Studies of protein structure
and
genome evolution

by Huan-Lin WU

for the degree of Doctor of Philosophy
University of Bath
Department of Biology and Biochemistry
April 2006

COPYRIGHT

Attention is drawn to the fact that copyright of this thesis rests with its author.

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the prior written consent of the author.

This thesis may be made available for consultation within the University Library and may be photocopied or lent to other libraries for the purposes of consultation.

[Signature]

1
Abstract

Just as hieroglyphic, demotic, and Greek with identical texts were inscribed on the Rosetta Stone, the information of life is conveyed by three different vehicles. It is known how individual messages translate from the language of nucleotide bases in DNA or RNA into the language of amino acids in a peptide. The RNA language is written in an alphabet of four letters \(\{A, C, G, U\}\). Arbitrary sets of three letters are grouped into words called triplets or codons. Each of the 64 codons specifies one of 20 amino acids or a stop punctuation mark. Moreover linear amino acid sequences fold into non-linear structures to express their final meanings.

In the first part of this study, a new model is proposed to explain the evolution of the genetic code to the current triplet code from a highly degenerate triplet code that was functional first as a singlet code and later as a doublet code that used two types of prefix and suffix codons. This model explains multiple features of the genetic code such as the origin of the pattern of four-fold degenerate and two-fold degenerate triplet codons, the origin of its error minimising properties, and why there are only 20 amino acids.

In the second part the three-dimensional structure of the catalytic domain of BopE (BopE residues 78-261) by NMR spectroscopy is presented. BopE is a type III secreted protein from \textit{Burkholderia pseudomallei}, the aetiological agent of a severe emerging infection of humans and animals called melioidosis. BopE is a guanine nucleotide exchange factor for the Rho GTPases Cdc42 and Rac1. To our knowledge this is the first effector protein from \textit{Burkholderia pseudomallei} to be characterised in this way.

Two articles concerning mathematical biology are presented in the third part. Cross-species genome comparison is a fundamental method for identifying biologically essential elements. Using this method, a considerable number of
conserved non-genic sequences (CNGs) have previously been identified in vertebrates. However, others have shown that deletion of gene deserts containing a total of 1243 CNGs in mice resulted in no detectable phenotypic changes. A mathematical explanation for the existence of many CNGs is given herein. Finally, many arguments on whether evolution is predictable have been posed since Darwin proposed the concept of evolution by natural selection. This big question was again posed at Kavli's theoretical physics conference in October 2004. Amongst 25 forward-looking questions suggested by the meeting's attendees, this nagging question was more precisely rephrased: can the theory of evolution be quantitative and predictive? Here it is demonstrated that the evolution of genomes cannot be properly predicted no matter how genomes are quantified.
Acknowledgments

I would like to thank my supervisors, Dr. Stefan Bagby and Dr. Jean ven der Elsen, for their guidance through the fields of protein structure and genome evolution.

Special thanks to my father who has been a constant source of love, support and encouragement, and to my sister’s family that helps me to take care of my father when I study in the UK. Having two newborn nieces is the most joyful thing during my PhD.

Abhishek Upadhyay is gratefully acknowledged for his important contribution to this work through protein purification. Thanks also to Didier Philippe and Christopher Williams, whose help and ideas have made this an enjoyable and fulfilling undertaking.

Finally, I would like to thank my friends for their encouragement: Kenny, Kevin and the many others.
# Table of Contents

Abstract.................................................................2
Acknowledgments.............................................................4
Table of Contents...........................................................5
Abbreviations..............................................................9

## Part One. Evolution of the Genetic Triplet Code via Two Types of Doublet Codons

Chapter One. Introduction to the Origin of the Genetic Code ..................13
  1.1 Cracking the code......................................................13
  1.2 Previous hypotheses for genetic code evolution .....................16
  1.3 The RNA world.......................................................21

Chapter Two. Evolution of the Genetic Triplet Code via Two Types of Doublet Codons .................................................27
  2.1 Abstract..............................................................27
  2.2 Hypothesis...........................................................27
    2.2.1 The evolution of triplet codons from singlet codons via two types of doublet codons ........................................27
    2.2.2 From a singlet code to a doublet code with two types of codons ..................................................31
  2.3 A primordial three-base reading frame..................................35
2.4 Why are there 64 codons and only 20 amino acids? 39
2.5 Codon ambiguity: overlapping prefix and suffix codons 39
2.6 Codon patterns 40
2.7 Relicts of a doublet code 43
  2.7.1 Four-fold degenerate codons: relicts of prefix (x_) doublet codons? 43
  2.7.2 The evolution of aminoacyl-tRNA synthetases: evidence for suffix(_x) doublet codon expansion? 46
2.8 Genetic code variations and the emergence of a ‘true’ triplet code 49
2.9 An ancestral doublet code 51
2.10 Conclusion 52

Part Two. The NMR Structure of *Burkholderia Pseudomallei* Type III Secreted Protein BopE 54

Chapter Three. Introduction to *Burkholderia Pseudomallei* Type III Secreted Protein BopE 55
  3.1 Abstract 55
  3.2 Classification of *Burkholderia* 56
  3.3 *Burkholderia pseudomallei* 58
  3.4 Melioidosis 59
  3.5 BopE 61
  3.6 Type III Secretion system 62
  3.7 Rho GTPases 66
  3.8 Guanine nucleotide exchange factors (GEFs) 68
  3.9 Structure of SopE-CDC42 complex 73
Chapter Four. Materials and Methods

4.1 Methods and experiments

4.2 Side chain assignment

4.3 Structural restraints

4.4 Structure calculation

Chapter Five. The Structure of BopE

5.1 Abstract

5.2 Structure analysis

5.3 Structure determination

5.4 Three-dimensional structure of BopE_{78-261}

5.5 Sequence and secondary structure comparison of *B. pseudomallei* BopE_{78-261} with *Salmonella* SopE_{78-240} and SopE_{69-240}

5.6 Tertiary structure comparison of *B. pseudomallei* BopE_{78-261} with *Salmonella* SopE_{78-240} and SopE_{69-240}

5.7 Implications of tertiary structure differences between BopE_{78-261} and SopE_{78-240} / SopE_{69-240} for interaction with Rho GTPases

