Understanding the molecular interplay between bacterial pathogens and hosts: an evolutionary approach

Thesis submitted to The University of Dublin for the degree of

Doctor of Philosophy

2012

Xiaowei Jiang

B.Eng., M.Sc.

Supervisor: Dr. Mario A. Fares

Department of Genetics

Trinity College Dublin

Declaration

This thesis is submitted by the undersigned for the degree of Doctor of Philosophy at the University of Dublin and has not been submitted as an exercise for a degree at this or any other University. Except where otherwise stated, it is entirely the candidate's own work. The candidate agrees that the Library may lend or copy the thesis upon request. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Summary

The emergence of ecological adaptations is a fundamental conundrum in evolutionary biology. A magnificent and astonishing diversity of life forms occupy a myriad of ecological niches, from stable intra-cellular nutrition-rich environments to harsh ones. How do these forms emerge? The aim of my PhD is to contribute to unearthing the evolutionary rules of ecological innovation by focusing on particular sets of molecules that mediate the interaction of organisms with the environment. We focused on the understanding of two questions: (i) what is the role of secretion systems in the emergence of ecological adaptations in prokaryotes, and in particular in pathogenic bacteria? and (ii) how does the environment react to these adaptations, particularly how do hosts counteract the different adaptation mechanisms of pathogens? The interplay between the plasticity of the environment, be it a chemical environment or an organism with a greater biological complexity, and that of the occupying organisms has generated a complex dynamic of evolution and counter-evolution responsible for the great biological diversity we see today on Earth.

To understand the role of secretion systems in the emergence of evolutionary novelties, we hypothesise that functional diversification of bacterial secretory machineries is an important step for subsequent ecological innovation, as secreted components establish the link with the environment, therefore playing an essential role in the localisation of proteins that mediate cell-environment or pathogen-host interaction. Based on the above assumption, we focused on the understanding of the evolution of secretion systems in bacteria by looking for signatures of protein functional diversification in the transition to a pathogenic lifestyle. To study functional diversification at the proteome level, in chapter 2 we developed a novel method that identifies signatures of functional divergence (FD)–understood as the process by which new genes and functions originate through the modification of existing ones. Applying this method made it possible to analyse the FD of 750 bacterial proteomes. This permitted the direct testing of signatures of FD linked to bacterial lifestyle because the set of bacterial proteomes tested included psychrophiles, thermophiles, pathogens and symbionts. We demonstrated that proteins that belong to carbohydrate transport and metabolism and cell wall/membrane/envelope biogenesis were the most functionally divergent, which is in line with our previous hypothesis that secreted proteins play an important role in mediating ecological adaptation. In contrast to this, proteins that belong to bacterial secretion systems showed relatively weak FD. To further address the role of secretion systems in ecological adaptation, in chapter 3 and 4 we analysed, using our previously developed approach, two evolutionarily conserved secretion systems in prokaryotic and eukaryotic cells: the general secretory (Sec) system and the twin-arginine translocation (Tat) pathway. We revealed two general patterns of FD that affected the most important Sec system proteins, the motor protein SecA and the channel of Sec translocase SecY. Our results suggest that FD has affected SecA in a way related to its ability to utilise adenosine triphosphate (ATP), with this having dramatic consequences on the transport of a variety of proteins, including short to medium sized proteins and extremely large cell surface proteins. With regard to the FD patterns in SecY, this mainly affected the SecY cytoplasmic regions (C1, C4 and C6), that play an important role in SecA-dependent posttranslational translocation and ribosome-dependent cotranslational translocation. These patterns of FD may also contribute to translocate particular sets of proteins for lineage-specific niche adaptations. Finally, we demonstrated that in halophilic and thermophilic archaea the FD of the core components of the Tat pathway, TatC and TatA, respectively, play an essential role in their ecological adaptation, as also evidenced by their Tat-dependent secretome analysis. Moreover, FD and secretome analysis of pathogenic bacteria revealed novel insights into bacterial evasion of host immune response.

To address how hosts in particular counteract bacterial invasions, we aimed to identify signatures of reciprocal molecular evolution (also known as coevolution) between hosts and pathogens. To do so, we focused on the proteins mediating the main host immune response mechanisms: the protein complex formed by the major histocompatibility complex (that includes two major classes: MHCI and MHCII), antigenic peptides derived from pathogens (p), T cell receptor (TCR) and co-receptor CD8/CD4 expressed on T cell surface (TCR/pMHCI/CD8 or TCR/pMHCII/CD4). This complex of proteins plays a central role in host adaptive immunity, that performs two types of effector mechanisms of the immune system: one is the cytotoxic T cell mediated apoptosis of host cells infected with intracellular pathogens, while the other is the antibody mediated host immunity based on the activation of B cells by helper T cells. In chapter 5 we identified the coevolutionary patterns within and between human leukocyte antigen class I (HLAI, human MHCI) and class II (HLAII, human MHCII) allele protein groups and we showed that most of the identified coevolving residues are grouped within the antigen recognition site and have undergone strong positive Darwinian selection. Our results revealed that there are correlated amino acid changes across the whole HLA proteins, from TCR-antigen binding domain to CD8 and CD4 coreceptor binding domains to transmembrane and cytoplasmic domains. This demonstrates the complex coevolutionary dynamics on MHCI and MHCII molecules. Our results suggest that the coevolutionary patterns operating on HLAI and HLAII molecules are the result of host–pathogen coevolution and cooperative binding of TCR, antigenic peptides, and CD8 and CD4 to HLA molecules. Moreover, our results point to that disruption of the optimised interactions between and among coevolving residues could lead to different levels of disease resistance and susceptibility (e.g., progression to AIDS).

The combined results of this project paves the way for future projects that aim to study the interaction and coevolution of pathogens and their hosts.

Acknowledgements

Firstly, I would like to thank the Irish Research Council for Science, Engineering and Technology (IRC-SET) for awarding me the EMBARK scholarship, which allows me to complete my postgraduate training and this thesis at The University of Dublin, Trinity College.

Secondly, I thank Dr. Mario A. Fares for giving me the opportunity to work in his lab; and I appreciate his trust and support during my PhD project. I also would like to thank my colleagues Tom and Brian for their helpful discussion and other current and former members of the Fares lab, Jenny, Orla, Christina,Valentine and Damien. Brian and Eiseart helped me proofread my first chapter.

I would especially like to thank my wife Wei for her endless support while writing this thesis. My daughter Yuhang was born in the second year of my PhD. Wei has been taking care of the baby, for which I should take at least equal share of parenting responsibility and I was not able to. I am forever in debt to Wei.

Finally, I would like to thank my parents and my mother- and father-in-law for their help with the baby and support during this project.

Xiaowei Jiang
Trinity College Dublin
2011

To my parents

To my wife and daughter

Contents

List of Tables

List of Figures

“Therefore one hundred victories in one hundred battles is not the most skillful. Seizing the enemy without fighting is the most skillful.”

Sun Tzu, The Art of War

Chapter 1

Introduction

1.1 Symbiosis: The basis of combinatorial evolution

Darwin acknowledged the slow pace at which natural selection acts: "that natural selection will always act with extreme slowness I fully admit" (Darwin, 1859). This observation was based upon the fact that selection allows fixing of new living forms in a gradual way. The astonishing diversity of living forms on earth is striking given the slow speed at which evolution occurs. Darwin recognised that new forms emerge as the result of the modification of ancestral ones and this established a quasi-linear relationship between diversity and time. How then do complex forms emerge so subtly in nature?

In 1879, Anton de Bary first used the term "symbiosis" to refer to different organisms that live together. This denotation was fundamental to conceive evolution and the emergence of biological complexity as the outcome of combining living forms with different complexities into one form. Indeed, symbiosis has been a major player in the emergence of complexity and has been the hallmark of eukaryotic cells' evolution. How does symbiosis emerge and what kind of interactions can take place between organisms that could drive complex evolution? The definition of symbiosis by Anton de Bary did not distinguish between the three organismal relations: parasitism, commensalism and mutualism, in the sense that the three relationships were included as particularities of symbiosis (Paracer & Ahmadjian, 2000). The three relationships, however, involve different types of organismal interactions.

Commensalism is established between two organisms when the relationship between them is neither harmful nor beneficial. Mutualistic symbiosis takes place when both of the symbiontic partners get some benefit of the relationship, in many cases such benefit leads to the metabolic marriage of the two or more involved organisms (Moya et al., 2008). Finally, parasitism refers to a relation in which one of the symbiotic organisms benefits at the cost of the other, although a parasitic organism may not necessarily be pathogenic to its host (Paracer & Ahmadjian, 2000). Pathogens are disease-causing agents to their hosts including viruses, bacteria, fungi, protozoa and worms (such as *Paragordius tricuspidatus*). In this thesis, I will mainly refer to bacterial pathogens unless otherwise specified. Nonetheless, much attention has been devoted to bacterial mutualistic symbiosis during the last two decades. For example, symbiosis of bacteria with their insects hosts has received special attention because of the direct implication of these symbiotic relationships with the ecological diversification of their hosts. For example, many of the hosts known to harbor symbiotic bacteria feed on diets that are poor in essential amino acids (Hansen & Moran, 2011) and vitamins (this is the case of tsetse flies) (Wu et al., 2006; McCutcheon et al., 2009). Several studies have shown that both, the bacterium and the host, have established an intimate metabolic association such that the bacterium could complement the diet of the host with the lacking components (Moran, McCutcheon, & Nakabachi, 2008), hence allowing the colonization of novel unexplored ecological niches.

In the next three sections, I will first introduce the concepts of molecular coevolution and functional divergence, which play a central role in host-pathogen coevolution and which are the central concepts of this thesis. Secondly, I will briefly discuss the host immune response and the functioning of the immune system, a prerequisite to understanding the complex host-pathogen interactions and the underlying patterns of genetic variation. Thirdly, I will introduce the concepts of bacterial pathogens and elaborate on a group of specialised molecular systems that bacteria use to interact with their environments. The way such molecular systems can contribute to environmental adaptation of prokaryotes will be also the focus of my discussion. Finally, I will summarise current progress in prokaryote genome projects, which are the foundation of the analyses carried out throughout this thesis and, indeed, in the last decade of research in comparative genomics of prokaryotes.

1.2 Host-pathogen coevolution

In evolutionary biology, the central dogma of species coevolution is that interacting species undergo reciprocal evolutionary changes through natural selection (Thompson, 1982). In this sense, the coevolutionary process between species is considered to play a fundamental role in shaping Earth's biodiversity (Thompson & Cunningham, 2002). Although this may seem a novel concept in ecology, the interaction between species, both within and between species, and its importance on species's "struggle for existence" was, however, already noticed by Charles Darwin in his book *"On the Origin of Species"* (Darwin, 1859; Thompson, 1994). For example, Charles Darwin understood that "a corollary of the highest importance may be deduced from the foregoing remarks, namely, that the structure of every organic being is related, in the most essential yet often hidden manner, to that of all other organic beings, with which it comes into competition for food or residence, or from which it has to escape, or on which it preys." (Darwin, 1859).

Ever since the acceptance of the evolutionary theory, evolutionary biologists have strived to understand how interactions between species could result in coevolution (Thompson, 1994). The interaction between species could lead to specialisation, in which the relationships between species are well established. Specialisation includes the three types of relationships in symbiosis that were introduced at the beginning of this chapter: commensalism, mutualism and parasitism (Paracer & Ahmadjian, 2000). These specialised interactions lead inevitably to coevolutionary processes that maintain a selective advantage for one or two of the partners of the interactions and that shape the diversity of these interacting species. The interaction between pathogens and their host(s) could lead to antagonistic coevolution, well described in the Red Queen hypothesis. According to the Red Queen hypothesis, first proposed by the evolutionary biologist Leigh Van Valen in 1973 (Van Valen, 1973), the fitness increase of one species could cause fitness decrease of another species sharing the same ecological (interacting species) niche and vice versa. For example, in terms of a pathogenic relationship of a bacterium with its host, the conflict of fitnesses between the pathogen and the host could result in the adaptation of the pathogen to its host and the consequent counter-adaptation of the host to its pathogen (Paterson et al., 2010). This process was thought to drive the molecular evolution of the interacting species through natural selection (Van Valen, 1974). However, the main factors responsible for the interaction between pathogen and host are those proteins involved in infection and resistance to infection (Paterson et al., 2010).

These proteins determine the virulence of the pathogen and the susceptibility of the host. For instance, in the case of pathogenic bacteria these proteins are frequently found to be associated with their secretion systems, whereas in the human host such proteins are key components of the innate and adaptive immune responses. Therefore, if we want to investigate how host-pathogen interaction can lead to the reciprocal evolutionary changes in the classic view of coevolutionary theory, then we have to demonstrate how this specilisation results in molecular coevolution. In the case of host-pathogen interaction, I define specialisation as a condition in which pathogens undergo functional divergence in genes encoding specialised proteins that specifically interact with their hosts to improve its biological fitness. As a case in point, in chapter 3 I show how pathogenic bacteria use two homologous secretion systems to adapt to their hosts. One set of secretion proteins performs the housekeeping functions important for the bacterium's viability, while the other is specialised in secreting a particular set of proteins responsible for the bacterium's potential to colonize the host. Identifying proteins in pathogens that have been diverging from their homologous proteins in closely related free-living organisms and that have specialised their functions in adapting to its host would make it possible to understand the coevolutionary dynamics between hosts and pathogens. Therein, the process whereby proteins may shift in function as a result of changes in protein sequence composition after a process of lifestyle change or gene duplication will be termed functional divergence.

Functional divergence involves a change in a protein's function and its departure from the ancestral function. Functional divergence is most likely to occur after particular processes that shift the selective constraints (change the strength of natural selection), such as gene duplication or a change in the ecological conditions. Gene duplication is a much rarer process in bacteria compared to eukaryotes. The metabolic plasticity of bacteria increases their ability to adapt to a wide spectrum of environmental conditions, hence allowing a change in the ecological conditions and providing opportunity for functional specialisation of pre-existing proteins. The two main evolutionary forces driving the combinatorial evolution in life is therefore coevolution of molecules from different organisms and indeed the specialisation of proteins to new ecological conditions by changes in their original functions. Because of their central role in the emergence of combinatorial organismal interactions, molecular coevolution and functional divergence of protein functions are at the very centre of this thesis and support my study of the interaction between pathogens and hosts.

1.2.1 Molecular coevolution

Amino acid covariation at positions across the same or homologous protein sequences can be attributed to non-independent evolution. This means changes in the amino acid states at one position may change the constraints of amino acid changes at other position(s) of the same protein or of biologically or physically interacting protein. The mechanisms that cause such correlated changes are elusive (Chakrabarti & Panchenko, 2009). At the molecular level, we can define such non-independent evolution as molecular coevolution, which can be further defined at different levels, namely, intramolecular and intermolecular coevolution. Intramolecular coevolution refers to coevolution of amino acids within the protein, whereas intermolecular coevolution refers to these changes among proteins (Fares & McNally, 2006; Galtier & Dutheil, 2007; Codoner & Fares, 2008). As shown in equation (1) molecular coevolution C_{ij} between two amino acid sites i and j can be decomposed into phylogenetic coevolution ($C_{phylogeny}$), structural coevolution ($C_{structure}$), functional coevolution ($C_{function}$), coevolution due to structural and/or functional interactions ($C_{interaction}$) and stochastic coevolution ($C_{stochastic}$) (Atchley et al., 2000).

$C_{ij} = C_{phylogeny} + C_{structure} + C_{function} + C_{interaction} + C_{stochastic}$ (1)

Thus, at the molecular level coevolution is a very complex process governed by many factors that are not observable at the species level. Coevolution due to structural and/or functional reasons is known to arise as a result of the action of natural selection because changes causing structural or functional instabilities are filtered by natural selection as long as they exert important negative effects on the biological fitness of organisms. Therefore, our primary goal in coevolutionary analyses would be identifying coevolutionary amino acid sites that are the result of the structural and/or functional interactions between amino acids, while removing those coevolutionary relationships that are due to phylogenetic and stochastic processes.

Many methods have been developed to detect sites under coevolution based on multiple sequence alignments (MSA). These methods generally fall into two categories: non-parametric and parametric methods. Among the non-parametric methods that are extensively used are those relying on the calculation of the Mutual Information Content (MIC) of amino acid sites in the MSA, without a priori knowledge of the relationships between residues. In contrast to these non-parametric methods, parametric methods assume explicit probabilistic models, whereby they estimate parameters that can weight the coevolutionary relationships between amino acid sites and test their statistical significance (Codoner & Fares, 2008). Fares et al. developed a method that can identify coevolving amino acid sites after removing phylogenetic and stochastic effects (Fares & Travers, 2006). This method makes use of an empirical amino acid substitution matrix to identify statistically significant coevolving amino acid sites. It is able to distinguish positively and negatively correlated coevolving sites and yields highly accurate results when the MSA are allele data from the same species.

It has become increasingly clear that molecular coevolution is instrumental to determine the nature of the relationships between proteins in the same species or between different species in the biosphere. Coevolution analyses at the molecular level are particularly interesting for human health and pathogens-caused diseases because many pathogens are known to be specialised in infecting humans and/or closely related mammals, with which they present interesting coevolutionary dynamics (Thompson, 1999; Woolhouse et al., 2002; Fares et al., 2011). In the last decades much attention has been devoted to understanding the dynamics of pathogens, specially of those with high immune-escape mutation variants and with short generation times (RNA viruses for instance). Much less attention has been directed towards deciphering how pathogens and their hosts interact at the molecular level, owing to the inherent complexity of such dynamics. Molecular parameters limiting or defining the interaction between these two biological systems are, however, of great importance. Many difficulties arise in such an endeavour, mainly due to the complex population dynamics of the pathogens within the host and the difficulty inherent to the social networks and treatment dynamics of the host. Identifying molecular signatures that could define the nature of the interaction between pathogens and host can shed light on an otherwise obscure problem. For instance, coevolutionary analyses may help us identify the patterns of correlated variation between immune proteins and proteins involved in bacterial and viral pathogenesis. These patterns could be paramount to our understanding of the molecular mechanisms of pathogenesis, infectious diseases and host immunity (Apanius et al., 1997; Borghans et al., 2004; De Boer et al., 2004; Hailing-Brown et al., 2008; Hedrick, 2002; Vogel et al., 1999).

1.2.2 Functional divergence

The concept of functional divergence has been used to describe how protein functions change after gene duplication events, which is considered to be a fundamental step in evolution and that could lead to the origin of new species (Ohno, 1970; Lynch & Conery, 2000; Conant & Wolfe, 2008; Innan & Kondrashov, 2010). Although much attention has been given to gene duplication as a source of genetic and functional innovation, two evolutionary events could drive to their functional diversification (Ohno, 1970; Soskine & Tawfik, 2010): gene duplication (Ohno, 1970) and speciation (Gu, 2003; Studer & Robinson-Rechavi, 2009, 2010). The biological phenomena that lead to functional innovation are partially understood, the mechanisms enabling such functional innovation, however, remain elusive.

Functional innovation is a rather improbable phenomenon because most innovative mutations are destabilising–that is, they compromise ancestral functions or protein stability. Some mechanisms have been proposed that may be successful at enabling the fixation of innovative, slightly deleterious, mutations. Key among these mechanisms is whole-genome duplication that results in new genetic material likely to be preserved for long evolutionary periods owing to the preservation of gene dosage balance. Other mechanisms, such as those that buffer the deleterious effects of innovative mutations, may result in novel functions, such as the action of molecular chaperones and chaperonins–folding machines that are ubiquitous and important as sentinels of the proper cell proteome function. Regardless, the mechanism enabling the fixation of innovative mutations, it is obvious that this evolutionary phenomenon is a likely outcome for one gene copy after duplication (paralogs) and homologous genes from species sharing a common ancestor (orthologs) (Koonin, 2005; Studer & Robinson-Rechavi, 2009).

It is generally considered that orthologs are functionally similar and paralogs are more functionally divergent (Studer & Robinson-Rechavi, 2009). Because of the potential to offer opportunity for functional diversification, gene duplication is considered to be a main source of evolutionary innovation (Lynch & Conery, 2000). Contrary to this, a recent study demonstrated that patterns of functional divergence between orthologs and paralogs in animal genes are not significantly different (Studer & Robinson-Rechavi, 2010), indicating that studying functional divergence of a protein family (a group of homologous proteins) is more appropriate since it includes both orthologs and paralogs (Gu, 2001, 2003).

Only recently have large scale analyses become possible with the increase of genomic data and the development of appropriate statistical methods (Gu, 1999, 2001, 2006; Toft et al., 2009). Most of these methods work under the reasonable assumption that amino acid sites conserved evolutionarily in homologous proteins are so because of their functional importance (Kimura, 1983). This assumption is based upon the fact that natural selection operates to preserve protein functions provided these functions optimise organismal fitness, hence functional amino acid sites are more constrained by natural selection than non-functional sites in a protein. The conservation pattern of amino acid sites in homologous proteins is used by most functional divergence detection methods to look for two types of functional divergence of amino acid sites: Type I and Type II functional divergence (Gu, 1999).

In Type I functional divergence amino acid sites are compared between two clusters of phylogenetically close protein sequences. In one cluster of protein sequences amino acids at a particular site are highly conserved, whereas in the second cluster amino acids at that site have much greater stochasticity, pointing to the acquisition of a function at these sites in one paralog (the conserved one) or the loss of the function in the other. Conversely, in Type II functional divergence amino acid sites are conserved in both groups, which implies divergence of the two functionally important clusters of homologous proteins. Recently Fares' group has developed a fast and simple statistical method to detect functional divergence, which can be applied to analyse genome-wide or protein family functional divergence (Toft et al., 2009; Williams et al., 2010). Analysing functional divergence of protein families and proteins at the genome level can reveal patterns of functional diversification that contributed to organism ecological adaptations. The main constraints of functional diversification of these proteins should largely come from the environment in which the organism thrives. Because duplicated genes were shown to play an important role in the ecological adaptation of prokaryotes (Gevers et al., 2004; Sanchez-Perez et al., 2008; Rose et al., 2002), this biological phenomenon is considered as a main source of functional divergence. In the context of bacterial pathogenesis, the main constraints should come from the host (e.g. human) environment, and mainly from the host immune responses. Consequently, pathogenic bacteria may be expected to utilise duplicate genes or orthologous genes with slight functional shifts to gain new functions that are crucial in host exploitation as evidenced by previous studies (Braunstein et al., 2003; Hou et al., 2008; Bensing & Sullam, 2009; Toft & Andersson, 2010).

1.3 Host immune response

Mammalian hosts have a variety of ways to stop the invading pathogens, most of which rely on the physiological barriers (Figure 1.1A) and the immune system of the host. The immune system has evolved specialised cells and proteins to perform immune effector functions thereby responding to pathogenic invasions. When pathogens pass the physiological barriers of the host, the immune system must be able to recognise such invasions and then eliminate them through effective immune mechanisms. In most situations, the induced immune responses successfully recognise and neutralise the invading pathogens and the agents they produce. However, in some situations the immune system fails to do so due to genetic defects or other complicated factors (Casadevall & Pirofski, 2001). For example, a human pathogen can acquire a highly virulent trait and it can increase the virulence of the pathogen. Consequently, this pathogen could be more lethal to susceptible human populations. Similarly, an outbreak of *Escherichia coli* infection has recently occurred in Germany. It has claimed the lives of 39 people in less than two months. This new *E. coli* strain (O104:H4) presents virulence characteristics of both Shiga-toxin-producing *E. coli* (STEC) and enteroaggregative *E. coli* (EAEC). This demonstrates

Cell	Primary Function
Macrophage	Phagocytosis, bactericidal, antigen presentation,
Dendritic cell	Antigen uptake in peripheral sites, antigen presentation in lymph nodes
Neutrophil	Phagocytosis, bactericidal
Eosinophil	Killing of antibody-coated parasites
Basophil	Antigen presentation (Sokol et al., 2009)
Mast cell	Inflammatory response
Natural killer (NK) cell	Killing of infected cells
B cell	Generating antibodies
T cell	Killing of infected cells, helping other cells respond to infection

Table 1.1: Typical cells involved in the immune response and their functions.

that a pathogenic *E. coli* strain can acquire both adhesive and Shiga-toxin-producing genes and this makes it extremely lethal to susceptible populations (Bielaszewska et al., 2011).

Mammalian hosts rely on two types of immune responses to fight infection, namely, the innate and adaptive immune responses (Murphy et al., 2007). The two types of immune responses have different speeds and specificities (Murphy et al., 2007). In general, the innate immune response is rapidly activated once the immune system detects the invading pathogens. Such a response is very broad and general, and acts by recognising evolutionary conserved molecular patterns among invading pathogens through germline-encoded receptors (proteins). If the innate immune response can not eliminate the pathogens, which normally takes hours to days, then the adaptive immune response will be activated at a later stage after invasion. This activation is highly specific and orders of magnitude more effective than the innate immune response, particularly if the invading pathogen has been encountered throughout the history of the host.

As mentioned above, an effective immune response to invading pathogens certainly requires coordinated action of components of the immune system itself. The mammalian host immune system can be perceived from two levels, the cellular and molecular level. At the cellular level, all cells of the immune system come from the bone marrow, they subsequently differentiate into different types of immune cells, which include B cells, T cells, natural killer (NK) cells, macrophages, dendritic cells and neutrophils. They are collectively called white blood cells or leukocytes. These cells perform various important functions in the innate and adaptive immune responses. For instance, macrophages, dendritic and neutrophil cells (Table 1.1), known as antigen presenting cells, can engulf invading pathogens, such as viruses and bacteria (Murphy et al., 2007). At the molecular level, there are a plethora of proteins involved in the innate and adaptive immune responses, which perform a vari-

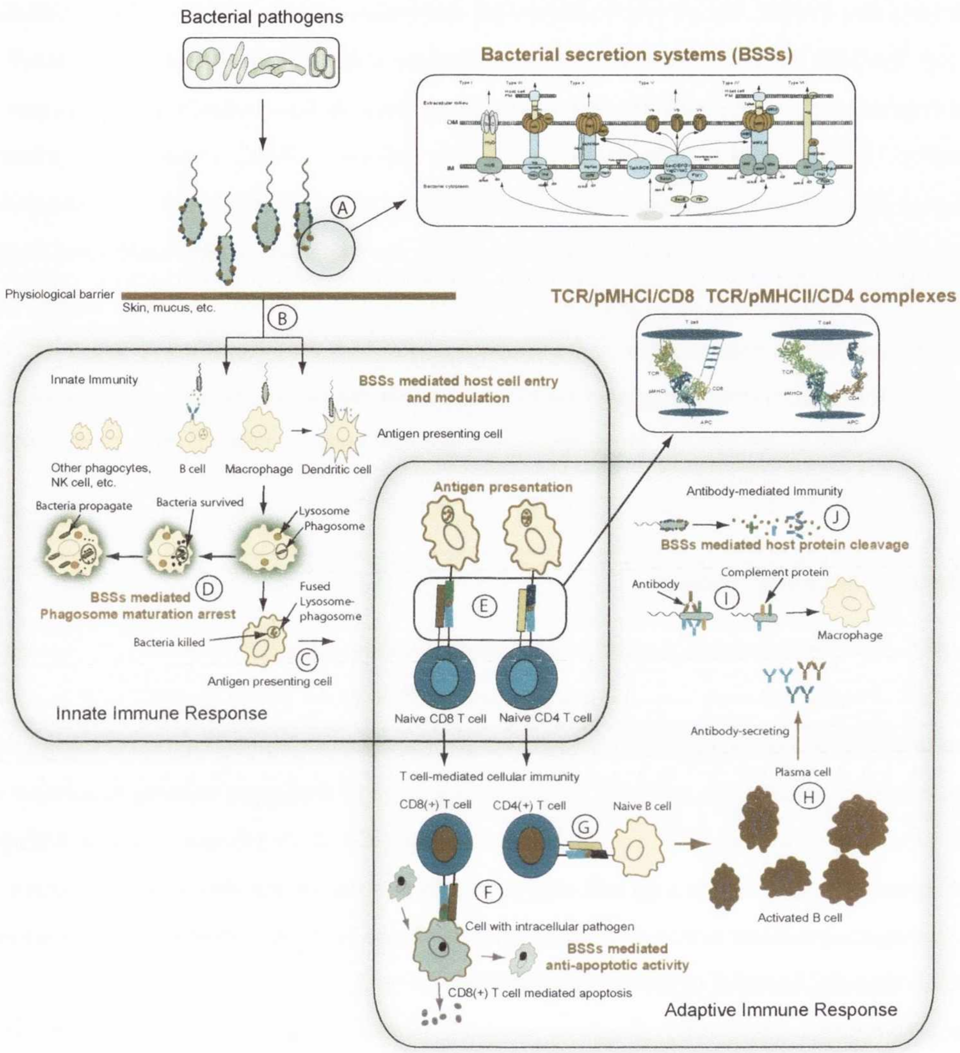


Figure 1.1: Schematic illustration of host immune response to pathogenic bacteria and bacterial secretion systems mediated immune evasion. Major processes involved in host immune response to pathogenic bacteria and pathogenic bacteria host invasion and immune evasion are illustrated and labeled accordingly. (A) Bacteria use their specialised secretion systems (see Figure 1.3 for details) to transport cytosolically synthesised proteins to their membranes and surrounding environments to interact with host. (B) Once pass the physiological barrier of the host, bacterial pathogens will encounter their first wave of defense of the innate immune response. They will be engulfed and killed by phagocytic cells such as macrophages, B cells, dendritic cells, etc. (C) These cells (APCs) also present antigenic peptides from the pathogens once they are killed. (D) If bacteria successfully evade their killing by bacterial secretion systems (BSSs), then they will establish a focus of infection. (E) APCs present antigenic peptides and form a complex TCR/pMHCI/CD8 or TCR/pMHCII/CD4 (see Figure 1.2 for details) with T cells to activate T cells to initiate adaptive immune response. Activated T cells will perform two major effector functions: (F) killing host cells infected with the intracellular pathogens; and (G) activating naive B cells to become (H) antibody-secreting plasma cells, which further facilitate adaptive immune response to kill invading bacteria. (I) The complement system, together with phagocytes (e.g., macrophages), can clear bacterial pathogens with or without antibody involvement. (J) Bacteria are able to evade antibody-mediated host immunity by BSSs secreted host protein proteases.

ety of functions that include, but are not limited to, the recognition of pathogen-associated molecular patterns (PAMPs, e.g. bacterial lipopolysaccharide) and neutralisation of invading pathogens; and cell-cell communication and regulation of gene expression. These proteins include toll like receptors (TLR), immunoglobulins, T cell receptors (TCRs), major histocompatibility complex molecules (MHC), cytokines(e.g. interleukin(IL)), tumor necrosis factor (TNF), and many other related proteins. Among these proteins, immunoglobulins have been extensively studied. Immunoglobulins have two forms: the soluble form (antibodies) and the membrane-bound form (B cell receptors: BCRs), and this is the foundation of humoral immunity. In sum, the immune system should present two main features: i) Its function should be tightly regulated so that it will not damage the host; and ii) It should be effective against future invasions by the same pathogen encountered throughout its existence–that is, keep in memory the characteristics of the pathogens invading it (Kindt et al., 2007; Murphy et al., 2007).

1.3.1 Innate immune system

The innate immune system is the first line of defence against pathogens that pass the host physiological barriers. Many cellular and molecular components are involved in the innate immune response and their action is fine-regulated and tuned to provide an efficient response against non-self agents (Murphy et al., 2007). Briefly, as shown in Figure 1.1B, the main cellular components of the innate immune system are phagocytic cells, such as macrophages, dendritic cells and other immune cells. Macrophages as a typical phagocytic cell can engulf invading pathogens through cell surface receptors, such as Toll-like receptors (TLR), which can recognise pathogen associated with particular molecules. These molecules include bacterial lipopolysaccharide, bacterial flagellin, bacterial peptidoglycan and viral RNAs.

Once macrophages engulf the pathogen, a series of pathways are activated (Murphy et al., 2007). First, the pathogen contained phagosome will merge with lysosomes, which kills the pathogen by releasing the content from the lysosomes. Second, activated macrophages secrete a wide range of cytokines to signal other immune cells to the site of infection and adjust the physiological state of the host to contain the infection and subsequently eradicate it. These cells also play an important role as antigen presenting cells (APCs). APCs are crucial in linking innate immune response and adaptive immune response to fight infection as shown in Figure 1.1C (Henderson & Oyston, 2003). Third, another important component of innate immune system, the complement system, is also involved in killing the invading bacterial pathogens before they are engulfed by macrophages. Complement system includes three major pathways, the classical pathway, the lectin pathway and the alternative pathway. These pathways have a variety of proteins, which can directly attack the pathogen surface with or without the aid of antibodies (Figure 1.1I). They also facilitate the pathogen uptake by antigen presenting cells, which contributes to host adaptive immune response. Finally, one of the important components of innate and adaptive immunity is the mammalian NF-kB pathway. NF-kB is a gene transcription factor complex formed by three components, inhibitor of NF-kB (IkB), p50 and p52, which regulates gene expressions of various pathways involved in immune system once activated. These pathways range from development and survival of lymphocytes and lymphoid organs to the control of immune responses and malignant transformation (Vallabhapurapu & Karin, 2009).

1.3.2 Adaptive immune system

The adaptive immune system (AIS) plays a central role in host immunity as many resistant pathogens can only be cleared by adaptive immune response. The emergence of AIS is a hallmark of vertebrate evolution, and is the core of the host defence mechanisms against a plethora of dangerous microbial pathogens, thereby maintaining its self-existence (Litman et al., 2010). The two major players in the AIS are the T cells and B cells. As shown in Figure 1.1E, the T cells have two major types named and classified according to two particular cell surface proteins CD8 and CD4: CD4 T cell (helper cell), and CD8 T cell (cytotoxic cell). During infection, T cells use their cell surface proteins T cell receptors (TCRs) to bind to the antigenic peptides presented by MHC (major histocompatibility complex) molecules on their antigen presenting cell surface. This interaction between T cells and antigenic peptides activates the T cells that go from a naive state to an active state. Activated T cells perform two major effector functions in the adaptive immune response (Murphy et al., 2007). First, activated cytotoxic T cells can kill host cells which are infected with intracellular pathogens as illustrated in Figure 1.1F. Second, activated helper T cells can activate B cells to become plasma cells, which secrete highly specific antibodies that play a fundamental role in humoral immunity as illustrated in Figure 1.1G-I. These antibodies will effectively attack pathogens and the pathogen produced agents and exert an effective neutralising action.

Both T cells and B cells have unique abilities that contribute to the successful neutralisation of invading pathogens. TCRs expressed on T cell surface can bind to a countless number of antigenic peptides derived from pathogens in a MHC restricted way, which allows the host to virtually respond to any pathogen invasion (Murphy et al., 2007). AIS uses a powerful mechanism to create a substantial amount of diversity of TCRs and immunoglobulins (antibodies and BCRs), where the diversity is created by the recombination of variable (V), diversity (D) and joining (J) gene segments: the so-called V(D)J recombination. The diversity of B cell receptor is further enhanced by class-switch recombination, somatic hypermutation and terminal differentiation. The total diversity of B cell and T cell receptors can reach an astonishingly high number: $\sim 5 \times 10^{13}$ and $\sim 10^{18}$, respectively (Murphy et al., 2007).

Pathogens are constantly evolving, which frequently generates immune-escape mutants (Kawashima et al., 2009; Tan et al., 2010; van der Woude & Baumler, 2004). For example, bacteria and viruses can change their molecular patterns of surface proteins, which include allelic versions of the same protein. Such changes would preclude antibodies generated from previous infections of the same pathogen to recognise the new allelic protein (Keck et al., 2009; Tan et al., 2009) leading to the successful colonization of the host by the pathogen. However, as mentioned above, the astonishing diversity of TCRs and BCRs can minimize the amount of variant that could escape immune response. Complex coevolutionary dynamics between host and pathogens is therefore established.

1.3.3 Major histocompatibility complex molecules

The major histocompatibility complex (MHC) plays a vital role in mediating innate and adaptive immune responses. MHC is a cluster of genes that were first found to be responsible for the strong immune response to transplanted tissues. MHC genes were first discovered in mouse by Peter A. Gorer and George D. Snell in 1936, and this set of genes was initially named histocompatibility-2 (H-2). Its homologous genes in humans were first discovered by Jean Dausset in 1950s and were later named human leukocyte antigen (HLA) (Klein, 1986). Sequentially, MHC were discovered in other mammals, such as rat, pig and so on (Klein, 1986). MHC genes have been extensively studied after their discovery and are regarded as the most variable genes in jawed vertebrates. Besides the initial discovered role in the rejection of transplanted tissues, MHC was shown to play a critical role in host immunity for several reasons. First, antigenic peptides from pathogens are constantly presented by APCs in a MHC restricted way to T cells, thereby playing a central role in initiating host adaptive immune response. Second, several notorious human infectious diseases are found to be MHC-associated with strong evidence, including acquired immune deficiency syndrome (AIDS, caused by human immunodeficiency virus (HIV) infection), chronic viral hepatitis (caused by hepatitis B and C virus infection) and mycobacterial infections (leprosy caused by *Mycobacterium leprosy* infection and tuberculosis caused by *Mycobacterium tuberculosis* infection) (Blackwell et al., 2009).

In humans, there are generally two main classes of MHC molecules: MHC class I (also known as human leukocyte antigen (HLA) in humans, HLAI) and MHC class II (MHCII, HLAII) molecules. MHC is located on chromosome 6 in humans and chromosome 15 in mice. MHCI and MHCII are further classified into three subclasses: MHCI has A, B and C three subgroups; while MHCII is generally classified in the subgroups DP, DQ and DR. Both MHCI and MHCII proteins share very similar overall structures consisting of two functional regions: the antigen and TCR binding and coreceptor CD8/CD4 binding regions (Figure 1.2). As introduced above, antigenic peptides are constantly presented by MHC molecules of antigen presenting cells from the innate immune system to T cells of the adaptive immune system, and this allows recognition and activation of two types of crucial immune cells: the CD8 T cell and the CD4 T cell. During antigen presentation to T cells, a complex is formed by the T cell receptor (TCR), an antigenic peptide bound to a MHC molecule (pMHC) and coreceptor CD8 and CD4, where CD8/CD4 stabilises the interaction between TCR and pMHC. This complex

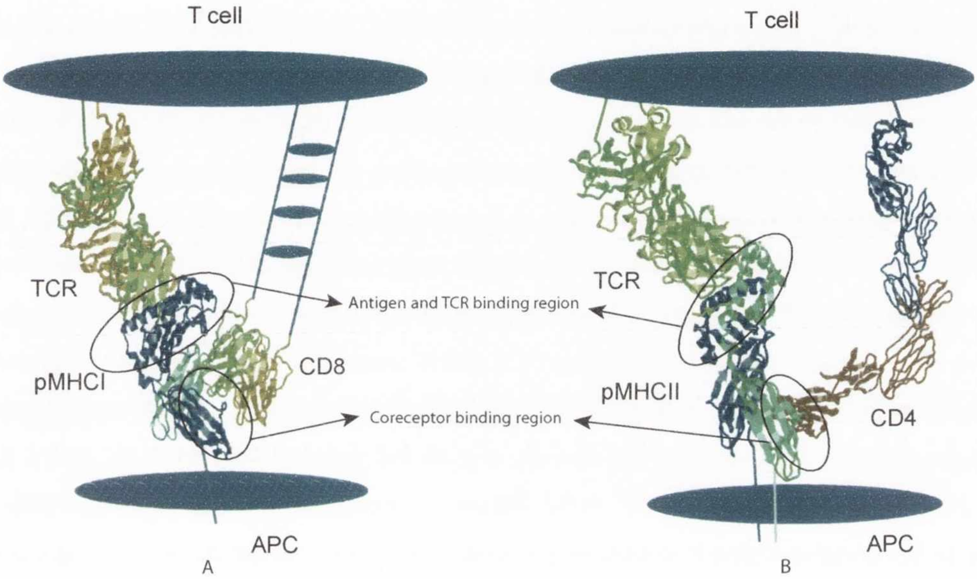


Figure 1.2: Crystal structures of the two critical complexes (TCR/pMHCI/CD8 and TCR/pMHCII/CD4) formed between an antigen presenting cell and a T cell. (A) TCR/pMHCI/CD8 complex is formed between T cell and antigen presenting cell (APC), where the alpha chain and the β_2 microglobulin of MHC are colored blue and cyan, respectively. TCR has an alpha chain (colored green) and a beta chain (colored orange). Homodimer CD8 has two alpha chains (colored green and orange respectively). (B) TCR/pMHCII/CD4 complex is formed between T cell and APC, where the alpha chain and the beta chain are colored blue and cyan, respectively. CD4 has four domains (D1-D4). D1 and D2 are colored red, while D3 and D4 are colored blue. This figure is produced with protein crystal structures from TCR/pMHCI-A complex (protein data bank ID: 1BD2), CD8/pMHCI-A complex (protein data bank ID: 1AKJ), TCR/pMHCII-DR complex (protein data bank ID: 1FYT) and CD4 (protein data bank ID: 1WIO).

allows T cell recognition and activation (Rudolph et al., 2006), which is the prerequisite of initiating subsequent adaptive immune response. If the antigenic peptide is derived from pathogens in the cytosol of the host cell, then it is loaded to MHCI molecules and presented to CD8 T cell. Conversely, antigenic peptides derived from phagosomes are loaded into MHCII molecules and presented to CD4 T cell (Murphy et al., 2007).

Two complexes are formed between APCs and T cells (Figure 1.2): the first complex TCR/pMHCI/CD8 (Figure 1.2A), MHCI molecule comprises a heavy chain alpha (colored blue on pMHCI) and a light chain beta-2 microglobulin (colored cyan on pMHCI), with the alpha chain being anchored in the cell membrane (Bjorkman et al., 1987); the second complex TCR/pMHCII/CD4 (Figure 1.2B), MHCII molecule comprises an alpha chain (colored blue on pMHCII) and a beta chain (colored cyan on pMHCII) (Brown et al., 1993) and both are anchored in the cell membrane. MHCI molecules are expressed in almost all nucleated cells, while MHCII are only expressed in cells of the immune system. This complex plays a fundamental role in host immunity, arguably it has been extensively studied during the last two decades (Rudolph et al., 2006; Gao et al., 2002). Two key functional regions have been extensively analysed: antigen binding groove, where the antigenic peptide is loaded and where TCR also binds to (Marrack et al., 2008), and the coreceptor binding region that, when bound to CD8/CD4, stabilises TCR-pMHC interaction and contributes to successful T cell activation (Rudolph et al., 2006). TCR has an enormous diversity generated by V(D)J recombination, while MHC has the most variable genes and this diversity is germ-line encoded (Murphy et al., 2007). Moreover, antigenic peptides as integral part of MHC molecules are also highly variable (Zhang et al., 1998). The rich variability of this complex makes it an ideal system to test hypotheses concerning host-pathogen coevolution owing to the rich evolutionary information that such variability contains.

Coevolutionary analysis of TCR/pMHCI/CD8 and TCR/pMHCII/CD4 complex would yield invaluable information for future projects more centred on the understanding of biomedical problems. First, this complex plays a fundamental role in the adaptive immune response to all kinds of pathogens, which directly links pathogens and immune systems, as antigenic peptides bound in the MHC antigen binding groove are derived from pathogens. Impairment of functions of this complex would have a drastic effect on downstream immune functions, such as T cell mediated cellular immunity (killing of intracellularly infected host cells by CD8(+) T cells) and antibody mediated humoral immunity (antibody production by B cells) (Figure 1.1). Second, from a mathematical point of view, this complex is a "battlefield of variation" as extensive evidence from last three decades indicates that: i) variation in antigenic peptides affects the function of this complex and leads to all sorts of infectious diseases; and ii) variation in MHC molecules affects the function of this complex, which all contribute to the evasion of T cell-mediated host immunity and thus adaptive immune response of the host. Studying the molecular coevolution between antigenic peptides, MHC, TCR and coreceptor CD4 and CD8 would be unrealistic since antigenic peptides can be derived from literally all pathogens, and TCRs have a prohibitive diversity as it could reach $\sim 5 \times 10^{13}$ different combinations and variants. To circumvent this issue, we can identify intramolecular coevolutionary patterns in MHC molecules since the diversity of MHC is limited and current computing resources are capable of handling such an analysis.

1.4 Pathogenic bacteria

Infectious diseases cause serious morbidity and mortality in both developed and developing countries. One of the causative agents of infectious diseases is bacteria. Bacteria are ubiquitous to all Earth environments. They live abundantly in soil and water across diverse ecological niches. We can find pathogens among all bacterial phylogenetic clades but not all clades are equally populated by pathogens, and they live in physical proximity to their hosts. When a pathogenic bacterium interacts with a host, it can live either outside or inside the host cell. The ability of bacteria to cause disease is regarded as virulence. Such ability depends on the genetic background of the host immune system, which also varies from individual to individual (Casadevall & Pirofski, 2001).

1.4.1 Bacterial pathogenesis

Bacterial pathogenesis involves the recognition of host cells, invasion of the host, evasion of the host immune response, persistence of the infection and several other mechanisms that are responsible for the progress and emergence of symptoms of the infection in the host (Ohl & Miller, 2001). To go through the different pathogenic steps, bacteria need at least four groups of effector proteins. First, they need to attach to the surfaces of host cells. Second, they need to actively and/or passively invade the cells of the host. Third, once they encounter the host immune response, they should be able to challenge the effector mechanisms of the host immune response and survive. For instance, during macrophage phagocytosis, the bacterium is internalised in the phagosome by macrophages, the subsequent step would involve the merge of phagosome and lysosome to kill the bacterium (Murphy et al., 2007). It is known that many bacteria can secrete some special proteins (e.g. superoxide dismutases (Braunstein et al., 2003)) to interfere with this process and avoid destruction successfully. These secreted proteins enable the bacteria to establish a focus of infection and colonise the host (Flannagan et al., 2009). Fourth, after the successful colonisation of the host, pathogenic bacteria could secrete toxic products to the host. These toxins may benefit the infecting bacterium but cause tissue damage to the host (Masignani et al., 2006). In spite of being synthesized in the cytosol, all four groups of bacterial effectors function extracellularly. Therefore, they need to be targeted to their final destinations to perform various functions contributing to the pathogenesis of bacterial infection. How do bacteria target these proteins? This can be achieved using bacterial secretion systems (Harper & Silhavy, 2001; Lee & Schneewind, 2001).

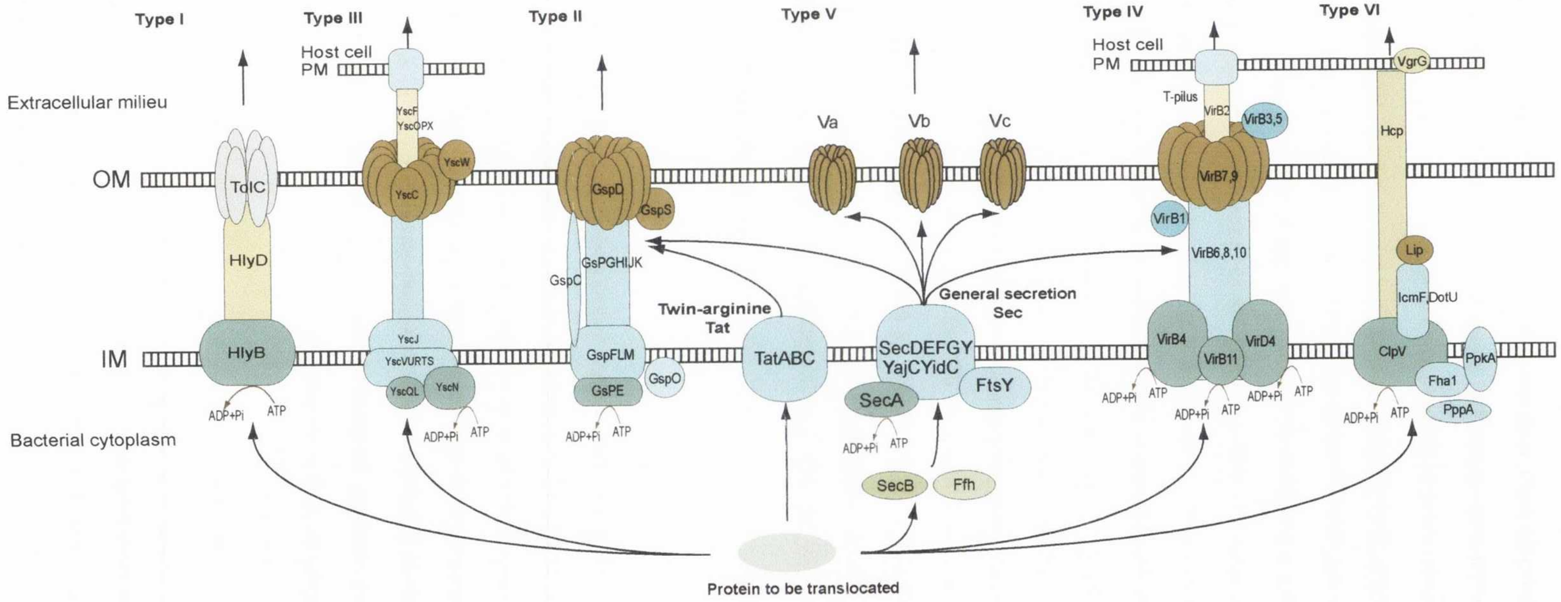


Figure 1.3: Schematic illustration of bacterial secretion systems. Two membranes are shown, the cytoplasmic (inner) membrane (IM) and the outer membrane (OM), which represent membranes in gram-positive and gram-negative bacteria, respectively. The Sec and Tat secretion systems are shown on the cytoplasmic membrane. Other secretion systems (Type I, Type II, Type III, Type IV, Type V and Type VI) are shown in a membrane spanning configuration between cytoplasmic and periplasmic membranes. Protein components for each secretion system are also labeled. Redrawn from `http://www.genome.jp/kegg-bin/show_pathway?map03070`

1.4.2 Bacterial secretion system

Bacterial secretion systems are molecular machines that are specialised in transporting proteins to subcellular locations important to bacterial physiology. Protein secretion is very different between the two major groups of bacteria: the gram-positive and the gram-negative bacteria. In the former group a bacterium has a single lipid bilayer membrane, while in the later group a bacterium has two lipid bilayer membranes. Consequently, in gram-positive bacteria secreted proteins need to pass through one lipid bilayer membrane and can be targeted to at least three subcellular locations by secretion systems: cytoplasmic membrane, cell wall and extracellular milieu (Figure 1.3). Conversely, in gram-negative bacteria secreted proteins need to pass through two lipid bilayer membrane and can be targeted to four subcellular locations: cytoplasmic membrane, space between cytoplasmic membrane and periplasmic membrane,and extracellular milieu (Figure 1.3). Translocation of secreted proteins in bacteria generally requires a signal peptide at the amino-terminus of the protein. The signal peptide is secretion system specific, with only those proteins that bear the required signal peptide being transported (Natale et al., 2008).

The general Secretion (Sec) pathway and the twin-arginine translocation (Tat) pathway are the only two secretion systems known to translocate proteins across or into the cytoplasmic membrane in all three domains of life (Yuan et al., 2010). About 25% to 30% of bacterial proteins were thought to be translocated through the Sec secretion system (Driessen & Nouwen, 2008). However, it is generally considered that a smaller fraction of bacterial proteins are translocated by the Tat pathway (Lee et al., 2006). Specifically, as shown in Figure 1.3, the core components of Sec and Tat secretion system are evolutionarily conserved and are found in the cytoplasmic membrane of bacteria and archaea. In contrast to prokaryotes, the Sec translocon is located in the endoplasmic reticulum (ER) membrane of eukaryotic cells (Pohlschroder et al., 1997), while the Tat components are located in the thylakoid membrane of chloroplast and mitochondria of plant cells (Lee et al., 2006).

The core of Sec translocon of the Sec pathway consists of three proteins: SecY, SecE and SecG. SecY, which forms the peptide conducting channel of the translocon, was shown to play a fundamental role in transporting unfolded proteins across the membrane and integration of membrane proteins. How does the cell target newly synthesised cytosolic proteins to the membrane through the SecY channel? There are two pathways used in bacteria: the posttranslational and the cotranslational translocation pathways. In the posttranslational translocation pathway SecB and/or SecA, through its chaperone activity, binds to the newly synthesised peptide to avoid premature folding and target it to the membrane (Figure 1.3), whereas SecA pushes the peptide through the SecY channel under ATP hydrolysis. In the cotranslational translocation pathway, SRP (signal recognition particle, it has subunit Ffh) can bind to the signal peptide and guide the ribosome to the membrane, where SRP binds to its receptor FtsY and the ribosome can then push the emerging peptide into the SecY channel during protein synthesis (Figure 1.3). Archaea were shown to use both pathways to translocate their secreted proteins in the Sec secretion. However, it is not clear how archaea prevent proteins from folding after their synthesis in the posttranslational translocation pathway since archaea have no SecB and SecA in their Sec system (Irihimovitch & Eichler, 2003).

The Sec-independent Tat pathway has three components: TatA, TatB and TatC (Figure 1.3). TatA and TatC comprise the minimum Tat system, which is found in most gram-positive bacteria and archaea. Conversely, most gram-negative bacteria have all three Tat components. During protein translocation TatC(B) first forms the complex for substrate protein recognition (secreted proteins), which subsequently recruits TatA to form pore-like structures with various dimensions on the membrane to allow folded substrate proteins (secreted proteins) pass through the membrane (Lee et al., 2006).

In addition to the Sec and Tat secretion systems, there are six specialised secretion systems found in gram-negative bacteria, as shown in Figure 1.3, namely, the Type I, II, III, IV, V and VI secretion systems (denoted by T1SS, T2SS, T3SS, T4SS, T5SS and T6SS, respectively) (Boyer et al., 2009; Wooldridge, 2009). These remarkable systems are preferentially used by gram-negative bacteria to communicate with their surrounding environments and are instrumental to their ecological adaptation. T1SS is a relatively simple system as only three proteins are reported to be involved in secretion (TolC, HlyD and HlyB, as shown in Figure 1.3), which can secrete toxins, proteases and lipases (Jenewein et al., 2009). T2SS is Sec and Tat-dependent, which can secrete intracellular toxins, cholera toxins, exotoxin A, heat-liable (LT) toxin and cell wall-degrading enzymes in the host environment (Michel & Voulhoux, 2009). These T2SS-dependent secreted proteins play an important role in host colonisation. T3SS consists of at least 15 proteins (Figure 1.3) and recent studies have shown that T3SS performs several important functions in immune evasion. For example, T3SS can directly injects (secretes) effector proteins into host cells, which allow bacteria to be resistant to phagocytosis (Sorg & Cornelis, 2009). T3SS also inhibits NF-*k*B, blocking secretion of certain cytokines important to recruit other immune cells to the site of infection (Sorg & Cornelis, 2009). T4SS is another secretion system that is able to directly secrete effector proteins into host cells. T4SS transports DNA or DNA-protein complexes across bacterial membrane to the extracellular milieu or host cells (Figure 1.3) and has at least 12 protein components in model organism *Agrobacterium tumefaciens* (Baron, 2009). T5SS is Sec-dependent and allows bacteria to secrete immunoglobulin A1 (IgA1) proteases to evade antibody-mediated host immunity (Scott-Tucker & Henderson, 2009), which can cleave IgA at the host mucosal surface as illustrated in Figure 1.1J. Besides T3SS, T4SS and T5SS, T6SS is the third secretion system that is able to directly deliver effector proteins into host cells (Hayes et al., 2010). Recently, the Type VII secretion system was proposed based on a specialised secretion system found in *Mycobacteria tuberculosis*, which is possibly conserved in other gram-positive bacteria (Abdallah et al., 2007).

1.4.3 Secretion system and ecological adaptation

Mounting evidence suggests that secretion systems and their substrate proteins (secreted proteins) play an important role in prokaryotes environmental adaptation (Rigel & Braunstein, 2008; Wooldridge, 2009; Pickering & Oresnik, 2010). Although several mechanisms have been proposed to explain microbe's environmental adaptation, no studies have demonstrated how a particular biological system can contribute to ecological adaptation in a systematic way across broad phylogenetic lineages. For example, it has been proposed that specialized paralogs play a role in bacterial adaptation to environmental changes (e.g. salinity and temperature) (Sanchez-Perez et al., 2008), and large scale studies have also shown that there are biased functional retention of paralogous genes (e.g. genes related to amino acid transport and metabolism) in bacterial genomes (Gevers et al., 2004).

Secretion systems mediate the interaction with environment and are universal and ubiquitous among microbes with different ecological traits, which include adaptation to extreme halophilic environment (salt close to saturation), extreme hot environment (temperature close to water boiling point), mammalian hosts (formidable defense lines including innate and adaptive immune responses), etc. How can the Tat and Sec secretion systems contribute to the emergence of novel ecological adaptations? One example comes from the extreme halophilic prokaryotes, which are adapted to an environment where salt concentration is close to saturation. High salt concentrations in the environment can produce high osmotic pressure in the cell. This can cause serious problems to the cell because cytosolic water can freely pass the membrane to the surrounding environment due to high salinity. To counterbalance such pressure, two strategies are used: the "salt in" strategy and the "low salt-in" strategy (Oren, 2008). The "salt in" strategy is used by extreme halophilic archaea (e.g. *Halobacterium sp. NRC-1*) and some extreme halophilic bacteria (e.g. *Salinibacter ruber*), where the cell accumulates high inorganic ion concentrations (e.g. K^+) in the cytoplasm. In contrast, the "low salt-in" strategy prevents inorganic ions from entering the cytoplasm. However, halophiles that use the "low salt-in" strategy still need to accumulate organic solutes to main osmotic pressure while actively pumping ions out (Oren, 2006). Although the two strategies are apparently different in halophiles, prokaryotes that use the two mechanisms presumably both have to invest a lot in inorganic ion (e.g. $K^+$and Na^+) transport and metabolism. It has been shown that haloarchaea (*Halobacterium sp. NRC-1*) extensively use the Tat pathway instead of the Sec secretion system to transport most of their secreted proteins (Rose et al., 2002). If such shift in protein secretion is due to adaptation to extreme halophilic environments, then to what extent this shift in secretion contributes to inorganic ion transport and metabolism? Do halophilic bacteria show similar patterns? Could functional divergence of the Tat pathway contribute to adaptation to extreme halophilic environment. If this is true, could functional divergence of the Tat pathway contribute to the adaptation of other prokaryotes to other environments (e.g.

hosts)? We address these questions in Chapter 4 of this thesis.

