Genomic patterns and phenotypic plasticity in prokaryotes analyzed within an ecological framework

Tesi Doctoral

Juan Antonio García

Centre d'Estudis Avançats de Blanes - CSIC

Departament
d'Ecologia Continental

Departament
de Genètica i Microbiologia

Genomic patterns and phenotypic plasticity in prokaryotes analyzed within an ecological framework

Juan Antonio García

Tesi Doctoral

Centre d'Estudis Avançats de Blanes
Departament d'Ecologia Continental

Universitat Autònoma de Barcelona
Departament de Genètica i Microbiologia

Genomic patterns and phenotypic plasticity in prokaryotes analyzed within an ecological framework

Juan Antonio García
2009

Director
Dr. Emilio Ortega Casamayor
Investigador titular CEAB-CSIC

Tutor
Dr. Jordi Mas Gordi
Professor titular UAB

A la meva mare i a la Carme

"Viure és provar-ho infinites vegades"
Màrius Sampere

Agraïments

Resulta obvi, malgrat això s'ha de dir que aquesta tesi és el resultat de la tasca d'un equip de persones que s'han guanyat el meu reconeixement i als quals voldria donar les gràcies públicament.

La petita història d'aquesta tesi té el seu origen en una carambola que es va donar fa uns cinc anys i que va permetre un primer encontre amb el que ha sigut el meu director, l'Emilio, al CEAB de Blanes. Des del primer moment em va demostrar total confiança, fet que ha sigut clau per a que aquest projecte finalment vegi la llum. No se m'han passat per alt la seva paciència, el seu bon humor sempitern, la seva manca de complexos i la seva habilitat per gestionar un equip tan divers com el que ha conjuntat. Tot plegat ha fet que la meva tasca fos molt més fàcil. El seu coratge de l'atreviment i entusiasme encomanadís ens va portar a sentir-nos una mica aventurers explorant el món de la genòmica i l'ecologia microbiana des de punts de vista insòlits pels biòlegs tradicionals.

La necessitat d'adquirir nous coneixements que ens permetessin orientar-nos dintre del camí fosc a on ens havíem ficat van portar-me directament a les fonts de coneixement. Així, vaig fer un primer salt a la Universitat Autònoma de Barcelona.

Allà hi ha un grup de recerca que em va tractar d'una manera exquisida durant la petita col·laboració que vam protagonitzar mentre feia els meus primer passos d'aquesta aventura. Són el grup de Biomatemàtica de Recerca de la Unitat de Bioestadística. Ells em van transmetre, d'una manera elegant, uns coneixements avançats de matemàtica i estadística pels quals senten veritable passió, i que van començar a enlairar la qualitat de la tesi. Xavier, Joan, David, Jesús, aquesta tesi també porta el vostre nom.

El segon salt va resultar ser el més llarg, Boston.

I am indebted to the members of the Center for Polymer Studies at Boston University for providing an exciting atmosphere of scientific discussion and amazing possibilities of research. I would specially like to express my gratitude to the director of the center, pioneer of interdisciplinary science and the quintessential bridge builder within sophisticated mathematics and important problems in the physical and life sciences, Professor Eugene Stanley (Gene). I would also like to thank my colleagues in the lab; particularly Plamen, for providing me with statistical physics advice at times of critical need, José and Alfonso who contributed with constructive critics and helped me to enrich the experience.

El següent destí va ser Braunschweig (Alemanya).

I am profoundly thankful to Dr. Vitor dos Santos for receiving me in the Environmental Microbiology Laboratory at the German Research Center for Biotechnology (GBF) and for introducing me into a new interdisciplinary study field, the systems biology. Thanks for your patience and for the continuing support and guidance provided to me. I would also like to gratefully acknowledge the people from GBF who kindly helped me during the six months that I have spent in Braunschweig and Wolfenbüttel. The detailed comments, suggestions and insight of Jacek and Amit have been of great value to me, as well as their provided valuable reviews. I could not have wished for better companions than Roberto and Pepe. They were a great support to me. I also greatly value the exchanges of knowledge with Dahay and Xinxing. Finally, I will be always indebted to the computer expert Miguel for his skills, his accurate and unselfish help and our discussions.

I cannot forget the European Network of Excellence Marine Genomics (MGE) and all the people involved in it. The network, which was devoted to the investigation of the biology marine organisms, gave me facilities to participate in bioinformatics, evolutionary and ecological training courses besides provide financial support for travels, conferences and meetings. I am particularly thankful to Frank Oliver and Marga for their hospitality at the Max Planck Institute for Marine Microbiology in Bremen; Jomuna, Alexander and the training team from Cebitec for the bioinformatics course in Bielefeld; Wiebe, Jeanine, Gabriele and Filip for

the great organization of the evolutionary course in the Stazione Zoologica of Naples; Ulrika for her excellent coordination of the network's events; Irene, Barbara, Maria Inês and Joana for their kindness and sharing their knowledge with me and finally, Francesco for his friendliness and hospitality.

El fantàstic equip del Wallex Life System va demostrar que un tot no és només la suma dels individus, i que els brainstormings no tenen secrets per a ells degut a la seva imaginació desbordant. Treballar durant les jornades doctorials de 2006 a Collbató amb un grup tan heterogeni (filòlegs, arqueòlegs, físics quàntics,...) em va enriquir d'una manera gairebé obligada i sobretot divertida ja que el bon humor va estar present durant totes les jornades. Alfons, Alicia, Rosa, Jaume, Eduard, Joonwoo, Sara i Alfredo, do you wallex?

Una de les conclusions que he tret d'aquesta tesi és que el Centre d'Estudis Avançats de Blanes (CEAB) es un lloc privilegiat tant per la seva ubicació com per la gent que li dóna vida cada dia. Voldria esmentar un grapat d'aquesta gent amb la qual he tingut el plaer d'interrelacionar-me durant anys i que han deixat la seva empremta, d'alguna manera, en aquest treball.

La primera persona que em va guiar els primers passos, al despatx i als laboratoris del CEAB, va ser la Isabel. Malgrat que vam coincidir breument a Blanes, la seva ajuda va ser bàsica per a situar-me i començar a endinsar-me al món dels microorganismes. El Jordi Catalan i el Fede van fer un primer tomb en el desenvolupament de la tesi posant sobre la taula els seus coneixements de física estadística i estadística (sense física) que ens van engrescar a ficar-nos a l'estudi dels procariotes per un camí fins aleshores inexplorat, el camí fosc del que parlava al principi. Així, ells van mostrar-nos una part del camí que ha recorregut aquesta tesi. Recordo que durant els primer mesos a Blanes sempre que vaig necessitar un seient lliure per viatjar de Blanes a Barcelona, i viceversa, disposava d'un als cotxes de la Paula i de l'Andrea. No vull oblidar-me del Juanma, el Guillermo, el Javier i la Patricia, que van arribar abans que jo al CEAB i que durant el seu terennar diari sempre trobaven temps per una ajuda o una paraula d'ànim, accions que es valoren quan s'arriba a un lloc nou. Un altre que ja hi era quan vaig arribar i que espero que hi sigui per molts anys pel bé de tota la gent del CEAB és el Ramón. No

només el recurs més valuós quan es tractava de posar a punt els ordinadors sinó també una persona sempre disposada a ajudar.

L'equip d'ecologia molecular microbiana ha sigut el nucli que m'envoltava i em feia suport durant bona part de la tesi. Voldria tornar a ressaltar l'excel·lent tasca de l'Emilio, aquesta vegada no com a director sinó com a un company més del grup. L'Antoni és d'aquelles persones que voldria tenir sempre al meu equip, per la seva autoconfiança (impossible is nothing), el seu altruisme, la seva curiositat i la seva capacitat de treball. L'Albert i el Jean Christophe han afegit una injecció de qualitat a aquesta tesi, sense la qual hauria quedat coixa. També he de destacar la col·laboració i la bona disposició que han tingut en tot moment la Bego, l'Anna, la Natalya, la Carmen i el Pierre.

Una persona indispensable per què aquesta tesi tingui sentit, que ens ha donat suport i que ha portat a terme una feina silenciosa des de la distància ha sigut el meu tutor a la Universitat Autònoma de Barcelona, el Jordi Mas.

Els responsables del Centre de Supercomputació de Catalunya ens han facilitat els seus ordinadors des de on hem pogut treballar amb els genomes i realitzar càlculs estadístics. A més, el tracte rebut ha sigut excel·lent i les incidències que hem tingut les han solucionades sempre amb eficàcia i solvència.

L'agraïment no seria complet sense la llista de protagonistes que tenen el mèrit d'haver tingut la paciència de suportar-me no només durant aquests últims cinc anys sinó des de ja fa uns quants lustres. En primer lloc, un grup que va fer del meu pas per la universitat una experiència fantàstica i amb el qual he compartit, des de llavors, un bon grapat d'estones agradables. Els bons sentiments que m'han transmès, com la generositat i l'atenció de la Gemma, els ànims i l'optimisme de l'Esther, l'empatia de l'Anna, l'energia i capacitat de lluita de la Mayra, l'esperit d'aventura de la Ruth, l'alegria de la Silvia o l'altruisme de la Sandra, estan reflectits en aquest treball.

My friend Marc is present in this thesis through his accurate English corrections and suggestions of some chapters. He and his kindness are often in my thoughts.

A la meva professora de matemàtiques a l'institut, l'Ana, li dec bona part de la capacitat d'entendre aquest llenguatge amb que està construït aquest treball. El seu compromís i la seva exigència van ser fonamentals per transmetre eficaçment els seus coneixements.

Els amics, els de sempre. Paco, Vicente, Francisco, Angel i Juan Antonio s'han guanyat a pols durant uns quants anys la seva presència aquí. M'heu ensenyat un camí i encara continuem endavant.

Finalment la Mar, els tiets, els cosins, el Fernando i la meva mare, la meva família. I l'última persona que mereix el meu reconeixement, at my most beautiful, la Carme, que m'ha suportat i m'ha entès i ha sigut sempre pacient amb mi durant més de cinc llargs anys. No sabria no estimar-te.

A tots vosaltres, ha sigut un honor arribar de la vostra mà fins aquí. Gràcies.

Contents

Contents

Summary

Summary

Genomic diversity of microorganisms is the result of the combined effects of past evolutionary roads and ecological events. Thus, the specific genome structure in a bacterium is consequence of the selective pressure by the interactions between microorganisms and environment along evolution. Therefore, DNA is predicted to contain more structural information than would be expected from nucleotide bases composition alone.

The general aim of this PhD thesis was to develop a theoretical framework based on genometric, statistic and mathematic modelling to study the relationship between genome structure, lifestyle and metabolism of prokaryotic microorganisms. To unveil the relationship between genome structure and lifestyle, a large set of genomes were analyzed by means of a statistical physics methodology which reduces the prokaryotic genomic complexity to a single parameter —the intrinsic long-range correlation that is related directly to the fractal structure of the DNA sequence— which can be further used for comparative genomics and ecological purposes.

DNA walk and Detrended Fluctuation Analysis (DFA) were the methods used for the study of long-range correlations in genomes. DNA walking is a genometric method based on a derivative function of the sequential position for each nucleotide along a DNA sequence. The resulting "walk" is representative of the DNA "landscape" and enables the simultaneous comparison among different genomes. DFA method provides a single quantitative parameter —the scaling exponent α— to represent correlation properties of a sequence. The sequential approach DNA walk–DFA was combined with a functional approach (distribution of clusters of orthologous genes, COG) showing that both, correlations and COG distribution in genomes may be originated by similar factors such as expansions and contractions in the genomic repertoire or adaptation to extreme habitats.

Relationships between lifestyle and metabolism were examined by means of a comprehensive comparative genomics study of two marine bacteria that exclusively use hydrocarbons as carbon and energy sources in different environmental scenarios, *Alcanivorax borkumensis* and *Oleispira antarctica*. The genomic bases of the unusual ecophysiological features of these microorganisms were studied to help for a better understanding of the influence of temperature on the oil-degrading based bacterial growth.

Finally, a functional genomics approach using mathematical modelling for the whole metabolism network codified in the genome of *Alcanivorax borkumensis* was carried out in order to look into the relationship between genome composition and metabolic phenotype. The whole set of genes, proteins, reactions and metabolites that participated in the metabolic activity were identified, categorized and interconnected to form a network through *in silico* metabolic reconstruction. This metabolic network allowed, by means of constraint-based methods and Flux Balance Analysis (FBA), to characterize the peculiar ecophysiologic features of this microorganism and to predict mutant cellular phenotypes. The modelling of carbon versus nitrogen fluxes allowed the discovery of conditions in which the excess carbon available in hydrocarbons was not directly translated into bacterial biomass but carbon overflow was diverted to the production of polyhydroxyalkanoates (bioplastics). The predictions showed a potential in the use of the model as a high-throughput analysis *in silico* tool for detailed studies on the growth of *A. borkumensis*.

Resum

La diversitat genòmica dels microorganismes és resultat de la combinació de processos evolutius i d'esdeveniments ecològics. Així, l'estructura específica dels genomes en bacteris és conseqüència de la pressió selectiva deguda a interaccions entre microorganismes i ambient al llarg de l'evolució. Aquestes evidències fan intuir que el DNA conté més informació estructural del que s'esperaria si només es mirés la seva composició de bases nucleotídiques.

El propòsit general d'aquesta tesi doctoral és desenvolupar un marc teòric basat en la genometria, l'estadística i el modelatge matemàtic per tal d'estudiar la relació entre estructura del genoma, estil de vida i metabolisme de microorganismes procariòtics. Per a desvelar la relació entre l'estructura del genoma i l'estil de vida, s'han analitzat un gran nombre de genomes mitjançant un mètode de física estadística, el qual aconsegueix reduir la complexitat genòmica procariota a un únic paràmetre —la correlació intrínseca de llarg abast, que es relaciona directament amb l'estructura fractal de la seqüència del DNA— que pot ser utilitzat en genòmica comparativa i anàlisis ecològiques.

Els mètodes utilitzats per a l'estudi de correlacions de llarg abast en genomes han sigut els "passejos" de DNA i Detrended Fluctuation Analysis (DFA). Els "passejos" de DNA és un mètode de genometria basat en una funció derivada de la posició seqüencial de cada nucleòtid al llarg d'una seqüència de DNA. El "passeig" resultant és representatiu del "paisatge" del DNA i permet la comparació simultània entre diferents genomes. El DFA proporciona un senzill paràmetre quantitatiu —l'exponent d'escala α— que representa les propietats de correlació d'una seqüència. La combinació de l'enfocament seqüencial "passejos" de DNA–DFA amb una aproximació funcional (distribució de gens ortòlegs, COG) mostra que tant les correlacions com la distribució de COGs del genoma poden tenir el seu origen en factors similars com expansions i contraccions en el repertori genòmic o adaptació a hàbitats extrems.

La relació entre l'estil de vida i el metabolisme s'ha examinat mitjançant un estudi de genòmica comparativa de dos bacteris marins que utilitzen exclusivament hidrocarburs com a fonts de carboni i d'energia en hàbitats diferents, *Alcanivorax borkumensis* i *Oleispira antarctica.* S'han estudiat les bases genòmiques dels seus inusuals trets ecofisiològics per tal de millorar la comprenssió de la influència de la temperatura en el creixement bacterià basat en la degradació d'hidrocarburs.

Finalment, s'ha portat a terme una aproximació de genòmica funcional utilitzant el modelatge matemàtic de la xarxa metabòlica codificada al genoma d'*Alcanivorax borkumensis* per tal d'aprofundir en la relació entre la composició del genoma i el fenotip metabòlic. El conjunt de gens, proteïnes, reaccions i metabòlits que participen a l'activitat metabòlica s'ha identificat, classificat i interconnectat per a reconstruir una xarxa metabòlica *in silico*. Aquesta reconstrucció metabòlica ha permès, mitjançant la utilització de mètodes basats en la restricció de fluxos i en l'Anàlisi d'Equilibri de Flux (FBA), caracteritzar el peculiar tret ecofisiològic d'aquest microorganisme i pronosticar fenotips mutants viables de la cèl·lula. El modelatge dels fluxos de carboni envers els de nitrogen ha permès el descobriment de condicions específiques en les quals l'excedent de carboni disponible en els hidrocarburs no es traduïa directament a biomassa sinó que es desviava cap a la producció de polyhydroxyalkanoates (bioplàstics). Les prediccions van mostrar un potencial en l'ús del model com a eina d'anàlisi *in silico* per a estudis detallats del creixement de *A. borkumensis*.

Resumen

La diversidad genómica de los microorganismos es resultado de la combinación de procesos evolutivos y de acontecimientos ecológicos. Así, la estructura específica de los genomas de bacterias es una consecuencia de la presión selectiva debida a interacciones entre microorganismos y ambiente a lo largo de la evolución. Estas evidencias hacen intuir que el DNA contiene más información estructural de lo que se esperaría mirando sólo la composición de las bases nucleotídicas.

El objetivo general de esta tesis doctoral es desarrollar un marco teórico basado en la genometria, la estadística y el modelado matemático para estudiar la relación entre estructura del genoma, estilo de vida y metabolismo de microorganismos procariotas. Para desvelar la relación entre la estructura del genoma y el estilo de vida, se han analizado un gran número de genomas mediante un método de física estadística que consigue reducir la complejidad genómica procariota a un solo parámetro —la correlación intrínseca de largo alcance, que está relacionada directamente con la estructura fractal de la secuencia de DNA— que puede ser utilizado en genómica comparativa y en estudios ecológicos.

Los métodos escogidos para el estudio de las correlaciones de largo alcance en genomas han sido el “paseo” de DNA y el Detrended Fluctuation Analysis (DFA). El “paseo” de DNA es un método de genometria basado en una función derivada de la posición secuencial de cada nucleótido a lo largo de una secuencia de DNA. El "paseo" resultante es representativo del "paisaje" del DNA y permite la comparación simultánea entre diferentes genomas. El DFA proporciona un sencillo parámetro cuantitativo —el exponente de escala α— que representa las propiedades de correlación de una secuencia. La combinación del enfoque secuencial “paseos” de DNA–DFA con una aproximación funcional (distribución de genes ortólogos, COG) muestra que tanto las correlaciones como la distribución de COGs del genoma pueden estar originadas por factores similares como extensiones y

contracciones en el repertorio genómico o la adaptación a hábitats extremos.

La relación entre el estilo de vida y el metabolismo se ha examinado mediante un estudio de genómica comparativa de dos bacterias marinas que utilizan exclusivamente hidrocarburos como fuentes de carbono y energía en hábitats diferentes, *Alcanivorax borkumensis* y *Oleispira antárctica*. Se han estudiado las bases genómicas de sus inusuales rasgos ecofisiológicos para mejorar la comprensión de la influencia de temperatura sobre el crecimiento bacteriano basado en la degradación de hidrocarburos.

Finalmente, se ha realizado una aproximación de genómica funcional utilizando el modelado matemático de la red metabólica codificada en el genoma de *Alcanivorax borkumensis* con el fin de profundizar en la relación entre la composición del genoma y el fenotipo metabólico. El conjunto de genes, proteínas, reacciones y metabolitos que participan en la actividad metabólica se ha identificado, clasificado e interconectado para reconstruir una red metabólica *in silico*. Dicha reconstrucción metabólica ha permitido, mediante el uso de métodos basados en la restricción de flujos y en el Análisis de Equilibrio de Flujo (FBA), caracterizar el peculiar rasgo ecofisiológico de este microorganismo y pronosticar fenotipos mutantes viables de la célula. El modelado de los flujos de carbono y de nitrógeno permitió el descubrimiento de condiciones específicas en las cuales el excedente de carbono disponible en los hidrocarburos no se traducía directamente a biomasa sino que se desviaba hacia la producción de polyhydroxyalkanoates (bioplásticos). Las predicciones mostraron un potencial en el empleo del modelo como un instrumento de análisis *in silico* para estudios detallados sobre el crecimiento de *A. borkumensis*.

I.

Introduction

I Introduction

There have been many important recent developments in the knowledge of the breadth of prokaryote diversity, the understanding of the driving forces behind that diversity, and its significance for fundamental biogeochemical processes on earth. It has become clear that the microorganims we know are actually just the minority. In fact, the majority of microbes are unculturable on laboratory at present. However, there is a growing appreciation that without microbes fundamental ecological processes would not be balanced and understood.

A major advance to understand the extent and nature of microbial diversity has been the development of in vitro genome sequencing. In parallel, there has been the development of *in silico* tools to allow whole-genome comparisons. This has facilitated the study of microbial diversity and evolution, such as allowing the tracking of unculturable microorganisms or the study of organisms from extreme environments. Genomic comparison has helped to identify core genes, horizontal transfer of genomic islands, phenotypic innovation and metabolic pathway evolution. Undoubtedly, genomics has influenced key concepts in microbiology and will place microbial systematics on a much more sound footing (Ward and Fraser, 2005).

1.1 A general perspective of prokaryotes

The first living things on earth are thought to be single cell prokaryotes. The oldest ancient fossil microbe-like objects are dated to be 3.5 billion years old, just a few hundred million years younger than earth itself (Wilde et al., 2001; Schopf et al., 2002). By 2.4 billion years, the ratio of stable isotopes of carbon, iron and sulfur shows the action of living things on inorganic minerals and sediments (Hayes and Waldbauer, 2006; Archer and Vance, 2006) and molecular biomarkers indicate photosynthesis, demonstrating that life on earth was widespread by this time (Cavalier-Smith et al., 2006; Summons et al., 2006).

Prokaryotes became the dominant force of life on the planet for a long time. They grew and diversified at a relatively fast rate and quickly adapted new ways of obtaining energy. The most important group was the cyanobacteria which was able to harness the power of the sun to derive energy resulting in a net oxygen release into the atmosphere. This gradual release of oxygen, over several billion years, into the atmosphere was a key factor in altering the earth's atmosphere into one where oxygen was a major element. This change laid the groundwork for the world we know today. It changed the environment in which life could evolve and introduced a new range of selection pressures that forced life to adapt to an oxygen-rich atmosphere. This was perhaps the most important change in climate in the history of the planet (Olson, 2006; Herrero and Flores, 2008).

The prokaryotes are generally a group of unicellular microorganisms which lack a cell nucleous that encloses the genetic material, or any other membrane-bound organelles. The genetic material of a prokaryote cell consists of one or more DNA chromosomes sometimes accompanied by plasmids and compacted in the cytoplasm and protected by a cell membrane and usually by a cell wall. Prokaryotes include the Archaea and Bacteria domains (Woese et al., 1990) on the basis of differences in 16S rRNA genes. These two groups and the eukaryotes each arose separately from an ancestor with poorly developed genetic machinery, often called a progenote (Fig. 1.1).

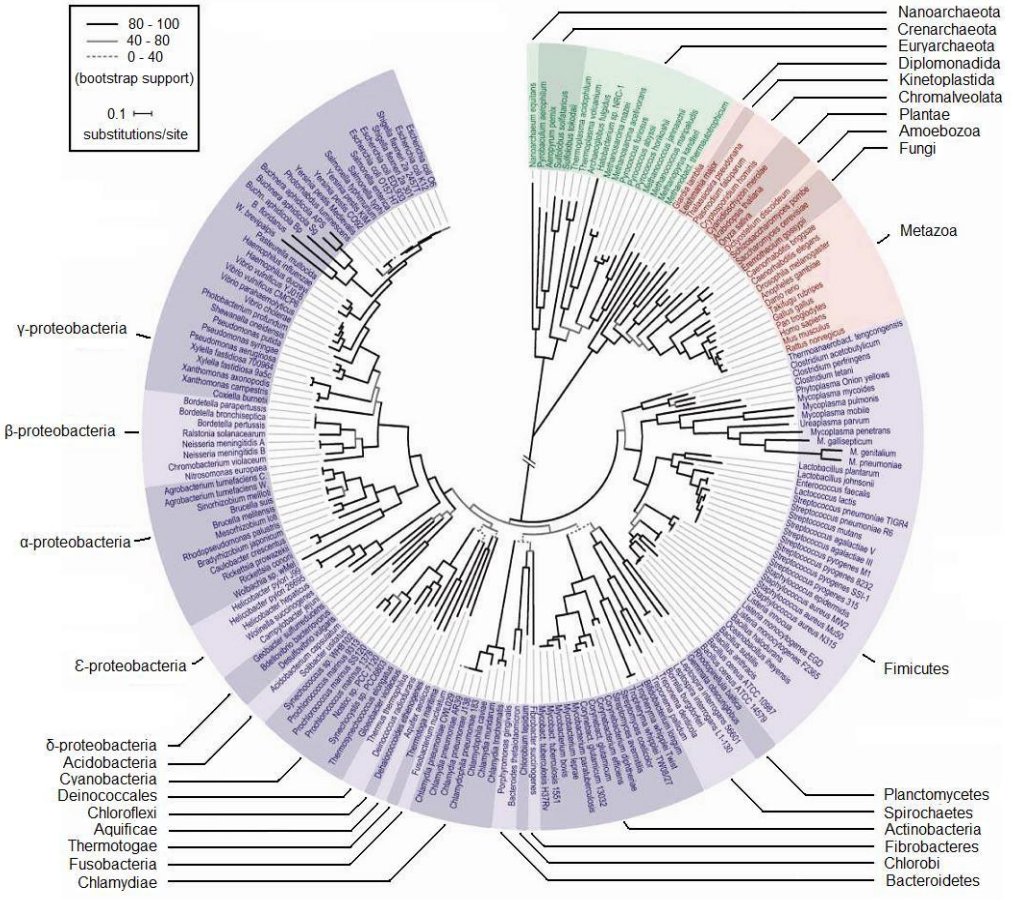


Figure 1.1. Schematic representation of the phylogeny of fully sequenced organisms. The phylogenetic tree covers 191 species whose genomes have been fully sequenced. Green section, Archaea; red, Eukaryota; blue, Bacteria. Labels and color shadings indicate various frequently used subdivisions. Adapted from Ciccarelli et al., 2006.

The exact relationships of the three domains are still being debated, as is the position of the root of the tree. It has also been suggested that due to lateral gene transfer, a tree may not be the best representation of the genetic relationships of all organisms. For instance some genetic evidence suggests that eukaryotes evolved from the union of some bacteria and archaea (one becoming an organelle and the other the main cell) (Blanchard and Lynch, 2000; Alberts et al., 2002). Undoubtedly, knowledge of prokaryotes contributes greatly to the studies of genetics and evolution.

Archaea are extensively similar to Bacteria. Most of the metabolic pathways, which comprise the vast majority of any organism's gene repertoire, are common between them (Koonin et al., 1997). However, Archaea differ from Bacteria in some translation initiation factors and in their cell membrane and cell wall composition (Woese et al., 1990; Woese, 1987; Kandler and König, 1998). Moreover, Archaea tend to adapt quickly to extreme environments, such as high temperatures, high acids or high sulfur. This includes adapting to use a wide variety of food sources.

1.1.1 Importance of prokaryotes

The prokaryotes are ubiquitous in virtually every habitat on earth, including thermal vents, ice sheets, soil, acidic hot springs, radioactive waste, water and deep in the earth's crust, as well as in organic matter and the live bodies of plants and animals. There are approximately 5×10^{30} bacteria on earth (Whitman et al., 1998), forming much of the world's biomass.

The accumulated knowledge of prokaryotes indicates that they play a major role in global biogeochemical cycles, including CO_2 respiration and decomposition (McGrady-Steed et al., 1997) and nitrogen cycling (Horz et al., 2004). Prokaryotes help produce CO_2, which plants take from the atmosphere. The carbon cycle continues when Bacteria help convert the material of which those organisms are made back into CO_2. Bacteria secrete enzymes that partially break down dead matter. Final digestion of this matter takes place within cells by the processes of fermentation and respiration. The CO_2 released by this action escapes

back into the atmosphere to renew the cycle. Some bacteria convert nitrogen in earth's atmosphere into the nitrogen compound ammonia, which plants take up to grow. Bacteria are the only organisms able to carry out this biochemical process known as nitrogen fixation.

From an anthropological point of view, knowledge of prokaryotes greatly contributes to its application in fields as bioremediation, food, mineral extraction or bioengineering. Bioremediation refers to the use of microorganisms to return toxic chemical elements to their natural cycles in nature. It may provide an effective method of environmental cleanup, which is one of the major challenges facing human society today. Bacteria contribute to the fermentation of many products as yogurt or cheese which are produced by bacterial fermentation of milk by the production of lactic acid. An interesting industrial process carried out by bacteria is the recovery of valuable minerals such as copper from ores. Microorganisms of the genera *Thiobacillus* and *Sulfolobus* are able to oxidize sulfides —that is, cause a chemical reaction of sulfides with oxygen— yielding sulfuric acid. This action produces the acid conditions necessary to remove the copper from the ores. Prokaryotes have been also at the center of recent advances in biotechnology thanks to the recombinant DNA technology. Microorganisms became factories for producing multiple copies of proteins in a short time. Bacteria play a role in the environmentally friendly production of industrial components such as polyhydroxyalkanoates or bioplastics, many bulk chemicals, including ethanol, a form of alcohol made from fermented corn and others enzymes used in detergents. They also produce many antibiotics, such as streptomycin and tetracycline.

1.1.2 Prokaryotic diversity

The true extent of prokaryote diversity, encompassing the spectrum of variability among bacteria and archaea, remains unknown. Current research efforts focus on understanding why prokaryote diversification occurs, its underlying mechanisms and its likely impact. The dynamic nature of the prokaryotic world and continuing advances in the technological tools available make this an important area.

Traditionally, microbial identification required both, the isolation of pure cultures and multiple physiological and biochemical tests. This approach used to explore the diversity of microbial communities was biased because of the limitations of the culture methods used. Moreover, the methodology was cumbersome and, as a result, only about 5000 species have been described. The vast majority of prokaryotic microorganisms cannot yet be successfully cultured (Amann et al., 1995). In fact, most microbial species in the environment have yet to be described and, therefore, global microbial diversity is only beginning to be mapped. Comparison of the cultivable bacteria with total cell counts from different habitats showed enormous discrepancies (Amann et al., 1995). Thus, alternative approaches have been developed to complement traditional microbiology. The most common of these approaches involve using information from genetic markers to make inferences regarding diversity. Usually, the rRNA gene sequences are used as indicators of microbial diversity. The use of these molecular techniques and their drawbacks and biases has been reviewed in detail elsewhere (von Wintzingerode et al., 1997). These molecular approaches have enabled the detection of non-culturable species and allowed a more complete and detailed picture of prokaryotic communities (Head et al., 1998; Mlot, 2004).

Additionally, modern genomic techniques as metagenomics have emerged as a powerful tool for analyzing microbial communities, regardless of the ability of member microorganisms to be cultured in the laboratory. Metagenomics is based on the genomic analysis of microbial DNA extracted directly from communities in environmental samples. Essentially, metagenomics provides genomics on a huge scale and enables surveys of various microorganisms present in a specific environment. Thus, the massive uncultured microbial diversity present in the environment can be investigated in detail that was not possible previously.

Genetic methods have shown clearly that culture-based studies of the past almost completely overlooked the vast majority of microbial diversity (Ward et al., 1990; Øvreås, 2000; Floyd et al., 2005). As microbial diversity is sampled more deeply and widely, a far greater appreciation of microbial diversity has been gained (Schloss and Handelsman, 2004; Venter et al., 2004). Thanks to these technological advances, genomics research provides a continuously increasing amount

of information from sequencing experimental data. Thus, the number of analyzed genomes in the public databases has grown steadily over the past eight years (Fig. 1.2). The complete list of sequenced genomes in public databases includes 1126 bacteria and 73 archaea at the time of writing.

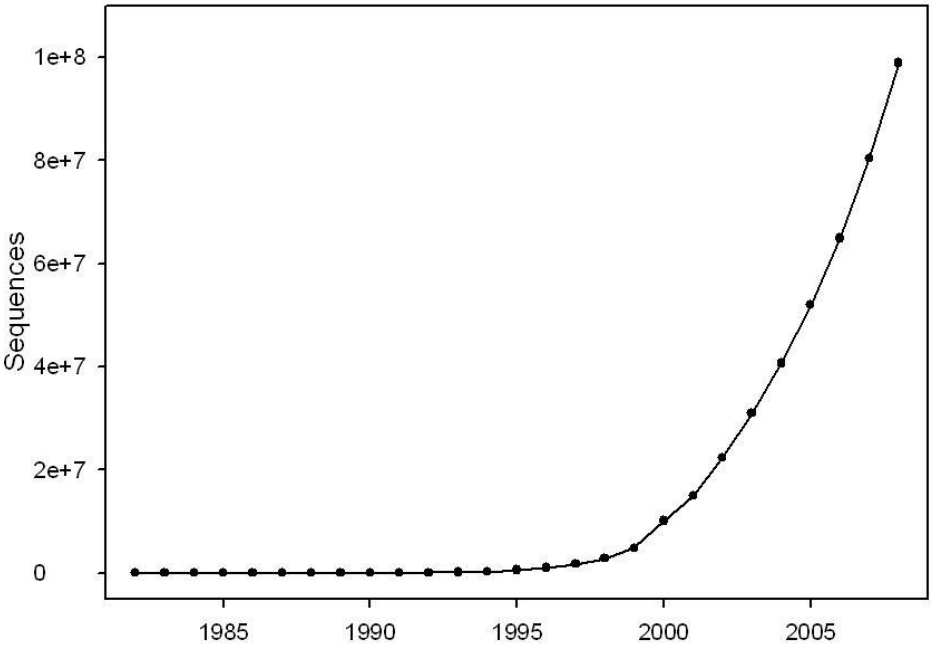


Figure 1.2. Growth of the released sequences in the public database Genbank during the 1982–2008 period.

It becomes clear that due to the constantly increasing rate of raw genomic data, further biological discovery will be limited not by the availability of biological data but by the lack of available tools to analyze and interpret these data. Thus, new tools and ways of thinking for a sequence-based approach to microbial systematic will be required.

The study of microbial biodiversity patterns is still in its infancy. We do not yet know how widespread such patterns are. Ultimately, the study of patterns in microbial biodiversity will shed light on the relative importance of the processes that generate and maintain diversity, such as diversification, extinction, dispersal and species interactions, as well as the potential importance of these processes for maintaining ecosystem functions.

1.2 Genotype – phenotype relationship

Genomic diversity of microorganisms is the result of effects of past evolutionary and ecological events. Thus, the specific genome structure for each strain is a consequence of the selective pressure by the interactions between microorganisms and environment during evolution. Therefore, DNA is predicted to contain more structural information than would be expected from base composition alone. Consequently, the analysis of the relationship between the organization and repeated patterns of prokaryotic genomes structure with their lifestyles and metabolisms would provide valuable information about prokaryotic genome complexity and microbial diversity.

The diversity of genomic structures in prokaryotes is the base of the observed diversity of lifestyles and metabolism, since the genotype and the phenotype are related and inseparable concepts. The **genotype** is defined like the genetic constitution of a cell and becomes a major influencing factor in the development of its phenotype. Genotypic variation is a fundamental prerequisite for evolution, since natural selection affects the genetic structure of microorganisms. The **phenotype** is any observable characteristic or trait of the microorganism, such as its morphology, lifestyle, biochemical or metabolical properties. Phenotypes result from the expression of an organism's genes as well as the influence of environmental factors and possible interactions between the two.

Despite its seemingly straightforward definition, the concept of the phenotype has some hidden subtleties. First, most of the molecules and structures coded by the genetic material are not visible in the appearance of an organism, and are thus part of the phenotype. Second, the

phenotype is not simply a product of the genotype, but is influenced by the environment to a greater or lesser extent. Thus, the concept of phenotypic plasticity describes the degree to which an organism's phenotype is determined by its genotype. A high level of plasticity means that environmental factors have a strong influence on the particular phenotype that develops. If there is little plasticity, the phenotype of an organism can be reliably predicted from knowledge of the genotype, regardless of environmental peculiarities during development. Third, the interaction between genotype and phenotype has often the influence of not only environment but also random variation. Therefore, the relationship between the genotype and the phenotype is complex, highly non-linear and cannot be predicted from simply cataloguing and assigning functions to genes found in a genome.

Comprehensive understanding of the relationship between the genome with the cellular lifestyle and metabolism requires integrated consideration of many interacting components. Mathematics, genometrics, statistics or bioinformatics provides a powerful way of handling such information and allows to effectively develop appropriate frameworks that account for these complexities.

1.3 Genometry and statistics analysis

Genometrics encompasses biometric analyses of chromosomes in order to identify features inherent to chromosome functioning and organization at the level of the whole genome. The word "genometrics" stresses the application of statistical methods to the study of genomic data. Genometric methods allow the study of stochastic properties of nucleotide sequences and provide a quantitative measure that can be used as genomic signature for characterizing and classifying strains. The prokaryotic chromosome could be considered a stochastic process because its sequence of nucleotides is in general non-deterministic since the subsequent nucleotide is rather determined by random elements than by any specific process.

One of the most interesting approaches to the stochastic properties of DNA molecules is the quantitative measure of its intrinsic long-range

correlation, one of the main features related to the whole genome structural composition. The long-range correlation is related directly to the fractal structure of the DNA sequence or self-similarity. The concept of self-similar processes (Kolmogorov, 1961) was introduced into mathematics through the influential work on fractals (Mandelbrot, 1982). A sequence is defined as self-similar if its fragments can be rescaled to resemble the original sequence itself. A scaling exponent —also called the *self-similarity parameter*— can be defined by this rescaling process. A stationary sequence with long-range correlations can be integrated to form a self-similar process. Therefore, measurement of the self-similarity scaling exponent of the integrated series can tell us the long-range correlation properties of the original sequence. Thus, a long-range correlated sequence suggests the existence of repetitive patterns inside it. The search for intrinsic patterns, correlations and parameters measuring self-similarity by scaling exponents has been carried out in past years by statistical methods (Bernaola-Galvan et al., 2002; Peng et al., 1992 and 1995; Chatzidimitriou-Dreismann and Larhammar, 1993).

