Causes of Substitution Frequency Variation in Pathogenic Bacteria

WAGIED DAVIDS
Dissertation presented at Uppsala University to be publicly examined in Lindahlsalen, EBC, Uppsala, Friday, April 22, 2005 at 12:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

**Abstract**


Estimating substitution frequencies at sites that influence (Ka) and do not influence (Ks) the amino acid sequence is important for understanding the dynamics of molecular sequence evolution and the selective pressures that have shaped genetic variation.

The aim of this work was to gain a deeper understanding of the driving forces of substitution frequency variation in human pathogens. *Rickettsia prowazekii*, the causative agent of epidemic typhus and *Helicobacter pylori*, which has been implicated in gastric diseases were used as model systems. A specific focus was on the evolution of orphan genes in *Rickettsia*. Additionally, adaptive sequence evolution and factors influencing protein evolutionary rates in *H. pylori* were studied.

The comparative sequence analyses of orphan genes using Typhus Group (TG) and Spotted Fever Group (SFG) *Rickettsia*, indicate that orphan genes in the SFG correspond to pseudogenes in the TG and that pseudogenes in the SFG correspond to extensively degraded gene remnants in the TG. The analysis also showed that ancestral gene sequences could be reconstructed from extant gene remnants of closely related species. The studies of split genes in *R. conorii* indicate that many of the small fragmented ORFs are probably pseudogenes. Analysis of the *H. pylori* carbamoyl phosphate synthetase provided an opportunity to understand natural selection acting on a protein undergoing adaptive evolution. Factors such as network properties, protein-protein interactions, gene essentiality and chromosomal position on protein evolutionary rates in *H. pylori* were studied, of which antigenicity and gene location were identified as the strongest factors.

In conclusion, high Ka/Ks ratios in human pathogens may reflect either adaptive sequence evolution or gene deterioration. Distinguishing between the two is an important task in molecular evolution and also of great relevance for medical microbiology and functional genomics research.

**Keywords:** Comparative genomics, Molecular evolution, Bioinformatics, Human pathogens, *Rickettsia, Helicobacter pylori*

Wagied Davids, Department of Evolution, Genomics and Systematics, Molecular Evolution, Norbyv. 18C, Uppsala University, SE-75236 Uppsala, Sweden

© Wagied Davids 2005

ISSN 1651-6214
ISBN 91-554-6181-6

urn:nbn:se:uu:diva-4838 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-4838)
List of Papers

This thesis is based on the following papers, which will be referred to in the text by their roman numerals.


Reprints were made with the permission of the publishers.
**Contents**

1 Introduction..................................................................................................9

1.1 *Rickettsia* ................................................................................................10
   1.1.1 A historical perspective ..............................................................10
   1.1.2 The genus *Rickettsia* .............................................................10

1.2 Comparative genomics of *Rickettsia* .................................................12

1.3 Reductive genome evolution..............................................................14

1.4 Deletional bias in *Rickettsia* ...........................................................16

1.5 Sequence evolution in *Rickettsia* ....................................................16
   1.5.1 Evolution of noncoding sequences .............................................16
   1.5.2 Pseudogenes in *Rickettsia* .....................................................17
   1.5.3 Split genes in *R. conorii* .........................................................18
   1.5.4 Evolution of coding sequences ................................................19
   1.5.5 Those mysterious little ORFans in *Rickettsia* .........................20

1.6 *Helicobacter pylori* .............................................................................21
   1.6.1 Comparative Genomics of *H. pylori* .........................................22

1.7 What determines the rate of protein evolution? .................................22
   1.7.1 Chromosomal neighbours..........................................................23
   1.7.2 Biological networks...................................................................23
   1.7.3 Protein essentiality/lethality.......................................................24
   1.7.4 Phylogenetic conservation profiles............................................24

2 Aims...........................................................................................................25

3 Methodology..............................................................................................26

3.1 The importance of sequence alignments in biology ...........................26
   3.1.1 Types of sequence alignments ....................................................26
   3.1.2 Assessment of sequence alignments ...........................................27

3.2 Gene prediction ..................................................................................27
   3.2.1 Extrinsic approaches to gene prediction .....................................28
   3.2.2 Intrinsic approaches to gene prediction .......................................28

3.3 Detecting and resolving orthologs......................................................29
   3.3.1 Homologs, paralogs and orthologs ..............................................29
   3.3.2 Gene families..............................................................................30
   3.2.3 Sequence clustering using the TRIBE-MCL algorithm..............30
3.4 Evolutionary substitution rates ...........................................................31
  3.4.1 What is Ka/Ks? ...........................................................................31
  3.4.2 Methods of estimating Ka/Ks .....................................................32
  3.4.3 Applications of Ka/Ks .................................................................34
3.5 Reconstruction of ancestral sequences ...............................................34
  3.5.1 Methods of ancestral sequence reconstruction ...........................35
  3.5.2 Uses of ancestral sequences ........................................................35
3.6 Construction of biological interaction networks ................................36
3.7 Comparative protein structure modelling...........................................37
  3.7.1 Mapping mutations onto protein structure.................................38
4 Results........................................................................................................39
  4.1 Gene degradation in Rickettsia ..........................................................39
    4.1.1 Small RNAs in Rickettsia- are they functional? .........................39
  4.2 Birth and death of ORFan genes in Rickettsia ....................................40
    4.2.1 Coding potential of intergenic regions in Rickettsia species ......41
    4.2.2 ORFan genes in the SFG correspond pseudogenes in the TG ....41
    4.2.3 ORFans as short, internal fragments of deteriorating genes ......41
    4.2.4 ORFans as short, fused fragments of deteriorating genes..........42
    4.2.5 Deletions and insertions in the TG and SFG Rickettsia ..........42
    4.2.6 Putative function of reconstructed genes ....................................44
  4.3 Positive selection scanning of H. pylori .............................................45
  4.4 Factors underlying protein evolutionary rates in H.pylori .................46
    4.4.1 Evolutionary rates of linked genes..............................................46
    4.4.2 Evolutionary rates of interacting proteins...................................46
    4.4.3 Evolutionary rates of essential vs nonessential genes.................48
5 Discussion..................................................................................................49
  5.1 Transition to intracellular lifestyles is often accompanied by genome size reduction ...............................................................49
  5.2 Acceleration of nucleotide substitution rates following partial loss of function? .................................................................50
  5.3 Maintenance of nongenic DNA .......................................................51
  5.4 Gene deletions mediated by repeated sequences ............................52
  5.2 An evolutionary perspective on ORFan genes ..................................52
    5.2.1 Characteristics of ORFan genes..................................................52
    5.2.2 Origin and formation of new gene families and the acquisition of new functions .................................................................53
    5.2.3 On the origin and functions of ORFans .......................................53
    5.2.4 Do ORFans correspond to real genes? .........................................54
    5.2.5 Are ORFans essential proteins? ..................................................56
    5.2.6 Prioritising ORFan studies .........................................................56
  5.3 Adaptive evolution in H. pylori: Carbamoyl Phosphate synthetase ...57
  5.4 Investigating substitution rate variation in pathogenic bacteria .......57
Abbreviations

bp  Base pair
BLAST  Basic Local Alignment Search Tool
BLOSUM  BLOcks Substitution Matrix
CDD  Conserved Domain Database
CPS  Carbamoyl Phosphate Synthetase
EIN  Enzyme Interaction Network
DNA  Deoxyribonucleic acid
GLIMMER  Gene Locator and Interpolated Markov Modeler
Ka  Number of nonsynonymous substitutions per nonsynonymous site
kb  Kilobases
Ks  Number of synonymous substitutions per synonymous site
nt  Nucleotides
Mb  Megabases
Myrs  Million years
ORF  Open Reading Frame
PAM  Point Accepted Mutations
PCR  Polymerase Chain Reaction
PDB  Protein Data Bank
PIN  Protein Interaction Network
PSI-BLAST  Position Specific Iterative BLAST
RNA  Ribonucleic acid
SFG  Spotted Fever Group Rickettsia
TG  Typhus Group Rickettsia
1 Introduction

The main emphasis of this thesis is on sequence evolution in human pathogenic bacteria. In particular using information derived from sequence analysis to infer evolutionary events in presenting examples such as genome degradation and formation of new genes, adaptive evolution, but also to gain a deeper understanding of the driving forces that underlie substitution frequency variation in pathogenic bacteria.

The thesis concentrates on two of my favourite bacteria, *Rickettsia* and *Helicobacter*. The first part of the thesis is mostly dedicated to *Rickettsia* and will presents the biological background to the genus *Rickettsia* in order to give the reader the opportunity to become familiar with these strange bacteria. This is followed by an introduction of the evolutionary forces that drive genome degradation and in particular gene intermediaries which characterise different steps of the gene degradation process. The latter part of the introductory chapter, introduces the reader to comparative genomics of *Helicobacter pylori*. Studies of various factors which may influence protein evolutionary rates are also highlighted. To help gain a better understanding and insight to the results generated, an overview of the various methodologies employed are presented. A summary of the results from the papers is followed by a presentation of known cases of genome degradation. A glimpse is given into the strange and exciting world of ORFan genes in the hope of unravelling some of their mysteries. The goal is to give the reader a coherent view of reductive genome evolution, but also an insight into the evolutionary dynamics of the birth and death of genes in bacteria. Adaptive evolution as a force in determining substitution frequency variation in *H. pylori* is also highlighted together with an overview of factors influencing substitution frequency variation.
1.1 *Rickettsia*

1.1.1 A historical perspective

Historically, the genus *Rickettsia* has always been associated with wars and human disasters afflicting mankind. It is reported that the plague of Athens (430-426 BC), may have been caused by a typhus epidemic (Retief et al., 1998). This devastating disease has also been known to have caused the deaths of at least 3 million people during the First World War in Eastern Europe and Russia.

*Rickettsia prowazekii*, the causative agent of epidemic typhus was discovered in 1909 and was named after its discoverers, Howard Ricketts and Stanislauw Von Prowazek. Charles Nicolle at the Pasteur Institute in Tunis, that same year also demonstrated that epidemic typhus is transmitted by the human body louse, *Pediculus humanus corpis* (Nicolle et al., 1909). A hunt for a vaccine against this deadly pathogen was underway with a vaccine being developed in 1930 by Rudolf Weigl. Unfortunately, most of Weigl's lab members as well as others, such as Ricketts and Von Prowazek died as a result of rickettsial infections. The only surviving member of these pioneering scientists, Charles Nicolle, was later honoured with the Nobel Prize in 1928 for his work "on the mode and transmission of epidemic typhus" (Raju, 1998).

Although, no major outbreaks of epidemic typhus have been reported, it is still considered a threat in developing countries by the World Health Organisation (WHO Report, 1997).

1.1.2 The genus *Rickettsia*

*Rickettsia* are rod-shaped, gram-negative, vector-transmitted, obligate intracellular parasites which belong the α-proteobacteria (Figure 1). Rickettsial genome sizes are small (1.0-1.6 Mb) and consists of a single circular chromosome (Roux et al., 1992). It is believed that the ancestors of these bacteria, may have possessed much larger genomes containing genes essential for maintaining their free-living status. From an evolutionary perspective, the ancestors of *Rickettsia* may have initiated the seminal event that lead to the formation of modern day mitochondria (Andersson S.G.E. et al., 1998b; Gray, 1998; Muller and Martin, 1999). The genomes of *Rickettsia*, may therefore tell tales of phylogenetic events characterising the road travelled to intracellular lifestyles.

Members of the genus *Rickettsia* are classified based on phylogenetic analysis into two main groups, namely the Typhus Group (TG) and Spotted Fever Group (SFG). The existence of a previous third group, the Scrub Typhus Group (STG), consisting solely of *R. tsutsugamushi*, is known to be phylogenetically distinct from the other *Rickettsia* to warrant its own genus,
Orientia, within the tribe Rickettsiae (Tamura et al., 1995). Some species previously belonging to the TG or SFG Rickettsia, such as R. bellii and R. canada, have been reclassified and shown to be phylogenetically close, but distinct from the other two groups (Roux et al., 1995).

1.1.2.1 The Typhus Group (TG)
The TG Rickettsia consists of two members, namely R. typhi and R. prowazekii which share many features in common such as 16S rRNA sequence similarity, strict intracytoplasmic localisation, antigenic properties and G+C content (29-30%) (Tyeryar et al., 1973), suggesting a close phylogenetic relationship (Roux et al., 1995; Baxter, 1996). The membership of R. canada is however unclear, since it shares some features in common with

Figure 1. Phylogenetic relationship of Rickettsia (taken from Sekeyova et al., 2001).
the TG such as serological properties and G+C content, but also display some characteristic properties of the SFG e.g. cytoplasmic and nuclear growth, ticks as arthropod vectors as well as transovarial transmission in ticks.

Both *R. prowazekii* and *R. typhi* are pathogenic to their human host. *R. prowazekii* is the causative agent of epidemic typhus and *R. typhi*, causes endemic murine typhus (Raoult et al., 1997). *R. typhi* primarily infects rodents and is transmitted to humans by fleas and results in a milder form of typhus.

