UNIVERSITÉ D'EVRY-VAL-D'ESSONNE

U.F.R DES SCIENCES

TNT: 2010EVRY0036

THÈSE

Pour obtenir le grade de

Docteur de l'Université d'Evry-VAL-d'ESSONNE

Discipline : Bioinformatique

Présentée par

Tingzhang WANG

Titre :

Make inferences about bacterial gene functions with the concept of neighborhood *in silico*

Directeur de thèse:

Professeur Antoine Danchin

Docteur Claudine Médigue

Soutenue le 15 Decembre 2010 devant le JURY composé de :

Mr Bernard PRUM	Président
Mme Frédérique LISACEK	Rapporteur
Mme Agnieszka SEKOWSKA	Rapporteur
Mr Stéphane CRUVEILLER	Examinateur
Mr Eric FOURMENTIN	Examinateur
Mr Antoine DANCHIN	Directeur de thèse
Mme Claudine MEDIGUE	Directeur de thèse

Acknowledgements

I am dedicating this thesis to China and France.

Writing a thesis was just the final, albeit long and painful, step of the process of earning my Bioinformatics Doctorate. I am very grateful to many people that helped me obtain my degree.

I wish to start giving my sincere thanks to my supervisor Prof. Antoine Danchin. I could not have finished this thesis without his help. I thank him for opening the door of France for me; for finding the financial support for me, even using his own money to support me during the last several months in 2009; for opening the door of science for me, especially directing me in several sub-fields in bioinformatics; for giving me so much freedom during my studies; for his patience and interest in both my results and writings, no matter negative or positive. Prof. Antoine Danchin is the greatest scientist and person that I have known. I do wish that I could always stay in his lab and learn from him, or at least communicate with him from time to time to obtain the greatest ideas.

I give my sincere thanks to my supervisor Prof. Claudine Médigue. I could not have finished this thesis without her help also. I thank her for helping me settle down in France which was particularly important for me since at that time I just knew several French words without any listening skills; for opening the door of team-work in science for me; and for her patience and interest in both my results and writings. Prof. Claudine Médigue is also a great scientist and kind person that I have known. I do wish that I could always stay in her team and learn from all the members there.

I give my sincere thanks to my previous supervisor Prof. Qingzhong XUE. I thank him for giving me the chance to study aboard; thank him for his constant interest in my research, especially evaluating my Chinese writings carefully and helping me persist in my Ph.D. studies. I do wish I could discuss with him frequently.

I give my sincere thanks to Prof. Huanming YANG for giving me the chance to study in Pasteur Institut as well as for his interest in my research.

I give my sincere thanks to Gang FANG for helping me during my studies. I do wish I could discuss with him more often.

I give my sincere thanks to Conghui YOU. I thank her for helping me settle down in France; for the interesting talking at tea-time; for her interest in my research and polishing on my writings. I do wish I could discuss with her from time to time.

Sincere thanks to Yanjiong CHEN. I thank her for helping me settle down in France; thank her for the interesting talks in tea-time; thank her for her concern in both my life and research. Sincere thanks to Jing WANG. I thank her for helping me studying in Database; thank her for her

concern on both my life and research. Sincere thanks to Dr. Undine MECHOLD. I thank her for her concern on both my research and life; thank her for the interesting talks in tea-time; thank her for the polishing on my writings. Sincere thanks to Grégory SALVIGNOL. I thank him for helping me settle down in France; thank him for fruitful discussion in my studies; thank him for his concern on both my research and life. Sincere thanks to Dr. David Vallenet. I thank him for his concern on my life; thank him for his helpful guidance in many bioinformatics techniques; Sincere thanks to Dr. Cruveiller Stéphane. I thank him for his helpful suggestions in multi-alignment and clustering techniques. Sincere thanks to Dr. Agnieszka SEKOWSKA. I thank her for her concern on both my research and life. Sincere thanks to Dr. David ROCHE, Dr. Calteau Alexandra, and also Dr. Cruveiller Stéphane. I thank them for our collaboration on the project of Genomic Islands. Sincere thanks to Dr. Ivan Moszer. I thank him for his helpful explanation in my studying of Correspondence Analysis. Sincere thanks to Dr Eduardo P. C. ROCHA. I thank him for his concern and helpful suggestions on my research. Sincere thanks to Rafik. I thank him for his discussion with me on the topic of Co-evolution. Sincere thanks to all the people in the unité of GGB and AGC, who compose the excellent and efficient research group and have given me help in different ways. Sincere thanks to all the people that ever helped me in both Institut Pasteur and GenoScope. Sincere thanks to Mita GUHA, Andrew MARTENS, Xin'ai ZHAO, Yingfeng LUO, Dongliang YU, Rongdong YU, Guohui, DING. I thank them for careful checking on my writings.

Besides neighbors who are important for me inside of the field of science, I would like to thank my neighbors outside of the field of science who helped me keep progress on the way to discovery.

I wish to start by giving my sincere thanks to my parents. I wish to thank them for giving me the freedom to envision new horizons and the encouragement to pursue my dreams. Without their initial momentum and their continuous support, it would have been impossible for me to get where I stand. I want to mention also my sisters Yuzhang and Yuqin who were always there for me. Although we were far apart, they always believed in me and they were always ready to help me in any way possible and they helped me to take care of my parents.

I would like to show my deepest gratitude to my dear wife Yinhua XIA. She joined my life in those dark days during my studies, and has become my most important support on the way to unknowns, my guiding star in the dark nights, gives me the strength to keep going.

I give my sincere thanks to the Fondation Fourmentin-Guilbert. I thank her and its creator Jean Fourmentin-Guilbert for giving me financial support during my studying in France.

I give my sincere thanks to Dr. Mita GUHA. I thank her for providing me an English class in Evry University and helping me find several international English classmates.

I give my sincere thanks to Secretary Annie Beaudeux, Florence HAMON, Anne-Sophie PEYSSON and Chantal SYLVAIN. I thank them for helping me during my studies in France.

I give my sincere thanks to Secretary Aihua CHEN. I thank her for helping me during my studying in university.

I give my sincere thanks to Ping XIE and Peng DENG. I thank them for supporting me during my studying in Middle school.

Sincere thanks to all the friends I met. I thank them for their involvements in my life or research in different ways.

I will finish this acknowledgement here. Then I can continue on the way to an unknown future with the concept of neighborhood in mind.

Abstract

With more and more genomes being sequenced, the organization of those raw data and the derived data, and the extraction of information and knowledge from these data has become a challenge. A key concept in this field is that of the neighborhood, especially with respect to the organization of data in relational databases. To extract information from bulk data, different kinds of neighborhoods were studied and each show interesting results in current study. Firstly, through the Correspondence Analysis (CA) and later Model Based Clustering (MBC), two kinds of neighbors i.e. the genes (proteins) and amino acids were analyzed respectively, and it was found that proteins from *Psychromonas ingrahamii* are clustered into six classes, and there is strong opposition between asparagine (N) and the oxygen-sensitive amino acids. Secondly, the relationship between genomic islands and core genome (i.e. two closely linked neighbors with large range on the chromosome) was studied by a new method combining composition, GI features and synteny break. On applying to *E. coli* and *B. subtilis* it was revealed that this new method can extract some meaningful regions not published before. Thirdly, the relationship between upstream and coding regions of thrS gene (i.e. a case for two closely linked neighbors with small range on the chromosome) was studied extensively. It was found that these two regions associated to one gene, behaved differently in the evolutionary history. Some of the upstream regions bearing non-essential function (i.e. regulation of gene expression) evolved more slowly than the coding region.

Keywords: neighborhood, Correspondence Analysis (CA), Model Based Clustering (MBC), Genomic Islands (GI), regulation of gene expression, threonyl-tRNA synthetase (thrRS)

Contents

Acknowledgements i
Abstract iv
Contents v
List of Figures vii
List of Tables viii
List of Acronyms ix
1. INTRODUCTION 1
1.0 Opening words 2
1.1 Role of bioinformatics 2
1.2 Concept for Data Organization 6
1.3 Three ways to make inferences 10
1.3.1 Abduction 11
1.3.2 Deduction 12
1.3.3 Induction 13
1.3.4 Relationship between three kinds of reasoning 14
1.4 Analysis neighbors to make inference and gain new knowledge 15
1.4.1 Neighborhoods defined along the chromosome 17
1.4.2 Neighbors defined in the perspective of evolution 18
1.4.3 Neighborhoods defined in terms of metabolic pathways 20
1.4.4 Neighborhoods defined in statistical spaces 23
1.5 Questions to be answered in this thesis 24
2. RESULTS 25
2.1 Correspondence Analysis (CA) and Model Based Clustering (MBC) 26
2.1.1 Literature overview 26
2.1.1.1 Concept and Application of Correspondence Analysis 26
2.1.2 Applying CA and MBC in the study of genome: *Psychromonas ingrahamii* 29
2.1.2.1 Study strategy and result summary 29
Article 1 30
2.2 Genomic Islands 31
2.2.1 Literature Review 31
2.2.1.1 Characteristics of Genomic Islands (GIs) 31
2.2.1.2 Function and classification of the Genomic Island 32
2.2.1.3 Prediction of Genomic Island 34
2.2.2 A combined approach for identification of genomic islands in prokaryotic sequences 39
2.2.2.1 Study Strategy and Results Summary 39
Poster 1 40
Article 2 41
2.3 The upstream and coding region of thrS gene 42
2.3.1 Literature overview 42

2.3.1.1 Regulation of expression for Aminoacyl-tRNA Synthetase 42
2.3.1.2 Regulation of aaRS expression in Bacillus subtilis 44
2.3.1.3 Transcriptional regulation of aaRS in *E. coli* 46
2.3.1.4 Translational regulation of theronyl-tRNA synthetase (thrRS) in E. coli 49
2.3.2 the study of upstream and coding regions of thrS gene 52
2.3.2.1 Study Strategy and Results Summary 52
Article 3 54
3. DISCUSSION 55
3.1 Neighbors defined in the statistical space, in the case of CA clouds 56
3.2 Neighbors defined along the chromosome, in the case of genomic islands 58
3.3 Neighbors defined along the chromosome, in the case of upstream and coding region of thrS 58
4. CONCLUSION 61
5. PROSPECTS 63
6. APPENDIX 65
6.1 The implementation of the computation of CA 66
6.2 Formula about Model Based Clustering (MBC) 73
7. REFERENCES 79
Résumé 1

List of Figures

Figure 1.1 Gene neighbors on the chromosome.....21
Figure 1.2 Gene neighbors in the statistical space 22
Figure 2.1 Study strategy for Correspondence Analysis in protein sequences.....29
Figure 2.2 Schematic diagram of Genomic Island/Islet structure 31
Figure 2.3 Classification of genomic islands according to the variety of function they contributed to the hosts. 33
Figure 2.4 The study strategy for genomic islands prediction 39
Figure 2.5 Central dogma and Protein biosynthesis.....43
Figure 2.6 Rho-independent transcription termination 45
Figure 2.7 The termination and antitermination formation of thrS leader region 45
Figure 2.8 Alternative structures of the leader regions of PheS-PheT operon. 48
Figure 2.9 Structure of thrS leader region and $tRNA^{Thr}$.....51
Figure 2.10 Study strategy for the study of upstream and coding regions of thrS gene.....52
Figure 3.1 Annotations in 5 genomes of *Salmonella* strains 59

List of Tables

Table 1.1 **Definitions of data, information and knowledge** 4
Table 6.1 The primitive matrix of Correspondence analysis 66
Table 6.2 the matrix of row profile 67
Table 6.3 the matrix of column profile 67
Table 6.4 Geometric characteristics of 10 covariance matrix 75

List of Acronyms

DIK	Data-Information-Knowledge
INSDC	International Nucleotide Sequence Database Collaboration
DDMS	DataBase Management System
GI	Genomic Islands
CA	Correspondence Analysis
MBC	Model Based Clustering
COG	Clusters of Orthologous Groups
PCA	Principal Component Analysis
ThrRS	Threonyl-tRNA Synthetase
AF	the Absolute codon Frequency
RF	the Relative codon Frequency
RSCU	the Relative Synonymous Codon Usage
WCA	Within-group CA
DR	Direct Repeat
IS	Insertion Sequence
PAI	PAthogeneicity Island
UPEC	Urinary tract Pathogenic *E. Coli*
SI	Symbiosis Island
FI	Fitness Island
MI	Metabolic Island
RI	Resistance Island

HPI	High Pathogenicity Island
IVOM	Interpolated Variable Order Motif
PFGE	Pulsed-Field Gel Electrophoresis
CGI	Comparative Genomic Indexing
aaRS	aminoacyl-tRNA synthetases
GRDC	Growth Rate-Dependent Control
ACSL	Anti-Codon Stem Loop
BBH	Bidirectional Best Hit
HGT	Horizontal Gene Transfer
SVD	Singular Value Decomposition
EM	Expectation-Maximization
BIC	Bayesian Information Criterion

1.INTRODUCTION

1.0 Opening words

Franklin, a scientist in the field of cytogenetics, said in 1993: "twenty-five years ago most biologists were able to follow their fields by regularly scanning a core of specific journals; in my own field of cellular genetics I was then quite confident that I missed ‘nothing’ by reading some 10 or so titles. However, change came fast, and some 4 years into ‘research’, a fairly slavish addiction to ISI's ‘Current Contents’ means that I had become more nervous about ‘missing’ that vital publication" [1].

1.1 Role of bioinformatics

This anecdote tells us that the information in the field of biology is accumulating at an uncontrollable pace. Meanwhile, we remain unable to give a precise definition to the concept of information. Based on different contexts, there is a wealth of potential meanings for information. Simply speaking, the concept of information includes at least the following aspects:

1) the most direct meaning for information is associated to that of a message, communicated between agents. For example, we can obtain news by reading the newspaper and/or watching TV programs.

2) Another meaning for information can be seen as related to sensory inputs to an organism [2]. For example, while light input is metabolically necessary to plants, it is not necessary as such to animals, yet it can still be used as a kind of information by animals. For example, the colored light reflected from flowers is too weak to do photosynthetic work, but it can be detected by the visual system of bees, propagated through the nervous system of bees, and can guide the bees to the flowers, where the bees can find nectar or pollen [3].

3) In perceptual and cognitive space, information, which is an important node in the popular hierarchy model of “Data-Information-Knowledge (DIK)” for classifying the human understanding, is the result of data processing (Table 1.1) [4]. Despite the lack of an agreeable set of definitions of data, information and knowledge, there is a general consensus that data is not information, and information is not knowledge. The relationship among data, information and knowledge can be illustrated through a simple example. Mendel's experiment on plant hybridization for the seed shapes collected the following raw data: 253 hybridization experiments, 7324 seeds, 5474 seed round, and 1850 wrinkled seed. This data sample records the number of hybridization experiments, the number of seeds, and seeds shapes respectively. Nothing is obtained if one just looks at the raw data separately (“data gazing”). However, after the appropriate treatment, Mendel wrote sentences that placed the data in context: "From 253 hybrids, 7, 324 seeds were obtained in the second trial year. Among them there were 5, 474 round or roundish seeds, and 1,850 angular wrinkled ones" [5] . After using some basic statistical techniques, Mendel obtained further information: “Therefore the ratio 2.96:1 is deduced” [5]. The ratio between round (including roundish) and wrinkled seeds is close to 3:1. Mendel also got similar results in his experiments, noting another six characteristics (the color of the seed albumen (endosperm), the color of seed-coat, the form of the ripe pods, the color of the unripe pods, the position of the flowers, and the length of the stem) [5]. This piece of information was eventually interpreted in the context of the science of heredity, and has been treated as one of most basic laws of genetics, Mendel’s law (or Mendelian inheritance or Mendelian genetics). According to this general law, we can predict the distribution of phenotypes of the offspring from ancestors with a pair of differentiating characters. In detail, the first generation (or F_1 generation)

from parents P1 (with dominant characteristic gene denoted as AA) and P2 (with the recessive one denoted as aa) show only the dominant trait as its parent P1; and after among the individuals of F_2 generation from the self-fertilized the F_1 generation, some show traits like P1, and the others show traits like P2, and the ratio between the two traits is close to 3:1.

4) In computer science, information is a term used extensively and often interrelated to data, e.g., information processing and data processing; information management and data management. From a system perspective, data is referred to as bits and bytes stored on or communicated via a digital medium. Thus, any computerized representations, including information and also knowledge representations, are types of data. Similar to the above, in computer science, although there is no agreeable boundary between these three terms, data is not information and information is not knowledge. Table 1. 1 summarizes the definitions on these three terms from the perspective of computer science [6].

Table 1.1 **Definitions of data, information and knowledge**

Category	Definition in perceptual and cognitive space	Definition in computational space
data	Symbols	computerized representations of models and attributes of real or simulated entities
information	the results of data processing, providing answers to 'who', 'what', 'where' and 'when' questions	the results of a computational process, such as statistical analysis, for assigning meanings to the data, or the transcripts of some meanings assigned by human beings
knowledge	application of data and information, providing answers to 'how' questions	the results of a computer-simulated cognitive process, such as perception, learning, association, and reasoning, or the transcripts of some knowledge acquired by human beings

5) Information is further developed within the theory of communication (transmission of signals through noisy channels) and in general through a variety of theories, usually collected

under the name “Theory of information” (see Cover and Thomas, 1991 Cover T, Thomas J (1991) Elements of information theory. Wiley, New York).

In recent decades, there has been a significant increase of the amount of available data in biology, and bioinformatics emerged to deal with the situation concerning both computers and biology. Particularly after genome sequencing technology was invented, genome analysis became an important branch of computer sciences, distinct from the branches developed before, and was named bioinformatics (after “informatique”, the French word for computer sciences) or biology *in silico*, to refer to the standard component of computers [7]. Previously, bioinformatics consisted of simulating biological physiological models, analyzing medicine images, or determining the structure of biological objects through X-ray diffraction pattern (or map of Nuclear Magnetic Resonance (NMR) recently). Now, besides the experiments *in vivo*, genome analysis, the novel technique for analyzing of DNA sequences *in silico*, can reveal many facts about genomic objects, such as the description of collective signals, the essential traits of a gene or protein and phylogenetic relationships.

If there had been no development of information technology, the available genomic text, obtained from the mass data produced by genome sequencing, would remain tough to interpret. The development of information technology itself is reflected in both software and hardware. In terms of hardware, computers have kept becoming more and more powerful in terms of computing speed and storage capacity. In terms of software developments, algorithms have been ever-increasing in depth, and their efficiencies have kept pace with the growing capacity of the hardware that runs them. Information technology is involved at several distinct levels in whole genome sequencing programs, to different degrees and they form clearly identifiable domains

linked to each other. The first one is related to data acquisition, by locating the radioactive or fluorescent bands on the sequencing gels or spots on slides, and by overlapping the sequences of fragments appropriately to obtain contigs or a complete genome. The second one is related to data analysis, by exploiting the data. For example via exploration of the biological meaning of the sequences, new genes could be found. Informatics is also involved in data management and storage. A genome sequence is usually a few millions or even tens of millions of bytes; the task to store them has become very easy in the last decade, but was almost unrealistic in the late 80s. At present, this task is mainly done through the database and the Internet technologies, both of which make the task convenient for users.

To summarize, one role of bioinformatics is to extract valuable information from the bulk of data, and further obtain knowledge. Now that we are in the data-driven era, we are faced with massive amounts of data. We should follow certain rules to organize our data, to do some data mining and explore the information, and later to obtain knowledge through making meaningful inferences, which in turn can guide biological experiments more effectively. During my Ph.D. work, I have been exploring different sub-projects involving the whole process about “data-information-knowledge” around the concept of "neighborhood".

1.2 Concept for Data Organization

The term database first appeared in June 1963, when the System Development Corporation launched a conference entitled "Development and Management of a Computer-centered Data Base". Later, database was used as an integrated word in Europe in 1970s. With a similar

meaning, "data bank" was first adopted in the Washington Post in 1966. Although both terms appeared at the same time, strictly speaking, there is a big difference between the two terms. Knowledge from any field could be put together in the form of reviews, which is often in a file system, including cards, pictures (such as charts), or a voice on a tape. These records have their intrinsic values, like bank records, and for this reason, the place with the corresponding collection of information is called data bank. If it is compared with the text in a book, it can be also called data libraries. A person who has experienced in using a library understands that the classification system is the key to a library. If people want to find the target books quickly in a large library, it would be great help to have some kind of catalog to organize similar books together. Similarly for a data bank, a most essential way to speed up the search is indexing records in the database with keywords. To this end, according to a data schema, the data are structured, and the related information is linked to each other. When knowledge is organized in a proper way, valuable information is actually added to the data bank. A data bank constructed in this manner is called a database.

Often, when information comes from a wide range of sources in the field, it should be further divided into a series of specific databases. For example: the international nucleotide sequence database collaboration (INSDC [8], strictly speaking, this is a data library, organized into different databases) collects all kinds of nucleotide sequences from different organisms, while there are also many databases interested in a certain genome or class of genes, and the number of this kind of databases has been increasing continuously. Also, through special fields in the database, the database itself also contains links between different entries, such as: "Sequence name", "identification number", "sequence", “length”, and “gene name”. The database designer

invents data schema through creating the appropriate fields and useful links between them, so users can carry out different queries based on their own criteria. Query is realized through a system called “database management system”, which usually manages the links (or relationships), and taking this into consideration, the relevant database is also known as a relational database.

Bovine insulin, which contains 51 amino acid residues, was the first protein to be completely sequenced in late 50s [9]. Subsequently, Dayhoff collected all the data of protein sequences into a database that was a book "Atlas of protein sequence and structure" published in 1965 [10]. At that time, the amount of data was not so large, and databases were often constructed and maintained by the corresponding laboratories in the form of documents. But with the accumulation of genomic sequence data and the following increase of protein sequences, protein 3D structures, the number of biological databases literally exploded. Among them, the most commonly used databases are several primary ones. The first is nucleotide international sequence database collaboration (INSDC), which has three access points: Genbank (http://www.ncbi.nih.gov/Genbank/) in United States, EMBL in Europe and DDBJ in Japan[11]. The second is the protein sequence database UniProt [12], which contains two parts: one is SwissProt manually annotated, and the other is TrEMBL formed through automatic translation of nucleotide sequence of the INSDC. The third is protein structure database PDB (Database Protein Data Bank)[13]. And the fourth is literature database Medline[14]. These databases can be accessed by end-users with a friendly web interface, of which Medline is accessed through the PubMed interface (http://www.ncbi.nlm.nih.gov/pubmed).

In addition to the general database, there are a large number of specific biological databases, where a comprehensive knowledge of specific organisms can be found. These databases extract

and combine information from ongoing or completed genome sequencing projects. Especially the database for model organism genome was already started much earlier, for example, in 1989, Kroger and colleagues collected *Escherichia coli* sequence from the EBI/Genbank [15]. Médigue recommended the establishment of a relational database for *E. coli* genome (Colibri), which at the beginning adopted DataBase Management System (DDMS) 4th Dimension to manage different tables, later the Sybase DBMS, and now the MySql database management system [16]. Médigue's work has also been extended to other genomes, for example *Bacillus subtilis* [17]. With a similar data scheme, an edition of Médigue's work (Genocore) published in 2005 integrated 17 genomes that were created during the beginning of the years 2000s and functionning well [18].

In addition to the general databases specific to a certain genome, there are several databases focusing on specific aspects of a model organism. For example RegulonDB focuses on *E. coli* transcriptional regulation [19]. EcoCyc (http://ecocyc.org/) started from the literature, and described regulation of transcription, transport of proteins and metabolic networks in *E. coli* [20]. Other good examples include PUMA, WIT created by Selkov and colleagues [21, 22], and KEGG created by Kanehisa and colleagues [23].

Derived information obtained from primary database constitutes a secondary database, which contains conserved sequences, sequence tags and active sites of a protein family determined by multiple sequence alignments. A special structure extracted from the PDB database could also form a secondary structure database. Structures could be classified according to whether alpha and/or beta structures or the conserved secondary structure motifs are contained. Examples of these databases are: SCOP maintained by the University of Cambridge[24], CATH

constructed by London School[25], PROSITE constructed by the Swiss Institute of bioinformatics[26] and eMOTIF constructed by Stanford[27].

In some cases, when a single database cannot answer a complex question, it is necessary to integrate information from several databases to provide an effective solution. To meet with this kind of need, there is also a kind of database represented by SRS [28], Entrez [29], and also Ensembl [30], which combine many different database resources and allow the user to simultaneously search for different resources. In particular, Entrez can access more than 20 interconnected databases [29].

The main function of above-mentioned databases is to query information stored there, but genome annotation is a mammoth task, which often requires collaboration from scientists around the world. During the collaboration, there should be frequent exchanges of information. By means of the network, this kind of communication has become much easier. For example Vallenet et al built MaGe systems, which combines the function of database integration and the function of data annotation [31].

1.3 Three ways to make inferences

At this point, we have a general idea of the way biology data could be organized and stored with the concept of “Neighborhood” in the background, implemented via a series of Relational Databases. However, this is not the end for a bioinformatics researcher. His role is to extract useful information from the data stored in the databases and obtain some knowledge that could be used as guidance to direct experiments in laboratory. Similar to other scientific fields, this

process is combined through the following three types of reasoning: Abduction, Deduction and Induction.

1.3.1 Abduction

Abduction, as Peirce put forward, can be seen as present behind exploratory data analysis: to find a pattern from the observed phenomenon and to propose a hypothesis trying to explain how the phenomenon emerged [32]. Abduction is a pre-requisite to construct an axiomatic theory, while induction (see below) is the process that proposes novel hypotheses. Although this kind of reasoning has a long history, it was not so popular in the research of “Logic and Methodology” that focuses more on formal logic statements. In terms of reasoning, two kinds can be proposed: one can be recorded with normative symbolic expression, while the other, represented by abduction, is just a kind of key thinking. It is difficult to find proper symbols to represent Abduction, therefor the following simple symbols are used to represent Abduction: 1) Observe an interesting phenomenon X; 2) In a series of assumptions, such as A, B and C, A is the most likely explanation of the observed phenomenon X; 3), then to continue with the exploration of the assumption A is much more reasonable. In the field of biology, there are many examples of successful abduction, such as: Mendel's experiments on pea hybridization, DNA double helix and deciphering the genetic codon. In 1950s, the understanding of DNA structure was still in the hypothetical stage, mainly represented by Pauling's triple helix structure [33, 34], but a reasonable model could not be established for this structure. Benefitting from DNA X-ray diffraction photos obtained by Wilkins and Franklin, Watson and Crick proposed a model of double helix, which is formed by the intertwining of two chains linked by hydrogen bonding and

with phosphate as the backbone[35]. Now this double helix model has become well known today. For abductive reasoning on the double helix model, one can use the simple symbols described as follows: 1) Watson and Crick observed interesting DNA X-ray diffraction photographs; 2) among the hypothesis of third helix, double helix and other models, the double helix model was the best to explain this picture; 3) to continue with the double helix model in research is much more reasonable and this forms the first axiom of the “Central Dogma” of molecular biology, which is talen into account replication of the genetic program. And indeed, it has been proved that the finding of double helix structure opened the era of molecular biology, brought the study about biological macromolecules into a new level, permitted genetic research be achieved at the molecular level, and let people make clear what genetic information and its transferring pathway is.

1.3.2 Deduction

Deduction is a way of thinking during which the conclusions for a particular instance are derived from general concepts, principles, and more specifically axioms. Universal principles are common properties about a certain kind of objects or knowledge of some necessities. Once this kind of knowledge has been obtained, one can extend it to any individual belonging to the same kind, and make conclusions for each individual. Let us review Mendel's experiments on pea hybridization, according to Mendelian genetics, the crossing between a pair of relative traits (“A” as a dominant trait, and “a” as a recessive trait) leads to a child generation (F_1 generation) with heterozygous genotype (Aa) but dominant phenotype; subsequently after the self-fertilized the F_1 generation, individuals from the second generation (F_2 generation) show both dominant

traits and recessive traits with a ratio close to 3:1. These were confirmed by seven kinds of comparative traits selected by Mendel [4], and later have also been confirmed in many experiments, especially studies from Hugo de Vries [36] and Carl Correns [37]. Once we mastered Mendelian genetics, we could predict the results of an experiment with two or more independent traits, and also the phenotypes distribution of the n^{th} generation from the crossing of paired differentiating traits.

1.3.3 Induction

Induction is a way of thinking which refers to the general concepts, principles or conclusions abstracted from a number of individual things. Induction is meant to generate novel hypotheses. Induction can be divided into full induction and incomplete induction. Once all individuals in a certain set have been considered in the drawing of a general conclusion, this process is called complete induction. In biology, when the target set of a study is composed of limited individuals, this kind of reasoning can be adopted, for example, the determination of triplet sequences for 64 standard codons [38-43]. Incomplete induction refers to the process in which a conclusion is drawn only from some individuals, and there is no exception when appling the conclusion to some other individuals. In practice, investigators are always faced with concrete instances, first obtaining the knowledge of these individuals, and then concluding a general knowledge based on these particulars. However, the most taboo in induction is making a generalization from isolated incidents, forming a conclusion only based on a small number of individuals, and considering this conclusion as inevitable, which puts the original into doubt.

1.3.4 Relationship between three kinds of reasoning

Induction and deduction, which are the earliest and most widely used ways of thinking, reflect two opposite directions of the thinking process during the recognition of things: the former is from the individual to the general, while the latter is from the general to the individual. Induction and deduction are not isolated from each other, but interrelated and mutually conditioned. On the one hand, no induction means no deduction, induction is the basis of deduction, provides a general prerequisite for the deduction. The general principles as a starting point for deduction often arise from induction. On the other hand, deduction provides guidance for the induction. Induction summarizes the individuals, whose guiding principle is often the result of another deduction.

For example, Darwin observed and summarized a large amount of experimental materials, and produced the conclusion of "biological evolution". A prerequisite of this was that he had accepted the idea of biological evolution from Lamarck and others, and the idea of geological evolution from Gilbert Ryle. And based on these ideas previously formed, Darwin thought inductively on his observations. Yet, the missing point in the hypothesis is that there is no axiomatic theory yet that comprises evolution at its core, in particular because the principle of natural selection has not been defined in an explicit manner in the world of physics. Abductive reasoning, which is last way of thinking carried out by human beings, provides a starting point in scientific research. When people are faced with an interesting phenomenon, abduction generates new ideas or assumptions based on the knowledge gained from a previous induction process. Subsequently, deduction makes the assessment for the hypothesis generated in abductive reasoning and induction confirms the assumptions with empirical data, and creates new

knowledge at last [44]. A central point, therefore, of the present research in biology is to provide a physical status to natural selection. This will ask for yet another stage of induction, where natural selection will appear as an authentic principle of physics.

1.4 Analysis neighbors to make inference and gain new knowledge

From genome sequencing, we can get a genome sequence, then usually we want to know what gene it is, whether there are any other similar genes, what control signals it preserves, which and how many articles described this, and so on. In order to better handle the links between knowledge from different areas, and make comprehensive analysis, a good idea is to put together all information related to a target gene and its function. Benefitting from the development of databases and Internet, through Internet browsing, and tracing related links, we can track information including pictures and even voices related to the topic, then we can do further analysis on the topic.

When searching through literature, usually we are not sure what keywords to use, but we can use the concept of “neighborhood” (or relationship) to improve the efficiency. Given a document database, in addition to including the title, author, abstract, each article can also be attached to keywords. The frequencies of these keywords can form a huge two-dimensional table, or vectors, with article records as rows and the keywords as columns. By comparing the similarity between keywords attached to articles and those keywords given by a searcher, one can quickly find the target article. Also a similarity comparison can be taken between article records. Both mean extracting the "neighbors" for an article and/or keywords used in searching.

These neighbors could be automatically extracted easily from a structured database through a proper computer program. Of course here, the "neighbor" is not only referring to people next to each other, and it means more than in the sense of geography and structure. It should be understood in a broadest possible way: any object must stay in a relationship with some other objects, and for each kind of relationship, there must be some objects which stay more closely to each other than others. As the saying from the Book of Changes documented 3000 years ago goes “Birds of a feather flock together.” It means things are not isolated from each other, but are connected to each other according to “like attracts like.” Biological objects are also not isolated from each other; but involved in various relationships regarding space, or time, or other forms. Taking biology into account, neighbors are a set of biological objects which are close to each other when considering a certain relationship, and it will be helpful to obtain knowledge for an unknown genomic object through studying different forms of "neighbors" in the genome.

Web Search Engines, which have become popular since the 1990s, also incorporate the concept of neighbors. For example, when using "E. coli" or "database" independently to search in Google or other search engines, generally we can not find the pages dealing both with "E. coli" and "database" at the same time. However, using "E. coli" and "database" to search, we can easily find links related to both. Here, the link is referring to the server, where the text vector of the interface is the closest neighbor of the text vector representing both "E. coli" and "database".

Two genes on the chromosome can be seen as geographic neighbors if they are close to each other. The concept of operon or the concept of gene islands in a greater level can clearly reflect the similarity of such relationship in terms of geography measured by distance.

1.4.1 Neighborhoods defined along the chromosome

On chromosomes, genes are arranged linearly. For example, the *lac* operon discovered by Jacob and Monod contains three structural genes *lacZ*, *lacY* and *lacA* [45]. Among them, *lacZ* encodes beta galactosidase, which is responsible for breaking lactose down into glucose and galactose; *lacY* encodes beta-galactoside permease, which is responsible for transporting environmental lactose into the cell; and *lacA* encodes thiogalactoside transacetylase, which is responsible for transferring an acetyl group from acetyl-CoA to beta galactosides. The *trp* operon contains five structural genes (*trpA*, *trpB*, *trpC*, *trpD* and *trpE*). All of them are involved in the tryptophan biosynthesis. This indicates that the functions of neighbor genes are often related, and this can help us to guess the function of one gene from other genes if they come from the same operon.

However, we have to notice that it would be a huge project to delimit the operons on a chromosome through laboratory experiments. At present most of the operons are predicted *in silico* [46, 47], for example, among operon records in RegulonDB, more than half (1402) are marked as "inferred computationally without human oversight". Also unfortunately, there are numerous operons that consist of only one gene. Following is a summary from RegulonDB database, nearly 70% of the predicted operons contain only one gene, about 20% contain two genes, and the operons with multiple genes are very limited. Both of the above limits make it more difficult to predict functions of genes just from the perspective of an operon, and we need to find other kinds of neighborhood to help the process of prediction. In addition to the operon, another genomic object (genomic island) defined according to the distance in a larger unit of DNA fragments is also very effective for exploring the function of unknown gene, and this will

be described in detail in the literature review of Chapter 3.

1.4.2 Neighbors defined in the perspective of evolution

During evolution, an organism inherits many genes from its common ancestors and these genes and also their products preserve some similarities. From the evolutionary point, genes encoding similar proteins constitute a gene family. Usually these neighbors are called homologous genes. In more detail, the neighbors are called orthologs if they are from different species (genomes), while the neighbors are called paralogs if they are from the same species (genome). Sequence similarity often leads to structural similarity, and sometimes even function similarity. According to the sequence similarity, Fang and the others classified the genes from a genome into two categories: one group is composed by Persistent genes which are preserved in the majority of species, and the other group is composed by Orphan genes which exist only in a few species [48]. To some extent, the persistent genes in bacteria are involved in intermediary metabolism, RNA metabolism, information transfer related to RNA, all of which are responsible for the biological processes common to most of organisms, while orphan genes make great contribution to organisms in opening up new habitat [49].

From the perspective of evolution, another relationship exists between different genes, referred to as co-evolution. The concept of co-evolution originates in the work of Charles Darwin, who mentioned it in "Origin of Species", and described it in detail in "Fertilisation of Orchids". Biological co-evolution refers to "a change in a biological object will lead to changes in the relevant biological objects." During evolution, biological objects are not static, but change dynamically. Each biological object cannot exist in isolation; it needs to play its own roles, albeit

in contact, tight or loose, with the surrounding biological objects in its environment. For example in nature, the viruses, fungi, bacteria, nematodes, insects, mammals and other organisms may exert selective pressure on plants, and make plants generate new resistance materials to reduce the attack from enemies; these new resistant materials are poisonous to the plant natural enemies, sometimes even lethal to them, and this in turn makes the plant enemies evolve new traits to overcome the resistance of the plant [50-52]. There are many examples about co-evolution, such as host and virus [53-55], host and parasite [56]. All examples above reflect the co-evolutions at the level of the species. At a microscopic level, numerous examples have also been found about the co-evolution between biological molecules. For example such relationship exists between tRNA 3' end sequence and its processing enzyme [57], and $tRNA^{Asn}$ and tRNA-dependent aminotransferase in Archaea [58]. An amino acid change in a protein may result in changes in its interacting proteins to complete some kind of cooperation between the two proteins [59-63]. A similar phenomenon of co-evolution also exists in a metabolic pathway, for example sialic acid biosynthesis [64].

In a word, co-evolution can be found at all levels of biology, such as species and species, species and the environment [65, 66], two (or more) organs [67], two (or more) metabolic pathways[68, 69], and two (or more) macromolecules (like proteins, RNA, DNA). It needs specific analysis according to the concrete situation when using this kind of neighbors from the view of co-evolution.

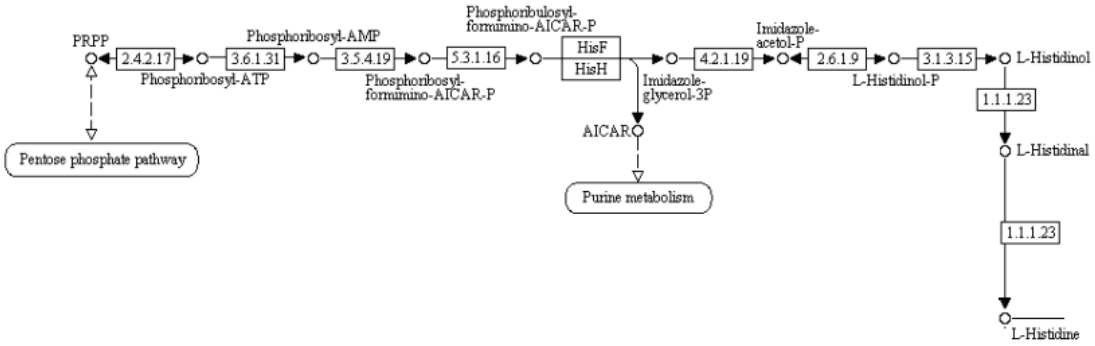
There are already some databases devoted to this kind of work, such as Gene Ontology (GO) database, which gives the description for each gene about the subcellular location (or cell components), biological processes and molecular functions [70];

Clusters of Orthologous Groups of proteins (COG) collect protein sequence appearing in at least three species with completely genome sequencing [71]; KEGG Orthologs collects orthologous genes not only considering sequence similarity but also considering the metabolic pathway and molecular complexes [72]; Ribosomal Database Project (RDB) collects ribosome related data and services, including online analysis, alignment and annotation for 16s rRNA from Bacteria and Archaea [73].

1.4.3 Neighborhoods defined in terms of metabolic pathways

A series of chemical reactions occurring in live cell and catalyzed by enzymes is called a metabolic pathway. The initial material input into a metabolic pathway is modified by these chemical reactions. For example, a metabolic pathway converting substance A to X, might be carried out in the form of (A-> B-> C-> ... X), where the initial substance A become the end product X through intermediate metabolites such as B, C and so on. Here different genes encoding enzymes involved in the same metabolic pathways can also be seen as a kind of neighbors. In some cases, this kind of neighborhood is represented in a simple form. For example, genes encoding enzymes involved in histidine biosynthesis pathway are clustered in an operon (Figure 1.1) [74].

a.



b.

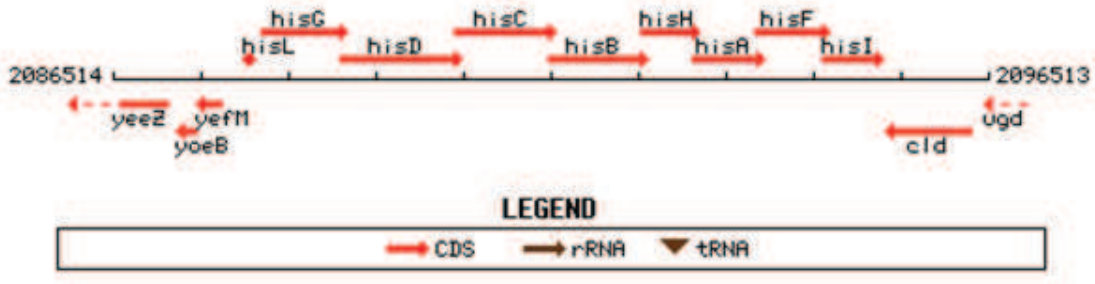


Figure 1.1 Gene neighbors on the chromosome
a. Enzymes participating in histidine biosynthesis pathway modified from the KEGG database ([23]); b Distribution of genes encoding enzymes involved in histidine biosynthesis on the chromosome (obtained from the Colibri database [16])

There are three fates for metabolites produced from enzymatic reactions: they might be stored in the cells; they might be input into the next chemical reaction as an intermediate metabolite; they might be an initiator to start a new metabolic pathway. For this reason, some metabolic pathways are much more complex than above-mentioned histidine metabolic pathways, and several pathways link with each other to form a metabolic network. In this network, there is another kind of neighborhood for enzymes according to which pathway they are involved in and which reaction they catalyze. And these neighbors have some similar known or unknown common characters. For example for enzymes participating in the biosynthesis of aromatic

amino acids (phenylalanine, tyrosine and tryptophan), their encoded genes are not arranged in an operon on the chromosome, but their codon usages have a special relationship, since they are not randomly distributed in the image of Correspondence Analysis (CA), but approximately in a straight line (Fig. 1.2) [75].

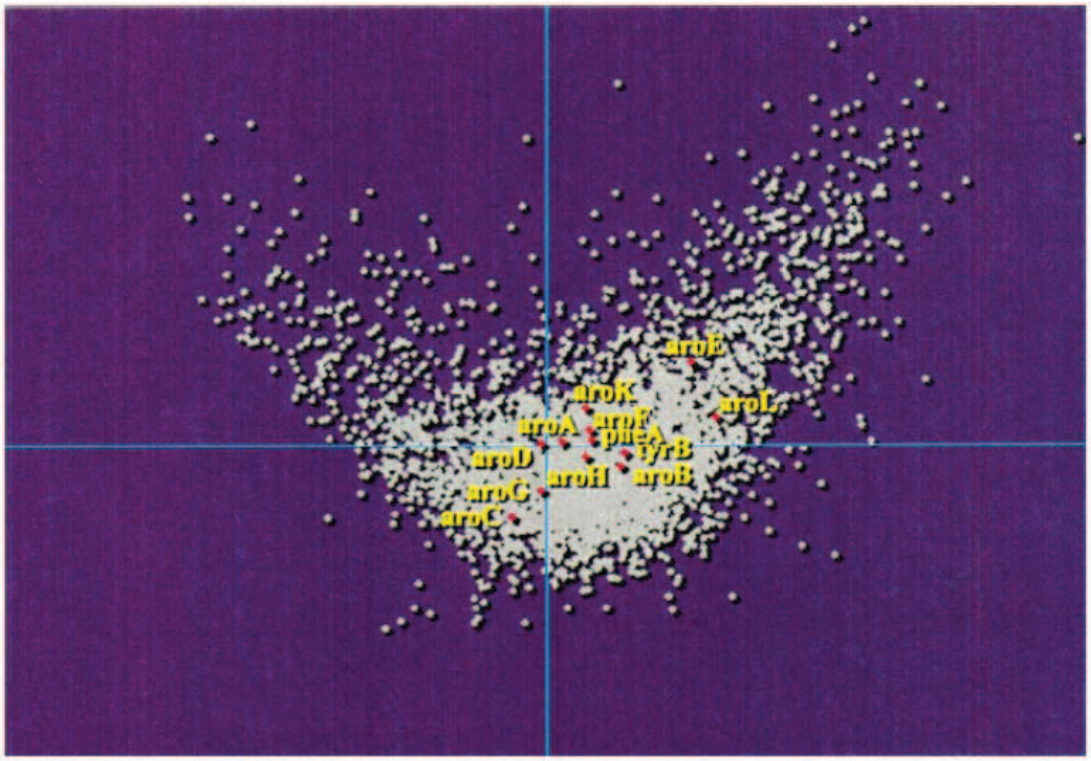


Figure 1.2 Gene neighbors in the statistical space
Distribution of genes involved in biosynthesis of aromatic amino acids in the cloud of factor correspondence analysis of the codon usage

KEGG PATHWAY is one of the most widely used databases for metabolic pathways, which describes the network formed by the interactions and reactions among molecules, and with the manually generated pictures for illustration of these entities [76], and this database is divided into seven component databases, namely, Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes, Organismal Systems, Human Diseases and Drug Development. This would be a good starting point to look for neighbors in metabolic pathways.

1.4.4 Neighborhoods defined in statistical spaces

From a statistical point of view, some techniques can summarize the commonalities for the bulk of objects with some measurable characteristics, though it is impossible to achieve this just by intuition. Usually it is easy to judge intuitively whether two objects are neighbors or not in a two- or three-dimensional space just at a glance. However, it is impossible to see how it appears in a multi-dimensional space just by looking, and indeed many biological objects belong to multi-dimensional spaces. For example, with 61 codons out of the possible 64 triplets, each gene can be represented in a 61-dimensional space, appropriately normalized, with each axis representing the frequency of use of the corresponding codon. And each protein can be represented in a 20-dimensional space, appropriately normalized, with each axis representing the frequency of the corresponding amino acid. The task to identify neighbors in multi-dimensional space could be completed through statistical analysis tools at some point. For example, through the principal component analysis (PCA), correspondence analysis (CA) and other multivariate analysis techniques, we can extract the main factors from multi-dimensional data. And to some extent, through multivariate cluster analysis, we can see how several objects are close enough to form a cluster. For example, by correspondence analysis on codon usage in *E. coli*, several factors were extracted, and three clusters were obtained through the dynamic programming algorithm [77]. Also by correspondence analysis on amino acids usage in *E. coli*, the first three main factors were extracted, and two clusters were obtained through dynamic programming algorithm[78, 79]. And to know whether the classification partitioning is meaningful in the biological sense, we could get some ideas from the genes' (or proteins') function in the same cluster determined by Cluster analysis. Obviously it is possible, a posteriori, to label the genes in

each of the classes discovered. It is this, and only this, that enable us to justify the distribution discovered by this "data-driven" approach and to give it a biological meaning. For example, in the correspondence analysis on codon usage in *E. coli*, three clusters have their respective biological meanings: The first two classes encompass gene expressed either continuously at a high level, or at a low level; the third class consists of genes corresponding to surface elements of the cell, genes coming from mobile elements as well as genes resulting in a high fidelity of DNA replication [77]. And from the correspondence analysis on the amino acid usage, the proteins integrated in the inner membrane can be easily identified from the bulk [78, 79].

1.5 Questions to be answered in this thesis

Starting from the concept of neighborhood, I have tried to explore the bacterial genome through the following three sub-projects:

1. Correspondence Analysis of a bacteria proteome through Model Based Clustering

2. Genomic Islands

3. Phylogenetic study of the upstream regulating and coding region of ThrRS (Threonyl-tRNA synthetase)

2.RESULTS

2.1 Correspondence Analysis (CA) and Model Based Clustering (MBC)

2.1.1 Literature overview

2.1.1.1 Concept and Application of Correspondence Analysis

Correspondence analysis (CA) is an exploratory data analysis technique to analyze two-dimensional or multidimensional data with implied relationships between the rows and columns [80]. Traditional hypothesis testing is used to verify a priori hypothesis about relations between variables. Unlike traditional hypothesis testing, exploratory data analysis is used to identify systematic relations between variables when there are no prior expected relationships or incomplete ones from the original data. Simplification of data provides useful information about the data with less expense of computer time. Correspondence analysis is one of such statistical technique which can greatly simplify the complexity of data.

To some extent, CA may be defined as a special case of principal components analysis (PCA), especially for those cross-tabulations. In general, PCA is more useful in dealing with continuous data, while CA deals with discrete data (i.e. contingency tables) more effectively. It will transform a table of numerical information into a graphical display, in which each row and column from the primitive table is represented by a point. The picture produced by CA can reveal relationships (or neighborhoods) from data in three aspects: the intra-relationship between different rows; the intra-relationship between different columns, and also the inter-relationships between rows and columns.

There are several features distinguishing CA from other methods of data analysis. One is the multivariate treatment of the data through simultaneous consideration of multiple categorical variables. Another important feature is the graphical display of row and column points in a biplot, which can help in detecting structural relationships among variable categories and objects.

Correspondence analysis is particularly suitable for data in large matrix, with only one restriction that the data cannot contain any negative entity. And so far, CA has been widely applied in geology, agriculture and bio-medicine. For example, Salgueiro applied correspondence analysis in the assessment of mine tailings dam breakage risk in the Mediterranean region[81]; Faye et al. applied correspondence analysis to evaluate the interrelationships between herd management practices and udder health status[82]; Vanoeteren et al. applied correspondence analysis to evaluate the trace elements in human lung tissue[83]. Similarly, in the sub-domain related to genomics and proteomics, correspondence analysis has been used to analysis the codon usage and amino acid composition. Guerdoux-Jamet et al. applied correspondence analysis in codon usage of 4285 genes (at that moment 1815 genes with known functions) in *Escherichia coli*, and thought the E. coli outer membrane is a patchwork of products from different genomes[84]. There are different approaches to calculate codon usage, and have formed different types of correspondence analysis, such as the absolute codon frequency (AF), the relative codon frequency (RF), the relative synonymous codon usage (RSCU) and within-group CA (WCA) [85-88]. Suzuki et al. evaluated and compared the above four CA methods by applying them to 241 bacterial genomes, and the results indicate that WCA is more effective than the other three methods since it can reveal sources that were previously unnoticed in some genomes; e.g. synonymous codon usage related to replication strand skew was detected

in *Rickettsia prowazekii* [89]. As early as 1994, Lobry et al. applied correspondence analysisi on E. coli available data, and found hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 E. coli chromosome-encoded gene[90]. Later, with more genome sequences available, correspondence analysis was applied to more model organisms, such as *Buchnera*[91] and *Bacillaceae*[92]; Dumontier et al. applied correspondence analysis to 360, 000 open reading frames from Archaea, Bacteria and Eukaryotes, and found that there are species-specific and environment residue preference among different organisms [93]; Also there are several studies using the mixed data from multi-organisms as the input for correspondence analysis [94, 95]. Pascal et al. have checked whether the rules of amino acid composition might exist while placing them in a functional genomics perspective, at first in three model proteomes, e.g. *E. coli* K12, *B. subtilis*, and *Methanococcus jannaschii* [78] and later in another 28 prokaryotic proteomes[79].

2.1.2 Applying CA and MBC in the study of genome: *Psychromonas ingrahamii*

2.1.2.1 Study strategy and result summary

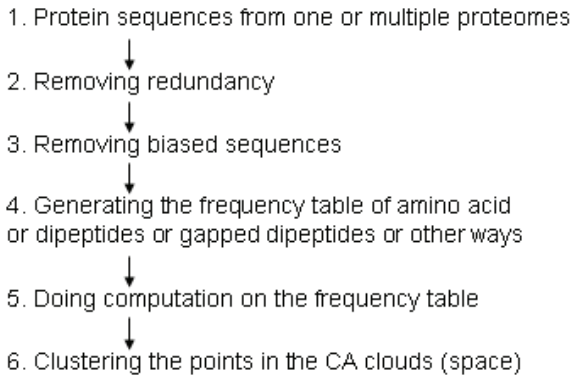


Figure 2.1 Study strategy for Correspondence Analysis in protein sequences

The study strategy for CA is summarized in Fig. 2.1. Due to different ways of calculating the frequency table, such as amino acid, dipeptide, or gapped dipeptide, different CA computation could be carried out. In Article 1, the protein sequences from genome *P. ingrahamii* were analyzed with the frequency table calculated by amino acid, and the results can be viewed from Figure 2 in Article 1, and summarized as following: (1) there are 6 classes of proteins clustered from Model Based Clustering, (2) integral inner membrane proteins are not sharply separated from bulk proteins, (3) there is strong opposition between asparagine (N) and the oxygen-sensitive amino acids methionine (M), arginine (R), cysteine (C) and histidine (H) and (4) one of the previously unseen clusters of proteins has a high proportion of “orphan” hypothetical proteins.

Article 1

Genomics of an extreme psychrophile, *Psychromonas ingrahamii*

Monica Riley, James T Staley, Antoine Danchin, **Tingzhang WANG**, Thomas S Brettin, Loren J Hauser, Miriam L Land and Linda S Thompson

BMC Genomics 2008, 9:210

BMC Genomics

BioMed Central

Research article

Open Access

Genomics of an extreme psychrophile, *Psychromonas ingrahamii*

Monica Riley*[1], James T Staley[2], Antoine Danchin[3], Ting Zhang Wang[3], Thomas S Brettin[4], Loren J Hauser[5], Miriam L Land[5] and Linda S Thompson[4]

Address: [1]Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA 02543, USA, [2]University of Washington, Seattle, WA 98195-7242, USA, [3]Genetics of Bacterial Genomes, CNRS URA2171, Institut Pasteur, 28 rue du Dr Roux, 75015 Paris, France, [4]DOE Joint Genome Institute, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA and [5]Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

Email: Monica Riley* - mriley@mbl.edu; James T Staley - jtstaley@u.washington.edu; Antoine Danchin - adanchin@pasteur.fr; Ting Zhang Wang - wangtz@pasteur.fr; Thomas S Brettin - brettin@lanl.gov; Loren J Hauser - hauserlj@ornl.gov; Miriam L Land - landml@ornl.gov; Linda S Thompson - lthompsonnm@comcast.net

* Corresponding author

Published: 6 May 2008

BMC Genomics 2008, **9**:210 doi:10.1186/1471-2164-9-210

Received: 3 September 2007
Accepted: 6 May 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/210

Abstract

Background: The genome sequence of the sea-ice bacterium *Psychromonas ingrahamii* 37, which grows exponentially at -12C, may reveal features that help to explain how this extreme psychrophile is able to grow at such low temperatures. Determination of the whole genome sequence allows comparison with genes of other psychrophiles and mesophiles.

