in the cell cycle it is important to determine properly the origin and terminus regions. Therefore, the methods, tools and databases for predicting the origins and terminuses of replication were reviewed and some complementary analyses to reinforce these predictions were proposed. These analyses include finding the *dnaA* gene and its binding sites; making BLAST analyses of the intergenic sequences compared to related species; studying the gene order around the origin sequence; and studying the distribution of the genes encoded in the leading versus the lagging strand. All these analyses we applied to correct the *Bacteroides thetaiotaomicron* origin prediction. This is a clear case where the genome project of this bacterium did not go in detail in the origin prediction and they gave a wrong origin prediction.

- BACKGROUND AND OBJECTIVES -

They located the origin on the opposite site of the circular chromosome. This subject is addressed in *Chapter 1*.

- Determination of the overlapping gene structures among prokaryote genomes.

Overlapping genes are a conserved feature of prokaryote genomes. Actually the overlapping pairs appear to be more conserved than the non-overlapping genes. The proportion of non-degenerate sites is higher in overlapping genes than in non-overlapping genes, thus reducing the proportion of synonymous mutations out of the total number of mutations. Therefore, a mutation could affect two proteins at the same time resulting in the loss of two functions in the cell. Generally, the overlapping lengths tend to be short because of the selective pressure against long overlaps, as the existence of overlapping reading frames increases the risk of deleterious mutations. The overlapping genes have preferred and prohibited overlapping lengths and there are allowed and not allowed overlapping phases. The overlaps seem to have a role in the transcriptional and translational regulation of gene expression and can potentially influence the evolution of genes. The significance and evolution of this conserved feature among prokaryotes has been well studied. Here we analyze that phenomenon in terms of genome organization and genome structure, as well as the relationship between the overlaps with the SD sequence presence and location. A good knowledge of the overlapping gene structures and the SD locations can help to improve genome annotation and may contribute to functional prediction. This subject is addressed in *Chapter 2* and *5*.

- Analysis of the reliability of large gene overlaps

The exponentially increasing amount of sequence information has spurred the need for automated and accurate large-scale prediction and functional annotation of genes. A new generation of technologies is speeding up

- BACKGROUND AND OBJECTIVES -

the sequencing even more, but this comes at the price of some biases and an increased error rate. Thus, it is important to investigate unexplained phenomena for systematic errors. One such phenomenon is a large number of annotated genes with long overlaps. While there is plenty of evidence that small gene overlaps of several nucleotides enhance coordinated transcription of functionally related genes, it is not known whether long overlaps are the product of special functional constraints or simply of large-scale misannotations. A number of previous studies of overlapping microbial genes suggested that annotation errors such as misprediction of start codons, loss of termination codons as well as the misidentification of the entire open reading frames (ORFs) can influence the statistics of overlapping genes and hence their analysis. Nevertheless none of the previous studies has attempted to quantify and characterize rigorously these possible misannotations to be able to study gene overlaps more reliably. In this thesis we analyse long overlaps between well-characterized genes to discriminate true events from misannotations and to use this knowledge to develop rules for improving gene annotation. This subject is addressed in *Chapter 3*.

- Identification and location of the SD sequence.

As has been mentioned above, the SD sequence is a motif, 5'-GGAGGU-3', located in the 5' of the initiation codons and is complementary of the sequence, 3'-CCUCCA-5', located at the 16S rRNAs tail. The prokaryote species seem to have preferred distances between the SD and the start codon and this distance varies among the species. The conservation of this distance is important to assure an efficient translation initiation. It has been postulated that when the SD resides within the 4 nucleotides from the initiation codon or when is located as far as 13 nucleotides from the initiation codon, gene expression is decreased drastically. Under the scenario of short intergenic distances and overlaps between genes, which are extremely common in prokaryote genomes, it is interesting asses the SD presence and location between genes. Since the

SD sequence seems to not vary its strength and its distance to the start codon because of the close proximity of the prokaryote genes, the intergenic regions and the stop codon may adapt themselves to the presence of a SD motif. This subject is addressed in *Chapter 2, 4* and *5*.

- Construction of a database to store and analyse the overlapping genes and the spacers between genes among the prokaryote genomes.

Because of the huge amount of data extracted from the analysis of the intergenic regions and the overlapping genes among prokaryote genomes we thought in build a database in order to store all the data generated. On one hand, the location of the SD can help to correct the gene annotations and could influence the spacing length and the stop codon usage. Therefore, a database dedicated to study the intergenic regions and their relationship with the SD locations seems to be very useful.

On the other hand, across the fully sequenced microbial genomes there are thousands of examples of overlapping genes. It is a consistent and worth studying phenomenon among prokaryotes and is often studied by the scientific community. The overlaps seem to have a role in the transcriptional and translational regulation of gene expression and can potentially influence the evolution of genes. Therefore, databases that can provide users with useful information about overlapping genes appear to be desirable. This subject is addressed in *Chapter 5*.

REFERENCES

1. Fleischmann R, Adams M, White O, Clayton R, Kirkness E, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496-512.
2. Koonin E, Wolf Y (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.*
3. Guzmán E, Romeu A, Garcia-Vallve S (2008) Completely sequenced genomes of pathogenic bacteria: a review. *Enferm Infecc Microbiol Clin* 26: 88-98.
4. Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet* 17: 589-596.
5. Casjens S (1998) The diverse and dynamic structure of bacterial genomes. *Annu Rev Genet* 32: 339-377.
6. Fraser C, Gocayne J, White O, Adams M, Clayton R, et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270: 397-403.
7. Kunst F, Ogasawara N, Moszer I, Albertini A, Alloni G, et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 390: 249-256.
8. Goffeau A, Barrell B, Bussey H, Davis R, Dujon B, et al. (1996) Life with 6000 genes. *Science* 274: 546, 563-547.
9. Lander E, Linton L, Birren B, Nusbaum C, Zody M, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
10. Venter J, Adams M, Myers E, Li P, Mural R, et al. (2001) The sequence of the human genome. *Science* 291: 1304-1351.
11. Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, et al. (2002) Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Res* 30: 4264-4271.
12. Frank AC, Lobry JR (2000) Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics* 16: 560-561.
13. Mira A, Pushker R (2005) The silencing of pseudogenes. *Mol Biol Evol* 22: 2135-2138.
14. Fuxelius H, Darby A, Cho N, Andersson S (2008) Visualization of pseudogenes in intracellular bacteria reveals the different tracks to gene destruction. *Genome Biol* 9: R42.
15. Marczynski GT, Shapiro L (1993) Bacterial chromosome origins of replication. *Curr Opin Genet Dev* 3: 775-782.
16. Mott ML, Berger JM (2007) DNA replication initiation: mechanisms and regulation in bacteria. *Nat Rev Microbiol* 5: 343-354.
17. Rocha EP, Danchin A, Viari A (1999) Universal replication biases in bacteria. *Mol Microbiol* 32: 11-16.
18. Rocha E (2004) The replication-related organization of bacterial genomes. *Microbiology* 150: 1609-1627.

- BACKGROUND AND OBJECTIVES -

19. Rudner R, Karkas JD, Chargaff E (1968) Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. *Proc Natl Acad Sci U S A* 60: 921-922.
20. Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol* 40: 318-325.
21. Lobry JR (1995) Properties of a general model of DNA evolution under no-strand-bias conditions. *J Mol Evol* 40: 326-330.
22. Mrazek J, Karlin S (1998) Strand compositional asymmetry in bacterial and large viral genomes. *Proc Natl Acad Sci U S A* 95: 3720-3725.
23. Frank AC, Lobry JR (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* 238: 65-77.
24. Tillier ER, Collins RA (2000) The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J Mol Evol* 50: 249-257.
25. Kowalczyk M, Mackiewicz P, Mackiewicz D, Nowicka A, Dudkiewicz M, et al. (2001) DNA asymmetry and the replicational mutational pressure. *J Appl Genet* 42: 553-577.
26. Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* 13: 660-665.
27. Grigoriev A (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res* 26: 2286-2290.
28. Lobry JR (1996) Origin of replication of *Mycoplasma genitalium*. *Science* 272: 745-746.
29. Francino MP, Ochman H (1997) Strand asymmetries in DNA evolution. *Trends Genet* 13: 240-245.
30. Brewer BJ (1988) When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. *Cell* 53: 679-686.
31. Karlin S (1999) Bacterial DNA strand compositional asymmetry. *Trends Microbiol* 7: 305-308.
32. Puigbò P, Romeu A, Garcia-Vallvé S (2008) HEG-DB: a database of predicted highly expressed genes in prokaryotic complete genomes under translational selection. *Nucleic Acids Res* 36: D524-527.
33. Johnson ZI, Chisholm SW (2004) Properties of overlapping genes are conserved across microbial genomes. *Genome Res* 14: 2268-2272.
34. Barrell BG, Air GM, Hutchison CA, 3rd (1976) Overlapping genes in bacteriophage phiX174. *Nature* 264: 34-41.
35. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, et al. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265: 687-695.
36. Normark S, Bergstrom S, Edlund T, Grundstrom T, Jaurin B, et al. (1983) Overlapping genes. *Annu Rev Genet* 17: 499-525.
37. Harrington ED, Singh AH, Doerks T, Letunic I, von Mering C, et al. (2007) Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proc Natl Acad Sci U S A* 104: 13913-13918.

- BACKGROUND AND OBJECTIVES -

38. Fukuda Y, Washio T, Tomita M (1999) Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. *Nucleic Acids Res* 27: 1847-1853.
39. Fukuda Y, Nakayama Y, Tomita M (2003) On dynamics of overlapping genes in bacterial genomes. *Gene* 323: 181-187.
40. Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, et al. (2002) Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet* 18: 228-232.
41. Sakharkar KR, Chow VT (2005) Strategies for genome reduction in microbial genomes. *Genome Inform* 16: 69-75.
42. Krakauer DC (2000) Stability and evolution of overlapping genes. *Evolution* 54: 731-739.
43. Sakharkar KR, Sakharkar MK, Verma C, Chow VT (2005) Comparative study of overlapping genes in bacteria, with special reference to *Rickettsia prowazekii* and *Rickettsia conorii*. *Int J Syst Evol Microbiol* 55: 1205-1209.
44. Lillo F, Krakauer DC (2007) A statistical analysis of the three-fold evolution of genomic compression through frame overlaps in prokaryotes. *Biol Direct* 2: 22.
45. Chen SM, Takiff HE, Barber AM, Dubois GC, Bardwell JCA, et al. (1990) Expression and characterization of RNase-III and Era proteins - products of the *rnc* operon of *Escherichia coli*. *Journal of Biological Chemistry* 265: 2888-2895.
46. Inokuchi Y, Hirashima A, Sekine Y, Janosi L, Kaji A (2000) Role of ribosome recycling factor (RRF) in translational coupling. *Embo Journal* 19: 3788-3798.
47. Cock PJ, Whitworth DE (2007) Evolution of gene overlaps: relative reading frame bias in prokaryotic two-component system genes. *J Mol Evol* 64: 457-462.
48. McCarthy JE (1990) Post-transcriptional control in the polycistronic operon environment: studies of the *atp* operon of *Escherichia coli*. *Mol Microbiol* 4: 1233-1240.
49. Shine J, Dalgarno L (1974) The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A* 71: 1342-1346.
50. Hui A, de Boer H (1987) Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in *Escherichia coli*. *Proc Natl Acad Sci U S A* 84: 4762-4766.
51. Jacob W, Santer M, Dahlberg A (1987) A single base change in the Shine-Dalgarno region of 16S rRNA of *Escherichia coli* affects translation of many proteins. *Proc Natl Acad Sci U S A* 84: 4757-4761.
52. Stormo G, Schneider T, Gold L (1982) Characterization of translational initiation sites in *E. coli*. *Nucleic Acids Res* 10: 2971-2996.
53. Chen H, Bjerknes M, Kumar R, Jay E (1994) Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the

- BACKGROUND AND OBJECTIVES -


- translation initiation codon of *Escherichia coli* mRNAs. *Nucleic Acids Res* 22: 4953-4957.
54. Ringquist S, Shinedling S, Barrick D, Green L, Binkley J, et al. (1992) Translation initiation in *Escherichia coli*: sequences within the ribosome-binding site. *Mol Microbiol* 6: 1219-1229.
 55. Ma J, Campbell A, Karlin S (2002) Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol* 184: 5733-5745.
 56. Kozak M (1999) Initiation of translation in prokaryotes and eukaryotes. *Gene* 234: 187-208.
 57. Schurr T, Nadir E, Margalit H (1993) Identification and characterization of *E. coli* ribosomal binding sites by free energy computation. *Nucleic Acids Res* 21: 4019-4023.
 58. Osada Y, Saito R, Tomita M Analysis of base-pairing potentials between 16S rRNA and 5' UTR for translation initiation in various prokaryotes. *Bioinformatics* 15: 578-581.
 59. Starmer J, Stomp A, Vouk M, Bitzer D (2006) Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLoS Comput Biol* 2: e57.
 60. Slupska M, King A, Fitz-Gibbon S, Besemer J, Borodovsky M, et al. (2001) Leaderless transcripts of the crenarchaeal hyperthermophile *Pyrobaculum aerophilum*. *J Mol Biol* 309: 347-360.
 61. Chang B, Halgamuge S, Tang S (2006) Analysis of SD sequences in completed microbial genomes: non-SD-led genes are as common as SD-led genes. *Gene* 373: 90-99.
 62. Tolstrup N, Sensen CW, Garrett RA, Clausen IG (2000) Two different and highly organized mechanisms of translation initiation in the archaeon *Sulfolobus solfataricus*. *Extremophiles* 4: 175-179.
 63. Weiner J, 3rd, Herrmann R, Browning GF (2000) Transcription in *Mycoplasma pneumoniae*. *Nucleic Acids Res* 28: 4488-4496.
 64. Boni I, Artamonova V, Tzareva N, Dreyfus M (2001) Non-canonical mechanism for translational control in bacteria: synthesis of ribosomal protein S1. *EMBO J* 20: 4222-4232.
 65. Kolev V, Ivanov I, Berzal-Herranz A, Ivanov I (2003) Non-Shine-Dalgarno initiators of translation selected from combinatorial DNA libraries. *J Mol Microbiol Biotechnol* 5: 154-160.
 66. Eyre-Walker A, Bulmer M (1993) Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Res* 21: 4599-4603.
 67. Sharp P, Bulmer M (1988) Selective differences among translation termination codons. *Gene* 63: 141-145.
 68. Eyre-Walker A (1996) The close proximity of *Escherichia coli* genes: consequences for stop codon and synonymous codon use. *J Mol Evol* 42: 73-78.
 69. Natale DA, Galperin MY, Tatusov RL, Koonin EV (2000) Using the COG database to improve gene recognition in complete genomes. *Genetica* 108: 9-17.

ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGACAGCTGATAGG
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGGCGCTGGTCAGGCAGCGCAGCATGGGAAAA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
GATCAGATGACAGATAGATAACCACAGAGACATG
CGCGCC**C**GACGCATGATTGATGATCAGATGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGATGATA**H**AGTGATTAGATAGATGGTGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTTT
GAGAGATAGAGAG**A**TAGGATCGCGCTCGAGCGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AACCGGCCAACCGGTGG**P**TTAGGATAGATGATGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AAAGTGTGTGTGACAGACAG**T**TGATGATAGTACA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGA**E**AGCTGATAGC
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCT
CGCGCTCGCTCGAGCGCTAGCTCGAT**R**GATC
CGCGCTCAAACGAGCGCTAGCTCGATCGAT
GATAGGTACGCGAAAATGGCAGTAGCTAGCTAGCT
ATAGGTACGCGGATGAATGGCAGT
ATCGATCGATCGATCGATCGCGCG
AGCATGACACACACACATG/
CCAGGCAGCATAAAGCA
AGCTGGGTGGTAGGAG/
TCGATCGATCGAT



IN SILICO PREDICTION OF THE ORIGIN OF REPLICATION
AMONG BACTERIA: A CASE STUDY OF *BACTEROIDES*
THETAIOAOMICRON

ACGCGAAAATGGC/
JATAGAGATACAGAA
JTACGCGAAAATGGCAG
CGCTCGAGCGCTAGCTCGAT
AGGTACGCGAAAATGGCAGTAC
ATCGATCGATCGATCGATCGCGCG
TCAGCATGACACACACACATGATA
AGTGCCAGGCAGCATAAAGCAGACG/
ACCAGCAGCTGGGTGGTAGGAGTGATG/
GCAGTGCCAGGCAGCATAAAGCAGACGAC
JCACCAGCAGCTGGGTGGTAGGAGTGATGATC



IN SILICO PREDICTION OF THE ORIGIN OF REPLICATION
AMONG BACTERIA: A CASE STUDY OF *BACTEROIDES*
THETA IOTAOMICRON

Albert Pallejà*, **Eduard Guzman**, **Santiago Garcia-Vallvé**, **Antoni Romeu**

All the authors belong to the same affiliation and have the same address:
Evolutionary Genomics Group, Department of Biochemistry and Biotechnology,
Rovira i Virgili University, Tarragona, Catalunya, Spain

Rovira i Virgili University
Department of Biochemistry and Biotechnology
Campus Sescelades
C/ Marcel·lí Domingo, s/n
E-43007 Tarragona
Catalunya – Spain
Telephone: 0034 977558486
Fax: 0034 977558232

Email addresses: albert.palleja@urv.cat (*corresponding author)
santi.garcia-vallve@urv.cat
eduardo.guzman@urv.cat
antoni.romeu@urv.cat

OMICS A Journal of Integrative Biology
Volume 12, Number 3, 2008

OMICS: A Journal of Integrative Biology 2008, 12, (3): 201-210

© Mary Ann Liebert, Inc.

DOI: 10.1089/omi.2008.0004

Keywords: origin of replication; termination of replication; DNA compositional asymmetry; Bacteroides; origin predictions

Abstract

The initiation of chromosomal replication occurs only once during the prokaryote cell cycle. Some origins of replication have been experimentally determined and have led to the development of *in silico* approaches to find the origin of replication among other prokaryotes. DNA base composition asymmetry is the basis of numerous *in silico* methods used to detect the origin and terminus of replication in prokaryotes. However the composition asymmetry does not allow us to locate precisely the positions of the origin and terminus. Since DNA replication is a key step in the cell cycle it is important to determine properly the origin and terminus regions. Therefore, here we have reviewed the methods, tools and databases for predicting the origins and terminuses of replication and we have proposed some complementary analyses to reinforce these predictions. These analyses include finding the *dnaA* gene and its binding sites; making BLAST analyses of the intergenic sequences compared to related species; studying the gene order around the origin sequence; and studying the distribution of the genes encoded in the leading versus the lagging strand.

Introduction

Replication is the part of the DNA metabolism in which genetic information is transmitted from one generation of cells to the next (Kornberg and Baker 1991; Liberi and Foiani 2004). The entire genome must be replicated precisely once for every cell division. Replication can start from one origin of replication (*ori*) among bacteria and some archaea such as *Pyrococcus abyssi*

(Matsunaga et al. 2001), from two *oris* among some other archaea such as *Sulfolobus solfataricus* (Robinson et al. 2004) or from multiple *oris* like in eukaryotes. In bacteria, the initiation of DNA replication is a complex process that starts at a unique site only once per each cell division. This process involves several regulated steps such as the binding of the DnaA protein (initiator protein) to specific binding sites located within the *ori* sequence; the unwinding of the DNA from an AT-rich region located at the beginning of the *ori* sequence; and the binding of some helicases and other proteins required to form the replication forks (Baker and Bell 1998; Kornberg and Baker 1991). These replication forks start at the *ori* site, proceed bidirectionally (Marczynski and Shapiro 1993) and move around the genome at approximately the same speed to a meeting point. Some origins of replication have been experimentally determined such as those of *Escherichia coli* (Oka et al. 1980), *Bacillus subtilis* (Moriya et al. 1992), *Mycobacterium smegmatis* (Qin et al. 1997), *Mycobacterium tuberculosis* (Salazar et al. 1996), *Mycobacterium capricolum* (Fujita et al. 1992), *Streptomyces coelicolor* (Calcutt and Schmidt 1992), *Caulobacter crescentus* (Brassinga and Marczynski 2001), *Pseudomonas putida* (Yee and Smith 1990), *Sinorhizobium meliloti* (Sibley et al. 2006), *Thermus thermophilus* (Schaper et al. 2000) and *Thermotoga maritima* (Lopez et al. 2000). These experimental determinations have led to the development of *in silico* approaches to find the origin of replication among other species. *In silico* methods are used to predict the location of the chromosomal origin of replication, even though these do not allow for precise origin localization. Since DNA replication is a key step in the cell cycle it is important to determine properly the *ori* site among the sequenced genomes to date. Usually, the origin of replication is used to annotate the first nucleotide of a bacterial circular chromosome that sometimes leads to not enough accurate predictions (Worning et al. 2006). The aim of this work is to review the available methods, tools and databases associated with *ori* prediction, as well as to suggest some clues and complementary analyses in order to provide a “standard procedure” for the biologist that needs to locate the *ori* and *ter* on a genome sequence.

OMICS: A Journal of Integrative Biology 2008, 12, (3): 201-210

DNA strand compositional asymmetry

DNA base composition asymmetry is the basis of numerous *in silico* methods for detecting the origin and terminus of replication in prokaryotes (Frank and Lobry 2000; Grigoriev 1998; Salzberg et al. 1998; Worning et al. 2006; Zhang and Zhang 2002). E. Chargaff experimentally determined the approximated equimolarities $A \sim T$ and $C \sim G$ for long, single stranded DNA molecules (Rudner et al. 1968). These equalities are observed in the lack of bias between the two DNA strands for mutation and selection (Lobry 1995; Sueoka 1995). However, in prokaryotic genomes there are local and systematic deviations for many reasons, even though the main reason appears to be the different mutational pressure associated with the different mechanisms for replication between leading and lagging strands (Frank and Lobry 1999; Kowalczyk et al. 2001; Mrazek and Karlin 1998; Tillier and Collins 2000). This asymmetry generally divides the chromosome into two regions with opposite signs for base composition skews (Lobry 1996a). The skew parameter is the difference in base composition between leading and lagging strands. The switch in skew direction generally occurs at the origin and terminus of replication (Francino and Ochman 1997; Grigoriev 1998; Lobry 1996a, b; Rocha et al. 1999). In bacteria the leading strand for replication is enriched in keto (G or T) bases while the lagging strand is enriched in amino bases (C or A) (Perriere et al. 1996; Rocha et al. 1999). This compositional asymmetry has been widely studied using DNA walks, mononucleotide skews and skewed oligomers. Lobry adapted vectorial representations of sequences that were first introduced by Mizraji and Ninio (Mizraji and Ninio 1985; Ninio and Mizraji 1995) to do a DNA walk along the DNA sequence (Lobry 1995). This DNA walk was done by reading the sequence in the third codon positions and walking into the plane according to four directions defined as: C: North; G: South; A: West; T: East. The *ori* is close to the reverse turn of the trajectory of the DNA walk. In circular chromosomes there are two reverse turns, the second is the *ter*. The OriLoc

program (Frank and Lobry 2000) is based on this method and it offers us both the *ori* and *ter* position. However the switch in asymmetry does not always exactly correspond to the position of a functional *ori* as is the case with the *in silico* predicted *ori* of *Helicobacter pylori* (Zawilak et al. 2001). Another method widely used by the authors is the mononucleotide skews that represent the difference in nucleotides between leading and lagging strands. The different skews are assessed in a given strand as: GC skew $((G-C)/(G + C))$, AT skew $((A-T)/(A + T))$, purine skew $((G + A-T-C)/N)$, and keto skew $(G + T-A-C)/N$ (McLean et al. 1998). The keto and purine skews are correlated with each other as well as with the GC and AT skew. The keto and purine skews can provide us with better predictions than the single nucleotide skews (Freeman et al. 1998). The sum of the skew parameter in adjacent sliding windows is the cumulative skew which approximately changes the polarity at the origin and terminus of replication. Such plots may not always be very illustrative due to visible fluctuations in small windows, while larger windows may hide the precise coordinates of polarity switches. Therefore we strongly recommend using different window lengths and analysing the different plots obtained. While there is a tendency for more G's on the leading strand and for more C's on the lagging strand the AT skew direction varies between species and phyla. For instance the *B. subtilis* has more G's and A's on the leading strand, whereas *Bacteroides thetaiotaomicron* has more G's and T's on the leading strand. Therefore, for both species the GC skew will have a negative minimum around the *ori* site. Nevertheless, for *B. subtilis* the AT skew will have a negative minimum and for *B. thetaiotaomicron* the AT skew will have a positive maximum. Some studies have related the positive AT skew to the presence of two genes in a chromosome which code for DNA polymerases alpha subunits (*polC* and *dnaE*) (Worning et al. 2006). Nevertheless, a recent paper stated that the presence of the polymerases alpha subunits is not enough to predict the sign of AT skew on the leading strand (Necsulea and Lobry 2007). The information in the different single nucleotide skews can be combined into a three-dimensional curve, the Z-curve, which has been used to predict origins in both bacterial and archaeal

chromosomes (Zhang and Zhang 2002). Going beyond mononucleotide skews, a third method was created by Salzberg based on short oligomers whose orientation is preferentially skewed around the origin (Salzberg et al. 1998). These skewed seven-base or eight-base sequences occur much more often on the leading strand than in the lagging strand and also are significantly overrepresented among the chromosome. The point around which these oligomers are skewed is the *ori* site. However, in a circular chromosome the skew occurs at two points, the *ori* and *ter*. Since Salzberg's algorithm does not distinguish between these two points, additional evidence such as *dnaA* gene proximity and the distribution of DnaA boxes must be used to determine which one is the *ori*. Worning *et al.* have developed a new method that, instead of using nucleotide skews or eight-base oligomers, involves all the oligomers up to eight nucleotides (Worning et al. 2006). This method is more sensitive than existing ones and provides a quantitative measure for the difference between the leading and the lagging strand. The results of applying this method to a large group of chromosomes are available in The Genome Atlas database (Hallin and Ussery 2004). Table 1 shows other tools and databases related to predicting the replication origin in bacteria (Arakawa et al. 2003; Frank and Lobry 2000; Gao and Zhang 2007; Hallin and Ussery 2004; Salzberg et al. 1998; Thomas et al. 2007).

Although compositional asymmetry methods do not allow us to establish a precise position for the *ori* and *ter* sequences and also require exhaustive human inspection, it brings us near to those sequences, which tend to be intergenic (Frank and Lobry 2000), although there are some cases where the *ori*, or more specifically, the DnaA boxes overlap a gene such as *E. coli* chromosome (Salzberg et al. 1998). One thing that can help us make a prediction is knowing beforehand the *ori* prediction of related species, and it is even more helpful if that *ori* has been experimentally determined in these species.

Name	Description	Web adress	Authors
Skewed oligomers	Program developed for the prediction of origins of replication among prokaryotes based on find short oligomers whose orientation is skewed around the origin.	http://www.cceb.umd.edu/~salzberg/	Salzberg et al. 1998
Oriloc	Program developed for the prediction of bacterial replication origins based on DNA walks.	http://pbil.univ-lyon1.fr/software/Oriloc/oriloc.html	Frank and Lobry 2000
The Genome Atlas database	A database that contains several genome features such as origin and terminus of replication predictions among a collection of complete microbial genomes.	http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/origin/	Hallin and Ussery 2004
GraphDNA	GraphDNA is an easy to use Java Web Start application designed to display and compare DNA sequences graphically using three available methods: DNA walks, cumulative purine/keto skews and cumulative dinucleotide skews.	http://athena.bioe.uvic.ca/	Thomas et al. 2007
Doric database	Doric contains bacterial <i>oriC</i> s that are identified based on a systematic method comprising the Z-curve analysis for nucleotide distribution asymmetry, DnaA box distribution, genes adjacent to candidate <i>oriC</i> s and phylogenetic relationships.	http://tubic.tju.edu.cn/doric/	Gao and Zhang 2007
G-language system	G-language Genome Analysis Environment provides a greater variety of useful genome analysis tools, for instance tools for predicting origin and terminus of replication.	http://www.g-language.org/	Arakawa et al. 2003

Table 1. Useful tools for origin predictions

Genes related to replication origin and the DnaA boxes

DNA asymmetry is the most widespread method for identifying *ori* in bacterial chromosomes, but it should be applied together with other methods in order to make better predictions. Some genome projects directly annotate the origin of replication around the region where the skews switch polarity, close to the *dnaA* gene, without finding the DnaA sequence specific binding sites (DnaA boxes). The *ori* sequence coordinates provided by these projects must be revised, such as the genomes of *Fusobacterium nucleatum* and *B. thetaiotaomicron* (Worning et al. 2006). Therefore, the second step is to check whether the genes related to replication and to DnaA boxes are distributed close to the region where the skews have changed the polarity. These genes are *rnpA* (ribonuclease P), *rmpH* (ribosomal protein L34), *dnaA* (chromosome replication initiator protein), *dnaN* (DNA polymerase III beta), *gyrB* (DNA gyrase, subunit B), *gyrA* (DNA gyrase, subunit A), and, in some cases, *recF* (recombination protein). Furthermore, a significant number of DnaA boxes are present around the center of this gene distribution. The genomes that follow this pattern are

close to a classic progenitor origin region (Yoshikawa and Ogasawara 1991) such as *B. subtilis*, *Borrelia burgdorferi*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae* and *Treponema pallidum* (Salzberg et al. 1998). Unlike these genomes, *Haemophilus influenzae* and *E. coli oris* are associated with the genes *gidA* (glucose inhibited protein A) surrounded by DnaA boxes. In *E. coli* *gidB* (glucose inhibited protein B) is also present. In other genomes, such as *H. pylori*, *Synechocystis* sp. and *Methanobacterium thermoautotrophicum*, the genes associated with the *ori* are scattered throughout the chromosome. Therefore, although the DnaA and its binding sites are well conserved throughout the bacterial kingdom, the distribution of genes associated with the *ori* along the chromosome can change depending on the species (Mott and Berger 2007). Nevertheless, *dnaA* genes have not been found in *Wigglesworthia glossinidia* and *Blochmannia floridanus*, although this reflects the dependency of the host on these reduced-in-size endosymbionts (Gil et al. 2003). Although in most cases *dnaA* gene is close to the *ori* site, this gene is not always adjacent to the *ori* region. In fact, the *dnaA* gene does not need to be close to the *ori* sequence for it to function properly and we cannot rule out that the replication may be initiated by a different protein in other species, such as has been demonstrated in mutants of *Synechocystis* sp. with an inactivated *dnaA* gene (Richter et al. 1998). However, the proximity of this gene to the *ori* would permit the DnaA protein to associate with the *ori* as soon as it is synthesized, as well as minimizing the possibility of disrupting the cooperation between *dnaA* gene and *ori* region. In *E. coli* a long studied model system for replication initiation, *oriC* is approximately 250 base pairs (bps) in length and contains multiple DnaA boxes (Fuller and Kornberg 1983; Matsui et al. 1985; Tabata et al. 1983). However replication origins from different species show differences in the overall length, number and arrangements of DnaA boxes (Zawilak-Pawlik et al. 2005). Each one of these DNA motifs (5'-TTATCCACA-3' as a DnaA box consensus sequence) can act as a DnaA binding site. Accompanying these DnaA boxes we can find I sites which differ subtly from DnaA boxes consensus and are found interspersed among the DnaA boxes

and AT-rich region elements composed of three 13 bps repeats. Both motif sequences also bind to DnaA protein. The interactions of DnaA with the different binding sites define how replication initiation progresses (Mott and Berger 2007). Therefore, finding the DnaA boxes as well as these sequence motifs mentioned above is a key factor for predicting the origin of replication. However, it must be taken into account that the preferred DnaA box sequences of a given bacteria can be slightly different from the *E. coli* perfect DnaA box. There are two examples to illustrate this: i) in high GC content organisms such as *Mycobacterium*, *Streptomyces* and *Micrococcus* the third position of the perfect *E. coli* consensus DnaA box is substituted by G or C, ii) the affinity of the DnaA protein from *H. pylori* with the TCATTCACA sequence is higher than with the *E. coli* perfect DnaA box (Mackiewicz et al. 2004). Thus, the possibility of different consensus sequences for the DnaA box should be considered, whilst searching for the putative DnaA specific binding sites. However, bases in the second, fourth, seventh, eighth and ninth positions are well conserved (Messer 2002; Schaefer and Messer 1991).

Complementary analysis

Once the *ori* sequence position is predicted in the chromosome, we could make another complementary analysis to validate and reinforce our prediction. To do this we could use the following means: i) by studying the gene order adjacent to the *ori* sequence, that means checking if the genes around the *ori* sequence are conserved across the related species. Evidently the conservation of the intergenic sequences and gene order across the related species is not a random process. Furthermore, it has been reported that genomic rearrangements rarely occur close to the *ori* site (Kelman and Kelman 2003). However, this first complementary analysis must be taking into account if in the intergenic region, putative to be the *ori*, we have found the DnaA boxes, because evidently, if one compares closely related species there will find other regions with conserved gene order; ii) by making a BLAST analysis (Altschul et

al. 1990) of the intergenic sequences of the chromosome against all the intergenic sequences of a related species' chromosome. This analysis, which could be made at the beginning of our prediction, could provide us with candidates which have the right length and are similar enough to be the *ori* sequence. This is even more useful if one of the species with which you compare your intergenic sequence has already had its *ori* sequences predicted; iii) by checking if our prediction corresponds to an expected distribution of genes encoded in the leading or lagging strand. In fact, once the location of the *ori* and *ter* position have been determined by Oriloc (Frank and Lobry 2000), the program computes the percentage of coding sequences on the leading strand. If the percentage is less than 50% the prediction is probably incorrect because, in bacteria, there is selective pressure to concentrate the coding sequences on the leading strand (Brewer 1988). The DNA replication mechanism is an essential factor in the organization of prokaryotic genomes. In bacterial circular chromosomes, both the *ori* and the *ter* location determine whether the semicircle of a given strand is a leading or a lagging strand. In addition, it is known that highly expressed genes are overrepresented in the leading strand while alien genes are encoded mainly in the lagging strand (Karlin 1999). Therefore if our prediction is correct we can expect a higher presence of genes in leading strand, especially among highly expressed genes. Evidently, to calculate the genes distributed in leading versus lagging strands we need to know both the prediction of the *ori* position and the prediction of the *ter* position. A *ter* region was described in a model organism *E. coli* (Masters and Broda 1971) where replication forks appeared to terminate in a region corresponding to ~5% of the chromosome (de Massy et al. 1987) located opposite the *ori*. However, termination does not seem to depend solely on the two replication forks arriving coincidentally at the same time at the meeting point. Some studies focused on this identified the *ter* sites (Hill et al. 1987; Hill et al. 1988; Pelletier et al. 1988) as the sequence elements which can stall replication forks only when Tus protein is bound there (Hidaka et al. 1989; Kobayashi et al. 1989). A model (Hill 1992) was proposed whereby the 'inner-most' *ter* sites act as a replication trap,

OMICS: A Journal of Integrative Biology 2008, 12, (3): 201-210

where forks could enter but not leave. However, in *E. coli*, deleting the *ter* sequences or mutational inactivation of Tus has no apparent effect (Hill 1992; Roecklein et al. 1991; Skokotas et al. 1994). Furthermore, the Tus protein is not conserved across species, only in relatives of *E. coli*. On the other hand, replication forks are arrested by an analogous but not structurally homologous system, which is the *ter* site/rtp protein system on *B. subtilis* chromosome (Bussiere and Bastia 1999; Wake 1997). This system is also restricted phylogenetically. Recently, some authors have proposed that *ter* sites participate in halting replication forks originating from DNA repair events and not those originating at the chromosomal *ori* (Hendrickson and Lawrence 2007). They also found that the *ter* position is most likely to be at or near the *dif* site in gamma-proteobacteria and Firmicutes and they provided us with a consensus sequence for the *dif* site in each taxonomic group. The XerCD recombinase is the protein which acts at the *dif* site to resolve chromosome catemers following replication termination. Thus when predicting the *ter* position we can use the change in the skew polarity opposite to the *ori* sequence, the *ter* sites that may be involved in halting the replication forks, and the *dif* site that seems to be close to the termination of the replication. However, it is difficult to predict the *ter* position among prokaryotic chromosomes due to the variability of the consensus sequences (*ter* sites or *dif* sites) across species.

Finally it should not be forgotten that we will never be absolutely confident that we have found the *ori* sequence until we have experimental confirmation.