1.1.2.2 The Spotted Fever Group (SFG)
The SFG *Rickettsia* represent a geographically diverse group, containing currently 13 pathogenic species and another 20 potential members having been identified, but not yet assigned any human disease symptoms. All SFG *Rickettsia* are grouped together in the same phylogenetic cluster with members sharing a G+C content of 32-33% (Tyeryar et al., 1973).

Pathogenic and non-pathogenic members are present, with the most common diseases representing typhus-like rickettsial diseases such as Rocky Mountain spotted fever caused by *R. rickettsii*, African tick typhus and rickettsial pox. *R. conorii* causes a disease with similar severity to *R. rickettsii*, known as Mediterranean spotted fever, with the other pathogenic species responsible for causing milder disease symptoms. Humans are considered only incidental hosts, although as many as 10 isolates are known to be human pathogens.

*Rickettsia* are recycled and maintained in nature by ticks via transovarial transmission from infected ticks to infected ova which then later hatch. The infected larval offspring mainly infect rodents which then completes the so-called "transovarian passage" (Azad et al., 1998). Although *Rickettsia* normally multiply directly within the host cell cytoplasm, some species of the SFG *Rickettsia* are also capable of dividing in the cell nucleus.

1.2 Comparative genomics of *Rickettsia*
Comparative analyses of the published *Rickettsia* genomes, *R. prowazekii* (Andersson S.G.E. et al., 1998b), *R. conorii* genome (Ogata et al., 2001), *R. sibirica* (Malek et al., 2004) and *R. typhi* genome (McLeod et al., 2004), provide some useful insights into the mechanisms and modes of genome evolution, lifecycles and pathogenicity which characterise the genus *Rickettsia*. Members of the TG and SFG *Rickettsia*, are thought to have diverged from their common ancestor between 40-80 Myrs ago, and thus also represent an interesting view into the molecular details which separate members of the typhus *Rickettsia* from that of the spotted fever *Rickettsia* group.
The genomes exhibit large differences in genome size, gene and G+C content with the *R. prowazekii* genome essentially appearing to be a subset of the larger *R. conorii* genome (Table 1). *R. prowazekii* has accumulated substantially more pseudogenes than *R. conorii*, which may reflect that it has suffered more severe gene decay or have been subjected to reductive forces for a greater period. Despite both genomes showing obvious signs of gene deterioration, the overall gene order between *R. conorii* and *R. prowazekii* genomes are remarkable similar, except for small rearrangements near the DNA replication terminus, indicating overall that no severe gene rearrangements have occurred since their divergence from their common ancestor.

The genome of *R. conorii* exhibits a much higher density of interspersed repetitive DNA than that of *R. prowazekii*, with 10 families of repeated DNA elements being identified. The repeat fraction varies in size and is G+C-rich (40%) and constitute 3.2% of the entire genome. The distribution of repeated elements is essentially random throughout the genome.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome size (bp)</th>
<th>No. of genes</th>
<th>Coding region (% of genome)</th>
<th>% G+C for the:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Genome</td>
<td>Coding regions</td>
</tr>
<tr>
<td><em>R. typhi</em></td>
<td>1,111,496</td>
<td>877</td>
<td>76.27</td>
<td>28.92</td>
</tr>
<tr>
<td><em>R. prowazekii</em></td>
<td>1,111,523</td>
<td>872</td>
<td>76.24</td>
<td>29.00</td>
</tr>
<tr>
<td><em>R. conorii</em></td>
<td>1,268,755</td>
<td>1,412</td>
<td>81.45</td>
<td>32.44</td>
</tr>
<tr>
<td><em>R. sibirica</em></td>
<td>1,250,021</td>
<td>1,234</td>
<td>77.76</td>
<td>32.47</td>
</tr>
<tr>
<td><em>R. rickettsii</em></td>
<td>1,257,710</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 1. Comparison of genome statistics of Rickettsia (taken from McLeod et al., 2004).

The genome of *R. typhi* is nearly identical to its close relative *R. prowazekii* and highly similar to *R. conorii* and other SFG bacteria. The few differences between the two TG *Rickettsiae* include a 12 kb insertion in the genome of *R. prowazekii*, a large inversion close the origin of replication with no loss of
genes in the region, and the fact that \textit{R. typhi} has lost the complete cytochrome c oxidase system. In addition, \textit{R. typhi} has several pseudogenes for which functional homologs are found in \textit{R. prowazekii} (McLeod et al. 2004).

1.3 Reductive genome evolution

Bacteria have various mechanisms at their disposal to increase their genetic content such as gene duplication and horizontal gene transfer. It is known that horizontal gene transfer is quite rampant in nature, contributing not only raw genetic material, but also increasing bacterial fitness through acquisition of novel genes (Lawrence, 1999). Although, horizontal gene transfer contributes to the steady inflow of genetic material, bacterial genomes remain compact and small in size, indicating that reductive forces must constantly be operating to prevent accumulation of potential harmful genetic parasites such as transposons and bacteriophages, but also other useless noncoding DNA (Lawrence and Ochman, 1997; Lawrence et al., 2001). Thus, the size and coding content of bacterial genomes reflect the balance between the inflow and outflow of genetic material, how much of each process can be detected in a genome at any given time depends on the rate of horizontal gene transfer events versus the rate of gene inactivation and elimination events (Figure 2) (Petrov, 2002; Mira et al., 2001).

![Figure 2](image)

\textit{Figure 2.} Mutational mechanisms responsible for genome size evolution (taken from Mira et al., 2000).

Now we consider the molecular mechanisms that mediate the degradation of genomes. Transition to intracellular lifestyles has frequently been correlated with a reduction in genome size, genes loss, changes in genome content and
base composition of bacteria (Stepkowski et al., 2001; Moran, 2002). Indeed, the genomes of intracellular organisms such as *Rickettsia* (Andersson S.G.E. et al., 1998b) and *Buchnera* (Tamas et al., 2002) are usually small, ranging in size from 0.5-1.9 Mb, exhibit extreme AT-richness and contain numerous pseudogenes.

Various hypotheses have been suggested to explain the reduction in genome sizes of intracellular bacteria, most of which center on selection for small genome sizes or selection against an increase in genome size expansion (Petrov, 2000; Petrov, 2001; Mira et al. 2001; Moran 2002). Some are listed below:

1. Selection favours small genome sizes. This could potentially be due to a need for faster replication or energy savings. However, this does not explain why some bacteria with smaller genomes such as *Rickettsia* are slow growing in comparison to free-living bacteria with larger genomes such as *Escherichia coli* (Andersson S.G.E. et al., 1995).
2. An increase in the rate of deletions or in the degree of deletional bias of novel mutations drives the reduction. High deletion rates could be advantageous, removing sequences which are detrimental such as genetic parasites (Lawrence et al., 2001).
3. Genome-wide decrease in selection across many loci results in a large proportion of the genome that is effectively neutral, and these regions are eliminated by deletional bias in the mutational pattern.

Forces acting on a population level can also affect genome evolution and organismal fitness. It is believed that obligate intracellular parasites conditioned by a rich intracellular milieu provided by their eukaryotic hosts, could potentially render some genes functionally redundant and therefore expendable. Due to their secluded intracellular lifestyle, not only do these organisms suffer genetic isolation, limiting the acquisition of genetic material via horizontal gene transfer, but also frequently experience recurrent bottlenecks during host transmission from one generation to the next, resulting in smaller effective population sizes. These factors eventually result in low recombination rates and accumulation of slightly deleterious mutations due to relaxed selection constraints which eventually become fixed within the population resulting in reduced fitness. This effect is known as Muller's ratchet and thought to operate in small asexual population such as the endosymbiont *Buchnera* (Moran, 1996) and presumably *Rickettsia* (Andersson S.G.E. et al., 1998b).
1.4 Deletional bias in *Rickettsia*

Genome size reduction in bacteria is thought to be a consequence of a bias for deletions (Andersson and Andersson 1999a, 1999b, 2001; Mira et al., 2001). This phenomenon is supposed to have played a significant role in balancing genome size expansion in free-living as well as obligate intracellular bacteria. Deletional bias may result in the elimination of large segments of DNA by homologous recombination (rapid gene loss) or step-wise degradation and subsequent elimination of small segments of DNA (gradual gene loss). For deletional bias to be effective in reducing genome size, conditions must exist under which selection coefficients are low or small population sizes exist i.e. Muller's ratchet.

A dramatic deletion mechanism and one that has left a more obvious signature on highly derived genomes is that of intrachromosomal recombination at repeated sequences (Ogata et al., 2001; Amiri et al., 2002; Frank et al., 2002). Such deletions leave at least two signatures. First, they lead to loss of intervening sequences between repeated sequences. Second, they lead to rearrangements of the flanking sequences surrounding the original repeat sequences. Such rearrangements may be detected in descendants of the deleted genome as the loss of highly conserved sequences.

From analyses of pseudogenes in *Rickettsia*, deletions are known to predominate over insertions with respect to frequencies, as well as number of nucleotides affected by event occurring in an apparent random manner (Andersson and Andersson, 1999a, 1999b). This trend was also verified in a larger study containing a set of 26 pseudogenes and also indicated the existence of a strong deletional bias that drives neutrally evolving sequences towards elimination (Andersson and Andersson, 2001).

1.5 Sequence evolution in *Rickettsia*

1.5.1 Evolution of noncoding sequences

Bacterial genomes are conceptually regarded as compact and efficient in genomic design, with most bacterial genomes consisting of roughly 90% coding content and minimal amount dedicated to "junk" or noncoding DNA, except for essential regulatory regions. The evolution of noncoding regions appear to be determined primarily by the selective pressure to minimise the amount of nonfunctional DNA, while maintaining essential regulatory signals (Rogozin et al., 2002). For most bacterial genomes this description is quite apt, however the discovery that the *R. prowazekii* genome contains 24% noncoding DNA, the largest thus far detected, raised questions of what lies buried in their contents (Andersson S.G.E. et al., 1998b). On deeper inspection, it was found that these noncoding segments represent decaying remnants of ancestral coding genes in their final stages of degradation.
1.5.2 Pseudogenes in *Rickettsia*

Most bacterial genomes contain very few pseudogenes, most noticeable exceptions are genomes of intracellular parasites such as *M. leprea* (Cole et al., 2001), *R. prowazekii* (Andersson S.G.E. et al., 1998b; Andersson and Andersson, 1999a, 1999b, 2001) and *R. typhi* (McLeod et al., 2004), which are estimated to contain 1116, 12 and 41 pseudogenes, respectively. The discovery of pseudogenes in *Rickettsia* has provided a wonderful opportunity of case studies representing neutral sequence evolution.

Pseudogenes are generally defined as disabled copies or decayed remnants of genes that display similarity to full-length functional genes, but are nonfunctional due to accumulation of disruptive mutations (Petrov et al., 2000). Disablements may be due to frameshift mutations, creation of premature termination codons or disablements of regulatory regions. These inactivated gene sequences are thought to evolve under relaxed or no functional constraints and therefore accumulate mutations in a neutral manner acting as "molecular fossils" which measure the overall mutation processes and the stability of genomic sequences.

The existence of pseudogenes in genomes can theoretically be explained by two alternatives. Pseudogenes may be the result of inactivation of recently acquired foreign genes via horizontal gene transfer or alternatively due to the inactivation of resident genes. The prior alternative seems unlikely due to the secluded lifestyle of *Rickettsia* and reductive mode of sequence evolution. In addition, studies based on the nucleotide frequencies and G+C content values at synonymous third codon positions suggests that pseudogenic sequences and unique genes have resided in *Rickettsia* genome since long before the divergence of the TG and SFG (Andersson and Andersson, 1999a, 1999b, 2001). It thus seems more plausible that pseudogenes mark the gradual nature of reductive forces acting on resident genes, targeting genes for further degradation after which gene degradation will ultimately ensue.

The cost of pseudogenes inactivation may reflect the relative ease with which a gene function can be made redundant and the extent to which such a loss of functionality can be tolerated. Although, it is generally perceived that pseudogenes are functionally inactive and therefore do not contribute to organismal fitness, it may be argued that they can still exert their effect due to shear bulk within the genome or due to location effects on neighbouring genes (Petrov et al., 2000). It is not un conceivable that if a pseudogene forms part of an intricate metabolic pathway, its absence may also lead to deterioration of its interacting partners or the complete loss of the associated pathway.

The *R. prowazekii* genome is known to contain at least 12 pseudogenes, with the most detailed studies done on the *metK* gene coding for AdoMet synthetase (Andersson and Andersson, 1999a, 1999b). This essential gene
involved in the biosynthesis of S-adenosylmethionine (SAM), contains a
termination codon at a central position disabling its function. Comparative
analyses have shown that this gene has been inactivated several times inde-
pendently in different Rickettsia lineages. Reconstruction of the ancestral
metK gene sequence using multiple alignment of several Rickettsia se-
quences, revealed that most mutations were deletions of a single or a few
nucleotides, with deletion generally predominating over insertions (Anders-
son and Andersson, 1999a). A model explaining metK pseudogene forma-
tion, hypothesises that the common ancestor between the TG and SFG had a
functional metK gene, which was subsequently inactivated due to the inven-
tion of a novel import system for Ado-Met.