Results: Correspondence analysis of the composition of all *P. ingrahamii* proteins showed that (1) there are 6 classes of proteins, at least one more than other bacteria, (2) integral inner membrane proteins are not sharply separated from bulk proteins suggesting that, overall, they may have a lower hydrophobic character, and (3) there is strong opposition between asparagine and the oxygen-sensitive amino acids methionine, arginine, cysteine and histidine and (4) one of the previously unseen clusters of proteins has a high proportion of "orphan" hypothetical proteins, raising the possibility these are cold-specific proteins.

Based on annotation of proteins by sequence similarity, (1) *P. ingrahamii* has a large number (61) of regulators of cyclic GDP, suggesting that this bacterium produces an extracellular polysaccharide that may help sequester water or lower the freezing point in the vicinity of the cell. (2) *P. ingrahamii* has genes for production of the osmolyte, betaine choline, which may balance the osmotic pressure as sea ice freezes. (3) *P. ingrahamii* has a large number (11) of three-subunit TRAP systems that may play an important role in the transport of nutrients into the cell at low temperatures. (4) Chaperones and stress proteins may play a critical role in transforming nascent polypeptides into 3-dimensional configurations that permit low temperature growth. (5) Metabolic properties of *P. ingrahamii* were deduced. Finally, a few small sets of proteins of unknown function which may play a role in psychrophily have been singled out as worthy of future study.

Conclusion: The results of this genomic analysis provide a springboard for further investigations into mechanisms of psychrophily. Focus on the role of asparagine excess in proteins, targeted phenotypic characterizations and gene expression investigations are needed to ascertain if and how the organism regulates various proteins in response to growth at lower temperatures.

Background

Well over half of the earth's surface is cold: deep oceans, mountains, polar regions. Likewise, Earth's solar system contains many planets and planetary bodies that are also cold. The cold environments on Earth are teeming with life [1] offering hope that other cold environments in our solar system such as Mars and Jupiter's moon, Europa, may harbor life [2]. For this reason it is surprising that so little is know about the lifestyle, particularly of microbial psychrophiles at low temperatures.

Psychrophiles have been studied primarily to understand biological mechanisms of adaptation to extreme conditions. Microbial physiologists have long been interested in psychrophiles as they employ mechanisms allowing them to maintain life processes at temperatures where rates of reactions and molecular properties present challenges. In reaching for an understanding of how life processes work at extremes of temperature, most of the focus to date has been on the properties of enzymes of extremophiles (reviewed by [3,4]). No single consistent answer has emerged to account for adaptation to temperature extremes. To date, no single type of modification is uniformly found in the enzymes of psychrophiles; instead numerous small and subtle differences appear to account for their increased flexibility thereby enabling them to function at low temperatures.

Recently whole genome sequences have been determined for psychrophiles *Colwellia psychrerythraea* 34 H [5], *Idiomarina loihiensis* L2TR [6], and *Pseudoalteromonas haloplanktis* TAC125 [7]. We now add the genomic sequence of the extreme species, *Psychromonas ingrahamii* 37 which grows at even colder temperatures. Availability of complete genome sequences provides the opportunity to search all of the proteins of the organisms for similarities and differences that might have bearing on the ability of the organism to grow at low temperatures.

The extreme psychrophile, *Psychromonas ingrahamii* was isolated from sea ice from the Arctic. It grows exponentially with a doubling time of 240 hours at -12 °C and may well grow at even lower temperatures [8]. These temperatures do not necessarily solidify salt water or cytoplasm into ice. Liquid water has been shown to exist at grain contacts as low as -20C [2].

Results and Discussion

The P. ingrahamii *genome*

The single, circular chromosome of 4.56 Mb constituting the genome of *P. ingrahamii* 37 was sequenced as a set of contigs by the DOE Joint Genome Institute Production Genomics Facility, 2800 Mitchell Drive, Walnut Creek, CA 94598, and finished at DOE Joint Genome Institute, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545. It was annotated at Oak Ridge National Laboratory, Oak Ridge, TN 37831 and deposited as GenBank file CP000510.1. Altogether 3708 genes were identified, 3545 of which were proteins of 83 residues or longer. A second round of annotation is described below.

Properties of proteins

Size

One can ask whether the sizes of proteins of a psychrophile differ from those of a mesophile. The sequences of all proteins of *P. ingrahamii* were compared to sequences of all proteins of three other bacteria: *Shewanella oneidensis* MR-1, *Vibrio cholerae* and *Escherichia coli* K-12 MG1655. 916 protein sequences were conserved among all four bacteria. The great majority of the conserved proteins were enzymes. The distribution of lengths of the 916 orthologous proteins were compared, revealing that the distribution was about the same for the comparable proteins in all four bacteria (Figure 1). Ideas that proteins that are required to function at low temperatures would be either shorter or longer than those of mesophiles are not borne out.

Amino acid composition

Amino acid composition of an organism's proteins is affected by nucleotide composition of the DNA. The GC content of *P. ingrahamii* DNA was determined experimentally to be 40% [8], verified by the composition of the total genome nucleotide sequence (40.1%). Overall amino acid content of the encoded proteins compared to those of *V. cholerae, S. oneidensis* and *E. coli* is shown in Table 1 (P. Sharp, personal comunication). The amino acids isoleucine, asparagine, lysine, phenylalanine and tyrosine all are present in higher percentage in *P. ingrahamii*.

However, like the genome, the codons for these residues are GC rich, a factor that must be taken into account. Codon usage for *P. ingrahamii* is shown in Table 2 for all CDSs (coding DNA sequences) and for the highly expressed genes *tuf, tsf, fus*A+*RP* genes (P. Sharp, personal communication). Thus when amino acid content is examined as a function of GC3s, correcting for GC content at synonymously variable third positions, the overall amino acid composition of *P. ingrahamii* was found not to be remarkable. At the value of 34.2 determined for *P. ingrahamii*, amino acid contents fall on the curves generated for GC3s dependency in 80 other organisms (P. Sharp, personal communication) (data not shown) [9].

Correspondence analysis (CA)

Lobry and Chessel [10] examined datasets of amino acid content and codon usage in thermophilic and mesophic bacteria and found evidence that the amino acid composition of thermophilic proteins was under the control of a

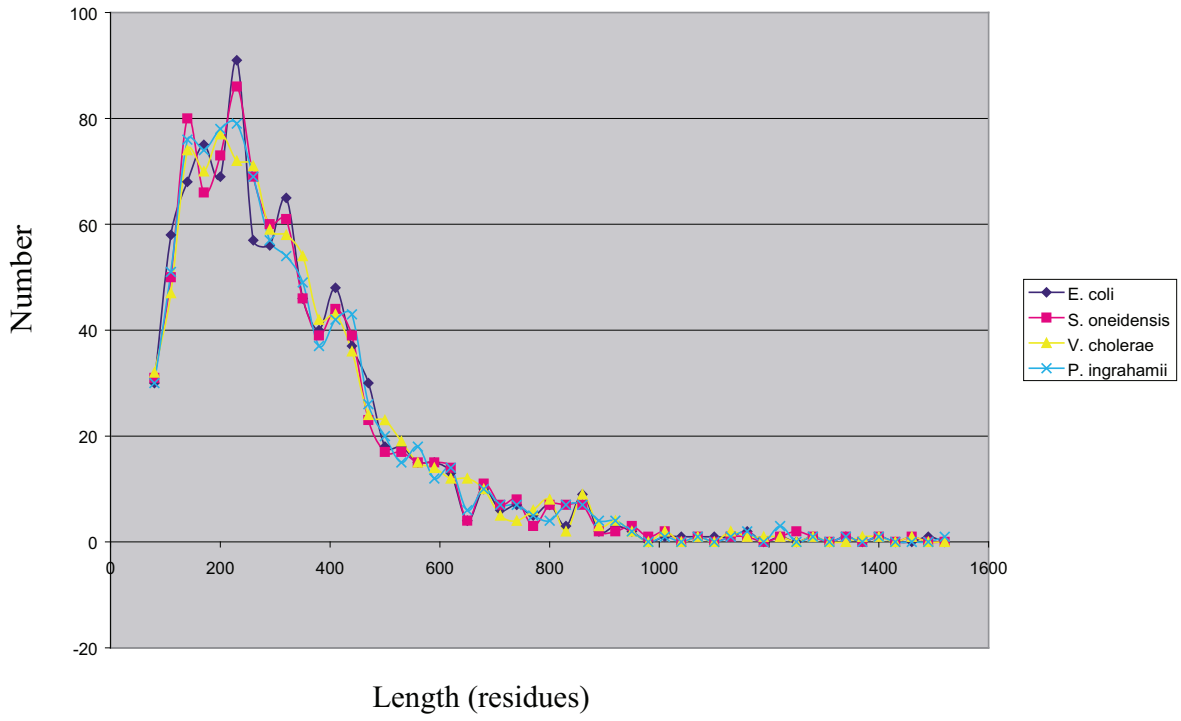


Figure 1
Distribution of lengths of 916 orthologous proteins in four bacteria. Distribution of lengths of proteins as numbers of amino acid residues, ranging from 83 to 1501, in increments of 30. Black diamond = *E. coli*, red square = *S. oneidensis*, yellow triangle = *V. cholerae*, blue cross = *P. ingrahamii*.

pressure at the nucleic acid level, not a selection at the protein level. An extended study [11] produced similar results of no connection. However, the authors identified the most discriminating codon as being AGG for arginine, present in many thermophiles, not in either mesophiles or a psychrophile. *P. ingrahamii* fits this observation in that it does not use the AGG codon (Table 2).

We analyzed the amino acid compositions of all individual proteins of *P. ingrahamii* by CA [12] after trimming the first 10 and last 5 residues because there is a strong nucleotide bias at these positions. Data for each protein is presented in Additional file 1: "Ping Correspondence Analysis.xls". Clustering of groups of proteins with similar composition was performed using a bayesian approach as proposed by Bailly-Bechet *et al*. [13] Figure 2 presents a

Table 1: Amino acid composition (%) of total proteins of 4 bacteria

Organisms*	S	L	R	P	T	V	A	G	I	F
Esccol	5.83	10.65	5.54	4.43	5.41	7.1	9.49	7.37	6	3.9
Shewone	6.47	10.97	4.64	4.06	5.37	6.76	9.42	6.79	6.03	3.97
Vibcho	6.33	10.9	4.95	4.02	5.19	7.07	9.15	6.68	6.04	4.08
Psying	4.82	10.99	4.08	3.71	5.49	6.57	8.49	6.68	7.53	4.47
	Y	C	H	Q	N	K	D	E	W	M
Esccol	2.85	1.17	2.27	4.43	3.95	4.41	5.14	5.75	1.53	2.8
Shewone	3.05	1.09	2.33	4.93	4.12	5.14	5.29	5.74	1.28	2.54
Vibcho	2.96	1.05	2.4	5.17	3.9	4.93	5.02	6.2	1.32	2.63
Psying	3.16	1.11	2.13	5.98	4.92	6.25	5.47	4.54	1.16	2.45

*Esccol = *Escherichia coli* MG1655, Shewone = *Shewanella oneidensis* MR-1, Vibcho = *Vibrio cholerae*, Psying = *Psychromonas ingrahamii* 37

Table 2: Number of codons in highly expressed and in all genes of Psychromonas ingrahamii

		High	All			High	All			High	All			High	All
Phe	UUU	107	41275	Ser	UCU	139	14451	Tyr	UAU	47	25785	Cys	UGU	35	8165
Phe	UUC	103	10226	Ser	UCC	3	7455	Tyr	UAC	86	10612	Cys	UGC	5	4681
Leu	UUA	128	50155	Ser	UCA	73	16739	Ter	UAA	32	2366	Ter	UGA	1	571
Leu	UUG	44	17564	Ser	UCG	16	7210	Ter	UAG	7	608	Trp	UGG	31	13405
Leu	CUU	150	20860	Pro	CCU	98	13581	His	CAU	56	17366	Arg	CGU	408	17987
Leu	CUC	9	7631	Pro	CCC	5	9132	His	CAC	67	7202	Arg	CGC	74	11043
Leu	CUA	87	9170	Pro	CCA	110	10292	Gln	CAA	167	29958	Arg	CGA	12	5132
Leu	CUG	52	21318	Pro	CCG	32	9734	Gln	CAG	50	22366	Arg	CGG		3939
Ile	AUU	180	48789	Thr	ACU	189	17262	Asn	AAU	91	39859	Ser	AGU	48	19260
Ile	AUC	208	20430	Thr	ACC	31	19451	Asn	AAC	128	16887	Ser	AGC	60	14213
Ile	AUA	22	17588	Thr	ACA	113	16391	Lys	AAA	452	57148	Arg	AGA	14	6404
Met	AUG	188	28244	Thr	ACG	41	10226	Lys	AAG	134	14939	Arg	AGG	0	2583
Val	GUU	347	29210	Ala	GCU	328	26162	Asp	GAU	228	49174	Gly	GGU	412	33290
Val	GUC	32	13155	Ala	GCC	41	21828	Asp	GAC	83	13878	Gly	GGC	119	20816
Val	GUA	198	15326	Ala	GCA	273	30437	Glu	GAA	354	47661	Gly	GGA	33	12381
Val	GUG	52	18068	Ala	GCG	62	19459	Glu	GAG	91	21337	Gly	GGG	6	10497

plot of the first three most informative axes: hydrophobicity, aromaticity and asparagine content.

The proteins fell into six classes whereas most bacteria contain at most five well separated clusters with bayesian clustering and at most four with dynamic clouds clustering [14]. Proteins of two bacterial proteomes examined to date fall into four clusters, *Aeropyrum pernix* and *Thermoplasma acidophilum* [14]. Subsequently it was seen that psychrophilic *P. haloplanktis* has 5 clusters [7], exceeded now by *P. ingrahamii*, a more extreme psychrophile, which resolves into six. In *P. ingrahamii*, the bulk of the proteins fall into three clusters (clusters 1, 2, 6). The integral inner membrane proteins (IIMPs) fall largely into clusters 3 and 5 (Figure 2). In other bacteria the IIMPs are distinctly separate from the bulk of other proteins [14]. In *P. ingrahamii*, however, separation is poor and there is some continuity between the IIMPs and the bulk proteins. Evidently the property of hydrophobicity is not distributed in psychrophile proteins as in mesophiles.

Not seen in any other bacteria viewed to date by CA is the sixth group, cluster 4. This group is of 57 proteins characterized by an excess of threonine that forms a bleb protruding from one of the core protein clusters. Examining the annotations of these proteins reveals that almost half are hypothetical proteins with no homologs in other bacteria examined at threshold Pam 150. It is tempting to suggest this group of proteins could be involved in facilitating low temperature growth.

As to amino acid content, one sees that asparagine occurs more frequently than expected for a random distribution, and on the other hand that cysteine, methionine, arginine and histidine are relatively rare (Figure 2). This same asparagine-driven bias has been seen in other psychrophiles, notably *P. haloplanktis* [7]. Deamidation via cyclization into aspartate threatens integrity of asparagine, a process which is sensitive to higher temperature, providing a rationale for an asparagine excess in psychrophiles. The corresponding lower amounts of cysteine, methionine, arginine and histidine can be understood as decreasing proportions of these oxygen-sensitive residues. Oxygen concentrations are higher in the liquid medium at low temperatures. Although similar composition gradients were seen in another psychrophile, *P. haloplanktis*, they are stronger in *P. ingrahamii*, perhaps correlated with its lower growth temperatures.

Thus, the main features that emerged from the CA of *P. ingrahamii* protein compositons (Figure 2) are that (1) there are more classes of proteins than have been seen in other bacteria, (2) one of these classes, cluster 4, has a high proportion of "orphan" hypothetical proteins, (3) IIMPs merge into bulk proteins rather than occupying a separate space, possibly due to IIMPs having lower hydrophobic character, and (4) one notes the strong opposition between asparagine, sensitive to heat, and the amino acids methionine, arginine, cysteine and histidine, sensitive to oxygen.

Annotation

Of the 3545 genes for proteins of length 83 or more, 41 are fusions of two genes which in other organisms are separate and independent. They are distributed as 21 fused enzymes, 9 fused regulator components, 4 fused ABC

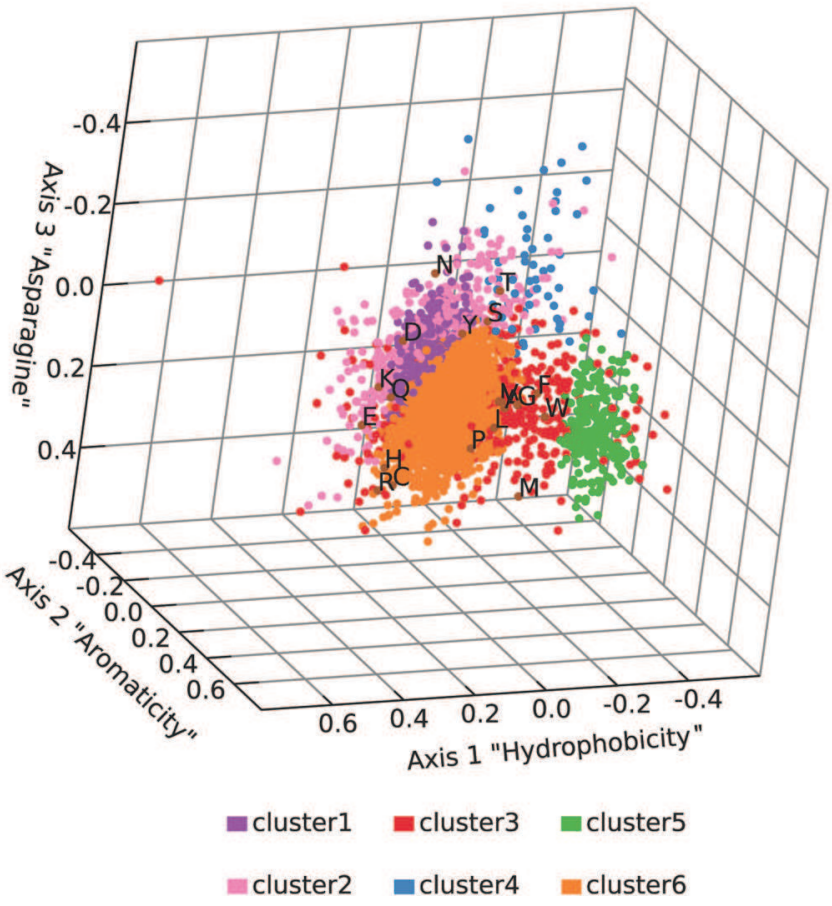


Figure 2
Correspondence analysis of amino acid content of *P. ingrahamii* proteins. Proteins over 100 residues were subjected to correspondence anlysis by amino acid content, clustered and plotted on first three most informative axes. Amino acid frequencies are superimposed. (See Methods).

transporter components, 4 fused phosphotransferase system (PTS) components and 3 mixed enzyme-regulator combinations. Fused genes can cause problems with annotation based on results of sequence similarity algorithms. To avoid such problems, we split fused genes for purposes of sequence comparisons in order to be able to identify orthologs of both parts independently.

The first round of annotation was carried out at Oak Ridge National Laboratory, posted for public access August 2006 [15]. We have here supplemented this data with manual analysis using the dynamic search programs of the Darwin system [16]. In this system after first approximations of sequence similarities, amino acid substitution tables are recalculated appropriate to degree of similarity and to the codon usage, and then pairwise alignments are produced. Degree of similarity between two sequences is expressed as percent identity and as Pam values [The Pam score (point accepted mutations) is an inverse measure of sequence differences] [17].

We first processed the *P. ingrahamii* protein sequences in relation to the then-completed 111 bacterial genomes. We extracted from this data the match with best (lowest) Pam score for each *P. ingrahamii* protein. The descriptions of the orthologs were retrieved from RefSeq and/or Genbank at the National Center for Biotechnology Information (NCBI) Web site [18]. The results allowed us to add some predicted protein products to the initial JGI annotation results.

Since the set of 111 bacterial genomes we first used was not balanced in respect to types of bacteria, we also identified orthologs in a reference set of 53 genomic sequences of organisms which were chosen to span the breadth of bacterial species. To include other sequenced psychrophiles, the set includes two additional marine species, *C. psychrerythraea* 34 H and *I. loihiensis*. *C. psychrerythraea* 34 H grows over the range -1C to 10C [5]; *I. loihiensis* has a broad temperature range from 4C to 46C [6].

Annotations and function information are attached in Additional file 2: "Ping Annotations 2.xls". The rank order of similar sequences and number of "best hits" are shown in Table 3.

Surprisingly, the organism with greatest similarity is *V. cholerae*. It is surprising because neither is *V. cholerae* one of the psychrophiles nor is it any member of the Order *Alteromonadale*s. *P. ingrahamii* is a member of the Family *Alteromonadacae* in the Order *Alteromondales*. *Psychromonas* is related to other members of this family such as *Shewanella*, *Colwellia* and *Idiomarina* [20]. Like some other members of these families, *P. ingrahamii* is a marine organism and has gas vesicles. We found there are extensive similarities with proteins of other Alteromonads as expected, but there are even more similarities with *V. cholerae*, even though the the Order *Vibrionales* and Family *Vibrionaceae* are separate from the *Alteromonadales* [19]. This observation may be explained by the selected conservation of genes from a common ancestor between these two genera and their loss by their closest relatives or it may indicate confusion due to unexpected horizontal gene transfer of 16S rDNA within these lineages.

In manually curating, we used practices aimed both at discovery and at caution. Whenever a "best match" was annotated as a hypothetical protein, yielding no information, we looked to the next best match. Sometimes the next best match was a very good one which provided useful information based on the annotated gene product of that ortholog.

Table 3: Organisms with similarity to greatest number of *P. ingrahamii* proteins*

Organism	Number of "Best Hits"
Vibrio cholerae	697
Shewanella oneidensis	539
Colwellia psychrerythraea	499
Escherichia coli	285
Idiomarina loihiensis	143
Pseudomonas aeruginosa	137

* sizes greater than 83 residues

It is widely appreciated in the genomics community that annotation by transfer of stipulated annotations from other organisms becomes ever more problematic at lower levels of similarity and as the number of sequential steps of sequence matches increases between the query and an experimentally demonstrated gene product. We were careful to be conservative in attributing a gene product if the match was not very close. When Pam scores were low (excellent match), we transferred directly the annotation of the match; but when in a middle range, 75–125, we sometimes generalized, removing specificity (i.e. we stipulated "an aminotransferase" instead of "aspartate aminotransferase"); and when Pam scores were high, over 125, we often used the word "predicted" to indicate that the assignment was based on a less rigorous threshold than for other assignments.

When formulating words of description of the gene products, we adopted another practice. We made an effort to name similar products similarly. We have tried to standardize product descriptions to some extent in order to make information on like proteins easier to find, thus making a list of alphabetized product names useful.

Assignments

From the above two sets of ortholog matches combined with the work of the JGI scientists, we constructed a table of *P. ingrahamii* gene numbers and corresponding best-guess annotated gene products available in Additional file 2: "Ping Annotations 2". Included in the table are the gene number, the type of gene product (enzyme, regulator, etc) and the name of the protein. (Gene number throughout this report is the number of the locus_tag which, as submitted December 2006 to Genbank, CP000510.1, has the prefix "Ping"). Altogether 217 proteins originally characterized as unknown gained an annotation by the manual process, some suggestive, others substantive.

Kinds of proteins

Table 4 gives classification of all proteins of *P. ingrahamii* by imputed function. The distribution of sizes of classes is similar to that for *Escherichia coli* K-12 [20] with enzymes being the largest class followed by transporters, then regulators, the rest divided into smaller categories.

Horizontal transfer

Eighty-one genes were identifiable as currently known horizontally transferrable loci such as transposases of insertion sequences and integrases of phages. Thus clearly horizontal exchange of mobile elements into the chromosome has taken place. Among the bacteria tested, most such loci of *P. ingrahamii* matched elements of *Shewanella oneidensis* MR1. As is the case with so many genome sequences, we cannot know with currently available information how much of the genome of *P. ingrahamii* was

Table 4: Distribution of functional types among all *P. ingrahamii* proteins*

Type of function	Number
Enzymes	1317
Unknown	761
Transporters	443
Regulators	252
Domains known	220
Factors	203
Structural	122
Horizontal	81
Carriers	47
Membrane	38
Cell process	38
Lipoproteins	21
Total	3545

* sizes greater than 83 residues

formed by horizontal transfer and how much vertically inherited.

RNA

There are 10 ribosomal RNA clusters containing 5S, 16S, 23S RNAs (see Additional file 2). The relatively high number could reflect the need for high capacity of translation at cold temperatures and an ability to adapt quickly to changing conditions of nutrient availability [21]. 86 tRNA genes support translation.

Gene clusters

There are 100 contiguous clusters of two or more related genes such as subunits of an enzyme or pathway-related proteins. The largest cluster, for septum formation and peptidogylcan synthesis enzymes, comprises 14 genes. These are genes 1140 through 1153. Altogether 850, or 25% of CDSs, reside in clusters related by function.

Paralogous groups

Enumeration of sets of paralogs within a genome requires definition of degree of relatedness. In the initial analysis done at Oak Ridge National Laboratory, 1799 proteins were identified as belonging to 510 paralogous clusters. Using Darwin analysis [22] and a more conservative threshold of relatedness (Pam =< 175), 965 proteins were identified as grouped into 273 paralogous families of sizes ranging from two to 66. This threshold is comparable to that used to enumerate *E. coli* paralogs [23] and permits comparison. Functions of paralogous groups of size 7 or above are shown in Table 5. Just as had been found previously for *E. coli*, the largest paralogous groups are transporters and regulators. Even though enzymes are present in the genome in the largest numbers, they fall into smaller, more differentiated families. For transporters and also regulators, evidently a limited number of mechanisms have evolved for these functions, creating large groups of similar proteins of either transporters or regulators that differ in specificity but not in mechanism of action. Enzymes are more diverse, use a larger variety of mechanisms, thus they fall into a larger number of smaller groups of similar proteins.

Table 5: Distribution of functional types among largest *P. ingrahamii* paralogous groups

Group size	Protein function
66	ATP-binding subunits of ABC transporters
51	Cyclic-diGMP regulation, diguanylate cyclases
24	Transcriptional regulators, LysR type
23	Substrate-binding subunits of ABC transporters
15	Two-component response regulators
15	Transcriptional regulators, LacI type
12	Peptide-binding subunits of ABC transporters
11	ATP-dependent RNA helicases
11	Short-chain alcohol dehydrogenase family
10	Tripartite C4-dicarboxylate transporter, DctM-type subunits
9	Unknown
9	Fused ATP-binding/substrate-binding subunits of ABC transporters
9	IS4 transposases
8	Two-component sensor histidine kinases
7	IS30 transposases
7	Aldehyde dehydrogenases
7	Oxidoreductases
7	Tripartite C4-dicarboxylate transporter, DctQ-type subunits
7	Crotonase-like epimerase/dehydratases
7	Extracellular amino acid-binding subunits of ABC transporters
7	Aminotransferases

Intermediary metabolism

Almost half of all enzymes annotated in *P. ingrahamii* (634/1317) were identified as enzymes of small molecule metabolism. Not all enzymes of every pathway were found, not unexpected since some orthologs may not reach the threshold of similarity employed, nevertheless there is strong evidence for standard pathways of intermediary metabolism being present.

P. ingrahamii is a facultative anaerobe capable of both respiratory and fermentative metabolism [24]. In agreement we find in its enzyme sequences that pathways are present for fermentation, glycolysis, pentose phosphate pathway, the TCA cycle and gluconeogenesis. For respiration, electron transfer agents are present such as iron-sulfur centers, flavodoxin and flavoproteins, ubiquinone, and cytochromes; for anaerobic respiration, sequences of reductases for fumarate, nitrate, nitrite and sulfite are present (Table 6). Nitrate reduction has been observed experimentally [24]. In addition there are 10 members of the short-chain oxidoreductase family and 8 oxidoreductases identified by domain. Any one of these could be either a primary dehydrogenase or a terminal reductase not yet characterized (Table 7). The varieties of oxidants that appear to be used by *P. ingrahamii* suggest that its capabilities in this regard are comparable to that of *Shewanella* species.

Enzymes are present for pathways of utilization of both carbohydrates and amino acids as carbon and energy sources. Glycerol was the carbon source provided in the medium for the below-freezing temperature growth experiments [8]. In agreement, genes for glycerol uptake (genes 3166, 3169), glycerol kinase (3168) and dehydrogenase (3207) are present.

Carbon sources besides glucose that were found experimentally to be utilized by *P. ingrahamii* [24] were compared with the list of enzyme orthologs. One finds in *P. ingrahamii* genes for enzymes for utilization of fructose (971, 3552), galactose (2016, 2017), mannitol (89), N-acetylglucosamine (488–490), ribose (344) and sucrose (974), in agreement with experimental findings. We also found genetic evidence for the capability to utilize lactose (2019) and glucuronate (130,131, 132 and 136).

However comparing to *E. coli* enzymes, many of the sequences of enzymes for utilization of other carbohydrates are not present in *P. ingrahamii*. These include enzymes for utilization of arabinose, xylose, sorbitol or galactitol. Again the absence is in agreement with experimental observations [24]. Also missing are orthologs of *E. coli* enzymes for utilization of tagatose, fucose, rhamnose, glucarate, galactarate, altronate or idonate. These carbon sources have not yet been tested in culture. This picture is similar to the one found in *S. oneidensis*, a relatively poor capacity to utilize a variety of monomeric carbohydrates [25].

Although capability for utilization of carbohydrates is restricted, by contrast *P. ingrahamii* does have sequences for enzymes for utilization of most amino acids, and many of these have been verified experimentally. Transaminases convert to the corresponding keto acids, decarboxylases to amines. The amino group can be removed with dehydratases or lyases.

Fatty acids, breakdown

The four major enzymes of degradation of fatty acids are present, capable of generating acetyl-CoA for general metabolism (Table 7).

Fatty acids, biosynthesis

It has long been known that proteobacteria at low temperatures adjust fatty acid composition to the more flexible unsaturated and/or branched types. *P. ingrahamii* has genes for enzymes of fatty acid biosynthesis (Table 7). Close similarity is found for all *E. coli* enzymes starting with malonyl-CoA, proceeding via malonyl-ACP through the fatty acid biosynthesis cycle, each cycle adding 2-carbon moieties. At the 10-carbon level, the pathway of unsaturated fatty acids commences. Orthologs for all *E. coli* enzymes of the unsaturation pathway are present except for the last enzyme, FabI. Final steps for synthesis of unsaturated fatty acids by *P. ingrahamii* must differ from those in *E. coli*. The principal fatty acids that were detected in the organism experimentally are the saturated 16:0 (18.7%) and the singly unsaturated 16:1ω7c (67.0%) [24].

P. ingrahamii has sequences of the four subunits of a polyunsaturase closely similar to the polyunsaturase in psychrophile *C. psychrerythraea* 34H (Table 8). It is reasonable to suppose that at the very low temperatures of the environment *P. ingrahamii* might require fatty acids with more than one unsaturated bond. Also branched chain fatty acids are often found at cold temperatures. Based on enzyme sequences, precursors for synthesis of the branched-chain fatty acids could be generated as intermediates in breakdown of the amino acids leucine, isoleucine and valine. However, neither polyunsaturated nor branched chain fatty acids were detected in *P. ingrahamii* cultures grown at about 6 to 8°C (i.e. refrigerator) conditions [25]. Perhaps the polyunsaturated fatty acids are produced at lower temperatures or were not detected by the fatty acid analysis procedure used.

Interestingly, two polyunsaturated fatty acids, 20:5 (eicosapentaenoic acid) and 22:6 (docosahexaenoic acid), have

Table 6: Genes and enzymes of glucose and energy metabolism

Pathway	Gene	Enzyme
Glycolysis		
	591	6-phosphofructokinase
	1316	6-phosphofructokinase
	669	enolase
	3617	fructose-1,6-bisphosphatase, class II
	2682	fructose-bisphosphate aldolase
	372	fructose-bisphosphate aldolase, class II
	2359	glucokinase
	2382	glucokinase
	2932	glucokinase, thermoresistant
	324	glucose-6-phosphate isomerase
	2004	glyceraldehyde-3-phosphate dehydrogenase,
	2367	glyceraldehyde-3-phosphate dehydrogenase,
	3636	glyceraldehyde-3-phosphate dehydrogenase,
	879	phosphoglucomutase
	769	phosphoglucomutase
	371	phosphoglycerate kinase
	249	phosphoglycerate mutase
	2320	phosphoglycerate mutase
	3211	phosphoglycerate mutase, cofactor-independent
	2361	pyruvate kinase
	2879	pyruvate kinase
	2199	triose-phosphate isomerase
Metabolic connections		
	3617	fructose-1,6-bisphosphatase, class II
	93	phosphoenolpyruvate carboxykinase
	537	malic enzyme
	304	isocitrate lyase and phosphorylmutase
	303	malate synthase
Pentose pathway		
	2752	6-phosphogluconate dehydrogenase, decarboxylating
	2937	6-phosphogluconate dehydrogenase, decarboxylating
	2753	6-phosphogluconolactonase
	167	ribulose-phosphate 3-epimerase
	2754	glucose-6-phosphate 1-dehydrogenase
	3554	ribose 5-phosphate isomerase
	601	ribose 5-phosphate isomerase
	2054	transaldolase
	86	transaldolase B
	339	transketolase
	3086	glucose dehydrogenase
	2936	gluconate kinase
Pyruvate dehydrogenase		
	2782	pyruvate dehydrogenase complex, E1 beta subunit
	3602	pyruvate dehydrogenase complex, E1 beta subunit
	3601	pyruvate dehydrogenase complex, E1 acetate transfer subunit
	3603	pyruvate dehydrogenase complex, E2 subunit
	2779	dihydrolipoamide dehydrogenase E3 subunit
	2780	dihydrolipoamide dehydrogenase E3 subunit
	2925	dihydrolipoamide dehydrogenase E3 subunit
Tricarboxylic acid cycle		
	2927	2-oxo-acid dehydrogenase E1 subunit
	2252	2-oxoglutarate dehydrogenase, E1 subunit
	2926	2-oxoglutarate dehydrogenase E2 subunit
	2251	2-oxoglutarate dehydrogenase, E2 subunit
	2899	aconitate hydratase
	2120	aconitate hydratase 1
	800	adenylyl-sulfate kinase
	2257	citrate synthase I
	2617	citrate synthase I
	1738	fumarate hydratase
	1977	fumarate hydratase
	983	isocitrate dehydrogenase, NADP-dependent
	297	malate dehydrogenase, NAD-dependent

Table 6: Genes and enzymes of glucose and energy metabolism *(Continued)*

	3376	oxaloacetate decarboxylase alpha subunit
	3375	oxaloacetate decarboxylase, beta subunit
	2253	succinate dehydrogenase catalytic subunit SdhB
	2254	succinate dehydrogenase, flavoprotein subunit SdhA
	2256	succinate dehydrogenase, cytochrome b-binding subunit sdhC
	2255	succinate dehydrogenase, cytochrome b-binding subunit sdhD
	2249	succinyl-CoA synthetase, alpha subunit
	2250	succinyl-CoA synthetase, beta subunit
Anaerobic respiration		
	3279	fumarate reductase iron-sulfur subunit
	3281	fumarate reductase, D subunit
	3278	fumarate reductase, flavoprotein subunit
	3280	fumarate reductase, subunit C
	2175	nitrate reductase accessory periplasmic protein NapD
	2172	nitrate reductase periplasmic cytochrome c-type protein NapC
	2173	nitrate reductase periplasmic cytochrome c-type subunit NapB
	2174	nitrate reductase, periplasmic large subunit
	1024	nitrite reductase [NAD(P)H], large subunit
	1023	nitrite reductase [NAD(P)H], small subunit
	3435	sulfite reductase (NADPH) hemoprotein, beta-component
	3434	sulfite reductase [NADPH] flavoprotein, alpha chain
	3436	phosphoadenosine phosphosulfate reductase (PAPS reductase)
Oxidoreductases of unknown substrate		
	45	short-chain dehydrogenase/reductase SDR
	223	short-chain dehydrogenase/reductase SDR
	951	short-chain dehydrogenase/reductase SDR
	989	short-chain dehydrogenase/reductase SDR
	1000	short-chain dehydrogenase/reductase SDR
	1973	short-chain dehydrogenase/reductase SDR
	2106	short-chain dehydrogenase/reductase SDR
	2109	short-chain dehydrogenase/reductase SDR
	2778	short-chain dehydrogenase/reductase SDR
	3154	short-chain dehydrogenase/reductase SDR
	272	oxidoreductase FAD/NAD(P)-binding domain protein
	3187	oxidoreductase FAD/NAD(P)-binding domain protein
	2122	oxidoreductase alpha (molybdopterin) subunit
	1244	oxidoreductase domain protein
	1809	oxidoreductase domain protein
	2666	oxidoreductase domain protein
	3535	oxidoreductase domain protein
	576	oxidoreductase, molybdopterin binding
Fermentation indications		
	91	fermentative D-lactate dehydrogenase
	2123	formate dehydrogenase, subunit FdhD
	1217	hydrogenase, NADP-reducing subunit C
		many alcohol dehydrogenases

been reported from two other species of *Psychromonas*, *P. kaikoae* and *P. marina*[25].

Glycogen storage

There are genes for 6 glucose-1-phosphate adenyltransferases (299, 1296, 2063, 3033,3034 and 3464) any one of which could serve for the first step in synthesis of glycogen, and there are 2 glycogen/starch synthases (2348 and 3035). One or more of the 15 glycosyl transferases could be involved in synthetic reactions.

Digestion of macromolecules

Like other marine organisms, *P. ingrahamii* appears to have the capability of utilizing macromolecules in the environment for nutrition and energy. *P. ingrahamii* has genes for a relatively large number of 48 peptidases and proteases (Table 8). Some of these are no doubt required for internal turnover, but some seem likely to be exported out of the cell in order to hydrolyze environmental proteins, thus providing small molecular weight nutrients for uptake. *P. ingrahamii* has a complete general secretion system capable of excreting such degradative enzymes to the environment. To take up the digestion products of proteolysis, there are ABC-type transporters for peptides, many for amino acids. Not consistent with this prediction is the experimental observation that gelatin is not hydrolysed [24].

Storage glycogen as well as external polysaccharides including starch could be hydrozysed for production of

Table 7: Metabolism of fatty acids

Pathway	Gene	Enzyme
Degradation	3600	acetyl-CoA synthetase
	2603	acyl-CoA dehydrogenase domain protein
	1208	enoyl-CoA hydratase/isomerase
	2604	fused 3-hydroxyacyl-CoA dehydrogenase, NAD-binding and enoyl-CoA hydratase/isomerase
	2401	acyl-CoA thiolase (acetyl-CoA transferase)
Synthesis	1090	acyl carrier protein
	1088	malonyl CoA-acyl carrier protein transacylase
	1995	3-oxoacyl-(acyl-carrier-protein) synthase I
	1087	3-oxoacyl-(acyl-carrier-protein) synthase III
	1997	3-oxoacyl-[acyl-carrier-protein) synthase III
	1091	beta-ketoacyl synthase
	1982	beta-hydroxyacyl-(acyl-carrier-protein) dehydratase
	1089	3-oxoacyl-(acyl-carrier-protein) reductase
	188	lauroyl (or palmitoleoyl)-ACP acyltransferase
	2965	(3R)-hydroxymyristoyl-ACP dehydratase
Unsaturation	1684	polyunsaturated fatty acid synthase
	1685	polyunsaturated fatty acid synthase
	1686	polyunsaturated fatty acid synthase
	1687	polyunsaturated fatty acid synthase

sugars to supply energy. For utilization of some polysaccharides there are amylases, glucosidases, debranching enzymes, and glycosyl hydrolases (Table 8), some of which may be intracellular and others extracellular. There are 7 lytic transglycosylases which, in cleaving peptidoglycan links could be involved in modellingof the cell or could break up environmental cell wall fragments. Capability to hydrolyse fats also exists as there are 3 lipases, 5 phospholipases and a lyso-lipase. There are many enzymes hydrolyzing nucleic acids with different specificities and functions, many with vital internal metabolic roles. In addition, some may be used to hydrolyze external nucleic acid debris (Table 8).

Other hydrolases are encoded in the genome whose physiological roles are not currently known, for example there are 11 HAD-superfamily hydrolases, 7 alpha/beta hydrolase fold proteins, 5 metal-dependent phosphohydrolases.

Chaperones and stress proteins

Multiple chaperone proteins are encoded in *P. ingrahamii*, suggesting that folding of proteins is an important process (Table 9). There are 4 proteins like DnaK, 4 like DnaJ, 3 GroEL monomers and 2 GroES monomers There are 12 peptidylprolyl isomerases (trigger factors that act at nascent polypeptide chains), and a ClpB protein disaggregating complex. One can speculate that the the role of the chaperones is to guide nascent polypeptides into functional three-dimensional configurations permitting activity at low temperatures. Future characterization of some of these chaperones could reveal what kinds of folding are required to retain protein function at sub-zero temperatures. Ferrer *et al*. [26] found that GroEL of *Oleispira antartica* RB8 functioned as a single ring of 7 units at cold temperatures, but as a double ring of 7 over 7 at warm temperatures. They pinpointed two residues as critical to the transition from double to single ring. However on inspection these residues do not occur at the comparable positions in *P. ingrahamii* GroEL proteins. Actions of *P. ingrahamii* chaperones at cold temperatures remain to be explored.

P. ingrahamii has genes for a variety of known types of stress proteins: There are 12 cold shock proteins, 9 heat shock proteins, 7 UspA-type stress proteins as well as 9 "tellurite resistance" proteins now known to protect against superoxide formation [27] (Table 9). Experimental work will be needed to determine which if any of these have functions directed specifically at living in cold temperatures and if there are other types of stress proteins or any other cold-associated functions among the open reading frames of unknown function.

Transporters

Compared to *E. coli*, *P. ingrahamii* has few transporters specific for sugars and sugar alcohols, but does have many transporters for amino acids (see Additional file 2). In this respect, it seems that *P. ingrahamii* is like *S. oneidensis* and other *Alteromondales* in a capacity to utilize environmental amino acids as carbon (and nitrogen) sources, contrasted to less capacity for sugars.

Table 8: Enzymes of macromolecule hydrolysis

Macromolecule	Gene	Enzyme
Peptides, protein	3027	D-alanyl-D-alanine carboxypeptidase, serine-type
	3293	D-alanyl-D-alanine carboxypeptidase/D-alanyl-D-alanine-endopeptidase
	314	O-sialoglycoprotein endopeptidase
	3325	prepilin peptidase type 4
	894	aminoacyl-histidine dipeptidase
	1331	aminopeptidase
	2545	aminopeptidase M24
	2344	aminopeptidase N
	1777	aspartyl aminopeptidase
	1034	carboxypeptidase thermostable
	2765	deacylase/carboxypeptidase family member, Zn-dependent
	409	dipeptidase
	395	leucyl aminopeptidase
	2026	metallopeptidase M24 family
	3003	methionine aminopeptidase, type I
	212	oligopeptidase A
	633	oligopeptidase B
	2690	peptidase
	1673	peptidase C1A, papain
	2671	peptidase C26
	268	peptidase M14, carboxypeptidase A
	2457	peptidase M14, carboxypeptidase A
	2444	peptidase M15B
	3480	peptidase M16 domain protein
	2127	peptidase M19, renal dipeptidase
	2301	peptidase M22, glycoprotease
	1544	peptidase M23B
	3212	peptidase M23B
	3225	peptidase M23B
	676	peptidase M23B, peptidoglycan-binding
	2784	peptidase M24
	1800	peptidase M48, Ste24p
	926	peptidase M48, Ste24p, Zn-dependent, TPR repeats
	686	peptidase M50
	3180	peptidase M50
	1350	peptidase M56, BlaR1
	2144	peptidase M6, immune inhibitor A
	2723	peptidase S1 and S6, chymotrypsin/Hap
	994	peptidase S16, LON domain protein
	925	peptidase S49
	1972	peptidase S49, N-terminal domain protein
	1301	peptidase U32
	2189	peptidase U32 family
	2460	peptidase dimerization domain protein
	410	peptidase domain protein
	478	prepilin peptidase dependent protein D
	606	proline aminopeptidase P II
	253	proline iminopeptidase
Polysaccharides	1954	alpha amylase, catalytic region
	3067	predicted glucoamylase I (alpha-1,4-glucan glucosidase)
	2381	alpha-D-1,4-glucosidase
	2383	dextran glucosidase
	554	glucan endo-1,3-beta-D-glucosidase
	893	glycogen debranching enzyme
	2363	glycogen debranching enzyme
	3070	glycogen debranching enzyme
	558	glycoside hydrolase family
	2014	glycoside hydrolase family
	2529	glycoside hydrolase family
	2841	glycoside hydrolase family
Murein	1782	gamma-D-glutamate-meso-diaminopimelate muropeptidase
	1075	lytic murein transglycosylase
	499	lytic murein transglycosylase, catalytic
	293	lytic transglycosylase, catalytic protein
	3319	lytic transglycosylase, catalytic protein
	293	lytic transglycosylase, catalytic protein

Table 8: Enzymes of macromolecule hydrolysis *(Continued)*

	367	lytic transglycosylase, catalytic protein
	3319	lytic transglycosylase, catalytic protein
Lipids	1779	esterase/lipase/thioesterase family protein
	1892	lipase, class 3
	2631	lipase-like
	2470	lipase/acylhydrolase family protein, GDSL-lilke
	258	phospholipase
	3493	phospholipase A(1)
	2455	phospholipase D/transphosphatidylase
	1334	phospholipase family, patatin-like protein
	1844	phospholipase family, patatin-like protein
	3290	predicted lysophospholipase
Nucleic acids	2776	5'-3' exonuclease
	201	ATP-dependent endonuclease of the OLD family
	501	DNA mismatch repair endonuclease mutH
	2451	HNH endonuclease
	317	TatD-related deoxyribonuclease
	807	TatD-related deoxyribonuclease
	716	crossover junction endodeoxyribonuclease
	1096	deoxyribonuclease of TatD family
	732	endonuclease III
	1020	endonuclease/exonuclease/phosphatase
	1518	endonuclease/exonuclease/phosphatase
	2456	endonuclease/exonuclease/phosphatase
	2694	endonuclease/exonuclease/phosphatase
	3327	endonuclease/exonuclease/phosphatase
	416	endoribonuclease L-PSP
	2129	endoribonuclease L-PSP
	2646	endoribonuclease L-PSP
	369	excinuclease ABC, A subunit
	2085	excinuclease ABC, A subunit
	1082	excinuclease ABC, B subunit
	1193	excinuclease ABC, C subunit
	2436	exodeoxyribonuclease I
	1319	exodeoxyribonuclease III
	1460	exodeoxyribonuclease V, alpha subunit
	1459	exodeoxyribonuclease V, beta subunit
	1458	exodeoxyribonuclease V, gamma subunit
	2951	exodeoxyribonuclease VII, large subunit
	2238	exodeoxyribonuclease VII, small subunit
	1303	exonuclease SMC domain protein
	1302	exonuclease SbcCD, D subunit
	2586	exonuclease of the beta-lactamase domain protein
	2270	exoribonuclease II
	2580	extracellular deoxyribonuclease
	53	formamidopyrimidine-DNA glycosylase
	2029	predicted endoribonuclease L-PSP
	2233	predicted exonuclease
	3302	single-stranded-DNA-specific exonuclease
	1283	uracil-DNA glycosylase
	1819	predicted ribonuclease BN
	3482	predicted ribonuclease BN
	1668	ribonuclease D
	496	ribonuclease H
	2962	ribonuclease HII
	640	ribonuclease III
	3609	ribonuclease P protein component
	3479	ribonuclease PH
	3417	ribonuclease R
	2450	ribonuclease T
	1126	ribonuclease, Rne/Rng family
	2208	ribonuclease, Rne/Rng family
	2214	tRNA-guanine transglycosylase
	2567	nuclease (SNase domain protein)
	2838	nuclease (SNase domain protein)
	265	exonuclease, RNase T and DNA polymerase III
	968	exonuclease, RNase T and DNA polymerase III
	3335	exonuclease, RNase T and DNA polymerase III

Table 9: Chaperones and stress proteins

Category	Gene numbers	Protein
Chaperones	917, 1232, 1233, 1328	DnaK-like
	918, 1039, 2621, 2499	DnaJ-like
	843, 2494	GroES
	844, 2493, 2791	GroEL
	919, 1049, 1080, 1469, 1619, 1856, 1917, 2185, 3116, 3199, 3257, 3269	Peptidyl prolyl isomerases (trigger factors)
	1040, 3623	ClpB disaggregator
Stress proteins	279, 755, 1097, 1881, 1953, 2158, 2543, 2698,2701, 3095, 3098, 3704	Cold Shock
	3, 95, 202, 956, 1039, 1051, 1246, 1533, 1806, 1916, 2499, 2692	Heat shock
	125, 930, 954, 955, 959, 1234, 2734	Universal stress proteins
	378, 1265, 1264, 1265, 1266, 2005, 2574, 2575	Tellurite resistance (anti_superoxide)

As to types of transporters, the ABC type of ATP-driven multisubunit transporter is most common in the *P. ingrahamii* genome, as is the case for other bacteria. Also, as for many other bacteria, conventional secondary transporters follow in frequency. However, *P. ingrahamii* differs from many in the fact that in there are 11 sets of the *dct*M, *dct*P and *dct*Q genes for the tripartite ATP-independent periplasmic transporter systems (TRAP) [28,29] (Table 10), more than are found in the three mesophiles whose whole genomes were compared (*E. coli*, *S. oneidensis* and *V. cholerae*). TRAP systems specialize in transport of C4-dicarboxylic organic acids such as fumarate, perhaps for anaerobic respiration purposes. The number of 3-gene TRAP systems in bacteria is variable. *E. coli* has one, *Pseudomonas aeruginosa* has 6 whereas *P. ingrahamii* has 11 of these three-protein systems. Recently 15 TRAP systems have been identified in *Sinorhizobium meliloti* 1021 [30]. TRAP transporters use a proton motive force energized system, simpler and possibly more primitive than the ATP-utilizing ABC systems. A connection between the TRAP transporters and low temperature growth is not currently known.

Regulators

Many types of regulation mechanisms have been identified in *P. ingrahamii*: transcriptional activators and repressors, cyclic-AMP regulation, chemotaxis systems, two-component sensor-response regulators of several types including the twin-arginine translocation (Tat) pathway signal sequence domain proteins, also synthesis/breakdown of cyclic-diGMP signalling second messengers associated with GGDEF and/or EAL domains (see Additional file 2). There are 61 regulators of the cyclic-diGMP signalling second messenger in the genome, compared to 29 in *E. coli* K-12 MG1655. Cyclic-diGMP concentrations are controlled by either a diguanylate cyclase or a specific phosphodiesterase or both, together governing synthesis or hydrolysis of the cyclic-GMP. The types of genes and corresponding physiology regulated by cyclic diguanylate systems so far identified are motility, adhesion factors, fimbriae and biofilm formation.

Given the finding of a large number of cyclic-diGMP signalling systems, we might guess that *P. ingrahamii* lives within the matrix of a biofilm. Altogether 16 glycosyl transferases are encoded (by genes 326, 327, 328, 329,

Table 10: Tripartite ATP-independent C4-dicarboxylate transporters

DctM-like	DctQ-like	DctP-like
IIM* subunit	IIM* subunit	Periplasmic subunit
Gene	Gene	Gene
133	134	135
538	539	540
572	571	570
646	647	648
710	709	708
2033	2034	2035
2595	2595	2594
2935	2934	2933
3148	3149	3150
3544	3545	3546
3673	3674	3675

*IIM = integral inner membrane

331, 336, 440, 449, 454, 456, 457, 779, 792, 1794, 3458, and 3647), suggesting that polysaccharide, perhaps exopolysaccharide, is a major synthetic product. For export, many efflux proteins are present. An ortholog is present (1200) of the quorum sensing HapR (LuxR) regulator that is involved in controlling biofilm formation in *V. cholerae* [31,32]. An extracellular matrix, a major physiological feature of the related *S. oneidensis* [33], could well be important to life in the cold, providing stability and resilience to the population. Since *P. ingrahamii* lives in sea ice, it is possible that extracellular polysaccharide (EPS) may be part of the sea ice microbial community (SIMCO) biofilm although it should be noted that *P. ingrahamii* was isolated from the ice column above the major biofilm of the SIMCO.

Alternatively, the production of EPS may serve a role in sequestering water from ambient saltwater at lower temperatures or actually lowering the freezing point. In this regard it is interesting to note that in the -12°C growth experiments, none of the tubes froze after the first week of growth following inoculation suggesting that a product, possibly EPS, produced by the bacterium may have lowered the freezing point of the growth medium.

Regulation of expression of certain classes of genes is moderated by the sigma factors of the RNA polymerase holoenzyme. *P. ingrahamii* is well endowed, appearing to have genes for sigma factors 24 (RpoE), 32 (RpoH), 38 (RpoS), 54 (RpoN) and 70(RpoD) [gene numbers are, respectively, 65, 626, 677, (424, 712, 2892 and 3175 for sigma-54) and (310, 946 and 995 for sigma-70)]. RpoE and RpoH are both stress-responding factors. RpoS operates in stationary phase and also under stress. RpoN is concerned with nitrogen metabolism and in *V. cholerae* regulates flagellin gene transcription. RpoD for sigma 70 is the primary factor in *E. coli*.