The *Bacteroides thetaiotaomicron* origin prediction case

In order to illustrate how to predict the origin of replication we have chosen to study the origin of replication in *B. thetaiotaomicron*, a human gut organism involved in numerous metabolic activities due to its high percentage of carbohydrate active enzymes (CAZY) (<http://afmb.cnrs-mrs.fr/CAZY/>) (Coutinho and Henrissat 1999). The *B. thetaiotaomicron* genome was sequenced by Xu et

al. (Xu et al. 2003) and they predicted the origin by using only the skew diagrams around 1100000 base pairs (bps). Later Mackiewicz *et al.* (Mackiewicz et al. 2004) as well as Worning *et al.* (Worning et al. 2006) predicted the origin of replication on the opposite site of the circular chromosome of this bacterium at around a little more than 4000000 bps. Examining the cumulative diagrams of *B. thetaiotaomicron* we can see that there are two points where there is a change in the DNA compositional asymmetry which are around 1050000 and 4000000 bps (Figure 1). Since the skew switches polarity near the origin and terminus of replication, the genome regions around those points of the circular chromosome are candidates for the *ori* or *ter* of replication. Thus we have two regions which may be the origin of replication and we have to check which one contains the *ori* sequence and then, locate precisely its coordinates in the chromosome.

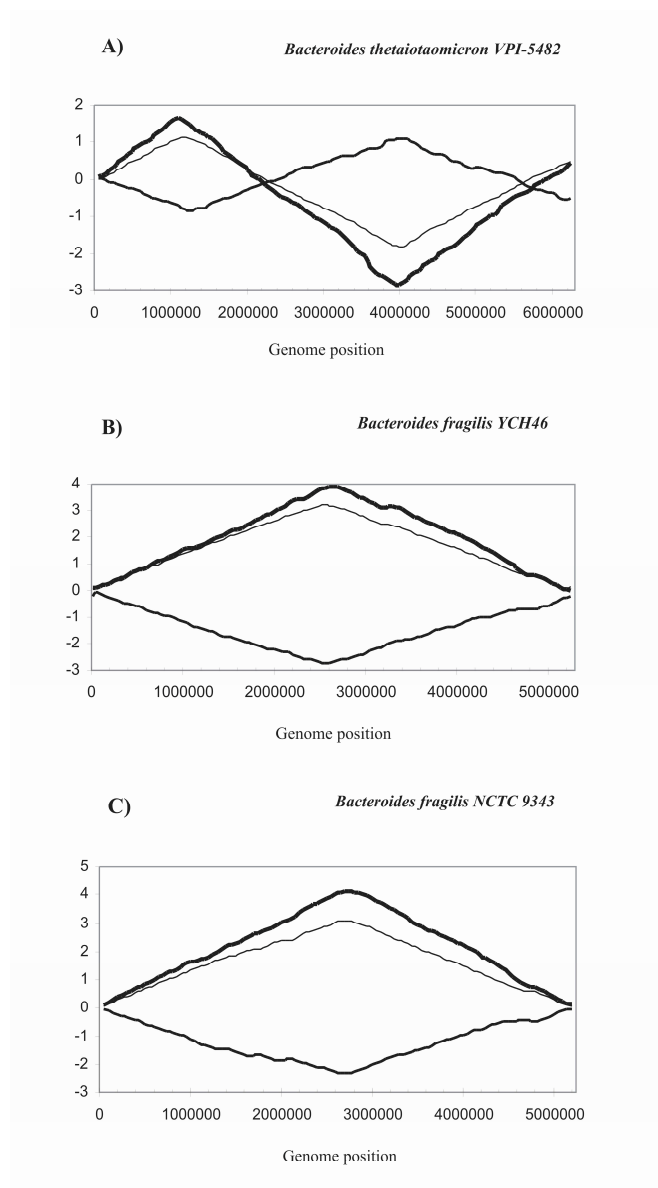


FIG. 1. Cumulative skew plots of *B. thetaiotaomicron* VPI-5482 (A), *B. fragilis* YCH46 (B), and *B. fragilis* NCTC 9343 (C). These diagrams were built taking windows of 100 kb and sliding 50 kb along the chromosome. The thinnest line represents the cumulative keto excess, the medium line represents the cumulative AT skew, and the thickest line represents the cumulative GC skew.

In this case we are helped by the fact that *Bacteroides fragilis* has already been predicted in The Genome Atlas Database. They have predicted that the origin of replication in *B. fragilis* YCH46 is around the position 5277274 bp, within the intergenic sequence located between the end of the last annotated gene and the beginning of the first annotated gene (Hallin and Ussery 2004). Around this region is approximately where we can see a GC skew minimum (Figure 1). The origin of *B. fragilis* YCH46 was used as a query in a BLAST search against all the completed sequenced genomes. The only retrieved sequences were the previously described origin of replication in other *B. fragilis* NCTC 9343 and one intergenic sequence of *B. thetaiotaomicron*. This *B. thetaiotaomicron* intergenic sequence is located between 4035393 and 4035883 bps of the chromosome, where there was approximately one change in skew polarity. Therefore it seems that *B. thetaiotaomicron ori* sequence could be located in this intergenic sequence, but it is necessary to go beyond and check both the presence of the *dnaA* gene and genes related with replication around this intergenic region, as well as the presence of DnaA boxes along the intergenic sequence.

We observed five DnaA boxes with several changes from those of *E. coli* in the pairwise alignment between the origin of replication of the two strains of *B. fragilis* and the *B. thetaiotaomicron* intergenic sequence (Figure 2). Three of the DnaA boxes found differed by only one position from the *E. coli* DnaA boxes respectively, whereas two of them differed by two positions. Therefore, like the *E. coli* origin, the *Bacteroides*' origins also contain DnaA binding sites which are a key factor for initiating chromosomal DNA replication (Bramhill and Kornberg 1988; Matsui et al. 1985). Furthermore the 13 bps sequences found in *E. coli* origin could be related to the 5'-extreme conserved regions in *Bacteroides* origins. These 5'-regions are enriched in AT where, in the replication, the DNA should start to unwind. However, a significant difference between the origins of replication in *Bacteroides* and *E. coli* is that in *E. coli* fourteen GATC palindromic sequences are located in the *oriC* and neighbouring regions, instead of the one or two expected in a 300 bps interval. GATC is a

methylation site from Dam methylase enzyme, where adenine is the methylated base. In contrast, *Bacteroides* origins do not show any GATC *dam* methylation site, even though two CCGG sequences (another palindromic DNA methylation site) have been observed. On the other hand the *dnaA* gene is not close to the intergenic sequence candidate to be the *ori* sequence. The other *E. coli* genes related to replication are also not found around this intergenic region. However, assuming a 360° chromosomal map in the three *Bacteroides* genomes, the distances, expressed in degrees, are conserved between the genes involved in replicating to the predicted origin. Although in *B. thetaiotaomicron* as well as in *B. fragilis* strains the genes related to replication are neither close nor adjacent to the *ori* sequence, it seems that the positions on the circular chromosome of these genes are conserved across the *Bacteroides* genomes. Therefore in these genomes the switch in the skew polarity and the DnaA boxes are close but the *dnaA* gene has been transferred to another region of the chromosome such as happens in the *Rickettsia prowazekii* Madrid E chromosome (Mackiewicz et al. 2004).

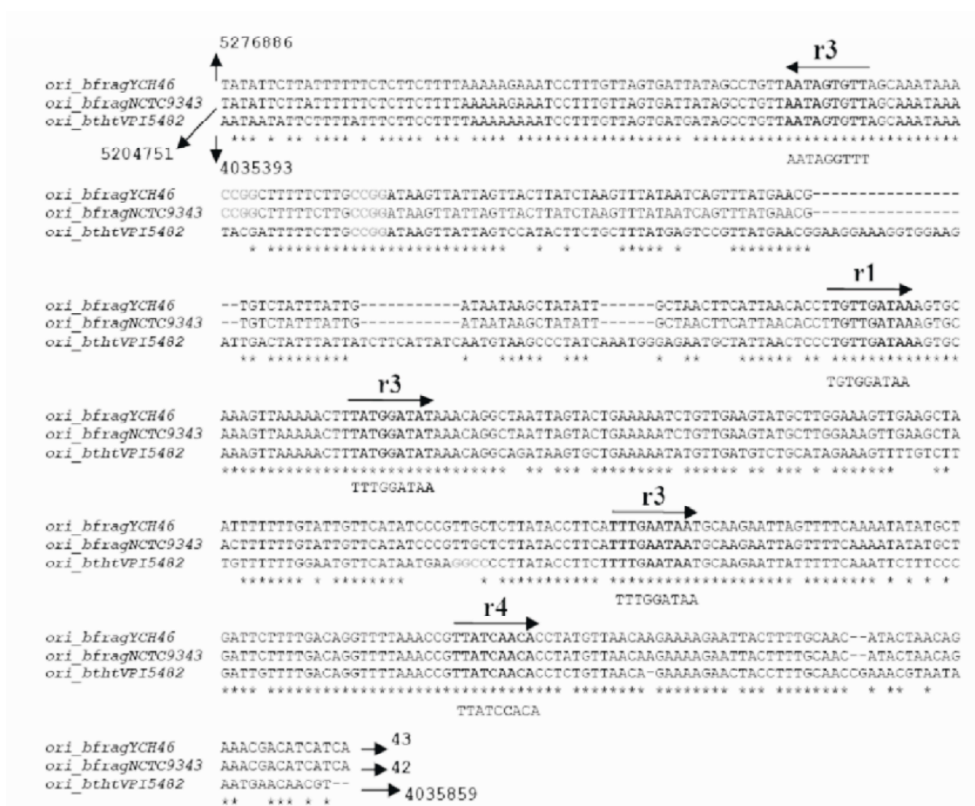


FIG. 2. Pairwise alignment of the putative origin region of *B. thetaiotaomicron* VPI-5482 (ori_btthtVPI5482) and both predicted origin regions of *B. fragilis* YCH46 (ori_bfragYCH46) and *B. fragilis* NCTC9343 (ori_bfragNCTC9343). Coordinates of the sequences in the chromosome are denoted in bps. Identities are marked with an asterisk. DnaA boxes are in bold. Sequence. DnaA box (r1–r4) orientations are indicated by an arrow. Below the boxes is the *E. coli* DnaA box sequence (nucleotide differences are denoted in bold). Methylation sites (CCGG) are denoted in gray.

In terms of gene topology, Bacteroides' *ori* sequences are located in a conserved chromosomal region (Figure 3). On one hand, the upstream gene is a quinolinate synthetase A with 88% similarity between *B. thetaiotaomicron* and

B. fragilis YCH46, and these genes in both genomes are located in the lagging DNA strand. On the other hand, the downstream gene is a tRNA/rRNA methyltransferase with 96% similarity, located in the leading DNA strand. Thus the *ori* sequence appears to be within a conserved region among *Bacteroides* species.

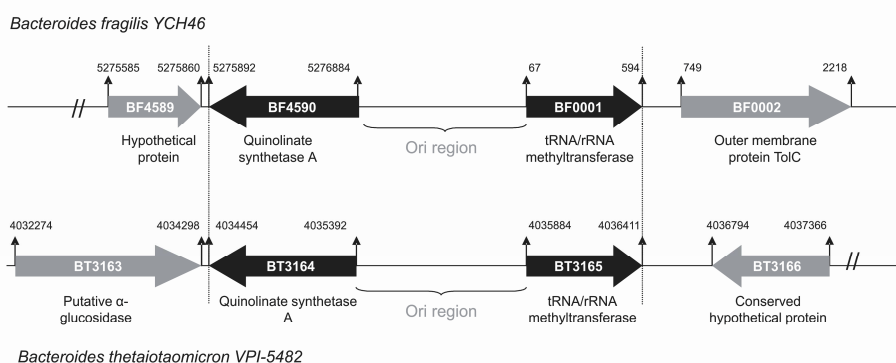


FIG. 3. Topology of the chromosome fragments of *B. fragilis* YCH46 and *B. thetaiotaomicron* VPI-5482 where the origins of replication are located.

A local BLAST between all the intergenic regions of *B. fragilis* YCH46 against all the intergenic regions of *B. thetaiotaomicron* confirmed the BLAST results above regarding the *ori* position. Furthermore, it indicated a *B. thetaiotaomicron* intergenic region between 1041337 and 1041592 bps that matches with a remarkably low E-value with a *B. fragilis* YCH46 intergenic region located between 2726093 and 2726571 bps. Both intergenic sequences were located in the regions of the respective chromosomes, where there were switches of polarity in the cumulative skew diagrams. This suggests that these intergenic regions in both species opposite to the *ori* sequence coordinates could be involved in the terminus of replication, therefore it could be the region where approximately the DNA polymerases meet and the replication is halted.

Although, the sequence alignments (data not shown) between these intergenic regions reveal a high sequence similarity, they do not show any pattern related to the terminus mechanisms, such as the classical *ter* sites of *E. coli* or the *dif* sites patterns for gamma-proteobacteria, Firmicutes or Actinobacteria. In addition, the XerCD protein found in these taxonomic groups is not found within Bacteroides genomes and neither is the FtsK translocase found within Bacteroides. This protein activates and delivers the XerCD protein. On the other hand, in the high conserved intergenic regions among Bacteroides, repetitions of 47 bps that are probably old fragments of DNA coding for Proline tRNAs were observed. These repetitions are present not only in Bacteroides but in a large number of prokaryotic genomes.

Finally once the *ori* has been predicted (between 4035393 and 4035883 bps) and assuming that the *ter* may be around the second switch in the skew polarity where similarity has been observed between Bacteroides sequences (between 1041337 and 1041592 bps), it is possible to check in *B. thetaiotaomicron* if the distribution of the genes between leading versus lagging strands agrees with the prediction. In the *B. thetaiotaomicron*, 58.1% of the total genes are encoded by the leading strand. We specifically analysed the DNA strand distribution of highly expressed genes (Puigbò et al. 2008), gene coding for ribosomal proteins and carbohydrate active enzymes (genes overrepresented among Bacteroides). In all groups a clearly higher percentage of genes encoded in the leading strand (69%, 87%, and 62% respectively) was observed. Therefore, this prediction agrees with the strand distribution of the genes in leading versus lagging strands. In conclusion, this reinforces this prediction and confirms the fact that the *ori* in *B. thetaiotaomicron* is not where it was indicated to be in the published sequence (Xu et al. 2003), it is where Worning *et al.*, who used a more accurate *in silico* method, determined it to be (Worning et al. 2006). Actually, in this case, simply a more accurate analysis of the nucleotide skew plots had been enough to give a correct prediction for this genome. Especially if one looks at the direction of the GC skew as well as the distribution of the genes on the leading strand. Thus, this is one example that

illustrates the importance of going beyond predictions made from the skew diagrams or at least analyse it properly, especially in genomes which are not close to species whose origin of replication has been experimentally determined.

Acknowledgments

This work has been supported by projects BIO02003-07672 and AGL2007-65678/ALI of the Spanish Ministry of Education and Science. We also want to thank Christian Brassington from the Language Service of the Rovira i Virgili University for their help in writing the manuscript.

References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.
- Arakawa, K., Mori, K., Ikeda, K., Matsuzaki, T., Kobayashi, Y. and Tomita, M. (2003). G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining. *Bioinformatics* **19**, 305-306.
- Baker, T.A. and Bell, S.P. (1998). Polymerases and the replisome: machines within machines. *Cell* **92**, 295-305.
- Bramhill, D. and Kornberg, A. (1988). A model for initiation at origins of DNA replication. *Cell* **54**, 915-918.
- Brassinga, A.K. and Marczyński, G.T. (2001). Replication intermediate analysis confirms that chromosomal replication origin initiates from an unusual intergenic region in *Caulobacter crescentus*. *Nucleic Acids Res* **29**, 4441-4451.

- Brewer, B.J. (1988). When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. *Cell* **53**, 679-686.
- Bussiere, D.E. and Bastia, D. (1999). Termination of DNA replication of bacterial and plasmid chromosomes. *Mol Microbiol* **31**, 1611-1618.
- Calcutt, M.J. and Schmidt, F.J. (1992). Conserved gene arrangement in the origin region of the *Streptomyces coelicolor* chromosome. *J Bacteriol* **174**, 3220-3226.
- Coutinho, P. and Henrissat, B. (1999). Carbohydrate active enzymes: an integrated database approach. In "*Recent Advances in Carbohydrate Bioengineering*". (H.J. Gilbert, G. Davies, B. Henrissat and B. Svensson eds., The Royal Society of Chemistry, Cambridge), pp. 3-12.
- de Massy, B., Bejar, S., Louarn, J., Louarn, J.M. and Bouche, J.P. (1987). Inhibition of replication forks exiting the terminus region of the *Escherichia coli* chromosome occurs at two loci separated by 5 min. *Proc Natl Acad Sci U S A* **84**, 1759-1763.
- Francino, M.P. and Ochman, H. (1997). Strand asymmetries in DNA evolution. *Trends Genet* **13**, 240-245.
- Frank, A.C. and Lobry, J.R. (1999). Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* **238**, 65-77.
- Frank, A.C. and Lobry, J.R. (2000). Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics* **16**, 560-561.
- Freeman, J.M., Plasterer, T.N., Smith, T.F. and Mohr, S.C. (1998). Patterns of Genome Organization in Bacteria. *Science* **279**, 1827.

- Fujita, M.Q., Yoshikawa, H. and Ogasawara, N. (1992). Structure of the *dnaA* and DnaA-box region in the *Mycoplasma capricolum* chromosome: conservation and variations in the course of evolution. *Gene* **110**, 17-23.
- Fuller, R.S. and Kornberg, A. (1983). Purified *dnaA* protein in initiation of replication at the *Escherichia coli* chromosomal origin of replication. *Proc Natl Acad Sci U S A* **80**, 5817-5821.
- Gao, F. and Zhang, C.T. (2007). DoriC: a database of *oriC* regions in bacterial genomes. *Bioinformatics* **23**, 1866-1867.
- Gil, R., Silva, F.J., Zientz, E., Delmotte, F., Gonzalez-Candelas, F., Latorre, A., et al. (2003). The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9388-9393.
- Grigoriev, A. (1998). Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res* **26**, 2286-2290.
- Hallin, P.F. and Ussery, D.W. (2004). CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data. *Bioinformatics* **20**, 3682-3686.
- Hendrickson, H. and Lawrence, J.G. (2007). Mutational bias suggests that replication termination occurs near the *dif* site, not at Ter sites. *Mol Microbiol* **64**, 42-56.
- Hidaka, M., Kobayashi, T., Takenaka, S., Takeya, H. and Horiuchi, T. (1989). Purification of a DNA replication terminus (*ter*) site-binding protein in *Escherichia coli* and identification of the structural gene. *J Biol Chem* **264**, 21031-21037.

Hill, T.M. (1992). Arrest of bacterial DNA replication. *Annu Rev Microbiol* **46**, 603-633.

Hill, T.M., Pelletier, A.J., Tecklenburg, M.L. and Kuempel, P.L. (1988). Identification of the DNA sequence from the *E. coli* terminus region that halts replication forks. *Cell* **55**, 459-466.

Hill, T.M., Henson, J.M. and Kuempel, P.L. (1987). The terminus region of the *Escherichia coli* chromosome contains two separate loci that exhibit polar inhibition of replication. *Proc Natl Acad Sci U S A* **84**, 1754-1758.

Karlin, S. (1999). Bacterial DNA strand compositional asymmetry. *Trends Microbiol* **7**, 305-308.

Kelman, L.M. and Kelman, Z. (2003). Archaea: an archetype for replication initiation studies? *Mol Microbiol* **48**, 605-615.

Kobayashi, T., Hidaka, M. and Horiuchi, T. (1989). Evidence of a *ter* specific binding protein essential for the termination reaction of DNA replication in *Escherichia coli*. *EMBO J* **8**, 2435-2441.

Kornberg, A. and Baker, T. (1991). *DNA Replication*, 2nd ed. (W.H. Freeman and Company, New York).

Kowalczyk, M., Mackiewicz, P., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Dudek, M.R., et al. (2001). DNA asymmetry and the replicational mutational pressure. *J Appl Genet* **42**, 553-577.

Liberi, G. and Foiani, M. (2004). Initiation of DNA replication: a new hint from archaea. *Cell* **116**, 3-4.

- Lobry, J.R. (1996a). Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* **13**, 660-665.
- Lobry, J.R. (1996b). Origin of replication of *Mycoplasma genitalium*. *Science* **272**, 745-746.
- Lobry, J.R. (1995). Properties of a general model of DNA evolution under no-strand-bias conditions. *J Mol Evol* **40**, 326-330.
- Lopez, P., Forterre, P., le Guyader, H. and Philippe, H. (2000). Origin of replication of *Thermotoga maritima*. *Trends Genet* **16**, 59-60.
- Mackiewicz, P., Zakrzewska-Czerwinska, J., Zawilak, A., Dudek, M.R. and Cebrat, S. (2004). Where does bacterial replication start? Rules for predicting the *oriC* region. *Nucleic Acids Res* **32**, 3781-3791.
- Marczynski, G.T. and Shapiro, L. (1993). Bacterial chromosome origins of replication. *Curr Opin Genet Dev* **3**, 775-782.
- Masters, M. and Broda, P. (1971). Evidence for the bidirectional replications of the *Escherichia coli* chromosome. *Nat New Biol* **232**, 137-140.
- Matsui, M., Oka, A., Takanami, M., Yasuda, S. and Hirota, Y. (1985). Sites of *dnaA* protein-binding in the replication origin of the *Escherichia coli* K-12 chromosome. *J Mol Biol* **184**, 529-533.
- Matsunaga, F., Forterre, P., Ishino, Y. and Myllykallio, H. (2001). In vivo interactions of archaeal Cdc6/Orc1 and minichromosome maintenance proteins with the replication origin. *Proc Natl Acad Sci U S A* **98**, 11152-11157.

- McLean, M.J., Wolfe, K.H. and Devine, K.M. (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J Mol Evol* **47**, 691-696.
- Messer, W. (2002). The bacterial replication initiator DnaA. DnaA and oriC, the bacterial mode to initiate DNA replication. *FEMS Microbiol Rev* **26**, 355-374.
- Mizraji, E. and Ninio, J. (1985). Graphical coding of nucleic acid sequences. *Biochimie* **67**, 445-448.
- Moriya, S., Atlung, T., Hansen, F.G., Yoshikawa, H. and Ogasawara, N. (1992). Cloning of an autonomously replicating sequence (ars) from the *Bacillus subtilis* chromosome. *Molecular microbiology* **6**, 309-315.
- Mott, M.L. and Berger, J.M. (2007). DNA replication initiation: mechanisms and regulation in bacteria. *Nat Rev Microbiol* **5**, 343-354.
- Mrazek, J. and Karlin, S. (1998). Strand compositional asymmetry in bacterial and large viral genomes. *Proc Natl Acad Sci U S A* **95**, 3720-3725.
- Necsulea, A. and Lobry, J.R. (2007). A new method for assessing the effect of replication on DNA base composition asymmetry. *Mol Biol Evol* **24**, 2169-2179.
- Ninio, J., and Mizraji, E. (1995). Perceptible features in graphical representations of nucleic acid sequences. In "*Visualizing Biological Information*". (Pickover, C.A., ed., World Scientific, Singapore), pp. 33-42.
- Oka, A., Sugimoto, K., Takanami, M. and Hirota, Y. (1980). Replication origin of the *Escherichia coli* K-12 chromosome: the size and structure of the minimum DNA segment carrying the information for autonomous replication. *Mol Gen Genet* **178**, 9-20.

- Pelletier, A.J., Hill, T.M. and Kuempel, P.L. (1988). Location of sites that inhibit progression of replication forks in the terminus region of *Escherichia coli*. *J Bacteriol* **170**, 4293-4298.
- Perriere, G., Lobry, J.R. and Thioulouse, J. (1996). Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid sequences. *Comput Appl Biosci* **12**, 519-524.
- Puigbo, P., Romeu, A. and Garcia-Vallve, S. (2008). HEG-DB: a database of predicted highly expressed genes in prokaryotic complete genomes under translational selection. *Nucleic Acids Res* **36**, D524-D527.
- Qin, M.H., Madiraju, M.V., Zachariah, S. and Rajagopalan, M. (1997). Characterization of the oriC region of *Mycobacterium smegmatis*. *J Bacteriol* **179**, 6311-6317.
- Richter, S., Hagemann, M. and Messer, W. (1998). Transcriptional analysis and mutation of a dnaA-like gene in *Synechocystis sp.* strain PCC 6803. *J Bacteriol* **180**, 4946-4949.
- Robinson, N.P., Dionne, I., Lundgren, M., Marsh, V.L., Bernander, R. and Bell, S.D. (2004). Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*. *Cell* **116**, 25-38.
- Rocha, E.P., Danchin, A. and Viari, A. (1999). Universal replication biases in bacteria. *Mol Microbiol* **32**, 11-16.
- Roecklein, B., Pelletier, A. and Kuempel, P. (1991). The tus gene of *Escherichia coli*: autoregulation, analysis of flanking sequences and identification of a complementary system in *Salmonella typhimurium*. *Res Microbiol* **142**, 169-175.

- Rudner, R., Karkas, J.D. and Chargaff, E. (1968). Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. Proc Natl Acad Sci U S A **60**, 921-922.
- Salazar, L., Fsihi, H., de Rossi, E., Riccardi, G., Rios, C., Cole, S.T., et al. (1996). Organization of the origins of replication of the chromosomes of *Mycobacterium smegmatis*, *Mycobacterium leprae* and *Mycobacterium tuberculosis* and isolation of a functional origin from *M. smegmatis*. Mol Microbiol **20**, 283-293.
- Salzberg, S.L., Salzberg, A.J., Kerlavage, A.R. and Tomb, J.F. (1998). Skewed oligomers and origins of replication. Gene **217**, 57-67.
- Schaefer, C. and Messer, W. (1991). DnaA protein/DNA interaction. Modulation of the recognition sequence. Mol Gen Genet **226**, 34-40.
- Schaper, S., Nardmann, J., Luder, G., Lurz, R., Speck, C. and Messer, W. (2000). Identification of the chromosomal replication origin from *Thermus thermophilus* and its interaction with the replication initiator DnaA. J Mol Biol **299**, 655-665.
- Sibley, C.D., MacLellan, S.R. and Finan, T. (2006). The *Sinorhizobium meliloti* chromosomal origin of replication. Microbiology **152**, 443-455.
- Skokotas, A., Wroblewski, M. and Hill, T.M. (1994). Isolation and characterization of mutants of Tus, the replication arrest protein of *Escherichia coli*. J Biol Chem **269**, 20446-20455.
- Sueoka, N. (1995). Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol **40**, 318-325.

- Tabata, S., Oka, A., Sugimoto, K., Takanami, M., Yasuda, S. and Hirota, Y. (1983). The 245 base-pair oriC sequence of the *E. coli* chromosome directs bidirectional replication at an adjacent region. *Nucleic Acids Res* **11**, 2617-2626.
- Thomas, J.M., Horspool, D., Brown, G., Tcherepanov, V. and Upton, C. (2007). GraphDNA: a Java program for graphical display of DNA composition analyses. *BMC bioinformatics* **8**, 21.
- Tillier, E.R. and Collins, R.A. (2000). The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J Mol Evol* **50**, 249-257.
- Wake, R.G. (1997). Replication fork arrest and termination of chromosome replication in *Bacillus subtilis*. *FEMS Microbiol Lett* **153**, 247-254.
- Worning, P., Jensen, L.J., Hallin, P.F., Staerfeldt, H.H. and Ussery, D.W. (2006). Origin of replication in circular prokaryotic chromosomes. *Environ Microbiol* **8**, 353-361.
- Xu, J., Bjursell, M.K., Himrod, J., Deng, S., Carmichael, L.K., Chiang, H.C., et al. (2003). A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. *Science* **299**, 2074-2076.
- Yee, T.W. and Smith, D.W. (1990). Pseudomonas chromosomal replication origins: a bacterial class distinct from Escherichia coli-type origins. *Proc Natl Acad Sci U S A* **87**, 1278-1282.
- Yoshikawa, H. and Ogasawara, N. (1991). Structure and function of DnaA and the DnaA-box in eubacteria: evolutionary relationships of bacterial replication origins. *Mol Microbiol* **5**, 2589-2597.

Zawilak, A., Cebrat, S., Mackiewicz, P., Krol-Hulewicz, A., Jakimowicz, D., Messer, W., et al. (2001). Identification of a putative chromosomal replication origin from *Helicobacter pylori* and its interaction with the initiator protein DnaA. *Nucleic Acids Res* **29**, 2251-2259.

Zawilak-Pawlik, A., Kois, A., Majka, J., Jakimowicz, D., Smulczyk-Krawczynsyn, A., Messer, W., et al. (2005). Architecture of bacterial replication initiation complexes: orisomes from four unrelated bacteria. *Biochem J* **389**, 471-481.

Zhang, R. and Zhang, C.T. (2002). Single replication origin of the archaeon *Methanosarcina mazei* revealed by the Z curve method. *Biochem Biophys Res Commun* **297**, 396-400.

Reprint requests to:

Albert Pallejà

Rovira i Virgili University

Department of Biochemistry and Biotechnology

Campus Sescelades

C/ Marcel·lí Domingo, s/n

E-43007 Tarragona

Catalunya, Spain

E-mail: albert.palleja@urv.cat

ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGACAGCTGATAGG
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGGCGGTGGTCAGGCAGCGCAGCATGGGAAAA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
GATCAGATGACAGATAGATAACCACAGAGACATG
CGCGCC**C**GACGCATGATTGATGATCAGATGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGATGATA**H**AGTGATTAGATAGATGGTGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTTT
GAGAGATAGAGAG**A**TAGGATCGCGCTCGAGCGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AACCGGCCAACCGGTGG**P**TTAGGATAGATGATGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AAAGTGTGTGTGACAGACAG**T**TGATGATAGTACA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGA**E**AGCTGATAG
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
TCGCGCTCGCTCGAGCGCTAGCTCGAT**R**GAT
TCGCGCTCAAACGAGCGCTAGCTCGATCGA
TGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
TATAGGTACGCGGATGAATGGCAGTAGCTAGCTT
TATCGATCGATCGATCGATCGCGCTAGCTAGCTT
TACGCATGACACACACACATCGATGATGATGAT
TCCAGGCAGCATAAAGCAGCTGGGTGGTAGGA
ATCGATCGATCGAT



OVERLAPPING GENE STRUCTURES AMONG PROKARYOTE GENOMES

.ACGCGAAAATGGC/
.GATAGAGATACAGAA/
.GTACGCGAAAATGGCAG/
TCGCTCGAGCGCTAGCTCGA/
AGGTACGCGAAAATGGCAGTA/
ATCGATCGATCGATCGATCGCGCG/
ATCAGCATGACACACACACATGATA/
CAGTGCCAGGCAGCATAAAGCAGACG/
ACCAGCAGCTGGGTGGTAGGAGTGATGT/
GCAGTGCCAGGCAGCATAAAGCAGACGAC/
GCACCAGCAGCTGGGTGGTAGGAGTGATGAT/



OVERLAPPING GENE STRUCTURES AMONG PROKARYOTE GENOMES

Albert Pallejà*, Santiago Garcia-Vallve, Antoni Romeu.

Department of Biochemistry and Biotechnology, Rovira i Virgili University,
Tarragona, Catalunya, Spain

*Corresponding author

Rovira i Virgili University
Department of Biochemistry and Biotechnology
Campus Sescelades
C/ Marcel·lí Domingo, s/n
E-43007 Tarragona
Catalunya – Spain

Email address: albert.palleja@urv.cat (corresponding author)
santi.garcia-vallve@urv.cat
antoni.romeu@urv.cat

Submitted to GENE

Abstract

Overlapping genes are genes that share either a part of or the whole coding sequence. Overlapping genes are a conserved feature of prokaryote genomes and represent the 17% of the gene pairs in these genomes. In this paper we analyze, in terms of genome organization and genome structure, the overlapping lengths, the preferred and prohibited phases, the permitted and non-permitted patterns as well as the presence and location of the Shine-Dalgarno (SD) sequence among the overlapping genes from 678 prokaryote genomes. The overlaps among the three transcriptional orientations (co-directional, convergent, and divergent) have preferred and prohibited lengths due to the restrictions imposed by the genetic code. The preferred lengths are overlaps of 1 and 4 bps among co-directional overlaps, 4 bps among convergent overlaps and 2 bps among divergent overlaps. Some of the overlapping patterns, such as ATGA in co-directional overlaps, are extremely common, but some of them are the result of wrong annotation, ribosomal frameshifting, or truncated genes. The frequency of the overlaps has a phase bias in all the three orientations and there is even a prohibited phase (phase 0) among the co-directional overlaps. Although in the co-directional and divergent overlapping genes the SD motif should be found within the coding sequence, a high percentage of overlapping genes indicate the presence of SD. Even in the divergent overlaps, the high SD presence indicates functional relevance. A good knowledge of the overlapping gene structures and the SD locations could help to improve genome annotation and may contribute to functional prediction.

Keywords: overlapping genes, overlapping lengths, overlapping phases, overlapping constraints, genome organization, Shine-Dalgarno location

1. Introduction

Overlapping genes were originally discovered in the late seventies in viruses, mitochondria and other extra chromosomal nuclear elements (Barrell et al., 1976; Sanger et al., 1977). Nowadays, they are a well-known and accepted feature of bacteriophages, animal viruses and mitochondria genomes, as well as being found in all prokaryotic genomes sequenced to date (Barrell et al., 1976; Sanger et al., 1977; Normark et al., 1983; Johnson and Chisholm, 2004). Overlapping genes have been classified into three types according to their transcriptional direction (Normark et al., 1983; Fukuda et al., 1999; Rogozin et al., 2002; Fukuda et al., 2003). These are: i) co-directional (genes in the same strand overlapping an upstream gene tail and a downstream gene head), ii) convergent (genes in opposite strands overlapping the gene tails) and iii) divergent (genes in opposite strands overlapping the gene heads). The co-directional and convergent overlaps can be caused by the loss of a stop codon in either gene, resulting in the elongation of the 3'-end of the gene's coding region. More specifically, the loss of a stop codon may result of one from the following events: i) deletion of the stop codon, ii) point mutation at the stop codon or iii) frameshift at the end of the coding region (Fukuda et al., 1999). The co-directional and divergent overlaps could arise when the downstream gene adopts a new start codon within the upstream coding sequence (Cock and Whitworth, 2007). The overlapping phenomena implies that among the co-directional and divergent overlapping genes, the regulatory signals such as the Shine-Dalgarno (SD) sequence (Shine and Dalgarno, 1974), which are needed for efficient translation among prokaryotes (Hui and de Boer, 1987; Jacob et al., 1987), must be within the coding region of the upstream gene (Eyre-Walker, 1996). Although this can hardly constrain the end coding regions of the co-directional overlapping genes, Ma and coworkers (2002) demonstrated that co-directional genes in close proximity to upstream genes are significantly higher in SD presence (Ma et al., 2002).

It has been hypothesized that overlapping genes, as a conserved feature among the prokaryote genomes, are (i) involved in compressing the maximum amount of genetic information because of evolutionary pressure to minimize genome size and increase the density of genetic information (Normark et al., 1983; Krakauer, 2000; Sakharkar and Chow, 2005; Sakharkar et al., 2005; Lillo and Krakauer, 2007); and (ii) are a mechanism for regulating gene expression through the translational coupling of functionally related polypeptides (Normark et al., 1983; Chen et al., 1990; Inokuchi et al., 2000; Johnson and Chisholm, 2004; Lillo and Krakauer, 2007). In fact, the number of genes that overlap in an organism clearly correlates to the number of ORFs in the chromosome (Fukuda et al., 2003; Johnson and Chisholm, 2004). Therefore, overlapping genes are maintained at a uniform rate across the species (Fukuda et al., 2003). In addition, the pairs of genes that overlap are better conserved across the species than the genes that do not overlap (Rogozin et al., 2002). The proportion of non-degenerate sites is higher in overlapping genes than in non-overlapping genes, thus reducing the proportion of synonymous mutations out of the total number of mutations (Rogozin et al., 2002). Therefore, a mutation could affect two proteins at the same time resulting in the loss of two functions in the cell. The significance and evolution of this conserved feature among prokaryotes has been well studied (Rogozin et al., 2002; Johnson and Chisholm, 2004; Cock and Whitworth, 2007; Kingsford et al., 2007; Lillo and Krakauer, 2007; Sabath et al., 2008). Here we analyze this phenomenon in terms of genome organization and genome structure. Therefore, the aim of this paper is to analyze the overlapping lengths, the preferred and prohibited phases, the permitted and non-permitted patterns as well as the presence and location of SD among the overlapping genes.