1.5.3 Split genes in R. conorii
The analysis of the R. conorii putative ORFs revealed numerous instances of
consecutive ORFs matching consecutive segments of a single longer ORF in
other species (Ogata et al., 2001) (Figure 3). 37 Split genes scattered across
105 ORFs were identified. Most of the ORFs retained the statistical prop-
ties such as coding potential and codon bias of normal coding regions and a
good similarity with intact protein orthologs, with the authors advocating the
more neutral term, "split genes". Analysis of their transcription patterns re-
ealed that some of these fragmented genes are still being expressed, which
may indicate continued usage of the promoter of the original gene, suggest-
ing that split genes may have retained some of their original functions.
1.5.4 Evolution of coding sequences

Although, Rickettsia have received much attention for studies on degradative evolutionary processes and degrading genes, studies on protein-coding sequences should however not be underscored and could be an interesting means to understand the general pervasiveness of degradative processes and how they affect coding sequence evolution.

In general, it is accepted that reductive evolution is a major mode of sequence evolution in Rickettsia genomes. This was evident from the complete genome sequence of R. prowazekii and later from the R. conorii and R. typhi genomes. Although, massive gene decay and loss have been reported for these genomes, it is still not clear whether genome shrinkage also affects coding sequence length.

A comparative analysis between the orthologous gene sets of the aphid endosymbiont Buchnera and the free-living bacterium Escherichia coli, has shown that Buchnera genes are generally shorter than their their free-living counterparts (Charles et al., 1999). This was attributed due to the occurrence of short deletions of approximately 8 bp. However, it was concluded that this "gene shortening" phenomenon was unlikely the cause for the massive ge-
nome size reduction in *Buchnera*, accounting only for approximately 0.8% of the estimated genome size reduction.

Results from a global survey done on deletional bias on microbial genomes, surprisingly indicate that the average sequence coding length of small, intracellular genomes are similar to that of larger, free-living genomes (Mira et al., 2001). The average length of coding sequences in the *R. prowazekii* genome is in good agreement with that reported for most bacterial genomes of approximately 1000 bp (Mira et al., 2001). However, *R. conorii* displays on average shorter coding sequences (746 bp), reflecting the bias due to the presence of small open reading frames or so-called "split genes" in this genome.

1.5.5 Those mysterious little ORFans in *Rickettsia*

It is estimated that about one third of each completed genome are home to genes with no sequence similarity to any previously characterised gene or protein. Since these genes are generally thought to be responsible for genus-specific characteristics, an investigation into their evolutionary origin might resolve not only their identity, but also their authenticity as coding sequences.

Presumably, ORFan genes do not have any evolutionary parents or relatives. Their presence in the growing number of completed genomes have thus recently sparked the interest of researchers in more ways imaginable, to such an extent that their enigmatic presence in genomes could be considered unsettling and even troublesome (Skovgaard et al., 2001; Ochman, 2002). Troublesome in the sense, that ORFan genes are not easily detected by current gene prediction programs and annotation systems, and unsettling in the sense that their coding potential remains largely unknown. However, recent studies are beginning to shed light on their existence in genomes, sparking interest into the nature of gene evolution and the process whereby genomes are degraded, but also raising fundamental questions of practical value, on how to validate potential protein-coding sequences generated from genome sequencing projects (Schmidt et al., 2001; Skovgaard et al., 2001; Ochman, 2002).

Presumably, the large number of ORFan genes may be due to either (i) restricted phylogenetic distribution to certain evolutionary lineages as a result of lineage-specific gene loss, (ii) rapid divergence because the proteins they encode are unconstrained in their sequence evolution (iii) ORFans may represent membrane proteins (iv) they are purely fortuitous ORFs resulting from the messy nature of the transcription process (Finta et al., 2001).

In characterising genes with unknown function, caution should be exercised in making judgements about their coding status. In general, ORFan genes are shorter than genes for which a functional homolog can be clearly identified (Mira et al., 2002). This potentially also complicate methods
which rely on sequence length such as codon usage for validating their coding status, earning them their reputation as ELFs, or in short 'evil little fellows' (Basrai et al., 1997; Ochman, 2002; Lawrence, 2003).

Many annotated ORFs generated from sequencing projects are considered to be no more than putative reading frames, represented by an in-frame start codon separated by an ample distance from a stop codon. Sequence length, thus comprises one of the key criteria used for annotating ORFs in many genome projects (Salzberg et al., 1998). Since the length of coding DNA segments are known to be under both functional and structural constraints, but also compositional constraints and the base composition of translation stop codons (TAG, TAA and TGA) are known to be biased towards a low GC content, a differential density of these termination signals is expected in random sequences of different base composition (Oliver et al., 1996; Li, 1999; Carpena et al., 2002). The expected length of reading frames in random sequences is thus a function of GC content i.e. the higher the GC content, the lower the density of stop codons and therefore the longer, the expected reading frames. Genomes which thus exhibit a strong AT-bias, such as that of intracellular bacteria e.g. M. genitalium, Rickettsia and Buchnera are therefore expected to have on average a higher stop codon density than GC-rich organisms, which may also aid in the misidentification and thus over-annotation of genome sequences (Skovgaard et al., 2001).

A recent survey done on unknown putative ORFs in microbial genomes indicate that the vast majority of ORFs studied, even those that are short, are indeed genuine protein-coding regions (Ochman, 2002). However, this study also eluded to the presence of unknown small, high Ka/Ks ORFs, with atypical codon usage patterns, indicating that some of these may not be genuine protein-coding regions.

1.6 Helicobacter pylori

Evolutionary studies of bacteria show that they have evolved many mechanisms to survive and prosper in unfavourable environments. Survival in the gastric environment seems to represent one of the most important and complex traits of the phenotype specific to H. pylori.

Helicobacter pylori is a micro-aerophilic, gram-negative, slow-growing, spiral-shaped and flagellated organism which belongs to the epsilon proteobacteria (Tomb et al., 1997). H. pylori is probably the most common chronic bacterial infection of humans, present in almost half of the world’s population. The presence of the bacterium in the gastric mucosa is associated with chronic active gastritis, and is also implicated in more severe gastric diseases such as peptic ulcers and mucosa-associated lymphoid tissue lymphomas (Cover and Blaser, 1996).
*H. pylori* is unusual among pathogenic bacteria in its ability to colonise host cells in an environment of high acid pH, while only transiently being subjected to extreme pH (pH~2). The ability to establish a positive inside membrane potential and subsequently to modify its microenvironment is one of the crucial factors for its survival. Its most characteristic enzyme is a potent multisubunit urease that plays a vital role in both survival at acidic pH and successful colonisation of the gastric environment (Tomb et al., 1997).

### 1.6.1 Comparative Genomics of *H. pylori*

The completion of the genome sequences for *H. pylori* 26695 (Tomb et al., 1997) and subsequently *H. pylori* J99 (Alm et al., 1999) marked the importance and value of *H. pylori* as a human pathogen. In addition, the availability of two complete genome sequences can provide valuable insights into pathogenesis, acid tolerance and antigenic variation which characterises the genus *Helicobacter*.

From the comparative genomics study of the two *H. pylori* strains, it was concluded that the overall genomic organisation, gene order and predicted proteosome of the two strains are quite similar except for minor inversions and translocation (Alm et al., 1999). Between 6-7% of the genes are specific to each strain, with almost half of these genes being clustered in a single hypervariable region.

In brief, some of the most noticeable highlights of the comparison revealed the essential pathogenic character of *H. pylori* for example cagPAI-the cag pathogenicity island genes which are associated with the CagA antigen and upregulation of interleukin (IL-8) in gastric epithelial cells, vacA-vacuolating cytotoxin protein which is known to induce multiple effects on epithelial and lymphatic cells such as vacuolation with alterations of endolysosomal function, anion-selective channel formation, mitochondrial damage, and the inhibition of primary human CD4(+) cell proliferation (Alm et al., 1999; Wada et al., 2004).

Spurred on by the availability of genomic, protein-protein interaction (Rain et al., 2001) and gene expression data (Thompson et al., 2003; Merrel et al., 2003), *H. pylori* may be regarded as an excellent model organism for evolutionary genomics studies.

### 1.7 What determines the rate of protein evolution?

Finding the determinants for the rate of a protein’s evolution remains one of the most interesting tasks in the study of molecular evolution. The idea of a relationship between rates, mode of evolution, and function have always intrigued researchers and have recently gained momentum due to the large volume and variety of data available from various genome and “omics”-
related projects. To this end, various studies have sought to explain the evolutionary rate of proteins by examining gene location (Williams and Hurst, 2000), gene expression (Wagner 2000; Hooper and Berg, 2003; Subramanian and Kumar, 2004), phyletic distribution/phylogenetic conservation profiles (Krylov et al., 2003), connectivity in biological interaction networks (Fraser et al., 2002; Jordan et al., 2003; Hahn et al., 2004), and essentiality or lethality (Hirsh and Fraser, 2001; Jordan et al., 2002; Yang et al., 2003).

1.7.1 Chromosomal neighbours

There is increasing evidence to suggest that the gene order in organisms is not always random. It is known that proteins of linked genes evolve at comparable rates (Williams and Hurst, 2000), and that natural selection may promote the conservation of linkage of coexpressed genes (Hurst et al., 2002). Evidence of coevolution of gene order and recombination rate have also been investigated and it was found that essential genes are clustered in regions of low recombination (Pal and Hurst, 2003). However, evidence has been lacking for an indication that favourable gene arrangements are preserved by selection more than expected by chance.

A study correlating gene mutation rates with their location in the genome confirmed the existence of regional mutation rates with certain classes of genes showing a tendency to congregate in mutational “hot spots” (i.e. regions with high mutation rates) which mostly include genes involved in immune system response while others gravitate towards “cold spots” (i.e. regions with relatively low mutation rates) containing “housekeeping” genes (Chuang and Li, 2004). It thus, appears that natural selection can also operate at the level of gene location, segregating genes according to their function.

1.7.2 Biological networks

Studies correlating protein connectivity with evolutionary rates have been rather controversial. At first glance it seems to propose an elegant explanation to the evolutionary rate variation of proteins by taking a network perspective on biology. The first study by Fraser and co-workers, seemed to imply a correlation between protein connectivity and evolutionary rates (Fraser et al., 2002). However, others have reported “a no simple dependence” between connectivity and evolutionary rates, stating only that either the most prolific interactors tend to evolve the slowest (Jordan et al., 2003), or that the observed correlation tended to be function-specific with genes involved in transcription and cell cycle showing significant correlations (Hahn et al., 2004).

Although evidence seems to be lacking, the idea is still compelling that genes with products that are involved in numerous protein-protein interac-
tions should be more constrained in evolution and thus tend to evolve more slowly. To that extent, it has been found that the most highly connected proteins in the yeast interaction network include a higher proportion of essential gene products than do proteins with fewer interactions (Jeong et al., 2001), and that these so-called “hub” proteins tend to be lost during evolution much less readily than proteins with fewer interaction partners (Wagner, 2001).

1.7.3 Protein essentiality/lethality
Studies have suggested that rates of protein evolution are correlated with protein dispensability and fitness effect (Hirsch and Fraser, 2001; Yang et al., 2003). Under the assumption that protein evolution is primarily caused by slightly deleterious amino acid substitutions (Ohta 1998), rates of evolution are predicted to be lowest with genes with the largest fitness contribution. Not surprising it has been found that genes with the largest individual fitness contributions are those whose products are highly connected in interaction networks. Those proteins which are highly connected proteins are also known to be lethal (Jeong et al., 2001).

In brief, studies indicate that essential genes are more evolutionary conserved than nonessential genes and thus expected to evolve significantly slower than nonessential genes (Jordan et al., 2002).

1.7.4 Phylogenetic conservation profiles
The phylogenetic distribution of genes can also give an indication of their evolutionary and functional importance (Pellegrini et al., 1999). Genes which are known to be phylogenetically conserved across different phyla are thought to form the core set of genes which are essential for life. These genes are thought to be the most conservatively evolving genes.

The tendency of a gene to be lost and sequence evolutionary rate can be considered two variables that characterise the evolutionary conservation of a gene. Genes that have a lower propensity to be lost during evolution are known to accumulate fewer substitutions and tend to be essential for organism viability. These genes also tend to be highly expressed and have many interaction partners (Krylov et al., 2003). If the most connected genes are the most conserved across species, then “speciation genes” should preferentially be found amongst those of low connectivity, which are expected to be the predominant targets of diversifying selection (Aris-Brosou, 2004). Indeed, specie-specific genes are less connected genes (Jeong et al., 2000).
In order to gain a better understanding of the causes of substitution frequency variation in human pathogenic bacteria, examples will be presented of:

- degradative forces shaping genome evolution using *Rickettsia* as our model system.
- evolutionary forces responsible for creation and destruction of ORFan genes in *Rickettsia*.
- adaptive evolution in *H. pylori* using the enzyme, carbamoyl phosphate synthetase as a case study.
- factors determining protein evolutionary rates in pathogenic bacteria such as *H. pylori*.