Ribosomes

There are 58 ribosomal proteins annotated, 5 of which are duplicated, therefore 53 unique. Of these, 38 were found as orthologs with Pam values less than 40 among the two sets of genomes we examined. Organisms with closest matches were, in descending order, *V. cholerae*, *V. parahaemolyticus*, *S. oneidensis*, *H. influenzae* = *I. loihiensis*, *C. psychrerythraea*. Again, the close relationship of *P. ingrahamii* with Vibrio species is evident.

Osmotic stability

P. ingrahamii grows well over the range 1 to 12% NaCl [25]. To manage the potential for osmotic imbalance, the genes for enzymes to synthesize the osmolyte glycine betaine from choline are present (genes 2071, 2072), as well as a transporter to take up choline (gene 969) or, bypassing synthesis, specific ABC transport systems for uptake of glycine betaine are present, (genes 614–616, 2073–2075). This capability may explain how the organism is able to survive and perhaps grow in the salt pockets that are formed within the sea ice.

Motility

There is a large cluster of flagellar genes in *P. ingrahamii* in the region between gene numbers 3562 to 3598 and four copies of sigma factor rpoN which is known in *V. cholerae* to regulate flagellin genes. Yet the bacteria in culture have been observed to be non-motile [8]. There may be a defect in one of the essential flagellar proteins or in the expression or assembly processes. Alternatively, the organism may not always express flagella formation and motility. Interestingly, the original description of the genus indicates that other members of the genus are motile [34].

Gas vesicles

Two kinds of gas vesicles have been observed in the cell under culture (24). Genes orthologous to known gas vesicle genes are present in two clusters, the ranges 1248 to 1262 and 1748 and 1750. Although the two types can be expressed simultaneously in cells, they may be differentially expressed under preferential conditions in the environment.

Different kinds of enzyme orthologs in different bacterial relatives

Of the 53 bacteria chosen to represent the breadth of the currently characterized eubacterial world, the three bacteria bearing the largest number of protein sequences scored as "best matches" with *P. ingrahamii* proteins are *V. cholerae*, *S. oneidensis* and *C. psychrerythraea* 34H (Table 3).

About 3/4 of the *P. ingrahamii* proteins annotated as enzymes, transporters or regulators have orthologs at the level of Pam <= 125 in either *V. cholerae*, *S. oneidensis* or *C. psychroerythraea*. By contrast, only 19% of *P. ingrahamii* CDSs of unknown function have orthologs in these bacteria. That the vast majority, 4/5, of *P. ingrahamii* CDSs of unknown function do not have good matches in the most closely related organisms suggests that many proteins in *P. ingrahamii* are qualitatively different. New, unique functions will not be revealed by annotations that depend on similarity to known proteins, but their existence suggests that further experimental study of this extremophile is worthwhile and could bring new biology to light.

Looking within the enzyme category, it is striking that most of the best hit homologs of enzymes of central metabolism are proteins of *V. cholerae*. This is particularly puzzling since *V. cholerae* is a mesophile and it is phylogenetically in a different Order and Family from *P. ingrahamii*. Could this be an example of massive horizontal transfer of genes from *Vibrio* to *Psychromonas*? It seems unlikely since the many genes for central metabolism are

not in a few clusters similar to pathogenicity islands, rather they map throughout the chromosome, and the majority of the transposase-type genes have closest orthologs in *S. oneidensis*. Nevertheless, in spite of expectations, we have found that the sequences of enzymes of basic metabolism that function at low temperature are more similar to those of a mesophile than they are to those of another psychrophile, and even more surprising, a mesophile of a different taxonomic Order and Family. A feature that may suggest global genome reorganization and account for an apparent scrambling of otherwise conserved genes is that the Vibrios usually comprise two chromosomes, in contrast to *P. ingrahamii* which has only one.

Different from *V. cholerae*, one sees that enzyme homologs in *S. oneidensis* and *C. psychrerythraea* 34H are largely the enzymes of peripheral and macromolecule metabolism, not the enzymes of central metabolism.

We narrowed our question, looking for a commonality in enzymes among the psychrophiles. We identified the 566 proteins in *C. psychrerythraea* and *I. hoiliensis* that have highest similarity to *P. ingrahamii* proteins at Pam values =< 125. In this set, there are few enzymes important to central metabolism; many more are involved in macromolecule synthesis and maintenance. There are enzymes dealing with nucleic acids: exo and endo-nucleases, RNA helicases, DNA polymerase and helicases, transcriptional and translational factors, and recombinases. Psychrophily may require differences in the proteins of nucleic acid metabolism since even though the salt content of the sea ice and the low GC content of the DNA (40.1%) should act to lower the melting point of double stranded polynucleotides. However, handling nucleic acid structures at such cold temperatures may also require differences in the interacting proteins. Similarly, transcription and translation factors of a particular kind may be required to enable the processes at low temperatures.

Unique unknown proteins

We identified a few small groups of proteins that could perform novel functions related to psychrophily, warranting further investigation in the future. A contiguous set of nine genes that could comprise an operon, gene numbers 3053 through 3061 have no orthologs in current databases with Pam value <150. Four of the proteins are paralogous, with similar sequences among themselves. As we are looking for functions in psychrophiles not found in other bacteria, experimental characterization of this group of proteins might be worthwhile.

There are 3 groups of genes that could be contiguous operons whose members reside either in cluster 4 of the Correspondence Analysis (*vide supra*), or are unknown in other bacteria to date, or both. These are clusters 1672 through 1676, 1960 through 1964, and 2315 through 2319. Could any or all of these relate to the ability to live and grow at -12C or lower temperatures? We identify these seemingly unique sets of proteins with the thought that when results of proteomic expression experiments on this organism are available, it might be useful to characterize the proteins and to know under what circumstances they are synthesized.

Conclusion

The *P. ingrahamii* genome, although it is 25% smaller than the genome of *E. coli* K-12 MG1655 and has many more CDSs labeled hypothetical, is similar to *E. coli* in the distribution of functions among annotated proteins, with enzymes by far the largest category followed by transporters, then regulators. Other categories follow distantly. Similar to *E. coli*, the largest paralogous groups within the genome were transporters and regulators, the smaller ones enzymes. Although transporters and regulators are fewer in number than enzymes, they form larger paralogous groups. They belong to fewer families employing fewer mechanisms than is the case for enzymes which belong to smaller families of greater variety.

Unexpectedly, *P. ingraham*ii protein sequences were similar to more *V. cholerae* proteins than to proteins from *S. oneidensis* or *C. psychrerythraea*. This is in spite of the fact that *Vibri*o species belong to a different Family and Order than do Psychromonadaceae, Shewanellaceae and Colwelliaceae, and in spite of the fact that *C. psychrerythraea* is also a psychrophile.

In a comparison of gross properties of all the proteins of *P. ingrahamii* differences were not found from those of three mesophiles in several respects: distribution of lengths, total amino acid composition, codon usage when corrected for genomic GC content.

However, correspondence analysis (CA) of the amino acid content of the proteins showed that they cluster in ways that differ from most other bacteria, falling into more clusters that are not as well separated from one another as in other bacteria. This may be a consequence of an unusual distribution of hydrophobicity. One of the clusters is composed almost half of unidentified, unknown CDSs. *P. ingrahamii* contains proteins with relatively high asparagine content, and low content of amino acids potentially sensitive to the higher concentration of oxygn present in cold waters. These properties would seem to be appropriate for an extreme psychrophile.

As to metabolism, *P. ingrahamii* is a facultative anaerobe capable of both fermentation and respiration. In agreement, proteins similar to the corresponding metabolic pathways and proteins for synthesis and harvest of glyco-

gen storage compound were identified. Most enzymes of central small molecule metabolism were most closely related to those of *V. cholerae*. Enzymes of macromolecular synthesis and maintenance are most closely related to those of *S. oneidensis* and *C. psychrerythraea*. There seems to be a preference of amino acids over sugars as sources of carbon and energy, perhaps harvested by extracellular degradation of environmental proteins by exported peptidases. To create a more flexible lipid layer adjusted to cold temperatures, sequences of a heteropolymeric polyunsaturase were found. However, the enzyme may not always be active as no polyunsaturated fatty acids were detected in culture. Perhaps, however, in nature they are expressed at temperatures lower than the lowest normally used for laboratory cultivation, 6–8°C.

In addition to transporters of types common to other proteobacteria were 11 three-component TRAP systems for C4-dicarboxylic acid transport. In addition to regulators of types common to other proteobacteria were 61 regulators of cyclic-diGMP second messengers, suggesting that biofilms and their regulation could be an important part of the life style of this psychrophile. Alternatively, *P. ingrahamii* may produce EPS to lower the freezing point in the surrounding environment, thereby making water available for low temperature growth.

Looking for particular proteins that might be unique to this psychrophile but are not yet in the genomic databases we searched, we have pointed out some sets of gene products that could be starting points for discovering new functions or new types of proteins. We noticed 9 contiguous genes that are all unknown hypotheticals, 4 of them sequence-related to one another. We also noted three apparent contiguous operons, members of which are unknown hypotheticals and/or members of the unique set of proteins in cluster 4. In future, we recommend that proteomic experiments should be used to explore the shift to cold temperature with a view of keeping an eye out for expression of any of these proteins, potentially part of the mechanisms of life at very cold temperatures.

Methods

Genome sequencing and finishing

The genome of *Psychromonas ingrahamii* was sequenced at the Joint Genome Institute Lawrence Laboratories, Walnut Creek, CA using a combination of (4 kb, 6.8 kb and 36 kb) DNA libraries. All general aspects of library construction and sequencing performed can be found on line [35]. Draft assemblies were based on 53473 total reads. All three libraries provided 11X coverage of the genome. For closing and finishing, the Phred/Phrap/Consed software package [36,37] was used for sequence assembly and quality assessment [38,39]. After the shotgun stage, 53473 reads were assembled with parallel phrap (High Performance Software, LLC). Possible mis-assemblies were corrected with Dupfinisher [41] or transposon bombing of bridging clones (Epicentre Biotechnologies, Madison, WI). Gaps between contigs were closed by editing in Consed, custom primer walks, or PCR amplification. A total of 952 primer walk reactions and 3 transposon bombs were necessary to close gaps, to resolve repetitive regions, and to raise the quality of the finished sequence. The completed genome sequences of *Psychromonas ingrahamii* contain 53,592 reads, achieving an average of 11-fold sequence coverage per base with an error rate less than 1 in 100,000. The sequence of *P. ingrahamii* can be accessed using the GenBank accession number CP000510.1.

Annotation

Annotation was initially carried out at Oak Ridge National Laboratory using methods detailed in [41,42]. Further annotation and analysis of protein sequence similarities used the Darwin system (Data Analysis and Retrieval With Indexed Nucleotide/peptide sequence package), version 2.0, developed at the ETHZ in Zurich, Switzerland [16,22]. Pairwise sequence alignments and scores were generated using the AllAllDb program of Darwin. Maximum likelihood alignments are generated with an initial global alignment by dynamic programming (Smith and Waterman algorithm) followed by dynamic local alignments (Needleman and Wunsch algorithm). A single scoring matrix is used for these steps. After the initial alignment, the scoring matrix is adjusted to fit the approximate distance between each protein pair to produce the minimum Pam value. Pam units are defined as the numbers of point mutations (base pair differences) per 100 residues [17]. The Darwin system's ability to apply scoring matrices according to the distance between each protein pair ensures a data set of highly accurate similarity calculations for distantly as well as closely related protein pairs. The identification of distantly related proteins is valuable in finding divergent but related protein functions.

To extract homolog matches from initial data, we required that the alignments with *P. ingrahamii* proteins be at least 83 residues long [43], and that the alignment must represent over 40% of both proteins. Unless otherwise stated we required Pam scores to be 150 or less. To assemble groups of paralogs, the Pam threshold for pairs was raised to 250, and then pairs were grouped by a transitive process as previously described [23].

Correspondence Analysis

The amino acid composition of the proteins of *P. ingrahamii* was analyzed using correspondence analysis (CA) [12-14] with the FactoMineR R package [44]. Each protein was truncated for the first 10 and last 5 amino acid residues and each is represented by its normalized aminoacid

content. The representation is in a 19-dimension space where specific statistical distances are measured by the chi-square method. The 3 most informative dimensions are used to plot the position of each individual protein. The individual amino acid distributions are superimposed in this same space. The axes are numbered in order of the amount of information they carry. A bayseian method, as in Bailly-Bechet et al. [41], was used to cluster proteins of similar composition using the Mclust R package [45], and the best classification, which is associated with the largest BIC (Bayesian Information Criterion) value, was selected for further analysis.

Codon usage and selection bias were determined by methods of Sharp *et al*. [9].

Authors' contributions

MR used Darwin-generated data for further annotation and analysis and wrote the paper. JTS made biological determinations and made significant contributions to writing the paper. AD and TZW carried out and interpreted the corresponence analysis. AD made significant contributions to writing the paper. JST and TSB closed and finalized the sequence. MLL and LJH carried out the initial annotation and submitted to GenBank.

Additional material

Additional file 1

***Ping Correspodence Analysis**. Correspondence analysis of amino acid content of* Psychromonas ingrahamii *proteins. Gene number, Cluster number, Product*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-9-210-S1.xls]

Additional file 2

***Ping Annotations 2**. Further annotation of* Psychromonas ingrahamii *proteins. Gene number, Gi identifier, predicted protein product, type of product*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-9-210-S2.xls]

Acknowledgements

Thanks to Dr. Paul M. Sharp for carrying out the codon usage analyses. Thanks to Dr. Margrethe Serres for suggestions on reading the manuscript and, with Daniella Wilmot, for database and library support. Dr. John L. Ingraham generated the list of 53 representative bacteria. MR acknowledges support from DE-FG02-04ER63940. JTS acknowledges the support from the University of Washington NASA NAI program and the NSF Astrobiology IGERT program. TZW acknowledges support from a grant from the Fondation Fourmentin-Guilbert and AD acknowledges support from the European Union BioSapiens Network of Excellence, Grant LSHG CT-2003-503265. Sequencing and first round of annotation was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University ofCalifornia, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231, Los Alamos National Laboratory under Contract No. W-7405-ENG-36, and Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725.

References

1. Margesin R, Neuner G, Storey KB: **Cold-loving microbes, plants, and animals-fundamental and applied aspects.** *Naturwissenschaften* 2006, **94:**77-99.
2. Jakosky BM, Nealson KH, Bakermans C, Ley RE, Mellon MT: **Subfreezing activity of microorganisms and the potential habitability of Mars' polar regions.** *Astrobiology* 2003, **3:**343-50.
3. Marx J-C, Collins T, D'Amico S, Feller G, Gerday C: **Cold-adapted enzymes from marine antarctic microorganisms.** *Marine Biotech* 2006, **9:**293-304.
4. Siddiqui KS, Cavicchioli R: **Cold-adapted enzymes.** *Annu Rev Biochem* 2006, **75:**403-33.
5. Methe BA, Nelson KE, Deming JW, Momen B, Melamud E, Zhang X, Moult J, Madupu R, Nelson WC, Dodson RJ, Brinkac LM, Daugherty SC, Durkin AS, DeBoy RT, Kolonay JF, Sullivan SA, Zhou L, Davidsen TM, Wu M, Huston AL, Lewis M, Weaver B, Weidman JF, Khouri H, Utterback TR, Feldblyum TV, Fraser CM: **The psychrophilic lifestyle as revealed by the genome sequence of Colwellia psychrerythraea 34H through genomic and proteomic analyses.** *Proc Natl Acad Sci USA* 2005, **102:**10913-10918.
6. Hou S, Saw JH, Lee KS, Freitas TA, Belisle C, Kawarabayasi Y, Donachie SP, Pikina A, Galperin MY, Koonin EV, Makarova KS, Omelchenko MV, Sorokin A, Wolf YI, Li QX, Keum YS, Campbell S, Denery J, Aizawa S, Shibata S, Malahoff A, Alam M: **Genome sequence of the deep-sea gamma-proteobacterium Idiomarina loihiensis reveals amino acid fermentation as a source of carbon and energy.** *Proc Natl Acad Sci USA* 2004, **101:**18036-18041.
7. Medigue C, Krin E, Pascal G, Barbe V, Bernsel A, Bertin PN, Cheung F, Cruveiller S, D'Amico S, Duilio A, Fang G, Feller G, Ho C, Mangenot S, Marino G, Nilsson J, Parrilli E, Rocha EP, Rouy Z, Sekowska A, Tutino ML, Vallenet D, von Heijne G, Danchin A: **Coping with cold: the genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC125.** *Genome Res* 2005, **15:**1325-1335.
8. Breezee J, Cady N, Staley JT: **Subfreezing growth of the sea ice bacterium "Psychromonas ingrahamii".** *Microbial Ecology* 2004, **47:**300-304.
9. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE: **Variation in the strength of selected codon usage bias among bacteria.** *Nucleic Acids Res* 2005, **33:**1141-1153.
10. Lobry J, Chessel D: **Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria.** *J Appl Genet* 2003, **44:**235-261.
11. Lobry JR, Necsulea A: **Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes.** *Gene* 2006, **385:**128-136.
12. Hill MO: **Correspondence analysis: a neglected multivariate method.** *Appl Statist* 1974, **23:**340-354.
13. Bailly-Bechet , Danchin A, Iqbal M, Marsili M, Vergassola M: **Codon usage domains over bacterial chromosomes.** *PloS Computational Biology* 2006, **2:**e37.
14. Pascal G, Medigue C, Danchin A: **Persistent biases in the amino acid composition of prokaryotic proteins.** *Bioessays* 2006, **28:**726-738. renumber
15. **Computational Biology at Oak Ridge National Laboratory** [http://genome.ornl.gov/microbial/ping/]
16. Gonnet GH, Hallett MT, Korostensky C, Bernardin L: **Darwin v. 2.0: an interpreted computer language for the biosciences.** *Bioinformatics* 2000, **16:**101-103.
17. Schwartz RM, Dayhoff MO: *Atlas of Protein Sequence and Structure Volume 5*. Issue Suppl 3 Edited by: Dayhoff MO. Washington: National Medical Research Foundation; 1978:353-358.
18. **National Center for Biotechnology Information** [http://ncbi.nlm.nih.gov/genomes/]
19. Ivanova EP, Flavier S, Christen R: **Phylogenetic relationships among marine Alteromonas-like proteobacteria: emended description of the family Alteromonadaceae and proposal of**

Pseudoalteromonadaceae fam. nov., Colwelliaceae fam. nov., Shewanellaceae fam. nov., Moritellaceae fam. nov., Ferrimonadaceae fam. nov., Idiomarinaceae fam. nov. and Psychromonadaceae fam. nov. *Int J Syst Evol Microbiol* 2004, **54:**1773-88.
20. Bowman JP, McMeekin TA: **OrderX. Alteromonadales ord. nov.** In *Bergey's Manual of Systematic Bacteriology The Proteobacteria Volume 2*. 2nd edition. Edited by: Garrity GM. New York: Springer-Verlag; 2005:443-491.
21. Klappenbach JA, Dunbar JM, Schmidt TM: **rRNA operon copy number reflects ecological strategies of bacteria.** *Appl Environ Microbiol* 2000, **66:**1328-1333.
22. **Computational Biochemistry Research Group** [http://www.cbrg.ethz.ch/darwin]
23. Serres MH, Riley M: **Gene fusions and gene duplications: relevance to genomic annotation and functional analysis.** *BMC Genomics* 2005, **6:**33.
24. Auman AJ, Breezee JL, Gosink JJ, Kämpfer P, Staley JT: **Psychromonas ingrahamii sp. nov., a novel gas vacuolate, psychrophilic bacterium isolated from Arctic polar sea ice.** *Int J Syst Evol Microbiol* 2006, **56:**1001-1007.
25. Serres MH, Riley M: **Genomic analysis of carbon source metabolism of Shewanella oneidensis MR-1: Predictions versus experiments.** *J Bacteriol* 2006, **188:**4601-4609.
26. Ferrer M, Lunsdorf H, Chernikova TN, Yakimov M, Timmis K, Golyshin PN: **Functional consequences of single:double ring transitions in chaperonins: life in the cold.** *Mol Microbiol* 2004, **53:**167-182.
27. Perez JM, Calderon IL, Arenas FA, Fuentes DE, Pradenas GA, Fuentes EL, Sandoval JM, Castro ME, Elias AO, Vasquez CC: **Bacterial toxicity of potassium tellurite: unveiling an ancient enigma.** *PLoS ONE* 2007, **2:**e211.
28. Kelly DJ, Thomas GH: **The tripartite ATP-independent periplasmic (TRAP) transporters of bacteria and archaea.** *FEMS Microbiol Rev* 2001, **25:**405-424.
29. Janausch IG, Zientz E, Tran QH, Kroger A, Unden G: **C4-dicarboxylate carriers and sensors in bacteria.** *Biochim Biophys Acta* 2002, **1553:**39-56.
30. Mauchline TH, Fowler JE, East AK, Sartor AL, Zaheer R, Hosie AH, Poole PS, Finan TM: **Mapping the Sinorhizobium meliloti 1021 solute-binding protein-dependent transportome.** *Proc Natl Acad Sci USA* 2006, **47:**17933-8.
31. Hammer BK, Bassler BL: **Quorum sensing controls biofilm formation in Vibrio cholerae.** *Mol Microbiol* 2003, **50:**101-104. Erratum in: Mol Microbiol. 2004 51:1521
32. Mueller RS, McDougald D, Cusumano D, Sodhi N, Kjelleberg S, Azam F, Bartlett DH: **Vibrio cholerae strains possess multiple strategies for abiotic and biotic surface colonization.** *J Bacteriol* 2007, **189:**5348-60.
33. Thormann KM, Duttler S, Saville RM, Hyodo M, Shukla S, Hayakawa Y, Spormann AM: **Control of formation and cellular detachment from Shewanella oneidensis MR-1 biofilms by cyclic di-GMP.** *J Bacteriol* 2006, **188:**2681-2691.
34. Mountfort DO, Rainey FA, Burghardt J, Kaspar HF, Stackebrandt E: **Psychromonas antarcticus gen. nov., sp. nov., A new aerotolerant anaerobic, halophilic psychrophile isolated from pond sediment of the McMurdo ice shelf, antarctica.** *Arch Microbiol* 1998, **169:**231-238.
35. **Joint Genomes Institute** [http://www.jgi.doe.gov/]
36. **Codon Code Corporation** [http://www.phrap.com]
37. Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing.** *Genome Research* 1998, **8:**195-202.
38. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Research* 1998, **8:**186-194.
39. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Research* 1998, **8:**175-185.
40. Han CF, Chain P: **Finishing repeat regions automatically with Dupfinisher.** In *Proceeding of the International Conference on Bioinformatics & Computational Biology* Edited by: Arabnia HR, Valafar H, Las Vegas NV. CSREA Press; 2006:141-146. 26-29 June 2006
41. Scott KM, Sievert SM, Abril FN, Ball LA, Barrett CJ, Blake RA, Boller AJ, Chain PS, Clark JA, Davis CR, Detter C, Do KF, Dobrinski KP, Faza BI, Fitzpatrick KA, Freyermuth SK, Harmer TL, Hauser LJ, Hugler M, Kerfeld CA, Klotz MG, Kong WW, Land M, Lapidus A, Larimer FW, Longo DL, Lucas S, Malfatti SA, Massey SE, Martin DD, McCuddin Z, Meyer F, Moore JL, Ocampo LH Jr, Paul JH, Paulsen IT, Reep DK, Ren Q, Ross RL, Sato PY, Thomas P, Tinkham LE, Zeruth GT: **The genome of deep-sea vent chemolithoautotroph Thiomicrospira crunogena XCL-2.** *PLoS Biol* 2006, **12:**e383.
42. **Integrated Microbial Genomes system** [http://img.jgi.doe.gov/]
43. Altschul SF: **Amino acid substitution matrices from an information theoretic perspective.** *J Mol Biol* 1991, **219:**555-565.
44. **FactoMineR R package** [http://factominer.free.fr/]
45. **Mclust R package** [http://www.stat.washington.edu/mclust/]

2.2 Genomic Islands

2.2.1 Literature Review

2.2.1.1 Characteristics of Genomic Islands (GIs)

Among bacteria with close relationship in evolution, even from various strains of the same species, it was common that 10% of their genomic regions were not conserved [96-98]. When comparing the strains K12 and O157 of *Escherichia coli*, the strain-specific fragments account for 12% of the K12 genome and 26% of the O157 genome respectively [99]. Among these fragments, the regions less than 10kb are called genomic islets, while the regions larger than 10kb are treated as genomic islands (GI) [100, 101], and some of the genomic islands may even span over 500kb [102]. Compared with the conserved core genomic regions, the genomic islands/islets have different features such as different nucleotide compositions and codon usage pattern. It has always been found that tRNAs, direct repeats (DRs) and insertion sequences (ISs) flank both sides of the genomic islands/islets, and integrases, transposase, other mobile genes and also some functional genes are located in the genomic islands/islets (Fig. 2.2) [103-105].

Figure 2.2 Schematic diagram of Genomic Island/Islet structure
Note: Core genome is the conserved region among closely related genomes; DR is direct repeat; int is integrase; Gene 1 (2) is one of functional genes; IS is an insertion sequence

Some of the characteristics of genomic islands listed above also appear in other genomic

elements, such as the integrated plasmid, phage and gene locus for synthesizing extracellular polysaccharide [105, 106]. Genomic islands (islets) can be easily distinguished from Phages and plasmids since the former does not contain the automatic replication initiation site [107].

2.2.1.2 Function and classification of the Genomic Island

In some pathogenic bacteria, some exogenous DNA fragments contain genes enhancing bacterial virulence through the encoding of toxins, adhesion, invasion and other virulent factors. For example, a pathogenicity island (PAI) in urinary tract pathogenic *E. coli* (UPEC) carries pathogenetic factors alpha-haemolysin and virulence determinant P fimbriae [108-110]. Similar to the Pathogenicity Island in pathogenic bacteria, large fragments of exogenous DNA can also be found in non-pathogenic bacteria. For example, an integrated plasmid region in *Mesorhizobium loti* contain genes functioning in the process of nitrogen fixation [111]; an integrated plasmid region in *Pseudomonas putida*, contains genes functioning in the degradation of chlorophenols [112]; an integrated region in the *Rhizobium* gives the species ability to form symbiose with legumes [113]; mecA in *Staphylococci* is a foreign origin region responsible for methicillin-resistant genes [114]; some foreign DNA regions in *Samonella senftenberg* contain genes participating in important metabolism genes like sucrose absorption [115]. All above mentioned DNA regions are referred as genomic islands, which contain genes improving the existing function of host bacteria, or adding novel functions to bacteria, and thereby increase the bacteria adaptability. Based on these characteristics contributed to the host bacteria, these genomic regions (or genomic islands) are classified into pathogenicity island (PAI), symbiosis island (SI), fitness Island (FI), metabolic islands (MI), resistance islands (RI) and so on (Fig. 2.3)

[101]. Accordingly, the above mentioned region integrated to *Rhizobium* is a symbiosis island (SI), and mecA in *Staphylococci* is a resistance island (RI).

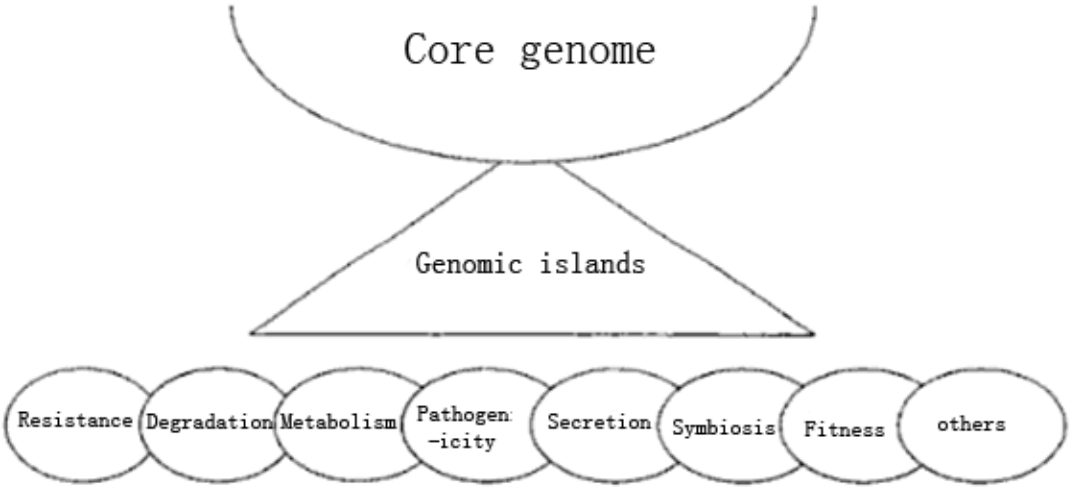


Figure 2.3 Classification of genomic islands according to the variety of function they contributed to the hosts.

The boundaries among different kinds of genomic islands are not clear. For example a pathogenicity island (high pathogenicity island, HPI) in *Yersinia* contains a collection of genes involved in iron uptake system [116]. However, this kind of region can also be found in non-pathogenic *E. coli* [117]. In non-pathogenic bacteria, the "High pathogenicity island (HPI)" can help host absorb iron and thus should be treated as a fitness island, which increase the competitiveness of their host strain, when compared to the other microbes and also the eukaryotes from the same environment. Yoon et al. detected candidate pathogenicity islands from six non-pathogen strains. Those candidate pathogenicity islands encode various gene functions, such as ABC transporter (*Bacillus halodurans*), flagellar proteins (*Bacillus subtilis*), iron transport and fimbrial proteins (*E. coli* K-12), transmembrane sensors and outer membrane efflux proteins (*Nitrosomonas europaea*), nodulation proteins (*Bradyrhizobium japonicum*) and proteins involved in type III secretion system (*Mesorhizobium loti*)[118]. Grozdanov et al. demonstrated

that factors such as adhesions, iron uptake systems, and proteases can make great contributions to the fitness and adaptability of nonpathogenic bacteria and thus do not necessarily have to be considered as virulence-associated factors [119]. Additionally, the normal non-pathogenic strains of *E. coli* also contain some regions coding for P fimbriae, the determinant of PAIs in urinary tract pathogenic *E. coli* (UPEC) [117]. Pathogenicity Island in pathogens enhances the host toxicity, while in other hosts they are called Fitness Island (FI) since they can enhance the host viability. For example, the genomic islands utilize a codon usage pattern different from the host housekeeping gene, which can be of benefit for the host to express the genomic islands genes in special environment [117].

2.2.1.3 Prediction of Genomic Island

In last century, there were few genomes sequenced, the pathogenicity islands were mainly identified through subtractive hybridization [120], signature tagging [121] and other experiments. By hybridization experiments with *E. coli* K12 strains, pathogenicity islands were identified from strain *E. coli* K1 [122]. By the application of island probe, pathogenicity islands were identified from *Shigella flexneri* [123]. And by PCR, molecular cloning and nucleotide sequencing, the fourth pathogenicity island, which contains a Type I secretion system related to toxin-secretion, was identified from *Salmonella* [124].

Subsequently, benefitting from large-scale genome sequencing, more and more whole genome sequences were available in the world, and the work of searching genomic islands gradually shifted from experiments to computers. First consider the different nucleotide compositions between genomic islands and the genome core regions. In practice, an

implementation is the assessment of GC content among regions from a chromosome (or genome). Typically, at first, one can choose a proper "sliding box" (or window) moving along the genome, and then calculate the total occurrences of G and C in all windows determined in the previous process. But in this method, the size of "sliding box" is difficult to determine, because a large box results in low resolution, while a small box results in statistical fluctuations. To overcome this shortcoming, Z-curve [125] and the cumulative GC profile [126-128] method adopt non-sliding-box technology, which is through computing the difference between AT content and GC content from the initiation site to any position. According to the curve, one can view the nucleotide composition changes at both boundaries of a region, and then determine whether the region is a genomic island or not.

In addition to using GC content to analyze the nucleotide content, one can also use the codon usage bias, dinucleotide frequency, oligonucleotide composition and other factors to identify genomic islands [129-132]. Compared to a fixed length of nucleotides in calculating the nucleotide composition, Interpolated Variable Order Motifs (IVOMs) integrates different lengths of nucleotides (consider calculating the time, generally considered from the first order to the 8th-order), to identify genomic islands in the genome [133].

In the core part of some bacteria genomes, highly expressed genes also adopt special codon usage different from others [134]. Using Hidden Markov Model Analysis of codon usage preferences, SIGI-HMM can exclude the possible highly expressed genes, and predict possible genomic islands [135].

In most predictions, whether the difference of a region from the core part of genome is sufficient to be considered as a genomic island is usually judged by researchers' empirical

experiences. Program Design-Island (an abbreviation for Detection of Statistically Significant Genomic Islands) introduced a precise statistical theory and reliable P-value to observe the significance level of a genomic island, which improved the sensitivity and accuracy of genomic island identification [136].

Besides the genome composition, we can also consider those regions with several characteristics related to genomic island as the candidates of genomic islands. IslandPath is a graphical web service program displaying genomes for Bacteria and Archaea, in which the image integrates several DNA signals and genome annotations, and can help identify genomic islands [137]. Islander is the database of the genomic island for prokaryotic genomes; particularly it records down integrases and their specific DNA sites [138]. In addition to tRNAs, tmRNAs and the small RNAs are also the hot spot or potential sites for genomic island integration, and these kinds of genomic islands can be identified by software tRNAcc [139, 140].

A pre-requirement for above listed methods in prediction of Genomic Islands is the availability of the corresponding whole genome sequence. These methods can not be applied to those stains without whole genome sequences. MobilomeFINDER is an interactive online tool to make up this gap [141]. Based on chip probes, ArrayOme can estimate theoretically the size of a genome, which is normally different from the physical size determined by pulsed-field gel electrophoresis (PFGE). Those genomes with significant differences between the two types of sizes contain many “novel” (or undiscovered) genes which have not been included in microarray-based Comparative Genomic Indexing (CGI) [142]. Subsequently PCR and tRNAcc will analyze the possible regions containing these novel genes, and determine whether they are genomic islands or not [141].

When horizontal gene transfer events occurred between species with similar or identical nucleotide composition, the differences of nucleotide compositions between genomic island and the core of host genome would be little or none [117]. When the horizontal gene transfer events occurred very anciently, as mentioned before, the complex evolutionary events might have wiped out the differences between genomic islands and core genome. To overcome these limitations, the following characteristics specific to genomic islands could be considered: they are large fragments inserted by horizontal gene transfer and they are distributed in limited species, which means they are present in some species, but absent from other closely related species. Therefore without considering the structure of genomic islands, it is possible to find genomic islands just according to their limited distribution among species. This kind of idea has been applied to several *Streptococcus* strains [143]. And for this, Vernikos gave the process: Firstly, select the genomic regions distributed in limited species; then annotate these regions with structural features; and later determine the contributions of each module to genomic islands using machine learning methods [144].

To detect the origin of a genomic island, the software Compare_Islands compares between different genomic islands and between genomic islands and other related genomes, then judges whether several genomic islands have similar origins, and even their possible sources [145].

Pathogenicity Island (PAI) is a small part of the genomic islands. Normally, the first thing to identify PAI is to identify genomic islands, and later those islands containing toxic genes are considered as PAIs. For example: through RPSBLAST, the software PredictBias compares those regions with significant difference from the core genome to the database of toxic factors, and considers a region with one or more genes related to toxicity as PAI [146]. In another article, the

authors collected some 207 integral or partial known PAI, and the possible genomic islands were identified through abnormal levels GC content and codon usage bias, and further the pathogenicity islands were determined from these genomic islands once they contained virulent genes [118].

2.2.2 A combined approach for identification of genomic islands in prokaryotic sequences

2.2.2.1 Study Strategy and Results Summary

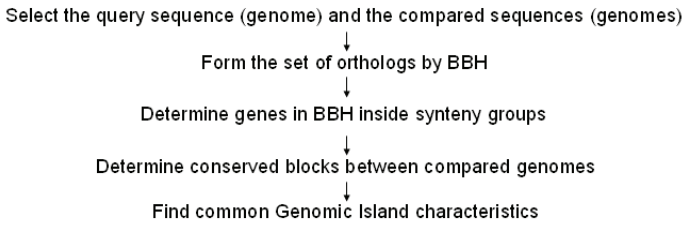


Figure 2.4 The study strategy for genomic islands prediction

The study strategy for genomic islands prediction was summarized in Figure 2.4, and described in Poster 1. This new method of prediction of Genomic Islands has been applied in two genomes, specifically *E. coli* described in Poster 1 and *B. subtilis* described in Article 2. In Poster 1, focusing on the comparison of ExPEC and commensal *E. coli*, we have identified new specific regions that haven't been published before. These specific regions correspond mainly to pathogenicity islands and strain-specific phage insertions, and carry several virulence-related features. In Article 2, genomic islands (regions) were found based on the newly sequenced *B. subtilis* str 168 with several other *Bacillus* strains as the comparing groups. It appeared that these genomic regions sometimes had a composite structure, e.g. they were made of regions partially conserved or found in different synteny groups (i.e. in different genomic locations) in the different *Bacillus* genomes. The predicted RGPs (region of genomic plasticity) have therefore been further manually curated to define subregions called modules. Associated with those genomic islands, we have found a variety of genes that keep signatures of widespread functions permitting gene transfer, and other features are related to genes present in Eukarya.

Poster 1

A combined approach for identification of genomic islands in prokaryotic sequences: an application to pathogenic *Escherichia coli* strains

Roche David, Calteau Alexandra, **Wang Tingzhang**, Cruveiller Stéphane, Médigue Claudine

A combined approach for identification of genomic islands in prokaryotic sequences: an application to pathogenic *Escherichia coli* strains

Roche David[1], Calteau Alexandra[1], Wang Tingzhang[1,2], Cruveiller Stéphane[1], Médigue Claudine[1]

1 CEA/ Institut de Génomique / Genoscope, Atelier de Génomique Comparative, 2 rue Gaston Cr émieux, 91057 EVRY Cedex – FRANCE
2 Institut Pasteur, Génétique des Génomes Bactériens, 28 rue du Dr Roux, 75015 Paris - FRANCE

Introduction

Acquisition of DNA by horizontal gene transfer (HGT) is an effective way of generating diversity between bacterial species. If the newly acquired DNA has a selective advantage for the organism, it may be retained and stably integrated into the host genome through the process of natural selection [1, 2] (Figure 1). Genomic Islands (GIs; Figure 2) are large chromosomal regions (5 – 100 kb in length) flanked by direct repeat sequences, and located near an insertion sequence and/or a tRNA gene. Furthermore, GIs frequently have a GC composition different from other parts of the genome (compositional bias). These chromosomal regions cluster functionally related genes and include pathogenicity islands (PAIs), symbiotic islands (SYIs), metabolic islands (MEIs), antibiotic resistance islands GREIs) and secretion system islands (SEIs). HGT is then believed to be essential for adaptative evolution of bacterial species.

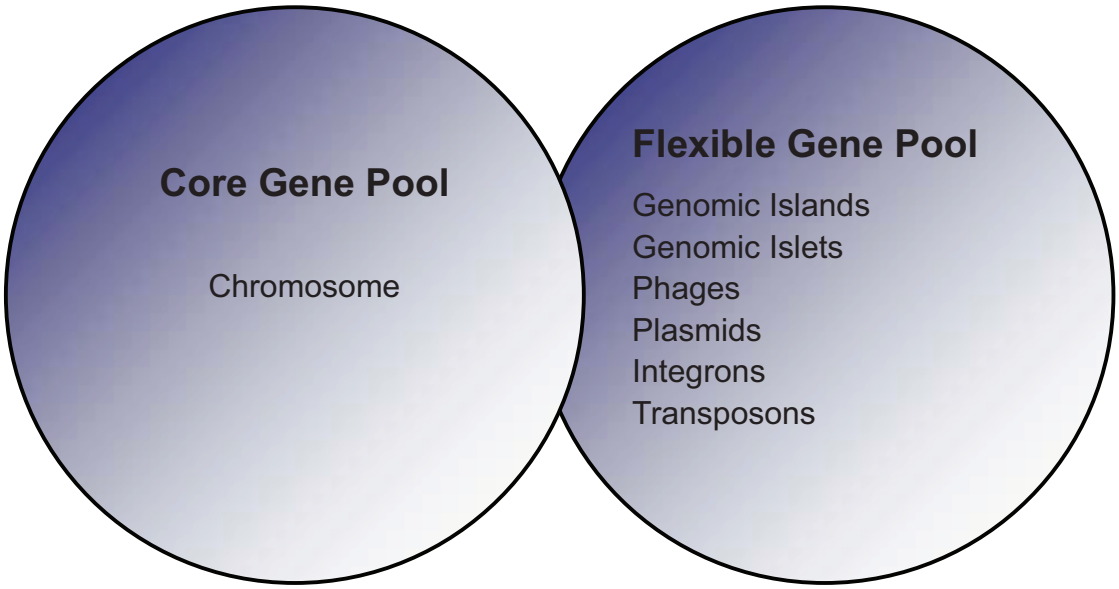


Figure 1: M**odel of the DNA pools in prokaryotic genomes [1]**

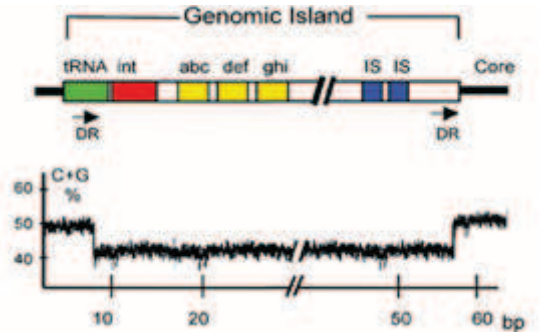


Figure 2: G**eneral characteristics of Genomic Islands [1]**

A significant number of methods have been developed in order to detect Genomic Islands from analysing genomic sequences [3-6]. Intrinsic methods based on compositional criteria: atypical G+C content, codon bias, dinucleotide frequency differences, and different genomic signatures. Extrinsic methods based on similarity and detection of common GIs features: tRNA (insertion sites), mobility genes (integrase), direct repeats and so on.

Because none of these methods are sufficient to define a GI when the donor and the recipient genomes are similar, or if the age of the HGT event is reasonably old and because some GI have no common criteria, it will be advantageous to use comparative genomics on the increasing number of sequences available. We have developed a new method which combines conservation of synteny groups (gene colocalisation) of orthologous genes between related bacteria with compositional biases and common GIs features such as tRNA, IS and direct repeats to analyse Genomic Regions. We thus extended the definition of Genomic Islands to all regions which differ in related species.

We applied this new method to the data from the ColiScope project (coordinator Erick Denamur, Inserm U722)

Method

To find potential **Genomic Regions** in a sequence we delineate the core gene pool from the flexible gene pool of a query sequence (conserved backbone) by comparing this sequence to a selected set of genomes. This analysis strongly depends of the set of organisms chosen by the user.

We first delineate the set of corresponding genes between the query and all other organisms, i.e. the set of orthologs, by identifying BBH (Bidirectional Best Hits, Figure 3). However, this straightforward method can introduce false negatives in cases of pseudo-orthology (i.e., genes that actually are paralogs but appear to be orthologs due to differential lineage-specific gene loss) and xenology (homologous genes acquired via HGT by one or both of the compared species but appearing to be orthologs in pairwise genome comparisons). So, as we are looking for regions, we have introduced the use of synteny (local conserved gene structure between organisms; Figure 3) as a second criterium. Genes in BBH inside synteny groups between all compared organisms are more likely to be part of the query sequence backbone.

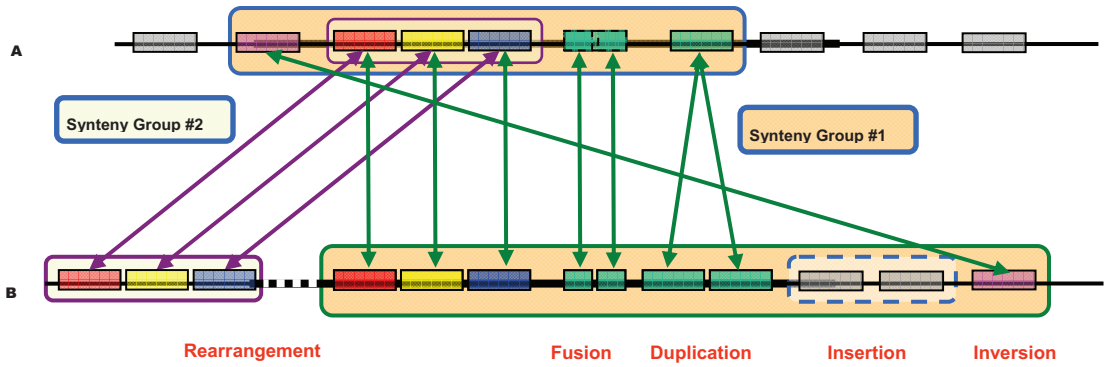


Figure 3: S**ynteny groups and BBH in prokaryotic genomes**

Then, we identify conserved blocks between compared genomes: they are defined by a minimum of 3 genes in BBH and synteny. To delineate potential Genomic Regions, we consider

regions above 5 kb in length between two conserved blocks in the query sequence. It is important to note that the lack of a conserved block in only one compared sequence is sufficient to define/extend a predicted region on the query genome (see Figure 4). These Genomic Regions are then analysed to find some common Genomic Island characteristics such as tRNA, integrase, atypical GC content (sliding window), tRNA repeats or combinations of these features.

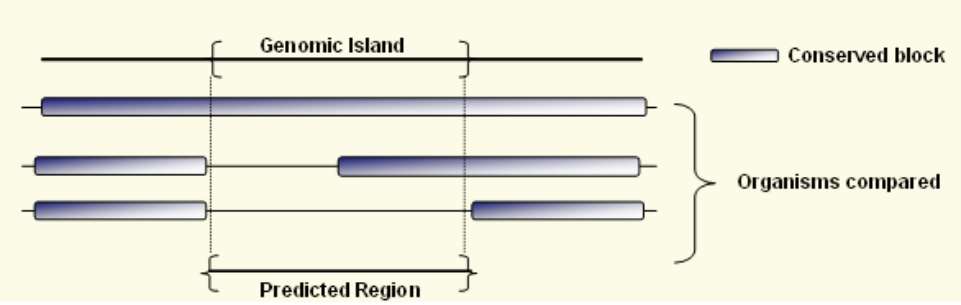
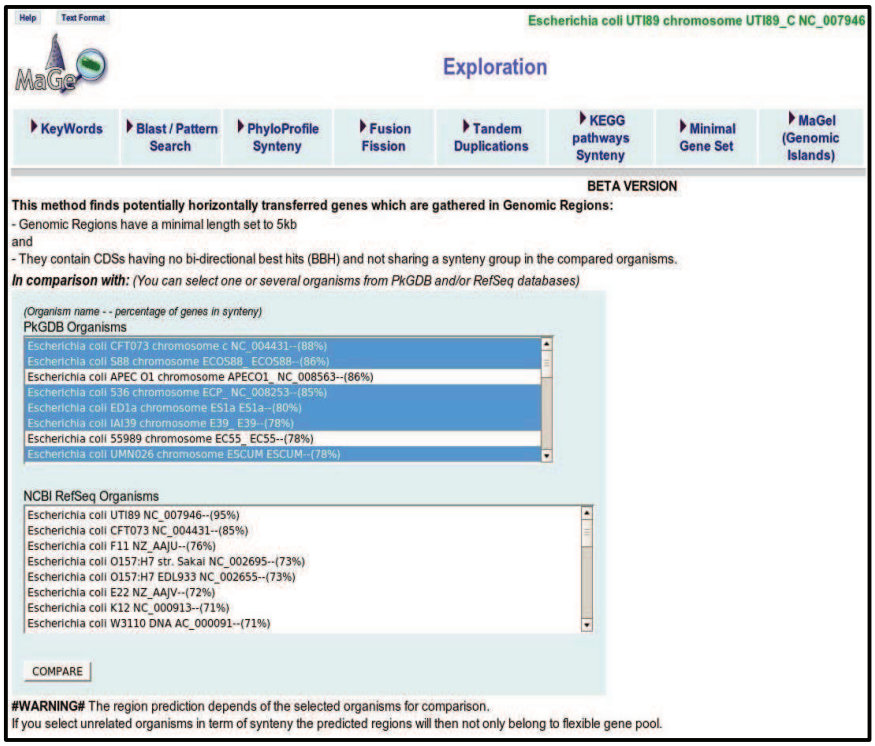


Figure 4: Example of a predicted region

MaGeI tool



Step 1: Selection of a set of genomes to compare through the MaGe interface of the MicroScope platform (http://www.genoscope.cns.fr/agc/mage)

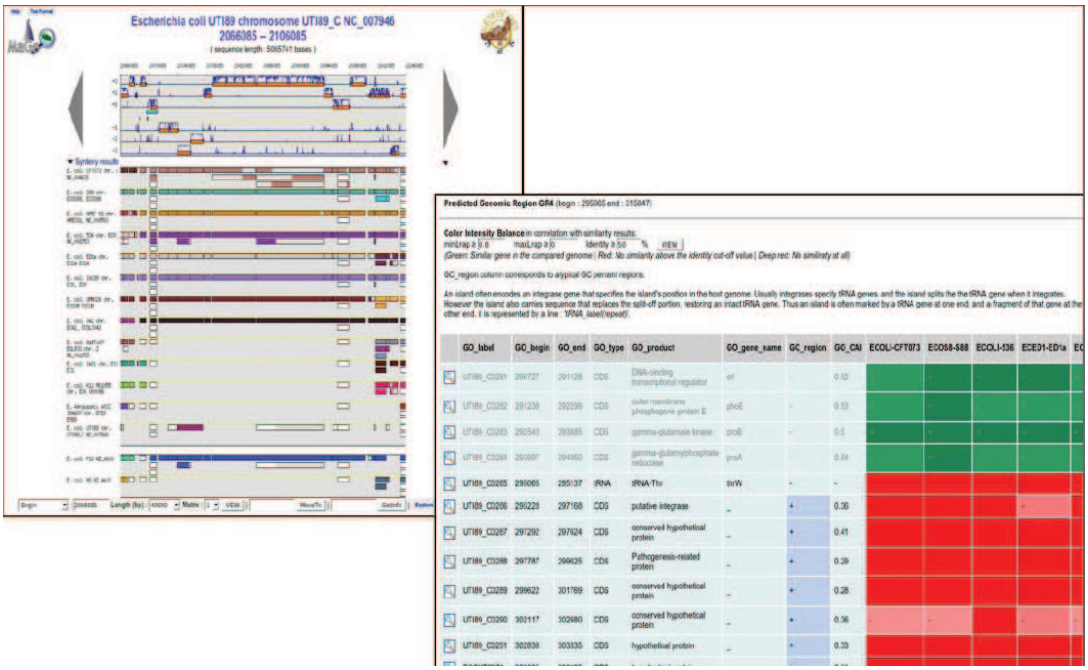
75 predicted regions in Escherichia coli UTI89 chromosome UTI89_C NC_007946 264 by comparison with:

Replicon Name	% CDS in Syntons	Origin
Escherichia coli CFT073 chromosome c NC_004431	88	PKGDB
Escherichia coli S88 chromosome ECOS88_ ECOS88	86	PKGDB
Escherichia coli 536 chromosome ECP_ NC_008253	85	PKGDB
Escherichia coli ED1a chromosome ES1a ES1a	80	PKGDB
Escherichia coli IAI39 chromosome E39_ E39	78	PKGDB
Escherichia coli UMN026 chromosome ESCUM ESCUM	78	PKGDB
Escherichia coli IAI1 chromosome EI1 EI1	73	PKGDB
Escherichia coli K12 MG1655 chromosome ECK U00096	72	PKGDB

Table Legend:
The feature field correspond to common Genomic Islands features.
For each compared organism (column) a specificity % is computed. It corresponds to the number of CDS not in synton in a region.
Table Sort:
by Begin by Features by Specificity Sort

	GR_label	GR_begin	GR_end	GR_length	feature [Left border] [Inside] [Right border]	ECOLI-CFT073	ECOS8-S88	ECOLI-536	ECED1-ED1a	ECIAI-IAI39	ECUMN-UMN026	ECOIA-IAI1	ECOLI-K12 MG1655
	GR4	295065	315847	20783	[tRNA/int][tRNA/int - GC][none]	54	58	58	85	65	77	92	77
	GR29	2109296	2168395	59100	[tRNA][tRNA - tRNA/int - GC][pseudo]	6	87	0	84	84	87	81	90
	GR28	2068085	2101820	33736	[tRNA][tRNA/int - GC][none]	6	6	6	18	6	18	88	94

Step 2: Visualisation of the results



Step 3: Selection of a specific region

All predicted genomic regions are summarized in a table. The “feature” column indicates to the user if the region harbors one or several features typical of genomic islands. The specificity percentages indicate the number of genes that are not in synteny between the query and the compared genomes. From the result window it is possible to: (1) view one specific region into the main MaGe window access and (2) obtain a detailed table of a specific region.