One of the most appropriated methods proposed in recent years for the study of long-range correlations in genomes is the combination of DNA walk model (Peng et al., 1992) and Detrended Fluctuation Analysis (DFA) (Peng et al., 1994). DNA walking is a genometric method based on a derivative function of the sequential position for each nucleotide along a DNA sequence. The resulting "walk" can be projected on a two-dimensional plot representative of the DNA "landscape" (see examples in Fig. 1.3) and enables the simultaneous comparison among different genome landscapes (Lobry, 1999). The defining feature of such a landscape is the statistical self-similarity. On the other hand, the basic idea underlying the DFA method is to provide a single quantitative parameter —the scaling exponent α— to represent correlation properties of a sequence and the characteristic length scale of repetitive patterns. The main DFA advantage over other methods is that it detects long-range correlations embedded in seemingly nonstationary series (conventional methods such as spectral analysis or root mean square fluctuation can be applied reliably only to stationary sequences).

Why statistical approach to genomics? The applied statistics methods explains a number of statistical properties of genomic DNA sequences such as the distribution of strand-biased regions —those with an excess of one type of nucleotide— as well as local changes in the slope of the

correlation exponent α. The generalized DNA walk model together with DFA simultaneously accounts for the long-range correlations in DNA sequences. Moreover, statistical methods offer theoretical and technical frameworks which allow the study of the biological significance of long-range correlations in genomic DNA sequences.

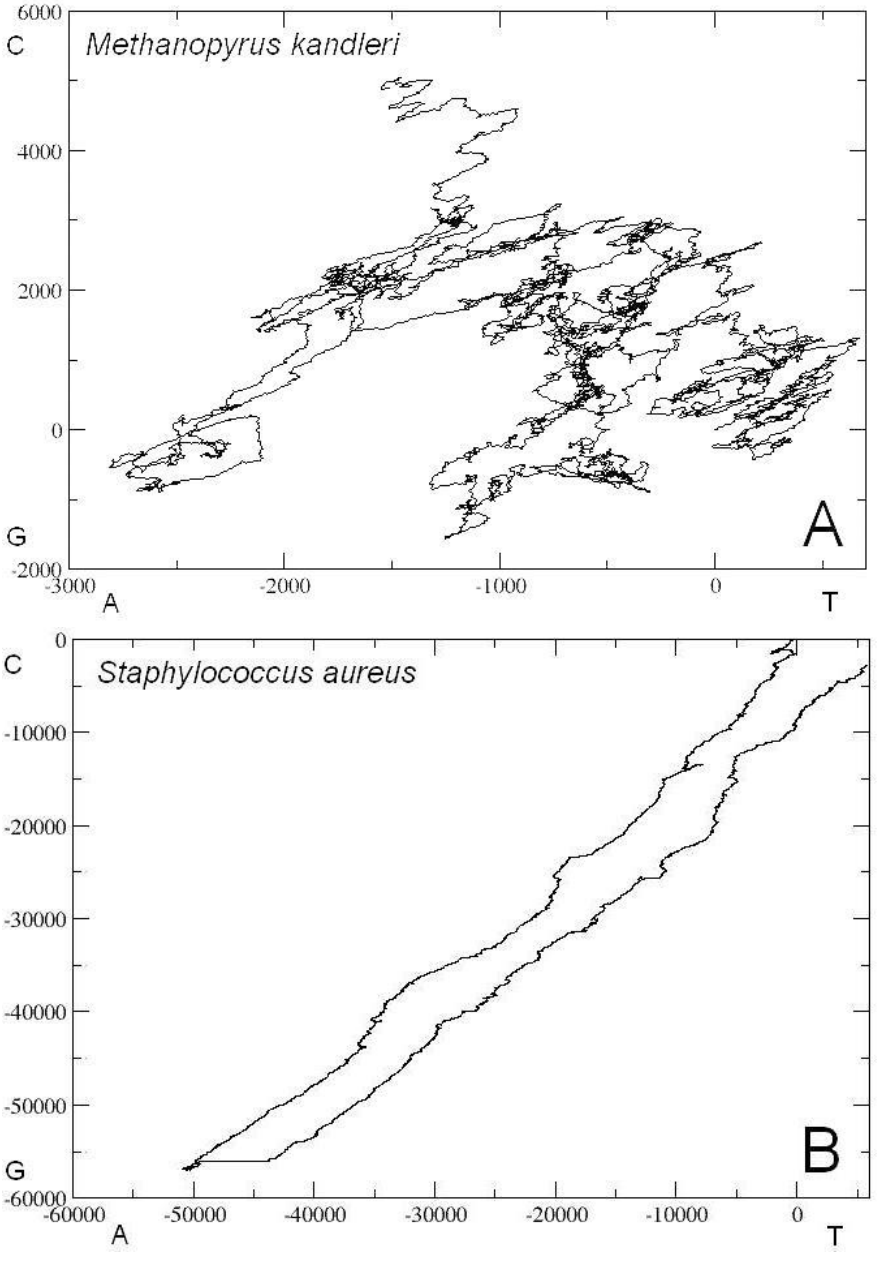


Figure 1.3. Examples of DNA walks of the archaea *Methanopyrus kandleri* and the bacteria *Staphylococcus aureus*.

1.4 Modelling a genome-scale metabolism

The genome–metabolism relationship in prokaryotes has lately become an active area of research (Edwards et al., 2002). The fields of bioinformatics and theoretical biology are now moving to the forefront of biological discovery from the flood of information now readily available, through fields as metabolic network reconstruction. Thus, genomic information, associated with biochemical knowledge, has been used to reconstruct whole-cell metabolic networks for sequenced microorganisms (Edwards and Palsson, 1999; Schilling and Palsson, 2000). Metabolic reconstruction is a process through which the genes, proteins, reactions and metabolites that participate in the metabolic activity of a biological system are identified, categorized and interconnected to form a network. Metabolic phenotypes can be defined in terms of flux distributions through a metabolic network with the help of mathematical modelling and computer simulation. Most often, the system is a single cell of interest and, by using the genomic sequence as a scaffold, reconstructions can incorporate hundreds of reactions that approximate the entire metabolic activity of a cell. Currently, several mathematical approaches exist for the dynamic analysis of cellular metabolism and its regulation (Shuler and Domach, 1983; Liao, 1993; Palsson and Lee, 1993; Barkai and Leibler, 1997; Bailey, 1998; Tomita et al., 1999; Varner and Ramkrishna, 1999).

With the growing availability of genome sequences, genome-scale metabolic reconstructions have been performed for organisms across all three of these domains (Reed et al., 2006). However, even though biological information is growing rapidly, there is not still enough information to describe mathematically the whole cellular metabolism for a single prokaryotic cell (Bailey, 2001).

Flux Balance Analysis

To overcome the lack of biological information, rather than attempting to calculate and predict exactly what a metabolic network does, the range of possible phenotypes that a metabolic systems can display based on the successive imposition of governing physicochemical constraints should be narrow (Palsson, 2000).

Introduction

Within the prokaryotic domains in particular, metabolic reconstructions have been analyzed using constraint-based methods, which simulate our current understanding of the structure and function of metabolic reaction networks in microorganisms (Covert et al., 2004). Constraint-based methods enforce cellular limitations on biological networks such as thermodynamics (e.g., effective reversibility or irreversibility of reactions), physico–chemical constraints, spatial or topological constraints, environmental constraints or gene regulatory constraints (Price et al., 2004). The addition of these constraints results in the definition of a bounded solution space wherein every possible flux distribution, or every possible metabolic phenotype of the cell, must lie. Although the exact flux distribution cannot be calculated, the properties of the constraint-defined solution spaces can be studied. Experimental measurements can be incorporated as constraints to aid in the calculation of the entire metabolic flux distribution (Vallino and Stephanopoulos, 1993; Wiechert and de Graaf, 1996; Sauer et al., 1997).

The effectiveness of metabolic modelling using constraint-based methods has been demonstrated in predicting the outcomes of gene deletions (Duarte et al., 2004), identifying potential drug targets (Yeh et al., 2004), engineering optimal production strains for bioprocessing (Burgard et al., 2003) and elucidating cellular regulatory networks (Covert et al., 2004). By calculating and examining optimal flux distributions under various conditions, it is possible to generate quantitative hypotheses *in silico* that may be tested experimentally.

One specific example of metabolic modelling using a constraint-based approach is Flux Balance Analysis (FBA). FBA uses linear optimization to determine the steady state reaction flux distribution in a metabolic network based on the systemic stoichiometric, thermodynamic and reaction capacity constraints by maximizing an objective function, such as growth rate (Kauffman et al., 2003; Varma and Palsson, 1994a; Bonarius et al., 1997; Edwards, 1999; Edwards et al., 1999; Feist et al., 2006). FBA has been used in many different applications for over 15 years, mainly to study cellular metabolism extensively, most thoroughly for bacterial genomes.

1.5 Objectives and structure

The general aim of this PhD thesis is to develop a theoretical framework based on genometrics, statistics and mathematic modelling to study the relationship between genome structure, lifestyle and metabolism of prokaryotic microorganisms. The novelty of our approach is the application of a statistical physics methodology to reduce the prokaryotic genomic complexity to a single parameter which will be used for comparative genomics and ecological purposes. The methodology was applied to the whole completed sequenced prokaryotic genomes for looking into general relationships between genome, lifestyle and metabolism.

Regarding the relationship between lifestyle and metabolism, we carried out a comprehensive study of comparative genomics between two marine bacteria, *Alcanivorax borkumensis* and *Oleispira antarctica* which share similar metabolism but growth under different environmental conditions.

Finally, a functional genomics approach using a mathematical model of the whole metabolism network of *Alcanivorax borkumensis* based on its complete set of genes was performed in order to look into the relationship between genome composition and metabolic phenotype. To achieve these objectives we carried out four main studies, which define the main parts of the work presented here.

1.5.1 Ecophysiological significance of scale-dependent patterns in prokaryotic genomes

We combined genometric (DNA walks) and statistical (Detrended Fluctuation Analysis) methods on 456 prokaryotic chromosomes from 309 different bacterial and archaeal species to look for specific patterns and long-range correlations along the genome and relate them to ecological lifestyles.

DNA walking and DFA are genometric methods based on a derivative function of the sequential position for each nucleotide along a DNA sequence. DNA walks provides a two-dimensional plot representative of the DNA "landscape" whereas DFA is a scaling analysis method

providing a single quantitative parameter —the scaling exponent α— to represent correlation properties of a sequence and the characteristic length scale of repetitive patterns.

These techniques allow us to infer statistical properties of the genomes and give a view on the mechanisms that are behind the structure of the microbial DNA, as well as the relationship between these structures and the specificity of habitat or metabolism. Thus, conclusions about the adaptation of the genotype to its natural habitat and the evolutionary process from an ecological perspective can emerge. Different features in the DNA landscapes among genomes from different ecological and metabolic groups of prokaryotes appeared with the combined analysis. Transition from hyperthermophilic to psychrophilic environments could have been related to more complex structural adaptations in microbial genomes, whereas for other environmental factors such as pH and salinity this effect would have been smaller. Prokaryotes with domain-specific metabolisms, such as photoautotrophy in Bacteria and methanogenesis in Archaea, showed consistent differences in genome correlation structure. Overall, we show that, beyond the relative proportion of nucleotides, correlation properties derived from their sequential position within the genome hide relevant phylogenetic and ecological information. This can be studied by combining genometric and statistical physics methods, leading to a reduction of genome complexity to a few useful descriptors.

1.5.2 Genome-scale proteins functions shape the genome of prokaryotes

We examined the links between DNA structure and phylogenetic, genometric, ecological and functional genomic information in microbial genomes analyzing 372 prokaryotic chromosomes from 260 different bacterial and archaeal species by means of a combination of Detrended Fluctuation Analysis (DFA) and canonical analysis.

DFA reduces the complexity of a DNA sequence to a simple integrative parameter representing the correlation properties of the sequence, the named long-range correlation. This correlation value is directly related to the DNA structure and reflects evolutionary footprints left in prokaryotic

genomes. Canonical correspondence and redundancy analysis were applied to study the relationships between long-range correlations with phylogenetic and lifestyle factors.

Thereby, the analyzed chromosomes showed a significant correlation between DNA structure and functional genomic level, besides others correlations between structure and ecology and phylogeny. That feature highlighted a significant relationship between functionality and primary nucleotidic sequence. Functional genomic information was referred to the distribution of individual genes in functional categories according to the data extracted from NCBI Clusters of Orthologous Groups (COG) database.

1.5.3 Whole genome comparison of two hydrocarbon degrading bacteria

An extensive work of comparative genomics of hydrocarbon-degrading microorganisms *Alcanivorax borkumensis* and *Oleospira antarctica* was carried out using both, genometric methods such as GC skew, DNA walk or Detrended Fluctuation Analysis (DFA) and comparisons of the genes contained in each genome. The information given by these methods enabled the simultaneous comparison among the different genomes. Moreover, genes contained in genomes provided essential information for understanding evolutionary relationships and ecological adaptations in these microorganisms.

Alcanivorax borkumensis is a rod-shape mesophilic γ-proteobacterium that uses exclusively hydrocarbons as sources of carbon and energy. This bacterium is found in many marine habitats worldwide and play a globally important role in bioremediation of petroleum oil contamination in marine ecosystems. It is present in low numbers in unpolluted environments, but becomes the dominant microbe in oil-polluted waters (Yakimov et al., 1998; Harayama et al., 1999; Kasai et al., 2001). *Alcanivorax borkumensis* is thus a paradigm of cosmopolitan hydrocarbonoclastic bacterium.

Oleospira antarctica is an aerobic psychrotrophic γ-proteobacterium that uses petroleum oil hydrocarbons as sources of carbon and energy. It is

found in low numbers in marine systems of all geographical areas and in high numbers in oil-polluted waters. Two strains, RB-8T and RB-9, were isolated from hydrocarbon-degrading enrichment cultures obtained from Antarctic coastal marine environments (Rod Bay, Ross Sea) (Yakimov et al., 2003). The isolates share many traits with the recently described genera of marine hydrocarbonoclastic bacteria *Alcanivorax*, *Marinobacter* and *Oleiphilus*, including isolation from a marine environment, purely respiratory metabolism (i.e., lack of fermentative metabolism), relatively restricted nutritional profiles, with a strong preference for aliphatic hydrocarbons.

The goal of this comparative and functional genomics project is to characterize the genomic basis of the unusual ecophysiological features and environmentally significant properties of these microorganisms, and to establish a knowledge base that could help to a better understanding of the influence of temperature on the oil-degrading bacterium grown.

1.5.4 A Genome-scale metabolic model of *Alcanivorax borkumensis*

The last part of the thesis is focused on central questions for understanding microbial diversity as the ecophysiological mechanisms related with the specificity of habitat or the study of the underlying bases of the metabolic diversity. Thus, the interrelation between microbial genomes and environmental parameters is believed to determine the specificity of habitat and the microbial diversity.

A genome-scale metabolic model of *Alcanivorax borkumensis* SK2 was constructed from genome sequence annotation, biochemical and physiological data. The reconstructed network represents the first metabolic model from a marine petroleum oil-degrading bacterium. Framework based on constraint-based modelling and FBA were used to explore the metabolic capabilities, under different environmental and genetic conditions, of the *in silico* model. Model predictions were validated by comparison to experimental measurements of *A. borkumensis* growing on different substrates. The predicted growth parameters were in reasonable agreement with experimental findings. The modelling of carbon fluxes versus those of nitrogen allowed the

discovery of conditions in which the excess carbon available in hydrocarbons is not directly translated into bacterial biomass but carbon overflow is diverted to the production of polyhydroxyalkanoates (bioplastics). This reconstructed metabolic network can help to a better understanding of oil-degrading bacterium grown, to predict cellular phenotypes and further uncover the metabolic characteristics of bioplastics formation.

The goal of the project of functional genomics is to characterize the genomics bases of the peculiar ecophysiology feature and of the environmental significance of this microorganism, and to establish a basis of knowledge that can be explored by the development of applications to accelerate the degradation of hydrocarbons in polluted coastal environments.

II.

Ecophysiological significance of scale-dependent patterns in prokaryotic genomes unveiled by a combination of statistic and genometric analyses

II Ecophysiological significance of scale-dependent patterns in prokaryotic genomes unveiled by a combination of statistic and genometric analyses

2.1 Introduction

Prokaryotes constitute, by far, the largest reservoir of life and encompass the major part of physiological and phylogenetic diversity. A large number of studies have been devoted to exploring microbial biodiversity by 16S rRNA analyses (e.g., Casamayor et al., 2002 and references therein) and, recently, with genomic tools (e.g., De Long, 2004). The present capacity to produce genomic information from both laboratory cultures and complex microbial field assemblages widely surpasses the available technical and intellectual skills to analyze and interpret such huge amounts of data into an ecological and evolutionary context. Due to the present size and constantly increasing rate of new raw data, microbiologists and microbial ecologists need new and integrative ways of thinking about microbial genomes to check quickly for similarities and differences among them and to explore and track interactions among genotypes, phenotypes and the environment. Several authors have recently highlighted the need for new computational tools to analyze and interpret the large amount of nucleotide sequences available in databases (De Long, 2004; Nelson, 2003; Streit and Schmitz, 2004). Genes contained in genomes provide essential information for understanding evolutionary relationships and ecological adaptations in microorganisms and, although there is a wide repertoire of bioinformatics tools, both

further manual checking and lack of close relatives in databases are the main limitations. Conversely, genome size and GC content are two integrative parameters that have been explored by comparative analyses offering interesting information (Muto and Osawa, 1987; Hurst and Merchant, 2001; Marashi and Ghalanbor, 2004; Foerstner et al., 2005; Musto et al., 2006). However, DNA is predicted to contain more structural information than would be expected from base composition alone (Pedersen et al., 2000).

One of the main features of a DNA sequence related to the whole genome structural composition is the long-range correlation, a scale invariant property of DNA. In a correlated sequence, occurrence of a nucleotide in a specific position depends on the previous nucleotides (memory). The long-range correlation is related directly to the fractal structure of the DNA sequence or self-similarity. A sequence is defined as self-similar if its fragments can be rescaled to resemble the original sequence itself. Thus, a long-range correlated sequence suggests the existence of repetitive patterns inside it. The search for intrinsic patterns, correlations and parameters measuring self-similarity by scaling exponents has been carried out in past years by statistical methods (Roten et al., 2002; Bernaola-Galvan et al., 2002; Peng et al., 1992 and 1995; Chatzidimitriou-Dreismann and Larhammar, 1993; Stanley et al., 1996). Peng et al. (1992) studied correlation properties in DNA sequences using a fractal landscape or DNA walk model. DNA walking is a genometric method based on a derivative function of the sequential position for each nucleotide along a DNA sequence. The resulting "walk" can be projected on a two-dimensional plot representative of the DNA "landscape" and enables the simultaneous comparison among different genome landscapes (Elston and Wilson, 1990; Lobry, 1999). From a different perspective, spectral and fractal analyses have been used to unveil long-range correlations in DNA sequences. Li and Kaneko (1992) found long-range correlation by means of spectral analysis in the DNA sequence. Fractal analysis has proven useful for revealing complex patterns in natural objects (Berthelsen et al., 1992; Vieira, 1999), and genome fragments have been classified according to their fractal properties (Anh et al., 2002). Finally, a prokaryotic phylogenetic tree based on fractal analyses has been proposed (Yu et al., 2003).

One of the most appropriated methods proposed in recent years for the study of long-range correlations in genomes is the Detrended Fluctuation

Analysis (DFA) (Peng et al., 1992 and 1994). DFA is a scaling analysis method providing a single quantitative parameter —the scaling exponent α— to represent correlation properties of a sequence and the characteristic length scale of repetitive patterns. It is a method specifically adapted to handle problems associated with nonstationary sequences. DFA takes into account differences in local nucleotide content (heterogeneity) and can be applied to the entire sequence. It shows linear behavior in log–log plots for all length scales, and the long-range correlation property is characterized by the scaling exponent (α), i.e., the log–log slope. DFA has two clear advantages over other methods. First, it detects long-range correlations embedded in seemingly nonstationary series (conventional methods such as spectral analysis or root mean square fluctuation can be applied reliably only to stationary sequences). Second, it also avoids the spurious detection of apparent long-range correlations that are an artifact of nonstationary sequences and differentiates local patchiness —excess of one type of nucleotide in a specific region— from long-range correlations. Conventional methods such as Markov models have limitations in coping with dependencies at multiple scales, although they are more appropriate for analyzing short-range nucleotide correlations. The case of the Fast Fourier Transform (FFT) method is strongly affected at high frequencies analysis by short-range correlations related to codon structure, whereas at low frequencies the signal is distorted by artifacts of the method. The scaling exponent values performed by FFT at midfrequency, however, are close to the values reported by DFA (Buldyrev et al., 1995).

DFA may help characterize different complex systems according to its different scaling behavior. One of the already shown potentials of DFA is a change in the quantification of genome complexity with evolution (Peng et al., 1995). Thus, an increase in the self-similarity —fractal structure— of DNA sequences with evolution has been reported (Voss, 1992), and links between long-range correlations and higher order structure of the DNA molecule have been suggested (Grosberg et al., 1993). It has been shown that scale-independent correlations offer the best compromise between efficient information transfer and immunity to errors on all scales (Voss, 1992), whereas the information theory suggests that one can package the largest amount of information into characters of constant length when a sequence is self-similar (Nagai et al., 2001).

In this work, we propose a combination of DNA walking and DFA methods to help decipher the biological significance of long-range correlations in microbial genomes and the influence of lifestyle in the DNA structure. First, we computed a DNA walk for 456 prokaryotic genome sequences to translate the DNA base sequence into a numerical sequence of Euclidean distances. Next, we used DFA to represent and characterize the correlation properties of the numerical sequence. The specific patterns and long-range correlations were related to phylogenetic, ecological and metabolic information, providing a combined window to look into prokaryotic genome complexity and microbial biodiversity.

2.2 Material and methods

Four hundred fifty-six completely sequenced closed genomes from 309 different species of prokaryotes were downloaded from GenBank (National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/Genbank) in May 2007. The prospected genomes belonged to three archaeal kingdoms (28 chromosomes) and 20 bacterial classes (for more details see Tables 2.1 and 2.2). The run for DNA walks started at position 0 of the annotated sequence. For comparative purposes, we constructed three artificial genomes as controls by randomly mixing the order of bases from original genomes. The following conditions of length and GC percentage were chosen: control 1 had the 1,197,687 bases of *Anaplasma marginale* in random order and 50% GC content. Control 2 was the randomly ordered strain of *Mycoplasma mycoides* (1,211,703 bases in length and 24% GC content). Finally, control 3 had the same length (1,849,735 bases) and GC percentage (70%) as the *Thermus thermophilus* chromosome.

DNA walks

We analyzed the sequential distribution of individual nucleotides along the genomes by the DNA walk method (Appendix A). Here, we have used two types of representations. First, we translate the original nucleotide sequence onto a one-dimensional numerical series grouping the bases in pairs following the hydrogen bond energy rule (SW). Then, the resulting SW DNA walk series were mapped onto an orthogonal plane and fitted by linear regression. The slopes of the regression lines were used as variables for subsequent analysis (SW DNA walk slope). For the second representation, we performed a two-dimensional (2D DNA walk) map where each nucleotide defines one direction in a plane formed by two orthogonal axes (i.e., C versus G and T versus A).

Detrended Fluctuation Analysis (DFA)

Detrended Fluctuation Analysis (Appendix B) was used to calculate the scaling exponents from the two types of DNA walks. On the one hand, scaling exponents were calculated directly from SW DNA walks as they consist of one-dimensional numerical series. On the other hand, the two-dimensional series of 2D DNA walks were transformed into one-dimensional ones by replacing every original x–y point, representing a step of the walk, with its Euclidean distances from the origin of the graph. The resulting one-dimensional series were then used to calculate the scaling exponents.

Discriminant analysis (Afifi et al., 2004) was used to construct the Fisher discriminant function —a linear combination of the variables whose coefficients make maximum the distance between the populations— for species classification into one of two or more groups on the basis of the 2D DFA slope and SW DNA walk slope variables. Computations were carried out with SAS/STAT release 9.1 statistical package (SAS Institute, Inc., Cary, NC, USA).

2.3 Results and discussion

Within the 456 microbial strains analyzed we covered a wide range of both genome lengths and GC content from several phylogenetic lineages. The range of lengths was between 0.16 Mb in *Candidatus carsonella ruddii* and 9.97 Mb in *Solibacter usitatus*. The percentage of GC content ranged between 16.56% in *Candidatus carsonella ruddii* and 74.90% in *Anaeromyxobacter dehalogenans*. Genome length and percentage of GC content were also heterogeneous within each phylogenetic group. For example, the 33 strains analyzed for Actinobacteria differed by up to one order of magnitude in length, whereas the largest difference in GC content was found within the γ-Proteobacteria (up to fourfold difference). We also covered microorganisms with different ecophysiological lifestyles related with optimal growth temperature, pH, salinity and metabolism, according to information from the taxonomy database at NCBI (www.ncbi.nlm.nih.gov) and *Bergey's Manual of Systematic Bacteriology* (Garrity et al., 2001). For more details see Tables 2.1 and 2.2.

	Number of genomes	Genome length (Mb) min	max	range	GC content (%) min	max	range
Archaea							
Crenarchaeota	5	1.67	2.99	1.32	32.79	56.31	23.52
Euryarchaeota	22	1.54	5.75	4.21	31.43	67.91	36.48
Nanoarchaeota	1	0.49		-	31.56		-
Bacteria							
Actinobacteridae	33	0.92	9.97	9.05	46.31	72.83	26.52
Aquificales	1	1.55		-	43.48		-
Bacteroides	7	2.34	6.26	3.92	36.61	66.22	29.61
Chlamydiales	11	1.04	2.41	1.37	34.72	41.31	6.59
Chlorobia	4	2.15	3.13	0.98	44.28	57.33	13.05
Dehalococcoides	2	1.40	1.47	0.07	47.03	48.85	1.82
Cyanobacteria	19	1.66	7.75	6.09	30.80	62	31.20
Deinococci	5	1.85	4.12	2.27	66.64	69.52	2.88
Bacilli	69	1.78	5.41	3.63	32.09	52.09	20
Clostridia	10	2.55	5.73	3.18	28.25	55.80	27.55
Mollicutes	16	0.58	1.36	0.78	23.77	40.01	16.24
Fusobacterales	1	2.17		-	27.15		-
Planctomycetacia	1	7.14		-	55.40		-
α-Proteobacteria	60	0.86	9.10	8.24	27.48	69.01	41.53
β-Proteobacteria	46	1.06	5.34	4.28	48.49	68.99	20.50
δ-Proteobacteria	13	1.46	9.14	7.68	33.28	74.90	41.62
ε-Proteobacteria	10	1.55	2.20	0.65	30.31	48.46	18.15
γ-Proteobacteria	105	0.16	7.22	7.06	16.56	67.53	50.97
Spirochaetales	13	0.30	4.33	4.03	28.30	52.77	24.47
Thermotogales	1	1.86		-	46.25		-

Table 2.1. Archaeal kingdoms and bacterial classes prospected in this work. The number of genomes, length and percentage of GC content within groups is shown. The minimum and maximal values, as well as the amplitude between these two values are also indicated.

	Bacteria	
Hyperthermophiles	*Aquifex aeolicus* *Thermoanaerobacter tengcongensis* *Thermotoga maritima*	
Thermophiles	*Acidothermus cellulolyticus* *Chlorobium tepidum* *Geobacillus kaustophilus* *Moorella thermoacetica* *Rubrobacter xylanophilus*	*Symbiobacterium thermophilum* *Synechococcus sp* *Thermobifida fusca* *Thermosynechococcus elongatus* *Thermus thermophilus*
Psychrophiles	*Colwellia psychrerythraea* *Desulfotalea psychrophila* *Photobacterium profundum*	*Pseudoalteromonas haloplanktis* *Psychrobacter arcticus* *Psychrobacter cryohalolentis*
Halophiles	*Alkalilimnicola ehrlichei* *Synechococcus elongatus* *Synechococcus sp*	*Synechocystis sp* *Thermus thermophilus* *Salinibacter ruber*
Acidophiles	*Acidobacteria bacterium* *Acidothermus cellulolyticus* *Solibacter usitatus*	
Alkalophiles	*Alkalilimnicola ehrlichei* *Bacillus clausii* *Bacillus halodurans*	*Bacillus licheniformis* *Oceanobacillus iheyensis*
Phototrophs	*Anabaena variabilis* *Bradyrhizobium japonicum* *Chlorobium chlorochromatii* *Chlorobium phaeobacteroides* *Chlorobium tepidum* *Erythrobacter litoralis* *Gloeobacter violaceus* *Nostoc* sp *Pelodictyon luteolum*	*Prochlorococcus marinus* *Rhodobacter sphaeroides* *Rhodospirillum rubrum* *Rhodopseudomonas palustris* *Roseobacter denitrificans* *Synechococcus sp* *Synechococcus elongatus* *Synechocystis sp* *Thermosynechococcus elongatus*
Methanogens	*Syntrophomonas wolfei*	
Nitrogen fixers	*Agrobacterium tumefaciens* *Azoarcus sp* *Bradyrhizobium japonicum* *Mesorhizobium loti* *Mesorhizobium sp* *Nostoc sp*	*Rhizobium etli* *Rhizobium leguminosarum* *Rhodopseudomonas palustris* *Rhodospirillum rubrum* *Sinorhizobium meliloti*
Sulfur oxidizers	*Thiomicrospira crunogena*	
Iron reducers	*Magnetococcus sp* *Pelobacter carbinolicus* *Pelobacter propionicus*	*Rhodoferax ferrireducens* *Shewanella frigidimarina*

	Archaea	
Hyperthermophiles	*Aeropyrum pernix* *Archaeoglobus fulgidus* *Methanococcus jannaschii* *Methanopyrus kandleri* *Methanosaeta thermophila* *Methanothermobacter thermautotrophicus* *Nanoarchaeum equitans*	*Pyrobaculum aerophilum* *Pyrococcus abyssi* *Pyrococcus furiosus* *Pyrococcus horikoshii* *Sulfolobus acidocaldarius* *Sulfolobus solfataricus* *Sulfolobus tokodaii* *Thermococcus kodakaraensis*
Thermophiles	*Picrophilus torridus* *Thermoplasma acidophilum* *Thermoplasma volcanium*	
Halophiles	*Halobacterium* *Haloarcula marismortui*	
Acidophiles	*Methanococcus jannaschii* *Methanococcus maripaludis* *Thermoplasma acidophilum* *Thermoplasma volcanium*	*Sulfolobus acidocaldarius* *Sulfolobus tokodaii* *Sulfolobus solfataricus* *Picrophilus torridus*
Alkalophiles	*Natronomonas pharaonis*	
Methanogens	*Methanococcus jannaschii* *Methanococcus maripaludis* *Methanopyrus kandleri* *Methanococcoides burtonii* *Methanosarcina acetivorans*	*Methanosarcina mazei* *Methanosphaera stadtmanae* *Methanosaeta thermophila* *Methanothermobacter thermautotrophicus*
Sulfur Oxidizers	*Pyrococcus abyssi* *Pyrococcus furiosus* *Pyrococcus horikoshii*	*Sulfolobus solfataricus* *Sulfolobus tokodaii* *Sulfolobus acidocaldarius*

Table 2.2. Prokaryotes analyzed in the present work that matched relevant metabolic and ecological grouping.

DNA walk architecture

For each genome we run an SW (strong–weak pairing) DNA walk and a 2D DNA walk (see Appendix A and Fig. 2.1) as reported in previous works (Roten et al., 2002; Lobry, 1999).

Because a direct relationship exists between %GC and slope in the SW plot (correlation coefficient 0.998), SW slopes were used as the equivalent variable for the percentage of G+C bases: positive slopes indicated dominance of GC, whereas negative slopes reflected dominance of AT. The complete set of genomes fit the previously reported assumption that large genomes have a tendency to be richer in GC (Heddi et al., 1998; Moran, 2002; Rocha and Danchin, 2002) and therefore they showed higher SW slopes (Tables 2.3 and 2.4). This has been related to the fact that random mutations are mainly from C to T and from G to A and to the lack of repair mechanisms in reduced genomes that would lead to a TA enrichment (Heddi et al., 1998; Moran, 2002).

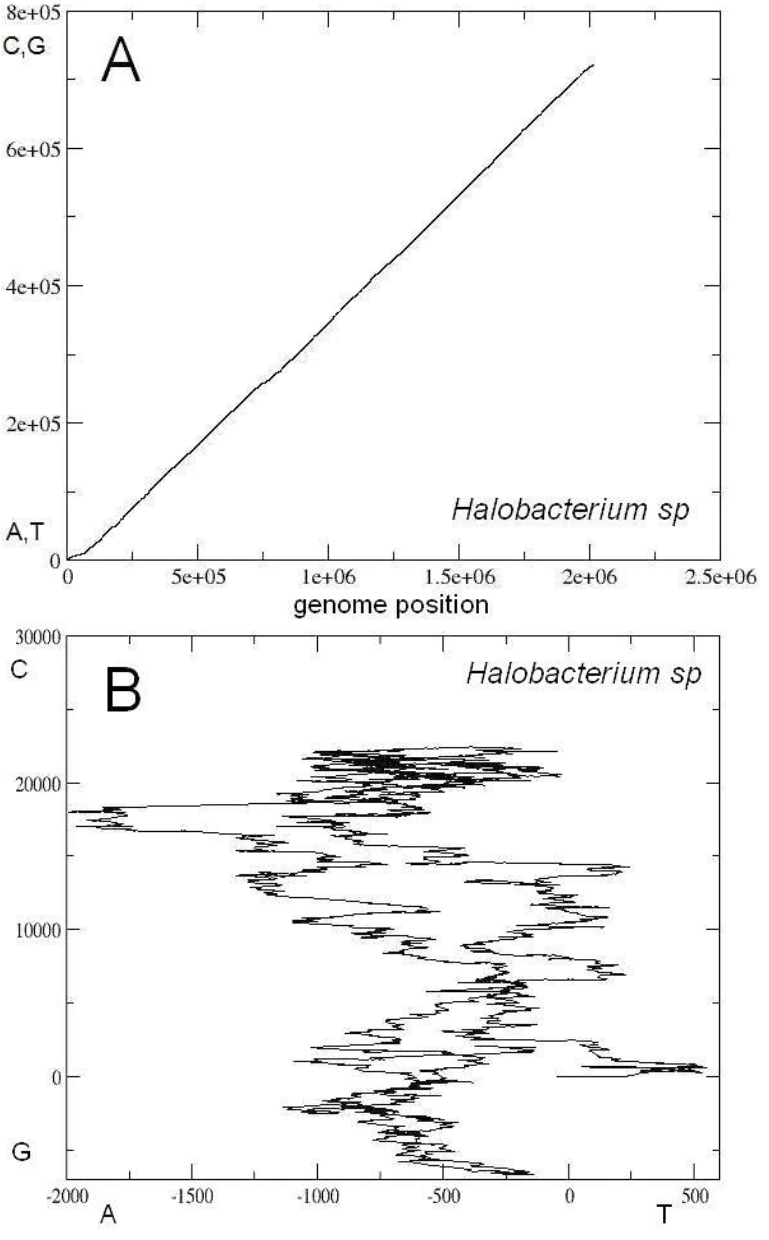


Figure 2.1. SW one-dimensional space (A) and two-dimensional space (B) DNA walk representations of the *Halobacterium* genome. Note the positive slope of the SW DNA walk indicating the dominance of GC along the genome. The run origin starts at the coordinates 0,0.

	SW DNA walk slope			2D DFA slope		
	min	max	range	min	max	range
Archaea						
Crenarchaeota	–0.34	0.14	0.48	0.67	0.69	0.02
Euryarchaeota	–0.45	0.36	0.81	0.62	0.77	0.15
Nanoarchaeota	–0.37		-	0.74		-
Bacteria						
Actinobacteridae	–0.08	0.46	0.54	0.56	0.68	0.12
Aquificales	–0.13		-	0.72		-
Bacteroides	–0.27	0.32	0.59	0.58	0.62	0.04
Chlamydiales	–0.30	–0.17	0.13	0.57	0.60	0.03
Chlorobia	–0.12	0.14	0.26	0.61	0.63	0.02
Dehalococcoides	–0.06	–0.02	0.04	0.56	0.58	0.02
Cyanobacteria	–0.39	0.24	0.63	0.57	0.67	0.10
Deinococci	0.33	0.39	0.06	0.58	0.62	0.04
Bacilli	–0.36	0.04	0.40	0.55	0.66	0.11
Clostridia	–0.44	0.12	0.56	0.58	0.71	0.13
Mollicutes	–0.53	–0.20	0.33	0.64	0.73	0.09
Fusobacterales	–0.46		-	0.74		-
Planctomycetacia	0.10		-	0.59		-
α-Proteobacteria	–0.45	0.39	0.84	0.54	0.73	0.19
β-Proteobacteria	–0.04	0.38	0.42	0.56	0.69	0.13
δ-Proteobacteria	–0.33	0.50	0.83	0.56	0.61	0.05
ε-Proteobacteria	–0.40	–0.03	0.37	0.58	0.72	0.14
γ-Proteobacteria	–0.66	0.35	1.01	0.55	0.72	0.17
Spirochaetales	–0.43	0.06	0.49	0.57	0.67	0.10
Thermotogales	–0.07		-	0.70		-

Table 2.3. SW DNA walk slope and DFA scaling exponent values for the different phylogenetic groups. The minimum and maximal values for each variable are indicated, as well as the range of values within the groups.

	Genomes	SW DNA walk slope			2D DFA slope		
		min	max	range	min	max	range
Hyperthermophiles	19	–0.44	0.23	0.67	0.65	0.77	0.12
Thermophiles	15	–0.27	0.41	0.68	0.58	0.72	0.14
Psychrophiles	8	–0.24	–0.06	0.18	0.58	0.61	0.03
Acidophiles	12	–0.44	0.35	0.79	0.59	0.77	0.18
Halophiles	15	–0.06	0.39	0.45	0.57	0.64	0.07
Alkalophiles	7	–0.29	0.35	0.64	0.57	0.66	0.09
Phototrophs	33	–0.39	0.39	0.78	0.55	0.63	0.08
Methanogens	11	–0.45	0.23	0.68	0.64	0.77	0.13
Nitrogen fixers	18	–0.17	0.19	0.36	0.55	0.61	0.06
Sulfur oxidizers	7	–0.34	–0.10	0.24	0.59	0.70	0.11
Iron reducers	5	–0.17	0.19	0.36	0.58	0.65	0.07

Table 2.4. DNA walk and DFA ranges found for some ecologic and metabolic groups of microorganisms. Values found for each group and the range between the maximum and minimum value for each variable are indicated.