The aim of this work is therefore to gain a deeper understanding of the driving forces that underlie substitution frequency variation in pathogenic bacteria by examining information derived from sequence analysis.
3 Methodology

3.1 The importance of sequence alignments in biology

A sequence alignment can be considered an arrangement of two or more biological sequences such as protein or DNA sequences highlighting their similarity. In performing a sequence alignment, the main idea is to align sequences such that columns within an alignment contain identical or similar characters and gaps are inserted where possible to maximise the number of aligned characters. The underlying assumption of a sequence alignment is mean to reflect a biological hypothesis about the common evolutionary origin of the sequences involved. Mismatches in the alignment is meant to signify mutations and gaps- insertions or deletions.

3.1.1 Types of sequence alignments

Basically, sequence alignments can be performed in a pairwise or multiple sequence alignment fashion. Methods employing pairwise sequence alignments are concerned with finding the best-matching piecewise (local or global) alignments considering only two sequences at a time. A multiple alignment could be considered as an extension of pairwise alignment incorporating several sequences and highlighting regions common between them all.

Another important issue that needs to be considered is whether a global or local alignment needs to be performed. A global alignment between two sequences considers all of the characters in both sequences, aligning characters from end-to-end. Global alignments are useful mostly for finding closely related sequences. Local alignment methods find related regions within sequences, consisting of only a subset of the characters within each sequence and are thus suitable for detecting distantly related sequences. Local alignment as opposed to global alignment methods are more sensitive and has the advantage of finding short conserved regions such as domains.

Examples of the most frequently used pairwise local alignment search tools include BLAST (Altschul et al., 1990; Altschul et al., 1997) and FASTA (Pearson and Lipman, 1988). For multiple sequence alignments, the ClustalW (Thompson et al., 1994) package remains the favourite. In addition, more sensitive methods such as PSI-BLAST (Altschul et al., 1997) and HMMER (Eddy, 1998) based on constructing sequence profiles of multiple sequence alignments are also available.
3.1.2 Assessment of sequence alignments

Two important questions arise when assessing the significance of sequence alignments, namely 1) how is the best alignment between two sequences (or regions of sequence) chosen? 2) how are alignments between the query and the numerous sequences in the database ranked? These two questions are in part related to each other.

It is important to realise that the actual biological meaning of any alignment can never be absolutely guaranteed. However, statistical methods can be used to assess the likelihood of finding an alignment between two regions (or sequences) by chance, given the size of the database and its composition. The first can be addressed by developing a model of how likely certain changes between characters in the sequences are. These models are derived empirically using related sequences, and are expressed as substitution matrices e.g. PAM (point accepted mutations) (Dayhoff et al., 1978) and BLOSUM (blocks substitution) (Henikoff and Henikoff, 1992) matrices. These matrices are used by the algorithms to give each possible alignment between two sequences a score. The highest-scoring alignments possible are reported by the algorithm. The actual biological quality of the alignments thus depends upon the evolutionary model used to generate the score.

The second question is purely statistical. It is generally accepted that the scores of alignments between random sequences follow an extreme value distribution (Altschul et al., 2001). Pairwise alignment programs such as BLAST estimate the parameters of this distribution for a particular parameter set (consisting of the query, database, substitution matrix and certain other parameters) using simulation methods. Alignments can thus be given a statistical significance value, allowing judgements on possible relationships between sequences to be inferred.

3.2 Gene prediction

The importance of a comprehensive description of the information content of a genome is obvious. Therefore, one of the first steps in the analysis of a microbial genome is the identification of all its genes. Since microbial genomes typically contain 90% coding sequence, we can rely on computational methods to differentiate the putative genes in each of the six possible reading frames. Computer-aided approaches to gene prediction can be divided into “intrinsic” and “extrinsic” approaches (Borodovsky et al., 1994). The intrinsic approach includes evaluation of certain properties of DNA sequences without explicit referral to other sequences. Some of the relevant properties include the length of the ORF, codon usage, the presence or absence of the ribosomal binding site (Shine-Dalgarno) sequence an appropriate distance upstream of the initiation codon, and various more subtle characteristics that
are believed to be typical of expressed genes as opposed to noncoding regions. The extrinsic approach to gene prediction includes a comparison of the putative encoded amino acid sequence with protein sequence databases and a search for functional motifs. If a putative ORF product shows significant sequence similarity to one or more proteins in the databases, it is almost certain that the ORF in question is a real gene. In practice, the extrinsic approach is usually regarded as the first line of recourse in genome annotation.

3.2.1 Extrinsic approaches to gene prediction

Sequence similarity has always been considered a strong indicator of gene coding potential. The most reliable way of identifying a gene in a new genome is to find a close homolog from another organism. This can be done today very effectively using sequence similarity search programs such as BLAST and FASTA to search all the entries in current databases such as Genbank. The computer program BLASTX available in the family of BLAST programs, can perform conceptual translation of a nucleotide query sequence against a protein database in one programmatic step. BLASTX has been considered appropriate for use in moderate and large scale sequencing projects at the earliest opportunity for the recognition of coding regions even in the presence of substitution, insertion and deletion errors in the query sequence and to sequence divergence. In addition, it also aids in the reliable identification of reading frames.

Although this method is accurate for homologous genes that have been identified, novel genes will not easily be found. Many genes in new genomes still have no significant homology to known genes. For these genes, we must rely on computational methods of scoring the coding region to identify the genes.

3.2.2 Intrinsic approaches to gene prediction

Markov models are well-known tools for analysing biological sequence data, and the predominant model for microbial sequence analysis is a fixed-order markov chain (Eddy, 1998; Krogh et al., 1994a, 1994b; Bessemer et al., 2001). A fixed order markov model predicts the next base in a sequence according to immediate observing a fixed number of previous bases in the sequence. To use a markov model to find genes in microbial DNA, we need to build at least six submodels, one for each other the possible reading frames (three forward and three reverse), and also a seventh model for non-coding regions, although this is may not be strictly necessary.

The GLIMMER system identifies potential genes using a combination of interpolated markov models (IMM) from 1st-8th order, weighting in each model according to its predictive power (Salzberg et al., 1998; Delcher et al., 1999). The IMM in GLIMMER does this by combining probabilities from
contexts (oligomers) of varying lengths to make predictions only using those contexts for which there is sufficient data. This means that when the statistics on longer oligomers are inadequate to make good estimates, an IMM can use the shorter oligomers to make predictions. After creating IMMs for all six reading frames plus noncoding DNA, we can produce an algorithm for finding genes. Simply score every ORF using all seven models, and choose the model with the highest score. All ORFs longer than a certain length and score obtained are taken into account and those that score higher than a certain threshold are processed further. They are examined for overlaps of a certain length. If one is found, the overlap region is scored again separately and the highest score of the overlap is combined with the highest score of the putative gene.

Although the GLIMMER system has a good reported overall accuracy and sensitivity, pitfalls do exist. For example the reliable identification of start codons of some organism is not well-understood and inputted appropriately in GLIMMER and therefore detection of of putative genes will not be as successful. The GLIMMER system only finds the genes that conform to the rules established by your IMM. Any exceptions or unusual genes, such as polycistronic genes will not be well characterized by GLIMMER.

An approach that advocates combining intrinsic and extrinsic approaches to gene prediction therefore has the potential of enhancing the reliability of the results obtained with each of them and is important for extracting the maximum amount of information from genome sequences. Here we present the results of our analysis of the complete set of unannotated sequences located between annotated genes using five *Rickettsia* spp strains using GLIMMER for predicting putative ORFs and BLAST as well as motif searches for detecting sequence similarities to proteins in current databases. The combination of these two approaches is featured as a strategy for identifying new genes in bacterial genomes (Papers II, IV).

### 3.3 Detecting and resolving orthologs

#### 3.3.1 Homologs, paralogs and orthologs

An important concept in molecular evolution is that of homology. We use homology in the evolutionary sense of the word: two sequences or structures are homologous if and only if they acquired that state directly from their common ancestor. A key point in determining whether a feature is homologous requires knowledge of the evolutionary relationship among the species having that feature. If two or more species are closely related, then the simplest interpretation is that they share features due to homology rather than sharing the feature due to independent evolution. Thus in some sense, different genes within a gene family maybe homologous because they are descen-
dants of the same gene. Further, we can distinguish between two basic types of homology among genes: orthology and paralogy. Paralogous genes are descendants of an ancestral gene that have undergone one or more duplications. If there are no duplications in the gene tree then the sequences are orthologous i.e. genes which are related by a speciation event. Orthologous proteins are generally assumed to have the same function, whilst this may not be the case for paralogous proteins which may diverge in function after a gene duplication event (Lynch and Connelly, 2000a, 2000b; Lynch and Force, 2000).

3.3.2 Gene families
Many genes are members of gene families - suites of genes that are the descendants of an ancestral gene. It is well known that members of the same protein family may possess similar or identical biochemical function (Hegyi and Gerstein, 1999). Protein families can be defined as those groups of molecules which share significant sequence similarity. To detect a protein family, algorithms should take into account all similarity relationships in a given arbitrary set of sequences, a process that is defined as “sequence clustering”. This approach is usually based on grouping homologous proteins together via a similarity measure obtained from direct sequence comparison. Ideally, the resulting clusters should correspond to protein families, whose members are related by a common evolutionary history. Well-characterised members within a family can hence allow one to reliably assign functions to family members whose function are not known or well understood. Many methods are currently available for clustering proteins into families. These methods generally rely on sequence similarity measures such as those obtained by BLAST or other database search methods.

3.2.3 Sequence clustering using the TRIBE-MCL algorithm
TRIBE-MCL algorithm has been reported to be an efficient and reliable method of sequence clustering for large data sets (Enright et al., 2002). TRIBE-MCL is based on Markov cluster (MCL) algorithm previously developed for graph clustering using flow simulation. An ideal clustering method would require sequence similarity relationships as input and be able to rapidly detect clusters solely related using this information. Traditionally, most methods deal with similarity relationships in a pairwise manner, while graph theory allows the classification of proteins into families based on global treatment of all relationships in similarity space simultaneously.

In this section, we describe how the algorithm relates to the clustering of proteins into protein families. Biological graphs may be represented as nodes i.e. sets of proteins and edges representing the similarity between these proteins. Edges can be weighted according to a sequence similarity score ob-
tained from BLAST. A markov matrix is constructed representing transition probabilities from any protein in the graph to any other protein for which a similarity has been detected. Each column represents a protein, and each entry within a column represents a similarity between this protein and another protein. The entries in the markov matrix are probabilities generated from weighted sequence similarity scores.

In a biological sense, we expect that members of a protein family will be more similar to each other than to proteins in another family. Ideally the grouping of genes into families should be based on their function and evolutionary history.

3.4 Evolutionary substitution rates

3.4.1 What is $Ka/Ks$?

Imagine you align the sequences of the same gene from two species. There will usually be differences between the sequences i.e. evolution! Some of these will lead to differences in the amino acids of the encoded protein (nonsynonymous changes) and some, because of the degeneracy of the genetic code, leave the protein unchanged (synonymous, or silent changes). Counting the number of each and the number of synonymous and nonsynonymous sites and adjusting these figures, one can calculate two normalized values, $Ka$ (or $dN$), number of nonsynonymous substitutions per nonsynonymous site, and $Ks$ (or $dS$), number of synonymous substitutions per synonymous site. Due to the degenerate nature of the code, only about 25% of the possible changes in our sequence are synonymous. If selection does not act on silent sites, from the neutral theory of evolution, $Ks$ should therefore be proportional to the mutation rate of the gene.

However, as sequences diverge over time, the observed number of changes underestimates the real number of changes. Fortunately, the extent of real divergence can be estimated from the total observed amount of divergence using so-called “multi-hit correction” methods. However, none of these methods can work miracles: as the number of changes increases, the amount of information from the alignment decreases and we approach saturation, in which case the data becomes useless. So, if corrections are made for unequal nucleotide frequencies, codon bias and degeneracy of the code, I should have a method that reports the number of nonsynonymous changes at each possible nonsynonymous site as the same as the number of synonymous changes per synonymous site, i.e. $Ka/Ks= 1$. Deviations from a ratio of one will then tell me something about the selective forces acting on the protein, given that $Ks$ is telling me the background rate of evolution (Hurst, 2002).

In practice, the types of substitutions which changes a protein are less likely to be different between two species than those which are silent; most
of the time selection acts to eliminate deleterious mutations, keeping the protein as it is. However, in a few instances (often when immune system genes co-evolve with parasites), we find that $K_a$ is much greater than $K_s$ (i.e. $K_a/K_s >> 1$). This is strong evidence that selection has acted to change the function of the protein (positive selection/ adaptive evolution).

3.4.2 Methods of estimating $K_a/K_s$

Estimation of synonymous and nonsynonymous substitution rates is important in understanding the dynamics of molecular sequence evolution and determining the selection pressures that have shaped genetic variation. Traditionally, synonymous and nonsynonymous substitution rates are defined in the context of comparing two DNA sequences i.e. a pairwise approach. However, if more sequences are available, a multiple sequence alignment and a phylogenetic tree can be constructed and used for estimating synonymous and nonsynonymous substitution rates.