Detection of Genomic Regions in *E. coli* strains

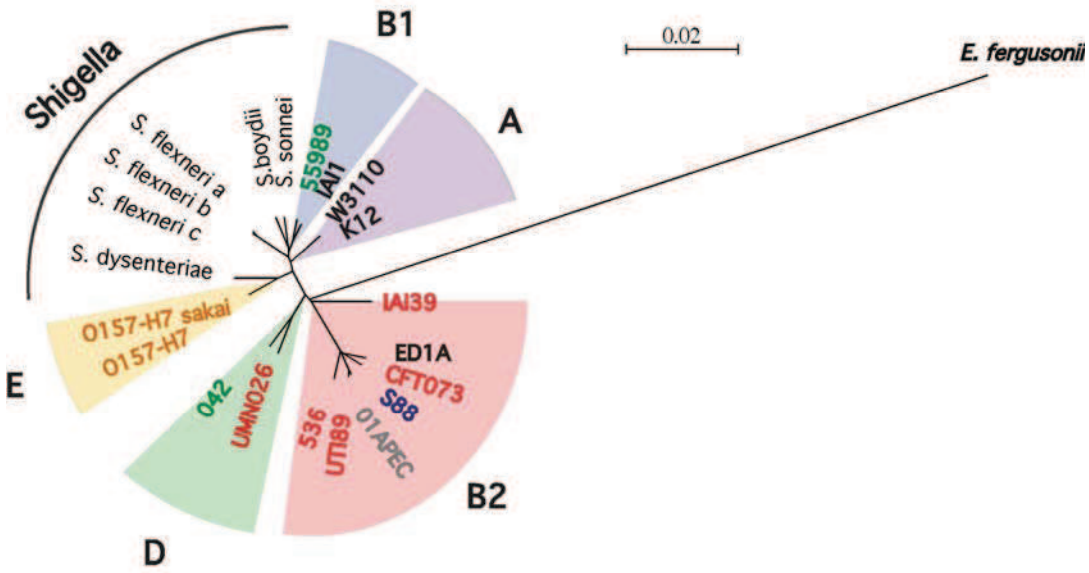


Figure 5: Maximum-likelihood phylogeny of *E. coli* strains using *E. fergusonii* as an outgroup. The different colors indicate the different pathotypes: uropathogenic (red), meningitis-associated (blue), enteroaggregative (green), enterohaemorrhagic (orange), avian pathogenic (grey), shigellosis (thin black), commensal (bold black). (courtesy of Marie Touchon)

To evaluate the performance of our method, we applied it on a set of pathogenic and commensal *E. coli* genomes. In particular, we have decided to focus our analysis on the comparison of ExPEC and commensal *E. coli* genomes (Figure 5). *E. coli* strains capable of causing diseases outside the gastrointestinal tract refer to Extra-intestinal Pathogenic *E. coli* (ExPEC). Uropathogenic *E. coli* (UPEC), a prominent member of the ExPEC family, are responsible for up to 90% of uncomplicated urinary tract infections. Among the ExPEC strains, three complete *E. coli* UPEC genomes (CFT073, UTI89 and 536) have been sequenced [7-9]; two other UPEC genomes (IAI39 and UMN026) and one MeNingitic E. coli genome (MNEC, S88) have been sequenced at Genoscope, but not yet published. Three commensal strains were added to

this list of pathogenic *E. coli*: K12 MG1655[10], and IAI1 and ED1a (both sequenced at Genoscope).

All ExPEC species have been investigated with our method and in each case all genomic regions have been manually inspected and corrected if necessary (modification of the limits of the regions, split of one region into several separate regions). The results are shown in Table 1. We have identified new UPEC-specific regions that haven't been published before, as well as MNEC-specific regions. These specific regions correspond mainly to pathogenicity islands and strain-specific phage insertions. They carry several virulence-related features. Conversely, non-specific regions correspond mainly to complete operons coding for genes involved in metabolism activities or for transporter systems.

Table 1: **Classification of genomic regions identified in ExPEC coli strains.**

	UPEC					MNEC
	536	**UTI89**	**CFT073**	**IAI39**	**UMN026**	**S88**
Taille (kb)	4938	5065	5231	4800	5100	4980
Number of regions detected by MaGeI	68	75	73	84	69	65
Atypic regions	38	42	43	55	32	35
Normal regions	30	33	30	29	37	30
Strain specific regions	4	5	10	15	22	9
UPEC specific regions	9	4	3	7	5	
ExPEC specific regions	16	18	23	26	31	18
Other	37	40	47	47	36	22

Only regions above 5 kb have been classified. ExPEC specific regions include strain-specific and UPEC specific regions. Atypical regions correspond to those that have at least one common feature of genomic islands (tRNA, GC, integrases ...)

Our preliminary analyses have also shown that some of the genomic regions identified have a modular structure. Moreover, some of these modules can be identified as part of large genomic regions or as independent regions. A more careful study of these modules will allow us to have an insight into the dynamics of the flexible gene pool among different *E. coli* strains.

Reference

1. Hacker, J. and E. Carniel, *Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes.* EMBO Rep, 2001. **2**(5): p. 376-81.
2. Dobrindt, U., et al., *Genomic islands in pathogenic and environmental microorganisms.* Nat Rev Microbiol, 2004. **2**(5): p. 414-24.
3. Hsiao, W.W., et al., *Evidence of a large novel gene pool associated with prokaryotic genomic islands.* PLoS Genet, 2005. **1**(5): p. e62.
4. Ou, H.Y., et al., *A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria.* Nucleic Acids Res, 2006. **34**(1): p. e3.
5. Mantri, Y. and K.P. Williams, *Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities.* Nucleic Acids Res, 2004. **32**(Database issue): p. D55-8.
6. Yoon, S.H., et al., *A computational approach for identifying pathogenicity islands in prokaryotic genomes.* BMC Bioinformatics, 2005. **6**: p. 184.
7. Welch, R.A., et al., *Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli.* Proc Natl Acad Sci U S A, 2002. **99**(26): p. 17020-4.
8. Chen, S.L., et al., *Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach.* Proc Natl Acad Sci U S A, 2006. **103**(15): p. 5977-82.
9. Brzuszkiewicz, E., et al., *How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic Escherichia coli strains.* Proc Natl Acad Sci U S A, 2006. **103**(34): p. 12879-84.
10. Blattner, F.R., et al., *The complete genome sequence of Escherichia coli K-12.* Science, 1997. **277**(5331): p. 1453-74.

Article 2

From a consortium sequence to a unified sequence: the Bacillus subtilis 168 reference genome a decade later

Valérie Barbe, Stéphane Cruveiller, Frank Kunst, Patricia Lenoble, Guillaume Meurice, Agnieszka Sekowska, David Vallenet, **Tingzhang Wang**, Ivan Moszer, Claudine Médigue and Antoine Danchin

Microbiology (2009), 155, 1758-1775

Microbiology (2009), **155**, 1758–1775 DOI 10.1099/mic.0.027839-0

From a consortium sequence to a unified sequence: the *Bacillus subtilis* 168 reference genome a decade later

Valérie Barbe,[1]† Stéphane Cruveiller,[2]† Frank Kunst,[1] Patricia Lenoble,[1] Guillaume Meurice,[3] Agnieszka Sekowska,[4] David Vallenet,[2] Tingzhang Wang,[4] Ivan Moszer,[3] Claudine Médigue[2] and Antoine Danchin[4]

Correspondence
Antoine Danchin
antoine.danchin@normalesup.org

[1]CEA, Institut de Génomique, Génoscope, 2 rue Gaston Crémieux, 91057 Évry, France

[2]CEA, Institut de Génomique, Laboratoire de Génomique Comparative/CNRS UMR8030, Génoscope, 2 rue Gaston Crémieux, 91057 Évry, France

[3]Institut Pasteur, Intégration et Analyse Génomiques, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France

[4]Institut Pasteur, Génétique des Génomes Bactériens/CNRS URA2171, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France

Comparative genomics is the cornerstone of identification of gene functions. The immense number of living organisms precludes experimental identification of functions except in a handful of model organisms. The bacterial domain is split into large branches, among which the Firmicutes occupy a considerable space. *Bacillus subtilis* has been the model of Firmicutes for decades and its genome has been a reference for more than 10 years. Sequencing the genome involved more than 30 laboratories, with different expertises, in a attempt to make the most of the experimental information that could be associated with the sequence. This had the expected drawback that the sequencing expertise was quite varied among the groups involved, especially at a time when sequencing genomes was extremely hard work. The recent development of very efficient, fast and accurate sequencing techniques, in parallel with the development of high-level annotation platforms, motivated the present resequencing work. The updated sequence has been reannotated in agreement with the UniProt protein knowledge base, keeping in perspective the split between the paleome (genes necessary for sustaining and perpetuating life) and the cenome (genes required for occupation of a niche, suggesting here that *B. subtilis* is an epiphyte). This should permit investigators to make reliable inferences to prepare validation experiments in a variety of domains of bacterial growth and development as well as build up accurate phylogenies.

Received 26 January 2009
Revised 25 February 2009
Accepted 25 February 2009

INTRODUCTION

Bacillus subtilis has been a model for Gram-positive bacteria for more than a century. Generally Recognized As Safe (GRAS), it is a ubiquitous ingredient of food supplies and a typical member of the A+T-rich Firmicutes, a major clade of the bacterial domain of life. It has been used as the reference model for cell differentiation, and many studies have analysed its well-defined sporulation programme (for reviews see Errington, 2003; Piggot & Hilbert, 2004; Yudkin & Clarkson, 2005). Whereas it has long been accepted that model organisms play a crucial role in the way they coordinate the research of many investigators, some now minimize the input of laboratory organisms that, historically, have been chosen in a more or less random way (Hobman *et al.*, 2007). Yet, we need references, and comparative genomics is based on approaches that have much in common with the way hieroglyphics were understood using the Rosetta stone. This was an obvious reason for the setting up of a *B. subtilis* genome sequencing programme in 1987. At that early time of genomics, cloning pieces covering the entire genome as well as sequencing was a tedious task. Furthermore, it was obvious, if one wished to couple sequencing to functional knowledge, that the various

†These authors contributed equally to this work.

Abbreviations: AdoMet, *S*-adenosylmethionine; CDS, coding sequence; IIMP, integral inner-membrane protein; MTR, methylthioribose; RGP, regions of genomic plasticity; ROS, reactive oxygen species.

The GenBank/EMBL/DDBJ accession number for the sequence reported in this paper is AL009126.

Four supplementary tables are available with the online version of this paper.

expertises of many investigators should be put together in one common programme. This was at the core of the European effort to set up two major consortia for sequencing microbial genomes, the *Saccharomyces cerevisiae* consortium and the *B. subtilis* consortium (Simpson, 2001). Joined by Japanese investigators in 1990, the consortium extended to some 30 groups (including two US and a Korean group, joining later on) which sequenced and annotated chromosomal segments covering the whole *B. subtilis* genome (Harwood & Wipat, 1996).

Whereas this mode of organization had obvious benefits in terms of creation of scientific knowledge, it had technical drawbacks. Indeed, putting together different laboratories implied different local practices and different performances in the quality of the final sequence. Furthermore, because of the difficulty in sequencing at that early time, some regions which had been sequenced in previous years were not resequenced (see Results and Discussion). Finally, the sequencing techniques were quite time-consuming and involved cloning into a variety of vectors, which could result in alteration of the original sequence, especially given that *B. subtilis* DNA is often toxic in *Escherichia coli* (Frangeul *et al.*, 1999). Being aware of this problem, the *B. subtilis* consortium decided to use a first final draft of the sequence to identify regions potentially containing inaccuracies, to PCR out and resequence 500 bp fragments in these regions (Kunst *et al.*, 1997; Medigue *et al.*, 1999). However, there was no doubt that errors still remained in the published sequence. Naturally, sequencing was the prelude to an explosion of new genetic, physiological and biochemical analyses, and genes which were initially of unknown function were constantly being discovered. The first reference database, SubtiList, displaying sequence and annotations, was replaced six years ago by an update (Moszer *et al.*, 2002). However, no further recent updates of the sequence and the annotation repository are available and this should become a matter of concern as *B. subtilis* remains widely used in automatic annotation procedures.

The 1997 sequence was used as a reference until the present time. Since then sequencing techniques have improved dramatically. They no longer require cloning steps, which, in the case of A+T-rich Firmicutes (excluding Mollicutes, which do not express well most of their DNA in *E. coli* because of their use of a UGA codon encoding tryptophan), counterselect regions that are expressed at high levels in the library hosts (Frangeul *et al.*, 1999). Furthermore, it has been repeatedly observed that bacterial strains evolve fast in laboratories. For example *E. coli* strains MG1655 and W3110 are significantly different (including a large inversion around the origin of replication) (Hayashi *et al.*, 2006; Herring & Palsson, 2007). It has also been found that isolates of MG1655 from different laboratories may display considerable variations (Soupene *et al.*, 2003). These observations should be extended to *B. subtilis*, and it was therefore timely to resequence the genome to have a reliable reference.

Per se a genome sequence is of limited interest. What is important for the scientific community is the identification of the genomic objects present in the sequence, associated with their functions experimentally identified or predicted *in silico*. In fact, while we do not have exact measures of the impact of annotations, we know how errors percolate in a very dangerous process (Gilks *et al.*, 2002). It is interesting to note that many genes and gene products retained the 'y' name created during the first sequencing programmes, testifying that the genome project has indeed been seminal to many discoveries, even if no reference to the original annotation work was provided. We therefore decided to reannotate the sequence entirely using a recently developed platform, MaGe [Magnifying Genomes (Vallenet *et al.*, 2006)], systematically proceeding by inference and using all kinds of neighbourhoods (including *in biblio*, with the systematic use of selected PubMed references, and using PubMed Central as much as possible), as described previously (Nitschke *et al.*, 1998). To keep in phase with the international community, a jamboree was organized by the Swiss Institute of Bioinformatics in Geneva, to harmonize our annotation with that of the HAMAP project (Lima *et al.*, 2009).

The knowledge of the genome sequence allows one to explore the consistency of annotations and, in particular, the organization of the genome into several distinct functional processes: what sustains life, what perpetuates life, and what permits life in a particular niche (Danchin *et al.*, 2007). The latter is an essential motivation for the choice of an organism as a useful model. Does analysis of the *B. subtilis* genome fit with its biotope? Back in 1859, Louis Pasteur thought he had clearly demonstrated the absence of spontaneous generation in broths, usually found to become rapidly full of living microbes. His (in)famous competitor Félix Pouchet fought back heartily, showing that he found evidence contrary to that of Pasteur, and a bitter fight ensued. What was Pouchet's evidence? Pasteur used heated yeast extracts while Pouchet showed that he recovered micro-organisms after having boiled vessels using 'l'eau de foin' (hay water) (Roll-Hansen, 1979). We now know the reason why this is so: hay is the normal niche of a bacterium identified in 1885 as the 'hay bacterium', *B. subtilis*, which makes heat-resistant spores, and is found as a major component in the process of water retting of plants used, in particular, for the production of linen threads (Tamburini *et al.*, 2003).

In this work we used the complete resequencing of the genome to emphasize some new features of the genome, the genes and their products, leaving more standard information readily available in two databases: a new updated release of SubtiList, now integrated in a multi-genome framework, GenoList, and BacilluScope, the overview of the annotation platform MaGe.

METHODS

Origin of strain and DNA. The *Bacillus subtilis* strain used in this work is the one described by Anagnostopoulos and Spizizen

(Anagnostopoulos & Spizizen, 1961; Zeigler *et al.*, 2008). It is the same as the one which was distributed to the consortium involved in the sequencing of strain 168, beginning in 1987 (Kunst *et al.*, 1997). Many strains labelled '168' exist in many laboratories in the world, and it is expected that there is significant polymorphism in these strains. Indeed, recent work has identified variations, some of which differ both from the published sequence and from the update presented in this work (Srivatsan *et al.*, 2008).

B. subtilis strain 168 had been distributed to the sequencing consortium initially in 1989–1990. However, it has already been noticed that after 20 years of passages in many laboratories the growth phenotype of the strain differed from place to place. In particular this was evidenced by variations in the growth rate on minimal media supplemented with ammonium as a nitrogen source (E. Presecan & A. Sekowska, unpublished observations; see Results and Discussion). For this reason a new culture derived from the original collection of *B. subtilis* strain 168 conserved by C. Anagnostopoulos was used for the functional analysis programme (Kobayashi *et al.*, 2003). The genome of this same isolate has been sequenced in the present work.

B. subtilis chromosomal DNA was purified as described by Saunders *et al.* (1984).

Sequencing data: directed vs *de novo* sequence assembly. Two strategies might have been used to assemble the *B. subtilis* 168 genome sequence. On the one hand, the assembly could have been achieved via a directed assembly procedure which would have used the original sequence of strain 168 as a template to orient and organize the contigs produced by the assembler software. However, this approach, which *de facto* presumes collinearity between the reference sequence and the newly sequenced material, would have certainly prevented the uncovering of large genomic rearrangements, if any. On the other hand, during a *de novo* assembly procedure, contigs are assembled but their respective orientation and organization cannot be determined, unless paired-end reads are used. Although in this case no assumption is made for the organization of the scaffold, this may lead to the construction of chimaeric contigs. Even though both problem types can be solved easily via PCR experiments, problematic areas on the scaffold had to be pinpointed first. We thus processed the sequencing data as follows.

A total of 346 189 (average length 230 nt) valid single reads were produced using Roche/454-GSFLX technology (Margulies *et al.*, 2005) representing approximately 19× coverage of the final molecule. *De novo* assembly was then performed with the Newbler 2.0 software, leading to 51 contigs (larger than 500 nt) for a total length of approximately 4.2 Mb. *De novo* assembled contigs were remapped using the Mummer3 software (Kurtz *et al.*, 2004) onto the reference sequence and no rearrangement or chimaera cases were revealed since all of them aligned perfectly with the original strain 168 sequence. A further stack of 1542 reads were generated by sequencing PCR products to validate the contig organization and orientation as well as to fill potential remaining sequence holes which did not encompass the location of rDNA clusters.

As they correspond to repetitive regions, rDNA clusters were logically discarded during the assembly procedure and their sequences were determined using a sequencing technology which generates longer fragments than 454-GSFLX. All but one rDNA cluster were successfully resequenced using Sanger technology. The large cluster comprising the three rDNA *H*, *G* and *I* clusters could not be properly resolved (i.e. one rDNA cluster was found instead of three). This raised two hypotheses: (i) a significant genomic rearrangement occurred during *B. subtilis* 168 evolution (e.g. recombination between clusters *I* and *G*) or (ii) PCR could not be successfully performed due to particular DNA secondary structures in this region (stem–loop structures). However, since spacers between rDNA subunits often contain tRNAs, it was possible to validate or invalidate one of the hypotheses by searching in sequence reads for tandems or, if any, triads of tRNAs that are characteristic of the large cluster *rrnIHG* (see Supplementary Table S1, available with the online version of this paper). One such triad theoretically exists (tRNA*arg*–tRNA*gly*–tRNA*thr*), but its length (370 bp) greatly exceeds the average length of reads produced by pyrosequencing in the present work. Luckily, tandems tRNA*arg*–tRNA*gly* and tRNA*gly*–tRNA*thr*, which are still specific to the *rrnIHG* cluster, were found in 42 reads out of 305 containing two complete tRNAs. This strongly suggested that PCR experiments failed for that particular region; hence we incorporated the original sequence of *rrnIHG* in our final assembly.

Consensus sequence correction. To identify possible sequencing errors, a script based on the ssaha2 aligner (Ning *et al.*, 2001) was written. This script takes as input the reference molecule as well as the sequencing data (i.e. sequence reads and associated base quality), remaps reads onto the reference sequence and reports a list of potential variations between this reference sequence and the newly sequenced material. The events reported can be either simple substitution or insertion or deletion events. This analysis led to the detection of more than 2000 potential differences (1589 substitutions and 578 indels) between the two *B. subtilis* genomes. All the positions spotted by this approach were manually checked using the consed software (Gordon *et al.*, 1998) and a large proportion of them turned out to be true variations.

However, it is now well known that pyrosequencing technology has trouble in correctly resolving homopolymer runs (Huse *et al.*, 2007). For this reason, we performed two further Solexa runs (Bentley, 2006), which produced 3 214 055 single reads for a 27× coverage of the molecule. Using an alignment process similar to that used for GSFLX reads, 42 positions only could be corrected. This indicated that GSFLX performed remarkably correctly and that the remaining variations were likely to be true.

Annotation procedures. Gene prediction was performed using the AMIGene software (Bocs *et al.*, 2003), using the gene models built with the published version of *B. subtilis* annotations. The predicted coding sequences (CDSs) were assigned a 'locus_tag' similar to that of the previous *B. subtilis* annotations, i.e. 'BSUxxxxx': the CDS number remained the same for identical genes, even in the case of variation in the total length of the alignment between the old and new versions of a gene (corresponding CDS shorter or longer). In the case of additional gene predictions and of fissions or fusions, we used the remaining numbers (1 to 9, as the *B. subtilis* CDSs were numbered from 0 to 90 by tens). For example, the gene fission BSU33220 is now replaced by two new locus_tags: BSU33221 + BSU33222. The fusion observed with BSU01840 + BSU01850 now corresponds to the CDS BSU01845 (see Supplementary Table S2). The sets of predicted genes were submitted to automatic functional annotation, as described previously (Vallenet *et al.*, 2006). The initial functional assignation was based on the transfer of the *B. subtilis* annotations (Moszer *et al.*, 2002) between gene products presumably differing as the result of sequencing errors, i.e. 85 % identity on at least 80 % of the length of the smallest protein. Sequence data for comparative analyses were obtained from the NCBI database (RefSeq section, http://www.ncbi.nlm.nih.gov/RefSeq). Putative orthologues and synteny groups (i.e. conservation of the chromosomal colocalization between pairs of orthologous genes from different genomes) were computed between the newly sequenced genome and all the other complete genomes as described previously (Vallenet *et al.*, 2006). Each CDS was individually analysed by searching for its name (and the name of synonyms) in the PubMed and PubMed Central data libraries, as well as in Google. PubMed identifiers (PMIDs) were included in the MaGe platform as an information validating the status of the annotation, with particular emphasis on experimental validation of activity. All

these data (syntactic and functional annotations, and results of comparative analysis) were stored in a relational database, called BacilluScope. A complete manual validation of the automatic annotation by several users in different locations was performed using the MaGe web-based interface.

Finding regions of genomic plasticity. Potential genomic islands in the *B. subtilis* chromosome were searched for with the RGP (region of genomic plasticity) tool implemented in MaGe (Vallenet *et al.*, 2006), which is based on the synteny breaks between genomes compared in parallel (here *B. subtilis*, *Bacillus amyloliquefaciens* FZB42, *Bacillus licheniformis* ATCC 14580 and *Bacillus pumilus* SAFR-032). The predicted regions were checked manually and compared to the chromosomal regions of potentially foreign origin previously detected using hidden Markov models (Nicolas *et al.*, 2002) and Multiple PairWise Test (Merkl, 2004) (Supplementary Table S3). It appeared that these genomic regions sometimes had a composite structure, e.g. they were made of regions partially conserved or found in different synteny groups (i.e. in different genomic locations) in the different *Bacillus* genomes. The predicted RGPs have therefore been further manually curated to define subregions called modules.

Metabolic reconstruction. Based on the new functional annotations of the *B. subtilis* genome, the reconstruction of metabolic pathways at work in this bacterium was performed using the Pathway Tools software suite (Paley & Karp, 2006). We first built a BioCyc Pathway/Genome database, which linked annotation data to MetaCyc data, a set of canonical metabolic pathways (Caspi *et al.*, 2008). Then a number of specific sections of the metabolism went through manual curation (see Results and Discussion). Data compiled in the framework of an independent metabolism modelling project were also used, in collaboration with their authors (Goelzer *et al.*, 2008). Altogether, the SubtiliCyc database produced in this way comprises approximately 280 metabolic pathways (both curated and automatically inferred) and 1300 reactions.

RESULTS AND DISCUSSION

Distribution of variants between the present sequence and the consortium sequence

Fig. 1 summarizes the distribution of single nucleotide polymorphisms (SNPs) and other variants along the reference sequence together with coloured boxes spanning the regions that have been assigned to the different groups of the consortium. There is clearly a highly non-random distribution of variations, which correlates with the different regions corresponding to different groups. First, we noticed several regions with only very few differences from the present sequence, and these are essentially very small variations (~0.860–0.950 Mb; 2140–2310 Mb; ~3480–3710 Mb). Comparing with variations in other related genomes, or in regions independently sequenced by various groups in the course of their own specific work on a gene or a group of genes, these variations appear most often not to be SNPs, but sequencing errors (see below). Some regions are heterogeneous, with subregions devoid of errors and subregions with a substantial amount of variations: typically this fits well with the way the sequencing programme has been organized at its onset in some laboratories, with pieces of 15–20 kb sequenced by students, who were obviously not all able to ensure the same level of accuracy (Glaser *et al.*, 1993). We also noted that some regions, which have not been resequenced by the consortium but taken as published at the time, carry a higher level of errors (this is the case for ribosomal protein operons; for example see the region around 0.120 Mb, where there is a concentration of variations, while the upstream region is almost without errors).

B. subtilis 168 is a laboratory strain which is highly competent for DNA uptake. Its popularity originates from the ease with which it can be genetically manipulated; the exact origin of its increased competence is not well established, as it may well have derived from some physiological change when the wild-type Marburg parent strain was domesticated at Yale University (for reviews, see Earl *et al.*, 2008; Zeigler *et al.*, 2008). This strain and its derivatives have been further subjected to extensive mutagenesis by irradiation or chemical mutagenesis (e.g. using nitrosoguanidine or ethyl methanesulfonate) to obtain a large number of genetic markers required for the establishment of a genetic map. Many additional mutants of *B. subtilis* 168 were also constructed and exchanged between laboratories for other research purposes, such as the study of different cell processes, including sporulation, competence, cell division, metabolism, motility, chemotaxis, swarming, etc. The overall result is that most laboratory derivatives of *B. subtilis* 168 contain chromosomal segments that may differ from the ancestral parental clone of strain 168 whose genome has been sequenced. This may account for some of the differences between the initial (uncorrected) sequence and the revised sequence presented here. However, comparison with the many genomes of Firmicutes we now possess, permitting comparison of conserved regions in proteins, strongly argues against polymorphism generated in the various laboratories of the consortium, and supports the idea that most if not all variations between the published sequence and the present one are sequencing errors (some of which derived from cloning artefacts). Our present work is therefore an update of the sequence annotation of *B. subtilis* strain 168.

Global features of the genome and the proteome

Genome programmes use the established genome sequence to list the major genomic objects, protein and RNA-coding genes, together with general features, including functional annotation. To this standard list (available in BacilluScope and in GenoList) we have now added new features such as riboswitches, small regulatory RNAs [srRNAs, including a series of seven srRNAs recently identified (Saito *et al.*, 2009)] as well as more elaborate features: genomic islands, local codon usage biases, protein amino acid composition and distribution along the chromosome.

Overall the systematic reassessment of syntactic gene locations and the integration of sequence corrections led to many feature updates, summarized in Table 1. A detailed listing of modified genes is given in Supplementary

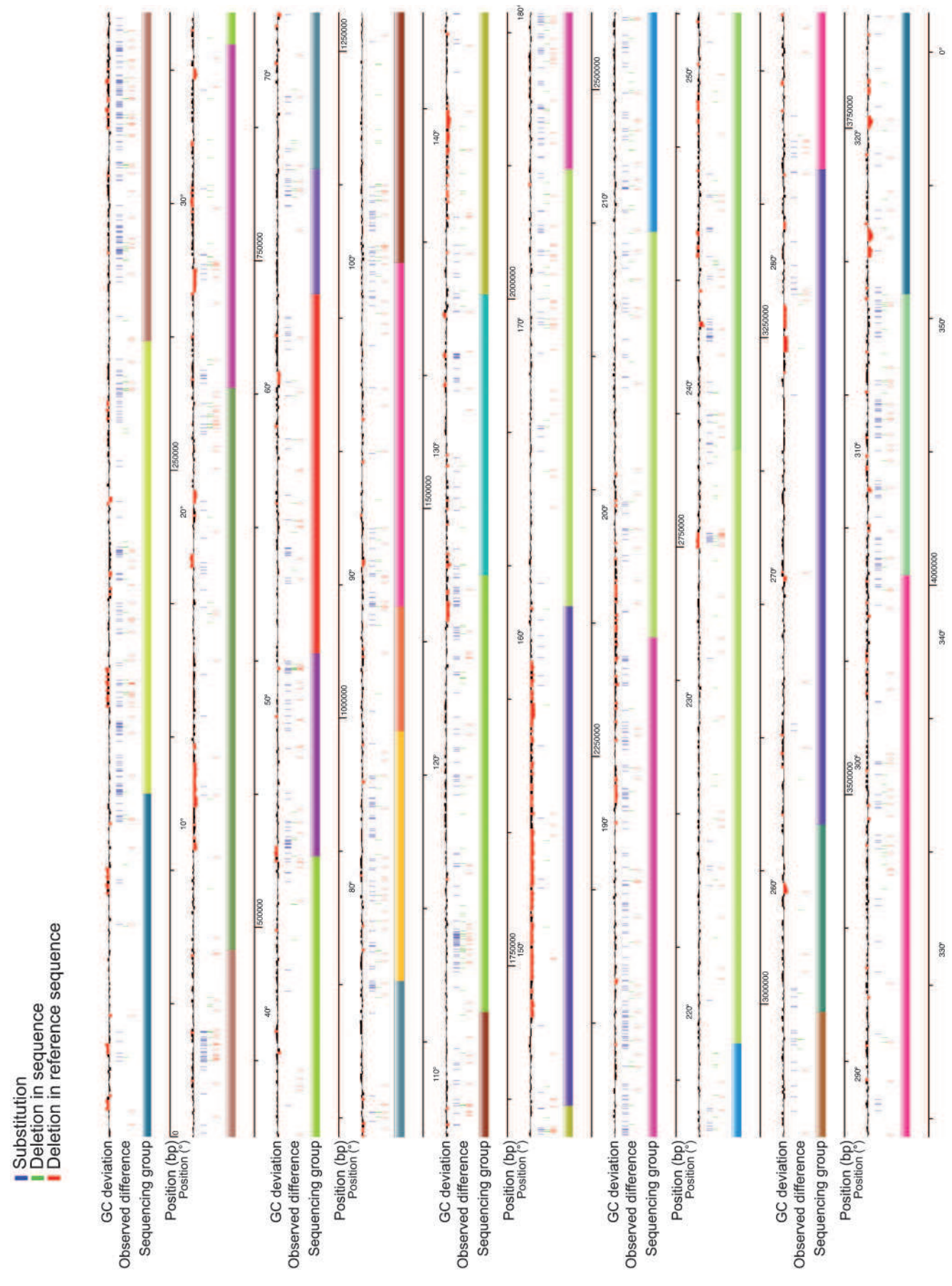

Substitution
Deletion in sequence
Deletion in reference sequence
GC deviation
Observed difference
Sequencing group
Position (bp)
Position (°)

Fig. 1. Comparison between the previously published sequence of strain 168 and the strain resequenced without cloning. SNPs and indels are as indicated, as well as the uneven distribution of G+C nucleotides in the sequence. Under the line representing the genome are displayed the positions of the different regions attributed to the various members of the sequencing consortium. It can be seen that the amount of variation is dependent on the sequencing group, not on the nucleotide composition of the genome. In some regions there is precious little variation, compared with the present sequence, despite the fact that the techniques used between 10 and 20 years ago were very different from those used today.

Table 1. Comparing the new sequence with the old one

Gene comparisons between the old and the new version of the annotations	No. (per cent) of genes*
Identical genes	3323 (78.3 %)
Amino acid variations	426 (10.0 %)
Adjusted start codons	50 (1.2 %)
C-terminal variations only	221 (5.2 %)
N-terminal variations only	4 (0.09 %)
C-terminal and N-terminal variations	11 (0.26 %)
Fusions	20 (0.47 %)
Fissions	20 (0.47 %)
Newly annotated genes	171† (4.0 %)

*In the case of fusion/fission events, the number of new genes resulting from the event is indicated.
†Including 48 pseudogenes or gene remnants.

Table S2. A fairly large number of new, and essentially small genes were annotated (171), amongst them approximately 30 % of pseudogenes or gene remnants, and a few of them resulting from fusions or fissions of previously existing genes. Several sequence corrections corresponded to compensating frameshifts, leaving gene boundaries untouched but generating new and better-conserved internal amino acid motifs. Being short, and resulting in an apparently unbroken open reading frame, errors of this type could not have been detected by the procedure meant to identify possible regions in error (Medigue *et al.*, 1999).

Correction of errors and manual annotation of small CDSs was analysed by investigating the gene size distribution in the genome (Fig. 2a). The distribution is quite even, and does not suggest an abnormal distribution of a particular class of lengths, which is the hallmark of spurious CDS identification (Yamazaki *et al.*, 2006).

The relatively small number of error corrections in the present sequence (~2000, see Methods), relative to the length of the genome implies that the overall analysis of words in the genome does not significantly deviate from our previous studies. In particular, *B. subtilis* is remarkably poor in repeats longer than 25 nt (Rocha *et al.*, 1998, 1999a). As in other bacterial genomes, the constraints placed on the processes of translation initiation and termination result in local compositional biases (Rocha *et al.*, 1999b). Using systematic comparison with CDSs extracted from other genomes (Firmicutes, and *Bacillus* species most often), as well as prediction of exported protein signals, if necessary, we tried to reassign translation start sites in all CDSs. Most are unambiguous, with an excellent ribosome-binding site (RBS, variation on AAGGAGGT) located 4–13 nt upstream of the ATG start codon. In some cases, it is difficult to be absolutely sure of the correct site. In a few cases (*infB*, *lysC*, *pgsB*), a given open reading frame harbours two or more authentic CDSs. This fact needs to be remembered when typical RBSs with a correct start site are found within a long CDS, as other similar cases might have been overlooked.

Our previous work had shown that the genome is heterogeneous when considering many features such as base composition, codon usage biases or functional consistency, with up to ten islands ascribed to prophages or prophage remnants (Bailly-Bechet *et al.*, 2006; Kunst *et al.*, 1997; Moszer *et al.*, 1999). Using the 'RGP-Finder' module in the MaGe annotation platform, which allows one to search for regions of genomic plasticity in bacterial genomes, we compared *B. subtilis* with *B. amyloliquefaciens* FZB42, *B. licheniformis* ATCC 14580 and *B. pumilus* SAFR-032. Eighty genomic regions (RGPs) ranging from 5171 nt to 141 234 nt were identified in *B. subtilis* 168 (Supplementary Table S3). Among those, 22 regions covered or overlapped with islands predicted using hidden Markov models (Nicolas *et al.*, 2002) and Multiple PairWise test (Merkl, 2004). These sequences include all of the 10 well-documented genomic regions of phage origin (Kunst *et al.*, 1997), of which only three are

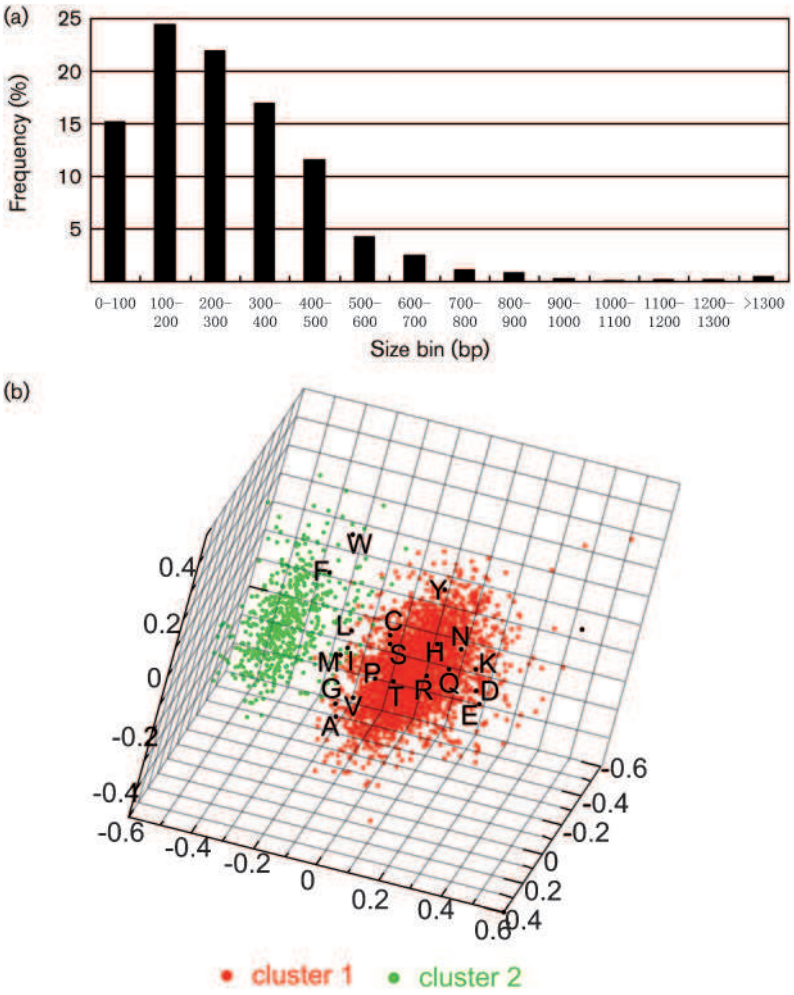


Fig. 2. (a) Distribution of gene length in the *B. subtilis* 168 genome. The absence of any overrepresentation of short CDSs supports the view that most if not all gene sequences predicted in the present annotation are authentic. (b) Correspondence analysis of the proteome of *B. subtilis*. Proteins in the proteome can be separated into two well-identified classes. The green cloud corresponds to proteins that are integral inner-membrane proteins (IIMPs). Note that the IIMP cloud is driven by the opposition between charged amino acids (D, E and K) and hydrophobic ones (F, L, M, W).

integrated at a tRNA gene location and/or contain mobility genes. However, apart from the PBSX phage and region P4, these prophagic regions harbour a significant GC deviation (Fig. 3). Six other predicted RGPs contain at least two specific genomic island features [tRNAs, integrases, mobility-associated genes and pseudogenes (Dobrindt *et al.*, 2004)] and can be named GI-like (genomic island-like). This is the case for GR17, which is made up of two modules (see Methods), one completely specific to *B. subtilis* (three genes: sporulation control gene and two hypothetical proteins), and one absent in *B. pumilus* only (several genes involved in sulfur transport and metabolism). The rest of the predicted RGPs were found using the synteny break point criteria only (Supplementary Table S3), and most often these regions harbour genes coding for enzymic activities and/or transporters. For example, GR48 is made of two modules distinct in terms of functional role: the first one is involved in rhamnogalacturonan transport and utilization, and it is absent in *B. amyloliquefaciens*; the second one contains genes coding for biotin synthase and lysine-8-amino-7-oxononanoate aminotransferase, and it is absent in *B. pumilus*.

Using correspondence analysis we have reinvestigated the distribution of amino acids in the proteome. As reported previously (Pascal *et al.*, 2005), this allowed us to create a list of plausible integral inner-membrane proteins (IIMPs, Fig. 2b and Supplementary Table S4). We also annotated exported proteins with signal peptides and lipoprotein signal peptides.

An overview of the cell's organization

Three major processes permit the development of life: sustaining life while combating ageing, propagating life, and living in a particular environment. The first two processes require presumably ubiquitous functions. The third one corresponds to large pools of horizontally transferred genes that are shared by the individual strains of a given species. Ubiquitous functions cannot be derived in any straightforward way from genome comparisons, as they often result from dissimilar structures recruited in the course of evolution. However, the structure of the descent of living organisms implies that there is a tendency for a gene coding for a particular function to be conserved over

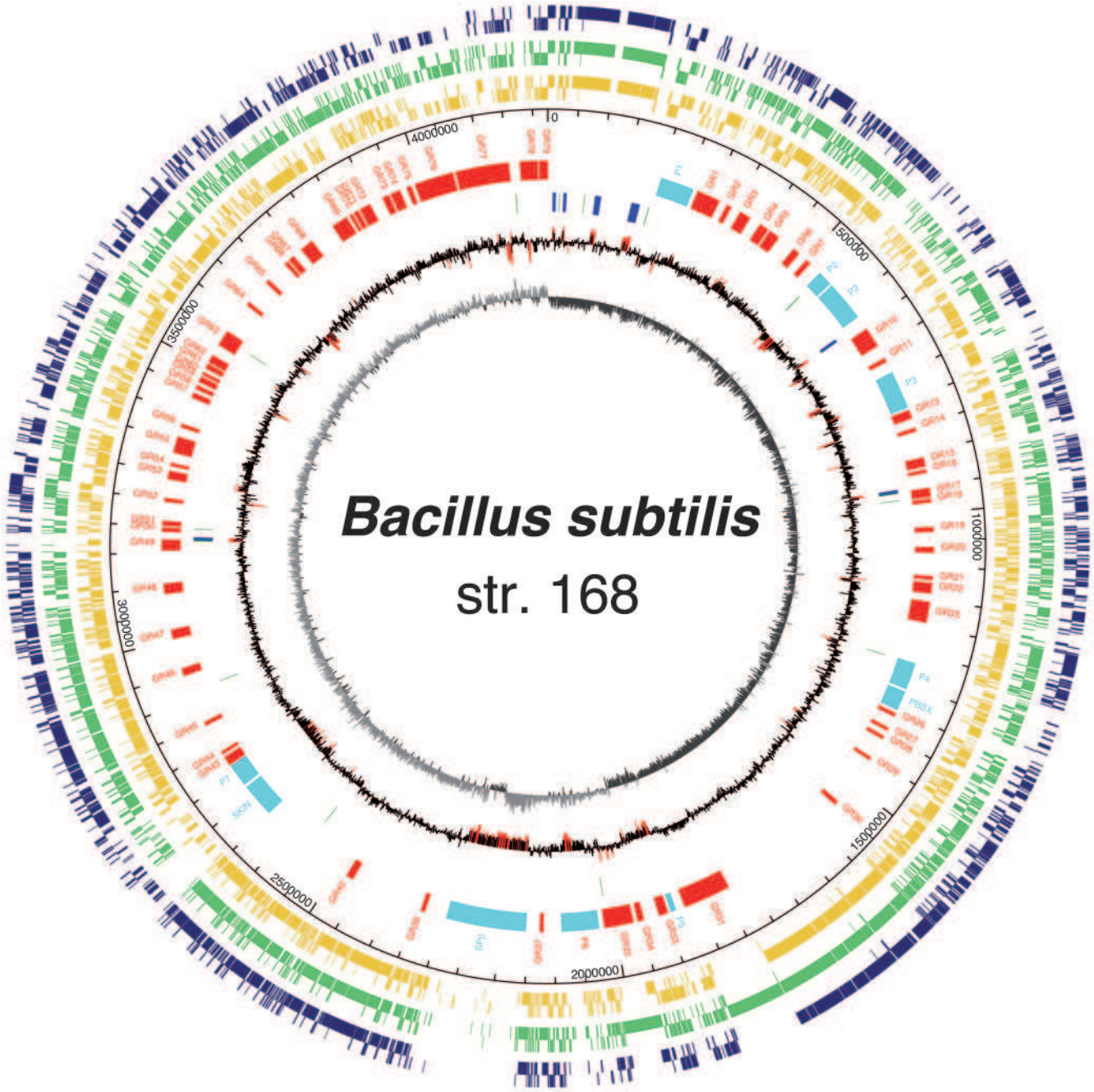


Fig. 3. Circular representation of the *B. subtilis* 168 genome for several specific genome features. Circles display the following, from the inside out. (1) GC skew (G+C/G−C using a 1 kb sliding window). (2) GC deviation (mean GC content in a 1 kb window − overall mean GC). Red areas indicate that deviation is higher than 1.5 standard deviation. (3) tRNA (dark green) and rDNA (blue). (4) Location of genomic regions with specific features differentiating them from the average sequence. Boxes coloured in light blue indicate regions of phage origin. The nonsymmetrical distribution (right and left halves of the circle) is to be emphasized. (5) Scale. (6, 7, 8) Genes having a presumed orthologue in other *Bacillus* species (*B. licheniformis*, *B. amyloliquefaciens* and *B. pumilus* respectively).

generations, leading not to ubiquity but to significant persistence. Comparative genomics of bacterial genomes identified a set of persistent genes in two major bacterial clades, the gamma-Proteobacteria (with *E. coli* as the model) and the Firmicutes (with *B. subtilis* as the model organism) (Fang *et al.*, 2005). Detailed analysis of conservation of proximity of genes in genomes showed that both persistent genes and rare genes tend to stay

clustered together, making two highly consistent families of genes, separated by a large twilight zone (Danchin *et al.*, 2007).

Remarkably, the connection network of the persistent genes coding for ubiquitous functions is reminiscent of a scenario of the origin of life, forming the *paleome* (from παλαιος, ancient). Two further splits must be made among the functions of the paleome. Some are essential for permitting formation of a colony on plates supplemented by rich medium (Kobayashi *et al.*, 2003); some, while coded by persistent genes, do not have this property (Fang *et al.*, 2005). This particular feature has to be superimposed on a third split, which separates reproduction from replication (Dyson, 1985). Taken together, this view of the paleome opens a novel way to consider genomes and evolution, where management of the creation of information is the central issue when a young organism is born from an aged one. Repeated invention of energy-dependent processes required to make room while accumulating information in a ratchet-like manner probably accounts for the remarkable diversity of the structures involved in the process (Danchin, 2008). The organization of the paleome makes a separation between the machine (which is compartmentalized and sustains metabolism), and the program (which replicates and is expressed both constitutively and under specific conditions).

Finally, bacteria need not only to survive and to perpetuate life, but also to occupy a particular niche. This capability corresponds to a very large class of genes, forming the *cenome* [after κοινος, common, as in biocenose (Danchin *et al.*, 2007)].

We therefore explored the *B. subtilis* genome sequence considering first the genes involved in making, maintaining and repairing the cell, the paleome; and then its cenome, allowing the cell to occupy a specific niche, which defines the features of the organism's biotope as well as those used in industrial applications for instance.

Compartmentalization

The cell membrane. Phospholipid synthesis and turnover is managed by a variety of processes, encoded by genes *cdsA*, *des*, *dgkA*, *fapR*, *glpQ*, *gpsA*, *lipC*, *mprF*, *pgsA*, *phoH*, *plsC*, *tagA*, *yodM*, *ytlR* and *ytpA*, some of which are essential (Kobayashi *et al.*, 2003). The important process of distribution of phospholipids in the outer layer of the membrane is performed by flippases, which have been at least partially characterized (EpsK, SpoVB, YgaD, YwjA).

The cell wall and the cell shape. The shape of the cell is determined by a variety of processes, combining an internal cytoskeleton coded in particular by *mre*-related genes (for a recent review see den Blaauwen *et al.*, 2008) with synthesis of the murein sacculus (Hayhurst *et al.*, 2008) and its associated teichoic acids (Formstone *et al.*, 2008). The *fts* *mur* islands are organized in a way compatible with the overall bacillus shape of the organism, as remarked by Tamames *et al.* (2001) in an attempt to correlate the architecture of the cell with the organization of the genome (Danchin, 2009b).

Transport. The genome of *B. subtilis* includes four major classes of transporters: ABC-transporters driven by ATP hydrolysis, phospho*enol*pyruvate-dependent transport systems (PTSs), electrochemically driven permeases (importer and antiporter involving charged substrates) and facilitators. While some have been experimentally analysed [in particular all those related to sucrose transport (Fouet *et al.*, 1987)] many have been annotated by inference (Saier *et al.*, 2002). It should be stressed here that the utmost caution should be exerted when using purely *in silico* analyses in this domain, as it is quite difficult to distinguish related metabolite transporters, including ions. Furthermore there is sometimes fairly wide specificity in the nature of the transported metabolites [e.g. the genes named *tcyJKLM* at the present time transport substrates considerably deviating from cysteine alone (Burguiere *et al.*, 2004; Sekowska *et al.*, 2001)].

Information transfer

Replication, recombination and repair. Split into three phases, initiation, elongation and termination, these processes are among the best-studied features of the organism (Frenkiel-Krispin & Minsky, 2006; Noirot-Gros *et al.*, 2002) and we did not expect to find much new information from sequence annotation. There is a strong bias of gene expression in the leading strand, and this correlates not only with the presence of two DNA polymerases (DnaE and PolC) as already noticed (Rocha, 2002), but also with that of a series of genes, some of which are of unknown function. Besides this remarkable fusion of two DNA replication apparatuses apparently coming from different origins, an interesting observation was that we repeatedly found association between part of sulfur metabolism and RNA (DNA) degradation in sets of genes that are often associated with poorly understood functions. This substantiates the observation that nanoRNase NrnA (YtqI) is also a 3′-phosphatase that hydrolyses 3′,5′-adenosine bisphosphate, a product of sulfur assimilation, into 5′-AMP (Mechold *et al.*, 2007).

Transcription and translation. Both these processes also follow the standard course: initiation, elongation and termination. Most steps involved in transcription in *B. subtilis* have been analysed for a long time, in particular transcription initiation (Haldenwang, 1995; Kazmierczak *et al.*, 2005; Kunst *et al.*, 1997). We wish however to emphasize the coupling between transcription and DNA repair, which probably needs to be explored much further than the identification of Mfd, the transcription repair coupling factor.

Much progress has been made at the level of translation. Beside a sequence error in protein S12 of the ribosome, previously identified (Carr *et al.*, 2006), a feature of the ribosome is worth noticing. Indeed, several genes code for homologues of previously identified ribosomal proteins, suggesting either involvement during maturation of the ribosome, or involvement in spore ribosomes: *rpmGA*/*rpmGB* (L33), *rplGA*/*rplGB*(*ybxF*) (L7A), *rpsNA*/*rpsNB* (S14) and *rpmEA*/*rpmEB* (L31). A newly identified ribosomal protein, somewhat similar to protein L14E in Archaea, and present in many Firmicutes, including *Clostridium* species, appears to be coded by gene *ybzG*. Ribosomal protein S12 is thiomethylated at a conserved aspartate (which was a – wrong – asparagine in the published sequence) by protein RimO (YqeV) (Anton *et al.*, 2008).

Several genes code for methylases and acetyltransferases, probably involved in ribosomal protein modification, suggesting that much more work must still be performed on the process of translation in *B. subtilis*. Ribosome assembly is also directed by a variety of energy-dependent proteins such as the YsxC GTPase. Several ATP- or GTP-dependent enzymes may be involved in the process, especially during temperature shifts.

The process of translation also involves proper folding of nascent proteins (Tig and PpiB prolyl isomerases play an important role in the process) as well as degradation of incomplete or chemically altered proteins. We note that we did not find a counterpart for the system repairing isoaspartate in other organisms, suggesting either that its counterpart belongs to the proteins of still unidentified function, or that there is an efficient degradation pathway recognizing isoaspartate (Danchin, 2008).

Anabolism and salvage

Coenzymes. *B. subtilis* synthesizes all major co-enzymes or prosthetic groups found in free-living bacteria except for coenzyme B12. We shall only stress here features that are original to this bacterium or very recently characterized.

The synthesis of biotin in *B. subtilis* is somewhat unusual. It uses a transaminase with lysine as the amino-group donor, not *S*-adenosylmethionine (AdoMet) as in reference pathways (Van Arsdell *et al.*, 2005).

The metabolism of pyridoxal phosphate is original in *B. subtilis*, as part of the pathway is not similar to that in Bacteria such as *E. coli*, but is rather highly similar to that found in plants and fungi (Raschle *et al.*, 2005).

The general metabolism of thiamin has not been completely unravelled. An alternative pathway to the standard route exists in *E. coli*, but it is not understood. Begley and co-workers have shown that in addition to *de novo* synthesis many salvage pathways exist to scavenge thiamin precursors or derivatives from the environment. The TenA–TenI system associated with regulating the production of extracellular proteases is in fact a widespread salvage system (Begley *et al.*, 2008; Jenkins *et al.*, 2008).

Most of the steps involved in menaquinone/ubiquinone biosynthesis can be identified in the sequence of the genome. However, the counterpart of *ubiG* is not easily uncovered among the possible genes of the pathway. Gene *yrrT*, located upstream of *mtnN* and genes involved in scavenging homocysteine (Andre *et al.*, 2008) could code for the corresponding function yielding *S*-adenosylhomocysteine, but more work needs to be performed to challenge this hypothesis (A. Sekowska, unpublished observations).

Carbon metabolism. *B. subtilis* displays a textbook organization of its carbon metabolism (for reviews see Sauer & Eikmanns, 2005; Sonenshein, 2007). The phenomenon of catabolite repression has been studied in much detail (for recent references see Singh *et al.*, 2008). While the main catabolite repressor CcpA has been identified, the process is far from being completely unravelled (the function of most genes conserved in synteny with *crh* is not understood yet). We hope that the tracks suggested by our annotation will help further discoveries in the domain.

We must also note that some genes involved in carbon metabolism are essential for unexpected reasons. This is the case for *eno*, *fbaA*, *pgm*, *tkt* and *tpi* in the main glycolytic pathway and *odhAB*, coding for ketoglutarate dehydrogenase, for example. Interestingly, analysis of co-evolution of RNase genes suggests the existence of a degradosome in *B. subtilis*, which would be functionally linked to the main glycolytic pathway (co-evolution with *tpi* and *eno* in particular) (Danchin, 2009a).