2. Materials and Methods

2.1 Determining the overlaps and their features

Submitted to GENE

The complete genome sequences of 678 prokaryote genomes were downloaded from the NCBI ftp site (<ftp://ftp.ncbi.nlm.nih.gov/genomes/>). Scripts in Perl language were implemented to extract and analyze the overlaps between adjacent genes. We decided to work with all the genes without excluding the ones that code for hypothetical proteins. This was because we observed the tendencies described here equally well independently of whether we removed the hypothetical ones or not. Furthermore, we found it interesting to take into account the hypothetical ones in some of our analyses. Also, we decided to include genomes with different genetic codes (Bacteria and Plant Plastic Code and Mycoplasma's code). This enriched the discussion because it enabled us to compare the overlapping patterns used with genomes that use different genetic codes.

The overlapping length distributions were represented graphically in their three possible orientations. The overlapping patterns were tabulated and examined manually. In order to study the phase bias in overlapping genes, as other authors have previously done (Cock and Whitworth, 2007; Kingsford et al., 2007; Lillo and Krakauer, 2007), we defined three overlapping phases: (i) phase 0 where the downstream gene is in the same reading frame with the upstream gene (lengths $n = \dots, -12, -9, -6, -3$), (ii) phase 1 where the downstream gene is in the reading frame +1 relative to the upstream gene frame (lengths $n = \dots, -11, -8, -5, -2$) and (iii) phase 2 where the downstream gene is in the reading frame +2 relative to the upstream gene frame (lengths $n = \dots, -10, -7, -4, -1$). Also, we used Perl scripting to obtain the frequencies of codons at the gene heads (from 1 to 10 codons) and at the gene tails (from -10 to -1 codons) of the non-overlapping genes. We focused our attention specifically on the pattern NTT, NTC, NCT followed by an ANN codon (N being any nucleotide) that causes a stop codon in the opposite strand.

2.2 Locating the SD sequences and determining their strength

We extracted the 16S rRNA sequences of the 678 prokaryote genomes analyzed from the NCBI ftp site (<ftp://ftp.ncbi.nlm.nih.gov/genomes/>). Then for each organism we looked at the 5' direction to find the first instance of the three letter motif, 5'-GAT-3', which has been consistently found at the 5'-end of all the 16S rRNA sequences. The location of this motif was used to define the 3' tail of the 16S rRNA of each organism. All the 16S rRNA tails of the 678 organisms were examined manually. If the different tails from one species did not follow a consensus, then we used the most frequent 16S rRNA tail. Following the method described by Starmer and co-workers (2006), we used free energy calculations to determine the best bindings between the 16S rRNA tails and the mRNA of the overlapping genes in order to identify and locate SD sequences. The scripts for calculating the energies were downloaded from <http://sourceforge.net/projects/freetobind> and were included in our programs. We located the SD sequence by the position of the lowest ΔG° value, which was calculated from 30 bps upstream from the initiation codon to 20 bps downstream from the initiation codon. The gene was assumed not to have a SD sequence if $\Delta G^{\circ} > -3.4535$ Kcal/mol. This threshold is based on the work of Ma and coworkers (2002). We defined as strong binding SD when between the mRNA and the 3' 16S rRNA tail the $\Delta G^{\circ} \leq -8.4$ Kcal/mol. This is the value obtained from the optimal base-pairing between the rRNA and the SD consensus sequence 5'-GGAGGU-3' (Starmer et al., 2006). For this analysis we only considered genes with a COG definition (Tatusov et al., 2000) taken from genomes that have at least 100 genes with a COG definition. Therefore we determined the SD of 90,870 overlapping pairs from 388 prokaryotic genomes.

3. Results and Discussion

3.1 Overlaps between genes

Seventeen per cent of the 1,956,294 gene pairs in the 678 genomes analyzed were overlapping. Some of the overlaps may be because the gene ends were not correctly annotated (Natale et al., 2000), particularly the long overlaps (Palleja et al., 2008). In order to minimize the effect of annotation errors, hypothetical proteins tend to be excluded when overlapping genes are analyzed. However, if we remove these genes, we observe the same tendencies shown here. Therefore, the annotation errors are only responsible for some of the overlaps between genes, although overlaps are a consistent feature of prokaryote genomes.

Among the overlaps, most of them are co-directional in orientation (87%), while fewer are convergent (11%) or divergent in orientation (3%). This shows that the most common gene orientation is co-directional because of the tendency of bacteria genes to be grouped on the same strand in operons (Normark et al., 1983; Dandekar et al., 1998; Overbeek et al., 1999; Salgado et al., 2000). In addition, the divergent orientation for an overlap is the most constrained because the regulatory sequences of both genes are overlapping (Rogozin et al., 2002; Fukuda et al., 2003). As we have mentioned above, co-directional overlaps can arise through 3'-end extensions as well as 5'-end extensions. The 3'-end extension needs a mutation that disrupts the stop codon, whereas the 5'-end extension can occur without any mutation, and only needs the downstream gene to adopt an upstream start codon in frame. Therefore, the change in the 5'-end extensions is a simpler mechanism that could gradually occur through evolution (Cock and Whitworth, 2007) and which also contributes to the fact that the co-directional overlaps are the most frequent. Convergent overlaps can only arise through a 3'-end extension, which is a more complex mechanism. Divergent orientation potentially places the SD sequences and the start codons of both genes in the overlapping region. This is highly constraining for the divergent overlaps and thus these overlaps occur least frequently.

Figure 1 shows the distribution of overlapping lengths in the three transcriptional orientations. As a general trend, the number of overlaps decreases as the overlapping length increases, according to the selective



pressure against long overlaps. It is also worth mentioning, as other authors have previously noted (Fukuda et al., 2003; Johnson and Chisholm, 2004; Cock and Whitworth, 2007; Lillo and Krakauer, 2007), that overlaps of one and four nucleotides are extremely common, especially the 4 bps co-directional overlap which includes, in the overlapping region, the start and the stop codon of both genes favoring translational coupling (McCarthy, 1990).

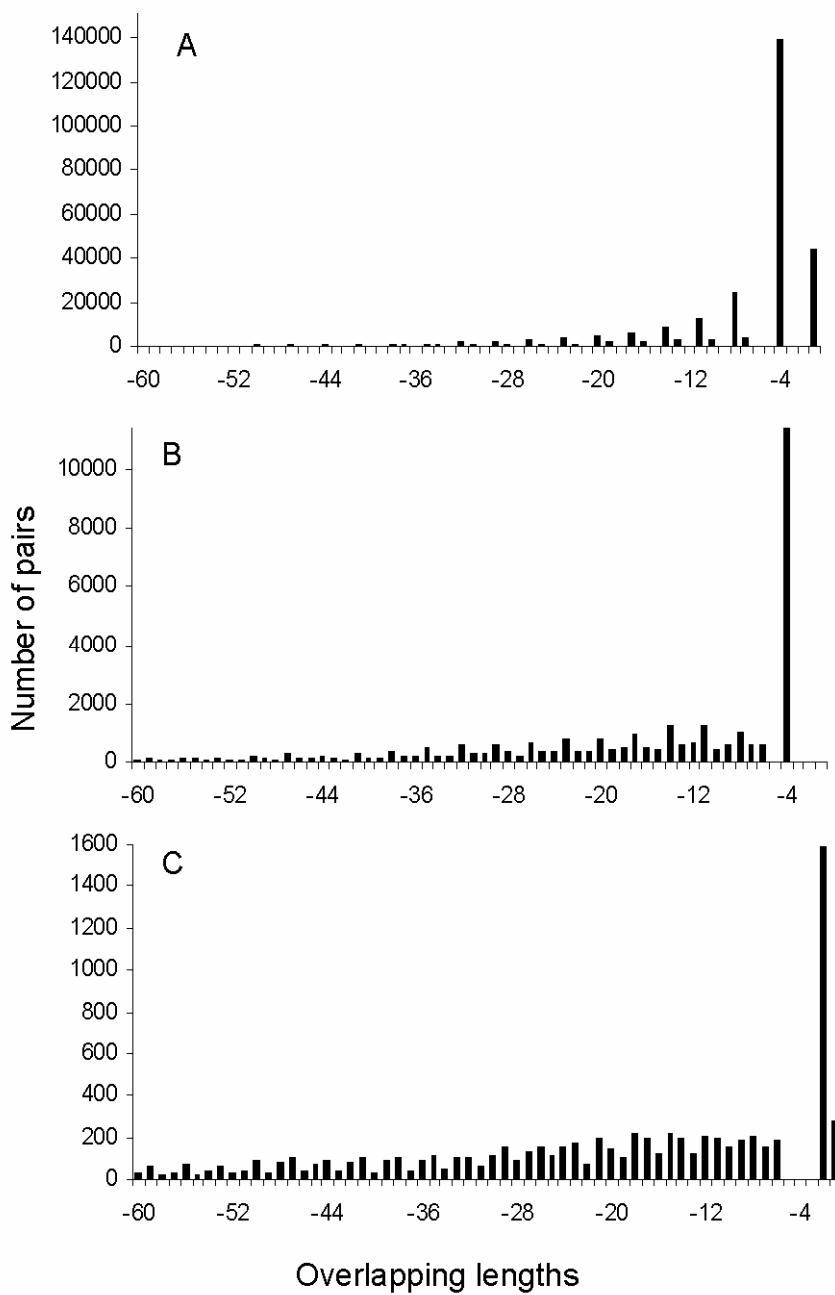


Fig. 1. Overlapping lengths distribution

Submitted to GENE

Overlapping lengths distribution in each orientation: (A) Co-directional, (B) Convergent, (C) Divergent.

3.1.1 Co-directional overlaps

Fig. 1A shows the distribution of co-directional overlapping lengths. These overlaps are the most affected by annotation errors because they are associated with the misprediction of the start codons at the 5'-ends (Rogozin et al., 2002; Kingsford et al., 2007; Lillo and Krakauer, 2007; Pallejà et al., 2008). In fact, excluding hypothetical or putative genes from an analysis does not guarantee that the gene starts have been correctly predicted (Pallejà et al., 2008). Thus, we obtained a similar distribution (data not shown) when we only analyzed genes with a predicted function, i.e. genes with a KEGG id (Kanehisa et al., 2006). Among co-directional overlaps there is a spike at 4 bps (Fig. 1A). This overlap is the most frequent and most likely to arise through random mutations, leading to the formation of a 4 bps overlap which includes the downstream gene start and the upstream gene stop codon (Johnson and Chisholm, 2004; Lillo and Krakauer, 2007). Thus, a ribosomal frame shifting is then required for the transcription of the downstream gene (Lillo and Krakauer, 2007).

3.1.1.1 Analysis of co-directional overlapping lengths and patterns

The most common overlapping pattern for 4 bps is ATGA (Fig. 2). This overlapping pattern has few constraints because a) it allows the penultimate amino acid in the upstream sequence to be one of the following 12 amino acids: L, S, P, Q, R, I, T, K, V, A, E and G and b) because it allows the second amino acid in the downstream sequence to be one of the following seven amino acids: I, M, T, N, K, S and R (Kozlov, 2000). Figure 2 shows other 4 bps overlapping patterns such as GTGA and TTGA. These patterns include the GTG and TTG initiation codons, which are the most frequent non-canonical start codons (Kozak, 1983; Fotheringham et al., 1986; Golderer et al., 1995; Nolling et al.,

1995; Sazuka and Ohara, 1996; Blattner et al., 1997; Genser et al., 1998; Wang et al., 2003). Other 4 bps overlapping patterns are CTGA, ATAA and ATAG (Figure 2). These patterns require NTG or ATN as their initiation codon, which are allowed in Archaea and Bacteria (Polard et al., 1991; Liveris et al., 1993; Sazuka and Ohara, 1996; Binns and Masters, 2002). Rare initiation codons might be used by low translated genes or in ribosomal frameshifting translations. Therefore, the use of non-canonical initiation codons could be related to the regulation of gene expression (Ma et al., 2002; Starmer et al., 2006). Some of the 4 bps overlapping patterns (TTAA, CTAA, ATAG and TTAG) are used mainly by Mycoplasmas because these species are able to use other codons as start codons. Other unexpected 4 bps overlapping patterns are AAGG, GTAA and GTAG. These patterns are extremely uncommon and correspond to fragments of genes and doubtful coding sequences (in gene pairs NC_006513.ebD66 & NC_006513.ebD67, NC_003116.NMA1601a & NC_003116.NMA1601b, NC_002952.SAR1930 & NC_002952.SAR1930a, respectively). We can not be sure without experimental data whether these genes have rare initiation or stop codons or are merely the product of annotation errors.

Co-directional overlapping pattern schemas	Overlapping lengths (bps)	Number of pairs	Phase	Overlapping constraints	Actual Overlapping patterns and number of pairs of each
	1	43,886	2	Small constraints. However TTG cannot be used as start codon	A (42,095) G (1,790)
	2	7*	1	Upstream stop codons give dinucleotides that provide rare start codons for a downstream gene	* overlaps of truncated genes, hypothetical fragments or genes participating in frameshifts
	3	0	0	Not allowed. Stop codon can not be read as a start codon	-
	4	138,775	2	Small constraints. However only one stop codon is allowed (TGA). D: A, G or T	ATGA (106,538) GTGA (25,430) TTGA (6,507) CTGA (180) ATAA (84) ...
	5	28*	1	Rare start codon (finished in a thymine) for a gene. X: F, S, Y, C, I, P, H, R, I, T, N, S, V, A, D or G	* Most of them use Mycoplasma's genetic code. Others are truncated genes and possible annotation errors
	6	0	0	Not allowed. X: M, V or L	-
	7	3,695	2	Adenine is not allowed at the fourth overlapping base. D: A, G or T; B: C, G, or T; X: C or W; X': V or L	ATGCTGA (612) ATGGTAA (439) ATGTAA (353) ATGGTGA (311) ATGTTAG (244) ...
	8	24,106	1	Small constraints. D: A, G or T; X: V, A, D, E or G; X': F, S, Y, C, L, P, H, R, I, T, N, S, V, A, D or G	ATGAATAA (1,751) ATGGCTGA (1,404) ATGAGTAA (1,199) ATGACTGA (1,177) ATGAGTGA (1,051) ...
	9	0	0	Not allowed. D: A, G or T; X: M, V or L; X': all the aminoacids	-
	10	2,868	2	Adenine is not allowed at the fourth overlapping base. D: A, G or T; B: C, G, or T; X: C or W; X': all the aminoacids; X'': F, L, S, Y, C, W, L, P, H, Q, R, V, A, D, E or G; X''': M, L, I or V	ATGGAACCTGA (43) ATGCAACTGA (37) ATGGAATAA (33) ATGCCTCTGA (29) ATGTGCATAG (29) ...
	11	12,494	1	Small constraints. D: A, G or T; X: V, A, D, E or G; X': all the aminoacids; X'': F, S, Y, C, L, P, H, R, I, T, N, S, V, A, D or G	ATGAAAAATAA (109) ATGAAAAATTA (106) ATGAAAAATTA (69) ATGAAAAATTA (66) ATGAAAGCTAA (53) ...

Fig. 2. Co-directional overlapping patterns

Co-directional overlapping patterns observed in genes from prokaryote species. For each overlapping length up to 11 bps, the figure shows the overlapping pattern schemas, the overlapping lengths, the numbers of pairs that overlap that length, the overlapping

phases, the overlapping constraints of that overlap and the real overlapping patterns. Notice that the overlaps which are not allowed have their overlapping lengths and the overlapping phases highlighted in red.

The 1 bps overlap is the second most frequent co-directional overlap (Fig. 1A). For this overlap, only 2 patterns (A or G) seem to be allowed. The most frequent one is A, the upstream gene stop codon being either TGA (TG[A]TG) or TAA (TA[A]TG) and the downstream gene start codon being the canonical one (ATG). In the second most frequent pattern, the stop codon of the upstream gene can only be TAG (TA[G]TG) and the downstream gene start codon must be GTG, which is much less used than the canonical one. This explains the lesser frequency of the G pattern. Surprisingly, we found one case of 1 bps overlap with the T pattern (CC[T]TT). This case corresponds to two fragments of hypothetical genes (NC_006513.ebD57 & NC_006513.ebD58) found in *Aromatoleum aromaticum* EbN1.

According to Cock and Whitworth (2007), overlaps of either 2 or 5 bps are not permitted because of the structure of the genetic code. However, we found some 2 and 5 bps overlaps (Fig. 2). On one hand, among the 2 bps overlaps, one or both overlapping genes are truncated, hypothetical fragments of a whole gene or genes participating in frameshifts (NC_003116.NMA0344 & NC_003116.NMA0344A, NC_003116.NMA0644 & NC_003116.NMA0644a, NC_003116.NMA1907A & NC_003116.NMA1907B, NC_006513.ebD55 & NC_006513.ebD56, NC_003888.SCO2680 & NC_003888.SCO2681, NC_009482.SynRCC307_1935 & NC_009482.SynRCC307_1936 and NC_009482.SynRCC307_2248 & NC_009482.SynRCC307_2249). These 2 bps overlapping genes are only found in the genomes of *Neisseria meningitidis* Z2491, *Aromatoleum aromaticum* EbN1, *Streptomyces coelicolor* A3(2) and *Synechococcus* sp. RCC307. On the other hand, among the 5 bps overlaps, we found 2 overlapping patterns such as ATTGA and ATTAG which are mainly used by *Mycoplasma* species that can use ATT as a start codon. We also found

2 patterns such as ATGGG and GCGCA which correspond to the overlapping pairs NC_003295.RSc0869 & NC_003295.RSc0870 and NC_003116.NMA1775 & NC_003116.NMA1776 from *Ralstonia solanacearum* GMI1000 and *Neisseria meningitidis* Z2491 respectively, with very rare stop codons in both and one rare start codon in the *N. meningitidis* Z2491 pair. Both pairs could also be truncated genes or the product of annotation errors.

In summary, we observed the expected overlapping patterns, taking into account the constraints of the genetic codes (i.e. bacterial, plant plastid and *Mycoplasma* codes). However, other overlapping patterns are found in gene pairs that could be wrongly annotated, could be participating in frameshifting translations or could be small fragments of a whole gene. The genomes that have rarer overlapping patterns are *Neisseria meningitidis* Z2491, *Aromatoleum aromaticum* EbN1 and *Synechococcus* sp. RCC307.

3.1.1.2 Analysis of co-directional overlapping phases

There are only 2 reading frames (phase 1 and 2) where co-directional gene pairs can overlap (Fig. 1A and Fig. 2). There are no 3 bps or multiples of 3 bps co-directional overlaps. In fact, a 3 bps overlap between a co-directional gene pair does not make sense because the upstream stop codon cannot work as a downstream start codon. Short overlaps of a multiple of 3 bps would generate downstream peptides which would be too small and without any known function or would require an evolutionarily unstable stop codon read-through (Keese and Gibbs, 1992; Krakauer, 2000). Long overlaps of a multiple of 3 bps would correspond to a redundant gene prediction, i.e. they are embedded genes. These observations suggest that phase 0 is not allowed between co-directional overlapping genes. Regarding the other two phases, recent studies have demonstrated that for long overlaps (7 bps or longer) overlapping pairs are more frequent in phase 1 than in phase 2 (Johnson and Chisholm, 2004; Cock and Whitworth, 2007; Sabath et al., 2008) although it has been predicted that there is no phase preference for co-directional overlaps (Krakauer, 2000). Our data agrees with this phase bias (Table 1) because phase 1 is more prevalent

than phase 2 by a factor of almost 4 if we only take into account overlaps no longer than 60 bps. For long overlaps, in phase 2 we observe one restriction more than in phase 1. The overlapping fourth base could not be an adenine (see the overlaps of 7 and 10 bps in Fig. 2). This could help us to understand the phase bias towards phase 1. Sabath and coworkers (2008) have shown recently that there is a higher frequency of alternative start codons in phase 1 than in phase 2. This higher frequency of alternative start codons explains the preference for long overlaps in phase 1, if we take into account that there are more co-directional overlaps originated by 5'-end extensions. Since there is a selective pressure against long overlaps, the low frequency of start codons in phase 2 constrains the number of overlaps created in that phase, leading to the phase bias (Sabath et al., 2008).

Another fact worth mentioning is that within the phase 1 we found long overlaps such as 8 or 11 bps that were quite common and have fewer constraints than the short overlaps (1 and 4 bps overlaps) found in phase 2. Furthermore the 1 and 4 bps overlaps have constraints related to the either the start or the stop codon that long overlaps do not have (Figure 2). However, the 8 and 11 bps overlaps are less represented probably due to both the easier formation of 1 and 4 bps by neutral point mutations and the selective pressure against long overlaps.

3.1.2 Convergent overlaps

3.1.2.1 Analysis of convergent overlapping lengths and patterns

Convergent overlaps, as with co-directional overlaps, show a spike in 4 bps overlaps (Fig. 1B). The tetramers CTAG, TTAA, TTAG and CTAA are the most frequent among the convergent overlapping patterns (Fig. 3). In contrast with the most frequent co-directional overlapping tetramers, the convergent overlapping tetramers do not include the stop codon TGA. Instead, the 4 bps convergent overlapping genes use the TAA and TAG stop codons, in addition to either a C or a T to generate patterns that contain a stop codon for both

overlapping genes. The patterns TTAA and CTAG are the most abundant and curiously generate palindromic sequences. As in the co-directional overlaps, several convergent short overlaps, for example 1, 2, 3 and 5 bps, are not allowed because of the constraints imposed by the structure of the genetic code. In such overlaps there is an incompatibility between the forward stop codons (TAA, TAG and TGA) and their reverse complementary sequences (TTA, CTA and TCA) which do not form any stop codon (Fig. 3). Among the 1, 2, 3 and 5 bps overlaps, we found only one convergent 1 bp overlap (NC_003143.YPO4025 & NC_003143.YPO4026) that overlapped the A nucleotide. One of the genes of that pair is a gene remnant carrying the rare stop codon CCT. We found fewer cases of disallowed patterns among convergent than among co-directional overlaps because stop codon usage (only TAA, TAG and TGA) is stricter than the start codon usage among prokaryotes. Also, this might be because the misprediction of start codons occurs more easily than the misprediction of stop codons (Rogozin et al., 2002).

Convergent overlapping pattern schemas	Overlapping lengths (bps)	Number of pairs	Phases	Overlapping constraints	Actual Overlapping patterns and number of pairs of each
	1	1*	2	Not allowed The last nucleotide of the stop codon in the upstream gene does not provide a proper stop codon in the opposite strand. R: A or G and Y: C or T	*overlap of a gene remnant
	2	0	1	Not allowed The last two nucleotides of the stop codon in the upstream gene does not provide a proper stop codon in the opposite strand. Y: C or T	-
	3	0	0	Not allowed None stop codon in the upstream gene causes a stop codon in the downstream gene. Y: C or T	-
	4	11,397	2	Only one stop codon is not allowed (TGA). R: A or G; Y: C or T	CTAG (3,557) TTAA (3,315) TTAG (2,279) CTAA (2,245)
	5	0	1	Not allowed The first nucleotide of the stop codon in the upstream gene does not provide a proper stop codon in the opposite strand. Y: C or T	-
	6	571	0	Small constraints: R: A or G; Y: C or T; X: P, L or S	TTATAA (116) TTATGA (73) TCATGA (73) CTATAG (67) TCATAA (51) ...
	7	601	2	Small constraints: R: A or G; Y: C or T X: H, Q, or Y	CTACTGA (84) TTACTGA (80) TCAGTAG (72) TCAGTGA (60) TCACTGA (57) ...
	8	1,029	1	Small constraints: R: A or G; Y: C or T; X: M, I, T, N, K, S or R	TCAGCTGA (77) TTAAATAA (40) TTATTAA (33) TCAAGTAG (26) TCAGTTGA (26) ...
	9	611	0	Small constraints: R: A or G; Y: C or T; X: all the aminoacids; X': P, L or S	TCAGGCTGA (15) TCAGCATGA (14) TTAAAATAA (14) TTATTTAA (11) TCAGCTGA (10) ...
	10	459	2	Small constraints: R: A or G; Y: C or T; X: all the aminoacids; X': H, Q, or Y	TCAGGCTGA (15) TCAGGCTGA (14) TTACGCTGA (9) TTATTTAA (9) CTATATCTGA (9) ...
	11	1,267	1	Small constraints: R: A or G; Y: C or T; X: all the aminoacids; X': M, I, T, N, K, S or R	TCAGGCTGA (9) CTATGCTATAA (9) TCAGGCTGA (9) TCAGGCTGA (9) CTATTTAA (8) ...

Fig. 3. Convergent overlapping patterns

Convergent overlapping patterns observed among genes from prokaryote species. For each overlapping length up to 11 bps, the figure shows the overlapping pattern schemas, the overlapping lengths, the numbers of pairs that overlap that length, the overlapping phases, the overlapping constraints of that overlap and the real overlapping patterns.

Submitted to *GENE*

Notice that the overlaps which are not allowed have both their overlapping lengths and their overlapping phases highlighted in red. The features of one of the overlapping genes are represented in green to highlight that the genes are in the opposite strand.

3.1.2.2 Analysis of convergent overlapping phases

Another difference between convergent and co-directional overlaps is that convergent overlaps can overlap in the three different phases. However, as we can see in Fig. 3 and Table 1, there is also a phase bias. Among the short overlaps (no longer than 7 bps) phase 2 is more prevalent than phase 1 because the 4 bps overlaps belong to phase 2. Phase 0 is allowed in convergent overlaps because no forward-stop codon generates a stop codon, in frame, in the reverse strand. Although phase 0 is used by short and long overlaps, it is the least preferred phase. Among the long overlaps (lengths ≥ 7 bps) phase 1 is more prevalent than phase 2, which in turn is more prevalent than phase 0 (Kingsford et al., 2007). It has been suggested that the overlapping lengths and the phase bias in convergent orientation may be explained by the frequency of the stop codons (Kingsford et al., 2007). Similarly, as with the different distribution of start codons within the phases in the co-directional overlaps, since there is a purifying selection against long overlaps, the long overlaps would be more likely to accumulate in those phases where start codons are more frequent and therefore more likely to be closer, after a mutation in the stop codon (Kingsford et al., 2007). However, the overlapping lengths and the phase bias among the convergent overlaps can also be explained by selection, as proposed by Rogozin and coworkers (2002). They explained the prevalence of phase 1 because the second codon positions, in which all mutations lead to amino acid replacements, are located opposite to the degenerate third codon positions of the complementary coding region (123:132). This ensures the most independent evolution of the two coding sequences (Rogozin et al., 2002). Both explanations would help to explain the phase 1 > phase 2 > phase 0 bias observed among long convergent overlaps (Table 1).

Co-directional overlaps	Phase 0	Phase 1	Phase 2	Total
Overlapping lengths < 7 bps	0	35	182,661	182,696
Overlapping lengths ≥ 7 bps	0	74,054	19,945	93,999
Total	0	74,089	202,606	276,695
Convergent overlaps	Phase 0	Phase 1	Phase 2	Total
Overlapping lengths < 7 bps	571	0	11,398	11,969
Overlapping lengths ≥ 7 bps	4,655	10,282	5,395	20,332
Total	5,226	10,282	16,793	32,301
Divergent overlaps	Phase 0	Phase 1	Phase 2	Total
Overlapping lengths < 7 bps	188	1,585	277	2,050
Overlapping lengths ≥ 7 bps	2,095	2,308	1,322	5,725
Total	2,283	3,893	1,599	7,775

Table 1. Distribution of the co-directional, convergent and divergent overlaps among the phases

Number of overlapping pairs < or ≥ 7 bps found in phases 1 and 2 among the co-directional, convergent and divergent overlaps. For the overlaps ≥ 7 bps we considered the overlaps up to 60 bps.

Interestingly, within phase 1, the overlaps of 11 and 14 bps are more frequent than those of 8 bps, and this is an issue that remains unsolved (Kingsford et al., 2007; Lillo and Krakauer, 2007). To solve it, we analyzed the frequency of the codon combinations between NTT or NTC or NCT and ANN (N being any of the 4 nucleotides ACTG) at the tail of non-overlapping genes (Table 2). These are the combinations at the end of a gene tail that generate a stop codon on the reverse strand in phase 1. Table 2 shows that these codon combinations occur less frequently at codons -3 & -2 (the forward stop codon corresponds to position -1) than at codons -4 & -3, -5 & -4 and -6 & -5. The highest frequency of the codon combinations that generates a stop codon on the reverse strand occurs at codons -4 & -3 and to a lesser extent at codons -5 & -4 (Table 2). This explains why the convergent overlaps of 11 and 14 bps are more frequent than those of 8 bps (Table 2). Beyond 14 bps overlaps, the number of

overlaps decreases due to both the selective pressure against long overlaps and to the decrease in the frequency of the codon combinations that generate a stop codon in the reverse strand in phase 1.

Codon -7	Codon -6	Codon -5	Codon -4	Codon -3	Codon -2	Codon -1	Overlapping length (bps)	Number of patterns which may generate stop codons
NNN	NNN	NNN	NNN	NTT NTC NCT	ANN	TAA TGA TAG	8	68,887
NNN	NNN	NNN	NTT NTC NCT	ANN	NNN	TAA TGA TAG	11	82,939
NNN	NNN	NTT NTC NCT	ANN	NNN	NNN	TAA TGA TAG	14	70,484
NNN	NTT NTC NCT	ANN	NNN	NNN	NNN	TAA TGA TAG	17	69,570
NTT NTC NCT	ANN	NNN	NNN	NNN	NNN	TAA TGA TAG	20	62,125

Table 2. Distribution of the pattern NTT or NTC or NCT and ANN at the gene tails

Distribution of the pattern NTT or NTC or NCT and ANN among the codons of the gene tails that not overlap. We only counted the patterns that could generate a stop codon in the opposite strand in case of overlap (NTT or NTC or NCT and ANN). Codon -1 represents the stop codon.

3.1.3 Divergent overlaps

In contrast to the distribution of the co-directional and convergent overlapping lengths, the most common divergent overlap is 2 bps (Fig. 1C), the dinucleotide AT being the most frequent divergent overlapping pattern. This pattern provides the beginning of the canonical start codon (ATG) in both strands (Fig. 4). The 3, 4 and 5 bps divergent overlaps are not allowed because

of the structure of the genetic code. In these overlaps, there is an incompatibility between the forward start codons (ATG, GTG, and TTG) and their reverse complementary sequences (CAT, CAC and CAA), which do not form any start codon (Figure 4). Only two cases of 4 bps overlaps were found with the rare overlapping patterns ATAT and CCAC. One was found in *Rickettsia massiliae* MTU5 (NC_009900.RMA_1363 & NC_009900.RMA_1364) and the other in *Mycobacterium tuberculosis* H37Rv (NC_000962.Rv2810c & NC_000962.Rv2811), where one of the genes is probably a fragment of a transposase. The second most frequent divergent overlap is 1 bps, A or T being the most frequent patterns. These patterns provide the beginning of the start codons ATG or TTG in both strands. However we found 4 cases which overlapped the nucleotide G in the species *Methanopyrus kandleri* AV19 (NC_003551.MK0975 & NC_003551.MK0976), *Synechococcus* sp. WH 8102 (NC_005070.SYNW2161 & NC_005070.SYNW2162), *Mycobacterium avium* subsp. *paratuberculosis* K-10 (NC_002944.MAP3621c & NC_002944.MAP3622) and *Mycobacterium tuberculosis* F11 (NC_009565.TBFG_12824 & NC_009565.TBFG_12825). This overlap is extremely rare because it generates the uncommon start codon CTG.

Divergent overlapping pattern schemas	Overlapping lengths (bps)	Number of cases	Phases	Overlapping constraints	Actual Overlapping patterns and number of pairs of each
	1	275	2	Small constraints. Only one start codon is not allowed (GTG). W: A or T	A (150) T (121)
	2	1585	1	Small constraints. However, only the canonical start codon (ATG) is allowed.	AT (1585)
	3	0	0	Not allowed None start codon in the upstream gene provides a start codon in the opposite strand. H: A, C or T	-
	4	2*	2	Not allowed The last two nucleotides of the start codon in the upstream gene do not provide a proper start codon in the opposite strand. D: A, T or G; H: A, C or T	*overlaps of a fragment of a gene and a gene coding for an hypothetical protein
	5	0	1	Not allowed The last nucleotide of the start codon in the upstream gene does not provide a proper start codon in the opposite strand. D: A, T or G; H: A, C or T	-
	6	188	0	Small constraints. D: A, T or G; H: A, C or T; X: H or Q	ATGCAT (53) TTGCAT (31) ATGCAA (29) GTGCAT (23) ATGCAC (20) ...
	7	159	2	Small constraints. D: A, T or G; H: A, C or T; X: S,P,T or A	ATGGCAT (13) ATGCCAT (12) ATGTCAT (11) ATGACC (10) TTGCCAT (10) ...
	8	204	1	Small constraints. D: A, T or G; H: A, C or T; X: F,S,Y,C,L,P,H,R,I,T,N,S,V,A,D or G	TTGTCAT (13) ATGAGCAT (7) ATGGACAT (7) ATGGCCAT (7) ATGTCCAT (7) ...
	9	188	0	Small constraints. D: A, T or G; H: A, C or T; X: all the aminoacids; X': H or Q	TTGTCAT (6) ATGAAGCAT (4) ATGAATCAT (4) ATGGCCAT (4) ATGGAACAT (4) ...
	10	151	2	Small constraints. D: A, T or G; H: A, C or T; X: all the aminoacids; X': S,P,T or A	ATGCCATCAT (10) ATGATAGCAC (8) ATGATGGCAT (8) ATGACCACAT (4) ATGACAGCAT (4) ...
	11	194	1	Small constraints. D: A, T or G; H: A, C or T; X: all the aminoacids; X': F,S,Y,C,L,P,H,R,I,T,N,S,V,A,D or G	ATGACTAACAA (5) ATGTTATTCAT (5) ATGAAAAACAT (2) ATGAAAGCCAA (2) ATGATGAGCAA (2) ...

Fig. 4. Divergent overlapping patterns

Divergent overlapping patterns observed among genes from prokaryote species. For each overlapping length up to 11 bps, the figure shows the overlapping pattern schemas, the overlapping lengths, the numbers of pairs that overlap that length, the overlapping phases, the overlapping constraints of that overlap and the real overlapping patterns.

Submitted to GENE

Notice that the overlaps which are not allowed have both their overlapping lengths and their overlapping phases highlighted in red. The features of one of the overlapping genes are represented in green to highlight that the genes are in the opposite strand.

The phase bias in divergent overlaps is slightly different than in co-directional or convergent overlaps. In the short (no longer than 7 bps) overlaps, phase 1 is more prevalent than phases 0 and 2 (Table 1). This is because the most frequent divergent overlap is the 2 bps overlap belonging to phase 1. Nevertheless, in the co-directional and convergent overlaps, phase 1 was prohibited (or was allowed in only co-directional overlaps of some species that use the *Mycoplasma*'s genetic code). Long overlaps show the following bias: phase 1 > phase 0 > phase 2. According to the hypothesis of Sabath and coworkers (2008), the frequency of start codons could also explain the phase bias for the divergent overlaps. Since the divergent overlaps arise by 5'-end extensions and there is a selective pressure against long overlaps, the long overlaps would be more likely to accumulate in those phases where start codons are more frequent and therefore more likely to be closer. Therefore, it could be hypothesized that the start codon frequency follows the bias phase 1 > phase 0 > phase 2.