Methods for estimation of $K_a$ and $K_s$ between sequences can be classified into counting-based methods (Li, Wu and Luo, 1985; Nei and Gojobori, 1986; Li, 1993; Pamilo and Bianchi, 1993; Ina, 1995; Yang and Nielsen, 2000a, 2000b) and maximum likelihood methods under codon substitution models (Goldman and Yang, 1994; Muse and Gaut, 1994; Muse, 1996; Nielsen and Yang, 1998). Counting methods usually involve three steps:

1. counting the number of synonymous and nonsynonymous sites in the two sequences
2. counting the synonymous and nonsynonymous differences between the two sequences
3. correcting for multiple substitutions at the same site to calculate the numbers of synonymous ($K_s$) and nonsynonymous ($K_a$) substitutions per site between the two sequences.

Although this strategy appears to be simple, well-known features of DNA sequence evolution, such as unequal nucleotide or codon frequencies, make it a real challenge to count sites and differences correctly. In general, counting methods are safe to use if codon usage (especially at the 3rd codon position) is uniform, the sequences are not very divergent, and transition and transversion rates are similar. In addition they are computationally fast and thus suitable for large-scale analyses.

Likelihood methods that employ codon-based models have been developed that describe the evolution of coding sequences in terms of both DNA substitutions and the selective forces acting on the protein product (Goldman and Yang, 1994; Muse, 1994; Nielsen and Yang, 1998; Yang et al., 2000a, 2000b). Likelihood methods use a likelihood score as an optimality criterion, calcu-
lated according to a specific model of evolution. An explicit model allows the incorporation of knowledge of the mechanisms and constraints acting on coding sequences, as well as the possibility of comparing the performance of different models. Obviously, choosing a more realistic evolutionary model will result in more accurate reconstruction. Goldman and Yang developed a likelihood method for estimating the rates of synonymous and nonsynonymous substitution considering a nucleotide substitution model of 61 sense codons (excluding the 3 nonsense codons) (Goldman and Yang, 1994). In their codon-based model, provision is made for three categories of amino acid sites:

1. conserved sites
2. sites evolving neutrally
3. sites that are targets of positive selection.

Yang’s method uses the likelihood ratio test to compare the performance of different likelihood models, and to determine if the $Ka/Ks$ ratio for the lineage of interest is elevated compared to other lineages in the phylogeny. In brief, likelihood ratio tests allow the comparison of different models that are nested with respect to one another, with more complex models fitting the sequence data better than simpler models as judged by the likelihood score of each model. The original codon-based models assumed equal nonsynonymous and synonymous rate ratios among sites and lineages. Subsequent models have allowed the ratio to vary across lineages and among sites in the protein, and have even incorporated unequal frequencies of different types of nonsynonymous substitutions based on the amino acids involved (Nielsen and Yang, 1998).

When performing $Ka/Ks$ analysis, some caveats have to be kept in mind:

1. results based on pairwise comparisons of sequences does not take a phylogeny into account and may therefore be less powerful
2. synonymous sites are assumed to be neutral which may not be the case if constraints are imposed on codon usage or secondary structure.
3. the requirement that $Ka/Ks >> 1$ may be too strict for detecting cases of positive selection in which selection acts only on a few amino acids rather than the entire protein sequence.

Methods for estimation of $Ka$ and $Ks$ that take a phylogeny into account, are usually regarded as more powerful since it allows selection to be identified at a single amino acid level and allows detection of rate variation along branches of a phylogenetic tree (Suzuki and Gojobori, 1999). These methods also have the advantage of inferring ancestral sequences as an intermediate step in estimation of substitution rates. However, methods that employ maximum likelihood are also computationally more intensive.
3.4.3 Applications of $Ka/Ks$

The ratio of $Ka/Ks$, is generally regarded as an indicator of the relative strength of selection acting on protein-coding sequences, and can therefore be used for determining the functional status of a sequence. This comparative approach can also provide a rapid means to establish whether an ORF is evolving in a typical manner of a protein-coding region. Because selective constraints differ between synonymous and nonsynonymous sites, ORFs that specify functional proteins tend to have a much higher divergence at synonymous sites than nonsynonymous sites. Similarly, in sequences that lack functional constraints such as pseudogenes or misannotated regions, values of $Ka$ and $Ks$ are expected to be the same, and for protein-coding regions undergoing adaptive evolution (positive selection), $Ka/Ks$ ratios will exceed one.

Traditionally, the $Ka/Ks$ ratio has most frequently been used in studies for detecting genes under positive selection (Endo et al., 1996). Innovative application of the uses of the $Ka/Ks$ ratio have also been reported. For example, Nekrutenko and co-workers have championed the use of $Ka/Ks$ ratios as an aid in gene recognition and identification for reliably predicting protein-coding regions (Nekrutenko et al., 2002). Ochman suggested the use of differential rates of evolution between nonsynonymous and synonymous sites to assist in identification of genes in bacterial genomes (Ochman, 2001). These method can clearly be of use when the genome sequence of two or more closely related organisms are available for analysis, increasing the reliability of annotation of small ORFs and thus aid in distinguishing protein-coding sequences from fortuitous ORFs.

In our studies, we have used the $Ka/Ks$ test for inferring the functional status and authenticity of genuine protein-coding sequences and ORFans, but also for distinguishing real ORFans from small fortuitous open reading frames (Papers I, II). The $Ka/Ks$ ratio test was also applied for detecting potential cases of genes under positive selection (Paper V). In our studies, we have mostly used the PAML package for estimation of evolutionary rates (Yang, 1997).

3.5 Reconstruction of ancestral sequences

By tracing the evolutionary history of molecular changes using phylogenetic methods, we can obtain powerful insights into how and why molecules work the way they do, but also begin to understand the evolutionary events from which they have arisen. Thus, present-day protein sequences can serve important roles to infer and resurrect their ancestral sequences.
3.5.1 Methods of ancestral sequence reconstruction

Several statistical methods have been developed to infer ancestral amino acid sequences from sequences of present day species when the phylogenetic relationship is known (Yang et al., 1995; Koshi and Goldstein, 1996; Pupko et al., 2000; Pupko et al., 2002). Among these, the maximum parsimony method has been used most frequently. Parsimony methods evaluate the phylogenetic relationships and ancestral state assignments based on the evolutionary changes along the branches of the phylogenetic tree, specifically trees or ancestral states that require the fewest changes are preferred. In the parsimony method, each amino acid site at each interior node of the tree is determined so as to make the total number of amino acid changes at the site minimal. Because no information about the branch lengths and the pattern of substitution is used, the inferred ancestral sequences may not be that reliable.

To infer the ancestral sequences more accurately, Yang and co-workers introduced a maximum likelihood method (Yang et al., 1995). In this method, the branch lengths of the phylogenetic tree are estimated by the maximum likelihood method, and the posterior probability of each assignment at ancestral nodes is computed at each site. The amino acid assignment with the highest posterior probability is chosen as the best set of ancestral amino acids.

The parsimony method is quite effective in inferring ancestral sequences when sequence divergence is low. However, when comparing distantly related sequences it gives several possible sequences for the ancestral protein, thus making it difficult to determine the most probable ancestral sequence. Since no consideration is given to branch lengths, erroneous results may be reached as the branch lengths increase with increased number of substitutions. When the extent of sequence divergence is high, the maximum likelihood approach may however give more accurate results.

In certain cases where the sequence divergence is extremely low, it may suffice to use the consensus sequence of an alignment. Although, this approach does not explicitly take an evolutionary model into account, it does follow Occam’s razor i.e. choosing the simplest model making the least assumptions which forms the basis for the principle of parsimony. However when sequences are highly divergent, the consensus sequence will contain too much uncertainty at particular sites for a reliable estimation.

3.5.2 Uses of ancestral sequences

Knowledge of ancestral sequences is helpful for understanding evolutionary processes, protein classification and biological function of protein families. Additionally, reconstructed ancestral protein sequences could serve to fill in the sequence space thus aiding remote homology inference and functional prediction (Collins et al., 2003; Cai et al., 2004). For homology detection,
ancestral sequences represent an enlargement of the sequence space around native sequences, which could improve detection of distant homologs in databases.

In our analyses, we have used contemporary nonfunctional sequences to reconstruct ancestral sequences which we presume to have previously been functional in their evolutionary history (Paper II). This situation is however more difficult since these sequences frequently have accumulated deletions and/or insertions during their course of degradative evolution, thus removing all existence of an intact reading frame and coding potential. We proceeded by translating sequences in all reading frames, and searching for matching full-length orthologous proteins and accommodating for deletions and insertion in such a way so that the aligned protein sequences matched as closely as possible. Deletions and insertions received equal weighting. This was performed for both pseudogenes and short ORFs. For pseudogenes a corresponding full-length gene was used for reconstruction. However, for segments containing short overlapping but nonidentical ORFs, no full-length gene was available, thus the longest ORF with the smallest number of mutations in each of the aligned sequences were used. Reconstructed genes having no sequence similarity were labeled as fossil-ORFs.

In conclusion, we have effectively used reconstructed ancestral sequences to infer evolutionary events leading to gene inactivation and eventual degradation, but also for homology detection and comparative protein structure modelling (Paper II).

3.6 Construction of biological interaction networks

A graph is a pictorial representation of a mapping between a set of nodes (or vertices) and a set of links (or edges) that connect pairs of distinct nodes. Graphs can be unweighted or weighted i.e. the links may contain additional information such as a distance metric representing the strength of an association between the respective nodes. Depending on the nature of the interactions, graphs may also be undirected or directed i.e. the links may have a well-defined direction in which case we distinguish between a source node and destination node. In terms of representing graphs computationally, two data structures frequently used are the adjacency matrix and adjacency list representation. In the adjacency matrix representation, a square matrix is constructed between columns and rows that represent the interacting nodes. In the adjacency list representation, a list is constructed between nodes which interact. The adjacency matrix is regarded as efficient in terms of space storage when dealing with densely connected graphs. However for sparse graphs, the adjacency list may be a more efficient data structure.

In our study, metabolic networks were reconstructed based on data obtained from the LIGAND database (Gotoh et al., 1998). In brief, for the
metabolic interaction network a directed graph was constructed consisting of
nodes representing enzymes that are connected by sharing a common me-
tabolite in an enzymatic reaction. For the *H. pylori* protein-protein interac-
tion network, an undirected graph was reconstructed using data obtained
from published yeast two hybrid results (Rain et al., 1999). Biological inter-
actions networks were visualised using PAJEK (Batagelj and Mrvar, 1998)
(Paper VI).

3.7 Comparative protein structure modelling

Proteins from different biological sources and sometimes diverse biological
functions can have similar sequences, and it is generally accepted that high
sequence similarity is reflected by distinct structural similarity (Wood and
Pearson, 1999). Comparative protein structure or so-called homology modelling
requires at least one sequence of known three dimensional structure, a
template sequence, with significant similarity to the target sequence for
which the desired model is to be built. The comparative modelling approach
mainly involves the extrapolation of the structure for the target sequence
from the known 3D-structure of related templates family members. In order
to determine if a modelling request can be made, one compares the target
sequence with a database of sequences derived from the Brookhaven Protein
Data Bank (PDB), using sequence similarity search programs such as
FASTA or BLAST to obtain a list of candidate templates. A structurally
corrected multiple sequence alignment is then generated by aligning each
residue of the reference structure to a residue of template structure if their Ca
atoms are located within 3.0 Å. In brief, the steps which are generally fol-
lowed during protein structure modelling are:

1. model framework generation
2. construction of loops
3. completion of the backbone and side chains
4. refinement of the primary model to ultimately yield the final model
structure.

Fortunately, nowadays most of these step can be automated and various
online proteins structure modelling servers are available for submitting your
protein sequence for structure prediction. In our comparative protein struc-
ture modelling of the anykyrin-repeat protein (Paper II, figure 5), we have
used the Structure Prediction Meta Server (Bujnicki et al., 2001) for protein
structure prediction of our reconstructed fossil ORF. The reconstruction of
an ancestral protein sequence combined with tertiary structure prediction
provided an invaluable approach for validating our assumption of previously
existing functional ancestral protein sequences.
3.7.1 Mapping mutations onto protein structure

The ability to visualise the effect of mutations on protein structure represents an invaluable tool to decipher and gain understanding of the effect of natural selection.

Substitutions between amino acids can be classified into radical or conservative substitutions based on the physico-chemical properties of the amino acids involved. For example, an arginine to lysine substitution would represent a conservative amino acid substitution preserving the charge, since both amino acids are positively charged. However, an arginine to aspartic acid substitution would represent a radical amino acid substitution since a positively charged amino acid is now exchanged with a negatively charged amino acid aspartic acid. Based on the assumption that substitutions between similar amino acids are more readily tolerated than between different classes of amino acids, Grantham constructed a matrix containing the difference in physico-chemical property values between amino acids (Grantham, 1974).