Nitrogen metabolism. The best nitrogen source for *B. subtilis* is glutamine. We have observed that bacteria grown in minimal medium with ammonium as a nitrogen source grow slowly and evolve rapidly to fast growers (Sekowska, 1999 and unpublished observations). This fits with the demonstration by Belitsky & Sonenshein (1998) that strain 168 can adapt to rapid growth on ammonium or glutamate by a reversible spontaneous duplication or deletion of a 9 bp sequence in the alternative glutamate dehydrogenase gene, *gudB*. This feature may have had significant consequences in the sequencing project by creating unwanted polymorphism, as some members of the consortium were familiar with growth on ammonium, unaware of the possibility that this original phase variation might have consequences in terms of mutation selection. Nitrate can also be used as a nitrogen source, as *B. subtilis* has both a respiratory and an assimilatory nitrate reductase (Nakano & Zuber, 1998). Several salvage pathways for purines and pyrimidines (Christiansen *et al.*, 1997; Tozzi *et al.*, 2006), including salvage of energy-rich nucleotides (Danchin, 2009a), exist in the organism. Adenine deaminase AdeC has activity demonstrated in the purine salvage pathway. The *yerA* paralogue could be a missing dihydropyrimidinase. It is also required for scavenging derivatives of AdoMet. There is another link between

nitrogen (via arginine and polyamines) and sulfur metabolism (Sekowska *et al.*, 2001). Further, cysteine protects ArgG against reactive oxygen species (ROS) (Hochgrafe *et al.*, 2007), thus leading to arginine derepression under conditions of excess cysteine or ROS production.

In diaminopimelate synthesis, there is a requirement for the conversion of *N*-acetyl-2,6-diaminopimelate to acetamido-6-oxoheptanoate. Since this reaction belongs to non-essential reactions in *B. subtilis*, there are probably several genes that encode proteins showing this enzymic activity. Transaminase PatA could be a candidate for this activity, as it is located in a region coding for activities involved in chemotaxis, sporulation and control of cell envelope synthesis. Unfortunately, the biochemical data collected on this enzyme only show that it is not involved in methionine transamination (Berger *et al.*, 2003). Other enzymes such as SpsC or NtdA might also display some of the missing activity. Particular attention should be paid to *N*-acetylornithine aminotransferase (EC 2.6.1.11), encoded by gene *argD*. This gene might need better characterization of its enzyme activity, as it might display two related functions, involving both the metabolism of arginine and the metabolism of diaminopimelic acid. Indeed, in *E. coli*, ArgD has both functions despite the fact that the diaminopimelate substrate is succinylated, not acetylated. This would make the *B. subtilis* enzyme particularly prone to have both activities. This might account for the particular feature of the transcription of this gene observed in transcriptome experiments (Sekowska *et al.*, 2001). Biochemical work is needed to substantiate this point.

Sulfur metabolism. Specificities of sulfur metabolism as established from the *B. subtilis* sequence have been described in a review article (Sekowska *et al.*, 2000). However, with the new complete annotation of the genome, novel features emerge that are worth describing here. While synthesis of cysteine and methionine is fairly standard for a Gram-positive organism, the reverse transsulfuration pathway, allowing growth on methionine, is unusual and not completely understood yet (Hullo *et al.*, 2007). Methionine salvage results from recycling of catabolites of AdoMet and from recycling of the start of proteins where the second residue is small, via two methionine aminopeptidases, MapA and MapB (You *et al.*, 2005). AdoMet is mainly used to transfer methyl groups (52 methyltransferases) and results in the production of *S*-adenosylhomocysteine, which is further metabolized by MtnN into *S*-ribosylhomocysteine and adenine. AdoMet is used as the precursor of polyamine after decarboxylation, yielding methylthioadenosine, which is recycled via methylthioribose (MTR) by a complete methionine salvage pathway where the carbon atoms of methionine derive from the ribose moiety, not from TCA (tricarboxylic acid) cycle intermediates (Sekowska *et al.*, 2004). AdoMet is used to make queuosine, a complex modified base near the anticodon of several tRNAs (gene *yqeE*). It is also used as a radical in several reactions coded by genes *bioB*, *hemN*, *hemZ*, *kamA*, *moaA*, possibly *skfB*, *splB*, possibly *ycnL* for a membrane protein, *yfkA* (fused from *yfkA* and *yfkB* in the 1997 sequence), *yloN*, *ymcB*, *yutB*(*lipA*), possibly *yuzB* and *yydG* (Frey *et al.*, 2008). The resulting product of the reaction, 5′-deoxyadenosine, might perhaps be recycled by the MTR salvage pathway (Sekowska *et al.*, 2004).

Genes involved in formation of iron–sulfur clusters have been predicted (Sekowska *et al.*, 2000) and some of them have been discovered experimentally (Kiley & Beinert, 2003). Analysis of the genome predicts several genes involved in the process of construction of iron–sulfur clusters: *nifS*, *yutI*, *iscU*(*yurV*), *sufC*(*yurY*), and possibly *ygaC* and *yneR*.

General maintenance and protection

Reactive oxygen species, nitric oxide. *B. subtilis* possesses three superoxide dismutases, including one exported lipoprotein. As described above, sulfur metabolism appears to be important for protection against ROS. A nitric oxide (NO) synthase (Chartier & Couture, 2007) uses YkuN and YkuP flavodoxins for electron transfer (Wang *et al.*, 2007). The NsrR response regulator monitors NO (Nakano *et al.*, 2006), suggesting a specific role of this gas in protection against ROS.

Two methionine sulfoxide reductases, MsrA and MsrB, are involved in oxidized methionine repair in a process that is at least in part regulated by the novel regulator Spx (see below) (You *et al.*, 2008).

Temperature. The *B. subtilis* niche implies considerable fluctuations in temperature. The organism harbours a standard arsenal of cold-shock proteins and RNA helicases. It has also two thymidylate synthases, ThyA and ThyB, permitting it to grow at temperatures as high as 55 °C (Montorsi & Lorenzetti, 1993).

Salt. *B. subtilis* is fairly resistant to desiccation and, in parallel, to fairly high levels of sodium. Some experiments suggested that this was mediated by the DegS–DegU system (Dartois *et al.*, 1998) and the sigma-B regulon (Petersohn *et al.*, 2001). However these systems are very general and might not be specific in the process. A significant part of the corresponding adaptation is mediated by synthesis of glutamate as a precursor of proline and may interfere with formation of iron–sulfur clusters (Hoper *et al.*, 2006). In any event, synthesis and uptake of compatible solutes (such as glycine betaine) permit adaptation to high osmolarity.

Regulation

Genes do not operate in isolation: RNA polymerase activity is regulated at a global level by sigma factors and anti-sigma factors (for reviews see Gruber & Gross, 2003; Helmann, 1999, 2002; Kazmierczak *et al.*, 2005; Kroos &

Yu, 2000; van Schaik & Abee, 2005). In some cases there is a coupling with the ribosome via the Nus factors and specific factors such as the ribosome-associated sigma-54 modulation protein YvyD. Integration and sensing is associated in two-component systems, which have been widely discussed previously (see for example Bisicchia *et al.*, 2007; Joseph *et al.*, 2002; Kobayashi *et al.*, 2001).

Specific identification of the regulators (often small molecules) of putative transcription factors is difficult and in most cases it is still unknown. Progress is slow in this domain; yet the interesting case of regulators with pyridoxal phosphate (PLP) sites (*ycxD*, *gabR*(*ycnF*), *ydeF*, *ydeL*, *ydfD*, *yhdI*, *yisV*) suggests coupling between regulation and enzyme activity. Indeed, there are some situations where a regulator is directly involved in an enzyme activity (e.g. biotin biosynthesis: Chapman-Smith *et al.*, 2001) or associated with it (Tanous *et al.*, 2008).

Several regulators often display a global behaviour. This is the case for the Lrp family coded by genes *azlB*(*yrdG*), *lrpA*, *lrpB*, *lrpC*, *yezC*, *yugG*(*alaR*) and *ywrC*, global regulators CcpA, CodY, TnrA (Sonenshein, 2007) and Spx (Beck *et al.*, 2007; Reyes & Zuber, 2008; You *et al.*, 2008). A paralogue of the latter, MgsR (YqgZ), is a transcriptional regulator of stress and modulates the sigma-B response (Reder *et al.*, 2008).

Cyclic di-GMP is a regulatory molecule involved in many processes controlling collective behaviour in bacteria (Sinha & Sprang, 2006); three proteins (YdaK, YkoW and YtrP) have motifs that suggest synthesis of cyclic di-GMP in *B. subtilis*. They are sometimes associated with sites that suggest a role in sensing environmental cues (in YkoW, an additional PAS domain is found between the MHYT and GGDEF domains, suggesting a role in sensing dioxygen, carbon monoxide or NO); protein YybT is a phosphodiesterase-like protein, which has a modified GGDEF motif, suggesting that it could act as the phosphodiesterase involved in cyclic di-GMP control, and YpfA could also hydrolyse cyclic di-GMP, while YdaN is a regulator that has a site which could bind cyclic di-GMP and relay its gene expression control activity in particular during biofilm synthesis. In short, it seems likely that cyclic di-GMP plays a role in gene expression in *B. subtilis*.

Occupying a niche

The cell's standard phenotypes. Described earlier as *Vibrio subtilis*, *B. subtilis* was reproducibly identified by Zopf in 1885 as isolated from hay soaked in a small volume of water at 36 °C for 4 h, then filtered and boiled for 1 h. Most often, the process ended with a pellicle formed at the surface of water after incubation for 1 day at 36 °C. It was usually exclusively formed by *B. subtilis* bacteria.

In Bergey's manual, Sneath noted that 'It is generally not possible to draw any conclusions from the site of isolation of a *Bacillus* strain as to its natural habitat' (Sneath, 1986). However, repeated isolation from hay indicates that *B. subtilis* is an epiphyte, with the phylloplane (and the rhizoplane as a consequence) as the preferred niche. Identification of many genes in the genome strongly supports this observation, and in particular supports the idea that the surface of leaves is a preferred niche of the organism (Table 2).

Strain 168 is auxotrophic for tryptophan. This character is derived from some of the initial mutagenic events associated with improvement of the organism as a laboratory workhorse. For the same reason it is also lacking surfactin production while restoration of pseudogene *sfp*

Table 2. Genes of the cenome suggesting that *B. subtilis* is an epiphyte

Response to dioxygen	Response to light	Maceration of leaves	Unusual sulfur metabolism	Plant-related genes	Swimming, swarming and miscellanea
spx	*ytvA*	*yesLMNOPQRS*	*ytlI ytmI tcyJKLMN ytmO ytnIJ rbfK ytnLM*	*pdxS pdxT*	Locus *tnrE* locus *sfr*
perR		*rhgT yesUVWXYZ*	*yxeIJKLMNOPQ*	*pyrD pyrK*	*swrAA swrAB swrB yabR*
ypoP msrAB		*yetA lplABCD*	CymR and CysL regulons	*thrB thrC*	*aprE bpr mpr nprE vpr wprA*
hemAT mtnD		*pelC yoaJ*		*mtnW*	*yhfL salA sinIR tasA*
nosA		*xsa lacA xynA xynD yvfM araABDLMNPQ abfA araE yvfO araA abnA yxiA*			*ydaM yfiQ ykfABCD ylbF ymcA yoaW ypfA yqhH yqxM yulF ywqH yxaM yxjH yydFGHIJ*
ydfO yetH yodE yrkC yrpB yubC		*bglC bglH bglS*			*epsABCDEFGHIKLMNO*
catDE cdoA mhqA qodI		*yoaJ (exlX)*			*ecsB comP*
		Glucomannan utilization operon (*gmuBACDREFG*, formerly *ydhMNOPQRST*)			

into a functional frame allows bacteria to produce efficiently this molecule, acting both as an antibiotic and as a surfactant used for swimming and swarming (Julkowska *et al.*, 2005; Kunst *et al.*, 1997; Reuter *et al.*, 1999). Leaves often produce H_2S to avoid overaccumulation of sulfur: this may be scavenged directly by the bacteria, accounting for a complex sulfur metabolism network, enabling good growth on *S*-methylcysteine, for example. Consistent with the phylloplane as a preferred niche, *B. subtilis* grows best with vigorous aeration and this mistakenly placed it for a long time among the obligate aerobes (Nakano & Zuber, 1998). It has for this reason an efficient arsenal of genes combating the effects of dioxygen (including many dioxygenases). Interestingly, it is light sensitive, with a sensor, YtvA (Gaidenko *et al.*, 2006), involved in regulation of gene expression. Finally it has a considerable amount of genes involved in plant maceration, and it is able to swim, swarm and make complex biofilms, which is consistent with strong association with plant leaves.

Antibiotics and quorum sensing. As noticed previously, *B. subtilis* synthesizes a variety of complex molecules via the non-ribosomal peptide synthesis pathway as well as via maturation of peptides (Kunst *et al.*, 1997). These molecules play the role of antibiotics (and this probably accounts for the pure cultures isolated from hay infusions), and are also used as surfactants, permitting smooth development on planar surfaces. While the quorum-sensing AI-2 pathway appears to exist, many other pathways, using short peptides as signals, are coded in the genome. This is consistent with the complex environment of the plant, which produces oxygen in the light of the day time, then CO_2 during the night, with a concomitant considerable change in temperature and humidity. This is further made more complex by the alternation of seasons, with decay of leaves followed by burgeoning and maturation. Among the processes permitting adaptation to changing conditions is sporulation (see below), but there may also exist some specific chemical adaptations such as synthesis of hopanoids, which would permit resistance to desiccation: YhfL is similar to a squalene hopene cyclase, present in *Thermus thermophilus*.

All these plant-related features require specific adaptation processes, which must work more or less orthogonally to each other. This implies original regulatory setups, marking the gene setup of *B. subtilis* as a goldmine for the construction of novel synthetic biology devices (de Lorenzo & Danchin, 2008).

The *B. subtilis* differentiation programmes. *B. subtilis* has been chosen as a model organism for its original differentiation programme, making spores, which was proposed as representative of the rules controlling differentiation in general (Aguilar *et al.*, 2007). Indeed the study of sporulation has involved thousands of scientists all over the world. This heavily trodden area therefore does not need to be documented further here except to stress the concept of cannibalism, which has recently been emphasized in the context of the way bacterial colonies may behave collectively (Claverys & Havarstein, 2007; Nandy *et al.*, 2007; Ellermeier *et al.*, 2006).

Competence. The discovery of genetic transformation by Avery and his colleagues placed the process of transformation of DNA into cells at the core of the early efforts to construct experimental processes based on the use of the DNA molecule as a genetic tool. This drove the isolation of a *B. subtilis* strain amenable to easy transformation. Gamma-ray mutagenized derivatives of the Marburg strain of *B. subtilis* at Yale were brought to Western Reserve University by Yanofsky, where Spizizen tried a number of strains for efficient transformation and settled on a tryptophan auxotroph of strain 168 (Anagnostopoulos & Spizizen, 1961). Competence was studied by several groups in detail (449 references at PubMed on the topic), and the new annotation of the genome now reveals novel features such as requirement for LipA (YutB), an AdoMet radical enzyme, for establishment of competence (Ogura & Tanaka, 2009), or further details of the DNA uptake machinery, such as involvement of DprA (Smf) (Tadesse & Graumann, 2007).

Sporulation. The process of sporulation is the best-documented behaviour of *B. subtilis*. As a matter of fact this organism has been used as a model for cell differentiation, with the study of sporulation as the paradigm. The keyword 'subtilis' is associated with 768 articles with keyword 'spor*' in PubMed at the time of writing, with 21 review articles since 1993. One observation may be worth mentioning: the analysis of the fine structure of the proteome using correspondence analysis suggests that some proteins involved in the process of sporulation are related to proteins performing phage functions. This is in line with the observation that the SinR repressor, a key regulator in sporulation, has a domain that is similar in tertiary structure to that of a lambdoid repressor (Lewis *et al.*, 1998). A thorough phylogenetic study should explore the conjecture that some phage functions might have been recruited at the origin of sporulation.

Swimming, swarming and forming biofilms. More recently, the collective behaviour of *B. subtilis* has been analysed in terms of other differentiation processes such as those permitting swimming, swarming and forming biofilms (see Table 2) and the literature is growing fast in the domain, in particular with systems biology approaches (Rajagopala *et al.*, 2007). While cellulose is now well understood as a core component of biofilm structures, the role and synthesis of polyglutamate has more recently been emphasized, including for industrial applications (Meerak *et al.*, 2008).

Horizontal gene transfer

As discussed above, *B. subtilis* harbours many genomic islands which display a variety of specific features in terms of DNA composition, codon usage biases and alteration of syntenies. Several are the hallmark of horizontal gene transfer, most often via integration of prophages, which can either remain functional (such as SPβ), retain some activity (PBSX, *skin*), or are in the process of genetic decay (Kunst *et al.*, 1997). Associated with these genomic islands we find a variety of genes that keep signatures of widespread functions permitting gene transfer. *whiA* (*yvcL*) is a distant homologue of LAGLIDADG homing endonucleases that retained only DNA binding (Knizewski & Ginalski, 2007). It is present in Gram-positive bacteria (both A+T- and G+C-rich) and present in mycoplasmas, suggesting that some sort of retrotranscription existed very early on. Other features, such as those displayed by genes *tilS* (lysidine synthesis), *smc* (chromosome segregation) *divIVA*, *sepF*(*ylmF*) and others, are related to genes present in Eukarya. This is consistent either with a common ancestry of the corresponding functions at some point in evolution, or with the 'phagocytosis' scenario of the origin of living cells where phagocytic eukaryotes acting as predators would have predated the appearance of Bacteria, endowed with complex envelopes permitting them to escape predation, or Archaea, which would have fled to harsh niches where they could not be reached by predators (Kurland *et al.*, 2007).

Sequence and annotations for all

As a result of the functional reannotation presented above, gene products and gene names used in SubtiList were significantly revised. In particular, 407 genes whose previous name started with the letter 'y' (meaning that their function was unknown) were given a biologically significant name, on the basis of experimental evidence found in the literature. In addition, over 3000 new bibliographical references were imported in the database and linked to the relevant genes.

The new sequence and annotation information is integrated in an updated version of SubtiList (Moszer *et al.*, 2002). This genome database is now part of a multi-genome framework, GenoList – http://genolist.pasteur.fr/GenoList (Lechat *et al.*, 2008) – which holds genome information of more than 700 prokaryotic organisms imported from the Genome Reviews repository (Sterk *et al.*, 2006), and replaced in a few cases by manually curated genome data, such as those described in this work (accessible at http://genolist.pasteur.fr/GenoList/Bacillus subtilis 168). GenoList enables comparative genome analysis together with browsing and query capabilities similar to those of the previous version of SubtiList (Moszer *et al.*, 2002). In addition, metabolic pathway reconstruction data are available in a BioCyc-type web server, SubtiliCyc (http://genocyc.pasteur.fr), which is dynamically linked to GenoList. The sequence and annotation is also available on the MaGe platform (http://www.genoscope.cns.fr/agc/mage) as project BacilluScope (Vallenet *et al.*, 2006).

The EMBL accession number for the sequence reported in this paper is AL009126.

Conclusion

Model bacteria play an essential role in defining reference knowledge, after which a considerable fraction of the knowledge on bacteria is accumulated. *Bacillus subtilis* 168 is the reference organism for the Firmicutes and it is therefore essential to have a particularly accurate sequence of the organism, with up-to-date annotations. In the PubMed reference library 23 000 articles refer to some aspect of *B. subtilis* biology, while one finds 2 million pages at Google and 208 000 at Google Scholar, showing that this bacterium indeed plays the role of a model bacterium. We hope that the present effort in resequencing and reannotating the genome will benefit the international community of microbiologists. In the annotation process we used as much as possible the full content of the articles present in data libraries (Open Access publications and publications at PubMed Central). This permitted us to propose a few educated guesses about gene functions that will need to be experimentally validated. A general updated metabolic schema of the organism, SubtiliCyc, is available via the GenoList environment. Among the interesting metabolic features of the organism are many pathways directly associated with interaction with plants. As an example of the guesses we propose to the reader, triggered by our interest in sulfur metabolism (Sekowska *et al.*, 2000), is the prediction of YoaDC and YoaE as being involved in cysteine degradation, with transfer of sulfur to the C-terminal end of YoaD producing glycerate, which is subsequently modified to phosphoglycerate by YoaC; YoaD could be an important sulfur donor for construction of Fe–S clusters or sulfur-containing coenzymes. Using cysteine directly, it could be used for cyanide detoxification in plants. Many other examples of this type can be found in the present annotation of the genome sequence. We hope this will act as an incentive to trigger further work to substantiate these inferences.

ACKNOWLEDGEMENTS

While this article was in press Frank Kunst, who had been instrumental in the whole setting up of the *Bacillus subtilis* genome project, deeply affected by the unjust laws that force active scientists to retire, suddenly passed away. He had helped us polish the present article, which we dedicate to his memory.

From its onset this work was supported by Jean Weissenbach, whom we wish to warmly thank here. This work was supported by the BioSapiens Network of Excellence, grant LSHG CT-2003-503265, as an essential prerequisite to reannotation of the genome of *Mycoplasma pneumoniae*. It was also a prerequisite for accurate annotation of Firmicute pathogens [Network of Excellence EuroPathogenomics (grant LSHB-CT-2005-512061, supporting G. M. for the initial part of the work)]. Reannotation

was at the root of the paleome/cenome distinction, which is central to the European Union programmes Probactys (grant CT-2006-029104, supporting A. S.) and TARPOL (grant KBBE-2007-212894, supporting G. M. for the final part of the work). The Fondation Fourmentin-Guilbert supported the work of T. W. This work was also supported by a grant from MRT/ANR PFTV 2007, MicroScope project. We thank our colleagues Hans Kruegel, Beth Nelson, Michael Rey, Preethi Ramaiya and Naotake Ogasawara for their interest and support at the onset of this work. Particular thanks are due to Elisabeth Coudert, Tania Lima, Anne Morgat, Catherine Rivoire and the SIB team for their help in creating a consistent nomenclature and annotation of the genome.

REFERENCES

Aguilar, C., Vlamakis, H., Losick, R. & Kolter, R. (2007). Thinking about *Bacillus subtilis* as a multicellular organism. *Curr Opin Microbiol* **10**, 638–643.

Anagnostopoulos, C. & Spizizen, J. (1961). Requirements for transformation in *Bacillus subtilis*. *J Bacteriol* **81**, 741–746.

Andre, G., Even, S., Putzer, H., Burguiere, P., Croux, C., Danchin, A., Martin-Verstraete, I. & Soutourina, O. (2008). S-box and T-box riboswitches and antisense RNA control a sulfur metabolic operon of *Clostridium acetobutylicum*. *Nucleic Acids Res* **36**, 5955–5969.

Anton, B. P., Saleh, L., Benner, J. S., Raleigh, E. A., Kasif, S. & Roberts, R. J. (2008). RimO, a MiaB-like enzyme, methylthiolates the universally conserved Asp88 residue of ribosomal protein S12 in *Escherichia coli*. *Proc Natl Acad Sci U S A* **105**, 1826–1831.

Bailly-Bechet, M., Danchin, A., Iqbal, M., Marsili, M. & Vergassola, M. (2006). Codon usage domains over bacterial chromosomes. *PLOS Comput Biol* **2**, e37.

Beck, L. L., Smith, T. G. & Hoover, T. R. (2007). Look, no hands! Unconventional transcriptional activators in bacteria. *Trends Microbiol* **15**, 530–537.

Begley, T. P., Chatterjee, A., Hanes, J. W., Hazra, A. & Ealick, S. E. (2008). Cofactor biosynthesis – still yielding fascinating new biological chemistry. *Curr Opin Chem Biol* **12**, 118–125.

Belitsky, B. R. & Sonenshein, A. L. (1998). Role and regulation of *Bacillus subtilis* glutamate dehydrogenase genes. *J Bacteriol* **180**, 6298–6305.

Bentley, D. R. (2006). Whole-genome re-sequencing. *Curr Opin Genet Dev* **16**, 545–552.

Berger, B. J., English, S., Chan, G. & Knodel, M. H. (2003). Methionine regeneration and aminotransferases in *Bacillus subtilis*, *Bacillus cereus*, and *Bacillus anthracis*. *J Bacteriol* **185**, 2418–2431.

Bisicchia, P., Noone, D., Lioliou, E., Howell, A., Quigley, S., Jensen, T., Jarmer, H. & Devine, K. M. (2007). The essential YycFG two-component system controls cell wall metabolism in *Bacillus subtilis*. *Mol Microbiol* **65**, 180–200.

Bocs, S., Cruveiller, S., Vallenet, D., Nuel, G. & Medigue, C. (2003). AMIGene: Annotation of MIcrobial Genes. *Nucleic Acids Res* **31**, 3723–3726.

Burguiere, P., Auger, S., Hullo, M. F., Danchin, A. & Martin-Verstraete, I. (2004). Three different systems participate in L-cystine uptake in *Bacillus subtilis*. *J Bacteriol* **186**, 4875–4884.

Carr, J. F., Hamburg, D. M., Gregory, S. T., Limbach, P. A. & Dahlberg, A. E. (2006). Effects of streptomycin resistance mutations on posttranslational modification of ribosomal protein S12. *J Bacteriol* **188**, 2020–2023.

Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S. Y., Shearer, A. G. & other authors (2008). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* **36**, D623–D631.

Chapman-Smith, A., Mulhern, T. D., Whelan, F., Cronan, J. E., Jr & Wallace, J. C. (2001). The C-terminal domain of biotin protein ligase from *E. coli* is required for catalytic activity. *Protein Sci* **10**, 2608–2617.

Chartier, F. J. & Couture, M. (2007). Substrate-specific interactions with the heme-bound oxygen molecule of nitric-oxide synthase. *J Biol Chem* **282**, 20877–20886.

Christiansen, L. C., Schou, S., Nygaard, P. & Saxild, H. H. (1997). Xanthine metabolism in *Bacillus subtilis*: characterization of the *xpt-pbuX* operon and evidence for purine- and nitrogen-controlled expression of genes involved in xanthine salvage and catabolism. *J Bacteriol* **179**, 2540–2550.

Claverys, J. P. & Havarstein, L. S. (2007). Cannibalism and fratricide: mechanisms and raisons d'être. *Nat Rev Microbiol* **5**, 219–229.

Danchin, A. (2008). Natural selection and immortality. *Biogerontology*, doi:10.1007/s10522-008-9171-5

Danchin, A. (2009a). A phylogenetic view of bacterial ribonucleases. *Prog Nucleic Acid Res Mol Biol* **85**, 1–41.

Danchin, A. (2009b). Bacteria as computers making computers. *FEMS Microbiol Rev* **33**, 3–26.

Danchin, A., Fang, G. & Noria, S. (2007). The extant core bacterial proteome is an archive of the origin of life. *Proteomics* **7**, 875–889.

Dartois, V., Debarbouille, M., Kunst, F. & Rapoport, G. (1998). Characterization of a novel member of the DegS-DegU regulon affected by salt stress in *Bacillus subtilis*. *J Bacteriol* **180**, 1855–1861.

de Lorenzo, V. & Danchin, A. (2008). Synthetic biology: discovering new worlds and new words. *EMBO Rep* **9**, 822–827.

den Blaauwen, T., de Pedro, M. A., Nguyen-Disteche, M. & Ayala, J. A. (2008). Morphogenesis of rod-shaped sacculi. *FEMS Microbiol Rev* **32**, 321–344.

Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* **2**, 414–424.

Dyson, F. J. (1985). *Origins of Life*. Cambridge, UK: Cambridge University Press.

Earl, A. M., Losick, R. & Kolter, R. (2008). Ecology and genomics of *Bacillus subtilis*. *Trends Microbiol* **16**, 269–275.

Ellermeier, C. D., Hobbs, E. C., Gonzalez-Pastor, J. E. & Losick, R. (2006). A three-protein signaling pathway governing immunity to a bacterial cannibalism toxin. *Cell* **124**, 549–559.

Errington, J. (2003). Regulation of endospore formation in *Bacillus subtilis*. *Nat Rev Microbiol* **1**, 117–126.

Fang, G., Rocha, E. & Danchin, A. (2005). How essential are nonessential genes? *Mol Biol Evol* **22**, 2147–2156.

Formstone, A., Carballido-Lopez, R., Noirot, P., Errington, J. & Scheffers, D. J. (2008). Localization and interactions of teichoic acid synthetic enzymes in *Bacillus subtilis*. *J Bacteriol* **190**, 1812–1821.

Fouet, A., Arnaud, M., Klier, A. & Rapoport, G. (1987). *Bacillus subtilis* sucrose-specific enzyme II of the phosphotransferase system: expression in *Escherichia coli* and homology to enzymes II from enteric bacteria. *Proc Natl Acad Sci U S A* **84**, 8773–8777.

Frangeul, L., Nelson, K. E., Buchrieser, C., Danchin, A., Glaser, P. & Kunst, F. (1999). Cloning and assembly strategies in microbial genome projects. *Microbiology* **145**, 2625–2634.

Frenkiel-Krispin, D. & Minsky, A. (2006). Nucleoid organization and the maintenance of DNA integrity in *E. coli*, *B. subtilis* and *D. radiodurans*. *J Struct Biol* **156**, 311–319.

Frey, P. A., Hegeman, A. D. & Ruzicka, F. J. (2008). The radical SAM superfamily. *Crit Rev Biochem Mol Biol* **43**, 63–88.

Gaidenko, T. A., Kim, T. J., Weigel, A. L., Brody, M. S. & Price, C. W. (2006). The blue-light receptor YtvA acts in the environmental stress signaling pathway of *Bacillus subtilis*. *J Bacteriol* **188**, 6387–6395.

Gilks, W. R., Audit, B., De Angelis, D., Tsoka, S. & Ouzounis, C. A. (2002). Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics* **18**, 1641–1649.

Glaser, P., Kunst, F., Arnaud, M., Coudart, M. P., Gonzales, W., Hullo, M. F., Ionescu, M., Lubochinsky, B., Marcelino, L. & other authors (1993). *Bacillus subtilis* genome project: cloning and sequencing of the 97 kb region from 325 degrees to 333 degrees. *Mol Microbiol* **10**, 371–384.

Goelzer, A., Bekkal Brikci, F., Martin-Verstraete, I., Noirot, P., Bessieres, P., Aymerich, S. & Fromion, V. (2008). Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*. *BMC Syst Biol* **2**, 20.

Gordon, D., Abajian, C. & Green, P. (1998). Consed: a graphical tool for sequence finishing. *Genome Res* **8**, 195–202.

Gruber, T. M. & Gross, C. A. (2003). Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol* **57**, 441–466.

Haldenwang, W. G. (1995). The sigma factors of *Bacillus subtilis*. *Microbiol Rev* **59**, 1–30.

Harwood, C. R. & Wipat, A. (1996). Sequencing and functional analysis of the genome of *Bacillus subtilis* strain 168. *FEBS Lett* **389**, 84–87.

Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., Ohtsubo, E., Baba, T., Wanner, B. L. & other authors (2006). Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Mol Syst Biol* **2**, 2006.0007.

Hayhurst, E. J., Kailas, L., Hobbs, J. K. & Foster, S. J. (2008). Cell wall peptidoglycan architecture in *Bacillus subtilis*. *Proc Natl Acad Sci U S A* **105**, 14603–14608.

Helmann, J. D. (1999). Anti-sigma factors. *Curr Opin Microbiol* **2**, 135–141.

Helmann, J. D. (2002). The extracytoplasmic function (ECF) sigma factors. *Adv Microb Physiol* **46**, 47–110.

Herring, C. D. & Palsson, B. O. (2007). An evaluation of Comparative Genome Sequencing (CGS) by comparing two previously-sequenced bacterial genomes. *BMC Genomics* **8**, 274.

Hobman, J. L., Penn, C. W. & Pallen, M. J. (2007). Laboratory strains of *Escherichia coli*: model citizens or deceitful delinquents growing old disgracefully? *Mol Microbiol* **64**, 881–885.

Hochgrafe, F., Mostertz, J., Pother, D. C., Becher, D., Helmann, J. D. & Hecker, M. (2007). *S*-Cysteinylation is a general mechanism for thiol protection of *Bacillus subtilis* proteins after oxidative stress. *J Biol Chem* **282**, 25981–25985.

Hoper, D., Bernhardt, J. & Hecker, M. (2006). Salt stress adaptation of *Bacillus subtilis*: a physiological proteomics approach. *Proteomics* **6**, 1550–1562.

Hullo, M. F., Auger, S., Soutourina, O., Barzu, O., Yvon, M., Danchin, A. & Martin-Verstraete, I. (2007). Conversion of methionine to cysteine in *Bacillus subtilis* and its regulation. *J Bacteriol* **189**, 187–197.

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8**, R143.

Jenkins, A. L., Zhang, Y., Ealick, S. E. & Begley, T. P. (2008). Mutagenesis studies on TenA: a thiamin salvage enzyme from *Bacillus subtilis*. *Bioorg Chem* **36**, 29–32.

Joseph, P., Fichant, G., Quentin, Y. & Denizot, F. (2002). Regulatory relationship of two-component and ABC transport systems and clustering of their genes in the *Bacillus*/*Clostridium* group, suggest a functional link between them. *J Mol Microbiol Biotechnol* **4**, 503–513.

Julkowska, D., Obuchowski, M., Holland, I. B. & Seror, S. J. (2005). Comparative analysis of the development of swarming communities of *Bacillus subtilis* 168 and a natural wild type: critical effects of surfactin and the composition of the medium. *J Bacteriol* **187**, 65–76.

Kazmierczak, M. J., Wiedmann, M. & Boor, K. J. (2005). Alternative sigma factors and their roles in bacterial virulence. *Microbiol Mol Biol Rev* **69**, 527–543.

Kiley, P. J. & Beinert, H. (2003). The role of Fe–S proteins in sensing and regulation in bacteria. *Curr Opin Microbiol* **6**, 181–185.

Knizewski, L. & Ginalski, K. (2007). Bacterial DUF199/COG1481 proteins including sporulation regulator WhiA are distant homologs of LAGLIDADG homing endonucleases that retained only DNA binding. *Cell Cycle* **6**, 1666–1670.

Kobayashi, K., Ogura, M., Yamaguchi, H., Yoshida, K., Ogasawara, N., Tanaka, T. & Fujita, Y. (2001). Comprehensive DNA microarray analysis of *Bacillus subtilis* two-component regulatory systems. *J Bacteriol* **183**, 7365–7370.

Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S. & other authors (2003). Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci U S A* **100**, 4678–4683.

Kroos, L. & Yu, Y. T. (2000). Regulation of sigma factor activity during *Bacillus subtilis* development. *Curr Opin Microbiol* **3**, 553–560.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A. & other authors (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256.

Kurland, C. G., Canback, B. & Berg, O. G. (2007). The origins of modern proteomes. *Biochimie* **89**, 1454–1463.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12.

Lechat, P., Hummel, L., Rousseau, S. & Moszer, I. (2008). GenoList: an integrated environment for comparative analysis of microbial genomes. *Nucleic Acids Res* **36**, D469–D474.

Lewis, R. J., Brannigan, J. A., Offen, W. A., Smith, I. & Wilkinson, A. J. (1998). An evolutionary link between sporulation and prophage induction in the structure of a repressor : anti-repressor complex. *J Mol Biol* **283**, 907–912.

Lima, T., Auchincloss, A. H., Coudert, E., Keller, G., Michoud, K., Rivoire, C., Bulliard, V., de Castro, E., Lachaize, C. & other authors (2009). HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res* **37**, D471–D478.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J. & other authors (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380.

Mechold, U., Fang, G., Ngo, S., Ogryzko, V. & Danchin, A. (2007). YtqI from *Bacillus subtilis* has both oligoribonuclease and pAp-phosphatase activity. *Nucleic Acids Res* **35**, 4552–4561.

Medigue, C., Rose, M., Viari, A. & Danchin, A. (1999). Detecting and analyzing DNA sequencing errors: toward a higher quality of the *Bacillus subtilis* genome sequence. *Genome Res* **9**, 1116–1127.

Meerak, J., Yukphan, P., Miyashita, M., Sato, H., Nakagawa, Y. & Tahara, Y. (2008). Phylogeny of gamma-polyglutamic acid-producing *Bacillus* strains isolated from a fermented locust bean product manufactured in West Africa. *J Gen Appl Microbiol* **54**, 159–166.

Merkl, R. (2004). SIGI: score-based identification of genomic islands. *BMC Bioinformatics* **5**, 22.

Montorsi, M. & Lorenzetti, R. (1993). Heat-stable and heat-labile thymidylate synthases B of *Bacillus subtilis*: comparison of the nucleotide and amino acid sequences. *Mol Gen Genet* **239**, 1–5.

Moszer, I., Rocha, E. P. & Danchin, A. (1999). Codon usage and lateral gene transfer in *Bacillus subtilis*. *Curr Opin Microbiol* **2**, 524–528.

Moszer, I., Jones, L. M., Moreira, S., Fabry, C. & Danchin, A. (2002). SubtiList: the reference database for the *Bacillus subtilis* genome. *Nucleic Acids Res* **30**, 62–65.

Nakano, M. M. & Zuber, P. (1998). Anaerobic growth of a "strict aerobe" (*Bacillus subtilis*). *Annu Rev Microbiol* **52**, 165–190.

Nakano, M. M., Geng, H., Nakano, S. & Kobayashi, K. (2006). The nitric oxide-responsive regulator NsrR controls ResDE-dependent gene expression. *J Bacteriol* **188**, 5878–5887.

Nandy, S. K., Bapat, P. M. & Venkatesh, K. V. (2007). Sporulating bacteria prefers predation to cannibalism in mixed cultures. *FEBS Lett* **581**, 151–156.

Nicolas, P., Bize, L., Muri, F., Hoebeke, M., Rodolphe, F., Ehrlich, S. D., Prum, B. & Bessieres, P. (2002). Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models. *Nucleic Acids Res* **30**, 1418–1426.

Ning, Z., Cox, A. J. & Mullikin, J. C. (2001). SSAHA: a fast search method for large DNA databases. *Genome Res* **11**, 1725–1729.

Nitschke, P., Guerdoux-Jamet, P., Chiapello, H., Faroux, G., Henaut, C., Henaut, A. & Danchin, A. (1998). Indigo: a World-Wide-Web review of genomes and gene functions. *FEMS Microbiol Rev* **22**, 207–227.

Noirot-Gros, M. F., Dervyn, E., Wu, L. J., Mervelet, P., Errington, J., Ehrlich, S. D. & Noirot, P. (2002). An expanded view of bacterial DNA replication. *Proc Natl Acad Sci U S A* **99**, 8342–8347.

Ogura, M. & Tanaka, T. (2009). The *Bacillus subtilis* late competence operon *comE* is transcriptionally regulated by *yutB* and under post-transcription initiation control by *comN* (*yrzD*). *J Bacteriol* **191**, 949–958.

Paley, S. M. & Karp, P. D. (2006). The Pathway Tools cellular overview diagram and Omics Viewer. *Nucleic Acids Res* **34**, 3771–3778.

Pascal, G., Medigue, C. & Danchin, A. (2005). Universal biases in protein composition of model prokaryotes. *Proteins* **60**, 27–35.

Petersohn, A., Brigulla, M., Haas, S., Hoheisel, J. D., Volker, U. & Hecker, M. (2001). Global analysis of the general stress response of *Bacillus subtilis*. *J Bacteriol* **183**, 5617–5631.

Piggot, P. J. & Hilbert, D. W. (2004). Sporulation of *Bacillus subtilis*. *Curr Opin Microbiol* **7**, 579–586.

Rajagopala, S. V., Titz, B., Goll, J., Parrish, J. R., Wohlbold, K., McKevitt, M. T., Palzkill, T., Mori, H., Finley, R. L., Jr & Uetz, P. (2007). The protein network of bacterial motility. *Mol Syst Biol* **3**, 128.

Raschle, T., Amrhein, N. & Fitzpatrick, T. B. (2005). On the two components of pyridoxal 5′-phosphate synthase from *Bacillus subtilis*. *J Biol Chem* **280**, 32291–32300.

Reder, A., Hoper, D., Weinberg, C., Gerth, U., Fraunholz, M. & Hecker, M. (2008). The Spx paralogue MgsR (YqgZ) controls a subregulon within the general stress response of *Bacillus subtilis*. *Mol Microbiol* **69**, 1104–1120.

Reuter, K., Mofid, M. R., Marahiel, M. A. & Ficner, R. (1999). Crystal structure of the surfactin synthetase-activating enzyme Sfp: a prototype of the 4′-phosphopantetheinyl transferase superfamily. *EMBO J* **18**, 6823–6831.

Reyes, D. Y. & Zuber, P. (2008). Activation of transcription initiation by Spx: formation of transcription complex and identification of a *cis*-acting element required for transcriptional activation. *Mol Microbiol* **69**, 765–779.

Rocha, E. (2002). Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol* **10**, 393–395.

Rocha, E. P., Viari, A. & Danchin, A. (1998). Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons. *Nucleic Acids Res* **26**, 2971–2980.

Rocha, E. P., Danchin, A. & Viari, A. (1999a). Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. *Mol Biol Evol* **16**, 1219–1230.

Rocha, E. P., Danchin, A. & Viari, A. (1999b). Translation in *Bacillus subtilis*: roles and trends of initiation and termination, insights from a genome analysis. *Nucleic Acids Res* **27**, 3567–3576.

Roll-Hansen, N. (1979). Experimental method and spontaneous generation: the controversy between Pasteur and Pouchet, 1859–64. *J Hist Med Allied Sci* **34**, 273–292.

Saier, M. H., Jr, Goldman, S. R., Maile, R. R., Moreno, M. S., Weyler, W., Yang, N. & Paulsen, I. T. (2002). Transport capabilities encoded within the *Bacillus subtilis* genome. *J Mol Microbiol Biotechnol* **4**, 37–67.

Saito, S., Kakeshita, H. & Nakamura, K. (2009). Novel small RNA-encoding genes in the intergenic regions of *Bacillus subtilis*. *Gene* **428**, 2–8.

Sauer, U. & Eikmanns, B. J. (2005). The PEP-pyruvate-oxaloacetate node as the switch point for carbon flux distribution in bacteria. *FEMS Microbiol Rev* **29**, 765–794.

Saunders, C. W., Schmidt, B. J., Mirot, M. S., Thompson, L. D. & Guyer, M. S. (1984). Use of chromosomal integration in the establishment and expression of *blaZ*, a *Staphylococcus aureus* beta-lactamase gene, in *Bacillus subtilis*. *J Bacteriol* **157**, 718–726.

Sekowska, A. (1999). *Une rencontre du métabolisme du soufre et de l'azote; le métabolisme des polyamines chez Bacillus subtilis*. PhD thesis, Université de Versailles Saint-Quentin-en-Yvelines.

Sekowska, A., Kung, H. F. & Danchin, A. (2000). Sulfur metabolism in *Escherichia coli* and related bacteria: facts and fiction. *J Mol Microbiol Biotechnol* **2**, 145–177.

Sekowska, A., Robin, S., Daudin, J. J., Henaut, A. & Danchin, A. (2001). Extracting biological information from DNA arrays: an unexpected link between arginine and methionine metabolism in *Bacillus subtilis*. *Genome Biol* **2**, RESEARCH0019.

Sekowska, A., Denervaud, V., Ashida, H., Michoud, K., Haas, D., Yokota, A. & Danchin, A. (2004). Bacterial variations on the methionine salvage pathway. *BMC Microbiol* **4**, 9.

Simpson, A. J. (2001). Genome sequencing networks. *Nat Rev Genet* **2**, 979–983.

Singh, K. D., Schmalisch, M. H., Stulke, J. & Gorke, B. (2008). Carbon catabolite repression in *Bacillus subtilis*: quantitative analysis of repression exerted by different carbon sources. *J Bacteriol* **190**, 7275–7284.

Sinha, S. C. & Sprang, S. R. (2006). Structures, mechanism, regulation and evolution of class III nucleotidyl cyclases. *Rev Physiol Biochem Pharmacol* **157**, 105–140.

Sneath, P. H. A. (1986). Endospore-forming Gram-positive rods and cocci. In *Bergey's Manual of Systematic Bacteriology*, pp. 1105–1139. Edited by P. H. A. Sneath, N. S. Mair, M. E. Sharpe & J. G. Holt. Baltimore: Williams & Wilkins Co.

Sonenshein, A. L. (2007). Control of key metabolic intersections in *Bacillus subtilis*. *Nat Rev Microbiol* **5**, 917–927.

Soupene, E., van Heeswijk, W. C., Plumbridge, J., Stewart, V., Bertenthal, D., Lee, H., Prasad, G., Paliy, O., Charernnoppakul, P. & Kustu, S. (2003). Physiological studies of *Escherichia coli* strain MG1655: growth defects and apparent cross-regulation of gene expression. *J Bacteriol* **185**, 5611–5626.

Srivatsan, A., Han, Y., Peng, J., Tehranchi, A. K., Gibbs, R., Wang, J. D. & Chen, R. (2008). High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. *PLoS Genet* **4**, e1000139.

Sterk, P., Kersey, P. J. & Apweiler, R. (2006). Genome reviews: standardizing content and representation of information about complete genomes. *OMICS* **10**, 114–118.

Tadesse, S. & Graumann, P. L. (2007). DprA/Smf protein localizes at the DNA uptake machinery in competent *Bacillus subtilis* cells. *BMC Microbiol* **7**, 105.

Tamames, J., Gonzalez-Moreno, M., Mingorance, J., Valencia, A. & Vicente, M. (2001). Bringing gene order into bacterial shape. *Trends Genet* **17**, 124–126.

Tamburini, E., Leon, A. G., Perito, B. & Mastromei, G. (2003). Characterization of bacterial pectinolytic strains involved in the water retting process. *Environ Microbiol* **5**, 730–736.

Tanous, C., Soutourina, O., Raynal, B., Hullo, M. F., Mervelet, P., Gilles, A. M., Noirot, P., Danchin, A., England, P. & Martin-Verstraete, I. (2008). The CymR regulator in complex with the enzyme CysK controls cysteine metabolism in *Bacillus subtilis*. *J Biol Chem* **283**, 35551–35560.

Tozzi, M. G., Camici, M., Mascia, L., Sgarrella, F. & Ipata, P. L. (2006). Pentose phosphates in nucleoside interconversion and catabolism. *FEBS J* **273**, 1089–1101.

Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C. & Medigue, C. (2006). MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res* **34**, 53–65.

Van Arsdell, S. W., Perkins, J. B., Yocum, R. R., Luan, L., Howitt, C. L., Chatterjee, N. P. & Pero, J. G. (2005). Removing a bottleneck in the *Bacillus subtilis* biotin pathway: BioA utilizes lysine rather than *S*-adenosylmethionine as the amino donor in the KAPA-to-DAPA reaction. *Biotechnol Bioeng* **91**, 75–83.

van Schaik, W. & Abee, T. (2005). The role of sigmaB in the stress response of Gram-positive bacteria – targets for food preservation and safety. *Curr Opin Biotechnol* **16**, 218–224.

Wang, Z. Q., Lawson, R. J., Buddha, M. R., Wei, C. C., Crane, B. R., Munro, A. W. & Stuehr, D. J. (2007). Bacterial flavodoxins support nitric oxide production by *Bacillus subtilis* nitric-oxide synthase. *J Biol Chem* **282**, 2196–2202.

Yamazaki, S., Yamazaki, J., Nishijima, K., Otsuka, R., Mise, M., Ishikawa, H., Sasaki, K., Tago, S. & Isono, K. (2006). Proteome analysis of an aerobic hyperthermophilic crenarchaeon, *Aeropyrum pernix* K1. *Mol Cell Proteomics* **5**, 811–823.

You, C., Lu, H., Sekowska, A., Fang, G., Wang, Y., Gilles, A. M. & Danchin, A. (2005). The two authentic methionine aminopeptidase genes are differentially expressed in *Bacillus subtilis*. *BMC Microbiol* **5**, 57.

You, C., Sekowska, A., Francetic, O., Martin-Verstraete, I., Wang, Y. & Danchin, A. (2008). Spx mediates oxidative stress regulation of the methionine sulfoxide reductases operon in *Bacillus subtilis*. *BMC Microbiol* **8**, 128.

Yudkin, M. D. & Clarkson, J. (2005). Differential gene expression in genetically identical sister cells: the initiation of sporulation in *Bacillus subtilis*. *Mol Microbiol* **56**, 578–589.

Zeigler, D. R., Pragai, Z., Rodriguez, S., Chevreux, B., Muffler, A., Albert, T., Bai, R., Wyss, M. & Perkins, J. B. (2008). The origins of 168, W23, and other *Bacillus subtilis* legacy strains. *J Bacteriol* **190**, 6983–6995.

Edited by: D. W. Ussery

2.3 The upstream and coding region of thrS gene

2.3.1 Literature overview

Most organisms contain 20 aminoacyl-tRNA synthetases (aaRSs), each specifically connecting a cognate amino acid to its corresponding tRNA carrying an anticodon adapted to the rule of the genetic code [147-150]. aaRSs are essential enzymes for protein synthesis in cells. The reaction process of aaRSs is usually completed in two steps: firstly synthetase binds to ATP and its corresponding amino acids to form an aminoacyl-AMP:aaRS complex with a release of an inorganic pyrophosphate (Reaction 1); Lastly, the aa-AMP:aaRS complex interacts with an appropriate tRNA, and the amino acid is transferred from aa-AMP to the 2' or 3'-OH of the last base (A76) of tRNA at 3' end, while aaRS is restored to a free state (see Reaction 2) [151].

Reaction 1: Amino acid (aa) + ATP + aaRS → aa-AMP:aaRS + PPi

Reaction 2: aa-AMP:aaRS + tRNA + ATP → aa-tRNA + AMP + aaRS

Total Reaction: Amino acid (aa) + tRNA + ATP $\xrightarrow{aaRS}$ aa-tRNA + AMP + PPi

Where ATP is adenosine triphosphate; aaRS is Aminoacyl-tRNA synthetase; aa-AMP:aaRS is the complex of Aminoacyl-adenosine monophosphate and Aminoacyl-tRNA synthetase; PPi is inorganic pyrophosphate; tRNA is transfer RNA; aa-tRNA is Aminoacyl-tRNA.

2.3.1.1 Regulation of expression for Aminoacyl-tRNA Synthetase

In order to maintain normal growth, cells usually regulate the expression level of their genes to consume resources while quickly adapting to the environment. In detail, normally, if a protein's amount is insufficient to meet the needs for growing, its level will be up-regulated; in

contrast, if the protein accumulates too much, its level will be down-regulated to save resources.

Protein biosynthesis is a complex process: the gene is firstly transcribed into mRNA with DNA as template; then mRNA is translated into amino acid chains (peptides) with the help of ribosome, tRNA and other enzymes; and the amino acid chains (peptides) are eventually folded into functional proteins after a series of post-translational modifications (Fig. 2.5).

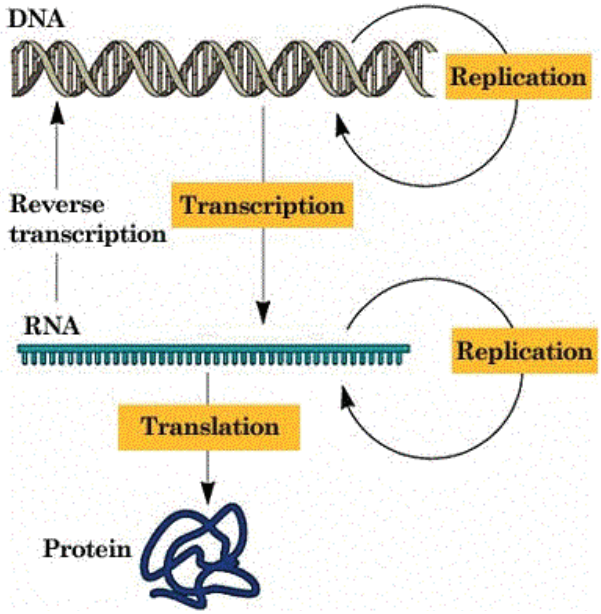


Figure 2.5 Central dogma and Protein biosynthesis

The "central dogma" describes the transfers of genetic information in a cell. Information transferred from DNA to proteins is achieved through two steps: the first is from DNA to the RNA called transcription, and the second is from RNA to protein called translation. In addition, there are some other types of information transfer, such as, by DNA replication, information can be passed from parent cell to child cells; by RNA self-replication, RNA can be duplicated in cell; by RNA reverse transcription, the information can be passed from RNA to DNA.

In eukaryotes, transcription and translation are separated from each other in both space and time: transcription is completed inside of the nucleus, while translation unfolds in the cytoplasm outside the nucleus [152, 153] and it starts once transcription is completely done. However, in prokaryotes, the two processes are closely linked. In fact, the translation of a bacterial mRNA begins during the formation of its transcripts. In bacteria, the majority of aminoacyl-tRNA synthetases (aaRS) are regulated at the transcriptional level. But threonyl-tRNA^{Thr} synthetase

(thrRS) is a special case because its expression can be regulated at either transcriptional level (in *Bacillus subtilis* [154]) or translational level (in *Escherichia coli* [155]). Normally, the leader regions of aaRS genes can fold into two or three forms. Among them, at least one is beneficial to transcription or translation, and also at least one represses either transcription or translation.

2.3.1.2 Regulation of aaRS expression in Bacillus subtilis

The expression of most aaRSs genes is regulated by an anti-termination mechanism during transcription [156]. aaRS genes have a highly structured and untranslated leader region about 300 nucleotides (nt), followed by a Rho-independent transcription terminator located just upstream of the translation initiation site (Fig. 2.6). Although the DNA sequences of aaRS genes' leader region are usually dramatically different from one another; they contain a highly conserved sequence about 14 nt, known as the T-box that can fold into similar secondary structures (Fig. 2.7). Typically, the leader region can be folded into a secondary structure with the appearance of a transcription terminator, which is a hairpin structure, resulting in the Rho-independent termination of transcription (Fig. 2.6). In the 5' strand of the terminator stem, there is a sequence complementary to the T box (Fig. 2.7). The complementation of these two sequences indicates that the leader region can be folded into an alternative secondary structure, the anti-terminator structure, which is different from the structure of terminator. Upon starvation of the cognate amino acid, the expression of the corresponding aaRS can be induced by transcription read-through when the leader region is in the anti-termination conformation in order to ensure the normal level of aminoacyl-tRNA and protein synthesis in cell [157-161].