How do secretion systems contribute to mammalian host adaptation of pathogenic bacteria? It is known that there are many mechanisms that pathogenic bacteria use to invade host and survive host immune response. For example, it has been observed that many known intracellular bacterial pathogens (e.g. *Helicobacter pylori, Legionella pneumophila* and *Vibrio cholerae*) can survive amoebae ingestion (Casadevall, 2008), and it has been suggested that this may provide a "training ground" for pathogens to adapt to higher eukaryotes (Toft & Andersson, 2010). Secretion systems have been shown to play important roles in bacterial adaptation to particular niches (Wooldridge, 2009). For example, it has been shown that the hallmark of many life-threatening infections such as endocarditis, (ID) disseminated intravascular coagulation (DIC), immune thrombocytopenic purpura (ITP) and myocardial infarction (MI) or stroke caused by *Staphylococcus* and *Streptococcus* species is the interaction of bacteria with host cells through the Sec-dependent secretion of adhesins expressed on bacterial cell surface (Fitzgerald et al., 2006). Knockout of genes that encode the secretion system responsible for translocating these adhesins was shown to block the bacteria colonization of host cells, which indicates that these secreted proteins contribute to the pathogenesis of the above diseases (Siboo et al., 2008; Rigel & Braunstein, 2008).

It is now possible to study microbes with the two secretion systems (Sec and Tat, Figure 1.3) from a wide range of environments and find patterns that could be associated to ecological adaptations for two reasons. First, the conservation of the Sec and Tat secretion systems in all three domains of life indicates that the two systems are ancient and they are shared by many microbes adapted to diverse ecological niches (Albers et al., 2006; Natale et al., 2008; Bohnsack & Schleiff, 2010; Reynolds et al., 2011). Second, similar to the Tat pathway, we can identify molecular patterns in the Sec system that may contribute to bacterial pathogenesis from a host-pathogen coevolution perspective. This can be unraveled by using molecular coevolution and functional divergence analyses. Moreover, reliable computational tools have been developed and can predict substrate proteins (secreted proteins) transported by the two secretion systems (Emanuelsson et al., 2007; Bendtsen et al., 2005b).The study of secretion systems and their secreted proteins from bacteria that directly interact with host and host immune response will provide insight into the molecular basis responsible for bacterial pathogenesis. Identification of secretion systems and their secreted proteins is a complicated task and requires interrogating the proteome through appropriate approaches. One way to achieve this is using comparative genomics.

Comparative genomic studies of prokaryotes with different biological features (e.g., intracellular pathogen and free-living non-pathogenic bacteria or even pathogens with facultative cellular life) can aid in the identification of the molecular patterns that are associated with specific prokaryotic lifestyles. Moreover, linking molecular evolutionary patterns of proteins from the secretion systems and their dependent-substrate proteins (secretome) can contribute to our understanding of the pathogenic potential of bacteria. To date (by February

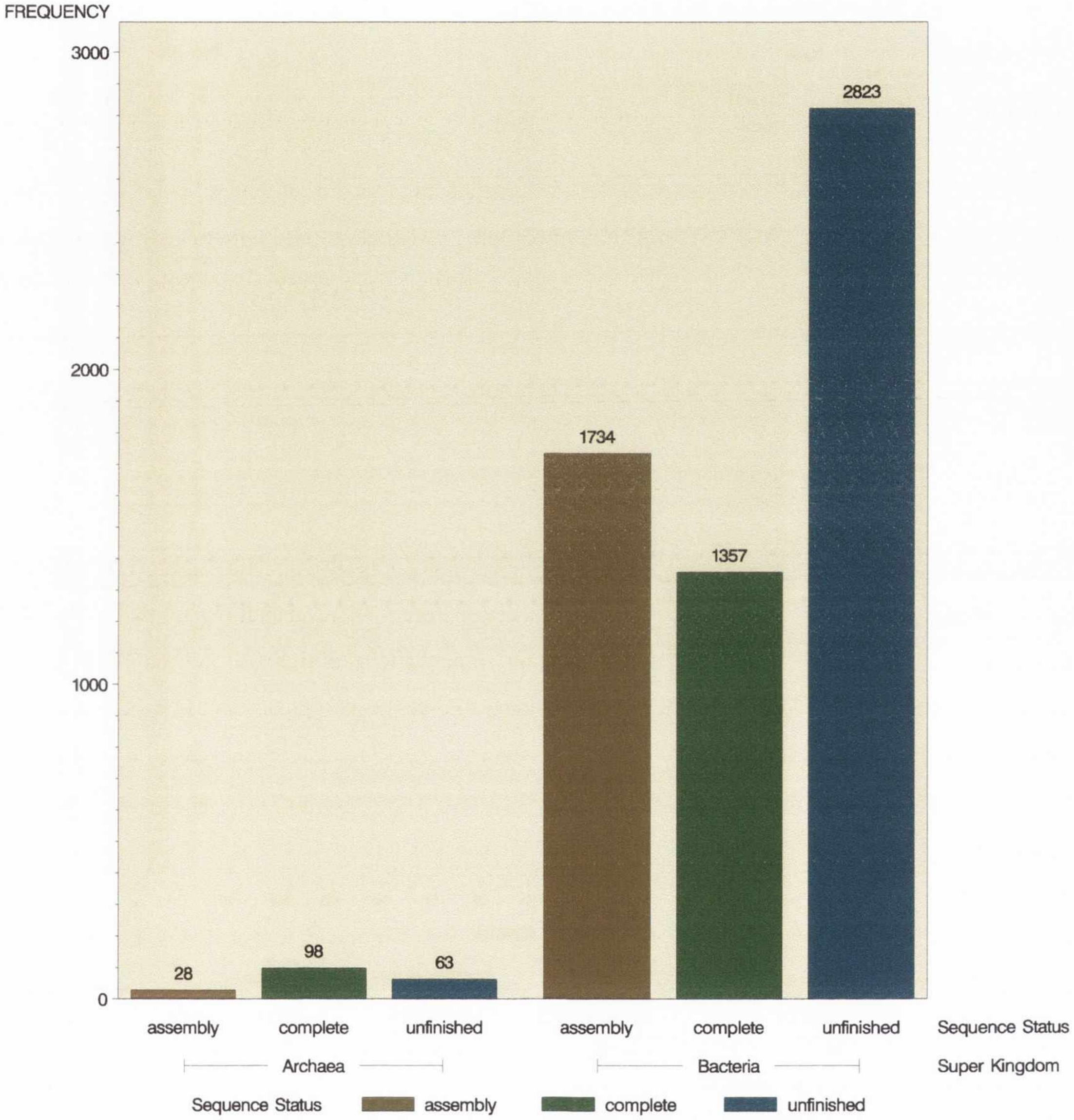


Figure 1.4: Bar chart from the National Center for Biotechnology Information (NCBI) prokaryotic genome projects. Initiated prokaryotic genome projects to date (23-02-2011) are plotted. Archaea and bacterial genome projects are summarized according to their status: assembly, complete and unfinished, respectively.

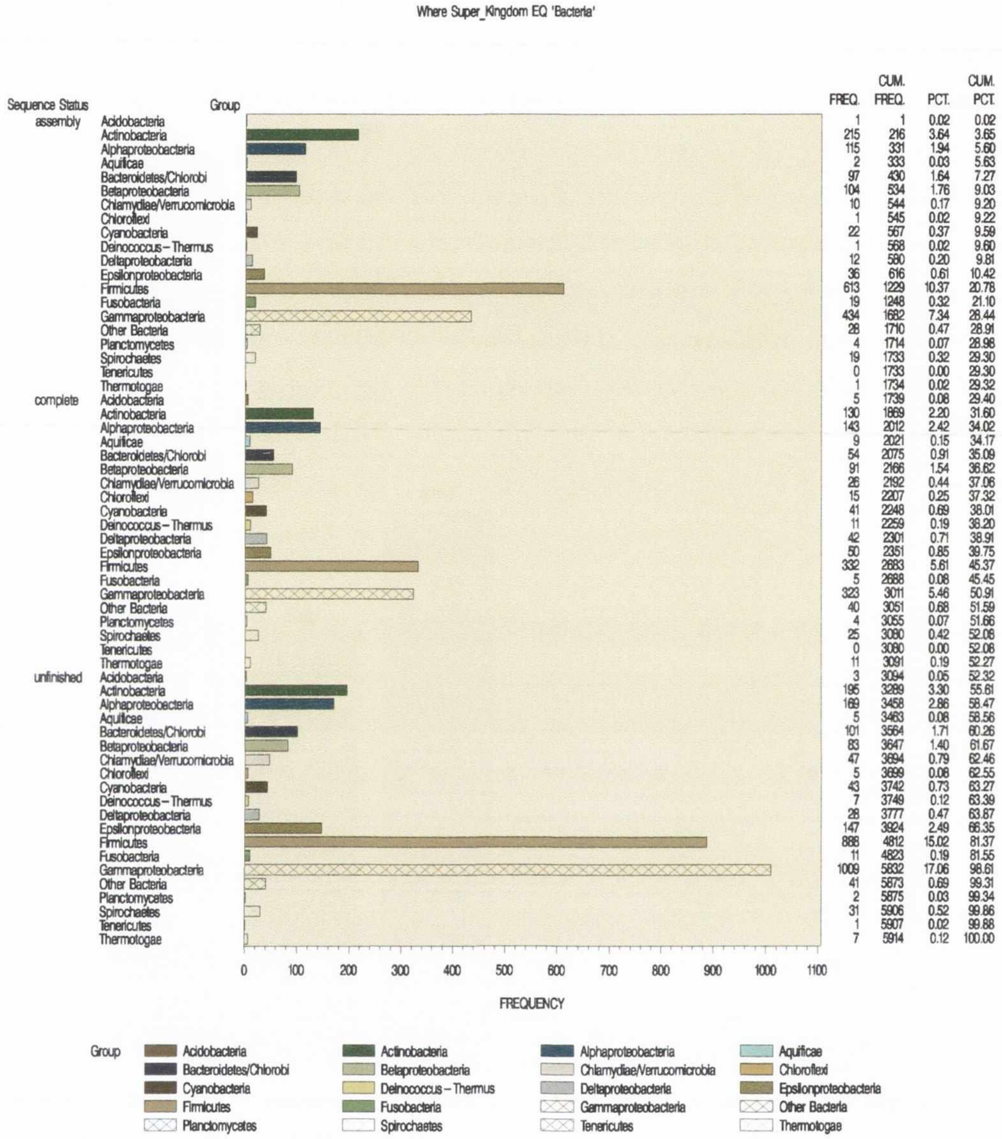


Figure 1.5: A summary of currently initiated bacterial genomes (23-02-2011). The summary is presented in a bar chart from NCBI bacterial genome projects. Right columns denote the frequency (FREQ.) of initiated genomes, cumulative frequency (CUM. FREQ.), percentage (PCT.) and cumulative percentage (CUM. PCT.) for corresponding left bar. Genome projects are categorised according to their sequencing status: assembly, complete and unfinished, respectively, where genome projects are further classified into major phylogenetic groups as shown at the bottom of the figure.

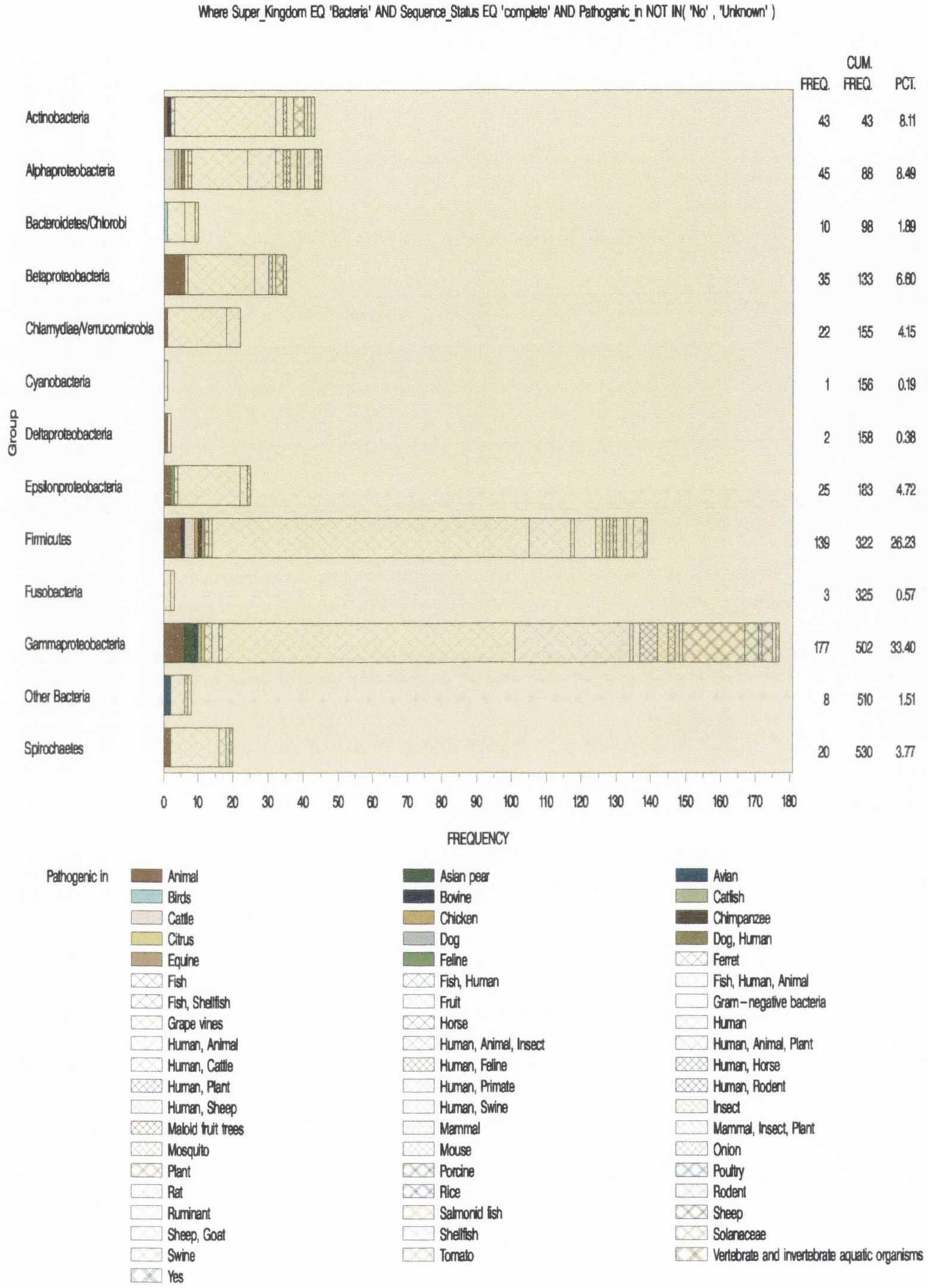
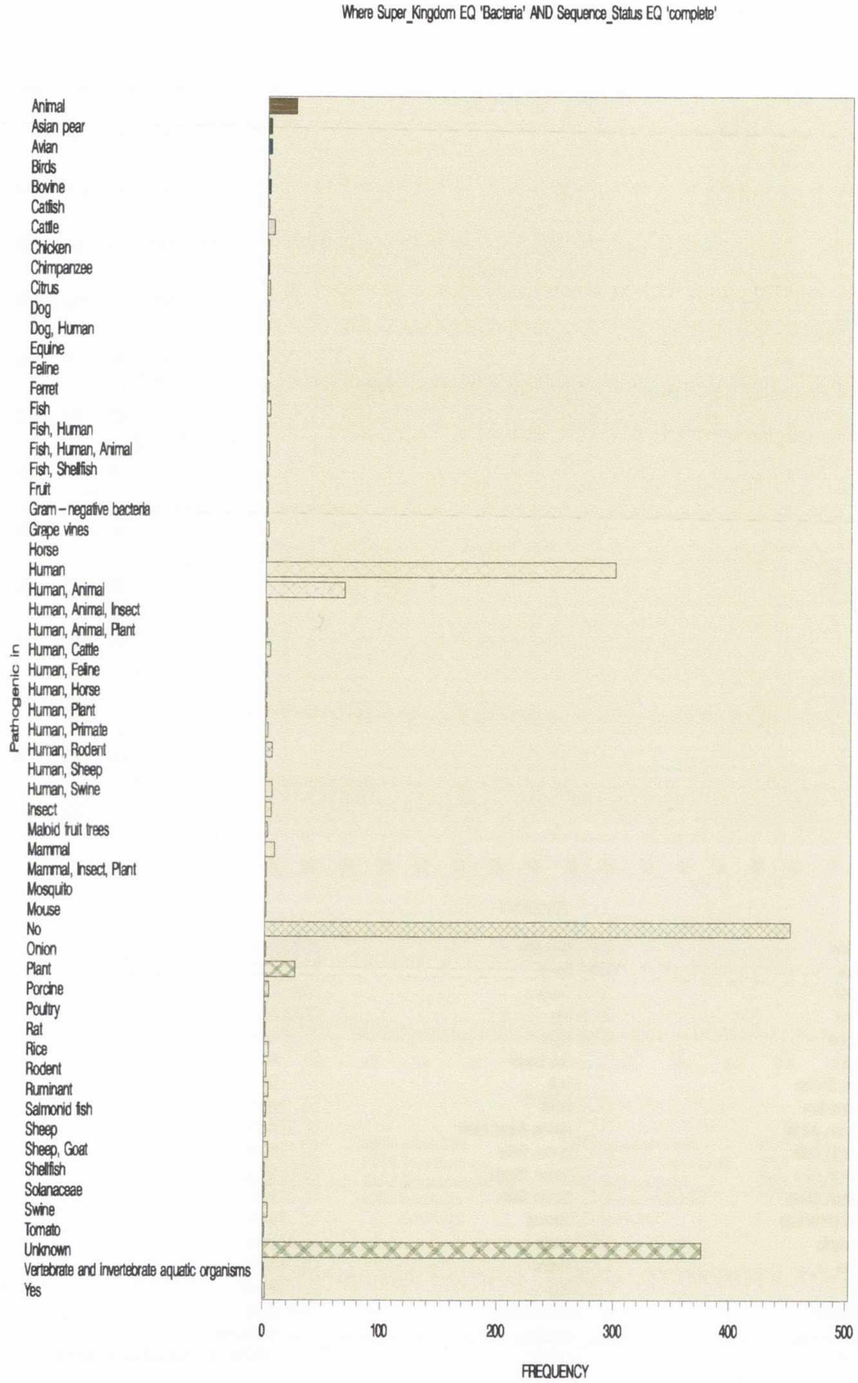


Figure 1.6: Bar chart stacked with different host categories for sequenced pathogenic bacteria groups to date. The left column is the sequenced bacteria group. The right columns denote the frequency (FREQ.), cumulative frequency (CUM. FREQ.), percentage (PCT.) and cumulative percentage (CUM. PCT.) of sequenced bacterial genomes for corresponding left bar (data source from NCBI Genome project database).



Pathogenic in	FREQ.	CUM. FREQ.	PCT.	CUM. PCT.
Animal	24	24	1.77	1.77
Asian pear	3	27	0.22	1.99
Avian	3	30	0.22	2.21
Birds	1	31	0.07	2.28
Bovine	2	33	0.15	2.43
Catfish	1	34	0.07	2.51
Cattle	6	40	0.44	2.95
Chicken	1	41	0.07	3.02
Chimpanzee	1	42	0.07	3.10
Citrus	2	44	0.15	3.24
Dog	1	45	0.07	3.32
Dog, Human	1	46	0.07	3.39
Equine	1	47	0.07	3.46
Feline	1	48	0.07	3.54
Ferret	1	49	0.07	3.61
Fish	3	52	0.22	3.83
Fish, Human	1	53	0.07	3.91
Fish, Human, Animal	2	55	0.15	4.05
Fish, Shellfish	1	56	0.07	4.13
Fruit	1	57	0.07	4.20
Gram-negative bacteria	1	58	0.07	4.27
Grape vines	2	60	0.15	4.42
Horse	1	61	0.07	4.50
Human	299	360	22.03	26.53
Human, Animal	67	427	4.94	31.47
Human, Animal, Insect	1	428	0.07	31.54
Human, Animal, Plant	1	429	0.07	31.61
Human, Cattle	4	433	0.29	31.91
Human, Feline	1	434	0.07	31.98
Human, Horse	1	435	0.07	32.06
Human, Plant	1	436	0.07	32.13
Human, Primate	2	438	0.15	32.28
Human, Rodent	6	444	0.44	32.72
Human, Sheep	1	445	0.07	32.79
Human, Swine	6	451	0.44	33.24
Insect	5	456	0.37	33.60
Maloid fruit trees	2	458	0.15	33.75
Mammal	8	466	0.59	34.34
Mammal, Insect, Plant	1	467	0.07	34.41
Mosquito	1	468	0.07	34.49
Mouse	1	469	0.07	34.56
No	451	920	33.24	67.80
Onion	1	921	0.07	67.87
Plant	26	947	1.92	69.79
Porcine	4	951	0.29	70.08
Poultry	1	952	0.07	70.15
Rat	1	953	0.07	70.23
Rice	4	957	0.29	70.52
Rodent	2	959	0.15	70.67
Ruminant	4	963	0.29	70.97
Salmonid fish	2	965	0.15	71.11
Sheep	2	967	0.15	71.26
Sheep, Goat	4	971	0.29	71.55
Shellfish	1	972	0.07	71.63
Solanaceae	1	973	0.07	71.70
Swine	4	977	0.29	72.00
Tomato	1	978	0.07	72.07
Unknown	376	1354	27.71	99.78
Vertebrate and invertebrate aquatic organisms	1	1355	0.07	99.85
Yes	2	1357	0.15	100.00

Figure 1.7: Bar chart for sequenced bacterial genomes with host information to date. The left column is the hosts that bacteria are pathogenic in. The right columns denote the frequency (FREQ.), cumulative frequency (CUM. FREQ.), percentage (PCT.) and cumulative percentage (CUM. PCT.) of sequenced bacterial genomes for corresponding left bar.

23, 2011), there are 5914 bacterial genome projects initiated (1357 completed, 1734 in assembly and 2823 unfinished, Figure 1.4) and more are expected to come. There are however less sequenced archaea genomes (98 completed, 28 in assembly and 63 unfinished). Among finished bacterial genomes (Figure 1.5) the proteobacteria group has the most sequenced bacterial genomes to date, which includes 143, 91, 42, 90, 323 sequenced genomes for alpha, beta, gamma, delta and epsilon subgroups, respectively. The majority of sequenced bacterial pathogens are human pathogens (299 genomes, Figure 1.6 and Figure 1.7). In the completed 1375 bacterial genomes, there are 451 non-pathogenic, 530 pathogenic and 376 pathogenic status unknown bacterial genomes (Figure 1.7). To this end, the genomes of major bacterial pathogens are sequenced and the diseases they cause are summarized in the supplementary information of this chapter (Supplementary Table S1.1). This wealth of information of pathogens offers unprecedented opportunities for conducting detailed research to unearth the molecular and genetic basis of pathogenesis.

Chapter 2

Proteome-wide Analysis of Functional Divergence Reveals the Molecular Basis of Adaptations in Bacteria

2.1 Related manuscript

Williams, T. A., Caffrey, B. E., Jiang, X., Toft, C., and M. A. Fares. (2009). Phylogenomic inference of functional divergence. BMC Bioinformatics 10 (Suppl. 13): P4.

Caffrey, B.E.*, Williams, T.A.*, Jiang, X., Toft, C., Hokamp, K.& Fares, M.A. (2011). Proteome-wide analysis of functional divergence reveals the molecular basis of ecological adaptations in bacteria. Submitted.

* denotes joint first authors.

This is a joint work. FMA conceived the project and designed the original algorithm. WTA, TC and FMA wrote the original Perl version of the software, which was used in an early published work (see Introduction for details). CBE wrote the C++ version of the software Clusterfunc (Clusters of functional categories) for this project. WTA, CBE and JX prepared the data. FMA, TAW, CBE and JX analysed the data. FMA, WTA, CBE, and JX wrote the manuscript. FMA and HK provided supervision of the project.

2.2 Abstract

Functional divergence is the process by which new genes and functions originate through the modification of existing ones. Both genetic and environmental factors influence the evolution of new functions. Novel functions occur at the expense of ancestral functions and require changes in selective constraints to become fixed. Detecting shifts in selective forces at constrained protein regions can aid in uncovering changes in functions. Identification of functional divergence is crucial in the understanding of the interaction between environmental and genetic factors and the generation of biological innovation. Bacteria are a paradigm of biological diversification enabled through an astonishing metabolic repertoire of functions. Understanding how shifts in selection constraints have contributed to the functional divergence in these bacteria may uncover the rules of their metabolic innovation. Using a novel approach we infer events of functional divergence from gene trees, and use patterns of divergence across gene functional categories to cluster bacterial species. We apply our method to 750 complete bacterial proteomes to identify high level patterns of functional divergence and link these patterns to ecological adaptations. We show that different types of genes have different propensities for functional change, that the evolution of pathogenic and symbiotic bacteria is constrained by their association with the host, and also identify unusual events of functional divergence even in well-studied bacteria such as *Escherichia coli*. We present a novel approach for comparing the roles of phylogeny and ecology in functional divergence at the level of entire proteomes.

2.3 Introduction

Most new genes, functions, and activities originate through the modification of existing ones (Ohno, 1970). The evolutionary process that gives rise to functional differences between related genes is called functional divergence (Lynch & Conery, 2000; Conant & Wolfe, 2008). At the species level, functional diversification is primarily associated with adaptive radiations, when a single ancestor differentiates into multiple descendant species, each adapting by natural selection to one of a new set of ecological niches (Schluter, 2000). Following this theory, environmental variation triggers divergent natural selection, leading to the emergence of niche specialists. In many cases, species under the same ecological conditions differ in their ability to adapt to new niches, even when they stem from the same ancestor (Gillespie, 2004; Pinto et al., 2008). Therefore, other factors such as genetic constraints also play an important role in the process of functional divergence.

The process of functional divergence, or departure of a gene from its ancestral function, is constrained by the requirement to maintain the original function: mutations that confer a new function are likely to interfere with the ancestral function and therefore are eliminated by negative selection. This constraint can be relaxed when selection for the ancestral function is weakened, either through gene duplication (and therefore redundancy), or through changes to the environment inhabited by the organism. After gene duplication, one copy of the gene may be free to evolve in a new direction if the other continues to perform the ancestral function (neofunctionalization). Alternatively, ancestral functions can be partitioned between the two gene copies, potentially leading to later specialization or subfunctionalization (Ohno, 1970; Lynch & Conery, 2000; Lynch & Katju, 2004; Conant & Wolfe, 2008; Innan & Kondrashov, 2010). Major changes in the environment or ecological niche can also lead to a relaxation of selective constraint on ancestral functions, although this process is less well characterized. For example, endosymbiotic bacteria have lost many of the genes their free-living relatives need to obtain nutrients from the environment (Moran, 2002), but have also experienced functional divergence in certain genes (Toft et al., 2009).

Prokaryotes are extraordinarily rich in biological diversity, whether measured in terms of number of species (Dykhuizen, 1998; Gans et al., 2005), habitat range (Pikuta et al., 2007), or the breadth of energy sources and biochemical pathways they can exploit in order to survive (Pace, 1997). Even photosynthesis and oxidative phosphorylation–the mainstays of eukaryotic energy metabolism–are bacterial inventions acquired by endosymbiosis during early eukaryote evolution (Dyall et al., 2004). How did this prokaryotic diversity evolve, particularly when the fixation of gene duplications appears to be somewhat more frequent in eukaryotes (Zhang, 2003)? Adaptive evolution in prokaryotes is promoted by at least three main factors: first, a high strength of selection relative to eukaryotes, on account of their generally large population sizes (Lynch & Conery, 2003); second, their ability to obtain genes by horizontal gene transfer (HGT), which enables the sharing of niche-relevant functions between distantly-related microbes living in the same environment (Ochman et al., 2000); and third, their use of stress-induced hypermutation (McKenzie et al., 2000), which may increase the production of adaptive variants as a "last gasp" response to a challenging environment.

Although we know that these processes can drive ecological adaptation in prokaryotes, identifying the fraction of genetic variation that is associated with these functional changes remains a challenging problem. In the case of bacteria, whole genome analyses must take into account widespread HGT, which means that different genes often disagree on the overall species tree (Dagan & Martin, 2006). This is a considerable problem for analyses of functional divergence, which require a tree in order to determine the branch upon which a particular trait arose. The rationale for previous methods, and indeed the new approach described here, derives from the neutral theory of Kimura (Kimura, 1983), which predicts that residues important for the function of a protein will be under strong functional constraint and therefore evolve slowly. These considerations have motivated the development of a number of methods for identifying changes in selective constraints on protein-coding genes and on single amino acid sites and lineages in a phylogenetic tree (Goldman & Yang, 1994; Nielsen & Yang, 1998; Suzuki & Gojobori, 1999; Yang & Bielawski, 2000; Fares et al., 2002a; Yang & Nielsen, 2002; Suzuki, 2004a,b; Zhang, 2004; Berglund et al., 2005; Zhang et al., 2005). At the protein level, Gu (Gu, 1999, 2001, 2006) developed a Bayesian approach to identify functional divergence, which has become the most widely used. Comparisons of amino acid site-specific evolutionary rate or residue conservation between two homologous clades can therefore be used to identify amino acid sites at which selective constraints have changed, potentially indicating functional divergence.

Recently, we have developed a new distance-based method which explores a bifurcating phylogenetic tree, testing for functional divergence at each node by comparing the two downstream clades to an outgroup in order to identify sites at which selective constraints have shifted (Toft et al., 2009; Williams et al., 2010). Similar to other methods, our approach was limited to tests of one gene at a time, unless the phylogeny of all genes could be fixed in advance. In the present study, we have optimised our method to (i) handle analyses of functional divergence that include hundreds of complete proteomes, (ii) address the fact that the phylogenies of individual proteins do not necessarily agree with the true phylogeny, as is often the case with organisms that acquire genes through HGT, (iii) provide an intuitive probability assignment for each test which takes the underlying phylogeny of the sequences into account and (iv) explore all levels of each gene tree, testing for functional divergence at each node. We infer patterns of radical change for each protein individually, and then cluster species according to the functional categories (derived from COG (Tatusov et al., 2003)) in which they exhibit significant evidence of functional divergence. We provide a fast, open source implementation of our method in the C++ program Clusterfunc (Clusters of functional categories). We perform an analysis of functional divergence on 750 bacterial proteomes. This set includes bacteria from various different ecological niches and therefore provides a good dataset for identifying ecology-related functional divergence. Our approach (i) reveals striking patterns of convergent evolution in phylogenetically distinct but ecologically related groups of bacteria, including pathogens, endosymbionts, and thermophiles, (ii) provides additional support for the view that bacteria have a conserved set of core functions, with a more variable metabolic layer and (iii) provides a detailed picture of how individual species of unusual bacteria have diverged from their closest relatives.

2.4 Materials and Methods

Our analysis of functional divergence, the individual steps of which are detailed below, is summarized in Figure 2.1.

2.4.1 Sequences, homology, and alignment

The first step in a whole-proteome analysis of functional divergence is the grouping of orthologs within the species of interest. We leave orthology assignment for which a number of tools are already in use (Altenhoff

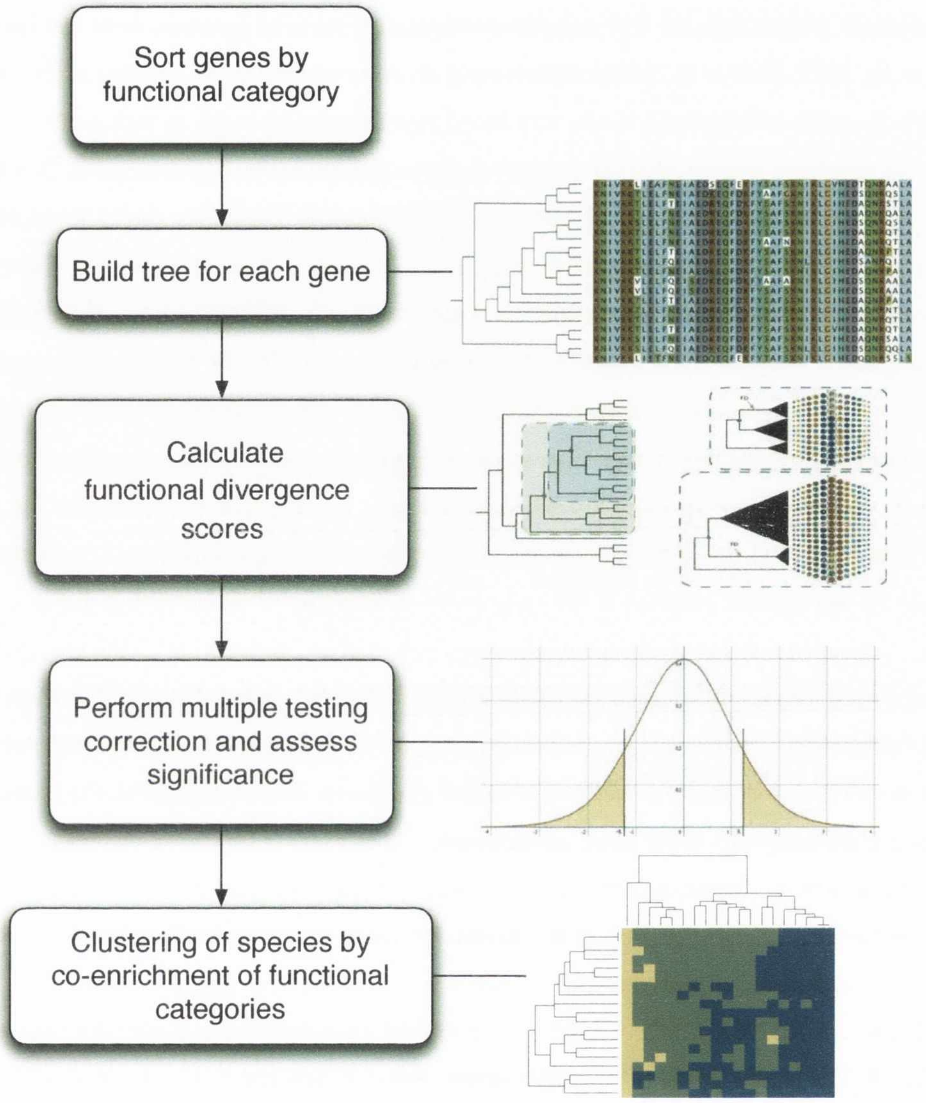

Sort genes by functional category
Build tree for each gene
Calculate functional divergence scores
Perform multiple testing correction and assess significance
Clustering of species by co-enrichment of functional categories

Figure 2.1: Clusterfunc program workflow. After alignments have been built for each gene in the analysis, the alignments are sorted by functional category. In this case, the COG system was used (Tatusov et al., 2003), but the user is free to pick an ontology of their choice. Trees are built for each gene using BIONJ (Gascuel, 1997) and the JTT substitution model (Jones et al., 1992), and sites are scored for functional divergence on each branch. Significance is assessed by simulating a distribution of test scores under a model of neutral evolution, taking the real phylogeny into account and using the False Discovery Rate approach to correct for multiple testing. For each species and functional category, we use chi-squared tests to evaluate whether the species is enriched, impoverished, or neither enriched nor impoverished for functional divergence in that category, and then cluster species according to similarities in their profile across all 19 categories. This approach enables us to account for HGT while identifying interesting and atypical patterns of functional change in the data, as discussed in the main text.

& Dessimoz, 2009), to the users' choice according to their own needs. For the present analysis, we retrieved pairwise orthology assignments for 750 completely-sequenced bacterial genomes from the OMA database (Schneider et al., 2007; Roth et al., 2008), representing all bacterial data in the October 2009 revision of the database. We chose the OMA project for its very broad phylogenetic coverage, as well as the favourable performance of its algorithm against other current orthology assignment methods (Altenhoff & Dessimoz, 2009). In addition to providing pairwise orthology calls, the OMA algorithm assembles strict orthologous groups in which every member is directly orthologous to every other. The rationale for this strict approach to grouping is the exclusion of paralogs, which is important for a number of potential applications of the OMA database, such as phylogenetic analysis. Unfortunately, these groups are unsuitable for functional divergence analysis across large phylogenetic distances because lineage-specific gene duplications tend to break up genuine orthologs into multiple, overlapping groups (that is, clustering problems arise because pairwise orthologies are not necessarily transitive). Using these groups in our analysis would result in multiple testing of the same clade, each time with overlapping but incomplete sampling of downstream sequences. The inclusion of both orthologs and lineage-specific paralogs in the same group is, however, of no concern in our per species comparison of divergence between different functional categories of genes, because our method relies on individual gene trees and not a single "species tree" to detect functional divergence (see below). Therefore, we decided to build our own groups from the pairwise orthology assignments in OMA, with the less stringent requirement that a chain of pairwise orthologies connect all members of a group. This strategy produces groups containing all orthologs and paralogs for a given gene, as appropriate for analysis of functional divergence. However, the approach is vulnerable to erroneous orthology calls in the original database, because a single false call will cause two unrelated groups of sequences to be merged. To assess the possible effect of false OMA orthology assignments on our dataset, we used the relevant genomic data at NCBI to assign COG ontology tags to each sequence (Tatusov et al., 2001, 2003). We then calculated the frequency of the modal COG tag in each group (see Figure 2.2). The largest group (4,788 alignments) contained only one COG tag each, validating our approach to grouping homologs (since COG categories are relatively broad, related sequences are expected to be annotated with the same tag). To avoid ambiguity in the clustering of functional categories, we only analyzed these single-tag alignments. We then filtered out poorly characterized groups (annotated with the ambiguous R or S COG categories) and any group containing less than 9 sequences, which we chose as the minimum number required for analysis. This resulted in a final dataset of 3,813 groups, which were then analyzed with Clusterfunc. Sequence alignments were built for each group with MUSCLE (Edgar 2004), using the default parameters. Data on the ecological niches occupied by the species included in the analysis was retrieved from HAMAP (Lima et al., 2009) and from the Genome database at NCBI. A typical alignment of 78 sequences takes 2 minutes and 40 seconds to analyze on a standard desktop computer, including NJ tree-building. At the other extreme, the large-scale analysis reported below (44,416 tests of functional divergence/3,813 alignments) took 60 hours on a 40-node cluster.

2.4.2 Building gene trees

When analyzing entire proteomes for functional divergence, the use of a species tree to infer events on each branch is problematic: extensive horizontal gene transfer (HGT), particularly among prokaryotes, means that genomes may not be related in a tree-like way (Dagan & Martin, 2006). We therefore calculated a tree for each gene (set of homologous sequences) in the dataset using BIONJ (Gascuel, 1997), under the JTT model of protein sequence evolution (Jones et al., 1992). Calculations for that gene were then made exclusively using the resulting tree.

Scoring functional divergence

We here define functional divergence as the potential departure of the derived protein function from its ancestral one as a result of amino acid changes at important functional sites. Therefore functional divergence is detected on the basis of shifts in selective constraints on proteins. This analysis of functional divergence can be used to provide a list of candidate genes for further experimental testing. Our method identifies statistically significant changes in the selective constraints on particular amino acid sites of a protein in each of the lineages of a tree. The method steps through the phylogenetic tree and calculates functional divergence scores at each of the inner nodes. For each site of the protein, our approach compares the amino acid composition between two clades to that of an outgroup. This comparison is performed using BLOSUM62 amino acid substitution matrix (Henikoff & Henikoff, 1992), indeed any substitution matrix can be used. BLOSUM62 and related matrices provide an empirical measure of the likelihood of the transition of one amino acid to any of the other 20 (including its conservation). Scores of functional divergence (FD score) for each column are given by:

$$FD_{score} = \frac{\overline{X}_1 - \overline{X}_2}{S_{X_1 - X_2}} (1)$$

where $\overline{X}_1, \overline{X}_2$ are the mean substitution scores for the transition from clades on either side of the bifurcation in the phylogenetic tree relative to the outgroup and $S_{X_1 - X_2}$, the standard error for unequal sample sizes with unequal variances is given by:

$$S_{X_1 - X_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}(2)}$$

Significance testing

To test the significance of functional divergence events, we simulated multiple sequence alignments of the same size as the real alignment but in which proteins were evolved under a neutral evolution model. Because functional divergence was tested in protein alignments, the seed ancestral sequence was protein based and this evolved under the JCprot model that assumes equal probability transitions among the 20 amino acids (Jukes & Cantor, 1969). For our simulations, we used the gene-specific tree topology and branch lengths calculated above. We built at least 1000 such simulated alignments (more simulations are created if mean and standard deviation have not converged after 1000), in each of which we searched for functional divergence and calculated a score according to equation (1). This search resulted in a null distribution of the test score against which P-values for the real data were calculated. These values were then corrected for multiple testing by the False Discovery Rate method (Benjamini & Hochberg, 1995) using an alpha value of 5% as the threshold of significance. Following this procedure, branches on the tree that still possess at least one significant site were considered to be under functional divergence for the purposes of enrichment and clustering.

2.4.3 Enrichment analysis

Once all alignments were analyzed, we performed three different enrichment tests to ask three different biological questions. These are based on a chi-squared test:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} (3)$$

Where Oi is the observed frequency of genes/alignments under functional divergence, E_i is the expected frequency and n is the number of possible outcomes of each event. We used the enrichment tests to identify (i) species and (ii) categories of genes that experienced significantly more (enriched: $O_i - E_i > 0$) or significantly less (impoverished: $O_i - E_i < 0$) functional divergence when compared to the background level (that is, P < 0.05 in a chi-squared test). We then calculated (iii) the enrichment status of each category within each species, in order to identify lineage-specific shifts in the pattern of functional divergence.

2.4.4 Hierarchical clustering

We created a heatmap from the enrichment status of functional categories within species to help visualize the structure in our large dataset. To do this we used the heatmap.2 function from the *gplots* library in R (R Development Core Team, 2010). This function performs two-dimensional hierarchical clustering according to Euclidean distance and outputs a heatmap together with a corresponding dendrogram. Visualizing the results

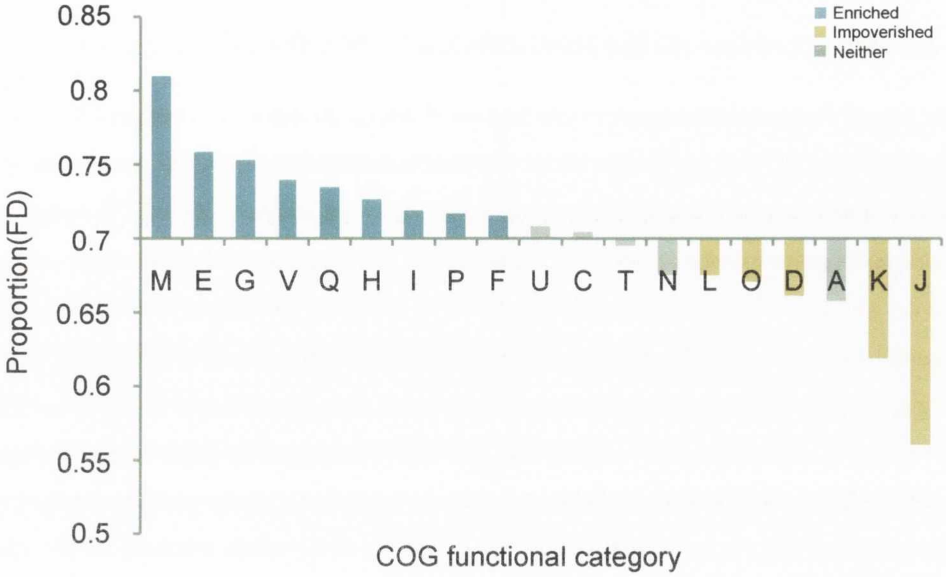


Figure 2.2: Different categories of genes experience different levels of functional divergence. Proportion (FD) is the proportion of tested branches with at least one functionally divergent site across all gene trees in a particular functional category. Categories are labeled according to the COG ontology system (Tatusov et al., 2001, 2003). Fourteen of the nineteen categories fall into two groups: significantly enriched or impoverished. Most information processing genes (K, J, L) fall into the latter group, while metabolic functions (E,F,G,H,P,Q) and genes involved in defense (V) or found on the cell surface (M) are enriched for radical change. RNA processing and modification (A) genes are not significantly impoverished, despite a relatively low proportion of positive tests for functional divergence; this may be due to the small number of tests in this category. Full table of enrichment for functional divergence in the different categories is provided in Table S2.1.

of the analysis in this way allows identifying unusual patterns of functional divergence in particular functional categories or convergent functional divergence among phylogenetically unrelated sequences.

2.4.5 Implementation

Clusterfunc was implemented in C++ and is available under the GNU General Public License v.3 for Linux, Mac and Windows. The code was written using the GNU Scientific Library and the Bio++ libraries (Dutheil et al., 2006). The program is accompanied by full documentation and enables the user to perform several different kinds of analysis, including the identification of lineage-specific functional divergence in a gene-of-interest (such as that reported by (Williams et al., 2010)) and the kind of multi-proteome investigation reported here. The latest version of the code and documentation is available at `http://www.bioinformatics.org/clusterfunc`.

2.5 Results and Discussion

2.5.1 Cell envelope proteins are the most affected by functional divergence in bacteria

An obvious sign of functional divergence would be a set of homologs that spans multiple COG categories. In this study we focus only on those alignments where all sequences have the same COG annotation. This represents the majority of homologs and is a reflection of the relatively broad character of the COG categories. The kinds of functional shifts that we detect on the basis of conserved, radical amino acid substitutions are therefore more subtle and not noticeable from simply comparing the COG classifications across homologous sequences. We used chi-squared tests to evaluate the differences in functional divergence between COG gene categories in our dataset (see Figure 2.2). We compared the proportion of positive tests for functional divergence within each of the 19 COG categories to the background expectation, which was calculated by combining all categories. If genes in different functional categories have similar propensities to undergo functional divergence, we would expect the proportion of positive tests in each category to be similar to the mean, resulting in few significant cases of enrichment. However, fourteen of the nineteen categories were either enriched or impoverished for functional divergence, while only five categories failed to deviate significantly from the background expectation. To test whether this polarization of our dataset was simply due to an artifact–for instance, the use of a non-conservative enrichment test–we performed simulations in which the genes in our original dataset were randomly assigned to one of the 19 COG categories before testing for enrichment. In these simulations, events of functional divergence were much more evenly distributed among the categories, so that 93% of categories were neither enriched nor impoverished for divergence relative to the background level. This result indicates that the probability of functional change is not evenly distributed among the real categories: there is a stark division between enriched and impoverished categories. This supports the idea that bacterial proteomes comprise a relatively unchanging core (that is, genes in impoverished categories) coupled with a set of more variable functions (enriched categories) (Lake et al., 1999; Makarova et al., 1999; Mushegian & Koonin, 1996).

The impoverished categories are almost exclusively those involved with information storage and processing, including DNA replication, recombination, and repair (L); transcription (K), ribosome biogenesis (J); and cell division (D). Metabolic genes were among those enriched for functional divergence, including genes involved in the metabolism of coenzymes (H), secondary metabolites (Q), carbohydrates (G), amino acids (E) and nucleotides (F). Along with these metabolic categories, cell wall and envelope genes (M) and cellular defense mechanisms (V) were among the most enriched categories in our analysis, highlighting the critical role of the environment in directing lineage-specific episodes of functional change. Taken together, our results agree with a number of previous reports indicating that proteins involved in information processing are more conserved across large evolutionary distances than those involved in metabolism (Azuma & Ota, 2009; Lake et al., 1999;

Makarova et al., 1999; Mushegian & Koonin, 1996). An additional point bears emphasizing here: since our method controls for the level of conservation at each node on the tree, the significance of a particular substitution pattern depends on the background evolutionary rate: in slow-evolving proteins, relatively conservative substitutions can be detected as significant events of functional divergence, whereas only very unusual substitution patterns will attain significance in fast-evolving proteins. Therefore, our results indicate that information processing genes are not only more conserved than others purely in terms of evolutionary rate, but that they also experience less functional change even taking this low rate of sequence evolution into account. Why are informational genes under greater functional constraint than the rest of the proteome? One possibility, which follows Crick's concept of the "frozen accident" (Crick, 1968), is that too many other genes depend on the basic functions of translation, transcription, and repair: functional changes in these genes would disrupt many other systems in the cell. This hypothesis is supported by the observation that the COG category containing protein trafficking and chaperones (O) is also impoverished: the core activities of generalist chaperones such as GroEL and DnaK are required for the proper folding of many different proteins in bacterial cells (Lund, 2009).

2.5.2 Host interactions constrain functional change in pathogenic and symbiotic bacteria

Does the ecological niche of an organism influence the pattern of functional change it experiences? To answer this question, we evaluated the enrichment of functional divergence in each species relative to the others in our dataset. To calculate the enrichment status of each species, we used the same statistical strategy as employed for enrichment by functional category: we calculated a background proportion of successful tests for functional divergence over all species, and then compared this to the proportion for each species individually using chi-squared tests. We also used chi-squared tests to identify associations between these three enrichment patterns (enrichment, impoverishment, or neither) and organism lifestyle, as is summarized in Table 2.1. While there was no statistically significant difference between psychrophiles and mesophiles in terms of functional divergence (chi-squared = 0.9762, P = 0.6138), thermophilic bacteria were significantly more likely to have proteomes enriched for functional divergence in comparison to mesophiles (chi-squared = 38.1, P = 1.2 x 10-7). This enrichment in the set of phylogenetically scattered thermophiles can be explained by the convergent adaptation of their proteomes to higher temperatures, a process that requires changes in amino acid composition (Singer & Hickey, 2003; Zeldovich et al., 2007). In support of this, thermophiles present a contrasting pattern of functional divergence in comparison to the general pattern, with two COG categories being enriched for functional divergence in thermophiles while being under-enriched in general. These categories are directly related with the survival of cells under heat stress: category K, which comprises mostly transcription factors,

Lifestyle	Comparison	Enriched	Neither	Impoverished	Significance
Psychrophile	Mesophile	2/74	5/393	2/93	N.S.
Thermophile	Mesophile	17/74	20/393	0/93	*** (+/over)
Pathogen	Non-pathogen	19/101	234/289	80/50	*** (-/under)
Intracellular pathogen	Other pathogen	0/19	26/189	4/72	N.S.
Symbiont	Non-symbiont	4/116	34/470	15/111	* (-/under)
Intracellular endosymbiont	All others	1/119	18/486	14/112	*** (-/under)
All interactors	Free-living	34/86	310/194	100/26	*** (-/under)

Table 2.1: Effect of organism lifestyle on functional divergence. Associations between lifestyle and enrichment for functional divergence: the numbers of genomes in each category are given in the form Lifestyle/Comparison. Significance was assessed with Yates-corrected chi-squared tests, or Fisher tests when the expected count was lower than 5 for any one cell in the contingency table. Significance codes: N.S. = $P > 0.05$; * = $P < 0.05$, ** = $P < 0.01$, *** = $P < 0.001$. If an association was significant, "+" or "-" denote the direction of the shift associated with the lifestyle being tested. For instance, interactors are significantly impoverished (-) compared to free-living bacteria.

and category L, which is involved in DNA replication, recombination and repair. Above certain temperature threshold, molecular pathways undergo dramatic temperature induced alterations that drive to cytotoxicity, radiosensitization and thermotolerance (Laszlo, 1992; Kampinga et al., 2004). Among all the responses that take place in the cell under high temperatures, inhibition of DNA, RNA and protein synthesis is the response that involves a complex and fine-tuning of regulation mechanisms, mainly orchestrated by transcription factors (Laszlo, 1992). One such important regulated mechanism is the induction of heat-shock proteins, particularly involved in mitigating the cytotoxic effects due to the non-specific aggregation of unfolded and denatured proteins (Lepock, 1997; Kregel, 2002).

Interestingly, we found that all bacteria that interact with a host as an integral part of their lifestyle (including pathogens, parasites, symbionts and commensals) were significantly impoverished for functional divergence in comparison to their free-living relatives (see Table 2.1). This result is somewhat surprising because pathogens and symbionts generally experience higher rates of evolution than free-living bacteria, although much of the increase can be attributed to heightened genetic drift (Moran, 2002). Our results suggest that once the overall conservation level of proteins is accounted for, these bacteria have undergone less functional change than their free-living relatives. This result can be explained by greater ecological constraints on host-associated bacteria, which must adapt to the highly specific environment of their host. In particular, pathogenic and symbiotic bacteria preferentially lose metabolic genes as they no longer require the capacity to exploit as wide a range of nutrient sources as free-living bacteria (Moran, 2002). Since these are precisely the kind of genes that are most amenable to functional change (Figure 2.2), their loss from host-associated bacteria explains the relative impoverishment of functional divergence in these proteomes. Also, the remaining genes are under strong constrains imposed by the specialized environment they are in, limiting therefore any opportunity for functional divergence (Toft et al., 2009). However, variability in genome size is a complicating factor in this analysis because host-associated bacteria tend to have smaller genomes than their free-living relatives (For instance, endosymbiotic bacteria of insects underwent substantial reduction in the gene content, with genomes sizes ranging between 144 kb and 792 kb depending on the host (in comparison, *E. coli K12* has a genome size of 4.639Mb) (See for example (Degnan et al., 2005; Gil et al., 2002; Nakabachi et al., 2006; Perez-Brocal et al., 2006; Shigenobu et al., 2000; Tamas et al., 2002; van Ham et al., 2003). Since functional divergence often follows gene duplication (Conant & Wolfe, 2008), it might be expected that larger genomes would be enriched for new functions in comparison to smaller ones. Does genome size alone account for the observed differences between host-associated and free-living bacteria? To test this possibility, we modeled genome enrichment and impoverishment for functional divergence as a function of lifestyle (host-associated vs. free-living) and genome size (in nucleotides) using a generalized linear model (see Supplementary Table S2.3). Both terms were significant, with host-associated bacteria significantly more likely to be impoverished (P = 1.28 x 10-13) and, perhaps surprisingly, a modest tendency towards impoverishment in larger genomes (P = 0.0212). Therefore, variation in genome size does not account for the observed differences in functional divergence between host-associated and free-living bacteria.

To better define the effect of lifestyle on functional divergence, we identified the functional categories with the greatest consistent differences in enrichment status between host-associated and free-living bacteria. Interestingly, genes involved in vesicular transport and secretion systems (U) were enriched for functional divergence in host-associated bacteria but neither enriched nor impoverished in free-living bacteria, while signal transduction genes (T) were impoverished in host-associated bacteria but enriched in their free-living relatives. This pattern can be readily understood in terms of the lifestyles of host-associated bacteria, as pathogens use elaborate secretion systems for delivering toxins and other virulence factors to their host (Baron, 2010), while symbionts provision their hosts with nutrients as part of their mutually beneficial relationship (Douglas, 1998; Sandstrom et al., 2000). In addition, the impoverishment in host-associated signal transduction genes may reflect their adaptation to a relatively constant host environment, which is considerably more stable than the fluctuating conditions experienced by their free-living relatives.

2.5.3 Two-dimensional clustering reveals the general patterns of functional divergence in bacteria

In order to visualize the results of our functional divergence analysis, we performed two-dimensional hierarchical clustering on the enrichment status (enriched, impoverished, or neither) associated with each species and functional category–that is, we clustered species according to similarities in their enrichment status across the 19 functional categories, resulting in the heatmap and dendrogram in Figure 2.3 and Figure 2.4. This is a powerful and intuitive way to represent our results because it reveals the overall patterns in the data–such as the extreme conservation among informational genes, particularly those involved in ribosome biogenesis (J) - while also highlighting individual, lineage-specific exceptions to the general trends. In this section, we demonstrate the utility of this approach by using the heatmap to identify species that have undergone major functional shifts.

Although top-level bacterial groups (such as the divisions of the proteobacteria, the *Firmicutes*, *Actinobacteria*, and so on) are not resolved in our dendrogram of functional divergence (Figure 2.4), family and genus-level relationships often are, presumably because of close phylogenetic relatedness, shared gene content, and similarity of ecological niche. This allows us to identify individual species with atypical patterns of functional divergence. A particularly striking case is that of the *Bartonella* genus (Figure 2.4b), which are a group of intracellular parasites that infect and replicate in erythrocytes (Anderson & Neuman, 1997). Of the four

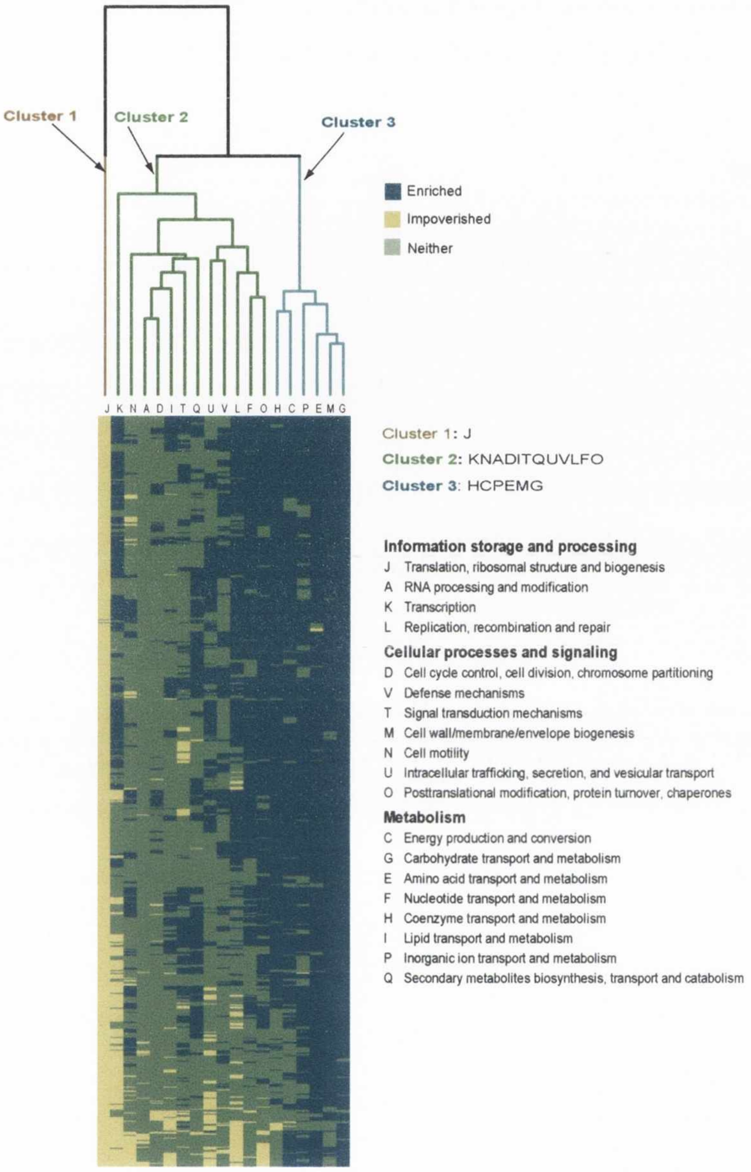


Figure 2.3: Hierarchical clustering of enrichment status of 19 functional categories across 750 bacterial proteomes. The complete heatmap, with a dendrogram corresponding to functional category clustering across the top, and species clustering on the left hand side (not shown here, a larger version of the complete heatmap is available as Supplementary Figure S2.1). 19 functional categories are clustered into three groups, which can be clearly distinguished by their colors (Blue: Enriched; Yellow: Impoverished; Grey: Neither). The first group (cluster 1) contains only protein sets associated with ribosome biogenesis (J), which is evident as they are all yellow across 750 bacteria. The second group (cluster 2) have 12 functional categories (KNADITQUVLFO), which seems to have a mix of all three enrichment status (Mostly grey and blue). The third group (cluster 3) have 6 functional categories (HCPEMG), which seems to be predominately represented by blue color (Enriched for functional divergence). In particular, proteins for cell wall/membrane/envelope biogenesis (M) and carbohydrate transport and metabolism (G) are extremely enriched for functional divergence.