The 2D DNA walk for the complete set of genomes was also within the expected results (Roten et al., 2002; Grigoriev, 1998). These plots are characterized by the so-called mutational strand bias (Lobry, 1999). Many microorganisms show a preference for G over C and T over A in the leading strand and C over G in the lagging strand because of several factors including proofreading efficiencies for the different types of DNA polymerases (Rocha 2002; Worning et al., 2006 and references therein). A simple model for explanation is based on the spontaneous deamination of cytosine that induces mutations from C to T. The rate of this deamination is highly increased in single-stranded DNA, such as the leading strand during DNA replication. This causes prevalence of G over C in the leading strand relative to the lagging strand (Lobry, 1999). Most of the chromosomes analyzed (~80% of total) showed strong strand bias that resulted in a symmetric chromosomal inversion in the 2D DNA walks, in which one-half on the genomic sequence was persistently enriched in two of the bases and the other half was enriched in the complementary ones. Both halves commonly split after an inversion point at which the walk changed direction to return back to the run origin

(see an example in Fig. 2.2A). The remaining chromosomes (~20%) showed weak strand asymmetry (Fig. 2.2B). Artificial controls run for the different genomes lost the observed architecture and fit a single linear path (see inner plots in Fig. 2.2).

Biological significance of long-range correlations

The 2D DNA landscapes were translated to a numerical series of Euclidean distances (see methods) for running the DFA. The resulting curves showed scaling exponents within $\alpha = 0.5417$ (*Brucella melitensis*), the lowest, and $\alpha = 0.7714$ (*Methanococcus jannaschii*), the highest, (Fig. 2.3). We found for each prokaryotic genome a specific scaling exponent with small variations among them. In all the cases, DFA scaling exponents were higher than 0.5, indicating persistent long-range correlations. DFA run for artificial control genomes always had scaling exponents up to 0.50 as expected for uncorrelated sequences (Fig. 2.3). Therefore, long-range correlations in the genome landscape indicate the existence of selective pressures modelling the architecture along the whole prokaryotic genomes (Stanley et al., 1996; Yu et al., 2003 and references therein).

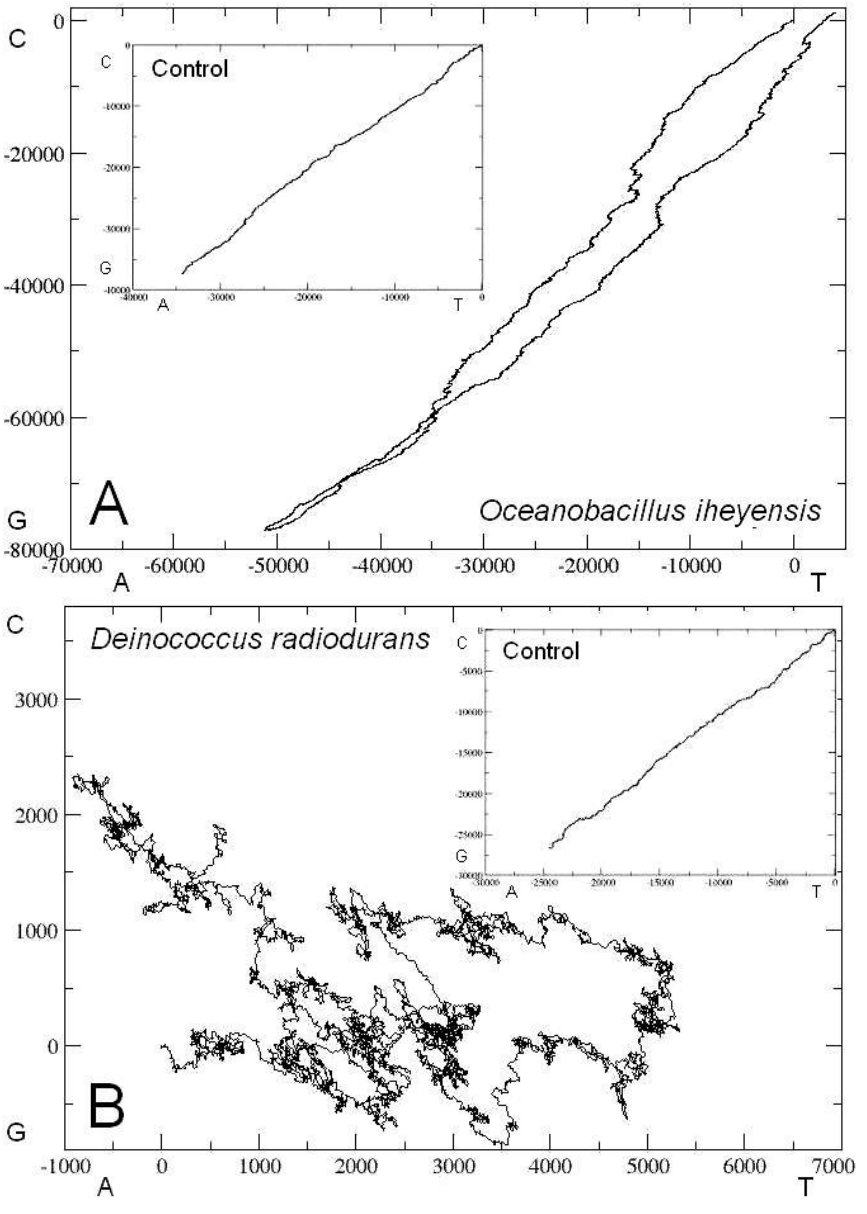


Figure 2.2. DNA walk representations in two-dimensional space (A vs. T and C vs. G in each direction, respectively). Plot A shows a strong strand-biased genome exemplified by *Oceanobacillus iheyensis*, whereas plot B is an example of weak strand-biased genome by *Deinococcus radiodurans*. For each genome a control sequence with the same length and percentage of bases was made.

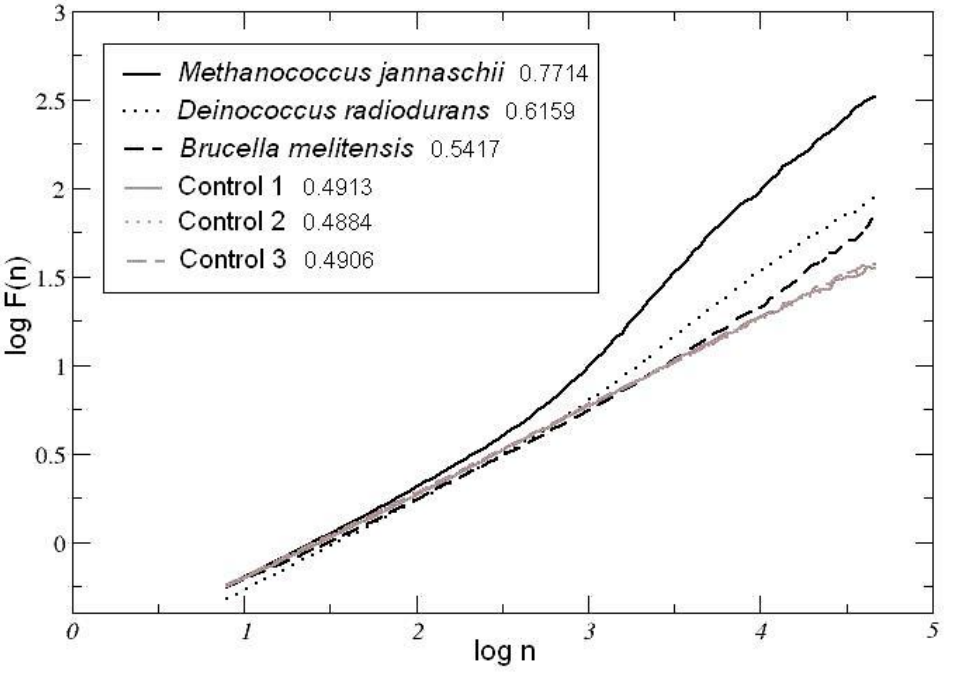


Figure 2.3. DFA run on two-dimensional DNA walks for three selected genomes and controls. The scaling exponents found for the genomes ranged from 0.5417 of *Brucella melitensis* to 0.7714 of *Methanococcus jannaschii*. The three controls were close to 0.5 as expected for random sequences.

The observed long-range correlations in all the DNA sequences may be due to two factors. On the one hand the elongation of the molecule by repetitive structures added inside the genomes (Li and Kaneko, 1992). The fact that long-range correlations were persistent —independent of the scale— means that repetitive structures with different lengths along the genome were present. These repetitive structures may be generated possibly by two important biological mechanisms for evolution: first, elongation of sequences by gene duplication (Li and Kaneko, 1992) and second, elongation and repetition in the genomes by massive lateral transfer of genes from other genomes. On the other hand, long-range correlation can also be related to asymmetric DNA replication along the whole microbial genome, as discussed earlier (Li et al., 1994; Mackiewicz et al., 2002).

We found significant differences in scaling exponents (α) between prokaryotes with weak and strong strand bias (t-student, $p < 0.0001$) for the complete set of genomes analyzed (i.e., genomes with weak strand bias consistently had a higher scaling exponent than strong strand-biased genomes).We also found a significant negative correlation between α value and GC content ($R = -0.474$, $p < 0.005$) (Fig. 2.4). Weak strand asymmetry has been related to the presence of multiple origins of replication (Mrázek and Karlin, 1998) in both Archaea and Bacteria (Worning et al. 2006; Kelman and Kelman, 2004). However, along the complete set of genomes we found weak strand asymmetry in archaeal species with single (e.g., *Methanobacterium thermoautotrophicum* and *Archaeoglobus fulgidus*) as well as multiple origins of replication (e.g., *Methanocaldococcus jannaschii* and *Sulfolobus solfataricus*). Conversely, strong strand biases were observed in Archaea with single (*Methanosarcina mazei*) and multiple origins of replication (*Halobacterium* NCR-1). This suggests that processes acting in genomes with weak strand asymmetry are somehow different from those that occur in the other genomes. Weak mutational bias appeared mainly in the genomes from hyperthermophiles and acidophiles. It is possible that adaptations to environmental stresses in extremophiles may minimize strand asymmetries. The rates of spontaneous mutation —hydrolytic depurination or hydrolytic deamination— are greatly accelerated at extremely high temperatures (Lindahl, 1993). In consequence, hyperthermophiles should have very efficient molecular strategies for repairing DNA under these conditions of chemical instability, since mutation rates in hyperthermophiles are not significantly different from those observed in mesophiles (Jacobs and Grogan, 1997).

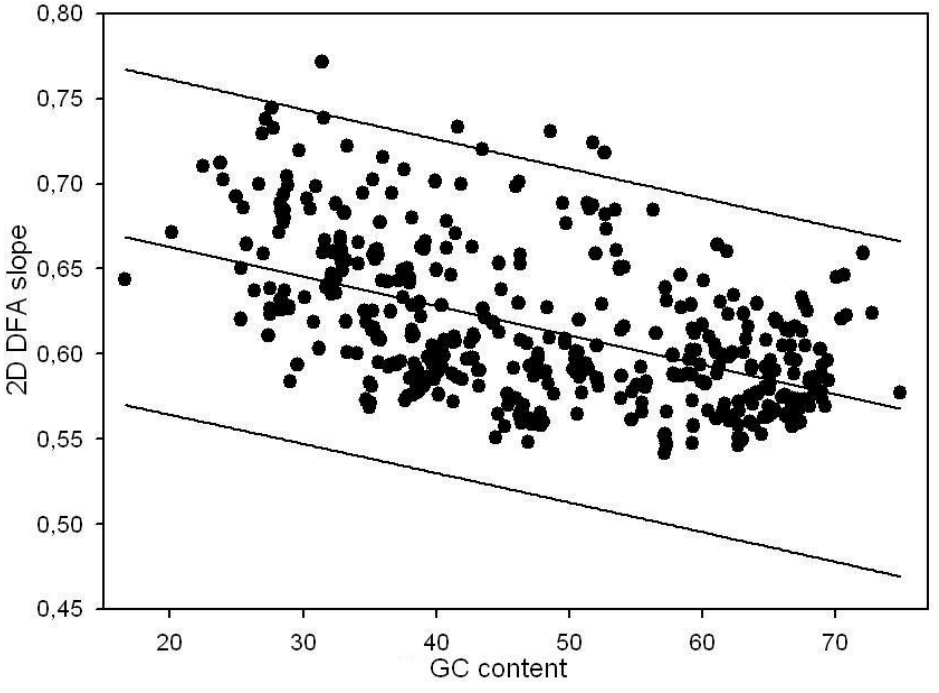


Figure 2.4. Significant correlation ($p < 0.05$) of scaling exponent with GC content for all the tested genomes (99% confidence).

Grouping genomes by phylogeny and lifestyle

It has been previously described that the raw genome sequence harbors a phylogenetic signal (Yu et al., 2003). On the one hand, over- and under-representation of oligonucleotide frequencies have been used by Pride et al. (2003) and Teeling et al. (2004), and more recently by McHardy et al. (2007) for whole genome phylogeny and classification of genomic fragments. On the other hand, the genomic GC content may change faster than previously thought and seems to be globally and actively influenced by environmental conditions (Foerstner et al., 2005 and references therein). Therefore, the combination of DFA and SW slopes should capture these phylogenetic, ecological and metabolic signals.

First, we looked for differences at the phylogenetic level. We plotted the combined graph between the DFA scaling exponent and the SW DNA walk slope (Fig. 2.5B) against the single percentage of each of the four bases (A, T, C and G) obtained by a PCA (Principal Components Analysis) using the covariance matrix (Fig. 2.5A). The combination of DFA values —a quantification of the self-similarity or presence of repetitive patterns over all the length scales contained in the genomes— and SW slopes —directly proportional to the GC content— clearly split prokaryotic chromosomes and controls into two different clusters and showed differences between bacterial and archaeal genomes (Fig. 2.5B). Controls clearly were on the left part with the lowest slopes, close to 0.5, as expected for randomly ordered sequences —the position of one nucleotide was completely uncorrelated with any previous nucleotide—, and separated along the y-axis (SW DNA slope) in agreement with their GC content. On average, Archaea had the highest scaling exponents (DFA slopes > 0.62) and were located on the right part of the plot. Bacteria appeared mainly in the middle zone of the plot (DFA slopes between 0.54 and 0.74). Discriminant analysis showed a correct prediction in 96% of archaea and 85% of bacteria. Conversely, Archaea and Bacteria, as well as the control genomes, were mixed in the quantitative PCA (Fig. 2.5A).

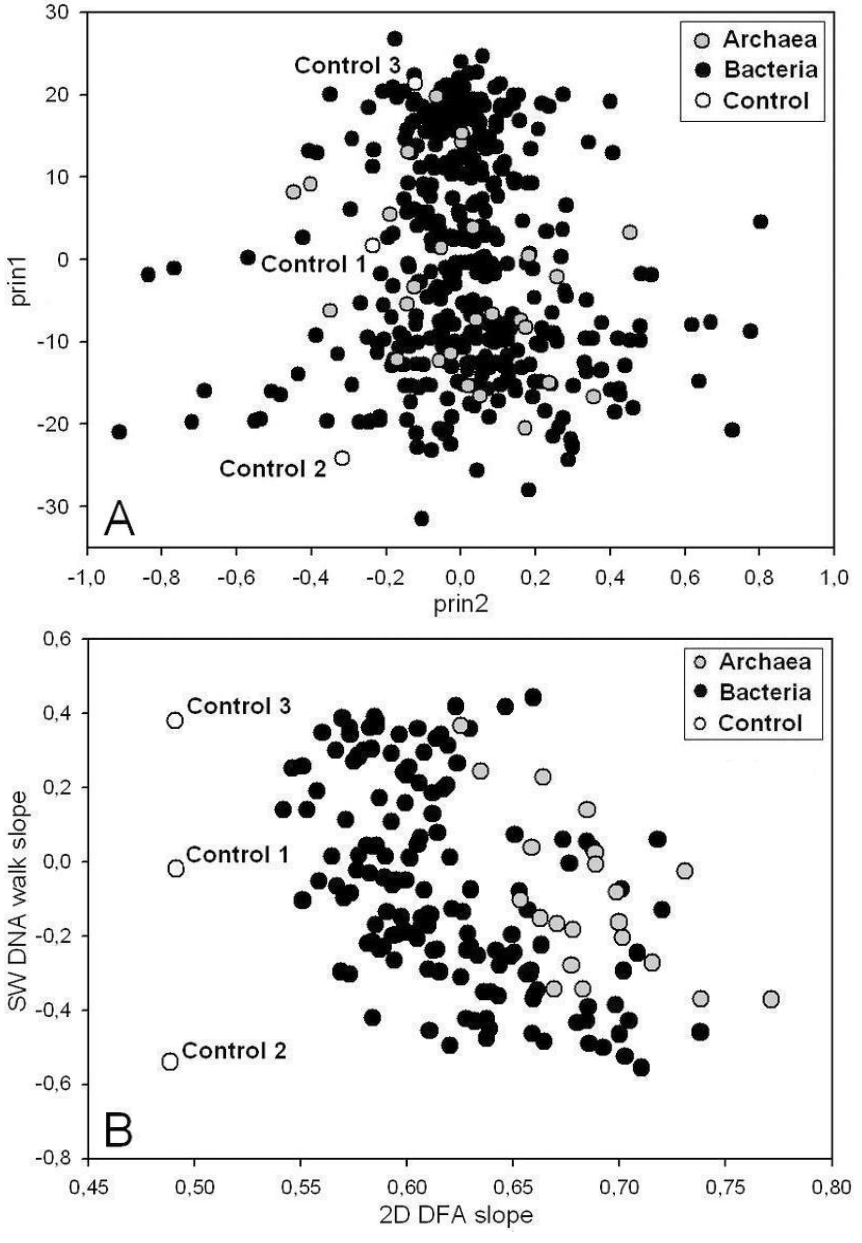


Figure 2.5. Results for the whole set of genomes and controls. Plot A shows Principal Component Analysis (PCA) using percentage of bases. The first principal component (prin1) is represented in the y-axis, whereas the x-axis represents the second principal component (prin2). Plot B shows data combination after genometric (SW DNA walk slope) and statistic (DFA scaling exponents) analyses on the sequential position of nucleotides.

Second, when we focused on ecological lifestyle, some of the groups clustered separately according to DFA and SW slope values (Fig. 2.6). For instance, looking at the optimal growth temperature (Topt), hyperthermophiles showed higher scaling exponents than thermophiles and psychrophiles. The three thermophiles placed within the hyperthermophiles were microorganisms with the highest Topt within their group (close to 60 °C). Psychrophiles were discriminated according to GC content in the low scaling exponent values region (Fig. 2.6A). The discriminant analysis correct prediction was 79% for hyperthermophiles, 80% for thermophiles and 100% for psychrophiles. Correlation between Topt and GC content in prokaryotes has been the focus of a recent controversy. Musto et al. (2004 and 2006) found in a limited number of genomes (ca. 20 genomes) that GC content increased at higher Topt. Conversely, several authors (Hurst and Merchant, 2001; Marashi and Ghalanbor, 2004; Galtier and Lobry, 1997; Wang et al., 2006) concluded that high GC content is not an adaptation to high temperatures and argued that the correlation between both variables is not robust. The data calculated in our survey (456 microbial genomes) indicate that a tendency to the low GC content exists in hyperthermophiles, but examples of genomes with high GC content are present as well. The decrement of GC content in parallel with Topt is very clear between thermophiles and psychrophiles. Thus, it appears that the transition from a hyperthermophilic to a psychrophilic environment would imply a structural adaptation in microbial genomes both in the GC content and in the sequential position of the nucleotides along the genome.

We also observed various clusters related to salinity and pH (Fig. 2.6B). Halophiles showed low scaling exponents (< 0.65) and high GC content. In opposition, most acidophiles presented high scaling exponents and low GC content, although examples of lower DFA values and higher GC contents were also detected. Alkalophiles showed intermediate values of both DFA slopes and GC contents. Therefore pH itself does not seem to have enough separation power. The true prediction calculated using discriminant analysis was 75% for acidophiles, 83% for alkalophiles and 87% for halophiles. Most of the acidophiles were hyperthermophilic archaea and a biased effect with temperature and phylogeny may be present in these cases. In fact, the acidophilic thermophilic bacterium *Acidothermus cellulolyticus* showed low scaling exponent (0.58) and high GC content, in agreement with moderate thermophiles. This example illustrates that temperature is an environmental factor that might

have stronger influence in the microbial genomic structure than pH. Another outlier was the genome of the alkalophilic and moderate halophilic bacterium *Natronomonas pharaonis*. This genome showed higher GC content than the remaining alkalophiles and again pH would have smaller influence on the genomic structure than another environmental factor such as salinity. Finally, photoautotrophs and methanogens were classified into two distinct groups with no overlap in their respective DFA slopes (Fig. 2.6C). Discriminant analysis showed a correct prediction in 91% of methanogens and 100% of phototrophs. Photoautotrophy is an exclusive bacterial metabolism that implies complex enzymatic pathways and no representatives with similar photosystem equipment have been described within archaea. On the other hand, methanogenesis is a feature present only in the archaeal world.

Similarly, nitrogen fixers and sulfur oxidizers showed opposite behavior in both DFA and SW DNA slopes (Fig. 2.6D), although both Bacteria and Archaea are able to carry out both processes. Discriminant analysis showed correct prediction of 94% for nitrogen fixers, 80% for iron reducers and 86% for sulfur oxidizers. In fact, we detected two outliers from the general trend shown by both clusters, one from each: first, the cyanobacterium *Nostoc*, which located away from the remaining nitrogen fixers at the center bottom of the graph —higher DFA slope and lower GC content than the remaining bacteria, mostly from soils—, and second, the mesophilic bacterium *Thiomicrospira crunogena*, which separated from the remaining sulfur oxidizers (all of them were archaea and thermophilic). Therefore, these conclusions could be biased for the limited number of nitrogen-fixer and sulfur-oxidizer genomes still available, but it seems that phylogeny and Topt have stronger influence than these metabolic features in the genomic properties detected.

Overall, the combination of genometrics and physical statistic methods captured intrinsic ecological and phylogenetic patterns present in the likelihood that one nucleotide will be followed by the same nucleotide along the entire prokaryotic genome, offering clues to deciphering their biological significance. Although the application of fractal and time series analyses (e.g., self-similarity and fractional dimensionality) to genome data has been carried out for several years already, these techniques have not seen broad usage in genomics. The application of self-similarity parameters as a measure of persistent, long-range correlations in the DNA sequence relative to different ecophysiological

lifestyles and other biological parameters would help to link physicists and statisticians' approaches with genomic microbiology aims. This work and other recent approaches (e.g., Foerstner et al., 2005; McHardy et al., 2007) will provide microbial ecologists new tools for a better understanding of the naturally occurring genomic structure and variation and, together with detailed studies of the gene content, may help them to follow and understand the genetic adaptations to specific environments and the magnitude of the genetic reservoir present in the microbial world.

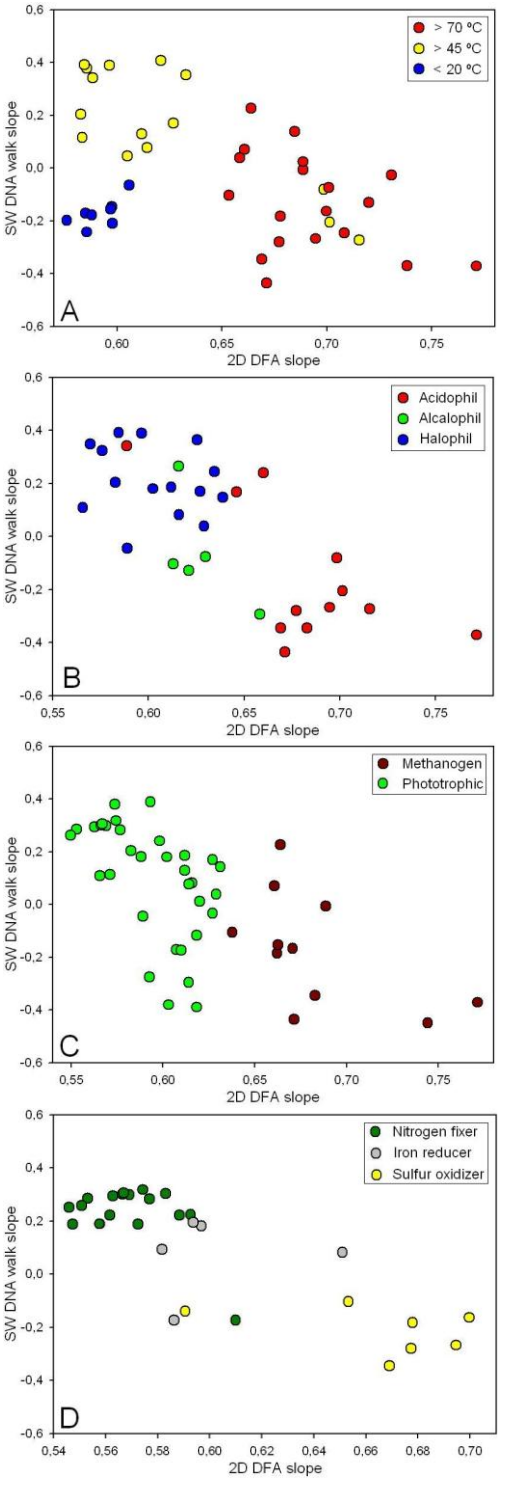

> 70 °C
> 45 °C
< 20 °C
SW DNA walk slope
2D DFA slope
A
Acidophil
Alcalophil
Halophil
B
Methanogen
Phototrophic
C
Nitrogen fixer
Iron reducer
Sulfur oxidizer
D

Figure 2.6. Some ecologic and metabolic clusters formation when the GC content (y-axis) and long-range correlation (x-axis) are plotted in a graph.

III.

Genome-scale proteins functions shape genometric structure in the genome of prokaryotes

III Genome-scale proteins functions shape genometric structure in the genome of prokaryotes

3.1 Introduction

With the accelerating discovery rate of new genomic information from diverse species and environmental sources, comparative genetic analyses have become common practice to obtain clues on the links between functional genomics, evolution and lifestyle. In the case of microorganisms, an increasing number of studies explore microbial biodiversity with genomic tools (e.g., De Long, 2004) since prokaryotes encompass the major part of physiological and phylogenetic diversity. Genes contained in genomes provide essential information for understanding evolutionary relationships, ecological and functional adaptations in microorganisms. Many studies have classified sets of ortholog sequences among different species, and therefore many databases of ortholog groups are available. NCBI Clusters of Orthologous Groups (COG) database for unicellular organisms (Tatusov et al., 2003) contains putative ortholog groups, mostly of prokaryotes. Consistent information about orthology provides the basis for inferring phylogenetic relationships (Tatusov et al., 1997).

Each COG consists of individual ortholog proteins which delineation is achieved by comparison of proteins encoded in different complete genomes from major phylogenetic lineages and elucidation of consistent patterns of sequence similarities. In order to extract the maximum

amount of information from the large set of prokaryotic genomic sequences, COGs allow us to classify conserved genes according to their homologous relationships. Orthologs typically have the same function, allowing the transference of functional information from one member to an entire COG. This relation automatically yields a number of functional predictions for poorly characterized genomes. Each COG represents a functional pathway and changes in COGs content will actually determine changes in the ecological lifestyle. Thus, from the inception of the COG methodology, COGs have the potential for straightforward evolutionary genomic applications in prokaryotes. One of these is the construction of gene-content trees whereby the phyletic patterns of COGs are converted into a distance matrix between the analyzed genomes (Makarova et al., 2007). Moreover, the differences in expression between the COGs from different prokaryotes provide some insight on the lifestyles and habitat, as recently reported for three *Frankia* strains (Sen et al., 2008).

Recently we have shown that correlation properties derived from the position of each single nucleotide within the genome hide relevant ecological information (García et al., 2008). These properties were extracted using Detrended Fluctuation Analysis (DFA) (Peng et al., 1992 and 1994) on the sequential distribution of individual nucleotides along the genomes. DFA is a scaling analysis method providing a single quantitative parameter —the scaling exponent α— to represent long-range correlation properties of a sequence and the characteristic length scale of repetitive patterns connected with self-similarity —fractal structure. An increase in the self-similarity of DNA sequences with evolution has been reported (Voss, 1992), and links between long-range correlations and higher order structure of the DNA molecule have been suggested (Grosberg et al., 1993).

In the present work, we extracted structural (DFA analyses) and functional (COGs analyses) information from 372 prokaryotic chromosomes that were combined with phylogenetic and ecological information by means of canonical analysis. Previous to DFA application we analyzed the sequential distribution of individual nucleotides along the genomes by the DNA walk method (Lobry, 1996a). DNA walks are graphical representations of the fluctuations in nucleotide series and provide quantification on internal deviations of individual nucleotides along the genome. Every genome produces a specific DNA walk which graphical representation can be achieved by several rules for plotting

genomic landscapes (Grigoriev, 1998). Here, we applied the complete set of rules for plotting genomic landscapes and for each of them we ran a DFA. Significant correlations were found between DNA structure (i.e., sequential distribution of individual nucleotides) and its functionality (i.e., the distribution of individual genes in functional categories) with close links to microbial lifestyle.

3.2 Material and methods

We downloaded 460 complete chromosomes from 304 different prokaryotic species available in GenBank (National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/Genbank) in November 2007. Within the prospected prokaryotic genomes, we analyzed 30 archaea from 3 main archaeal phyla and 417 bacteria from 17 bacterial phyla. Overall, we covered a wide range of phylogenies, different ecophysiological lifestyles related with optimal growth temperature, pH, respiration and metabolism, according to information obtained from the taxonomy database at NCBI (www.ncbi.nlm.nih.gov) and *Bergey's Manual of Systematic Bacteriology* (Garrity et al., 2001). For more details see supplementary file 3.1.

Clusters of orthologous genes derived information

For 372 of the total number of chromosomes, we obtained the distribution of encoded proteins within functional categories according to the implemented clusters of orthologous genes (COGs) database (http://www.ncbi.nlm.nih.gov/COG/). For the remaining species this information was not available in COG database. COG, is a systematic grouping of gene families where each COG contains individual ortholog proteins or ortholog sets of paralogs from at least three different lineages and therefore matching an ancient conserved domain. Thus, COGs are grouped according to cell functionality characters. The main functional groups are divided in 4 categories: (i) information storage and processing genes, (ii) cellular processes coding genes, (iii) metabolic genes and (iv) genes of poorly characterized proteins. Each category contains different

subgroups such as genes associated to translation, transcription and DNA replication within the information storage and processing group, or genes related with cell division, post-translational modification, cell envelope biogenesis, cell motility, ion transport and signal transduction mechanisms for the cellular processes category. The metabolism category includes genes associated with energy production, carbohydrate, amino acid, nucleotide, coenzyme and lipid metabolism, and the poorly characterized category includes genes with unknown functions. These groups are formulated by comparing protein sequences of known source to those proteins coded in the genomes which have been extensively studied, have a phylogenetic lineage and have properly been annotated.

DNA walks and Detrended Fluctuation Analysis (DFA)

The original nucleotide sequences of the 460 downloaded genomes were translated onto numerical series using three types of one-dimensional DNA walks: The hybrid rule (KM) where the nucleotides were grouped according to their amino (A or C) or keto forms (G or T); the purine–pyrimidine rule (RY) that groups separately the pyrimidine (C or T) and the purine (A or G) and the hydrogen bond energy rule (SW) that groups the nucleotides in strongly bonded pair (G or C) and weakly bonded pair (A or T). Being n_i the i nucleotide of the genomic sequence and y_i the DNA walk value for the nucleotide n_i, if n_i is a keto forms, pyrimidine or a strongly bonded pair then $y_i = +1$ and if n_i is an amino forms, purine or a weakly bonded pair then $y_i = -1$. Once the numerical series were calculated, their sequential distributions were plotted. See Appendix A for more details.

The three scaling exponents, corresponding with the three types of DNA walk, were calculated for each genome by the Detrended Fluctuation Analysis (Appendix B). In addition, we added to each genome three new variables as controls (KMc, RYc and SWc) which were assigned with random values comprised between the lowest and the highest scaling exponent values calculated by DFA using KM, RY and SW rules, respectively. Thus, controls were used to emphasize the influence of factors as phylogeny, ecology or functional genomics in the chromosome structure.

Fig. 3.1 represents an example of the plots generated by the three types of DNA walks (KM, RY, SW) for the genome of the bacteria *Oceanobacillus iheyensis*, as well as the corresponding scaling exponents associated to each walk.

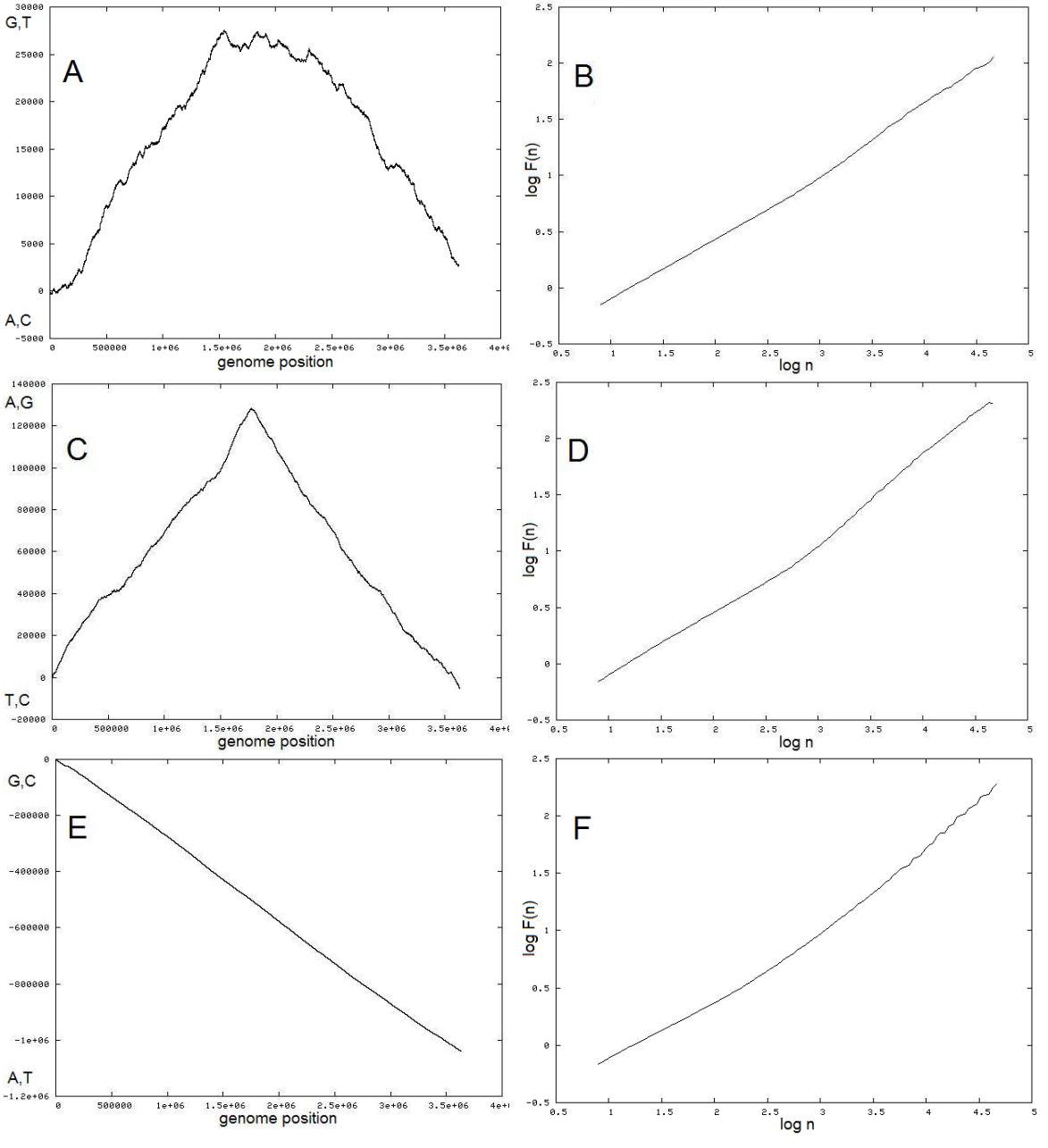


Figure 3.1. DNA walk and DFA graphical representation of the bacteria *Oceanobacillus iheyensis*. The plots A, C and E represent the KM, RY and SW DNA walks, respectively, while the corresponding DFAs are plotted in B, D and F.

Multivariate analysis

For each of the 372 genomes with available COG information we defined up to 58 variables grouped in six different categories: phylogeny, ecology, genometry, genome, COG and DFA (Table 3.1). Within the variables, we included taxonomy hierarchy from domain to specie, oxygen requirement, optimal growth temperature (Topt), pH, salinity, COGs, genome length, percentage of each nucleotide, number of chromosomes, proteins and RNAs, coding and non-coding length mean and the three types of DFAs. Using the whole data set we created a multivariate dataset (see supplementary file 3.2). Two canonical analyses, Canonical Correlation Analysis (CCA) and Redundancy Analysis (RDA), were carried out to explore the relationships between the three types of scaling exponents and the remaining variables.

Taxonomy	Genometry	Genome
superkingdom, phylum, class, order, family, genus, gram, species	length, A, T, G, C, G+C, slope SW, 1stGC, 2ndGC, 3rdGC	chromosomes, plasmids, total prot, total RNA, perc53, coding mean, non coding mean, gen density, CDS
Ecology	**COG**	**DFA**
salinity, oxygen, habitat, Topt, metabolism, pH, biofilm	J, K, L, D, V, T, M, N, U, O, C, G, E, F, H, I, P, Q, R, S, not	KM, RY, SW

Table 3.1. Parameters prospected in this work classified in six categories or tables.