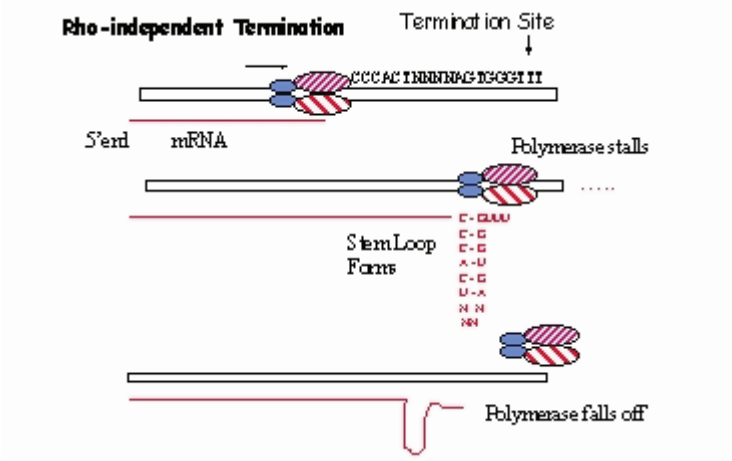


Figure 2.6 Rho-independent transcription termination

Note: Part of DNA around the termination site contains a sequence rich in G (guanylate) and C (cytosine), which can be easily folded into a stable stem-loop during the formation of the mRNA, and a pyrimidine chain of urine residues (UUU), that forms less stable nucleotide pairs with adenylate residues (AAA). A protein associated with RNA polymerase can tightly bind to the stem-loop structure. This forces the polymerase to stall at the termination site, RNA and DNA chains separate from each other because of the unstable AU pairs in the RNA-DNA dimer, and both RNA and DNA chains detach from the RNA polymerase.

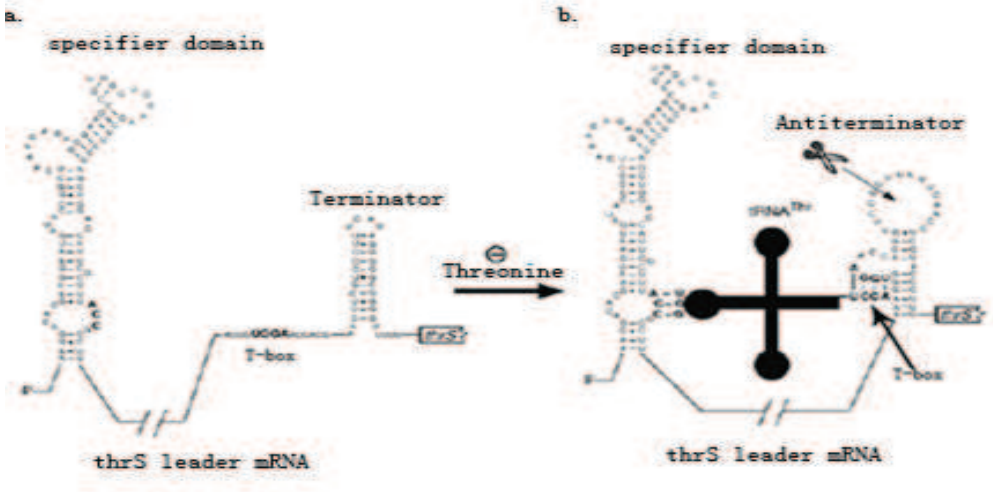


Figure 2.7 The termination and antitermination formation of thrS leader region

Notes: a. the termination structure model of leader region contains three major domains (specifier domain, T box and the terminator structure); b. anti-terminator structure model of leader region: uncharged tRNA interacts with specifier domain and complementary sequences in T-box. The scissor picture represents a possible splice site located in anti-terminator structure. This picture is modified from http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=eurekah&part=A19024.

The anti-terminator structure formed by the complementary sequence between T box and the 5' end sequence of terminator stem is unstable, but it can be stabilized by the uncharged

cognate tRNA [162]. Uncharged tRNA interacts with the leader region at two sites at least: one is the classical cooperation site of codon:anti-codon between specifier domain in the leader region and tRNA; the other is the unpaired CCA sequence located in the acceptor arm of uncharged tRNA at the 3' end and the complementary sequence (UGG), located in T box and bulging out from the stem in the anti-terminator conformation (Fig. 2.7) [163]. The charged cognate tRNAs act as competitors, since the CCA sequences in charged tRNAs have been attached to the corresponding amino acid, the charged tRNAs lose the ability to communicate with T box, and have no effect on stabilizing the conformation of anti-terminator.

In addition to the conversion between anti-terminator and terminator conformation, there is also a possible splice site existing in the loop of anti-terminator structure (Fig. 2.7). In the case of threonine starvation, the shear frequency on this site is specifically increased, leading to more read-through for transcription, and generating more thrS mRNA, and increasing thrRS level in the cell finally [154].

2.3.1.3 Transcriptional regulation of aaRS in *E. coli*

Compared to *Bacillus subtilis*, in which all the aaRSs use the same mechanism to adjust their proteins' final level, there are at least two different molecular mechanisms in *Escherichia coli* to achieve this effect at the transcriptional level.

The first model is represented by alanyl-tRNA synthetase. It represses its own gene transcription through the binding to a palindromic sequence flanking the gene's transcription start site. The alanine effect is caused by direct association of the ligand with the synthetase that in turn mediates tighter binding to the DNA. In this case, transcription repression is greatly

enhanced as the concentration of cognate amino acid increases [164].

The second model is represented by Phenylalanyl-tRNA synthetase (PheRS). Both *in vivo* and *in vitro* experiments show that the expression of the PheS-PheT operon is controlled through transcription attenuation. During the transcription process, the extension of mRNA of the structural gene is controlled by the aminoacylation level of charged cognate $tRNA^{Phe}$, and thus results in a proper concentration of the enzyme itself. Genes with this regulation mechanism are preceded by a precursor gene encoding peptide rich in the cognate amino acid. In the case for the PheS-PheT operon, it contains 5 phenylalanine residues in the leader peptide. As stated before, translation of mRNA in bacteria is immediately started when the transcription begins. Different stages in translation of the leader peptide can result in different folding of the subsequent sequence in the mRNA, and Rho-independent transcription terminator is an important form that appears among these different conformations. In the absence of any leader peptide translated, the transcription of the PheS-PheT operon is terminated because of the formation of a stem-loop between regions 3 and 4 (Fig. 2.8A). When the concentration of Phe-$tRNA^{Phe}$ is higher, the ribosome will halt at the first or the second phenylalanine codon, and regions 1 and 2 will form a stem loop, and regions 3 and 4 can also form a terminator structure to prevent new PheRS mRNA generation (Fig. 2.8B). When phenylalanine concentration is much higher, the first and second phenylalanine residues in the leader peptide sequence can be expressed successfully, but the ribosome will halt at the last two phenylalanine residues codons. In this case regions 1 and 2 can not form a stem and loop, but regions 2 and 3 can form into the anti-terminator stem loop resulting in the transcription read-through for PheS (Fig. 2.8C). When phenylalanyl-$tRNA^{Phe}$ has been accumulated, the leader peptide can be translated successfully and quickly, and there is no

chance for region 2 and 3 to form the anti-terminator stem loop, as a result, regions 3 and 4 can form into terminator stem loop once again and prevent the transcription of PheS (Fig. 2.8D).

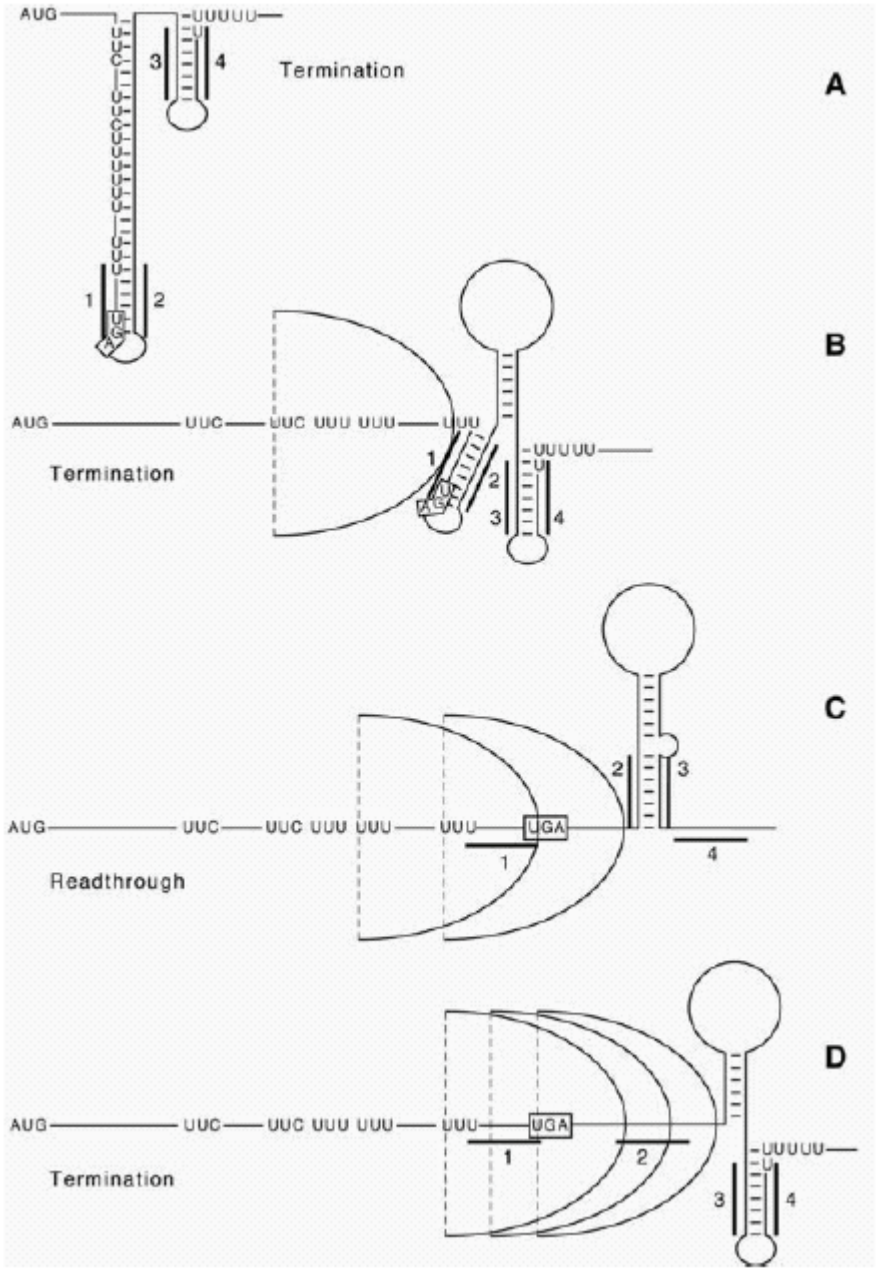


Figure 2.8 Alternative structures of the leader regions of the PheS-PheT operon.
The initiation codon, the five phenylalanine codons of the leader peptide and the U stretch of the terminator are written out, the stop codon is boxed. Important regions 1, 2, 3, and 4 are indicated. This picture is modified from
http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=eurekah&part=A19024

2.3.1.4 Translational regulation of theronyl-tRNA synthetase (thrRS) in E. coli

Although the expression of the majority of genes is controlled at the transcriptional level in *E. coli*, there are some genes, like those encoding ribosomal proteins (r proteins), that regulate their own expression by a negative feedback mechanism at the translational level[165]. Each r-protein operon contains a gene encoding a specific r-protein which is responsible for the feedback. These control r-proteins are generally thought to inhibit translation by binding to their own mRNA. Since a ribosome consists of rRNAs and proteins and its composition is not changed with growth rate during the exponential phase, some mechanism must exist to keep the ratio between rRNA and r-protein constant. It was proposed that growth rate-dependent control (GRDC) of r-protein synthesis results from the increase of rRNA synthesis with growth rate[166]. Under fast growth conditions, when rRNA synthesis is high, the set of control r-proteins (one per r-protein operon) will preferentially bind to rRNA and participate in ribosome assembly, i.e. they will not bind to their respective mRNAs and the expression of the r-protein operons will be derepressed. Under slow growth conditions, when rRNA synthesis is low, the control r-proteins will bind to their respective mRNAs and cause repression [166, 167]. Thus, the regulation of r-protein synthesis depends on whether the particular r-protein specifically binds to rRNA or itself [168].

In addition to ribosomal proteins, threonyl-tRNA synthetase (thrRS), a component of the translation machinery, was also shown to regulate its own mRNA translation both *in vivo* and *in vitro* [169]. The adjustment mechanism is similar to that displayed in the case of ribosomal proteins. ThrRS combines to its own leader mRNA which is a region of about 120 nts called the

operator and consists of four structural domains: Shine-Dalgarno (SD) sequence and translation start codon (domain 1); two anticodon Stem-Loops (ACSLs, special kind of hairpins with specific anticodon in the loop region, domain 2 and domain 4); single-stranded region (domain 3) (Fig. 2.9). Ribosome binds to thrS mRNA in two disconnected sites: the domain 1 and domain 3. These two domains are brought close to each other by the stem-loop structure formed by the domain 2. While the two stem loops was recognized by the enzyme through a simulating the anti-codon arm of $tRNA^{Thr}$. Although thrRS and ribosome bind to different domains, the competition of binding thrS mRNA between thrRS and ribosomes can be introduced by space hindrance [170]. The binding between domain 2 or 4 and thrRS prevents the normal translation of thrS mRNA, while the uncharged $tRNA^{Thr}$, which also binds to thrRS as a competitor to domain 2 and 4, is a derepressor. By default, cells contain the appropriate amount of thrRS to maintain the appropriate supply of Thr-$tRNA^{Thr}$ to synthesize proteins, at this point, since there is no need of new thrRSs, thrS mRNA translation is inhibited or partially inhibited through the binding between thrRS and domain 2 or 4. While in the case of threonine starvation, considering the above mentioned reaction 1, the concentration of threonine (amino acid) is considerably decreased. Normally the supply of threonyl-$tRNA^{Thr}$, which is the ultimate product of the above mentioned 2-step reactions, will be considerably decreased, but in order to synthesize proteins normally, the cell should keep aminoacyl-tRNA at a stable level. In the case of threonine starvation, to achieve this, the cell can increase the content of thrRS to start reaction 1, and eventually maintain a stable supply of threonyl-$tRNA^{Thr}$. It is through the derepression of thrS mRNA translation that the thrRS content is increased in *E. coli*. During this process, the relative surplus of uncharged $tRNA^{Thr}$ binds to thrRS which originally interacted with the leader regions,

and ribosome binds to domain 1 and 3 of the leader region smoothly, making thrS mRNA translation proceed smoothly. Domain 2 and 4 are not completely identical. The affinity of domain 2 binding to thrRS is higher than domain 4, thus domain 2 is essential for the whole control process while domain 4 plays a supporting role in the control [171, 172].

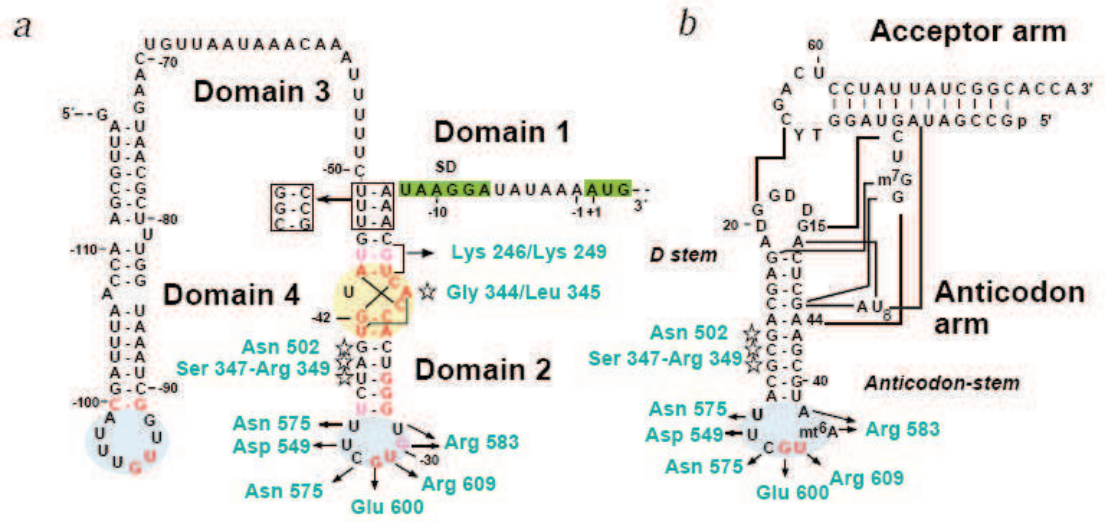


Figure 2.9 Structure of thrS leader region and $tRNA^{Thr}$

a. Four domains in thrS leader region: Domain 1, 2, 3 and 4 respectively; b. two structural domain of tRNA: Acceptor arm and Anticodon arm. The positions of amino acids that interact with Domain 2 and Anticodon arm respectively were marked out from the enzyme of thrRS.

2.3.2 the study of upstream and coding regions of thrS gene

2.3.2.1 Study Strategy and Results Summary

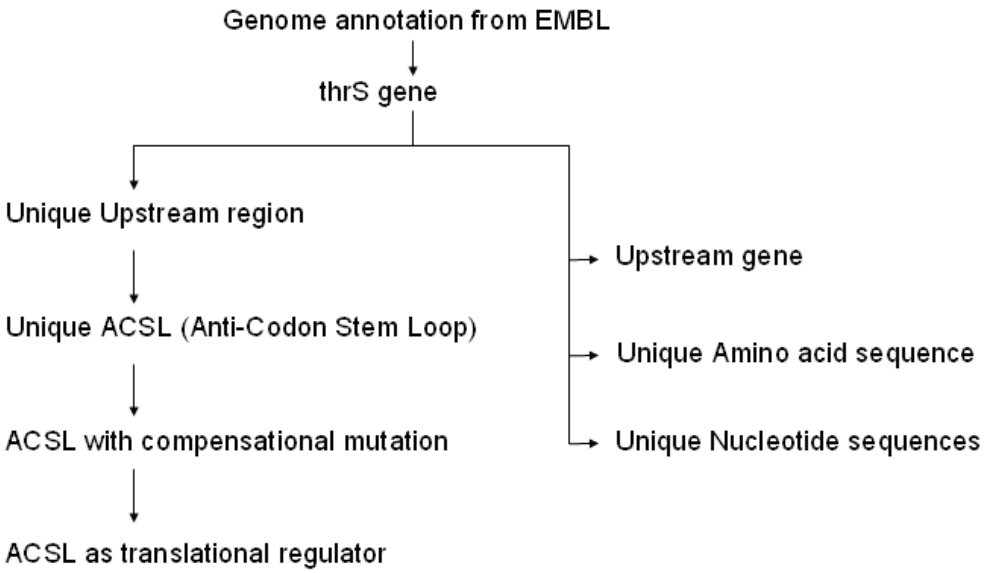


Figure 2.10 Study strategy for the study of upstream and coding regions of thrS gene

The study strategy for the comparison of upstream and coding regions of thrS gene is summarized in Fig. 2.10. In Article 3, firstly, 837 genomes with annotations were collected from EMBL entry point of INSDC, and thrS genes were found in 832 of them according to BBH search with the seed from *E. coli* K12. Secondly, the nucleotide sequences of both upstream regions and coding regions were extracted from the corresponding genome sequence, and the amino acid sequences were generated for the corresponding coding regions. Their unique sequences were obtained through the process of removing redundancy. Thirdly, the possible secondary structures were predicted for each unique upstream region by *mfold*, and the ACSL (Anti-codon Stem Loop) were analyzed, especially the stability of the ACSL and the compensational mutations that occurred in evolution. Fourthly, The evolutionary profiles were

analyzed according to these unique sequences, and found that (1) the upstream regions without functions evolved much faster than the coding region, (2) the upstream regions with functions evolved more slowly than the coding region, (3) there was an evolutionary jump if the upstream region had a non-essential function. Fifthly, the possible evolution events about chromosome rearrangement were evaluated through checking the upstream genes of thrS, and many different genes were found and have orthologs in *E. coli* K12, and some of them are related to phage. Lastly since the translational regulator can only be definitely determined through experiments, I just discussed the possibility of an ACSL as a functioning translational regulator.

Article 3

Title: the evolution of the upstream and coding region of thrS in Bacteria

Abstract

This paper described the regulating and coding regions of thrS genes in the domain of Bacteria undergoing different evolutionary pressure but not in full independent ways. The analysis of the upstream region of thrS coding sequences in 837 bacteria genomes demonstrated that the Anti-codon Loop Stem (ACLS) exists in a set of Gammaproteobacteria, Alphaproteobacteria and also some Firmicutes genomes with the validation of compensational mutation, and possible in other genomes without validation of compensational mutation partly due to the current limited sequenced genomes in the corresponding taxa. These phenomena reflect that the translational regulation mechanisms were invented several times independently by different bacteria. Analysis of the first upstream genomic objects, especially the first possible coding sequences, demonstrated that the upstream of thrS were hot spot sites for evolution, and showed different profiles between in distant genomes and in closely related genomes.

Introduction (or Background)

Regulation of gene expression can help a cell maintain homeostasis[1], and different mechanisms can be adopted for orthologs in the same class by different organisms [2, 3], such as Threonyl-tRNA synthetase (ThrRS) [2]. As a member of the 20 aminoacyl-tRNA synthetases (aaRSs), ThrRS is in charge of the joining of threonine to its cognate $tRNA^{Thr}$ carrying the

anticodon ACN (any one of the four codons ACU, ACC, ACA and ACG). This process is crucial in cells, since it is important to the normal synthesis of protein [4]. In condition of threonine starvation, to maintain the normal level of Thr-tRNA^{Thr} for synthesis of protein, at the very beginning the cell could achieve this by increasing the content of ThrRS, and there are at least two mechanisms to regulate the level of ThrRS in bacteria. One belongs to the classical transcriptional regulation of gene expression and is represented by *Bacillus subtilis* [5], and the other belongs to the translational level and is represented by *Escherichia coli* [6].

For regulation at the translational level, anti codon stem loop (ACSL), that is the hairpin loop containing tRNA^{Thr} anticodon (NGT, where N could be any of T, C, A and G), is very important. In E. coli, there are two such ACSLs with base triplets (CGU or UGU), and they can be bound by ThrRS. The regions flanking the first ACSL form the ribosome-binding site. These two bindings cannot exist simultaneously because of the stereo hindrance, and the regulation of thrS expression is realized through the switching of these two bindings [2].

In the present study, thrS, together with 16S rDNA and 3 housekeeping genes (*atpD*, *dnaJ* and *tuf*) were collected from the 837 fully sequenced bacterial genomes. The upstream regions of thrS were analyzed to see the possible ACSL. Combined with the preliminary phylogenetic analysis based on the above referred 5 genes, distribution of ACSL in bacteria were described and the inference about translational level and was proposed. The evolution of upstream region was analyzed through the inter-genetic region between thrS and its upstream gene, and was compared with its coding region.

Materials and Methods

Source of Data

Bacteria genomes and the corresponding preliminary annotations were downloaded from the EBI entry point (http://www.ebi.ac.uk/genomes/bacteria.html) of International Nucleotide Sequence Database Collaboration (INSDC) on June 26, 2009 [7]. The evident plasmid genomes, genomes with no CDS annotations and environment sample genomes were excluded during the analysis, and lastly 837 bacteria genomes were adopted in present study (Table 1, and Supplementary Table 1).

Identification of orthologs

BBH (Bidirectional Best Hit) Searching

Firstly, through keywords searching, obtain amino acid or nucleotide sequence of the target gene (marked as gene a) in reference genome (marked as genome A). For example, in present study, 16S rDNA, *thrS*, *atpD*, *dnaJ* and *tuf* sequences (for protein coding gene, both amino acid sequence and nucleotide sequence were considered) are collected in reference genome E. coli K12 (U00096) using keywords “16s ribosomal RNA”, “thrS”, “atpD”, “dnaJ” and “tufA” respectively. In the second step, collect all the paralogs of gene “a” in the reference genome “A” through the similarity searching programs such as blastp and blastn [8]. In the present study, two genes were considered as paralogs if their identity was more than 90%, and length difference was smaller than 20%. In the third step, use sequence of gene “a” from genome “A” as query sequence to fish out the best similar gene (marked as gene “b”) from another genome “B”. In the present study, two genes were considered as similar genes between different genomes if their identity was more than

30%, and length difference was smaller than 20%, and gene “a” from genome “A” may have several different best similar genes in genome “B”, only one was kept arbitrarily. In the fourth step, use sequence of gene “b” from genome B as the query sequence to fish out the best similar gene (marked as “c”) from genome A. The criterion here is the same to that adopted in the second step. Finally, check whether gene c is gene “a” or its paralogs, if it is, then gene “a” from genome “A” and gene “b” from genome “B” are a pair of orthologs.

Prediction of Anti Codon Stem Loop (ACSL)

RNA secondary structures were predicted using *mfold* 3.5, and reported △G values reflect *efn2* refinement (the *mfold* free-energy computation incorporating coaxial stacking and the Jacobson-Stockmeyer theory for multi-branched loops) [9, 10]. In the case where multiple structures were predicted, the most stable structure was chosen.

In bacteria, the folding of mRNA into secondary structure was started immediately during its formation, to check the existence of ACSL, we intuitively observed the folding results from two kinds of sequences with different length (specifically the longer one with 120-nt, and the shorter one with 60-nt around ACSL predicted from the longer one). And those ACSLs that appeared in both results were kept for further analysis.

The stability of an ACSL is quite affected by the size of the loop, the length of stem and the number of GC pairs in the stem, and the position of anticodon (NGT) can affect the possible interaction with thrRS during the translational regulation. All these four parameters are recorded and used as references to infer the possibility of the existence and stability of ACSL, and further function as a translational operator.

The possibility of the existence of the obtained secondary structure was furtherly analyzed through the concept of compensational mutation. Compensational mutation, which is preferred in evolution, refers to a mutation of normal base pairs which can be restored by the other Watson-Crick base pair or by the incidental unstable G:U pair [11]. We aligned the related sequences using Clustal W[12], and marked the stem and loop regions according to the prediction, and adjusted manually to view the possible compensational mutations.

Comparison between ACSL, upstream region and coding region

Comparison of different regions (ACSL, upstream region and coding region) related to thrS was carried out in two different ways. For the closely related organisms, since there are some sequences from different genomes identical for the three regions respectively, and they are not consistent for the three regions. For an example, see strains in *E. coli-Shigella* group described in the part of results. We defined a redundant group for a specific region if their sequences from different genomes are identical between each other. And the sequence in a redundant group is called unique sequence for the corresponding redundant group, specifically, the unique ACSL, unique upstream sequence and unique coding sequence. And to analyze whether two regions have identical sequences or not, we extracted sequences from the ACSL and its 5 nucleotides flanking at both sides, the upstream 120-nt sequences to the translation initiation site, and both the amino acid sequences and nucleotide sequences for the whole coding region.

To compare the distant related organisms, we first aligned the “unique sequences” through

CLUSTAL W. And the obtained multi-alignment were used as the input for *dnadist* and *protdist* programs in PHYLIP [13] to compute the similarity matrix for both nucleotide sequences and amino acid sequences. Later the comparisons were carried out using the pair-wised sequence differences in R environment.

Result

1.ThrS Ortholog Identification

Through BBH searching, with the ThrRS sequence from *E. coli* K12 as the query sequence, 832 among 837 bacteria genomes have thrS gene annotations. For *Ralstonia solanacearum* IPO1609 genome, later nucleotide sequence analysis shows the thrS gene split into two parts, and the BBH searching picked up the larger part. In *Eubacterium eligens* ATCC 27750, the best similar gene identified by ThrRS (K12) fished out ProRS in strain K12 as the best similar gene in the third step of BBH searching, and was not treated as a thrS ortholog. While in *Candidatus Sulcia muelleri* GWS and other 4 genomes, no similar gene could be identified by ThrRS (K12) (Supplementary Table 2).

2. ACSL prediction from the upstream sequence of thrS

After removing the redundant sequences from our 832 collected thrS-containing genomes, 645 unique upstream sequences were left, and among them except for *Nautilia profundicola* AmH and *Onion yellows phytoplasma* OY-M, the 120 nts sequences upstream to the translation initiation codon in the other genomes contain anti-codon (NGT) numbered from 1 to 13 (Supplementary

Table 3). Among the other 643 unique upstream sequences, 231 of them were predicted to fold into a secondary structure containing one or two ACSLs by *mfold* software, and the total number of ACSLs predicted is 280. 141 of them appeared at the similar positions to Domain 2 in *E. coli*, and were marked as ACSL I, 117 of them at the similar positions to Domain 4 in *E. coli*, and were marked as ACSL II; and 22 ACSLs left appeared at the similar positions to Domain 3, and were marked as ACSL III (Supplementary Table 3).

Although there were 280 confirmed ACSLs from different 120-nt upstream regions, some of the ACSLs share the same sequences, for example, ACSL IIs from three representative strains (namely Angola、biovar Microtus str. 91001 and YPIII) of *Yersinia pestis*, and ACSL Is from two representative strains (K-12 and 536) of *E. coli* and the representative strain (Ss046) of *Shigella sonnei*. After reducing the redundancy, 215 unique (non-redundant) ACSL patterns, whose four parameters namely loop size, NGT position, stem size, and GC pairs in stem were obtained in current study (Supplementary Table 4).

Among the 215 unique ACSLs from the prediction of mfold, the smallest loops, composed by the exact anticodon (Fig1. a); the largest loop, composed by 20 nucleotides (Fig1. b); and more than half (143) of the ACSLs contain a loop with size from 6 to 9, for example the two ACSLs from *E. coli* (Fig1. c). And in 83, 35 and 42 ACSLs, the anticodon started from 3, 2 and 1 nt respectively from the 5' half stem (Fig1. c); in another 31 ACSLs, the anticodon started immediately from the 5' half-stem (Fig1. a); while in the ACSL from *Ehrlichia chaffeensis* str. Arkansas, the anticodon started from 12 nt from 5' half-stem (Fig1. b).

Among the 215 unique ACSLs predicted by *mfold*, ACSL II from *Aromatoleum aromaticum* EbN1contains only 2 base pairs (Fig1. d); ACSL II from *Lactobacillus salivarius* UCC118

contains the most base pairs, that is 21 base pairs (Fig1. e); more than half (114) ACSLs contains 5 to 7 base pairs, for example, ACSL I and II from *E. coli* contains 6 base pairs (Fig 1. c). According to the predictions, there is no G:C base pair in the ACSLs from 5 genomes, *Buchnera aphidicola* str. 5A (Acyrthosiphon pisum), *Brucella abortus* bv. 1 str. 9-941, *Nitrosomonas europaea* ATCC 19718, *Staphylococcus aureus* subsp. aureus JH1 and *Thermosipho africanus* TCF52B; whereas ACSL from *Acidiphilium cryptum* JF-5 contains 9 G:C base pairs and the majority of ACSLs contains 2 to 5 G:C base pairs in the stem region.

3. Compensational mutation observed in ACSLs

All 215 unique ACSLs and their corresponding 5 flanking nucleotides were extracted from DNA sequences and were aligned together, and adjusted manually. Possible compensational mutations could be found in several groups from Fig. 2. Take ACSL I and II from *E. coli* as an example:

The well documented *E. coli* ACSL II and other 6 unique ACSLs from Enterobacteriales constitute a group with high similarity. Take the first sequence in the multi-alignment in this group (and this criterion has also been taken for the group below once there is a description for multi-alignment), counted from the anti codon loop, the first C:G pair in the stem are identical in all ACSLs; the 2nd base pairs which are G:C in three strains, A:T in another three strains, and G:T in one strain; the 3rd base pairs which are A:T in four strains, G:T in two strains, and one un-paired nucleotides (G-G) appeared in *Serratia proteamaculans*; the 4th base pairs which are T:A in 6 strains, and C:G in one strain; the 5th base pair which are T:G in 3 strains, C:G in 2 strains, T:A in one strain and un-paired nucleotides (A-A) appeared in *Yersinia enterocolitica*; the 6th base pairs

which are T:A in 5 strains, T:G and A:T in 1 strains respectively; the 7th base pairs, which are all A:T basepairs and the flanking regions are also conserved in some of the strains.

The well-documented *E. coli* ACSL I appears in a large group of strains (specifically 41 strains) from 8 orders in Gammaproteobacteria. Except for the 6th position found to be constituted by identical T:A base pair, mutations can be viewed at the other 6 sites in the stem: the 1st base pair which were found mainly to be T:G and C:G pairs (specifically in 21 and 17 strains respectively) but T:A in *Marinobacter aquaeolei* in Alteromonadales and unpaired nucleotides in 2 strains in Chromatiales; the 2nd base pair which were found mainly to be C:G pairs (specifically in 31 strains) but G:C pairs were found in *Alteromonas macleodii* species and 4 other Vibrionales strains, such as *Vibrio vulnificus*, G:T was found in *Pseudoalteromonas atlantica* speices, and A:T pairs were found in the remaining 4 strains; the 3rd base pair, which were found mainly to be T:G and C:G pairs (specifically in 10 and 23 strains respectively), and G:C pair was presented in *Pectobacterium atrosepticum* species from Enterobacteriales and T:A pairs were presented in the remaining 7 strains; the 4th base pair which were found mainly to be A:T and G:T pairs (specifically in 16 and 11 strains respectively), and C:G, G:C and T:A pairs were found in 5, 4 and 4 strains respectively, and leaving a couple of unpaired nucleotides (A-C) from strain; the 5th base pair which were found mainly to be G:C pairs (specifically in 35 strains), and A:T were found in 4 strains and C:G pairs were found in 2 strains respectively; the 7th base pair which were found mainly to be G:C pairs (in 39 strains), and A:T pairs were found in Enterobacter sp. 638 and Acinetobacter sp. ADP1. And from the loop region in this kind of ACSL, the first and last nucleotides in most of the strains are both T, but mutated into T:G base pairs in 10 strains in Altermononadales, Pseudomonadales, Aeromonadales, and Alteromonadales.

In the Alphaproteobacteria class, there are two major groups about ACSL with compensational mutations that can be seen in 10 and 6 unique ACSLs respectively. Also, there are other ACSLs that can be found with evidence of compensational mutations in Alphaproteobacteria and Gammaproteobacteria, but with limited numbers of unique ACSLs. And unique ACSL could also be viewed in Betaproteobacteira, Deltaproteobacteria and Epsilonproteobacteria, even in Actinobacteria, Firmicutes, but some groups contain few observations, and it is impossible to find the confident compensational mutations.

4. Unique sequence and its representing redundant group

Two or more genomes were treated to constitute a redundant group if they had identical sequences related to ACSL, upstream region and coding region respectively. For example, in 30 strains from *E. coli-Shigella* group, all the ACSL Is and IIs are identical respectively, i.e. there is only one redundant group for each kind of ACSL; according to the upstream 120-nt sequences, these strains could be classified into 3 groups; according to the amino acid sequences from the coding region, they could be classified into 6 groups; while according to the nucleotide sequences of coding region, they could be classified into 17 groups.

For all the 832 thrS orthologs, according to the 120-nt sequences of thrS upstream region, amino acid sequences of thrS coding region, and nucleotide sequences of thrS coding region, there were 645, 695 and 745 non-redundant groups respectively (Supplementary Table 5). And for 280 ACSLs, there are 215 redundant groups.

5. Several comparisons between unique sequences

GC contents counted from ACSL, 120-nt upstream sequences and according to different codon positions of coding region are differentiated from each other (Fig. 3). GC contents from ACSL correlated with the upstream sequence more highly than others (Fig 3a), and among different ways for calculating GC contents in coding region, the GC content from the upstream sequence correlated with the third codon position the most highly (Fig. 3b).

For ACSL, upstream region and coding region, their sequences are different among organisms at different ranges respectively (Fig. 4). From the comparisons of pair-wised sequences (between ACSL and upstream region, between ACSL and coding region, and between upstream region and coding region respectively) in 84 orders, it was shown that the upstream region did not show similar trend of variation with the coding region, but it could be seen that the upstream region evolved much faster than the coding region, and showed a similar way to the third codon position which was not significant (Fig. 4). Results did not change too much when comparisons were concentrated on the family of Gammaproteobacteria (Fig. 5). When carrying out the comparison between the 25 species in Enterobacteria, the coefficient elevated great, but was still not significant (Fig. 6).

6. The genes and inter-genetic regions upstream of thrS

These genomic objects which have stayed upstream of thrS genes are different from one genome to another genome. According to the genome annotations at the time of downloading, they were annotated as CDS, tRNA or repeat region, pseudogene, and rRNA in 716, 87, 24 and 5 genomes respectively (Supplementary Table 6).

Among the previous mentioned upstream genomic objects, 35 of them had overlapped nucleotide sequences with thrS coding region, for example the overlapped sequences reach to 70 nt in osb251; another 18 of them were located at more than 1000 nt upstream to thrS, for example, the distances from b750p1419 and b346p1528 to thrS gene were 6704 and 2594 nts respectively; among the remaining 779 genomic objects, most of them have a distance less than 400 nts to thrS (Fig. 7).

Except for pseudogene and rRNA genes, among the remaining 803 genomes, the first CDSs upstream of thrS from 778 genomes have annotations in their corresponding genomes, and 765 CDSs have similar genes in *E. coli* K12. One large group of the first upstream CDSs has 51 members annotated as *dnaI* according their corresponding genome (they were annotated as "dnaI" in the field of gene, or "primosomal" related or "dnaI" related keywords in the field of product or function), but their best similar genes in *E. coli* annotated as *ydaV*, *cysC*, *minD* and others. This kind of gene is related to DNA replication with prophage origin. Another large group has 23 members with the best similar genes in E. coli annotated as *relA*, the GTP pyrophosphokinase. Also the numbers of annotations in different fields are listed in Supplementary Table 7.

Discussion

1. Prediction of Translational Operator in silico

Strictly speaking, it is only the experiments in the lab that can demonstrate whether the thrS gene is regulated at the translational level or not. But it is not practical to carry out experiments in

hundreds of genomes, for example in the 837 genomes collected in this study. An alternative choice is to select the most possible candidate genomes through the prediction *in silico*. The results of computational prediction can be rendered more reliable by strengthening through careful analyses after prediction (or computation). During the process of computational prediction, much biological knowledge, such as compensational mutation, can be used to judge whether the prediction is reliable or not. The results from computation can give the answer. In the case of thrS study, we carefully analyzed the possibilities of the existence of those predicted secondary structures.

The first consideration is how likely a computational way can confirm secondary structures. Our results from mfold prediction showed that two ACSLs functioning in E. coli translational regulation and the transcriptional terminator functioning in B. subtilis transcriptional regulation are all predicted by mfold, and to some extent this indicates that the computation can be used in the preliminary exploration of the secondary structures formed in the upstream of thrS gene. Secondly, if an ACSL is very similar to either one in the E. coli, it can be stated with great confidence that the related genomes are regulated at the translational level. Our results showed that ACSL I from E. coli has similar counterparts from another 40 genomes, and ACSL II from another 6 genomes, and compensational mutations have been found in both cases. This indicates that the related genomes should be considered first to say they were regulated at the translational level. Thirdly, these ACSLs should be considered with similar features as both experimentally characterized in E. coli. In this regard, two groups with 10 and 6 genomes respectively in Alphaproteobacteria are good examples, and compensational mutations can be found in each group, while when it comes to whether they were regulated at translational level, we still suggest

checking the experimental results. Fourthly, we should pay more attention to those genomes with other secondary structures formed besides ACSL. In some cases, these additional structures can result in another form of regulation. This is the case for B. subtilis in Firmicutes (for a detailed discussion, see below).

2. Different Evolutionary Traces between ACSL, upstream and Coding Region

For the distantly related organisms, the comparison between representatives from different orders indicates that the upstream region behaved like the 3rd codon position in the coding region, the upstream region evolved much faster than the coding region, and might have stayed in a saturated state of mutation. For the closely related organisms, the analysis of redundant groups indicates that the upstream region, especially those ACSLs formed in this region, evolved slower than the coding region since fewer differences could be found between genomes.

The redundant groups established with nucleotide sequences in the coding regions were more than those established with amino acid sequences. This can be easily explained, since 64 codons correspond to 20 amino acids. Any change of one codon would give a new redundant group at the nucleotide level, while this change may not result in the change at the amino acid level, and lead to no increment of the redundant group at the amino acid level, or even result in the change at the amino acid level, there should be only one amino acid, and also lead to one increment of redundant group at the amino acid level. This indicates that the change of redundant groups at amino acid level is smaller than that at nucleotide level.

For the change of upstream region, this should be considered under two different conditions.

The first is that the upstream region bears a certain function besides the connection of two genes. Especially if the function is essential to an organism and highly associated with secondary or higher structures, perhaps mutation, here the mutation is not only the change of nucleotides, but also may be the insertion or deletion of nucleotides, the formation of the structure is affected, then its function, the related individual can not survive. While if the function is not essential one, then the mutation in this region just affects the efficiency of survival, and the related mutation can be kept during evolution. On focussing on E. coli and its closely related genomes, the upstream regions of thrS in these genomes are functioning with well characterized secondary structures that are Domain 1, 2, 3 and 4. Among these domains, Domain 1 and 3 are responsible for the communication with ribosomes, and the differences between 3 unique upstream sequences in this group are one nucleotide difference in either of these two domains, and leading to no effect on the communication with ribosomes. Both Domain 2 and 4 are responsible for the communication with thrRS with different affinities. While in *Escherichia fergusonii*, Domain 4 does not exist since the nucleotide sequences is totally different from E. coli and cannot fold into ACSL. This indicates that there is a deletion in *E. fergusonii* and since this deletion is not lethal and then it could be kept.

In the second condition, the upstream region has no function except the connecting of two genes. Mutation in such region has less effect on its connecting role. Obviously, it is similar to the 3rd codon position. In many cases, the change of the 3rd codon position does not change the ultimate protein product, which means to some extent it is free for the mutation on 3rd codon position. Similar explanation could be applied to those non-functioning upstream regions. Although in B. subtilis and E. coli, the upstream regions of thrS have regulating roles, this region

may have no important function in other bacteria. If this is true, it is easy to explain why the pair-wised sequences differences are larger than those observed between the coding regions.

3. The Evolutionary Mechanism for upstream region

After checking the upstream genomic objects, especially the CDS of thrS, we think of the upstream region as a hot-spot site exposed to mutation, even the deletion and insertion of large fragments. According to the current genome annotations, and also their best similar genes in *E. coli* K12, from 24 genomes, these CDSs are annotated as "pseudogene" that are the results of some mutations; some CDSs are phage origin, which means these regions are still unstable; and in 87 genomes, the first upstream genomic object is documented as tRNA or repeat region, which are the preferred sites of new insertion of sequences.

And these mutations that happened in the upstream region may affect the function of regulation possessed in some taxa. There is a most attractive phenomenon viewed in three species from the family of Enterobacteriaceae. That is, from amino acid sequences from coding region, the pair-wised sequences differences for *Escherichia fergusonii vs. E. coli* and *E. fergusonii* and S. enterica are 99% and 96.5% respectively, while if viewed from the unique ACSL in the upstream region, *E. fergusonii* only contains ACSL I, and it is identical to 3 ACSL Is from Salmonella. In this regard, we can infer that both inherited ACSL I from the ancestor of *E. coli* and *E. fergusonii*, and there are some mutations in *E. coli*, while the *E. fergusonii* kept the original one; but *E. coli* also inherited ACSL II, while *E. fergusonii* did not.

4. thrS gene in the phylum of Firmicutes

In 50 genomes, BBH determined there is a thrS gene which could be confirmed by its corresponding annotation. What's more, according to the annotation, there are another 1 or 2 genes annotated as threonyl-tRNA synthetases in each genome. And among these 50 genomes, 30 belong to the phylum of Firmicutes. In our current study, the designed technique in BBH strategy only detects one of them, and with some modifications, it should be detect both (study in progress).

According to the experiments carried out in *B. subtilis*, the regulation of thrS happens at the transcriptional level, and the key domains are transcriptional terminator, T-box and the specifier domain. From the sequences, the first two domains are very close to each other, but the specifier domain is something too distant. Many studies have described the secondary structures in these three domains in detail, but there is little information about the secondary structure formed in the long inter-domain sequence between specifier domain and T-box. From the prediction of *mfold*, we found there is an ACSL formed in that region, besides the formation of transcriptional terminator and specifier domain. And the predicted ACSL might be stable, since it contains 4 GC pairs among 7 base pairs in the stem and 8 nucleotides in the loop. If compared with that ACSL in *E. coli*, there are 3 and 2 GC pairs among both 7 base pairs in the stem of ACSL I and II respectively, and 7 and 8 nucleotides in the loop of ACSL I and II.

Although there is a possibility that ACSL could be formed in the upstream region of thrS gene in *B. subtilis*. But the chance for translational regulation in this organism is still small. Since for translational regulation, the switch of induction and repression of thrS expression is caused by the simple steric hindrance. In the case for *B. subtilis*, it does not have a similar environment: firstly, it is quite far from the translation initiation codon, that means if it indeed can cooperate

with thrRS, the steric hindrance caused between the complexes of thrRS-upstream and ribosome-upstream is not significant; secondly, other secondary structures are much more evident than ACSL in *B. subtilis*, for example both the length and stability of the stem of transcriptional terminator are more significant than ACSL in *B. subtilis*, and this may reduce the chance of the combination between thrRS and ACSL. So we do think there should be no or little chance to function as the translational operator in *B. subtilis*. And I think this ACSL is indeed an interesting genomic object, whether it exists in truth, and if it has indeed existed, what is the role, whether the ACSL can function at the translational level in a specific human incubated environment. And all these should be studied in more detail in experiments.

Conclusion

1, Translational regulation may exist at least in Gammaproteobacteria and Alphaproteobacteria.

2, ACSL is an essential secondary structure for translational regulation but not sufficient.

3, the upstream region and coding region evolved in different ways. Specifically, if the upstream region functions with an important stable secondary structure, it will evolve more slowly than the coding region, while if that region does not play roles with stable structure, it will evolve much faster than the coding region.

4, the upstream region may have suffered a lot of mutations when it executes a non-essential function.

Reference

1. Lindsley, J.E. and J. Rutter, *Nutrient sensing and metabolic decisions.* Comp Biochem Physiol

B Biochem Mol Biol, 2004. **139**(4): p. 543-59.
2. Torres-Larios, A., et al., *Structural basis of translational control by Escherichia coli threonyl tRNA synthetase.* Nat Struct Biol, 2002. **9**(5): p. 343-7.
3. Lee, Y.S., et al., *Molecular basis of cyclin-CDK-CKI regulation by reversible binding of an inositol pyrophosphate.* Nat Chem Biol, 2008. **4**(1): p. 25-32.
4. Williams, T.A., K.H. Wolfe, and M.A. Fares, *No rosetta stone for a sense-antisense origin of aminoacyl tRNA synthetase classes.* Mol Biol Evol, 2009. **26**(2): p. 445-50.
5. Luo, D., et al., *In vitro and in vivo secondary structure probing of the thrS leader in Bacillus subtilis.* Nucleic Acids Res, 1998. **26**(23): p. 5379-87.
6. Moine, H., et al., *Messenger RNA structure and gene regulation at the translational level in Escherichia coli: the case of threonine:tRNAThr ligase.* Proc Natl Acad Sci U S A, 1988. **85**(21): p. 7892-6.
7. Kulikova, T., et al., *EMBL Nucleotide Sequence Database in 2006.* Nucleic Acids Res, 2007. **35**(Database issue): p. D16-20.
8. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
9. Mathews, D.H., et al., *Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.* Journal of Molecular Biology, 1999. **288**(8): p. 911-940.
10. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Research, 2003. **31**(13): p. 3406-3415.
11. Gutell, R.R., N. Larsen, and C.R. Woese, *Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective.* Microbiol Rev, 1994. **58**(1): p. 10-26.
12. Larkin, M.A., et al., *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. **23**(21): p. 2947-8.
13. Felsenstein, *PHYLIP-Phylogeny Inference Package (Version 3.2).* Cladistics, 1989. **5**: p. 164-166.

FIGURE LEGENDS

Figure 1. Some examples of ACSL predicted from mfold

ACSL was predicted from mfold with the 120-nt sequences from the upstream region of thrS gene and default parameters.

a. ACSL in *Sulfurihydrogenibium azorense* Az-Fu1(os724II Fig1a.pdf) and *Candidatus Ruthia magnifica* str. Cm (os172IIFig1a.pdf); b. ACSL in *Ehrlichia chaffeensis* str. Arkansas (os282II Fig1b.pdf); c. ACSLs in *E. coli*(os291I Fig1c.pdf); d. ACSL in *Bifidobacterium longum* DJO10A (os92I Fig1d.pdf); e. *Lactobacillus salivarius* UCC118 (os92I Fig1d.pdf) (note: I will merge all these sub files into an integrated one)

Figure 2. Multi-alignment of all ACSLs

The sequences of all the unique ACSLs predicted from mfold together with 5 nucleotides flanking the stem loop are extracted, and aligned initially in Clustal W, and manually checked and adjusted. The regions in blue represent the stem regions, and those sequences between two blue are the corresponding loops. Taxon names are as described in Figure 2.

Figure 3. GC contents comparison between 120-nt upstream region and coding region, and between ACSL and upstream region and coding region respectively

Each sequence was plotted as a point with x representing the GC content calculated from the coding region in different ways as showed in the x labels, and with y representing the GC content calculated from the 120-nt upstream region. And the regression line was also added, with the expression stated below x labels and R squares shown above the corresponding Figure.

Figure 4. Pair-wised sequence differences plot between 120-nt upstream region and coding region, and between ACSL and upstream region and coding region respectively

Each pair was plotted as a point with x representing the differences calculated from the coding region in different ways as showed in the x labels, and with y representing the differences calculated from the 120-nt upstream region. And the regression line was also added, with the expression stated below x labels and R squares shown above the corresponding Figure. This comparison includes all the representatives at the level of order.

Figure 5. Pair-wised sequence differences plot between 120-nt upstream region and coding region in Gammaproteobacteria

Pair-wised sequences differences plot are described in Figure 6. And this comparison is also at the level of order, with just all the representatives from the calss of Gammaproteobacteria.

Figure 6. Pair-wised sequence differences plot between 120-nt upstream region and coding region in Enterobacteriales

Pair-wised sequence differences plot are described in Figure 6. And this comparison is calculated at the level of species with all the representatives from the order of Enterobacteriales.

Figure 7. Length distribution of distance from the upstream genomic object to thrS gene

In 35 genomes, the upstream genomic objects are overlapped with thrS, in another 18 genomes, the upstream genomic objects are located too far from thrS with length larger than 1kb, and these distances were not reflected in this figure.

Tables

TABLE 1 Summary of genomes adopted in current study

Taxon level	Number
Kingdom	1
Phylum	22
Class	39
Order	157
Family	280
Genus	547
Species	830
Strain	831
Genome	837
Chromosome	914

This table is associated to the supplementary Table 1, and the numbers of adopted entities at different taxon level are listed.

Supplementary Files

Supplementary File 1
Filename: SupplementaryTable1.xls
File Format: Microsoft Excel document file
Title of data: Chromosomes adopted in present study
Description of data: Organism ID and Chromosome ID are assigned during the present study for operation convenience; the names for main taxon level, namely: kingdom, phylum, class, order, family, genus, species and strain, were checked by hand with the references from NCBI taxon database. If an unclassified taxon for a genome was viewed, it was assigned upper level's name with the first letter in lower case.

Supplementary File 2
Filename: SupplementaryTable2.xls
File Format: Microsoft Excel document file
Title of data: Searching of Orthologs
Description of data: Keyword searching results for thrS gene, and BBH searching results for thrS, 16S rRNA, *atpD*, *dnaJ* and *tuf* are documented. Organism ID is that assigned in the present study as described in Supplementary Table 1. The names of Strains are extracted from the genome annotation. Keywords' searching was carried out in the gene field with "thrS" as keyword, and in both product and function fields with "thrRS", "threonyl-tRNA synthetase", "threonine-tRNA synthetase" and "threonine-tRNA ligase" as keywords from the corresponding genome annotation, and the number of resulting genes are recorded down in the "gene" and "product/function" columns in the table respectively. BBH searching results were characterized in three levels, namely: 0 stands for that the query gene "a" from reference genome "A" can not fish out a best

similar gene “b” from genome “B”; 1 stands for that the best similar gene “b” detected by query gene “a” can not fish out a best similar gene “c” from the reference genome “A” or the detected gene “c” is neither the gene “a” nor a paralogs of gene “a”; 2 stand for that the query gene “a” can successfully detected its ortholog (that is gene “b”) from genome “B”.

Supplementary File 3
Filename: SupplementaryTable3.xls
File Format: Microsoft Excel document file
Title of data: Results of ACSLs from mfold prediction
Description of data: Results of ACSLs predicted from 120-nt sequences and 60-nt sequences from thrS upstream region were documented. Organism ID and Strain are similar to the description id “Supplementary Table 2”. Number of NGT was counted for the 120-nt sequences upstream to the translation initiation start site of thrS gene. dG is the free energy calculated for the secondary structure predicted from *mfold*. ACSL I, II, and III means that they were predicted from the regions of Domain 2, 4 and 3 respectively defined in the reference [2]. The field marked as “first time” means it was predicted based on the 120-nt sequences, and “Yes” in the cell means an appearance of ACSL in the corresponding region, “No” for no ACSL predicted; and the field marked as “second time” means it was predicted based on the 60-nt sub-sequences as described in the text, and “Confirmed” in the cell means a confirmation of that ACSL predicted according to the 120-nt sequences, “not Confirmed” means no ACSL was or different ACSL was predicted in the corresponding 60-nt sub-sequences, and “-” means there was no prediction for the correspondence organism since no ACSL was predicted according to the 120-nt sequences..