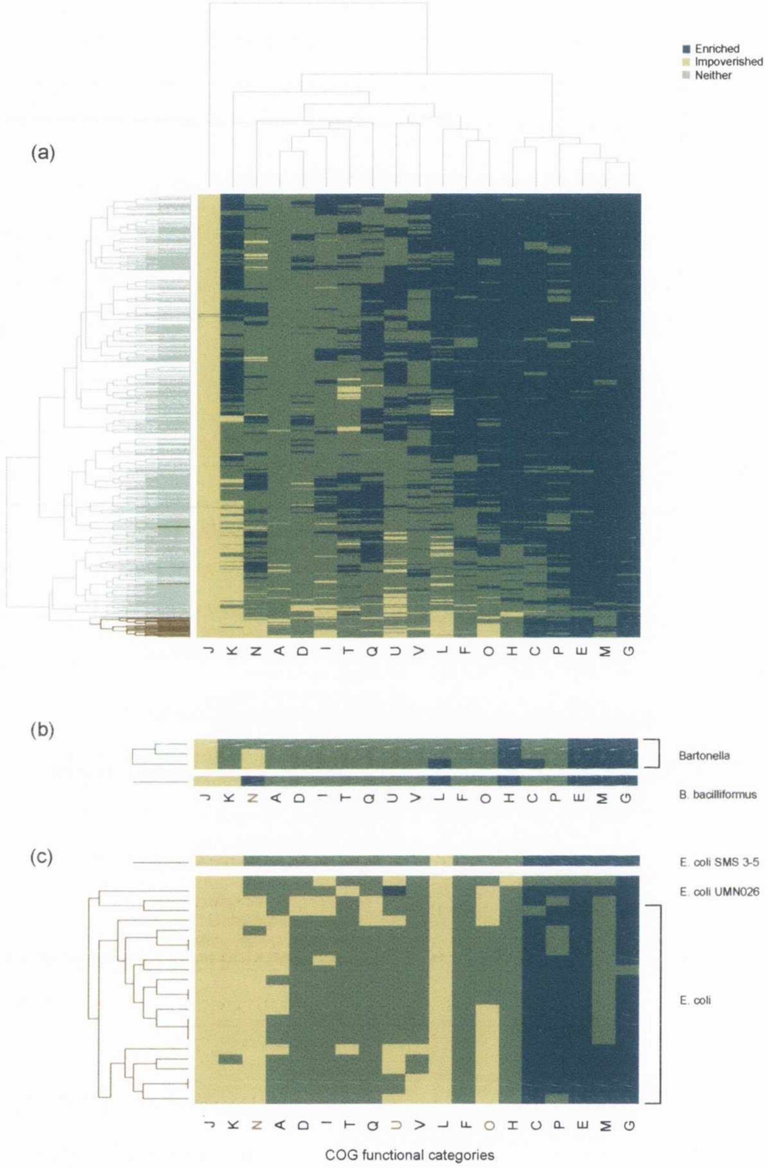


Figure 2.4: Visualizing high-level patterns of functional divergence. We used hierarchical clustering to reveal the main patterns of functional divergence in our dataset of 750 bacterial proteomes. (a) The complete heatmap, with a dendrogram corresponding to category clustering across the top (see Figure 2.3 for details), and species clustering on the left hand side. (b) Lineage-specific events of functional divergence picked out from the heatmap (dendrogram colors denote the regions expanded upon–a larger version of the complete heatmap is available as Supplementary Figure S2.1). Unlike other *Bartonella* species, *B. bacilliformus* is impoverished for divergence in cell motility genes (N), and is unique among *Bartonella* species in using a flagellum to infect erythrocytes. (c) Two strains of *E. coli*–SMS 3-5 and UMN026–have phylogenetically atypical patterns of functional divergence: the constraints on cell motility (N) and chaperone (O) genes are among those that have relaxed relative to the other strains in SMS 3-5, while UMN026 is uniquely enriched for secretion system (U) genes.

Bartonella species in our dataset, only one–*Bartonella bacilliformus*–is enriched for functional divergence in cell motility genes (N), with the others being impoverished (2 species) or neither enriched nor impoverished (1 species). Remarkably, this is the only member of the genus that possesses flagella (Brenner et al., 1991). Since erythrocytes lack an active cytoskeleton, they cannot be induced to take up external bacteria by invagination (Dramsi & Cossart, 1998). Instead, erythrocyte invasion by *Bartonella* species is an active process (Dehio, 2001). The mechanism employed by *Bartonella bacilliformus* involves the use of its flagella (Scherer et al., 1993) and is more efficient than that of other *Bartonella* species, with up to 80% of erythrocytes infected (Dehio, 2001; Ihler, 1996). This appears to be a clear case where our approach has identified an interesting, lineage-specific case of adaptation to a specialized ecological niche.

Our heatmap turns up surprises even among relatively well-characterized species (Figure 2.4c). As expected, closely related *E. coli* and *Shigella* strains cluster together at the bottom of the dendrogram (Figure 2.4a), with one exception: *E. coli SMS 3-5*, a multi-drug-resistant, heavy-metal tolerant strain isolated from a polluted industrial environment (Fricke et al., 2008). This bacterium is distinguished from other *E. coli* strains on the basis of an overall relaxed functional constraint, particularly in the categories of cell motility (N) and protein modification, turnover and chaperones (O): most others are impoverished for functional divergence, while SMS 3-5 is not. This profile correlates well with what is known about the biology of this strain, which is unique among sequenced *E. coli* genomes in possessing a second, intact lateral flagellar system called Flag-2, in addition to the normal peritrichous flagella found in other *E. coli* strains (Fricke et al., 2008; Ren et al., 2005). This system was originally characterized in a different strain, 042, where it has been rendered nonfunctional by a frame shift mutation in one of the component genes (Ren et al., 2005), although it appears to be complete in SMS 3-5 (Fricke et al., 2008). Interestingly, the Flag-2 gene cluster also contains flagellum-specific chaperones and proteins involved in posttranslational modification, providing an explanation for the functional shift in the O category.

Another *E. coli* proteome with an unusual pattern of functional divergence is O17:K52:H18 (strain UMN026), a multi-drug-resistant strain that causes urinary tract infections (Manges et al., 2001). Unique among *E. coli* and *Shigella* species, this strain is enriched for functional divergence among genes involved in secretion (U). Investigation of the genes underlying this enrichment revealed functional divergence in the VirB8 and VirB9 genes, which encode core proteins in a Type IV secretion system found only in two *E. coli* strains–UMN026 and 018 (ED1a), although the latter species is not enriched in this category. In other bacteria, Type IV systems are involved in the exchange of DNA with the environment, as well as the delivery of effector proteins to host cells (Cascales & Christie, 2003). Since these two proteins are important components of the Type IV secretion systems of other bacteria, functional divergence in these genes may be involved in adapting the system to an UMN026-specific role (see Figure 2.5). To gain further insight into the possible implications of the UMN026-

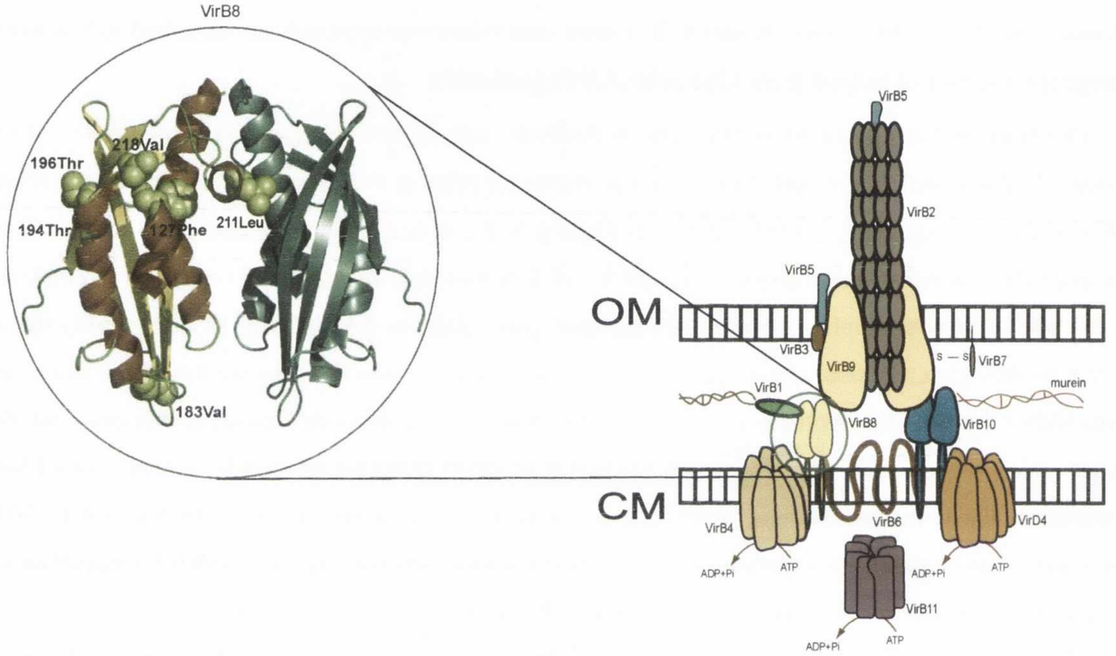


Figure 2.5: Amino acid residues under functional divergence in *E. coli UMN026* VirB8. Right: the structure of a Type IV secretion system found only in two strains of *E. coli*. CM = cytoplasmic membrane, OM = outer membrane. The complex structure is based on that of Baron (2006). In UMN026, the central complex proteins VirB8 and VirB9 are under functional divergence. Left: Of the six sites detected by Clusterfunc that could be mapped to the VirB8 crystal structure (Bailey et al, 2006), there is evidence that five are involved in forming the VirB8 homodimer, so that functional divergence at these positions might alter the quaternary structure of the complex.

specific changes in these proteins, we mapped the specific residues under functional divergence in VirB8 (also output by Clusterfunc) onto the *Agrobacterium tumefaciens* crystal structure (Bailey et al., 2006). Of the 15 sites under functional divergence (see Supplementary Table S2.5), 6 could be mapped onto the crystallized region of the protein. Of these 6, 5 are at or close to positions previously shown to be of functional importance. Thr-194 (residues numbered according to the *A. tumefaciens* sequence), detected as being under functional divergence, is 3.8 Angstroms from Thr-192. Mutating Thr-192 to Met results in a variant that is stable, but cannot complement a VirB8 deletion mutant (Kumar & Das, 2001), indicating that the function of the protein is sensitive to changes at this site. In addition, Thr-194, itself moderately conserved among VirB8 homologs, is directly adjacent to two highly conserved sites, Ala-195 and Thr-196. Thr-196, which Clusterfunc also detected as being under functional divergence in UMN026, is directly involved in the stabilization of the VirB8 homodimer (Bailey et al., 2006), as is Leu-211, another functionally divergent site. Finally, two additional sites identified by Clusterfunc are at positions that suggest they may have an indirect role in dimerization. Val-218 is located between two other residues (Leu-217 and Val-219) that are involved in dimer formation, while Phe-127 is adjacent to Ser-128, a conserved residue that stabilizes the interaction surface on VirB8. The function of the other site detected under functional divergence, Val-183, is currently unknown. Taken together, these results indicate that functional divergence in *E. coli UMN026* VirB8 has occurred at residues important in forming the homodimer, which may have important implications for the overall structure and function of the complex. With no crystal structure available for VirB9, it is more difficult to evaluate the functional significance of the sites detected there. Further, we detected functional divergence on branches of the VirB9 tree, suggesting that this protein experiences a more general pattern of radical change.

2.6 Conclusion

The identification of functional divergence and ecological adaptation from sequence data is an interesting and important goal in evolutionary biology, with the potential to deepen our understanding of the evolution of individual traits and species, as well as the processes of evolution as a whole. Bacteria display an astonishing capacity for adaptation to different lifestyles and ecological niches, but investigating the evolution of these traits is problematic because their phylogenetic context is often unclear. Here, we have circumvented this problem by evaluating functional divergence on gene trees and then clustering species by gene functional category. Our approach revealed the overall patterns that have characterized functional change during the evolution of bacteria, including strong constraint on information storage and processing genes and also constraints induced by the host on pathogenic and symbiotic bacteria. It also identified lineage-specific events of atypical functional divergence, such as the use of flagella by *Bartonella bacilliformus* to invade host erythrocytes and residue-level changes in the VirB8 protein of *E. coli UMN026*. This is, to our knowledge, the first method that can be used to identify functional divergence at the level of entire proteomes. Although used here to perform a large-scale analysis on bacteria, our Clusterfunc software is extremely flexible and can be applied to individual genes, complexes, lineages, or groups of proteomes using any ontology system in order to investigate functional divergence at every level of biological organization.

Chapter 3

Two Bacterial SecA/SecY Systems with Distinct Roles in Adaptations

3.1 Related manuscript

Jiang, X.W., & M.A. Fares (2011). Two bacterial SecA/SecY systems with distinct roles in adaptations. In preparation

3.2 Abstract

Bacteria present an astonishing metabolic diversity, which enables them occupy a wide spectrum of unrelated environments. This metabolic capacity is due to the potential of bacteria to interact with a diverse set of environments mediated by a set of bacterial secreted proteins, which are translocated across the cell membrane. For example, translocation of virulence effector proteins is key for the interaction of pathogenic bacteria with their host. The general secretion (Sec) pathway is responsible for the translocation of a substantial fraction of these bacterial proteins from the cytosol to the cell membrane and beyond. Sec pathway is therefore central to bacterial metabolism and ecology. In bacteria, the core of Sec system is formed by a complex of proteins (SecA-SecYEG) key among which are the motor ATPase SecA and the membrane channel SecY. We sought to investigate whether evolutionary differences in proteins from the Sec system could be associated to differences in protein secretion and bacterial lifestyles. To identify changes in the evolutionary constraints among bacterial groups, we conducted an analysis of functional divergence (FD)–the process by which new genes and functions originate through the modification of existing ones–of Sec proteins from different groups of bacteria,

archaea and plant chloroplasts. Importantly, we find the strongest signature of evolutionary change to have occurred in lineages leading to pathogenic bacteria. Amino acid regions affected by FD in SecA and SecY are involved in ATP binding and hydrolysis activity and in the translocation of proteins across the membrane, respectively. Furthermore, functional analysis of secreted proteins from 207 bacterial proteomes suggests differential protein secretion in those bacterial lineages that present strong functional divergence. We show that FD of key Sec components has played a fundamental role in the adaptations of bacteria and archaea and in bacterial pathogenesis in particular.

3.3 Introduction

Bacteria use two major secretion systems to translocate proteins across or into the plasma membrane: the general secretion (Sec) pathway and the Sec-independent Twin-arginine translocation (Tat) pathway. In the Sec pathway proteins are translocated in an unfolded state (preproteins), while in the Tat pathway preproteins can be translocated in a partially or fully folded state. Approximately one quarter to one third of bacterial proteins are initially synthesized in the cytosol and subsequently targeted either to the cell membrane or to the extra-cellular space of the cell (Driessen & Nouwen, 2008). Most of these translocated proteins possess a Sec signal peptide at their amino terminals, which plays a central role in targeting the substrate preprotein to its final destination by the Sec pathway. Due to the fundamental role these substrate proteins play in bacterial environmental adaptation, Sec system is central to bacterial metabolism and ecology (Silhavy et al., 2010).

In bacteria, Sec system has two independent pathways directing synthesized proteins to their final destination: the post-translational and the co-translational pathways. The substrates (preproteins) of both translocation pathways converge at the cytoplasmic membrane, across which they are subsequently translocated by an evolutionarily conserved heterotrimeric protein complex channel: the SecYEG complex. Generally, in the post-translational pathway, the exported preprotein is first synthesized by ribosomes in the cytosol of the bacterium and, then targeted by SecB and/or SecA to the plasma membrane. SecB does not perform this role in all bacteria because this protein is absent in most gram-positive bacteria .

SecB is a secretion-dedicated chaperone that binds to preprotein and prevents it from undergoing premature folding (Bechtluft et al., 2010). SecB interacts with SecA thereby passing its substrate preprotein to SecA. SecA was also shown to have chaperone activity and some evidence suggests that it can complement SecB function (McFarland et al., 1993; Eser & Ehrmann, 2003). The main role of SecA, however, has been shown to push the preprotein across the translocation channel SecY in an ATP-dependent manner (Yuan et al., 2010) using a “clamp” (Zimmer et al., 2008).

In the co-translational pathway of bacteria, SRP (signal recognition particle) binds to the polypeptide emerging from the ribosome and then directs the ribosome-nascent chain complex to the plasma membrane, where the complex binds to the SRP receptor FtsY. Subsequent to this, the emerging polypeptide passes through the SecY translocation channel (Yuan et al., 2010). Both pathways exist in archaea. However, archaea lack chaperone SecB and ATPase SecA. So it is not clear how post-translational translocation is achieved in archaea (Irihimovitch & Eichler, 2003).

Of all the proteins forming the SecYEG-SecA complex, SecA and SecY have been extensively studies because they are directly responsible for the translocation of a large fraction of preproteins through the membrane (Bieker et al., 1990; Bost & Belin, 1997). SecA has been proposed to mediate the translocation of preproteins by coordinating ATP (adenosine triphosphate) binding and hydrolysis (Rapoport, 2007). This protein is conserved among bacteria and some plants (possibly located on the thylakoid membrane of chloroplast (Aldridge et al., 2009)), and plays a critical role in the protein export pathway, being therefore essential to cell viability (Bieker et al., 1990). SecA is about 100kDa and consists of six protein domains (Figure 3.1), namely the nucleotide-binding domain 1 (NBD1) and 2 (NBD2), the polypeptide-cross-linking domain (PPXD), the helical scaffold domain (HSD), the helical wing domain (HWD) and C-terminal domain (CTL, it is partially shown in Figure 3.2A and colored white because the structure CTL domain hasn't been completely solved.). The conformational changes that SecA requires to translocate preproteins are suggested to be modulated by ATP binding and hydrolysis in the NBD domain (Hunt et al., 2002), whereas PPXD, NBD2 and part of HSD are thought to form a "clamp" (Figure 3.2C, G and H) for preprotein peptide binding and translocation (Zimmer et al., 2008). Interestingly, SecA dramatically increases its ATPase activity only when it binds to SecY (Robson et al., 2009).

SecY presents a clam shell-like symmetrical arrangement and consists of ten transmembrane (TM) segments (TM1-TM10) (Figure 3.3B), which form a gate at the front side and are clamped together at the back by SecE (Zimmer et al., 2008). The role of SecG has been deemed not essential (Smith et al., 2005), although early studies demonstrated that it may be involved in mediating the interaction of SecA and SecY translocon and may have a role in the SecA-dependent secretion in some bacteria (Hanada et al., 1994; Douville et al., 1995; Nishiyama et al., 1996; Sugai et al., 2007).

Two SecY/SecA systems are found in bacteria, one is termed SecA1/SecY1 (canonical SecA/SecY, interchangeably used in this paper), which has a higher degree of sequence similarity to that of *Escherichia coli* and *Bacillus subtilis*, and the other is called (accessory) SecA2/SecY2, which has a lower degree of sequence similarity to that of *E. coli* and *B. subtilis* (Rigel & Braunstein, 2008). SecA1 and/or SecY1 has been shown to perform "housekeeping" functions in bacterial physiology, whereas SecA2 and/or SecY2 is indispensable in many cases for species specific virulence traits (Rigel & Braunstein, 2008). Some gram-positive bacteria and mycobacteria are known to have an accessory SecA2 system, such as *Staphylococcus aureus*, *Strepto-*

coccus gordonii and *Mycobacterium tuberculosis*. How does the accessory SecA2/SecY2 system contribute to the overall bacterial physiology, transcription and secretion? This fundamental problem has only been recently addressed by Ian Siboo and colleagues (Siboo et al., 2008) and they showed that the accessory Sec system in *Staphylococcus aureus* might solely secrete one serine-rich glycoprotein SraP (serine-rich adhesin for platelets). Disruption of SecA2, SecY2 and other members of the accessory Sec system resulted in almost complete abolishment of cell surface expression of SraP. They also showed that disruption of the accessory Sec system had no effect on the mutants' growth in vitro compared with the parent strains or on the canonical Sec system. Moreover, by comparing the mutants and parent strains using microarray, SDS-PAGE and 2D-DIGE experiments, these authors observed that the disruption resulted in no perturbation of the major regulatory systems, no significant changes of bacterial physiological state, and no stress response.

Bacteria that possess an accessory SecA2/SecY2 system include highly pathogenic and antibiotic resistant bacteria such as *Staphylococcus aureus* (meticillin-resistant, vancomycin-susceptible, vancomycin-intermediate resistance strains) and *Streptococcus pneumoniae* etc. Although not essential for pathogens survival (Bensing & Sullam, 2002; Rigel & Braunstein, 2008) this system has been found to be solely responsible for secreting a set of virulence factors (Rigel & Braunstein, 2008). It has been shown that SecA2/SecY2 system is involved in the secretion of posttranslationally modified glycoproteins in pathogenic *Streptococcus* and *Staphylococcus* species (Chen et al., 2004; Takamatsu et al., 2004a,b; Bensing et al., 2005; Siboo et al., 2005; Takamatsu et al., 2005; Wu et al., 2007; Siboo et al., 2008; Mistou et al., 2009; Zhou et al., 2010). Given that these heavily glycosylated preproteins are exceptionally long (with thousands of amino acid residues) we hypothesise that in these pathogens SecA2 and SecY2 have undergone functional divergence in comparison with the canonical proteins that enabled two important changes: greater conformational changes in SecA2 through an optimization of the regulation of ATP binding and hydrolysis that would allow interacting with bigger proteins and interaction with SecY2 channel (Chen et al., 2004; Takamatsu et al., 2004a,b; Bensing et al., 2005; Siboo et al., 2005; Takamatsu et al., 2005; Wu et al., 2007; Siboo et al., 2008; Mistou et al., 2009; Zhou et al., 2010), while SecY2 channel had to cope with the translocation of preproteins with heavily glycosylated amino acid residues.

All these evidence point to that the two SecA/SecY systems have distinct functional roles in different microorganisms, hinting speculation that these proteins may have undergone dramatic functional shifts to enable microorganisms adapt to the various ecological requirements. These functional shifts and the consequences associated to the molecular changes along the evolutionary history of bacterial natural diversification remain largely elusive .

Previous studies support the functional divergence in molecules from the secretion systems (Gevers et al., 2004; Sanchez-Perez et al., 2008; Das & Oliver, 2011). To identify molecular changes in Sec proteins and understand how these changes relate to bacterial lifestyle, we conducted an analysis of the molecular changes to explain the divergence of functions of key Sec proteins across the bacterial phylogeny. We performed a thorough functional analysis of all secreted proteins within those bacterial clades with strong evidence for functional divergence. Our results suggest that strong functional divergence in bacterial Sec secretion system could indeed be associate with their various ecological adaptations.

3.4 Materials and Methods

3.4.1 Sequence selection and alignments

For functional divergence analysis, homologous protein sequences for SecA, SecB, SecE, SecG and SecY were retrieved from KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database (updated December 1, 2009) using KEGG API (Application Programming Interface) function with Perl programming (Kanehisa et al., 2008). These homologous protein sequences come from 944 bacteria, 68 archaea and 10 plants (Supplementary Table S3.1, S3.2, S3.3, S3.4 and S3.5). Pathogenic bacterial strains and their hosts are also identified from KEGG (`http://www.genome.jp/files/org2key.xl`).

To identify protein sequences of SecA1/SecY1 and SecA2/SecY2, we followed a strategy used by a previous study (Rigel & Braunstein, 2008). Briefly, in bacteria with two SecA/SecY homologs, SecA/SecY with a slightly higher sequence similarity to the canonical SecA/SecY of *E. coli* and *B. subtilis* are termed SecA1/SecY1, while the other SecA/SecY with a lower sequence similarity to SecA and SecY of E. coli and B. subtilis are termed SecA2/SecY2. This strategy was used to identify SecA1/SecY1 and SecA2 And SecY2 in functional divergence analysis and intermolecular coevolutionary analyses.

To analyse intermolecular coevolution of the SecA2 and SecY2 of *Staphylococcus* and *Streptococcus* species, protein sequences for SecA2/SecY2 were retrieved from KEGG database by FASTA search service following the aforementioned strategy, where 15 bacterial strains were found to have both SecA2 and SecY2. We then retrieved the 15 protein sequences for SecA2 and SecY2, respectively. Similarly, to analyse intermolecular coevolution between SecA and SecY of bacteria that have SecB, we retrieved 91 protein sequences, respectively. To analyse intermolecular coevolution between SecA1 and SecB, we retrieved 83 protein sequences, respectively. All protein sequences were aligned by Muscle 3.7 (Edgar, 2004). Alignments were manually checked using Jalview (Clamp et al., 2004).

3.4.2 Phylogenetic reconstruction

RaxML Pthreads version 7.2.4 (Stamatakis, 2006) was used to build the Maximum likelihood (ML) phylogenetic trees after estimation of the appropriate amino acid substitution model by Prottest version 2.4 (Abascal et al., 2005). Then, the best ML trees for SecA, SecB, SecE, SecG and SecY were computed based on the LG substitution model (Le & Gascuel, 2008) with GAMMA model of rate heterogeneity among sites and empirical amino acid frequencies.

3.4.3 Identifying functional divergence

Functional divergence is a term used to refer to the divergence in protein function in a group of organisms as a result of mutations at particular amino acid sites. This functional divergence can take place at different degrees, going from slight changes of the ancestral function to complete acquisition of novel functions. Functional divergence can be classified in two groups (Gu, 1999, 2006) i) FD type I: refers to the acquisition of a functional role of an amino acid site in a group of bacteria that are phylogenetically close and therefore that amino acid site will be highly conserved within this group but variable among groups. This site will evolve neutrally outside the group; and ii) FD type II in which the amino acid site is highly conserved in two different groups of bacteria but have different functional roles. Details of the method to identify functional divergence and its application have been recently published in two case studies (Toft et al., 2009; Williams et al., 2010). Briefly, the method uses a protein sequence alignment and a phylogenetic tree including paralogs and orthologs. The method then compares a pair of clades in the tree sharing a common ancestral origin to their closest phylogenetic outgroup. The comparison is performed for each amino acid site in the protein and the strength or likelihood of the amino acid state is evaluated using the appropriate BLOSUM matrix (BLOcks of Amino Acid SUbstitution Matrix) (Henikoff & Henikoff, 1992). BLOSUM matrices comprise the scores for the transition between the 20 amino acids, with positive scores meaning the transition is more frequent than expected, negative means that the transition is less frequent than expected and 0 means that these transitions are as frequent as expected. Using BLOSUM as a score matrix for the transitions between amino acids allows scoring both variable sites and conserved sites: one clade may seem variable yet the amino acid transitions between orthologous sequences within the clade may have taken place between biochemically close amino acids (positive scores). In the comparison between two clades to an outgroup (clade 1 and clade 2), functional divergence in clade 1 would be detected if the BLOSUM scores were positive within clade 1, negative between clade 1 and the outgroup and positive between clade 2 and the same outgroup. To test the significance of the variances of these score , we calculated the mean and standard error of the scores for clade 1 ($\overline{C}_1$ and SE_1) and clade 2 ($\overline{C}_2$ and SE_2) (Toft et al., 2009). The significance of the difference in the means was obtained by calculating the Z-score for the comparison of samples with unequal sizes as:

$$Z = \frac{\overline{C}_1 - \overline{C}_2}{SE_{C_{1,2}}}$$

Where $\overline{C}_{1,2}$ are the mean substitution scores for the transition from clades on either side of the bifurcation in the phylogenetic tree relative to the outgroup and $SE_{C_{1,2}}$ is the combined standard error of clade 1 and clade 2.

In order to identify the patterns of sites under FD among protein domains in SecA/SecYEG proteins, we first ranked each of the clades according to the number of sites under FD in comparison with the rest of clades. We then counted the number of sites under FD within each protein domain for that particular protein. Third, we used these numbers to build a data matrix for each protein, with each row representing the FD enrichment profile for the top 10% clades most enriched for FD, while columns referred to the FD profiles within each of the protein domains. Finally, we used this data matrix as input for the clustering of the rows and columns in order to search for similarities in the FD profiles of protein domains and/or clades, which was performed using heatmap.2 function in R (`http://www.r-project.org/`). The heatmap.2 function scales each element in a row in the data matrix based on a normalised Z-score, which is calculated as follows:

$$z_{ij} = \frac{x_{ij} - \overline{x}_i}{\sigma_i}$$

Here, z_{ij} is the Z-score calculated for element x_{ij} in the matrix, $\overline{x}_i$ is the mean of the row i , σ_i is the standard deviation of all the elements in the row i .

3.4.4 Prediction of potential Sec and Tat substrate proteins

To identify potential SecA1/SecY1 dependent substrate proteins, we follow three steps. First, we use SignalP (Bendtsen et al., 2004) to screen a proteome for potential Sec-dependent secreted proteins (secretome). Second, we use TatP (Bendtsen et al., 2005a) to screen that proteome for potential Tat-dependent secreted proteins. Finally, we remove those protein predicted by TatP in the SignalP results to minimize false positives. To identify potential SecA2/SecY2 dependent substrate proteins, the reciprocal best-best FASTA hits from sequenced *Staphylococcus* and *Streptococcus* strains are retrieved from KEGG database using the SraP protein, an experimentally verified SecA2-dependent substrate (and it is the only protein that was found to be SecA2 dependent), of *Staphylococcus aureus N315* (Siboo et al., 2008). We then use SignalP (Bendtsen et al., 2004) to predict if these proteins have potential signal peptides. These proteins will be further analysed by SecretomeP if they don't have any signal peptides. SecretomeP can predict if these proteins are secreted even if they don't have an obvious signal peptide. The final set that is predicted to be non-secretory was analysed further by TatP (Bendtsen et al., 2005b), which can predict if these proteins can be secreted through the Twin-arginine secretion pathway. We also analyze the original dataset by TatP to see if they have conflicting results predicted by SignalP. In all cases, the length of truncated sequences for prediction is set to 200, as indicated by functional studies of GspB protein of *Streptococcus gordonii* (Bensing et al., 2007).

Predicted Sec and Tat substrates were grouped according to their COG functional categories. To determine if the numbers of substrates for the Tat-dependent secretion in each of the categories were significantly higher (enriched category for secreted substrates) or lower (impoverished category) than expected by chance, a χ^2 test was performed on each COG functional category with one degree of freedom. The statistical test was performed according to the following equation:

$$\chi_i^2 = \frac{(O_i - E_j)^2}{E_i}$$

$$E_i = n_i \frac{N_{predicted}}{N_{COG}}$$

O_i is the observed number of predicted substrate proteins in COG category i. while E_i is the expected number of such proteins in that category i. The total number of proteins in category i is indicated by n_i. $N_{predicted}$ is the total number of predicted substrate proteins, while N_{COG} is the total number of proteins in the proteome. Only proteins assigned within single COG functional categories were analysed (ambiguous categories R and S were not included).

3.4.5 Intermolecular coevolutionary analysis

Intermolecular (protein-protein) coevolution analyses were performed using the software CAPS version 1.0 (Fares & McNally, 2006) based on a published method (Fares & Travers, 2006). CAPS has been compared with other methods in a previous study to identify coevolution between membrane proteins and was placed on the top ranking, indicating the validity and power of this method to identify true coevolution cases under stringent statistical conditions (Fuchs et al., 2007). CAPS algorithm uses a normalization strategy for correcting heterogeneity in the data, which uses Poisson distance between protein sequences as a correction parameter. This correction method is reasonable and fast, and performs well when the analyzed data are polymorphic protein sequences (allelic proteins) from the same species (e.g. human immunodeficiency virus env gene protein products gp120 and gp41 (Travers et al., 2007), human leukocyte antigen molecules (Jiang & Fares, 2010)). The working flow chart of this method has been discussed (Travers & Fares, 2007). CAPS can classify pairs of coevolving amino acid sites that share the same site into groups. Therefore, a group of coevolving sites for protein-protein coevolution analyses means that all the amino acid sites in one protein coevolve with all sites in the other protein and vice versa. These coevolving sites may imply allosteric long distance interactions as well as close interactions between coevolving amino acid sites from two interacting proteins. In this way, we can identify structurally and/or functionally coevolving sites and link these sites to relevant functional or structural roles provided that structural and functional data are available the (Travers & Fares, 2007). We set the alpha-value to 0.05 when performing the statistical test, and the random sampling size to 100 million. To ensure convergence in the results (e.g. that the same amino acid sites are consistently identified as coevolving) the analysis were performed three times with the same parameters, and all three results were consistent.

3.4.6 Structural analysis of SecYEG translocon and SecA ATPase

To define protein domains of SecA, E.coli SecA domain regions (Sharma et al., 2003; Driessen & Nouwen, 2008) were mapped to SecA alignment. Domain information obtained here was also used for other purposes, such as mapping sites under FD and coevolving sites to protein domains. To define protein domains of SecY, the transmembrane domain (T1-T10) and cytoplasmic domains (C1-C6) and periplasmic domains (P1-P5) were obtained using SMART (Simple Modular Architecture Research Tool) service (Schultz et al., 1998; Letunic et al., 2009) with *E.coli* SecY protein sequence.

To understand structural and functional interactions between amino acid sites under FD and coevolution, protein X-ray crystallographic structures for SecA, SecYEG-SecA complex were retrieved from Protein Data Bank (PDB, `http://www.pdb.org`) using keyword SecA. Only structures which were solved by X-ray diffraction method and had complete NBD1, NBD2, PPXD, HWD and HSD domains were kept for further analysis. The PDB codes of these structures are 1M6N (Hunt et al., 2002), 1NKT (Sharma et al., 2003), 1TF2 (Osborne et al., 2004), 2IBM (Zimmer et al., 2006), 2IPC (Vassylyev et al., 2006), 3JUX (Zimmer & Rapoport, 2009), 3JV2 (Zimmer & Rapoport, 2009) and 3DIN (Zimmer et al., 2008). SecA from SecYEG-SecA (3DIN, Figure 3.2H) was used as the comparison base when performing the structure based alignments through CEalign in PyMOL (Figure 3.2A-H). PyMOL Version 1.3 (`http://pymol.sourceforge.net/`) was used to visualize X-ray crystallographic structures, and the structure alignment was performed using CEalign plug-in (`http://pymolwiki.org/index.php/Cealign`) in PyMOL (Shindyalov & Bourne, 1998). The distance in the protein structure between two amino acids under FD was calculated by their shortest Euclidean atomic distance. For example, the distance between the two atoms with three-dimensional coordinates (x^A, y^A, z^A) and (x^B, y^B, z^B) from amino acid "A" and amino acid "B" respectively in the X-ray crystallographic structure was calculated as:

$$d = \sqrt{(x^A - x^B)^2 + (y^A - y^B)^2 + (z^A - z^B)^2}$$

We did the same for every pair of atoms in both proteins, and then we took the shortest atomic distance as the amino acid distance between both of the amino acids. We followed the same procedure to calculate the shortest amino acid distances between domains and the distance between drug target amino acid sites identified

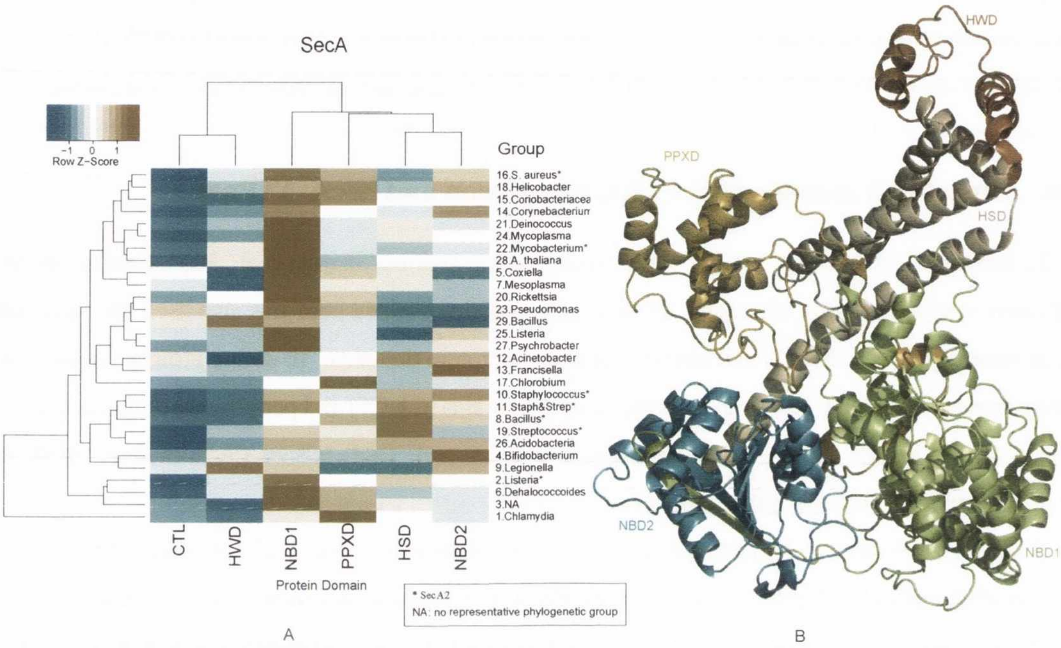


Figure 3.1: Two-way clustering of the top ten percent branches under functional divergence of SecA. (A) Protein domains are clustered according to the number of sites under functional divergence after Z-score standardisation, group number represents the rank of number of sites under functional divergence in each analysed branch, and groups are also clustered. (B) Protein domains are shown in SecA. * denotes functional divergence results for SecA2, representative phylogenetic group in each group is also shown.

in (Chen et al., 2010) and amino acid sites under FD identified in this study.

3.5 Results

3.5.1 Evidence of functional divergence of the motor and translocon pore of pathogenic bacteria

In this study we identified 298, 87, 191, 199 and 268 clades that have at least one amino acid site under FD (Supplementary Table S3.1, S3.2, S3.3, S3.4 and S3.5 for SecA, SecB, SecE, SecG and SecY, respectively). In our results, we found 271 (35% of all species under FD), 89 (31% of all species under FD), 189 (29% of all species under FD), 199 (30% of all species under FD), 219 (29% of all species under FD) pathogenic bacterial

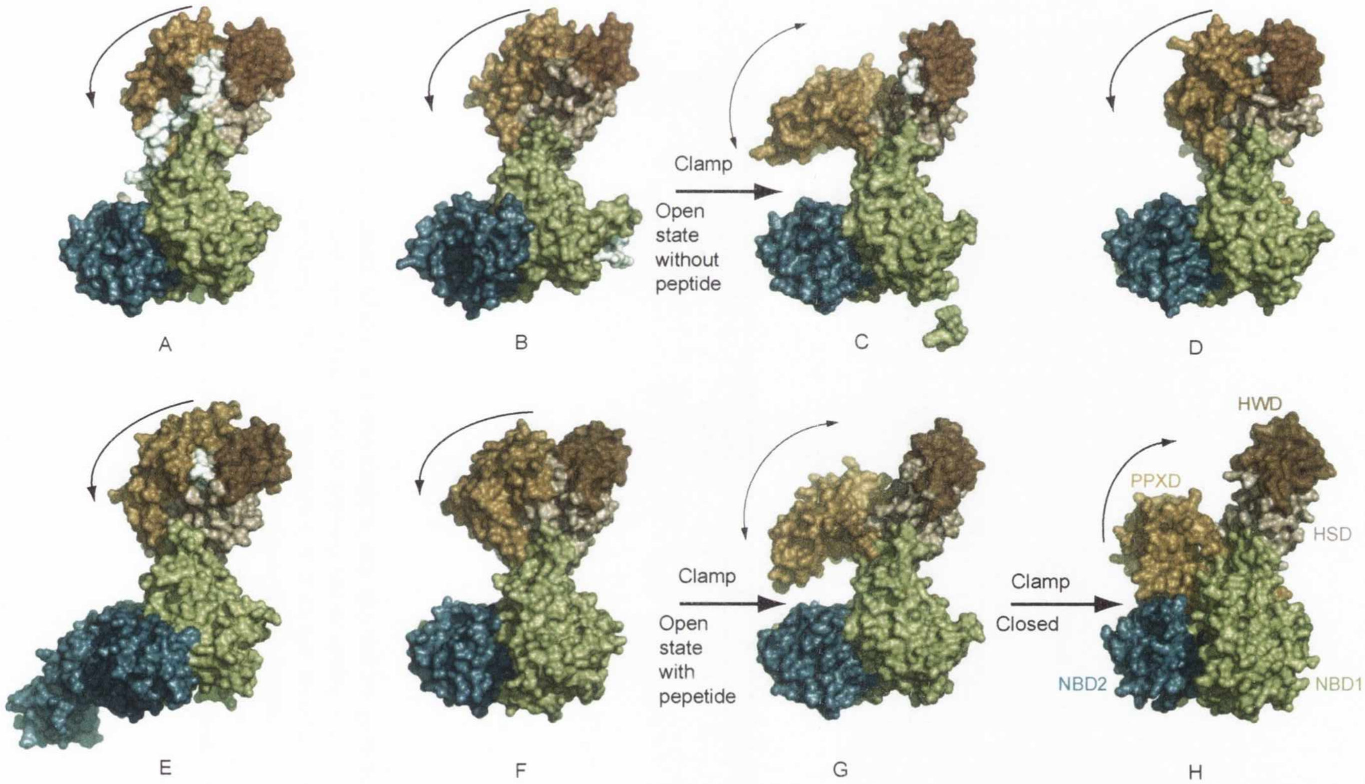


Figure 3.2: Structural alignment of SecA proteins. Solved SecA crystal structures are aligned based on (H) (Zimmer et al., 2008) for (A) 1M6N (Hunt et al., 2002), (B) 1NKT (Sharma et al., 2003), (C) 1TF2 (Osborne et al., 2004), (D) 2IBM (Zimmer et al., 2006), (E) 2IPC (Vassylyev et al., 2006), (F) 3JUX (Zimmer & Rapoport, 2009), (G) 3JV2 (Zimmer & Rapoport, 2009). The possible trend of domain movement of PPXD is denoted by arch lines with arrow.

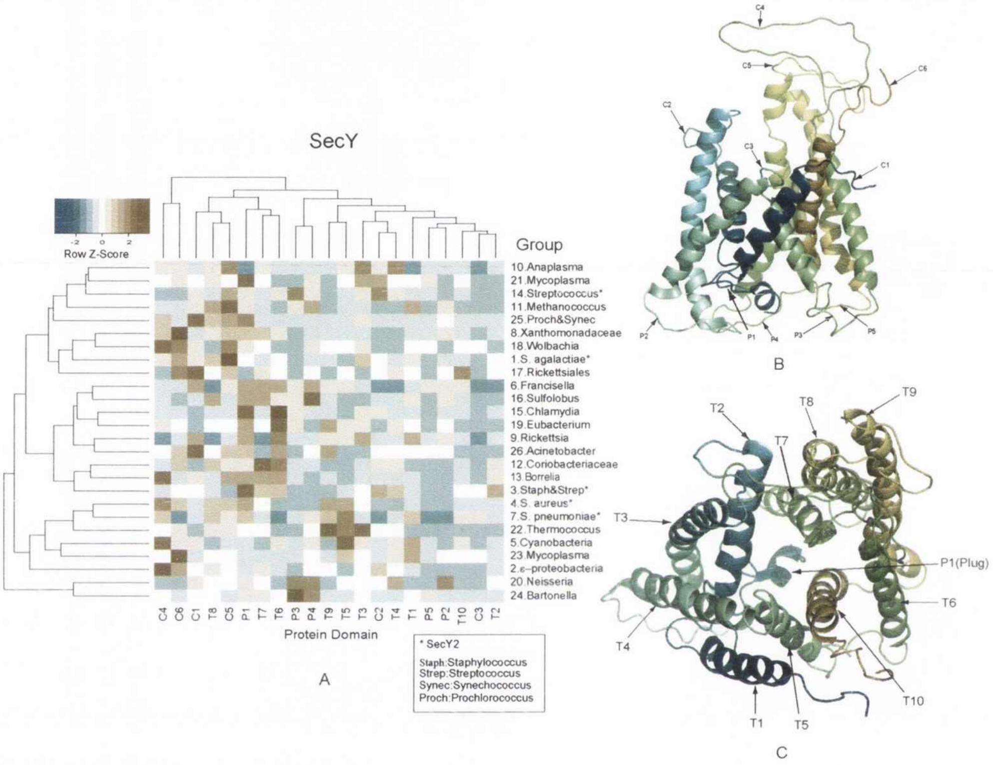


Figure 3.3: Two-way clustering of the top ten percent branches under functional divergence of SecY. (A) Protein domains are clustered according to the number of sites under functional divergence after Z-score standardisation, group number represents the rank of number of sites under functional divergence in each analysed branch, and groups are also clustered. (B) Cytoplasmic (C1-C6) and periplasmic (P1-P5) domains are shown in SecY. (C) Transmembrane (T1-T10) domains are shown in SecY. The so-called "plug" domain is also shown (P1) in the centre. * denotes functional divergence results for SecY2, representative phylogenetic group in each group is also shown.

strains to have at least one amino acid site under FD for SecA, SecB, SecE, SecG and SecY, respectively. Moreover, all pathogenic bacterial strains in the original data were identified to have at least one site under FD in the results. We focused on those top 10% clades with the strongest signal of FD (clades are ranked by the number of sites within the proteins identified to be under FD). Interestingly, we found that there are clinically important bacterial pathogens in the top 5 most functionally divergent clades (Supplementary Table S3.1, S3.2, S3.3, S3.4 and S3.5), such as *Chlamydia*, *Listeria*, *Staphylococcus* and *Streptococcus*.

We asked how protein domains are affected by functional divergence within SecA/SecY across large phylogeny. To identify sites under FD in protein domains, we followed the steps desribed in Materials and Methods and visualised the results in the heatmaps shown in Figure 3.1A and Figure 3.3A. As shown in Figure 3.1A, SecA, it is striking that the most affected domain by FD is NBD1 domain, which is consistent with our previous hypothesis that claimed that functional divergence of ATP binding domain may play a significant role in SecA evolution. However, NBD1 and PPXD domains present similar heatmap patterns for FD compared to other domains. HSD and NBD2 domains of SecA were clustered within the same group due to their similarities in site numbers under FD, as did CTL and HWD domains, which were the two domains most impoverished for FD (Figure 3.1A), indicating that their functions have been conserved during SecA evolution. Taking all clades into account instead of the 10% most enriched ones made no difference to our results (Supplementary Figure S3.1). NBD2 and HSD domains shared similar profiles of FD, probably due to the possible modulation of HSD domain function or structure by NBD2. A recent study showed that PPXD, NBD1 and HSD also interact with signal peptide of the preprotein during the translocation process (Auclair et al., 2010). As shown in Table 3.1, amino acid sites under FD between the PPXD domains and the remaining domains (NBD2, NBD1 and HWD) deviates the most in terms of mean atomic distance, with the distance to the sites from the NBD2 being the largest. The conformational changes between PPXD and other protein domains in SecA here are consistent with previous in vivo experiments (Keramisanou et al., 2006; Erlandson et al., 2008b; Zimmer & Rapoport, 2009).

In SecY (SecY1 and SecY2) we observed a strong and similar enrichment of domains C4 and C6 for FD, being consequently joined in the same cluster (Figure 3.3A), which is in support of previous studies highlighting their equally important functional role (Baba et al., 1994; Smith et al., 2005). Like in SecA, SecY showed similar clustering patterns when the full set of bacterial lineages was considered (Supplementary Figure S3.2). Domains C1, T8, C5, P1, T7 and T6 form the second cluster, and P3, P4, T9, T5, T3, C2, T4, T1, P5, P2, T10, C3 and T2 form the third cluster. These clustering may indicate possible functional dependencies during the evolution of SecY in these bacterial lineages. For example, studies have shown that T8, T7 and T6 play roles in the plug domain P1 displacement for subsequent peptide translocation upon ribosome (Gumbart et al., 2009) or SecA (Zimmer et al., 2008) binding to C1 (C-terminal tail) and C5 domains of SecY. Next, we aimed

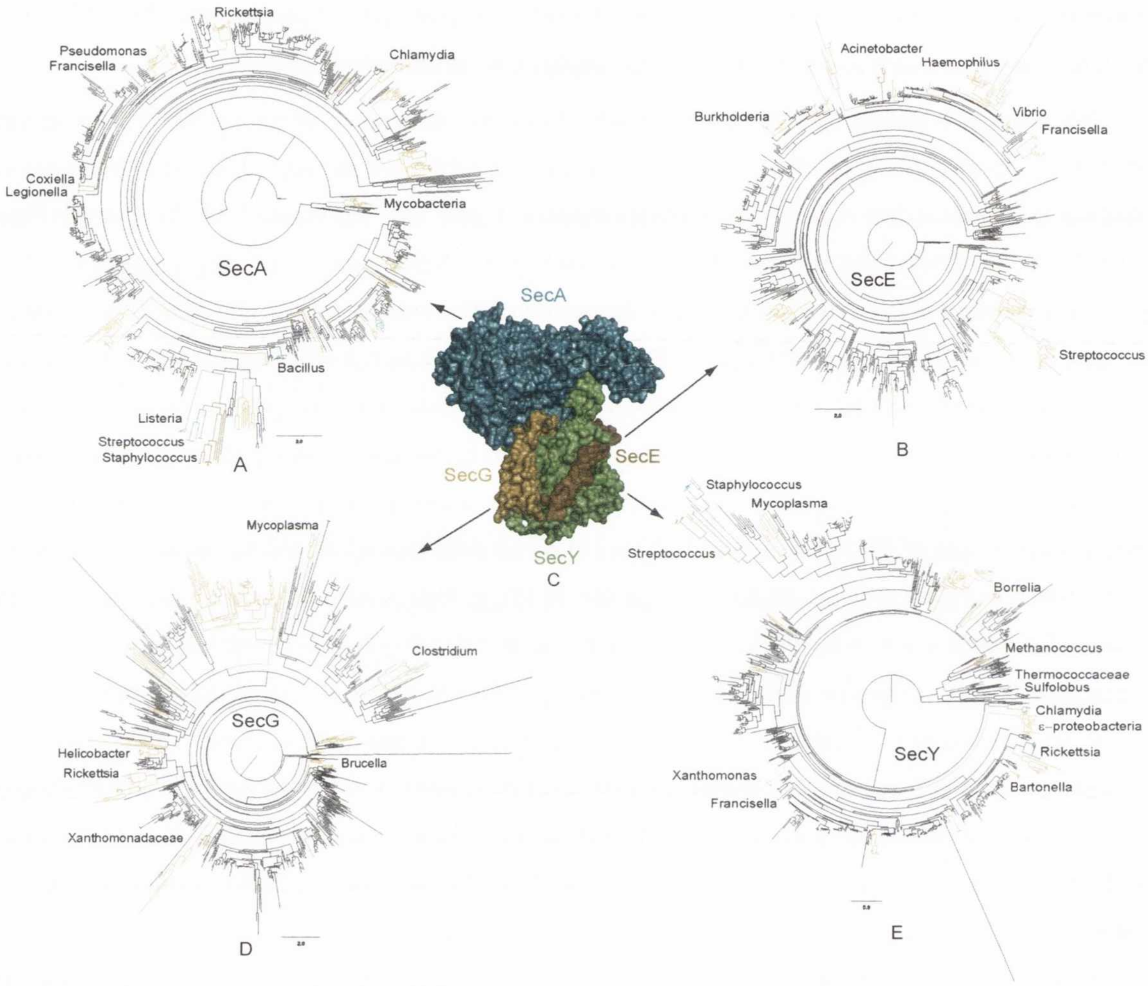


Figure 3.4: Pathogenic bacteria and archaea that have strong functional divergence in their Sec components. Protein components SecA (A), SecE (B), SecG (D) and SecY (E) under functional divergence of the SecA-SecYEG complex (C) and their corresponding phylogenetic trees, where the top 10% most functionally divergent branches (see Supplementary Table S3.1, Table S3.3, Table S3.4 and Table S3.5 for details) are color coded on their corresponding phylogenetic trees. Moreover, clinically interesting pathogenic bacteria are labelled.

		Mean Amino Acid Distance (*Å*)			
Domain1	**Domain2**	**IM6N (Figure 3.2A)**	**1TF2 (Figure 3.2C)**	**3JV2 (Figure 3.2G)**	**3DIN (Figure 3.2H)**
NBD2	NBD1	37.78	37.21	37.1	37.19
NBD2	HWD	75.32	74.52	74.25	74.22
NBD2	PPXD	54.99	42.21	38.98	31.18
NBD2	HSD	48.82	47.94	48.08	46.3
NBD1	HWD	56.3	57.5	56.8	63.69
NBD1	PPXD	49.11	46.11	44.44	40.44
NBD1	HSD	41.11	41.58	41.6	44.48
HWD	PPXD	29.66	43.96	45.35	47.87
HWD	HSD	28.97	29.73	29.35	31.42
PPXD	HSD	25.42	29.3	28.8	29.17

Table 3.1: Mean domain distance of amino acid sites under functional divergence. The mean distances between Amino acid sites under functional divergence from different protein domains (NBD1, PPXD, NBD2, HSD and HWD) of SecA are calculated. PDB IDs of SecA structures are IM6N, 1TF2, 3JV2 and 3DIN, and corresponding SecA structures are shown in Figure 3.2.

at identifying unique patterns pertaining to the FD of canonical and accessory Sec systems.

3.5.2 Functional divergence of SecA1 and SecY1

What are the FD pattern profiles on the canonical SecA/SecY? We found that not two rows in the heat map of Figures 3.1A and 3.3A showed the same FD profile for SecA and SecY protein domains. Nonetheless, the clustering of non-sister bacterial lineages within the same group according to the similarity in their FD profile hints the possibility that they present similar ecological traits. For instance, the pathogenic *Chlamydia* form a monophyletic group (group 1, Figure 3.1A and 3.4E) that showed the strongest signal of FD. This group also presented a unique FD profile for the different protein domains with the greatest proportion of sites under FD being located in the PPXD domain, a slightly lower proportion of sites located at the NBD1 domain, a medium proportion of sites located at the HSD domain and a low proportion of sites located at the NBD2 domain. The fifth most functionally divergent group comprises five strains of *Coxiella burnetii*, which is clustered together with group seven (group five and seven for SecA, Figure 3.1A) including *Mesoplasma florum*, *Mycoplasma capricolum* and *Mycoplasma mycoides*.

ε–proteobacteria presents a particular functional divergence profile in the translocon core protein SecY. We identified 40 sites under FD between ε- (group 2 in Figure 3.3A) and other proteobacteria groups (α, β, γ, δ), 11 sites between α-, β-, γ- (group 102 in Supplementary Table S3.5) and δ-proteobacteria groups, eight sites between α- (group 101 in Supplementary Table S3.5) and β-, γ-proteobacteria groups, and six sites between β- (group 185 in Supplementary Table S3.5) and γ- proteobacteria group (Figure 3.5). Sequenced bacteria from Epsilonproteobacteria (ε) represent the second most functionally divergent group for SecY protein compared with other major sequenced proteobacteria (α, β, γ, δ) groups. As shown in Figure 3.3A, Epsilonproteobacteria (group 2) showed a distinctive molecular pattern of domain FD compared with other branches, where group 2 did not cluster with any of the other groups (Figure 3.3A). To understand the relative functional importance of this pattern, we mapped sites under FD of Epsilonproteobacteria to the C4 loop formed between T7 and T8 (Figure 3.6). The tip of C4 domain is embedded in the "clamp" formed by PPXD, NBD2, NBD1 and part of the HSD (Figure 3.6A). 250Gly, 243Gln and 248Val have been identified under FD in this bacterial group and due to their position in C4 (Figure 3.6B) they are likely to interact with the preproteins and participate in their translocation. Also, residues 255Gln, 256Gly and 257Ala of SecY interact with PPXD domain (Figure 3.6C and 3.6D) and may therefore contribute to the movement of C4 loop.