3.2 Location and presence of Shine-Dalgarno (SD) sequences

Both co-directional and divergent overlapping genes must have their regulatory sequences within the overlapping region. The SD sequence must be at the end of the upstream gene in co-directional overlapping genes, whereas it must be at the beginning of the upstream gene in divergent overlapping genes (Eyre-Walker, 1996; Ma et al., 2002). The convergent overlapping genes are not constrained by the SD presence because they overlap at their 3'-ends. Table 3 shows that overlapping genes tend to keep their SD sequences. Almost 73 % of co-directional overlapping genes with a short overlap (overlapping length \leq 60 bps) have a predicted SD sequence. Usually these SD sequences are, as

expected, upstream from the initiation codon of the downstream gene. However, 26.8 % of the co-directional overlapping genes have a predicted SD sequence downstream from the initiation codon. This unexpected location is probably due to a mis-annotation of the initiation codon (Starmer et al., 2006). Among the long co-directional overlaps (overlapping lengths > 60 bps), which are even more likely to result from mis-annotations because of the selective pressure against long overlaps (Pallejà et al., 2008), the SD presence decreases to around ~4 % compared to the short overlaps. Again, a possible explanation for this decrease is that many of the genes with a long overlap are incorrectly annotated (Ma et al., 2002). Indeed, we observed that among the long overlaps, the overlapping genes that keep their SD sequence upstream to the initiation codon decrease, while those that keep their SD sequence downstream from the initiation codon increase substantively to around 15 % (Table 3). Therefore, although the genome annotation errors can affect both the short and the long co-directional overlaps, there are more mispredictions of the initiation codons among the long overlaps. Also, it was worth studying whether the SD strength was compromised because it was in a coding region. We observed that among the co-directional overlapping genes (short and long overlaps) with a predicted SD, ~18 % of genes show a strong SD (Table 3). This percentage is similar to the percentage of non-overlapping genes that have a strong SD. Therefore, the strength of the SD does not depend on overlapping or the overlapping lengths. These results agree with the finding that both the SD strength and its relative position to the initiation codon are not affected by the fact that SD are located within a coding region (Eyre-Walker, 1996).

	<i>Co-directional overlaps</i>	
	<i>SD presence</i>	
	<i>Short overlaps (<= 60 bps)</i>	<i>Long overlaps (> 60 bps)</i>
<i>SD sequence</i>	72.8 %	69.2 %
<i>Upstream SD</i>	73.2 %	58.3 %
<i>Downstream SD</i>	26.8 %	41.7 %
<i>Strong SD</i>	17.7 %	18.4 %
<i>No SD sequence</i>	27.2 %	30.8 %
	<i>Divergent overlaps</i>	
	<i>SD presence</i>	<i>% of TRX structure</i>
<i>None gene with SD</i>	15.5 %	3.0 %
<i>One gene with SD</i>	41.9 %	7.3 %
<i>Both genes with SD</i>	42.6 %	10.8 %

Table 3. SD presence among the co-directional and divergent overlaps

A percentage of the SD presence among the short co-directional (lengths ≤ 60 bps) and the long co-directional overlaps (longer than 60 bps). The upstream SD category is the percentage of genes whose SD sequence is predicted upstream from the initiation codon, whereas the downstream SD category is the percentage of genes whose SD sequence is predicted downstream from the initiation codon. The second part of the table shows the percentages of overlapping pairs that conserve none, one or both SD sequences among the divergent overlaps. Also there is a column dedicated to the percentage of TRX structures (in which one gene encodes a transcriptional regulator, TR, and the other gene, X, encodes any other class of protein) among the three gene sets of divergent overlaps described above.

Divergent overlaps have more constraints imposed by regulatory signals such as the SD sequences. Therefore we would expect a low frequency of SD sequences among these overlaps. However, among the divergent overlapping pairs, 42.6 % of them have a predicted SD sequence in both genes and a 41.9 % have a predicted SD sequence in at least one of the genes of the pair (Table 3). Therefore, SD must be present in divergent overlapping genes, even in the high constraints they show. In an analogy with operons, coregulation is an evolutionary constraint that could maintain divergent gene pairs (Korbel et al.,

2004). Therefore, within divergent gene pairs, there is a strong enrichment of pairs in which one gene encodes a transcriptional regulator (TR) and the other gene (X) encodes any other class of protein (Korbel et al., 2004). This TRX structure is found in a high percentage of the divergent overlaps, particularly in overlaps where both genes show a SD sequence (10.8 %) and the overlaps where only one gene of the pair shows a SD sequence (7.8 %) (Table 3). In contrast, the TRX structure is very rare among the co-directional overlaps, only 0.1 % of them show this structure. This strong evolutionary conservation of divergently transcribed gene pairs, which is also evident in the divergent overlaps, implies biological relevance.

4. Conclusions

Gene overlaps arise in all the three transcriptional orientations with extremely common and prohibited overlapping lengths resulting from the structure of the genetic code and strong selective pressure against long overlaps. The majority of the overlapping patterns can be properly explained by the restrictions of the genetic code of bacteria or Mycoplasma. However some overlapping patterns seem to be the product of mispredicting gene ends, genes participating in frameshifts, truncated genes or possible fragments of genes. Both the correct and non-correct overlapping patterns should be taken into account for subsequent genome annotations. Overlaps may have preferred or prohibited phases, such as phase 0 in co-directional overlaps. The phase bias seems to result from the frequencies of initiation and termination codons within the three phases (Kingsford et al., 2007; Sabath et al., 2008), even though selection also influences the phase bias (Krakauer, 2000; Rogozin et al., 2002; Lillo and Krakauer, 2007). Here we have tried to explain some of the constraints imposed by the structure of the genetic code as well as the codon usage of the gene flanks.

The overlapping genes have predicted SD sequences just as the non-overlapping ones do. In addition, a relevant percentage of overlapping genes

are predicted to have a strong SD. This means that genes may overlap regardless of their expression level. The long co-directional overlaps show an overall decrease in the SD presence. However, they show an increase in SD prediction downstream from the initiation codon. This means that the long overlaps may be caused by annotation errors as we have previously pointed out (Pallejà et al., 2008). These annotation errors are exposed by the SD predictions (Starmer et al., 2006). A high proportion of the divergent overlapping genes have predicted SD sequences, even though these genes are the most constrained because of the location of the regulatory sequences within a coding region. Within divergent overlaps, the gene structure in which one gene encodes a transcriptional regulator (TR) and the other gene (X) encodes any other class of protein, is overrepresented. Therefore, instead of a rare genome feature, divergent gene pair overlaps may be conserved structures of coregulated genes where a transcriptional regulator regulates their overlapping gene.

Understanding the overlapping phase bias, the preferred and prohibited overlapping patterns and lengths could be a powerful tool for functional prediction as well as for improving genome annotation.

Acknowledgments

This work has been supported by projects BIO02003-07672 and AGL2007-65678/ALI of the Spanish Ministry of Education and Science. We would like to thank the Language Service of the Rovira i Virgili University for their help in writing the manuscript.

References

- Barrell, B.G., Air, G.M. and Hutchison, C.A., 3rd Overlapping genes in bacteriophage phiX174. *Nature* **264** (1976), pp. 34-41.
- Binns, N. and Masters, M. Expression of the Escherichia coli pcnB gene is translationally limited using an inefficient start codon: a second chromosomal example of translation initiated at AUU. *Mol Microbiol* **44** (2002), pp. 1287-98.
- Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J.,

- Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. The complete genome sequence of *Escherichia coli* K-12. *Science* **277** (1997), pp. 1453-74.
- Chen, S.M., Takiff, H.E., Barber, A.M., Dubois, G.C., Bardwell, J.C.A. and Court, D.L. Expression and characterization of RNase-III and Era proteins - products of the *rnc* operon of *Escherichia coli*. *Journal of Biological Chemistry* **265** (1990), pp. 2888-2895.
- Cock, P.J. and Whitworth, D.E. Evolution of gene overlaps: relative reading frame bias in prokaryotic two-component system genes. *J Mol Evol* **64** (2007), pp. 457-62.
- Dandekar, T., Snel, B., Huynen, M. and Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23** (1998), pp. 324-8.
- Eyre-Walker, A. The close proximity of *Escherichia coli* genes: consequences for stop codon and synonymous codon use. *J Mol Evol* **42** (1996), pp. 73-8.
- Fotheringham, I.G., Dacey, S.A., Taylor, P.P., Smith, T.J., Hunter, M.G., Finlay, M.E., Primrose, S.B., Parker, D.M. and Edwards, R.M. The cloning and sequence analysis of the *aspC* and *tyrB* genes from *Escherichia coli* K12. Comparison of the primary structures of the aspartate aminotransferase and aromatic aminotransferase of *E. coli* with those of the *pig* aspartate aminotransferase isoenzymes. *Biochem J* **234** (1986), pp. 593-604.
- Fukuda, Y., Nakayama, Y. and Tomita, M. On dynamics of overlapping genes in bacterial genomes. *Gene* **323** (2003), pp. 181-7.
- Fukuda, Y., Washio, T. and Tomita, M. Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. *Nucleic Acids Res* **27** (1999), pp. 1847-53.
- Genser, K.F., Renner, G. and Schwab, H. Molecular cloning, sequencing and expression in *Escherichia coli* of the poly(3-hydroxyalkanoate) synthesis genes from *Alcaligenes latus* DSM1124. *J Biotechnol* **64** (1998), pp. 125-35.
- Golderer, G., Dlaska, M., Grobner, P. and Piendl, W. TTG serves as an initiation codon for the ribosomal protein MvaS7 from the archaeon *Methanococcus vannielii*. *J Bacteriol* **177** (1995), pp. 5994-6.
- Hui, A. and de Boer, H. Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in *Escherichia coli*. *Proc Natl Acad Sci U S A* **84** (1987), pp. 4762-6.
- Inokuchi, Y., Hirashima, A., Sekine, Y., Janosi, L. and Kajii, A. Role of ribosome recycling factor (RRF) in translational coupling. *Embo Journal* **19** (2000), pp. 3788-3798.
- Jacob, W., Santer, M. and Dahlberg, A. A single base change in the Shine-Dalgarno region of 16S rRNA of *Escherichia coli* affects translation of many proteins. *Proc Natl Acad Sci U S A* **84** (1987), pp. 4757-61.

- Johnson, Z.I. and Chisholm, S.W. Properties of overlapping genes are conserved across microbial genomes. *Genome Res* **14** (2004), pp. 2268-72.
- Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **34** (2006), pp. D354-7.
- Keese, P.K. and Gibbs, A. Origins of genes: "big bang" or continuous creation? *Proc Natl Acad Sci U S A* **89** (1992), pp. 9489-93.
- Kingsford, C., Delcher, A.L. and Salzberg, S.L. A unified model explaining the offsets of overlapping and near-overlapping prokaryotic genes. *Mol Biol Evol* **24** (2007), pp. 2091-8.
- Korbel, J., Jensen, L., von Mering, C. and Bork, P. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat Biotechnol* **22** (2004), pp. 911-7.
- Kozak, M. Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. *Microbiol Rev* **47** (1983), pp. 1-45.
- Kozlov, N.N. Analysis of a set of overlapping genes. *Dokl Biochem* **373** (2000), pp. 119-22.
- Krakauer, D.C. Stability and evolution of overlapping genes. *Evolution* **54** (2000), pp. 731-9.
- Lillo, F. and Krakauer, D.C. A statistical analysis of the three-fold evolution of genomic compression through frame overlaps in prokaryotes. *Biol Direct* **2** (2007), p. 22.
- Liveris, D., Schwartz, J.J., Geertman, R. and Schwartz, I. Molecular cloning and sequencing of infC, the gene encoding translation initiation factor IF3, from four enterobacterial species. *FEMS Microbiol Lett* **112** (1993), pp. 211-6.
- Ma, J., Campbell, A. and Karlin, S. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol* **184** (2002), pp. 5733-45.
- McCarthy, J.E. Post-transcriptional control in the polycistronic operon environment: studies of the atp operon of Escherichia coli. *Mol Microbiol* **4** (1990), pp. 1233-40.
- Natale, D.A., Galperin, M.Y., Tatusov, R.L. and Koonin, E.V. Using the COG database to improve gene recognition in complete genomes. *Genetica* **108** (2000), pp. 9-17.
- Nolling, J., Pihl, T.D., Vriesema, A. and Reeve, J.N. Organization and growth phase-dependent transcription of methane genes in two regions of the Methanobacterium thermoautotrophicum genome. *J Bacteriol* **177** (1995), pp. 2460-8.
- Normark, S., Bergstrom, S., Edlund, T., Grundstrom, T., Jaurin, B., Lindberg, F.P. and Olsson, O. Overlapping genes. *Annu Rev Genet* **17** (1983), pp. 499-525.

- Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. and Maltsev, N. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96** (1999), pp. 2896-901.
- Palleja, A., Harrington, E.D. and Bork, P. Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? *BMC Genomics* **9** (2008), p. 335.
- Polard, P., Prere, M.F., Chandler, M. and Fayet, O. Programmed translational frameshifting and initiation at an AUU codon in gene expression of bacterial insertion sequence IS911. *J Mol Biol* **222** (1991), pp. 465-77.
- Rogozin, I.B., Spiridonov, A.N., Sorokin, A.V., Wolf, Y.I., Jordan, I.K., Tatusov, R.L. and Koonin, E.V. Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet* **18** (2002), pp. 228-32.
- Sabath, N., Graur, D. and Landan, G. Same-strand overlapping genes in bacteria: compositional determinants of phase bias. *Biol Direct* **3** (2008), p. 36.
- Sakharkar, K.R. and Chow, V.T. Strategies for genome reduction in microbial genomes. *Genome Inform* **16** (2005), pp. 69-75.
- Sakharkar, K.R., Sakharkar, M.K., Verma, C. and Chow, V.T. Comparative study of overlapping genes in bacteria, with special reference to *Rickettsia prowazekii* and *Rickettsia conorii*. *Int J Syst Evol Microbiol* **55** (2005), pp. 1205-9.
- Salgado, H., Moreno-Hagelsieb, G., Smith, T.F. and Collado-Vides, J. Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci U S A* **97** (2000), pp. 6652-7.
- Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M. and Smith, M. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265** (1977), pp. 687-95.
- Sazuka, T. and Ohara, O. Sequence features surrounding the translation initiation sites assigned on the genome sequence of *Synechocystis* sp. strain PCC6803 by amino-terminal protein sequencing. *DNA Res* **3** (1996), pp. 225-32.
- Shine, J. and Dalgarno, L. The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A* **71** (1974), pp. 1342-6.
- Starmer, J., Stomp, A., Vouk, M. and Bitzer, D. Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLoS Comput Biol* **2** (2006), p. e57.
- Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28** (2000), pp. 33-6.
- Wang, G., Nie, L. and Tan, H. Cloning and characterization of *sanO*, a gene involved in nikkomycin biosynthesis in *Streptomyces ansochromogenes*. *Lett Appl Microbiol* **37** (2003), pp. 452-7.

A manuscript number has been assigned: GENE-D-08-00537

Ms. Ref. No.: GENE-D-08-00537

Title: Overlapping Gene Structures among Prokaryote Genomes
Gene

Dear PhD student Albert Pallejà,

Your submission entitled "Overlapping Gene Structures among Prokaryote Genomes" has been assigned the following manuscript number: GENE-D-08-00537.

You may check on the progress of your paper by logging on to the Elsevier Editorial System as an author. The URL is <http://ees.elsevier.com/gene/>.

Thank you for submitting your work to this journal.

Kind regards,

Peter Geraghty
Journal Manager
Gene

Submitted to GENE

ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGACAGCTGATAGG
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGGCGGTGGTCAGGCAGCGCAGCATGGGAAAA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
GATCAGATGACAGATAGATAACCACAGAGACATG
CGCGCC**C**GACGCATGATTGATGATCAGATGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGATGATA**H**AGTGATTAGATAGATGGTGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTTT
GAGAGATAGAGAG**A**TAGGATCGCGCTCGAGCGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AACCGGCCAACCGGTGG**P**TTAGGATAGATGATGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AAAGTGTGTGTGACAGACAG**T**TGATGATAGTACA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGA**E**AGCTGATAG
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
TCGCGCTCGCTCGAGCGCTAGCTCGAT**R**GAT
TCGCGCTCAAACGAGCGCTAGCTCGATCGA
TGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
TATAGGTACGCGGATGAATGGCAGTAGCTAGCTT
TATCGATCGATCGATCGATCGCGCTAGCTAGCTT
TACGCATGACACACACACATCGATGATGATGAT
TCCAGGCAGCATAAAGCAGCTGAGCTAGCTAGCTT
TAGCTGGGTGGTAGGAATCGATCGATCGAT



**LARGE GENE OVERLAPS IN PROKARYOTIC
GENOMES: RESULT OF FUNCTIONAL
CONSTRAINTS OR MISPREDICTIONS?**

ACGCGAAAATGGC
GATAGAGATACAGAA
GTACGCGAAAATGGCAG
CGCTCGAGCGCTAGCTCGA
AGGTACGCGAAAATGGCAGTA
ATCGATCGATCGATCGATCGCGC
ATCAGCATGACACACACATGATA
CAGTGCCAGGCAGCATAAAGCAGACG
ACCAGCAGCTGGGTGGTAGGAGTGATGT
GCAGTGCCAGGCAGCATAAAGCAGACGAC
GCACCAGCAGCTGGGTGGTAGGAGTGATGAT



**LARGE GENE OVERLAPS IN PROKARYOTIC GENOMES: RESULT
OF FUNCTIONAL CONSTRAINTS OR MIS PREDICTIONS?**

Albert Pallejà^{1,2§}, Eoghan D. Harrington² and Peer Bork^{2,3}.

¹Biochemistry and Biotechnology Department.

Rovira i Virgili University.

C/Marcel·lí Domingo s/n,

43007 Tarragona,

Catalunya, Spain

²European Molecular Biological Laboratory

Meyerohofstrasse, 1

69012 Heidelberg

Germany

³Max Delbrück Centre for Molecular Medicine

Berlin-Buch

Robert-Rössle-Strasse 10

D-13092 Berlin

Germany

[§]Corresponding author

Email addresses:

AP: albert.palleja@urv.cat

EDH: harringt@embl.de

PB: bork@embl.de

Published: 15 July 2008

Received: 10 March 2008

Accepted: 15 July 2008

BMC Genomics 2008, **9**:335 doi:10.1186/1471-2164-9-335

This article is available from: <http://www.biomedcentral.com/1471-2164/9/335>

© 2008 Pallejà et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Across the fully sequenced microbial genomes there are thousands of examples of overlapping genes. Many of these are only a few nucleotides long and are thought to function by permitting the coordinated regulation of gene expression. However, there should also be selective pressure against long overlaps, as the existence of overlapping reading frames increases the risk of deleterious mutations. Here we examine the longest overlaps and assess whether they are the product of special functional constraints or of erroneous annotation.

Results: We analysed the genes that overlap by 60 bps or more among 338 fully-sequenced prokaryotic genomes. The likely functional significance of an overlap was determined by comparing each of the genes to its respective orthologs. If a gene showed a significantly different length from its orthologs it was considered unlikely to be functional and therefore the result of an error either in sequencing or gene prediction. Focusing on 715 co-directional overlaps longer than 60 bps, we classified the erroneous ones into five categories: i) 5'-end extension of the downstream gene due to either a mispredicted start codon or a frameshift at 5'-end of the gene (409 overlaps), ii) fragmentation of a gene caused by a frameshift (163), iii) 3'-end extension of the upstream gene due to either a frameshift at 3'-end of a gene or point mutation at the stop codon (68), iv) Redundant gene predictions (4), v) 5' & 3'-end extension which is a combination of i) and iii) (71). We also studied 75 divergent overlaps that could be classified as misannotations of group i). Nevertheless we found some convergent long overlaps (54) that might be true overlaps, although an important part of convergent overlaps could be classified as group iii) (124).

Conclusions: Among the 968 overlaps larger than 60 bps which we analysed, we did not find a single real one among the co-directional and divergent orientations and concluded that there had been an excessive number of misannotations. Only convergent orientation seems to permit some long overlaps, although convergent overlaps are also hampered by misannotations.

We propose a simple rule to flag these erroneous gene length predictions to facilitate automatic annotation.

Background

The exponentially increasing amount of sequence information has spurred the need for automated and accurate large-scale prediction and functional annotation of genes. A new generation of technologies is speeding up the sequencing even more, but this comes at the price of some biases and an increased error rate [1, 2]. Thus, it is important to investigate unexplained phenomena for systematic errors. One such phenomenon is a large number of annotated genes with long overlaps. Overlapping genes are frequently observed in microbial chromosomes. Although they were initially found in the genomes of bacteriophages, animal viruses and mitochondria [3-5], they currently represent an important part of the genes in the fully sequenced prokaryotic genomes [6]. Furthermore, it is already known that overlapping pairs are conserved across species [7], and it is likely they have more homologs than genes that do not overlap. This makes the overlapping gene pairs highly valuable as a tool for function prediction as other structural prokaryotic features such as well-conserved operons, conserved distances between adjacent genes, COG groups or KEGG pathways have been used to infer functions in genomic and metagenomic data [8, 9]. However, they still remain strongly affected by sequencing and annotating errors [10]. Among the fully sequenced microbial genomes, thousands of overlapping gene pairs have been predicted in all three transcriptional directional classes (co-directional (->->), convergent (-><-) and divergent (<->-) [5, 11, 12]. The overlaps can arise when the 3'-end of one of the genes in a pair is extended because a stop codon has been deleted, or because the stop codon has been disrupted by a point mutation or a frameshift mutation [7, 11, 13]. However, the overlaps can also arise through the elongation of the 5'-end of a gene because an alternative upstream start codon has been used [13-15]. While there is plenty of evidence that small gene overlaps of several

nucleotides enhance coordinated transcription of functionally related genes [6-8, 11, 13, 15], it is not known whether long overlaps are the product of special functional constraints or simply of large-scale misannotations. For bacterial genomes it has been reported that overlaps longer than 20 bps have a reduced Shine-Dalgarno (SD) prediction percentage [16]. This regulatory motif appears to work in concert with the start codons as part of an elaborate regulatory system for gene expression. Therefore, one possible explanation for this low percentage is that many of these genes are incorrectly annotated.

A number of previous studies of overlapping microbial genes suggested that annotation errors such as misprediction of start codons, loss of termination codons as well as the misidentification of the entire open reading frames (ORFs) can influence the statistics of overlapping genes and hence their analysis [6, 7, 11-15] (Table 1). These studies used to exclude from their analysis both the genes coding for hypothetical proteins and the genes whose start codons have been assigned differently by the annotation programs and have therefore been deposited with different coordinates in the databases. On the other hand, the authors tend to accept the gene pairs that are conserved in the COG database [17]. Only Rogozin *et al.* [14] have tried to find out how the overlapping genes evolve and have examined some long convergent overlaps. Nevertheless none of the previous studies has attempted to quantify and characterize rigorously these possible misannotations to be able to study gene overlaps more reliably. Here we analyse long overlaps between well-characterized genes to discriminate true events from misannotations and to use this knowledge to develop rules for improving gene annotation.

Reference	Objective	Excluded genes	Accepted gene set	Annotation errors suggested
Fukuda <i>et al.</i> ,	Comparison study of	Homologous genes whose	Authentic ORFs, thus	Misprediction of the

1999 [11]	overlapping genes in two	start codons was assigned	genes not annotated as	start codons
Fukuda et al., 2003 [7]	Mycoplasma genomes. Study of overlapping genes in bacterial genomes	differently and genes coding for hypothetical or putative proteins	hypothetical or putative proteins and conserved in COG database	
Rogozin et al., 2002 [12]	Study of non-coding DNA in prokaryotic genomes	Genes coding for hypothetical proteins and overlapping more than 90 bps	Gene pairs not annotated as hypothetical or putative proteins and conserved in COG database	Misprediction of start codons, falsely predicted genes and missed genes, frameshifts
Rogozin et al., 2002 [14]	Analysis of the purifying and directional selection in overlapping prokaryotic genes	Genes not conserved in COG database and neither co-directional nor divergent overlapping pairs nor overlapping gene pairs not conserved in two or more species	Convergent overlapping genes conserved in both the COG database and in two or more than two genomes	Misprediction of start codons (affecting co-directional and divergent overlaps) and loss of termination codons (affecting co-directional and convergent overlaps)
Johnson and Chisholm, 2004 [6]	Study of the properties of the overlapping genes in microbial genomes	Genes coding for hypothetical proteins	Gene pairs not annotated as hypothetical or putative proteins	Misidentification of coding sequences
Sakharkar et al., 2005 [13]	Comparison study of overlapping genes in two Rickettsia genomes	Genes coding for hypothetical proteins	Gene pairs not annotated as hypothetical or unknown proteins	Incorrectly annotated ORFs
Cock and Withworth, 2007 [15]	Study of the relative reading frame bias in Prokaryotic Two-component system genes which use to overlap	Genes with ambiguous locations	Two component system gene pairs well located in the chromosome	Invalid bacterial start codons or premature stop codons

Table I - Analysing previous overlapping genes reports

Comparison of previous overlapping genes studies. Columns referring to the authors, the authors' objectives, the genes excluded from their study, the genes accepted for their study, and the misannotations which they suggest are present in prokaryotic chromosomes.

Results and Discussion

Usually, adjacent genes in prokaryotic chromosomes tend to be separated by a short intergenic distance or overlap by some base pairs in a preferred phase [6, 12, 14, 15]. Particularly common are overlaps where the stop codon of the upstream gene is overlapping with the start codon of the downstream gene (overlaps of 1 or 4 bps) [6, 7, 11, 14, 15, 18]. Overlapping

genes among prokaryotes represented around 17% (173,663 overlapping pairs) out of the total gene pairs contained in 338 microbial genomes (1,016,129 gene pairs). Although it is lower percentage than some authors have reported before [6], those overlapping genes are a consistent feature of the prokaryotic chromosomes and are worthy of study. Of these 173,663 overlaps we selected 42,055 where both genes were well-characterized for our study. Among the prokaryotic overlaps, those with co-directional overlaps were clearly the most frequent, reflecting the fact that this is the most common orientation of two adjacent prokaryotic genes [18]. Furthermore, the genes in the prokaryotic chromosomes tend to be grouped into operons of functionally related genes and usually, those genes of a given operon are on the same strand [19-24]. In fact, co-directional overlaps represented around 92% (38,563 overlaps) of the well-characterized overlaps considered here, while convergent overlaps represented 7% (3,035) and divergent overlaps 1% (457). Of these overlaps, we chose a set of 968 overlaps longer than 60 bps that had consistent coordinates in three different databases.

Types of misannotation

We were looking for functional overlaps among the 968 overlaps longer than 60 bps. Every gene of the overlapping pairs was compared with its orthologs. If there is a difference in gene length between the gene and its orthologs the overlap is probably unreal and caused by a sequencing or annotation error in one of the genes of the overlap. This difference in gene length could also mean that the overlap is real though unconserved and therefore, not functional. Although we can not definitively distinguish between these two facts, by categorizing the long overlaps manually, we can notice patterns that provide us with hints. For a list of all the overlaps manually analysed here see Additional file 1.

First of all, we manually analyzed 715 co-directional overlaps longer than 60 bps. Surprisingly all of them fell into the following categories (Figure 1):

i) 5'-end extension of the downstream gene due to either a mispredicted start codon or a frameshift at 5'-end of the gene. The upstream gene had the same length as its orthologs, while the downstream gene was longer than its orthologs at the 5'-end. Furthermore, in all the 409 cases classified, the downstream gene had alternative start codons which were downstream of the predicted initial codon, which could produce a product with a similar or even an equal length to its orthologs. These cases represented around 57% of the co-directional overlaps longer than 60 bps analysed. Therefore this suggests that the most important cause of long overlaps is a misprediction of the start codon of a gene;

ii) Fragmentation of a gene caused by a frameshift. In these cases the upstream gene was longer than its orthologs at the 3'-end and the downstream gene was clearly shorter than its orthologs. Furthermore, in these 163 cases both members of the overlapping pair could be mapped to a single gene in a closely related species, suggesting that a frameshift mutation/sequencing error fragmented one gene into an overlapping pair. These cases represented around 23% of the co-directional overlaps longer than 60 bps analysed and therefore, this is the second most important group of misannotations.

iii) 3'-end extension of the upstream gene due to either a frameshift at 3'-end of gene or point mutation at the stop codon. The upstream gene was longer than its orthologs at the 3'-end, whereas the downstream gene had a similar length to its orthologs. Either a frameshift at the 3'-end or a point mutation at the stop codon may cause the loss of the stop codon, thus extending the reading frame to the next in-frame stop codon. We found 68 cases (9,5% of the co-directional overlaps analysed) that showed this pattern.

iv) Redundant gene prediction where the genes overlap entirely or almost entirely and are in the same reading frame. This is a really strange case and actually we only found 4 gene pairs (0,5%), most of them labelled as putative genes.

v) 5' & 3'-end extension which is a combination of i) and iii). The upstream gene is longer than its orthologs at the 3'-end as well as the

downstream gene being longer than its orthologs at the 5'-end. We classified in this group 71 overlaps (10%).

Regarding the overlapping lengths, the overlapping mean length of the 5', 3' and 5' & 3'-end extension groups was 104, 121 and 106 bps respectively. Nevertheless, the overlapping mean length of the fragmentation type was 162 bps, therefore this type of misannotations appears to cause longer overlaps. In order to know what type of misannotations causes the longest overlaps, we did not take into account the lengths of the overlaps caused by redundant gene prediction, because the gene pair is overlapping entirely or almost entirely and actually this type of misannotations occurs very rarely.

Although we extensively focused on the co-directional orientation, we also examined the long overlaps in the other orientations, specifically, 75 divergent overlaps and 178 convergent overlaps longer than 60 bps. All the divergent long overlaps belonged to group i), which means that all of them were misannotations due to a 5'-end extension of one or both genes of the divergent overlap. However, among the convergent overlaps we found putative true overlaps. Actually, as other authors have reported before [14], conserved convergent overlaps are affected by annotation errors to a lesser extent because they are not affected by the high rate of misannotated start codons. However, we could classify 124 convergent overlaps into group iii) as misannotations. Therefore, the misannotations are also affecting convergent overlaps, particularly those misannotations caused by a 3'-end extension in one or both genes of the pair. The other 54 convergent overlaps might be real, although most of them are only conserved in very close species.

Thus, we can now suggest ways to correct 914 gene pairs and clear the respective overlaps that are the result of misannotations. These overlaps caused by misannotations represent around the 2 % of the overlaps of well characterized genes (42,055). Therefore, this is worth taking into account in the annotation processes.

Co-directional overlaps



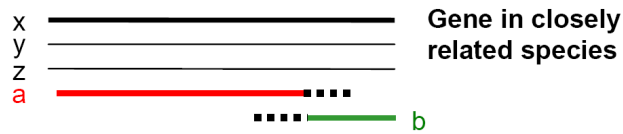
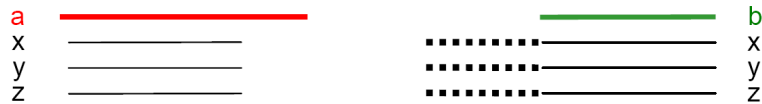
1) 5'-end extension

409 pairs → 57%



2) Fragmentation

163 pairs → 23%



3) 3'-end extension

68 pairs → 9,5%



4) Redundant gene prediction

4 pairs → 0,5%



5) 5' & 3'-end extension

71 pairs → 10%



Figure 1 - Types of misannotation

Schema of the five categories of putative misannotations. Both the number and the percentage of co-directional overlapping pairs longer than 60 bps classified in each group is shown. Gene a represents the upstream gene, while gene b represents the downstream gene. In Fragmentation type gene x, y and z represent the orthologs of gene a and b.

Misannotations in prokaryotic genomes

As expected, the number of overlaps decreases with an increasing overlap length (Figure 2). Equally expected is the avoidance of multiples of 3 bps overlaps for adjacent co-directional genes [6, 14, 15]. Although Figure 2 shows multiples of 3 bps convergent and divergent overlaps, none co-directional overlap was found with an overlapping length of multiple of 3 bps. We also studied in co-directional overlaps whether some particular genomes stood out in terms of overlaps because of their annotation protocols. Indeed, in some genomes large overlaps are more abundant with *Brucella melitensis* 16M leading with 38 likely misannotated events. Interestingly, 25 of those pairs were due to fragmentations [see Additional file 2]. Second in the list is *Rhodopirellula baltica* SH1, which has a really strange genome. It contains 28 misannotated overlaps, 26 of them are due to 5' or 5' & 3'-end extensions and it is the genome which has more divergent overlaps misannotated. Also we have observed that *Xanthomonas* genomes accumulated a high number of misannotations. Probably, the initial mispredictions in the first *Xanthomonas* genomes sequenced were propagated within this taxon due to the high sequence similarity among their genomes. For a list of 27 genomes with high number of overlaps see Additional file 3.

We tried to further identify reasons that might cause frameshifts and misannotations in the genome projects [see Additional file 3]. The genomes that accumulate a high number of errors are not the longest in size or the highest in gene content. For instance, the *Brucella melitensis* 16M chromosome has 3294931 nucleotides and 3198 predicted genes and accumulated 38

misannotations, whereas the *Vibrio vulnificus* YJ016 chromosome has 5211578 nucleotides and 5098 predicted genes but accumulated only 12 annotation errors. A high AT content could be related to a high number of mispredictions of start codons. However, no correlation between a high number of misannotations and a high percentage of AT was observed. We also did not observe any clear bias to any sequencing or annotation method, though 6 out of the 28 genomes worst annotated were done by Glimmer predictor [25] exclusively. However, the use of a determined gene predictor or a combination of different gene predictors, does not assure us that we will avoid the types of misannotations described here. The number of misannotations could also be related to the sequencing date. On one hand, an early sequencing date could be related to a high number of misannotations because less matured technologies and tools were used. On the other hand, a recent sequencing date could be related to a high number of misannotations due to lower coverage and a higher degree of automation. However, no trend was observed in the number of misannotations regarding the sequencing date.

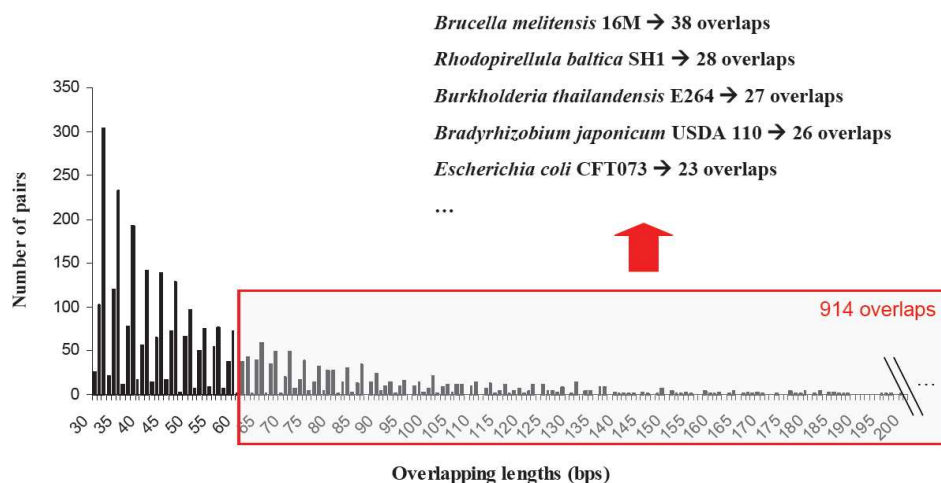


Figure 2 - Distribution of the overlapping pairs with respect to the overlapping length

The longest overlaps selected for manual analysis are indicated by the red box. Several species contribute a disproportionate number of overlapping pairs to the misannotations. In the figure we can see the 5 species that accumulate more misannotations.

Mispredicted start codons

5'-end extensions clearly have the highest number of misannotations because of mispredictions of start codons or upstream frameshifts whereby the former is clearly dominant (data not shown). Therefore we can say that the main problem in the annotation of real genes is the misprediction of start codons. Most genes tend to start with AUG while the alternatives GUG and UUG are used sparingly [16]. AUG is a more potent initiator than GUG or UUG [26], which are considered weak start codons. To quantify the observed effect regarding start codon usage, we compared the start codons of potentially misannotated genes with those from randomly chosen microbial genes. The genes which have putative mispredicted start codons (the genes with a 5'-end extension from wrong categories i), v) and from misannotated divergent overlaps group) had alternative start codons (AUG, GUG or UUG) downstream in the sequence. This could indicate that a gene with a mispredicted start codon has an additional correct one nearby. Furthermore, we observed statistical differences ($P < 0.0001$, Chi square analysis) which were extremely significant among the start codon usage between genes with a putative mispredicted start codon and a random set of genes. It seems that the use of the weak start codons (GUG, UUG) is overrepresented among the genes with putative mispredicted start codons [see Additional file 4]. We found that from the 579 genes, which potentially could have a mispredicted start codon, 270 start with AUG, whereas 172 and 133 with GUG and UUG respectively. In contrast, among the random sets of genes around ~462 start with AUG, whereas only around ~77 and ~38 with GUG and UUG respectively. Therefore, long overlaps, in conjunction with the use of weak start codons could be a sign that the 5'-end of an ORF has been mispredicted and must be taken into account by the annotation algorithms.