Supplementary File 4
Filename: SupplementaryTable4.xls
File Format: Microsoft Excel document file
Title of data: Unique ACSLs predicted from *mfold*
Description of data: A unique ACSL pattern is named by a representative ACSL from a redundant group. Four parameters “loop size”, “NGT position”, “Stem size” and “number of G:C (or C:G) pairs” are documented, and for detailed explanation, see text. And the other ACSLs in the redundant group are listed. If the sequences from the ACSL region and the 5 nucleotides flanking to the ACSL are both identical for two or more ACSLs from different strains adopted in present study, these ACSLs are defined as a redundant group.

Supplementary File 5
Filename: SupplementaryTable5.xls
File Format: Microsoft Excel document file
Title of data: Unique 120-nt upstream region and unique coding region
Description of data: Organism ID and strain fields are defined similar to Supplementary Table 1. Redundancy group is marked by a strain it contains, and the strain is randomly chosen in theory, but for convenience, the one attached with the smallest organism ID was used as the marker. And the field of “Origin of Redundancy Group” describes the redundancy group the corresponding strain belongs to and in the field of “Redundancy Members”, all strains belonging to the same redundancy group were collected together and listed out at the line of marker strain. The upstream

120-nt sequences of all thrS orthologs were compared and defined 645 redundancy groups. The redundancy group of coding region is presented at two levels: one is at the amino acid level, and the other is at the nucleotide level.

Supplementary File 6
Filename: SupplementaryTable6.xls
File Format: Microsoft Excel document file
Title of data: Annotations of the upstream genomic object to thrS gene
Description of data: Organism ID and strain fields are defined in the same manner as Supplementary Table 1. Inter-genetic length is counted from the last nucleotide of the upstream genomic object to the first nucleotide of initiation codon of thrS gene. Type of upstream genomic object is directly extracted from the genome annotation. “Gene Name in K12” and “product/function in K12” are the annotations in K12 about the ortholog of the upstream genomic object of thrS from the corresponding genome. “Gene Name in its own genome” and “product/function in its own genome” are annotations about the upstream genomic object of thrS in the genome indicated by Organism ID.

Supplementary File 7
Filename: SupplementaryTable7.xls
File Format: Microsoft Excel document file
Title of data: Number of **a**nnotation keywords related to the upstream genomic object of thrS
Description of data: Number of genomes is counted for the corresponding annotation type, specifically for “gene” field in *E. coli* K12 genome, “function” or “product” field in *E. coli* K12 genome, “gene” field in its own genome, “function” or “product” field in its own genome.

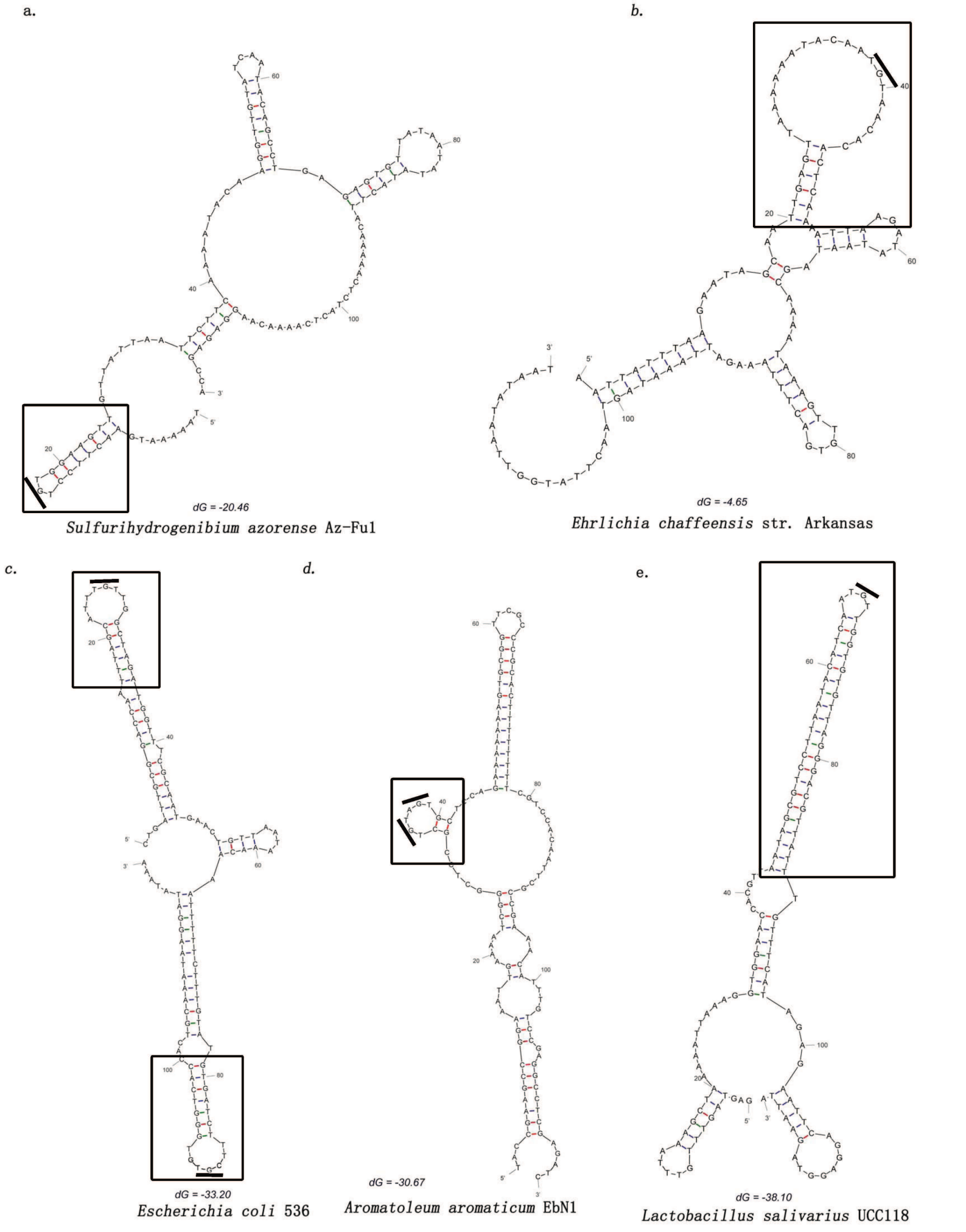
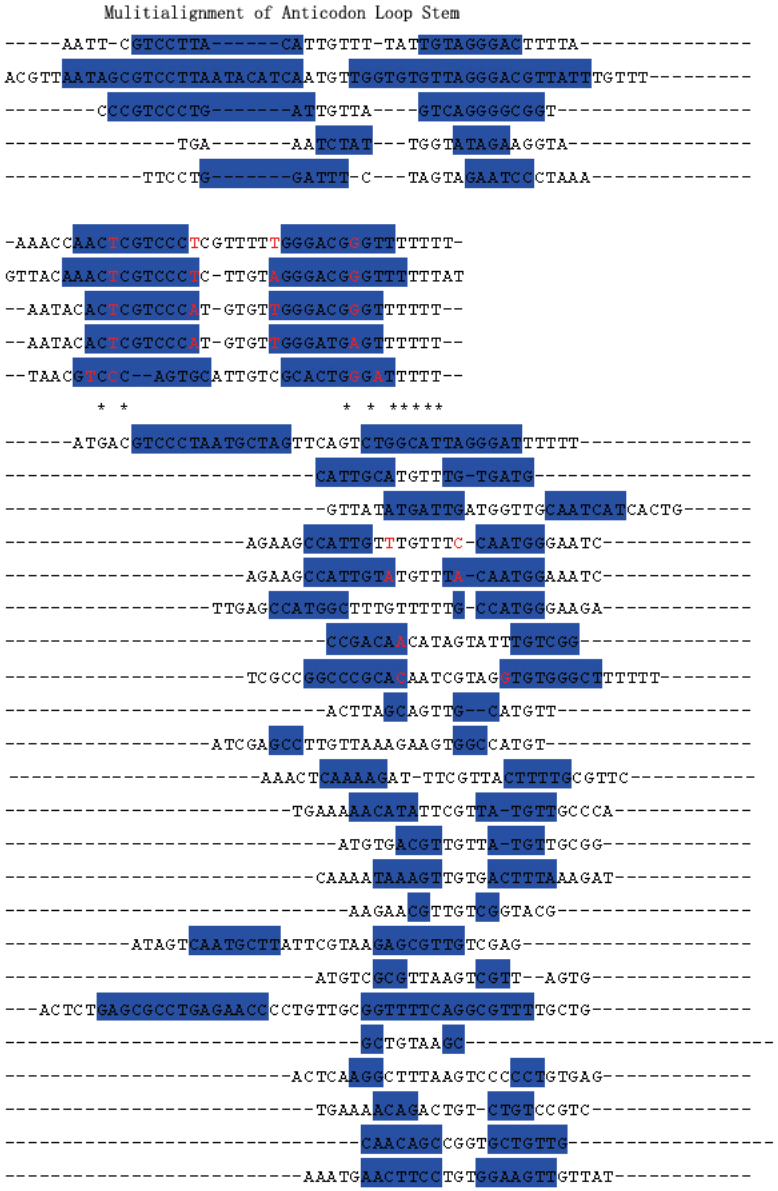
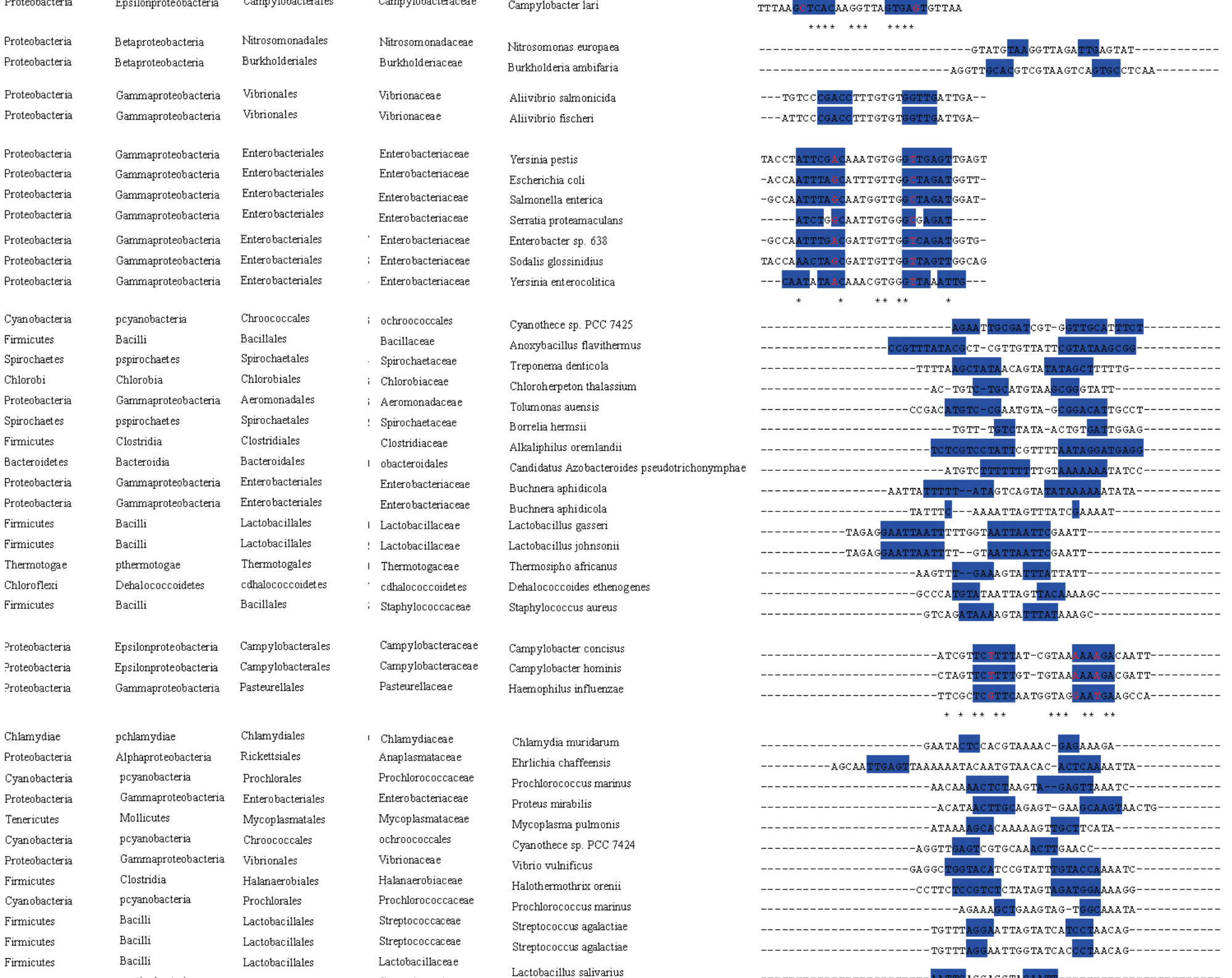

a.
dG = -20.46
Sulfurihydrogenibium azorense Az-Fu1
b.
dG = -4.65
Ehrlichia chaffeensis str. Arkansas
c.
dG = -33.20
Escherichia coli 536
d.
dG = -30.67
Aromatoleum aromaticum EbN1
e.
dG = -38.10
Lactobacillus salivarius UCC118

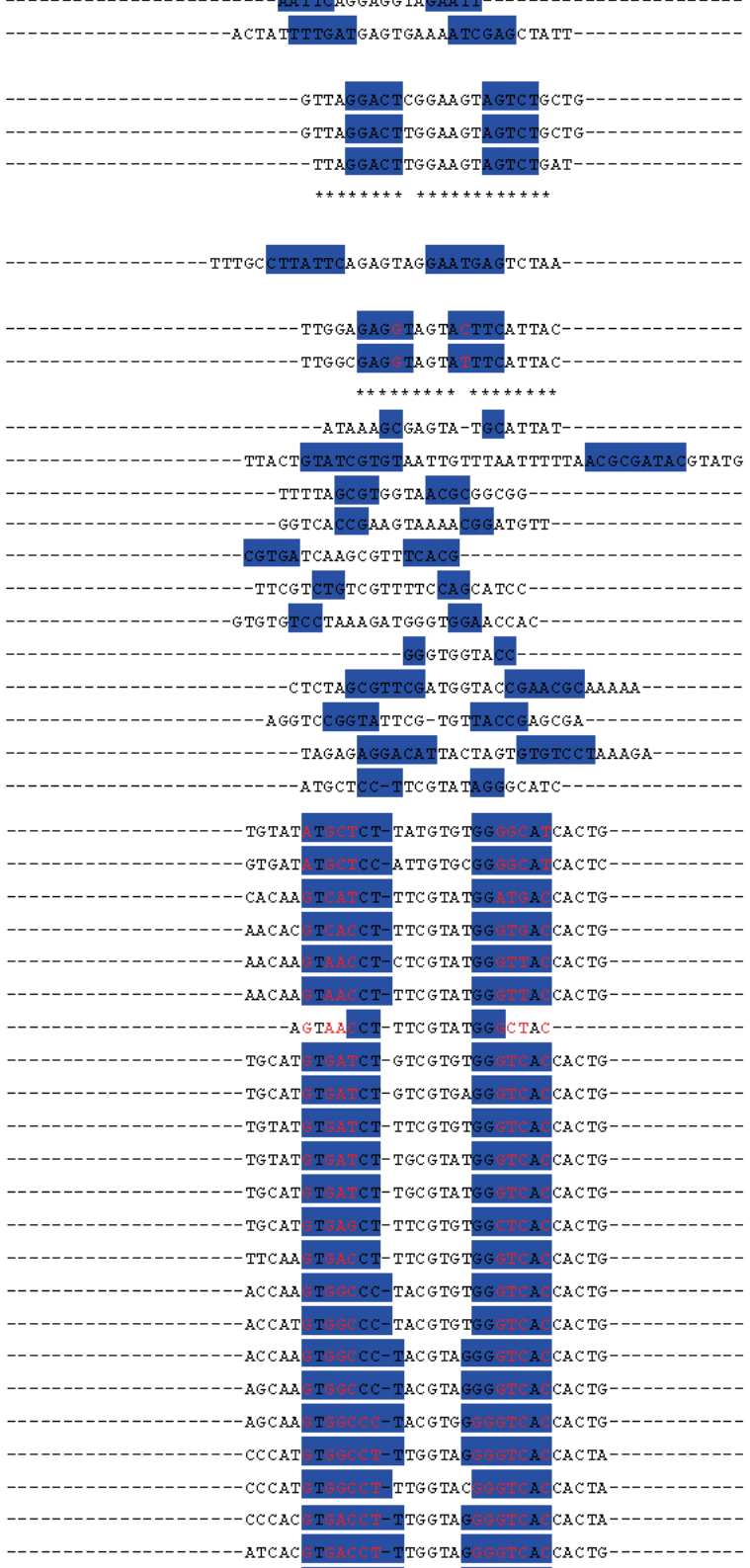
Figure 1

Phylum	Class	Order	Family	Species
Tenericutes	Mollicutes	Acholeplasmatales	Acholeplasmataceae	Acholeplasma laidlawii
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus salivarius
Firmicutes	Clostridia	Clostridiales	Peptococcaceae	Pelotomaculum thermopropionicum
Proteobacteria	Alphaproteobacteria	Rickettsiales	prickettsiales	Candidatus Pelagibacter
Proteobacteria	Gammaproteobacteria	Legionellales	Legionellaceae	Legionella pneumophila
Firmicutes	Bacilli	Bacillales	Bacillaceae	Anoxybacillus flavithermus
Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus haemolyticus
Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Clostridium botulinum
Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Clostridium botulinum
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus acidophilus
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus plantarum
Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus anthracis
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Psychrobacter arcticus
Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus cereus
Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus cereus
Firmicutes	Bacilli	Bacillales	Bacillaceae	Geobacillus thermodenitrificans
Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus subtilis
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Cellvibrio japonicus
Firmicutes	Clostridia	Halanaerobiales	Halanaerobiaceae	Halothermothrix orenii
Thermotogae	pthermotogae	Thermotogales	Thermotogaceae	Thermotoga neapolitana
Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides thetaiotaomicron
Bacteroidetes	Sphingobacteria	Sphingobacteriales	Flexibacteraceae	Cytophaga hutchinsonii
Proteobacteria	Alphaproteobacteria	Rickettsiales	Rickettsiaceae	Rickettsia prowazekii
Proteobacteria	Alphaproteobacteria	Rickettsiales	Anaplasmataceae	Ehrlichia chaffeensis
Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Ralstonia eutropha
Proteobacteria	Gammaproteobacteria	Alteromonadales	Pseudoalteromonadaceae	Pseudoalteromonas haloplanktis
Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	Bordetella avium
Deinococcus-Thermus	Deinococci	Deinococcales	Deinococcaceae	Deinococcus radiodurans
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Citrobacter koseri
Proteobacteria	Gammaproteobacteria	Thiotrichales	Piscirickettsiaceae	Thiomicrospira crunogena
Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	Gluconobacter oxydans
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus casei
Aquificae	paquificae	Aquificales	Hydrogenothermaceae	Sulfurihydrogenibium azorense
Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	Campylobacter jejuni
Actinobacteria	pactinobacteria	Actinomycetales	Corynebacteriaceae	Corynebacterium jeikeium

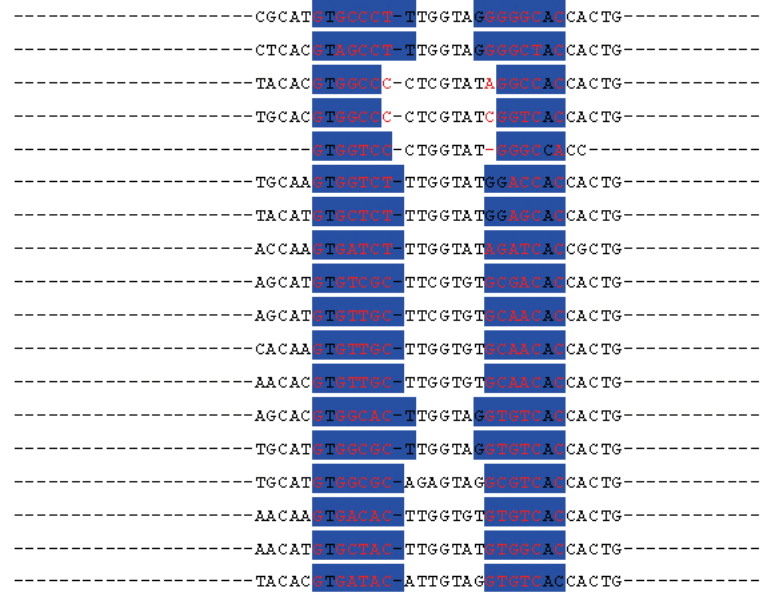



Proteobacteria Epsilonproteobacteria Campylobacterales Campylobacteraceae Campylobacter lari
Proteobacteria Betaproteobacteria Nitrosomonadales Nitrosomonadaceae Nitrosomonas europaea
Proteobacteria Betaproteobacteria Burkholderiales Burkholderiaceae Burkholderia ambifaria
Proteobacteria Gammaproteobacteria Vibrionales Vibrionaceae Aliivibrio salmonicida
Proteobacteria Gammaproteobacteria Vibrionales Vibrionaceae Aliivibrio fischeri
Proteobacteria Gammaproteobacteria Enterobacteriales Enterobacteriaceae Yersinia pestis
Proteobacteria Gammaproteobacteria Enterobacteriales Enterobacteriaceae Escherichia coli
Proteobacteria Gammaproteobacteria Enterobacteriales Enterobacteriaceae Salmonella enterica
Proteobacteria Gammaproteobacteria Enterobacteriales Enterobacteriaceae Serratia proteamaculans
Proteobacteria Gammaproteobacteria Enterobacteriales Enterobacteriaceae Enterobacter sp. 638
Proteobacteria Gammaproteobacteria Enterobacteriales Enterobacteriaceae Sodalis glossinidius
Proteobacteria Gammaproteobacteria Enterobacteriales Enterobacteriaceae Yersinia enterocolitica
Cyanobacteria pcyanobacteria Chroococcales ochroococcales Cyanothece sp. PCC 7425
Firmicutes Bacilli Bacillales Bacillaceae Anoxybacillus flavithermus
Spirochaetes pspirochaetes Spirochaetales Spirochaetaceae Treponema denticola
Chlorobi Chlorobia Chlorobiales Chlorobiaceae Chloroherpeton thalassium
Proteobacteria Gammaproteobacteria Aeromonadales Aeromonadaceae Tolumonas auensis
Spirochaetes pspirochaetes Spirochaetales Spirochaetaceae Borrelia hermsii
Firmicutes Clostridia Clostridiales Clostridiaceae Alkaliphilus oremlandii
Bacteroidetes Bacteroidia Bacteroidales obacteroidales Candidatus Azobacteroides pseudotrichonymphae
Proteobacteria Gammaproteobacteria Enterobacteriales Enterobacteriaceae Buchnera aphidicola
Proteobacteria Gammaproteobacteria Enterobacteriales Enterobacteriaceae Buchnera aphidicola
Firmicutes Bacilli Lactobacillales Lactobacillaceae Lactobacillus gasseri
Firmicutes Bacilli Lactobacillales Lactobacillaceae Lactobacillus johnsonii
Thermotogae pthermotogae Thermotogales Thermotogaceae Thermosipho africanus
Chloroflexi Dehalococcoidetes cdhalococcoidetes cdhalococcoidetes Dehalococcoides ethenogenes
Firmicutes Bacilli Bacillales Staphylococcaceae Staphylococcus aureus
Proteobacteria Epsilonproteobacteria Campylobacterales Campylobacteraceae Campylobacter concisus
Proteobacteria Epsilonproteobacteria Campylobacterales Campylobacteraceae Campylobacter hominis
Proteobacteria Gammaproteobacteria Pasteurellales Pasteurellaceae Haemophilus influenzae
Chlamydiae pchlamydiae Chlamydiales Chlamydiaceae Chlamydia muridarum
Proteobacteria Alphaproteobacteria Rickettsiales Anaplasmataceae Ehrlichia chaffeensis
Cyanobacteria pcyanobacteria Prochlorales Prochlorococcaceae Prochlorococcus marinus
Proteobacteria Gammaproteobacteria Enterobacteriales Enterobacteriaceae Proteus mirabilis
Tenericutes Mollicutes Mycoplasmatales Mycoplasmataceae Mycoplasma pulmonis
Cyanobacteria pcyanobacteria Chroococcales ochroococcales Cyanothece sp. PCC 7424
Proteobacteria Gammaproteobacteria Vibrionales Vibrionaceae Vibrio vulnificus
Firmicutes Clostridia Halanaerobiales Halanaerobiaceae Halothermothrix orenii
Cyanobacteria pcyanobacteria Prochlorales Prochlorococcaceae Prochlorococcus marinus
Firmicutes Bacilli Lactobacillales Streptococcaceae Streptococcus agalactiae
Firmicutes Bacilli Lactobacillales Streptococcaceae Streptococcus agalactiae
Firmicutes Bacilli Lactobacillales Lactobacillaceae Lactobacillus salivarius

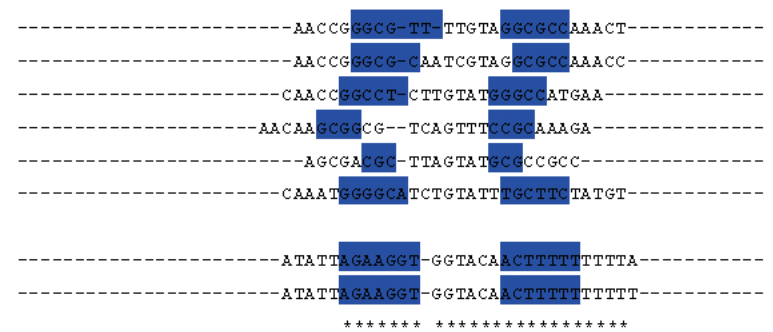
Actinobacteria	pactinobacteria	Actinomycetales	Corynebacteriaceae	Corynebacterium diphtheriae
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus pneumoniae
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus pneumoniae
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus pneumoniae
Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Pasteurella multocida
Proteobacteria	Gammaproteobacteria	cgammaproteobacteria	cgammaproteobacteria	Candidatus Vesicomyosocius okutanii
Proteobacteria	Gammaproteobacteria	cgammaproteobacteria	cgammaproteobacteria	Candidatus Ruthia magnifica
Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus pumilus
Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	Chlorobium chlorochromatii
Proteobacteria	Gammaproteobacteria	Oceanospirillales	Hahellaceae	Hahella chejuensis
Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Polynucleobacter necessarius
Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides thetaiotaomicron
Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Ralstonia metallidurans
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus reuteri
Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus saprophyticus
Proteobacteria	Gammaproteobacteria	Alteromonadales	Idiomarinaceae	Idiomarina loihiensis
Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Vibrio harveyi
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus reuteri
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter baumannii
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Enterobacter sp. 638
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter sp. ADP1
Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Actinobacillus succinogenes
Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus influenzae
Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus ducreyi
Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus parasuis
Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus somnus
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Serratia proteamaculans
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Yersinia pestis
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia coli
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia fergusonii
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Proteus mirabilis
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Pectobacterium atrosepticum
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Photorhabdus luminescens
Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella baltica
Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella frigidimarina
Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella halifaxensis
Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella woodyi
Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella sediminis
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Azotobacter vinelandii
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas stutzeri
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas entomophila
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Cellvibrio japonicus



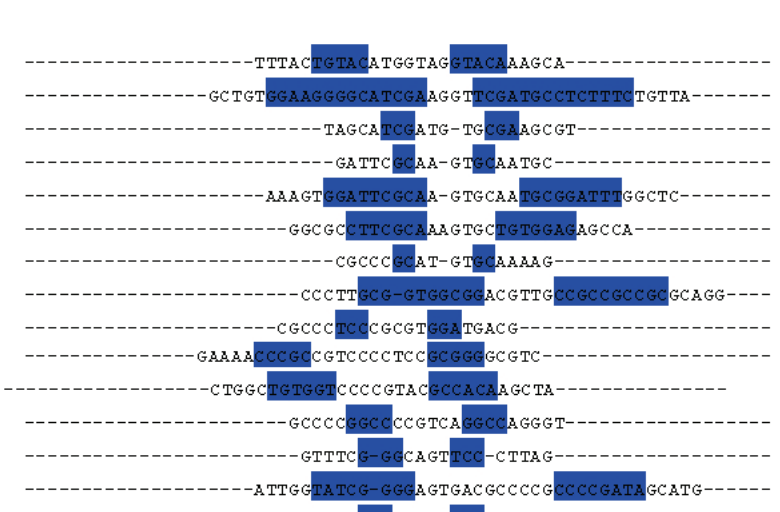
Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	Aeromonas hydrophila
Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	Saccharophagus degradans
Proteobacteria	Gammaproteobacteria	Chromatiales	Ectothiorhodospiraceae	Alkalilimnicola ehrlichii
Proteobacteria	Gammaproteobacteria	Chromatiales	Ectothiorhodospiraceae	Halorhodospira halophila
Proteobacteria	Gammaproteobacteria	Oceanospirillales	Halomonadaceae	Chromohalobacter salexigens
Proteobacteria	Gammaproteobacteria	Oceanospirillales	Hahellaceae	Hahella chejuensis
Proteobacteria	Gammaproteobacteria	Oceanospirillales	Oceanospirillaceae	Marinomonas sp. MWYL1
Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	Marinobacter aquaeolei
Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Aliivibrio salmonicida
Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Aliivibrio fischeri
Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Vibrio cholerae
Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Vibrio vulnificus
Proteobacteria	Gammaproteobacteria	Alteromonadales	Idiomarinaceae	Idiomarina loihiensis
Proteobacteria	Gammaproteobacteria	Alteromonadales	Pseudoalteromonadaceae	Pseudoalteromonas atlantica
Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	Alteromonas macleodii
Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	Tolumonas auensis
Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Photobacterium profundum
Proteobacteria	Gammaproteobacteria	Alteromonadales	Pseudoalteromonadaceae	Pseudoalteromonas haloplanktis



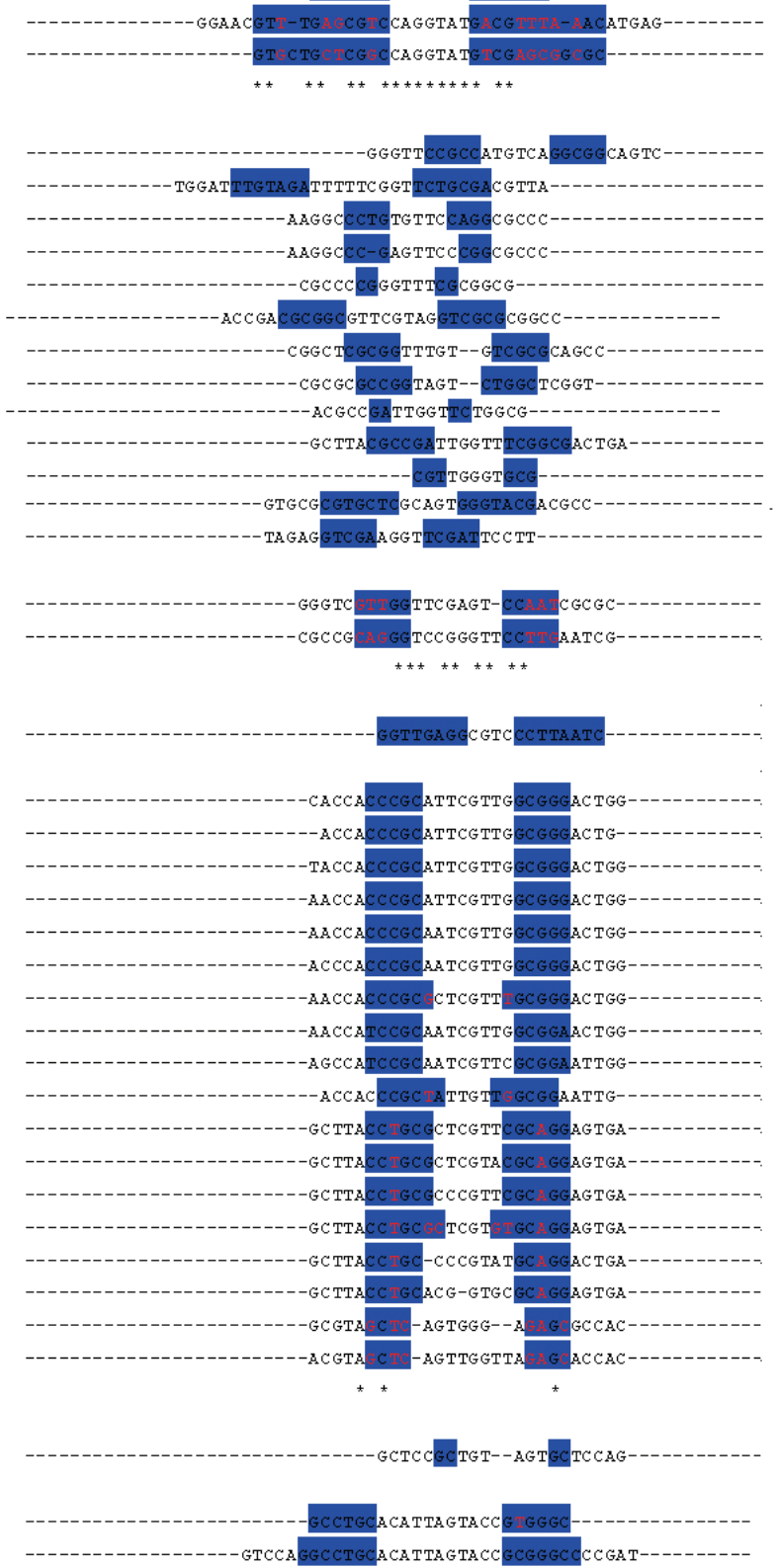
Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Jannaschia sp. CCS1
Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Rhodobacter sphaeroides
Proteobacteria	Gammaproteobacteria	Chromatiales	Chromatiaceae	Nitrosococcus oceani
Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	Chloroherpeton thalassium
Proteobacteria	Gammaproteobacteria	Alteromonadales	Colwelliaceae	Colwellia psychrerythraea
Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Vibrio splendidus
Proteobacteria	Alphaproteobacteria	Rickettsiales	Anaplasmataceae	Ehrlichia ruminantium
Proteobacteria	Alphaproteobacteria	Rickettsiales	Anaplasmataceae	Ehrlichia ruminantium



Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Actinobacillus succinogenes
Proteobacteria	Deltaproteobacteria	Desulfuromonadales	Geobacteraceae	Geobacter sulfurreducens
Actinobacteria	pactinobacteria	Actinomycetales	Pseudonocardiaceae	Saccharopolyspora erythraea
Proteobacteria	Epsilonproteobacteria	Campylobacterales	Helicobacteraceae	Helicobacter pylori
Proteobacteria	Epsilonproteobacteria	Campylobacterales	Helicobacteraceae	Helicobacter pylori
Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	Aeromonas hydrophila
Proteobacteria	Alphaproteobacteria	Sphingomonadales	Erythrobacteraceae	Erythrobacter litoralis
Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	Acidiphilium cryptum
Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Rhizobium etli
Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	Methylobacterium chloromethanicum
Actinobacteria	pactinobacteria	Actinomycetales	Micrococcaceae	Arthrobacter aurescens
Deinococcus-Thermus	Deinococci	Thermales	Thermaceae	Thermus thermophilus
Proteobacteria	Deltaproteobacteria	Desulfobacterales	Desulfobacteraceae	Desulfobacterium autotrophicum
Actinobacteria	pactinobacteria	Actinomycetales	Mycobacteriaceae	Mycobacterium vanbaalenii
Actinobacteria	pactinobacteria	Actinomycetales	Micrococcaceae	Kocuria rhizophila
Actinobacteria	pactinobacteria	Actinomycetales	Corynebacteriaceae	Corynebacterium glutamicum



Actinobacteria	pactinobacteria	Actinomycetales	Corynebacteriaceae	Corynebacterium diphtheriae
Actinobacteria	pactinobacteria	Actinomycetales	Corynebacteriaceae	Corynebacterium aurimucosum
Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	Bordetella petrii
Elusimicrobia	pelusimicrobia	pelusimicrobia	pelusimicrobia	envSample Termite group 1
Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Stenotrophomonas maltophilia
Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Stenotrophomonas maltophilia
Verrucomicrobia	Opitutae	Opitutales	Opitutaceae	Opitutus terrae
Acidobacteria	pAcidobacteria	pAcidobacteria	pAcidobacteria	Acidobacteria bacterium
Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	Marinobacter aquaeolei
Actinobacteria	pactinobacteria	Actinomycetales	Microbacteriaceae	Clavibacter michiganensis
Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	Beijerinckia indica
Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	Methylocella silvestris
Tenericutes	Mollicutes	Entomoplasmatales	Entomoplasmataceae	Mesoplasma florum
Actinobacteria	pactinobacteria	Actinomycetales	Nocardiaceae	Rhodococcus erythropolis
Firmicutes	Clostridia	Thermoanaerobacterales	Thermodesulfobiaceae	Coprothermobacter proteolyticus
Proteobacteria	Gammaproteobacteria	Chromatiales	Chromatiaceae	Nitrosococcus oceani
Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Rhodopseudomonas palustris
Cyanobacteria	pcyanobacteria	Chroococcales	ochroococcales	Synechococcus elongatus
Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Agrobacterium tumefaciens
Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Rhizobium sp. NGR234
Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Sinorhizobium medicae
Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Agrobacterium radiobacter
Proteobacteria	Alphaproteobacteria	Rhizobiales	Brucellaceae	Brucella abortus
Proteobacteria	Alphaproteobacteria	Rhizobiales	Brucellaceae	Ochrobactrum anthropi
Proteobacteria	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	Mesorhizobium loti
Proteobacteria	Alphaproteobacteria	Rhizobiales	Bartonellaceae	Bartonella bacilliformis
Proteobacteria	Alphaproteobacteria	Rhizobiales	Bartonellaceae	Bartonella henselae
Proteobacteria	Alphaproteobacteria	Rhizobiales	Bartonellaceae	Bartonella quintana
Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Bradyrhizobium
Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Rhodopseudomonas palustris
Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Oligotropha carboxidovorans
Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Rhodopseudomonas palustris
Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	Methylobacterium radiotolerans
Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	Methylobacterium nodulans
Proteobacteria	Deltaproteobacteria	Desulfobacterales	Desulfobulbaceae	Desulfotalea psychrophila
Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	Neisseria gonorrhoeae
Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Aromatoleum aromaticum
Actinobacteria	pactinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium longum
Actinobacteria	pactinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium longum



Actinobacteria	pactinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium animalis	----------GTCCAGGTCTGCACCTT-GTACCGCGGACCCAGAT----------
					****** ** ****** ** *
Actinobacteria	pactinobacteria	Actinomycetales	Micrococcaceae	Kocuria rhizophila	----------GGGACGCCCGC-CCCCGGTACGGGCGGGGCACCCA----------
Proteobacteria	Deltaproteobacteria	Myxococcales	Polyangiaceae	Sorangium cellulosum	----------ACACACCTGA-CGCTCGTACAGTCGGGGGAAA----------
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas putida	----------GGCT-TCTGCCACTGTGGGGCAGG-CTTC----------
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas putida	----------GGCT-TCTGCCACTGTGAGGCAGG-CTTC----------
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas fluorescens	----------AGGCT-CCTGCCGCTGTAAGGCAGG-CGTTC----------
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas stutzeri	----------CAGGTGCCTGCCATTGTGCGGCAGGCCTGCT----------
					* * ***** *** ****** *
Proteobacteria	Deltaproteobacteria	Syntrophobacterales	Syntrophobacteraceae	Syntrophobacter fumaroxidans	----------ATCTTCCCTTGTAGGG-CAGGG----------
Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia cenocepacia	----------GGTCTGCCGCCCACTGTCGGCATCAC----------
Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia cenocepacia	----------TGGTCTGCCGCCCACTGTCGGCATCACG----------
Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia ambifaria	----------TGGTTTGCCGCCCACTGTCGGCATCACG----------
					*** **********
Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	Desulfovibrio desulfuricans	----------AAAGCAGGGGGAGGCCCGTAGCCTCCCCCTTTTTG----------
Actinobacteria	pactinobacteria	Actinomycetales	Nocardiaceae	Rhodococcus erythropolis	----------CACTCGGTTGTCTCACCATGTC----------
Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Acidovorax citrulli	----------CCCCCGAAAAAGCGCGGACATCGTTCCGCGTTTTTTTTCGCC---
Proteobacteria	Betaproteobacteria	Hydrogenophilales	Hydrogenophilaceae	Thiobacillus denitrificans	----------CCCCGAAAAGCGCGG---CCGGTCCGCGCTTTTTTTCT----
Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Acidovorax sp. JS42	----------CTTACAAGCGCGGAC-CTGTACCGCGCTTTTTGC-------
Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Diaphorobacter sp. TPSY	----------CCCATAAAAGCGCGGAC-CTGTACCGCGCTTTTTGCTT----
Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Polaromonas sp. JS666	----------TTATCAAAAGCGCGG-C-CGGTTCCGCGCTTTTTGTTT----
Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Verminephrobacter eiseniae	----------CTCCGAAAAGCGCGAGC-CTGTTTCGCGCTTTTTCGTT----
Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Polaromonas naphthalenivorans	----------ATTGCAAAGCGCGGCT--TGGTCCGCGCTTTTTTAT------
					******* **********
Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella baltica	----------GCGGCTAAGACTCCTGTACTGTCTTAACCTA----------
Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella oneidensis	----------GCGGCTAAGACTCCTGTACTGTCTTAACCTC----------
Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella sediminis	----------TAAGACTCTTGTAATGGCTTAT----------
Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella halifaxensis	----------AACACAGACTCCAGTAAGGTCTAGTCC----------
Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella pealeana	----------AACACAGACTACTGTAAGGTCTAGTCC----------
Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella piezotolerans	----------AACATAGACTCCTGTACGGTCTTGTCC----------
Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella loihica	----------TAGATCAAAACTCCGGTAATGTTTTGCGCTT----------
					* *** *** * *

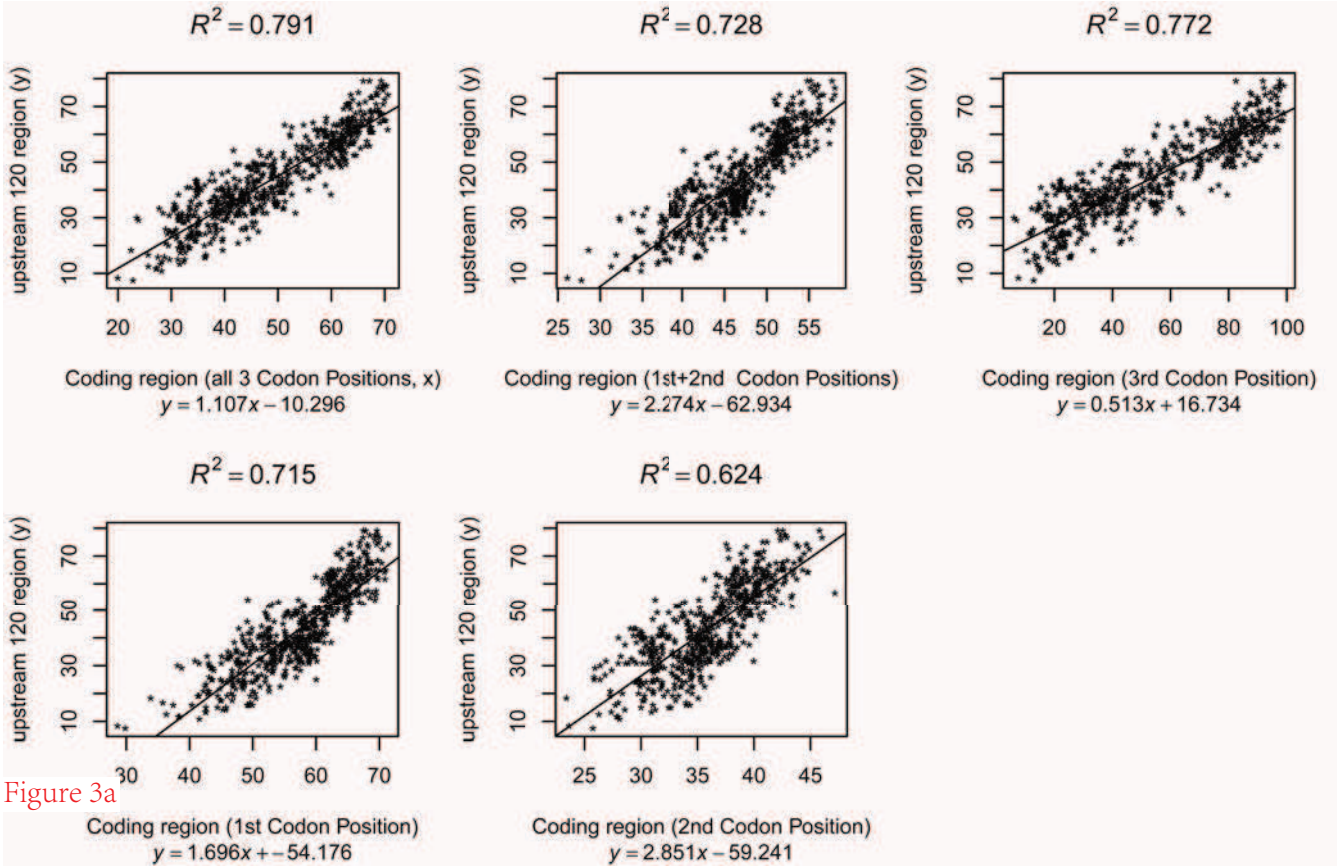

$R^2 = 0.791$
upstream 120 region (y)
Coding region (all 3 Codon Positions, x)
$y = 1.107x - 10.296$
$R^2 = 0.728$
upstream 120 region (y)
Coding region (1st+2nd Codon Positions)
$y = 2.274x - 62.934$
$R^2 = 0.772$
upstream 120 region (y)
Coding region (3rd Codon Position)
$y = 0.513x + 16.734$
$R^2 = 0.715$
upstream 120 region (y)
Coding region (1st Codon Position)
$y = 1.696x + -54.176$
$R^2 = 0.624$
upstream 120 region (y)
Coding region (2nd Codon Position)
$y = 2.851x - 59.241$

Figure 3a

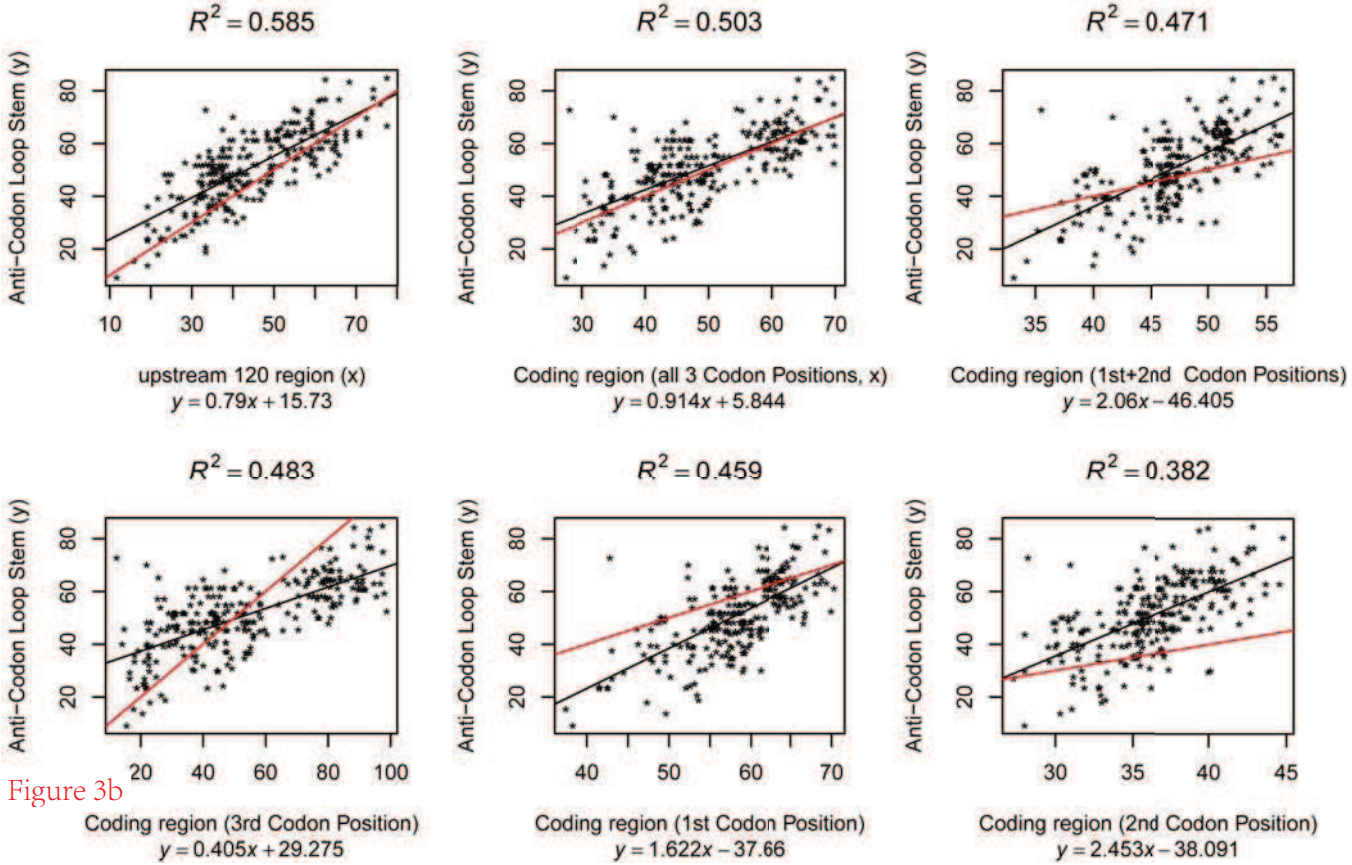

$R^2 = 0.585$
Anti−Codon Loop Stem (y)
upstream 120 region (x)
$y = 0.79x + 15.73$
$R^2 = 0.503$
Coding region (all 3 Codon Positions, x)
$y = 0.914x + 5.844$
$R^2 = 0.471$
Coding region (1st+2nd Codon Positions)
$y = 2.06x - 46.405$
$R^2 = 0.483$
Coding region (3rd Codon Position)
$y = 0.405x + 29.275$
$R^2 = 0.459$
Coding region (1st Codon Position)
$y = 1.622x - 37.66$
$R^2 = 0.382$
Coding region (2nd Codon Position)
$y = 2.453x - 38.091$

Figure 3b

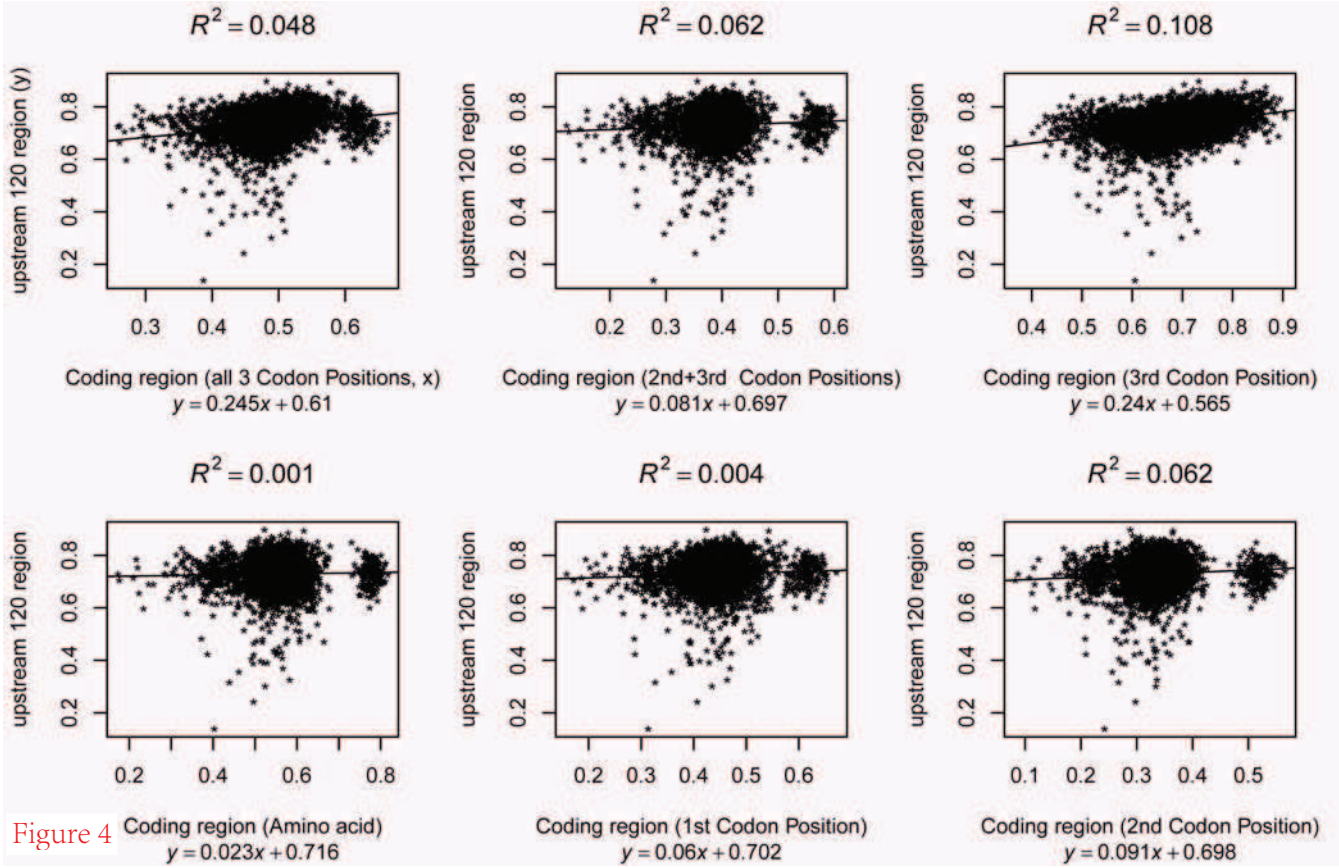

$R^2 = 0.048$
upstream 120 region (y)
Coding region (all 3 Codon Positions, x)
$y = 0.245x + 0.61$
$R^2 = 0.062$
upstream 120 region
Coding region (2nd+3rd Codon Positions)
$y = 0.081x + 0.697$
$R^2 = 0.108$
upstream 120 region
Coding region (3rd Codon Position)
$y = 0.24x + 0.565$
$R^2 = 0.001$
upstream 120 region
Coding region (Amino acid)
$y = 0.023x + 0.716$
$R^2 = 0.004$
upstream 120 region
Coding region (1st Codon Position)
$y = 0.06x + 0.702$
$R^2 = 0.062$
upstream 120 region
Coding region (2nd Codon Position)
$y = 0.091x + 0.698$