5.8 Conclusions

Part Three. Mathematical Biology

Chapter Six. The Origin of Conserved Non-Genic Sequences

6.1 Abstract

6.2 Statistical evidence

6.3 Ramsey Theory model

6.4 Assessment of likelihood that CNGs have arisen by chance
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D</td>
<td>Two-dimensional</td>
</tr>
<tr>
<td>3D</td>
<td>Three-dimensional</td>
</tr>
<tr>
<td>AA</td>
<td>Amino acid(s)</td>
</tr>
<tr>
<td>aaRS</td>
<td>Aminoacyl-tRNAs synthetases</td>
</tr>
<tr>
<td>BMRB</td>
<td>BioMagResBank</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair(s)</td>
</tr>
<tr>
<td>Bsa</td>
<td><em>B. pseudomallei</em> type III protein secretion apparatus</td>
</tr>
<tr>
<td>BopE</td>
<td><em>Burkholderia</em> outer protein E</td>
</tr>
<tr>
<td>CCH</td>
<td>Coding coenzyme handle</td>
</tr>
<tr>
<td>CD</td>
<td>Circular dichroism spectroscopy</td>
</tr>
<tr>
<td>CF</td>
<td>Cystic fibrosis</td>
</tr>
<tr>
<td>CNG</td>
<td>Conserved non-genic</td>
</tr>
<tr>
<td>D$_2$O</td>
<td>Deuterium oxide</td>
</tr>
<tr>
<td>Da</td>
<td>Dalton</td>
</tr>
<tr>
<td>DH</td>
<td>Dbl homology domain</td>
</tr>
<tr>
<td>GAP</td>
<td>GTPase activating protein</td>
</tr>
<tr>
<td>GDI</td>
<td>Guanine dissociation inhibitor</td>
</tr>
<tr>
<td>GDP</td>
<td>Guanosine-5’-diphosphate</td>
</tr>
<tr>
<td>GEF</td>
<td>Guanine nucleotide exchange factor</td>
</tr>
<tr>
<td>GTP</td>
<td>Guanosine-5’-triphosphate</td>
</tr>
<tr>
<td>HRP</td>
<td>Hypersensitive response and pathogenicity</td>
</tr>
<tr>
<td>HSQC</td>
<td>Heteronuclear single quantum correlation</td>
</tr>
<tr>
<td>kb</td>
<td>Kilo base pair(s)</td>
</tr>
<tr>
<td>ms</td>
<td>Millisecond</td>
</tr>
<tr>
<td>NC</td>
<td>Needle complex</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear magnetic resonance</td>
</tr>
<tr>
<td>NOE</td>
<td>Nuclear Overhauser effect</td>
</tr>
<tr>
<td>NOESY</td>
<td>Nuclear Overhauser enhancement spectroscopy</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>ns</td>
<td>Nanosecond</td>
</tr>
<tr>
<td>PEV</td>
<td>Position effect variegation</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>PH</td>
<td>Pleckstrin homology domain</td>
</tr>
<tr>
<td>PIs</td>
<td>Pathogenicity islands</td>
</tr>
<tr>
<td>ppm</td>
<td>Parts per million</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root mean square deviation</td>
</tr>
<tr>
<td>SopE(2)</td>
<td><em>Salmonella</em> outer protein E</td>
</tr>
<tr>
<td>SPI</td>
<td><em>Salmonella</em> pathogenicity island</td>
</tr>
<tr>
<td>TOCSY</td>
<td>Total correlation spectroscopy</td>
</tr>
<tr>
<td>TTS(S)</td>
<td>Type three secretion (system)</td>
</tr>
</tbody>
</table>
The standard genetic code.

<table>
<thead>
<tr>
<th>1st position (5' end)</th>
<th>2nd position</th>
<th>3rd position (3' end)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U</strong></td>
<td><strong>C</strong></td>
<td><strong>A</strong></td>
</tr>
<tr>
<td>Phe</td>
<td>Ser</td>
<td>Tyr</td>
</tr>
<tr>
<td>Phe</td>
<td>Ser</td>
<td>Tyr</td>
</tr>
<tr>
<td>Leu</td>
<td>Ser</td>
<td>Stop</td>
</tr>
<tr>
<td>Leu</td>
<td>Ser</td>
<td>Stop</td>
</tr>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
</tr>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
</tr>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>Gln</td>
</tr>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>Gln</td>
</tr>
<tr>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
</tr>
<tr>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
</tr>
<tr>
<td>Ile</td>
<td>Thr</td>
<td>Lys</td>
</tr>
<tr>
<td>Ile</td>
<td>Thr</td>
<td>Lys</td>
</tr>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
</tr>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
</tr>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Glu</td>
</tr>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Glu</td>
</tr>
</tbody>
</table>

Nomenclature of Common Biological Amino Acids

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>Asp</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>Glu</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
</tr>
</tbody>
</table>
PART ONE

Evolution of the Genetic Triplet Code via Two Types of Doublet Codons
1.1 Cracking the code

Since the 1953 discovery by James Watson and Francis Crick of the structure of DNA, many ingenious 'genetic code' schemes have been proposed. The first coding scheme was the diamond code proposed by a physicist George Gamow (1956). It was suggested that double-stranded DNA acted directly as a template for assembling amino acids into proteins. The various combinations of bases along one of the grooves in the double helix could form distinctively shaped cavities. He proposed that the side chains of amino acids might fit in these cavities. When all the amino acids were lined up in the correct order along the groove, an enzyme would come along to polymerize them (Figure 1.1).

Figure 1.1. George Gamow's diamond code assumed that proteins form directly on a DNA template, (Hayes, 1989).
Gamow's diamond code could also explain why there were 64 codons but 20 amino acids used only. Since most amino acid side chains are symmetrical, it is thought that the diamonds could be flipped end-to-end, such as the triplet CTG becomes GTC, or flopped side-to-side, such as CTG becomes CAG in which the middle T is replaced by the complementary A, without changing its meaning. When all such symmetries are taken into account, there is a total of 20 different meanings left.