CCA measures the linear relationship between two multidimensional variables (i.e., DFA versus COG tables). It finds two vectors, one for each set of variable, in such a way that the correlation between the set of variables is optimized (Hotelling, 1936). The CCA was performed using the procedure cancorr from SAS/STAT release 9.1 statistical package (SAS Institute, Inc., Cary, NC, USA).

RDA explains the variance of a table of response variables —in our case the three DFA variables— based on a table of explanatory variables —each of the tables corresponding to the other five categories—, (Rao, 1964). RDA seeks the combinations of explanatory variables that best explain the variation of the response variables. It is therefore a constrained ordination process. A constrained ordination produces as many canonical axes as there are explanatory variables, but each of these axes is a linear combination —a multiple regression model— of all explanatory variables. The RDA was carried out using the CANOCO software package, version 4.5 (Ter Braak, 1988). The reported p-values are based on 999 Monte Carlo permutations under the null model.

Finally, discriminant analysis was used to construct the Fisher discriminant function —a linear combination of the variables whose coefficients make maximum the distance between the populations— for species classification into one of two or more groups on the basis of its canonical variables (Afifi et al., 2004). Computations were carried out with SAS/STAT release 9.1 statistical package (SAS Institute, Inc., Cary, NC, USA).

3.3 Results

Scaling exponents by Detrended Fluctuation Analysis

We run three types of DNA walks according to the KM, RY and SW genometric rules (see methods) for each of the 460 bacterial and archaeal downloaded genomes. The walks were translated to a numerical series in order to calculate the scaling exponents for each DNA walk series using the previously reported DFA method (García et al., 2008). DFA values for each genome are shown in supplementary file 3.2. In all the cases, DFA scaling exponents were higher than 0.5 independently of the used rule, indicating persistent long-range correlations for each prokaryotic genome (García et al., 2008). For instance, the resulting DFA curves after applying the RY rule ranged between scaling exponents 0.54, the lowest, in *Mycobacterium leprae* TN and 0.78, the highest, in *Methanococcus jannaschii*.

Relating scaling exponents with phylogenetic, genomic and ecological patterns

The RY rule clearly split the DFA values for Archaea and Bacteria domains whereas the KM and SW rules did not offer enough resolution power (Fig. 3.2A). Furthermore, the RY was the best rule for discriminating ecological groups according to the optimal growth temperature, i.e., the hyperthermophilic species showed higher RY DFA values than the remaining microbial species (Fig. 3.2B). Therefore, we used the RY rule to explore further differences within each domain. Fig. 3.2C shows the scaling exponent averages for 17 archaeal and bacterial phyla. Most of the bacterial DFA were within the range 0.60 – 0.65 whereas Archaea showed values > 0.70. Three bacterial phyla (Aquificae, Fusobacteria and Thermotogae) showed, however, scaling exponents closer to those of archaea.

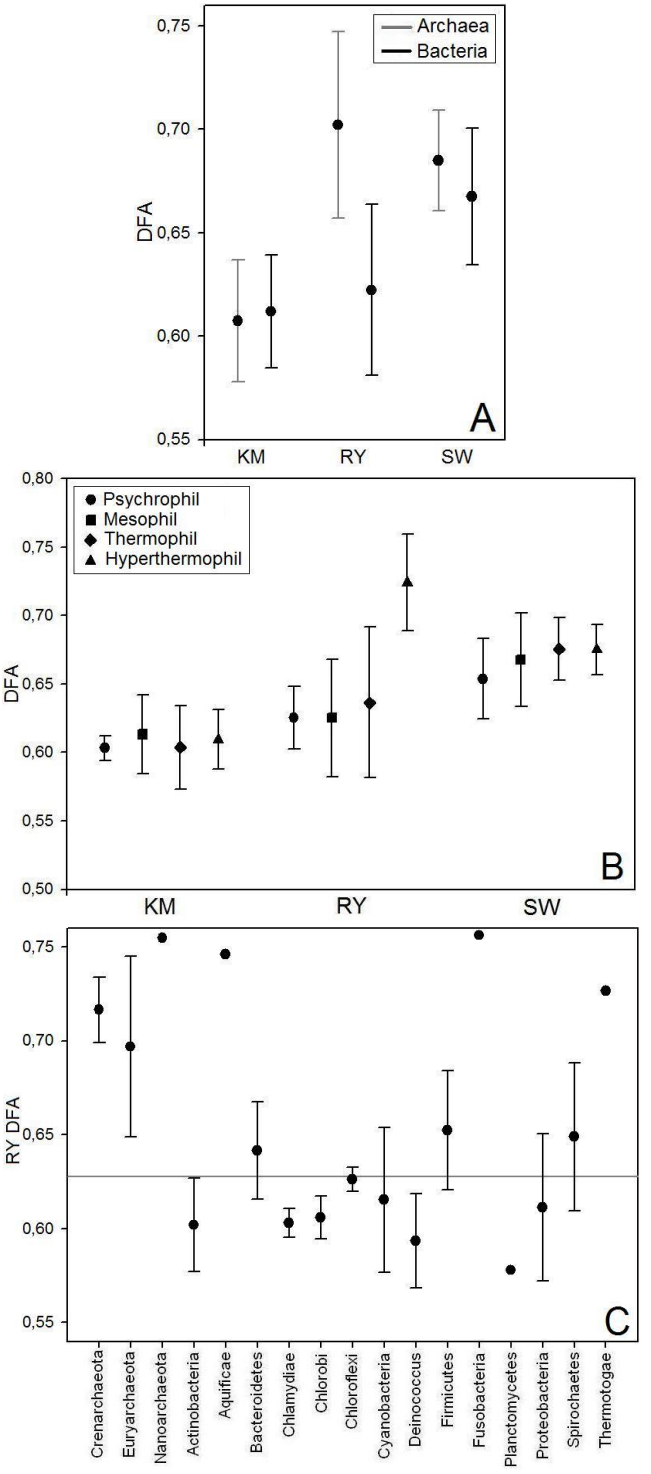

Archaea
Bacteria
DFA
0,75
0,70
0,65
0,60
0,55
KM
RY
SW
A
Psychrophil
Mesophil
Thermophil
Hyperthermophil
0,80
0,75
0,70
0,65
0,60
0,55
0,50
B
RY DFA
Crenarchaeota
Euryarchaeota
Nanoarchaeota
Actinobacteria
Aquificae
Bacteroidetes
Chlamydiae
Chlorobi
Chloroflexi
Cyanobacteria
Deinococcus
Firmicutes
Fusobacteria
Planctomycetes
Proteobacteria
Spirochaetes
Thermotogae
C

Figure 3.2. Mean and standard deviation of DFAs for the three different rules (KM, RY and SW). Panel A shows the mean and standard deviation of the Archaea and Bacteria separately. Note that RY DFA is the best rule to accentuate the differences between the two domains. Panel B shows the mean and standard deviation of groups formed according temperature. RY DFA is again the best rule to differentiate the groups. In general, the three classes belonging to Archaea own higher RY DFAs than bacterial classes (panel C).

The results of the CCA and RDA calculations between DFA and the whole set of variables shown in Table 3.1 (i.e., taxonomy, genometry, genome, ecology, COGs) are summarized in Table 3.2. We centered the CCA on the first canonical correlation —the correlation between the first pair of canonical variables— because this value represents the highest possible correlation between any linear combination of the two variables. The highest correlation value was with genometry (0.83), closely followed by functional genes composition (COGs, 0.78), and genome traits and taxonomy (0.60 and 0.58 respectively). The complete set of ecological parameters showed the lowest correlation value (0.33), although individual ecological variables had high correlation with DFA (see below). This indicated a close link not only between DFA value and genometry (as expected) but surprisingly also between DFA and functional genes composition.

	CCA		RDA	
	DFA	Random DFA (control)	DFA	Random DFA (control)
Taxonomy	0.58	0.23	19.2%	2.7%
Genometry	0.83	0.23	42.7%	2.6%
Genome	0.60	0.23	22.6%	3.4%
Ecology	0.33	0.21	6.6%	2.5%
COGs	0.78	0.30	43.6%	6.5%
All			69.5%	17.2%

Table 3.2. Relationship, obtained by CCA and RDA, between both, DFA and random DFA (control) and the variables included in the five categories. All the first canonical correlations obtained by CCA involving DFAs were statistically significantly different from zero, while, all the canonical correlations involving random DFAs were not statistically significantly different from zero. The total DFA and random DFA variance explained by each of the five categories and by the sum of all the variables obtained by RDA is also shown. All the RDA results for DFA presented statistically significant p-values (< 0.05), while the p-values for controls (random DFAs) were not statistically significant (> 0.05).

The canonical variables for each genome were represented in canonical axes. The original variables were weighted through canonical coefficients. As expected, the random DFA variables used as controls showed low canonical correlation with all the variables (Table 3.2). The F-test output from SAS tested the hypothesis that the first canonical correlation was equal to zero. The F statistic was less than 0.0001 indicating that first canonical correlations were statistically significantly different from zero for all genomic DFA analysis. Conversely, all the F-test from random DFAs showed values higher than 0.05 indicating that first canonical correlations were not statistically significantly different from zero.

RDA showed similar results as CCA but the relationship found between DFAs and COGs was even stronger. Thus, COG was the category that explained most of DFAs variance (up to 43.6%). Genometry also explained a high percentage of DFA variance (42.7%), as expected. Genome traits, taxonomy and ecology accounted for a lower percentage of DFA variance (22.6%, 19.2% and 6.6% respectively). Altogether, the variance of DFA explained by the whole set of variables was close to 70%. Furthermore, all the RDA results for DFA variable presented significant p-values. Conversely, non-significant p-values were achieved when the RDA was run for random DFA variables.

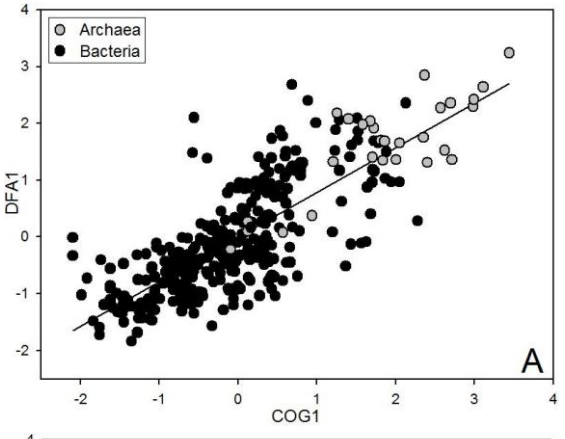
COG analysis

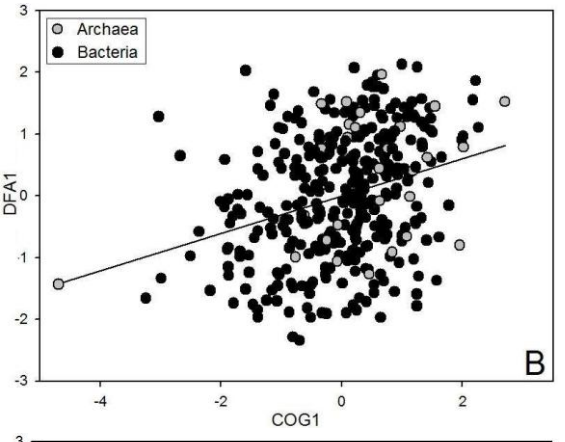
Looking into the main functional categories of COGs, the correlation between DFA and the percentage of genes belonging to each functional category (information storage, cellular processes and metabolism) were similar (Table 3.3). Only the "poorly characterized genes" category showed low correlation with DFAs . Moreover, each of the three main cited categories explained a high percentage of the DFA variance according to the RDA results (Table 3.3).

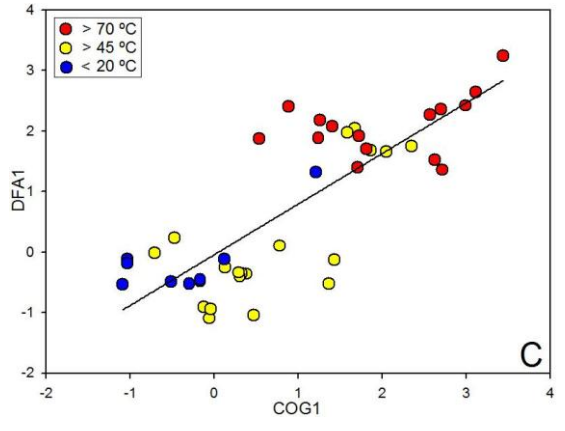
Canonical correlation between DFA and COG is represented in Fig. 3.3 together with the first canonical correlation analysis between random DFA values and COGs. Consistent differences between DFA and COG canonical variables between Archaea and Bacteria were detected (Fig. 3.3A). In general, Archaea showed higher canonical variables than Bacteria. The plots showing the clusters related with temperature and pH (Fig. 3.3C and 3.3E respectively) were, logically, nearly equivalent than the plot regarding the canonical correlation between DFA and genometry (data not shown) since its canonical correlations were similar. The discriminant analysis correct prediction was 94% for hyperthermophiles, 61% for thermophiles and 89% for psychrophiles. The true prediction calculated using discriminant analysis was 75% for acidophiles and 100% for alkalophiles. The three controls did not show canonical correlation either with the genometric values or with ecology (Fig. 3.3B, 3.3D and 3.3F).

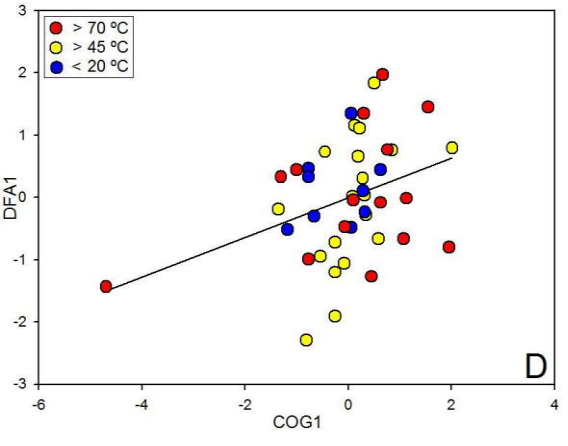
	CCA		RDA	
	DFA	Random DFA (control)	DFA	Random DFA (control)
Information storage	0.54	0.14	15.5%	1.1%
Cellular processes	0.64	0.16	26.3%	1.5%
Metabolism	0.61	0.18	25%	1.7%
Poorly characterized	0.36	0.19	3.9%	1,7%

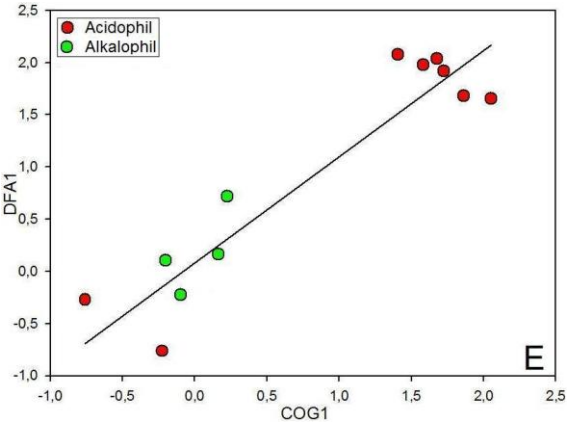
Table 3.3. Relationship, obtained by CCA and RDA, between DFA and random DFA (control) and the percentage of individual cogs belonging to the four main functional categories. The first canonical correlations of all the DFAs were statistically significantly different from zero, while, the canonical correlations of all the random DFAs were not statistically significantly different from zero. The total DFA and random DFA variance explained by each of the four categories of COGs obtained by RDA is also shown. All p-values for DFA results presented statistically significant values (< 0.05), while the p-values for controls (random DFAs) were not statistically significant (> 0.05).


Archaea
Bacteria
DFA1
COG1
A


Archaea
Bacteria
DFA1
COG1
B


> 70 °C
> 45 °C
< 20 °C
DFA1
COG1
C


> 70 °C
> 45 °C
< 20 °C
DFA1
COG1
D


Acidophil
Alkalophil
DFA1
COG1
E

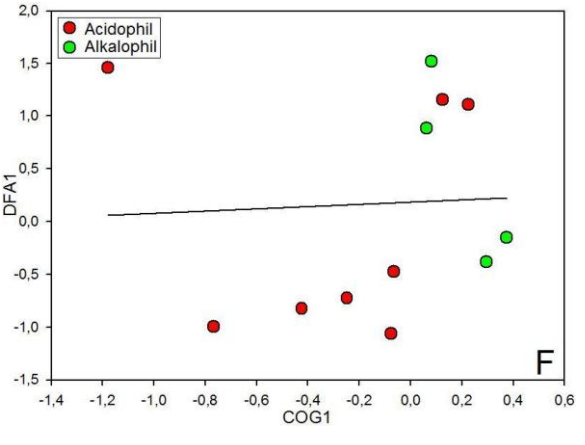

Acidophil
Alkalophil
DFA1
COG1
F

Figure 3.3. Graphical representation of the correlation between the first canonical variables concerning the DFA and COG variables calculated by Canonical Correlation Analysis. CCA was also run with random DFAs values used as a control. Panels A and B show the distribution of Bacteria and Archaea domains running CCA with DFA and random DFA variables, respectively. Panels C and D represent the canonical variables, using DFA and random DFA, respectively, with the groups highlighted according to their Topt. Panels E and F represent canonical variables using DFA and random DFA, respectively. The groups are highlighted according to their pH.

Fig. 3.4 showed that the standard deviation of the percentage of COGs for each genome increases with the DFA average of the three rules. Briefly, the more heterogeneous distribution of genes in each functional category, the higher the scaling exponent of the genome. Contrarily, the random DFA average did not present any relation with the heterogeneity of COGs.

Furthermore, Table 3.4 shows that the groups with higher scaling exponents exhibited higher standard deviation in COG percentage than the groups with low scaling exponent. Thereby, Archaea presented more standard deviation than Bacteria, and hyperthermophiles and acidophiles the highest standard deviation within the group according temperature and pH, respectively.

	Std COG
Archaea	5.09
Bacteria	4.73
Hyperthermophiles	4.89
Thermophiles	4.53
Mesophiles	4.78
Psychrophiles	4.15
Acidophiles	4.97
Mesophiles	4.76
Alkalophiles	4.26

Table 3.4. Standard deviation of percentage of COG for the chromosomes analyzed. The differences between microorganisms grouped according to their domain, Topt and pH are indicated.

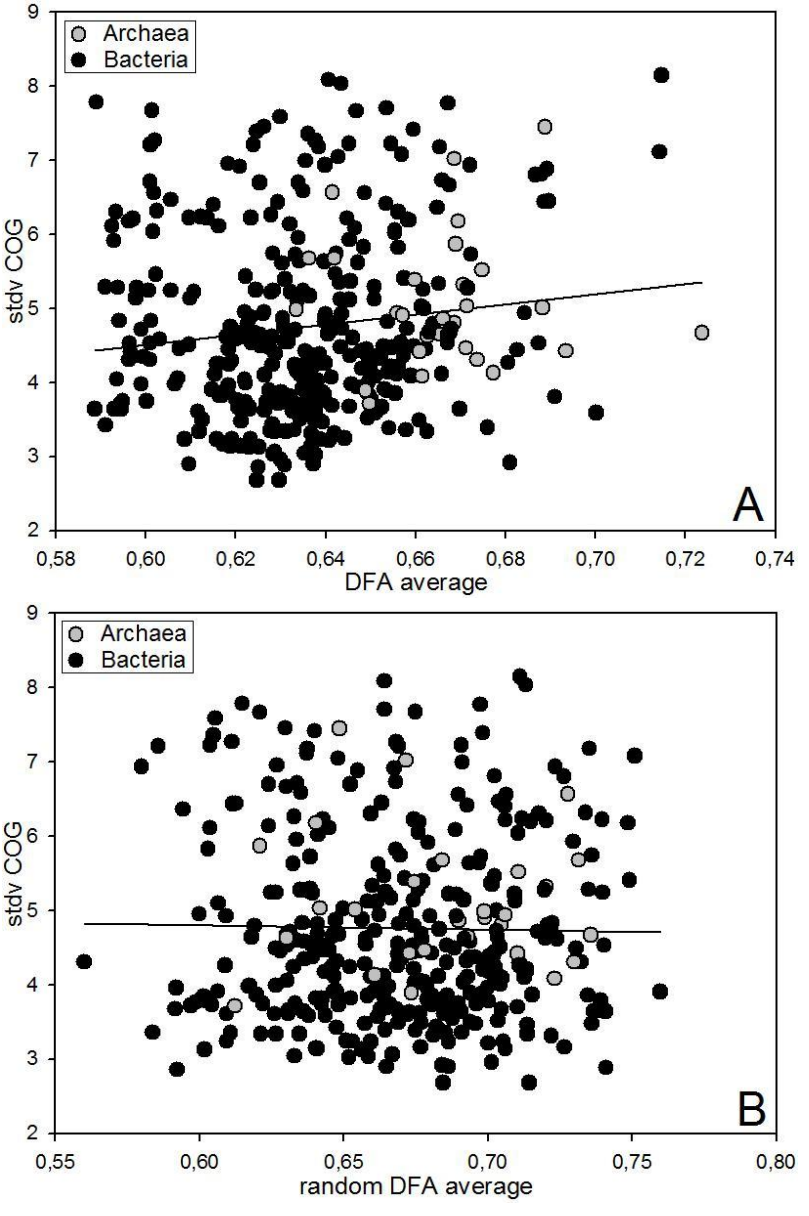


Figure 3.4. Relationship between DFA values and standard deviation of COG for all the chromosomes analyzed. Panel A shows the DFA average of the three values (KM, RY and SW) while panel B is used as a control since it shows the average of the three random DFAs.

3.4 Discussion

DFA – COG relationship

The most outstanding result obtained by canonical analysis was the high canonical correlation between DFA and COG, as well as the high DFA variance explained by COGs indicating that the distribution of the nucleotidic sequence along the genome is somehow related with the genome functionality (COG). As the heterogeneity of percentage of COGs is directly correlated with the canonical variable COG1, it is suggested from Fig. 3.3 that heterogeneity is also related with high DFA1 values and, in consequence, with high scaling exponents.

It has been proposed that the genomic properties of bacteria are greatly conditioned by their specialist or generalist lifestyle (Pushker et al., 2004). Thus, each prokaryotic genome has a specific distribution of its genes grouped into functional COGs as a result of different factors like its evolutionary story and lifestyle; examples of deletion of genes no longer required in a specialized environment have been recently reported (Toh et al., 2006). The emergence of new families of genes in individual lineages, the clade-specific gene loss and the horizontal gene transfer have been recognized as the major evolutionary factors of prokaryotic evolution (Koonin et al., 1997; Aravind et al., 1998; Doolittle, 1999a; Doolittle, 1999b; Logsdon and Faguy, 1999; Nelson et al., 1999; Makarova et al., 2007). It has been also suggested that the inferred frequencies of deletions, duplications and horizontal gene transfers depend on bacterial lifestyle features (Boussau et al., 2004). One of the main features of the COG —its evolutionary plasticity— has been previously reported (Koonin, 2000). This plasticity may be also due to the mentioned forces involved in prokaryotic evolution. Furthermore, gene duplication and gene lateral transfer were proposed to be the responsible to high scaling exponents in prokaryotic genomes (García et al., 2008). In consequence, both genomic features, COG distribution and scaling exponent, may be caused by equivalent factors and this fact could explain the high correlation between DFAs and COGs.

Functional categories associated with environmental interactions (e.g., energy metabolism, transport and regulation) were found to be the most variable among bacteria with different lifestyles. Thus, expansions and contractions in the genomic repertoire have mostly affected genes

involved in environmental interactions. In turn, basic information processes such as transcription and translation were distributed more homogeneously (Boussau et al., 2004).

An example of a selective advantage for survival in a specific habitat by means of the variation on the percentage of COG categories in three strains of *Frankia* has been recently detailed (Sen et al., 2008). Thus, the increase in number of predicted highly expressed genes in several COG groups may increase the ability of some strains to compete and survive in new habitats. For instance, the high levels of predicted highly expressed genes in transcription (K) and signal transduction mechanisms (T) would be advantageous by increasing its ability to respond to signals and regulate gene expression. Regarding the relationship between COG and lifestyle, comparative analysis of the genomes of Archaea and Bacteria have revealed that some ecological groups presented specific COGs that were absent in the rest of the groups. Some of these groups included thermophiles (Koonin et al., 2000) and hyperthermophiles (Makarova et al., 2003; Omelchenko et al., 2005). Another example of expansion-contraction of specific gene families with depth was previously reported: the shallow bathytypes were enriched in genes for energy production and conversion, while the deep bathytypes had a higher percentage of genes involved in cell motility and secretion, intracellular trafficking, secretion, translation, ribosomal structure and DNA replication and repair (Simonato et al., 2006).

Since we observed that microorganisms living in extreme habitats own high scaling exponents and heterogeneity in the percentage of functional genes, these two features could be essential requirements for genetic adaptation to extreme habitats. On the contrary, the microorganisms adapted to non-extreme environments seemed to share homogenously its genes into different functional groups. This fact seems to be in concordance with the relation between the gene family size and the heterogeneity of the habitat that has been previously pointed out (Pushker et al., 2004). It was suggested that bacteria with reduced gene family size might be adapted to homogeneous and non-extremophile habitat like the marine environment, in contrast to other extended gene family size of bacteria which have the ability to survive in much more heterogeneous habitats, such as soil. Moreover, the correlation between extremophiles and high scaling exponents was shown to be consistent (García et al., 2008).

Regarding the relationship between COGs composition and phylogeny, two different clusters were formed in Fig. 3.3A belonging to Archaea and Bacteria. Archaea appeared to have a higher genomic portion devoted to energy production and conversion, coenzyme metabolism, and poorly characterized categories than Bacteria. Nevertheless, Archaea also had relatively fewer genes involved in carbohydrate transport and metabolism, cell envelope and membrane biogenesis, and inorganic ion transport and metabolism (Konstantinidis et al., 2004).

Canonical analysis

The highest first canonical variable achieved by CCA was, logically, the corresponded to the correlation between DFA and genometry, since DFA depends directly on the percentage of the different nucleotides forming the DNA sequence. Moreover, DFA also showed high canonical correlation with genomic and phylogenic factors. This can be explained by the fact that DFAs value are related with the structure of the sequence and the genomic category includes some variables related also with the structure of the genome as number of proteins, coding sequence mean or gene density. The canonical correlation between phylogeny and DFA could be related by the capacity of DFA to capture phylogenetic signals (García et al., 2008).

The poor correlation between DFA and ecology is probably due to the heterogeneous nature of the variables included in the ecology category. This result could be explained by the fact that we added to the ecological category some factors that were not correlated with scaling exponents (e.g., biofilm formation or aerobic/anaerobic respiration). Nevertheless, the rest of factors (salinity, habitat, temperature, metabolism and pH) were somehow related with scaling exponents as discussed previously (García et al., 2008). Note the high values, reported in results section, achieved by discriminant analysis for all the ecologic clusters analyzed in the different canonical axes.

The variance of DFAs explained by the different categories, obtained by the RDA method, was in concordance with the canonical correlations. Furthermore, all the variables together accounted for a high variance of DFAs, contributing new information to demonstrate that the scaling

exponent feature of prokaryotic genomes is related with the genometry, the phylogeny and the ecology of the microorganism.

RY DFA

As showed in Fig. 3.2A and 3.2B, scaling exponent was more efficient discriminating phylogenetic and ecologic clusters when grouping the nucleotides that constitute the chromosomes in purine and pyrimidine (RY) than when grouping the nucleotides by its number of hydrogen bonds used to bind the base pairs or when grouping the bases by its keto or amino forms. This data highlights the importance of the heterocyclic aromatic structure to find phylogenetic and ecological signals inside the genome. From the biochemical point of view purine and pyrimidine are connected with most important properties of nucleotides (Akberova and Yu, 1996). Furthermore, RY has been detailed as a precise rule both, to reveal the replication origins of the prokaryotic chromosome and to study the evolution of DNA sequences as a distance measure (Akberova and Yu, 1996). Additionally, RY rule are more likely than the SW to be consistent with evolution under stationary, reversible and homogeneous conditions (Ho et al., 2006).

The exceptional RY DFA values presented by the bacterial phyla Aquificae, Fusobacteria and Thermotogae in Fig. 3.2C, similar to those from Archaea, can be explained by the fact that the three phyla form a monophyletic clade closely rooted to Archaea. Moreover, Aquificae phylum is a diverse collection of bacteria that live in harsh environmental settings like hot springs, sulfur pools and thermal oceanic vents sharing habitat with some archaea. Some of its members have an OGT between 85 to 95 °C. Thermotoga are also found in extreme environments, they are thermophiles and hyperthermophiles bacteria and some species present salt and oxygen tolerance. Fusobacteria contains only the genus Fusobacterium and its high scaling exponent could be a consequence of its unusual low GC content (26 – 34%), since both parameters are highly correlated as showed in Fig. 2.4. See Chapter II for more details.

Overall, the canonical analysis highlighted the intrinsic correlation between the structure of the prokaryotic genomes and its functionality summarized by the scaling exponent captured by the DFA and the

percentage of COGs, respectively. This work aims at helping for a better understanding the genetic adaptations to specific environments by means of the variation of COG percentages relative to the adaptation to a specific habitat and lifestyle. Nevertheless, the translation of the obtained results by this approach to their biological significance was not straightforward.

IV.

Whole genome comparison of the hydrocarbon-degrading bacteria *Alcanivorax borkumensis* and *Oleispira antarctica*

IV Whole genome comparison of the hydrocarbon-degrading bacteria *Alcanivorax borkumensis* and *Oleispira antarctica*

4.1 Introduction

Several millions of tons of oil are discharged into oceans every year. Some of it seeps from natural oil fields, but the bulk of the discharge comes as a result of anthropogenic activities. Maintenance of sustainable marine and coastal ecosystems requires the development of effective measures to reduce oil pollution and mitigate its environmental impact.

Hydrocarbon-degrading microorganisms usually exist in very low abundance in the absence of oil pollution. A pollution event is rapidly followed by a bloom of these microorganisms, the populations of which expand to nearly complete dominance of the viable microbial community during the period of contamination (Margesin and Schinner, 1999; Harayama et al., 1999). The properties of hydrocarbon compounds depend on the ambient temperature. Short-chain alkanes become less volatile and more water-soluble at low temperatures, whereas longer-chain compounds precipitate under cold conditions as waxes rendering them bioavailable and inaccessible to microbes, respectively. Such behaviour obviously reflects the establishment of specific oil-based marine microbial communities at low temperatures that are somehow different from those observed in a temperate climate. The most important permanently cold habitat is the ocean, since the temperature of more than 90% of the seawater volume is below 5 ºC. Recently several genera and

families of hydrocarbonoclastic microorganisms have been described within the γ-proteobacteria such as *Alcanivorax* (Yakimov et al., 1998), *Cycloclasticus* (Dyksterhouse et al., 1995), *Marinobacter* (Gauthier et al., 1992), *Oleiphilus* (Golyshin et al., 2002) and *Oleispira* (Yakimov et al., 2003). The first hydrocarbonoclastic genomes completely sequenced and annotated were the 3,120,143 bp of *Alcanivorax borkumensis* (Schneiker et al., 2006) and the 4,406,383 bp of *Oleospira antarctica* (unpublished data).

Alcanivorax borkumensis is a rod-shaped marine mesophilic gram-negative γ-proteobacterium that uses exclusively hydrocarbons as sources of carbon and energy. Its ubiquity and unusual physiology indicate that it is pivotal in the removal of hydrocarbons from polluted marine systems. It is found in low numbers in all oceans of the world and in high numbers in oil-contaminated waters community (Harayama et al., 1999; Kasai et al., 2002). *Alcanivorax borkumensis* is thus a paradigm of cosmopolitan hydrocarbonoclastic bacterium.

Oleospira antarctica is a marine psychrotrophic γ-proteobacterium that uses petroleum oil hydrocarbons as sources of carbon and energy. It is found in marine systems of all geographical areas and at all depths. The presence in unpolluted environments is poor, but becomes the dominant microbe in oil-polluted waters. Two strains, RB-8T and RB-9, were isolated from hydrocarbon degrading enrichment cultures obtained from Antarctic coastal marine environments (Rod Bay, Ross Sea). These bacteria, which form a monophyletic line within the γ-proteobacteria, are aerobic, gram-negative, have polar flagella and an optimal growth temperature (Topt) between 2–4 ºC. The isolates share many traits with the recently described genera of marine hydrocarbonoclastic bacteria *Alcanivorax*, *Marinobacter* and *Oleiphilus*, including isolation from a marine environment, purely respiratory metabolism (i.e. lack of fermentative metabolism), and relatively restricted nutritional profiles with a strong preference for aliphatic hydrocarbons.

In this paper an extensive work of comparative genomics between *Alcanivorax borkumensis* and *Oleospira antarctica* is presented to gain insight into the basis of its hydrocarbonoclastic metabolisms, marine lifestyle, its genomic responses to environmental stresses, the ability to degrade a range of oil hydrocarbons and its competitive advantage in oil-polluted environments. The comparison was carried out by means of

both, genometric methods such as GC skew, DNA walk or Detrended Fluctuation Analysis (DFA) and comparative analysis of the genes contained in each genome.

Genome GC content is an integrative parameter that has been explored by comparative analyses offering interesting information (Muto and Osawa, 1987; Hurst and Merchant, 2001; Marashi and Ghalanbor, 2004; Foerstner et al., 2005; Musto et al., 2006). However, DNA is predicted to contain more structural information than would be expected from base composition alone (Pedersen et al., 2000). One of the main features of a DNA sequence related to the whole genome structural composition is the long-range correlation, a scale invariant property of DNA. In a correlated sequence, occurrence of a nucleotide in a specific position depends on the previous nucleotides (memory). The long-range correlation is related directly to the fractal structure of the DNA sequence or self-similarity. Peng et al. (1992) studied correlation properties in DNA sequences using a fractal landscape or DNA walk model. DNA walking is a genometric method based on a derivative function of the sequential position for each nucleotide along a DNA sequence. The resulting "walk" can be projected on a two-dimensional plot representative of the DNA "landscape" and enables the simultaneous comparison among different genome landscapes (Lobry, 1999). Additionally, one of the most appropriated methods proposed in recent years for the study of long-range correlations in genomes is the DFA (Peng et al., 1992 and 1994). DFA is a scaling analysis method providing a single quantitative parameter —the scaling exponent α— to represent correlation properties of a sequence and the characteristic length scale of repetitive patterns. DFA takes into account differences in local nucleotide content (heterogeneity) and can be applied to the entire sequence. It shows linear behavior in log–log plots for all length scales, and the long-range correlation property is characterized by the scaling exponent (α), i.e., the log–log slope. One of the already shown potentials of DFA is a change in the quantification of genome complexity with evolution (Peng et al., 1995). Thus, links between long-range correlations and higher order structure of the DNA molecule have been suggested (Grosberg et al., 1993). Combination of DNA walk and DFA methods has been demonstrated to help to decipher the biological significance of long-range correlations in microbial genomes and the influence of lifestyle in the DNA structure. The specific patterns and long-range correlations were related to phylogenetic, ecological and metabolic information (García et al., 2008). Furthermore, genes

contained in genomes provide essential information for understanding evolutionary relationships and ecological adaptations in microorganisms.

The goal of this comparative and functional genomics project is hence to characterize the genomic basis of the unusual ecophysiological features and environmentally significant properties of these two microorganisms, and to establish a knowledge base that may help to a better understanding of the influence of temperature on the growth of oil-degrading bacteria and predicting cellular phenotypes.

4.2 Methods

Complete sequenced genome and genomic annotation from *A. borkumensis* were downloaded from GenBank (National Center for Biotechnology Information, www.ncbi.nlm.nih.gov/Genbank), while the sequenced and annotated genome from *O. antarctica* were provided by the Environmental Microbiology department from the Helmholtz Centre for Infection Research for comparative purposes. The information was complemented with biochemical information and cell physiology data in order to complete the knowledge of the whole metabolism of the microorganism.

GC skew and DNA walk

The GC skew (Lobry, 1996b) gives a measure of the deviations from the base frequencies A = T and G = C. It is usually stronger than the AT skew, so we will focus on the GC skew from both genomes calculated as the ratio (C – G)/(C + G), which gives the percentage of excess of C over G.

We also analyzed the distribution of individual nucleotides along the genomes by the DNA walk method (Appendix A). Here, we have used four types of representations. First, we translate the original nucleotide sequence into a one-dimensional numerical series grouping the bases in pairs following the three types of rules (KM, RY and SW). Then, the

resulting DNA walk series for each rule were mapped onto an orthogonal plane. The slopes of the regression lines from the SW DNA walk were used as variables for subsequent analysis. For the last representation, we performed a two-dimensional (2D) map where each nucleotide defines one direction in a plane formed by two orthogonal axes (i.e., C versus G and T versus A). (see Chapters II and III for more details).

Detrended Fluctuation Analysis (DFA)

Detrended Fluctuation Analysis (Appendix B) was used to calculate the scaling exponents from the four types of DNA walks. On the one hand, scaling exponents were calculated directly from one-dimensional DNA walks (KM, RY and SW). On the other hand, the two-dimensional (2D) series of DNA walks were transformed into one-dimensional ones by replacing every original x–y point, representing a step of the walk, with its Euclidean distances from the origin of the graph. The resulting one-dimensional series were then used to calculate the scaling exponents.

Genome alignment

A whole-genome alignment was performed using MUMmer, a software developed at the Institute for Genomic Research (Delcher et al., 1999; Kurtz et al., 2004).

Genome visualization

Dot-plotting is the best way to see all of the structures in common between two sequences. Dot plot was calculated using the software Gepard (GEnome PAir Rapid Dotter) (Krumsiek et al., 2007), which allows the calculation of dot plots even for large sequences like bacterial genomes.