Figure 4

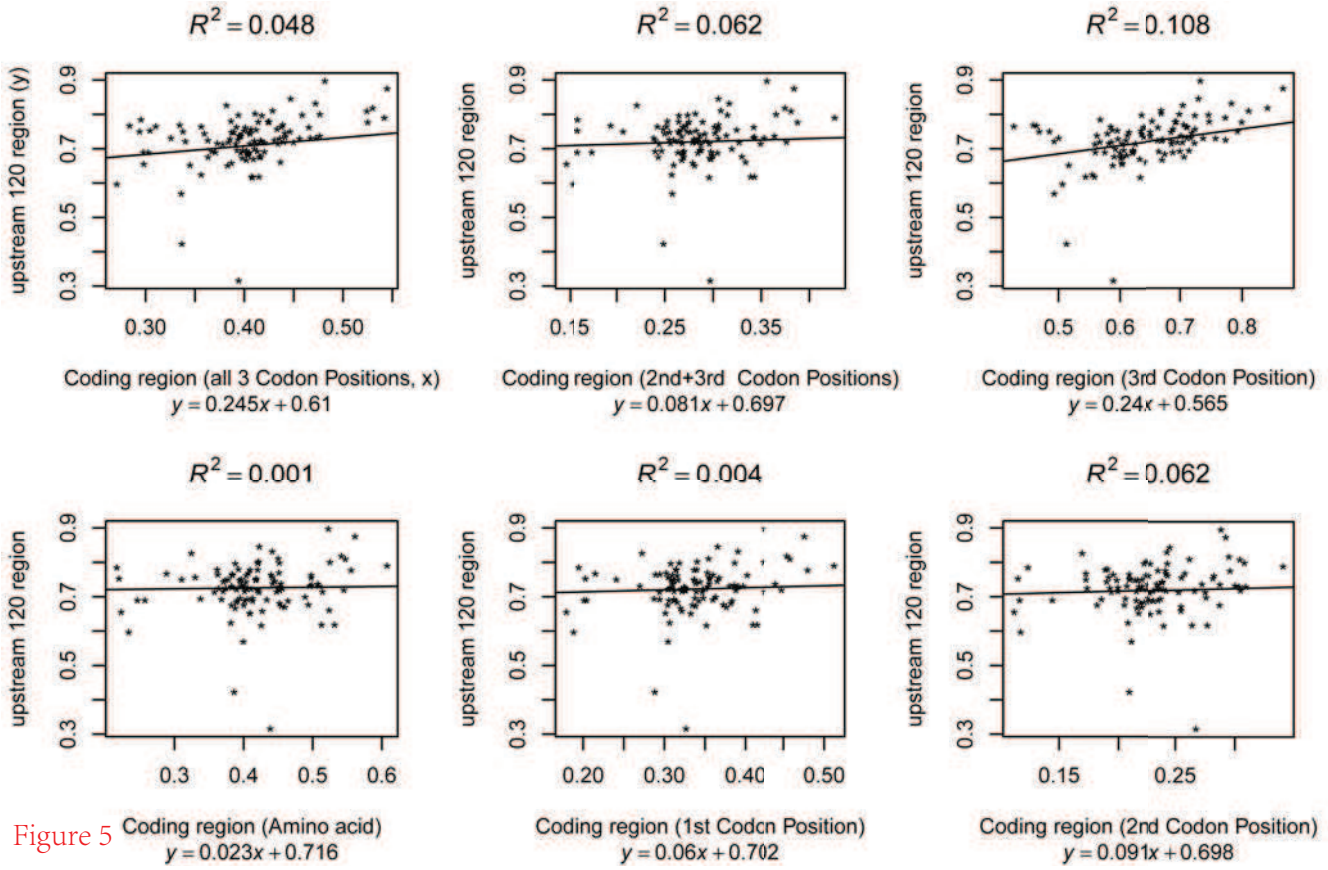

$R^2 = 0.048$
upstream 120 region (y)
Coding region (all 3 Codon Positions, x)
$y = 0.245x + 0.61$
$R^2 = 0.062$
upstream 120 region
Coding region (2nd+3rd Codon Positions)
$y = 0.081x + 0.697$
$R^2 = 0.108$
upstream 120 region
Coding region (3rd Codon Position)
$y = 0.24x + 0.565$
$R^2 = 0.001$
upstream 120 region
Coding region (Amino acid)
$y = 0.023x + 0.716$
$R^2 = 0.004$
upstream 120 region
Coding region (1st Codon Position)
$y = 0.06x + 0.702$
$R^2 = 0.062$
upstream 120 region
Coding region (2nd Codon Position)
$y = 0.091x + 0.698$

Figure 5

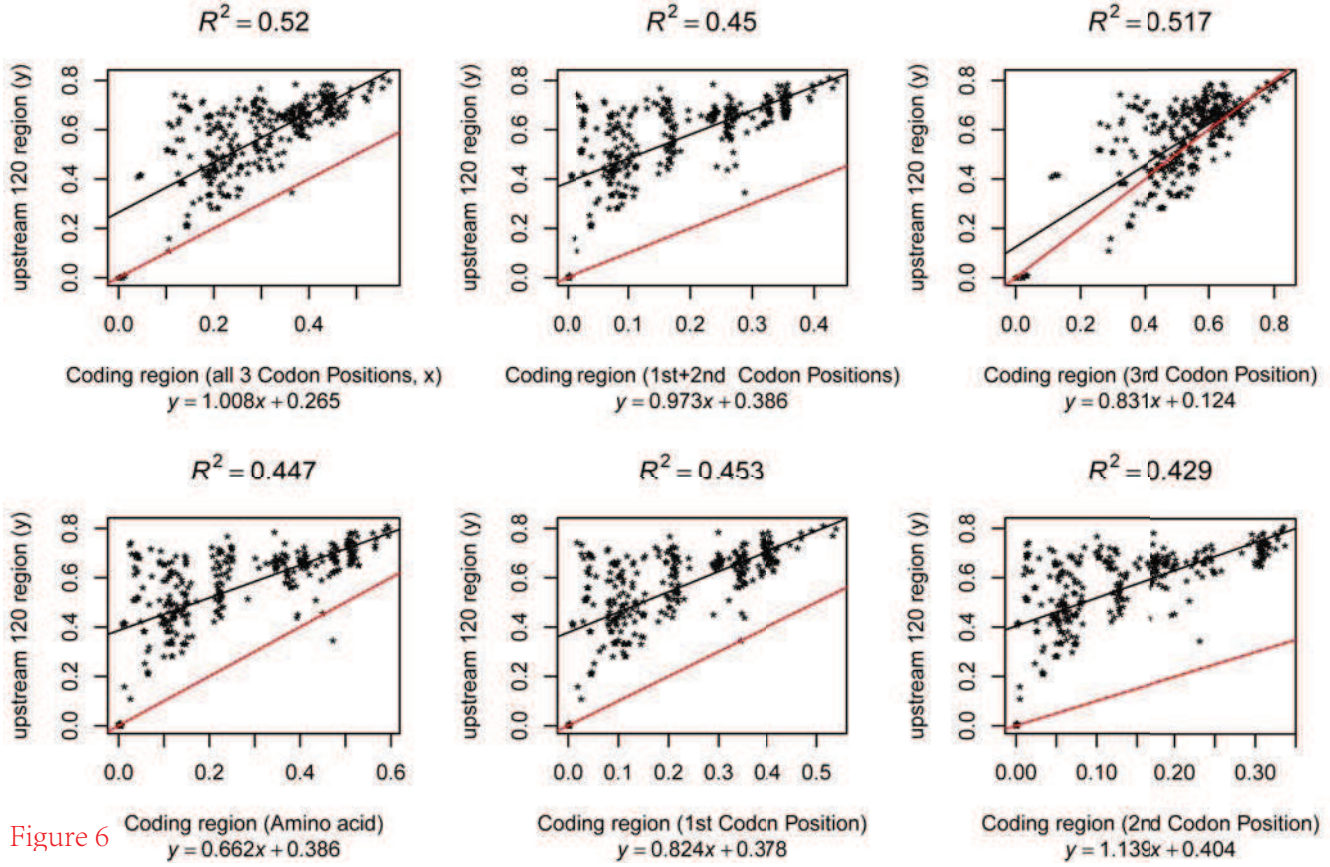

$R^2 = 0.52$
upstream 120 region (y)
Coding region (all 3 Codon Positions, x)
$y = 1.008x + 0.265$
$R^2 = 0.45$
upstream 120 region (y)
Coding region (1st+2nd Codon Positions)
$y = 0.973x + 0.386$
$R^2 = 0.517$
upstream 120 region (y)
Coding region (3rd Codon Position)
$y = 0.831x + 0.124$
$R^2 = 0.447$
upstream 120 region (y)
Coding region (Amino acid)
$y = 0.662x + 0.386$
$R^2 = 0.453$
upstream 120 region (y)
Coding region (1st Codcn Position)
$y = 0.824x + 0.378$
$R^2 = 0.429$
upstream 120 region (y)
Coding region (2nd Codon Position)
$y = 1.139x + 0.404$

Figure 6

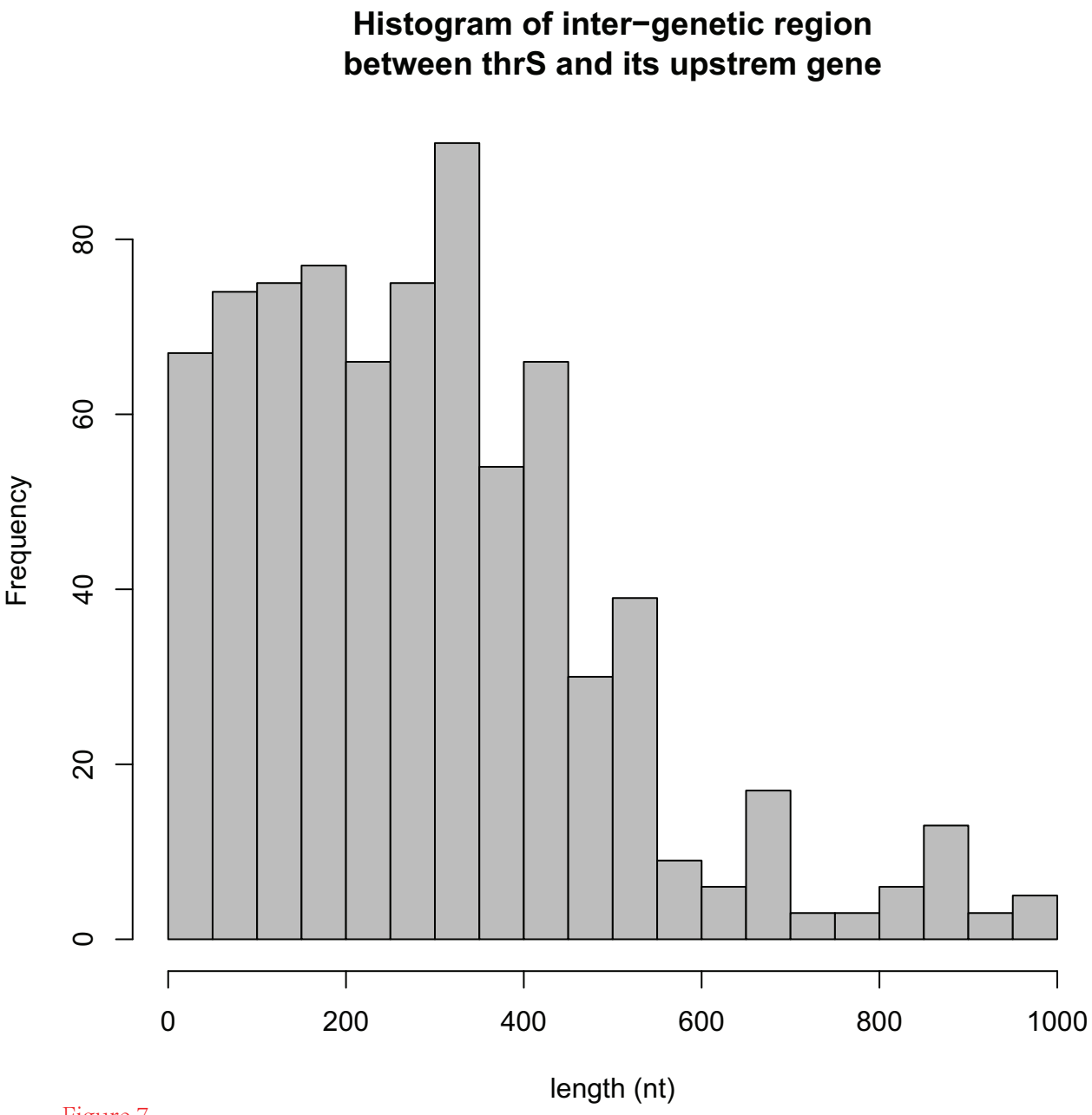

Histogram of inter−genetic region
between thrS and its upstrem gene
Frequency
0
20
40
60
80
0
200
400
600
800
1000
length (nt)

Figure 7

3. DISCUSSION

Now we are in a genomic era, to deal with so many genomes, we need some guiding concept to organize the messy raw data, then analyze, do inference and extract the useful information to form knowledge at the end. Inspired from the concept of neighborhood described in “the Delphic Boat”[173] and Indigo[75], this thesis focused on several kinds of neighborhoods deduced from genomes, specifically neighbors of the same class along the chromosome (Poster 1, Article 2 for genomic islands), neighbors of points generated by statistical techniques (Article 1 for points in Correspondence Analysis (CA) clouds), and neighbors of different classes along the chromosome (Article 3 for the upstream and coding regions of thrS).

3.1 Neighbors defined in the statistical space, in the case of CA clouds

Correspondence Analysis is a useful statistical technique to reduce multi-dimensional data to low dimensional space with the main relationship captured. That is to say the dimensions kept by CA are much more informative. Usually the inertia kept by one dimension is 3 times more than the next one (data not shown), this posed a question as to how many dimensions should be kept for the space generated by CA. There could be several criteria: (1) the first 3 dimensions, (2) the first dimension whose inertia is 10 times more than that kept by the next one, (3) the last dimension to which the inertia summed from the first dimension is larger than 80% of all inertia, (4) the last dimension whose inertia is larger than the average one. Besides these four criteria, there are still many others. Obviously, criterion (1) is an arbitrary one without any statistical support, according to which much information might be lost in further analysis. For example, the

first three dimensions are ranged from 41% to 43% for *E. coli*, *B. subtilis* and *M. jannaschii* data [78]. Criterion (2) has introduced some statistical meaning, but is still arbitrary. Criterion (3) is somewhat statistical meaningful, but it is not practical, since according to the above mentioned three data, at least 10 dimensions should be kept in further analysis [78]. And criterion (4) is still an arbitrary one but bearing some statistical meaning and is practical. According to criterion (4), 6 dimensions were kept for *P. ingrahamii*.

Although CA facilitates the visualization through reducing the multi-dimensional data, it is still difficult to draw the boundaries from one cluster of points to another just by viewing. Clustering techniques, such as hierarchical clustering and K-means clustering, can help to resolve this problem. The assignment of each point to a specific cluster could be approached through dynamic programming, but how many clusters are suitable for the studied data is unresolved, and usually this is manually selected just by viewing. Also another aspect associated to the member of each cluster is to what degree the member is truly in the assigned cluster. Upon reflexion the members close to the boundaries might have different features from those close to the center. These two problems are partially resolved in this thesis by introducing Model Based Clustering (MBC) through partitioning those points in CA cloud into different clusters. Usually the partitioning with highest BIC value is the best resolution. The members are assigned to all clusters generated by MBC with different possibilities, and the classification table is determined by the members and its highest possible cluster.

3.2 Neighbors defined along the chromosome, in the case of genomic islands

As described above, genomic islands show different features and can be distinguished from their neighbors (the core genome). Based on these features, different experiments can be designed to identify genomic islands in new genome. The limits of this method are that it is difficult to include the whole genome in one experiment, and it is impossible to carry out all the experiments in every genome. Benefitting from the genome sequencing, whole genome about one species/strain can be easily retrieved from the public database, and the prediction of a whole set of genomic islands became possible. *In silico*, the first consideration is the composition difference by which Genomic Islands differ from the core genome, and the second consideration is the common GIs features, such as tRNA, IS and direct repeats. But both cannot detect a GI when the donor and the recipient genomes are similar, or the age of the HGT event is reasonably old and some GI have no common criteria. To fill in this gap, the conservation of synteny groups of orthologous genes between related bacteria should be considered. The three referred aspects above can only detect partial set of Genomic Islands in a genome, while our new method combines all the three aspects, and extends the definition of Genomic Islands to all regions which differ in related species.

3.3 Neighbors defined along the chromosome, in the case of upstream and coding region of thrS

With reference of the translational initiation codon (ATG) in most Bacteria, the upstream and

downstream (coding) regions are a pair of closely related neighbors on the chromosome. To see how these two neighbors behaved in the perspective of evolution, we intended to collect all the thrS genes from all the sequenced Bacteria genomes. Here, one thing should be referred is that to obtain the clean data is somewhat difficult in current public database. An example is annotation of thrS gene among several *Salmonella* strains. From the alignments (Fig. 3.1), it is clear that annotators treated parts of the upstream sequence as the coding sequences in arizonae strain, and that led to strain which missed the first ACSL in the further analysis, obviously that is not the truth after comparing those sequences from the closely related species and strains.

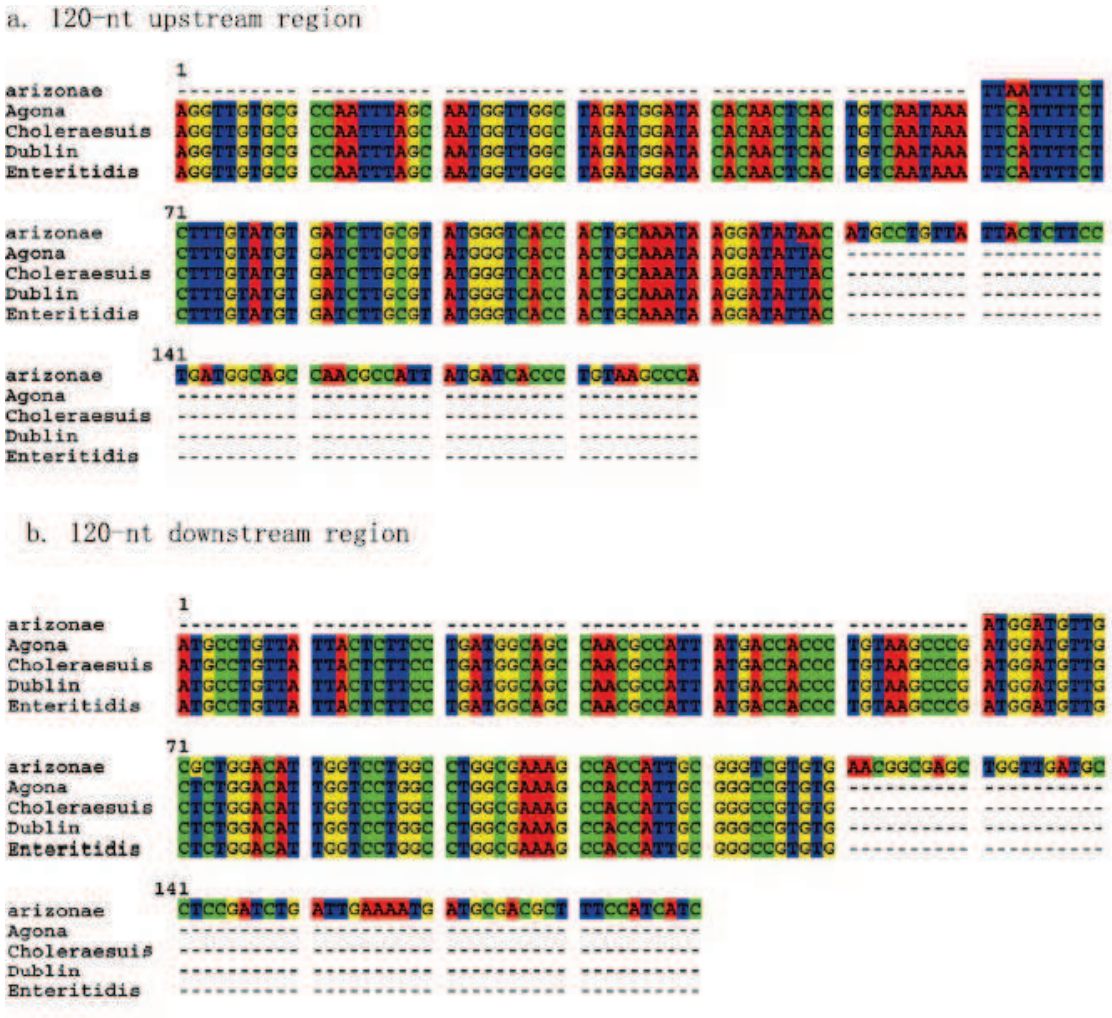


Figure 3.1 Annotations in 5 genomes of *Salmonella* strains

a. The 120 nucleotide sequences upstream of the annotated translational initiation codon "ATG";

b. The 120 nucleotide sequences downstream of the annotated translational initiation codon

"ATG".

Another thing should be emphasized again, the translational regulator in a genome can only be determined through experiments or when the associated ACSLs are closely related to those documented in E. coli with evidence of compensational mutations. While in other genomes, through the ACSLs and the flanking sequences, we can predict how likely the related strains are regulated at the translational level, and propose those most possible ones for experiment validation. When there is no other more significant secondary structure resulting in other kinds of regulation mechanism, for example the transcriptional terminator appeared in *B. subtilis*, the more stable the ACSL is, the closer to the translation initiation codon the ACSL is, the more possible the related strain is regulated at the translational level.

The evaluation of the unique upstream sequences, unique ACSL sequences, and both unique nucleotide sequences and amino acid sequences of coding regions reveals different evolutionary pressure on the two neighbors, i.e. the upstream region and coding region. For details see the discussion part in Article 3.

4. CONCLUSION

Faced with the messy genome text, different methods can be employed to decode the background information, and in this thesis, we proposed that the concept of neighborhood can help the researchers organize the raw genome and its corresponding annotation through relational database, extract useful information from the genome and its annotations, such as partitioning genes to clusters through CA and MBC, the prediction of genomic islands through different ways, evolutionary behavior of regulators through the comparison with the corresponding coding region.

In Article 1, we have included 6 dimensions each associating with inertia higher than the average one from the CA result so as to consider both avoiding the loss of information and simplifying the visualization. And in the next clustering analysis, we have introduced MBC to better understand the formed clusters.

In Poster 1 and Article 2, we have developed a new method that takes account for composition, GIs' features and synteny break among the comparing genomes to find genomic islands from a sequenced genome. When applying this new method in both *E. coli* and *B. subtilis*, we have identified some potential regions not described in previous literature.

In Article 3, after careful dealing with those upstream and coding regions of thrS gene, we found that the evolutionary behavior of the upstream regions was different from that of coding regions. In the closely related strains/species, when the upstream region bears a non-essential function, generally it evolved more slowly than the coding region (an example is those 30 strains in *E. coli-Shigella* group), but sometimes the jumping event happened in this region (an example is described in *E. fergusonii*, *E. coli* and *Samonella*).

5. PROSPECTS

Since there are too many kinds of neighborhoods related to genome, this thesis has just selected several points and shown that the concept of neighborhood can be used in this field, and they reveal some interesting information. The following studies will be carried out or are already ongoing:

1) Organize genome data with the consideration of the concept of persistence proposed by Gang and Danchin [48].

2) Neighbors in the perspective of evolution (in progress Article for *ppk* (pyrophosphate kinase) and its co-evolved genes).

3) Correspondence Analysis with the primitive frequency table calculated from the dipeptides and gapped dipeptides.

4) Incorporated Correspondence Analysis to the relational database genome to facilitate the ordinary biological researcher without much background in the field of statistics.

6. APPENDIX

6.1 The implementation of the computation of CA

Here an application of CA in proteomics is used to illustrate the general process. The input for CA is the number of occurrences of 20 amino acids in each protein. Take each amino acid as a column tag, and each protein as a row tag, and then a raw data matrix called the primitive matrix (N (I, J)) is formed. Each element ((n_{ij}) in the table represents the occurrence of amino acid *j* in the protein *i* (Table 6.1).

Table 6.1 The primitive matrix of Correspondence analysis

i/j	1	…	*j*	…	J
1	n_{11}		n_{1j}		n_{1J}
…		…			
i	n_{i1}		n_{ij}		n_{iJ}
…					
I	n_{I1}		n_{Ij}		n_{IJ}

where J=1…J, on behalf of 20 amino acids; I=1…I, on behalf of the proteins used in CA

From the primitive table, the row total can be calculated for each row in Table 1 (here this row total stands for the length of protein adopted in CA, if there is no modifications to the original protein). For the ith row (protein i), its total (n_{i+}) can be calculated by **Formula 1**. Similarly, the column total can also be calculated for each column in Table 1 (here this is the number of amino acids which occurred in the whole proteome). For the jth column (amino acid j), its total (n_{+j}) can be calculated by **Formula 2**. And the grand total, which is the sum of the whole table and is indicated by symbol n_{++}, can be calculated by **Formula 3**:

$$n_{i+} = \sum_{j=1}^{J} n_{ij} \qquad \textbf{(Formula 1)}$$

$$n_{+j} = \sum_{i=1}^{I} n_{ij} \qquad \textbf{(Formula 2)}$$

$$n_{++} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \qquad \textbf{(Formula 3)}.$$

Profile for each row (row profile) is a conditional density vector, and the ith row profile is calculated by **Formula 4**, and all row profiles constitute an I X J matrix R called row profile matrix (Table 6.2). Similarly, each column has its own profile, and the profile for jth column is calculated by **Formula 5**, and all column profiles constitute a J X I matrix C (Table 6.3). The concept of profile is a very important idea in CA, and row profiles and column profiles can be treated as independent points in the J-dimensional and I-dimensional spaces defined by their corresponding density vectors respectively. Since these profiles are seen as the points, the distances can be calculated between any two points. To calculate these distances, CA adopts a special way, which will be discussed later.

$r_i = (r_{i1}, r_{i2}, \ldots, r_{ij}, \ldots, r_{iJ}) = (n_{i1}/n_{i+}, n_{i2}/n_{i+}, \ldots, n_{ij}/n_{i+}, \ldots, n_{iJ}/n_{i+})$ (j=1, 2… J) **(Formula 4)**

$c_j = (c_{1j}, c_{2j}, \ldots, c_{ij}, \ldots, c_{Ij}) = (n_{1j}/n_{+j}, n_{2j}/n_{+j}, \ldots, n_{ij}/n_{+j}, \ldots, n_{Ij}/n_{+j})$ (i=1, 2… I) **(Formula 5)**

Table 6.2 the matrix of row profile

Rows	Columns			Total
	1	j	J	
1.	n_{11}/n_{1+}	n_{1j}/n_{1+}	n_{1J}/n_{1+}	1
.	.	.	.	.
I	n_{i1}/n_{i+}	n_{ij}/n_{i+}	n_{iJ}/n_{i+}	1
.	.			.
I	n_{I1}/n_{I+}	n_{Ij}/n_{I+}	n_{IJ}/n_{I+}	1

Table 6.3 the matrix of column profile

Columns	Rows			Total
	1	*i*	*I*	
1.	n_{11}/n_{+1}	n_{i1}/n_{+1}	n_{I1}/n_{+1}	1
.	.	.	.	.
J	n_{1j}/n_{+j}	n_{ij}/n_{+j}	n_{Ij}/n_{+j}	1
.				.
J	n_{1J}/n_{+J}	n_{iJ}/n_{+J}	n_{IJ}/n_{+J}	1

Another basic concept adopted in CA is the mass. The mass for the ith row ($m_{i\cdot}$) is the ratio between the row total (n_{i+}) and the grand total (n_{++}) (**Formula 6**), and the row mass vector for table 6.1, i.e., the primitive matrix, is illustrated in **Formula 7**. Similarly, the mass for jth column ($m_{\cdot j}$) is the ratio between the column total (n_{+j}) and the grand total (n_{++}) (**Formula 8**), and the column mass vector for table 6.1 is illustrated in **Formula 9**.

$m_{i\cdot}=n_{i+}/n_{++}$ (**Formula 6**)

$m_r=(m_{1\cdot}, m_{2\cdot}, \dots, m_{i\cdot}, \dots, m_{I\cdot})$ (**Formula 7**)

$m_{\cdot j}=n_{+j}/n_{++}$ (**Formula 8**)

$m_c=(m_{\cdot 1}, m_{\cdot 2}, \dots, m_{\cdot j}, \dots, m_{\cdot J})$ (**Formula 9**)

The average row profile is the weighted average of all row profiles (**Formula 10**).

$$\bar{r} = \sum_{i=1}^{I} \left(w_i r_i \right) \qquad \textbf{(Formula 10)}$$

Here the weight used for row is the corresponding mass of each row, i.e., the weight for the ith row (w_i) is the mass for ith row ($m_{i\cdot}$), and the average row profile can be calculated as illustrated in **Formula 11**. According to the result from **Formula 11**, it is easy to conclude that the average row profile is equal to the column mass vector (m_c). Similarly, from **Formula 12**, it is easy to obtain the average column profile, which is equal to the row mass vector (m_r).

$$\bar{r} = \sum_{i=1}^{I}(w_i r_i) = \sum_{i=1}^{I}(m_{i\bullet} r_i) = \sum_{i=1}^{I}\left(\left(\frac{n_{i+}}{n_{++}}\right)\left(\frac{n_{i1}}{n_{+j}}, \frac{n_{i2}}{n_{i+}}, \cdots, \frac{n_{ij}}{n_{i+}}, \cdots, \frac{n_{iJ}}{n_{i+}}\right)\right)$$

$$= \sum_{i=1}^{I}\left(\frac{n_{i+}}{n_{++}} * \frac{n_{i1}}{n_{i+}}, \frac{n_{i+}}{n_{++}} * \frac{n_{i2}}{n_{i+}}, \cdots, \frac{n_{i+}}{n_{++}} * \frac{n_{ij}}{n_{i+}}, \cdots, \frac{n_{i+}}{n_{++}} * \frac{n_{iJ}}{n_{i+}}\right)$$

$$= \sum_{i=1}^{I}\left(\frac{n_{i1}}{n_{++}}, \frac{n_{i2}}{n_{++}}, \cdots, \frac{n_{ij}}{n_{++}}, \cdots, \frac{n_{iJ}}{n_{++}}\right) = \left(\sum_{i=1}^{I}\frac{n_{i1}}{n_{++}}, \sum_{i=1}^{I}\frac{n_{i2}}{n_{++}}, \cdots, \sum_{i=1}^{I}\frac{n_{ij}}{n_{++}}, \cdots, \sum_{i=1}^{I}\frac{n_{iJ}}{n_{++}}\right) \quad \textbf{(Formula}$$

$$= \left(\frac{\sum_{i=1}^{I} n_{i1}}{n_{++}}, \frac{\sum_{i=1}^{I} n_{i2}}{n_{++}}, \cdots, \frac{\sum_{i=1}^{I} n_{ij}}{n_{++}}, \cdots, \frac{\sum_{i=1}^{I} n_{iJ}}{n_{++}}\right) = \left(\frac{n_{+1}}{n_{++}}, \frac{n_{+2}}{n_{++}}, \cdots, \frac{n_{+j}}{n_{++}}, \cdots, \frac{n_{+J}}{n_{++}}\right)$$

$$= \left(m_{\bullet 1}, m_{\bullet 2}, \cdots, m_{\bullet j}, \cdots m_{\bullet J}\right) = m_c$$

11)

$$\bar{c} = \sum_{j=1}^{J}(w_j r_j) = \sum_{j=1}^{J}(m_{\bullet j} c_j) = \sum_{j=1}^{J}\left(\left(\frac{n_{+j}}{n_{++}}\right)\left(\frac{n_{1j}}{n_{+j}}, \frac{n_{2j}}{n_{+j}}, \cdots, \frac{n_{ij}}{n_{+j}}, \cdots, \frac{n_{Ij}}{n_{+j}}\right)\right)$$

$$= \sum_{j=1}^{J}\left(\frac{n_{+j}}{n_{++}} * \frac{n_{1j}}{n_{+j}}, \frac{n_{+j}}{n_{++}} * \frac{n_{2j}}{n_{+j}}, \cdots, \frac{n_{+j}}{n_{++}} * \frac{n_{ij}}{n_{+j}}, \cdots, \frac{n_{+j}}{n_{++}} * \frac{n_{Ij}}{n_{+j}}\right)$$

$$= \sum_{j=1}^{J}\left(\frac{n_{1j}}{n_{++}}, \frac{n_{2j}}{n_{++}}, \cdots, \frac{n_{ij}}{n_{++}}, \cdots, \frac{n_{Ij}}{n_{++}}\right) = \left(\sum_{j=1}^{J}\frac{n_{1j}}{n_{++}}, \sum_{j=1}^{J}\frac{n_{2j}}{n_{++}}, \cdots, \sum_{j=1}^{J}\frac{n_{ij}}{n_{++}}, \cdots, \sum_{j=1}^{J}\frac{n_{Ij}}{n_{++}}\right) \quad \textbf{(Formula}$$

$$= \left(\frac{\sum_{j=1}^{J} n_{1j}}{n_{++}}, \frac{\sum_{j=1}^{J} n_{2j}}{n_{++}}, \cdots, \frac{\sum_{j=1}^{J} n_{ij}}{n_{++}}, \cdots, \frac{\sum_{j=1}^{J} n_{Ij}}{n_{++}}\right) = \left(\frac{n_{1+}}{n_{++}}, \frac{n_{2+}}{n_{++}}, \cdots, \frac{n_{i+}}{n_{++}}, \cdots, \frac{n_{I+}}{n_{++}}\right)$$

$$= \left(m_{1\bullet}, m_{2\bullet}, \cdots, m_{i\bullet}, \cdots m_{I\bullet}\right) = m_r$$

12)

The corresponding matrix P is defined as the primitive matrix divided by the grand total (n_{++}) (**Formula 13**), then the values for the entities in corresponding matrix P are the relative frequencies calculated by the corresponding true occurrences divided by the grand total (**Formula 14**). Correspondence matrix illustrates how the mass is distributed among the table. And the sum of each row (column) in the correspondence matrix is the corresponding row mass (column mass). The row profile matrix can also be written as the ratio of the correspondence

matrix and the row mass (**Formula 15**), while the column profile matrix can also be written as the ratio of the corresponding matrix P and the column mass (**Formula 16**).

$$P = \frac{1}{n_{++}} \bullet N \qquad \textbf{(Formula 13)}$$

$p_{ij}=n_{ij}/n_{++}$ **(Formula 14)**

Matrix of row profile $R=D_r^{-1}P$ **(Formula 15)**

where Dr is the diagonal matrix with row mass at the diagonal line.

Matrix of column profile $C=D_c^{-1}P$ **(Formula 16)**

where Dc is the diagonal matrix with column mass at the diagonal line.

Distances between profile points are calculated through the weighted Euclidean distance. For example, for two row profile points i and i', their profiles r_i and $r_{i'}$ are represented by the vectors $\left(n_{i1}/n_{i+}, n_{i2}/n_{i+}, n_{i3}/n_{i+}, \cdots \right)$ and $\left(n_{i'1}/n_{i'+}, n_{i'2}/n_{i'+}, n_{i'3}/n_{i'+} \cdots \right)$ respectively, and their distance, which is called weighted Chi-square distance, is calculated by **Formula 17**. Similarly, the distance between two column profile points j and j' is calculated from **Formula 18**.

$$d(i,i') = \sqrt{w(r_i - r_{i'})^2} = \sqrt{\sum_{j=1}^{J} \frac{1}{n_{+j}} \left(\frac{n_{ij}}{n_{i+}} - \frac{n_{i'j}}{n_{i'+}} \right)^2} \qquad \textbf{(Formula 17)}$$

where $1/n_{+j}$ is the weight for the j^{th} item of the row profile point

$$d(j,j') = \sqrt{\sum_{i=1}^{I} \frac{1}{n_{i+}} \left(\frac{n_{ij}}{n_{+j}} - \frac{n_{ij'}}{n_{+j'}} \right)^2} \qquad \textbf{(Formula 18)}$$

where $1/n_{+i}$ is the weight for the i^{th} item of the column profile point

In physics, for each object, there is a center of gravity (or centroid). Any part of the object has its own partial mass, and also the moment of inertia related to the center of gravity, which is the product of the mass and the square of the distance to the centroid (**Formula 19**). And the

moment of inertia for the whole body is the sum of the moment of inertia from all the components (**Formula 20**).

Moment of inertia for the particle of the object=md^2 （**Formula 19**）

Moment of inertia for the object=$\sum md^2$ （**Formula 20**）

Inertia is another statistic used in CA, which has a similar meaning to the "moment of inertia" in physics [174]. In row profile space, the point for the average row profile acts as the center of gravity in physics, and each row profile point has its own mass ($m_{i\bullet}$) and its distance, which is represented by the weighted Chi-square distance, to the average row profile. One can calculate the inertia for each row profile point in the similar way of calculating the moment of inertia for objects in physics (what should be noted is: In CA, inertia has the same concept as moment of inertia in physics). And the inertia for i^{th} row is calculated from **Formula 21**. Similarly, in the column profile space, the inertia attached to each column profile point can be obtained from **Formula 22**.

Inertia of i^{th} row profile = $$m_{i\bullet}\sum_{j=1}^{J}\left(\frac{r_{ij}-\bar{r}_j}{\sqrt{\bar{r}_j}}\right)^2$$ (**Formula 21**)

where r_{ij} is equal to n_{ij}/n_{i+}, while $\bar{r}_j$ is equal to n_{+j}/n_{++}.

Inertia of j^{th} column profile= $$m_{\bullet j}\sum_{i=1}^{I}\left(\frac{c_{ij}-\bar{c}_i}{\sqrt{\bar{c}_i}}\right)^2$$ (**Formula 22**)

where c_{ij} is n_{ij}/n_{+j}, and $\bar{c}_i$ is n_{i+}/n_{++}.

And each entity from the primitive matrix has its own contribution to the total inertia (**Formula 23**). If the rows and columns are completely independent of each other, or say amino acids are randomly distributed in a protein, the expected value for each item in the primitive

matrix, which is the expected frequency of amino acid j in protein i, can be calculated from the row total and column total respectively. The differences between actual and expected frequencies contribute to the total inertia, and these differences constitute a matrix A, with elements calculated by **Formula 24**.

Inertia of the cell= $\left(\frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} \right)^2$ (**Formula 23**)

$$a_{ij} = \left| \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} \right| \quad \text{(Formula 24)}$$

The next task of CA is the singular value decomposition (SVD) of matrix A, which is defined as the matrix A (IXJ) decomposed into three matrices U, Γ, and V in the form of **Formula 25**, where matrix Γ is a diagonal matrix with the diagonal members arranged in a decreasing order from up-left to bottom-right and greater than 0 (**Formula 26**), and matrix U and V are both orthogonal matrices (**Formula 27**). Γ defines a transformed space for both row profile points and column profile points, and reserves most of their information (or inertia in the case for CA). The coordinates of row profile points are contained in the following matrix: F=U Γ , and the coordinates of column profile points are contained in the following matrix of G=V Γ . γ_1, γ_2, … , γ_k are called the singular values, which correspond to eigenvectors representing those mutually independent factors (axes). The square of the singular value reflects the total inertia distributed along various factors. Normally several vectors corresponding to the largest singular values are selected as major factors for the whole data set.

$A=U \Gamma V^T$ (**Formula 25**)

$\gamma_1 \geq \gamma_2 \geq \cdots \gamma_k > 0$ (**Formula 26**)

where k is the order of matrix A

$U^TU=I$, $V^TV=I$ (**Formula 27**)

where U^T is the transposed matrix of U, and V^T is the transposed matrix of V.

In Summary:

The primitive matrix N (I X J) is the input data, the element (i.e. n_{ij}) here is the number of occurrences of the amino acid i in the protein j.

The row mass is n_{i+}/n_{++}, and the column mass is n_{+j}/n_{++}.

The row profile is a vector with J dimensions composed by n_{ij}/n_{i+}; the average row profile is the vector composed by n_{ij}/n_{i+}, which is equal to the mass for column j. The column profile is a vector with I dimensions composed by n_{ij}/n_{+j}, and the column average profile is the vector composed by n_{ij}/n_{+j}, which is equal to the mass for row i.

The elements for corresponding matrix P are the ratios between the original matrix elements and the grand total, e.g., $p_{ij}=n_{ij}/n_{++}$.

A new matrix A is the centralization of corresponding matrix P, e.g., $A=P-rc^T$, then the task is to find the principal axes, which retain most of the inertia for matrix A, through SVD: $A=U\,\Gamma\,V^T$.

Matrix UΓ holds the coordinates for the projection of row profiles on those principal axes defined by Γ, while matrix VΓ holds the coordinates for the projection of column profiles on those principal axes defined by Γ.

The squares of singular values correspond to the inertia reserved along the corresponding axis.

6.2 Formula about Model Based Clustering (MBC)

For more than two-dimensions, the multivariate normal density for random variable $\mathbf{X}^T = [X_1, X_2, \ldots X_p]$ is defined as **Formula 28** and can be abbreviated as $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ in p-dimensions with a

mean μ and a variance-covariance matrix Σ. The maximum likelihood estimators (mle) of μ and Σ are calculated by **Formula 29** and **Formula 30** respectively. The observed values (x and $\overline{x}$) can be obtained or derived from a dataset that follows a multivariate normal density. The shape and orientation of the contour of the multivariate normal distribution can be determined by the form of the variance-covariance matrixΣ.

$$f(\mathbf{x}) = (2\pi)^{-p/2} |\Sigma|^{-½} \exp[- ½ (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)] \quad \textbf{(Formula 28)}$$

$$\hat{\mu} = \overline{X} \quad \textbf{(Formula 29)}$$

$$\hat{\Sigma} = \frac{1}{n} \Sigma(X_j - \overline{X})(X_j - \overline{X})^T \quad \textbf{(Formula 30)}$$

where $-\infty < x_i < \infty$, i = 1,2,…,*p*.

The covariance matrix Σ can be expressed in terms of its eigenvalue decomposition [175], in the form of λDAD^T, where D is the orthogonal matrix of eigenvectors of Σ, A is the diagonal matrix with the normalized eigenvalues of Σ on the diagonal in decreasing order and |A| = 1. λ is a scalar [176]. D determines the orientation, and A determines the shape of the density contours, while λ determines the volume of the corresponding ellipsoid. Characteristics (orientation, volume and shape) of distributions are usually estimated from the data, and can be allowed to vary between clusters, or are constrained to be the same for all clusters [176-178]. This parameterization includes but is not restricted to well-known models such as equal-volume spherical variance ($\Sigma_k = \lambda I$) [179], constant variance[180], and unconstrained variance[181]. We will encounter 10 parameterizations of the covariance matrix, which are listed in Table 6.4 [182].

Table 6.4 Geometric characteristics of 10 covariance matrix

identifier	Model[a]	volume	shape	Orientation
λI	EII	equal	equal	NA
$\lambda_k I$	VII	variable	equal	NA
λA	EEI	equal	equal	coordinate axes
$\lambda_k A$	VEI	variable	equal	coordinate
λA_k	EVI	equal	variable	coordinate axes
$\lambda_k A_k$	VVI	variable	variable	coordinate axes
λDAD^T	EEE	equal	equal	equal
$\lambda D_k A D_k^T$	EEV	equal	equal	variable
$\lambda_k D_k A D_k^T$	VEV	variable	equal	variable
$\lambda_k D_k A_k D_k^T$	VVV	variable	variable	variable

the model column is the abbreviation of the description of the geometric characteristics of the model, for example, VEI denotes a model in which the volumes of clusters may vary (V), the shapes of the clusters are equal (E), and the orientations are identical (I).

Bayesian statistics, credited to Thomas Bayes, treats the data as known and the parameters (θ) of a population as random variables. In Bayesian strategy, θ is treated as a distribution of possible values, while the observed data provide some information on that distribution. Suppose $\mathbf{x}^T = (x_1,\ldots,x_n)$ is a vector of n observations with a probability distribution $p(\mathbf{x})$, and these observations come from the population with a probability distribution of $p(\boldsymbol{\theta})$, then the joint probability distribution $p(\mathbf{x},\boldsymbol{\theta})$ can be obtained in two ways: 1) $p(\mathbf{x},\boldsymbol{\theta})= p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ where $p(\mathbf{x}|\boldsymbol{\theta})$ is the conditional probability distribution of observations when the population probability distribution is $p(\boldsymbol{\theta})$, and 2) $p(\mathbf{x},\boldsymbol{\theta})= p(\boldsymbol{\theta}|\mathbf{x})p(\mathbf{x})$ where $p(\boldsymbol{\theta}|\mathbf{x})$ is the conditional probability distribution of population when the observation has the probability distribution of $p(\mathbf{x})$. So it is easy to conclude $p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\mathbf{x},\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{x})p(\mathbf{x})$.

Given the observed data x, the conditional distribution of $\boldsymbol{\theta}$ is $p(\boldsymbol{\theta}|x) = \frac{p(\mathbf{x}\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})} =$ $kp(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ where k is a normalizing constant that ensures $p(\boldsymbol{\theta}|x)$ integrates to 1. This is the Bayes' Theorem [183]. $p(\boldsymbol{\theta})$ is defined as the *prior* distribution of $\boldsymbol{\theta}$ and conveys what is known about $\boldsymbol{\theta}$ without knowledge of the data. $p(\boldsymbol{\theta}|\mathbf{x})$ is the *posterior* distribution of $\boldsymbol{\theta}$ given $\mathbf{x}$. Given data $\mathbf{x}$, $p(\mathbf{x}|\boldsymbol{\theta})$ can be regarded as a function of $\boldsymbol{\theta}$. In this regard, $p(\mathbf{x}|\boldsymbol{\theta})$ is called the *likelihood function* of $\boldsymbol{\theta}|\mathbf{x}$ and is designated as $l(\boldsymbol{\theta}|\mathbf{x})$. Bayes' Theorem states the probability distribution for $\boldsymbol{\theta}$ posterior to the data $\mathbf{x}$ is proportional to the product of the distribution for $\boldsymbol{\theta}$ prior to the data.

In general, there are three kinds of cluster analysis algorithms [184], namely: those based on an attempt to find the optimal partitioning for a given number of clusters; those based on a hierarchical attempt to discover cluster structure; and those based on a probabilistic model for the underlying clusters.

Describing data in terms of its underlying distribution (or density) function is a descriptive strategy. The probability density function is described in **Formula 31**. Since we are dealing with multivariate normal distributions, the component distribution, $f_k(x|\theta_k)$ in **Formula 31**, could be represented by $f_k(x|\mu_k,\Sigma_k)$. The clustering task is to take a set of observations and a pre-determined number of clusters and then work out each cluster's mean and variance. The basic approach is: 1) Given a data set D, determine how many clusters G which can be fitted to the data; 2) Choose parametric models for each of the G clusters (i.e., multivariate normal distributions); 3) Use the expectation-maximization (EM) algorithm [185] to determine the component parameters θ_k and probabilities π_k from the data; 4) Assign data to a cluster by assigning each point to a specific cluster with the highest probability.

$$f(x|\Phi_G) = \sum_{k=1}^{G} \pi_k f_k(x|\theta_k) \qquad \textbf{(Formula 31)}$$

where f_k are the component distributions, π_k represents the mixing proportions ($\pi_k > 0$, $\Sigma\pi_k = 1$), and G is the number of components/clusters. Φ_G represents all the unknown parameters (π_k, θ_k)[186].

Given *n* observations ($x_1,\ldots,x_n$), the task is to maximize the log-likelihood to obtain the mle $\hat{\Phi}_G$ (**Formula 32**). The first step in the EM algorithm is the E (Expectation) step: calculate π_k, which is the conditional probability that an object belongs to cluster *k* given an initial start for component population parameters (θ). And later is an M (maximization) step: compute parameter estimates (θ) given the partitions determined in the E step. The E and M steps switch from one to the other until convergence. Eventually, any observation will be associated with a cluster that provides the highest probability for its membership. Although the EM algorithm is guaranteed to converge to a maximum, this is a local maximum and not necessarily the global maximum. The EM algorithm should be repeated several times with different initial inputs for the parameter values to overcome this local optimum.

$$\log L(\Phi_G) = \sum_{j=1}^{n} \log f(x_j|\Phi_G) \qquad \textbf{(Formula 32)}$$

After the clusters formed through the concept of mixture models, and the assignment of each observation to different clusters, the next step is to determine how many clusters are appropriate –there could be several viable cluster partitions for the same dataset. Several measures have proposed for choosing the clustering model (parameterization and number of clusters) [187]. Among them, one is to use the Bayesian Information Criterion (BIC) [188], which adds a penalty to the loglikelihood based on the number of parameters, and has performed well in a number of applications [189, 190]. The BIC can be calculated by **Formula 33.** A small

difference between BIC values with less than 2 means a weak difference between two models, and difference between 2 and 6 means a minor difference between two models, and a difference between 6 and 10 means a very strong difference between two models, and a difference larger than 10 means a very significant difference the between models [189]. MCLUST [182] software can be used to perform the MVN clustering analysis. Each cluster is represented by a Gaussian model (**Formula 34**). Clusters are centered at the means (μ_k). The covariances, $\sum_k$, determine other geometric features.

$$BIC = 2\log L(\hat{\Phi}_G) - \nu_G \log(n) \qquad \textbf{(Formula 33)}$$

where $\log L(\hat{\Phi}_G)$ is the maximized loglikelihood for the model and data, ν_G is the number of independent parameters to be estimated in Φ_G, and n is the number of observations in the data.

$$f_k(\mathbf{x}|\mu_k, \Sigma_k) = (2\pi)^{-p/2} |\Sigma_k|^{-1/2} \exp[-\tfrac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k)] \qquad \textbf{(Formula 34)}$$

where **x** represents the data, and *k* is an integer subscript specifying a particular cluster.

7. REFERENCES

Résumé

Avec l'accroissment du nombre de génomes séquencés, l'organisation de ces données brutes et des données dérivées, l'extraction de l'information et des connaissances associées défie l'imagination. La notion de voisinage a été d'abord été introduite pour l'organisation des données dans des bases de données relationnelles. Pour extraire des informations pertinentes à partir de données massives, différents types de voisinages ont été étudiés ici. Tout d'abord, avec l'analyse des correspondances (CA) et en utilisant le regroupement supervisé ("model clustering" MBC), la proximité mutuelle des éléments formant deux entités biologiques centrales, les gènes (codant les protéines) et les acides aminés a été analysée. Nous montrons par exemple que les protéines de *Psychromonas ingrahamii*, bactérie psychrophile extrêmes, sont regroupées en six classes, et qu'il y a une forte opposition entre le comportement de l'asparagine (N) et des acides aminés sensibles à l'oxygène, ce que nous expliquons en terms de résistance au froid. Ensuite, nous avons analysé la répartition entre les îlots génomiques (GI) et le squelette du génome de base à partir d'une nouvelle méthode combinant composition en bases et en gènes, caractéristiques GI et de briser les synténies. L'application de cette approche à *E. coli* et *B. subtilis* a révélé que cette nouvelle méthode permet d'extraire certaines régions significative, non publiées auparavant. Enfin, pour illustrer un voisinage fin, la régulation de l'expression d'un gène et son évolution, nous avons étudié la relation entre les régions en amont du gène et la zone codante du gène *thrS* de façon approfondie. Nous avons constaté que ces deux régions associées à un gène, se sont comportés différemment dans l'histoire évolutive. Certaines des régions en amont porteuses de la fonction non-essentielle de régulation (qui contrôle l'expression de gène) ont évolué différemment de la région codante.

Mots-clés: voisinage, l'analyse des correspondances (CA), regroupement supervisé (MBC), îlots génomiques (GI), règulation de l'expression génétique, threonyl-tRNA synthetase (thrRS)