The diamond code is an overlapping triplet code. Each nucleotide base claims simultaneous membership in three adjacent codons. For example, the base sequence ATTGAGT consists of five overlapping triplets: ATT, TTG, TGA, GAG and AGT (Figure 1.2). This overlapping property, in which the density of information storage can be maximized, has been thought to be the reason why nature might have adopted this code. However, the concern with efficiency turned out to be misplaced; information density is not a very high priority for most organisms.

Another famous code is a non-overlapping code, the comma-free code, proposed by Francis Crick (1957).

The frame-shift problem does not arise with an overlapping code, because all three reading frames are simultaneously valid. However, with sequential codons, recognition of the right reading frame seemed to be a problem in the non-overlapping code. In order to explain how the translation machinery was guided to the right frame, Crick devised a solution that adaptor molecules might exist for only a subset of the 64 codons; if the codons in one reading frame are meaningful, then the overlapping triplet are nonsense (Figure 1.2). In such a code, only a subset of codons would be meaningful rendering the rest of the triplets “nonsense codons”. For example, AAA, CCC, GGG and UUU cannot appear in any comma-free code, since they cannot combine with themselves without generating reading-frame ambiguity. Here I give a brief proof.
Take AAA as an example. If AAA is a code word in a comma-free code, then (AAA)(AAA)=A(AAA)AA=AA(AAA)A. It implies AAA cannot be a code word.

Of the remaining 60 codons, Crick and his colleagues pointed out that 60 codons could be sorted into three groups, where the codons within each group are related by a cyclic permutation. For example, the codons AGU, GUA and UAG are from one of the three groups respectively. Since 60 codons (excluding AAA, CCC, GGG and UUU) are divided into 3 groups by a cyclic permutation, each of the groups has 20 codons. This result also answered the question why there are 20 amino acids.

Figure 1.2. An overlapping code packs 16 codons into 18 base pairs by exploiting triplets in all three phases, or reading frames. A comma-free code is constructed so that only the codons in one reading frame are meaningful; the overlapping triplets are nonsense, in black (Hayes, 1989).

Although many fascinating schemes were suggested, the genetic code was not cracked until 1965. This breakthrough was made by Marshall W. Nirenberg and J. Heinrich Matthaei. They announced that artificial RNAs could stimulate protein synthesis in a cell-free system. The first RNA they tried was poly-U, a long chain of repeating uracil units. In Crick's comma-free
code, UUU is a nonsense codon, but Nirenberg and Matthaei’s result implied that it codes for the amino acid phenylalanine (Nirenberg, 2004). When the table for the assigning of codons to amino acids was completed, it revealed that nature seemingly did not appreciate the ingenious ideas of humans.

1.2 Previous hypotheses for genetic code evolution

Given 64 codons and 20 amino acids plus a punctuation mark, there are $10^{83}$ possible genetic codes. Although the genetic code has been deciphered, explaining the apparent non-random codon distribution and the nature and number of amino acids in the “standard” genetic code remains a challenge.

In 1968 Francis Crick proposed his ‘frozen accident’ model. He argued that the code did not need to be special at all; it could be nothing more than a “frozen accident.” The assignment of codons to amino acids might have been subject to reshuffling and refinement in the earliest era of evolution, but further change became impossible because the code was embedded so deeply in the core machinery of life (Crick, 1968). A mutation that altered the codon table would also alter the structure of every protein molecule, and would therefore almost surely be lethal (Hayes, 2004). However the “frozen accident” model, where codon assignments are historical accidents that became fixed in the last common ancestor of all organisms, does not explain or predict the observed order of codons.

There is recent evidence that the genetic code is not quite frozen. Certain protozoa, bacteria and intracellular organelles employ genetic codes slightly different from the standard one (Figure 1.3), hinting that changes to codon assignments are not impossible after all. Moreover, a study has generated a completely autonomous bacterium with a 21 amino acid genetic code. This bacterium can biosynthesize a non-standard amino acid from basic carbon sources and incorporate this amino acid into proteins in response to the amber (UAG) nonsense codon (Mehl et al., 2003). This and other such organisms may provide an opportunity to examine the evolutionary
consequences of adding new amino acids to the genetic repertoire, as well as generate proteins with new or enhanced biological functions.

With the possibility of evolving genetic codes, various arguments have hence been put forth to explain these characteristics of the genetic code (Knight et al., 1999). It is thought that the genetic code evolved in two distinct phases. First, the 'canonical' code emerged before the last universal ancestor; subsequently, this code diverged in numerous nuclear and organelle lineages (Knight et al., 2001). Some of the varied arguments on how the genetic code evolved from the last common ancestor are described below.

**Figure 1.3.** Composite phylogeny of variant codes. (Knight et al., 2001).
Adaptive (statistical) argument: It was shown that information density is not a very high priority for genetic code. Some researchers suggested that the pattern of codon assignments was an adaptation that optimises some function, such as minimization of errors caused by mutation or mistranslation, or physicochemical property of amino acid, such as polar, size and charge. A variety of criteria have been used to assess whether the genetic code is in some sense optimal. Cynthia Alff-Steinberger was the first to try to quantify this error-tolerance possibility. With the computing limitations in the 1960s, she was able to test only 200 variant random codes and concluded that the natural code is more error-tolerant than a typical random code (Alff-Steinberger, 1969). A decade later, a study approached the same question from a different angle (Wong, 1980). Handcrafted codes, with the best substitution for each amino acid, were created and compared with the natural code. This time a totally different conclusion was arrived at: the substitutions generated by the natural code are less than half as close, on average, as the best-handcrafted codes. This result suggested that the genetic code has not evolved to maximize error-tolerance.