In fact, some previous SD studies agreed with this finding. Starmer *et al.* explained genome annotation errors with a bias in the start codon prediction towards the usage of GUG instead of AUG [27], whereas a previous study performed by Ma *et al.* [16] found in *E. coli* K12 a significant group of genes which started with GUG or UUG and which do not have an SD sequence and hence were erroneously annotated as putative or hypothetical proteins.

The longest real co-directional overlap

When studying co-directional overlaps below 60 bps, the longest real one we could identify was caused by two co-directional genes coding for the DNA polymerase psi subunit (*holD*) and an alanine acetyltransferase (*rimI*). Figure 3 shows the alignment of the C-terminal end of the DNA polymerase psi subunit and the N-terminal end of the alanine acetyltransferase as well as an arrangement of overlapping regions and amino acid conservation within the overlap among three representative Enterobacteria species. This figure highlights the high similarity among the Enterobacteria orthologs at the C-terminal end of the protein encoded in *holD* gene, at the N-terminal end of the protein encoded in *rimI* gene and within the overlapping region at the level of nucleotide sequence. This overlap was previously reported to be 32 bps long in *Escherichia coli* [28] which would correspond to around 10 overlapping amino acids; however orthologs gene pairs in the Yersinia and Salmonella genomes reached 56 bps, which would correspond to overlaps of about 18 amino acids. Although the exact gene length seems genus specific, this particular overlap is well conserved among Enterobacteria, and therefore unlikely to be due to a misannotation reported here.

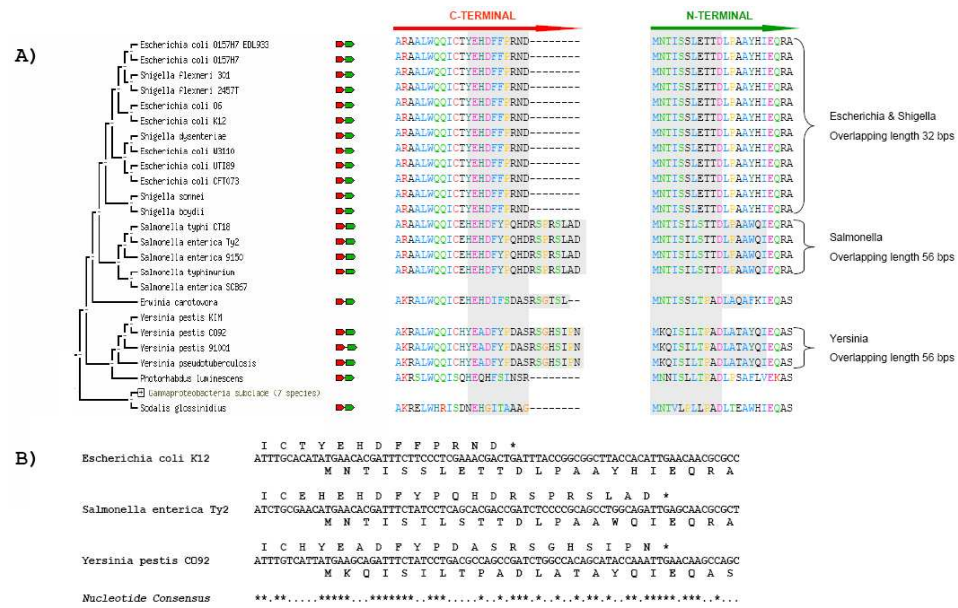


Figure 3 - Aligning a co-directional true overlap

Overlap between the *hoiD* (coding for a DNA polymerase psi subunit) and *rmlI* (coding for an alanine acetyltransferase) genes among Enterobacteria. A) Multiple alignment of the C-terminal of the DNA polymerase psi subunit and the N-terminal of the alanine acetyltransferase protein among Enterobacteria species. The grey boxes indicate the fragments that are encoded in the overlapping region between *hoiD* and *rmlI* genes. The alignments of *Escherichia* & *Shigella*, *Salmonella* and *Yersinia* are marked. B) Arrangement of overlapping regions and amino acid conservation within the overlap among *Escherichia coli* K12, *Salmonella enterica* Ty2 and *Yersinia pestis* CO92. The nucleotide consensus shows an asterisk for the conserved nucleotides and a dot for the not conserved. Although we chose one species of each group marked in part A (*Escherichia* & *Shigella*, *Salmonella* and *Yersinia*) we can observe the high similarity at the level of sample nucleotide sequences too.

Conclusions

Misannotation of real genes leading to artificial extensions of genes seems to be more frequent than previously anticipated and can lead to frequent gene overlaps. We could show here that all co-directional and divergent overlaps extending 60 bps are artificial due to misannotations that can be

classified into five categories. This clear-cut result enables us to propose a simple rule that can flag many thousand erroneous gene length predictions to facilitate automatic annotation. On the other hand, convergent orientation seems to allow longer overlaps than the other two orientations, although convergent long overlaps are also affected by misannotations.

The most common misannotation is the 5'-end extension, mostly caused by the misprediction of start codons. The respective genes carrying putative mispredictions of the start codon show an overrepresentation of weak start codons use. Thus genes with a 5'-end extension involved in long overlaps with predicted weak start codons must be checked by the annotation algorithms.

Although several species seemed to have a higher number of such potential misannotations, no correlation was found with genome size, gene content, GC content, sequencing or ORF prediction method, annotation team or sequencing date. Therefore these imprecise gene predictions have the potential to affect any microbial genome annotation process.

Methods

Overlapping genes were retrieved from the 338 microbial genomes in the STRING database release 7.0 [29]. As has been mentioned above, analysis of the overlapping genes is hampered by sequencing and annotation errors present in genomes [10]. Because of this concern, only well-characterized genes were analysed. We defined as well-characterized genes only those gene pairs where both members could be assigned to a KEGG pathway [30]. This means that only 42,055 overlaps out of the 173,663 overlapping gene pairs observed among 338 prokaryotic genomes were considered in our study. Of these, 38,563 were in the co-directional orientation, whereas 3,035 were in convergent orientation and 457 were in divergent orientation. We focused on long overlaps to identify unusual differences in length. In order to avoid work with overlaps originated by inconsistent data among the databases, we checked whether their coordinates were consistent in STRING database release 7.0,

Genome Reviews and RefSeq. We started analysing the longer overlaps and we stopped at 60 bps length because we observed conserved overlaps just below this cut-off.

After the application of all these restrictions commented on above, we eventually had 715 co-directional overlaps with overlapping lengths longer than 60 bps, which were examined manually. Each protein of these overlaps was compared to its corresponding orthologs, analogous to the consistency check used in the HAMAP project [31] for the SWISS-PROT protein validation. Therefore, for each member of an overlapping pair a multiple sequence alignment was constructed from the gene itself and its orthologs (as defined in the STRING [29] database) using Muscle [32]. These alignments were analysed by eye and if the overlapping genes showed significant differences in length, relative to their respective orthologs, we concluded that it was a misannotation. Then, these overlaps were placed into one of five categories based on putative sequencing or annotation errors that might have caused the artificial overlap. The convergent (178) and divergent (75) overlaps longer than 60 bps were also analysed manually. These overlaps were also placed into the categories previously defined with the exception of some of the convergent long overlaps.

We also examined whether certain species were associated with higher numbers of overlapping genes. In addition, we analyzed the correlation between the number of gene overlaps with genome size, gene content, GC content, sequencing or ORF prediction method, annotation team or sequencing date. We also analysed the misprediction of start codons using the genes that show 5'-end extensions among the groups 5'-end extension, 5' & 3'-end extension and the misannotated divergent overlaps, totalling 579 genes. The alternative start codons considered were AUG, GUG or UUG. The genes of genomes which use a different start codon to these three or a bacterial code different to the bacterial and plant plastic genetic code were classified as 'others' in the start codons table [see Additional file 4]. We checked the start codon in each case and how many times each of the three alternative start codons was used up to one third of the length of the gene. The figures were compared to normal gene sets

randomly selected with two restrictions (random set I, II, and III). In the first one, the normal genes had to have gene lengths similar to the misannotation gene set (around 1400 bps). In the second one, the number of genes in each set had to be the same (that is, 579 genes in each set). We took well-characterized non-overlapping genes randomly selected as our normal genes. Furthermore, a Chi square analysis was performed comparing the start codon usage of one normal gene set with the mispredicted gene set. Where necessary we used Perl programming language in all the steps of this work as well as PostgreSQL to communicate with the STRING [29] database.

Authors' contributions

AP performed the necessary Perl Scripts and sequence alignments and manually checked the overlaps. AP, EH and PB participated in the analysis and interpretation of the data. AP drafted the manuscripts and EH and PB revised it critically. Finally, all the authors read and approved the version to be published.

Acknowledgements

We would like to thank the DAAD for the scholarship grant (Referat 314) given to Albert Pallejà to develop this work. This work has also been supported by projects BIO02003-07672 and AGL2007-65678/ALI of the Spanish Ministry of Education and Science. Also we would like to thank the Language Service from the Rovira i Virgili University for their help in writing the manuscript. Finally, we would like to thank the anonymous reviewers that provided us with useful comments to improve this paper.

References

1. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bembien LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR,

- Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437**(7057):376-380.
2. Shendure J, Porreca G, Reppas N, Lin X, McCutcheon J, Rosenbaum A, Wang M, Zhang K, Mitra R, Church G: **Accurate multiplex polony sequencing of an evolved bacterial genome.** *Science* 2005, **309**(5741):1728-1732.
 3. Barrell BG, Air GM, Hutchison CA, 3rd: **Overlapping genes in bacteriophage phiX174.** *Nature* 1976, **264**(5581):34-41.
 4. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M: **Nucleotide sequence of bacteriophage phi X174 DNA.** *Nature* 1977, **265**(5596):687-695.
 5. Normark S, Bergstrom S, Edlund T, Grundstrom T, Jaurin B, Lindberg FP, Olsson O: **Overlapping genes.** *Annu Rev Genet* 1983, **17**:499-525.
 6. Johnson ZI, Chisholm SW: **Properties of overlapping genes are conserved across microbial genomes.** *Genome Res* 2004, **14**(11):2268-2272.
 7. Fukuda Y, Nakayama Y, Tomita M: **On dynamics of overlapping genes in bacterial genomes.** *Gene* 2003, **323**:181-187.
 8. Harrington ED, Singh AH, Doerks T, Letunic I, von Mering C, Jensen LJ, Raes J, Bork P: **Quantitative assessment of protein function prediction from metagenomics shotgun sequences.** *Proc Natl Acad Sci U S A* 2007, **104**(35):13913-13918.
 9. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM: **Comparative metagenomics of microbial communities.** *Science* 2005, **308**(5721):554-557.

BMC Genomics 2008, 9: 335

10. Natale DA, Galperin MY, Tatusov RL, Koonin EV: **Using the COG database to improve gene recognition in complete genomes.** *Genetica* 2000, **108**(1):9-17.
11. Fukuda Y, Washio T, Tomita M: **Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*.** *Nucleic Acids Res* 1999, **27**(8):1847-1853.
12. Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, Wolf YI, Yin J, Koonin EV: **Congruent evolution of different classes of non-coding DNA in prokaryotic genomes.** *Nucleic Acids Res* 2002, **30**(19):4264-4271.
13. Sakharkar KR, Sakharkar MK, Verma C, Chow VT: **Comparative study of overlapping genes in bacteria, with special reference to *Rickettsia prowazekii* and *Rickettsia conorii*.** *Int J Syst Evol Microbiol* 2005, **55**(Pt 3):1205-1209.
14. Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, Tatusov RL, Koonin EV: **Purifying and directional selection in overlapping prokaryotic genes.** *Trends Genet* 2002, **18**(5):228-232.
15. Cock PJ, Whitworth DE: **Evolution of gene overlaps: relative reading frame bias in prokaryotic two-component system genes.** *J Mol Evol* 2007, **64**(4):457-462.
16. Ma J, Campbell A, Karlin S: **Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures.** *J Bacteriol* 2002, **184**(20):5733-5745.
17. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41.
18. Eyre-Walker A: **The distance between *Escherichia coli* genes is related to gene expression levels.** *J Bacteriol* 1995, **177**(18):5368-5369.

19. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **Use of contiguity on the chromosome to predict functional coupling.** *In Silico Biol* 1999, **1**(2):93-108.
20. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 1998, **23**(9):324-328.
21. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: **Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context.** *Genome Res* 2001, **11**(3):356-372.
22. Cherry JL: **Genome size and operon content.** *J Theor Biol* 2003, **221**(3):401-410.
23. Price M, Arkin A, Alm E: **The life-cycle of operons.** *PLoS Genet* 2006, **2**(6):e96.
24. Moreno-Hagelsieb G, Collado-Vides J: **A powerful non-homology method for the prediction of operons in prokaryotes.** *Bioinformatics* 2002, **18 Suppl 1**:S329-336.
25. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: **Improved microbial gene identification with GLIMMER.** *Nucleic Acids Res* 1999, **27**(23):4636-4641.
26. Ringquist S, Shinedling S, Barrick D, Green L, Binkley J, Stormo GD, Gold L: **Translation initiation in Escherichia coli: sequences within the ribosome-binding site.** *Mol Microbiol* 1992, **6**(9):1219-1229.
27. Starmer J, Stomp A, Vouk M, Bitzer D: **Predicting Shine-Dalgarno sequence locations exposes genome annotation errors.** *PLoS Comput Biol* 2006, **2**(5):e57.
28. Carter JR, Franden MA, Aebersold R, McHenry CS: **Identification, isolation, and overexpression of the gene encoding the psi subunit of DNA polymerase III holoenzyme.** *J Bacteriol* 1993, **175**(17):5604-5610.

29. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P: **STRING 7--recent developments in the integration and prediction of protein interactions**. *Nucleic Acids Res* 2007, **35**(Database issue):D358-362.
30. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG**. *Nucleic Acids Res* 2006, **34**(Database issue):D354-357.
31. Gattiker A, Michoud K, Rivoire C, Auchincloss AH, Coudert E, Lima T, Kersey P, Pagni M, Sigrist CJ, Lachaize C, Veuthey AL, Gasteiger E, Bairoch A: **Automated annotation of microbial proteomes in SWISS-PROT**. *Comput Biol Chem* 2003, **27**(1):49-58.
32. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**:113.

Additional files

Additional file 1. The 968 overlaps manually analysed

The co-directional, convergent and divergent overlaps analysed. They are separated depending on the orientation of the pair. The genes identification is made joining the Taxonomy ID of the species which contains the gene and the gene name separated by a dot. The columns are the upstream and the downstream gene ids, the functions of the protein encoded in the genes, the orientation, the overlapping length and the type of misannotation. Notice that the types of misannotations are described at the end of each of the lists.

Because of its size this figure is not in the book. It can be downloaded following this link:

<http://www.biomedcentral.com/1471-2164/9/335/additional/>

Additional file 2. Number of misannotations per genome in each category

Summary of the mispredicted overlaps found within the genome of each species sorted by categories.

Because of its size this figure is not in the book. It can be downloaded following this link:

<http://www.biomedcentral.com/1471-2164/9/335/additional/>

TaxID	Taxname	Misannotations	Kingdom	Length	Gene content	GC content	Sequenced	Annotation tools	CreationDate
224914	<i>Brucella melitensis</i> 16M	38	bacteria	3,294,931	3,198	43%	whole genome shotgun strategy	Integrated Genomics and ERGO annotation tools, FASTA, Prosite, PFAM, COGS	Nov_13_2001
243090	<i>Rhodospirillum rubrum</i> SH 1	28	bacteria	7,036,071	6,743	38%	whole-genome shotgun combining with bridging shotgun	Glimmer and BLASTP	Jul_8_2003
271848	<i>Burkholderia thailandensis</i> E264	27	bacteria	6,123,972	5,634	33%	random shotgun method	Glimmer, BLASTP and comparison in GenBank, GenPept, Pfam, and SwissProt	Dec_17_2005
224911	<i>Bradyrhizobium japonicum</i> USDA 110	26	bacteria	9,055,828	88,17	36%	whole genome shotgun strategy combined with the bridging shotgun	Glimmer 2.02 and BLASTP and BLASTX	Dec_27_2002
198310	<i>Escherichia coli</i> CF7073	23	bacteria	5,231,428	5,379	50%	whole-genome shotgun strategy	MAGEPI, Glimmer and BLAST	Dec_9_2002
268835	<i>Mesorhizobium loti</i> IMAF303099	23	bacteria	7,036,071	6,743	38%	whole-genome shotgun combining with bridging shotgun	Glimmer and BLASTP	Dec_26_2000
320372	<i>Burkholderia pseudomallei</i> 1710b	22	bacteria	7,308,054	6,647	33%	Unpublished		Oct_3_2005
291331	<i>Xanthomonas oryzae</i> pv. <i>oryzae</i> KACC10331	22	bacteria	5,231,428	5,379	50%	whole genome shotgun strategy	MAGEPI, Glimmer and BLAST	Feb_4_2005
364106	<i>Escherichia coli</i> UT89	19	bacteria	5,655,741	5,044	50%	whole genome shotgun libraries	Unpublished	Apr_7_2006
56780	<i>Syntrophus acidithrophicus</i> SB	17	bacteria	3,179,300	3,169	49%	whole genome shotgun sequencing	Glimmer, Critica and Integrated Genomics and ERGO annotations	Jan_28_2006
196164	<i>Corynebacterium efficiens</i> YS-314	14	bacteria	3,447,080	3,020	63%	shotgun method	comparison with <i>C. glutamicum</i> ATCC 13022 and <i>C. diphtheriae</i> NCTC12129	Nov_15_2002
190486	<i>Xanthomonas axonopodis</i> pv. <i>citri</i> str. 306	14	bacteria	5,175,554	4,912	36%	shotgun methodologies	GeneMark and Glimmer, BLAST and KEGG	May_23_2002
196600	<i>Vibrio vulnificus</i> VJ016	12	bacteria	5,211,578	5,088	46.50%	whole-genome shotgun libraries	GeneMark and Glimmer, gene prediction programs	Oct_9_2003
273057	<i>Sulfolobus solfataricus</i> P2	11	archaea	2,902,245	2,977	36%	the genome was cloned and mapped by using cosmid libraries and bacterial artificial chromosome libraries and lambda and bacterial artificial chromosome libraries and sequenced	Annotation was performed by searching sequence databases of genomes and metabolic pathway enzymes at http://www.archaea-upslab.fhp.jax.org/ in combination with Meggie	Oct_3_2001
197221	<i>Thermosynerchococcus elongatus</i> BP-1	11	bacteria	2,933,857	2,475	54%	whole-genome shotgun strategy combined with the bridging shotgun method	Glimmer 2.02, BLASTP and BLASTX	Aug_19_2002
316273	<i>Xanthomonas campestris</i> pv. <i>vesicatoria</i> str. 85-10	11	bacteria	5,178,466	4,604	65%	Whole-genome sequencing	GeneDB 2.0. Briefly, a combined gene prediction strategy was applied on the assembled sequences using Glimmer and CRITICA	Oct_23_2005
177439	<i>Desulfotalea psychrophila</i> LSy-d4	10	bacteria	3,523,383	3,204	46%	whole genome shotgun strategy	GUMMER, CRITICA, DRPHUS	Aug_16_2004
267747	<i>Propionibacterium acnes</i> KP147102	10	bacteria	2,960,265	2,333	60%	several shotgun libraries were constructed	Initial gene prediction was accomplished using YACOP and annotation was done using the ERGO tool	Jul_30_2004
370553	<i>Streptococcus pyogenes</i> MG-AS2096	10	bacteria	1,800,555	1,979	38%	Whole-genome shotgun sequencing	Integrated Genomics and ERGO bioinformatics tools	May_9_2006
243277	<i>Vibrio cholerae</i> O1 biovar eltor str. N16961	10	bacteria	4,033,464	4,008	46.50%	whole genome random sequencing method	GUMMER and HMMER package	Sep_10_2001
224325	<i>Archaeoglobus fulgidus</i> VC-16	9	archaea	2,178,400	2,436	48.30%	Whole-genome random sequencing procedure	GeneSmith and CRITICA	Dec_17_1987
342108	<i>Magnetospirillum magnetotacticum</i> AMB-1	9	bacteria	4,867,148	4,611	65%	whole genome shotgun strategy	Xanagen and Xanegenome	Dec_7_2005
190485	<i>Xanthomonas campestris</i> pv. <i>campestris</i> str. ATCC 33913	9	bacteria	5,076,188	4,242	65%	shotgun methodologies	GeneMark and Glimmer	Nov_28_2001
251221	<i>Gloeobacter violaceus</i> FCC 7421	8	bacteria	4,659,019	2,333	62%	whole genome shotgun method in combination with the bridging shotgun strategy	Glimmer 2.02 and BLASTP and BLASTX	Oct_6_2003
290633	<i>Glucobacter oxydans</i> 621H	8	bacteria	2,702,173	2,432	61%	whole-genome shotgun approach using plasmid and cosmid libraries	Initial ORFs prediction with YACOP. Annotation with ERGO Software. All annotations were inspected manually through searches against PFAM, PROSITE, PRODOM, and COGS databases. In addition to the BLASTP versus GenBankEMBL and SWISSPROT databases	Jan_24_2005
84588	<i>Synechococcus</i> sp. <i>MH-8102</i>	8	bacteria	2,434,428	2,519	59.00%	Whole-genome shotgun libraries	Combination of three gene modelling programs—Critica, Glimmer and GeneMark was used in the determination of potential coding sequences	Aug_19_2003
243274	<i>Thermotoga maritima</i> IM858	8	bacteria	1,860,725	1,928	46%	Whole-genome random sequencing procedure	GUMMER	Sep_10_2001

Additional file 3. Misannotations related to some genome features

Table summarizing the genomes with more misannotations and some features of the genome such as genome length, gene content, GC content, sequencing method, annotating method and sequence date.

	mispredicted start codon	random set I	random set II	random set III
number of genes	579	579	579	579
AUG usage	270	470	466	452
GUG usage	172	76	68	86
UUG usage	133	31	44	40
other start codons usage	4	2	1	1
% AUG usage	46,6	81,2	80,5	78,1
% GUG usage	29,7	13,1	11,7	14,8
% UUG usage	23,0	5,4	7,6	6,9
% other start codons usage	0,7	0,3	0,2	0,2

Additional file 4. Start codons analysis

Study of the start codons usage found among the three normal gene sets (random set I, II and II), which contains well-characterized non-overlapping genes randomly selected, and within the mispredicted start codon gene set. The usage and percentage of usage of each alternative start codon considered (AUG, GUG, UUG, other) is shown in the rows.

ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGACAGCTGATAGG
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGGCGGTGGTCAGGCAGCGCAGCATGGGAAAA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
GATCAGATGACAGATAGATAACCACAGAGACATG
CGCGCC**C**GACGCATGATTGATGATCAGATGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGATGATA**H**AGTGATTAGATAGATGGTGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTTT
GAGAGATAGAGAG**A**TAGGATCGCGCTCGAGCGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AACCGGCCAACCGGTGG**P**TTAGGATAGATGATGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AAAGTGTGTGTGACAGACAG**T**TGATGATAGTACA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGA**E**AGCTGATAG
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CGCGCTCGCTCGAGCGCTAGCTCGAT**R**GAT
CGCGCTCAAACGAGCGCTAGCTCGATCGA
GATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ATAGGTACGCGGATGAATGGCAGTAGCTAGCTT
ATCGATCGATCGATCGATCGCGCG
CAGCATGACACACACACATC
CCAGGCAGCATAAAGC/
AGCTGGGTGGTAGGA
ATCGATCGATCGAT



**ADAPTATION OF THE SHORT CO-DIRECTIONAL
SPACERS TO THE SHINE-DALGARNO MOTIF IN
PROKARYOTE GENOMES**

.ACGCGAAAATGGC/
.GATAGAGATACAGAA
.GTACGCGAAAATGGCAG
TCGCTCGAGCGCTAGCTCGA
AGGTACGCGAAAATGGCAGTA
ATCGATCGATCGATCGATCGCGCG
ATCAGCATGACACACACATGATA
CAGTGCCAGGCAGCATAAAGCAGACG/
ACCAGCAGCTGGGTGGTAGGAGTGATG/
GCAGTGCCAGGCAGCATAAAGCAGACGAC
GCACCAGCAGCTGGGTGGTAGGAGTGATGATC



**ADAPTATION OF THE SHORT CO-DIRECTIONAL SPACERS TO
THE SHINE-DALGARNO MOTIF IN PROKARYOTE GENOMES**

Albert Pallejà*, Santiago García-Vallvé, Antoni Romeu.

Department of Biochemistry and Biotechnology, Rovira i Virgili University,
Tarragona, Catalonia, Spain.

*Corresponding author

Rovira i Virgili University
Department of Biochemistry and Biotechnology
Campus Sescelades
C/ Marcel·lí Domingo, s/n
E-43007 Tarragona
Catalonia – Spain

Email addresses: albert.palleja@urv.cat (corresponding author)
 santi.garcia-vallve@urv.cat
 antoni.romeu@urv.cat

Submitted to PLoS ONE

Abstract

Background

In prokaryote genomes most of the co-directional genes are in close proximity. Even the coding sequence or the stop codon of a gene can overlap with the Shine-Dalgarno (SD) sequence of the downstream co-directional gene. Here we analyze how the SD presence may influence the stop codon usage or the spacing lengths between co-directional genes.

Methodology/Principal Findings

The SD sequences for 530 prokaryote genomes have been predicted using computer calculations of the base pairing free energy between translation initiation regions and the 16S rRNA 3' tail. Genomes with a large number of genes with the SD sequence concentrate such a regulatory motif from 4 to 12 bps before the start codon. However, not all genes seem to have the SD sequence. Genes separated from 1 to 4 bps from a co-directional upstream gene show a high SD presence, although this regulatory signal is located towards the 3' end of the coding sequence of the upstream gene. Genes separated from 9 to 15 bps show the highest SD presence as they accommodate the SD sequence within an intergenic region. However, genes separated from around 5 to 8 bps have a lower percentage of SD presence and when the SD is present, the stop codon usage of the upstream gene changes to accommodate the overlap between the SD sequence and the stop codon.

Conclusions/Significance

When the SD sequence overlaps with the upstream coding sequence or stop codon, its strength and relative distance to the downstream start codon do not vary significantly. However, the SD presence may make the intergenic lengths from 5 to 8 bps less favored and cause an adaptation of the stop codon usage.

Introduction

Prokaryote genomes are considered compacted genomes with only a small fraction of the genomic DNA containing intergenic regions, which are thought to typically contain regulatory signals [1]. There are variations in percentage of non-coding DNA among the prokaryote genomes. These variations do not depend on the genome size nor the gene content, whereas the latter variables strongly correlate [2]. The spacers between a pair of genes were classified into three types according to their transcriptional direction: i) unidirectional, ii) convergent and iii) divergent [1]. Here we decided to use a co-directional instead of a unidirectional term. These three classes of spacers differ in the type of regulatory signals that they contain. The co-directional spacers may contain an upstream gene terminator, a promoter and an operator for a downstream gene. The convergent spacers may contain terminators for both genes while the divergent ones have only promoters and other upstream transcriptional signals. The different types of intergenic regions in prokaryotes, including the convergent and divergent ones (all of them inter-operonic) and the co-directional ones (largely intra-operonic), evolve under the same evolutionary pressures. The principal evolutionary force is the selective pressure to minimize the amount of non-functional DNA [1,2]. However, in prokaryotes, these intergenic regions must maintain a minimal extension to accommodate essential regulatory signals [1] and the DNA replication sequences [3,4]. According to the genomic compactness, prokaryote genomes have intergenic distances that are much shorter than the gene lengths and are relatively short compared to those in eukaryote genomes [5]. The eukaryote genomes show a much larger range of genome sizes and contain protein-coding genes that are typically, interrupted by introns, and have longer intergenic regions.

One of the regulatory sequences that are compromised by the short distances between prokaryote genes is the Shine-Dalgarno (SD) sequence. In 1974, Shine and Dalgarno found a sequence (5'-GGAGGU-3') at the 5' of the initiation codons in several messenger RNAs (mRNAs) of *Escherichia coli* that was

complementary to the sequence 3'-CCUCCA-5' located at the tail of the 3'-end of the 16S ribosomal RNA (rRNA) [6]. Although it is not mandatory in translation initiation, it has been suggested that a strong SD sequence may compensate for a weak start codon and counteract mRNA secondary structures that hinder access to the start [7,8]. Several studies have addressed the SD presence in prokaryotes [9,10,11,12]. Although the genes with SD sequence are widely found in prokaryote genomes, these studies also reflected that there is a significantly and previously underestimated population of genes without SD sequence. Furthermore, the exponential increase of the fully sequenced genomes has provided us with thousands of examples of leaderless genes or genes without SD sequence in the prokaryote genomes [13]. It has been suggested that the leaderless genes could use an independent pathway in their gene translation, while leader genes without SD sequence must use alternative unknown mechanisms in their translation initiation [14,15].

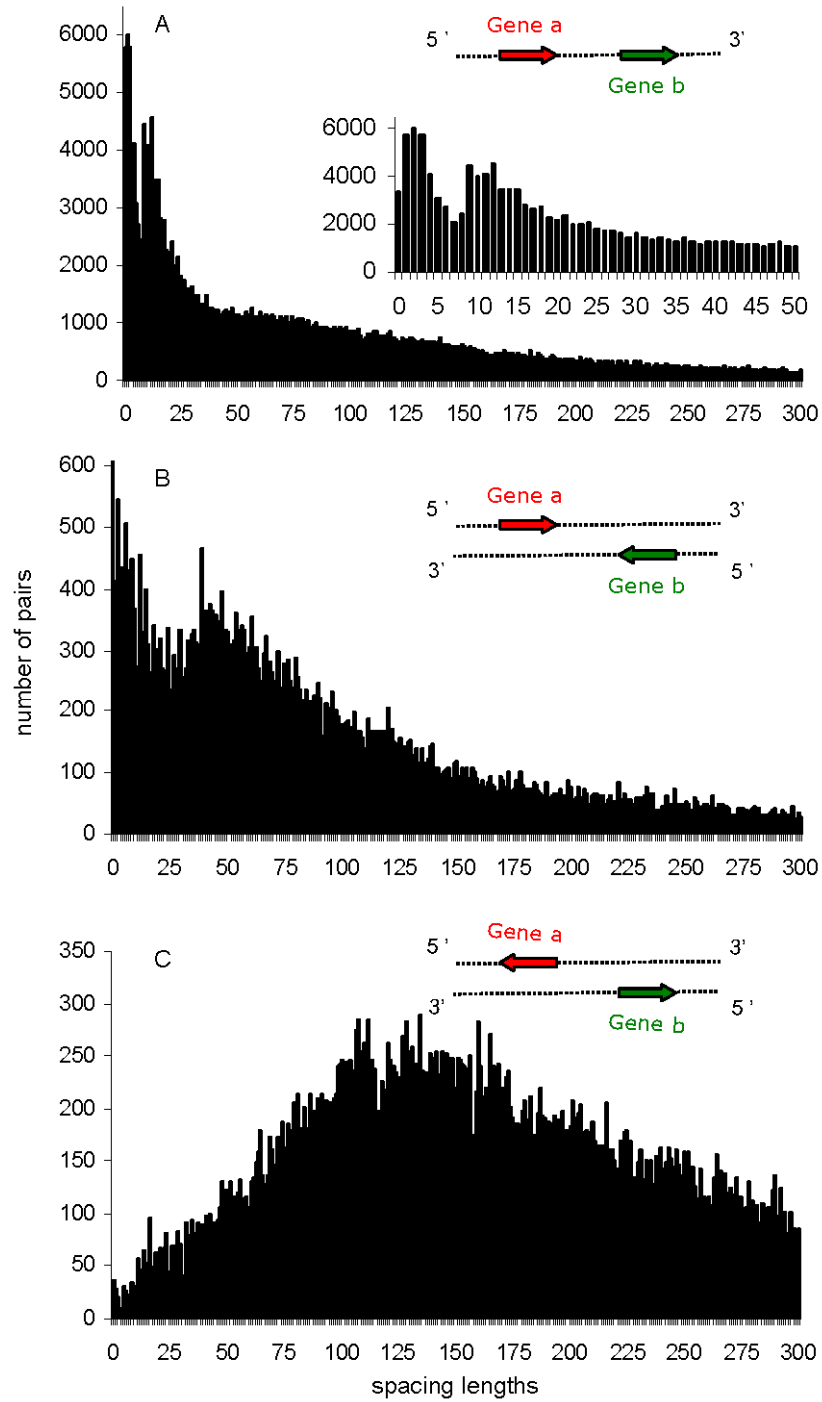
Among the genes that have the SD sequence, the ribosome does not need a perfect distance between the SD and the initiation codon for the initiation of translation. In spite of this, when the SD is located within four nucleotides from the initiation codon or when it is located as far as 13 nucleotides from the initiation codon, the gene expression is decreased drastically [16,17,18]. Therefore, there are apparently structural constraints that require an optimal space between the SD motif and the start codon. This sequence has been found mostly from 7 to 12 nucleotides upstream of the start codon [9,10,19]. Taking this into account the intergenic distances are an important feature of the prokaryote genomes that might correlate with the SD presence [10]. Many genes are sufficiently close together that the end of one gene can be overlapping either the SD or the coding sequence of the next gene. Eyre-Walker and Bulmer pointed out that there is a change in composition at the end of genes, which is consistent with selection against the formation of mRNA secondary structures around the SD sequence [20]. In addition, Eyre-Walker demonstrated that the strength and the SD location do not vary significantly because of the close proximity of the prokaryote genes [21]. Therefore it seems

that the spacing lengths and the stop codon usage would adapt themselves to the SD presence. Recently, in viruses which have compactly organized genomes, there has been found a preference for the TGA stop codon in the genes that overlap their stop codon with the SD sequence of the next gene, forming the pattern GGTGA as the SD motif [22]. In prokaryotes, it seems that there are some intergenic distances that would be less favored because of the SD location. Then a determined stop codon usage is required to form the SD motifs, as it has been previously described in viruses. The aim of this paper is to assess how the SD presence affects the spacing lengths between adjacent genes and the adaptation of the stop codon usage to the SD presence among prokaryotes.

Results and Discussion

Spacing lengths between prokaryote genes

The distribution of the spacing lengths among the three gene orientations is different, probably due to the different gene structures found in each orientation (Figure 1). The co-directional number of pairs found in each spacing length decreases as the spacing lengths increases, even though a long spacing length tail is observed (Figure 1A). The average spacing lengths among co-directional pairs is the lowest (163 bps) and the modal spacing length is 2 bps. The short modal spacing length reflects that the co-directional gene pairs tend to be grouped in operons and separated by short distances [23]. However, the long tail distances and the average spacing lengths of the co-directional pairs suggest that among the prokaryote genomes there is also a small minority of co-directional pairs that might be non-operonic. We have noticed that some prokaryote genomes have long intergenic regions that may be the result of pseudogenes accumulated in prokaryote genomes undergoing processes such as niche change, host specialization or weak selection strength [24]. The longest spacing lengths are found in *Mycobacterium leprae* and in the *Rickettsia* genomes, which appear to be in an extensive process of extensive genome



Submitted to PLoS ONE

Figure 1. Distribution of the spacing lengths

Distribution of the spacing lengths between genes in co-directional (A), convergent (B) and divergent (C) orientation. A representation for each transcriptional orientation is shown. A distribution of the short spacing lengths between co-directional genes is showed in detail (A).

degradation via pseudogenization [25]. The convergent number of pairs found in each spacing length decreases as the spacing lengths increase as happens in the co-directional spacing lengths, but with a much longer tail for long distances (Figure 1B). However, there is a remarkable increase of number of pairs at around 30 bps spacing length, but the explanation of this increase is not addressed in this paper. Although the modal spacing length is similar to the co-directional distribution (0 bps), the higher mean of the convergent spacing lengths (195 bps) indicates that the convergent spacing lengths tend to be longer than the co-directional ones, probably due to the fact that the convergent gene pairs are basically inter-operonics. In contrast, the distribution of the divergent spacing lengths is totally different. The divergent number of pairs increases gradually up to around 100 bps and remains high up to 175 bps, and then it decreases gradually with a long tail for long distances between genes (Figure 1C). The divergent distribution shows the highest mean of the spacing lengths (273 bps) and had a modal spacing length of 135 bps. These results indicate that the divergent gene pairs are basically inter-operonics, like the convergent ones. However, they require a longer space between them than the convergent and co-directional pairs, probably because of the accommodation of several upstream regulatory signals for both genes of the pair [26]. Therefore, maintaining of the upstream regulatory signals seems to constrain the compression of the DNA more than the operon structures or the termination signals. Also it is worth commenting that the convergent spacing lengths appeared to follow a phase bias, at least among the short spacing lengths (up to 30 bps) (Figure 1B). This phase bias is the result of the continuous creation and

elimination of overlaps that is reflected among the closely spaced genes. This uneven distribution of small separation distances arises from the non-uniform distribution of reverse-complement stop codons [27]. Phase 0 ($x = 0, 3, 6, 9, \dots$) which is prevalent, is the one that has more concentration of stop codons. Neither co-directional nor divergent pairs show any phase bias.