Using the Grantham physico-chemical distance matrix, we scored each amino acid substitution from a protein alignment representing CPS between \textit{H. pylori} J99 and 26695. Normalised values were then used to colour each amino acid from red to blue in the protein structure, which was then visualised using Rasmol viewer (Sayle and Milner-White et al., 1995) (Paper V).
4 Results

4.1 Gene degradation in *Rickettsia*

4.1.1 Small RNAs in *Rickettsia* - are they functional?

Studies of the transcription profiles of the split genes in *R. conorii* suggest that gene inactivation is a complex process. Analysis of gene expression patterns in *R. conorii* has shown that short ORFs inside deteriorating genes are occasionally transcribed into RNA with transcription sometimes re-initiated from inside the fragmented genes of *R. conorii* (Ogata et al., 2001). This suggests that promoters can either be created by mutations or recruited from existing sequences inside the fragmented genes. Transcription may however be less well-regulated. The use of new promoters inside deteriorating genes could lead to a temporary retention of a partial gene function, which could compensate for mutations. Alternatively, transcription of these fragments may be due to exposure of internal binding sites for RNA polymerase, with no functional consequence at the protein level. Since protein-coding sequence are thought to evolve under strong functional and structural constraints, analysing the ratio of nonsynonymous (*Ka*) to synonymous (*Ks*) substitutions will give an indication of the selective pressure operating on these sequences. To distinguish between these two alternatives, split genes were divided into the following categories listed below and we aimed to search for functional constraints on the expressed gene fragments in *R. conorii* by comparing substitution frequencies:

1. split genes vs full-length genes
2. split genes with different expression pattern characteristics
3. expressed vs unexpressed gene fragments.

To evaluate the functional status (and authenticity) of recognized ORFs, *Ka* and *Ks* values were computed for all orthologs. Based on analysis of 785 orthologous, full-length genes in *R. prowazekii* and *R. conorii*, we estimated the nonsynonymous substitution frequency to be 0.07 per site and the synonymous frequency per site to be 0.4. The corresponding *Ka* for 39 gene fragments derived from 13 split genes in *R. conorii* was estimated to be 0.16 substitutions per position (Paper I, table I).
1. Split genes have on average accumulated twice as many mutations at nonsynonymous sites as the full-length genes, suggesting that they have less functional constraints on evolution.

2. Split genes have two-fold higher fixation rates for mutations at nonsynonymous sites than the set of full-length orthologues. A higher substitution frequency seems to be a characteristics of split genes, irrespective of the pattern of transcription.

3. If the split genes are indeed not functional then we expect no difference in substitution frequencies between expressed and unexpressed ORFs inside fragmented genes.

To examine systematically whether there is a stronger selective constraint on the gene fragments that still produce mRNA, we compared the substitution frequencies for 26 expressed ORFs with those of 13 unexpressed ORFs. No difference was found between the two sets of genes, suggesting that the expressed genes have not been more functionally constrained than the unexpressed gene fragments.

However, a comparative analyses of orthologous genes between these two species indicate that the majority of protein-coding sequences in excess of 1500 bp (41%) are generally evolving under selective constraints ($Ka/Ks < 0.3$) and that ORFs displaying high $Ka/Ks$ values ($Ka/Ks > 0.4$) in general are represented by shorter protein-coding sequences or it may be that short ORFs in general may not be real genes.

Our analyses of the split genes show that substitution frequencies at nonsynonymous sites are similar for expressed and non-expressed parts of the fragmented genes, which supports the view that these gene fragments are no longer functional, but represent decaying remnants in their final stages of degradation. Their expression might be due to the formation of cryptic promoters, since bacterial promoter in general are known to be AT-rich. The observed expression may however be regarded as a temporary stage in the gene degradation process. We conclude that although transcription is maintained for some ORFs, these are accumulating mutations at the same high frequencies as the unexpressed ORFs.

4.2 Birth and death of ORFan genes in Rickettsia

To understand the origin and evolution of the ORFan sequences, we examined the evolutionary fate of a selected set of genes annotated as ORFans in R. conorii (Ogata et al., 2001). Here we have studied the rates and patterns of ORFan sequence evolution using Rickettsia as our reference system. We selected for analysis eight genomic segments covering 11 ORFans in R. conorii that are located at positions corresponding to intergenic regions in R.
prowazekii genome. These segments were amplified by PCR in *R. typhi* as well as *R. montana* and *R. rickettsii*.

A length distribution plot of 413 ORFans in *R. conorii* shows that these species-specific ORFs are much shorter than the 785 orthologous genes present in both *R. conorii* and *R. prowazekii* (Paper II, figure 1).

**4.2.1 Coding potential of intergenic regions in *Rickettsia* species**

A comparison of the aligned sequences of each of the eight segments revealed a considerable heterogeneity in size and coding content for these *Rickettsia* species (Paper II, table 1). The overall G+C content values of these segments were 23.7% in the TG and 30.7% in the SFG. The variation in the G+C content within as well as among species is related to the different coding potentials of these segments. It is also in accordance with previously observed differences in G+C content for genes and intergenic DNA in *Rickettsia* (Andersson S.G.E. et al. 1998b; Ogata et al., 2001).

**4.2.2 ORFan genes in the SFG correspond pseudogenes in the TG**

Three segments consists of known genes and ORFan genes in *R. conorii* that are conserved in size among members of the SFG (ampG, RCO295, RCO529), but are present as pseudogenes with multiple termination codons and frameshift mutations in the TG (Paper II, figures 2A-C). We reconstructed the putative gene sequences present in the common ancestor of the TG as the longest ORF, given the smallest number of mutations. These results suggest that ampG, RCO295, RCO529 were present in the common ancestor of *Rickettsia* but they have started to decay in *R. prowazekii* and *R. typhii*.

**4.2.3 ORFans as short, internal fragments of deteriorating genes**

Another set of ORFans in *R. conorii* display little or no sequence similarity to the corresponding fragments in the TG *Rickettsia* (Paper II, figures 2D-G). Most of these sequences correspond to short fragments of full-length genes with different parts of the ancestral gene being retained in different species. Taken together, these results suggest that lineage-specific recombination events at repeated sequences inside the ancestral gene have resulted in different short sets of species-specific ORFans.
4.2.4 ORFans as short, fused fragments of deteriorating genes
We have identified an ORFan in *R. montana* that consists of the 5′-terminal and 3′-terminal segments of two genes present in the genome of its ancestor (Paper II, figures 2G and 3C). This illustrates how a new, species-specific ORFan can be created by a single-deletion event that results in the fusion of the 5′- and 3′-flanking segments, which in this case consisted of two neighbouring genes.

4.2.5 Deletions and insertions in the TG and SFG *Rickettsia*
Estimating the frequencies of deletions in sequences for which no full-length copy is present in any of the modern species is even more difficult. In this case, the putative ancestral gene sequences were first reconstructed as the longest ORF given the smallest number of mutations in each species of the SFG *Rickettsia*. These inferred mutations were then counted as deletions and insertions in each individual species as compared to the reconstructed reference sequence.
In total, we estimated that less than 100 nt per kb (6-7%) have been eliminated from members of the SFG, with the exception of R. montana in which the deletion frequency may have been as high as 34%. Nevertheless the general trend is that there is a bias towards the deletion mutations in all the Rickettsia species examined. Furthermore, the data indicated recombination at short repeated sites has occurred at high frequencies (Figure 4).
4.2.6 Putative function of reconstructed genes

The genes reconstructed from the different pieces of short ORFans in the SFG display the characteristics G+C content statistics, suggesting that they were functional in the common ancestor of the SFG species. If so, what were the function of ancestral genes and why were they purged from the modern *Rickettsia* genomes? To resolve this questions, we searched the reconstructed protein sequences against GenBank and the Conserved Domain Database (CDD). The search against GenBank revealed a weak sequence similarity to a protein from *D. melonogaster* that contain multiple ankyrin repeats. Likewise, the search against CDD identified six ankyrin repeats in the reconstructed protein sequence.

![Figure 5. Representations of the three-dimensional protein structure of the consensus ankyrin repeat domain (A) and a comparative homology protein model of fossil-ORF F (B).](https://via.placeholder.com/150)

Comparative molecular modelling confirmed these results by revealing similarity to the crystal structure of the consensus ankyrin repeat domain that consists of a β-turn followed by two anti-parallel α-helices and a loop reaching the turn of the next repeat (Figure 5). The modelling approach suggest that the fossil-ORF F is a dimer of six ankyrin repeats with the overall protein topology being remarkably maintained, despite the low sequence identity (25%). The predicted ankyrin-repeat protein was predicted to contain 23 α-helices.
4.3 Positive selection scanning of *H. pylori*

We performed a survey in two pathogenic strains of *H. pylori* (J99 and 26695) in an effort to detect genes which may be under positive selection to maintain virulence. Numerous examples of potential candidate genes were identified based on estimates of nonsynonymous and synonymous substitution rate ratios (Paper V, table 1). We focused our analysis on a promising candidate, carbamoyl phosphate synthetase (CPS), which had a close *E. coli* protein structural homologue in PDB (< 120 PAM units).

*Figure 6.* Representation of the Cα backbone model of *E. coli* carbamoyl phosphate synthetase using (A) *H. pylori* 26695-J99 mutational differences and (B) *H. pylori*-*E. coli* mutational differences coloured according to their Grantham values.
E. coli CPS is a bifunctional enzyme whose two different but coupled functionalities are involved in arginine and pyrimidine biosynthesis. Using a combined approach based on information available from studies on structural biology and site-directed mutagenesis analysis, we reinterpreted our results from mutation rate analysis and found evidence indicating that one of these activities is under selective pressure (Figure 6). It may thus appear that natural selection may act to decouple its amidotransferase activity from synthetase activities.

4.4 Factors underlying protein evolutionary rates in H. pylori

4.4.1 Evolutionary rates of linked genes

To examine the effects of linkage on rate of evolution, we established a data set consisting of rates of evolution at nonsynonymous and synonymous sites for all orthologs between H. pylori (J99 and 26695) and also Bartonella (henselea and quintana) (Alsmark et al., 2004). The difference in $K$-values ($K_a/K_s$, $K_a$, $K_s$) were then calculated between physically adjacent (neighbouring) orthologous genes and used to compare the mean $\Delta K$ over all neighbouring pairs. To analyse the significance of this value, a permutation test was performed by selecting pairs of $K$-values (depending on the analysis) at random. The proportion of 10, 000 random data sets with lower mean difference is an estimate of the P-value that can be attached to the hypothesis that linked genes evolve at similar rates.

None of the 10, 000 random data sets showed a significant mean difference lower than that of the real data, indicating that genes which are physically located adjacent to each other evolve at similar rates much more than expected by chance (Paper VI, figure 5).

4.4.2 Evolutionary rates of interacting proteins

To identify possible sources for the variation in substitution frequencies among genes, and in particular to correlate these with network properties, we first inferred the protein interaction (PIN) and the enzyme interaction (EIN) networks of H. pylori (Figure 7).
Figure 7. Visualisation of *H. pylori* protein-protein interaction (above) and enzyme-enzyme (below) interaction networks.
To test the hypothesis that proteins with high connectivity should be structurally and/or functionally constrained to a greater extent than proteins with fewer interactions, we investigated the correlation between connectivity and substitution frequency for both PIN and EIN. No significant correlation was found between nonsynonymous substitution frequency and the connectivity for PIN, however a weak negative correlation was found for the EIN (Pearson correlation coefficient = -0.152; p-value = 0.009) (Paper VI, figure 2).

4.4.3 Evolutionary rates of essential vs nonessential genes

Additionally, we tested the hypothesis that genes required for viability and host colonisation evolve at a reduced rate compared to genes that are less essential. We found, however no significant difference between the mean substitution frequencies at nonsynonymous sites for genes required for gastic colonisation ($K_a = 0.017$) or viability ($K_a = 0.017$) compared to the complete set of orthologs ($K_a = 0.019$) (Paper VI, figure 4).
5 Discussion

Evolution acts on different time-scales. Some evolutionary events occur more frequently and faster than others such as substitutions and insertion/deletions (indels) of a single or few base pairs, having only a small effect on genome size. Other events however occur less frequently, act on slower time scales such as recombination and segmental deletions of large blocks of intervening DNA and thus have a more pronounced effect on shaping the architecture of genomes. By comparing genomes of closely related species in a phylogenetic context, it is not only possible to detect recent evolutionary events, but also ancient events. This approach may give clues indicating that some species are in a late state of reductive evolution where most of the genes that can be lost easily are already lost, while others have basically just begun their path of degradative evolution. Indeed, reductive evolution is a slow and ongoing process, the genetic events responsible have left footprints of various stages, such as pseudogenes and split genes which characterise the early onset of the degradative process, and gene remnants eroded to the extent that little or no detectable sequence similarity can be discerned exemplifying the late stages, and eventually- complete gene loss. For some genes, the ultimate elimination may indeed be the eventual scenario, but what if evolution is capable of re-using parts of existing genes in forming new genes. Could this be how ORFans originate?

In this thesis, we have used Rickettsia as our model system for studies on reductive genome evolution and on the origin and fate of ORFan genes in bacterial genomes.