Another study claimed that the genetic code is 'one in a million' (Freeland and Hurst 1998; Knight et al., 2000, 2003). In this study, semi-random codes were created with some constraints: these codes keeping the same blocks of synonymous codons found in the natural code but permuting the amino acids assigned to them. Subsequently the natural code was compared to those semi-random codes. It was concluded that only 114 of the million semi-random codes gave better substitutions than the natural code when evaluated with respect to polar requirement, but the natural code was not better in other respects, such as the amino acid’s size and charge. However, the result depended strongly on what criterion was chosen to judge the similarity of amino acids. Amino acids can be classified into different sets referring to varied physicochemical properties (Figure 1.4). There are still some questions whether polar requirement is the right criterion for estimating the similarity of amino acids and why polar requirement, not other properties of amino acids, is favoured by nature (Hayes, 2004).
Chemical argument: The chemical argument suggests that the assignment of certain codons was directly influenced by chemical interactions between RNA and amino acids. If this hypothesis is true, then some optimal adaptation exhibited in the genetic code may not have been shaped by natural selection, because this hypothesis implies that similar amino acids can bind to similar short RNA motifs and therefore have similar codons. Meanwhile, this hypothesis also implies that the genetic code is universal. However, it has recently become clear that certain protozoa, bacteria and intracellular organelles employ genetic codes slightly different from the standard one (Figure 1.3). This evidence suggests that because the

Figure 1.4. A Venn diagram representation of the properties of the amino acid residues (Mathews et al., 1999).
mechanisms that allowed recent changes in the genetic code might be entirely different from those that generated the code initially, the problem arises that all stereochemical theories might have dealt only with the canonical code found in the last common ancestor (Knight et al. 2000).

**Historical hypothesis:** The historical or co-evolution hypothesis proposes that the canonical genetic code evolved from a primitive ancestral form very early in evolution. This primitive code comprised a small number of amino acids encoded by 64 highly degenerate codons. The gradual expansion of the code is reflected in the pattern of codon assignments. Recently introduced amino acids would have been incorporated in the code by usurping codons of amino acids from related biosynthetic pathways (Amirnovin 1997; Wong 1975).

Historical theories propose a stepwise evolution from a system in which any tRNA required recognition of one base to tRNAs with specificities for the first and second bases. This evolution would have allowed more amino acids to be distinguished, and at a later time, when the system's specificity had improved, tRNAs could evolve with specificity for all three bases (Yockey 2000). In the light of this theory a new hypothesis is presented for the evolution of the genetic code to the current triplet code from a highly degenerate triplet code that was functional first as a singlet code and later as a doublet code, using two types of codons. This hypothesis accounts for the overall codon patterns in the canonical genetic code and the origin of its error-minimising properties; In addition, it explains why there are 64 codons and only 20 amino acids (Wu et al., 2005). This hypothesis will be presented in Part 1, Chapter 2.
1.3 The RNA world

Above I discussed the arguments about how the genetic code has evolved from a simpler code in the last common ancestor. Considering many crucial components involved in the transcription/translation mechanism (Figure 1.5), how could the transcription/translation mechanism ever have evolved initially? The RNA world hypothesis may provide a good solution. The RNA world (Walter Gilbert, 1986) was suggested to be the first life form on earth. This hypothesis is supported by RNA's ability to store, transmit, and duplicate genetic information, just like DNA does. Unlike DNA, RNA can also act as a ribozyme, which can catalyse as enzyme.

![Figure 1.5. The mechanism of transcription/translation.](image)

Transfer RNAs and aminoacyl-tRNA synthetases (aaRS) have recently been the center of attention in explaining the origin of the genetic code (Figure 1.6). This is because that code is established in aminoacylation reactions. In the aminoacylation reactions, each amino acid is joined to the tRNA that harbours the anticodon triplet of the code for that amino acid and the reaction is
catalysed by aminoacyl tRNA synthetases. Hence the question of the origin of the code has to deal at some point with tRNAs and aminoacyl-tRNA synthetases.
Figure 1.6. (A) Schematic representation of 3D conformation of a tRNA. (B) tRNA all have a characteristic cloverleaf type of secondary structure. In the acceptor stem, the 5' and 3' ends of the tRNA are base-paired. The amino acid specific for the tRNA is covalently attached to a 3' OH group on the terminal adenine nucleotide. The anticodon arm contains the anticodon triplet exposed in a loop region (de Duve 1989).

Despite their conserved mechanisms of catalysis, the aminoacyl-tRNA synthetases can be divided into two unrelated classes (I and II) based on mutually exclusive motifs that reflect distinct active site topologies (Figure 1.7; Figure 1.8; de Pauplana et al., 2001). No evidence supports the existence of a common ancestor for the two groups (Schimmel et al., 1996). Moreover, the class to which an enzyme belongs is fixed in evolution through all three kingdoms—bacteria, archae, and eukarya, except for Lysyl-tRNA synthetase. Lysyl-tRNA synthetase (LysRSs) are unique in being composed of unrelated class I and class II enzymes (Figure 1.7).

The hypothesis of coding coenzyme handles (CCH) suggests the genetic code might have evolved in a RNA world and preceded the existence of translation. It views the anticodon loop as stereochemical adapter, which was charged and used by primordial ribozymes to compensate for missing functional groups. Amino acids were covalently linked to particular oligonucleotides (handles), which could basepair with ribozymes. As peptide synthesis became more feasible, handles turned into tRNA, assignment ribozymes were replaced by protein aminoacyl-tRNA synthetases and many ribozymes became mRNA molecules while losing their original enzymatic activity (Szathmary, 1993, 1999). There is some evidence supporting this idea. Sequence analyses on the two types of Lysyl-tRNA synthetases (Figure 1.7) indicate that at least one of them appeared after the establishment of tRNA\textsuperscript{Lys}, because sequences of Lysyl-tRNA synthetases from class I belong to a single tree. This suggests that evolutionary patching on protein
synthetases happened around a tRNA population that was already present (Schimmel et al., 1996).