Whole genome comparison of *A. borkumensis* and *O. antarctica*

Visualization of inter-chromosomal relationship was performed using the software Circos (www.bcgsc.ca) which is open-source Perl application particularly suited for visualizing alignments, conservation and intra- and inter-chromosomal relationships.

4.3 Results

Genomic features

The general features of the *A. borkumensis* and *O. antarctica* genomes are listed in Table 4.1. The *A. borkumensis* genome is composed of one circular chromosome of 3.12 Mb harboring 3073 annotated genes, 513 of them with predicted metabolic functions, while the remaining genes (2560) were either non-metabolic genes or designated as hypothetical proteins or proteins of unknown function. In turn, the *O. antarctica* consists of one single chromosome of 4.41 Mb and encodes 3986 annotated genes, 687 of which were predicted as metabolic genes. The rest of the genes (3299) were either non-metabolic genes or designated as hypothetical proteins or proteins of unknown function. The G+C content is higher in *A. borkumensis* (54.73%) than in *O. antarctica* (42.16%).

For each genome we run the GC skew, the keto–amino (KM), purine–pyrimidine (RY) and strong–weak (SW) pairing DNA walk, and a two-dimensional (2D) DNA walk (see methods). The GC skew showed strong strand bias that resulted in symmetric plots (Fig. 4.1, panels A and B) which are characterized by a preference for G and C over T and A in the leading strand and vice versa in the lagging strand in both microorganisms. The inflection point of the GC skew coincides with the terminus of replication. *A. borkumensis* is likely to be more persistent in the enrichment of its respective bases in both strands since its plot seems to be smoother than the plot of *O. antarctica*. The RY (Fig. 4.1, panels C and D) and KM (data not shown) DNA walks also resulted in symmetric plots with preference for G over C in the leading strand and C over G in the lagging strand. The inflection point of both DNA walks also coincides with the terminus of replication. In agreement with the GC skew, *O. antarctica* DNA walks are rougher than *A. borkumensis* walks, indicating a greater persistence in the bases of the last microorganism.

The SW walks for both microorganisms were well fitted by a linear regression with positive and negative slopes for *A. borkumensis* and *O. Antarctica*, respectively. The strong strand bias and the persistence of the bases on both genomes were also showed by the 2D DNA walk (Fig. 4.1, panels E and F). Thus, *A. borkumensis* presented an enrichment of Gs in the leading strand and Cs in the lagging strand while *O. antarctica* showed a preference of Ts in the leading strand and As in the lagging strand. The complete set of DNA walk plots are attached in the supplementary file 4.1.

	A. borkumensis	*O. antarctica*
Size (bp)	3120143	4406383
Annotated genes	3073	3986
Metabolic genes	513	687
Non-metabolic genes	2560	3299
Metabolic processes	490	667
ORF	365	399
G+C (%)	54,73	42,16
(G – C)/(G + C)	0,00103	–0,00041
SW DNA walk slope	0.093	–0.159
A (%)	22,63	28,90
T (%)	22,64	28,94
G (%)	27,39	21,07
C (%)	27,34	21,09
α KM	0,660	0,691
α RY	0,610	0,644
α SW	0,661	0,667
α 2D	0.5779	0.6132
Opt. Temp. (ºC)	20–30	2–4
NaCl (%)	3–10	3–5

Table 4.1. General genomic features and optima of temperature and sodium chloride concentrations for *A. borkumensis* and *O. antarctica.*

Whole genome comparison of *A. borkumensis* and *O. antarctica*

The four different DNA walks were translated into a numerical series (see methods) for running the DFA. The resulting curves showed scaling exponents within $\alpha = 0.58$ (2D DFA *A. borkumensis*) and $\alpha = 0.69$ (KM DFA *O. antarctica*) (Table 4.1). In all the cases, DFA scaling exponents were higher than 0.5, indicating persistent long-range correlations. We found that the genome of *O. antarctica* had, on average, higher DFA scaling exponents than the strong strand-biased genome of *A. borkumensis*. Moreover, we also noticed a negative correlation between scaling exponents and GC content. The complete set of DFA plots is included in the supplementary file 4.2.

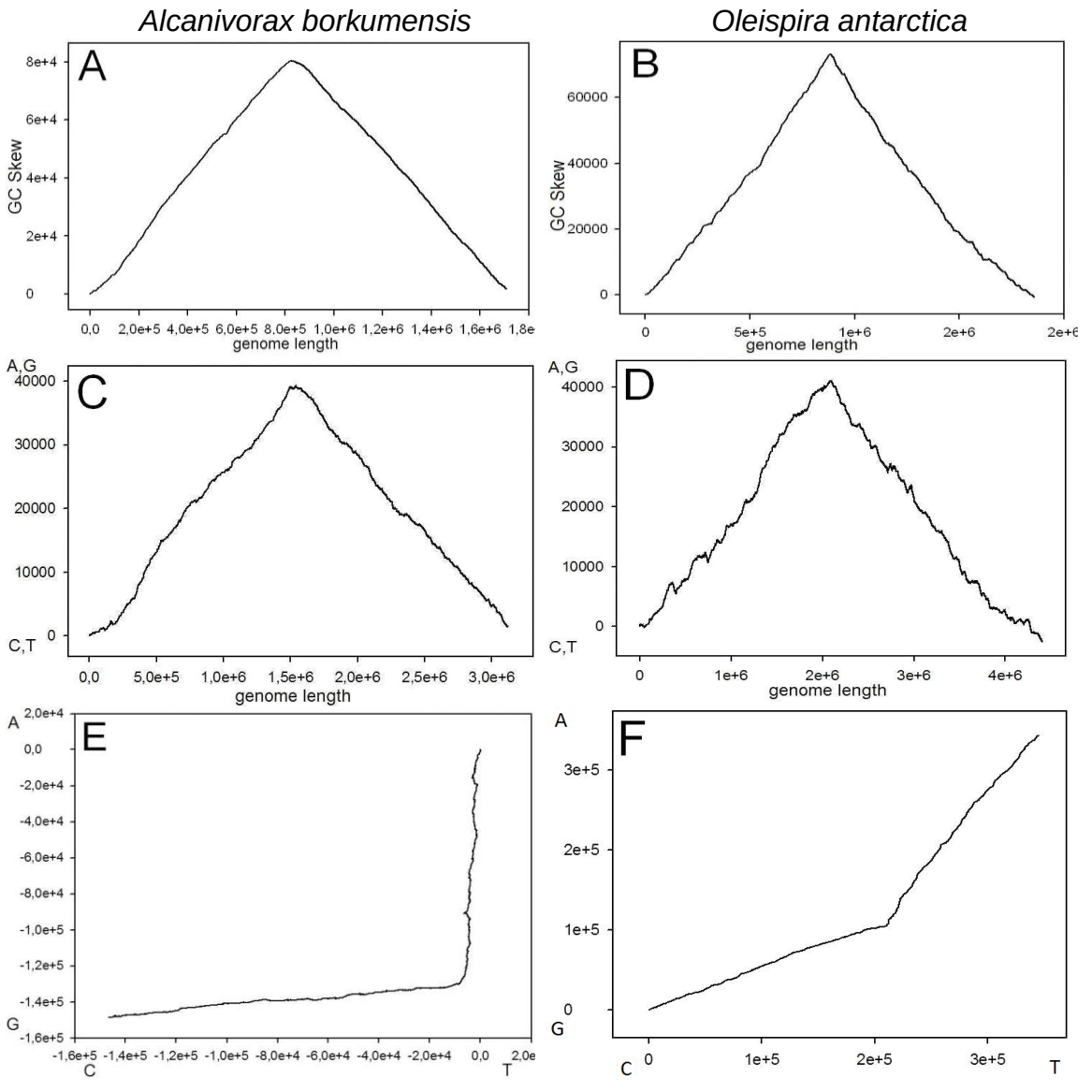


Figure 4.1. GC skew calculated as the ratio [(G – C)/(G + C)] (panels A,B), DNA walk samples using the purine–pyrimidine (RY) rule (panels C,D) and 2D DNA walks (panels E,F) for the genomes of *A. borkumensis* (left) and *O. antarctica* (right), respectively. In GC skew and RY DNA walks, abscissa represents the genomic sequence position from the beginning to the end of the genome. Not only base composition is not randomly distributed in the genomes but strong strand bias resulted in symmetric plots which are characterized by a preference for G and C over T and A in the leading strand and vice versa in the lagging strand. Note the *O. antarctica* skew and

RY walks are rougher than *A. borkumensis* ones. The inflection point of the GC skew and DNA walks coincides approximately with the terminus of replication. 2D DNA walks showed an enrichment of Gs in the leading strand and Cs in the lagging strand of *A. borkumensis* while *O. antarctica* showed a preference of Ts in the leading strand and As in the lagging strand.

Regarding the ecological lifestyle, we plotted both genomes into a previous published set of data (Chapter II) were some microbial groups with a wide range of temperature optima clustered separately according to the values of 2D DFA and SW slope (Fig. 4.2). For instance, looking at the Topt, *O. antarctica* joined, as expected, to the psychrophiles cluster while *A. borkumensis* was closer to moderate thermophiles than to psychrophiles.

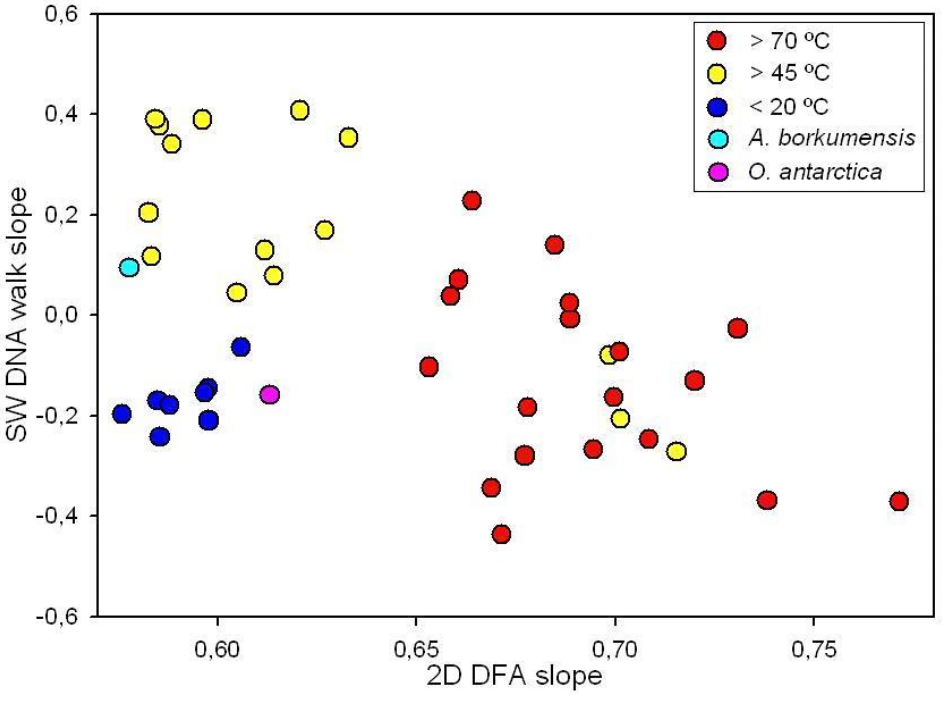


Figure 4.2. Ecological clusters formed after the SW DNA walk slope and the 2D DFA scaling exponent analyses were plotted in combination. *O. antarctica* was located within the psychrophiles whereas *A. borkumensis* was closer to the thermophiles than the psychrophiles.

Direct visual comparison of the whole genomes was performed by a DNA dot plot (Fig. 4.3), which shows the regions of close similarity between both genomes. Two discontinuous diagonals can be distinguished. The main diagonal starting at origin represents the direct matches between both sequences, whereas the reverse diagonal indicates partially palindromic areas, due to transposition of fragments from a specific location in genome to another in such a way that both locations are equidistant from the origin of replication. The remaining points off the main diagonals represent partial deletions, insertions or repetitive patterns between the sequences.

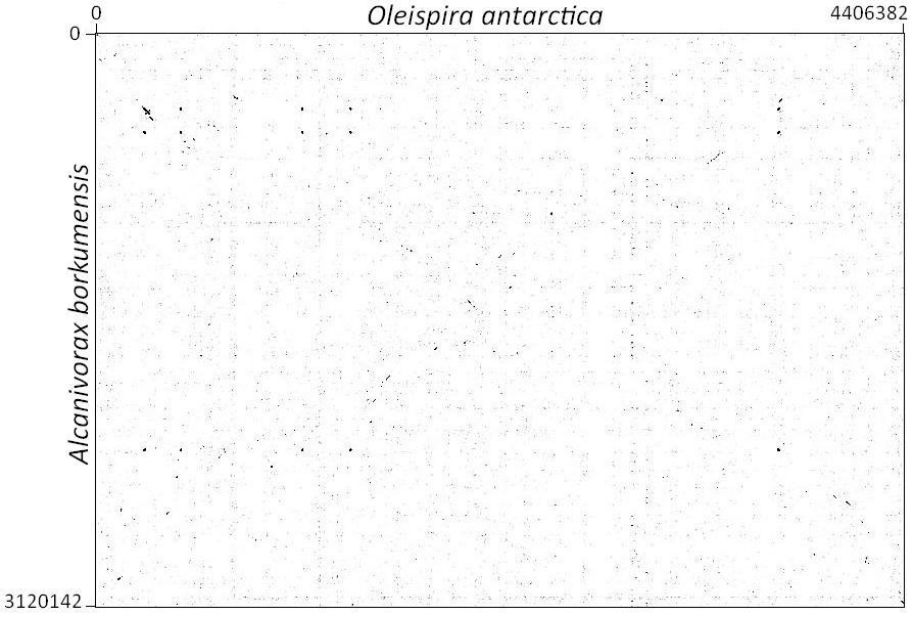


Figure 4.3. Dot plot of *A. borkumensis* plotted against *O. antarctica* generated with Gepard software.

Metabolic and functional comparative genomics

The distribution and arrangement of the metabolic genes along the whole genomes were analyzed by means of the annotation data. For instance, Fig. 4.4 shows the location of the genes involved in amino acids, nucleotides, cofactors and vitamins metabolisms. In agreement with the dot plot pattern, most of the genes were located over the two diagonals of the plot. The same distribution pattern was found for the rest metabolic genes (alkane degradation, carbohydrates, lipids, glycans and energy metabolism) (data not shown). The diagrams in Fig. 4.5 were generated by Circos software and compare the localization of genes involved in six different metabolic pathways (amino acids, carbohydrates, cofactors and vitamins, glycans, lipids and nucleotides metabolism) from both genomes. It can be appreciated not only the genes that have conserved a similar localization in both genomes but also the genes that have suffered a rearrangement due to translocations, duplications, deletions or lateral transfers. All the diagrams clearly show that rearrangements have happened in genes from all the metabolic pathways.

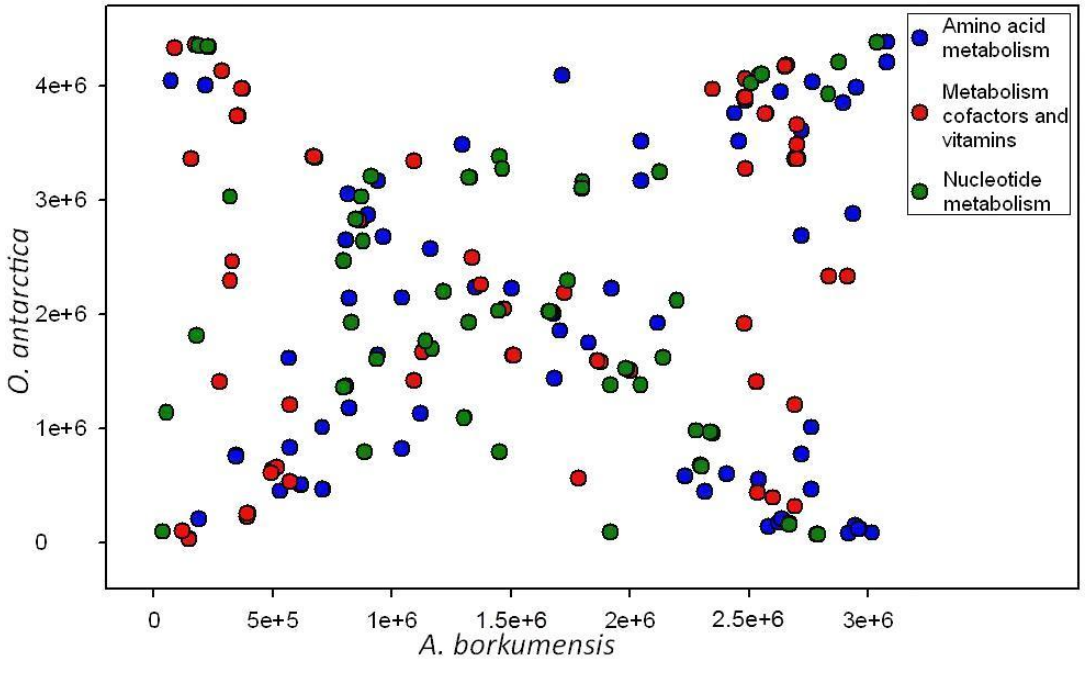


Figure 4.4. Dot plot of several metabolic genes of *A. borkumensis* plotted against the same genes of *O. antarctica*. Most of amino acid, cofactors, vitamins and nucleotide metabolisms are placed in the direct or reverse diagonal.

Whole genome comparison of *A. borkumensis* and *O. antarctica*

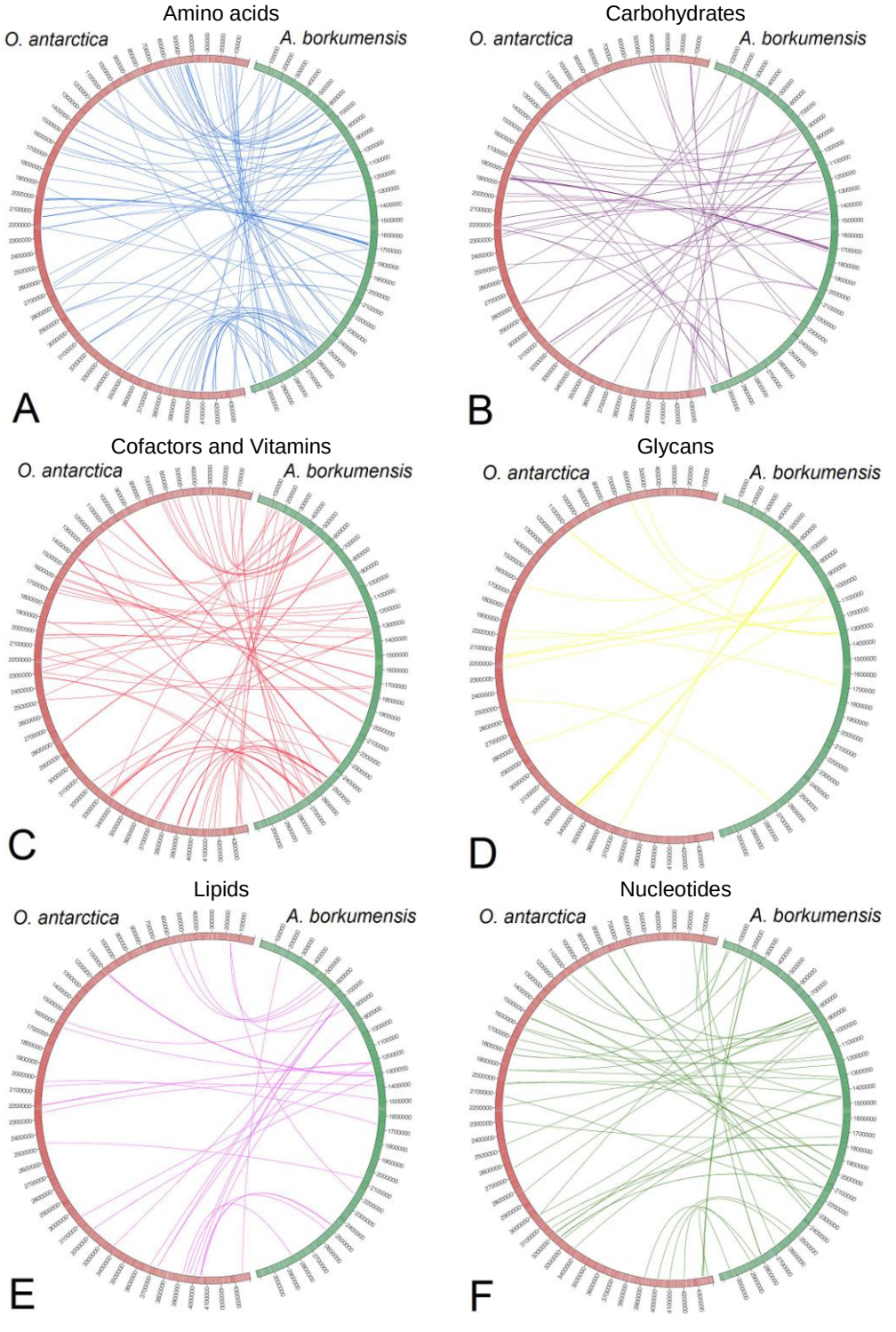


Figure 4.5. Circos representation of *A. borkumensis* (green) and *O. antarctica* (red) chromosomes and the position of their genes involved in several metabolic pathways. The pathways represented are amino acid metabolism and biosynthesis (A); carbohydrate metabolism, which includes central metabolism (B); metabolism of cofactors and vitamins (C); glycan biosynthesis and metabolism (D); lipid metabolism (E) and nucleotide metabolism (F).

Furthermore, classification of the metabolic genes into various pathways reveals that the genomes of these two bacteria were highly conserved with respect to the percentage of genes dedicated to the basic metabolic functions (Fig. 4.6). The highest percentages of metabolic genes in both bacteria were the devoted to amino acid biosynthesis and metabolism, metabolism of cofactors and vitamins, nucleotide, carbohydrate and lipid metabolism. The major difference in percentages between bacteria was the related with the nucleotide metabolism (12% in *A. borkumensis* and 18% in *O. antarctica*) followed by metabolism of cofactors and vitamins (17% versus 21%), lipid metabolism (13% versus 9%) and transport (6% versus 3%). Regarding the unique and shared metabolic ORFs, Table 4.2 and Fig. 4.7 show their distribution into the functional metabolic pathways. In general, *O. antarctica* appears to have more quantity of unique ORFs than *A. borkumensis* in concordance with its longer size. On the one hand, the pathways with the higher percentages of shared genes were the amino acid metabolism (71.34%), metabolism of cofactors and vitamins (66.67%), carbohydrate metabolism (62.26%), energy metabolism (100%), and alkane degradation (75%). On the other hand, the smaller percentages of shared genes were found in the nucleotide metabolism (47.15%), lipid metabolism (49.48%), glycan metabolism (34.38%) and transport (37.78%). Some remarkable values regarding *A. borkumensis* were the low percentage of unique genes in the amino acid and nucleotide metabolisms (2.55% and 2.44% respectively), as well as the inexistence of unique genes in metabolism of cofactors and vitamins. Moreover, significant scores were the 50.41% of unique nucleotide metabolic related genes in *O. antarctica* and the complete coincidence of genes involved in the energy metabolism between both genomes.

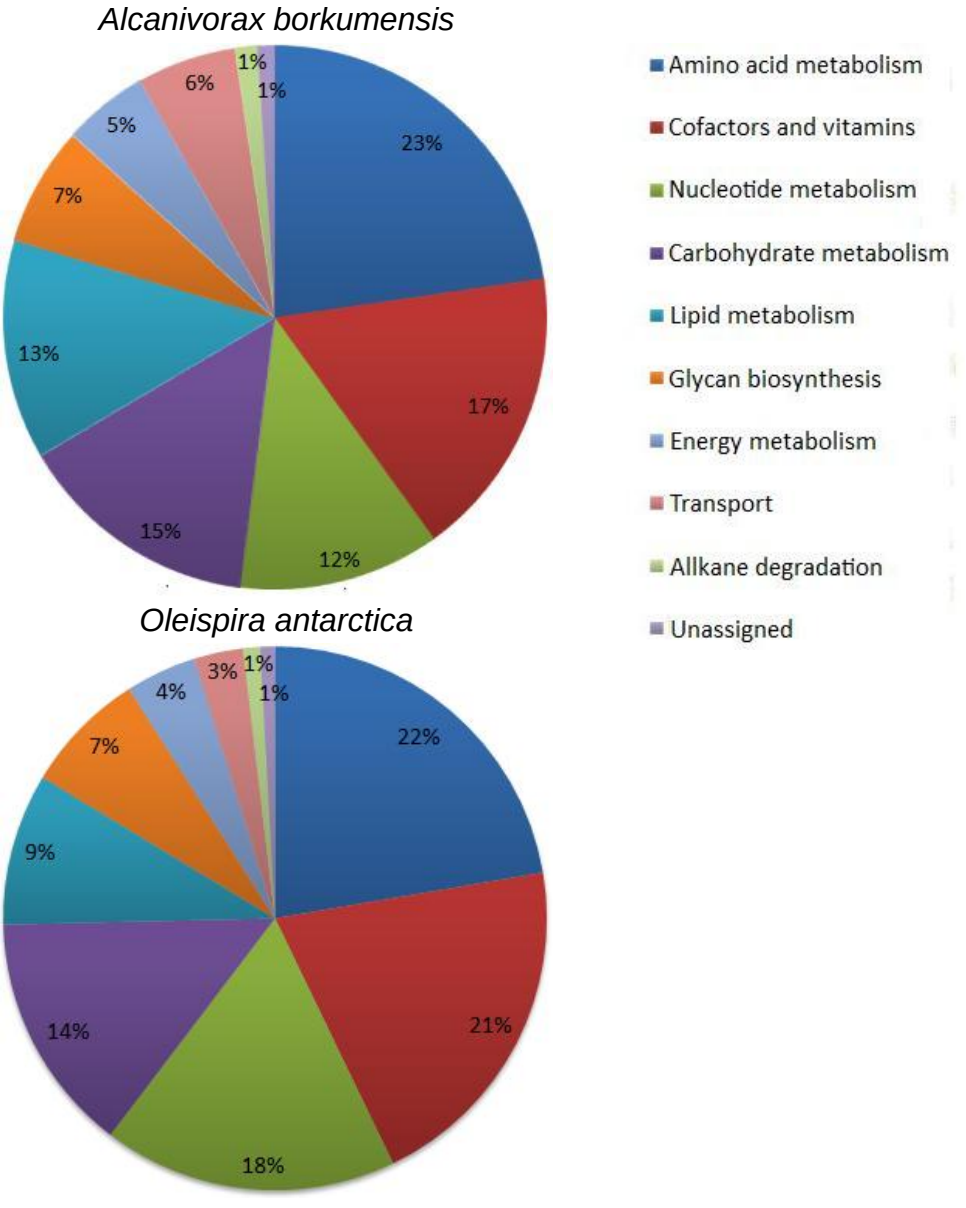


Figure 4.6. Distribution of the metabolic genes involved in the principal pathways of *A. borkumensis* and *O. antarctica*.

Metabolic pathway	*A. borkumensis*	Shared	*O. antarctica*
Amino acid metabolism	4	112	41
Cofactors and vitamins	0	94	47
Nucleotide metabolism	3	58	62
Carbohydrate metabolism	8	66	32
Lipid metabolism	19	48	30
Glycan biosynthesis	14	22	28
Energy metabolism	0	31	0
Transport	13	17	15
Alkane degradation	1	3	0
Unassigned	1	4	2

Table 4.2. Distribution of specific genes of *A. borkumensis* (first column), metabolic genes shared by *A. borkumensis* and *O. Antarctica* (second column) and specific genes of *O. Antarctica* (third column). The genes were classified according to its metabolic pathway.

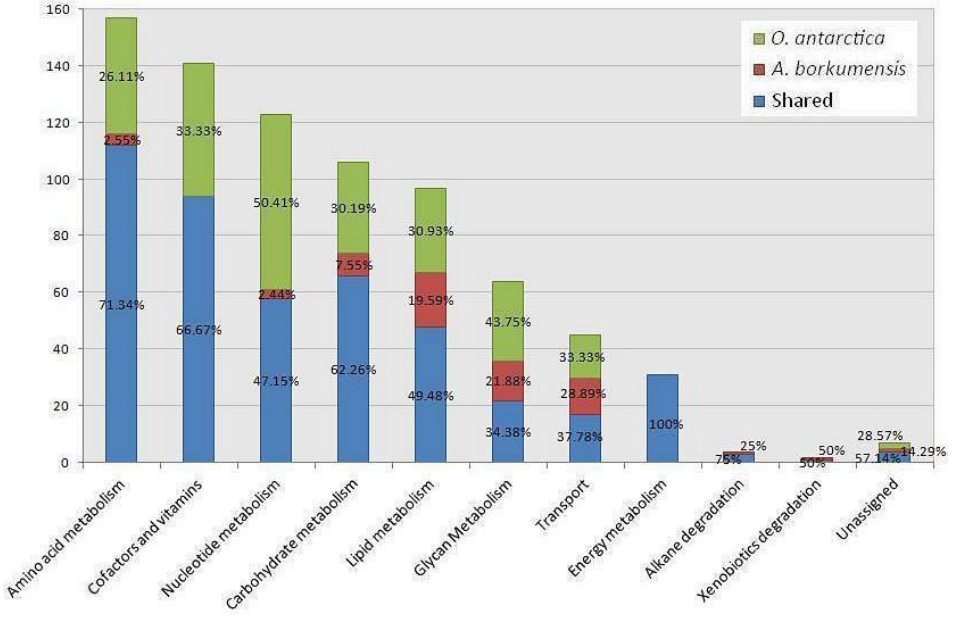


Figure 4.7. Distribution and percentage of specific genes of *A. borkumensis* (red), metabolic genes shared by *A. borkumensis* and *O. antarctica* (blue) and specific genes of *O. Antarctica* (green) over the principal pathways.

In addition, the distribution of the whole genes in various functional categories was also highly conserved in both bacteria (Fig. 4.8). In general, percentages of functionally assigned genes showed high identities. The largest functional categories corresponded to the oxidoreductases, transferases, hydrolases and transport. Furthermore, the greater difference in percentages between both bacteria was the related with the oxidoreductases and transport. Concerning the unique and shared functional ORFs, Table 4.3 and Fig. 4.9 show their distribution in both genomes. As expected, *O. antarctica* had more unique genes than *A. borkumensis*. Furthermore, the percentage of shared genes was low in all the categories. This was accentuated in the ORFs devoted to transport (7%). However, ribosomal RNAs and ligases presented significant more shared genes than the average (73.3% and 49.5% respectively). Another remarkable percentage was the low unique ribosomal RNAs found in *A. borkumensis* (2.7%). On the contrary, its ORFs related with the membrane proteins with no counterparts in *O. antarctica* were significantly high (61.5%). Finally, *O. antarctica* had a great number of unique sensors (83.9%) and transcriptional regulators (65.1%).

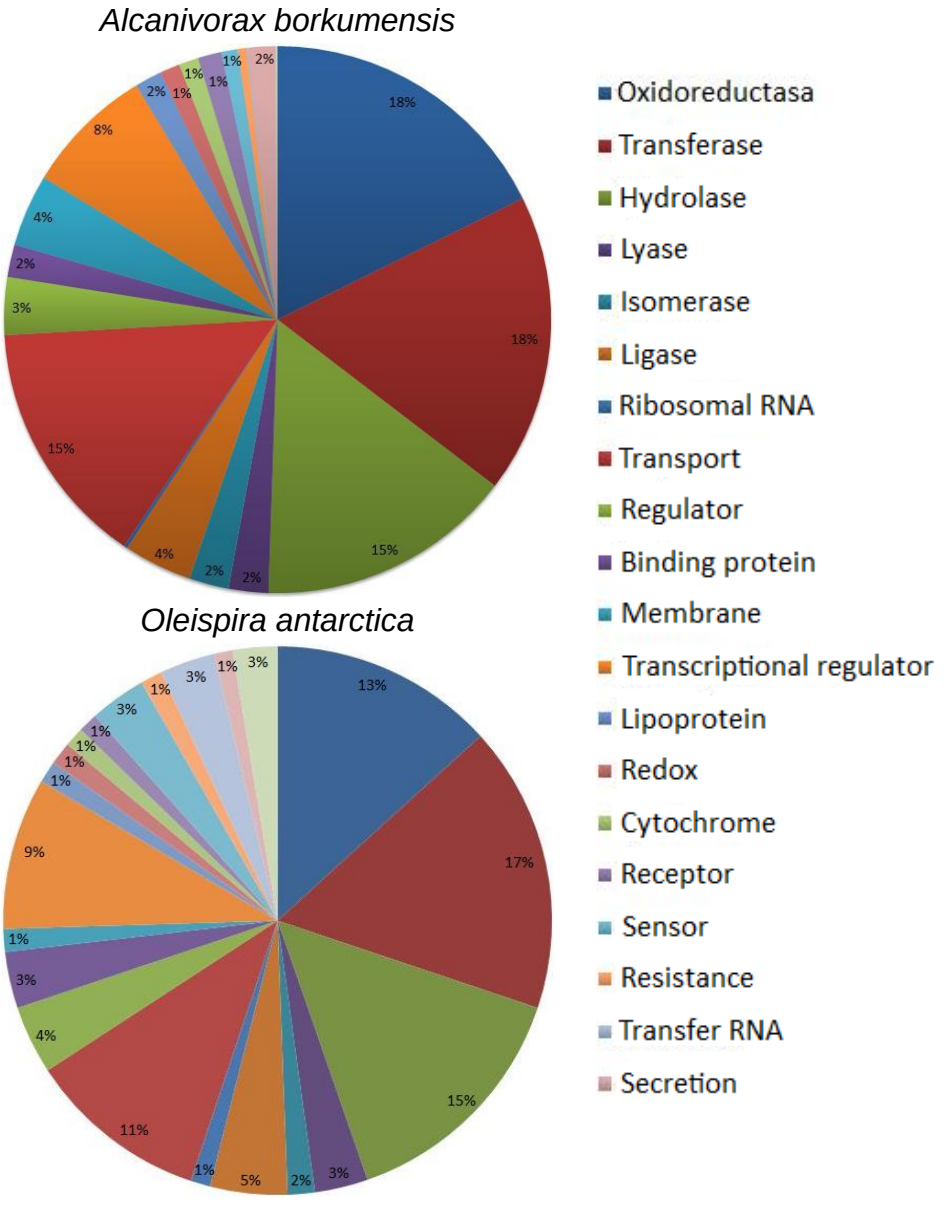


Figure 4.8. Functional classification and distribution of genes with known function in *A. borkumensis* and *O. antarctica.*

	A. borkumensis	Shared	*O. antarctica*
Oxidoreductases	166	77	207
Transferases	166	118	264
Hydrolases	142	84	227
Lyases	22	36	49
Isomerases	22	24	25
Ligases	38	107	71
Ribosomal RNA	2	55	18
Transport	137	23	167
Regulators	32	12	63
Binding proteins	18	11	52
Membrane	40	4	21
Transcription	72	3	140
Lipoproteins	15	3	20
Redox	11	2	21
Cytochromes	11	2	18
Receptors	13	1	18
Sensors	9	1	52
Resistance	5	1	20
Transfer RNA	1	1	51
Secretion	15	0	17
Flagel	1	0	41

Table 4.3. Distribution of unique or shared functional genes of *A. borkumensis* and *O. antarctica*. The genes were classified by functional pathways.

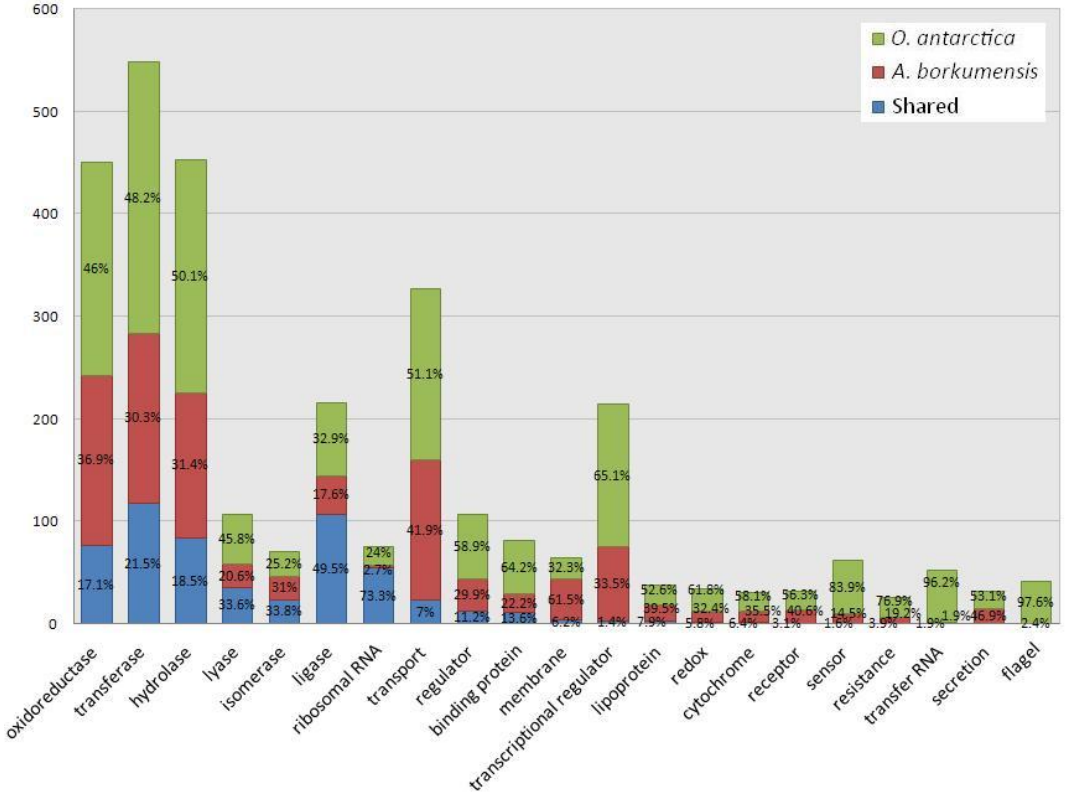


Figure 4.9. Distribution and percentage of unique and shared genes with known functions in *A. borkumensis* and *O. antarctica*.