Strains	**Gene Locus**	**Cleavage Site**	**Prediction Method**
Staphylococcus aureus N315	SA2447 (SraP)	90 and 91	SignalP
Staphylococcus aureus COL	SACOL2676	90 and 91	SignalP
Staphylococcus aureus ED98	SAAV_2725	90 and 91	SignalP
Staphylococcus aureus JH1	SaurJH1_2734	90 and 91	SignalP
Staphylococcus aureus JH9	SaurJH9_2678	90 and 91	SignalP
Staphylococcus aureus MSSA476	SAS2540	NA	SecretomeP
Staphylococcus aureus Mu3	SAHV_2638	90 and 91	SignalP
Staphylococcus aureus Mu50	SAV2654	90 and 91	SignalP
Staphylococcus aureus MW2	MW2575	NA	SecretomeP
Staphylococcus aureus NCTC8325	SAOUHSC_02990	90 and 91	SignalP
Staphylococcus aureus Newman	NWMN_2553	90 and 91	SignalP
Staphylococcus aureus USA300	SAUSA300_2589	90 and 91	SignalP
Staphylococcus aureus USA300 TCH1516	USA300HOU_2654	90 and 91	SignalP
Staphylococcus carnosus	Sca_2202	32 and 33	SignalP
Staphylococcus epidermidis ATCC 12228	SE2249	NA	SecretomeP
Staphylococcus epidermidis RP62A	SERP2281	NA	SecretomeP
Staphylococcus haemolyticus	SH0326	92 and 93 (17 and 18)	SignalP (TatP)
Staphylococcus saprophyticus	SSP0135	50 and 51	SignalP
Streptococcus agalactiae A909	SAK_1493	50 and 51	SignalP
Streptococcus agalactiae NEM316	gbs1529	NA	SecretomeP
Streptococcus gordonii	SGO_0966*	90 and 91	SecretomeP
Streptococcus pneumoniae 70585	SP70585_1816*	NA	SecretomeP
Streptococcus pneumoniae ATCC 700669	SPN23F_17820*	NA	SecretomeP
Streptococcus pneumoniae CGSP14	SPCG_1750*	NA	SecretomeP
Streptococcus pneumoniae Hungary19A 6	SPH_1885*	NA	SecretomeP
Streptococcus pneumoniae TIGR4	SP_1772	42 and 43	SignalP
Streptococcus sanguinis	SSA_0829*	NA	SecretomeP

Table 3.2: Orthologs of *Staphylococcus aureus N315* SraP from *Staphylococcus* and *Streptococcus* species. *: Predicted to be non-secretory by SignalP, SecretomeP and TatP. NA:No cleavage sites available.

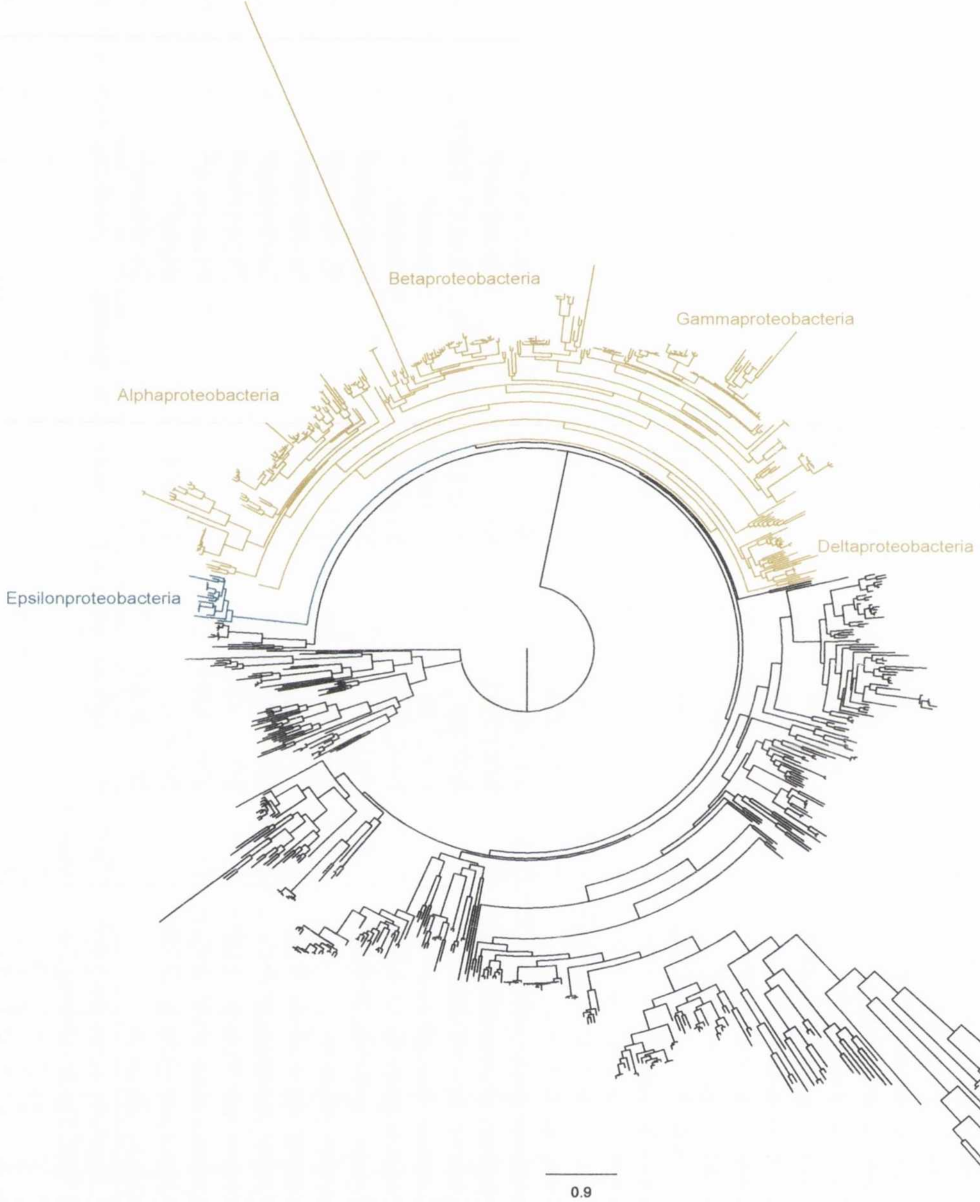


Figure 3.5: Maximum likelihood phylogenetic tree of SecY. Proteobacteria groups are mapped to tree, which are epsilonproteobacteria, alphaproteobacteria, betaproteobacteriam gammaproteobacteria and deltaproteobacteria.

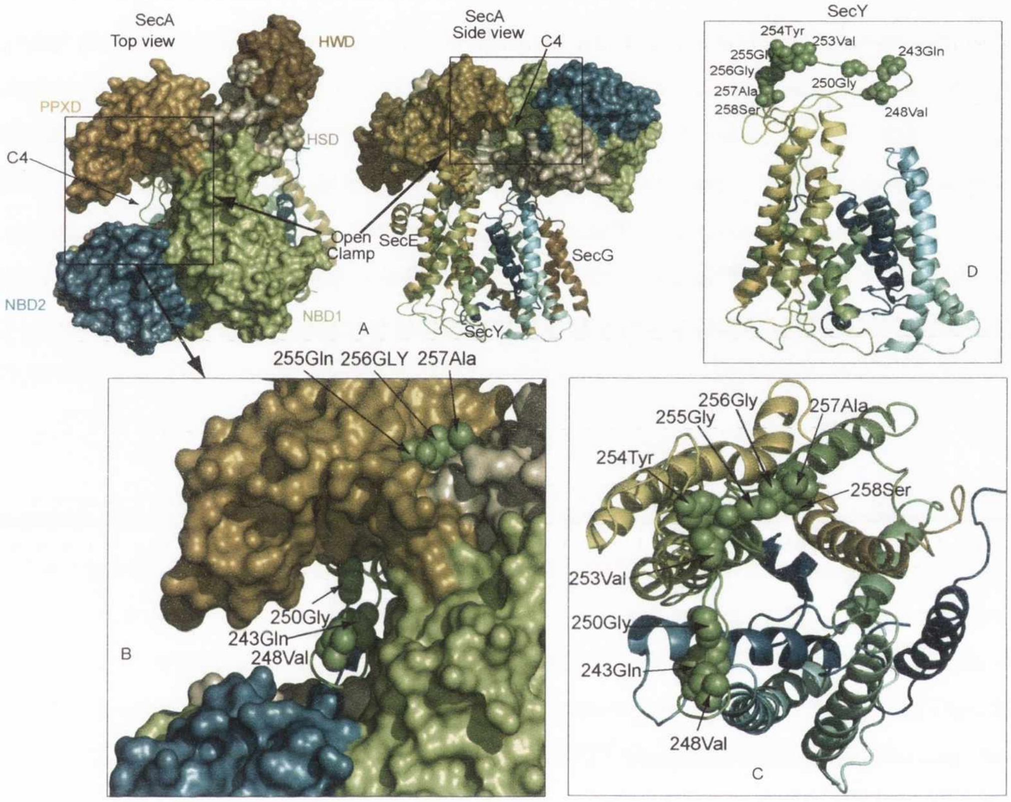


Figure 3.6: Amino acid sites under functional divergence mapped to the C4 domain of the SecY structure. Amino acid sites under functional divergence from ε–proteobacteria were mapped to the three-dimensional crystal structure of the SecYEG-SecA complex. (A) Three-dimensional crystal structure of SecA from *B. subtilis* (PDB: 3JV2) was superimposed to the SecYEG-SecA structure using CEalign in PyMOL, the original SecA structure was then removed from the SecYEG-SecA complex, where left figure top view and right side view clearly showed the C4 region in the polypeptide binding clamp. (B) A zoomed in figure of the C4 region clearly showed that mapped amino acid residues are in the clamp and also interact with the PPXD domain. (C) The same figure as shown in (B) without the SecA region, all sites from the C4 domain can be clearly observed from the top of the SecY structure. (D) The same sites are viewed from the SecY side.

3.5.3 Functional specialization of SecA2 and SecY2

In contrast with the canonical Sec system, here we found that all SecA2 proteins showed greater FD compared with the canonical SecA1 proteins in bacteria that have both SecA1 and SecA2 proteins, with the only exception of *Corynebacterium spp.*, in which the trend was the inverse to the previous one (group 14, Figure 3.1A). This is striking as a recent study has shown that both SecA1 and SecA2 proteins are essential in *Corynebacterium glutamicum* (Caspers & Freudl, 2008). As shown in Figure 3.1A and 3.3A, the SecA2/Y2 system of *Streptococcus spp.* and *Staphylococcus spp.* (group numbers 10, 11, 16, 19 and 1, 3, 4, 7, 14 for SecA2 (Figure 3.1A, Figure 3.4A) and SecY2 (Figure 3.3A, Figure 3.4E), respectively) are among the top 10% most functionally divergent bacterial lineages. This significant FD may have been required for the secretion of large serine-rich glycoproteins, which are heavily glycosylated in the cytosol after protein synthesis. These proteins include SraP and its homologs in *Staphylococcus* and *Streptococcus* species (Siboo et al., 2008) (Table 3.2).

In order to examine the functional locations of amino acid sites under FD in the protein structure, we mapped the corresponding amino acid sites of SecA2 and SecY2 from *Staphylococcus spp.* (group 10 in Figure 3.1A and group 4 in Figure 3.3A, respectively) to the recently solved SecYEG-SecA crystal structure (Zimmer et al., 2008). As shown in Figure 3.7, there are 14, 8, 16, 14, 3 and 0 sites mapped to the NBD1, PPXD, NBD2, HSD, HWD and CTL domains of SecA, respectively (gaps are not shown). Moreover, domains NBD2, HSD and NBD1 are specially enriched for sites under FD (Figure 3.1A), and some of these sites in SecA2 are located in two regions recently identified as functionally important. First, amino acids 774Glu, 775Ala, 786Pro and 793Glu are located in the two helix finger of the HSD domain (Figure 3.7D), and these four sites have been suggested to play an important role in moving polypeptide chains into the SecY channel (Erlandson et al., 2008b). Second, amino acid site 77Phe was identified in NBD1 (Figure 3.7E) and has been previously suggested to be a major target for SecA ATPase inhibitor (Li et al., 2008; Chen et al., 2010). Among the sites under FD, 35 were mapped to the C1-C6, T1-T10 and P1-P5 domains of SecY (sites located on unsolved structure regions are not shown), (Figure 3.8). Domains C4, C6, T8, C5, P1, T6 and T5 may play more important functions for niche adaptation as compared to other domains owing to their striking enrichment for FD (group 4 in Figure 3.3A). Among the bacteria with evidence for FD, 11 *Mycobacteria* without SecY2 (group number 22 in SecA2, Figure 3.1A) were clustered with group 28 consisting of seven plant species (*Physcomitrella patens subsp. patens*, *Sorghum bicolor*, *Vitis vinifera*, *Arabidopsis thaliana*, *Populus trichocarpa*, O*ryza sativa japonica* and *Ricinus communis*).

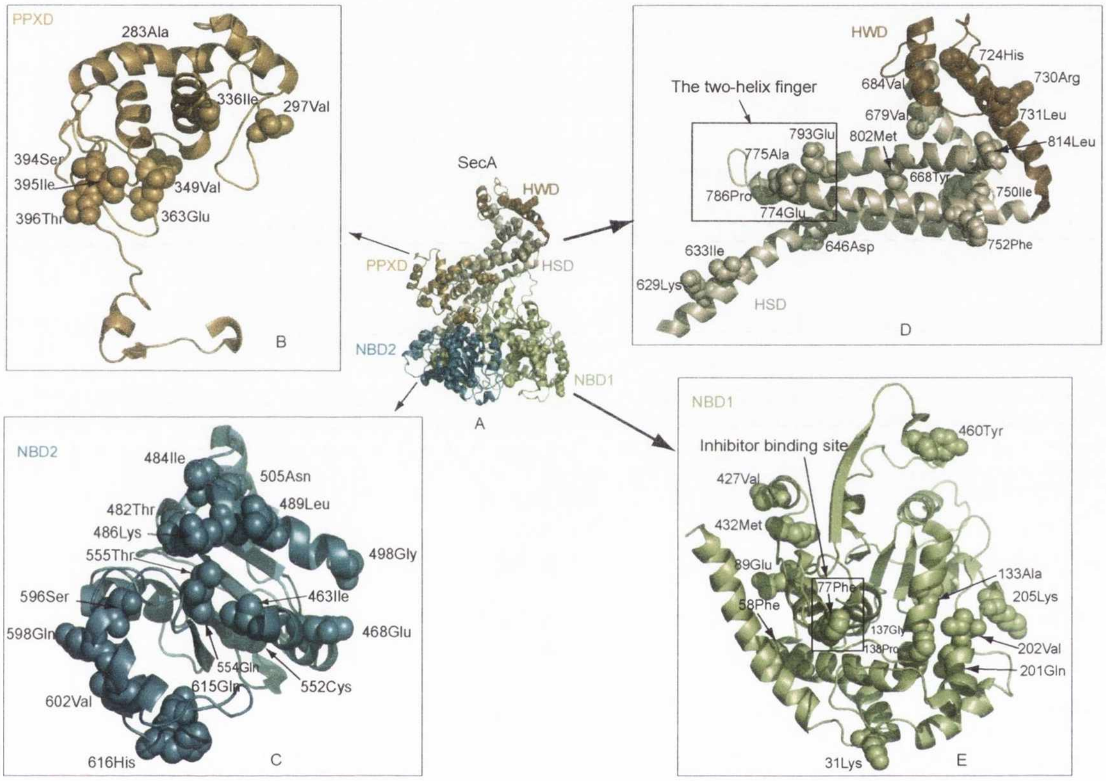


Figure 3.7: Amino acid sites under functional divergence of *Staphylococcus spp*. (group 10 in Figure 3.1) mapped to the SecA structure. (A) Sites are mapped to PPXD, NBD2, HSD, HWD and NBD1 of SecA. (B) Detailed view of the sites on PPXD. (C) Detailed view of the sites on NBD2. (D) Detailed view of the sites on HSD and HWD. (E) Detailed view of the sites on NBD1.

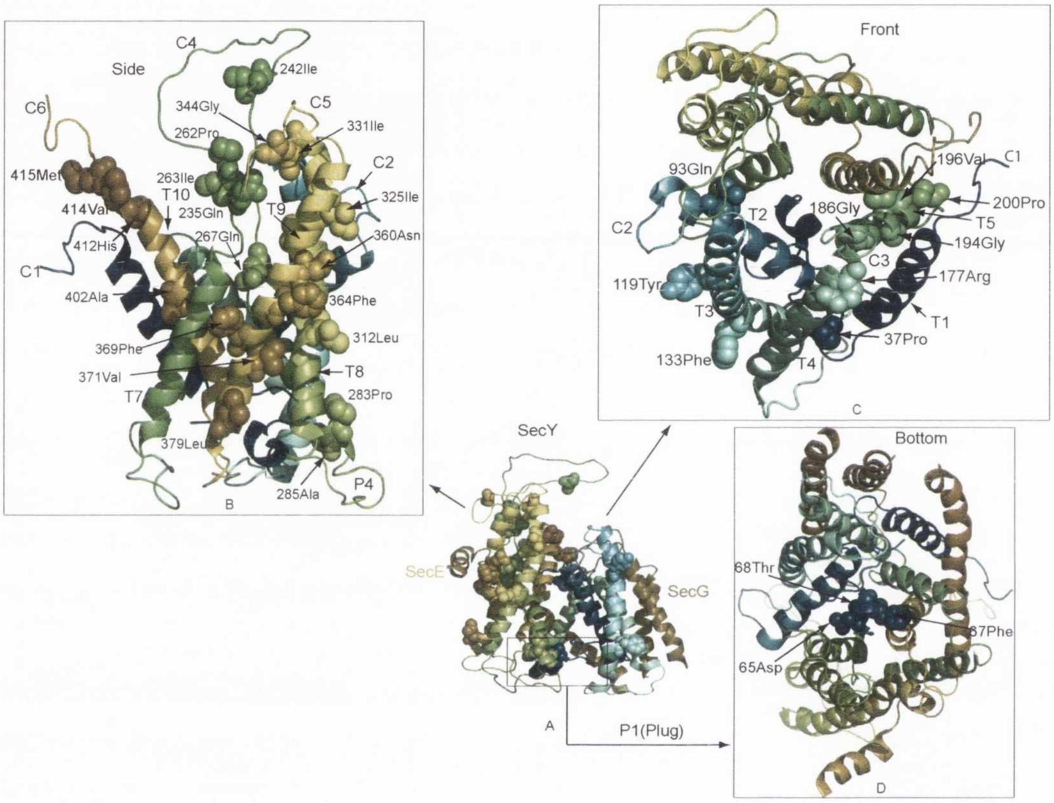


Figure 3.8: Amino acid sites under functional divergence of *Staphylococcus spp.* (group 4 in Figure 3.3) mapped to the SecY structure. (A) Sites are mapped to the SecYEG complex. (B) Detailed view of the sites from the SecY side. (C) Detailed view of the sites from the SecY front. (D) Detailed view of the sites from the SecY bottom.

3.5.4 Intermolecular coevolutionary patterns in the accessory SecA2/SecY2 system

Due to the importance of the interaction between SecA2 and SecY2 in the secretion of heavily glycosylated large serine-rich proteins, we aimed at identifying amino acids from both these proteins involved in such interactions using coevolutionary analyses. We define here coevolution as the correlated variation of two amino acids from the same protein or from two different proteins. It is based on the rationale that the coevolution of two amino acids is the result of the selective pressures that one amino acid imposes on the other amino acid so that breaking their correlated variation would result in a deleterious phenotype. To distinguish this type of coevolution from historical and stochastic covariation, we used a well-tested method to identify inter-protein coevolution that corrects for such phenomena (Fares & Travers, 2006), and that is implemented in the program CAPS (Fares & McNally, 2006) (see Materials and Methods for details).

It has long been of great interest to know how SecA and SecY translocon interact at the molecular level (Ito & Mori, 2009; Robson et al., 2009; Mori et al., 2010), and this has recently been advanced by the publication of the first X-ray crystallographic structure of the SecYEG-SecA translocon complex (Erlandson et al., 2008b; Zimmer et al., 2008; du Plessis et al., 2009; Zimmer & Rapoport, 2009; Auclair et al., 2010; Mori et al., 2010). However, to the best of our knowledge, no experimental studies have been performed so far to address this problem in the SecA2/SecY2 system, and therefore, it is not clear how SecA2 and SecY2 interact during the preprotein translocating process. An alternative way to characterize interactions between proteins and identify the involved amino acids is applying methods that identify common patterns of evolution among functionally/structurally linked residues.

Coevolution analysis identified two groups of intermolecular coevolving residues from the motor protein SecA2 and the core protein of the translocon SecY2, respectively, from 15 *Staphylococcus* species (see Materials and Methods section for details). As shown in Table 3.3 and Figure 3.9, group 1 has 8 coevolving residues, while group 2 has 19 coevolving residues (Figure 3.9B, Table 3.3). The two groups of coevolving residues share residues on T3, T8 and T10 of SecY2 (Table 3.3), which is in line with previous experimental studies conducted on the canonical SecA/SecY system demonstrating that these domains play an important role in the change of channel open and close states and in the displacement of the channel plug (P1 region). From Table 3.3 it is interesting to see that coevolving residues (465Glu and 593Ile) and 44Trp in group 2 are located in the NBD2 domain of the motor protein SecA2 and P1 domain (the plug domain) of SecY2, respectively (Figure 3.9B), where the plug domain is believed to play a role in sealing and opening the translocation channel SecY (Bostina et al., 2005; Tam et al., 2005; Li et al., 2007; Maillard et al., 2007; Saparov et al., 2007; Erlandson et al., 2008a; Gumbart & Schulten, 2008; Zimmer et al., 2008; Lycklama et al., 2010). Further analysis of coevolving residues identified that one amino acid site (396Thr) located on the PPXD domain has the high-

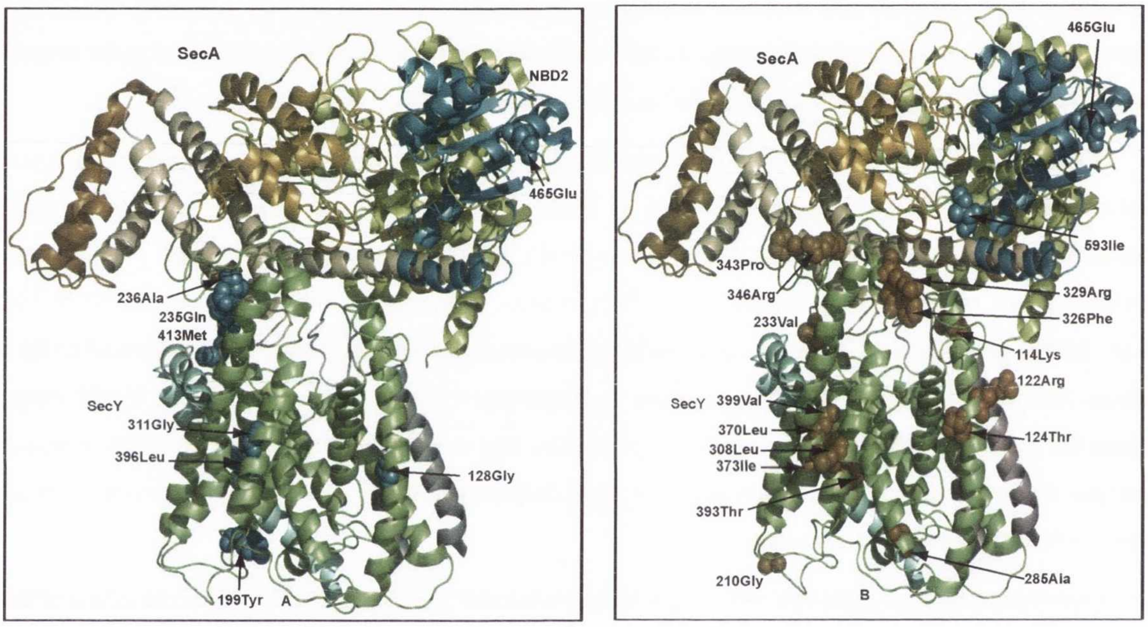


Figure 3.9: Coevolving residues of the accessory SecA2/SecY2 system mapped to the SecYEG-SecA complex. (A) Coevolving amino acid residues of group one (Table 3.3) are mapped to SecA (blue balls) and SecY (blue balls), respectively. (B) Coevolving amino acid residues of group two (Table 3.3) are mapped to SecA (blue balls) and SecY (red balls), respectively.

Group	AA Site	Domain	Protein
1&2	465Glu	NBD2	SecA2
2	593Ile	NBD2	SecA2
1	236Ala	C4	SecY2
1	235Gln	C4	SecY2
1	413Met	C6	SecY2
1	128Gly	T3	SecY2
1	199Tyr	T5	SecY2
1	311Gly	T8	SecY2
1	396Leu	T10	SecY2
2	4Ala*	C1	SecY2
2	114Lys	C2	SecY2
2	346Arg	C5	SecY2
2	329Arg	C5	SecY2
2	343Pro	C5	SecY2
2	44Leu*	P1(Plug)	SecY2
2	210Gly	P3	SecY2
2	393Thr	T10	SecY2
2	399Val	T10	SecY2
2	124Thr	T3	SecY2
2	122Arg	T3	SecY2
2	233Val	T6	SecY2
2	285Ala	T7	SecY2
2	326Phe	T8	SecY2
2	308Leu	T8	SecY2
2	373Ile	T9	SecY2
2	370Leu	T9	SecY2

Table 3.3: Functionally or structurally coevolving amino acid sites from the accessory SecA2/SecY2 system. Sites are mapped to domains and proteins and they are classified into two groups by CAPS analysis. *Amino acid sites are not available in the SecYEG-SecA crystal structure.

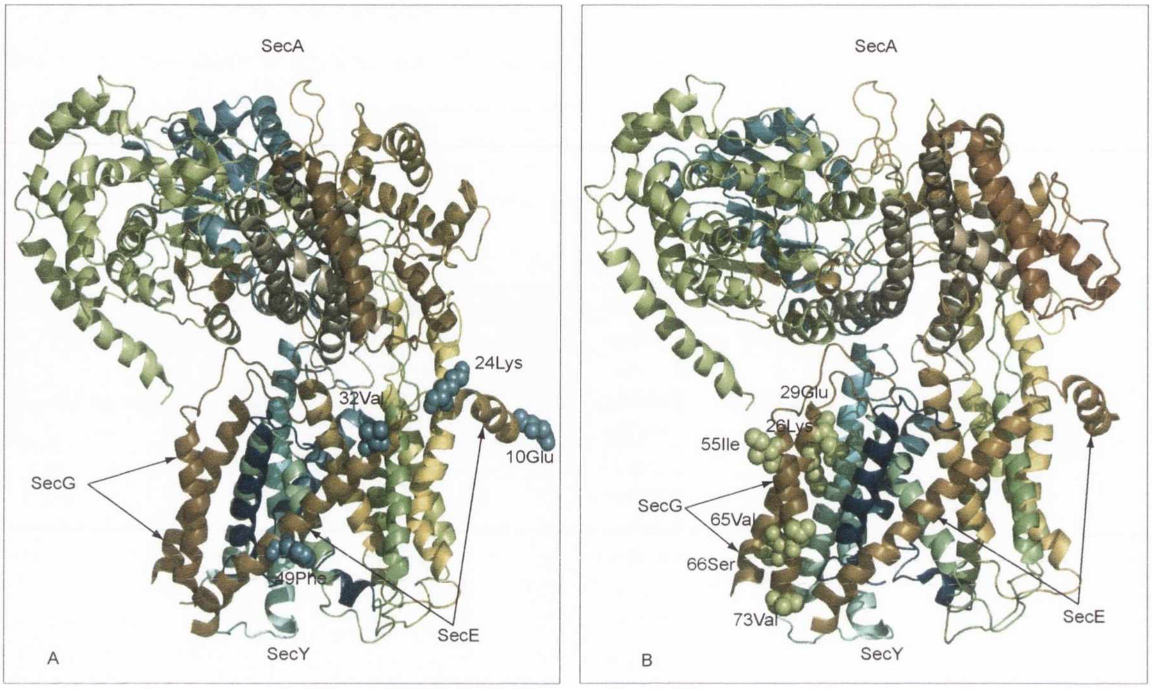


Figure 3.10: Amino acid sites under functional divergence (group 1 in Supplementary Table S3.3 and Table S3.4) of SecE and SecG mapped to the SecA-SecYEG structure. (A) Sites are mapped to SecE. (B) Sites are mapped to SecG.

est coevolving partners (326 amino acid sites), which indicate PPXD domain may play an important role in translocating SecA2-dependent substrate proteins.

3.5.5 Functional divergence in SecB

We identified a clade with the strongest FD that contains mostly pathogenic *Rickettsia spp.* (group 1 in Supplementary Table S3.2). In this clade 20 amino acid sites are identified to be under strong functional divergence. To further address the functions of these sites, we mapped these sites to *E. coli* SecB structure (Dekker et al., 2003). This SecB structure is structurally aligned to the *Haemophilus influenzae* SecB structure in which SecB is bound to the C-terminal 27 amino acids of *H. influenzae* SecA. This alignment is successful as the RMSD (root mean square deviation) is 3.2 angstroms, which indicates that the over-all protein structure of SecB may be highly conserved. Structural analysis shows that most of these FD sites are distributed among two SecA-binding regions on SecB (Figure 3.11). One is close to the C-terminal linker on SecA (Zhou & Xu, 2005) (Figure 3.11A). The other is close to SecB-binding site on SecA (Figure 3.11B). This SecB-binding site is highly conserved and located in the C-terminus of SecA. These functionally divergent sites on SecB may play a role in regulating SecA by binding to the conserved C-terminus region of SecA. This is consistent with our FD analysis of SecA, where we show that the CTL domain is impoverished for functional divergence (Figure 3.1). This suggests that the CTL domain may function in a different way in SecA's ATPase activity.

To address the relations between CTL and other protein domains in SecA, SecB and SecY, we performed two coevolutionary analyses between SecA and SecY and SecA and SecB, respectively. In the combined coevolutionary networks between SecA, SecB and SecY, we found that three amino acid sites have the highest number of coevolving partners that come from the CTL domain of SecA. These coevolving partners are distributed among all protein domains (except C2 and T2) in SecY and two regions mentioned above in SecB (Figure 3.11A and Figure 3.11B), which suggests that CTL domain may structurally/functionally regulate other protein domains during SecB, SecA and SecY interaction (Supplementary Figure S3.3).

3.5.6 Functional divergence in SecE and SecG

One clade containing two pathogenic *Acinetobacter sp.* is the most functionally divergent group of SecE (Figure 3.4B). Interestingly, we observed that SecG proteins of plant pathogens from *Xanthomonas sp.* and *Xylella fastidiosa* have accumulated the most radical changes (Figure 3.4C) indicating SecG may play a role in the pathogenesis of bacteria in plants. We mapped the sites under FD to SecE (Figure 3.10A) and SecG (Figure 3.10B) in the SecYEG-SecA structure. In SecE, sites 49Phe, 32Val and 24Lys are in the vicinity of the transmembrane domains of SecY, and they may consequently contribute to the proper function of SecY channel. Similarly, in SecG, sites 29Glu, 26Lys, 65Val, 66Ser and 73Val may contribute to the proper function of SecY and possible interaction with SecA.

3.5.7 Functional divergence of translocon core protein SecY in archaeal clades

There are 24 archaeal branches identified with at least one site under FD (Supplementary Table S3.5). Among the top 10% groups of archaeal species with FD, there are three archaeal groups (group number 11, 16 and 22, Figure 3.3A). These groups have accumulated the most radical amino acid changes compared with the rest of the 21 archaeal groups. The top three archaeal groups are *Methanococcus* and *Methanocaldococcus sp.* (group number 11), *Sulfolobus sp.* (group number 16), *Thermococcus sp.* and *Pyrococcus sp.* (group number 22). Interestingly, group 11 and 22 were not clustered with any other bacterial groups (Figure 3.3A), indicating a unique pattern of protein domains under FD. *Thermococcus sp.* and *Pyrococcus sp.* are hyperthermophilic archaea and they belong to the order *Thermococcales*. *Thermococcales* can be found in terrestrial, submarine hot vents and deep subsurface environments. Importantly, *Pyrococcus sp.* can only be found in marine environments and belong to a particular niche (Bertoldo & Antranikian, 2006). It is tempting to speculate that

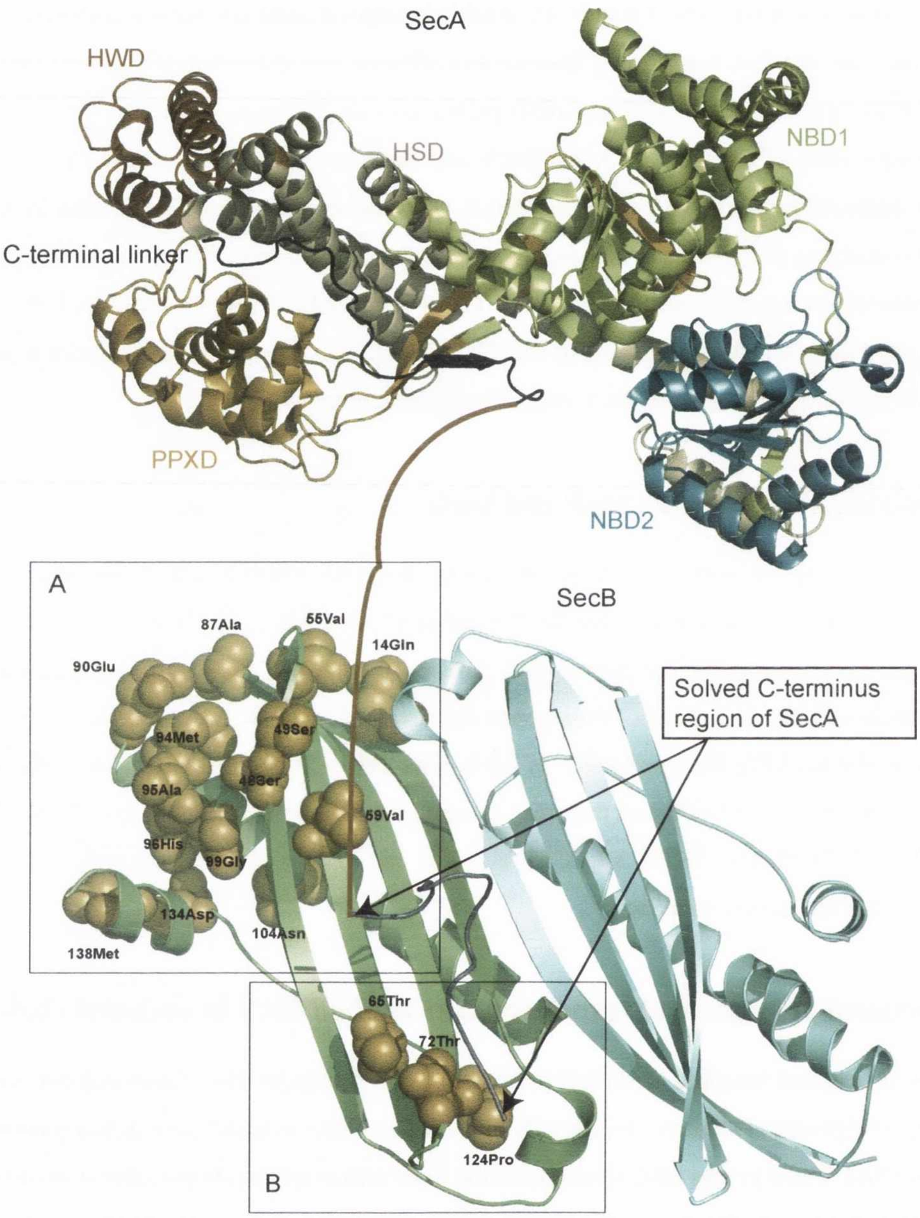


Figure 3.11: Amino acid sites under functional divergence of *Rickettsia spp.* mapped to the SecB structure. SecA structure is shown to demonstrate the possible interactions between the SecA CTL domain and the SecB dimer (two SecBs formed a dimer: the left one is colored green, and the right one is colored cyan. Sites are only mapped to the left one). (A) SecA-binding site close to the C-terminal linker (colored black on the SecA structure, PDB ID: 1M6N). Note that the unsolved CTL region is colored red, and it is manually linked to the solved C-terminus region. (B) SecA-binding site close to the highly conserved C-terminus region (colored magenta, the solved structure comes from PDB ID: 1QYN, 1OZB).

the above clustering pattern for group 11 and 22 may be a consequence of adaptation to their specific living environments. It is known that SecY is key in the translocation of most of their membrane proteins in these archaeal species (Yuan et al., 2010), and due to their unique living environments, preproteins translocated through archaeal SecY might be substantially different compared with bacterial preproteins.

3.5.8 Deferential protein secretion in functionally divergent bacteria

To find possible correlations between functional divergence and ecological adaptation, we identified the Sec- and Tat-dependent secreted proteins of the top 5 most functionally divergent clades for SecA, SecB, SecE, SecG and SecY, respectively (we analysed 207 secretomes, for complete analysis results, see Supplementary Table S3.6). To characterize the functions of these secreted proteins, we grouped these proteins into COG functional categories and statistically tested their significance. In this way, we can identify statistically enriched functional categories of secreted proteins. Next, we report representative COG functional categories that are significantly over-enriched (P<0.001) across all the secretomes we analysed. First, most secretomes were enriched in COG category M (Cell wall/membrane/envelope biogenesis).This is expected as the Sec system is a major secretion pathway for secreting membrane related proteins. Second, to our surprise, we found that COG category P (Inorganic ion transport and metabolism) is almost over-enriched in all the secretomes of pathogenic as well as non-pathogenic bacteria under functional divergence. This indicates the proteins responsible for inorganic ion transport and metabolism may play an important role in bacterial niche adaptation. Finally, COG category U (Intracellular trafficking, secretion, and vesicular transport) is enriched in most pathogenic bacteria, this is consistent with the conclusion from many previous studies that secretion systems play a central role in bacteria host adaptation (Wooldridge, 2009).

To illustrate how secretomes are analysed for their enrichment status, we showed the enrichment status of COG functional categories of secreted proteins of 5 representative species from clades that have the greatest number of sites under FD for SecA, SecB, SecE, SecG and SecY, respectively. In Figure 3.12, we ranked all the COG functional categories according to their numbers of secreted proteins, and then we labeled their enrichment status based on their statistical test. Interestingly, in Figure 3.12B, we identified that the obligate intracellular pathogen *Rickettsia prowazekii* has COG category O (Posttranslational modification, protein turnover, chaperones) over-enriched in its secretome (P<0.01). Chaperones are known to buffer the deleterious mutations in bacteria and therefore this can help *Rickettsia prowazekii* adapt to an obligate intracellular life style (Fares et al., 2002b). To demonstrate how strong functional divergence in SecA/SecY system of gram-positive bacteria *Streptococcus* and *Staphylococcus* affects their secretion, we showed three species from *Streptococcus* and *Staphylococcus*, respectively. As shown in Figure 3.13, all three *Streptococcus spp.* are over-enriched in COG category M, P, T and U. Strikingly, all three *Staphylococcus* species are over-enriched in COG category P (P<0.001) and they have the highest number of secreted proteins in category P compared to other categories. How secreted proteins in inorganic ion transport and metabolism contribute to *Staphylococcus spp.* niche adaptation? Surprisingly, *Staphylococcus* species was found to survive in high salt concentrations, and it was reported that these species can grow well at NaCl concentrations as high as 10–15% and even higher (Oren, 2006). To survive high salt concentrations, the cell must be able to counterbalance the osmotic pressure produced by this environment. Inorganic ions such as Na^+ and K^+ must be processed to maintain appropriate osmotic pressure in the cell. It was also known the pathogenic bacteria use inorganic ion concentrations to sense their location. This is important for *Streptococcus* and *Staphylococcus* species because they are frequently found to be part of the microbiota of the human skin, mucosal surface and so on (Cleary & Cheng, 2006; Goetz et al., 2006). And more importantly, many crucial ions that pathogenic bacteria require are located in the host cell and bound to host proteins. Therefore, these pathogens need to actively acquire these ions (Guiney, 1997).

3.6 Discussion

In this study we demonstrated that the SecYEG-SecA system of pathogenic bacteria is the most functionally divergent among the bacterial lineages. We also presented evidence supporting that bacteria lineages with the accessory SecA2/SecY2 system, which are mostly pathogenic, share largest proportion of sites under FD. Importantly, most of the amino acid sites affected by FD are located in the two-helix finger of the HSD domain as well as in the NBD1 ATP binding domain. This indicates that pathogenic bacteria with accessory SecA2/SecY2 system may have evolved new functions to secrete substrates specialised in mediating their interaction with hosts, as suggested by previous studies (Bensing & Sullam, 2002; Braunstein et al., 2003; Lenz et al., 2003; Bensing et al., 2004; Chen et al., 2004; Takamatsu et al., 2004a; Chen et al., 2006; Kurtz et al., 2006; Seifert et al., 2006; Bensing et al., 2007; Chen et al., 2007; Gibbons et al., 2007; Muraille et al., 2007; Wu et al., 2007; Caspers & Freudl, 2008; Hou et al., 2008; Rigel & Braunstein, 2008; Siboo et al., 2008; Bensing & Sullam, 2009; Hou et al., 2009; Mistou et al., 2009; Rigel et al., 2009). As a case in point, many life-threatening infections are caused by *Staphylococcus spp.* and *Streptococcus spp.* when they interact with their host cells through the SecA2-dependent secretion of adhesins expressed on bacterial cell surface (Fitzgerald et al., 2006).

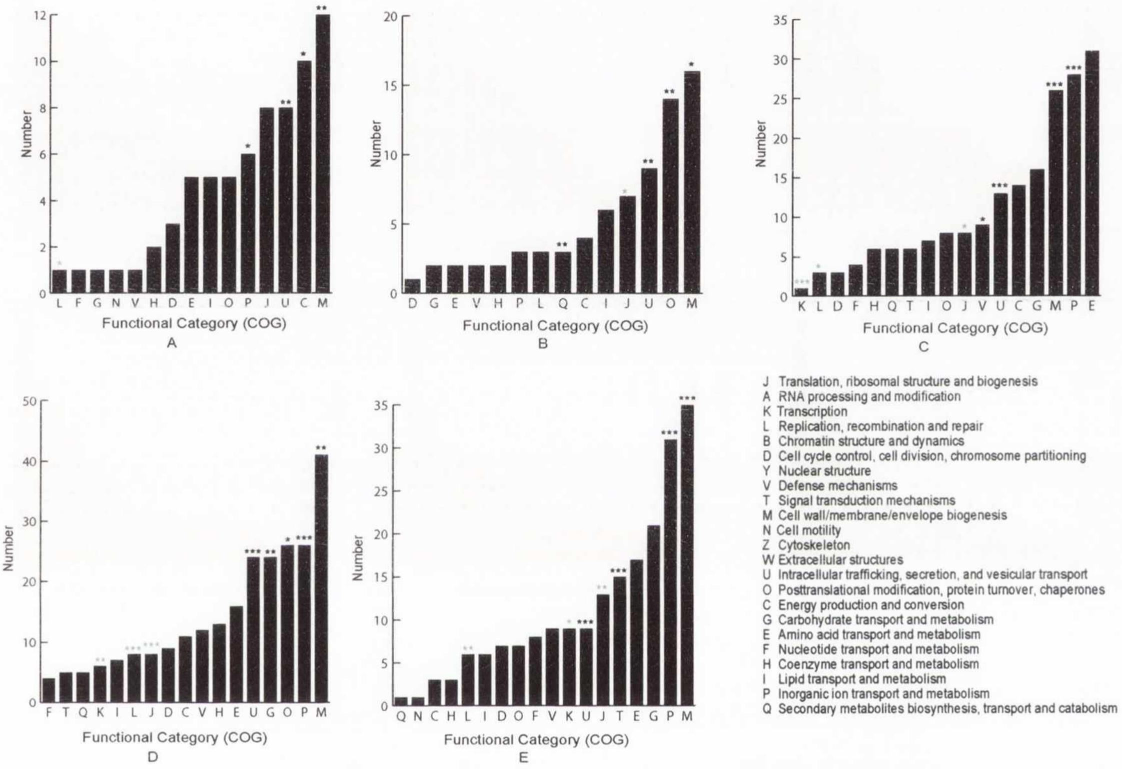


Figure 3.12: Analysis of Sec-dependent secreted proteins (secretome) of representative bacteria from clades with the strongest functional divergence. These clades are from SecA, SecB, SecE, SecG and SecY, respectively. Secreted proteins are grouped functionally according to the classification of proteins into the Cluster of Orhtologous Groups (COG). These functional categories were ranked and plotted according to the numbers of their secreted proteins. (A) *Chlamydia trachomatis*, (B) *Rickettsia prowazekii*, (C) *Acinetobacter baumannii*, (D) *Xylella fastidiosa*, (E) *Streptococcus agalactiae*. COG functional categories showing significant enrichment (black stars) or impoverishment (grey stars) of predicted substrate proteins are labelled by (* P<0.05, ** P<0.01, *** P<0.001).

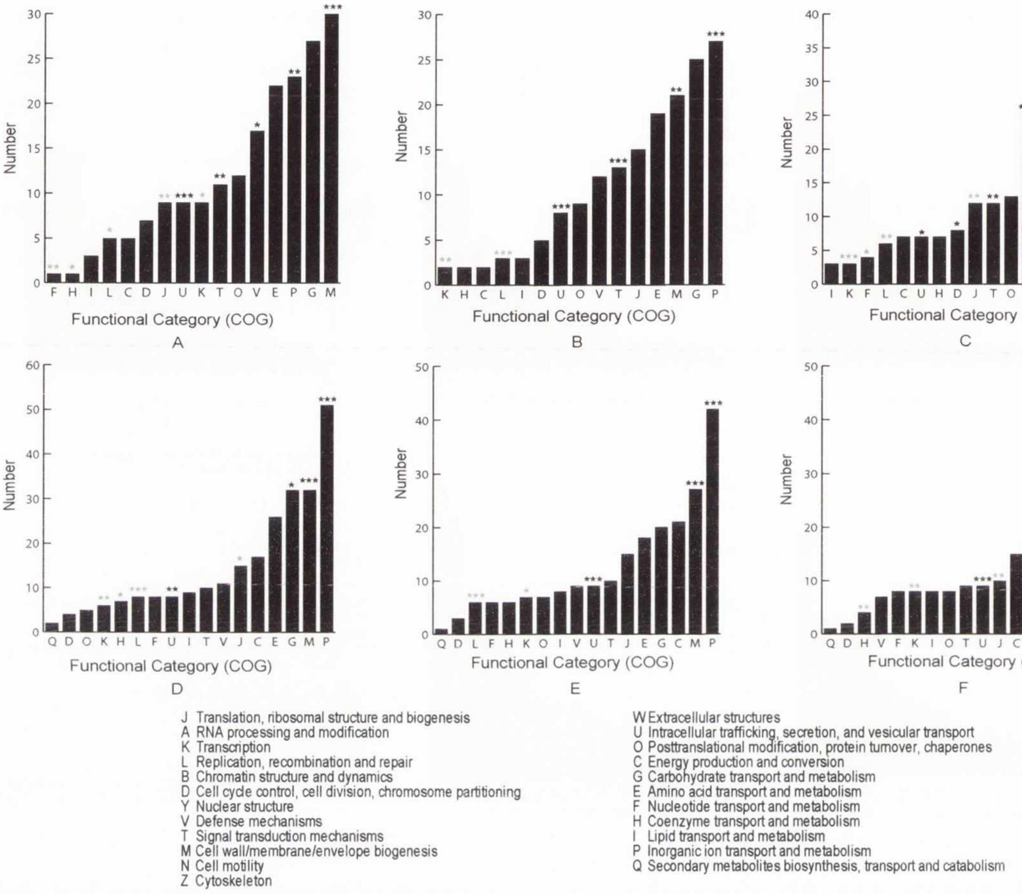


Figure 3.13: Analysis of Sec-dependent secreted proteins (secretome) from *Streptococcus* and *Staphylococcus*. Three species are chosen for *Streptococcus* and *Staphylococcus* from the clades with strong functional divergence, respectively. Secreted proteins are grouped functionally according to the classification of proteins into the Cluster of Orhtologous Groups (COG). These functional categories were ranked and plotted according to the numbers of their secreted proteins. (A) *Streptococcus gordonii*, (B) *Streptococcus pneumoniae*, (C) *Streptococcus sanguinis*, (D) *Staphylococcus aureus*, (E) *Staphylococcus epidermidis*, (F) *Staphylococcus haemolyticus*. COG functional categories showing significant enrichment (black stars) or impoverishment (grey stars) of predicted substrate proteins are labelled by (* P<0.05, ** P<0.01, *** P<0.001).

3.6.1 Functional divergence in the ATP binding domain (NBD1) of motor protein ATPase SecA

We revealed the main mechanisms that drive the evolution of the motor protein SecA. We show that NBD1 and PPXD are the most enriched domains for sites under FD. NBD1 contains the ATPase activity, which is required for the right conformational changes in SecA to take place. These conformational changes are crucial in the transfer of the preprotein into the transmembrane SecYEG complex. PPXD domain of SecA seems to be involved in protein substrates intake and release and is believed to be the most dynamic domain, as supported by structural alignments and distance measurements of the different states of all solved SecA structures (Figure 3.2 and Table 3.1). Because of the fundamental role of these two domains in the interaction with SecA substrates, their high enrichment for FD may be directly related to the ability of SecA to interact with substrates specific to the ecological niche of the cell. This conclusion would require, however, further investigation to directly link the functionally divergent amino acid replacements to protein secretion.

The first low micromolar inhibitors of bacterial SecA were recently synthesized following a preliminary in silico screening and have proven effective in vitro and in vivo against *E. coli* strains (Li et al., 2008; Chen et al., 2010). Authors of the two studies showed that the parent compound of the synthesized inhibitors form hydrogen bond specifically with a region located in the NBD1 domain of SecA (Figure 3.7E). They also showed that such a compound interacts with a residue (417Gly) located in the PPXD domain and a residue (534Arg) located in the NBD2 domain. The amino acid sites identified to have undergone FD in our study are in the vicinity of these regions (Supplementary Table S3.5) and hence they are promising targets for novel inhibitory compounds. Moreover, NBD2 and HSD domains shared similar profiles of FD, probably due to the possible modulation of HSD domain function or structure by NBD2, possibly through the two-helix finger, proposed previously to participate in moving the substrate polypeptide chain into the SecY channel (Erlandson et al., 2008b). It is worth mentioning that most studies on the development of therapeutic drugs examine highly conserved regions of proteins. We show here that variable regions may also have important roles in more profound protein functions and may indeed be directly responsible for the action of key molecules in bacteria responsible for their pathogenesis.

Evidence from a variety of literature sources and our own data on functional divergence support the the conclusion that NBD1 is the most functionally divergent domain in SecA from pathogenic bacteria. Similar molecular patterns of FD in the SecYEG-SecA protein complex were observed among major human and animal bacterial pathogens such as *Chlamydia*, *Streptococci*, *Staphylococci*, *Mycobacteria*, *Listeria*, *Legionnella* and *Mycoplasma*. *Chlamydial* species are obligate intracellular bacterial pathogens and have a unique lifestyle, requiring a special set of membrane proteins to interact with the host (Heinz et al., 2009). The secretion of these proteins may have been possible through the FD changes in SecA of *Chlamydia*.

Coxiella burnetti was amongst the most affected bacterial branches for FD. The range of hosts that *Coxiella* and *Mycoplasma* can infect is very wide and includes arthropods, fish, birds, and a variety of mammals (Heinzen & Samuel, 2006; Razin, 2006). The generalist capacity of these bacteria might have required specific protein sets to invade a variety of ecological niches (hosts), although the virulence factors involved in the host-pathogen interaction are poorly defined. We also found that the sequences of SecYEG-SecA of all the *Mycobacteria tuberculosis* strains so far sequenced are identical both at the protein and nucleotide levels and probably present similar pathogenic potential for Sec-dependent virulence. Interestingly, these bacteria have been proposed to be prone to acquire virulence factors by horizontal gene transfer (Smith et al., 2009), sparking speculation that gain and loss of Sec-dependent host colonization factors may have contributed to host specificity in bacteria such as in *Mycobacterium bovis* (host : cattle) and *M. tuberculosis* (host: human).

Some of the most affected pathogenic bacteria with FD are those that possess an accessory SecA2/SecY2 system and include highly pathogenic and antibiotic resistant bacteria such as *Staphylococcus aureus* (meticillin-resistant, vancomycin-susceptible, vancomycin-intermediate resistance strains) and *Streptococcus pneumoniae* etc. Although not essential for pathogens survival (Bensing & Sullam, 2002; Rigel & Braunstein, 2008) this system has been found to be responsible for secreting a set of virulence factors (Rigel & Braunstein, 2008). It has been demonstrated that some gram-positive bacteria have an accessory SecA2/SecY2 system (Rigel & Braunstein, 2008), suggested to be involved in translocation of posttranslationally modified (for example glycosylated) proteins in pathogenic *Streptococcus* and *Staphylococcus* species (Chen et al., 2004; Takamatsu et al., 2004a,b; Bensing et al., 2005; Siboo et al., 2005; Takamatsu et al., 2005; Wu et al., 2007; Siboo et al., 2008; Mistou et al., 2009; Zhou et al., 2010). The enrichment of secreted proteins sets in these bacteria for glycosylated large cell surface glycoproteins (about 2000-5000 amino acids) may have required larger conformational SecA changes than usual, hence optimised ATP binding and hydrolysis. Such changes may have been possible through the duplication of SecA and SecY, yielding SecA2 and SecY2, and the functional specialisation (functional divergence) of both these proteins to secrete large proteins (Chen et al., 2006; Hou et al., 2008; Rigel et al., 2009).

Clustering of bacteria according to FD has also brought some interesting results, among which we highlight the clustering of some *Mycobacteria* that lack SecY2 with a group consisting of seven plant species. Recent studies have shown that the secretion of superoxide dismutase A (SodA) is SecA2 dependent, and SodA may help *M. tuberculosis* survive the oxidative attack of macrophages and thus plays a role in *M. tuberculosis* pathogenesis (Zhang et al., 1991; Braunstein et al., 2003; Rigel et al., 2009). *M. tuberculosis* SodA belongs to the iron Sod (FeSOD) group, which has five homologs in *Arabidopsis thaliana*. It has been shown that two of the five homologs are FeSOD (fsd2 and fsd3) and located in the chloroplast, where they play essential roles in early chloroplast development (Myouga et al., 2008). Although it is not established whether the secretion of chloroplast FeSOD is SecA dependent (to the best of our knowledge), it is tempting to speculate that this might be the case. Such a remarkable convergent evolution between bacterial SecA2 and chloroplast SecA may have been the result of an adaptation to lifestyles that require the secretion of similar set of proteins.

3.6.2 Radical functional divergence in SecA ATPase may compensate for the lack of SecB chaperone

Gram-positive pathogens seem to lack the molecular chaperone SecB (for example using *E. coli* SecB as query against KEGG SSDB database returns no significant hits in these bacteria). This is surprising since SecB mediates the posttranslational translocation. Other proteins may therefore substitute SecB function, although these proteins remain to be identified. The identification of a chaperone activity for SecA raises the possibility that this protein has a SecB like role in posttranslational translocation (Eser & Ehrmann, 2003). We show that SecA protein from gram-positive bacteria has accumulated the most radical changes compared with other bacterial lineages supporting a different role of this protein in these bacteria. The fact that a mutation of *E. coli* SecA can partially compensate for the absence of SecB chaperone (McFarland et al., 1993), which hints that radical amino acid substitutions in SecA are likely to confer this protein the ability to directly bind preproteins and perform a chaperone-like activity.

3.6.3 Functional divergence and intermolecular coevolution of prokaryotic translocon channel and motor protein mediates ecological adaptation

The translocon core protein SecY (Sec61 in eukaryotes) is conserved within all three domains of life, supporting strong selective constraints on this protein within each of the domains. In addition, coevolutionary analysis on the accessory SecA2/SecY2 system indicates that there may be common mechanisms shared between the accessory and canonical SecA/SecY systems. For instance, the ATPase activity (NBD2 domain) of SecA2 seems to be coupled with the channel opening and sealing of SecY2, which was also evident in the canonical SecA/SecY system (Robson et al., 2009). Thus, these coevolving sites may play an important role in maintaining the stability of the plug displacement upon SecA2 binding to SecY2 as suggested in the canonical SecA1/SecY1 system (Zimmer et al., 2008), which would support that ATP hydrolysis (NBD2 domain) is important in the opening of the sealed accessory translocation channel SecY2. Given our results we propose four conditions to be met for the translocation of proteins through this complex. First, amino acid changes in the ATPase SecA2 may be required to help the PPXD domain to effectively interact with the highly glycosylated substrates. Second, a wider channel opening state may also be required to cope with the glycosylated amino acid residues. Third, the NBD2 domain on SecA2 may play an important role in tuning a "translocation competent" state of SecY2 protein for translocating these large glycosylated cell surface proteins. Lastly, sites on the plug domain (P1), together with others located on other SecY2 protein domains (e.g. T7) may cope with the binding of the long N-terminal signal peptide of large glycosylated preprotein in a coordinated way and therefore their coevolution is essential.

3.6.4 Functional divergence of the translocon channel core SecY contributes to the diversification of major proteobacteria groups

There are three main reasons that justify describing the evolution of the conserved SecY protein to understand the diversification of major proteobacteria groups (ε, α, β, γ, δ). First, SecY is evolutionarily conserved across all domains of life. Second, SecY plays a fundamental role in the biogenesis of membrane proteins and cell surface proteins of prokaryotes. Third, membrane proteins and cell surface proteins directly determine the lifestyle that a prokaryotic species can adopt, because these proteins are at the frontier of environment-bacteria interaction (Tjalsma & van Dijl, 2005; Zanen et al., 2006; Marques et al., 2008; Poetsch & Wolters, 2008). Therefore, diversification of membrane and cell surface proteins may contribute to the diversification of major proteobacteria groups, which may be reflected on the evolutionarily conserved SecY protein. Indeed, we show that the levels of FD differ greatly with the diversification of the major proteobacterial groups and link tightly with their ecological contexts.