<table>
<thead>
<tr>
<th>Class I</th>
<th>Class II</th>
</tr>
</thead>
<tbody>
<tr>
<td>la</td>
<td>2a</td>
</tr>
<tr>
<td>Leucine</td>
<td>Serine</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Threonine</td>
</tr>
<tr>
<td>Valine</td>
<td>Alanine</td>
</tr>
<tr>
<td>Arginine</td>
<td>Glycine</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Proline</td>
</tr>
<tr>
<td>Methionine</td>
<td>Histidine</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1b</th>
<th>2b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glutamate</td>
<td>Aspartate</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Asparagine</td>
</tr>
<tr>
<td>Lysine-I</td>
<td>Lysine-II</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1c</th>
<th>2c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyrosine</td>
<td>Phenylalanine</td>
</tr>
<tr>
<td>Tryptophan</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1.7.** Classification of aminoacyl-tRNA synthetases. The 20 aminoacyl-tRNA synthetases fall into one of two classes. Based on their mode of binding to the tRNA acceptor stem, both classes of tRNA synthetases have been subdivided into three subclasses. Lysyl-tRNA synthetase (LysRSs) are unique in being composed of unrelated class I and class II enzymes (de Pauplana et al., 2001).
Figure 1.8. Active site domains of (a) class I aminoacyl-tRNA synthetase, e.g. GlnRS, and (b) class II aminoacyl-tRNA synthetase, e.g. AspRS. Shown are ATP and the acceptor ends.
of cognate tRNAs (red). Two classes share no similarity in structure (Schimmel et al., 1996).

At the point I mentioned in this section, the primitive genetic code was surely smaller and simpler than the modern one at the origin of life. This primitive genetic code probably included only a set of fewer amino acids, or likely several sets of fewer amino acids. Moreover, the code may have functioned as a pure doublet code, ignoring the third base (or the first base) in each codon. I will explain this idea further in the next chapter (also see Wu et al., 2005).
PART ONE
Chapter Two
Evolution of the Genetic Triplet Code via
Two Types of Doublet Codons

2.1 Abstract
Explaining the apparent non-random codon distribution and the nature and number of amino acids in the 'standard' genetic code remains a challenge, despite the various hypotheses so far proposed. Here I propose a simple new hypothesis for code evolution involving a progression from singlet to doublet to triplet codons with a reading mechanism that moves three bases each step. It is suggested that triplet codons gradually evolved from two types of ambiguous doublet codons, those in which the first two bases of each three-base window were read ('prefix' codons) and those in which the last two bases of each window were read ('suffix' codons). This hypothesis explains multiple features of the genetic code such as the origin of the pattern of four-fold degenerate and two-fold degenerate triplet codons, the origin of its error minimising properties, and why there are only 20 amino acids.

2.2 Hypothesis:
2.2.1 The evolution of triplet codons from singlet codons via two types of doublet codons
The historical or co-evolution theory suggests that the genetic code was once simpler, with fewer types of amino acids. With increasing organismal complexity came the necessity for more amino acids. A system in which amino acids were encoded by one nucleotide could have evolved towards a two-base codon system. At a later time, tRNAs could have evolved with specificity for three bases, allowing more amino acids to be distinguished. But could an organism have survived a genetic code expansion? A change from a pure doublet codon to a pure triplet codon system, for instance, would
cause a frameshift that not only affects the encoded amino acid sequence, but also drastically reduces the number of amino acids in resulting polypeptide chains.

Figure 2.1. The evolution of triplet codons from singlet codons via two types of doublet codons. Illustrated is the hypothesis that the current triplet genetic code originates from a singlet code in which the reading mechanism moves three bases per step but in which only the middle base (shown in green) of each codon specifies the encoded amino acid. The subsequent doublet code is comprised of two types of codons: prefix codons (shown in the yellow box), in which the first two bases are read (as suggested by Crick), labelled \( \omega \); and suffix codons (shown in the violet box), in which the last two bases are read, labelled \( \omega \). In the doublet-to-triplet codon expansion, prefix doublet codons \( (\omega \omega) \) evolve to triplet.
codons in which the third base is read for the first time, and suffix doublet codons (\textunderscore \omega) evolve to triplet codons in which the first base is read for the first time. Triplet codons are shown as a mixture of yellow and violet.

On the subject of codon expansion, Francis Crick (1968) wrote: "It might be argued that the primitive code was not a triplet but that originally the bases were read one at a time (giving 4 codons), then two at a time (giving 16 codons) and only later evolved to the present triplet code. This seems highly unlikely, since it violates the Principle of Continuity. A change in codon size necessarily makes nonsense of all previous messages and would almost certainly be lethal." In the same paper, however, Crick suggests a possible change from two base to three base codons that would avoid violation of 'the Principle of Continuity': "...This is quite different from the idea that the primitive code was a triplet code (in the sense that the reading mechanism moved along three bases each step) but that only, say, the first two bases were read. This is not at all implausible." The codon expansion hypothesis described below is developed from Crick's 1968 statement: I hypothesise that the current triplet genetic code originates from a highly degenerate triplet code that is operational as a singlet code in which the reading mechanism moves three bases per step but in which only the middle base of each codon specifies the encoded amino acid (Figure 2.1). In the subsequent doublet code either the first two bases (as suggested by Crick) or the last two bases of each three are read. Here a generic doublet codon is labelled as \omega, with doublet codons in which the first two bases are read termed prefix codons (\omega\textunderscore) and those in which the last two bases are read termed suffix codons (\textunderscore\omega). As illustrated in Figure 2.1, a codon expansion from singlet to doublet codons leads to a division of the genetic code into two types of doublet codons, prefix codons and suffix codons. In the double-to-triplet codon expansion, prefix doublet codons (\omega\textunderscore) evolve to triplet codons in which the third base is read for the first time, and suffix doublet codons (\textunderscore\omega) evolve to triplet codons in which the first base is read for the first time.
Figure. 2. 2 (a) Codon tables of the doublet code that result from the expansion of a singlet code in which only the middle base, in a three-base reading mechanism, specifies the encoded amino acid. Prefix codons are depicted in light grey, and suffix codons in dark grey. The total coding capacity of the code increases from 4, in the singlet code, to a maximum of 32 codons ($2 \times 16$) in the doublet code model. (b) In a doublet code each of the prefix and suffix codons encodes one amino acid (or a termination signal), then, upon expansion to a triplet code, each doublet codon leads to four possible triplet codons; a prefix codon gives rise to $\omega A$, $\omega B$, $\omega C$, and $\omega D$, and a suffix codon gives rise to $A\omega$, $B\omega$, $C\omega$, and $D\omega$. Codon expansion from doublet to triplet codons, in both the prefix and the suffix codon halves of the double code, leads to the creation of two identical triplet codes.
From a singlet code to a doublet code with two types of codons