5.1.1 Transition to intracellular lifestyles is often accompanied by genome size reduction

A drastic reduction in genome size is often accompanied with a transition to intracellular lifestyles. Indeed, the genomes of free-living organisms are generally much larger than their intracellular counterparts. Invariably, gene loss is accompanied with the corresponding loss of function and is partly responsible for the observed differences in gene content of bacteria. It is thought that the nutrient-rich and constant cellular environment provided by the host cell, results in reduced selective pressure to maintain certain genes. The gradual process of host cell adaptation and integration inevitably leads to the loss of overlapping, dispensable genes in the parasite, indicating a
commitment to the eukaryotic intracellular environment. An intriguing question that one ponders at this stage is, whether there is a bias towards certain categories of genes to be lost? Studies on the metabolic profiles as inferred from the *Rickettsia* genome sequences, indicates that mostly genes involved in amino acid biosynthesis have been lost. It is thought that *Rickettsia* predates on the host production of amino acids, and consequently only a few amino acid biosynthetic genes have been identified. A strategy which *Rickettsia* has adopted to compensate for the lack of amino acid biosynthetic genes, has been expansion of transporter proteins. Indeed, as many as 15 genes encoding amino acid transporter have been identified in *R. prowazekii* genome (Andersson S.G.E. et al., 1998b).

It is however unclear whether ancestral genomes were also reduced. It may be that the smallest of these genomes do not represent the ancestral states, as was once believed, but are derived from larger genomes through massive loss of genes (Moran and Mira, 2001; Bossau et al., 2004). Thus, when lineages make the transition from potentially free-living lifestyles to obligate host-associated ones, genome reduction ensues. Comparative analysis of small and large genomes may eventually shed light on reconstructing the steps leading to genome reduction.

5.1.2 Acceleration of nucleotide substitution rates following partial loss of function?

If the rate of substitution is indeed inversely proportional to the stringency of the functional constraints as claimed by the neutral theory, then we should observe an increase in the rate of nucleotide substitution in genes that lost their function. This has already been known to be the case for pseudogenes, in which all constraints presumably have been removed, and thus are some of the fastest evolving sequences. Let us now examine what happens when selective constraints are only partially, rather than entirely removed. Such a phenomenon is called relaxation of selection, and may be presumed to be operative in sequences which have partially been fragmented, but not yet lost their expression. This is reported to have been the case of split genes found in *R. conorii*, and may represent cases of partial losses of function resulting in an increase in the rate of nucleotide substitutions due to relaxation of function (Paper I).

Comparative studies of gene conservation and substitution rates across closely related species can yield important clues about the functional significance of any observed RNA expression pattern. Our interpretation of these results is that the split genes in *R. conorii* are degraded genes in which mutations have started to accumulate, in spite of which transcription and translation may continue. The enhanced substitution frequencies at nonsynonymous sites suggest that split genes are no longer functional and that the expression driven by some of these fragments is most likely a temporary phenomenon.
5.1.3 Maintenance of nongenic DNA

It has previously been suggested that most of the intergenic sequences in *R. prowazekii* consists of decayed genes that are no longer active but are not yet fully eliminated from the genome. Whereas the complete loss of a function can only be indirectly inferred from the current gene repertoires, pseudogenes can be used to follow the deterioration process in “real evolutionary time”. *R. prowazekii* is known to contain at least 12 pseudogenes (Andersson S.G.E. et al., 1998a, 1998b) and 41 sequences that appear to be pseudogenes were discovered in *R. typhi* (McLeod et al., 2004). In general, the reported pseudogenes fall into different categories. Some contain a few insertions and deletions, substitutions that convert codons to stop codons as well as the occasional loss of start and/or stop codons. Others are much more extensively degraded and are recognized as pseudogenes only because of similarity to a gene or a less degraded pseudogene in closely related species. Interestingly, most of the pseudogenes in *R. typhi* were found in regions corresponding to deletions in *R. prowazekii* and *R. typhi* with respect to *R. conorii*, and in fact many appear to be remnants of the deletion event. The pseudogenes may be functional in other *Rickettsiae* or indicative of regions that are not functional in any of the *Rickettsiae*.

It is interesting to speculate why the *Rickettsiae* appear to maintain nonfunctional DNA for much longer periods than other bacteria. In doing so, we must address the question of what function this DNA may have, if any. A selectionist hypothesis asserts that the nongenic DNA performs some essential functions, such as global regulation of gene expression. According to this hypothesis, the excess of DNA is only apparent, and the DNA is wholly functional. Consequently, deletion of such DNA will have a deleterious effect on fitness. On the other hand, the neutralist hypothesis proposes that nongenic DNA fraction is genetically and physically inert, so-called “junk DNA” to emphasize its uselessness. This junk DNA is carried passively, merely due to physical linkage to functional genes (Petrov et al., 2000). According to this view, the excess DNA is an incidental result of evolutionary processes, and as long as it does not affect the fitness of the organism, it will be carried on from generation to generation. Or could it be that nongenic DNA may act as a repository or sink from which evolution tries to create new genes?

A more practical explanation that has been proposed is that slow growth times may also account for the long residence times of nonfunctional DNA in *Rickettsia*. However, there are many slow-growing bacteria that do not accumulate pseudogenes. Another, yet interesting possibility is that in *Rickettsia*, the pathway of removing nonfunctional DNA is somehow different due to their reductive mode of evolution. By comparing this genome to others, which loose nonfunctional genes much faster, we may discover the mechanism by which organisms loose DNA.
5.1.4 Gene deletions mediated by repeated sequences

Short repeated sequences located in close proximity to each other are known to play an important role in mediating recombination events in *Rickettsia* and other species (Frank et al., 2002; Amiri et al., 2004). Because repeated sequences allow clean excision of large segments of DNA per event, they are excellent mediators of deletions of sequences that are not selectively maintained in the population. As we have seen in this study, the outcome of these recombination events maybe fusions and partial losses of genic sequences (Paper II).

5.2 An evolutionary perspective on ORFan genes

It is said that “nature is a tinkerer and not an inventor”. True novelty is almost unheard of during evolution. Rather, the norm has been to use pre-existing genes and parts of genes that are then transformed to produce new functions in a process generally referred to as molecular tinkering. With this in mind, trying to understand the origins and fate of ORFan genes, it is worthwhile to take a few steps back and review some of the mechanisms that have been proposed for creation of new genes. It may be that the underlying theme of these processes may shed some light onto those responsible for the birth of ORFan genes in *Rickettsia*.

In this section, I will briefly summarise some of the salient features characterising ORFan genes and discuss some of the hypotheses put forward to explain the origins and functions of ORFans. We also bring interesting questions to surface, such as whether ORFans correspond to real genes and if they are, do they correspond to essential proteins?

5.2.1 Characteristics of ORFan genes

ORFans are generally shorter than sequences which have orthologous counterparts (Paper II). Although, long ORFans are likely to be actual coding sequences, short hypothetical ORFs must be viewed with caution (Ochman, 2002; Siew and Fischer, 2003a). Previous analyses of the species and strain-specific genes in bacteria, showed that such ORFan sequences tend to have lower G+C contents than genes with a wider distribution among the species. This has also been the case of ORFans studies in the *E. coli* lineage which are short, A+T rich, functional and quickly evolving, and can be differentiated based on their sequence properties both from laterally acquired genes (Daubin and Ochman, 2004)
5.2.2 Origin and formation of new gene families and the acquisition of new functions

The most common gene formation mechanism is the duplication of existing genes. The evolutionary significance of gene duplication was first recognized by Haldane (1932) and Muller (1935), who suggested that a redundant duplicate of a gene may acquire divergent mutations and eventually emerge as a new gene. Using evidence from various types of studies, Ohno (1970) put forward a view according to which gene duplication is the only mechanism by which a new gene can arise. Although, other means of creating new functions are now known, Ohno’s view remains largely valid.

A complete gene duplication produces two identical copies. How they will evolve varies from case to case. In principle, three possibilities exist. The copies may retain their original function, enabling the organism to produce a larger quantity of RNAs or proteins. Alternatively, one of the copies may be incapacitated by the occurrence of deleterious mutations and become a non-functional pseudogene. More important, however, is the third possibility: that the gene duplication may result in the emergence of genetic novelties or new genes. This will happen, if one of the duplicates retains its original function while the other accumulates molecular changes, diverge and subfunctionalisation can occur, making both of the copies essential (Lynch and Connery, 2000a, 2000b). Regardless of the fate of duplicates, their sequences may remain relatively similar or they may diverge beyond recognition.

5.2.3 On the origin and functions of ORFans

Turning our attention now to ORFan genes, many ORFans may have been generated as the result of a number of possible evolutionary events, which may include horizontal transfer, rapid evolution and gene loss (Siew and Fischer, 2003a). ORFans without selection pressure have been deleted through microbial deletion mechanisms, and thus microbial genomes are kept at 'reasonable sizes'. ORFans that have retained an important function are kept, thus creating new sequence families with a seed of a single ORFan. With time, and subsequent duplications, this family may expand to form a family of paralogues, or remain a singleton family. ORFans observed today in the sequenced genomes may thus be the result of a dynamic process that may have occurred in the past, or may be ongoing; the genome of a descendant may have some ORFans deleted, some new paralogous ORFans, or may contain a number of new ORFans. Not every functional ORFan will be the seed of a multi-membered paralogous family. Some ORFans may correspond to ancient proteins that do not proliferate and are likely to remain ORFans in the future. Depending on the order of the deletion events, the decaying DNA sequences may differ even among closely related strains and isolates, explaining the many strain-specific and species-specific ORFans. If
ORFans are the result of this process, then one need to explain why is it that they are so divergent from all other proteins, and why we do not see today any of the intermediate sequences that must have given rise to them. Possible explanations include rapid evolution and massive gene losses.

So far in our studies, the use of comparative sequence approaches in the study of ORFans has revealed a fascinating interplay between coding and noncoding sequences over time. As many as 412 ORFans were identified in the 1.3 Mb genome of \textit{R. conorii} and these are randomly distributed around the genome. If they represent remnants of full-length genes that once had identifiable functions, it is of interest to reconstruct these genes and try to infer their ancestral function. Here, our case studies of \textit{Rickettsia} offers promising insights. In this species, a large majority of ORFans represent short gene fragments and are atypically short (313 bp on average) compared with full-length orthologs present in both \textit{R. conorii} and \textit{R. prowazekii} (1030 bp on average). This observation per se suggests that many of the annotated ORFans are fragments of deteriorating genes. This also provides an explanation for the low average length, 775 bp of genes in the \textit{R. conorii} genome as compared with an average gene length of 947 bp in bacterial genomes (Mira et al., 2001).

Our comparative phylogenetic analyses of ORFan genes using strains contained within both the TG and SFG \textit{Rickettsia}, indicate that ORFan genes in SFG correspond to pseudogenes in the TG, with pseudogenes in SFG corresponding to extensively degraded gene remnants in the TG (Paper II). This evidence combined with their short lengths, strengthen the view that those unknown, mysterious little ORFans were most likely genuine protein-coding sequences in their previous life, but now represent degraded gene remnants on their way to degradation. The analysis also showed that full-length ancestral gene sequences could be reconstructed from extant gene remnants of very closely related species. This represents a novel approach to validate the authenticity of the ancestral genes, and to gain insights into the mechanism and modes of sequence evolution of these gene vagrants. To this end, we have reconstructed the putative full-length ancestral protein-coding sequences from a set of short, overlapping ORFan genes in multiple closely related \textit{Rickettsia} genomes and searched for evidence of their ancestral functional status. The analysis was based on phylogenetic approach that considered both comparative protein sequence and structure information.

5.2.4 Do ORFans correspond to real genes?

This could be regarded as the million dollar question! In deciding the authenticity of ORFans, a multilevel and detailed analysis at the sequence, structural and phylogenetic levels must be undertaken. However, ultimately biochemical and structural studies will provide the only definitive clues as to the functions of the many uncharacterised ORFans. In what follows, I pre-
sent evidence from sporadic studies conducted at the sequence level using evolutionary substitution rates, gene expression studies and evidence from both protein and gene interaction networks to support the claim that the majority of ORFans are indeed real protein-coding genes.

First, at the sequence level, studies from the analysis of evolutionary substitution rates have reported encouraging findings that the vast majority of putative ORFs, even those are short, are genuine protein-coding regions (Ochman, 2002; Daubin and Ochman, 2004). Although, a low Ka/Ks ratio can be strong evidence that a proposed ORF is protein coding, a high Ka/Ks ratio does not necessarily mean that an ORF is not protein coding (Lawrence, 2003). This may be the case for the many small ORFs that do not evolve in a typical manner as predicted of protein-coding regions and may erroneously dismiss small ORFs that are indeed translated, but where the entire primary sequence of the protein is not the information under selection. Examination of the relative amounts of sequence divergence, as well as features of the sequence itself, permit discrimination among these possibilities and can provide new insights into the role of uncharacterised sequences.