4.4 Discussion

Genomic features

DNA walks and GC skews are characterized by the so-called mutational strand bias (Lobry, 1999). Many microorganisms show a preference for G over C and T over A in the leading strand and C over G in the lagging strand because of several factors including proofreading efficiencies for the different types of DNA polymerases (Rocha, 2002; Worning et al., 2006) or the differential mutation in both strands as the result of asymmetry inherent to the DNA replication mechanism. A simple model for explanation is based on the spontaneous deamination of cytosine that induces mutations from C to T. The rate of this deamination is highly increased in single leading strand during DNA replication. This, combined with natural selection, leads to an observed base distribution that depends in part on the mutational pattern and in part on selection. It has been reported that the rates of spontaneous mutation are greatly accelerated and the strand asymmetries are minimized as a consequence of adaptations to extreme environments like low temperatures (Lindahl, 1993). This could be a possible explanation for the weaker strand bias and the consequent higher scaling exponents or long-range correlation in *O. antarctica* compared with *A. borkumensis*. Moreover, the difference in genome size between both genomes may be due to processes, such as gene duplication and gene lateral transfer. Both events add repetitive structures with different lengths inside the genomes. The consequence of these processes in the genome may be both, an increase of the genome size and a decrease of the strand bias according to previous published results (García et al., 2008). Thus, it seems like *O. antarctica* was more affected by spontaneous mutation, duplication and insertion events than *A. borkumensis*. Additionally, *A. borkumensis* harbored a smaller number of mobile genetic elements such as transposons and insertions elements (Schneiker et al., 2006) which might also explain the differences in strand bias and genome size. This paucity of mobile elements may be a consequence of the counter selection against variants with increased numbers of mobile elements.

The GC contents of *A. borkumensis* (54.73%) and *O. antarctica* (42.16%) reached a considerable variation. Nevertheless, both genomes did not fit the reported assumption that large genomes have a tendency to be richer in GC due to the fact that random mutations are mainly from C

to T and from G to A and the lack of repair mechanism in reduced genomes (Heddi et al., 1998; Moran, 2002). This disagreement could be explained by some different causes. On the one hand, an unbalanced supply of the essential external precursors for nucleic acid synthesis and repair since both bacteria grow in different environment. While *A. borkumensis* is ubiquitous in marine environments, *O. antarctica* mainly grows in a specific habitat such as Antarctic Ocean. On the other hand, a nucleotide bias in the mutational mechanism caused by a stressed psychrophilic lifestyle of *O. antarctica* might be the cause of the GC variation. However, the relation between GC content and Topt was in agreement with the previous reported tendency of GC content to decrease with the temperature when comparing thermophilic and psychrophilic microorganisms (García et al., 2008). Furthermore, *O. antarctica* fell into the previously reported cluster of psychrophiles when GC content and scaling exponent values were taken into account (Fig. 4.2), whereas *A. borkumensis* was placed in a GC content transition area between thermophiles and psychrophiles. Thus, adaptation to low temperatures would imply a structural modification in microbial genomes concerning the GC content.

The phylogenetic relationship between both bacteria is summarized in Fig. 4.10 and in the dot plot from Fig. 4.3. Despite the relative phylogenetic proximity between both bacteria, *O. antarctica* forms a distinct phyletic line within the γ-proteobacteria, with less than 89.6% sequence identity in the 16S rRNA gene sequence to their closest relatives (Yakimov et al., 2003). The homologous sequences arose in a discontinuous diagonal in the dot plot from the origin of replication. This pattern indicates that both bacteria shared a common ancestor, although the high number of interruptions on the diagonal is consequence of a long time of independent evolution with continued rearrangements occurring after their divergence. The reverse diagonal refers to fragments that were located in opposite places inside each of the two genomes with the peculiarity that the distance from the origin of replication to each fragment was the same. This distribution may be possible thanks to arrangement events occurred during evolution.

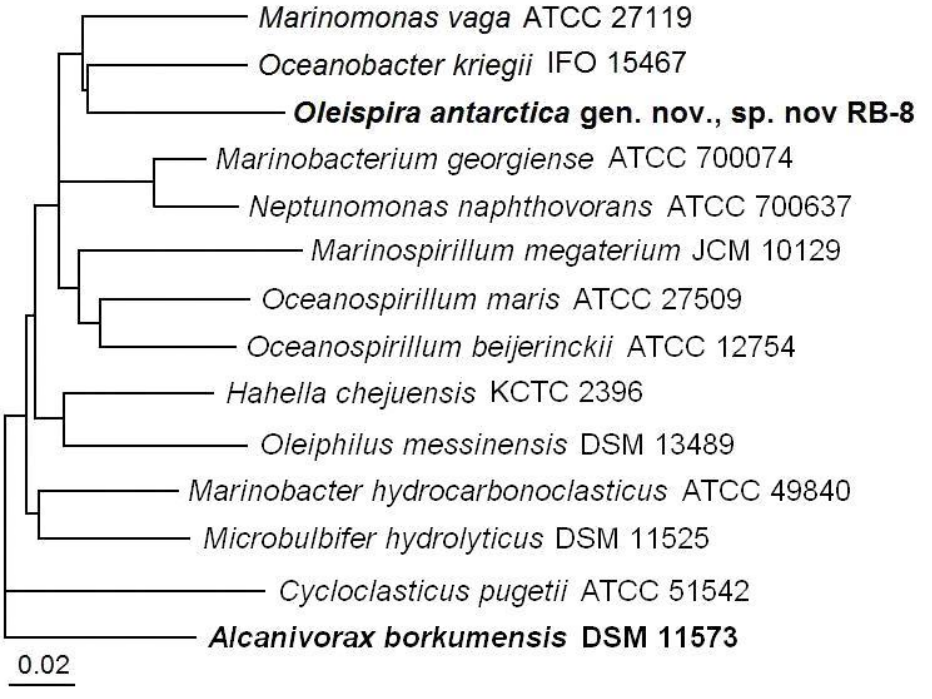


Figure 4.10. Estimated phylogenetic position of *A. borkumensis* and *O. antarctica* among the most closely related γ-proteobacteria, derived from 16S rRNA gene sequence comparisons. The three was constructed by the neighbour-joining method and nucleotide substitution rates and computed by Kimura's two-parameter model.

We concluded therefore that although the order of these fragments and its genes were not conserved in both bacteria, the quantity of proteins produced from the genes involved in these transpositions is similar due to its equidistant position from the origin. Although an important number of genes may be implicated in arrangement events, the quantity of proteins produced from them was highly conserved because the significant high number of points in the reverse diagonal as showed in Fig. 4.4 and 4.5.

Metabolic and functional comparative genomics

In concordance with its higher genome size, *O. antarctica* presented, in general, significant higher number of unique genes than *A. borkumensis*, many of which probably arose either via gene duplication or lateral transfer in order to adapt to extreme environmental conditions and may account for some of the observed physiological differences between the two strains. These unique genes might be excellent candidates to analyze the adaptation of bacteria to low temperatures. Furthermore, this result was also in agreement with the supposed higher genomic diversity of extremophilic microorganisms. The exceptions were the genes coding for membrane proteins, which were more abundant in *A. borkumensis* and the genes involved in energy metabolism that were equivalent in both microorganisms.

Regarding the distribution of metabolic genes in individual pathways, we found similar percentages between both bacteria (Fig. 4.6) and some differences in the shared and unique metabolic genes (Fig. 4.7). Thus, they appeared to have similar percentages of orthologs devoted to the carbohydrate metabolism and, as consequence, to the central metabolism as corresponded to related microorganisms. The amino acid metabolism was also conserved since both bacteria presented all of the *de novo* pathways to synthesize the 20 essential amino acids (Schneiker et al., 2006; Yakimov et al., 2003). Moreover, glycan and energy metabolism harbored similar percentages of genes in both genomes. In the case of glycan metabolism, the small amount of shared genes suggests a marked difference of lipopolysaccharides and pepdidoglycans in the composition of the cell envelope, whereas the entire energy metabolism pathways were shared by both bacteria, indicating a high level of conservation in the genes involved in pathways as oxidative phosphorilation and nitrogen metabolism. *A. borkumensis* and *O. antarctica* mainly differ from one another with respect to the metabolism of cofactors and vitamins, nucleotide and lipid metabolism and transport. Surprisingly, *A. borkumensis* lacked unique genes involved in cofactors and vitamins metabolism suggesting that these pathways were highly conserved in the strains. Furthermore, one third of the genes were unique in *O. antarctica* indicating that a psychrophilic environment is more exigent with respect to these components.

The nucleotide metabolism also showed a significant difference between genomes. Thus, less than a half of the genes were shared and most of them were unique in *O. antarctica*. The evident lack of genes of *A. borkumensis* involved in nucleotide metabolism could be compensated by its identified transporters of a wide variety of substrates which were suggested to be involved in the import of nucleic acid precursors, including ATP-Binding Cassette (ABC) and Major Facilitator Superfamily (MFS) primary active transporter (Pollack, 2002). Moreover, pyrimidine nucleotides may be produced by alternate routes from the corresponding nucleosides using a bifunctional enzyme —the cytidine/uridine kinase— and therefore, several genes could be removed without having loss of functionality.

As expected, lipid metabolism presented some differences between bacteria. The number of shared genes was less than a half and both genomes owned an important part of unique genes. Some explanations for this distribution could be that the genetic organization of the glucolipid biosynthesis still remains unclear in both bacteria. Moreover, psychrophilic marine bacteria modify their membrane lipids in order to adapt to low temperatures. Thus, psychrophiles are characterized by lipid cell membranes chemically resistant to the stiffening caused by extreme cold which protect their DNA even in temperatures below water's freezing point. It has been reported that the ratio of total unsaturated versus saturated fatty acids in the membrane lipids of some psychrophilic marine bacterial increased when the microorganism was grown at decreasing temperatures. This regulatory capacity appears to be functional in maintaining membrane fluidity at typical low sea temperatures. Particularly, *O. antarctica* modified the level of unsaturated fatty acids in order to achieve a homeostatic adaptation of the membrane in terms of its viscosity (Yakimov et al., 2003). Additionally, the modification of the phospholipid content in psychrophilic bacteria in response to a decreasing growth temperature has been described (Takada et al., 1991). For instance, phosphatidylethanolamine was replaced completely by a phosphoglycolipid and phosphatidylserine at low temperatures, neither of which was present at higher temperatures. In summary, the principal fatty acid in *A. borkumensis* were 16:0 and 18:0 (Yakimov et al., 1998), whereas in *O. antarctica* the most abundant were 16:0, 16:1 and 18:1 (Yakimov et al., 2003).

Both bacteria degrade a large range of alkanes. Presumably, the gene clusters that confer the assimilation of aliphatic hydrocarbons at *A. borkumensis*, are localized in two genome islands which were probably acquired from an ancestor of the *Yersinia* lineage (Reva et al., 2007). The alkane degradation pathway of *A. borkumensis* includes hydroxylases, oxidoreductases and dehydrogenases. However, the key enzymes of alkane catabolism, alkane hydroxylase and alkane monooxygenase were not detected in *O. antarctica*, which suggests a certain novelty in the structure of those enzymes (Yakimov et al., 2003). The significance of lacks will require further functional analyses.

Despite the lack of carbohydrate transporters in both strains, the genes devoted to transport showed a high specificity since the great number of unique genes. This result highlighted the important role played by transporters in the habitat adaptation. The differences were not only the quantity of transporters found in each genome, but also the substrate specificity. For instance, *A. borkumensis* encodes genes for a broad range of transport proteins, within them a high number of permeases. Many of the transport systems found in both bacteria were consistent with their marine lifestyle, like Na^+ pumps, sodium/protons antiporters as well as several Na^+ dependent symporters (Schneiker et al., 2006). Moreover, *A. borkumensis* presented transport systems that were not detected in *O. antarctica* like the system for the uptake of N and P mediated by a high-affinity ABC system or specific transport system for various oligoelements such as molybdenum, zinc, magnesium or cobalt.

Regarding the classification of all the genes according to its functionality, we found similar distribution of percentages between both bacteria (Fig. 4.8). Looking in more detail the distribution of genes, we appreciated, in general, a poor pool of conserved genes (Fig. 4.9). This might be due in part, to the great quantity of hypothetical genes included in the analysis. The genes devoted to lyases, isomerases, ligases and ribosomal RNA activities were the most conserved within the whole proteins. Lyase and ligase enzymes catalyze the joining and breaking, respectively, of various chemical bonds like carbon–oxygen, carbon–sulfur or carbon–nitrogen, therefore they should be expected to be conserved. Likewise, isomerases were supposed to be conserved as they catalyse only intramolecular structural rearrangement of isomers. In the case of ribosomal rRNAs, most of them were shared by both genomes since they are the central components of the ribosome and its function is to provide a mechanism

for decoding mRNA into amino acids and to interact with the tRNAs during translation. Furthermore, bacterial ribosomal rRNAs includes 16S, 23S and 5S rRNA genes which are typically organized as operon and the phylogenetic relatedness of both bacteria was usually deduced after 16S rRNA gene similarity.

Genes devoted to oxidoreductases, transferases and hydrolases were poorly conserved. With regard to oxidoreductases —which play important roles in hydrocarbon catabolism, among others—, some differences have been noticed between both genomes. For instance, *A. borkumensis* harbors various oxidoreductase genes clustered together or in operon-like structures (Schneiker et al., 2006), whereas in *O. antarctica* such operons were not described. Moreover, as discussed above, two key oxidoreductases like alkane hydroxylase and alkane monooxygenase were not detected in *O. antarctica*. However, they are likely to exist in a novelty structure, which agrees with the great number of unique oxidoreductases. In turn, several transferases seemed to take part on specific functions in each bacterium. Thus, transferases of *A. borkumensis* were reported to participate, amongst others, in biosurfactant production as well as formation of biofilms. On the contrary, neither the formation of biofilms nor the production biosurfactant by *O. antarctica* have been still published. The high unique hydrolases found could be related with the great variety of subclasses within this group of enzymes. No additional information about these enzymes for both bacteria has been reported before.

Finally, the discrepancy showed in the distribution of the remaining functional genes might be a consequence of the differences in the adaptation to specific habitats, the resistance factors or the physiological features. For instance, *A. borkumensis*, that thrives mostly in the upper layers of the ocean, hold a relatively low number of cold-shock proteins as well as several systems for detoxification of compounds like arsenate, mercury and other heavy metals (Schneiker et al., 2006), whereas *O. antarctica* is supposed to harbor far more cold-shock proteins as psychrophilic microorganism and any detoxification system has been described. Moreover, flagella is only present in *A. borkumensis*.

In summary, the differences in the optimal growth temperature between both bacteria lead to several genomic discrepancies. Adaptation to low temperatures implies lower strand asymmetries and higher scaling

exponents as showed by the DFA. The variation of GC contents could be explained by an unbalanced supply of the essential external precursors for nucleic acid synthesis and repair since both bacteria grow in different environment. Furthermore, a nucleotide bias in the mutational mechanism caused by a psychrophilic lifestyle might cause a GC decrease. The higher genomic diversity of extremophilic microorganisms was reflected by the higher size and the significant higher number of unique genes of *O. antarctica*. Some of these genes probably arose via gene duplication or lateral transfer in order to adapt to extreme environmental conditions. As expected, lipid metabolism presented some differences since psychrophilic marine bacteria modify their membrane lipids in order to adapt to low temperatures. For instance, the ratio of total unsaturated versus saturated fatty acids in the membrane lipids increases when the microorganism growths at low temperatures.

V.

A genome-scale metabolic model of *Alcanivorax borkumensis*, a paradigm for hydrocarbonoclastic marine bacteria

V A genome-scale metabolic model of *Alcanivorax borkumensis*, a paradigm for hydrocarbonoclastic marine bacteria

5.1 Introduction

Alcanivorax borkumensis SK2 is a gram-negative rod-shaped marine γ-proteobacterium that uses hydrocarbons as sole or principal sources of carbon and energy (Yakimov et al., 1998). It belongs to the genus *Alcanivorax,* a group of slow-growing marine hydrocarbonoclastic —hydrocarbons breaker— bacteria that preferentially use petroleum-derived aliphatic and aromatic hydrocarbons as carbon and energy sources. This bacterium is found in many marine habitats worldwide including the Mediterranean Sea, the Pacific Ocean, the Japanese and Chinese Seas and the Arctic Ocean (Golyshin et al., 2003; Harayama et al., 1999; Kasai et al., 2001; Röling et al., 2004; Syutsubo et al., 2001; Yakimov et al., 2005). It is present in low numbers in unpolluted environments, but becomes the dominant microbe in oil-polluted waters (Hara et al., 2003; Harayama et al., 1999; Kasai et al., 2001; Yakimov et al., 2005). The ubiquity and unusual physiology of *A. borkumensis* strongly suggest a pivotal role for hydrocarbons removal from polluted marine systems. *A. borkumensis* has been the first marine petroleum oil-degrading bacterium to be sequenced (Schneiker et al., 2006).

The relationship between the genotype and the phenotype is complex, highly non-linear and cannot be predicted from simply cataloguing and assigning gene functions to genes found in a genome. Comprehensive

understanding of cellular metabolism requires placing the function of every gene in the context of its role in attaining the set goals of a cellular function. This demands the integrated consideration of many interacting components. Mathematical modelling provides us a powerful way of handling such information and allows us to effectively develop appropriate frameworks that account for these complexities. Constraint-based metabolic models have been constructed for many microorganisms, including *Haemophilus influenzae* (Schilling and Palsson, 2000), *Methylobacterium extorquens* (Van Dien and Lidstrom, 2002), *Helicobacter pylori* (Schilling et al., 2002), *Escherichia coli* (Reed and Palsson, 2003; Feist et al., 2007), *Mannheimia succiniciproducens* (Hong et al., 2004), *Staphylococcus aureus* (Becker and Palsson, 2005; Heinemann et al., 2005), *Lactococcus lactis* (Oliveira et al., 2005), *Methanosarcina barkeri* (Feist et al., 2006), *Bacillus subtilis* (Oh et al., 2007), *Pseudomonas aeruginosa* (Oberhardt et al., 2008) and *Pseudomonas putida* (Nogales et al., 2008). In an attempt to capture some of the cellular metabolism complexities in *Alcanivorax borkumensis*, we present here a development of a genome-wide quantitative framework based on constraint-based modelling of its metabolic and transport network using one specific metabolic modelling approach named Flux Balance Analysis (FBA) (Kauffman et al., 2003). Drawing on annotated genome sequence data, biochemical information and strain-specific knowledge, we have made a reconstruction of this network, which currently comprises a total of 487 reactions, of which 30 are transport related.

The construction of comprehensive metabolic maps describes the metabolic capacities of *A. borkumensis* within the scope of given environmental constraints and provides a framework both, to study the consequences of alterations in the genotype and to gain insight into the phenotype–genotype relationship. Moreover, this analysis defines the entire metabolic space of the possible flux distributions and metabolic interactions within the network. A direct comparison of phenotypic spaces under different conditions will help in identifying evolutionary features and genetic plasticity. Such *in silico* models can be used as well to choose the most informative knockouts and to rationally design experiments relevant for the elucidation of the behavior of this environmentally important marine bacterium. For instance, the modelling of carbon fluxes versus those of nitrogen and phosphorus through a virtual metabolic network based on the genome sequence allowed us the

discovery of conditions in which the excess of carbon available in hydrocarbons was not directly translated into bacterial biomass. In fact, the optimal environmental agent would be the one that expresses a maximum of catalytic activity with a minimum buildup of biomass. Thus, carbon overflow is rationally diverted to the production of components like triacylglycerol, wax esters (Kalscheuer et al., 2007) or polyhydroxyalkanoates (bioplastics), an activity for which *A. borkumensis SK2* is genetically well endowed (Sabirova et al., 2006a). The metabolic model paves the way to optimize such carbon conversion processes. We present here the results and implications of this work for biotechnological and environmental applications, as well as for the metabolomic comparison of hydrocarbonoclastic bacteria.

5.2 Material and methods

Reconstruction of the metabolic network

The genomic sequence and annotation data of *A. borkumensis* strain SK2 (Schneiker et al., 2006) was used to assist in reconstructing its metabolic network. This information was complemented with biochemical information and cell physiology data in order to construct the set of all the specific reactions of the microorganism. The detailed process of metabolic reconstruction and model development has been previously reviewed (Covert et al., 2001). Briefly, the annotation of all the open reading frames (ORFs) of *A. borkumensis* pertaining to metabolic enzymes or membrane transporters were used as a framework on which translated metabolic proteins were assigned to form gen–protein–reaction (GPR) assignments. Most GPR assignments were made from the genome annotation and the model was manually constructed on a pathway basis. GPR associations were also directly made from biochemical evidence presented in journal publications and reviews. Next, biochemical database KEGG (Kanehisa et al., 2002), the previous constraint-based model constructed for the related microorganism *E. coli* (Reed and Palsson, 2003) and additional information from the literature pertaining to the phylogenetic related *Pseudomonas putida* were used as general guides for pathways reconstructions. The functionality of each gene in the genome of *A. borkumensis* was examined manually to find additional

reactions that were not present in the KEGG database. The fact that a reaction was present in any related microorganism or genomic data suggested that the reaction also occurs in A. borkumensis and, consequently, it was included in the model.

After assembling the network based on genomic data, missing functions were noted using two approaches. On the one hand, likely reactions were added to the model based on the research of information from the literature concerning the characterization of precise biochemical functions and the physiological data regarding *A. borkumensis* and related microorganisms. On the other hand, the BLAST algorithm (Altschul et al., 1997) was implemented to infer gene function for enzymes needed to complete pathways where no gene could be found in the *A. borkumensis* annotation. In order to determine if a not found gene exists in *A. borkumensis* for a given function, BLAST analysis was run against its genome annotation with the information of the missing enzyme from *E.coli* and *P. putida*. A reaction was putatively associated with a gene based on homology searches provided by BLAST and the annotation information for that gene from KEGG database. Only gene assignments with a high level of sequence similarity were selected for inclusion in the metabolic genotype. A BLAST e-value of 1e–04 was considered sufficient criterion to make an association when there were no close homologs. Each reaction included in the model represents the best determination of the biochemical reactions that the microorganism is believed to be capable of carrying out based on available data.

All of the reactions of the network were elementally balanced and were assigned either reversible or irreversible. Reversibility was determined from literature when the enzyme was characterized whereas reversibility of not characterized enzymes was assigned from thermodynamic considerations.

The model was created and maintained using ToBiN (Toolbox for Biochemical Networks, http://www.lifewizz.com).

Constraint-based modelling and FBA

Forty-nine metabolites were selected as required biomass constituents (supplementary file 5.1). Because no thorough biomass composition has been published for *A. borkumensis*, the relative production of metabolites required for growth was taken to be similar to that published for the related gram-negative γ-proteobacterium *E. coli* (Reed and Palsson, 2003). An output biomass reaction —exchange flux— was created that utilized these constituents in equal stoichiometric ratios to be used as a means to assess the ability of the network to produce all of the required demands based on particular substrate availability conditions. The complete ability of the network to produce all of the biomass constituents led to a positive flux value for this objective reaction. Biomass components were added to the objective function individually. Thus, when a simulation resulted in a positive net flux through the biomass, a subsequent component was added to the biomass and the simulation was run. When a biomass component added resulted in no flux, the network was manually updated. This process was continued until all of the biomass constituents in supplementary file 5.1 were included.

The reconstructed metabolic network and the defined biomass function allowed the calculation of network properties (including stoichiometry, thermodynamics and enzyme capacity) and optimal growth phenotypes through the use of Flux Balance Analysis (Appendix C).

Minimal medium determination

Each of the extracellular metabolites available to the metabolic network (alkane, NO_3, SO_4, O_2, P_i) was individually removed to determine if they were required for producing all of the biomass constituents. This determination was accomplished by constraining the exchange flux or uptake reaction of the metabolite to zero and optimizing for the biomass objective reaction. After thorough examination a set of metabolites was arrived at for which the removal of any of them would render the network unable to produce the biomass demands. This set of metabolites constitutes a defined minimal medium required for the *in silico* model to support growth of *A. borkumensis*.

Deletion studies

All the reactions in metabolic network were examined to determine whether they were essential to the model. The constraints on the particular reaction were set to zero in order to assess the consequences of deleting a reaction from the S matrix (see Appendix C). A simulation was then run to see if the network could support growth by optimizing for the biomass objective reaction without such reaction. If the network could not support growth, then the deleted reaction was deemed essential under the particular environment and medium conditions used in the simulation.

5.3 Results and discussion

Basic network properties

The metabolic reconstruction of *A. borkumensis* SK2 was generated using the procedure described in materials and methods. The model contains 462 metabolic genes associated with 487 reactions and 478 distinct metabolites. Basic properties of the reconstructed network are summarized in Table 5.1. The entire reaction and metabolite list were included in the supplementary files 5.2 and 5.3, respectively. Up to 52 reactions were included because either they have been reported in prior biochemical and physiological literature, or they were required to fill a gap in the reconstructed network (see materials and methods and supplementary file 5.4). These reactions were consequently unassociated with any gene product in the genome annotation. The reactions in *A. borkumensis* were subdivided into 10 functional categories based on the major metabolic roles of the cell. These subsystems included alkane (hydrocarbons) degradation, amino acid biosynthesis and degradation, carbohydrate metabolism, energy metabolism, glycan biosynthesis, lipid and cell envelope biosynthesis, vitamin and cofactor biosynthesis, nucleotide biosynthesis and degradation, transport and xenobiotics degradation (see Table 5.2). The largest number of reactions (120) involved amino acid metabolism and biosynthesis, probably because *A. borkumensis* contains all of the pathways required to synthesize *de novo*

the 20 common amino acids. A total of five reactions were not assigned to any subsystem because no clear information about them was found. Detailed information about subsystems can be consulted in supplementary file 5.2.

Annotated genes	3073
Metabolic processes	487
Irreversible	354
Reversible	136
Metabolites	478
Enzymes	462
ORF	410
added genes	52

Table 5.1. Main characteristics of the reconstructed metabolic network of *A. borkumensis.*

Metabolism	Reactions
Alkane degradation	**4**
Amino acid metabolism	**120**
Alanine and aspartate metabolism	4
Arginine and proline metabolism	20
Arginine, putriscine and spermidine biosynthesis	1
Cysteine metabolism	9
Glutamate metabolism	5
Glutathione biosynthesis	1
Glycine and serine metabolism	8
Histidine metabolism	10
Methionine metabolism	6
Tyrosine, tryptophan and phenylalanine metabolism	20
Threonine and lysine metabolism	12
Valine, leucine and isoleucine metabolism	24
Carbohydrate metabolism	**73**
Glyoxylate and decarboxylate metabolism	6
Butanoate metabolism	3
Propanoate metabolism	1
Glycolysis	16
Alternate carbon metabolism	17
Anaplerotic reactions	6
Citrate Cycle (TCA)	10
Methylglyoxal metabolism	2
Pentose Phosphate Cycle	7
Pyruvate metabolism	4
Inositol metabolism	1
Energy metabolism	**25**
Nitrogen metabolism	4
Methane metabolism	2
Oxidative phosphorylation	19
Glycan Biosynthesis	**36**
Cell envelope biosynthesis	33
Lipopolysaccharide biosynthesis	3
Lipid metabolism	**63**
Biosynthesis of steroids	7
Fatty acid metabolism	21
Isoprenoid biosynthesis	1
Membrane lipid metabolism	32
Synthesis and degradation of ketone bodies	2

Metabolism	Reactions
Metabolism of cofactors and vitamins	**88**
Cofactor and prosthetic group biosynthesis	53
Folate biosynthesis	12
NAD biosynthesis	3
Nicotinate and nicotinamide metabolism	1
Porphyrin and chlorophyll metabolism	5
Quinone biosynthesis	6
Riboflavin biosynthesis	2
Tetrapyrrole biosynthesis	1
Thiamin (vitamin B1) biosynthesis	2
Vitamin B6 (pyridoxine) biosynthesis	3
Nucleotide metabolism	**61**
Nucleotide salvage pathways	37
Purine and pyrimidine biosynthesis	24
Transport	**30**
ABC transporters	1
Membrane transport	15
PTS	14
Xenobiotics degradation	**2**
Unassigned	**5**

Table 5.2. Distribution of reactions in the network of *A. borkumensis*. The network was divided into 10 basic submetabolisms, each of them with a variable number of pathways.

In agreement with the fact that *A. borkumensis* uses hydrocarbons as carbon and energy sources, neither glucose nor monomeric sugar membrane transports systems were identified within the genomic annotation. Moreover, there were no hexokinase enzymes to introduce the glucose into the glycolysis through glucose-6-P. Alkanes were used as hydrocarbon components for simulation purposes. The results presented here were achieved using octadecane as specific alkane, similar results were obtained when using other alkanes (data not shown).

The network has a significant number of dead-end metabolites. These dead-ends are compounds that were either only produced or only consumed by reactions in the network. The hypotheses regarding the

presence of a dead-end metabolite in a reconstructed network have been reviewed previously (Becker and Palsson, 2005) and may be mainly motivated for missing enzymes or reactions. Taking into account the fact that FBA analysis must be done under steady state conditions, the accumulation or depletion of any compound cannot occur in the network. Thus, any reaction involving dead-end metabolites cannot be used in a computed network state. The total number of dead-end compounds in the network is 129 (see supplementary file 5.5), being 101 the reactions involved in these dead-end metabolites (see supplementary file 5.6). This number of reactions is similar to reconstructions previously done [e.g., *Staphylococcus aureus,* (Becker and Palsson, 2005)]. All of these reactions have an associated gene and were included because of genetic evidence that they are present in *A. borkumensis*. Such number of dead-end metabolites may be due in part because this is the first approach of *A. borkumensis* reconstruction. New additions to the model will likely close some of these gaps.

A reaction representing biomass formation, consisting of 49 metabolites required for cellular growth, has been defined (supplementary file 5.1). Key components of this reaction include amino acids, nucleotides, lipids, vitamins, cofactors, solutes and cell wall constituents. Because data describing the biomass composition of *A. borkumensis* could not be located in the literature, data from the phylogenetic related γ-proteobacterium *E. coli* was used where necessary.

Validation of the model

In order to validate the reconstructed metabolic network of *A. borkumensis*, computational predictions were compared with the available experimental results (Sabirova et al., 2006a and 2006b). Thus, comparisons between *in silico* predicted results and experimental data concerning biomass and polyhydroxybutyrate (PHB, the four carbon units polyhydroxyalkanoate) formation of *A. borkumensis* growing on both, pyruvate and octadecane carbon sources are showed in Table 5.3. Experimental biomass was calculated plotting optical density measurements under 600 nm wavelength versus time (Sabirova, 2006). Te considered time of reference for the biomass formation was 24 hours. The correlation factor *k* was then used to convert optical density units

into biomass units (gDW/l). The correlation factor, which is an intrinsic parameter of each prokaryotic string, was determined by dividing the cell dry weight by the optical density. Since the correlation factor in *A. borkumensis* has not been previously determined, we used the average from 6 different prokaryotic species: *Escherichia coli* (Nanchen et al., 2006; Sauer et al., 1999), *Bacillus subtilis* (Fischer and Sauer, 2005; Nijland et al., 2007; Tännler et al., 2008), *Corynebacterium glutamicum* (Lindner et al., 2007), *Bacillus licheniformis*, *Bacillus pumilus* and *Bacillus amyloliquefaciens* (Tännler et al., 2008). The resulted correlation factor averaged 0.4283. Because the PHB biosynthesis implies less biomass formation as PHB is a dead-end component and is not a biomass metabolite, optimizing the network for biomass formation causes null PHB formation. Thus, we set the network for non-zero PHB production rate; first, by fixing the biomass to the maximum value given by experimental data (i.e. 4.283 gDW/l) and then optimizing for PHB formation (i.e. 0.043 M PHB), and second, by fixing the PHB to experimental value (i.e. 0.0234 μM PHB) and then optimizing for biomass formation (i.e. 0.64 gDW/l). Regarding the biomass formation, the results were quite similar when *A. borkumensis* grew on pyruvate (the model predicted 0.64 gDW/l and the experimental value was 0.4283 gDW/l), whereas the growth under octadecane showed some differences (1.52 gDW/l versus 0.2142 gDW/l). This discrepancy could be caused by both, the specific high C/N ratio conditions used to growth the samples in the laboratory and experimental measurements errors in the optical density of the samples at a wavelength of 600 nm. Moreover, the predicted biomass of *A. borkumensis* was based on an averaged correlation factor from different bacteria. Thus, the knowledge of the specific correlation factor of *A. borkumensis* would allow a more accurate prediction of the biomass formation. The direct comparison of the PHB formation was not possible because experiment data referred to the polymer of PHB ($[C_4H_6O_2]_n$), whereas the *in silico* calculated PHB formation values referred to the monomer of PHB, the hydroxybutyrate. Logically, polymer molar concentration value was lower than the molar concentration of the monomer. Thus, the molar concentration of the PHB polymer was within the micromolar magnitude whereas the monomer concentration was within the molar range. Comparison between the model and experimental PHB formation was possible taking into account the percentage of PHB increment when growing with octadecane versus pyruvate as carbon source. Thus, the increment ratio of the polymer PHB for experimental data was 2.76 (0.0647/0.0234) while the predicted ratio

increment for the monomer PHB had a similar value, 2.35 (0.101/0.043). This discrepancy might be due to the difficulty to experimentally quantify the PHB bacterial production.

Substrate	**Biomass model** (gDW/l)	**Biomass exp.** (gDW/l)	**PHB yield model** (M)	**PHB yield exp.** (μM)
0.23 M pyr	0.64	0.4283	0.043	0.0234
0.06 M oct	1.52	0.2142	0.101	0.0647

Table 5.3. Comparison between computational predictions and experimental data for biomass and PHB production grown on pyruvate and octadecane as carbon sources.

Further comparisons between computational predictions and experimental data were carried out using more experimental results (Sabirova et al., 2006b). The experimental results presented some different expressed genes in *A. borkumensis* growing on pyruvate and octadecane as carbon sources. We run the network with the carbon concentration indicated in the published results and highlighted the differences in gene expression. The optimized objective function for calculating the predicted fluxes was, in this case, the biomass formation. Similar trends regarding the up- and down-expression were shown between computed and experimental results (Table 5.4). The only exception was the acetyl-CoA carboxylate which showed slight differences between prediction (2.38 times up-expressed in alkanes) and experimental result (1.9 times down-expressed in alkanes). As expected, the model predicted that oxidation of alkanes pathway and fatty acid oxidation complex were only expressed in alkanes in agreement with experimental results. Moreover, prediction of fatty acid biosynthesis genes up-expression in alkanes presented close values to that of experimental data. One specific and worth pathway for comparison purposes between bacterial grown on pyruvate and alkane was the glyoxylate bypass (Sabirova et al., 2006b). Briefly, during growth on alkanes as the sole carbon source, bacteria must generate all cellular precursor metabolites from acetyl-CoA. One mechanism to do this is the short-circuiting of the citric acid cycle through activation of the glyoxylate bypass, which routes acetyl-CoA to phosphoenolpyruvate via

isocitrate, glyoxylate and malate, by means of isocitrate lyase and malate synthase. Succinate produced via glyoxylate bypass is converted to malate by succinate dehydrogenase. Malate is converted to oxaloacetate by malate dehydrogenase or is used by malic enzyme in gluconeogenesis to produce pyruvate. Pyruvate is then converted by phosphoenolpyruvate synthase to produce phosphoenolpyruvate. The incomplete TCA cycle is associated with the alkane-induced down-regulation of isocitrate dehydrogenase, 2-oxoglutarate dehydrogenase and succinyl-CoA synthetase. The glyoxylate bypass route, induced by alkanes, adapts the cell to produce key cellular precursor metabolites directly from the fatty acids produced by alkane oxidation. The predicted data, regarding the enzymes involved in the glyoxylate bypass, were also in agreement with the experimental results.

These results reflected the potential of the reconstructed network to simulate different growth conditions and helped to validate the metabolic model.

Metabolism	Gene	Differential model	Differential experiment
Terminal oxidation of alkanes	Alcohol dehydrogenase	A	A
	Aldehyde dehydrogenase	A	A
	Alkane hydroxylase	A	A
Fatty acid oxidation	Fatty acid oxidation complex	A	A
Fatty acid and phospholipid biosynthesis	Acetyl-CoA carboxylate	2.38 up	1.9 down
	Fatty acid biosynthesis	2.38 up	6.4 up
	Cardiolipin Synthase	A	A
Amino acid biosynthesis	Dihydroxy-acid dehydratase	2,38 up	A
TCA, glyoxylate bypass and gluconeogenesis	Isocitrate dehydrogenase	1.22 down	2 down
	2-oxoglutarate dehydrogenase	P	P
	Isocitrate lyase	A	36 up
	Malate synthase	A	6.1 up
	Malic enzyme	3.44 up	3.1 up
	Succinate oxidoreductase	2.33 up	A

Table 5.4. Comparison between *in silico* and experimental differentially expressed enzymes grown on either alkane or pyruvate as carbon source. “A” means that the enzyme is solely expressed on alkane; “P” means that the enzyme is solely expressed on pyruvate; “down” means that the enzyme is down expressed on alkane; “up” means that the enzyme is up expressed on alkane.

Minimal media and growth requirements

We used FBA to determine fluxes leading to both, optimal growth and optimal PHB formation subject to constraints on the usage of each reaction. This principle allowed us to systematically predict a minimal media composition capable of supporting growth and PHB formation in *A. borkumensis*.

A. borkumensis was capable of synthesizing all amino acids since the complete set of reactions corresponding to the biosynthesis pathway were included in the model. Computationally predicted set of required substrates for growing includes oxygen, a phosphate source —inorganic phosphate—, sulfate as sulfur source, nitrate as nitrogen source and alkane as the only carbon source. Once the carbon from alkane was fed into central metabolism as pyruvate, the formation of the necessary precursor metabolites for widespread biosynthesis can be accomplished.