According to this hypothesis a single-to-doublet codon expansion results in a virtual division of the genetic code into two types of doublet codons, one in which the first two bases of each three are read ('prefix' codons, \( \omega \)) and the other in which the last two bases of each three are read ('suffix' codons, \( _\omega \)) (Figure 2.1). The total coding capacity of the code increases from 4, in the singlet code, to a maximum of 32 codons (2 \( \times \) 16) in this doublet code model (Figure 2.2a). An expansion of both types of doublet codons to triplet codons would hypothetically create two identical triplet codes (Figure 2.2b). The real consequence of this expansion is that the 16 ambiguous (overlapping) codon assignments from each half of the doublet code would have to compete for positions in the new triplet code.

I tested this doublet code model to see if a codon table pattern like the one observed in the standard genetic code is obtained by merging of the two sets of doublet codons from Figure 2.2a into one triplet codon table. Figures 2.3a–c show the theoretical patterns of triplet codon distributions in a codon table resulting from different possible routes of doublet to triplet codon expansion, based on differences in the specificity for the first base (in prefix codons) and the third base (in suffix codons). Upon expansion to a triplet code, the merging of prefix (\( \omega \)) doublet codons that discriminate between the four nucleotides (A, C, U, or G) at the first base position and suffix codons with no distinction between two purines (recognizing both A and G) or two pyrimidines (recognizing both C and T) at the third nucleotide position results in a codon table (Figure 2.3c) that resembles the pattern of fourfold and two-fold degenerate codons seen in various genetic codes (see also Figure 2.4).
Figure 2.3. Theoretical codon distribution patterns of triplet codons resulting from doublet codon expansion, based on differences in the specificity for the first base (in prefix codons) and the third base (in suffix codons). (a) In a ‘maximal’ doublet code where each of the 16 prefix and 16 suffix codons encodes one amino acid or a termination signal (with a total of 32 effectively encoded doublets), then, upon expansion to a triplet code, each doublet codon leads to four possible triplet codons: a prefix codon gives rise to triplet codons _coA, _coB, _coC, and _coD, and a suffix codon gives rise to _Aco, _Bco, _Cco, and _Dco. For example, upon expansion to a triplet codon, prefix doublet codon AA_ is replaced by four new triplet codon possibilities (AAA, AAB, AAC, and AAD), where A and B could represent purines, and C and D could represent pyrimidines, for example, potentially forming a ‘family box’ of four. Suffix codon _AA is replaced by four new triplet codon possibilities (AAA, BAA, CAA, DAA) that are vertically arrayed in the leftmost column of the codon table. If all triplet codons are uniquely encoded (with no degeneracy), then the
resulting codon table displays a random distribution of 64 triplet codons originating from both prefix and suffix doublet codons, based on the assumption that all possible codons have equal weighting. (b) A doublet code where codons are assigned based on pyrimidine versus purine discrimination in the first base position (in a prefix codon) and in the third base position (in a suffix codon) effectively encodes only 16 of a possible 32 doublets. A triplet codon table resulting from merging of prefix and suffix doublet codons is shown here. For example, in a doublet code with no distinction between two purine bases (for instance, A and B) or two pyrimidine bases (for instance, C and D) at the first base position, prefix codons AA and BA encode the same amino acid (or termination signal). Double-to-triplet expansion of the two doublet codons creates eight triplet codon possibilities (AAA, AAB, AAC, AAD, BAA, BAB, BAC, and BAD), potentially forming a 'family box' of eight. The expansion of suffix codons AA and AB (with no discrimination between two purines or between two pyrimidines at the third base position) creates codons AAA, BAA, CAA, DAA and AAB, BAB, CAB, DAB. The codon pairs AAA and AAB, BAA and BAB, CAA and CAB, DAA and DAB can occupy codon space in a vertical fashion (i.e., in the same column of the codon table). Again, it is assumed that all prefix and suffix codons merge into this triplet code with equal probability occupation of the emerging 64 codon slots. (c) The only triplet codon table that resembles the canonical genetic code originates from prefix doublet codons (ω) discriminating among the four nucleotides at the first base position (16 codons) and suffix doublet codons (ω) with no discrimination between purine (R) or between pyrimidine (Y) bases at the third base position (eight codons). As in (a), a prefix doublet codon gives rise to four new triplet codons, ωA, ωB, ωC, and ωD, potentially forming a family box of four codons in the codon table. From one pair of suffix doublet codons, four pairs of triplet codons can arise that occupy the new triplet codon table in
a vertically arrayed manner, as shown in (b). All family boxes shown in the upper panel originate from prefix doublet codons (shown in light grey). The codon table in the upper panel occurs if it is assumed that all prefix and suffix codons merge into the triplet code with equal weighting. It is expected all codon pairs to result either from suffix codons (shown in dark grey) or from ‘family boxes’ that are equally shared by prefix codons (shown in light grey) and suffix codons (shown in dark grey).