Insights into the short co-directional spacing lengths

We focused our attention in the fluctuations observed within the short co-directional spacing lengths (up to 15 bps). Apparently there is a decrease of the co-directional gene pairs separated by spacing lengths from around 5 to 8 bps (Figure 1A). In order to confirm such fluctuations we fit a smoothed decay function of the form

$$p_{\text{obs}}(x) = be^{ax}$$

to the observed distribution of co-directional spacing lengths x . We obtained values $a = 15,301.406$ and $b = -0.0288$ for the parameters by fitting a least-squares regression line to the logarithm of the values in the histogram of observed spacing lengths over the range $x = 0 \dots 50$. We used a function of this form because an exponential drop-off was expected due to primarily, the expectance of a great amount of short spacing lengths between co-directional genes (due to the operon structures presence [23]) and secondly, the selective pressure to reduce the non-coding DNA content [1,2]. Both facts combined may contribute to an exponential distribution of the spacing lengths, with a peak around short spacers and an exponential decay. Comparing the expected number of pairs and the observed number of pairs in every spacing length three areas of the Figure 2 were worth studying and could give us relevant biological information. The spacing length ranges from 1 to 4 bps and from 9 to 15 bps spacers showed overrepresentation of number of pairs, while in the range from 5 to 8 bps the number of pairs dropped off. Beyond 15 bps the number of pairs observed for every spacing length was more similar to the expected number. In order to investigate such fluctuations we studied these three areas mentioned above and we included two more in order to make a good comparison. These

additional areas were the spacing lengths beyond 15 bps and the whole spacing lengths. Although the tendency towards the reduction of non-coding DNA, the genomes must maintain a space between genes long enough to accommodate the regulatory signals such as the SD sequence, which is needed for an efficient translation of the gene. Therefore, SD presence may influence the length of the spacers because usually, it is located between the upstream gene stop codon and the downstream gene start codon of the pair, maintaining a proper distance between the SD sequence and the downstream gene start codon [10].

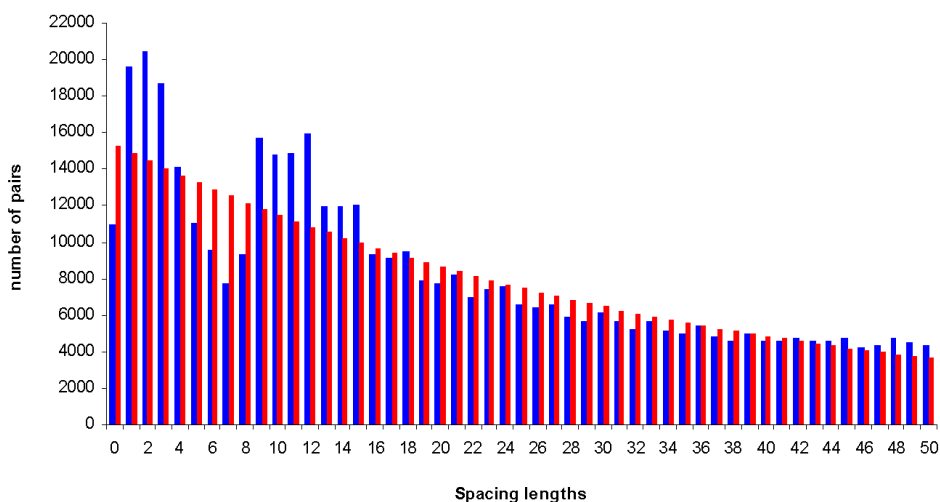


Figure 2. Observed vs. expected co-directional spacing lengths

Red bars show the expected number of gene pairs separated by each spacing length while blue bars show the observed number of gene pairs separated by each spacing length up to 50 bps long, among the co-directional spacing lengths.

SD presence within the prokaryote genomes

In order to detect the presence and the location of the SD sequence we used a free energy calculations approach described in the methods section [11]. The genomes that have more genes with the SD sequence are the ones belonging

to Firmicutes phylum, specially the Bacillales class. Interestingly, the genome that shows more genes with the SD sequence predictions is *Listeria innocua* Clip11262 with ~93% of the genes with the regulatory motif. In contrast, the genomes that have fewer genes with SD predictions are the genomes that the 16S rRNA tail is short and not very well defined (Table S1), being the *Mycobacterium avium* 104 the genome with less genes with SD sequence (0.59%). Apart from this one, the genomes with less percentage of SD presence, but with a conserved 16S rRNA tail, are two genomes of the Bacteroidetes phylum, the *Gramella forsetii* KT0803 (6.34%) and *Flavobacterium psychrophilum* JIP02/86 (3.07%). The fact that the number of genes with SD sequence varies from 0.59% to 93% implies that the populations of genes without SD or leaderless genes are really significant as other authors have pointed out [12]. Furthermore, from the 530 prokaryote genomes analysed here, there are 248 prokaryote genomes that less than 50 % of their genes do not have SD sequence and there are around 40 genomes with fewer than 20 % of the genes with SD sequence (Table S1). This might indicate that there are prokaryote genomes using alternative translation initiation processes to translate their genes, and there are even genomes that do not use the SD sequence to bind the ribosome in the translation initiation process. Finding the alternative processes to the SD guided one is an issue that remains still opened.

The genomes with a large number of genes with SD sequence seem to concentrate such a regulatory motif in a distance range from 4 to 12 bps before the start codon in the majority of the 530 prokaryote genomes analyzed (Table S1). This distance range that we have obtained is slightly different from the previously defined one (from 7 to 12 bps) [9,10,12]. The difference may be explained by the fact that we are calculating the distance from the base that binds the 5'A of the 16S rRNA tail sequence 3'-CCUCCA-5' to the first base of the start codon [11]. Other authors calculate the difference from the core of the SD sequence to the start codon and then they obtain longer distances. As the percentage of genes with SD decrease the distance range between the SD and the start codon becomes more scattered, particularly in the genomes with very

low percentages of genes with SD sequence. Also as the percentage of genes with SD sequence decreases, the number of possible mispredicted start codons or downstream start codon reflections (see Materials and Methods) increases, although the correlation is very bad ($R^2 = 0.255$). Therefore, it appears that the prokaryote genomes, whose translation initiation process is mainly guided by the SD binding the ribosome, have an optimal space conserved between the SD motif and the start codon, which can vary slightly depending on the species [10].

SD presence within the co-directional short spacers

We analyzed the SD presence within the spacing lengths from 0 up to 50 bps among the prokaryote genomes (Figure 3). The percentage of genes with SD decreases gradually from 60% to 50% in the genes with a spacer from 0 to 4 bps before it. The percentage of genes with SD is within the percentages range from 40% to 47% in the genes that have a spacer from 5 to 8 bps. The percentage of genes with SD sequence rises 60% in the genes with a spacer from 9 to 15 bps with a maximum at 12 bps (69%). Surprisingly the genes separated from 0 to 4 bps have a high SD presence, although these SD sequences overlap the previous coding sequence. The genes that have more SD presence are the ones that are separated from 9 to 15 bps probably due to the fact that these genes have a previous space long enough to accommodate the SD sequence properly. As the spacing lengths increase the proportion of genes with SD divided by genes without SD sequence is closer to 1, although the proportion of genes with SD sequence is frequently slightly higher. Beyond 50 bps the percentage of genes with SD sequence is close to the number of genes without SD sequence. After 44 bps we can find some spacing lengths with a percentage of genes without SD sequence higher than the percentage of genes with SD sequence. Distances between co-directional genes shorter than 40 bps are associated to genes belonging to the same operon structure [28]. Then we can say that within an operon there are more genes with SD than without SD sequence. Actually, Ma and co-workers pointed out that the SD sequence prevalence is significantly in genes within operons [10].

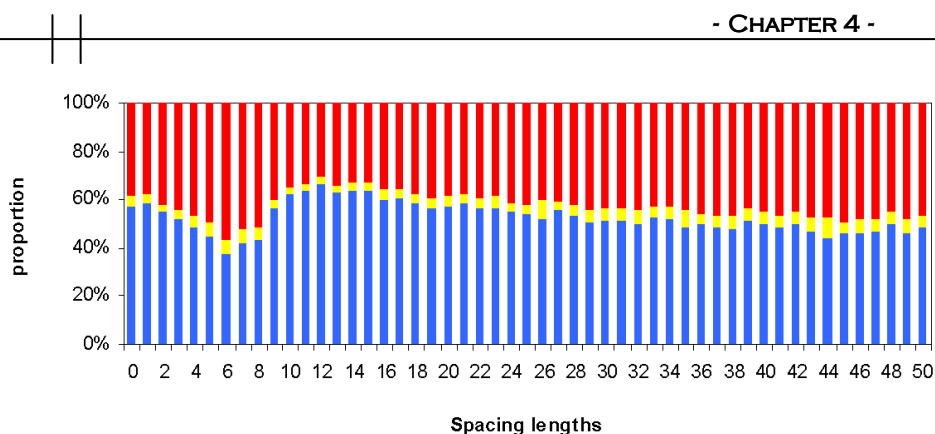


Figure 3. SD presence along the short co-directional spacers

Proportion of genes with SD sequence (blue color), with a *mispredicted start codon* or a *downstream start codon reflection* (yellow color) and *without SD sequence* (red color) in each of the spacing lengths up to 50 bps long.

Apart from the gene groups that have or do not have the SD sequence we defined another group that appears to have the SD sequence downstream to the start codon. This group that has the SD sequence a few bases downstream of the gene start may mean a mispredicted start codon if there is a sudden drop in ΔG° value at 1 bps or immediately close [11]. Another explanation for these downstream drops of ΔG° value may reflect the presence of downstream alternative start codons. The annotation algorithms could detect as a start codon part of the SD sequence. In fact, these genes with the SD sequence predicted downstream of the start codon usually have an overrepresentation of the GTG codon as start codon (Table 1 and see [11]), which fits very well within the SD motif core. The gene groups called SD sequence and downstream start codon reflection and mispredicted start codon or downstream start codon reflection (see Materials and Methods) show a percentage of GTG of 46.6% and 52.2% respectively (Table 1). These are high percentages of GTG usage in comparison to the upstream SD or the non SD groups. If the downstream prediction is due to a start codon misprediction this false start codon should actually be the SD

sequence, and the actual start codon would be a few codons downstream (Figure 4D). For instance, the *Clostridium tetani* E88 gene CTC00194 has a downstream prediction at 1 bps and the start codon is GTG, but the actual one would be the ATG codon located 4 codons downstream and the GTG one is part of the SD sequence (Figure 4D). The percentage of genes classified within the mispredicted start codon or downstream start codon reflection group does not vary significantly depending on the spacing lengths that separate a gene pair. Therefore, the mispredicted start codons can be found in all the genes independently of their spacing length before it.

Start codons	Start codon usage			
	<i>upstream SD sequence</i> (% genes)	<i>SD sequence and downstream start codon reflection</i> (% genes)	<i>mispredicted start codon or downstream start codon reflection</i> (% genes)	<i>without SD sequence</i> (% genes)
AUG	84.9%	49.4%	41.9%	80%
GUG	9.4%	46.6%	52.2%	11%
UUG	5.7%	4%	5.6%	8.6%
other	0.1%	0.1%	0.3%	0.3%

Table 1. Start codon usage among the SD populations

Percentages of start codon usage among genes with an *upstream SD sequence*, genes with a *SD sequence and downstream start codon reflection*, genes with a *mispredicted start codon or downstream start codon reflection* and genes *without SD sequence*.

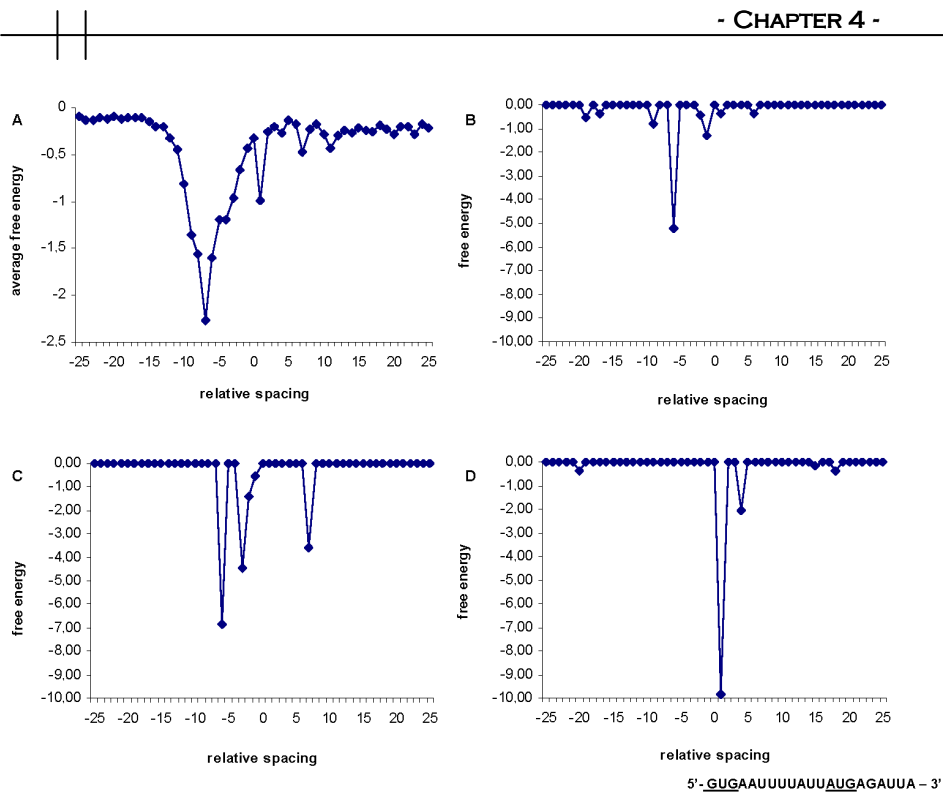


Figure 4. ΔG^0 values in the translation initiation region for the *C. tetani* E88 genes

For all the *C. tetani* E88 genes we calculated the average of ΔG^0 values in the translation initiation region for each relative spacing position (A). A dramatic drop in the ΔG^0 value before the start codon, especially from 4 to 11 nucleotides before the start codon, indicates presence of SD locations. The sudden drop in the ΔG^0 value immediately after the first base in the start codon may indicate potentially wrong SD predictions, while the sudden drop at, for instance, 7 bps may indicate downstream start codon reflections. A drop in the ΔG^0 values at 6 bps to the start codon of the CTC00136 gene indicates that it is a gene with an upstream SD sequence (B). The gene CTC00983 shows three drops in ΔG^0 value (C). The drop at 6 bps to the start codon falls within the optimal distance between the SD sequence and the start codon (from 4 to 11 bps), while the drop at 3 bps falls out of this optimal distance. Looking downstream of the start codon the drop in ΔG^0 value falls at 7 bps after the first base of the start codon, which may mean that there is a start codon reflecting a SD sequence around this position. A dramatic drop in ΔG^0 value is observed 1 bps after the first base of the GTG start codon of the gene CTC00194 (D).

Submitted to PLoS ONE

This drop is indicating a mispredicted start codon (GTG underlined in the sequence) and as it can be observed in the downstream sequence, which is denoted below the graph, this gene has an alternative start codon (ATG underlined in the sequence) only 4 codons downstream of the mispredicted one.

Location of the SD motif within the short co-directional spacers

We studied the presence or absence of the SD sequence within the co-directional gene pairs separated by the different spacing length ranges, which have been described above. In all of them we find SD presence (Figure 5A). The set of all the gene pairs shows more genes with SD than without SD sequence by a factor of 1.25. A similar proportion we obtain in the set of gene pairs separated by spacers longer than 15 bps (1.21). Although the translation in prokaryotes is mainly guided by the SD sequence that can bind the ribosome [10], it seems that there are only a slightly higher number of genes with SD than without SD sequence. These results agree with the idea that non-SD-led genes are as common as SD-led genes [12]. The gene pairs separated from 1 to 4 bps should have the SD sequence along the end of the coding sequence of the previous gene. Although this constrains the 3'-end of the upstream gene, as Eyre-Walker found (1996) [21], a higher number of genes with SD sequence and separated from the previous one by a spacing length from 1 to 4 bps were found. In this spacing length range, the proportion of genes with SD divided by the genes without SD sequence is slightly higher (1.28) than in the total gene set (1.25). Within the spacing lengths ranging from 9 to 15 bps we find the highest proportion of genes with SD divided by genes without SD sequence (1.85). This might indicate that within the spacing lengths ranging from 9 to 15 bps, generally, we find the optimal distances between genes that allow a least constrained accommodation of the SD motif, which is commonly found from 4 to 12 nucleotides to the start codon in prokaryotes (Table S1). In contrast, within the spacing lengths ranging from 5 to 8 bps a decrease of the genes with SD sequence is observed (Figure 5A) and even, within this range we find a proportion of genes with SD divided by genes without SD sequence lower than 1

(0.8). Therefore, it seems that the genes with SD may preferentially have an intergenic distance to the previous gene either shorter than 5 bps, with the SD sequence overlapping the upstream coding sequence, or longer than 8 bps, with the SD well accommodated within the intergenic region. In contrast, the intergenic distances from 5 to 8 bps may make the SD motif of a gene to overlap with the previous stop codon, constraining the spacing length and the stop codon usage of the previous gene.

The optimal distance range between the SD sequence and the start codon can vary depending on the genome (Table S1 and [10]). This variation can slightly change the spacing length ranges defined here. For instance, in *E. coli* the number of adjacent genes separated by spacing lengths from 1 to 4 bps and from 9 to 13 bps is overrepresented, while from 5 to 8 bps is underrepresented. The *E. coli* genome fits very well with the general spacing length distribution in prokaryote genomes (Figure 1A). Nevertheless, in the *Bacillus subtilis* genome, the number of adjacent genes separated by spacing lengths from 0 to 6 bps and from 11 to 18 is overrepresented, while from 7 to 10 bps is underrepresented. In *E. coli* genome the optimal distance between the beginning of the SD sequence and the start codon is from 3 to 10 bps (Table S1) with a maximum drop in ΔG° value at 5 bps before the start codon, while in *B. subtilis* genome the optimal distance is from 3 to 11 bps (Table S1) with a maximum drop in ΔG° value at 6 bps. These differences in respect to the location of the SD sequence contribute to the reason for the spacing lengths overrepresented and underrepresented in each prokaryote genome.

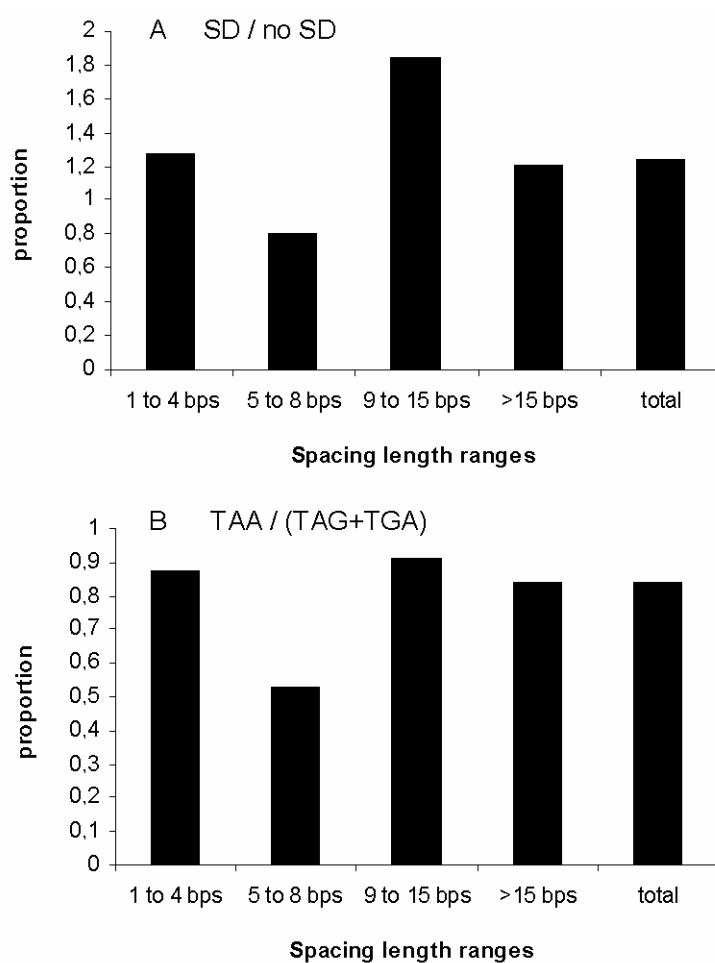


Figure 5. SD presence and stop codon usage among the co-directional gene pairs
Histogram of the proportion of number of genes with SD divided by the number of genes without SD (A) and histogram of the proportion of the TAA usage divided by TGA and TAG usage as a stop codon (B) among the ranges of spacing lengths analyzed. Each bar is related to the ranges of spacing length studied and previously defined in the text.

Adaptation of the stop codon usage and the spacing lengths to the SD presence

TAA is used in preference to TGA, which itself is used in preference to TAG [29]. TAA is the preferred stop codon because of the greater availability of TAA-cognate release factor(s) or lower levels of translational read-through [29]. TGA and TAG might be used when they have an additional function to the one of coding for a stop signal [21]. For instance the TGA stop codon is used in one of the extremely common overlaps found in prokaryotes, the co-directional overlap of 4 bps, which includes the start codon of an upstream gene (ATG, GTG or TTG) and the TGA stop codon of a downstream gene [30,31]. The proportion of TAA stop codon divided by the sum of the TGA and the TAG stop codons observed in each spacing length ranges when analyzed, is closer to 1 (from 0,84 to 0,91) with the exception of the genes separated from 5 to 8 bps (Figure 5B), whose proportion falls to 0,53. Therefore, it seems that an upstream gene which is in the distance from 5 to 8 bps to the next one may use in preference TGA or TAG as stop codon (Figure 5B); and the SD presence of the downstream gene decreases (Figure 5A). This adaptation of the stop codon usage of a gene could be the result of the SD sequence of the next gene overlapping its stop codon. The stop codons TGA and TAG may fit more easily within the SD motif. Recently, in a virus, whose genome is highly compacted, it was described that overlaps of a stop codon and the SD sequence resulted in a common motif GGTGA. This is a clear adaptation of the upstream gene stop codon to become part of the SD motif maintaining its function as stop codon.

In order to study the possible adaptation of the stop codon usage and the spacing lengths to the SD presence among the co-directional short spacers in prokaryote genomes, we built sequence logos for the intergenic regions of *E. coli* from 1 to 12 bps (Figure 6). From 1 to 4 bps we observe a high proportion of As and Gs before the upstream stop codon, which may indicate the SD presence along the end of the upstream gene coding sequence (Figure 6). Looking from 2 to -20, a drop in the ΔG^0 value is observed before these regions

of high frequency of As and Gs. The stop codon usage is biased to the TAA use and the proportion of genes with SD and without SD sequence is higher than 1 in each spacer (Table 2). From 5 to 6 bps we find few genes with SD sequence to build the logo (4 and 3 genes respectively). There are more genes without SD than with SD sequence in the genes separated by such spacing lengths (Table 2). Therefore the 5 and the 6 bp distances are the most compromised by the SD presence. From 7 to 8 bps the high frequency of As and Gs that could indicate the SD presence is overlapping the downstream gene stop codon and there is a drop in ΔG° value just before the stop codon (looking from 2 to -20). Interestingly, the SD sequence seems to adopt the TGAGG pattern when the E. coli genes are separated by 7 or 8 bps (Figure 6). From 9 to 12 bps the high frequency of As and Gs is between the upstream gene stop codon and the downstream gene start codon. The drop in ΔG° value is around the middle of the intergenic region and the TAA stop codon is used in preference. In fact, this stop codon could bind well with the end bases of the E. coli 16S rRNA tail (3'-AUU-5'), especially in a length of 9 and 10 bps (Figure 6).

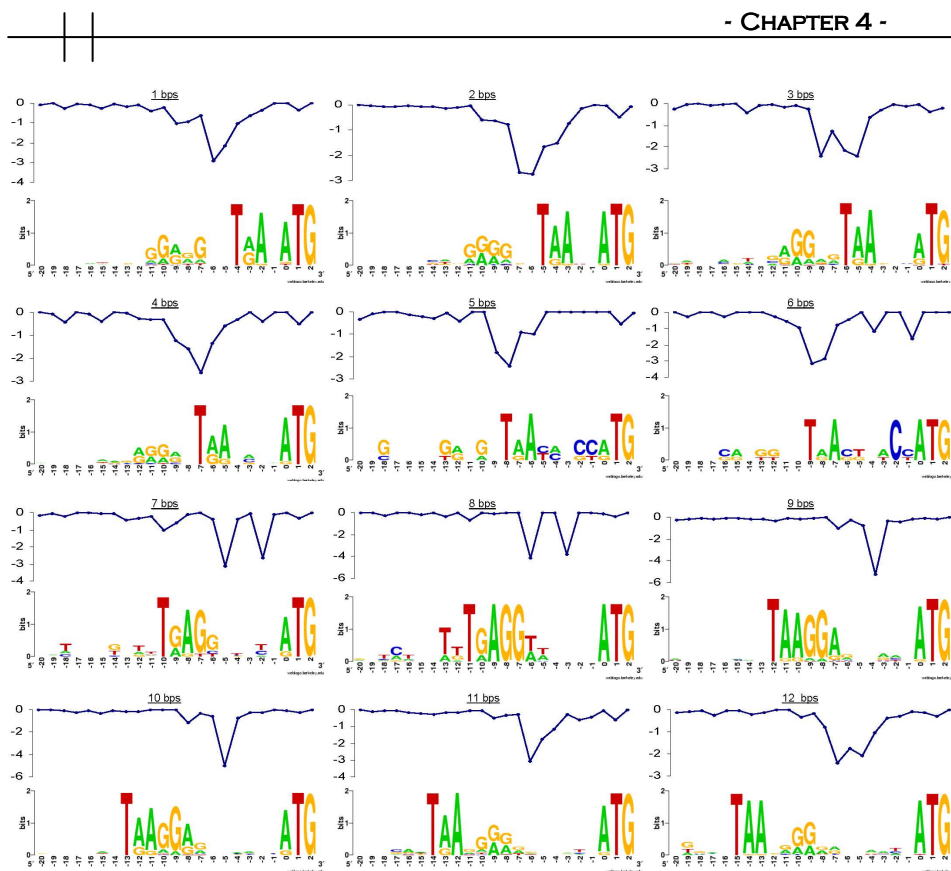


Figure 6. Sequence logos for *E. coli*

Sequence logos of the nucleotides between positions -20 and 2 of the genes with a spacer from 1 to 12 bps long. For each position, the sequence logo shows the amount of information content and the frequency of nucleotides. The blank positions mean that there is no information content, while those with information content contain a stack of nucleotides. The size of the nucleotide character is proportional to its frequency at that position. Each sequence logo has the average of ΔG^0 values from -20 to 2 bps of the genes separated by each of the spacers analysed. The drops in ΔG^0 values indicate the position where the 5' A of the 16S rRNA tail (3'-CCUCCA-5') can bind the SD sequence. These drops are before the regions with high frequency of As andGs.

When the SD sequence overlaps with the upstream coding sequence or stop codon, its strength and relative distance to the downstream start codon do not vary significantly (Figure 6, Table S1 and see [21]). However, the SD presence may make the intergenic lengths from 5 to 8 bps less favored and cause an adaptation of the stop codon usage. In *E. coli* this adaptation is reflected in the prevalence of TGA usage forming the TGAGG pattern for co-directional genes separated by 7 or 8 bps. However, the adaptation could be slightly different depending on the prokaryote species. For instance, in the *B. subtilis* genome, genes separated by 10 bps use TAG as stop codon instead of TGA, resulting in the SD motif TAGGAGG. Two mechanisms could cause the SD sequence overlaps with a TGA or a TAG stop codon. The first mechanism may consist of a deletion of a portion of an intergenic sequence followed by a mutation at the upstream stop codon that changes the most frequent TAA stop codon to TGA or TAG. The second mechanism may include merely one step, which is a deletion of a portion of an intergenic sequence when the stop codon of the upstream gene is already either a TGA or a TAG. This second mechanism seems a more parsimonious explanation. This adaptation might reflect the compression process of the genome size among prokaryote genomes.

Spacing lengths (bps)	Genes with SD	Genes without SD	Proportion genes with SD / genes without SD
1	22	17	1.29
2	28	22	1.27
3	31	16	1.93
4	14	13	1.08
5	4	8	0.50
6	3	10	0.30
7	14	11	1.27
8	10	6	1.67
9	48	8	6.00
10	48	6	8.00
11	39	9	4.33
12	30	10	3.00

Table 2. Genes with and without SD separated by short co-directional spacers in *E. coli*

Number of genes with SD and without SD sequence in each spacing length from 1 to 12 bps in the *E. coli* genome.

Materials and Methods

Data retrieval and study of the distribution of spacing lengths

The complete genome sequences of 678 prokaryote chromosomes were downloaded from the NCBI ftp site (<ftp://ftp.ncbi.nlm.nih.gov/genomes/>). Perl scripting was used to extract the overlaps and the spacers between adjacent genes. Unfortunately, in prokaryotes, all the analysis of intergenic regions are hampered by the annotation errors such as incorrect initiation codon prediction, falsely predicted genes and frameshifts [32,33,34]. Taking into account only the gene pairs that are assigned with COG category, we observed the same profiles in the Figures 1, 2, 3, 4 and in the Table 1 shown here. Since the mispredictions can affect both the well characterized genes and those that are not [35], we worked with all the genes contained in the prokaryote genomes. The spacing lengths were represented graphically and we focused our attention in the co-directional spacing lengths. The distribution of the co-directional spacers was analyzed using a smoothed decay function of the form $pobs(x) = beax$. With this function we compared the number of gene pairs observed and expected separated by each spacing length up to a length of 50 bps. This was useful to point out the short spacing lengths that are over or underrepresented. We studied the stop codon usage and the SD presence among the gene pairs separated by the ranges of spacing lengths from 1 to 4 bps, from 5 to 8 bps and from 9 to 15 bps. In order to make comparisons of the stop codon usage and the SD presence within the spacing lengths analyzed, we added two more spacer groups. These groups were the spacers longer than 15 bps and all the spacers between genes.

Stop codon usage analysis

Since we studied the co-directional spacing lengths we only considered the pairs of genes with an orientation (->->) or (<-<-). Taking into account the DNA direction from 5' to 3', in the case of orientation (->->) we looked at the stop codon of the upstream gene, while taking into account the DNA direction from 3' to 5', in the case of orientation (<-<-) we looked at the stop codon of the downstream gene. The region that involves the upstream gene stop codon, the possible downstream SD motif and the downstream gene start codon (from -20 to 2) was represented by WebLogo [36] in *E. coli* spacing lengths from 1 to 12 bps (Figure 6).

Location of the SD motif

Since SD sequence was discovered and characterized [6], two different approaches have been used to identify and locate the SD motif in prokaryotes. These approaches are based on either sequence similarity or free energy calculations. In this paper we used the Starmer and co-workers method based on energy free calculations [11]. We chose this method because it is based on thermodynamic considerations of the 30S binding to the mRNA and overcomes the limitations of sequence analysis [11]. We extracted the 16S rRNAs from the NCBI ftp site (<ftp://ftp.ncbi.nlm.nih.gov/genomes/>). For each 16S rRNA sequence of each organism we looked at in 5' direction for the first instance of the three letter motif, 5'-GAT-3', which was found consistently on the 5' end of the tails of the 16S rRNAs with known structure. The location of this motif was used to define, up to the end of the 3' tail, the 16S rRNA tail of each organism. For species that have two or more copies of the 16S rRNA gene, we calculated the consensus sequence of all the tails. If the different tails observed did not follow a consensus, then we used the majority of the 16S rRNA gene tails. All the 16S rRNA tails of the 678 organisms were examined manually. We only used the genomes that have a 16S rRNA tail close to the most conserved motif 5'-GAUACCUCUU-3'[37]. Therefore, we only considered the SD locations of the genes contained in the 530 prokaryote chromosomes that have the

conserved 16S rRNA tail (Table S1). The scripts to calculate the free energies of the 16S rRNA tail binding with the mRNA were downloaded from <http://sourceforge.net/projects/freetobind> and were included in our Perl scripts. We located the SD sequence by the position of the lowest ΔG^0 value calculated from 35 bps upstream to the initiation codon to 35 bps downstream from the initiation codon. The gene was assumed not to have a SD sequence if $\Delta G^0 > -3.4535$ Kcal/mol. The threshold used is based on the work of Ma and co-workers [10]. In order to pinpoint the exact SD position we used the relative spacing parameter [11], that means that we calculated the distance between the first residue of the start codon and the 5' A of the rRNA sequence 5'-ACCUCC-3' in the positions around the start codon. If the SD motif is located before the start codon the relative spacing will be negative, while if the SD motif is located after the start codon the relative spacing will be given as a positive number.

Classifying the SD motif signal

Among the prokaryote genes with SD sequence there are genes with drops in ΔG^0 value upstream to the gene start, genes with drops in ΔG^0 value upstream and downstream to the gene start and genes with drops in ΔG^0 value downstream to the gene start. The distance between an upstream SD sequence and the start codon was studied for all the 530 prokaryote chromosomes and was tabulated for each chromosome (Table S1). For instance, the *C. tetani* E88 genes tend to have their SD sequence from 4 to 11 nucleotides before the start codon. Taking into account the different drops in ΔG^0 value observed and the most frequent distance between the upstream SD sequence and the start codon observed in each genome, we classified the different SD motif signals in three groups. Figure 4 shows the average ΔG^0 values in the translation initiation region for *C. tetani* E88 genome (Figure 4A) and the three ΔG^0 values in the translation initiation region observed in three different *C. tetani* E88 genes, which contribute to the overall average ΔG^0 values of the genome. We consider that a gene has an upstream SD sequence if it has at least a clear drop in ΔG^0

value within the most frequent distance range between the SD sequence and the start codon of the genome (Figure 4B). If a gene has drops in ΔG^0 value upstream and downstream to the start codon and one of them falls within the most frequent distance range between the SD sequence and the start codon, we consider that the gene has SD sequence and a downstream start codon reflection (Figure 4C). The genes with only drops in ΔG^0 value downstream of the gene start may have either a mispredicted start codon (see above in Results and Discussion section or in [11]) or a downstream start codon reflection (Figure 4D). Gene cases shown in Figure 4B and 4C were considered to have SD sequence for the purpose of our analysis. Therefore, we distinguished the genes of each genome in three groups: the genes with SD sequence (upstream SD sequence and SD sequence and a downstream start codon reflection groups), the genes with a hypothetical mispredicted start codon or a downstream start codon reflection, and the genes without SD sequence. Only for start codon usage analysis (Table 1) we considered all the groups mentioned separately.

Acknowledgments

We would like to thank Joshua Starmer and co-workers for making available their programs for detecting Shine-Dalgarno motifs, and especially Joshua Starmer for his kind assistance. Also we would like to thank Richard Tuby for his help in writing the manuscript.

References

1. Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, et al. (2002) Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Res* 30: 4264-4271.
2. Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet* 17: 589-596.

3. Pallejà A, Guzman E, Garcia-Vallvé S, Romeu A (2008) In silico prediction of the origin of replication among bacteria: a case study of bacteroides thetaiotaomicron. *OMICS* 12: 201-210.
4. Frank AC, Lobry JR (2000) Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics* 16: 560-561.
5. Koonin E, Wolf Y (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.*
6. Shine J, Dalgarno L (1974) The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A* 71: 1342-1346.
7. de Smit M, van Duin J (1994) Translational initiation on structured messengers. Another role for the Shine-Dalgarno interaction. *J Mol Biol* 235: 173-184.
8. Weyens G, Charlier D, Roovers M, Piérard A, Glansdorff N (1988) On the role of the Shine-Dalgarno sequence in determining the efficiency of translation initiation at a weak start codon in the *car* operon of *Escherichia coli* K12. *J Mol Biol* 204: 1045-1048.
9. Osada Y, Saito R, Tomita M Analysis of base-pairing potentials between 16S rRNA and 5' UTR for translation initiation in various prokaryotes. *Bioinformatics* 15: 578-581.
10. Ma J, Campbell A, Karlin S (2002) Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol* 184: 5733-5745.
11. Starmer J, Stomp A, Vouk M, Bitzer D (2006) Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLoS Comput Biol* 2: e57.
12. Chang B, Halgamuge S, Tang S (2006) Analysis of SD sequences in completed microbial genomes: non-SD-led genes are as common as SD-led genes. *Gene* 373: 90-99.