3.7 Conclusion

In this study, we revealed the main evolutionary forces that drive the evolution of the SecA-SecYEG complex. In addition to this, we show that these proteins have diverged the most in pathogenic bacteria compared to non-pathogenic ones. We also show that SecA2 and SecY2 have coevolved possibly to drive the secretion of a more specialised set of proteins involved in pathogenesis. Finally, we unveil the molecular evolutionary mechanisms in these secretion proteins and their potential roles in niche specific adaptations, and we propose these proteins as possible drug targets for future therapeutic drugs.

3.8 Acknowledgments

This work was supported by Science Foundation Ireland, under the program of the President of Ireland Young Researcher Award to Mario A. Fares, and the IRCSET (Irish Research Council for Science, Engineering and Technology) Government of Ireland Postgraduate Scholarship in Science, Engineering and Technology to Xiaowei Jiang. This work was also supported by grant BFU2009-12022 funded by the Spanish Ministerio de Ciencia e Innovación to MAF. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We would like to thank some members of the Fares' lab for a critical reading of the manuscript.

Chapter 4

Functional Diversification of the Twin-Arginine Translocation Pathway Mediates the Emergence of Novel Ecological Adaptations

4.1 Related manuscript

Jiang, X., & Fares, M. A. (2010). Functional diversification of the twin-arginine translocation pathway mediates the emergence of novel ecological adaptations. Molecular Biology and Evolution, 28(11), 3183–3193.

4.2 Abstract

Microorganisms occupy a myriad of ecological niches that show an astonishing diversity. The molecular mechanisms underlying microbes' ecological diversity remain a fundamental conundrum in evolutionary biology. Evidence points to that the secretion of a particular set of proteins mediates microbes' interaction with the environment. Several systems are involved in this secretion, including the Sec secretion system and the Tat pathway. Shifts in the functions of proteins from the secretion systems may condition the set of secreted proteins and can, therefore, mediate adaptations to new ecological niches. In this manuscript we have investigated processes of functional divergence—a term used here to refer to the emergence of novel functions by the mod-

ification of ancestral ones—of Tat pathway proteins using a large set of microbes with different lifestyles. The application of a novel approach to identify functional divergence allowed us to distinguish molecular changes in the three Tat proteins among different groups of archaea and bacteria. We found these changes as well as the composition of secreted proteins to be correlated with differences in microbe's lifestyles. We identified major signatures of functional divergence in halophilic and thermophilic archaea as well as in pathogenic bacteria. The location of amino acids affected by functional divergence in functionally important domains of Tat proteins made it possible to find the link between the molecular changes in Tat, the set of secreted proteins and the environmental features of the microbes. We present evidence that links specific molecular changes in secretion mediating proteins of microbes to their ecological adaptations.

4.3 Introduction

Bacterial, archaeal and eukaryotic cells use two major secretion systems to transport proteins that are synthesised in the cytosol, namely the general secretion (Sec) pathway and the twin-arginine translocation (Tat) pathway. The Sec pathway transports proteins in an unfolded state, whereas the Sec-independent Tat pathway transports proteins that are partially or fully folded. The N-terminal signal peptide, as part of the substrate protein, plays a significant role in targeting proteins to the cytoplasmic membrane, where they can be integrated or alternatively transported across. This N-terminal peptide presents a tripartite configuration that includes an N-terminal region at the beginning, a hydrophobic region in the middle and a C-terminal region in the end. A typical Tat signal peptide has an arginine-arginine (twin-arginine) consensus motif in its N-terminal region (Yuan et al., 2010).

Two major differences exist in the mechanism of protein secretion of Sec-dependent and Tat systems. The Sec-dependent translocation pathway uses two modes of protein secretion: the cotranslational translocation and the posttranslational translocation. In the cotranslational translocation pathway, which is conserved in both archaea and bacteria, proteins are translocated across or into the plasma membrane during protein synthesis. In contrast to this, in the posttranslational translocation, proteins are first synthesized in the cytosol and stabilised in an unfolded state through the chaperone activity of SecB and/or SecA for their subsequent translocation across the membrane (McFarland et al., 1993; Eser & Ehrmann, 2003). The translocation of proteins occurs through the pore formed by SecY protein and favoured by the ATPase activity of SecA. Importantly, SecB and/or SecA mediated translocation pathway is only found in bacteria but not in archaea. The mechanism that archaea use to solve the problem of posttranslational translocation—that is stabilising proteins in an unfolded or partially folded states for their subsequent translocation across the membrane, remains elusive (Irihimovitch & Eichler, 2003).

The Sec-independent Tat pathway is conserved in bacteria, archaea, chloroplasts and plant mitochondria (Lee et al., 2006), pointing to its essentiality in the viability of these organisms. It consists of three major components, TatA, TatB and TatC, all of which are present in most gram-negative bacteria. In contrast, most gram-positive bacteria and archaea have only two Tat components, TatA and TatC (Eijlander et al., 2009a). Most of our knowledge regarding the mechanisms of substrate recognition and translocation by the Tat system is based on studies of the two model organisms, *Escherichia coli* and *Bacillus subtilis* (Panahandeh et al., 2009). These studies show that Tat translocase uses Tat(B)C complex as a receptor site for binding Tat substrate proteins. After substrate binding, Tat(B)C recruits multiple TatA proteins to form a pore-containing protein complex that allows passing substrates through the pore driven by a proton motive force (Lee et al., 2006; Tarry et al., 2009; Yuan et al., 2010). TatC was shown to determine the substrate specificity and thus plays a critical role in Tat-dependent protein secretion (Strauch & Georgiou, 2007; Eijlander et al., 2009a). What is the mechanism whereby the Tat pathway monitors the folding fidelity of proteins for their correct translocation? In one study, authors showed that indeed TatA presents a proofreading function of its substrate FeS proteins NrfC and NapC, so that when such substrates are incorrectly folded they undergo rapid degradation (Matos et al., 2008).

Although the mechanism is not entirely clear, the Tat pathway seems to mediate the emergence of new ecological adaptations in many bacteria and archaea. For example, recently an association between ecological adaptation and the Tat secretion pathway has been established in halophilic archaea *Halobacterium sp. NRC-1* (Rose et al., 2002). Studies showed that this group of archaea uses the Tat translocation pathway extensively to transport most of their secreted proteins (Dilks et al., 2005). Haloarchaea live in an environment with "near-saturation" salt concentrations, a harsh condition for the proper function and secretion of proteins whose functions are important in later stages outside the cytosol. Interestingly, this correlates with the fact that most sequenced haloarchaea have two TatC homologs (TatCo and TatCt). In contrast, therefore, to many bacteria in which Tat transports a small fraction of proteins and are not essential for cell viability (Lee et al., 2006), TatC may be essential to haloarchaea in buffering the environmental conditions.

Not only is the Tat pathway a major route for protein secretion in haloarchaea, but also it is essential for cell viability (Dilks et al., 2005; Thomas & Bolhuis, 2006). Some other prokaryotes are known to use the Tat pathway extensively, such as *Streptomyces* species (e.g. *S. coelicolor*) (Widdick et al., 2006). This difference in Tat essentiality between microbes seeks an explanation and can be crucial to understand ecological adaptation. What makes Tat essential and what molecular bases contribute to the different essentiality of Tat in different groups of organisms? Could we find molecular changes that are associated with environmental adaptations? Are there any other prokaryotes that present similar patterns in their Tat translocase? These questions remain largely unexplored.

Secretion systems play a significant role in prokaryotes environmental adaptation (Wooldridge, 2009) and, as such, understanding the evolution of these systems is a crucial aim in evolutionary biology. Unfortunately, however, contrary to cytosolic proteins, a large amount of secreted proteins are not structurally characterised, posing difficulties to perform detailed evolutionary analyses (Elofsson & von Heijne, 2007; Popot, 2010). Alternative approaches are therefore required to circumvent the limitations to studying the evolution of these systems.

Here we hypothesise that the Tat pathway has diverged its function from the ancestral ones in microbes with particular ecological traits. As a case in point, we focus in two main types of microbes: halophilic archaea, that live in environments saturated with salt, and pathogenic bacteria, that may require the secretion of proteins to interfere with the immune response of the host. We use prokaryotic genome sequences, together with abundant secondary structure data, a new computational tool to identify divergence in protein functions and a wide range of experimental studies reporting crucial information to test this hypothesis.

4.4 Materials and Methods

4.4.1 Sequences and genomes

Homologous protein sequences for TatA, TatB and TatC were retrieved using API (Application Program Interface) (`http://www.genome.jp/kegg/soap/`) service from KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database (Kanehisa et al., 2008). Genomes and protein table files were retrieved from NCBI (National Center for Biotechnology Information) Genome database (`ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.ptt.tar.gz`). Protein table files were later used for functionally annotating Tat-dependent substrates proteins with COG (Clusters of Orthologous Groups) (Tatusov et al., 2003). We also annotated all pathogenic species with their host (human, animal and plant) and the diseases they cause (ICD-10, International Statistical Classification of Diseases and Related Health Problems 10th Revision, `http://apps.who.int/classifications/apps/icd/icd10online`) wherever possible as this allowed us to classify bacteria as being pathogenic when appropriate. Amino acid sequences of all homologs were aligned using MUSCLE 3.7 (Edgar, 2004) and alignments were manually checked using Jalview (Waterhouse et al., 2009). Gaps are removed in the functional divergence analysis.

4.4.2 Phylogeny reconstruction

Amino acid substitution model and model parameters were chosen using Prottest (Abascal et al., 2005) for each of the Tat proteins: TatA, TatB and TatC. These parameters were then used by RaxML program

(Stamatakis, 2006) to estimate the best Maximum likelihood phylogenetic trees for the three proteins. The phylogenetic trees of TatA, TatB and TatC with their corresponding protein sequence alignments were used as starting data for the analyses of functional divergence.

4.4.3 Analysis of functional divergence in Tat proteins

Functional divergence (FD) is a term that refers to a change in function as a result of an amino acid substitution event within a protein. Here we used a relaxed definition of FD to describe shifts in the ancestral function of a protein as a result of important amino acid substitutions in that protein—these shifts might not involve the emergence of a completely novel function but the modification of an existing one. FD can therefore refer to complete change in function after gene duplication or slight shifts in the function of a protein after speciation.

Previous studies have used the term FD to exclusively refer to shifts in the function of a protein after gene duplication—that is to say, FD between paralogs. Under this view, they classified FD into two main categories, type I and type II FD (Gu, 1999). In type I FD amino acids are highly conserved at particular sites in one paralog while being highly stochastic in the other, indicating the acquisition of a function at these sites in one paralog (the conserved one) or the loss of the function in the other. In type II FD, amino acids are highly conserved in both paralogs at the same positions of the protein, indicating the divergence of two highly important functions between the paralogs.

Recently, a new method was implemented to identify type I FD at the genome level between orthologs (Toft et al., 2009) and was proved efficient in identifying conserved amino acid sites with FD in a large phylogenetic tree including orthologous as well as paralogous sequences (Williams et al., 2010). Briefly, the method uses a protein sequence alignment and a phylogenetic tree including paralogs and orthologs. The method then compares a pair of clades in the tree sharing a common ancestral origin to their closest phylogenetic outgroup. The comparison is performed for each amino acid site in the protein and the strength or likelihood of the amino acid state is evaluated using the appropriate BLOSUM matrix (BLOcks of Amino Acid SUbstitution Matrix) (Henikoff & Henikoff, 1992). BLOSUM matrices comprise the scores for the transition between the 20 amino acids, with positive scores meaning the transition is more frequent than expected, negative means the transition is less frequent than expected and 0 means these transitions are as frequent as expected. Using BLOSUM as a score matrix for the transitions between amino acids allows scoring both variable sites and conserved sites: one clade may seem variable yet the amino acid transitions between orthologous sequences within the clade may have taken place between biochemically close amino acids (positive scores). In the comparison between two clades to an outgroup (clade 1 and clade 2), functional divergence in clade 1 would be detected if the BLOSUM scores were positive within clade 1, negative between clade 1 and the outgroup and positive between clade 2 and the same outgroup. To test the significance of these score variances, we calculated the mean and standard error of the scores for clade 1 ($\overline{C}_1$ and SE_1) and clade 2 ($\overline{C}_2$ and SE_2) (Toft et al., 2009). The significance of the difference in the means was obtained by calculating the Z-score for the comparison of samples with unequal sizes as:

$$Z = \frac{\overline{C}_1 - \overline{C}_2}{SE_{C_{1,2}}}$$

Where $\overline{C}_{1,2}$ are the mean substitution scores for the transition from clades on either side of the bifurcation in the phylogenetic tree relative to the outgroup and $SE_{C_{1,2}}$ is the combined standard error of clade 1 and clade 2.

4.4.4 Protein secondary structure prediction

Two approaches were used to predict protein secondary structures for TatC, TatA and TatB, respectively. We first used the SMART database to predict protein domains in TatC (Letunic et al., 2009). We then used the Psipred to predict protein structures of TatA and TatB (Bryson et al., 2005). The latter method is based on artificial neural network approaches and can reach accurate protein secondary structure predictions (`http://bioinf.cs.ucl.ac.uk/index.php?id=779`). Protein domains of *E. coli* TatC were predicted first and then mapped back to the protein sequence alignment to define protein domains for all the proteins in the alignment. These domains included three cytoplasmic domains (C1-C3), six transmembrane domains (T1-T6) and three periplasmic domains (P1-P3). Similarly, protein sequences of TatA and TatB from *E. coli* were used as reference sequences. N-terminal region, transmembrane domain, hinge region, the amphipathic helix and C-terminal domain were predicted for *E. coli* TatA and TatB, respectively. We then used the *E. coli* TatA and TatB as reference and mapped these structural regions in their multiple sequence alignments to define the protein domains for TatA and TatB. After FD analysis, we mapped amino acid sites under FD to these protein domains. To understand the relevance of the results of FD analyses, we built a two dimensional matrix, with the rows representing clades under FD and the columns representing the number of sites under FD in each protein domain. Three such data matrices were made, one per protein: TatC, TatA and TatB, respectively. We then used the heatmap.2 function in R (`http://www.r-project.org`) to cluster the rows and columns to find common patterns of FD among protein domains and clades in the matrix. The heatmap.2 function scales each element in a row in the data matrix based on a normalised Z-score, which is calculated as follows:

$$z_{ij} = \frac{x_{ij} - \overline{x}_i}{\sigma_i}$$

Here, z_{ij} is the Z-score calculated for element x_{ij} in the matrix, $\bar{x}_i$ is the mean of the row i , σ_i is the standard deviation of all the elements in the row i .

4.4.5 Prediction of potential Sec and Tat substrate proteins

We used Tatp to predict potential Tat-dependent secreted proteins (secretome) (Bendtsen et al., 2005b) and SignalP to predict Sec-dependent secretome (Emanuelsson et al., 2007). In predicting Sec-dependent substrates, we first used SignalP and then removed those predicted also by TatP to minimise false positives. Predicted substrates from the Sec and the Tat pathways were used to assess how different the two secretion systems, Tat and Sec, were and how much they explained about ecological adaptations in our organisms. We preferred using TatP instead of Tatfind (Rose et al., 2002) for the prediction of Tat-dependent secretome as the former approach is based on artificial neural networks that largely outperforms other approaches in pattern recognition tasks pertaining variant Tat signal peptides. In contrast, Tatfind is a regular expression based approach, which does not allow recognition of variant Tat signal peptides (Bendtsen et al., 2005b).

Predicted Sec and Tat substrates were grouped according to their COG functional categories. To determine if the numbers of substrates for Tat-dependent secretion in each of the categories were significantly higher (enriched category for secreted substrates) or lower (impoverished category) than expected by chance, a χ^2 test was performed on each COG functional category with one degree of freedom. The statistical test was performed according to the following equation:

$$\chi_i^2 = \frac{(O_i - E_j)^2}{E_i}$$

$$E_i = n_i \frac{N_{predicted}}{N_{COG}}$$

O_i is the observed number of predicted substrate proteins in COG category i. while E_i is the expected number of such proteins in that category i. The total number of proteins in category i is indicated by n_i. $N_{predicted}$ is the total number of predicted substrate proteins, while N_{COG} is the total number of proteins in the proteome. Only proteins assigned within single COG functional categories were analysed (ambiguous categories R and S were not included).

4.5 Results and Discussion

4.5.1 Evidence for functional divergence in the Tat translocation system

In this study we analysed Tat homologs (orthologs and paralogs) retrieved from 944 bacterial genomes, 68 archaeal genomes and 11 chloroplasts. We estimated the amount of functional divergence in each of the Tat proteins in the different lineages of the phylogenetic tree (see Materials and Methods for details). Functional divergence (FD) refers here to the divergence of protein's function from its ancestral function by amino acid replacements in functionally important amino acid sites. We applied a method to identify functional divergence previously developed (Toft et al., 2009; Williams et al., 2010). Analysis of our tree allowed us to identify 204, 145 and 119 clades with at least one conserved amino acid site with evidence of FD for TatC, TatA and TatB, respectively (Supplementary information: Table S4.1 to Table S4.3 and Figure S4.1 to Figure S4.3). The number of clades under FD in each Tat phylogenetic tree (TatC: 204; TatA: 145; TatB: 119) seems to be correlated with TatC and TatA being the most important components in determining a minimum Tat translocase (Blaudeck et al., 2005; Panahandeh et al., 2009). In particular, a mutant TatA can complement a strain with a deleted TatB without compromising substrates translocation (Blaudeck et al., 2005).

In total, we identified 167 bacterial pathogens (100% of the pathogenic bacteria in our data) and 8 haloarchaea (100% of the analysed haloarchaea) to be involved in a cluster with at least one site under FD. To understand the relationship between FD and ecological adaptation, we focused on the 5 most functionally divergent clades of the tree, that have at least 4 sequences in clade 1—that is to say, organism clades that presented the largest number of amino acid sites with evidence of having diverged their functions from ancestral clades. We also show the relative phylogenetic positions of the clades under FD (Figure 4.1). Phylogenetic clades of halophilic archaea containing TatC homologs, TatCo and TatCt which resulted from a TatC gene duplication, have been identified to be the most functionally divergent of all clades (Figure 4.1A), namely the halophilic archaea have the greatest number of amino acid sites (32 sites for TatCo, TatCt has 24 sites) under FD compared with other lineages. This is interesting as haloarchaea had to adapt to an extreme environment and FD of TatC after duplication may have enabled such adaptations through the acquisition of novel functions or the modification of its original function. In contrast to TatC, thermophilic archaea *Sulfolobus* presented most of its FD signal in TatA (Figure 4.1B). Finally, *Corynebacteria* accumulated most of the functionally divergent changes in TatB (Figure 4.1C). Importantly, 8 out of the 11 clades with the strongest signal of FD included pathogenic bacterial strains (Figure 4.1), which suggests that Tat translocation pathway may play a role in bacterial pathogenesis. In summary, clades showing strong FD in Tat translocation system are unique as they include microorganisms living in extreme environments always challenged by ionic concentrations, temperature or host immune response (We discuss these results with detail in the following sections).

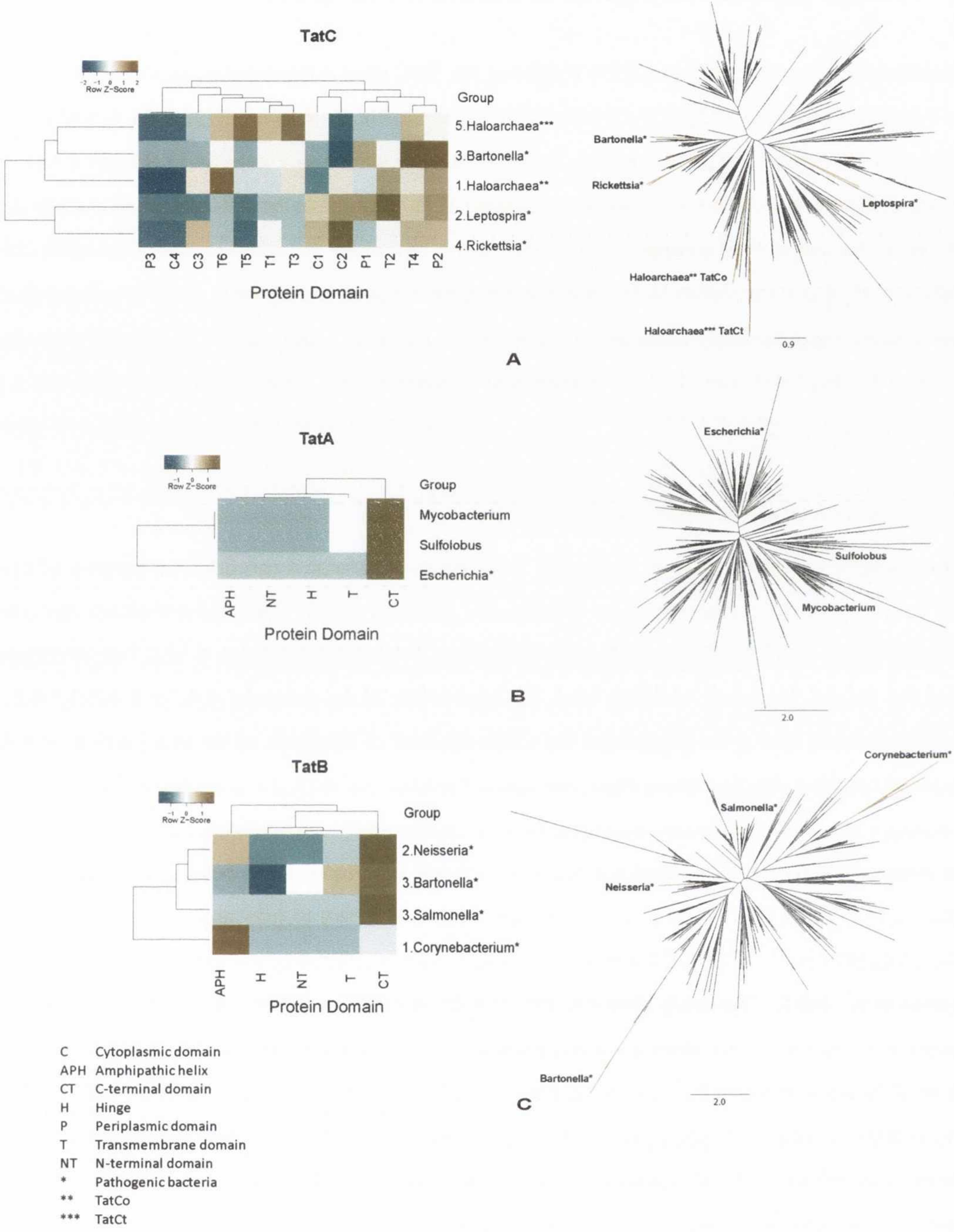


Figure 4.1: Clustering analysis of phylogenetic clades and protein domains according to functional divergence analyses. We examined enrichment for amino acid sites with functional divergence in the proteins of the Tat secretion pathway TatC (A), TatA (B) and TatB (C) under functional divergence. The relative position of the enriched clades for functional divergence is shown in red in the maximum-likelihood phylogenetic tree next to each of the heatmaps. The high resolution trees are available as Supplementary information (TatC: Figure S4.10; TatA: Figure S4.11; TatB: Figure S4.12).

4.5.2 Strong functional divergence in TatC from halophilic archaea

Haloarchaea presented substantial FD following the TatC gene duplication event that gave rise to TatCo (Table 4.1, amino acid sites under FD are mapped to Tat alignment, as shown in Figure S4.4) and TatCt (Table 2, amino acid sites under FD are mapped to Tat alignment, as shown in Figure S4.5) (Figure 4.1A). A close look to the distribution of amino acid sites with evidence of FD shows that the different proteins domain present significant differences in their content of FD (Figure 4.1A). In general, we could identify three main clusters of domains according to their content in functionally divergent amino acid sites (sites under functional divergence are shown in the Supplementary Table S4.4 to Table S4.6). The first cluster consists of 2 protein domains (from left to right: P3 and C4, Figure 4.1A) that were impoverished for FD. The second cluster included 5 protein domains (from left to right: C3, T6, T5, T1 and T3, Figure 4.1A) that showed modest number of amino acid sites under FD. Finally, the third cluster comprised 6 protein domains (from left to right: C1, C2, P1, T2, T4 and P2, Figure 4.1A) and was highly enriched for FD.

Interestingly, halophilic archaea TatCo and TatCt proteins presented two different patterns of FD: TatCt showed strong FD in the second cluster of FD domains, affecting mostly TatC transmembrane domains while TatCo presented most of its FD in the N-terminal region of the protein (Figure 4.1A). Non-overlapping FD between the protein homologs resulting from the duplication of an ancestral gene is a strong indicator of subfunctionalization after gene duplication, in which the sum of functions of the two paralog complements the ancestral function. We also found three pathogenic bacterial clades (*Leptospira*, *Bartonella* and *Rickettsia*) to be amongst the most functionally divergent ones (Supplementary information: Table S4.4), with the third cluster (comprising protein domains C1, C2, P1, T2, T4 and P2, Figure 4.1A) being enriched for FD in these bacteria (Figure 4.1A). The N-terminal half of TatC has been shown to play an important role in precursor binding (Holzapfel et al., 2007). C1 and C2 were shown to be important for Tat-dependent protein transport (Buchanan et al., 2002). One study demonstrated that the mutation R16A in TatC (also called TatCd, a TatC homolog) in *B. subtilis* could abolish the secretion of a Tat-dependent substrate PhoD, homologous site of which we detected to be under FD (Table 4.2, see Col. 226) in haloarchaea (Eijlander et al., 2009b). Moreover, mutations in the C2 and P2 domains were shown to generate a Tat version capable of translocating proteins with variations in the RR motif of Tat signal peptide (Kreutzenbeck et al., 2007; Strauch & Georgiou, 2007), a motif that needs to be generally conserved for the translocation to take place. These studies suggest that extensive FD in these domains of haloarchaea TatCo has contributed to the expansion of Tat substrates repertoire in this clade. The same rationale applies to the pathogenic bacteria *Leptospira*, *Bartonella* and *Rickettsia* (De Buck et al., 2008; Joshi et al., 2010; Reynolds et al., 2011).

TatC transmembrane domains have been shown to interact with other TatC or TatB to form complexes

Col.[1]	Z[2]	Residues (count) for **C1**/C2/Outgroup [3]
226	8.633***	**V(4) M(1) I(1) A(2)**/ R(4)/ R(1)
242	3.321***	**A(3) T(2) S(3)**/ I(2) M(2)/ V(1)
245	2.878*	**V(6) L(1) A(1)**/ A(2) S(1) I(1)/ S(1)
539	5.549***	**H(2) G(2) Q(2) N(2)**/ S(3) A(1)/ S(1)
543	198.49***	**L(7) T(1)**/ W(3) M(1)/ W(1)
551	34.199***	**V(8)**/ L(3) M(1)/ F(1)
553	5.26***	**T(3) G(1) S(3) A(1)**/ L(3) F(1)/ L(1)
569	10.404***	**T(6) S(2)**/ L(3) A(1)/ L(1)
577	197***	**M(8)**/ A(4)/ A(1)
596	73.907***	**Y(8)**/ F(3) L(1)/ L(1)
599	10.538***	**S(3) A(5)**/ I(1) V(3)/ I(1)
604	116.59***	**L(7) V(1)**/ F(3) L(1)/ Y(1)
607	19.379***	**G(2) S(2) A(4)**/ F(4)/ F(1)
611	2.798*	**V(5) I(2) L(1)**/ A(2) T(1) G(1)/ T(1)
629	43.133***	**Y(8)**/ Y(3) I(1)/ I(1)
709	4.966***	**I(3) V(3) L(1) T(1)**/ A(3) P(1)/ P(1)
808	34.199***	**L(8)**/ V(4)/ A(1)
811	3.125***	**T(3) A(1) G(2) R(1) S(1)**/ Q(3) G(1)/ E(1)
815	3.376***	**D(2) N(6)**/ S(4)/ T(1)
817	34.199***	**I(4) L(2) M(2)**/ V(4)/ A(1)
823	34.199***	**L(2) F(3) Y(3)**/ G(4)/ S(1)
839	115.6***	**M(8)**/ K(4)/ R(1)
844	13.071***	**T(7) S(1)**/ E(1) K(2) Q(1)/ K(1)
845	131.98***	**R(8)**/ Y(3) R(1)/ Y(1)
847	135.95***	**W(8)**/ T(4)/ T(1)
859	4.404***	**G(7) A(1)**/ V(1) A(2) L(1)/ L(1)
860	6.419***	**F(3) S(2) A(3)**/ I(3) L(1)/ I(1)
879	237.7***	**P(8)**/ Q(4)/ Q(1)
883	43.133***	**A(7) T(1)**/ A(3) L(1)/ L(1)
885	167.22***	**T(8)**/ L(3) V(1)/ L(1)
893	34.199***	**T(8)**/ S(4)/ G(1)
902	54.715***	**W(8)**/ L(3) F(1)/ I(1)

Table 4.1: Amino acid sites under functional divergence in haloarchaea TatCo. Amino acid residues are represented with one letter code. Numbers of the same amino acid residues are also calculated in each clade. Amino acid sites under functional divergence are also mapped to TatC alignment shown in Figure S4.4.

[1] Relative position of amino acid sites in TatC multiple sequence alignment.

[2] Score of Functional Divergence: *, ** and *** indicate $P<0.05$, $P<0.01$, $P<0.001$, respectively.

[3] C1 is the clade under functional divergence, C2 is the clade against which comparative analyses are made. Amino acid residues and their numbers are labelled in bold in C1.

Col.[1]	Z[2]	Residues (count) for **C1**/C2/Outgroup[3]
190	4.33***	**A(4) G(1) S(1) T(1)**/ E(3) P(1) M(1)/ P(1)
225	3.919***	**A(6) I(1)**/ L(4) F(1)/ L(1)
230	52.645***	**Q(7)**/ L(4) T(1)/ L(1)
236	4.798***	**F(6) W(1)**/ L(3) I(1) V(1)/ L(1)
244	4.53***	**G(5) A(2)**/ V(1) I(2) L(2)/ I(1)
536	61.242***	**A(6) T(1)**/ Y(3) Q(1) L(1)/ Y(1)
570	16.101***	**D(3) R(3) N(1)**/ Y(4) L(1)/ Y(1)
617	2.439*	**L(4) A(2) V(1)**/ L(3) F(1) M(1)/ F(1)
618	42.122***	**F(7)**/ M(5)/ V(1)
620	33.667***	**F(7)**/ L(4) F(1)/ V(1)
631	78.648***	**A(7)**/ L(3) Y(1) M(1)/ L(1)
707	5.134***	**F(5) L(2)**/ V(4) A(1)/ A(1)
815	4.52***	**Q(3) E(3) G(1)**/ S(3) N(2)/ T(1)
817	4.737***	**V(2) I(5)**/ I(2) A(2) V(1)/ A(1)
828	27.752***	**A(4) S(2) G(1)**/ F(5)/ F(1)
839	68.695***	**Y(7)**/ R(4) K(1)/ R(1)
844	2.905*	**P(4) A(1) Q(1) R(1)**/ K(3) P(1) Q(1)/ K(1)
853	56.751***	**W(7)**/ R(5)/ R(1)
857	173.78***	**V(5) I(2)**/ Y(5)/ Y(1)
861	7.867***	**F(5) A(1) Y(1)**/ M(1) L(3) F(1)/ L(1)
868	26.379***	**F(7)**/ V(2) I(2) T(1)/ T(1)
891	19.033***	**G(5) V(1) A(1)**/ E(5)/ E(1)
895	3.558***	**A(1) Y(4) G(1) Q(1)**/ V(3) L(2)/ L(1)
897	6.241***	**A(3) S(3) T(1)**/ V(4) L(1)/ L(1)

Table 4.2: Amino acid sites under functional divergence in haloarchaea TatCt. Amino acid residues are represented with one letter code. Numbers of the same amino acid residues are also calculated in each clade. Amino acid sites under functional divergence are also mapped to TatC alignment shown in Figure S4.5.

[1] Relative position of amino acid sites in TatC multiple sequence alignment.

[2] Score of Functional Divergence: *, ** and *** indicate $P<0.05$, $P<0.01$, $P<0.001$, respectively.

[3] C1 is the clade under functional divergence, C2 is the clade against which comparative analyses are made. Amino acid residues and their numbers are labelled in bold in C1.

for signal peptide and substrate binding (Alami et al., 2003; Behrendt et al., 2007; Punginelli et al., 2007). However, the exact function of these domains in substrate recognition and binding is less well understood. The observed FD in TatCt transmembrane domains may have led to the formation of a functionally different TatCt complex, which could interact with a specific group of Tat substrates in haloarchaea.

TatC was suggested to play a role in determining the specificity of Tat pathway dependent secretion (Jongbloed et al., 2000). In *B. subtilis* and *E. coli*, both TatCs were shown to serve as a primary RR motif recognition site (Mendel et al., 2008). Haloarchaea encodes two TatC paralogs, and both are among the most functionally divergent clades. Rose and colleagues first showed that *Halobacterium sp. NRC-1* extensively uses the twin-arginine translocation pathway, instead of Sec, to transport most of its secreted proteins (Rose et al., 2002). Such shift of protein transport from Sec to Tat may have been crucial to solve the protein folding problem faced by microbes in high salt concentration environments: folding proteins first in the cytosol and then exporting them from the cell. This provides an explanation to the essentiality of TatC proteins for cell viability in halophilic archaea (Dilks et al., 2005; Thomas & Bolhuis, 2006). Moreover, we analysed the secretomes of two extreme halophilic bacteria *Salinibacter ruber* and *Halorhodospira halophila* whose genomes are available. Strikingly, COG category P—that comprises proteins involved in inorganic ion transport and metabolism–is ranked as the largest group with known function in Sec-dependent secretome of *S. ruber* (χ^2=23.49, P< 0.001, Supplementary Figure S4.6C and Table S4.10). Conversely, in Tat-dependent secretome, functional category P is only ranked at 6th position ($\chi^2 = 0.18$, not statistically significant, Supplementary Figure S4.6D and Table S4.11). Similarly, P is ranked at the second ($\chi^2 = 9.48$, P< 0.01) and fourth ($\chi^2 = 1.14$, not statistically significant) positions in Sec- and Tat-dependent secretomes in *H. halophila*, respectively (Supplementary Figure S4.6A and B and Table S4.10 and Table S4.11). Not surprisingly, none of the two halophilic bacteria, which present both secretion systems Sec and Tat, showed any sign of strong FD in their Tat components. In the post-translational translocation pathway in bacteria, SecB and/or SecA, that present chaperone activity, can bind to the newly synthesized proteins and keep them in an unfolded state, which is required for their Sec-dependent translocation. In halophilic bacteria, SecB and/or SecA can act as holdases preventing the premature folding of the synthesized proteins, ensuring therefore correct protein translocation despite the high salt concentrations. Contrary to this, halophilic archaea do not have SecA and SecB proteins that could mitigate the effects of high salt concentrations on protein translocation (Dilks et al., 2005). In these archaea proteins need therefore to be translocated in a folded state, which is possible through the Tat pathway that does not require unfolded proteins. This, however, would be possible provided that the substrate-specificity of TatC has relaxed (for example, recognition of RR motif in a signal peptide of Tat-substrate proteins would be no longer a requirement), possibly through key amino acid substitutions that we detect to be under FD.

In conclusion, the distinct FD patterns observed here between TatCo and TatCt have played an important role in the adaptation of halophilic archaea to environments with high salt concentrations, which is a clear example of how FD of the Tat pathway has contributed to the ecological adaptation of halophilic archaea.

4.5.3 Functional divergence of thermophilic archaea *Sulfolobus* TatA

FD analyses of TatA highlighted thermophilic archaea *Sulfolobus* to be the clade with the strongest profile of FD (Figure 4.1B, amino acid sites under FD are mapped to TatA alignment shown in Figure S4.7). The other two clades containing non-pathogenic *Mycobacteria* and pathogenic *E. coli* strains have the same amount of amino acid sites under FD as *Sulfolous* (Figure 4.1B, Supplementary information: Table S4.5). In TatA we identified a main cluster joining domains APH, NT and H (Figure 4.1B). Importantly, the C-terminal domain showed strong signal of FD compared to the other domains (Figure 4.1B). In contrast to TatC, however, the amount of functional divergence in the different domains of TatA was low and we decided therefore to look at the phylogeny that included the entire set of clades to gain insight on FD in TatA. Clustering analyses based on functional divergence approach and taking all clades together for TatA allowed identifying a more general pattern with three well-distinguishable clusters (Supplementary information: Figure S4.2). The first cluster contained the C-terminal domain, being this the one most affected by functional divergence in TatA, the second cluster was almost exclusively represented by the amphipathic helix domain and the third cluster contained the transmembrane, the hinge and the N-terminal domains. This clustering pattern is consistent with previous functional studies that reported mutations in the first 42 residues of TatA, particularly in the amphipathic helix region, to affect significantly Tat-dependent secretion (Hicks et al., 2005). The C-terminal domain was suggested to play an important functional role in lineage specific substrate transport (Warren et al., 2009) . The low FD observed in TatA transmembrane domain points to the conserved role of this domain, mainly responsible for the formation and structural support of the pore containing complexes for protein transport (Leake et al., 2008; Warren et al., 2009).

4.5.4 Extensive functional divergence in TatB of pathogenic bacteria

Analysis of TatB showed extensive FD affecting clades mostly containing pathogenic bacteria. For example, the top five clades in terms of the amount of amino acid sites under functional divergence included *Corynebacterium*, *Neisseria*, *Bartonella* and *Salmonella*, well-known bacteria for their pathogenic lifestyle (Supplementary information: Table S4.6). Moreover, two human pathogens (*Neisseria gonorrhoeae* and *Neisseria meningitidis*) from the Neisseria clade were the second most functionally divergent (11 amino acid sites, Supplementary information Table S4.6, amino acid sites under FD are mapped to TatB alignment, as shown in Figure S4.8). FD affecting TatB was heterogeneously distributed amongst clades and protein domains: the amphipathic helix and the C-terminal domains were the most affected. Interestingly, we observed the same three-cluster pattern as in TatA (Supplementary information: Figure S4.3), suggesting that both proteins may share similar evolutionary histories. In support of this, a fusion protein consisting of the N-terminal region of TatA and the amphipathic helix and C-terminal domain of TatB can maintain a low level of Tat-dependent secretion (Lee et al., 2002). Moreover, the secretion of a tat substrate in E. coli can still be detected after mutating TatA in a TatB knock-out strain (Blaudeck et al., 2005). These data indicate that FD of TatB may serve to determine the specificity of some Tat-dependent substrates through the modulation of the interaction of TatB with TatA, which is important to bacterial virulence (De Buck et al., 2008).

4.5.5 Differential protein transport in functionally divergent prokaryotic lineages

We identified the Sec- and Tat-dependent secretome of three halophilic archaeal species, *Halobacterium sp. NCR-1*, *Haloferax volvanii* and *Halomicrobium mukohataei* using SignalP and TatP approaches, respectively (only the data for Tat, and not for Sec, are shown here). Tat substrates in the three secretomes were grouped based on COG functional categories, which were then plotted and ranked according to their numbers of substrates within each of the categories. These numbers were statistically tested for category enrichment in secretion by comparing the observed number of secreted proteins within each category to the expected number. Expected number of secreted proteins within each of the categories was calculated by assuming that the percentage of secreted proteins is proportional to the total number of proteins within COG categories for that organism. Interestingly, in *Halobacterium sp. NCR-1*, the functional category P—annotated as Inorganic ion transport and metabolism—was significantly enriched for secreted proteins and was ranked the largest group amongst all the COG functional categories (Figure 4.2A). Similar patterns were also observed in *H. volvanii* (Figure 4.2B) and *Halomicrobium mukohataei* (Figure 4.2C) and in Sec-dependent secretomes (data not shown). It is known that halophilic archaea accumulate potassium ions to counter balance the high salt concentration in their environment (Albers et al., 2006). It is not surprising therefore to see that a large fraction of Tat-dependent substrates are related to inorganic ion transport and metabolism (Supplementary information: Table S4.7). To determine if FD in TatC from haloarchaea may be correlated with its adaptation to halophilic environments and not to other indirect causes, we also identified the Sec- and Tat-dependent secretome of three species from the other three non-halophilic lineages under FD (Figure 4.1A): *Leptospira interrogans Lai*, *Bartonella quintana* and *Rickettsia rickettsii*. Conversely to the case of haloarchaea, Tat-dependent secretome analysis of these bacteria did not rank the COG P category amongst the best represented in the secretome (Figure 4.2D-F, Supplementary information: Table S4.7).

Sulfolobus clade was ranked among the top ones regarding the enrichment for FD in TatA (Figure 4.1B). *Sulfolobus* is known to live in hot environments with temperatures ranging between 60 and 90 °C (Huber &

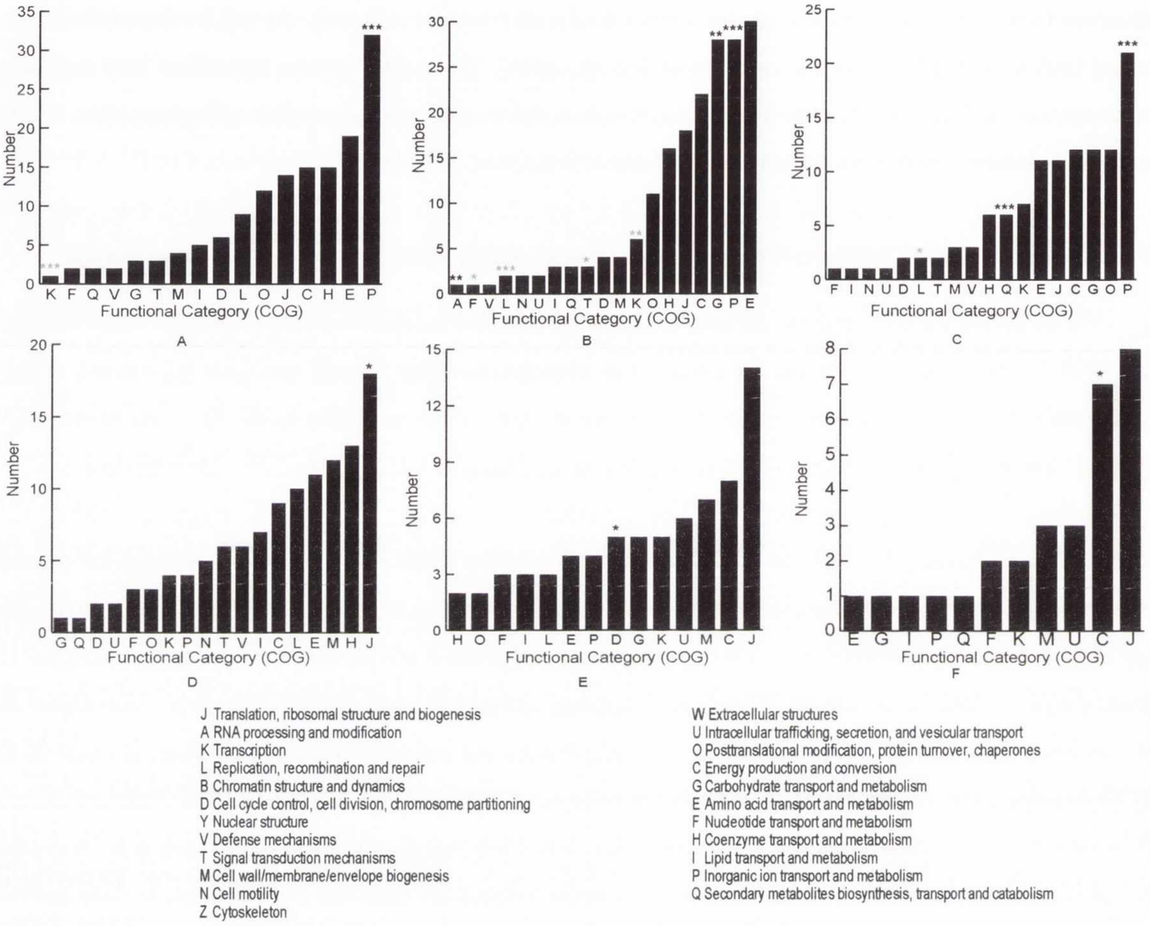


Figure 4.2: Analysis of Tat-dependent secreted proteins (secretome) of halophilic archaea and pathogenic bacteria. Substrates are grouped functionally according to the classification of proteins into the Cluster of Orhtologous Groups (COG). These functional categories were ranked and plotted according to the numbers of their substrates. (A) *Halobacterium sp. NCR-1*, (B) *Haloferax volvanii*, (C) *Halomicrobium mukohataei*, (D) *Leptospira interrogans*, (E) *Bartonella quintana*, (F) *Rickettsia rickettsii*. COG functional categories showing significant enrichment (black stars) or impoverishment (grey stars) of predicted substrate proteins are labelled by (*P<0.05; **P<0.01; ***P<0.001).

Prangishvili, 2006). Ribosomal proteins (RPs) are shown to bind zinc (Makarova et al., 2001) and mounting evidence point to their alternative extracellular location (Trost et al., 2005; Tjalsma et al., 2008; Joshi et al., 2010), in where they likely perform non-ribosomal functions (Warner & McIntosh, 2009). Interestingly, prediction of Tat-dependent substrate proteins from one of the *S. islandicus* strains (M.14.25) and subsequent COG grouping of the secreted proteins (Supplementary information: Table S4.8) identified the functional category J–that comprises proteins involved in translation, ribosomal structure and biogenesis—as the most enriched for secreted proteins from the secretome (χ^2=7.81, P<0.01). 13 ribosomal proteins (RPs) were predicted to be Tat substrate proteins, 6 out of which were predicted to contain zinc binding sites. Moreover, 4 of the 6 RPs were predicted to have multiple zinc binding sites (one has 4 zinc binding sites; three have 5 zinc binding sites) (Shu et al., 2008). To test the significance of Tat-dependent zinc containing RPs, we analysed the 64 RPs from *S. islandicus M.14.25*. We found that there are 19 RPs having at least one zinc-binding site, 4 of which have at least 5 zinc binding sites. Interestingly, 3 out of the 4 RPs are predicted to be Tat-dependent substrates (χ^2=5.89, P<0.05). The same analysis was performed in other clades to determine whether this observation was restricted to *Sulfolobus* or was general to thermophiles. Our results suggest that this pattern was unique to *Sulfolobus*. RP with binding zinc was proposed to be more stable, and therefore they are more frequently used in thermophilic bacteria (Makarova et al., 2001). In a thermophilic environment proteins are likely to go denaturation. Zn-binding proteins are more stable in such environments because binding Zn, as other metals, stabilizes proteins (Vetriani et al., 1998; Li & Solomon, 2001; Watanabe et al., 2005). Maintaining the homeostasis of cytosolic Zn(II) is critical to cell physiology as excess metal ions in the cytosol are toxic to the cell. Metallochaperones were shown to maintain homeostasis of Copper in eukaryotic cells (Robinson & Winge, 2010). However, no cytoplasmic metallochaperones for Zn(II) have been identified in any organism. A role for Zinc binding RPs in maintaining cytosolic Zn(II) homeostasis has recently been proposed (Gunasekera et al., 2009). In their model, Gunasekera and colleagues also proposed that Zinc binding RPs may participate in Zn(II) export. Interestingly, TatA was shown to have proofreading function of cofactor (Fe, Zn, etc) containing substrate proteins (Matos et al., 2008). FD of *S. islandicus* TatA may also contribute to the export of Zn(II) that is bound to Tat-dependent RPs, which can subsequently mediate the homeostasis of Zn(II) in the cytosol.

In pathogenic bacteria, COG functional category J was the largest, although not always significantly enriched, Tat-dependent substrate group with FD (Figure 4.2D-F). This was the case for all the pathogenic bacterial strains we analysed, such as *Leptospira borgpetersenii JB197* (TatC, Supplementary information: Figure S4.9A), *L. interrogans* (TatC, Figure 4.2D), *B. quintana* (TatB and TatC, Figure 4.2E), *Rickettsia prowazekii Madrid E* (TatC, Supplementary information: Figure S4.9B) and *R. rickettsii* (TatC, Figure 4.2F) (Supplementary information: Table S4.7), *Corynebacterium diphtheriae* (TatB, Supplementary information: Figure S4.9C), *Neisseria meningitides Z2491*(TatB, Supplementary information: Figure S4.9D). As shown above, zinc containing proteins are known to play vital role in bacterial physiology, including DNA polymerases, proteases and ribosomal proteins, among others. During pathogenesis, bacterial pathogen can experience zinc starvation, which can interfere with the normal function of zinc-containing proteins. A feasible solution would be using a homologous protein which does not need, or that would require very little, zinc to function properly (Panina et al., 2003). We found 3 out of 8 secreted RPs from the obligate intracellular bacterial pathogen *R. rickettsii* bearing predicted zinc binding sites, although prediction scores were weak. Among the identified Tat-dependent secreted proteins in *L. borgpetersenii* and *L. interrogans*, we found the 30S ribosomal protein S3 (RPS3). Recently, a study demonstrated that RPS3 is an active component of NF-*k*B transcription factor complex (Wan et al., 2007). NF-*k*B plays an important role in the host immune response to pathogens by regulating the expression of genes involved in different immune pathways (Vallabhapurapu & Karin, 2009). Bacterial secreted Tat substrate RPS3 can be important to interfere with the normal function of host RPS3, which would impair NF-*k*B dependent gene regulation and expression. These pathogenic bacteria may therefore evade host immune response by the Tat-dependent secretion of RPS3 and other ribosomal proteins to the immune cells. This conclusion, however, should be taken with caution and requires further investigation.

4.6 Acknowledgements

This work was supported by Science Foundation Ireland, under the Research Frontiers Program (10/RFP/Gen2685) and a grant from Ministerio de Ciencia e Innovaci6n (BFU2009-12022) to MAF. XJ is supported by Irish Research Council for Science, Engineering and Technology Government of Ireland Postgraduate Scholarship in Science, Engineering and Technology. We would like to thank both the editor and reviewers for their important contribution in helping us to improve the clarity and presentation of this manuscript. Finally, we are thankful to all our colleagues who contributed to improve the quality of the writing of this manuscript.

Chapter 5

Identifying Coevolutionary Patterns in Human Leukocyte Antigen (HLA) Molecules

5.1 Related manuscript

Jiang, X., & Fares, M. A. (2010). Identifying coevolutionary patterns in human leukocyte antigen (HLA) molecules. Evolution, 64 (5), 1429–1445.

5.2 Abstract

The antigenic peptide, major histocompatibility complex molecule (MHC; also called Human Leukocyte Antigen, HLA), coreceptor CD8 or CD4 and T-cell receptor (TCR) function as a complex to initiate effectors' mechanisms of the immune system. The tight functional and physical interaction among these molecules may have involved strong coevolution links among domains within and between proteins. Despite the importance of unraveling such dependencies to understand the arms race of host-pathogen interaction, no previous studies have aimed at achieving such an objective. Here we perform an exhaustive coevolution analysis and show that indeed such dependencies are strongly shaping the evolution and probably the function of these molecules. We identify intramolecular coevolution in HLA class I and II molecules (HLAI, HLAII) at domains important for their immune activity. Most of the amino acid sites identified to be coevolving in HLAI have been also

detected to undergo positive Darwinian selection highlighting therefore their adaptive value. We also identify coevolution among antigen binding pockets (P1-P9) and among these and TCR binding sites. Conversely to HLAI, coevolution is weaker in HLAII. Our results support that such coevolutionary patterns are due to selective pressures of host-pathogen coevolution and cooperative binding of TCRs, antigenic peptides, and CD8/CD4 to HLAI and HLAII.

5.3 Introduction

There are generally two main defence lines against microbial pathogens for jawed vertebrates, namely, the antibody mediated humoral immunity and T lymphocytes (T-cell) mediated cellular immunity. In the cellular immunity, CD8(+) cytotoxic T lymphocytes (CTL) and CD4(+) helper T cells are the two main effector cells. T cells use their membrane bound protein T-cell receptor (TCR) to detect antigens previously processed by antigen presenting cells and presented by the major histocompatibility complex (MHC) molecules, which is coordinated by coreceptors CD3, CD8 or CD4 and other costimulatory molecules (Rudolph et al., 2006). MHC (human leukocyte antigen (HLA) in humans) proteins are cell surface glycoprotein, which generally have two classes, namely, MHC class I (MHCI, HLAI) and MHC class II (MHCII, HLAII). MHCI molecule comprises a heavy chain alpha (molecular weight 44 kDa) and a light chain beta-2 microglobulin (12 kDa), with the alpha chain being anchored in the cell membrane (Bjorkman et al., 1987). Conversely, MHCII molecule comprises an alpha and a beta chain (molecular weight ~60kDa) (Brown et al., 1993) and both are anchored in the cell membrane. MHCI molecules are expressed in all nucleated cells, while MHCII are only expressed in cells of the immune system.

T cell activation involves a costimulatory signal from coreceptor CD8 or CD4, which stabilizes the interaction between TCR and peptide-MHC (pMHC) complex (Campanelli et al., 2002; Garcia et al., 1996; Wooldridge et al., 2005). Such a complex, together with other cell surface molecules, forms a tight junction between T cells and antigen-presenting cells (APCs) called immunological synapse (Davis et al., 2007), which leads to T cell recognition and activation. The structure basis of this complex has been previously explored to understand the processes involved in T cell recognition and activation (Bhasin et al., 2003; Davis et al., 1998; Deng et al., 2007; Garcia et al., 1998, 1999; Rudolph et al., 2006; Rudolph & Wilson, 2002; Wang & Reinherz, 2000, 2002). Moreover, recent biochemistry studies have characterized the binding interplay between TCR, CD4 and CD8 to pMHC and have shown that it is crucial to human TCR binding efficiency and T-cell activation (Hutchinson et al., 2003; Laugel et al., 2007; van den Berg et al., 2007). Therefore, MHC evolution is expected to be subject to functional and structural constraints from the molecules forming this complex (TCR/pMHCII/CD4 or TCR/pMHCI/CD8) and to present hence inter-domain evolutionary dependencies.

Because of its important function as a mediator of the immune response, MHC patterns of variation is an active research area due to the possible correlation between this variation and disease resistance and susceptibility (Carrington et al., 1999; Cohen, 2002; Gao et al., 2001; Hedrick, 2002; Madeleine et al., 2008; Thio et al., 2003). This correlation could be better addressed by studying the molecular evolution of vertebrate immune system (Bartl et al., 1994). For example, it has been shown that polymorphism in the alpha-3 domain of HLAI (Salter et al., 1989, 1990; Sekimata et al., 1993) as well as the beta-2 domain of HLAII (Cammarota et al., 1992; Fleury et al., 1995; Konig et al., 1992) affects the binding of coreceptor CD8 and CD4 respectively. Most effort has been invested in understanding the evolutionary history of MHC within and between different orders of mammals (Vogel et al., 1999; Yeager & Hughes, 1999). Most of these efforts however have been centred on either unveiling the relationship between amino acid variation in the antigen binding groove of MHC molecules and their antigen binding specificity (Matsumura et al., 1992) or on identifying amino acids under positive selection in MHCI (Yang & Nielsen, 2002) and MHCII molecules (Furlong & Yang, 2008) as an indication of a rapid evolution due to the host-pathogen arms race. Some modest attempts have been previously carried out to identify dependencies in HLA class II molecules (Nilsson et al., 1999). We however propose to bridge the gap of the possible dependency in the evolutionary patterns between different domains by conducting coevolutionary analyses that may yield further insights into the link of the pattern of variation in MHC molecules and the immune response (Fares & McNally, 2006). The increasing body of data available on the molecular biology and genetics of host-pathogen interactions makes it possible to use formal statistical and mathematical methods to detect such correlated molecular changes (Fares & McNally, 2006). Despite the vast number of available data for such molecules, coevolutionary analyses remain to be performed on proteins of the human immune system, among which MHC genes are the most polymorphic in jawed vertebrates. Analysis of these proteins is of paramount importance because of their implication in pathogen recognition and presentation by the host (host-pathogen coevolution) (Apanius et al., 1997; Borghans et al., 2004; De Boer et al., 2004; Hailing-Brown et al., 2008; Hedrick, 2002; Vogel et al., 1999) and tumour cell recognition (Algarra et al., 2000; Boon et al., 1994; Boon & vanderBruggen, 1996; Deng et al., 2007). Here we explore the evolutionary dependencies among MHC domains to understand the dynamics underlying MHC polymorphism and functional variation.

5.4 Materials and Methods

5.4.1 Sequence selection and alignment

Data obtained in this study is from the IMGT/HLA Database of EBI version 2.22.0 (Robinson et al., 2001). The data includes allele protein sequences from samples of world human population and it comprises world health organization (WHO) confirmed alleles of the alpha chain of HLAI molecules and the beta chain of HLAII molecules. These are believed to be responsible for the differential antigen binding properties. We included 339 (96 HLA-A, 177 HLA-B, 66 HLA-C) non-redundant complete protein sequences for HLAI alpha chain and 77 (15 HLA-DPB1, 11 HLA-DQB1, 51 HLA-DRB) non-redundant complete sequences for HLAII beta chain in our analysis. Although there are no sufficient complete protein sequences for HLA-DPA, HLA-DQA and HLA-DRA, we analyzed HLA-DQA alleles, which have 16 non-redundant complete sequences available.

Protein sequences are aligned by MUSCLE (Edgar, 2004) and CLUSTAL-X (Larkin et al., 2007; Thompson et al., 1997) and then manually edited using Jalview (Clamp et al., 2004), where only the leader sequence and gaps in the cytoplasmic domain are removed. Therefore, our actual analysis included the alpha-1, alpha-2, alpha-3, beta-1, beta-2, transmembrane and cytoplasmic domains for HLAI and HLAII alleles accordingly.