Deletion Study

In order to determine the effects of the deletion of a reaction from the network, as would occur in a gene knockout experiment, FBA was used with the additional constraint of the flux through a particular reaction reset to zero. This allowed for the rapid prediction of both, gene and reaction deletions. We calculated the effects of all single reaction deletions in the minimal media described above. We found that 259 reaction deletions were computationally predicted to be lethal. This number was close to the 230 lethal reaction deletions in *Staphylococcus aureus* (Becker and Palsson, 2005).

The number of predicted essential reactions in relation to the total number of reactions in each submetabolism allowed classifying them into flexible or rigid parts. Logically, all the alkane pathway's reactions were essential for *A. borkumensis*. Moreover, metabolism of amino acids, lipids and nucleotides seemed to be quite rigid and deletions could poorly be compensated. On the contrary, energy and glycan metabolisms presented a better compensation of deletions by means of alternative enzymes or pathways. Carbohydrate, cofactors and vitamins metabolisms showed intermediate compensation values. There does not exist a

comprehensive resource regarding gene essentiality in *A. borkumensis* for a comparison between experimental data and the predictions detailed herein. However, there is a comprehensive gene essentiality study for the microorganism *Staphylococcus aureus* (Heinemann et al., 2005). Our predictions presented some similarities to the *S. aureus* values, but there was a significant difference in the amino acid, nucleotide and glycan metabolisms probably caused by the different lifestyle of the two bacteria and the comparison between gram-negative and gram-positive bacteria. The predictions are detailed in Table 5.5 and supplementary file 5.7.

Submetabolism	Essential reactions	Total reactions	Percentage	Classification
Alkane degradation	4	4	100.00	Rigid
Amino acid metabolism	90	120	75.00	Rigid
Carbohydrate metabolism	33	72	45.83	Flexible
Energy metabolism	6	25	24.00	Flexible
Glycan biosynthesis	3	13	23.08	Flexible
Lipid metabolism	47	63	74.60	Rigid
Cofactors and vitamins	36	88	40.91	Flexible
Nucleotide metabolism	37	61	60.66	Rigid
Xenobiotics degradation	0	2	0.00	Not used for growth
Unassigned	0	5	0.00	Not used for growth

Table 5.5. Number and percentage of essential reactions in the submetabolisms, excluding transport reactions.

Polyhydroxyalkanoate biosynthesis

Under carbon-limited conditions, an increase in carbon allows an increase in bacterial growth rate until another growth limitation is reached. The appearance of alkanes in oligotrophic environments like most marine habitats allows *A. borkumensis* to grow until nitrogen or phosphorus limitation is experienced. Oil pollution constitutes a temporary condition of carbon excess coupled with limiting nitrogen. Under such conditions of high C/N ratios, many microbes synthesize carbon storage materials, like polyhydroxyalkanoates, triacylglycerol, wax esters or other cellular storage substances (Kalscheuer et al., 2007; Steinbüchel, 1991). The predicted quantitative influence of carbon, nitrogen and phosphorus limitation over the growth and polyhydroxyalkanoates synthesis of *A. borkumensis* is showed in Tables 5.6 and 5.7. Biomass and PHB formation were clearly more affected than respiration process by the limitation of either nitrogen or phosphorus sources. Thus, in an environment with excess of alkanes, these two compounds seem to be critical for *A. borkumensis* to grow. On the contrary, the aerobic respiration process was clearly more affected by the carbon limitation since the oxygen consumption and the carbon dioxide formation showed significantly lower values in comparison with the other two limited conditions. Similar conditions were imposed to *E. coli* to experimentally study its growth parameters under carbon and nitrogen limited conditions (Sauer et al., 1999). The reported results were in agreement with those presented in Tables 5.6 and 5.7 regarding the biomass formation and the respiration process.

Limitation	**Biomass** (gDW)	**Biomass yield on**		
		Octadecane (g g of oct.)	**NO_3** (g g of NO_3)	**P_i** (g g of Pi)
Alkane	0.83	13.91	4.7145	8.1962
NO_3	0.71	0.97	4.7142	8.1958
P_i	0.66	1.64	4.7143	8.1959

Table 5.6. Predicted biomass formation and biomass yield on octadecane, nitrate and phosphate under limited conditions in *A. borkumensis*.

Limitation	Octadecane consum. (M)	NO_3 consum. (M)	P_i consum. (M)	PHB form. (M)	O_2 consum. (M)	CO_2 form. (M)
Alkane	0.06	0.1770	0.1018	0.1428	0.3965	0.1394
NO_3	0.06	0.15	0.0863	0.1347	0.5306	0.2280
P_i	0.06	0.1391	0.08	0.1315	0.5848	0.2639

Table 5.7. Predicted growth parameters under alkane, nitrate and phosphate limited conditions in *A. borkumensis*. (consum - consumption; form - formation).

Regarding the relation between the carbon–nitrogen ratio consumption and the biomass and PHB formation, Table 5.8 shows that an increment in the C/N ratio caused both, biomass and PHB increment. However, the PHB increment was proportional to the C/N ratio increment over all the scale of values, whereas biomass grew proportionally to the C/N ratio solely for small values. When ratio was close to 25:1 or higher, there was an excess of carbon available for *A. borkumensis* which was not directly transformed into bacterial biomass, but derived into carbon storage materials, like PHB, triacylglycerol or wax esters. These trends can be also noticed in Fig. 5.1. Respiration process increased proportionally with the C/N ratio, as showed in Table 5.8, in agreement with the results discussed in Tables 5.6 and 5.7.

C/N ratio	**P_i consum.** (mmol $g^{-1}h^{-1}$)	**Biomass** (gDW)	**PHB form.** (mmol $g^{-1}h^{-1}$)	**O_2 consum.** (mmol $g^{-1}h^{-1}$)	**CO_2 form.** (mmol $g^{-1}h^{-1}$)
1:3	0.0017	0.0141	0.0023	0.0082	0.0038
1:2	0.0046	0.0377	0.0084	0.0651	0.0393
1:1	0.0057	0.0471	0.0180	0.1988	0.1191
2:1	0.1438	1.1786	0.8246	9.38	5.18
5:1	0.2301	1.8857	3.11	40.51	22.68
10:1	0.2876	2.3571	7.64	69.34	27.76
15:1	0.3451	2.8285	13.67	126.71	51.31
20:1	0.4026	3.2999	21.20	198.58	80.86
25:1	0.4314	3.5358	28.34	267.14	109.14
30:1	0.4371	3.5828	34.42	325.80	133.40
40:1	0.4429	3.6300	46.42	441.74	181.35

Table 5.8. Growth condition and PHB formation analysis under different C/N ratios using octadecane as carbon source.

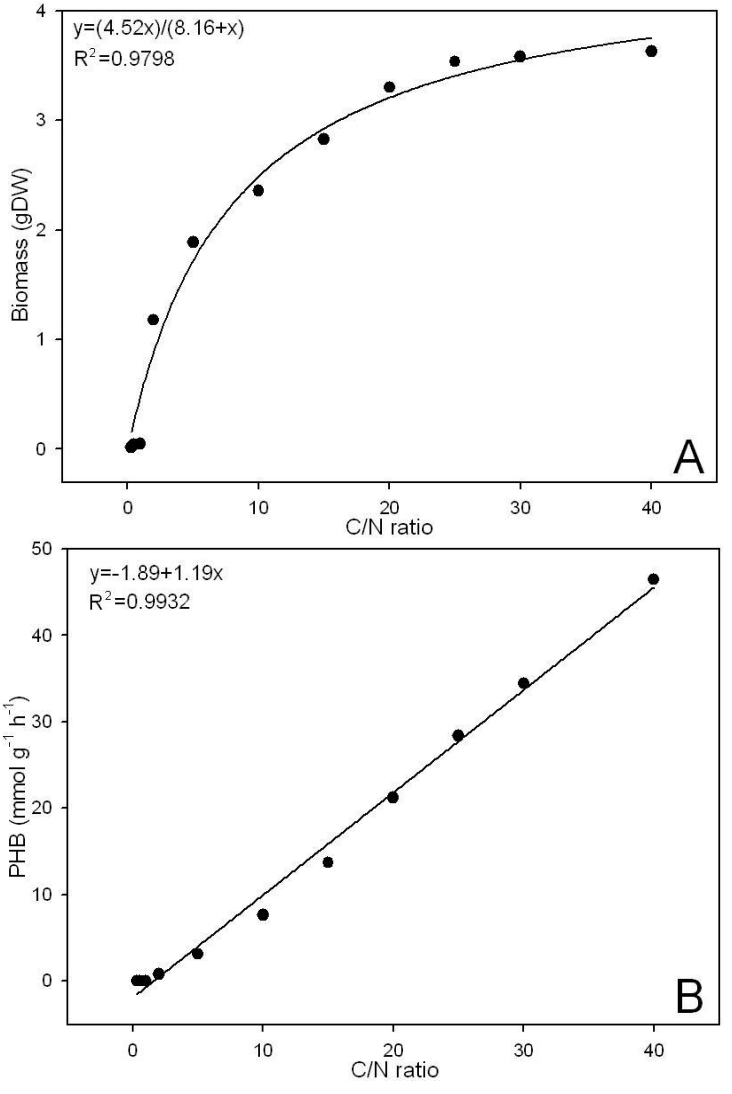


Figure 5.1. Plots A and B show the predicted relation between the carbon–nitrogen ratio consumption of *A. borkumensis* and its biomass and PHB formation, respectively when octadecane was used as carbon source. As observed, biomass grows proportionally to the C/N ratio for small values, whereas for C/N ratios of approximately 25:1 or higher, there is an excess of carbon available for *A. borkumensis* which is not directly translated into bacterial biomass. A fraction of the excess of carbon is directly transformed into PHB, whose growth is proportional to the C/N ratio increment for all the scales of values.

Moreover, large differences in PHB formation have been noticed between bacteria growing on pyruvate and octadecane. Thus, Fig. 5.2 shows that the amount of PHB produced from alkane (i.e., under conditions of a high C/N ratio) was three times higher than the amount of PHB produced during growth on pyruvate. This significant difference was produced by the specific pathways used to synthesize PHB depending on the carbon source. During growth on pyruvate as carbon source (Fig. 5.2A), PHB was formed from acetyl-CoA that was further converted to acetoacetyl-CoA and (S)-3-hydroxybutyryl-CoA, as precursor of PHB formation by the enzymes acetyl-CoA acyltransferase and (S)-3-hydroxybutyryl-CoA dehydrogenase, respectively. When *A. borkumensis* grew on alkane, like octadecane, as carbon source (Fig. 5.2B), the PHB was generated from octadecane which was converted, via alkane oxidation and fatty acid metabolism in the PHB precursor, (S)-3-hydroxybutyryl-CoA. Finally, the model confirmed the notable increment in PHB formation that was achieved when the enzyme tesB was constrained to 0 simulating an *A. borkumensis* tesB-like mutant (Fig. 5.2C). PHB formation in the tesB-like mutant growing on alkane was 20 times higher than in the WT strain under the same growth conditions. Thus, the lack of tesB gene implies higher PHB formation due to the fact that the enzyme tesB-like acyl-CoA thioesterase also uses (S)-3-hydroxybutyryl-CoA as substrate to produce 3-hydroxyalkanoic acid (Sabirova et al., 2006a).

In summary, the metabolic model allowed to gain insight into the basis of hydrocarbonoclastic, marine lifestyle, its genomic responses to environmental stresses, the ability to degrade a range of hydrocarbons and to dominate oil-degrading microbial communities, as well as the mechanisms that provide it with its remarkable oil-degrading abilities and its competitive advantage in oil-polluted environments. The modelling of alkane fluxes versus those of nitrogen and phosphorus through the metabolic network also allowed the discovery of conditions in which the excess carbon available in hydrocarbons was not directly translated into bacterial biomass. Instead of that, carbon overflow was diverted to the production of polyhydroxyalkanoates, an activity for which *A. borkumensis* SK2 showed to be genetically well endowed. However, it has been described that *A. borkumensis* probably also employs other types of storage compounds to serve as carbon/energy source storage during periods of carbon/energy limitation like triacylglycerol or wax

esters (Kalscheuer et al., 2007). Further refinement is necessary in order to analyze these storage compounds.

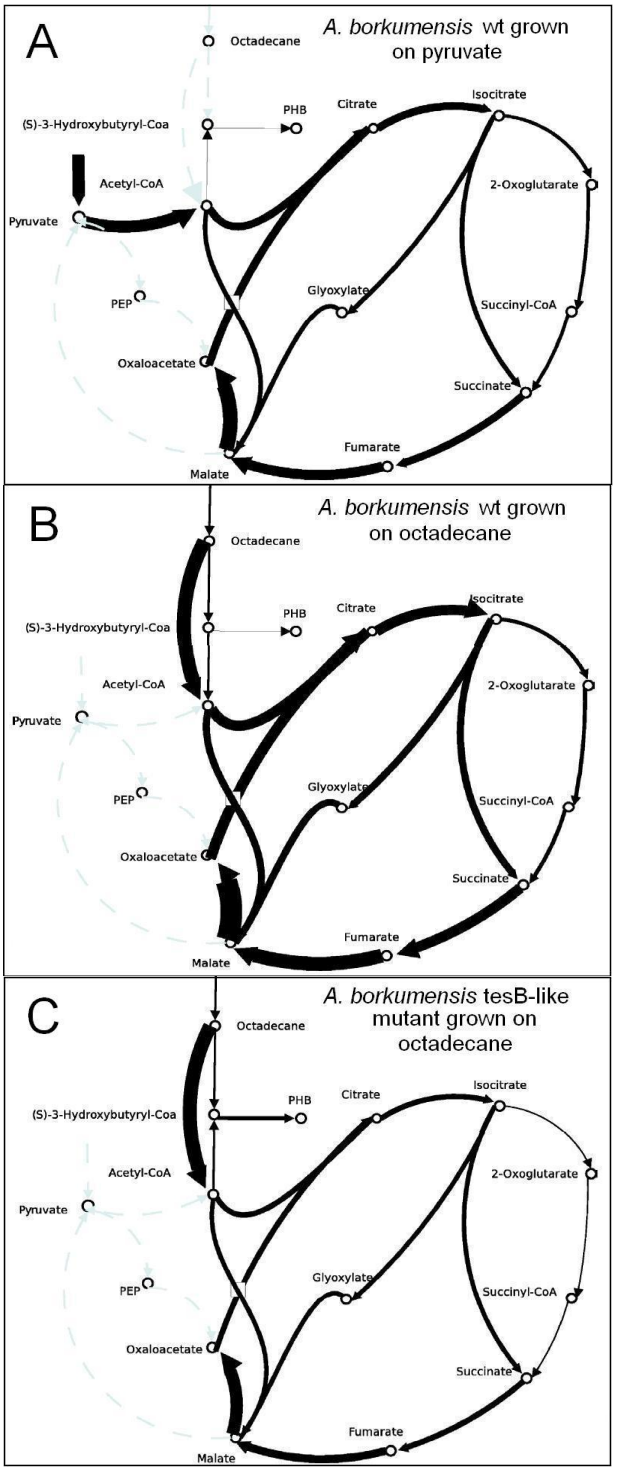

A
A. borkumensis wt grown on pyruvate
Octadecane
PHB
Citrate
Isocitrate
(S)-3-Hydroxybutyryl-Coa
Acetyl-CoA
2-Oxoglutarate
Pyruvate
Glyoxylate
PEP
Succinyl-CoA
Oxaloacetate
Succinate
Fumarate
Malate
B
A. borkumensis wt grown on octadecane
Octadecane
PHB
Citrate
Isocitrate
(S)-3-Hydroxybutyryl-Coa
Acetyl-CoA
2-Oxoglutarate
Pyruvate
Glyoxylate
PEP
Succinyl-CoA
Oxaloacetate
Succinate
Fumarate
Malate
C
A. borkumensis tesB-like mutant grown on octadecane
Octadecane
PHB
Citrate
Isocitrate
(S)-3-Hydroxybutyryl-Coa
Acetyl-CoA
2-Oxoglutarate
Pyruvate
Glyoxylate
PEP
Succinyl-CoA
Oxaloacetate
Succinate
Fumarate
Malate

Figure 5.2. Simulation of TCA cycle and fatty acid metabolism used by *A. borkumensis* to convert alkanes via terminal oxidation to fatty acids, acetyl-CoA and PHB. Schemes show PHB production in the WT (wild type) and tesB-like mutant strains of *A. borkumensis* under different sources of carbon. The arrows pointing down from octadecane mean fatty acid metabolism (beta-oxidation) and those pointing up from acetyl-CoA mean fatty acid biosynthesis and synthesis of (S)-3-hydroxybutyryl-CoA from acetyl-CoA. The thick arrow from octadecane shows total acetyl-CoA production in the 7 rounds of beta-oxidation between octadecane and (S)-3-hydroxybutyryl-CoA. The thin arrow shows production of (S)-3-hydroxybutyryl-CoA from octadecane. The former arrow is 7 times thicker than the latter due to the 7 rounds of beta-oxidation. Fig. 5.2A shows the basic metabolic fluxes of *A. borkumensis* WT strain when growing on pyruvate as carbon source. The PHB formation is possible, under this condition, thanks to the reversibility property of the first two enzymes of the fatty acid biosynthesis represented by the arrow pointing from acetyl-CoA to (S)-3-hydroxybutyryl-CoA. First, the enzyme acetyl-CoA acyltransferase converts acetyl-CoA to acetoacetyl-CoA and second, (S)-3-hydroxybutyryl-CoA dehydrogenase converts acetoacetyl-CoA to (S)-3-hydroxybutyryl-CoA which is the precursor of the PHB formation. Fig. 5.2B shows the metabolic flux in *A. borkumensis* WT from alkane to PHB and the TCA cycle. In this case, the PHB is generated from octadecane which is transformed, via alkane oxidation, in the corresponding alcohol, aldehyde and acid. Then, the acid is processed by the fatty acid metabolism, first until the PHB precursor (S)-3-hydroxybutyryl-CoA and later until acetyl-CoA and the TCA cycle. The PHB formation when the carbon source is an alkane is 3 times higher than the simulation when pyruvate is chosen. Fig. 5.2C shows the metabolic flux in *A. borkumensis* tesB-like mutant growing on alkane. The PHB formation in the mutant is 20 times higher than in WT strain under the same growth conditions. Note the change in the direction of acetyl-CoA acyltransferase and (S)-3-hydroxybutyryl-CoA dehydrogenase (arrow pointing from acetyl-CoA to (S)-3-hydroxybutyryl-CoA) in contrast with Fig. 5.2B. This change allowed the (S)-3-hydroxybutyryl-CoA formation from both, acetyl-CoA and octadecane (via fatty acid metabolism) and, in consequence, more PHB formation.

5.4 Conclusion

We have reconstructed the metabolic network of the poorly known microorganism *A. borkumensis* integrating genomic, biochemical and physiological data together in the context of an *in silico* model with the aim of generating a practical tool in the quest to understand the physiology of this microorganism. The generated functional information was possible thanks to genome sequencing efforts and it provides additional justification for new contributions in genome sequencing.

The model predictions, using constraint-based analysis, were in agreement with the experimental data, especially the related with growth phenotypes and PHB formation. It shows a potential in the use of the model as a high-throughput analysis tool for studying growth of *A. borkumensis*. Moreover, the overall modelling process can assist in accelerating the pace of biological discovery by generating experimentally testable hypotheses. It can also determine the redundancy or robustness of reactions in the network and predict the formation of products by WT and mutants under different media. Although the *A. borkumensis* metabolic model is a useful tool, frequent refinement and updating with new experimental data is necessary to improve its accuracy to predict cellular phenotypes and to provide the most concise representation of the microorganism's known functional capabilities. The generated model can serve as a starting point for additional hydrocarbonoclastic bacteria reconstructions and as an analysis platform for the study of natural oil-degrading bacteria.

VI.

Conclusions

VI Conclusions

Interactions between genomes of prokaryotic microorganisms and environment during evolution have been analyzed in different contexts. The study of the relationship between the prokaryotic genomic patterns with lifestyles and metabolisms provide valuable information about microbial diversity. Genometric and mathematical techniques have demonstrated to be suitable tools for deciphering genomic patterns and for analysing genotypic adaptations to specific environments. Next, an outline of the main conclusions reached is provided.

Chapter II

- The combination of DNA walk and Detrended Fluctuation Analysis (DFA) is a suitable method for capturing intrinsic phylogenetic, ecological, and metabolic signals in prokaryotic genomes.

- All the analyzed genomes presented persistent long-range correlations (i.e., DFA scaling exponents higher than 0.5). This specific feature in the prokaryotic genome landscape indicates the existence of selective pressures modelling the architecture along the whole genome.

Conclusions

- The observed long-range correlations may be related to two biological factors:

    1. Elongation of the molecule by repetitive structures added inside the genomes generated by both, gene duplication and massive lateral transfer of genes from other genomes.
    2. The asymmetric DNA replication along the whole microbial genome.

- There was a consistent correlation between extremophiles and high scaling exponents.

- The rates of spontaneous mutation are greatly accelerated at extreme environments. However, extremophiles should have very efficient molecular strategies for repairing DNA under these conditions of chemical instability since they presented weak mutational bias in their genomes.

- The decrement of GC content in parallel with Topt in thermophiles and psychrophiles suggest that the transition from a hyperthermophilic to a psychrophilic environment would imply a structural adaptation in microbial genomes both in the GC content and in the sequential position of the nucleotides along the genome.

Chapter III

- The distribution of the nucleotidic sequence along the genome appeared to be related with the genome functionality, as deduced from the high canonical correlation between DFA and COG, as well as the high DFA variance explained by COGs. In consequence, COG distribution and DFA scaling exponent are two closely related genomic features that may be originated by similar factors.

- Expansions and contractions in the genomic repertoire have affected genes mostly involved in environmental interactions (e.g., energy metabolism, transport and regulation). In turn, basic information processes such as transcription and translation are distributed more homogeneously.

- High scaling exponents and heterogeneity in the percentage of functional genes seem to be essential requirements for genetic adaptation to extreme habitats.

Chapter IV

- *A. borkumensis* and *O. antarctica* seem to have evolved during long time as independent evolutionary lines as deduced from their comparative dot plot, DNA walks and suite of genes.

- *O. antarctica* had higher DFA scaling exponents than *A. borkumensis*, suggesting that a large number of genes from this psychrophilic bacterium may have been implicated in arrangement events, such as lateral transfer and gene duplication, in order to better adapt to extreme low temperatures.

- A negative correlation between scaling exponents and GC content exists in both bacteria. *A. borkumensis* has a significantly higher GC content (54.73%) than *O. antarctica* (42.16%). A nucleotide bias probably caused by a stressed psychrophilic lifestyle might explain the GC variation.

- The genometric analyses suggested that *O. antarctica* was more affected by spontaneous mutation, duplication and insertion events than *A. borkumensis*, which harbored a small number of mobile genetic elements such as transposons and insertions elements. This factor might also explain the differences in scaling exponents and genome size.

Chapter V

- The reconstruction of the metabolic network of the marine microorganism *A. borkumensis* was possible thanks to the integration of genomic, biochemical and physiological data that were put together in the context of an *in silico* model. The *A. borkumensis* metabolic model allowed us to gain insight into the basis of hydrocarbonoclastic and marine lifestyle.

- The modelling of alkane versus nitrogen and phosphorus fluxes through the metabolic network unveiled conditions in which the excess of carbon available in hydrocarbons was not directly translated into bacterial biomass. Carbon overflow, instead, was diverted to the production of polyhydroxyalkanoates, triacylglycerol and wax esters.

- The constraint-based model predictions showed the large potential of the model to be used as a high-throughput analysis tool to study growth strategies of *A. borkumensis*.

Appendices

Appendices

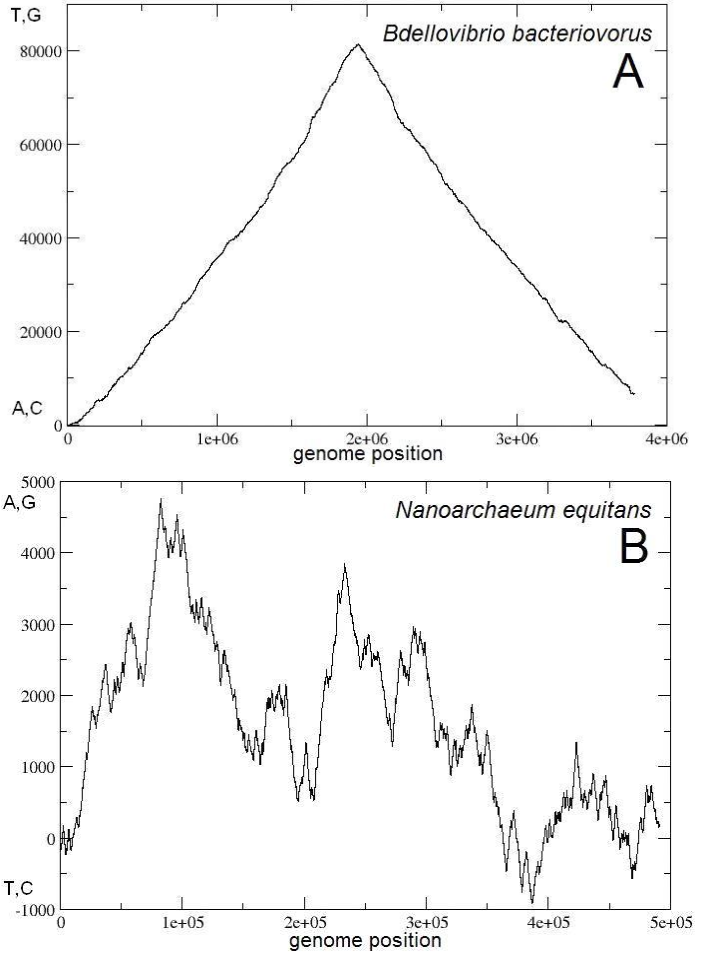
A. DNA walk

The DNA walk method (Lobry, 1996a and 1999; Grigoriev, 1998; Cebrat and Dudek, 1998) is a graphical representations of the fluctuations in nucleotide series which provide quantification on internal deviations of individual nucleotides along the genome. The run of DNA walks starts at position 0 of each sequence and every genome produces a specific DNA walk. There are several possibilities and rules to plot genomic landscapes (Buldyrev et al., 1995).

First, the original nucleotide sequence can be translated onto a one-dimensional numerical series. Since there are four different residues in a DNA sequence and the random walk has two possible directions, one needs to group the bases in pairs following three mapping rules; i) the hybrid rule (KM): being n_i the i nucleotide of the genomic sequence and y_i the DNA walk value for the nucleotide n_i, if n_i is a keto forms (G or T) then $y_i = +1$ and if n_i is an amino forms (A or C) then $y_i = -1$; ii) the purine–pyrimidine rule (RY): if n_i is a pyrimidine (C or T) then $y_i = +1$ and if n_i is a purine (A or G) then $y_i = -1$; iii) the hydrogen bond energy rule (SW): if n_i is a strongly bonded pair (G or C) then $y_i = +1$ and if n_i is a weakly bonded pair (A or T) then $y_i = -1$. These DNA walks generate an irregular graph resembling a fractal landscape. The defining feature for the landscape is the statistical self-similarity of the plots obtained at

various magnifications calculated with the Detrended Fluctuation Analysis (DFA) method (see Appendix B for details). The resulting DNA walk series can be mapped onto an orthogonal plane. Fig. A1 represents an example of the plots generated by the three types of DNA walks (KM, RY, SW) for three microbial (bacteria and archaea) genomes.

Another possible DNA walk representation is the named two-dimensional (2D) map, in which each nucleotide defines one direction in a plane formed by two orthogonal axes (i.e., C versus G and T versus A). In this walk, the walker moves 1 unit onto the plane according to the four senses defined by the nucleotide read. This 2D DNA walk generates an irregular graph resembling a fractal landscape (see example in Fig. A2).



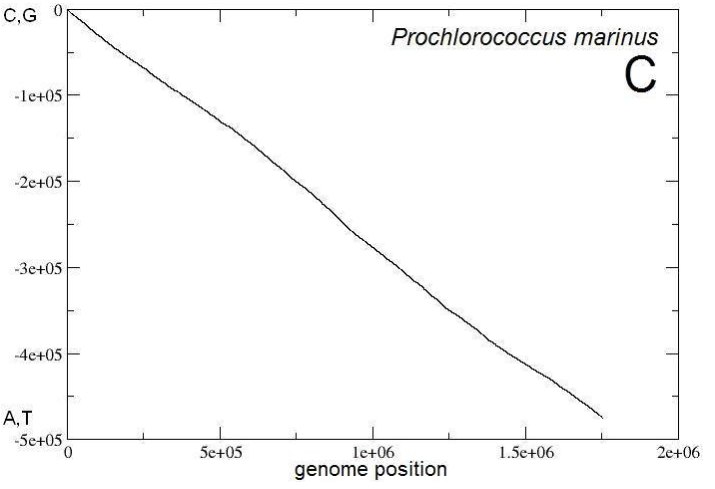


Figure A1. Typical DNA walk representations in one-dimensional space of three different genomes. Plot A illustrates the *Bdellovibrio bacteriovorus* DNA walk using the KM mapping rule. This is a typical symmetric representation in which one-half on the genomic sequence is persistently enriched in two of the bases and the other half is enriched in the complementary ones. Plot B represents the *Nanoarchaeum equitans* DNA walk using the RY rule. The resulted topography is an example of random DNA walk with no persistence along the genome. Finally, plot C shows the *Prochlorococcus marinus* DNA walk grouping the bases according to SW rule. In this case, the resulted walk shows persistence along the whole genome and it can be fitted by linear regression. The abscissa axes represent the genomic sequence position from the begining to the end of the genome in all the plots. The run of DNA walks starts at position 0 of each sequence. Note that the scales are different in each graph.

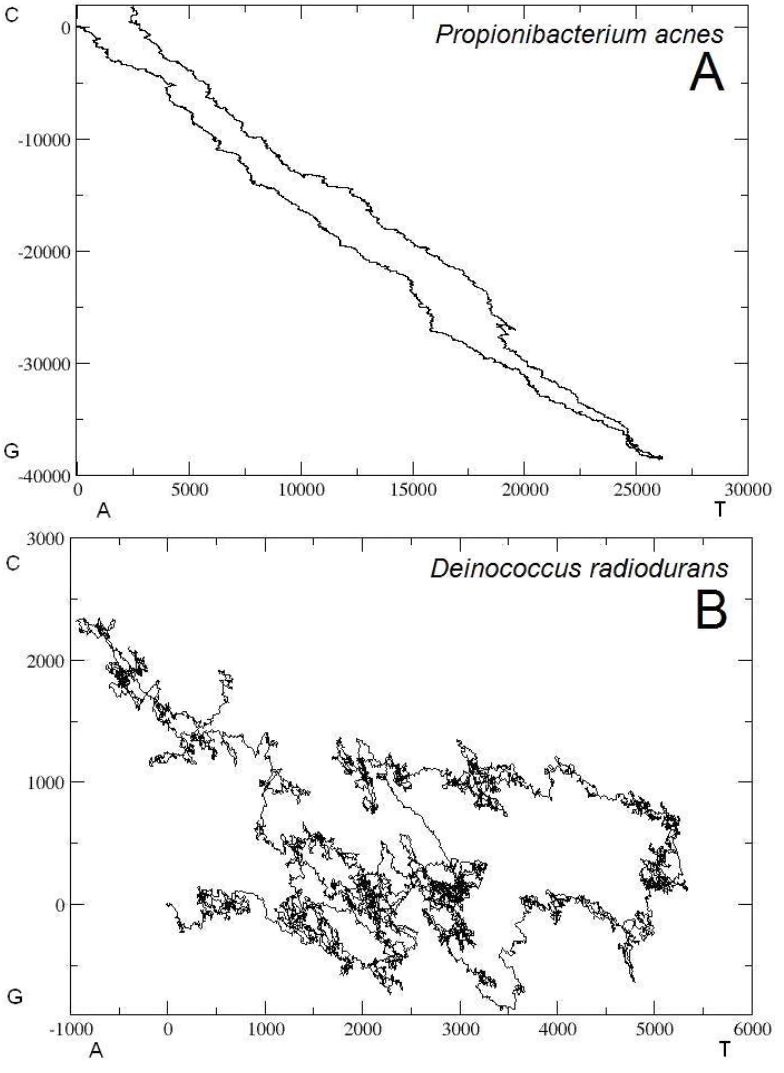


Figure A2. DNA walks representations in two-dimensional space. The positive and negative abscissa values are defined by thymine and adenine, respectivelly, whereas the positive and negative ordinate values represent cytosine and guanine, respectivelly. Plot A illustrates the *Propionibacterium acnes* DNA walk as an example of walk with a strong strand-biased genome. *Deinococcus radiodurans* DNA walk (B) is an example of weak strand-biased genome with different topography along the genome. The run of DNA walks starts at position 0 of each sequence. Note that the scales are different in each graph.

B. Detrended Fluctuation Analysis (DFA)

DFA is a scaling analysis method providing a simple integrative parameter —the scaling exponent α— to represent the correlation properties of numerical series. The scaling exponent is also called the self-similarity parameter. An object is self-similar if its subsets can be rescaled to resemble statistically the original object itself. A numerical sequence is considered stationary if the mean, standard deviation and correlation functions are invariant under space translation (Peng et al., 1992, 1994 and 1995). Sequences that do not fit these conditions are nonstationary. DFA allows detection of long-range correlations embedded in seemingly nonstationary series, and it avoids the spurious detection of apparent long-range correlations that are an artifact of nonstationarity (Hu et al., 2001). The scaling exponent quantifies the amount and range of the correlations. In a given sequence, a change in the scaling exponent indicates changes in the correlations through different scales.

Scaling exponents were calculated from the one- and two-dimensional DNA walks. In the case of one-dimensional DNA walk, the numerical series, obtained by each of the three mapping rules, were the direct input of the DFA method. On the other hand, the scaling exponents were calculated from the two-dimensional DNA walks using Euclidean distances from the origin of the graph to every x–y point representing a

step of the walk. Thus, the entire sequence of length N, understood as the numerical series obtained by each of the three mapping rules (one-dimensional walk) and the Euclidean distances for each step of the walk (two-dimensional walk), was then used to run the DFA method as follows: First, the entire sequence was divided into boxes of equal length, n, each containing l steps of the walk. We defined the "local trend" in each box by fitting a least squares linear model (proportional to the compositional bias in the box) to the data. Second, we defined the "detrended walk" as the difference between the original walk y(n) and the local trend. Next, we calculated both the variance of the detrended walk for each box and the average of these variances over all the boxes of size l, denoted F(n). Such computation was repeated over all scales (box sizes) to provide a relationship between F(n) and the box size n. Typically, F(n) increases with box size n (see example in Fig. B1). A linear relationship on a log–log graph indicates the presence of long-range correlations. Obtaining linear log–log plots of the integrated and detrended series versus "box size" (F(n) vs n) can help to establish the appropriateness of the DFA method to all nonstationary data encountered. Under these conditions, fluctuations can be characterized by the scaling exponent (α), i.e., the slope of the line relating log F(n) to log n. The minimum box size (n_{min}) does not depend on N. On the contrary, the maximum box size (n_{max}) scales as $n_{max} = N/10$ (Hu et al., 2001).

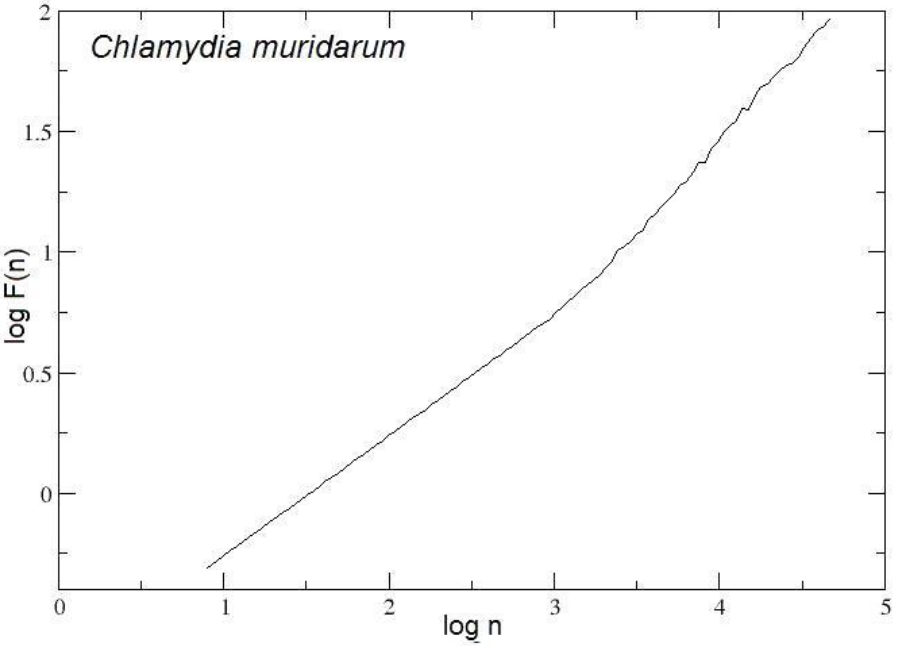


Figure B1. DFA representation calculated on the gram-negative bacterium *Chlamydia muridarum* DNA walk. The intregrated and detrended series versus "box size" (F(n) vs n) yield in a linear log–log plot whose slope is the scaling exponent α.

For an ideal sequence of infinite length, $\alpha = 0.5$ indicates the absence of long-range correlation (random walk), where the value of one nucleotide is completely uncorrelated with any previous values; whereas α different from 0.5 indicates long-range correlation. For a sample of finite length, statistical fluctuations due to finite size should be taken into account. Therefore, we considered a DNA sequence to exhibit long-range correlation only if a value was significantly different from the α value of the random finite control sequences. The α values in the range $0.5 < \alpha < 1$ indicate persistent long-range power-law correlations, suggesting the existence of repetitive patterns in the sequence and that finding a particular nucleotide on a sequential position depends on the previous nucleotides (memory).