<table>
<thead>
<tr>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>UU</td>
<td>Phe</td>
<td>Ser</td>
<td>UAU</td>
</tr>
<tr>
<td>UUC</td>
<td>Phe</td>
<td>Ser</td>
<td>UAC</td>
</tr>
<tr>
<td>UUA</td>
<td>Leu</td>
<td>Ser</td>
<td>UAG</td>
</tr>
<tr>
<td>UUG</td>
<td>Leu</td>
<td>Ser</td>
<td>UUG</td>
</tr>
<tr>
<td>CU</td>
<td>Leu</td>
<td>Pro</td>
<td>CAU</td>
</tr>
<tr>
<td>CUC</td>
<td>Leu</td>
<td>Pro</td>
<td>CAC</td>
</tr>
<tr>
<td>CUA</td>
<td>Leu</td>
<td>Pro</td>
<td>CAA</td>
</tr>
<tr>
<td>CUG</td>
<td>Leu</td>
<td>Pro</td>
<td>CAG</td>
</tr>
<tr>
<td>AU</td>
<td>Ile</td>
<td>Thr</td>
<td>AAU</td>
</tr>
<tr>
<td>AUC</td>
<td>Ile</td>
<td>Thr</td>
<td>AAC</td>
</tr>
<tr>
<td>AUA</td>
<td>Ile</td>
<td>Thr</td>
<td>AAA</td>
</tr>
<tr>
<td>AUG</td>
<td>Met</td>
<td>Thr</td>
<td>AAG</td>
</tr>
<tr>
<td>GU</td>
<td>Val</td>
<td>Ala</td>
<td>GAU</td>
</tr>
<tr>
<td>GUC</td>
<td>Val</td>
<td>Ala</td>
<td>GAC</td>
</tr>
<tr>
<td>GUA</td>
<td>Val</td>
<td>Ala</td>
<td>GAA</td>
</tr>
<tr>
<td>GUG</td>
<td>Val</td>
<td>Ala</td>
<td>GAG</td>
</tr>
</tbody>
</table>

Key:
- **Acidic**
- **Amide**
- **Hydroxyl containing**
- **Alkyl**
- **Aromatic**
- **Sulfur containing**
- **Basic**
- **STOP**
From these results I conclude that if triplet codons indeed evolved from prefix and suffix doublet codons, they must have originated from prefix codons \((\omega_\_)_\) discriminating between all four bases at the first base position (16 codons) and suffix codons \((_\omega\_\) recognising either purine (R) or pyrimidine (Y) bases at the third base position (eight codons), giving a maximum coding potential of 24.

As a corollary of the codon expansion model, the following issues need to be addressed.

### 2.3 A primordial three-base reading frame

The starting point of this doublet code model, a singlet codon genetic code with a three base reading frame in which the middle base (the only codon position that is non-degenerate) specifies the encoded amino acid, is an essential aspect of this model since it is the only possible starting point for the existence of prefix and suffix doublet codons. Initially the first and third bases of the triplet could have played a role in stabilising the anticodon–codon interaction before they were utilized to encode an expanding number of amino acids. Is there a reason to believe that a primordial code had some intrinsic capacity to define a three-base reading frame with a central encoding nucleotide?

Primordial tRNAs have been thought to be much simpler and evidence has been presented suggesting that a mini-RNA helix containing only the acceptor arm’s base pairing nucleotides was required for charging of the correct amino acid by the cognate aaRS (Figure 2.5; Maizels and Weiner 1994; Schimmel et al 1993).
Figure 2.5. The two halves of contemporary tRNA. The “top half” of tRNA is structurally and functionally independent and may be more ancient than the “bottom half” of the molecule (Maizels and Weiner 1994).

The acceptor stem-containing minihelix has been thought to predate the present-day tRNA structure, where the anticodon helix and loop might have been derived via duplication of the acceptor stem helix (Figure 2.5). The ‘informational’ code of the anticodon is believed to have arisen from specific nucleotides in the acceptor stem, encrypting the ‘second genetic code’. It has been postulated that the central base of the anticodon triplet originates from the second position base of the primordial acceptor minihelix (Rodin et al 1996; Schimmel 1996) and this second base complementarity appears to be conserved in contemporary tRNA. The acceptor–anticodon complementarity relationship for the second base appears to be a vestige of an ancestral genetic code and is in support of the encoding central base proposed in this model.

The interaction of the contemporary tRNA anticodon with an mRNA codon requires accessibility of the anticodon nucleotides in an unpaired region of the tRNA molecule and the triplet base reading frame perhaps results from the size of this unpaired loop containing the anticodon. The size of an RNA hairpin loop is dictated by its stability and by the accessibility of the anticodon nucleotides for base-pairing with the mRNA codon. The size of a canonical
tRNA anticodon loop is 7 nucleotides. Formation of loops with less than 5 bases requires more energy, likewise for loops with more than 10 bases. This indicates that loops with the canonical 7 nucleotides form stable hairpin structures in tRNA (all three tRNA loops are 7–8 nucleotides long) and could thereby determine that mRNA is translated in successive 3-nucleotides steps. Increasing the anticodon loop size from the canonical 7 nucleotides to 8 nucleotides induces frameshifting, suggesting that anticodon loop size is critical for frame maintenance, although expanding the anticodon does not disrupt the translational reading frame by forcing recognition of a larger codon (Farabaugh and Bjork 1999). The selection of a 3-nucleotide reading frame is perhaps also driven by thermodynamic considerations and/or may be an inherent feature of the ribosome (Ogle et al. 2002; Stahl et al. 2002). Analysis of frameshifting indicates that an interaction with mRNA involving fewer than 3 base pairs is insufficiently stable to exclude realignment of the tRNA molecule in the ribosomal P-site (Figure 2.6; Farabaugh and Bjork 1999). The normal 3-bp interaction is just stable enough to prevent frequent slippage. Selection of a 4-bp interaction, causing quadruplet translations as seen in suppressor tRNAs, may have been unfavourable because of codon–anticodon interactions that are too stable or even irreversible. Any need to encode more amino acids (256 codons in a 4-nucleotide code) may have been insufficient.
Figure 2.6. Proposed structural model for the 3-nucleotide translocation step. E, exit site; P, peptidyl site; A, aminoacyl site. The figure shows the complex rotation movement that Wilson and Noller (1998) have proposed occurs during translocation of the peptidyl-tRNA to the P-site of the ribosome. Top, the P- and A-sites are occupied by the peptidyl-tRNA (blue anticodon) and aa-tRNA (yellow anticodon). Middle, peptide transfer creates a deacyl-tRNA (blue) and peptidyl-tRNA (yellow). Bottom, the two tRNAs undergo rotational movements (indicated by the curved arrows in the middle panel) moving them to the E- and P-sites. The rotation is accompanied by movement of the mRNA 3 nucleotides (indicated by the angled arrow in the middle panel). The placement of the E-site tRNA is controversial and this is only one of several possible placements.
2.4 Why are there 64 codons and only 20 amino acids?
The numerical relationship between these numbers has been a matter of
dispute ever since the code was deciphered and has led to many abstract
theories without any biological basis or significance (reviewed by Hayes
1998). The primordial doublet code that forms the basis for the current triplet
genetic code, as hypothesised here, has space for 24 codons. Why the need
for a triplet code with 64 codons, when only 20 amino acids are used?