5.4.2 Intramolecular coevolutionary analysis

Due to the dependency among amino acid sites in a protein identification of their coevolutionary relationships is regarded as an important tool to determine their functional and/or structural relationships. Many methods have been devised therefore for such purpose, some based on the information theory (Gloor et al., 2005; Korber et al., 1993; Martin et al., 2005; Tillier et al., 2006; Tillier & Lui, 2003), others based on a more parametric modeling of the codependence between amino acid sites (Pollock et al., 1999; Poon et al., 2007b,a; Tuff & Darlu, 2000). Despite efforts to make these methods robust to violations of assumptions, most of them suffer from inaccuracies stemming from their inability to disentangle real coevolutionary signal from background noise (Codoner & Fares, 2008; Fares & Travers, 2006). In coevolutionary analyses the historical relationships among amino acid sites as well as stochastic amino acids covariation are the components of such noise. In spite of the many attempts to erase it, sensitivity of the majority of parametric and non-parametric methods remains to be significantly affected by this noise (Codoner & Fares, 2008; Halabi et al., 2009; Tillier & Lui, 2003). In a previous article, we developed a method that seems to outperform all the others and where historical signal is clearly distinguished from functional/structural coevolution (Fares & Travers, 2006). To identify coevolutionary patterns we used this parametric method implemented in the software CAPS version 1.0 (Fares & McNally, 2006). This method has proved to be successful in yielding meaningful results in several case studies, including those aimed at identifying coevolution of membrane proteins (Fuchs et al., 2007), HIV gp120 and gp41 proteins (Travers et al., 2007) as well as Hsp70-Hop-Hsp90 system (Travers & Fares, 2007). To estimate the probabilities and significance of the correlated evolutionary patterns among amino acid sites we used a large number of random samplings (1 million and 10 million random samples) and a small alpha value (0.001 in both MHCI and MHCII analysis) to minimize false positive rate (type I error). CAPS also implements the step-down permutational procedure as described previously (Travers & Fares, 2007; Westfall

& Young, 1993) to correct for multiple testing. The scores for the amino acid substitutions were obtained using a blocks substitution matrix (BLOSUM80) (Henikoff & Henikoff, 1992) instead of the BLOSUM62 used in previous studies due to the high similarity of our protein sequences (overall mean distance d=0.12, S.E. =0.01 for HLAI molecules, d=0.25, S.E.=0.01 for HLAII). Moreover, three dimensional structures are available for the alpha chain of HLAI (HLA-A, HLA-B, HLA-C) and the beta chain of HLAII (HLA-DQ, HLA-DR), which were later used in the three dimensional analysis of the functional and structural relationships between coevolving amino acid sites. Each allele group has its own reference sequence and crystal structure when performing the coevolution analysis. We performed group-specific analyses of coevolution using alleles HLA-A*020115(PDB ID:1HHG), HLA-B*080101(PDB ID:1AGB), HLA-C*04010101(PDB ID:1IM9), HLA-DPB1*1501, HLA-DQA1*060101(PDB ID:1UVQ), HLA-DQB1*060201(PDB ID:1UVQ) and HLA-DRB1*010101(PDB ID:1SJH) for HLA-A, HLA-B, HLA-C, HLA-DPB1, HLA-DQA1, HLA-DQB1 and HLA-DRB groups, respectively.

5.4.3 Visualisation of coevolutionary networks

We used the software Cytoscape (version 2.6.1) (Shannon et al., 2003) to visualize the coevolutionary networks identified by CAPS. Cytoscape was originally designed to visualize biomolecular interaction networks. This tool however can be used to visualize any data that describes interactions between objects. CAPS can produce four files containing information of coevolutionary networks and compensatory mutations. We used this program to generate the networks of correlation between coevolving amino acids and used the correlation coefficients generated in CAPS to determine the colouring patterns of the linking lines between nodes (amino acid residues).

5.4.4 Identifying correlated variation of amino acids properties

In order to determine the physico-chemical relationships between coevolving amino acids we also tested for the correlation in the variance of the hydrophobicity and molecular weight between the amino acids detected as coevolving in the previous analysis. The reason for choosing these properties is that they are easily measurable without much error, and they have been reported previously to be among the most important factors (the two most important) in explaining amino acids contribution to protein structure stability (Atchley et al., 2005). We determined the correlated variability in the physico-chemical properties of amino acids *i* and *j* by estimating the pairwise sequence variability in these properties (for example in hydropathy) for each of the amino acid columns in the alignments as:

$$\rho_{(H)_{ij}} = \frac{\sum\limits_{k=1}^{N} (H_x - H_y)_i (H_x - H_y)_j}{\sqrt{\sum\limits_{k=1}^{N} (H_x - H_y)_i^2 (H_x - H_y)_j^2}}$$

Here, H_x and H_y stand for the hydropathy index of amino acid states x and y, while N refers to the number of pairwise comparisons. The same calculation is applied for the case of the correlation in the variance in the molecular weight of the amino acids being compared. This analysis is a new addition to the previously published coevolution method (Fares & Travers, 2006) and yields insightful information about the nature of the coevolutionary process among amino acids. For example, we could test whether covariation between a particular pair of amino acid sites occur so that the volume of the local structure region occupied by both coevolving amino acids remains constant. We also visualized this information using the program Cytoscape.

5.4.5 Detection of conserved pathways connecting coevolving residues

Conserved amino acids throughout protein evolution (Agrawal-Gamse et al., 2009; Bozek et al., 2009) are expected to have critical effects on protein functions (Gu, 1999; Wooldridge et al., 2007). It is also known that variable sites may also be important in two ways. First, viral antigenic sites for example are often variable sites while functionally important per se because they provide the virus its ability to escape immune response while recognizing host cell receptors (Abecasis et al., 2009; Carlson et al., 2008; Keck et al., 2009; Kim & Sigalov, 2008; Rehermann, 2009; Rong et al., 2009; Travers et al., 2007; Tully & Fares, 2006; Zhong et al., 2006). Second, variable regions may be indirectly important because they fall in the vicinity of important amino acid sites and therefore their variability may dramatically affect functional sites. In the latter case, variable amino acid sites tend to coevolve to preserve the structural characteristics of the functional sites (Gloor et al., 2005). Based on these two reasons, it is expected that compensatory coevolution may occur either between amino acid sites three-dimensionally proximal or alternatively between sites apparently far apart from one another but in contact with functionally important sites. Amino acid sites proximal in the three-dimensional space (for example 4 angstroms (Å) distant from one another) are very likely to affect each other and to coevolve both structurally and functionally. However, if two coevolving residues are distant in the structure, it may be hard to tell. To determine if a compensatory relationship exists between any pairs of distantly located amino acid sites identified by our analyses as coevolving we followed an approach previously described (Tully & Fares, 2009). Briefly, for each one of the amino acid sites we draw spheres of 4Å radius. Then we estimated the Poisson average amino acid distance in the multiple sequence alignment for each of the amino acids falling within the sphere. Among the detected amino acids, we identified those showing significant conservation when compared to a distribution of 10000 amino acids randomly sampled from the protein. Once these amino acids detected, we draw new spheres around the conserved amino acids and continue with the same procedure as above until no further cases are detected. We tested whether we could connect the spheres for both coevolving amino acids by walking over the conserved amino acids of these spheres. In the case where that was possible we considered both amino acids to interact between each other indirectly. This allowed us to find a conserved pathway that connects two coevolving residues. In this analyses however, caution is required as three-dimensional protein structures used are meant to be static images of the protein once crystallized. It is possible therefore that contacting amino acids identified in these structures may not be in contact in the solution. Nonetheless, the likelihood for a pair of amino acids in contact in the crystallized protein to be far apart is very slim. This probability may increase though with increasing atomic distances among amino acids in a real structure. We also stress the fact that amino acids may establish either permanent contacts or transient contacts depending on how dramatic the conformational changes are when proteins interact with other proteins or cofactors. In either case however we would expect stronger coevolution than between never-contacting amino acid sites. Finally our results do not account for structural shifts among highly divergent sequences, although this is unlikely given the similarity levels among our amino acid sequences (below 20% divergence).

5.5 Results

We first conducted the coevolutionary analyses in each HLAI allele group and then we performed the same analysis for the three HLAII allele groups mentioned above. We finally analysed each subgroup (HLA-A, B, C; HLA-DP, DQ, DR) separately. The rationale for doing this is that group analysis would allow identifying group-specific coevolutionary patterns while the joined-group analyses would allow determining if there has been shifts in coevolutionary relationships. To gain more insights regarding the group-specificity of the coevolving pairs of amino acid sites identified, we compared the coevolutionary patterns between and among the subgroups and with the covariation patterns yielded by the combined group dataset. Figure 5.1 shows the different coevolving amino acid sites identified for the different analyses. In the HLAI proteins, we identified 17 amino acids as coevolving in the HLA-A group, 16 in the HLA-B group, 13 amino acids in the HLA-C group and 23 amino acids in the combined HLAI group (Figure 5.2) (details of the coevolution coefficients and parameters are given in the supplementary information contained in Table S5.1, Table S5.2, Table S5.3 and Table S5.4 of the supplementary information). We have also identified coevolving residues shared by different HLA alleles (for example, networks of residues coevolving in two groups such as HLA-A and B, Figure 5.1e) (Figure 5.1e to h). Combined analyses of coevolution in an alignment including sequences from all three groups shows that most of the sites were also identified in each of the groups separately (Figure 5.2 and supplementary Table

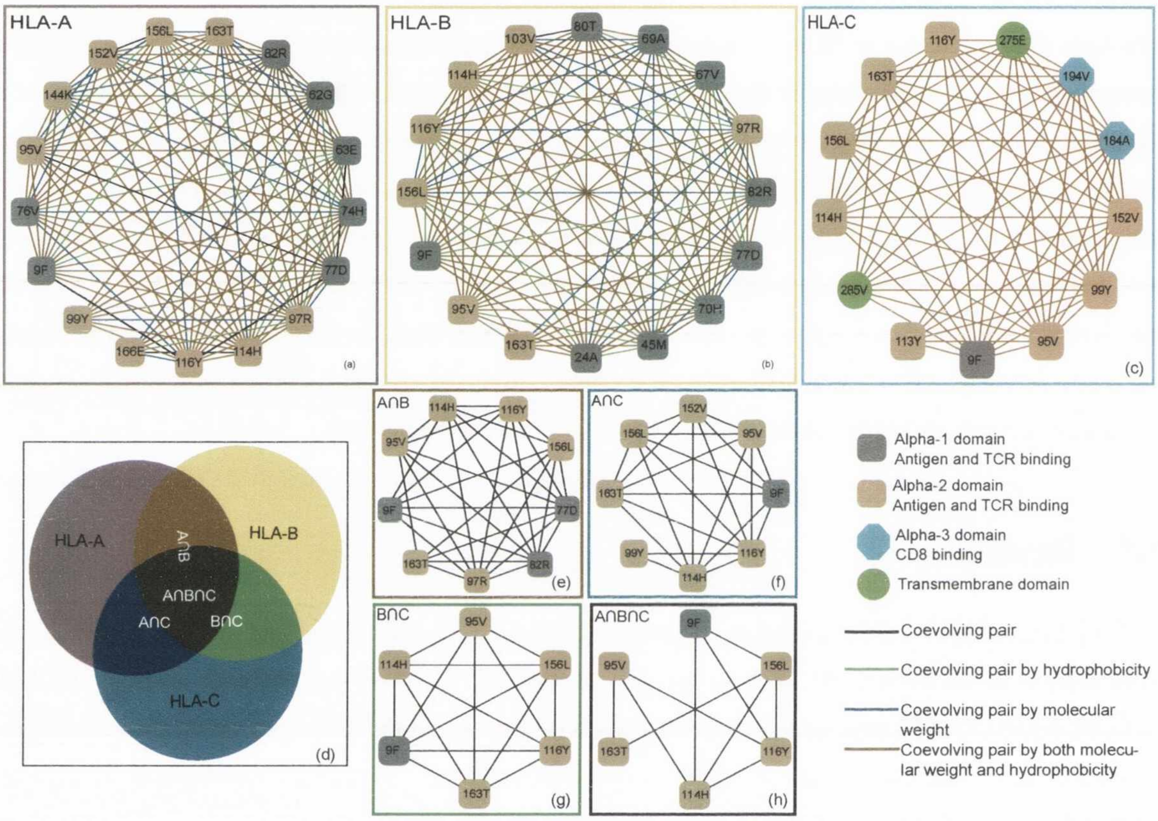


Figure 5.1: Coevolutionary networks in HLAI group alleles. Group specific coevolutionary networks for HLA-A (a), HLA-B (b), and HLA-C (c) are shown. Those pairs of coevolving residues that are shared by groups A and B (e), A and C (f), B and C (g) and A, B and C (h) are also presented. Residues in the networks are sorted clockwise in ascending order depending on the number of coevolutionary interactions each amino acid residue establishes. The domains to which these amino acid sites belong are color-coded. Nodes (amino acid residues) are connected through edges colored according to the nature and characteristics of their coevolution.

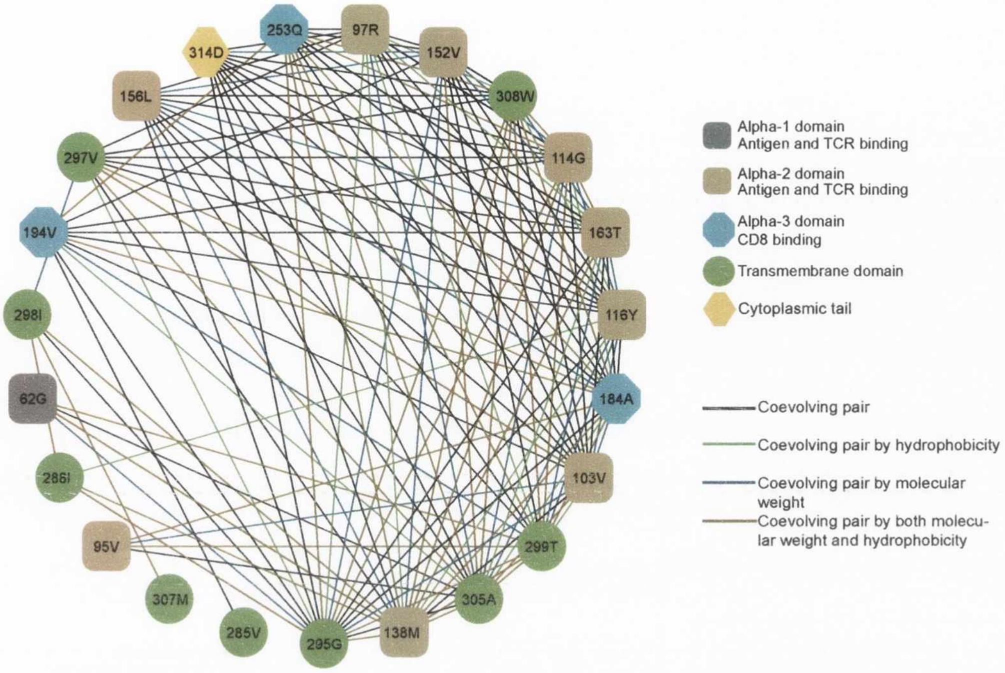


Figure 5.2: Coevolutionary networks in HLAI alleles joining groups A, B and C in one alignment. Residues in the networks are sorted clockwise in ascending order depending on the number of coevolutionary interactions each amino acid residue establishes. The domains to which these amino acid sites belong are color-coded. Nodes (amino acid residues) are connected through edges colored according to the nature and characteristics of their coevolution.

S5.4 in "HLA_class_I.xls" file). The combination of both analyses allows hence identifying lineage-specific coevolutionary events. A brief look at the figure suffices to realize that most of the mutations are falling within antigen and TCR biding domains. In the HLAII proteins we identified 5 amino acid sites as coevolving in the HLA-DP group, 9 amino acid sites in HLA-DQ and 14 amino acid sites in the HLA-DR group (Figure 5.3, supplementary information Table S5.5 to Table S5.8 available in the file "HLA_class_II"). However, we observe fewer coevolving residues and less dense coevolutionary networks compared with HLAI molecules. There are no shared coevolving networks for HLA-DPB1 HLA-DQB1 and HLA-DRB except one shared residue (71R) (Figure 5.3e). Nonetheless, there is one three-residue network shared by HLA-DPB1 and DRB (Figure 5.3f) and one two-residue network shared by HLA-DQB1 and DRB (Figure 5.3g). Conversely to the case of HLAI group, we did not identify coevolving amino acids in the combined HLAII group, possibly due to our stringent

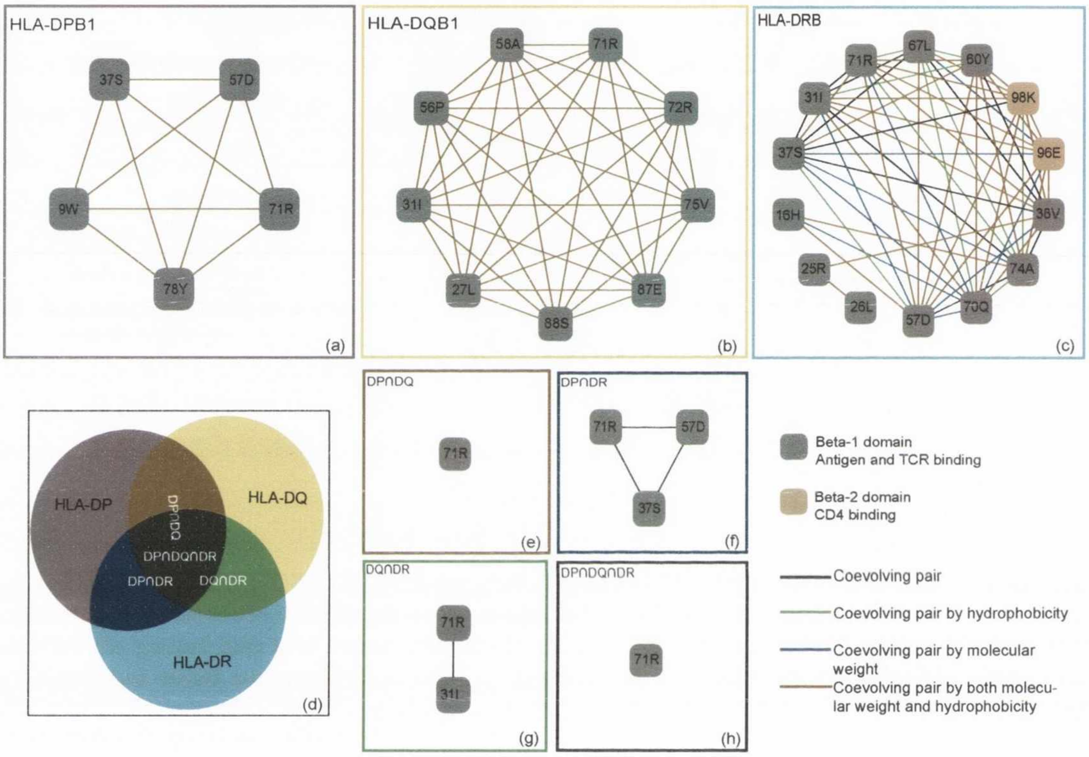


Figure 5.3: Coevolutionary networks in HLAII group alleles. Group specific coevolutionary networks for HLA-DPB1 (a), HLA-DQB1 (b), HLA-DRB(c), and combination of shared coevolutionary events (e to h) are shown. Residues in the networks are sorted clockwise in ascending order depending on the number of coevolutionary interactions each amino acid residue establishes. The domains to which these amino acid sites belong are color-coded. Nodes (amino acid residues) are connected through edges colored according to the nature and characteristics of their coevolution.

criterion to identify coevolving sites. To further understand the relationships between the amino acid sites detected as coevolving we performed analyses of correlation in two of their physico-chemical characteristics, including their covariation in molecular weight and in their hydrophobicity (ranked correlation coefficients plotted in Figure 5.4 and Figure 5.5 and detailed information is available in the supplementary tables provided in the excel file "HLA_class_I.xls") in HLAI and HLAII molecules. Doing this analysis we identified positive and negative correlation relationships among coevolving residues. Positive correlation indicates covariation of amino acids due to convergent constraints on functionally related regions of the protein. Conversely, negative correlation points to compensatory events between structurally related amino acid sites. For example, a mutation from a small amino acid to a big amino acid may require a compensatory mutation of a neighbour amino acid to a smaller amino acid in order to maintain stable volumetric characteristics of the region. We also detected amino acid sites in both groups of HLA that seemed to coevolve throughout the phylogenies, indicating their ancestral linkage. When we take these amino acid sites into account, we see that different sites present different densities of coevolution (here density refers to the percentage from the total number of coevolutionary events for a particular group represented by a particular amino acid site) in HLA-A to C (Figure 5.6A) as well as in HLA-DQ to DR (Figure 5.6B). In Figure 5.6 we provide a summary of the coevolutionary analyses of Figure 5.1 and Figure 5.3 as well as the clustering analyses to identify convergence in coevolutionary patterns between groups. The small graphs are a color-key graph where color indicates density of coevolution (going from red, meaning no coevolving sites detected, to white, meaning high percentage of coevolving sites). Also in these graphs we present a histogram (value) that indicates the frequency of the density color code in the analyses. For example, 152V in Figure 5.6A presents to coevolution in HLA-B (red color), it presents high density of coevolution in HLA-A (yellow color), which is correlated with its high frequency in the coevolution network (histogram) and presents the highest density in HLA-C (white color). We performed clustering analyses based on these data of the different HLA sequence groups using the complete linkage method where the distance between groups is estimated using the maximum Euclidean distance between members of two clusters. To calculate this distance we used a matrix score system where the elements of the matrix are the percentage of coevolution dependencies for a particular amino acid site within each group (the percentage is calculated as the total number of lines stemming from one site in the coevolutionary network of a group in Figure 5.1 and Figure 5.3 divided by the total number of lines of the group network). These data could be used to cluster the different groups of HLA assuming no convergent evolution. This clustering leads to groups HLA-A-B to the exclusion of HLA-C (Figure 5.6A) and HLA-DQ-DP to the exclusion of HLA-DR (Figure 5.6B).

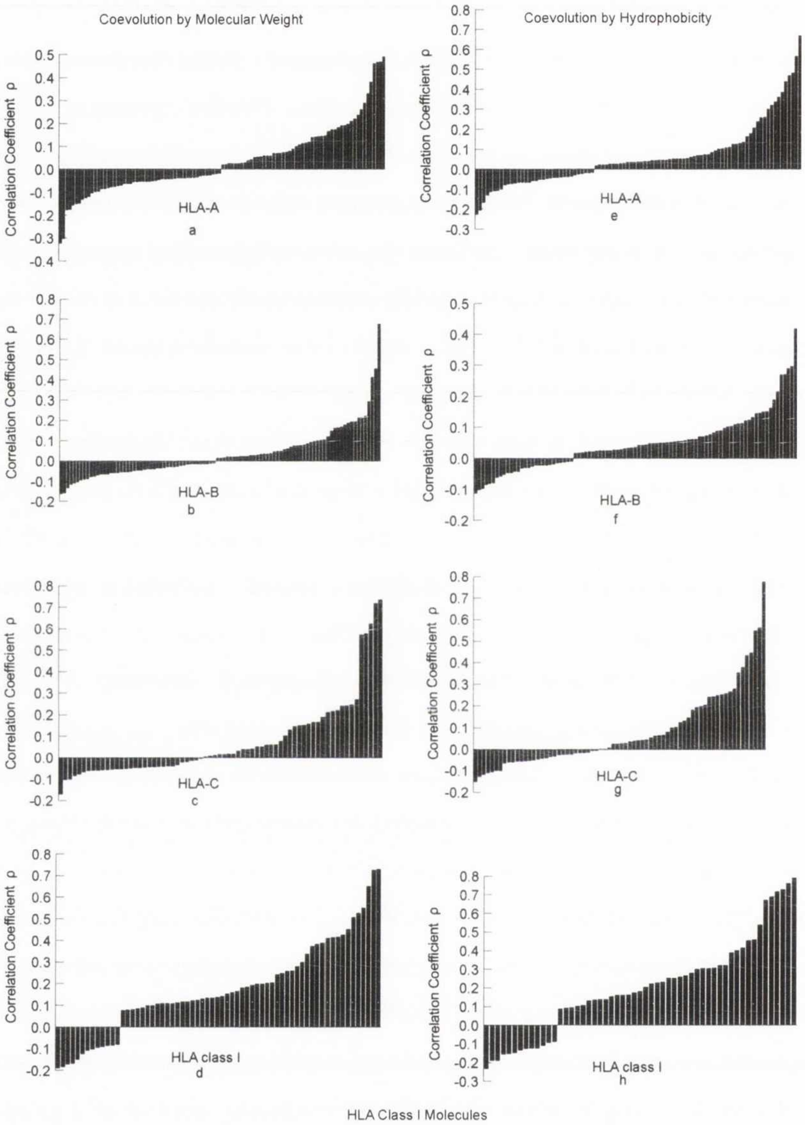


Figure 5.4: Ranked correlation coefficient for coevolving amino acid sites in HLA class I alleles. Plots of correlated variations in molecular weight (a to d) and hydrophobicity (e to h) among amino acid sites for HLA-A, HLA-B, HLA-C and the entire HLA class I allele groups respectively.

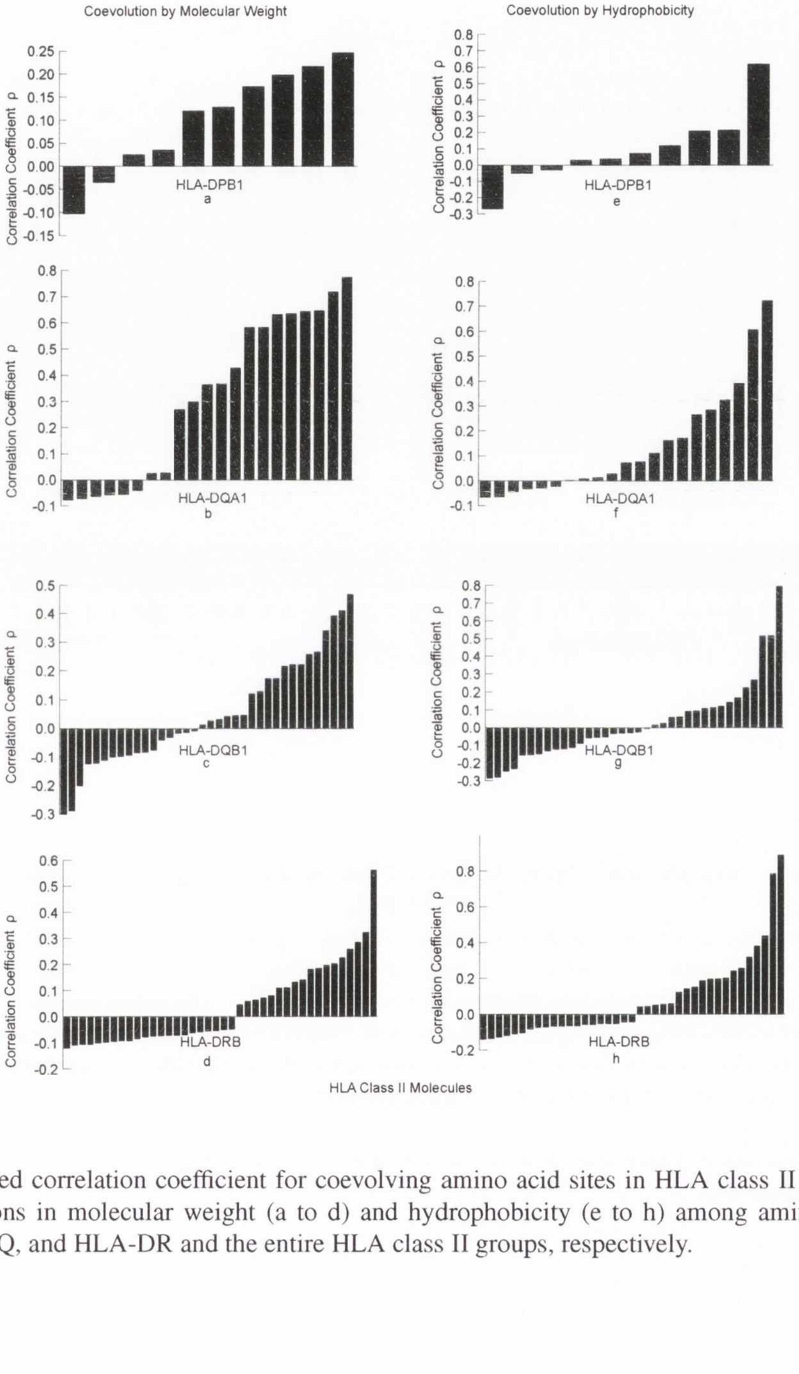


Figure 5.5: Ranked correlation coefficient for coevolving amino acid sites in HLA class II alleles. Plots of correlated variations in molecular weight (a to d) and hydrophobicity (e to h) among amino acid sites for HLA-DP, HLA-DQ, and HLA-DR and the entire HLA class II groups, respectively.

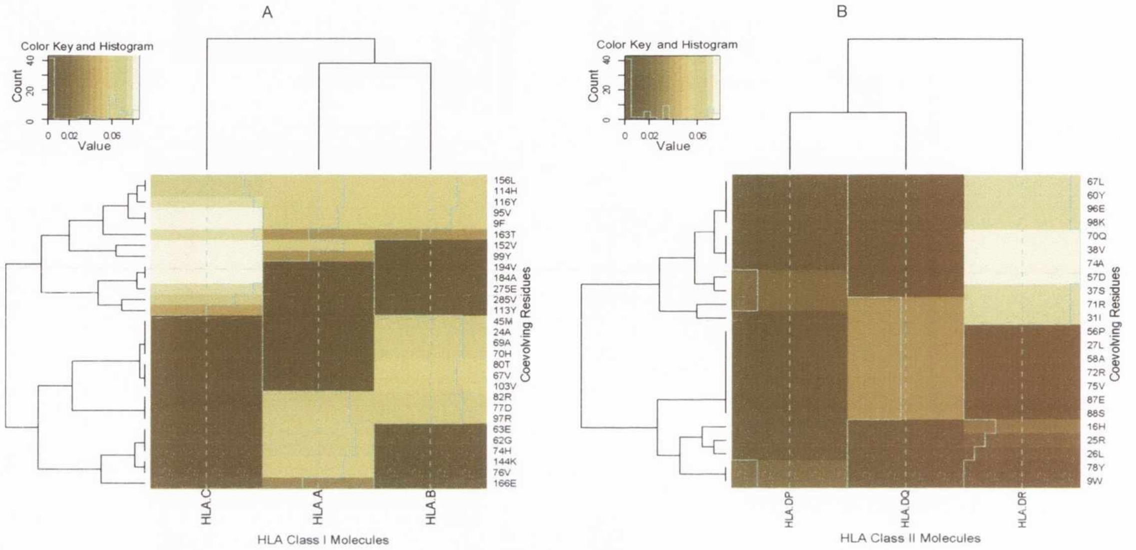


Figure 5.6: Heat map plot and hierarchical clustering of coevolving patterns of HLAI and HLAII allele groups. In each HLAI and HLAII allele group we calculated the percentage of interaction pairs for every coevolving residue as the number of coevolving events of that site with other residues divided by the total number of coevolutionary events within the molecule. We used then these data to build the heat maps where colors correspond to the amount of coevolution each amino acid presents (density). The small graphs represent the density of coevolution color-code (red meaning no coevolutionary events and white meaning maximum coevolutionary events) and the frequency of that density (value). These values were calculated for each amino acid site found to establish a coevolutionary relationship within a particular molecule and the results of frequencies were used to infer a matrix. Two matrices from two allele groups were used to calculate the maximum Euclidean distance between them, and this distance was taken to cluster alleles based on their coevolutionary density using the complete linkage method.

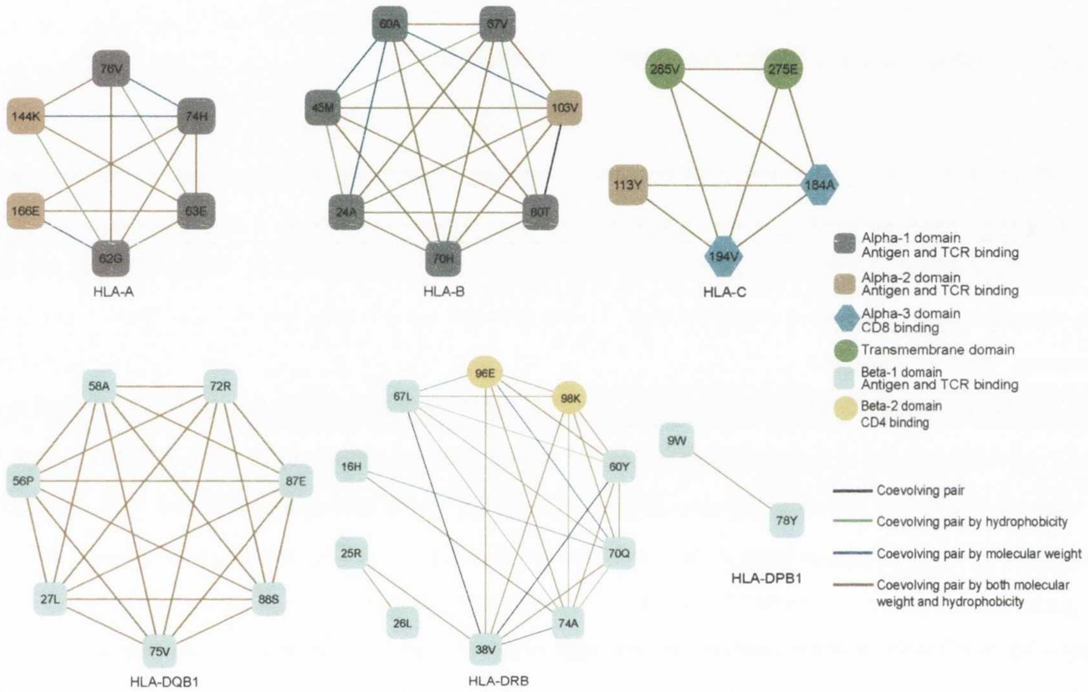

76V
144K
74H
166E
63E
62G
HLA-A
69A
67V
45M
103V
24A
80T
70H
HLA-B
265V
275E
113Y
184A
194V
HLA-C
Alpha-1 domain Antigen and TCR binding
Alpha-2 domain Antigen and TCR binding
Alpha-3 domain CD8 binding
Transmembrane domain
Beta-1 domain Antigen and TCR binding
Beta-2 domain CD4 binding
58A
72R
56P
87E
27L
88S
75V
HLA-DQB1
96E
67L
98K
16H
60Y
25R
70Q
26L
74A
38V
HLA-DRB
9W
78Y
HLA-DPB1
Coevolving pair
Coevolving pair by hydrophobicity
Coevolving pair by molecular weight
Coevolving pair by both molecular weight and hydrophobicity

Figure 5.7: Networks of coevolving residues in HLAI and HLAII allele groups within functionally important domains. This network represents a subset of the residues identified in Figure 5.1 and Figure 5.3. Residues in the networks are sorted clockwise in ascending order depending on the number of coevolutionary interactions each amino acid residue establishes. The domains to which these amino acid sites belong are color-coded. Nodes (amino acid residues) are connected through edges colored according to the nature and characteristics of their coevolution.

Correlation Type	Number of Coevolving Pairs (HLA)							
	A	B	C	Class I	DPB1	DQA1	DQB1	DRB
Hydrophobicity	94	94	70	55	10	21	36	39
Molecular weight	102	96	70	65	10	21	36	40
Hydrophobicity and molecular weight	82	80	70	41	10	21	36	30

Table 5.1: The nature of correlated variation of amino acids in HLA allele groups. This table presents the number of coevolving amino acid pairs significantly correlated by hydrophobicity, molecular weight or both for each HLA class I and class II allele groups.

5.5.1 Group specific intramolecular coevolution

We identified HLA group specific coevolutionary pairs of amino acid sites in order to determine events of rates shifts following HLA duplications. After removing coevolving network shared by HLA-A, B and C allele groups from the initial results (Figure 5.1 and Figure 5.3) we found 6, 7 and 5 amino acids to form a group of coevolution in HLA-A, HLA-B and HLA-C, respectively (Figure 5.7). While HLA-A and HLA-B are characterized by frequent coevolutionary events between amino acids belonging to alpha-1 and alpha-2 domains, HLA-C showed localized coevolution in the alpha-3 and transmembrane domains. This however may be due to the high conservation of alpha-3 domain in HLA-A and B. Interestingly, while not all the pairs were coevolving in the variation of their molecular weight and hydrophobicity in HLA-A and B groups, HLA-C showed clear covariations in physico-chemical properties of the amino acids (Figure 5.7). Similarly, we identified 2, 7 and 10 amino acids to be coevolving in HLA-DP, HLA-DQ and HLA-DR groups, respectively (Figure 5.7). Only HLA-DQ and HLA-DR presented a substantial number of amino acid sites to coevolve, while HLA-DP only showed evidence of coevolution for a single pair of amino acids (Figure 5.7). All the coevolving amino acids detected in these groups showed evidence of covariation in their physico-chemical properties (Figure 5.7). Coevolving amino acids are located in the beta-1 domain, with only two exceptions, site 96E and site 98K. Interestingly, the two sites that belong to the CD4 binding region of the beta-2 domain were coevolving with most of the amino acids identified as coevolving in the beta-1 domain.

5.5.2 Coevolution in the antigen and TCR binding domain (alpha-1, alpha-2 and beta-1)

HLA class I

Conventionally, there are seven antigen/peptide binding pockets consisting of different HLA residues in the antigen recognition site of HLA-A2 named site A to F (Saper et al., 1991). These define the antigen/peptide binding specificities of different HLA molecules (Garrett et al., 1989; Matsumura et al., 1992). Based on the structure and sequence analysis of HLA-A, HLA-B and HLA-C proteins, a less restrictive description of nine antigen/peptide binding pockets was proposed (named P1 to P9) (Chelvanayagam, 1996). The antigen that the nine antigen binding pockets bind to is a peptide with nine residues, so P1 to P9 here also correspond to the actual residue positions of the antigenic peptide. We used the information from the above study to classify the coevolving residues in the antigen recognition site (ARS) of HLAI proteins into nine antigen/peptide binding pockets (P1 to P9) (Table 5.2). In the HLA-A group, all the identified coevolving residues are in the antigen and TCR binding domain, with 14 residues being in the nine antigen binding pockets, with the exception of 82R, 144K and 166E (Figure 5.1a). For HLA-B group, of the 16 identified coevolving residues (Figure 5.1b) 14 (all but 82R and 103V) were mapped to the binding pockets, while 8 out of the 9 coevolving residues identified in HLA-C (all except 113Y) are located in these regions (Figure 5.1c). All the amino acid sites detected within the antigen binding pockets in each of the groups were found to coevolve between each other (summary of these relationships is available in Figure S5.1 of the supplementary information). We used the information derived from the crystallized structure for the binding of TCR and pMHC (Marrack et al., 2008; Rudolph et al., 2006) to identify protein binding residues that have undergone coevolution. We identified five residues out of those coevolving within HLA-A group (62G, 76V, 152V, 163T, 166E), three residues from HLA-B (69A, 70H, 163T) and two residues from HLA-C (152V and 163T) as involved in the physical interaction with a type of TCR sharing the two polypeptide chains α and β, respectively ($\alpha\beta$TCR) (Marrack et al., 2008; Rudolph et al., 2006). Moreover, one of the residues coevolving throughout the entire phylogeny (163T) was also detected to be involved in such interactions. Interestingly, most of the coevolving residues identified from ARS are in the nine antigen binding pockets (Table 5.2) and they coevolve with each other and with residues mentioned above that physically contact $\alpha\beta$TCR (Marrack et al., 2008; Rudolph et al., 2006).

To understand the role of these coevolutionary events in the preservation and cooperation in the different functions of binding between molecules it is necessary to test the selective forces underlying the fixation of mutations at these sites. Fixation of correlated mutations by adaptive evolution would point to the importance of such correlated mutations in the function or structure of the protein domains they belong to. It has been already shown that residues in the ARS of HLAI (Yang & Nielsen, 2002) and HLA-DR (Furlong & Yang, 2008) have been fixing mutations by positive selection. Interestingly, with the exception of three residues in the HLA-A group (82R, 144K and 166E), all the other amino acids identified as coevolving under our approach have been previously reported as evolving under adaptive evolution (Yang & Nielsen, 2002).

HLA class II

We found similar coevolutionary patterns in this class as compared with class I molecules. Firstly, in HLA-DPB1 molecules all the 5 coevolving residues are in the beta-1 domain (Figure 5.3a), which is also the case

Position/Pocket	Coevolving Residues		
	HLA-A	**HLA-B**	**HLA-C**
P1(A)	62G 63E 99Y 163T	163T	99Y 163T
P2(B)	9F 62G 63E 99Y 163T	9F 24A 45M 67V 70H 163T	9F 99Y 163T
P3(D)	9F 97R 99Y 114H 152V 156L 163T	9F 70H 97R 114H 156L 163T	9F 99Y 114H 152V 156L 163T
P4	62G 156L	69A 70H 156L	156L
P5	74H 97R 99Y 114H 116Y 152V 156L	69A 70H 97R 114H 116Y 156L	114H 116Y 152V 156L
P6(C)	9F 74H 97R 99Y 114H 116Y 152V 156L	9F 24S 69A 70H	9F 99Y 114H 116Y152V 156L
P7(E)	77D 97R 114H 116Y 152V 156L	77D 97R 114H 116Y 156L	114H 116Y 152V 156L
P8	76V 97R	77D 80T 97R	
P9(F)	74H 76V 95V 97R 114H 116Y 114H 116Y	70H 77D 80T 95V 97R 114H 116Y	95V 114H 116Y

Table 5.2: Functional location of coevolving residues in the HLAI protein. A subset of coevolving residues were mapped to nine antigen binding pockets (P1 to P9), where pockets A-F were defined by an early study (Saper et al., 1991). Numbers indicate the position of residues using the reference sequence HLA-A*020115(PDB ID:1HHG). Amino acids are coded using one-letter codes.

Position/Pocket	Coevolving Residues		
	HLA-DQ		HLA-DRB
	Alpha-1	Beta-1	Beta-1
P1	52R 53Q	87E	
P2			
P3			74A
P4		71R	26L 70Q 71R 74A
P5		71R	67L 70Q 71R 74A
P6			67L 71R 74A
P7		71R	67L 70Q 71R 74A
P8			60Y
P9			37S 57D 60Y

Table 5.3: Functional location of coevolving residues in the HLAII protein. A subset of coevolving residues were mapped to nine peptide binding environments (P1 to P9). The positions of amino acid residues involved in the coevolutionary networks are determined using the reference sequences HLA-DQA1*060101(PDB ID:1UVQ) and HLA-DRB1*010101(PDB ID:1SJH). Amino acids are coded using one-letter codes.

for HLA-DQB1 molecules (9 coevolving residues) (Figure 5.3b). Also, there are 12 coevolving residues out of the 14 coevolving residues identified in the beta-1 domain in HLA-DRB molecules (Figure 5.3c). Secondly, most of the coevolving pairs detected in this molecule are significantly correlated either in their hydrophobicity variance or molecular weight variance or both (details are shown in Table 5.1 and Table S5.5 to Table S5.8 of the supplementary information). Thirdly, similar peptide binding environments can also be derived from crystal structures, which are available for HLA-DQ (Baas et al., 1999) and HLA-DR (Chelvanayagam, 1997) showing the location of coevolving amino acids in the antigen-binding sites (Table 5.3). Fourthly, in the HLA-DRB group almost half of the coevolving residues (amino acid positions 37S, 57D, 67L, 71R, 74A and 96E) are identified as being under positive selection by one study (Furlong & Yang, 2008). Lastly, among these coevolving residues, we identified two residues (position 71R and 78Y in beta-1 domain) in the HLA-DPB1, one residue (position 72R in beta-1 domain) in the HLA-DQB1, three residues (position 60Y, 67L and 70Q in beta-1 domain) in the HLA-DRB, and one residue (position 68 in alpha-1 domain) in the HLA-DQA1 that physically contact the $\alpha\beta$TCR.

5.5.3 Coevolution between residues of antigen-TCR (alpha-1 and alpha-2, beta-1) and coreceptor binding domains (alpha-3, beta-2)

In HLA-C group, we identified 2 residues (184A and 194V) in the main coreceptor-binding domain, and these two residues coevolve with one residue (9F) in the alpha-1 domain and 8 residues in the alpha-2 domain (Figure 5.1c). Moreover, all the identified coevolving pairs above are significantly correlated by hydrophobicity and molecular weight. Conversely, in the HLA-DRB group we identified 2 residues in the beta-2 domain, which

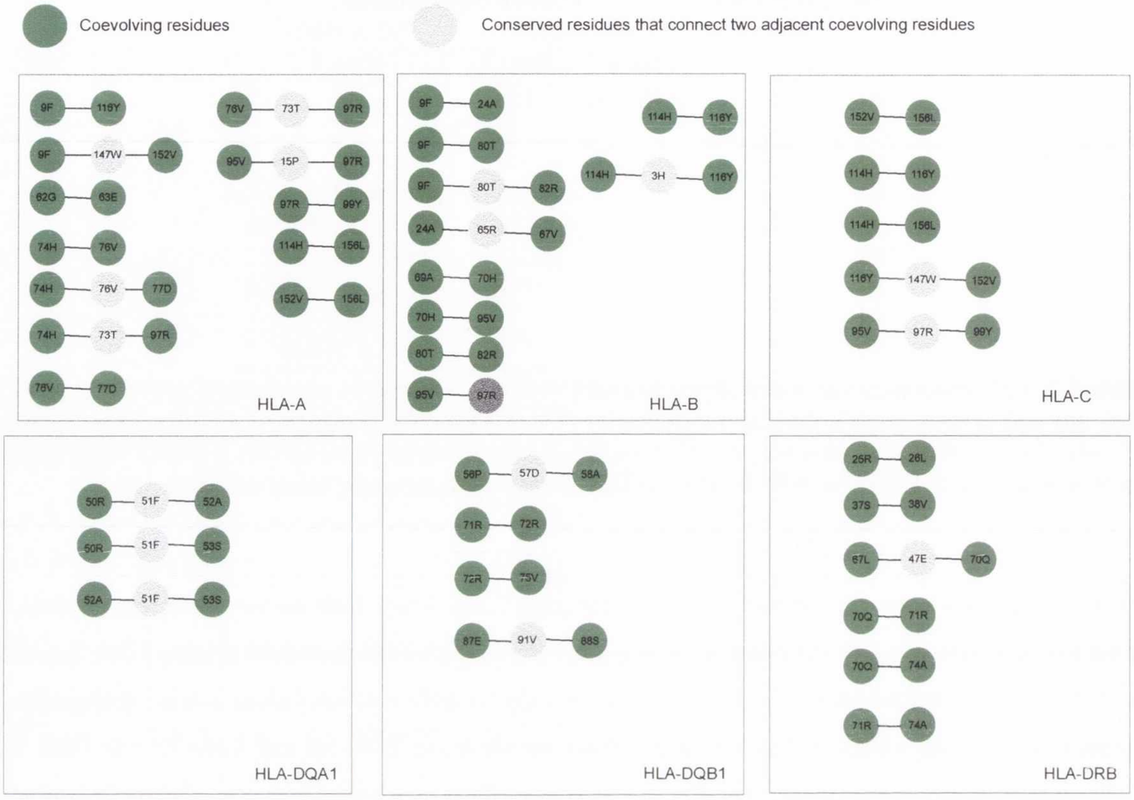


Figure 5.8: Networks of compensatory mutations in the HLA allele groups. Networks represent amino acid residues linked through edges if they were either structurally proximal in the protein or they were within 8 angstroms of a highly conserved amino acid site (light colored circles). Coevolving residue interaction networks for HLA-A, B and C are shown in the first row. Coevolving residue interaction networks for HLA-DP, DQ and DR are shown in the second row.

coevolve with 9 residues in the beta-1 domain (Figure 5.1b). Such correlated amino acid changes between functional domains (antigen-TCR and coreceptor binding) may underlie their functional importance in immune responses. In addition to these results, we also identified 2 residues from the transmembrane domain in the HLA-C group (Figure 5.1c), which coevolve with one residue in the alpha-1 domain, 7 residues in the alpha-2 domain and 2 residues in the alpha-3 domain. However there are few studies focusing on the transmembrane and cytoplasmic domains. One study has shown that these two domains are not involved in the signalling process of MHCI molecules. Yet, another earlier study has shown that mutations in the alpha chain or the beta chain of MHCII molecules can cause failure when forming some complexes and are retained in the endoplasmic reticulum (ER) (Cosson & Bonifacino, 1992).

5.5.4 Conserved amino acid pathways linking coevolving residues

We built networks (Figure 5.8) based on pathways of conserved residues (for example, a pathway is formed by the coevolving residues and the set of conserved amino acid sites interacting with coevolving residues as defined in Methods), which show how residues interact with each other. We identified 12, 10 and 5 pathways for HLA-A, B and C alleles, respectively (Figure 5.8a to c). Conversely, we identified only 3, 4 and 6 pathways for HLA-DQA1, DQB1 and DRB1, respectively (Figure 5.8d to f). In most pathways coevolving residues interact with each other. We however identified 13 pairs that are simply connected by only one residue.

5.6 Discussion

In this study, we show that coevolutionary networks on HLAI and HLAII molecules involve almost every domain in the protein, which puts forward the conclusion that there are complex forces shaping the evolution of HLA proteins. We demonstrate that there are group specific patterns as well as shared patterns of molecular coevolution among HLAI and HLAII allele groups. Remarkably, most of the coevolving residues in the ARS are under positive selection, and the polymorphic sites among antigen binding pockets coevolve with each other and with the sites physically contacting TCR and sites from the rest of the protein domains.

The residues that we identify in the coevolutionary networks on HLA-B alleles (corresponding to amino acid positions 77, 80, 114, 116 and 156) have been previously proposed as playing a critical role to the rate of progression to AIDS for HIV infected patients who carry HLA-B*35 alleles (Gao et al., 2001; McMichael et al., 1988; Nikoliczugic & Carbone, 1990; Wooldridge et al., 2007). We also identify residues positions 114, 116 and 156 as involved in the shared coevolutionary networks among HLA-A, B and C allele groups, indicating their functional interrelationship regardless the phylogenetic grouping. These residues and their corresponding coevolutionary network are conserved across all allele groups. Since changes of these residues may have drastic influence on antigen recognition of corresponding alleles, polymorphism at these sites may be associated to diseases. It become hence of special interest not only studying the polymorphism of these sites but also understanding the covariation patterns among them. It is also intriguing whether covariation networks are the key to enhance the immune response against pathogens and whether pathogens, such as HIV-1, overcome our defences when such networks are compromised.

5.6.1 Host-pathogen coevolution

Our exhaustive study of MHC protein sequences unearths complex coevolutionary patterns among different antigen binding pockets in ARS, supporting that the observed polymorphism may be the result of the host-pathogen coevolution. Moreover, there is high polymorphism across the whole MHC proteins, which include the traditionally interesting ARS or peptide-binding region (PBR) (in the alpha-1 and alpha-2 domains) and the transmembrane domain. It has been argued that MHC alleles have evolved to maintain a high degree of polymorphism in the human population. Authors have proposed two selective explanations, namely, the heterozygous advantage and the rare allele advantage (Jeffery & Bangham, 2000). The heterozygous advantage hypothesis maintains that heterozygous hosts can present more antigenic peptides from pathogens than homozygous hosts. Conversely, the rare allele advantage hypothesis argues that hosts with rare alleles can have more chance to survive in an epidemic because the hosts can respond better to the pathogen. Both these advantages may be involved in the evolution of MHC. One study demonstrates that heterozygote advantage alone fails to explain the high degree of polymorphism of the MHC (De Boer et al., 2004). Their subsequent study supports that the high polymorphism is mainly due to host-pathogen coevolution (Borghans et al., 2004).

5.6.2 MHC polymorphism affects antigen, TCR, coreceptor binding and underlying immune functions

Our study shows that polymorphic amino acid sites evolve in a correlated way forming networks that reveal complex evolutionary dynamics operating on MHC molecules. Among those coevolutionary networks, most of the coevolving residues are located in antigen binding pockets. Mutations at these regions can affect amino acid residues distantly located in the protein structure, including alpha-3 and transmembrane domains. This can also affect TCR binding through different intermediate residues in the coevolutionary networks. Presumably, our results reveal that different non-covalent antigen binding forces are orchestrated in the antigen binding pockets to respond to most pathogen antigens. The main structure destabilizing effects of pathogen-related mutations can then be compensated with correlated changes at nearby structural regions, as we show in this study.

During the last two decades, many studies demonstrated that polymorphism could indeed affect immune functions. One study has shown that polymorphism in the alpha-3 domain of HLA-A molecules affects binding of CD8 (Salter et al., 1989), and further studies have shown that residues 223-229 in the alpha-3 domain of HLA-A2 protein play a dominant role in CD8 coreceptor binding, which correlates with CTL recognition and activation (Salter et al., 1990; Sekimata et al., 1993). Later, similar patterns were observed for HLA-DR molecules (Cammarota et al., 1992; Fleury et al., 1995). Most effort has been invested to understand how single mutation can affect CTL or helper T-cell activities. One mutagenesis study demonstrated that residue 256 on the alpha-3 domain on MHC class I molecules can significantly affect the allorecognition by the T cell hybridomas (Tanabe et al., 1990), which is only 5.5 Å away from our coevolving residue 253Q in Figure 5.2. The study also shows residues 184, 193, 195, 197, 262, and 264 to have some effect on T-cell recognition. We have identified residue 184A as coevolving with other residues, while residues 193 and 195 are in the vicinity of our coevolving residue 194V (Figure 5.2). Another study shows that a Q115E substitution in the alpha-2 domain of HLA-A*0201 enhances CD8 binding by about 50% without altering TCR/pMHCI interactions (Wooldridge et al., 2007). This is a conserved residue across all allele proteins and it is bridging coevolving residues 114G and 116Y (Figure 5.2). Our coevolutionary data correlate therefore with functional analyses conducted elsewhere and show that covariation may be related to compensatory events to preserve functional and structural features of the molecule.

5.6.3 Binding affinity affects the detection of coevolutionary signals

The binding affinity between TCR and pMHC is extremely low, which can range between 1-50 μM, whereas the binding affinity between pMHC and CD4 or CD8 can be larger than 100 μM (Hutchinson et al., 2003). Moreover, the binding affinity between CD8$\alpha\alpha$ and MHC Class I A, B and C has been investigated and ranges between 90 and 220μM (Gao & Jakobsen, 2000). In our study, we found that there are no coevolving residues from the alpha-3 domain for HLA-A and B allele groups. However, there are 2 coevolving residues in alpha-3 domain in the HLA-C group. Presumably, the higher binding affinity may cause more structural constraints on the alpha-3 domain, which may explain the correlated variation in the physico-chemical properties among all the coevolving pairs identified in HLA-C allele group. Previous studies observed similar patterns in MHC class II molecules, however, the overall binding affinities are weaker than those for MHC class I molecules (Xiong et al., 2001). These binding differences may explain the weak coevolutionary signals detected in HLAII molecules by our method. Coevolutionary relationships have been identified previously using other methods (Nilsson et al., 1999). Although some of the sites identified by those authors coincide with those identified in our study, many other sites do not. The fundamental reason may be the fact that such authors did not use a formal statistical approach to identify mutations but set up a cut-off correlation value, which is generally problematic in distinguishing false from real positive and negative values.

In sum, our analyses provide a benchmark protocol to conduct more exhaustive experimental work in the MHC mediated immune response.

5.7 Conclusion

In this study, we show the molecular coevolving networks for HLAI and HLAII molecules. We identified the coevolutionary patterns within and between HLAI and HLAII allele protein groups and showed that most of the identified coevolving residues in the ARS are under strong positive selection. Our results reveal that there are correlated amino acid changes across the whole MHC proteins, from TCR-antigen binding domain to

CD8 and CD4 coreceptor binding domain to transmembrane and cytoplasmic domain. This demonstrates the complex coevolutionary dynamics on MHCI and MHCII molecules. Our results suggest that the coevolutionary patterns operating on HLAI and HLAII molecules are due to host-pathogen coevolution and cooperative binding of TCR, antigenic peptides, and CD8 and CD4 to HLA molecules. Moreover, our results point to that disruption of the optimized interactions between and among coevolving residues could lead to different levels of disease resistance and susceptibility (e.g. progression to AIDS). Insights derived from this work may also shed light on the question of why small changes in the TCR/pMHC interface in response to different altered peptide ligands can lead to drastically different biological outcomes (Rudolph et al., 2006).

5.8 Acknowledgement

This work was supported by Science Foundation Ireland, under the program of the President of Ireland Young Researcher Award to Mario A. Fares, and the IRCSET (Irish Research Council for Science, Engineering and Technology) Government of Ireland Postgraduate Scholarship in Science, Engineering and Technology to Xiaowei Jiang. We also thank the editor and two anonymous reviewers for their substantial contribution to the improvement of the quality of this manuscript.

Chapter 6

Discussion

6.1 Invited talk

Xiaowei Jiang. "Identifying coevolutionary patterns in HLA molecules: Insights into host-pathogen coevolution". Institute of theoretical physics, University of Cologne, Germany. Invited by Prof. Michael Lässig, December 1, 2010.

6.2 Evolution of bacterial secretion system and ecological adaptation

Bacterial secretion systems are a group of membrane associated proteins that are specialised in transporting cytosolically synthesised proteins to their final destinations for their proper function. These secretion systems generally differ in two main groups of bacteria: the gram-positive and gram-negative bacteria. As we introduced in the first chapter, gram-positive bacteria mostly use the Sec and Tat secretion systems, whereas gram-negative bacteria use, in addition to the Sec and Tat secretion systems, other six specialised secretion systems, the Type I-Type VI secretion systems as illustrated in Figure 1.3. Components of these specialised secretion systems span across two membranes in gram-negative bacteria and are specially involved in the export of bacterial proteins to their extracellular milieu. Interestingly, in a large scale study of bacterial proteomes (chapter 2 of this thesis) we demonstrate that bacterial secreted proteins together with proteins involved in carbohydrate transport and metabolism are amongst the most functionally divergent of the proteomes. These microbes come from diverse ecological niches and their secreted proteins under strong functional divergence seem related to bacterial capabilities to interact with the environment, hinting speculation that bacterial secreted proteins may play a fundamental role in bacterial ecological adaptations, which is consistent with previous studies. Moreover, we demonstrate that bacterial secretion systems, responsible for translocating these secreted proteins, have intermediate functional divergence, directly supporting a role for the secretion systems in ecological adaptation.