C. Constraint-based modelling and Flux Balance Analysis

Flux Balance Analysis (FBA) is a useful mathematical technique for analysis of metabolic capabilities of cellular systems. Living organisms transform the nutrients into molecules they can use through a complex set of chemical reactions. When the whole metabolism is approximately known, one can use the FBA to find out which set of metabolic fluxes maximizes the growth rate of the organism given some known available nutrients.

The metabolic reconstruction of well-known microorganism can be generated and refined using an iterative model building procedure (see Fig. C1). The metabolic network can be expressed in a stoichiometric matrix, S (m x n), where m is the number of metabolites in the reaction network and n is the number of reactions. The corresponding entry in the stoichiometric matrix, S_{ij}, represents the stoichiometric coefficient for the participation of the *i*th metabolite in the *j*th reaction. A particular flux distribution of the network, v, indicates the flux levels through each of the reactions. Based on principles of conservation of mass and the assumption of a steady state, the flux distribution through a reaction network can be represented by the following equation:

$$S \cdot ע = 0 \qquad (1)$$

where ע (n x 1) is the vector of reaction fluxes. Additionally, constraints are imposed on individual reactions that state the upper and lower bounds on the range of flux values that each of the reactions can have. This constraint is described in the following form:

$$\alpha_i \leq ע_i \leq \beta_i \qquad (2)$$

where α_i and β_i are the lower and upper limits placed on each reaction flux $ע_i$, respectively. For reversible reactions, $-\infty \leq ע_i \leq \infty$, and for irreversible reactions, $0 \leq ע_i \leq \infty$.

The genome-scale metabolic models are normally underdetermined systems; in consequence, there are multiple solutions for ע that satisfy equation 1. To find an optimal flux distribution for ע, an objective function must be defined as a linear equation and should be optimized in the linear system. Then, a solution that satisfies all the constraints of equations 1 and 2 is calculated. The result is the optimal flux distribution that will allow the highest flux through the chosen objective reaction. The ability to produce the required components of cellular biomass (e.g., amino acids, nucleotides, phospholipids, etc.) that enable the organism to grow and survive has been defined as the objective function. This growth objective is mathematically defined as an output flux using each biomass precursor metabolite as a substrate. The ability of the network to produce all of the biomass constituents led to a positive flux value for this objective reaction.

The reconstructed metabolic network and the defined biomass function allow the calculation of network properties and optimal growth phenotypes through the use of Flux Balance Analysis. FBA allows for computation of feasible steady state fluxes through a reaction network that maximizes a particular objective and satisfies various constraints, including stoichiometry, thermodynamics and enzyme capacity. The fundamentals of FBA have been previously reviewed (Bonarius et al., 1997; Edwards et al., 2002; Edwards et al., 1999; Schilling et al., 1999; Varma and Palsson, 1994a). Specifically, FBA uses the principles of linear programming (LP), which is a subset of convex analysis. FBA was then used to solve the linear programming problem for biomass

optimization under steady state criteria (Kauffman et al., 2003; Price et al., 2004; Varma and Palsson, 1994b).

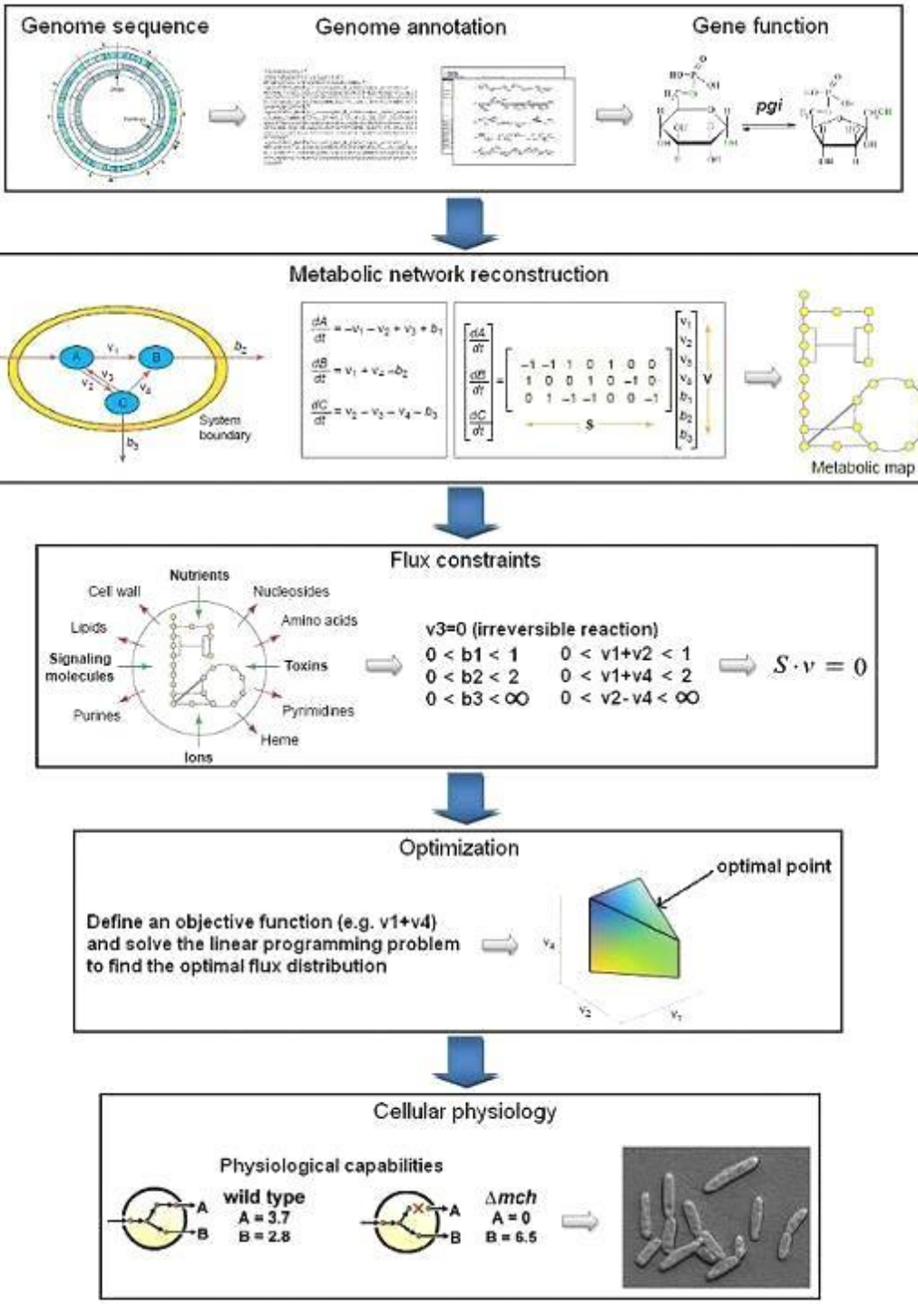

Genome sequence
Genome annotation
Gene function
pgi
Metabolic network reconstruction
System boundary
Metabolic map
Flux constraints
Nutrients
Cell wall
Nucleosides
Lipids
Amino acids
Signaling molecules
Toxins
Purines
Pyrimidines
Heme
Ions
v3=0 (irreversible reaction)
0 < b1 < 1
0 < b2 < 2
0 < b3 < ∞
0 < v1+v2 < 1
0 < v1+v4 < 2
0 < v2-v4 < ∞
$S \cdot v = 0$
Optimization
Define an objective function (e.g. v1+v4)
and solve the linear programming problem
to find the optimal flux distribution
optimal point
Cellular physiology
Physiological capabilities
wild type
A = 3.7
B = 2.8
Δmch
A = 0
B = 6.5

Figure C1. Integrated process of the microbial metabolic model construction. Such construction requires a comprehensive knowledge of the metabolism of an organism. High-throughput sequencing technology and automated genome annotation tools enable identification and functional assignment of most of the metabolic genes in an organism. From the annotated genome sequence and the experimentally determined biochemical and physiological characteristics of a cell, the metabolic reaction network can be reconstructed. Then, it can be subjected to methods such as FBA to quantitatively analyze, interpret and predict cellular behavior. Linear programming is used to determine optimal flux distributions based on objectives such as cell growth and metabolic by-product secretion. This network can be modified in the context of other physiological constraints or environmental factors to produce a mathematical model, which can be used to generate quantitatively testable hypotheses *in silico*. Modelling simulations were run under steady state conditions to determine the reaction flux distribution in the network.

D. Supplementary data

Supplementary data associated with this thesis can be found at http://nodens.ceab.csic.es/ecogenomics/docs/index.html.

Bibliography

Bibliography

Afifi A, Clark VA, May S. 2004. Discriminant analysis, Computer-aided multivariate analysis. Chapman & Hall/CRC, New York, pp. 249–279.

Akberova NI, Yu AL. 1996. Symmetrical structure in genetic texts of prokaryotes DNA replication origins. NetSci.

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. 2002. Molecular Biology of the Cell. Garland Science, New York.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Amann R, Ludwig W, Schleifer KH. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 59:143–169.

Anh VV, Lau KS, Yu ZG. 2002. Recognition of an organism from fragments of its complete genome. Phys. Rev. E 66:031910.

Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV. 1998. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet. 14:442–444.

Archer C, Vance D. 2006. Coupled Fe and S isotope evidence for Archean microbial Fe(III) and sulfate reduction. Geology 34:153–156.

Bailey JE. 1998. Mathematical modeling and analysis in biochemical engineering: Past accomplishments and future opportunities. Biotechnol. Prog. 14:8–20.

Bailey JE. 2001. Complex biology with no parameters. Nature Biotechnol. 19:503–504.

Barkai N, Leibler S. 1997. Robustness in simple biochemical networks. Nature 387:913–917.

Becker SA, Palsson BØ. 2005. Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. BMC Microbiol. 5:8.

Bernaola-Galvan P, Carpena P, Román-Roldán R, Oliver JL. 2002. Study of statistical correlations in DNA sequences. Gene 300:105–115.

Berthelsen CL, Glazier JA, Skolnick MH. 1992. Global fractal dimension of human DNA sequences treated as pseudorandom walks. Phys. Rev. A 45:8902–8913.

Blanchard JL, Lynch M. 2000. Organellar genes: why do they end up in the nucleus? Trends Genet. 16:315–320.

Bonarius HPJ, Schmid G, Tramper J. 1997. Flux analysis of underdetermined metabolic networks: The quest for the missing constraints. Trends Biotechnol. 15:308–314.

Boussau B, Karlberg EO, Frank AC, Legault BA, Andersson SGE. 2004. Computational inference of scenarios for α-proteobacterial genome evolution. PNAS 101:9722–9727.

Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Matsa ME, Peng CK, Simons M, Stanley HE. 1995. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. Phys. Rev. E 51:5084–5091.

Burgard AP, Pharkya P, Maranas CD. 2003. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol. Bioeng. 84:647–657.

Casamayor EO, Pedrós-Alió C, Muyzer G, Amann R. 2002. Microheterogeneity in 16S rDNA-defined bacterial populations from a stratified planktonic environment is related to temporal succession and to ecological adaptations. Appl. Environ. Microbiol. 68:1706–1714.

Cavalier-Smith T, Brasier M, Embley TM. 2006. Introduction: how and when did microbes change the world? Phil. Trans. R. Soc. B 361:845–850.

Cebrat S, Dudek MR. 1998. The effect of DNA phase structure on DNA walks. Eur. Phys. J. B, Cond. Matter Phys. 3:271–276.

Chatzidimitriou-Dreismann CA, Larhammar D. 1993. Long-range correlations in DNA. Nature 361:212–213.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward Automatic Reconstruction of a Highly Resolved Tree of Life. Science 311:1283–1287.

Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BØ. 2004. Integrating high-throughput and computational data elucidates bacterial networks. Nature 429:92–96.

Covert MW, Schilling CH, Famili I, Edwards JS, Goryanin II, Selkov E, Palsson BØ. 2001. Metabolic modeling of microbial strains *in silico*. Trends Biochem. Sci. 26:179–186.

Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. 1999. Alignment of whole genomes. Nucleic Acids Res. 27:2369–2376.

De Long EF. 2004. Microbial population genomics and ecology: the road ahead. Environ. Microbiol. 6:875–878.

Doolittle WF. 1999a. Lateral genomics.Trends Cell Biol. 9:M5–M8.

Doolittle WF. 1999b. Phylogenetic classification and the universal tree. Science 284:2124–2129.

Duarte NC, Herrgard MJ, Palsson BØ. 2004. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. Genome Res. 14:1298–1309.

Dyksterhouse SE, Gray JP, Herwig RP, Lara JC, Staley JT. 1995. *Cycloclasticus pugetii* gen. nov., sp. nov., an aromatic hydrocarbon degrading bacterium from marine sediments. Int. J. Syst. Bacteriol. 45:116–123.

Edwards JS. 1999. Functional genomics and the computational analysis of bacterial metabolism. San Diego Department of Bioengineering, University of California.

Edwards JS, Covert M, Palsson BØ. 2002. Metabolic modelling of microbes: the flux-balance approach. Environ. Microbiol. 4:133–140.

Edwards JS, Palsson BØ. 1999. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. J. Biol. Chem. 274:17410–17416.

Edwards JS, Ramakrishna R, Schilling CH, Palsson BØ. 1999. Metabolic flux balance analysis, in Metabolic engineering. Lee SY, Papoutsakis ET (eds.), Marcel Dekker, New York, pp. 13–57.

Elston RC, Wilson AF. 1990. Genetic linkage and complex disease: a comment. Genet. Epidemiol. 7:17–19.

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ. 2007. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol. Syst. Biol. 3:121.

Feist AM, Scholten JCM, Palsson BØ, Brockman FJ, Ideker T. 2006. Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. Mol. Syst. Biol. 2:2006.0004.

Fischer E, Sauer U. 2005. Large-scale in vivo flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. Nat. Genet. 37:636–640.

Floyd MM, Tang J, Kane M, Emerson D. 2005. Captured diversity in a culture collection: case study of the geographic and habitat distributions of environmental isolates held at the American Type Culture Collection. Appl. Environ. Microbiol. 71:2813–2823.

Foerstner KU, von Mering C, Hooper SD, Bork P. 2005. Environments shape the nucleotide composition of genomes. EMBO Rep. 6:1208–1213.

Galtier N, Lobry JR. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J. Mol. Evol. 44:632–636.

García JAL, Bartumeus F, Roche D, Giraldo J, Stanley HE, Casamayor EO. 2008. Ecophysiological significance of scale-dependent patterns in prokaryotic genomes unveiled by a combination of statistic and genometric analyses. Genomics 91:538–543.

Garrity GM, Boone DR, Castenholz RW. 2001. Bergey's manual of systematic bacteriology, 2nd ed. Springer, Berlin.

Gauthier MJ, Lafay B, Christen R, Fernandez L, Acquaviva M, Bonin P, Bertrand JC. 1992. *Marinobacter hydrocarbonoclasticus* gen. nov., sp. nov., a new, extremely halotolerant, hydrocarbon-degrading marine bacterium. Int. J. Syst. Bacteriol. 42:568–576.

Golyshin PN, Chernikova TN, Abraham WR, Lünsdorf H, Timmis KN, Yakimov MM. 2002. *Oleiphilaceae* fam. nov., to include *Oleiphilus messinensis* gen. nov., sp. nov., a novel marine bacterium that obligately utilizes hydrocarbons. Int. J. Syst. Evol. Microbiol. 52:901–911.

Bibliography

Golyshin PN, Martins Dos Santos VAP, Kaiser O, Ferrer M, Sabirova YS, Lünsdorf H, Chernikova TN, Golyshina OV, Yakimov MM, Pühler A, Timmis KN. 2003. Genome sequence completed of *Alcanivorax borkumensis*, a hydrocarbon-degrading bacterium that plays a global role in oil removal from marine systems. J. Biotechnol. 106:215–220.

Grigoriev A. 1998. Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. 26:2286–2290.

Grosberg A, Rabin Y, Havlin S, Neer A. 1993. Crumpled globule model of the three-dimensional structure of DNA. Europhys. Lett. 23:373–378.

Hara A, Syutsubo K, Harayama S. 2003. *Alcanivorax* which prevails in oil-contaminated seawater exhibits broad substrate specificity for alkane degradation. Environ. Microbiol. 5:746–753.

Harayama S, Kishira H, Kasai Y, Shutsubo K. 1999. Petroleum biodegradation in marine environments. J. Mol. Microbiol. Biotechnol. 1:63–70.

Hayes JM, Waldbauer JR. 2006. The carbon cycle and associated redox processes through time. Phil. Trans. R. Soc. B 361:931–950.

Head IM, Saunders JR, Pickup RW. 1998. Microbial evolution, diversity, and ecology: a decade of ribosomal RNA analysis of uncultivated microorganisms. Microb. Ecol. 35:1–21.

Heddi A, Charles H, Khatchadourian C, Bonnot G, Nardon P. 1998. Molecular characterization of the principal symbiotic bacteria of the weevil *Sitophilus oryzae*: a peculiar G+C content of an endocytobiotic DNA. J. Mol. Evol. 47:52–61.

Heinemann M, Kümmel A, Ruinatscha R, Panke S. 2005. *In silico* genome-scale reconstruction and validation of the *Staphylococcus aureus* metabolic network. Biotechnol. Bioeng. 92:850–864.

Herrero A, Flores E (eds.). 2008. The Cyanobacteria: Molecular Biology, Genomics and Evolution, 1st ed. Caister Academic Press.

Ho JWK, Adams CE, Lew JB, Matthews TJ, Ng CC, Shahabi-Sirjani A, Tan LH, Zhao Y, Easteal S, Wilson SR, Jermiin LS. 2006. SeqVis: Visualization of compositional heterogeneity in large alignments of nucleotides. Bioinformatics 22:2162–2163.

Hong SH, Kim JS, Lee SY, In YH, Choi SS, Rih JK, Kim CH, Jeong H, Hur CG, Kim JJ. 2004. The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens*. Nat. Biotechnol. 22:1275–1281.

Horz HP, Barbrook A, Field CB, Bohannan BJM. 2004. Ammonia-oxidizing bacteria respond to multifactorial global change. PNAS 101:15136–15141.

Hotelling H. 1936. Relations between two sets of variates. Biometrika 28:321–377.

Hu K, Ivanov PC, Chen Z, Carpena P, Stanley HE. 2001. Effect of trends on detrended fluctuation analysis. Phys. Rev. E 64:011114.

Hurst LD, Merchant AR. 2001. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc. Biol. Sci. 268:493–497.

Jacobs KL, Grogan DW. 1997. Rates of spontaneous mutation in an archaeon from geothermal environments. J. Bacteriol. 179:3298–3303.

Kalscheuer R, Stöveken T, Malkus U, Reichelt R, Golyshin PN, Sabirova JS, Ferrer M, Timmis KN, Steinbüchel A. 2007. Analysis of storage lipid accumulation in *Alcanivorax borkumensis*: evidence for alternative triacylglycerol biosynthesis routes in bacteria. J. Bacteriol. 189:918–928.

Kandler O, König H. 1998. Cell wall polymers in Archaea (Archaebacteria). Cell. Mol. Life Sci. 54:305–308.

Kanehisa M, Goto S, Kawashima S, Nakaya A. 2002. The KEGG databases at GenomeNet. Nucleic Acids Res. 30:42–46.

Kasai Y, Kishira H, Sasaki T, Syutsubo K, Watanabe K, Harayama S. 2002. Predominant growth of *Alcanivorax* strains in oil-contaminated and nutrient-supplemented sea water. Env. Microbiol. 4:141–147.

Kasai Y, Kishira H, Syutsubo K, Harayama S. 2001. Molecular detection of marine bacterial populations on beaches contaminated by the Nakhodka tanker oil-spill accident. Environ. Microbiol. 3:246–255.

Kauffman KJ, Prakash P, Edwards JS. 2003. Advances in flux balance analysis. Curr. Opin. Biotechnol. 14:491–496.

Kelman LM, Kelman Z. 2004. Multiple origins of replication in archaea. Trends Microbiol. 12:399–401.

Kolmogorov AN. 1961. The local structure of turbulence in incompressible fluids for very large Reynolds´ numbers, in Turbulence classic papers on statistical theory. Friedlander SK, Topper L (eds.), Wiley-Interscience, Nueva York.

Konstantinidis KT, James M, Tiedje JM. 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. PNAS 101:3160–3165.

Koonin EV. 2000. How many genes can make a cell: The minimal-gene-set concept. Annu. Rev. Genomics Hum. Genet. 1:99–116.

Koonin EV, Aravind L, Kondrashov AS. 2000. The Impact of comparative genomics on our understanding of evolution. Cell 101:573–576.

Koonin EV, Mushegian AR, Galperin MY, Walker DR. 1997. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. Mol. Microbiol. 25:619–637.

Krumsiek J, Arnold R, Rattei T. 2007. Gepard: A rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23:1026–1028.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5:R12.

Li W, Kaneko K. 1992. Long-range correlation and partial $1/f^{\alpha}$ spectrum in a non-coding DNA sequence. Europhys. Lett. 17:655–660.

Li W, Marr TG, Kaneko K. 1994. Understanding long-range correlations in DNA sequences. Physica D 75:392–416.

Liao JC. 1993. Modelling and analysis of metabolic pathways. Curr. Opin. Biotechnol. 4:211–216.

Lindahl T. 1993. Instability and decay of the primary structure of DNA. Nature 362:709–715.

Lindner SN, Vidaurre D, Willbold S, Schoberth SM, Wendisch VF. 2007. NCgl2620 encodes a class II polyphosphate kinase in *Corynebacterium glutamicum*. Appl. Environ. Microbiol. 73:5026–5033.

Lobry JR. 1996a. A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. Biochimie 78:323–326.

Lobry JR. 1996b. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13:660–665.

Lobry JR. 1999. Genomic landscapes. Microbiol. Today 26:164–165.

Logsdon JM, Faguy DM. 1999. Evolutionary genomics: *Thermotoga* heats up lateral gene transfer. Curr. Biol. 9:747–751.

Mackiewicz P, Kowalczuk M, Mackiewicz D, Nowicka A, Dudkiewicz M, Laszkiewicz A, Dudek MR, Cebrat S. 2002. Replication associated mutational pressure generating long-range correlation in DNA. Physica A 314:646–654.

Makarova KS, Sorokin AV, Novichkov PS, Wolf YI, Koonin EV. 2007. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. Biol. Direct 2:33.

Makarova KS, Wolf YI, Koonin EV. 2003. Potential genomic determinants of hyperthermophily. Trends Genet. 19:172–176.

Mandelbrot BB. 1982. The fractal geometry of nature. Freeman, San Francisco.

Mandelbrot BB, Van Ness JW. 1968. Fractional Brownian motions, fractional noises and applications. SIAM Rev. 10:422–437.

Marashi SA, Ghalanbor Z. 2004. Correlations between genomic GC levels and optimal growth temperatures are not 'robust'. Biochem. Biophys. Res. Commun. 325:381–383.

Margesin R, Schinner F. 1999. Biological decontamination of oil spills in cold environments. J. Chem. Technol. Biotechnol. 74:381–389.

McGrady-Steed J, Harris P, Morin P. 1997. Biodiversity regulates ecosystem predictability. Nature 390:162–165.

McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I. 2007. Accurate phylogenetic classification of variable-length DNA fragments. Nat. Methods 4:63–72.

Mlot C. 2004. Microbial diversity unbound: what DNA-based techniques are revealing about the planet's hidden biodiversity. Bioscience 54:1064–1068.

Moran NA. 2002. Microbial minimalism: genome reduction in bacterial pathogens. Cell 108:583–586.

Mrázek J, Karlin S. 1998. Strand compositional asymmetry in bacterial and large viral genomes. PNAS 95:3720–3725.

Musto H, Naya H, Zavala A, Romero H, Alvarez-Valín F, Bernardi G. 2004. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Lett. 573:73–77.

Musto H, Naya H, Zavala A, Romero H, Alvarez-Valín F, Bernardi G. 2006. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochem. Biophys. Res. Commun. 347:1–3.

Muto A, Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. PNAS 84:166–169.

Nagai N, Kuwata K, Hayashi T, Kuwata H, Era S. 2001. Evolution of the periodicity and the self-similarity in DNA sequence: a Fourier transform analysis. Jpn. J. Physiol. 51:159–168.

Nanchen A, Schicker A, Sauer U. 2006. Nonlinear dependency of intracellular fluxes on growth rate in miniaturized continuous cultures of *Escherichia coli*. Appl. Environ. Microbiol. 72:1164–1172.

Nelson KE. 2003. The future of microbial genomics. Environ. Microbiol. 5:1223–1225.

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM. 1999. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399:323–329.

Nijland R, Veening JW, Kuipers OP. 2007. A derepression system based on the *Bacillus subtilis* sporulation pathway offers dynamic control of heterologous gene expression. Appl. Environ. Microbiol. 73:2390–2393.

Nogales J, Palsson BØ, Thiele I. 2008. A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: iJN746 as a cell factory. BMC Syst. Biol. 2:79.

Oberhardt MA, Puchałka J, Fryer KE, Martins dos Santos VAP, Papin JA. 2008. Genome-scale metabolic network analysis of the opportunistic pathogen *Pseudomonas aeruginosa* PAO1. J. Bacteriol. 190:2790–2803.

Oh YK, Palsson BØ, Park SM, Schilling CH, Mahadevan R. 2007. Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. J. Biol. Chem. 282:28791–28799.

Oliveira AP, Nielsen J, Förster J. 2005. Modeling *Lactococcus lactis* using a genome-scale flux model. BMC Microbiol. 5:39.

Olson JM. 2006. Photosynthesis in the Archean era. Photosyn. Res. 88:109–117.

Omelchenko MV, Wolf YI, Gaidamakova EK, Matrosova VY, Vasilenko A, Zhai M, Daly MJ, Koonin EV, Makarova KS. 2005. Comparative genomics of *Thermus thermophilus* and *Deinococcus radiodurans*: divergent routes of adaptation to thermophily and radiation resistance. BMC Evol. Biol. 5:57.

Øvreås L. 2000. Population and community level approaches for analysing microbial diversity in natural environments. Ecol. Lett. 3:236–251.

Palsson BØ. 2000. The challenges of *in silico* biology. Nature Biotechnol. 18:1147–1150.

Palsson BØ, Lee I. 1993. Model complexity has a significant effect on the numerical value and interpretation of metabolic sensitivity coefficients. J. Theoret. Biol. 161:299–315.

Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH, Ussery DW. 2000. A DNA structural atlas for *Escherichia coli*. J. Mol. Biol. 299:907–930.

Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Simons M, Stanley HE. 1995. Statistical properties of DNA sequences. Physica A 221:180–192.

Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, Stanley HE. 1992. Long-range correlations in nucleotide sequences. Nature 356:168–170.

Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL. 1994. Mosaic organization of DNA nucleotides. Phys. Rev. E 49:1685–1689.

Pollack JD. 2002. The necessity of combining genomic and enzymatic data to infer metabolic function and pathways in the smallest bacteria: Amino acid, purine and pyrimidine metabolism in Mollicutes. Front. Biosci. 7:d1762–d1781.

Price ND, Reed JL, Palsson BØ. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat. Rev. Microbiol. 2:886–897.

Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. 2003. Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Res. 13:145–158.

Pushker R, Mira A, Rodríguez-Valera F. 2004. Comparative genomics of gene-family size in closely related bacteria. Genome Biol. 5:R27.

Rao CR. 1964. The use and interpretation of principal component analysis in applied research. Sankhya A 26:329–358.

Reed JL, Famili I, Thiele I, Palsson BØ. 2006. Towards multidimensional genome annotation. Nat. Rev. Genet. 7:130–141.

Reed JL, Palsson BØ. 2003. Thirteen years of building constraint-based *in silico* models of *Escherichia coli*. J. Bacteriol. 185:2692–2699.

Reva ON, Hallin PF, Willenbrock H, Sicheritz-Ponten T, Tümmler B, Ussery DW. 2007. Global features of the *Alcanivorax borkumensis* SK2 genome. Env. Microbiol. 10:614–625.

Rocha EP. 2002. Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol. 10:393–395.

Rocha EP, Danchin A. 2002. Base composition bias might result from competition for metabolic resources. Trends Genet. 18:291–294.

Röling WFM, Milner MG, Jones DM, Fratepietro F, Swannell RPJ, Daniel F, Head IM. 2004. Bacterial community dynamics and hydrocarbon degradation during a field-scale evaluation of bioremediation on a mudflat beach contaminated with buried oil. Appl. Environ. Microbiol. 70:2603–2613.

Roten CAH, Gamba P, Barblan JL, Karamata D. 2002. Comparative Genometrics (CG): a database dedicated to biometric comparisons of whole genomes. Nucleic Acids Res. 30:142–144.

Sabirova JS. 2006. Functional genome analysis of *Alcanivorax borkumensis* strain SK2: alkane metabolism, environmental adaptations and biotechnological potential. Technical University Carolo-Wilhelmin, Braunschweig.

Sabirova JS, Ferrer M, Lunsdorf H, Wray V, Kalscheuer R, Steinbuchel A, Timmis KN, Golyshin PN. 2006a. Mutation in a "tesB-like" hydroxyacyl-coenzyme A-specific thioesterase gene causes hyperproduction of extracellular polyhydroxyalkanoates by *Alcanivorax borkumensis* SK2. J. Bacteriol. 188:8452–8459.

Sabirova JS, Ferrer M, Regenhardt D, Timmis KN, Golyshin PN. 2006b. Proteomic insights into metabolic adaptations in *Alcanivorax borkumensis* induced by alkane utilization. J. Bacteriol. 188:3763–3773.

Sauer U, Hatzimanikatis V, Bailey JE, Hochuli M, Szyperski T, Wuthrich K. 1997. Metabolic fluxes in riboflavin-producing *Bacillis subtilis*. Nature Biotechnol. 15:448–452.

Sauer U, Lasko DR, Fiaux J, Hochuli M, Glaser R, Szyperski T, Wüthrich K, Bailey JE. 1999. Metabolic flux ratio analysis of genetic and environmental modulations of *Escherichia coli* central carbon metabolism. J. Bacteriol. 181:6679–6688.

Schilling CH, Covert MW, Famili I, Church GM, Edwards JS, Palsson BØ. 2002. Genome-scale metabolic model of *Helicobacter pylori* 26695. J. Bacteriol. 184:4582–4593.

Schilling CH, Edwards JS, Palsson BØ. 1999. Towards metabolic phenomics: analysis of genomic data using flux balances. Biotechnol. Prog. 15:288–295.

Schilling CH, Palsson BØ. 2000. Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. J. Theor. Biol. 203:249–283.

Schloss PD, Handelsman J. 2004. Status of the microbial census. Microbiol. Mol. Biol. Rev. 68:686–691.

Schneiker S, Martins Dos Santos VA, Bartels D, Bekel T, Brecht M, Buhrmester J, Chernikova TN, Denaro R, Ferrer M, Gertler G, Goesmann A, Golyshina OV, Kaminski F, Khanane AN, Lang S, Linke B, McHardy AC, Meyer F, Nechitaylo T, Puhler A, Regenhardt D, Rupp O, Sabirova JS, Selbitschka W, Yakimov MM, Timmis KN, Vorholter FJ, Weidner S, Kaiser O, Golyshin PN. 2006. Genome sequence of the ubiquitous hydrocarbon-degrading marine bacterium *Alcanivorax borkumensis*. Nat. Biotechnol. 24:997–1004.

Schopf JW, Kudryavtsev AB, Agresti DG, Wdowiak TJ, Czaja AD. 2002. Laser-Raman imagery of Earth's earliest fossils. Nature 416:73–76.

Sen A, Sur S, Bothra AK, Benson DR, Normand P, Tisa LS. 2008. The implication of life style on codon usage patterns and predicted highly expressed genes for three *Frankia* genomes. Antonie van Leeuwenhoek 93:335–346.

Shuler ML, Domach MM. 1983. Mathematical models of the growth of individual cells, in Foundations of biochemical engineering. Blanch HW, Papoutsakis ET, Stephanopoulos G (eds.), American Chemical Society, Washington, D.C.

Simonato F, Campanaro S, Lauro FM, Vezzi A, D'Angelo M, Vitulo N, Valle G, Bartlett DH. 2006. Piezophilic adaptation: a genomic point of view. J. Biotechnol. 126:11–25.

Stanley HE, Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Peng CK, Simons M. 1996. Scale invariant features of coding and non coding DNA sequences, in Fractal geometry in biological systems: an analytical approach. Iannaconne PM, Khokha M (eds.), CRC Press, Boca Raton, FL, pp. 15–30.

Steinbüchel A. 1991. Polyhydroxyalkanoic acids, in Biomaterials: novel biomaterials from biological sources. Byrom ID (ed.), MacMillan Publishers, Basingstoke, United Kingdom, pp. 123–213.

Streit WR, Schmitz RA. 2004. Metagenomics —the key to the uncultured microbes. Curr. Opin. Microbiol. 7:492–498.

Summons RE, Bradley AS, Jahnke LL, Waldbauer JR. 2006. Steroids, triterpenoids and molecular oxygen. Phil. Trans. R. Soc. B 361:951–968.

Syutsubo K, Kishira H, Harayama S. 2001. Development of specific oligonucleotide probes for the identification and in situ detection of hydrocarbon-degrading *Alcanivorax* strains. Environ. Microbiol. 3:371–379.

Takada Y, Fukunaga N, Sasaki S. 1991. Modification of membrane lipids of a psychrophilic marine bacterium. J. Fac. Sci. Hokkaido Uni. Ser. V 15:1–13.

Tännler S, Decasper S, Sauer U. 2008. Maintenance metabolism and carbon fluxes in *Bacillus* species. Microb. Cell Fact. 7:19.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278:631–637.

Teeling H, Meyerdierks A, Bauer M, Amann R, Glockner FO. 2004. Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ. Microbiol. 6:938–947.

Ter Braak CJF. 1988. CANOCO - a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal component analysis and redundancy analysis (version 2.1). Wageningen.

Toh H, Weiss BL, Perkin SAH, Yamashita A, Oshima K, Hattori M, Aksoy S. 2006. Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. Genome Res. 16:149–156.

Tomita M, Hashimoto K, Takahashi K, Shimizu TS, Matsuzaki Y, Miyoshi F, Saito K, Tanida S, Yugi K, Venter JC, Hutchison CA. 1999. E-CELL: software environment for whole-cell simulation. Bioinformatics 15:72–84.

Vallino JJ, Stephanopoulos G. 1993. Metabolic flux distributions in *Corynebacterium glutamicum* during growth and lysine overproduction. Biotechnol. Bioeng. 41:633–646.

Van Dien SJ, Lidstrom ME. 2002. Stoichiometric model for evaluating the metabolic capabilities of the facultative methylotroph *Methylobacterium extorquens* AM1, with application to reconstruction of C-3 and C-4 metabolism. Biotechnol. Bioeng. 78:296–312.

Varma A, Palsson BØ. 1994a. Metabolic flux balancing: basic concepts, scientific and practical use. Biotechnol. 12:994–998.

Varma A, Palsson BØ. 1994b. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. Appl. Environ. Microbiol. 60:3724–3731.

Varner J, Ramkrishna D. 1999. Metabolic engineering from a cybernetic perspective. 1. Theoretical preliminaries. Biotechnol. Prog. 15:407–425.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74.

Vieira MS. 1999. Statistics of DNA sequences: a low-frequency analysis. Phys. Rev. E 60:5932–5937.

von Wintzingerode F, Göbel U, Stackebrandt E. 1997. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiol. Rev. 21:213–229.

Voss RF. 1992. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys. Rev. Lett. 68:3805–3808.

Wang HC, Susko E, Roger AJ. 2006. On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors. Biochem. Biophys. Res. Commun. 342:681–684.

Ward DM, Weller R, Bateson MM. 1990. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature 345:63–65.

Ward N, Fraser CM. 2005. How genomics has affected the concept of microbiology. Curr. Opin. Microbiol. 8:564–571.

Whitman W, Coleman D, Wiebe W. 1998. Prokaryotes: the unseen majority. PNAS 95:6578–6583.

Wiechert W, de Graaf AA. 1996. In vivo stationary flux analysis by 13C labeling experiments. Adv. Biochem. Eng./Biotechnol. 54:109–154.

Wilde SA, Valley JW, Peck WH, Graham CM. 2001. Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago. Nature 409:175–178.

Woese CR. 1987. Bacterial evolution. Microbiol. Rev. 51:221–271.

Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucarya. PNAS 87:4576–4579.

Worning P, Jensen LJ, Hallin PF, Stærfeldt HH, Ussery DW. 2006. Origin of replication in circular prokaryotic chromosomes. Environ. Microbiol. 8:353–361.

Yakimov MM, Denaro R, Genovese M, Cappello S, D'Auria G, Chernikova TN, Timmis KN, Golyshin PN, Giluliano L. 2005. Natural microbial diversity in superficial sediments of Milazzo Harbor (Sicily) and community successions during microcosm enrichment with various hydrocarbons. Environ. Microbiol. 7:1426–1441.

Yakimov MM, Giuliano L, Gentile G, Crisafi E, Chernikova TN, Abraham WR, Lünsdorf H, Timmis KN, Golyshin PN. 2003. *Oleispira antarctica* gen. nov., sp. nov., a novel hydrocarbonoclastic marine bacterium isolated from Antarctic coastal sea water. Int. J. Syst. Evol. Microbiol. 53:779–785.

Yakimov MM, Golyshin PN, Lang S, Moore ER, Abraham WR, Lunsdorf H, Timmis KN. 1998. *Alcanivorax borkumensis* gen. nov., sp. nov., a new, hydrocarbon-degrading and surfactant-producing marine bacterium. Int. J. Syst. Bacteriol. 48:339–348.

Yeh I, Hanekamp T, Tsoka S, Karp PD, Altman RB. 2004. Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery. Genome Res. 14:917–924.

Yu ZG, Anh V, Lau KS, Chu KH. 2003. The genomic tree of living organisms based on a fractal model. Phys. Lett. A 317:293–302.