Gene expression analysis can also be used as proof strengthening the authenticity of hypothetical genes. Microarray studies provide a global perspective on gene expression in genomes. Not only can the data be used to monitor the global expression levels of genes, but also aid the functional inference of genes based on context analysis and functional interaction networks (Doerks et al., 2004). It is believed that genes which are co-expressed, generally belong to the same operon and therefore can be expected to interact in the same functional pathway (Overbeek et al., 1999). Using a microarray-based approach, the expression levels of 54 conserved hypothetical genes found in \textit{H. influenza}, were confidently detected both at the mRNA and protein levels and used for their functional annotation (Kolker et al., 2004). In another study using a technique called reverse transcriptase polymerase reaction (RT-PCR), the mRNA expression levels of poorly conserved hypothetical ORFs in the archeal genome, \textit{Halobacterium sp. NRC-1} (Shmuely et al., 2004), as well as \textit{E. coli} (Alimi et al., 2000) were successfully detected and monitored. ORFan genes of the aphid intracellular symbiont \textit{Buchnera sp. APS} have also been experimentally verified (Shimomura et al., 2002).

Interaction networks provide another powerful high-throughput method to gain insight into protein functions. Studies based on the physical protein-protein interaction network of \textit{H. pylori} (Rain et al., 2001), \textit{S. cerevisiae} (Uetz et al., 2000; Ito et al., 2001) and \textit{R. sibirica} (Malek et al., 2004) have also shed light not only on mapping the functional interaction network of annotated proteins, but also hypothetical proteins. Using this approach, it is possible to assign general functions to close homologues as well as uncharacterised hypothetical proteins based on their physical interactions (Kelley et al., 2004).
5.2.5 Are ORFans essential proteins?

Mutagenesis and gene knockout studies undertaken in various organisms to investigate essential genes have also identified among their list of genes, strangely enough- hypothetical genes! Organisms thus far investigated that harbour essential hypothetical genes include \textit{H. pylori} (Salama et al. 2004), \textit{E. coli} (Tong et al., 2004; Gerdes et al., 2003), \textit{H. influenza} (Akerley et al., 2002), \textit{P. aeruginosa} (Jacobs et al. 2003) and \textit{S. cerevisiae} (Giaever et al., 2002). A priori, one would assume that ORFans which are generally thought to be the least conserved would correspond to nonessential genes, since essential proteins are generally conserved. However, it may be that these hypothetical genes correspond to ORFans whose phyletic profiles are evolutionary conserved and are thus regarded as essential (Garbom et al., 2004). In addition, some of these essential hypothetical ORFans may be expressed under certain conditions and thus represent conditional lethal mutants.

5.2.6 Prioritising ORFan studies

Large-scale genome sequencing projects provide an enormous wealth of data to enable us to gain a different perspective on nature, not possible before. Thus, with the increasing use of sequence databases, it has become important to consider the accuracy of the information stored in them. Genome annotators have carefully and clearly marked the unconfirmed genes as “hypothetical”. However, many users of the databases assume that all annotated genes indeed correspond to real genes, and this can clearly lead to wrong conclusions.

Since ORFans lie beyond homology-modelling distance from any other protein, it is thought that may correspond to proteins with novel functions or 3D-structures, thus making them attractive targets for crystallization and further studies. Major international efforts such as the Structural Genomics Consortium may eventually help to provide detailed structural information for the many uncharacterised proteins, in the hope to achieve a complete coverage of structure and sequence space (Fischer, 1999; Siew and Fischer, 2003a, 2003b, 2003c; Siew and Fischer, 2004a; Galperin et al., 2004). To this extent, an effort has also been made to construct a database of ORFans (ORFanage), in the hope to alleviate some of these problems and help to address studies solely devoted to ORFans (Siew and Fischer, 2004b).

Eventually, a combination of computational and experimental studies directed towards unravelling ORFans will allow us to verify the validity of these models and will also provide answers to the questions of the origin of ORFans, and how many of them correspond to real genes, to essential genes or to proteins with novel functions and 3D-structures. These may provide excellent examples of the subtleties of how highly divergent sequences retain, lose or acquire functionality. In short, having possible explanations of the
origin of ORFans only makes them more interesting to study and still leaves us with an enormous amount of ORFans in the databases awaiting characterization.

5.3 Adaptive evolution in *H. pylori*: Carbamoyl Phosphate synthetase

In studying proteins and their functions we arrive at a fundamental principle, attributed to Francis Crick: “If you want to understand function, then you need to understand structure”.

In our study, we have identified candidate genes in *H. pylori* that are under positive selection based on an increased ratios of nonsynonymous/synonymous substitutions. Some of these genes may mediate specific adaptation to its environment. Our structure-function analysis of one such candidate, CPS, enabled us to perform a more detailed investigation of a protein under adaptive evolution and also provide rigorous tests of the structural, functional and by implication adaptive hypothesis in our study. Substitutions in *H. pylori* CPS map to regions which are known to decouple amidotransferase from synthetase activities, leading to increased amidotransferase activity and resultant ammonia production. It is thought that the increased ammonia production acts as a buffer to increase the local pH required for successful colonisation and growth in the gastric environment. Positive selection of CPS could potentially represent a mechanism of adaptation to the harsh gastric environment affording *H. pylori* a selective advantage in adapting to its human host. Although our case study exemplifies this approach, its potential promises to yield profound effects in bringing candidate genes to surface when new protein structures are determined.

5.4 Investigating substitution rate variation in pathogenic bacteria

We have undertaken a systematic attempt to investigate factors which may influence rate of protein evolution in *H. pylori*.

It is known that selection acts predominantly at the protein structure level, since this is the level at which proteins execute their function. Proteins with divergent sequences have previously been shown to have similar structures, indicating that changes at the sequence level may not result in significant changes at the protein structure level (Russel et al., 1997). Therefore, it may appear that sequence divergence may not be the most biologically relevant measure of evolutionary conservation of a gene.
Our investigation seems to confirm previously published results which echoes a “weak, but statistically significant” correlation with all variables examined. However, our study highlights the surprising finding that gene location comes out as a strong statistical predictor of evolutionary rate rather than any other factors examined. Evolutionary conservation of gene organisation into operons may be optimised to minimise diffusion of protein products for interaction and regulation by minimising path lengths in their respective interaction networks. The logic may be that “proper organisation eliminates later complications”.
6 Concluding remarks and future prospects

Genome reduction is thought to be the main force behind the evolution of parasitic and/or intracellular bacteria. Given their genetic isolation, it is tempting to postulate that *Rickettsia* had to rely on internal mechanisms such as duplication and recombination to acquire or modify some of the gene functions for adaptation to their niche.

During the progress of our studies, evidence of the very nature of the gradual reduction process were presented by identifying all intermediates from intact ORFs: transcribed split ORFs, further split ORFs no longer transcribed, fully decayed but still recognizable ORFs and complete gene disappearance. Results from studies on the split genes in *R. conorii* indicate that these small fragmented ORFs may be expressed despite their gene deterioration and eventual loss from the genome. The mutation spectrum of noncoding sequences from different species of *Rickettsia* were also studied. Such a comparative analysis showed that short deletions provide the dominant evolutionary mode in these sequences. Thus noncoding sequences that are thought to be mutation-degraded versions of nonessential coding sequences can slowly depart the genome by virtue of small deletions. We have also observed evidence implicating short repetitive sequences in mediating evolutionary events such as recombination in yielding new chimeric genes.

Our journey of exploration of the ORFan phenomenon, indicate that the existence of ORFans which was previously thought to have been a “mystery” may indeed not be that mysterious. Although numerous explanations have been offered to account for their occurrence and for the inability to classify these genes into existing protein families (Fischer and Eisenberg 1999), results from our studies indicate that these genes could be produced de novo from noncoding sequences and from fragments of existing genes which may be highly divergent between taxa.

This thesis has discussed some of the questions related to degradative genome evolution in *Rickettsia* and those surrounding the existence of ORFans. It is our perception that similar mechanisms probably occur in the evolution of bacterial species but have thus far remained undetected because of more active recombination and faster evolutionary rate. What remains clear is that ORFans will continue to represent untapped sources of research and can provide valuable insights into gene evolution, and thus warrants further computational and experimental studies.
Our detailed analysis at both the sequence and structural level of *H. pylori* carbamoyl phosphate synthetase provided an opportunity to understand and rationalise the effect of natural selection acting on a protein undergoing adaptive evolution. The overview of factors takes a holistic approach in trying to understand the factors which may play a role in determining protein evolutionary rates in *H. pylori*. Our studies have identified gene location as an important factor which may influence protein evolutionary rates.

In conclusion, it is evident that no single factor can solely explain the rate of a protein’s evolution, but that a multitude of factors are responsible for governing the rate of protein evolution, all of which have been fine-tuned during evolutionary history.
Patogena bakterier är orsaken bakom en stor mängd sjukdomar hos människor över hela världen. Kunskap om hur patogena bakterier evolverar är därför ytterst viktig, inte minst för att förstå, förutsäga och förebygga spridning av nya och existerande sjukdomar.

Genom att jämföra genomsekvenser i olika arter för regioner med samma ursprung kan man beräkna substitutionsfrekvenser, det vill säga en uppskattning av hur många nukleotider som har förändrats i sekvensen sedan arterna separerade. Jämförelser mellan substitutionsfrekvenser vid positioner som förändrar (Ka) och inte förändrar (Ks) aminosyrasekvensen är viktiga för att förstå hur genomiska sekvenser evolverar.

Syftet med det här arbetet var att förstå vilka drivkrafter som ligger bakom variationer i substitutionsfrekvenser i humana patogener. De bakterier som undersökts är framför allt Rickettsia prowazekii, som orsakar epidemic tyfus och fläckfeber, och Helicobacter pylori, som är inblandad i en rad magsjukdomar. Arbetet har särskilt fokuserat på gener i Rickettsia som saknar motsvarighet i någon annan känd art, så kallade "orphan genes". Adaptiv sekvensevolution och faktorer som påverkar proteiners evolutionshastighet har också studerats.


En viktig slutsats från arbetet är att höga Ka/Ks-värden i humana patogener kan bero antingen på adaptiv sekvensevolution eller på att generna är på väg att degraderas. Förmågan att särskilja mellan dessa två processer är en
viktig uppgift inom molekylär evolution, men är även av stor betydelse för medicinsk microbiologi och forskning inom funktionell genomik.
8 Acknowledgements

I would like to thank Siv for the educational opportunity and financial support to enable me to study at the department.

Although my time has rather been consumed with work, most of the people (past and present) at the department have not escaped my attention in some way or the other. To Karin Olsson, for assistance in moments of need, your effort is appreciated. To Ola- “Retouch” Lundström, Håkan- “the Bullet” Svensson. To Lisa, Hillevi, 2 Karens, Micke, Erika, Carolin, BjörnB, Hans-Henrik, Gustav, Johan, Eva, Mats, Magnus, Linus for “chatty” moments, Anna-Sofie and Kristina. To BjörnS- your help and intellectual and friendly discussions are appreciated. To Gabor- “a friend in need, ...a friend indeed!”.

To previous members of the department, Olof for settling me in during my initial period and preventing me from getting lost! A special thanks to the 2 B’s - Boris and Björn, for providing sparks of joy in my chaotic world. Ivaca for nice talking moments and Haleh- ...knowing you have made my time eventfull and fun. To “amazing” Alex for always keeping that friendly smile. To those who have sat in the same room as me: Maxim, Bastien, Dirk, Emmanuel and Maya...ooh and Gregor... it was an experience knowing you all!

On a personal note, to people from my past: Malek, Mustaffa, Beatrice and Brita, the Icelandic Håkan, Santosh, Rajni and Bo, the 2 Daves, the 2 Javiers, Ramy for your friendship. To Nadia, Rushana, Rizada- who knows what could have been and what will be...

I would like to show a special sign of thanks to the National Research Foundation (NRF), South Africa without whose financial support this effort would not be possible. At last, I would like to thank my mother for her eternal love and encouragement ..."you are the wind beneath my wings”....
9 References


Jacobs M.A., Alwood A., Thaipisuttikul I., Spencer D., Haugen E., Ernst S., Will O.,
Kaul R., Raymond C., Levy R., Chun-Rong L., Guenthner D., Bovee D., Olson
M.V., Manoil C. (2003) Comprehensive transposon mutant library of Pseudo-
monas aeruginosa. Proc Natl Acad Sci. USA. 100(24):14339-14344.
evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 2(6):962-968.
evolution rate and the number of protein-protein interactions: only the most pro-
PathBLAST: a tool for alignment of protein interaction networks. Nucleic Acids
Kolker E., Makarova K.S., Shabalina S., Picone A.F., Purvine S., Holzman T.,
Cherny T., Armbruster D., Munson R.S. Jr., Kolesov G., Frishman D., Galperin
M.Y. (2004) Identification and functional analysis of 'hypothetical' genes ex-
evolution of gene duplications. Genome Biology. 3(2): RESEARCH0008. Epub
2002 Jan 14.
in E. coli DNA. Nucleic Acids Res. 22(22):4768-78.
quence divergence, gene dispensability, expression level, and interactivity are
Lawrence J.G. (1999) Gene transfer, speciation, and the evolution of bacterial gen-
Lawrence J. (2003) When ELF's are ORFs, but don't act like them. Trends Genet.
19(3):131-132.
nonsynonymous rates of nucleotide substitution considering the relative like-
Li W.H. (1993) Unbiased estimation of the rates of synonymous and nonsynony-
Li W. (1999) Statistical properties of open reading frames in complete genome se-
Biotechnol. 21(7):298-300.


A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)