To further elucidate the role of bacterial secretion systems in the adaptation to novel ecological niches, in Chapter 3 and 4 we analysed two evolutionarily conserved secretion systems: the Sec and Tat systems. We demonstrate that functional diversification of secretion systems may condition the sets of proteins that the microbes can secrete. Furthermore, molecular coevolution between key components of the secretion system can maintain functional specificity of these proteins and also contributes to microbe's ecological adaptation. Based on the conclusions in our study, future studies may consider how other mechanisms contribute to microbe's adaptation to ecological niches in the context of protein secretion. It is conceivable that in order to adapt successfully to one or multiple ecological niches an organism must be able to i) acquire energy accordingly either heterotrophically or autotrophically or both; and ii) sense its environments either actively or passively and coordinate the cellular machineries accordingly. Certainly, all these activities require the intervention of components of the membrane, cell surface and secreted proteins to interact with its environments for acquiring metabolites, motility, sensing, defence, and so on. Arguably, translocated bacterial proteins must play important functional roles in determining bacterial life styles.

The evolution of new substrate proteins could arise by domain shuffling, acquisition of mobile genetic elements and convergent evolution, which would facilitate their adaptation to new niches (McCann & Guttman, 2008). However, little is known about how the protein secretion systems evolved (McCann & Guttman, 2008). Understanding the evolution of secretion systems can help us perceive eukaryogenesis and organelle (mitochondria and plastids) evolution (Bohnsack & Schleiff, 2010), yet this objective remains to be achieved.

Since the revealing by Günter Blobel of the secrets of protein localisation in eukaryotic cells mediated by signal peptides–the signal peptide guides the localisation of its mature protein part (Blobel, 2000), different and similar mechanisms of protein localisation have been discovered in cells of all three domains of life. Some of these mechanisms are found to be common among prokaryotic and eukaryotic cells. The striking conservation of the general Secretory (Sec) and the twin-arginine translocation pathways among cells of all three domains of life indicates that Sec- and Tat-dependent protein localisation machineries may have already emerged in their last common ancestor (Alcock et al., 2010; Bohnsack & Schleiff, 2010). Many evidence supports this. First, Sec components are found in all three domains of life, which includes the cytoplasmic membrane of bacteria and archaea and the ER membrane, chloroplast thylakoid membrane, and mitochondrial inner membrane of eukaryotic cells. Second, the Tat pathway components are also found in bacteria and archaea cytoplasmic membranes, chloroplast thylakoid membrane, and mitochondrial inner membrane of plant cells. Last, it was found that the insertion of TatC, the conserved core protein of the Tat pathway, into inner membrane is dependent on the Sec secretion system in *E. coli* (Yi et al., 2003), which may indicate that the emergence of a Sec like secretion system predates the Tat pathway. Moreover, phylogenetic analysis of the core components SecY and TatC of Sec and Tat pathways also supports current classifications of major phylogenetic groups of bacteria, archaea and eukaryota (Yen et al., 2002; Cao & Saier, 2003), and it is now generally accepted that eukaryotic organelles such as mitochondria and chloroplasts have evolved by endosymbiosis of a purple bacterium and a cyanobacterial progenitor, respectively (Schleiff & Becker, 2011). It is conceivable that endosymbiosis certainly played an important role in eukaryogenesis, which raises three interesting questions regarding protein translocation/secretion in mitochondria and plastids since they are membrane-bound organelles: (i) how did endosymbiosis affect protein translocation/secretion in the precursors of mitochondria and plastids (chloroplasts)? (ii) how did the shift in protein secretion in the precursors of mitochondria and plastids contribute to eukaryogenesis? and (iii) how did the shift in protein secretion in mitochondria and plastids contribute to eukaryotic cell function? The first two questions seem difficult to answer as we cannot find organelle genomes at that stage. However, we may find eukaryotic organelles that have characteristics of early eukaryotic evolution (Signorovitch et al., 2007; Tong et al., 2011) and find the functional patterns of protein secretion. Regarding the third question, in chapter 2 of the thesis, we showed a general trend in the bacterial evolution by the analysis of the functional divergence of 750 bacterial proteomes: bacteria have a functionally conserved core protein set, a group of protein sets with intermediate functional changes and peripheral protein sets with the most radical functional changes. Three distinctive clusters of functional categories (COG) were discovered (Figure 2.3). Although protein components belong to secretion machineries are not found to be the most functionally divergent (U), membrane proteins (M) are the most functionally divergent. This is in support of our claim that membrane proteins are the most important mediators of ecological adaptation, as membrane proteins are in the frontier of cell-environment interaction. Previous evidence suggests that secretion system has intrinsic molecular mechanisms that govern substrate protein recognition and transport (Mendel et al., 2008; Smith et al., 2005), affecting therefore protein translocation across or ingratiation into membrane. Functional diversification of secretion systems in different bacterial lineages would be a prerequisite for ecological innovation, as this would allow secretion of new substrate proteins important to particular bacterial life styles. Although organelles are no longer bacteria, would this also be true for the organelle protein secretion? For example, plant chloroplasts have two SecA/SecY systems, so would the two systems secrete two functionally different sets of proteins contributing differently to plant physiology? These interesting questions worth further investigation.

Understanding the evolution of secretion systems can also have medical benefits. Antibiotics resistant nosocomial bacterial infection is a major cause of morbidity and mortality of hospitalised patients (Rasko & Sperandio, 2010). The intense dedication of the scientific community to reviewing secretion pathways of prokaryotes and eukaryotes from different angles is testament to the broad interest of this topic, which has materialized in many recent reviews (Baron, 2010; Lee & Kessler, 2009; Yuan et al., 2010). This, in addition to the strikingly increasing pace at which antibiotic treatments fail (Walsh, 2000), has triggered a broad interest in looking for alternative therapeutic approaches based on the secretion systems. Mounting evidence suggests that bacterial secretion systems play a central role in bacterial pathogenesis (Desvaux & Hebraud, 2009; Evans et al., 2009; Hensel, 2009; McCann et al., 2009; Oldfield & Wooldridge, 2009; Thanassi et al., 2009; Turner et al., 2009; Worthington & Carbonetti, 2009), which ranges from evolutionarily conserved Sec and Tat secretion systems to more specialised secretion systems including the Type I –VI secretion systems in gram-negative bacteria. Many virulence factors are either secreted proteins or membrane proteins. Some pioneering studies have shown that targeting of specific virulence factors can indeed decrease bacterial virulence without posing immediate evolutionary pressure. Numerous anti-virulence compounds have been synthesised and are at different stages of drug development, which specifically target the Type III secretion system components and prevent its function in *Chlamydia, Shigella, Yersinia*, UPEC (Uropathogenic *E. coli*), *Salmonella* and *Pseudomonas* spp. (Rasko & Sperandio, 2010). To the best of our knowledge, no other compounds are reported to target other secretion system components in pathogenic bacteria. Identifying the molecular and evolutionary interactions taking place within Sec systems is paramount to the development of novel therapeutics against pathogenic bacteria, particularly against antibiotic resistant strains. In chapter 3 and 4, we showed how functional diversification of the Sec and Tat secretion systems contribute to ecological adaptation by comparing a large set of microbes with different lifestyles. We demonstrate that bacteria have two functionally distinctive SecA/SecY-dependent pathways, with the accessory SecA2/SecY2-dependent pathway being the most functionally divergent. In particular, we identified the most functionally important protein region in SecA–the nucleotide binding domain 1 (NBD1), to have diverged functionally and propose this domain as a potential drug target. The SecA2/SecY2-dependent pathway was shown to be a very important virulence factor of pathogenic *Streptococcus* and *Staphylococcus* species as we introduced in Chapter 1. Knocking out of SecA2/SecY2 was shown to block the secretion of relevant SraPs (serine-rich adhesin for platelets) without affecting the physiology of the bacterium, which suggests an anti-virulence strategy can be applied here to develop inhibitor drugs that specifically target NBD1 domain of SecA2. This is possible as low micromolar synthetic inhibitors were developed recently to be able to inhibit SecA function in *E. coli*. Regarding another important protein SecY, we demonstrated that the functional diversification of cytoplasmic domains of SecY contributes to posttranslational translocation and cotranslational translocation, which corresponds to SecA-dependent and ribosome-dependent translocation, respectively. A previous study demonstrated that cytoplasmic domain 6 (C6) is important for the SecA-dependent secretion (Chiba et al., 2002). Another study points to that C4 and C5 domains play an important role in ribosome-dependent cotranslational translocation (Gumbart et al., 2009). The Sec-independent Tat pathway is another important protein transport system, which plays an important role in prokaryotes ecological niche adaptation. As a particular case, we demonstrated that extensive functional diversification of duplicated TatC genes can contribute to the adaptation to extreme halophilic environment in archaea. Similarly, extensive functional divergence of pathogenic bacteria could have also contributed to their host adaptation. In particular, we showed that, in Tat-dependent secretome, pathogenic bacteria could secrete ribosomal proteins with extrarisbosomal functions (Warner & McIntosh, 2009), which can play a role in NK-*kB* related immune evasion. This is surprising although there is mounting evidence suggesting that the Tat pathway plays an important role in the pathogenesis of some bacteria. In sum, as we discussed above the emergence of Sec and Tat pathway is ancient in prokaryotes evolution, therefore, mechanisms of Sec- and Tat-dependent bacterial host invasion and evasion may have emerged earlier than other secretion systems (Type I-VI) based mechanisms. Functional molecular patterns identified here would provide invaluable insights into the design of novel therapeutics to fight bacterial infections.

6.3 MHC and host-pathogen coevolution

Major histocompatibility complex (MHC, also known as human leukocyte antigen (HLA) in humans) molecules play a central role in host immunity for several reasons. These proteins are products of germ-line encoded protein coding genes. Unlike T cell receptors (TCRs), they don't undergo any active somatic hypermutations to increase their variability. That is, their variability is germ-line encoded and heritable. Moreover, MHC molecules directly bind to antigenic peptides from pathogens processed by host immune and non-immune cells. In addition, a vital component of adaptive immune response, the T cell receptor (TCR) with stunning variability, has to function in a MHC restricted way, where TCR binds to peptide-MHC on antigen presenting cell with coreceptor CD8 or CD4 from T cell to form a complex (TCR/peptide-MHC/CD8(CD4)). Finally, based on research of last decades and genome wide association studies, we now know that individuals with certain MHC allele types have a propensity to develop underlying diseases during infection, such as, human immunodeficiency virus (HIV), hepatitis B virus (HBV) and hepatitis C virus (HCV), *Mycobacteria tuberculosis* and malaria (Blackwell et al., 2009).

It is conceivable the idea that effective host immune response against pathogens requires orchestrated action of all kinds of components of the innate and adaptive immune system at both molecular and cellular levels. The ability of the host to use these components to fight infection is regarded as immunity. Two kinds of host immunity are defined in the adaptive immune system: the antibody-mediated humoral immunity and the T cell mediated cellular immunity, which play a central role in host immune response against pathogens. One of the important contributions I made during this thesis is to reveal a general pattern in the evolution of HLA class I and class II molecules: amino acid sites from antigen binding pockets coevolve with T cell re-

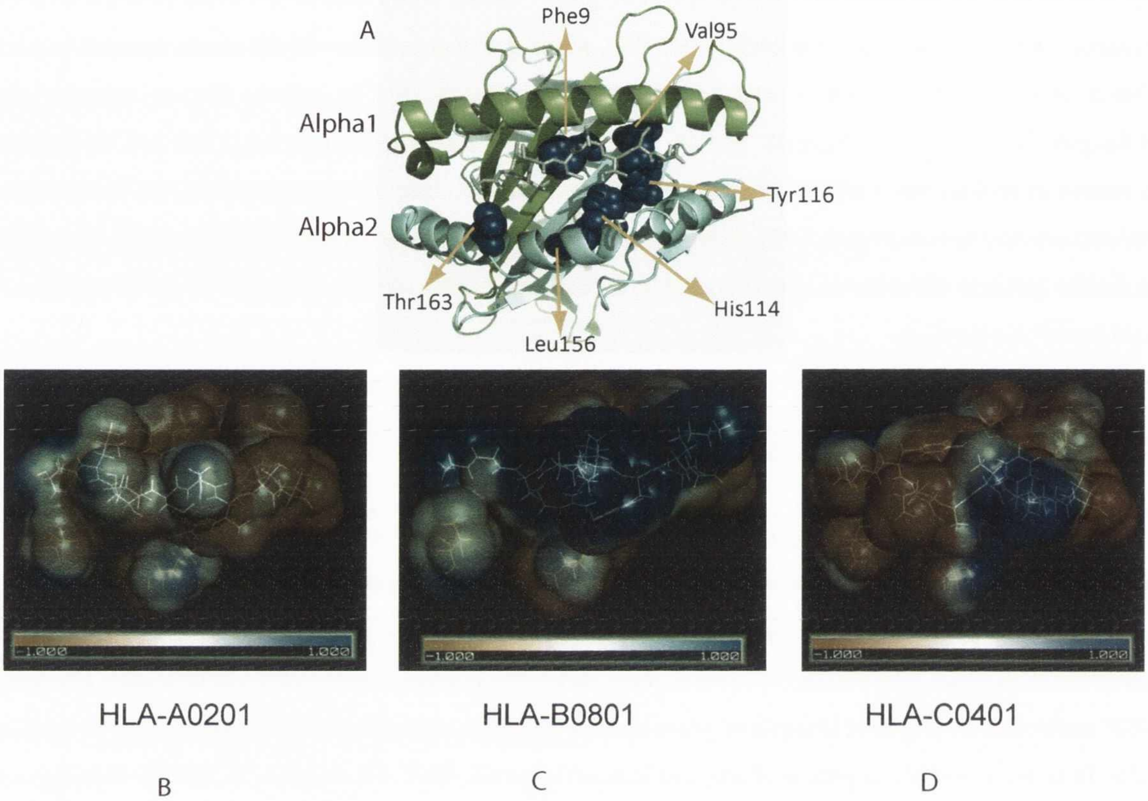


Figure 6.1: Electrostatic potential surfaces formed between a coevolving network and the corresponding antigenic peptides from the HLA-A, HLA-B and HLA-C alleles, respectively. (A) Amino acid residues from a coevolving network (Figure 5.1h) shared between the HLA-A, HLA-B and HLA-C molecules. (B) An antigenic peptide is in the middle with a white backbone in the background and bound by the coevolving residues in the network from the HLA-A0201 allele. (C) An antigenic peptide is in the middle with a white backbone and bound by coevolving residues in the network from the HLA-B0801 allele. (D) An antigenic peptide is in the middle with a white backbone and bound by coevolving residues in the network from the HLA-C0401 allele. (B)(C)(D) Amino acid residues are labelled according to the strength of the electrostatic potential calculated, where the electrostatic potential ranges from -1 (the most negative, red) to 1 (the most positive, blue). Crystal structure figures were produced by PyMOL (`http://sourceforge.net/projects/pymol/`) and the corresponding electrostatic potential surfaces were calculated by the APBS plugin in PyMOL (`http://www.poissonboltzmann.org/apbs/`).

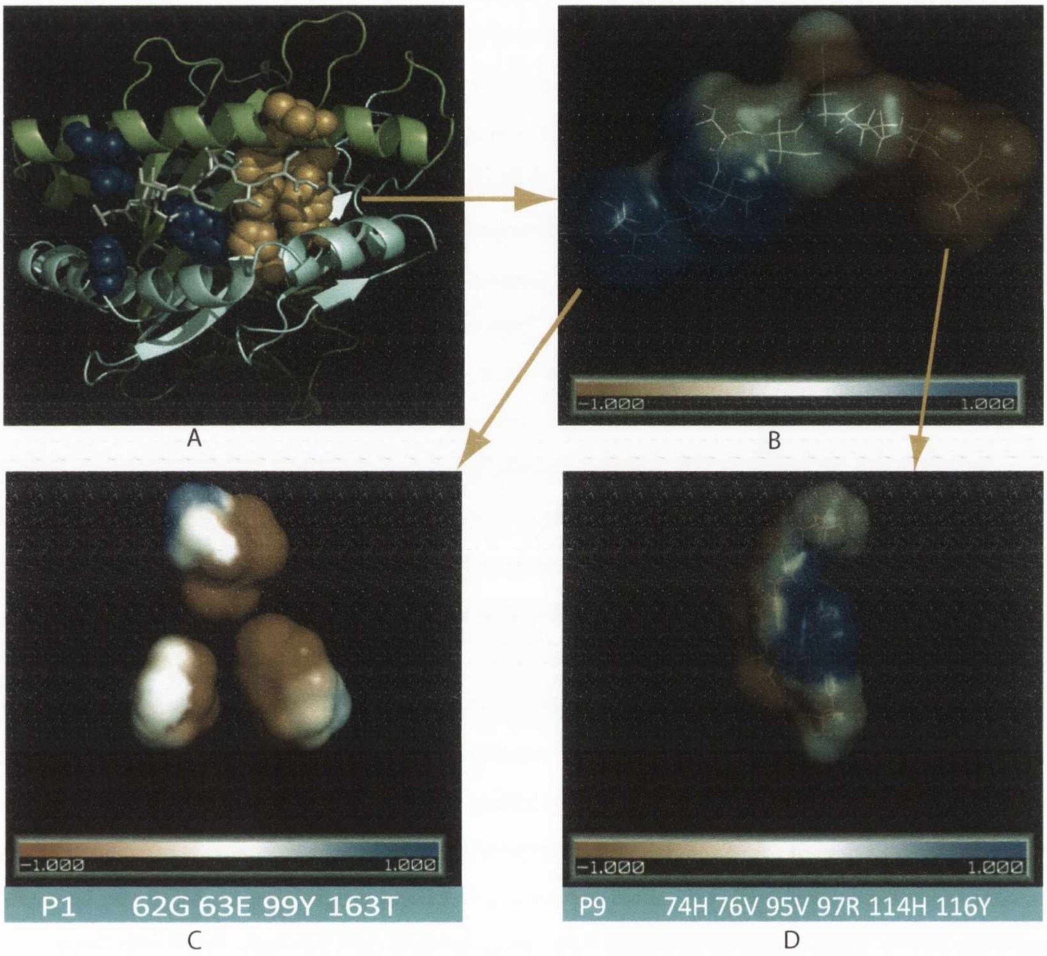


Figure 6.2: Electrostatic potential surfaces formed between an antigenic peptide and the antigen binding pocket 1 and 9. (A) An antigenic peptide is bound to the antigen binding groove of the HLA-A molecule, where amino acid residues in the antigen binding pocket 1(blue balls) and 9 (orange balls) are color coded with blue and orange, respectively. (B) Electrostatic potential was calculated for the antigenic peptide, where the N-terminal and C-terminal part of the peptide are colored with blue (positive electrostatic potential) and red (negative electrostatic potential), respectively, which corresponds to the antigen binding pocket 1 and 9, respectively. (C) Antigen binding pocket 1 is colored red in the core with negative electrostatic potential, which can firmly bind to the N-terminal part of the antigenic peptide. (D) Antigen binding pocket 9 is colored blue in the core with positive electrostatic potential, which can also firmly bind to the C-terminal part of the antigenic peptide. Crystal structure figures were produced by PyMOL 1.3 (`http://sourceforge.net/projects/pymol/`) and corresponding electrostatic potential was calculated by the APBS plug-in in PyMOL 1.3 (`http://www.poissonboltzmann.org/apbs/`).

ceptor binding sites, coreceptor CD8/CD4 binding sites and other regions of HLA molecules. We demonstrate that signals from pathogen antigenic variation can be directly sensed by host immune system and a unique complex (TCR/peptide- MHC/CD8(CD4)) is formed to aid host immune response to fight such pathogen. As shown in Figure 6.1, we identified a coevolving network (Figure 5.1h) shared by HLA-A, HLA-B and HLA-C molecules in chapter 5, in which the electrostatic binding forces are drastically different among HLA alleles when they bind to different antigenic peptides. Electrostatic potential formed between the antigen binding pockets, an antigenic peptide from pathogens and TCR binding site clearly demonstrates that disruption of the optimized interactions (stable binding formed by positive and negative electrostatic forces) between and among coevolving residues could lead to different levels of disease resistance and susceptibility (e.g. progression to AIDS) as illustrated in Figure 6.2, because these bindings determine how T cells are activated. In particular, we demonstrated that the coevolutionary patterns among HLA-A, HLA-B and HLA-C molecules are different in two ways. Firstly, HLA-A and B alleles showed similar patterns of coevolution. In the two proteins, coevolving amino acid sites are located in the antigen binding groove (alpha 1 and alpha 2 domains). Conversely, coevolving sites in HLA-C are not only located in the antigen binding groove, but also located in the coreceptor binding domain alpha 3, which indicates that coreceptor binding is more functionally important in HLA-C molecules. It was shown that the expression level of HLA-C molecules is the lowest among the three HLA types. Moreover, many studies demonstrated that coreceptor binding by CD8 is very important in cytotoxic T cell activation, which modulate sensitivity and specificity of T cell binding (Gao & Jakobsen, 2000; Rudolph et al., 2006; van den Berg et al., 2007). Therefore, our results suggest that HLA- C has the strongest sensitivity among HLA-A, B and C molecules, and higher expression level in natural population would increase their defence against infection, as it has recently been demonstrated in HIV infected patients. Authors demonstrate that association between certain HLA-C allele expression level and progression to AIDS can be established (Thomas et al., 2009). Thus, our study certainly helps us understand how our body fights infections.

Immune system is a master piece of evolution. The diversity of key immune proteins such as T cell receptor, immunoglobulin and MHC is stunning. Understanding the biological meaning of the variation in these molecules can help us appreciate how our immune system evolved to fight against pathogens and vice versa. Evidence from other studies suggest that host immune system and pathogens coevolve. However, demonstrating molecular coevolution is strikingly difficult (Woolhouse et al., 2002). It is understandable that direct demonstration of molecular coevolution between pathogens and host is extremely difficult, since at molecular level, hosts can be infected with a diverse array of pathogens and the coevolution is formed between a population of hosts and a mixed range of pathogens. Therefore, a molecular pattern that is observed in one protein of the host immune system could be attributed to the interaction of an array of proteins from many different pathogens. To test what molecular patterns are involved in the coevolution between host and pathogen, one would set up a controlled experimental system that includes only the host and its pathogen and sequence the whole genomes of both parties at defined time points. This certainly will be prohibitively expensive for sequencing large eukaryotic genomes not to mention the time points and sample size needed to achieve certain statistical power, such as mammalian genomes. Alternatively, it could be feasible to only sequence genes that are responsible for defence by the host and invasion/evasion of the pathogen if these genes are well defined for a particular type of infection model. For example, the interaction between HIV and human host has been extensively studied. The propensity of interactions between HIV proteins and host proteins can be summarised (e.g. regulation of cell cycle, apoptosis and intracellular signaling cascade, etc.) (Dickerson et al., 2010). However, up until now, no such study has been experimentally tested. Only recently, molecular coevolution was genuinely demonstrated by experimental evolution on very simple infection model–the bacteria-phage infection model, where authors showed that fast evolving genes of phage during coevolution is those that are responsible for host infection, and coevolution drives greater genetic divergence of both bacteria and phage between replicated populations (Paterson et al., 2010).

6.4 Conclusion

In this thesis, we seek to understand how bacterial pathogens adapt to their hosts from an ecological point of view. The rationale behind this objective has been that all the machineries used to invade their hosts and evade host immune responses by pathogens are also used by other non-pathogenic organisms in nature. Studying the same system from different organisms with different life styles would allow us to identify molecular patterns that are associated with host adaptation. We demonstrated that such strategy is indeed successful as we discussed in Chapter 2-5. The combined results of this project shed light on the way bacteria adapt to different environments through evolutionarily orchestrated molecular changes. Our results will set the ground and provide molecular targets for future projects that are aimed at studying host-pathogen interaction and coevolution.

Appendix A

Supplementary Information Index

Supporting files are provided in electronic format as most these files are not feasible to print on A4 papers. Therefore, all files described in the appendix are burned into an optical disk attached to this thesis. These files are also available at the website (`http://bioinf.gen.tcd.ie/~faresm/jxw_thesis_suppl_info.zip`). XLS/DOC file can be viewed by Microsoft Office Excel/Word or free software OpenOffice (download at `http://www.openoffice.org/`) . PDF file can be viewed by Adobe Reader (download at `http://get.adobe.com/reader/`).

Table	File name	Description
Table S1	Table_S1_1.xls	Currently sequenced human bacterial pathogens and the diseases they cause. Infectious diseases are annotated by ICD-10 (International Statistical Classification of Diseases and Related Health Problems 10th Revision).

Table A.1: List of supplementary information for chapter 1.

Figure/Table	File name	Description
Figure S2.1	Figure_S2_1.pdf	Two-dimensional clustering of 19 COG functional categories and 750 bacterial proteomes.
Table S2.1	Table_S2_1.doc	Enrichment of COG gene categories for functional divergence
Table S2.3	Table_S2_3.doc	Effect of bacterial lifestyle and genome size on functional divergence
Table S2.5	Table_S2_5.doc	Sites under functional divergence in VirB8.

Table A.2: List of supplementary information for chapter 2.

Figure/Table	**File name**	**Description**
Figure S3.1	Figure_S3_1.pdf	Two-dimensional clustering of protein domains and all species under functional divergence for SecA.
Figure S3.2	Figure_S3_2.pdf	Two-dimensional clustering of protein domains and all species under functional divergence for SecY.
Figure S3.3	Figure_S3_3.pdf	Networks of coevolving sites between SecA, SecB and SecY. Residues in the networks are sorted clockwise in ascending order depending on the number of coevolutionary interactions each amino acid residue establishes. Properties of amino acid sites are explained as follows. **Amino Acid**: amino acid sites are mapped to *E. coli* SecA. **Degree**: number of coevolving partners. **Domain**: protein domain in which a coevolving site locates. **Protein**: the protein where the coevolving sites locate.)
Table S3.1	Table_S3_1.xls	Species in all clades under functional divergence annotated with pathogenic status for SecA. All the clades are ranked according to their number of sites detected under functional divergence
Table S3.2	Table_S3_2.xls	Species in all clades under functional divergence annotated with pathogenic status for SecB. All the clades are ranked according to their number of sites detected under functional divergence
Table S3.3	Table_S3_3.xls	Species in all clades under functional divergence annotated with pathogenic status for SecE. All the clades are ranked according to their number of sites detected under functional divergence
Table S3.4	Table_S3_4.xls	Species in all clades under functional divergence annotated with pathogenic status for SecG. All the clades are ranked according to their number of sites detected under functional divergence
Table S3.5	Table_S3_5.xls	Species in all clades under functional divergence annotated with pathogenic status for SecY. All the clades are ranked according to their number of sites detected under functional divergence
Table S3.6	Table_S3_6.xls	Secretome analysis of 207 bacterial genomes. This table includes all COG functional categories that have been statistically tested to be over-enriched with three levels of P-values: $p<0.05$; $p<0.01$; $p<0.001$. Clades are also grouped to their closest phylogenetic group.
Table S3.7	Table_S3_7.xls	The mean amino acid atomic distance between amino acid sites that have been identified to be inhibitor binding in citation (Chen et al., 2010) and amino acid sites identified to be under FD in this study. Sites are located in SecA NBD1 domain, protein structure used for calculating the atomic distance is 3DIN chain A.

Table A.3: List of supplementary information for chapter 3.

Figure	File name	Description
Figure S4.1	Figure_S4_1.pdf	Two-dimensional clustering analysis of all clades with functional divergence for TatC.
Figure S4.2	Figure_S4_2.pdf	Two-dimensional clustering analysis of all clades with functional divergence for TatA.
Figure S4.3	Figure_S4_3.pdf	Two-dimensional clustering analysis of all clades with functional divergence for TatB.
Figure S4.4	Figure_S4_4.pdf	Amino acid sites of haloarchaea TatCo under functional divergence mapped to TatC alignment. Sites here from haloarchaea TatCo correspond to the sites in Table 4.1. Protein sequences in the alignment forming clade 1, clade 2 and outgroup are labelled accordingly.
Figure S4.5	Figure_S4_5.pdf	Amino acid sites haloarchaea of TatCt under functional divergence mapped to TatC alignment. Sites here from haloarchaea TatCt correspond to the sites in Table 4.2. Protein sequences in the alignment forming clade 1, clade 2 and outgroup are labelled accordingly.
Figure S4.6	Figure_S4_6.pdf	Sec- and Tat-dependent secretome analysis of two extreme halophilic bacteria. Substrates are grouped by COG functional category in each secretome. These functional categories are then ranked and plotted according to the numbers of their substrates. (A-B) *H. halophila SL1*, (C-D) *S. ruber DSM 13855*. COG functional categories showing significant enrichment (black stars) or impoverishment (grey stars) of predicted substrate proteins are labelled by (*: $P<0.05$; **: $P<0.01$; ***: $P<0.001$).
Figure S4.7	Figure_S4_7.pdf	Amino acid sites of *Sulfolobus* TatA under functional divergence mapped to TatA alignment. Sites here from *Sulfolobus* TatA correspond to the sites in Table S4.5. Protein sequences in the alignment forming clade 1, clade 2 and outgroup are labelled accordingly.
Figure S4.8	Figure_S4_8.pdf	Amino acid sites of *Neisseria* TatB under functional divergence mapped to TatB alignment. Sites here from *Neisseria* TatA correspond to the sites in Table S4.6. Protein sequences in the alignment forming clade 1, clade 2 and outgroup are labelled accordingly.
Figure S4.9	Figure_S4_9.pdf	Tat-dependent secretome analysis of pathogenic bacteria. Substrates are grouped by COG functional category in each secretome, these functional categories are then ranked and plotted according to the numbers of their substrates. (A) *L. borgpetersenii JB197*, (B) *R. prowazekii Madrid E*, (C) *C. diphtheriae*, (D) *N. meningitides Z2491*. COG functional categories showing significant enrichment (black stars) or impoverishment (grey stars) of predicted substrate proteins are labelled by (*: $P<0.05$; **: $P<0.01$; ***: $P<0.001$).
Figure S4.10	Figure_S4_10.pdf	The maximum likelihood phylogenetic tree for TatC, it is a high-resolution version of Figure 4.1A.
Figure S4.11	Figure_S4_11.pdf	The maximum likelihood phylogenetic tree for TatA, it is a high-resolution version of Figure 4.1B.
Figure S4.12	Figure_S4_12.pdf	The maximum likelihood phylogenetic tree for TatB, it is a high-resolution version of Figure 4.1C.

Table A.5: List of supplementary figures for chapter 4.

Table	File name	Description
Table S4.1	Table_S4_1.xls	Functional divergence analysis of TatC. We identified 204 clades with at least one amino acid sites under functional divergence. NA denotes Host or disease unknown.
Table S4.2	Table_S4_2.xls	Functional divergence analysis of TatA. We identified 145 clades with at least one amino acid sites under functional divergence. NA denotes Host or disease unknown.
Table S4.3	Table_S4_3.xls	Functional divergence analysis of TatB. We identified 119 clades with at least one amino acid sites under functional divergence. NA denotes Host or disease unknown.
Table S4.4	Table_S4_4.xls	Amino acid sites under functional divergence in TatC for haloarchaea, *Leptospira*, *Bartonella* and *Rickettsia*.
Table S4.5	Table_S4_5.xls	Amino acid sites under functional divergence in TatA for *Sulfolobus*, non-pathogenic *Mycobacteria* and pathogenic *Escherichia coli* .
Table S4.6	Table_S4_6.xls	Amino acid sites under functional divergence in TatB for *Corynebacterium*, *Neisseria*, *Bartonella* and *Salmonella*.
Table S4.7	Table_S4_7.xls	Predicted Tat-dependent substrates annotated with COG for *Bartonella henselae str. Houston-1*, *Bartonella quintana str. Toulouse*, *Rickettsia prowazekii str. Madrid E* and *Rickettsia rickettsii str. 'Sheila Smith'*.
Table S4.8	Table_S4_8.xls	Predicted Tat-dependent substrates annotated with COG for *S. islandicus M.14.25*, *E. coli O157:H7 str. EDL933* and *M. smegmatis str. MC2 155*.
Table S4.9	Table_S4_9.xls	Predicted Tat-dependent substrates annotated with COG for *N. meningitidis Z2491*, *B. henselae str. Houston-1*, *B. quintana str. Toulouse* and *S. enterica subsp. enterica serovar Typhi str. CT18*.
Table S4.10	Table_S4_10.xls	Predicted Sec-dependent substrates annotated with COG for two halophilic bacteria *H. halophila* and *S. ruber*.
Table S4.11	Table_S4_11.xls	Predicted Tat-dependent substrates annotated with COG for two halophilic bacteria *H. halophila* and *S. ruber.*

Table A.6: List of supplementary tables for chapter 4.

Figure/Table	File name	Description
Figure S5.1	Figgure_S5_1.pdf	Coevolutionary networks for coevolving residues in Table 5.1 and Table 5.3
Table S5.1	HLA_class_I.xls	Coevolution analyses of HLA-A allele sequences.
Table S5.2	HLA_class_I.xls	Coevolution analyses of HLA-B allele sequences.
Table S5.3	HLA_class_I.xls	Coevolution analyses of HLA-C allele sequences.
Table S5.4	HLA_class_I.xls	Coevolution analyses for the combined HLA class I allele sequence.
Table S5.5	HLA_class_II.xls	Coevolution analyses for the combined HLA class II dpb1 allele sequence.
Table S5.6	HLA_class_II.xls	Coevolution analyses for the combined HLA class II dqa1 allele sequence.
Table S5.7	HLA_class_II.xls	Coevolution analyses for the combined HLA class II dqb1 allele sequence.
Table S5.8	HLA_class_II.xls	Coevolution analyses for the combined HLA class II drb allele sequence.

Table A.7: List of supplementary information for chapter 5.

References

Abascal, F., Zardoya, R., & Posada, D. (2005). Prottest: selection of best-fit models of protein evolution. *Bioinformatics*, *21*(9), 2104–2105.

Abdallah, A. M., Van Pittius, N. C. G., Champion, P. A. D., Cox, J., Luirink, J., Vandenbroucke-Grauls, C., Appelmelk, B. J., & Bitter, W. (2007). Type vii secretion - mycobacteria show the way. *Nature Reviews Microbiology*, *5*, 883–891.

Abecasis, A. B., Vandamme, A.-M., & Lemey, P. (2009). Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. *Journal of Virology*, *83*(24), 12917–12924.

Agrawal-Gamse, C., Lee, F. H., Haggarty, B., Jordan, A. P., Yi, Y., Lee, B., Collman, R. G., Hoxie, J. A., Doms, R. W., & Laakso, M. M. (2009). Adaptive mutations in a human immunodeficiency virus type 1 envelope protein with a truncated v3 loop restore function by improving interactions with cd4. *Journal of Virology*, *83*(21), 11005–11015.

Alami, M., Luke, I., Deitermann, S., Eisner, G., Koch, H. G., Brunner, J., & Muller, M. (2003). Differential interactions between a twin-arginine signal peptide and its translocase in escherichia coli. *Molecular Cell*, *12*(4), 937–946.

Albers, S. V., Szabo, Z., & Driessen, A. J. M. (2006). Protein secretion in the archaea: multiple paths towards a unique cell surface. *Nature Reviews Microbiology*, *4*(7), 537–547.

Alcock, F., Clements, A., Webb, C., & Lithgow, T. (2010). Tinkering inside the organelle. *Science*, *327*(5966), 649–650.

Aldridge, C., Cain, P., & Robinson, C. (2009). Protein transport in organelles: Protein transport into and across the thylakoid membrane. *FEBS Journal*, *276*(5), 1177–1186.

Algarra, I., Cabrera, T., & Garrido, F. (2000). The hla crossroad in tumor immunology. *Human Immunology*, *61*(1), 65–73.

Altenhoff, A. M., & Dessimoz, C. (2009). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, *5*(1), e1000262–e1000262.

Anderson, B. E., & Neuman, M. A. (1997). Bartonella spp. as emerging human pathogens. *Clinical Microbiology Reviews*, *10*(2), 203–219.

Apanius, V., Penn, D., Slev, P. R., Ruff, L. R., & Potts, W. K. (1997). The nature of selection on the major histocompatibility complex. *Critical Reviews in Immunology*, *17*(2), 179–224.

Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W., & Dress, A. W. (2000). Correlations among amino acid sites in bhlh protein domains: An information theoretic analysis. *Molecular Biology and Evolution*, *17*, 164–178.

Atchley, W. R., Zhao, J., Fernandes, A. D., & Druke, T. (2005). Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(18), 6395–6400.

Auclair, S. M., Moses, J. P., Musial-Siwek, M., Kendall, D. A., Oliver, D. B., & Mukerji, I. (2010). Mapping of the signal peptide-binding domain of escherichia coli seca using forster resonance energy transfer. *Biochemistry*, *49*(4), 782–792.

Azuma, Y., & Ota, M. (2009). An evaluation of minimal cellular functions to sustain a bacterial cell. *BMC Systems Biology*, *3*, 111.

Baas, A., Gao, X. J., & Chelvanayagam, G. (1999). Peptide binding motifs and specificities for hla-dq molecules. *Immunogenetics*, *50*(1-2), 8–15.

Baba, T., Taura, T., Shimoike, T., Akiyama, Y., Yoshihisa, T., & Ito, K. (1994). A cytoplasmic domain is important for the formation of a secy-sece translocator complex. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(10), 4539–4543.

Bailey, S., Ward, D., Middleton, R., Grossmann, J. G., & Zambryski, P. C. (2006). Agrobacterium tumefaciens virb8 structure reveals potential protein-protein interaction sites. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(8), 2582–2587.

Baron, C. (2009). *Mechanistic and Structural Analysis of Type IV Secretion Systems*, chap. 6, (pp. 117–137). Norfolk, UK: Caister Academic Press.

Baron, C. (2010). Antivirulence drugs to target bacterial secretion systems. *Current Opinion in Microbiology*, *13*(1), 100–105.

Bartl, S., Baltimore, D., & Weissman, I. L. (1994). Molecular evolution of the vertebrate immune system.

Proceedings of the National Academy of Sciences of the United States of America, *91*(23), 10769–10770.

Bechtluft, P., Nouwen, N., Tans, S. J., & Driessen, A. J. (2010). Secb–a chaperone dedicated to protein translocation. *Molecular BioSystems*, *6*(4), 620–627.

Behrendt, J., Lindenstrauss, U., & Bruser, T. (2007). The tatbc complex formation suppresses a modular tatb-multimerization in escherichia coli. *FEBS Letters*, *581*(21), 4085–4090.

Bendtsen, J., Kiemer, L., Fausbøll, A., & Brunak, S. (2005a). Non-classical protein secretion in bacteria. *BMC Microbiology*, *5*(1), 58.

Bendtsen, J., Nielsen, H., Widdick, D., Palmer, T., & Brunak, S. (2005b). Prediction of twin-arginine signal peptides. *BMC Bioinformatics*, *6*(1), 167.

Bendtsen, J. D., Nielsen, H., von Heijne, G., & Brunak, S. (2004). Improved prediction of signal peptides: Signalp 3.0. *Journal of Molecular Biology*, *340*(4), 783–795.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Bensing, B. A., Lopez, J. A., & Sullam, P. A. (2004). The streptococcus gordonii surface proteins gspb and hsa mediate binding to sialylated carbohydrate epitopes on the platelet membrane glycoprotein ib alpha. *Infection and Immunity*, *72*(11), 6528–6537.

Bensing, B. A., Siboo, I. R., & Sullam, P. M. (2007). Glycine residues in the hydrophobic core of the gspb signal sequence route export toward the accessory sec pathway. *Journal of Bacteriology*, *189*(10), 3846–3854.

Bensing, B. A., & Sullam, P. M. (2002). An accessory sec locus of streptococcus gordonii is required for export of the surface protein gspb and for normal levels of binding to human platelets. *Molecular Microbiology*, *44*(4), 1081–1094.

Bensing, B. A., & Sullam, P. M. (2009). Characterization of streptococcus gordonii seca2 as a paralogue of seca. *Journal of Bacteriology*, *191*(11), 3482–3491.

Bensing, B. A., Takamatsu, D., & Sullam, P. M. (2005). Determinants of the streptococcal surface glycoprotein gspb that facilitate export by the accessory sec system. *Molecular Microbiology*, *58*(5), 1468–1481.

Berglund, A. C., Wallner, B., Elofsson, A., & Liberles, D. A. (2005). Tertiary windowing to detect positive diversifying selection. *Journal of Molecular Evolution*, *60*(4), 499–504.

Bertoldo, C., & Antranikian, G. (2006). The order thermococcales. In *The Prokaryotes*, (pp. 69–81). New York: Springer.

Bhasin, M., Singh, H., & Raghava, G. P. S. (2003). Mhcbn: a comprehensive database of mhc binding and non-binding peptides. *Bioinformatics*, *19*(5), 665–666.

Bieker, K. L., Phillips, G. J., & Silhavy, T. J. (1990). The sec and prl genes of escherichia coli. *Journal of Bioenergetics and Biomembranes*, *22*(3), 291–310.

Bielaszewska, M., Mellmann, A., Zhang, W., Kock, R., Fruth, A., Bauwens, A., Peters, G., & Karch, H. (2011). Characterisation of the escherichia coli strain associated with an outbreak of haemolytic uraemic syndrome in germany, 2011: a microbiological study. *The Lancet infectious diseases*, *11*(9), 671–676.

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L., & Wiley, D. C. (1987). Structure of the human class i histocompatibility antigen, hla-a2. *Nature*, *329*(6139), 506–512.

Blackwell, J. M., Jamieson, S. E., & Burgner, D. (2009). Hla and infectious diseases. *Clinical Microbiology Reviews*, *22*(2), 370–385.

Blaudeck, N., Kreutzenbeck, P., Muller, M., Sprenger, G. A., & Freudl, R. (2005). Isolation and characterization of bifunctional escherichia coli tata mutant proteins that allow efficient tat-dependent protein translocation in the absence of tatb. *Journal of Biological Chemistry*, *280*(5), 3426–3432.

Blobel, G. (2000). Protein targeting (nobel lecture). *Chembiochem*, *1*(2), 87–102.

Bohnsack, M. T., & Schleiff, E. (2010). The evolution of protein targeting and translocation systems. *Biochimica Et Biophysica Acta-Molecular Cell Research*, *1803*(10), 1115–1130.

Boon, T., Cerottini, J. C., Vandeneynde, B., Vanderbruggen, P., & Vanpel, A. (1994). Tumor antigens recognized by t lymphocytes. *Annual Review of Immunology*, *12*, 337–365.

Boon, T., & vanderBruggen, P. (1996). Human tumor antigens recognized by t lymphocytes. *Journal of Experimental Medicine*, *183*(3), 725–729.

Borghans, J. A. M., Beltman, J. B., & De Boer, R. J. (2004). Mhc polymorphism under host-pathogen coevolution. *Immunogenetics*, *55*(11), 732–739.

Bost, S., & Belin, D. (1997). prl mutations in the escherichia coli secg gene. *Journal of Biological Chemistry*, *272*(7), 4087–4093.

Bostina, M., Mohsin, B., Kuhlbrandt, W., & Collinson, I. (2005). Atomic model of the e-coli membrane-bound protein translocation complex secyeg. *Journal of Molecular Biology*, *352*(5), 1035–1043.

Boyer, F., Fichant, G., Berthod, J., Vandenbrouck, Y., & Attree, I. (2009). Dissecting the bacterial type vi secretion system by a genome wide in silico analysis: what can be learned from available microbial genomic resources? *BMC Genomics*, *10*, 14.

Bozek, K., Thielen, A., Sierra, S., Kaiser, R., & Lengauer, T. (2009). V3 loop sequence space analysis suggests different evolutionary patterns of ccr5- and cxcr4-tropic hiv. *PLoS One*, *4*(10), e7387.

Braunstein, M., Espinosa, B. J., Chan, J., Belisle, J. T., & Jacobs, W. R. (2003). Seca2 functions in the secretion of superoxide dismutase a and in the virulence of mycobacterium tuberculosis. *Molecular Microbiology*, *48*(2), 453–464.

Brenner, D. J., O'Connor, S. P., Hollis, D. G., Weaver, R. E., & Steigerwalt, A. G. (1991). Molecular characterization and proposal of a neotype strain for bartonella bacilliformis. *Journal of Clinical Microbiology*, *29*(7), 1299–1302.

Brown, J. H., Jardetzky, T. S., Gorga, J. C., Stern, L. J., Urban, R. G., Strominger, J. L., & Wiley, D. C. (1993). Three-dimensional structure of the human class ii histocompatibility antigen hla-dr1. *Nature*, *364*(6432), 33–39.

Bryson, K., McGuffin, L. J., Marsden, R. L., Ward, J. J., Sodhi, J. S., & Jones, D. T. (2005). Protein structure prediction servers at university college london. *Nucleic Acids Research*, *33*(suppl 2), W36–W38.

Buchanan, G., de Leeuw, E., Stanley, N. R., Wexler, M., Berks, B. C., Sargent, F., & Palmer, T. (2002). Functional complexity of the twin-arginine translocase tatc component revealed by site-directed mutagenesis. *Molecular Microbiology*, *43*(6), 1457–1470.

Cammarota, G., Scheirle, A., Takacs, B., Doran, D. M., Knorr, R., Bannwarth, W., Guardiola, J., & Sinigaglia, F. (1992). Identification of a cd4 binding site on the beta 2 domain of hla-dr molecules. *Nature*, *356*(6372), 799–801.

Campanelli, R., Palermo, B., Garbelli, S., Mantovani, S., Lucchi, P., Necker, A., Lantelme, E., & Giachino, C. (2002). Human cd8 co-receptor is strictly involved in mhc-peptide tetramer-tcr binding and t cell activation. *International Immunology*, *14*(1), 39–44.

Cao, T. B., & Saier, M. H. (2003). The general protein secretory pathway: phylogenetic analyses leading to evolutionary conclusions. *Biochimica Et Biophysica Acta-Biomembranes*, *1609*(1), 115–125.

Carlson, J. M., Brumme, Z. L., Rousseau, C. M., Brumme, C. J., Matthews, P., Kadie, C., Mullins, J. I., Walker, B. D., Harrigan, P. R., Goulder, P. J., & Heckerman, D. (2008). Phylogenetic dependency networks: inferring patterns of ctl escape and codon covariation in hiv-1 gag. *PLoS Computational Biology*, *4*(11), e1000225.

Carrington, M., Nelson, G. W., Martin, M. P., Kissner, T., Vlahov, D., Goedert, J. J., Kaslow, R., Buchbinder, S., Hoots, K., & O'Brien, S. J. (1999). Hla and hiv-1: Heterozygote advantage and b*35-cw*04 disadvantage. *Science*, *283*(5408), 1748–1752.

Casadevall, A. (2008). Evolution of intracellular pathogens. *Annual Review of Microbiology*, *62*, 19–33.

Casadevall, A., & Pirofski, L. A. (2001). Host-pathogen interactions: The attributes of virulence. *Journal of Infectious Diseases*, *184*(3), 337–344.

Cascales, E., & Christie, P. J. (2003). The versatile bacterial type iv secretion systems. *Nature Reviews Microbiology*, *1*(2), 137–149.

Caspers, M., & Freudl, R. (2008). Corynebacterium glutamicum possesses two seca homologous genes that are essential for viability. *Archives of Microbiology*, *189*(6), 605–610.

Chakrabarti, S., & Panchenko, A. (2009). Coevolution in defining the functional specificity. *Proteins: Structure, Function, and Bioinformatics*, *75*(1), 231–240.

Chelvanayagam, G. (1996). A roadmap for hla-a, hla-b, and hla-c peptide binding specificities. *Immunogenetics*, *45*(1), 15–26.

Chelvanayagam, G. (1997). A roadmap for hla-dr peptide binding specificities. *Human Immunology*, *58*(2), 61–69.

Chen, Q., Sun, B. M., Wu, H., Peng, Z. X., & Fives-Taylor, P. M. (2007). Differential roles of individual domains in selection of secretion route of a streptococcus parasanguinis serine-rich adhesin, fap1. *Journal of Bacteriology*, *189*(21), 7610–7617.

Chen, Q., Wu, H., & Fives-Taylor, P. M. (2004). Investigating the role of seca2 in secretion and glycosylation of a fimbrial adhesin in streptococcus parasanguis fw213. *Molecular Microbiology*, *53*(3), 843–856.

Chen, Q., Wu, H., Kumar, R., Peng, Z. X., & Fives-Taylor, P. M. (2006). Seca2 is distinct from seca in immunogenic specificity, subcellular distribution and requirement for membrane anchoring in streptococcus parasanguis. *FEMS Microbiology Letters*, *264*(2), 174–181.

Chen, W. X., Huang, Y. J., Gundala, S. R., Yang, H. C., Li, M. Y., Tai, P. C., & Wang, B. H. (2010). The first low ÎŒm seca inhibitors. *Bioorganic & Medicinal Chemistry*, *18*(4), 1617–1625.

Chiba, K., Mori, H., & Ito, K. (2002). Roles of the c-terminal end of secy in protein translocation and viability of escherichia coli. *Journal of Bacteriology*, *184*(8), 2243–2250.

Clamp, M., Cuff, J., Searle, S. M., & Barton, G. J. (2004). The jalview java alignment editor. *Bioinformatics*, *20*(3), 426–427.

Cleary, P., & Cheng, Q. (2006). Medically important beta-hemolytic streptococci. In *The Prokaryotes*, (pp. 108–148). New York: Springer.

Codoner, F. M., & Fares, M. A. (2008). Why should we care about molecular coevolution? *Evolutionary Bioinformatics*, *4*, 237–246.

Cohen, S. (2002). Strong positive selection and habitat-specific amino acid substitution patterns in mhc from an estuarine fish under intense pollution stress. *Molecular Biology and Evolution*, *19*(11), 1870–1880.

Conant, G., & Wolfe, K. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics*, *9*(12), 938–950.

Cosson, P., & Bonifacino, J. S. (1992). Role of transmembrane domain interactions in the assembly of class ii mhc molecules. *Science*, *258*(5082), 659–662.

Crick, F. H. (1968). The origin of the genetic code. *Journal of Molecular Biology*, *38*(3), 367–379.

Dagan, T., & Martin, W. (2006). The tree of one percent. *Genome Biology*, *7*(10), 118–118.

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray.

Das, S., & Oliver, D. B. (2011). Mapping of the seca.secy and seca.secg interfaces by site-directed in vivo photocross-linking. *Journal of Biological Chemistry*, *286*(14), 12371–12380.

Davis, M. M., Boniface, J. J., Reich, Z., Lyons, D., Hampl, J., Arden, B., & Chien, Y. H. (1998). Ligand recognition by alpha beta t cell receptors. *Annual Review of Immunology*, *16*, 523–544.

Davis, M. M., Krogsgaard, M., Huse, M., Huppa, J., Lillemeier, B. F., & Li, Q. J. (2007). T cells as a self-referential, sensory organ. *Annual Review of Immunology*, *25*, 681–695.

De Boer, R. J., Borghans, J. A. M., van Boven, M., Kesmir, C., & Weissing, F. J. (2004). Heterozygote advantage fails to explain the high degree of polymorphism of the mhc. *Immunogenetics*, *55*(11), 725–731.

De Buck, E., Lammertyn, E., & Anne, J. (2008). The importance of the twin-arginine translocation pathway for bacterial virulence. *Trends in Microbiology*, *16*(9), 442–453.

Degnan, P. H., Lazarus, A. B., & Wernegreen, J. J. (2005). Genome sequence of blochmannia pennsylvanicus indicates parallel evolutionary trends among bacterial mutualists of insects. *Genome Re-*

search, *15*(8), 1023–1033.

Dehio, C. (2001). Bartonella interactions with endothelial cells and erythrocytes. *Trends in Microbiology*, *9*(6), 279–285.

Dekker, C., de Kruijff, B., & Gros, P. (2003). Crystal structure of secb from escherichia coli. *Journal of Structural Biology*, *144*(3), 313–319.

Deng, L., Langley, R. J., Brown, P. H., Xu, G., Teng, L., Wang, Q., Gonzales, M. I., Callender, G. G., Nishimura, M. I., Topalian, S. L., & Mariuzza, R. A. (2007). Structural basis for the recognition of mutant self by a tumor-specific, mhc class ii-restricted t cell receptor. *Nature Immunology*, *8*(4), 398–408.

Desvaux, M., & Hebraud, M. (2009). *Listeria monocytogenes*, chap. 14, (pp. 313–345). Norfolk, UK: Caister Academic Press.

Dickerson, J. E., Pinney, J. W., & Robertson, D. L. (2010). The biological context of hiv-1 host interactions reveals subtle insights into a system hijack. *BMC systems biology*, *4*, 80.

Dilks, M., Gimenez, M. I., & Pohlschroder, M. (2005). Genetic and biochemical analysis of the twin-arginine translocation pathway in halophilic archaea. *Journal of Bacteriology*, *187*(23), 8104–8113.

Douglas, A. E. (1998). Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria buchnera. *Annual Review of Entomology*, *43*, 17–37.

Douville, K., Price, A., Eichler, J., Economou, A., & Wickner, W. (1995). Secyeg and seca are the stoichiometric components of preprotein translocase. *Journal of Biological Chemistry*, *270*(34), 20106–20111.

Dramsi, S., & Cossart, P. (1998). Intracellular pathogens and the actin cytoskeleton. *Annual Review of Cell and Developmental Biology*, *14*, 137–166.

Driessen, A. J. M., & Nouwen, N. (2008). Protein translocation across the bacterial cytoplasmic membrane. *Annual Review of Biochemistry*, *77*, 643–667.

du Plessis, D. J. F., Berrelkamp, G., Nouwen, N., & Driessen, A. J. M. (2009). The lateral gate of secyeg opens during protein translocation. *Journal of Biological Chemistry*, *284*(23), 15805–15814.

Dutheil, J., Gaillard, S., Bazin, E., Glemin, S., Ranwez, V., Galtier, N., & Belkhir, K. (2006). Bio++: a set of c++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics*, *7*(1), 188–188.

Dyall, S. D., Brown, M. T., & Johnson, P. J. (2004). Ancient invasions: from endosymbionts to organelles. *Science*, *304*(5668), 253–257.

Dykhuizen, D. E. (1998). Santa rosalia revisited: why are there so many species of bacteria? *Antonie Van Leeuwenhoek*, *73*(1), 25–33.

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797.

Eijlander, R. T., Jongbloed, J. D. H., & Kuipers, O. P. (2009a). Relaxed specificity of the bacillus subtilis tatadcd translocase in tat-dependent protein secretion. *Journal of Bacteriology*, *191*(1), 196–202.

Eijlander, R. T., Kolbusz, M. A., Berendsen, E. M., & Kuipers, O. P. (2009b). Effects of altered tatc proteins on protein secretion efficiency via the twin-arginine translocation pathway of bacillus subtilis. *Microbiology*, *155*, 1776–1785.

Elofsson, A., & von Heijne, G. (2007). Membrane protein structure: Prediction versus reality. *Annual Review of Biochemistry*, *76*, 125–140.

Emanuelsson, O., Brunak, S., von Heijne, G., & Nielsen, H. (2007). Locating proteins in the cell using targetp, signalp and related tools. *Nature Protocols*, *2*(4), 953–971.

Erlandson, K. J., Miller, S. B. M., Nam, Y., Osborne, A. R., Zimmer, J., & Rapoport, T. A. (2008a). A role for the two-helix finger of the seca atpase in protein translocation. *Nature*, *455*(7215), 984–987.

Erlandson, K. J., Or, E., Osborne, A. R., & Rapoport, T. A. (2008b). Analysis of polypeptide movement in the secy channel during seca-mediated protein translocation. *Journal of Biological Chemistry*, *283*(23), 15709–15715.

Eser, M., & Ehrmann, M. (2003). Seca-dependent quality control of intracellular protein localization. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(23), 13231–13234.

Evans, T. J., Perez-Mendoza, D., Monson, R. E., Stickland, H. G., & Salmond, G. P. C. (2009). *Secretion systems of the enterobacterial phytopathogen, Erwinia*, chap. 21, (pp. 479–503). Norfolk, UK: Caister Academic Press.

Fares, M. A., Elena, S. F., Ortiz, J., Moya, A., & Barrio, E. (2002a). A sliding window-based method to detect selective constraints in protein-coding genes and its application to rna viruses. *Journal of Molecular Evolution*, *55*(5), 509–521.

Fares, M. A., & McNally, D. (2006). Caps: coevolution analysis using protein sequences. *Bioinformatics*, *22*(22), 2821–2822.

Fares, M. A., Ruiz-Gonzalez, M. X., & Labrador, J. P. (2011). Protein coadaptation and the design of novel approaches to identify protein-protein interactions. *IUBMB life*, *63*(4), 264–271.

Fares, M. A., Ruiz-Gonzalez, M. X., Moya, A., Elena, S. F., & Barrio, E. (2002b). Endosymbiotic bacteria: groel buffers against deleterious mutations. *Nature*, *417*(6887), 398.

Fares, M. A., & Travers, S. A. A. (2006). A novel method for detecting intramolecular coevolution: Adding a further dimension to selective constraints analyses. *Genetics*, *173*(1), 9–23.

Fitzgerald, J. R., Foster, T. J., & Cox, D. (2006). The interaction of bacterial pathogens with platelets. *Nature Reviews Microbiology*, *4*(6), 445–457.

Flannagan, R. S., Cosio, G., & Grinstein, S. (2009). Antimicrobial mechanisms of phagocytes and bacterial evasion strategies. *Nature Reviews Microbiology*, *7*(5), 355–366.

Fleury, S., Thibodeau, J., Croteau, G., Labrecque, N., Aronson, H. E., Cantin, C., Long, E. O., & Sekaly, R. P. (1995). Hla-dr polymorphism affects the interaction with cd4. *Journal of Experimental Medicine*, *182*(3), 733–741.

Fricke, W. F., Wright, M. S., Lindell, A. H., Harkins, D. M., Baker-Austin, C., Ravel, J., & Stepanauskas, R. (2008). Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate escherichia coli sms-3-5. *Journal of Bacteriology*, *190*(20), 6779–6794.

Fuchs, A., Martin-Galiano, A. J., Kalman, M., Fleishman, S., Ben-Tal, N., & Frishman, D. (2007). Co-evolving residues in membrane proteins. *Bioinformatics*, *23*, 3312–3319.