13. Slupska M, King A, Fitz-Gibbon S, Besemer J, Borodovsky M, et al. (2001) Leaderless transcripts of the crenarchaeal hyperthermophile *Pyrobaculum aerophilum*. *J Mol Biol* 309: 347-360.
14. Boni I, Artamonova V, Tzareva N, Dreyfus M (2001) Non-canonical mechanism for translational control in bacteria: synthesis of ribosomal protein S1. *EMBO J* 20: 4222-4232.
15. Kolev V, Ivanov I, Berzal-Herranz A, Ivanov I (2003) Non-Shine-Dalgarno initiators of translation selected from combinatorial DNA libraries. *J Mol Microbiol Biotechnol* 5: 154-160.
16. Chen H, Bjerknes M, Kumar R, Jay E (1994) Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs. *Nucleic Acids Res* 22: 4953-4957.
17. Kozak M (1999) Initiation of translation in prokaryotes and eukaryotes. *Gene* 234: 187-208.
18. Ringquist S, Shinedling S, Barrick D, Green L, Binkley J, et al. (1992) Translation initiation in *Escherichia coli*: sequences within the ribosome-binding site. *Mol Microbiol* 6: 1219-1229.
19. Schurr T, Nadir E, Margalit H (1993) Identification and characterization of *E. coli* ribosomal binding sites by free energy computation. *Nucleic Acids Res* 21: 4019-4023.
20. Eyre-Walker A, Bulmer M (1993) Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Res* 21: 4599-4603.
21. Eyre-Walker A (1996) The close proximity of *Escherichia coli* genes: consequences for stop codon and synonymous codon use. *J Mol Evol* 42: 73-78.
22. Peng X (2008) Evidence for the horizontal transfer of an integrase gene from a fusellovirus to a pRN-like plasmid within a single strain of *Sulfolobus* and the implications for plasmid survival. *Microbiology* 154: 383-391.

23. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J (2000) Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci U S A* 97: 6652-6657.
24. Mira A, Pushker R (2005) The silencing of pseudogenes. *Mol Biol Evol* 22: 2135-2138.
25. Fuxelius H, Darby A, Cho N, Andersson S (2008) Visualization of pseudogenes in intracellular bacteria reveals the different tracks to gene destruction. *Genome Biol* 9: R42.
26. Pérez-Rueda E, Gralla J, Collado-Vides J (1998) Genomic position analyses and the transcription machinery. *J Mol Biol* 275: 165-170.
27. Kingsford C, Delcher AL, Salzberg SL (2007) A unified model explaining the offsets of overlapping and near-overlapping prokaryotic genes. *Mol Biol Evol* 24: 2091-2098.
28. Moreno-Hagelsieb G, Collado-Vides J (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* 18 Suppl 1: S329-336.
29. Sharp P, Bulmer M (1988) Selective differences among translation termination codons. *Gene* 63: 141-145.
30. Fukuda Y, Nakayama Y, Tomita M (2003) On dynamics of overlapping genes in bacterial genomes. *Gene* 323: 181-187.
31. Johnson ZI, Chisholm SW (2004) Properties of overlapping genes are conserved across microbial genomes. *Genome Res* 14: 2268-2272.
32. Devos D, Valencia A (2001) Intrinsic errors in genome annotation. *Trends Genet* 17: 429-431.
33. Natale DA, Galperin MY, Tatusov RL, Koonin EV (2000) Using the COG database to improve gene recognition in complete genomes. *Genetica* 108: 9-17.
34. Brenner S (1999) Errors in genome annotation. *Trends Genet* 15: 132-133.
35. Palleja A, Harrington ED, Bork P (2008) Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? *BMC Genomics* 9: 335.

36. Crooks G, Hon G, Chandonia J, Brenner S (2004) WebLogo: a sequence logo generator. *Genome Res* 14: 1188-1190.
37. Cannone J, Subramanian S, Schnare M, Collett J, D'Souza L, et al. (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* 3: 2.

Supporting Information

Table S1. Percentage of SD presence, preferred location for the SD motif and 16S rRNA tail used of each prokaryote genome.

Excel file that shows the percentage of the genes in each SD populations (see Materials and Methods), the optimal distance between the beginning of the SD motif and the first base of the start codon and the 16S rRNA tail used to calculate the binding between the 16S rRNA and the mRNA for each prokaryote chromosome. Taxonomical information of each prokaryote genome is also given. The 530 prokaryote chromosomes are sorted by percentage of genes with SD.

Because of its size you will find this table following this link:

<http://genomes.urv.cat/albert/TableS1>

09-PONE-RA-08141 Receipt of New Manuscript by PLoS ONE

Dear Dr. Albert Pallejà (and Santiago Garcia-Vallve, Antoni Romeu)

On January 15, 2009, we received your Research Article entitled "Adaptation of the Short Co-directional Spacers to the Shine-Dalgarno Motif in Prokaryote Genomes" by

Albert Pallejà (Rovira i Virgili University)
Santiago Garcia-Vallve (Rovira i Virgili University)
Antoni Romeu (Rovira i Virgili University)

Your manuscript has been assigned the manuscript #: 09-PONE-RA-08141.

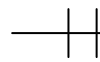
****Please ensure that you check the spelling of your name and institution above; if it appears incorrectly we ask that you log on to the PLoS ONE Journal Management System at <http://one.plosjms.org> and ensure that your profile details are correct.**

If you do not have a username or password; please use the 'Unknown/Forgotten Password' link provided on the log in page.**

Your manuscript will be assigned to an academic editor within the next few days. We will keep you informed about the progress of your manuscript or you can check the status of yourself by logging on to the PLoS ONE online manuscript tracking system at <http://one.plosjms.org>.

Please be aware that you will NOT be required to complete the 'Open-Access Agreement' field until your manuscript is accepted for publication.

Submitted to PLoS ONE



If you have any enquiries or other comments regarding this manuscript, please contact PLoSONE@plos.org.

Thank you for choosing PLoS ONE.

Best wishes,

Peter Binfield, PhD
Managing Editor, PLoS ONE

ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGACAGCTGATAGG
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGGCGGTGGTCAGGCAGCGCAGCATGGGAAAA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
GATCAGATGACAGATAGATAACCACAGAGACATG
CGCGCC**C**GACGCATGATTGATGATCAGATGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGATGATA**H**AGTGATTAGATAGATGGTGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTTT
GAGAGATAGAGAC**A**TAGGATCGCGCTCGAGCGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AACCGGCCAACCGGTGG**P**TTAGGATAGATGATGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AAAGTGTGTGTGACAGACAG**T**TGATGATAGTACA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGA**E**AGCTGATAG
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCT
CGCGCTCGCTCGAGCGCTAGCTCGAT**R**GAT
CGCGCTCAAACGAGCGCTAGCTCGATCGA
GATAGGTACGCGAAAATGGCAGTAGCTAGCTAG
ATAGGTACGCGGATGAATGGCAGTAGCTAGCTAG
ATCGATCGATCGATCGATCGCGCGGATCGATCG
CAGCATGACACACACACATCGATCGATCGATCG
CCAGGCAGCATAAAGCAGCTGGGTGGTAGGAGT
ATCGATCGATCGATCGATCGATCGATCGATCGAT



PAIRWISE NEIGHBORS: OVERLAPS AND SPACERS AMONG PROKARYOTE GENOMES

.ACGCGAAAATGGC/
.GATAGAGATACAGAA
GTACGCGAAAATGGCAG
TCGCTCGAGCGCTAGCTCGA
AGGTACGCGAAAATGGCAGTA
ATCGATCGATCGATCGATCGCGCG
ATCAGCATGACACACACATGATA
CAGTGCCAGGCAGCATAAAGCAGACG/
ACCAGCAGCTGGGTGGTAGGAGTGTAT/
CGAGTGCCAGGCAGCATAAAGCAGACGAC
GCACCAGCAGCTGGGTGGTAGGAGTGTATC



**PAIRWISE NEIGHBOURS DATABASE: OVERLAPS AND
SPACERS AMONG PROKARYOTE GENOMES**

Albert Pallejà[§], Tomàs Reverter, Santiago Garcia-Vallvé, Antoni Romeu.

Department of Biochemistry and Biotechnology, Rovira i Virgili
University, Tarragona, Catalunya, Spain.

[§]Corresponding author

Rovira i Virgili University

Department of Biochemistry and Biotechnology

Campus Sescelades

C/ Marcel·lí Domingo, s/n

E-43007 Tarragona

Catalonia – Spain

Email addresses:

AP: albert.palleja@urv.cat (corresponding author)

TR: tomas.reverter@urv.cat

SGV: santi.garcia-vallve@urv.cat

AR: antoni.romeu@urv.cat

Submitted to BMC Bioinformatics



Abstract

Background

Although prokaryote genomes live in a variety of habitats and possess different metabolic and genomic complexity, they have architectural features in common. The overlapping genes are a common feature of the prokaryote genomes. The overlapping lengths tend to be short because as the overlaps become longer they have more risk of deleterious mutations. The spacers between genes tend to be short too because of the tendency to reduce the non coding DNA among prokaryotes. However they must be long enough to maintain essential regulatory signals such as the Shine-Dalgarno (SD) sequence, which is responsible of an efficient translation.

Description

PairWise Neighbours is an interactive and intuitive database used for retrieving information about the spacers and overlapping genes among bacterial and archaeal genomes. It contains 1,956,294 gene pairs from 678 fully sequenced prokaryote genomes and is freely available at the URL <http://genomes.urv.cat/pwneigh>. This database provides information about the overlaps and their conservation across species. Furthermore, it allows the wide analysis of the intergenic regions providing useful information such as the location and strength of the SD sequence.

Conclusions

There are experiments and bioinformatic analysis that rely on correct annotations of the initiation site. Therefore, a database that studies the overlaps and spacers among prokaryotes appears to be desirable. PairWise Neighbours database permits the reliability analysis of the overlapping structures and the study of the SD presence and location among the adjacent genes, which may help to check the annotation of the initiation sites.

Submitted to BMC Bioinformatics

Background

The availability of fully sequenced genomes has grown exponentially over the past few years. There is a huge variety of environments for the prokaryote species, as well as different metabolic and genomic complexities. However, their genomes have common architectural principles [1]. The prokaryote genomes contain protein-coding genes, structural RNAs and spacers between genes which are thought to typically contain regulatory signals [2]. These spacers tend to be short because of the selective pressure to minimize the non-functional DNA in prokaryotes [2, 3]. It is a consistent feature of these genomes that the genes often overlap their coding sequences [4]. Under this scenario of genomic compactness due to their physically small environments, the overlapping genes follow the rules that impose the structure of the genetic code and the spacers between genes must adapt their lengths to the requirements of the regulatory signals [2].

One of the regulatory signals that we can find between genes, which is related to an efficient translation, is the Shine-Dalgarno (SD) sequence [5]. The SD sequence is a motif, 5'-GGAGG-3', located at the 5' of the initiation codons and is complementary to the sequence, 5'-CCUCC-3', located at the end of the 16S rRNAs [5]. The ribosome does not need a perfect distance between the SD sequence and the start codon for the initiation of translation. However, it has been studied that when the SD resides within the 4 nucleotides from the initiation codon or when is located as far as 13 nucleotides from the initiation codon, gene expression is decreased drastically [6-8]. The prokaryote species seem to have preferred distances between the SD and the start codon and these distances vary among the species [9], although this sequence has been found mostly from the 7th to the 12th base upstream from the start codon [9-11]. The location of the SD can help to correct the gene annotations [12] and could influence the spacing length and the stop codon usage [13].

Among the prokaryote genomes there is a huge amount of examples of overlapping genes [14-18]. The overlapping lengths tend to be short because of

the selective pressure against long overlaps, as the existence of long overlapping reading frames increases the risk of deleterious mutations. The co-directional overlaps are the most common overlaps, which reflects that this is the most common orientation for a gene pair due to the tendency to be grouped in operons in prokaryote genomes [19-21]. Among the co-directional overlaps the 4 bps overlap is extremely common [4, 14, 22, 23], which permits the upstream stop codon and the downstream start codon overlap and the gene pair is thought to be translationally coupled [24]. The co-directional and divergent overlapping genes can arise by 5'-end elongations when the downstream gene adopts a new start codon within the upstream coding sequence [22], while the co-directional and the convergent overlapping genes can arise by 3'-end extensions after a loss codon event [15]. Overlaps in prokaryotes have been hypothesized to be involved in reducing the genome size in order to increase the density of genetic information [16, 23, 25-27], and in regulating gene expression through translational coupling of functionally related polypeptides [4, 23, 25, 28, 29]. In addition, other authors have used the overlapping pairs as genetic markers for phylogenetic inferences due to its high conservation [30, 31]. Overlapping genes are better conserved across the species than non-overlapping genes [18]. The extent of conservation of the overlapping pairs correlates with the evolutionary distances between the pairs of species [14].

The overlapping genes, as a common structure of the prokaryote genomes, and the spacers between genes are structural features worth studying in prokaryotes. However, the analysis of both the overlapping genes and the spacers between genes is often affected by genome annotation errors [32-34]. An accurate annotation would facilitate the experiments as well as the bioinformatic analysis of gene regulation and gene structure [35]. In this interactive database is stored all the overlapping genes and the spacers of 678 fully sequenced prokaryote genomes. The aim of this database is to provide the users with useful information about the overlapping genes and the spacing lengths between adjacent genes. The conservation of the overlaps across the

species and the SD presence and location within the intergenic regions or the overlapping sequences can be analysed.

Construction and Content

Retrieval of the Spacing lengths and the Overlapping genes

The complete genome sequences of 678 prokaryote genomes were downloaded from the NCBI ftp site (<ftp://ftp.ncbi.nlm.nih.gov/genomes/>). Scripts implemented in Perl language were performed to extract and analyse the spacers and the overlaps between adjacent genes and all the information related (spacing & overlapping lengths, spacing & overlapping sequences, gene orientations, phases, protein functions, gene COGS and stop & start codons of the genes). The gene ids in this database have been formed joining the GenBank Accession Number with the gene name. For instance, the gene id for the HI0038 gene from *Haemophilus influenzae* Rd KW20 is NC_000907.HI0038. Furthermore, each overlap and spacer between adjacent genes has an internal id. The spacing lengths and the overlapping genes have been classified into three types according to their transcriptional direction [2, 15, 25]: i) unidirectional (genes in the same strand overlapping the 3'-end of an upstream gene and the 5'-end of a downstream gene), ii) convergent (genes in opposite strand overlapping the 3'-ends) and iii) divergent (genes in opposite strand overlapping the 5'-ends). In this database we use the term co-directional instead of the unidirectional term. In order to study the phases between adjacent genes, as other authors have previously done [4, 18, 22], we defined three overlapping phases: (i) phase 0 where the downstream gene is in frame with the upstream gene (lengths $n = \dots, -12, -9, -6, -3, 0, 3, 6, 9, 12, \dots$), (ii) phase 1 where the downstream gene is in the reading frame +1 relative to the upstream gene frame (lengths $n = \dots, -11, -8, -5, -2, 1, 4, 7, 10, \dots$) and (iii) phase 2 where the downstream gene is in the reading frame +2 relative to the upstream gene frame (lengths $n = \dots, -10, -7, -4, -1, 2, 5, 8, 11, \dots$).

Submitted to BMC Bioinformatics

Location of the SD sequence and determination of its binding strength

We extracted the 16S rRNAs from the NCBI ftp site (<ftp://ftp.ncbi.nlm.nih.gov/genomes/>). For each 16S rRNA sequence of each organism we looked at the 5' direction for the first instance of the three letter motif, 5'-GAT-3', which was found consistently on the 5' end tails of the 16S rRNAs with known structures. The location of this motif was used to define, up to the end of the 3' tail, the 16S rRNA tail of each organism. For species that have two or more copies of the 16S rRNA gene, we calculated the consensus sequence of all the tails. If the different tails observed did not follow a consensus, then we used the majority of the 16S rRNA gene tails. All the 16S rRNA tails of the 678 organisms were examined manually. The SD sequences for 678 prokaryote genomes have been predicted using computer calculations of the base pairing free energy between translation initiation regions and the 16S rRNA 3' tail. The method used was developed by Starmer and co-workers [12]; and the scripts to calculate the free energies were downloaded from <http://sourceforge.net/projects/freetobind> and were included in our Perl scripts. We located the SD sequence by the position of the lowest ΔG° value calculated from 35 bps upstream to the initiation codon to 35 bps downstream from the initiation codon. The gene was assumed not to have the SD sequence if $\Delta G^{\circ} > -3.4535$ Kcal/mol. The threshold used is based on the work of Ma and co-workers [9]. In order to point the exact SD position we used the relative spacing parameter [12], that means that we calculated the distance between the first residue of the start codon and the 5' A of the rRNA sequence 5'-ACCUCC-3' in each position around the start codon. If the SD motif is located before the start codon the relative spacing will be negative, while if the SD motif is located after the start codon the relative spacing will be given as a positive number. Regardless the gene pair orientation, the SD information and the graph of the ΔG° values is given for the upstream and the downstream gene.

Database Construction

Submitted to BMC Bioinformatics

The huge amount of data generated required a data model to make it possible to work with this data efficiently. The Entity-relationship model, showed in Figure 1, was designed and transformed in a MySQL database. A web application was developed using the framework web TurboGears. This Python framework MVC (Model-View-Controller) is an advanced tool to create data consulting systems quickly, efficiently and consistently. The BLAST search tool [36] was installed in our server and is used to study the conservation of the gene overlaps. All the graphs are generated at the user side by a Java Script library named PlotKit.

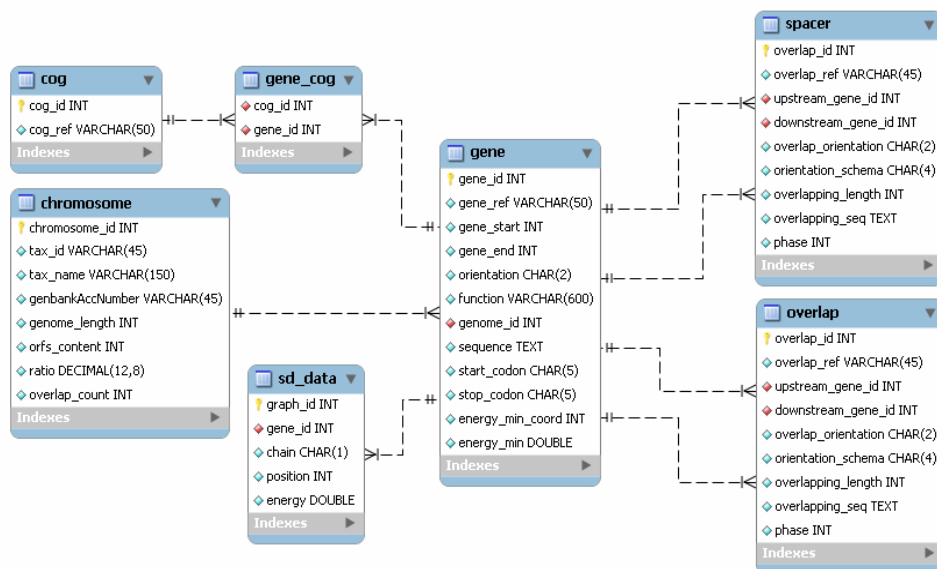


Figure 1. Entity-relationship model of the MySQL database.

Schema of the data model designed and translated to a MySQL database.

Utility

We have developed an interactive and intuitive database that currently contains 1,956,294 gene pairs from 678 fully sequenced microbial genomes. The database is freely available at the URL <http://genomes.urv.cat/pwneigh>.

Submitted to BMC Bioinformatics

Basically, this database provides information about the overlapping genes and the spacers between genes among the prokaryote genomes. Users access the information through three browsers that are described below. They can find information about the overlaps and the spacers with the species name or the GenBank Accession Number, with the gene id or with an internal id. While the users are typing the species name or any id the search engine helps to complete the name or the id. Interestingly by clicking on the database logo a tag cloud of the species contained in the database is obtained, which can be sorted by number of overlaps in a genome or by genome length. Furthermore, the database is able to provide the users with reports in CVS format at every step of their consultation.

The Genome browse

With this browser, users can find general information about the genomes and connect to the overlapping genes or the spacers between genes contained in the genome. They can access this information by typing the name of the species (by tax name) or the GenBank Accession Number (by genbank). If they do not remember the species name or the GenBank Accession Number by clicking on "Genome" the users can consult an exhaustive list of the species contained in this database and their GenBank Accession Numbers. Once the user has made a genome search, the first page obtained gives basic features of the genome including the Species name, the GenBank Accession Number, the TaxID, the genome length, the number of ORFs in the chromosome, the number of overlaps and spacers in the genome, the overlaps between ORFs ratio in the chromosome and the number of co-directional, convergent and divergent overlaps tabulated and represented graphically. By clicking the number of overlaps a list of the overlaps contained in the genome is displayed on a new page, while on clicking the number of spacers a list of the spacers contained in the genome is displayed on another new page.

The overlapping genes browse

Submitted to BMC Bioinformatics

The users can analyse the overlapping genes in a genome or a particular overlap of interest (by gene or by internal id). Once the user has made a genome search, the first page obtained has a list of the overlaps with the overlapping genes and their orientations as well as the distribution of the overlapping lengths represented graphically. The representation of the overlapping length distribution gives a general idea about the most common overlaps and the most common overlapping phases in the genome. Each overlap id leads to a detailed new page of the overlap including five labels that provide: overlap information, upstream gene information, upstream gene sequence, downstream gene information and downstream gene sequence. The overlap information label (*General Info* label) provides the internal id, chromosome name, the orientation, the overlapping phase, the overlapping length and the overlapping sequence. The upstream and downstream gene information labels (*Upstream Gene* and *Downstream Gene* label respectively) show the gene name, the gene function, the gene COG, the stop codon and the start codon. Also, on these labels is given information related to the SD location (position of the minimal ΔG° value and minimal ΔG° value) and the ΔG° values in the translation initiation region is represented graphically. The SD related information will be given in the upstream or in the downstream label depending on the gene pair orientation. The labels *Fasta Up* and *Fasta Down* contain the upstream and the downstream gene sequence in fasta format. Above the sequences there is a BLAST button. By clicking on it, the gene sequence is directly pasted in the BLAST local search engine and the conservation of one overlap across the species can be analysed. Interestingly, in the PairWise Neighbours database, the user can define the Expected threshold of the BLAST search engine among other features. Therefore the user can decide the threshold used to study the similarity among orthologous genes in order to analyse the overlapping pair conservation. In the BLAST results, by clicking on any hit, the information of the overlap is displayed on a new page.

The spacers browse

The users can analyse the spacers between adjacent genes in a genome or a particular spacer of interest (by gene or by internal id). If the user makes a genome search, a bar chart of the spacing lengths of the genome is shown and the user can have a first view of the most common spacers in the genome. Below a list of all the spacers in the genome is displayed, providing the internal id, the genes separated by the spacer and their orientation. By clicking any internal id all the information about the spacer is displayed on a new page. On this page there are three labels that give information about: the spacer, the upstream gene and the downstream gene. Basically the information given in the fields on a general information label (*General Info* label) is the same as the fields on a *General Info* label of an overlap. However, the user can find the Spacing length instead of the Overlapping length and Spacer sequence instead of Overlapping sequence. The information provided on the *Upstream* and *Downstream Gene* labels is the same as that on the overlap labels and the SD related information is also given depending on the gene pair orientation.

Discussion

In this Discussion section we give a few examples that we find interesting to illustrate the uses that can be attributed to the PairWise Neighbours database.

Conservation of gene overlaps

The first one is about the gene couple NC_000913.b0043 and NC_000913.b0044 of *E. coli* K12, which code for two proteins 4Fe-4S ferredoxin-type and have the COG ids COG0644C and COG2440C respectively (Figure 2). These genes are overlapping 4 bps. From the upstream and downstream sequence labels it is easy to study the conservation of the gene pair, using the BLAST button. The BLAST results show 24 genes with high similarity (E Value < $2e^{-7}$) to the NC_000913.b0043 gene and 33 genes with high similarity (E Value < $4e^{-5}$) to the NC_000913.b0044 gene (Figure 2). By

clicking on a gene id in the BLAST results, information about the overlap that involves the gene is displayed on a new page. Most of the genes similar to the NC_000913.b0043 gene have their adjacent gene in the group of similar genes to the NC_000913.b0044 gene and the majority of these gene pairs are overlapping 4 bps. Therefore it is a conserved overlap, particularly across the Enterobacteria species. Interestingly, we also find high conservation in the location of the SD sequence. Analysing the SD information for the NC_000913.b0044 gene (*Downstream Gene* label in Figure 2) we observe a drop in ΔG° value at 9 nucleotides to the start codon. This SD position is conserved among Enterobacteria species. Figure 2 shows the information for the NC_003197.STM0078 gene of *Salmonella typhimurium* LT2, which overlaps 4 bps with the NC_003197.STM0077 gene. These genes are similar to the *E. coli* K12 gene pair analysed. The NC_003197.STM0078 gene shows a drop in ΔG° value at 9 nucleotides to the start codon, as it happens in the NC_000913.b0044 gene of *E. coli* K12. This indicates that the SD sequence is located along the 3'-end of a previous coding sequence and it might suggest that the SD locations of conserved gene pairs can also be highly conserved.

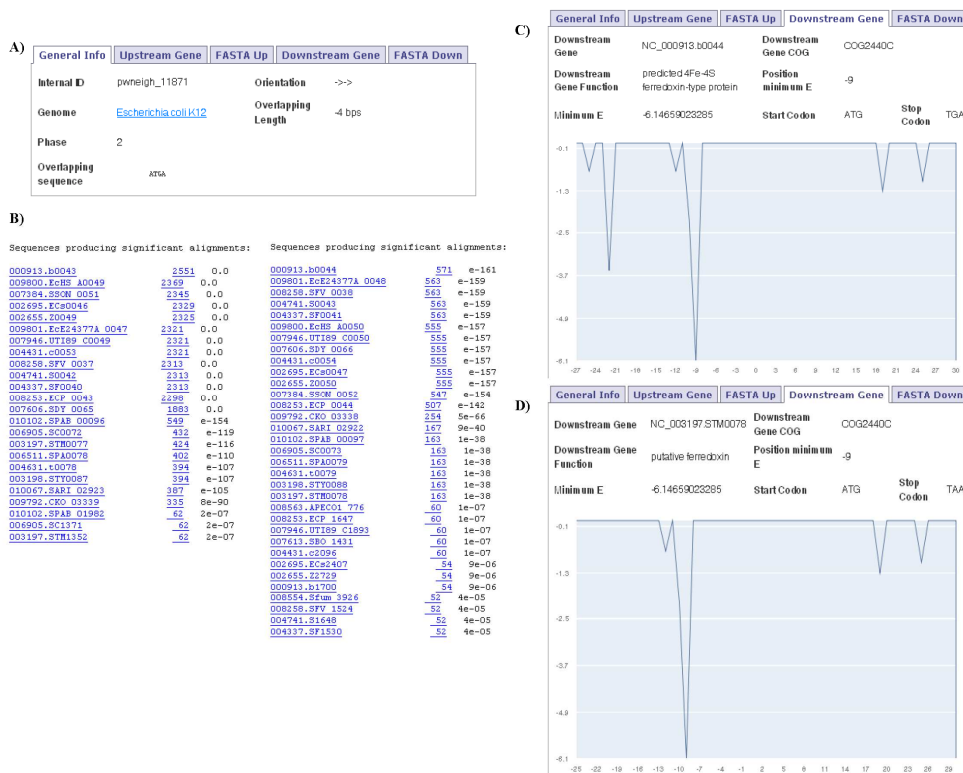


Figure 2. Study of a 4 bps overlap conservation

Compilation of images that the users can find when they are studying the conservation of an overlap. General Info label shows information about the 4 bps overlap between NC_000913.b0043 and NC_000913.b0044 genes (A). The BLAST results give an idea of the conservation of the overlap across the species (B). Information given on the NC_000913.b0044 *Downstream Gene* label provides gene details (gene function, gene COG, start and stop codon), SD related information (position of minimal ΔG^0 value, minimal ΔG^0 value) as well as a graph of the ΔG^0 values along translation initiation region.

The second example is about the gene couple NC_002947.PP_2780 and NC_002947.PP_2781 of *Pseudomonas putida* KT2440 that overlap 130 bps.

This overlap is the product of a misprediction of the start codon of the gene NC_002947.PP_2781 causing a 5'-end extension of the gene [34]. If we use the sequence of this gene as a query for the BLAST, we obtain as a first hit the orthologous gene NC_009512.Pput_2974, which is 127 bps shorter (compared with NC_002947.PP_2781) at the 5'-end and it is adjacent to the NC_009512.Pput_2975 gene (Figure 3). This gene pair (NC_009512.Pput_2974 and NC_009512.Pput_2975) belongs to *P. putida* F1 and overlaps only 4 bps, which is more reliable than the overlap of 130 bps. This is an example of a mispredicted overlap (NC_002947.PP_2780 and NC_002947.PP_2781) that could be corrected by just analyzing the BLAST results that we obtain automatically in this database. Furthermore, the SD prediction indicates that the NC_002947.PP_2781 gene has no SD sequence, while the NC_009512.Pput_2974 gene has the SD sequence at 7 nucleotides to the start codon of the gene (Figure 3). Therefore, the SD may help to expose the wrong start codon predictions [12].

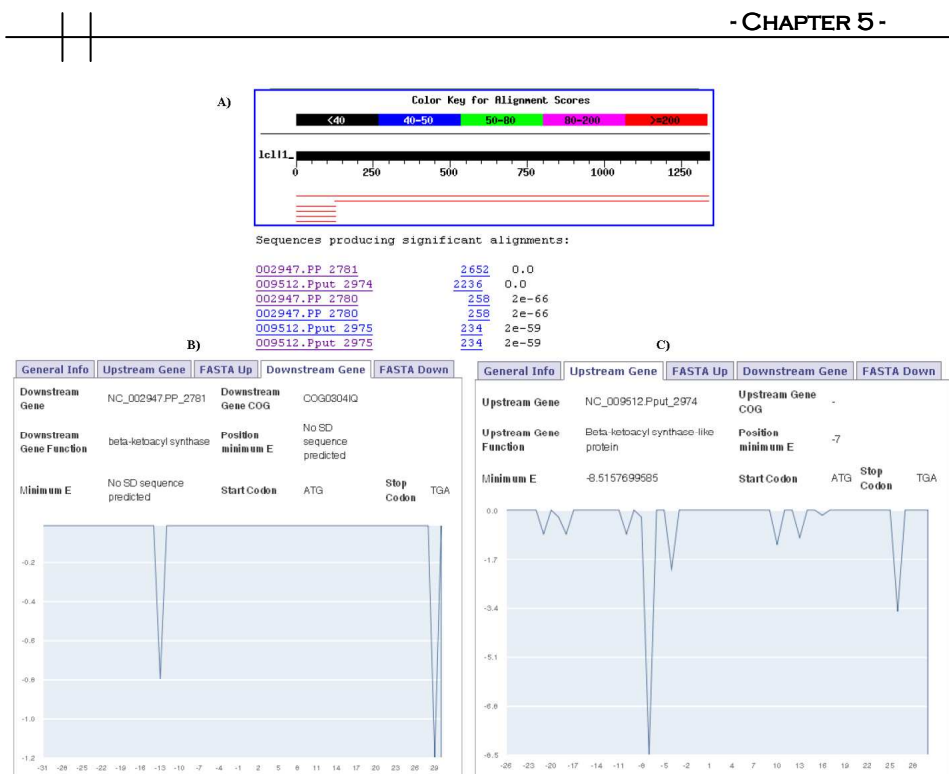


Figure 3. Study of a 130 bps incorrectly annotated overlap

The BLAST results show that the gene NC_002947.PP_2781 of *P. putida* KT2440 is longer than its orthologous gene NC_009512.Pput_2974 in *P. putida* F1 (A). This difference in length indicates that the 130 bps overlap between NC_002947.PP_2780 and NC_002947.PP_2781 is not conserved and thus not reliable. In the NC_002947.PP_2781 *Downstream Gene* label is shown that this gene has no SD sequence (B), while in the NC_009512.Pput_2974 upstream gene label is shown that this gene has the SD sequence at 7 nucleotides to the start codon (C).

Relationship between SD positions and the spacing lengths

The third example is about the genes NC_000913.b2644 and NC_000913.b4548 of *E. coli* K12. These genes are separated by 8 bps (Figure 4), which is a short intergenic distance for a co-directional gene pair. The NC_000913.b4548 label shows that there is a drop in ΔG° value at 6 bps to the

start codon (Figure 4). This means that the SD sequence of this gene is overlapping the upstream stop codon (TGA). If we join the upstream stop codon, the intergenic sequence and the downstream start codon we have the sequence TGAGGTATTACATG (Figure 4). The upstream stop codon is overlapping the SD motif resulting in the pattern TGAGGT that can bind with the SD sequence 3'-CCUCCA-5'. Therefore here we have detected a co-directional gene pair of *E. coli* K12 whose SD sequence for the downstream gene overlaps the upstream stop codon.



Figure 4. Study of the location of the SD sequence between a co-directional gene pair

Compilation of images that the users can find when they are studying the location of the SD sequence between the co-directional genes NC_000913.b2644 and NC_000913.b4548 separated by 8 bps. *General Info* label gives details about the spacer between this gene pair, which include the Spacing length and the Spacer sequence (A). The NC_000913.b2644 *Upstream Gene* label gives information about this gene (B), while the NC_000913.b4548 *Downstream Gene* label gives information about this gene as well

Submitted to BMC Bioinformatics

as SD related information and the corresponding graph of the ΔG° values along the translation initiation region (C).

SD presence among different gene sets

Other uses of the PairWise Neighbours database are to find out SD information of gene sets of interest, which have been labelled in other databases. For instance, the gene NC_000913.b3297 of *E. coli* K12 has been labelled as a *highly expressed gene* (HEG) in the HEG database [37]. This gene codifies for the 30S ribosomal protein S11 and has a strong SD sequence (the drop of ΔG° value is -11.44 Kcal/mol) at 10 nucleotides upstream to the start codon. If we analyse the SD presence in the *E. coli* K12 genes predicted as HEG in the HEG database (Table 1) [37], we find that the 81.03% of these genes have the SD sequence. This percentage is significantly higher compared with all the *E. coli* K12 genes (69.66%) and with the mean and standard deviation of the SD presence in 100 sets of 300 *E. coli* K12 genes randomly selected ($69.04\% \pm 2.58\%$) (Table 1). Therefore, as other authors have already found [9], the HEGs appear to have more SD presence. Another interesting gene set that can be analysed in this database is the *horizontally transferred genes* (HGTs). We studied the SD presence among the *E. coli* K12 genes predicted as HGTs in the HGT database [38]. The percentage of HGTs that have SD sequence (68.39%) is close to the percentage of SD presence found in all the *E. coli* K12 genes. This percentage falls within the range of the mean and the standard deviation of 100 sets of 300 genes randomly selected from *E. coli* K12 (Table 1). Therefore, it seems that the HGTs have an equal SD presence to the original genes of the species.

	Number of genes	Percentage of genes with SD	Percentage of genes without SD
--	-----------------	-----------------------------	--------------------------------

All <i>E. coli</i> genes	4,133	69.66	30.34
Highly expressed genes (HEG) from <i>E. coli</i> ⁽¹⁾	253	81.03	18.97
Horizontally transferred genes (HGT) from <i>E. coli</i> ⁽²⁾	310	68.39	31.61
Mean and standard deviation of 100 sets of 300 genes randomly selected from <i>E. coli</i>	300	69.04 ± 2.58	30.96 ± 2.58

Table 1. Genes with or without SD sequence in *E. coli* K12

Number of genes and percentage of genes with the Shine-Dalgarno motif from *E. coli* K12.