Furlong, R. F., & Yang, Z. (2008). Diversifying and purifying selection in the peptide binding region of drb in mammals. *Journal of Molecular Evolution*, *66*(4), 384–394.

Galtier, N., & Dutheil, J. (2007). Coevolution within and between Genes. *Gene and Protein Evolution*, *3*, 1–12.

Gans, J., Wolinsky, M., & Dunbar, J. (2005). Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science*, *309*(5739), 1387–1390.

Gao, G., Rao, Z., & Bell, J. (2002). Molecular coordination of [alpha][beta] T-cell receptors and coreceptors CD8 and CD4 in their recognition of peptide-MHC ligands. *Trends in Immunology*, *23*(8), 408–413.

Gao, G. F., & Jakobsen, B. K. (2000). Molecular interactions of coreceptor cd8 and mhc class i: the molecular basis for functional coordination with the t-cell receptor. *Immunology Today*, *21*(12), 630–636.

Gao, X. J., Nelson, G. W., Karacki, P., Martin, M. P., Phair, J., Kaslow, R., Goedert, J. J., Buchbinder, S., Hoots, K., Vlahov, D., O'Brien, S. J., & Carrington, M. (2001). Effect of a single amino acid change in mhc class i molecules on the rate of progression to aids. *New England Journal of Medicine, 344*(22), 1668–1675.

Garcia, K. C., Degano, M., Pease, L. R., Huang, M. D., Peterson, P. A., Teyton, L., & Wilson, I. A. (1998). Structural basis of plasticity in t cell receptor recognition of a self peptide mhc antigen. *Science, 279*(5354), 1166–1172.

Garcia, K. C., Scott, C. A., Brunmark, A., Carbone, F. R., Peterson, P. A., Wilson, I. A., & Teyton, L. (1996). Cd8 enhances formation of stable t-cell receptor mhc class i molecule complexes. *Nature, 384*(6609), 577–581.

Garcia, K. C., Teyton, L., & Wilson, L. A. (1999). Structural basis of t cell recognition. *Annual Review of Immunology, 17*, 369–397.

Garrett, T. P. J., Saper, M. A., Bjorkman, P. J., Strominger, J. L., & Wiley, D. C. (1989). Specificity pockets for the side chains of peptide antigens in hla-aw68. *Nature, 342*(6250), 692–696.

Gascuel, O. (1997). Bionj: An improved version of the nj algorithm based on a simple model of sequence data. *Molecular Biology and Evolution, 14*(7), 685–695.

Gevers, D., Vandepoele, K., Simillion, C., & Van de Peer, Y. (2004). Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends in Microbiology, 12*, 148–154.

Gibbons, H. S., Wolschendorf, F., Abshire, M., Niederweis, M., & Braunstein, M. (2007). Identification of two mycobacterium smegmatis lipoproteins exported by a seca2-dependent pathway. *Journal of Bacteriology, 189*(14), 5090–5100.

Gil, R., Sabater-Munoz, B., Latorre, A., Silva, F. J., & Moya, A. (2002). Extreme genome reduction in buchnera spp.: toward the minimal genome needed for symbiotic life. *Proceedings of the National Academy of Sciences of the United States of America, 99*(7), 4454–4458.

Gillespie, J. (2004). *Population genetics: a concise guide*. Maryland, USA: Johns Hopkins University Press.

Gloor, G. B., Martin, L. C., Wahl, L. M., & Dunn, S. D. (2005). Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry, 44*(19), 7156–7165.

Goetz, F., Bannerman, T., & Schleifer, K.-H. (2006). The genera staphylococcus and macrococcus. In *The Prokaryotes*, (pp. 5–75). New York: Springer.

Goldman, N., & Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular Biology and Evolution, 11*(5), 725–736.

Gu, X. (1999). Statistical methods for testing functional divergence after gene duplication. *Molecular Biology and Evolution*, *16*(12), 1664–1674.

Gu, X. (2001). Mathematical modeling for functional divergence after gene duplication. *Journal of Computational Biology*, *8*(3), 221–234.

Gu, X. (2003). Functional divergence in protein (family) sequence evolution. *Genetica*, *118*(2), 133–141.

Gu, X. (2006). A simple statistical method for estimating type-ii (cluster-specific) functional divergence of protein sequences. *Molecular Biology and Evolution*, *23*(10), 1937–1945.

Guiney, D. G. (1997). Regulation of bacterial virulence gene expression by the host environment. *Journal of Clinical Investigation*, *99*(4), 565–569.

Gumbart, J., & Schulten, K. (2008). The roles of pore ring and plug in the secy protein-conducting channel. *Journal of General Physiology*, *132*(6), 709–719.

Gumbart, J., Trabuco, L. G., Schreiner, E., Villa, E., & Schulten, K. (2009). Regulation of the protein-conducting channel by a bound ribosome. *Structure*, *17*(11), 1453–1464.

Gunasekera, T., Easton, J. A., Sugerbaker Stacy, A., Klingbeil, L., & Crowder Michael, W. (2009). *Zn(II) Homeostasis in E. coli*, vol. 1012 of *ACS Symposium Series*, chap. 6, (pp. 81–95). Washington, DC: American Chemical Society.

Hailing-Brown, M., Sansom, C. E., Davies, M., Titball, R. W., & Moss, D. S. (2008). Are bacterial vaccine antigens t-cell epitope depleted? *Trends in Immunology*, *29*(8), 374–379.

Halabi, N., Rivoire, O., Leibler, S., & Ranganathan, R. (2009). Protein sectors: Evolutionary units of three-dimensional structure. *Cell*, *138*(4), 774–786.

Hanada, M., Nishiyama, K., Mizushima, S., & Tokuda, H. (1994). Reconstitution of an efficient protein translocation machinery comprising seca and the three membrane proteins, secy, sece, and secg (p12). *Journal of Biological Chemistry*, *269*(38), 23625–23631.

Hansen, A. K., & Moran, N. A. (2011). Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(7), 2849–2854.

Harper, J. R., & Silhavy, T. J. (2001). *Germ Warfare: The Mechanisms of Virulence Factor Delivery*, chap. 2, (pp. 43–74). San Diego: Academic Press.

Hayes, C. S., Aoki, S. K., & Low, D. A. (2010). Bacterial contact-dependent delivery systems. *Annual Review of Genetics*, *44*, 71–90.

Hedrick, P. W. (2002). Pathogen resistance and genetic variation at mhc loci. *Evolution*, *56*(10), 1902–1908.

Heinz, E., Tischler, P., Rattei, T., Myers, G., Wagner, M., & Horn, M. (2009). Comprehensive in silico prediction and analysis of chlamydial outer membrane proteins reflects evolution and life style of the Chlamydiae. *BMC Genomics*, *10*(1), 634.

Heinzen, R., & Samuel, J. (2006). The genus coxiella. In *The Prokaryotes*, (pp. 529–546). New York: Springer.

Henderson, B., & Oyston, P. (2003). *Bacterial evasion of host immune responses*. Cambridge, UK: Cambridge University Press.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(22), 10915–10919.

Hensel, M. (2009). *Secreted Proteins and Virulence in Salmonella enterica*, chap. 11, (pp. 239–263). Norfolk, UK: Caister Academic Press.

Hicks, M. G., Lee, P. A., Georgiou, G., Berks, B. C., & Palmer, T. (2005). Positive selection for loss-of-function tat mutations identifies critical residues required for tata activity. *Journal of Bacteriology*, *187*(8), 2920–2925.

Holzapfel, E., Eisner, G., Alami, M., Barrett, C. M. L., Buchanan, G., Luke, I., Betton, J. M., Robinson, C., Palmer, T., Moser, M., & Muller, M. (2007). The entire n-terminal half of tatc is involved in twin-arginine precursor binding. *Biochemistry*, *46*(10), 2892–2898.

Hou, J. M., D'Lima, N. G., Rigel, N. W., Gibbons, H. S., McCann, J. R., Braunstein, M., & Teschke, C. M. (2008). Atpase activity of mycobacterium tuberculosis seca1 and seca2 proteins and its importance for seca2 function in macrophages. *Journal of Bacteriology*, *190*(14), 4880–4887.

Hou, J. M., D'Lima, N. G., Rigel, N. W., Gibbons, H. S., McCann, J. R., Braunstein, M., & Teschke, C. M. (2009). Atpase activity of mycobacterium tuberculosis seca1 and seca2 proteins and its importance for seca2 function in macrophages (vol 190, pg 4880, 2008). *Journal of Bacteriology*, *191*(12), 4051–4051.

Huber, H., & Prangishvili, D. (2006). Sulfolobales. In *The Prokaryotes*, (pp. 23–51). New York: Springer.

Hunt, J. F., Weinkauf, S., Henry, L., Fak, J. J., McNicholas, P., Oliver, D. B., & Deisenhofer, J. (2002). Nucleotide control of interdomain interactions in the conformational reaction cycle of seca. *Science*, *297*(5589), 2018–2026.

Hutchinson, S. L., Wooldridge, L., Tafuro, S., Laugel, B., Glick, M., Boulter, J. M., Jakobsen, B. K., Price, D. A., & Sewell, A. K. (2003). The cd8 t cell coreceptor exhibits disproportionate biological activity at extremely low binding affinities. *Journal of Biological Chemistry*, *278*(27), 24285–24293.

Ihler, G. M. (1996). Bartonella bacilliformis: dangerous pathogen slowly emerging from deep background. *FEMS Microbiology Letters*, *144*(1), 1–11.

Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, *11*(2), 97–108.

Irihimovitch, V., & Eichler, J. (2003). Post-translational secretion of fusion proteins in the halophilic archaea haloferax volcanii. *Journal of Biological Chemistry*, *278*(15), 12881–12887.

Ito, K., & Mori, H. (2009). *The Sec Protein Secretion System*, chap. 1, (pp. 3–22). Norfolk, UK: Caister Academic Press.

Jeffery, K. J. M., & Bangham, C. R. M. (2000). Do infectious diseases drive mhc diversity? *Microbes and Infection*, *2*(11), 1335–1341.

Jenewein, S., Holland, I. B., & Schmitt, L. (2009). *Type I Bacterial Secretion Systems*, chap. 3, (pp. 45–65). Norfolk, UK: Caister Academic Press.

Jiang, X., & Fares, M. A. (2010). Identifying coevolutionary patterns in human leukocyte antigen (hla) molecules. *Evolution*, *64*(5), 1429–1445.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences: CABIOS*, *8*(3), 275–282.

Jongbloed, J. D. H., Martin, U., Antelmann, H., Hecker, M., Tjalsma, H., Venema, G., Bron, S., van Dijl, J. M., & Muller, J. (2000). Tatc is a specificity determinant for protein secretion via the twin-arginine translocation pathway. *Journal of Biological Chemistry*, *275*(52), 41350–41357.

Joshi, M. V., Mann, S. G., Antelmann, H., Widdick, D. A., Fyans, J. K., Chandra, G., Hutchings, M. I., Toth, I., Hecker, M., Loria, R., & Palmer, T. (2010). The twin arginine protein transport pathway exports multiple virulence proteins in the plant pathogen streptomyces scabies. *Molecular Microbiology*, *77*(1), 252–271.

Jukes, T., & Cantor, C. (1969). Evolution of protein molecules. In H. Munro (Ed.) *Mammalian protein metabolism*, (pp. 21–123). New York: Academic Press.

Kampinga, H. H., Dynlacht, J. R., & Dikomey, E. (2004). Mechanism of radiosensitization by hyperthermia (> or = 43 degrees c) as derived from studies with dna repair defective mutant cell lines. *International Journal of Hyperthermia*, *20*(2), 131–9.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., & Yamanishi, Y. (2008). Kegg for linking genomes to life and the environment. *Nucleic Acids Research*, *36*, D480–D484.

Kawashima, Y., Pfafferott, K., Frater, J., Matthews, P., Payne, R., Addo, M., Gatanaga, H., Fujiwara, M., Hachiya, A., & Koizumi, H. (2009). Adaptation of hiv-1 to human leukocyte antigen class i. *Nature*, *458*(7238), 641–645.

Keck, Z. Y., Li, S. H., Xia, J., von Hahn, T., Balfe, P., McKeating, J. A., Witteveldt, J., Patel, A. H., Alter, H., Rice, C. M., & Foung, S. K. (2009). Mutations in hepatitis c virus e2 located outside the cd81 binding sites lead to escape from broadly neutralizing antibodies but compromise virus infectivity. *Journal of Virology*, *83*(12), 6149–6160.

Keramisanou, D., Biris, N., Gelis, I., Sianidis, G., Karamanou, S., Economou, A., & Kalodimos, C. G. (2006). Disorder-order folding transitions underlie catalysis in the helicase motor of seca. *Nature Structural & Molecular Biology*, *13*(7), 594–602.

Kim, W. M., & Sigalov, A. B. (2008). Viral pathogenesis, modulation of immune receptor signaling and treatment. *Advances in Experimental Medicine and Biology*, *640*, 325–349.

Kimura, M. (1983). *The neutral theory of molecular evolution*. London: Cambridge University Press.

Kindt, T. J., Goldsby, R. A., & Osborne, B. A. (2007). *Kuby Immunology*, (6 ed.). New York: W.H. Freeman.

Klein, J. (1986). *Natural history of the major histocompatibility complex*. New York ; Chichester: Wiley.

Konig, R., Huang, L. Y., & Germain, R. N. (1992). Mhc class ii interaction with cd4 mediated by a region analogous to the mhc class i binding site for cd8. *Nature*, *356*(6372), 796–798.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, *39*(1), 309–338.

Korber, B. T., Farber, R. M., Wolpert, D. H., & Lapedes, A. S. (1993). Covariation of mutations in the v3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(15), 7176–7180.

Kregel, K. C. (2002). Heat shock proteins: modifying factors in physiological stress responses and acquired thermotolerance. *Journal of Applied Physiology*, *92*(5), 2177–2186.

Kreutzenbeck, P., Kroger, C., Lausberg, F., Blaudeck, N., Sprenger, G. A., & Freudl, R. (2007). Escherichia coli twin arginine (tat) mutant translocases possessing relaxed signal peptide recognition specificities. *Journal of Biological Chemistry*, *282*(11), 7903–7911.

Kumar, R. B., & Das, A. (2001). Functional analysis of the agrobacterium tumefaciens t-dna transport pore protein virb8. *Journal of Bacteriology*, *183*(12), 3636–3641.

Kurtz, S., McKinnon, K. P., Runge, M. S., Ting, J. P. Y., & Braunstein, M. (2006). The seca2 secretion factor of mycobacterium tuberculosis promotes growth in macrophages and inhibits the host immune response. *Infection and Immunity*, *74*(12), 6855–6864.

Lake, J. A., Jain, R., & Rivera, M. C. (1999). Mix and match in the tree of life. *Science*, *283*(5410), 2027–2028.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2007). Clustal w and clustal x version 2.0. *Bioinformatics*, *23*, 2947–2948.

Laszlo, A. (1992). The effects of hyperthermia on mammalian cell structure and function. *Cell Proliferation*, *25*(2), 59–87.

Laugel, B., van den Berg, H. A., Gostick, E., Cole, D. K., Wooldridge, L., Boulter, J., Milicic, A., Price, D. A., & Sewell, A. K. (2007). Different t cell receptor affinity thresholds and cd8 coreceptor dependence govern cytotoxic t lymphocyte activation and tetramer binding properties. *Journal of Biological Chemistry*, *282*(33), 23799–23810.

Le, S. Q., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, *25*(7), 1307–1320.

Leake, M. C., Greene, N. P., Godun, R. M., Granjon, T., Buchanan, G., Chen, S., Berry, R. M., Palmer, T., & Berks, B. C. (2008). Variable stoichiometry of the tata component of the twin-arginine protein transport system observed by in vivo single-molecule imaging. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(40), 15376–15381.

Lee, P. A., Buchanan, G., Stanley, N. R., Berks, B. C., & Palmer, T. (2002). Truncation analysis of tata and tatb defines the minimal functional units required for protein translocation. *Journal of Bacteriology*, *184*(21), 5871–5879.

Lee, P. A., Tullman-Ercek, D., & Georgiou, G. (2006). The bacterial twin-arginine translocation pathway. *Annual Review of Microbiology*, *60*, 373–395.

Lee, V. T., & Kessler, J. L. (2009). Type iii secretion systems as targets for novel therapeutics. *Idrugs*, *12*(10), 636–641.

Lee, V. T., & Schneewind, O. (2001). Protein secretion and the pathogenesis of bacterial infections. *Genes & Development*, *15*(14), 1725–1752.

Lenz, L. L., Mohammadi, S., Geissler, A., & Portnoy, D. A. (2003). Seca2-dependent secretion of autolytic enzymes promotes listeria monocytogenes pathogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(21), 12432–12437.

Lepock, J. R. (1997). Protein denaturation during heat shock. *Advances in Molecular Cell Biology*, *19*(19), 223–259.

Letunic, I., Doerks, T., & Bork, P. (2009). Smart 6: recent updates and new developments. *Nucleic Acids Research*, *37*, D229–D232.

Li, M. Y., Huang, Y. J., Tai, P. C., & Wang, B. H. (2008). Discovery of the first seca inhibitors using structure-based virtual screening. *Biochemical and Biophysical Research Communications*, *368*(4), 839–845.

Li, W. K., Schulman, S., Boyd, D., Erlandson, K., Beckwith, J., & Rapoport, T. A. (2007). The plug domain of the secy protein stabilizes the closed state of the translocation channel and maintains a membrane seal. *Molecular Cell*, *26*(4), 511–521.

Li, X., & Solomon, B. (2001). Zinc-mediated thermal stabilization of carboxypeptidase a. *Biomolecular engineering*, *18*(4), 179–83.

Lima, T., Auchincloss, A. H., Coudert, E., Keller, G., Michoud, K., Rivoire, C., Bulliard, V., de Castro, E., Lachaize, C., Baratin, D., Phan, I., Bougueleret, L., & Bairoch, A. (2009). Hamap: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in uniprotkb/swiss-prot. *Nucleic Acids Research*, *37*(Database issue), D471–D478.

Litman, G. W., Rast, J. P., & Fugmann, S. D. (2010). The origins of vertebrate adaptive immunity. *Nature Reviews Immunology*, *10*(8), 543–553.

Lund, P. A. (2009). Multiple chaperonins in bacteria–why so many? *FEMS Microbiology Review*, *33*(4), 785–800.

Lycklama, A. N. J. A., Bulacu, M., Marrink, S. J., & Driessen, A. J. (2010). Immobilization of the plug domain inside the secy channel allows unrestricted protein translocation. *Journal of Biological Chemistry*, *285*(31), 23747–23754.

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, *290*(5494), 1151–1155.

Lynch, M., & Conery, J. S. (2003). The origins of genome complexity. *Science*, *302*(5649), 1401–1404.

Lynch, M., & Katju, V. (2004). The altered evolutionary trajectories of gene duplicates. *Trends in Genetics*, *20*(11), 544–549.

Madeleine, M. M., Johnson, L. G., Smith, A. G., Hansen, J. A., Nisperos, B. B., Li, S., Zhao, L. P., Daling, J. R., Schwartz, S. M., & Galloway, D. A. (2008). Comprehensive analysis of hla-a, hla-b, hla-c, hla-drb1, and hla-dqb1 loci and squamous cell cervical cancer risk. *Cancer Research*, *68*(9), 3532–3539.

Maillard, A. P., Lalani, S., Silva, F., Belin, D., & Duong, F. (2007). Deregulation of the secyeg translocation channel upon removal of the plug domain. *Journal of Biological Chemistry*, *282*(2), 1281–1287.

Makarova, K., Ponomarev, V., & Koonin, E. (2001). Two c or not two c: recurrent disruption of zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biology*, *2*(9), research0033.1–research0033.14.

Makarova, K. S., Aravind, L., Galperin, M. Y., Grishin, N. V., Tatusov, R. L., Wolf, Y. I., & Koonin, E. V. (1999). Comparative genomics of the archaea (euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Research*, *9*(7), 608–628.

Manges, A. R., Johnson, J. R., Foxman, B., O'Bryan, T. T., Fullerton, K. E., & Riley, L. W. (2001). Widespread distribution of urinary tract infections caused by a multidrug-resistant escherichia coli clonal group. *New England Journal of Medicine*, *345*(14), 1007–1013.

Marques, M. A. M., Neves-Ferreira, A. G. C., da Silveira, E. K. X., Valente, R. H., Chapeaurouge, A., Perales, J., Bernardes, R. D., Dobos, K. M., Spencer, J. S., Brennan, P. J., & Pessolani, M. C. V. (2008). Deciphering the proteomic profile of mycobacterium leprae cell envelope. *Proteomics*, *8*(12), 2477–2491.

Marrack, P., Scott-Browne, J. P., Dai, S., Gapin, L., & Kappler, J. W. (2008). Evolutionarily conserved amino acids that control tcr-mhc interaction. *Annual Review of Immunology*, *26*, 171–203.

Martin, L. C., Gloor, G. B., Dunn, S. D., & Wahl, L. M. (2005). Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, *21*(22), 4116–4124.

Masignani, V., Pizza, M., & Rappuoli, R. (2006). Bacterial toxins. In *The Prokaryotes*, (pp. 893–955). New York: Springer.

Matos, C., Robinson, C., & Di Cola, A. (2008). The tat system proofreads fes protein substrates and directly initiates the disposal of rejected molecules. *Embo Journal*, *27*(15), 2055–2063.

Matsumura, M., Fremont, D. H., Peterson, P. A., & Wilson, I. A. (1992). Emerging principles for the recognition of peptide antigens by mhc class i molecules. *Science*, *257*(5072), 927–934.

McCann, H. C., & Guttman, D. S. (2008). Evolution of the type iii secretion system and its effectors in plant-microbe interactions. *New Phytologist*, *177*(1), 33–47.

McCann, J. R., Kurtz, S., & Braunstein, M. (2009). *Secreted and Exported Proteins Important to Mycobacterium tuberculosis Pathogenesis*, chap. 12, (pp. 265–297). Norfolk, UK: Caister Academic Press.

McCutcheon, J. P., McDonald, B. R., & Moran, N. A. (2009). Convergent evolution of metabolic roles in bacterial co-symbionts of insects. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(36), 15394–15399.

McFarland, L., Francetic, O., & Kumamoto, C. A. (1993). A mutation of escherichia coli seca protein that partially compensates for the absence of secb. *Journal of Bacteriology*, *175*(8), 2255–2262.

McKenzie, G. J., Harris, R. S., Lee, P. L., & Rosenberg, S. M. (2000). The sos response regulates adaptive mutation. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(12), 6646–6651.

McMichael, A. J., Gotch, F. M., Santosaguado, J., & Strominger, J. L. (1988). Effect of mutations and variations of hla-a2 on recognition of a virus peptide epitope by cytotoxic t lymphocytes. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(23), 9194–9198.

Mendel, S., McCarthy, A., Barnett, J. P., Eijlander, R. T., Nenninger, A., Kuipers, O. P., & Robinson, C. (2008). The escherichia coli tatabc system and a bacillus subtilis tatac-type system recognise three distinct targeting determinants in twin-arginine signal peptides. *Journal of Molecular Biology*, *375*(3), 661–672.

Michel, G. P. F., & Voulhoux, R. (2009). *Type II Secretory System (T2SS) in Gram-negative Bacteria: a Molecular Nanomachine for Secretion of Sec and Tat-dependent Extracellular Proteins*, (pp. 67–92). Norfolk, UK: Caister Academic Press.

Mistou, M. Y., Dramsi, S., Brega, S., Poyart, C., & Trieu-Cuot, P. (2009). Molecular dissection of the seca2 locus of group b streptococcus reveals that glycosylation of the srr1 lpxtg protein is required for full virulence. *Journal of Bacteriology*, *191*(13), 4195–4206.

Moran, N. A. (2002). Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, *108*(5), 583–586.

Moran, N. A., McCutcheon, J. P., & Nakabachi, A. (2008). Genomics and evolution of heritable bacterial symbionts. *Annual Review of Genetics*, *42*, 165–190.

Mori, T., Ishitani, R., Tsukazaki, T., Nureki, O., & Sugita, Y. (2010). Molecular mechanisms underlying the early stage of protein translocation through the sec translocon. *Biochemistry*, *49*(5), 945–950.

Moya, A., Peretó, J., Gil, R., & Latorre, A. (2008). Learning how to live together: genomic insights into prokaryote–animal symbioses. *Nature Reviews Genetics*, *9*(3), 218–229.

Muraille, E., Narni-Mancinelli, E., Gounon, P., Bassand, D., Glaichenhaus, N., Lenz, L. L., & Lauvau, G. (2007). Cytosolic expression of seca2 is a prerequisite for long-term protective immunity. *Cellular Microbiology*, *9*(6), 1445–1454.

Murphy, K. P., Paul, T., Janeway, C., & Walport, M. (2007). *Janeway's Immunobiology*, (7 ed.). New York: Garland Science.

Mushegian, A. R., & Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(19), 10268–10273.

Myouga, F., Hosoda, C., Umezawa, T., Iizumi, H., Kuromori, T., Motohashi, R., Shono, Y., Nagata, N., Ikeuchi, M., & Shinozaki, K. (2008). A heterocomplex of iron superoxide dismutases defends chloroplast nucleoids against oxidative stress and is essential for chloroplast development in arabidopsis. *Plant Cell*, *20*(11), 3148–3162.

Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H. E., Moran, N. A., & Hattori, M. (2006). The 160-kilobase genome of the bacterial endosymbiont carsonella. *Science*, *314*(5797), 267.

Natale, P., Bruser, T., & Driessen, A. J. (2008). Sec- and tat-mediated protein secretion across the bacterial cytoplasmic membrane–distinct translocases and mechanisms. *Biochimica Et Biophysica Acta*, *1778*(9), 1735–1756.

Nielsen, R., & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the hiv-1 envelope gene. *Genetics*, *148*(3), 929–936.

Nikoliczugic, J., & Carbone, F. R. (1990). The effect of mutations in the mhc class i peptide binding groove on the cytotoxic t lymphocyte recognition of the kb-restricted ovalbumin determinant. *European Journal of Immunology*, *20*(11), 2431–2437.

Nilsson, A. M., Wijaywardene, M., Gkoutos, G., Wilson, K. M., FernÃ¡ndez, N., & Reynolds, C. A. (1999). Correlated mutations in the hla class ii molecule. *International Journal of Quantum Chemistry*, *73*, 12.

Nishiyama, K. I., Suzuki, T., & Tokuda, H. (1996). Inversion of the membrane topology of secg coupled with seca-dependent preprotein translocation. *Cell*, *85*(1), 71–81.

Ochman, H., Lawrence, J. G., & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, *405*(6784), 299–304.

Ohl, M. E., & Miller, S. I. (2001). Salmonella: A model for bacterial pathogenesis. *Annual Review of Medicine*, *52*, 259–274.

Ohno, S. (1970). *Evolution by gene duplication*. London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.

Oldfield, N. J., & Wooldridge, K. G. (2009). *Protein Secretion and the Pathogenesis of Campylobacter jejuni*, (pp. 299–311). Norfolk, UK: Caister Academic Press.

Oren, A. (2006). Life at high salt concentrations. In *The Prokaryotes*, vol. 2, (pp. 263–282). New York: Springer.

Oren, A. (2008). Microbial life at high salt concentrations: phylogenetic and metabolic diversity. *Saline Systems*, *4*, 2.
URL `http://dx.doi.org/10.1186/1746-1448-4-2`

Osborne, A. R., Clemons, W. M., & Rapoport, T. A. (2004). A large conformational change of the translocation atpase seca. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(30), 10937–10942.

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, *276*(5313), 734–740.

Panahandeh, S., Holzapfel, E., & Mueller, M. (2009). *The Twin-arginine Translocation Pathway*, (pp. 23–43). Norfolk, UK: Caister Academic Press.

Panina, E. M., Mironov, A. A., & Gelfand, M. S. (2003). Comparative genomics of bacterial zinc regulons: Enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(17), 9912–9917.

Paracer, S., & Ahmadjian, V. (2000). *Symbiosis: An Introduction to Biological Associations*. Oxford, UK: Oxford University Press.

Paterson, S., Vogwill, T., Buckling, A., Benmayor, R., Spiers, A., Thomson, N., Quail, M., Smith, F., Walker, D., & Libberton, B. (2010). Antagonistic coevolution accelerates molecular evolution. *Nature*, *464*(7286), 275–278.

Perez-Brocal, V., Gil, R., Ramos, S., Lamelas, A., Postigo, M., Michelena, J. M., Silva, F. J., Moya, A., & Latorre, A. (2006). A small microbial genome: the end of a long symbiotic relationship? *Science*, *314*(5797), 312–313.

Pickering, B. S., & Oresnik, I. J. (2010). The twin arginine transport system appears to be essential for viability in sinorhizobium meliloti. *Journal of Bacteriology*, *192*(19), 5173–5180.

Pikuta, E. V., Hoover, R. B., & Tang, J. (2007). Microbial extremophiles at the limits of life. *Critical Reviews in Immunology*, *33*(3), 183–209.

Pinto, G., Mahler, D. L., Harmon, L. J., & Losos, J. B. (2008). Testing the island effect in adaptive radiation: rates and patterns of morphological diversification in caribbean and mainland anolis lizards. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1652), 2749–2757.

Poetsch, A., & Wolters, D. (2008). Bacterial membrane proteomics. *Proteomics*, *8*(19), 4100–4122.

Pohlschroder, M., Prinz, W. A., Hartmann, E., & Beckwith, J. (1997). Protein translocation in the three domains of life: Variations on a theme. *Cell*, *91*(5), 563–566.

Pollock, D. D., Taylor, W. R., & Goldman, N. (1999). Coevolving protein residues: maximum likelihood identification and relationship to structure. *Journal of Molecular Biology*, *287*(1), 187–198.

Poon, A. F., Lewis, F. I., Pond, S. L., & Frost, S. D. (2007a). Evolutionary interactions between n-linked glycosylation sites in the hiv-1 envelope. *PLoS Computational Biology*, *3*(1), e11.

Poon, A. F., Lewis, F. I., Pond, S. L., & Frost, S. D. (2007b). An evolutionary-network model reveals stratified interactions in the v3 loop of the hiv-1 envelope. *PLoS Computational Biology*, *3*(11), e231.

Popot, J. (2010). Amphipols, nanodiscs, and fluorinated surfactants: three nonconventional approaches to studying membrane proteins in aqueous solutions. *Annual review of biochemistry*, *79*, 737–775.

Punginelli, C., Maldonado, B., Grahl, S., Jack, R., Alami, M., Schroder, J., Berks, B. C., & Palmer, T. (2007). Cysteine scanning mutagenesis and topological mapping of the escherichia coli twin-arginine translocase tatc component. *Journal of Bacteriology*, *189*(15), 5482–5494.

Rapoport, T. A. (2007). Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature*, *450*(7170), 663–669.

Rasko, D. A., & Sperandio, V. (2010). Anti-virulence strategies to combat bacteria-mediated disease. *Nature Reviews Drug Discovery*, *9*(2), 117–128.

Razin, S. (2006). The genus mycoplasma and related genera (class mollicutes). In *The Prokaryotes*, (pp. 836–904). New York: Springer.

Rehermann, B. (2009). Hepatitis c virus versus innate and adaptive immune responses: a tale of coevolution and coexistence. *Journal of Clinical Investigation*, *119*(7), 1745–1754.

Ren, C. P., Beatson, S. A., Parkhill, J., & Pallen, M. J. (2005). The flag-2 locus, an ancestral gene cluster, is potentially associated with a novel flagellar system from escherichia coli. *Journal of Bacteriology, 187*(4), 1430–1440.

Reynolds, M., Bogomolnaya, L., Guo, J., Aldrich, L., Bokhari, D., Santiviago, C., McClelland, M., Andrews-Polymenis, H., & Hensel, M. (2011). Abrogation of the twin arginine transport system in salmonella enterica serovar typhimurium leads to colonization defects during infection. *PLoS One, 6*(1), 18003–18006.

Rigel, N. W., & Braunstein, M. (2008). A new twist on an old pathway - accessory secretion systems. *Molecular Microbiology, 69*(2), 291–302.

Rigel, N. W., Gibbons, H. S., McCann, J. R., McDonough, J. A., Kurtz, S., & Braunstein, M. (2009). The accessory seca2 system of mycobacteria requires atp binding and the canonical seca1. *Journal of Biological Chemistry, 284*(15), 9927–9936.

Robinson, J., Waller, M. J., Parham, P., Bodmer, J. G., & Marsh, S. G. E. (2001). Imgt/hla database - a sequence database for the human major histocompatibility complex. *Nucleic Acids Research, 29*(1), 210–213.

Robinson, N. J., & Winge, D. R. (2010). Copper metallochaperones. *Annual Review of Biochemistry, 79*, 537–562.

Robson, A., Gold, V. A. M., Hodson, S., Clarke, A. R., & Collinson, I. (2009). Energy transduction in protein transport and the atp hydrolytic cycle of seca. *Proceedings of the National Academy of Sciences of the United States of America, 106*(13), 5111–5116.

Rong, R., Li, B., Lynch, R. M., Haaland, R. E., Murphy, M. K., Mulenga, J., Allen, S. A., Pinter, A., Shaw, G. M., Hunter, E., Robinson, J. E., Gnanakaran, S., & Derdeyn, C. A. (2009). Escape from autologous neutralizing antibodies in acute/early subtype c hiv-1 infection requires multiple pathways. *PLoS Pathogens, 5*(9), e1000594.

Rose, R. W., Bruser, T., Kissinger, J. C., & Pohlschroder, M. (2002). Adaptation of protein secretion to extremely high-salt conditions by extensive use of the twin-arginine translocation pathway. *Molecular Microbiology, 45*(4), 943–950.

Roth, A., Gonnet, G., & Dessimoz, C. (2008). Algorithm of oma for large-scale orthology inference. *BMC Bioinformatics, 9*(1), 518–518.

Rudolph, M. G., Stanfield, R. L., & Wilson, I. A. (2006). How tcrs bind mhcs, peptides, and coreceptors. *Annual Review of Immunology, 24*, 419–466.

Rudolph, M. G., & Wilson, I. A. (2002). The specificity of tcr/pmhc interaction. *Current Opinion in Immunology*, *14*(1), 52–65.

Salter, R. D., Benjamin, R. J., Wesley, P. K., Buxton, S. E., Garrett, T. P. J., Clayberger, C., Krensky, A. M., Norment, A. M., Littman, D. R., & Parham, P. (1990). A binding site for the t-cell co-receptor cd8 on the alpha 3 domain of hla-a2. *Nature*, *345*(6270), 41–46.

Salter, R. D., Norment, A. M., Chen, B. P., Clayberger, C., Krensky, A. M., Littman, D. R., & Parham, P. (1989). Polymorphism in the alpha 3 domain of hla-a molecules affects binding to cd8. *Nature*, *338*(6213), 345–347.

Sanchez-Perez, G., Mira, A., Nyiro, G., Pasic, L., & Rodriguez-Valera, F. (2008). Adapting to environmental changes using specialized paralogs. *Trends in Genetics*, *24*(4), 154–158.

Sandstrom, J., Telang, A., & Moran, N. A. (2000). Nutritional enhancement of host plants by aphids - a comparison of three aphid species on grasses. *Journal of Insect Physiology*, *46*(1), 33–40.

Saparov, S. M., Erlandson, K., Cannon, K., Schaletzky, J., Schulman, S., Rapoport, T. A., & Pohl, P. (2007). Determining the conductance of the secy protein translocation channel for small molecules. *Molecular Cell*, *26*(4), 501–509.

Saper, M. A., Bjorkman, P. J., & Wiley, D. C. (1991). Refined structure of the human histocompatibility antigen hla-a2 at 2.6 a resolution. *Journal of Molecular Biology*, *219*(2), 277–319.

Scherer, D. C., DeBuron-Connors, I., & Minnick, M. F. (1993). Characterization of bartonella bacilliformis flagella and effect of antiflagellin antibodies on invasion of human erythrocytes. *Infection and Immunity*, *61*(12), 4962–4971.

Schleiff, E., & Becker, T. (2011). Common ground for protein translocation: access control for mitochondria and chloroplasts. *Nature Reviews Molecular Cell Biology*, *12*(1), 48–59.

Schluter, D. (2000). *The ecology of adaptive radiation*. New York: Oxford University Press.

Schneider, A., Dessimoz, C., & Gonnet, G. (2007). Oma browser exploring orthologous relations across 352 complete genomes. *Bioinformatics*, *23*(16), 2180–2182.

Schultz, J., Milpetz, F., Bork, P., & Ponting, C. P. (1998). Smart, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(11), 5857–5864.

Scott-Tucker, A., & Henderson, I. R. (2009). *Type V Secretion*, (pp. 139–157). Norfolk, UK: Caister Academic Press.

Seifert, K. N., Adderson, E. E., Whiting, A. A., Bohnsack, J. F., Crowley, P. J., & Brady, L. J. (2006). A unique serine-rich repeat protein (srr-2) and novel surface antigen (epsilon) associated with a virulent lineage of serotype iii streptococcus agalactiae. *Microbiology, 152*, 1029–1040.

Sekimata, M., Tanabe, M., Sarai, A., Yamamoto, J., Kariyone, A., Nakauchi, H., Egawa, K., & Takiguchi, M. (1993). Different effects of substitutions at residues 224 and 228 of mhc class i on the recognition of cd8. *Journal of Immunology, 150*(10), 4416–4426.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research, 13*(11), 2498–2504.

Sharma, V., Arockiasamy, A., Ronning, D. R., Savva, C. G., Holzenburg, A., Braunstein, M., Jacobs, W. R., & Sacchettini, J. C. (2003). Crystal structure of mycobacterium tuberculosis seca, a preprotein translocating atpase. *Proceedings of the National Academy of Sciences of the United States of America, 100*(5), 2243–2248.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., & Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids buchnera sp. aps. *Nature, 407*(6800), 81–86.

Shindyalov, I. N., & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Engineering, 11*(9), 739–747.

Shu, N. J., Zhou, T. P., & Hovmoller, S. (2008). Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics, 24*(6), 775–782.

Siboo, I. R., Chaffin, D. O., Rubens, C. E., & Sullam, P. M. (2008). Characterization of the accessory sec system of staphylococcus aureus. *Journal of Bacteriology, 190*(18), 6188–6196.

Siboo, I. R., Chambers, H. F., & Sullam, P. M. (2005). Role of srap, a serine-rich surface protein of staphylococcus aureus, in binding to human platelets. *Infection and Immunity, 73*(4), 2273–2280.

Signorovitch, A. Y., Buss, L. W., & Dellaporta, S. L. (2007). Comparative genomics of large mitochondria in placozoans. *Plos Genetics, 3*(1), e13.

Silhavy, T., Kahne, D., & Walker, S. (2010). The bacterial cell envelope. *Cold Spring Harbor Perspectives in Biology, 2*(5), a000414.

Singer, G. A., & Hickey, D. A. (2003). Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene, 317*(1-2), 39–47.

Smith, M. A., Clemons, W. M., DeMars, C. J., & Flower, A. M. (2005). Modeling the effects of prl mutations on the escherichia coli secy complex. *Journal of Bacteriology, 187*(18), 6454–6465.

Smith, N. H., Hewinson, R. G., Kremer, K., Brosch, R., & Gordon, S. V. (2009). Myths and misconceptions: the origin and evolution of mycobacterium tuberculosis. *Nature Reviews Microbiology*, *7*(7), 537–544.

Sokol, C., Chu, N., Yu, S., Nish, S., Laufer, T., & Medzhitov, R. (2009). Basophils function as antigen-presenting cells for an allergen-induced t helper type 2 response. *Nature Immunology*, *10*(7), 713–720.

Sorg, I., & Cornelis, G. R. (2009). *The Type III Secretion System*, (pp. 93–116). Norfolk, UK: Caister Academic Press.

Soskine, M., & Tawfik, D. (2010). Mutational effects and the evolution of new protein functions. *Nature Reviews Genetics*, *11*(8), 572–582.

Stamatakis, A. (2006). Raxml-vi-hpc: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, *22*(21), 2688–2690.

Strauch, E. M., & Georgiou, G. (2007). Escherichia coli tatc mutations that suppress defective twin-arginine transporter signal peptides. *Journal of Molecular Biology*, *374*(2), 283–291.

Studer, R., & Robinson-Rechavi, M. (2009). How confident can we be that orthologs are similar, but paralogs differ? *Trends in Genetics*, *25*(5), 210–216.

Studer, R. A., & Robinson-Rechavi, M. (2010). Large-scale analysis of orthologs and paralogs under covarion-like and constant-but-different models of amino acid evolution. *Molecular Biology and Evolution*, *27*, 2618–2627.

Sugai, R., Takemae, K., Tokuda, H., & Nishiyama, K. I. (2007). Topology inversion of secg is essential for cytosolic seca-dependent stimulation of protein translocation. *Journal of Biological Chemistry*, *282*(40), 29540–29548.

Suzuki, Y. (2004a). New methods for detecting positive selection at single amino acid sites. *Journal of Molecular Biology*, *59*(1), 11–19.

Suzuki, Y. (2004b). Three-dimensional window analysis for detecting positive selection at structural regions of proteins. *Molecular Biology and Evolution*, *21*(12), 2352–2359.

Suzuki, Y., & Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, *16*(10), 1315–1328.

Takamatsu, D., Bensing, B. A., & Sullam, P. M. (2004a). Four proteins encoded in the gspb-secy2a2 operon of streptococcus gordonii mediate the intracellular glycosylation of the platelet-binding protein gspb. *Journal of Bacteriology*, *186*(21), 7100–7111.

Takamatsu, D., Bensing, B. A., & Sullam, P. M. (2004b). Genes in the accessory sec locus of streptococcus gordonii have three functionally distinct effects on the expression of the platelet-binding protein gspb. *Molecular Microbiology*, *52*(1), 189–203.

Takamatsu, D., Bensing, B. A., & Sullam, P. M. (2005). Two additional components of the accessory sec system mediating export of the streptococcus gordonii platelet-binding protein gspb. *Journal of Bacteriology*, *187*(11), 3878–3883.

Tam, P. C. K., Maillard, A. P., Chan, K. K. Y., & Duong, F. (2005). Investigating the secy plug movement at the secyeg translocation channel. *EMBO Journal*, *24*(19), 3380–3388.

Tamas, I., Klasson, L., Canback, B., Naslund, A. K., Eriksson, A. S., Wernegreen, J. J., Sandstrom, J. P., Moran, N. A., & Andersson, S. G. (2002). 50 million years of genomic stasis in endosymbiotic bacteria. *Science*, *296*(5577), 2376–2379.

Tan, C., Hsia, R. C., Shou, H. Z., Carrasco, J. A., Rank, R. G., & Bavoil, P. M. (2010). Variable expression of surface-exposed polymorphic membrane proteins in in vitro-grown chlamydia trachomatis. *Cellular Microbiology*, *12*(2), 174–187.

Tan, C., Hsia, R. C., Shou, H. Z., Haggerty, C. L., Ness, R. B., Gaydos, C. A., Dean, D., Scurlock, A. M., Wilson, D. P., & Bavoil, P. M. (2009). Chlamydia trachomatis-infected patients display variable antibody profiles against the nine-member polymorphic membrane protein family. *Infection and Immunity*, *77*(8), 3218–3226.

Tanabe, M., Minami, M., Kano, K., & Takiguchi, M. (1990). The allorecognition of h-2kb-specific cd4-cd8- t cell hybridomas is influenced by the substitution at residue 256 of mhc class i molecules. *Journal of Immunology*, *145*(6), 1968–1973.

Tarry, M. J., Schafer, E., Chen, S. Y., Buchanan, G., Greene, N. P., Lea, S. M., Palmer, T., Saibil, H. R., & Berks, B. C. (2009). Structural analysis of substrate binding by the tatbc component of the twin-arginine protein transport system. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(32), 13284–13289.

Tatusov, R., Fedorova, N., Jackson, J., Jacobs, A., Kiryutin, B., Koonin, E., Krylov, D., Mazumder, R., Mekhedov, S., Nikolskaya, A., et al. (2003). The COG database: an updated version includes eukaryotes. *BMC bioinformatics*, *4*(1), 41.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., & Koonin, E. V. (2001). The cog database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, *29*(1), 22–28.

Thanassi, D. G., Chapman, M. R., & Chakraborty, S. (2009). *Assembly and Secretion of Surface Fibres in Gram-Negative Bacteria*, (pp. 159–192). Norfolk, UK: Caister Academic Press.

Thio, C. L., Thomas, D. L., Karacki, P., Gao, X. J., Marti, D., Kaslow, R. A., Goedert, J. J., Hilgartner, M., Strathdee, S. A., Duggal, P., O'Brien, S. J., Astemborski, J., & Carrington, M. (2003). Comprehensive analysis of class i and class iihla antigens and chronic hepatitis b virus infection. *Journal of Virology*, *77*(22), 12083–12087.

Thomas, J. R., & Bolhuis, A. (2006). The tatc gene cluster is essential for viability in halophilic archaea. *FEMS Microbiology Letters*, *256*(1), 44–49.

Thompson, J. (1982). *Interaction and coevolution*. New York ; Chichester: Wiley.

Thompson, J. (1994). *The coevolutionary process*. Chicago ; London: University of Chicago Press.

Thompson, J., & Cunningham, B. (2002). Geographic structure and dynamics of coevolutionary selection. *Nature*, *417*(6890), 735–738.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., & Higgins, D. G. (1997). The clustal-x windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, *25*(24), 4876–4882.

Thompson, J. N. (1999). The evolution of species interactions. *Science*, *284*(5423), 2116–2118.

Tillier, E. R., Biro, L., Li, G., & Tillo, D. (2006). Codep: maximizing co-evolutionary interdependencies to discover interacting proteins. *Proteins*, *63*(4), 822–831.

Tillier, E. R., & Lui, T. W. (2003). Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics*, *19*(6), 750–755.

Tjalsma, H., Lambooy, L., Hermans, P. W., & Swinkels, D. W. (2008). Shedding & shaving: Disclosure of proteomic expressions on a bacterial face. *Proteomics*, *8*(7), 1415–1428.

Tjalsma, H., & van Dijl, J. M. (2005). Proteomics-based consensus prediction of protein retention in a bacterial membrane. *Proteomics*, *5*(17), 4472–4482.

Toft, C., & Andersson, S. G. (2010). Evolutionary microbial genomics: insights into bacterial host adaptation. *Nature reviews. Genetics*, *11*(7), 465–475.

Toft, C., Williams, T., & Fares, M. (2009). Genome-wide functional divergence after the symbiosis of Proteobacteria with Insects unraveled through a novel computational approach. *PLoS Computational Biology*, *5*(4), e1000344.

Tong, J., Dolezal, P., Selkrig, J., Crawford, S., Simpson, A. G., Noinaj, N., Buchanan, S. K., Gabriel, K., & Lithgow, T. (2011). Ancestral and derived protein import pathways in the mitochondrion of reclinomonas americana. *Molecular Biology and Evolution*, *28*(5), 1581–1591.

Travers, S. A. A., & Fares, M. A. (2007). Functional coevolutionary networks of the hsp70-hop-hsp90 system revealed through computational analyses. *Molecular Biology and Evolution*, *24*(4), 1032–1044.

Travers, S. A. A., Tully, D. C., McCormack, G. P., & Fares, M. A. (2007). A study of the coevolutionary patterns operating within the env gene of the hiv-1 group m subtypes. *Molecular Biology and Evolution*, *24*(12), 2787–2801.

Trost, M., Wehmh
"oner, D., K
"arst, U., Dieterich, G., Wehland, J., & J
"ansch, L. (2005). Comparative proteome analysis of secretory proteins from pathogenic and non-pathogenic Listeria species. *Proteomics*, *5*(6), 1544–1557.

Tuff, P., & Darlu, P. (2000). Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. *Molecular Biology and Evolution*, *17*(11), 1753–1759.

Tully, D. C., & Fares, M. A. (2006). Unravelling selection shifts among foot-and-mouth disease virus (fmdv) serotypes. *Evolutionary Bioinformatics*, *2*, 211–225.

Tully, D. C., & Fares, M. A. (2009). Shifts in the selection-drift balance drive the evolution and epidemiology of foot-and-mouth disease virus. *Journal of Virology*, *83*(2), 781–790.

Turner, D. P., Wooldridge, K. G., & Ala'Aldeen, D. (2009). *Protein Secretion and Pathogenesis in Neisseria meningitidis*, (pp. 347–360). Norfolk, UK: Caister Academic Press.

Vallabhapurapu, S., & Karin, M. (2009). Regulation and function of nf-kappa b transcription factors in the immune system. *Annual Review of Immunology*, *27*, 693–733.

van den Berg, H. A., Wooldridge, L., Laugel, B., & Sewell, A. K. (2007). Coreceptor cd8-driven modulation of t cell antigen receptor specificity. *Journal of Theoretical Biology*, *249*(2), 395–408.

van der Woude, M. W., & Baumler, A. J. (2004). Phase and antigenic variation in bacteria. *Clinical Microbiology Reviews*, *17*(3), 581–611.

van Ham, R. C., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U., Fernandez, J. M., Jimenez, L., Postigo, M., Silva, F. J., Tamames, J., Viguera, E., Latorre, A., Valencia, A., Moran, F., & Moya, A. (2003). Reductive genome evolution in buchnera aphidicola. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(2), 581–586.

Van Valen, L. (1973). A new evolutionary law. *Evolutionary Theory*, *1*(1), 1–30.

Van Valen, L. (1974). Molecular evolution as predicted by natural selection. *Journal of Molecular Evolution*, *3*(2), 89–101.

Vassylyev, D. G., Mori, H., Vassylyeva, M. N., Tsukazaki, T., Kimura, Y., Tahirov, T. H., & Ito, K. (2006). Crystal structure of the translocation atpase seca from thermus thermophilus reveals a parallel, head-to-head dimer. *Journal of Molecular Biology*, *364*(3), 248–258.

Vetriani, C., Maeder, D. L., Tolliday, N., Yip, K. S. P., Stillman, T. J., Britton, K. L., Rice, D. W., Klump, H. H., & Robb, F. T. (1998). Protein thermostability above 100 degrees c: A key role for ionic interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(21), 12300–12305.

Vogel, T. U., Evans, D. T., Urvater, J. A., O'Connor, D. H., Hughes, A. L., & Watkins, D. I. (1999). Major histocompatibility complex class i genes in primates: co-evolution with pathogens. *Immunological Reviews*, *167*, 327–337.

Walsh, C. (2000). Molecular mechanisms that confer antibacterial drug resistance. *Nature*, *406*(6797), 775–781.

Wan, F. Y., Anderson, D. E., Barnitz, R. A., Snow, A., Bidere, N., Zheng, L. X., Hegde, V., Lam, L. T., Staudt, L. M., Levens, D., Deutsch, W. A., & Lenardo, M. J. (2007). Ribosomal protein s3: A kh domain subunit in nf-kappa b complexes that mediates selective gene regulation. *Cell*, *131*(5), 927–939.

Wang, J. H., & Reinherz, E. L. (2000). Structural basis of cell-cell interactions in the immune system. *Current Opinion in Structural Biology*, *10*(6), 656–661.

Wang, J. H., & Reinherz, E. L. (2002). Structural basis of t cell recognition of peptides bound to mhc molecules. *Molecular Immunology*, *38*(14), 1039–1049.

Warner, J. R., & McIntosh, K. B. (2009). How common are extraribosomal functions of ribosomal proteins? *Molecular Cell*, *34*(1), 3–11.

Warren, G., Oates, J., Robinson, C., & Dixon, A. M. (2009). Contributions of the transmembrane domain and a key acidic motif to assembly and function of the tata complex. *Journal of Molecular Biology*, *388*(1), 122–132.

Watanabe, S., Kodaki, T., & Makino, K. (2005). Complete reversal of coenzyme specificity of xylitol dehydrogenase and increase of thermostability by the introduction of structural zinc. *The Journal of biological chemistry*, *280*(11), 10340–10349.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing*. New York: Wiley.

Widdick, D. A., Dilks, K., Chandra, G., Bottrill, A., Naldrett, M., Pohlschroder, M., & Palmer, T. (2006). The twin-arginine translocation pathway is a major route of protein export in streptomyces coelicolor. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(47), 17927–17932.

Williams, T. A., Codoner, F. M., Toft, C., & Fares, M. A. (2010). Two chaperonin systems in bacterial genomes with distinct ecological roles. *Trends in Genetics*, *26*(2), 47–51.

Wooldridge, K. (2009). *Bacterial Secreted Proteins: Secretory Mechanisms and Role in Pathogenesis*. Norfolk, UK: Caister Academic Press.

Wooldridge, L., Lissina, A., Vernazza, J., Gostick, E., Laugel, B., Hutchinson, S. L., Mirza, F., Dunbar, P. R., Boulter, J. M., Glick, M., Cerundolo, V., van den Berg, H. A., Price, D. A., & Sewell, A. K. (2007). Enhanced immunogenicity of ctl antigens through mutation of the cd8 binding mhc class i invariant region. *European Journal of Immunology*, *37*(5), 1323–1333.

Wooldridge, L., van den Berg, H. A., Glick, M., Gostick, E., Laugel, B., Hutchinson, S. L., Milicic, A., Brenchley, J. M., Douek, D. C., Price, D. A., & Sewell, A. K. (2005). Interaction between the cd8 coreceptor and major histocompatibility complex class i stabilizes t cell receptor-antigen complexes at the cell surface. *Journal of Biological Chemistry*, *280*(30), 27491–27501.

Woolhouse, M. E. J., Webster, J. P., Domingo, E., Charlesworth, B., & Levin, B. R. (2002). Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature Genetics*, *32*(4), 569–577.

Worthington, Z. E. V., & Carbonetti, N. H. (2009). *Bordetella pertussis*, chap. 18, (pp. 413–431). Norfolk, UK: Caister Academic Press.

Wu, D., Daugherty, S. C., Van Aken, S. E., Pai, G. H., Watkins, K. L., Khouri, H., Tallon, L. J., Zaborsky, J. M., Dunbar, H. E., Tran, P. L., Moran, N. A., & Eisen, J. A. (2006). Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. *Plos Biology*, *4*(6), e188.

Wu, H., Bu, S., Newell, P., Chen, Q., & Fives-Taylor, P. (2007). Two gene determinants are differentially involved in the biogenesis of fap1 precursors in streptococcus parasanguis. *Journal of Bacteriology*, *189*(4), 1390–1398.

Xiong, Y., Kern, P., Chang, H. C., & Reinherz, E. L. (2001). T cell receptor binding to a pmhcii ligand is kinetically distinct from and independent of cd4. *Journal of Biological Chemistry*, *276*(8), 5659–5667.

Yang, Z., & Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, *15*(12), 496–503.

Yang, Z., & Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, *19*(6), 908–917.

Yeager, M., & Hughes, A. L. (1999). Evolution of the mammalian mhc: natural selection, recombination, and convergent evolution. *Immunological Reviews*, *167*, 45–58.

Yen, M. R., Tseng, Y. H., Nguyen, E. H., Wu, L. F., & Saier, M. H. (2002). Sequence and phylogenetic analyses of the twin-arginine targeting (tat) protein export system. *Archives of Microbiology*, *177*(6), 441–450.

Yi, L., Jiang, F., Chen, M., Cain, B., Bolhuis, A., & Dalbey, R. E. (2003). Yidc is strictly required for membrane insertion of subunits a and c of the f(1)f(0)atp synthase and sece of the secyeg translocase. *Biochemistry*, *42*(35), 10537–10544.

Yuan, J. J., Zweers, J. C., van Dijl, J. M., & Dalbey, R. E. (2010). Protein transport across and into cell membranes in bacteria and archaea. *Cellular and Molecular Life Sciences*, *67*(2), 179–199.

Zanen, G., Antelmann, H., Meima, R., Jongbloed, J. D. H., Kolkman, M., Hecker, M., van Dijl, J. M., & Quax, W. J. (2006). Proteomic dissection of potential signal recognition particle dependence in protein secretion by bacillus subtilis. *Proteomics*, *6*(12), 3636–3648.

Zeldovich, K. B., Berezovsky, I. N., & Shakhnovich, E. I. (2007). Protein and dna sequence determinants of thermophilic adaptation. *PLoS Computational Biology*, *3*(1), e5.

Zhang, C., Anderson, A., & DeLisi, C. (1998). Structural principles that govern the peptide-binding motifs of class i mhc molecules. *Journal of Molecular Biology*, *281*(5), 929–947.

Zhang, J. (2004). Frequent false detection of positive selection by the likelihood method with branch-site models. *Molecular Biology and Evolution*, *21*(7), 1332–1339.

Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution*, *22*(12), 2472–2479.

Zhang, J. Z. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, *18*(6), 292–298.

Zhang, Y., Lathigra, R., Garbe, T., Catty, D., & Young, D. (1991). Genetic-analysis of superoxide-dismutase, the 23 kilodalton antigen of mycobacterium-tuberculosis. *Molecular Microbiology*, *5*(2), 381–391.

Zhong, J., Gastaminza, P., Chung, J., Stamataki, Z., Isogawa, M., Cheng, G., McKeating, J. A., & Chisari, F. V. (2006). Persistent hepatitis c virus infection in vitro: Coevolution of virus and host. *Journal of Virology*, *80*(22), 11082–11093.

Zhou, J. H., & Xu, Z. H. (2005). The structural view of bacterial translocation-specific chaperone secb: implications for function. *Molecular Microbiology*, *58*(2), 349–357.

Zhou, M. X., Zhu, F., Dong, S. L., Pritchard, D. G., & Wu, H. (2010). A novel glucosyltransferase is required for glycosylation of a serine-rich adhesin and biofilm formation by streptococcus parasanguinis. *Journal of Biological Chemistry*, *285*(16), 12140–12148.

Zimmer, J., Li, W. K., & Rapoport, T. A. (2006). A novel dimer interface and conformational changes revealed by an x-ray structure of b-subtilis seca. *Journal of Molecular Biology*, *364*(3), 259–265.

Zimmer, J., Nam, Y. S., & Rapoport, T. A. (2008). Structure of a complex of the atpase seca and the protein-translocation channel. *Nature*, *455*(7215), 936–943.

Zimmer, J., & Rapoport, T. A. (2009). Conformational flexibility and peptide interaction of the translocation atpase seca. *Journal of Molecular Biology*, *394*(4), 606–612.