⁽¹⁾ HEG extracted from the HEG-DB (<http://genomes.urv.es/HEG-DB>) [37]

⁽²⁾ HGT extracted from the HGT-DB (<http://genomes.urv.es/HGT-DB>) [38]

Abbreviations: SD, Shine-Dalgarno

Conclusions

The studies of the translation initiation mechanism, gene regulation and gene structure (such operon predictions) rely on correct annotations. With the growing number of fully sequenced prokaryote genomes, the databases that help the annotation processes are very desirable. PairWise Neighbours is an interactive and intuitive database for retrieving information about the spacers and overlapping genes among bacterial and archaeal genomes. With this information, on the one hand, it is possible to study the reliability of an overlap as well as its conservation across the species with a BLAST local system, which permits the user to study the conservation of an overlap applying their desired Expect threshold. On the other hand, with the information related to the SD sequence and the ΔG^0 values along the translation initiation region, the users

can analyse the intergenic regions widely. They can check the reliability of the initiation site prediction, the SD location and the SD strength or the relationship between SD location and the spacing lengths.

Authors' Contributions

AP performed the necessary Perl Scripts to obtain the raw data. TR built the MySQL database and designed the web application. AP, SGV and AR participated in the analysis and interpretation of the data. AP drafted the manuscripts and SGV and AR revised it critically. Finally, all the authors read and approved the version to be published.

Acknowledgements

This work has also been supported by projects BIO02003-07672 and AGL2007-65678/ALI of the Spanish Ministry of Education and Science. Also we would like to thank Richard Tuby for his help in writing the manuscript. Finally, we would like to thank Joshua Starmer and co-workers for making available their programs for detecting Shine-Dalgarno motifs, and especially thanks to Joshua Starmer for his kind assistance.

References

1. Koonin E, Wolf Y: **Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world.** *Nucleic Acids Res* 2008.
2. Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, Wolf YI, Yin J, Koonin EV: **Congruent evolution of different classes of non-coding DNA in prokaryotic genomes.** *Nucleic Acids Res* 2002, **30**(19):4264-4271.
3. Mira A, Ochman H, Moran NA: **Deletional bias and the evolution of bacterial genomes.** *Trends Genet* 2001, **17**(10):589-596.

4. Johnson ZI, Chisholm SW: **Properties of overlapping genes are conserved across microbial genomes.** *Genome Res* 2004, **14**(11):2268-2272.
5. Shine J, Dalgarno L: **The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites.** *Proc Natl Acad Sci U S A* 1974, **71**(4):1342-1346.
6. Chen H, Bjercknes M, Kumar R, Jay E: **Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs.** *Nucleic Acids Res* 1994, **22**(23):4953-4957.
7. Ringquist S, Shinedling S, Barrick D, Green L, Binkley J, Stormo GD, Gold L: **Translation initiation in Escherichia coli: sequences within the ribosome-binding site.** *Mol Microbiol* 1992, **6**(9):1219-1229.
8. Kozak M: **Initiation of translation in prokaryotes and eukaryotes.** *Gene* 1999, **234**(2):187-208.
9. Ma J, Campbell A, Karlin S: **Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures.** *J Bacteriol* 2002, **184**(20):5733-5745.
10. Osada Y, Saito R, Tomita M: **Analysis of base-pairing potentials between 16S rRNA and 5' UTR for translation initiation in various prokaryotes.** *Bioinformatics*, **15**(7-8):578-581.
11. Schurr T, Nadir E, Margalit H: **Identification and characterization of E.coli ribosomal binding sites by free energy computation.** *Nucleic Acids Res* 1993, **21**(17):4019-4023.
12. Starmer J, Stomp A, Vouk M, Bitzer D: **Predicting Shine-Dalgarno sequence locations exposes genome annotation errors.** *PLoS Comput Biol* 2006, **2**(5):e57.
13. Eyre-Walker A: **The close proximity of Escherichia coli genes: consequences for stop codon and synonymous codon use.** *J Mol Evol* 1996, **42**(2):73-78.
14. Fukuda Y, Nakayama Y, Tomita M: **On dynamics of overlapping genes in bacterial genomes.** *Gene* 2003, **323**:181-187.

15. Fukuda Y, Washio T, Tomita M: **Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*.** *Nucleic Acids Res* 1999, **27**(8):1847-1853.
16. Sakharkar KR, Sakharkar MK, Verma C, Chow VT: **Comparative study of overlapping genes in bacteria, with special reference to *Rickettsia prowazekii* and *Rickettsia conorii*.** *Int J Syst Evol Microbiol* 2005, **55**(Pt 3):1205-1209.
17. Harrington ED, Singh AH, Doerks T, Letunic I, von Mering C, Jensen LJ, Raes J, Bork P: **Quantitative assessment of protein function prediction from metagenomics shotgun sequences.** *Proc Natl Acad Sci U S A* 2007, **104**(35):13913-13918.
18. Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, Tatusov RL, Koonin EV: **Purifying and directional selection in overlapping prokaryotic genes.** *Trends Genet* 2002, **18**(5):228-232.
19. Moreno-Hagelsieb G, Collado-Vides J: **A powerful non-homology method for the prediction of operons in prokaryotes.** *Bioinformatics* 2002, **18 Suppl 1**:S329-336.
20. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J: **Operons in *Escherichia coli*: genomic analyses and predictions.** *Proc Natl Acad Sci U S A* 2000, **97**(12):6652-6657.
21. Ermolaeva MD, White O, Salzberg SL: **Prediction of operons in microbial genomes.** *Nucleic Acids Res* 2001, **29**(5):1216-1221.
22. Cock PJ, Whitworth DE: **Evolution of gene overlaps: relative reading frame bias in prokaryotic two-component system genes.** *J Mol Evol* 2007, **64**(4):457-462.
23. Lillo F, Krakauer DC: **A statistical analysis of the three-fold evolution of genomic compression through frame overlaps in prokaryotes.** *Biol Direct* 2007, **2**:22.
24. McCarthy JE: **Post-transcriptional control in the polycistronic operon environment: studies of the *atp* operon of *Escherichia coli*.** *Mol Microbiol* 1990, **4**(8):1233-1240.
25. Normark S, Bergstrom S, Edlund T, Grundstrom T, Jaurin B, Lindberg FP, Olsson O: **Overlapping genes.** *Annu Rev Genet* 1983, **17**:499-525.
26. Krakauer DC: **Stability and evolution of overlapping genes.** *Evolution* 2000, **54**(3):731-739.

27. Sakharkar KR, Chow VT: **Strategies for genome reduction in microbial genomes.** *Genome Inform* 2005, **16**(2):69-75.
28. Chen SM, Takiff HE, Barber AM, Dubois GC, Bardwell JCA, Court DL: **Expression and characterization of RNase-III and Era proteins - products of the rnc operon of Escherichia coli.** *Journal of Biological Chemistry* 1990, **265**(5):2888-2895.
29. Inokuchi Y, Hirashima A, Sekine Y, Janosi L, Kaji A: **Role of ribosome recycling factor (RRF) in translational coupling.** *Embo Journal* 2000, **19**(14):3788-3798.
30. Luo Y, Fu C, Zhang D, Lin K: **Overlapping genes as rare genomic markers: the phylogeny of gamma-Proteobacteria as a case study.** *Trends Genet* 2006, **22**(11):593-596.
31. Luo Y, Fu C, Zhang D, Lin K: **BPhyOG: an interactive server for genome-wide inference of bacterial phylogenies based on overlapping genes.** *BMC Bioinformatics* 2007, **8**:266.
32. Natale DA, Galperin MY, Tatusov RL, Koonin EV: **Using the COG database to improve gene recognition in complete genomes.** *Genetica* 2000, **108**(1):9-17.
33. Brenner S: **Errors in genome annotation.** *Trends Genet* 1999, **15**(4):132-133.
34. Palleja A, Harrington ED, Bork P: **Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions?** *BMC Genomics* 2008, **9**:335.
35. Hu G, Zheng X, Yang Y, Ortet P, She Z, Zhu H: **ProTISA: a comprehensive resource for translation initiation site annotation in prokaryotic genomes.** *Nucleic Acids Res* 2008, **36**(Database issue):D114-119.
36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403-410.
37. Puigbò P, Romeu A, Garcia-Vallvé S: **HEG-DB: a database of predicted highly expressed genes in prokaryotic complete genomes under translational selection.** *Nucleic Acids Res* 2008, **36**(Database issue):D524-527.
38. Garcia-Vallve S, Guzman E, Montero M, Romeu A: **HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes.** *Nucleic Acids Res* 2003, **31**(1):187-189.



**1469646768248269 PairWise Neighbours Database: Overlaps
and Spacers among Prokaryote Genomes**

Article title: PairWise Neighbours Database: Overlaps and Spacers among
Prokaryote Genomes

MS ID : 1469646768248269

Authors : Albert Pallejà, Tomàs Reverter, Santiago Garcia-Vallvé and Antoni
Romeu

Journal : BMC Bioinformatics

Dear Mr Pallejà

Thank you for submitting your article. This acknowledgement and any queries
below are for the contact author. This e-mail has also been copied to each
author on the paper, as well as the person submitting. Please bear in mind that
all queries regarding the paper should be made through the contact author.

A pdf file has been generated from your submitted manuscript and figures. We
would be most grateful if you could check this file and let us know if any aspect
is missing or incorrect.

http://www.biomedcentral.com/imedia/1469646768248269_article.pdf (863K)

For your records, please find below link(s) to the correspondence you uploaded
with this submission. Please note there may be a short delay in creating this file.

http://www.biomedcentral.com/imedia/2000886935248275_comment.pdf

If the PDF does not contain the comments which you uploaded, please upload
the cover letter again, click "Continue" at the bottom of the page, and then
proceed with the manuscript submission again. If the letter will not upload,

Submitted to BMC Bioinformatics

please send a copy to editorial@biomedcentral.com.

We will assign peer reviewers as soon as possible, and will aim to contact you with an initial decision on the manuscript shortly. The submitting author can check on the status of your manuscript in peer review at any time by logging into 'My BioMed Central' (<http://www.biomedcentral.com/my>).

In the meantime, if you have any queries about the manuscript you may contact us on editorial@biomedcentral.com. We would also welcome feedback about the online submission process.

You will be able to change details or submit revised versions of your manuscript by going to:

http://www.biomedcentral.com/manuscript/login/man.asp?txt_nav=man&txt_man_id=1469646768248269

Regards

The BioMed Central Editorial Team

Tel: +44 (0)20 7631 9921

Facsimile: +44 (0)20 7631 9923

e-mail: editorial@biomedcentral.com

Web: <http://www.biomedcentral.com/>

Submitted to BMC Bioinformatics

ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGACAGCTGATAGG
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGGCGCTGGTCAGGCAGCGCAGCATGGGAAAA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
GATCAGATGACAGATAGATAACCACAGAGACATG
CGCGCGGACGCATGATTGATGATCAGATGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGATGATAAAGTGATTAGATAGATGGTGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTTT
GAGAGATAGAGAGATAGGATCGCGCTCGAGCGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AACCGCCAACCGGTGGCTTAGGATAGATGATGA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
AAAGTGTGTGTGACAGACAGTTGATGATAGTACA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGACAGCTGATAGC
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CGCGCGCTCGCTCGAGCGCTAGCTCGATCGATCGA
TGGCGCTCAAACGAGCGCTAGCTCGATCGATCC
TGGATAGGTACGCGAAAATGGCAGTAGCTA
TGATAGGTACGCGGATGAATGGCAGTAG
GATCGATCGATCGATCGATCGCGCGA
TCAGCATGACACACACACATGAT
TGCCAGGCAGCATAAAGCAGAC
TCAGCTGGGTGGTAGGACG
GATCGATCGATCGATCG
TGCACAGACAGTTGA



ACGCGAAAATGGCA
TCTCGAGCGCTAGCTCG,
TACCGCGAAAATGGCAGCT

CONCLUSIONS

TAGCATGACACACACACATGAT,
TGCCAGGCAGCATAAAGCAGAC
TCAGCAGCTGGGTGGTAGGAGTGATG
TAGTGCCAGGCAGCATAAAGCAGACGA
TACCAGCAGCTGGGTGGTAGGAGTGATGTA

From the different chapters of this thesis the following conclusions can be extracted:

Chapter 1:

- Although the DNA compositional asymmetry analysis generally brings us close to the origin and terminus sequences, it is not enough to provide us precise predictions. The DNA compositional asymmetry analysis must be applied together with other methods such as finding the DnaA protein and its binding sites in order to make better predictions.
- We can improve and reinforce our origin and terminus predictions with other complementary tools. These include making BLAST analysis of intergenic sequences, studying the gene order around the origin sequence, locating both the *ter* sites and the *dif* sites of the genome studied and analyzing the distribution of genes encoded in the leading or lagging strand.
- The origin prediction of *Bacteroides thetaiotaomicron* is located between 4035393 and 4035883 bps, of the chromosome, where there was approximately one change in compositional skew polarity and we found several DnaA boxes. This location is not where it was supposed to be according to the published sequence.
- By having a more accurate analysis of the nucleotide skew plots was enough to give a correct origin prediction for *Bacteroides thetaiotaomicron*. Especially if one looks at the direction of the GC skew as well as the distribution of the genes on the leading strand.

Chapter 2:

- Gene overlaps arise in all the three transcriptional orientations with extremely common and prohibited overlapping lengths resulting from the

structure of the genetic code and strong selective pressure against long overlaps.

- The preferred overlapping lengths are 1 and 4 bps among co-directional overlaps, 4 bps among convergent overlaps and 2 bps among divergent overlaps.
- Some of the overlapping patterns are extremely common, such as ATGA in co-directional overlaps, which includes the stop codon for the upstream gene and the start codon for the downstream gene. However, some of them are the result of wrong annotation, ribosomal frameshifting, or truncated genes.
- The codirectional overlaps have a prohibited overlapping phase (Phase 0). Therefore, the co-directional gene pairs cannot overlap 3 bps or a multiple of 3 bps.
- The co-directional, convergent and divergent overlaps have a phase bias due to the restrictions that the genetic code imposes, the different frequency of start and stop codons within the phases, and the selection.
- The overlapping genes have Shine-Dalgarno (SD) sequences in the same way as the non-overlapping ones do. Even a relevant percentage of overlapping genes have a strong SD sequence. This means that genes may overlap regardless of their expression level.
- A high proportion of the divergent overlapping genes have SD presence, even though these genes are the most constrained because of the location of the regulatory sequences within a coding region.
- The divergent overlaps may be conserved structures of coregulated genes where one is a transcriptional regulator which regulates the other overlapping gene.

Chapter 3:

- Co-directional and divergent overlaps extending 60 bps are artificial due to misannotations that can be classified into five categories.
- The most common misannotation is the 5'-end extension, mostly caused by the misprediction of start codons. The respective genes carrying putative mispredictions of the start codon show an overrepresentation of weak start codon usage.
- Convergent orientation seems to allow longer overlaps than the other two orientations, although convergent long overlaps are also affected by misannotations.
- Although several species seemed to have a higher number of such potential misannotations, no correlation was found with genome size, gene content, GC content, sequencing or ORF prediction method, annotation team or sequencing date. Therefore these imprecise gene predictions have the potential to affect any microbial genome annotation process.

Chapter 4:

- The differences in respect to the location of the SD sequence could contribute to explaining the variations in the ranges of the spacing lengths overrepresented and underrepresented in the prokaryote genomes.
- The genomes with a large number of genes with SD sequence seem to concentrate such a regulatory motif in a range from 4 to 12 bps before the start codon in the majority of the 530 prokaryote genomes analyzed.
- Although the translation in prokaryotes is mainly guided by the SD sequence that can bind the ribosome, it seems that there are only a slightly higher number of genes with SD than without SD sequence.

- Genes separated from 1 to 4 bps from a co-directional upstream gene show a high SD presence, although this regulatory signal is located towards the 3' end of the coding sequence of the upstream gene.
- Genes separated from 9 to 15 bps from a co-directional upstream gene show the highest SD presence as they can accommodate the SD sequence within the intergenic region.
- When the SD sequence overlaps with the upstream coding sequence or stop codon, its strength and relative distance to the downstream start codon do not vary significantly. However, the SD presence may make the intergenic lengths from 5 to 8 bps less favored and cause an adaptation of the stop codon usage.
- For co-directional genes separated by 7 or 8 bps, in *Escherichia coli* K12, the TGA stop codon is prevalent and part of the TGAGG pattern that acts as a SD motif. However, the stop codon usage adaptation and the SD motif form could be slightly different depending on the prokaryote species.

Chapter 5:

- PairWise Neighbors database is an interactive and intuitive database for retrieving information about the spacers between genes and overlapping genes among bacterial and archaeal genomes.
- It is possible to study the reliability of an overlap as well as its conservation across the species.
- With the spacer information given and the ΔG^0 values along the translation initiation region graphs, the users can analyze the intergenic regions and there is a wide scope for analysis especially,

- CONCLUSIONS -

the SD location, the SD strength and the reliability of the initiation site prediction.

- It is possible to analyze the relationship between the SD location and both the spacing length and the stop codon usage.
- The analysis of the reliability of the overlaps and the SD information given can help the annotation processes.

ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGACAGCTGATAGG
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGGCGCTGGTCAGGCAGCGCAGCATGGGAAAA
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
GATCAGATGACAGATAGATAACCACAGAGACATG
CGCGCGGACGCATGATTGATGATCAGATGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CCCCGATGATAAAGTGATTAGATAGATGGTGGGAT
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTTT
GAGAGATAGAGAGATAGGATCGCGCTCGAGCGA

LIST OF PUBLICATIONS AND CONGRESS CONTRIBUTIONS

ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
ACGCACGATGATAGAGATACAGACAGCTGATAGC
ATGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
CGCGCGCTCGCTCGAGCGCTAGCTCGATCGATCGA
TGGCGCTCAAACCTCGCTCGATCGATCG
TGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
TGGATAGGTACGCGAAAATGGCAGTAGCTAGCTT
GATCGATCGATCGATCGATCGATCGATCGATCGA
TCAGCATGACACACACACATGATGATGATGATGAT
TGCCAGGCAGCATAAAGCAGACACACACACATGAT
TCAGCTGGGTGGTAGGAGTGATGATGATGATGATGAT
TATCGATCGATCGATCGATCGATCGATCGATCGATCGA
TTGACAGACAGTTGATGATGATGATGATGATGATGATGAT



TACGCGAAAATGGCA
CGCTCGAGCGCTAGCTCG,
GGTACGCGAAAATGGCAGT,
TCGATCGATCGATCGATCGCGG
TCAGCATGACACACACATGAT,
TGCCAGGCAGCATAAAGCAGAC
CCAGCAGCTGGGTGGTAGGAGTGATG
CAGTGCCAGGCAGCATAAAGCAGACGA
TACCAGCAGCTGGGTGGTAGGAGTGATGTA

- LIST OF PUBLICATIONS AND CONGRESS CONTRIBUTIONS -

PUBLISHER PAPERS

- *In Silico Prediction of the Origin of Replication among Bacteria: A Case Study of Bacteroides thetaiotaomicron.* **OMICS: A Journal of Integrative Biology** 2008, 12 (3): 201-210.
- *Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions?* **BMC Genomics** 2008, 9: 335.
Highly accessed

SUBMITTED PAPERS

- *Overlapping Gene Structures among Prokaryotic Genomes.* Submitted to **GENE**.
- *Adaptation of the Short Co-directional Spacers to the Shine-Dalgarno Motif in Prokaryote Genomes.* Submitted to **PLOS ONE**.
- *PairWise Neighbors: Overlaps and Spacers among Microbial Genomes.* Submitted to **BMC Bioinformatics**.

CONGRESS CONTRIBUTIONS

- *Large gene overlaps among prokaryotic genomes...result of functional constraints or mispredictions?*

A.Pallejà, E. D. Harrington, S. Garcia-Vallvé, A. Romeo, P. Bork

Oral communication

Congress: Reunió conjunta de Genòmica i Proteòmica de la Xarxa valenciana i la Xarxa catalana i de la secció de la SCB. Peníscola 2008.

- *Study of gene overlapping patterns and spacers in prokaryotic genomes*

A. Pallejà, S. Garcia-Vallvé, A. Romeu

Poster presentation

- LIST OF PUBLICATIONS AND CONGRESS CONTRIBUTIONS -

Congress: VII Spanish Symposium on Bioinformatics and Computational Biology. Zaragoza 2006.

Publication: VII Spanish Symposium on Bioinformatics and Computational Biology book. PP-640

- *Genome organization of genes coding for carbohydrate active enzymes in Bacteroides species*

A. Pallejà, P. Puigbò, E. Guzmán, J.M. Orellana, M. Marcet, M.A. Montero, S. Garcia-Vallvé, A. Romeu

Poster presentation

Congress: 31st Meeting of the Federation of the European Biochemical Societies (FEBS). Istanbul 2006.

Publication: FEBS Journal Volume 273 PP-640

- *Genes solapados y regiones no codificantes en genomas bacterianos. Modelo de estudio: Bacteroides thetaiotaomicron*

A. Pallejà, E. Guzmán, P. Puigbó, J.M. Orellana, M.A. Montero, M. Marcet, S. Garcia-Vallvé, A. Romeu

Oral communication

Congress: II Reunión Científica de la red Nacional de Genómica Bacteriana. Santander 2005.

- *Arrangement of sequences in the gamma-proteobacteria replication origin*

E. Guzman, A. Pallejà, P. Puigbó, J.M. Orellana, M.A. Montero, S. Garcia-Vallvé, A. Romeu

Poster presentation

Congress: 30th Meeting of the Federation of the European Biochemical Societies (FEBS). Budapest 2005.

Publication: A3-034P

- *Análisis de las regiones no codificantes de Bacteroides thetaiotaomicron*

- LIST OF PUBLICATIONS AND CONGRESS CONTRIBUTIONS -

A. Pallejà E. Guzman, P. Puigbó, J.M. Orellana, M. Marcet, M.A. Montero, S. Garcia-Vallvé, A. Romeu

Poster presentation

Congress: XXVII Congreso de la Sociedad Española de Bioquímica y Biología Molecular (SEBBM). Zaragoza 2005.

Publication: T1.9-4

- *Codon usage in bacteria: correspondence análisis and correlation with genomic parameters*

E. Guzman, A. Pallejà, P. Puigbó, S. Garcia-Vallvé, A. Romeu

Poster presentation

Congress: 29th Meeting of the Federation of the European Biochemical Societies (FEBS). Varsòvia 2004.

Publication: Reference P1.3-10

- *Predicción de genes altamente expresados en genomas de procariontes*

P. Puigbó, E. Guzmán, A. Pallejà, M.A. Montero, A. Romeu, S. Garcia-Vallvé

Poster presentation

Congress: XXVII Congreso de la Sociedad Española de Bioquímica y Biología Molecular (SEBBM). Lleida 2004.

Publication: Reference P05-2

- *Translational selection en los genes metabólicos de genomas procariontes*

P. Puigbó, E. Guzmán, A. Pallejà, M.A. Montero, A. Romeu, S. Garcia-Vallvé

Congress: I Reunión Científica de la Red Nacional de Genómica Bacteriana. Granada 2004.

- *Theodosius Dobzhansky (1900-1975). El naixement de la teoria sintètica*

A. Rojas, M.A. Montero, E. Guzman, A. Pallejà, P. Puigbó, S. Garcia-Vallvé, A. Romeu

- LIST OF PUBLICATIONS AND CONGRESS CONTRIBUTIONS -

Congress: VIII Trobada d'Història de la Ciència i de la Tècnica. Societat Catalana d'Història de la Ciència i de la Tècnica. Institut d'Estudis Catalans. Palma de Mallorca 2004.

- *Caracterització de genomes procariotes*

S. Garcia-Vallvé, E. Guzmán, A. Pallejà, P. Puigbó, A. Romeu

Poster presentation

Congress: II Reunió anual de la Red Catalana de Bioinformàtica. Les Avellanes, Lleida 2003.

UNIVERSITAT ROVIRA I VIRGILI
COMPUTATIONAL INSIGHTS INTO INTERGENIC REGIONS AND OVERLAPPING GENES AMONG PROKARYOTE GENOMES
Albert Pallejà Caro
ISBN:978-84-692-2150-1/DL:T-508-2009

- AGRAÏMENTS / ACKNOWLEDGEMENTS -

La meva tesi no hagués estat possible sense la confiança dipositada en mi del meu director de tesi i professor de la Universitat Rovira i Virgili, el Doctor Antoni Romeu. L'Anton em va donar la possibilitat de començar a familiaritzar-me amb la recerca durant algun estiu mentre feia la carrera de Bioquímica i finalment, m'ha donat la possibilitat de fer cinc anys de recerca. La seva confiança sempre ha estat absoluta. Agrair el seu suport, la seva preocupació i la seva ciència. També agrair els consells i les discussions científiques al professor de la mateixa Universitat i també Doctor, Santi García-Vallvé. Sempre ha tingut un moment per un intercanvi d'impressions sobre qualsevol tema científic o no científic (castellers, el Nàstic, el Barça, els blogs, etc...). Així mateix agrair als professors del Departament de Bioquímica i Biotecnologia tot el que m'han ensenyat durant la carrera i els cursos de doctorat. Un s'ha fet menys ignorant gràcies a vosaltres. En especial, al Doctor Gerard Pujades qui sempre s'ha interessat en el esdevenir de la meva tesi i el meu futur. Gràcies també al Tomàs Reverter per els teus ànims, la teva bona disposició i la teva important assistència tècnica. Also I would like to thank Richard Tuby for his pleasant English classes and helpful advices, for your friendliness and your kindness.

I want to thank all the members' jury of my thesis: Lluís Arola, Gerard Pujades, Toni Gabaldon, Alex Mira and Lars J. Jensen for their acceptance to be part of this thesis, their cooperation and their kindness. Thanks also to the substitute members of the jury Francesc Xavier Avilés i Enric Querol and to the external reviewers Peer Bork and Jeroen Raes for their assessment.

Vull començar recordant a aquells amb els que vaig començar aquesta petita aventura ara fa cinc anys. Amb ells vam riure, ens vam organitzar, ens vam ajudar i vam conviure cada dia durant molt de temps, que això ja és molt. Mai van faltar almenys cinc minuts de dedicació al company, que això ja és més. Per això moltes gràcies a l'Eduard (l'avi) per les seves mil i una històries,

- AGRAÏMENTS / ACKNOWLEDGEMENTS -

experiències i trucs informàtics. Difícilment et pots avorrir amb l'avi al costat. A la Montserrat Vaqué per la teva simpatia, transparència i bona disposició barrejada amb aquest ordre i organització propi. Llàstima que de lo últim no se m'hagi encomanat gaire. Al Pere per la teva amistat, ajuda i el teu lloable interès en la ciència i els teus companys. Però alhora mira que ets tossut també eh, jo sé que a tu les mandarines t'encanten! Al Pep Orellana agrair-li les grans converses que hem fet, fem i farem i el seu especial sentit de l'humor que sota cap situació s'acaba. I dins aquest bloc, la darrera però no menys important, la Marina. Gràcies Marina per ser aquella persona que no fa mai soroll però sempre està allí per si la necessites. Una abraçada pels respectius del Pep i la Montserrat, la Xana i el Puxeu, amb els que he compartit també bones estones. A tots vosaltres gràcies i sapiguen que de tots alguna cosa he après i que se us troba a faltar més del que us penseu.

Paral·lelament a aquest grup de bioinformàtics, quan vaig començar ja hi havia una gran remesa de doctorands. Aquests però, eren dels que es mullen les mans, vaja els de poyata! Entre aquests m'agradaria agrair al Josep tots aquests anys d'amistat. Sincerament, has estat un amic amb qui parlar de moltes coses. Més d'un cop hem tancat algun local o hem acabat algun concert i, en la intensitat de la conversa, no ens en hem adonat. Bona senyal! Agraïments també per la teva dona, la Noemí, per llargues canyetes, sopars i xerrades a la Plaça de la Font. Lo riu és vida! Sort a tots dos en els nous projectes! Més agraïments pels doctorands que vaig trobar en aquella antiga facultat, on els coloms venien sovint a visitar les campanes d'extracció. Grans moments he passat amb tots ells, el Cesc, el Nino, l'Àngel, la Vanessa i la Pinent, ja sigui en un sopar o en una conversa al passadís. A part, amb els nois hem compartit una de les coses que més m'agrada fer, jugar a futbol. Quan jugàvem jo sempre els vaig insistir en fer un joc maco de molt de toc, el *Fiu-Fiu* com ho van anomenar. El Josep i jo els hi vam fer més d'una demostració pràctica. Agraïments també pels vostres respectius o respectives, Núria Setó (ja

- AGRAÏMENTS / ACKNOWLEDGEMENTS -

no te tinc al costat però molts ànims en lo teu!), Sonia i Raulín, Bianca (com se't troba a faltar per aquí guapa!), Jordi i Òscar.

Va passar el temps i els becaris antics vam passar a ser nosaltres. El temps no perdona i passa factura. La nova fornada de doctorands va portar personatges peculiars, per dir-ho d'alguna manera. Molts d'ells ja prometien quan els hi feia classe de pràctiques. Moltes gràcies a tots per la vostra companyia a la 'facu', als dinars al Soteres i als sopars fora de la 'facu'! Comencem pel grup que es mullen les mans. La Gemma, una persona amb moltes coses a dins per donar, com indica aquest genial somriure que se't dibuixa a la cara sovint. Et sentis còmode o no, ets un encant. El David Pajuelo, que 'eres la alegría de la huerta!', amb això ja t'ho dic tot i per la teva bona educació. L'Helena Cheesecake 'con ese salero' que li poses a tot i aquest riure encomanadís. La Sabina (Vainilla), amb les seves interessants històries sobre cuina i Capoira. I a la Ximena, la Isa i el Mario per més d'una xerradeta i cafè llarg al matí. Guaitem qui estem ara en el laboratori sec i fa temps que estem formant un segon grup ben avingut. El Gerard (àlies Colombo), per la teva predisposició genètica a la conversa sobre qualsevol tema, en especial sèries mítiques de TV3. Sort que no vau coincidir masses vegades tu i l'avi en el laboratori de bioinformàtica. Fora conyes s'agraeix la teva companyia. L'Esther va ser com una arribada d'aire fresc al laboratori. Ets divertida i molt sensible al teu voltant, cosa que s'agraeix, tot i que estiguis com un llumí, no deixis que aquest llumí s'apagui mai. Les teves són virtuts difícils de trobar. La Laura Guasch i el seu especial sentit de l'humor. No ets molt xerradora però quan la deixes anar tela marinera. M'has fet riure molt! Bona sort per tu Cristina que acabes de començar amb l'epigenètica. Una altra gent que corre per la facultat són el Jose 'Viste!' que no té ni idea de futbol però 'algo sabe de música!' I la Vanessa que tot i no comprar-li l'encenedor de la seva promoció hem acabat sent bons amics (o això crec...). Gràcies també pel reportatge fotogràfic de tots els episodis que hem viscut! I would like to thank also two german guys that came to Catalunya to do a short research stage, Ali and Sven. It was a pleasure

- AGRÀIMENTS / ACKNOWLEDGEMENTS -

to meet you and to spend with you some days of your stage. Ich hoffe Sie das Beste!

During my thesis period I was in Heidelberg for 5 months. I would like to thank Peer Bork and his entire group for giving me the chance of joining Bork Group at EMBL and for their kind cooperation. Specially, thanks to Eoghan Harrington for his kind attention and his cooperation. Good luck in your American adventure! Thanks also to Lars J. Jensen, Chris Creevey and Yang Ping Yuan for his kindness and technical support. Finally, thanks to Konrad and Mani for your friendship and kindness. I wish you the best in your future! At EMBL I shared the visitor's room with really friendly people such as Adriano, Michelle Chan, my Australian dude Sean O'Donegheu and with almost my brother in Heidelberg Philippe Julien (I have to admit that your wine is very good but you must try the catalan one!). I had really nice parties, talks and fun situations with Philippe, Sean, Oriol and Marc at EMBL and outside EMBL. Thank you very much for your friendship! Also thanks to Maria, Irene and Cristina for the amazing parties and nice dinners in Heidelberg. We used to go together to the Sonderbar library! There we met with Ton and Jordi. All of you were like my family in Heidelberg! Gràcies per la vostra companyia i amistat! Se us troba molt a faltar, em vau fer passar grans estones! Vielen Danke to Frieder Hansen to show me your typical Baden-Württemberg home cooks and culture, and for your interest in my culture. Ich hoffe zu bis bald!

On the other hand, I would like to remember in these acknowledgements three Croatian girls that I met in the FEBS Congress held on Istanbul. So, thanks to Matea, Goga and Ivana for those wonderful days in Istanbul. Jako sam željela ići u svoju zemlju. A zagrliti! Also thanks Goga for the last crazy days in Heidelberg, it was really fun and you were very helpful for me in those last days in Germany! I wish you the best in your future.

- AGRAÏMENTS / ACKNOWLEDGEMENTS -

Per mitjà de l'Esther, he conegut una gent molt maca com són el Joan Oriol (bé aquest peça ja li havia fet classe anteriorment), el Gerard, la Laura, el Joanet, la Marta, el Roger, l'Olivan, la Cris, la Berta, el Xavi i els bessons (Ori i Carles). Gràcies per la vostra companyia i amistat.

Tinc altres amics a Vilafranca que vull recordar perquè han estat molt presents amb mi al llarg dels últims anys. La meva 'germaneta' Sandra, per molt lluny que vagis o molt dies que faci que no et veig, sempre et tinc molt present! Aviam si fan més concerts dels que tu i jo sabem i et puc veure més! Gràcies a la Laura i la Judit per la vostra simpatia i empenta. A tu Dani per la teva senzillesa, la teva facilitat per escoltar i explicar coses i el Metal que portes a dins, se't troba a faltar ja ho saps! I al Balaguer per la seva amistat i amabilitat inacabable.

Vull també recordar als grans amics que tinc a Barcelona i que lamentablement no els puc veure tant com voldria. Gràcies per molts anys d'escola que he passat amb molts de vosaltres, sortides de nit, primers concerts, excursions, dinars i sopars, etc... Àlex i Natàlia, Carles, Naxo i Lourdes i 'los argentinos' Francesc i Vero!

També un agraïment molt especial a una gent que hem va acollir molt bé ara fa cinc anys i on tinc grans amics, el CE Imperial. Amb ells he pogut jugar a futbol i riure una estona després de treballar. Ja se sap *'Mens sana in corpore sano'* *Décimo Junio Juvenal, poeta romà del segle I.*

I en general, gràcies a tots aquells que en algun moment del dia us atureu, fugiu de les vostres cabòries i sou capaços de dedicar-li unes paraules a un amic o company. En els temps que corren queda molt poca gent així i és una llàstima.

- AGRAÏMENTS / ACKNOWLEDGEMENTS -

"A vegades, una tarda qualsevol, la dolçor s'instal·la en les paraules"
Miquel Martí i Pol, poeta català 1929-2003.

Hi ha hagut unes persones molt importants en la meua vida i pels quals va dedicat aquest llibre. Ells sense fer soroll sempre estan allí amb la mà estesa per lo que calgui. Gràcies a vosaltres pares per la vostra lleial confiança i per creure en mi i en tots els projectes que he començat. Sense el vostre amor, suport i sacrifici res d'això hagués estat possible. Us estimo i us admiro! I gràcies extensives als meus avis i a la resta de la família, que la tinc a Barcelona i no veig sovint, però van sempre dins meu. Gràcies també als pares de la Lídia, Manolí i Esteve i al seu germà culè, Gabriel, per moltes estones divertides, entranyables i pel seu suport.

Per últim però no menys important he deixat a la Lídia. No hi ha prou fulls en aquest llibre per agrair tot lo que t'haig d'agrair, lo molt que m'has ensenyat ni per expressar com em fas sentir. M'has donat el teu amor, la teua confiança, el teu suport i m'has omplert d'alegria durant tots aquests anys. Ets la darrera persona que veuen cada dia els meus ulls i la primera que busquen al despertar. Les teves paraules són sempre presents i es barregen amb els meus pensaments. M'agrada formar part de la teua vida i deixar sorprendre'm per la teua encisadora personalitat. Res d'això tindria sentit sense el teu somriure, la teua mirada o la teua paraula en el rerefons. T'estimo.

UNIVERSITAT ROVIRA I VIRGLI
COMPUTATIONAL INSIGHTS INTO INTERGENIC REGIONS AND OVERLAPPING GENES AMONG PROKARYOTE GENOMES
Albert Pallejà Caro
ISBN:978-84-692-2150-1/DL:T-508-2009