In the standard genetic code, codons for amino acids leucine, serine, and
arginine occur in both the four-fold degenerate and the two-fold degenerate
halves of the canonical code (Figure 2.4). In this doublet code model this
pattern can occur when each of these amino acids are encoded by a prefix as
well as a suffix doublet codon (see also Section 2.6 below). This would limit
the coding capacity of the model to 21. The requirement for one or more stop
codons in this doublet code system would reduce the number of encoded
amino acids to 20 or fewer, 20 being the number of naturally occurring amino
acids.

2.5 Codon ambiguity: overlapping prefix and suffix codons
Coexistence of prefix and suffix codons within the last common ancestor of all
organisms implies that prefix and suffix doublet codons would have
overlapped, for example, DNA triplet ABC is enclosed within prefix codon AB_
and suffix codon pair (_BC, _BD). Would this overlap between prefix and
suffix codons have caused a translational crisis? Any overlapping prefix and
suffix codon encoding amino acids with similar physicochemical properties
would have had a selective advantage over a doublet code encoding
dissimilar amino acids. In the example of the ABC triplet, all prefix and suffix
codon overlaps are contained within a column of the translation table because
overlapping codons share the same second-position base of a triplet. Figure
2.4 shows that in the present day standard genetic code the second-position
base connects amino acids with comparable physicochemical properties (in
terms of polar requirement) in a remarkably similar fashion (Taylor and Coates 1989; Woese 1965a), a pattern which may well be a relict of an ambiguous doublet genetic code. Encoding amino acids that share the central base of a triplet would have significantly limited prefix and suffix codon ambiguity and consequential translational error. Recognition of an extra base in the three-base doublet codons, the advent of the ‘true’ triplet code, would have eliminated this doublet codon ambiguity.

The model described here therefore not only explains the origin of the error minimising potential of the genetic code, but also provides an explanation of why 20 amino acids are used. If the error minimising potential of the genetic code resulted from an ambiguous doublet code with a natural coding capacity of 20 amino acids, the current genetic code perhaps represents an optimal balance or a compromise between its vocabulary of natural amino acids and its error minimising efficiency.

2.6 Codon patterns

As in the proposed codon expansion model (see Figure 2.3c), the standard genetic code table (Figure 2.4) and its variations (see NCBI Taxonomy Database [Wheeler et al, 2000]) display a strong group feature, as initially recognized by Bertman and Jungck (1979), with the code divided into two octets, one of fourfold degenerate codon boxes and one of twofold codon boxes (the vertebrate mitochondrial code is a typical example of this division). The doublet code model, using prefix and suffix codons, presents a possible historical explanation for the combinations of four-fold and two-fold degenerate codons. The four-fold degenerate triplet codons in the model originate from prefix codons and the two-fold degenerate triplet codons originate from either prefix or suffix codons in the doublet code (see Figure 2.3c). Similar patterns, however, can emerge from a prefix code in which some doublets were four-fold degenerate and others two-fold degenerate, but such a prefix code pattern has 32 codon possibilities and therefore does not explain the number of amino acids in the natural set, nor does it explain the assignments of the paired codons additional to the ‘family box’ assignments.
for those amino acids with six synonyms. It has been explained the specific four-fold and two-fold codon assignments in the triplet genetic code as resulting from the existence of both a prefix and a suffix codon for these amino acids in this doublet code model. For leucine and arginine both four-fold and two-fold degenerate codons share the same second-position base and are therefore in agreement with the error-minimising capacity of this prefix-suffix codon ambiguity model, thereby limiting its amino acid vocabulary.

The assignment for the four-fold (UCN) and two-fold (AGY) serine codons (in different translation table columns), however, cannot be explained as load minimisation. Aminoacyl-tRNA synthetases (aaRSs) are involved in uniquely charging a specific tRNA molecule with its cognate amino acid. Most aaRSs, except SerRS, AlaRS, and LeuRS, recognise their cognate tRNA molecules using an anticodon recognition domain. Serine's codon promiscuity perhaps results from the lack of tRNA anticodon recognition by serine-tRNA synthetase, which does not recognize the anticodon of tRNA^{Ser} but instead recognizes the long extra arm found in all tRNA^{Ser} isoacceptors by means of a C-terminal coiled-coil domain. This would be expected for an enzyme whose substrate is a set of tRNAs that must recognize six different codons (Ibbà, Curnow and Soli 2000).

So far I have described this model and explained how I believe it is better than previous models that do not use two kinds of intermediate doublet codons (see Figure 2.7 for a synopsis). But is there any extant evidence for the evolution of triplet codons in the standard genetic code from primordial doublet codons? For example, can we find relicts of the existence of doublet codons that used either the first two bases (ω.) or the last two bases (._ω) of a three-base reading frame?
What the model explains:

## 20 amino acids

- 24 codons

- 3 Leu, Ser & Arg encoded by both prefix and suffix codons or Leu, Ser & Ala tRNAs lack anticodon recognition by aaRSs

- 1 Stop codon

### Error minimisation

codon ambiguity / overlap for example:

- *A A A* triplet is encoded by:
  - *prefix codon*
  - *suffix codons*

  overlapping codons share 2nd nucleotide in the triplet

### Codon patterns

Patterns of four-fold and two-fold degenerate codons result from the merging of prefix and suffix doublet codons

codonts that share the 2nd nucleotide encode amino acids with similar hydrophobicity

Evidence for the model:

### Prefix codons

* tRNA anticodon nucleotides recognised by:

- ThrRS
- ProRS
- ValRS

### Suffix codons

* tRNA anticodon nucleotides recognised by non-discriminating:

- AspRS-2
- *T. thermophilus*
- GluRS-2
- *H. Pylori*

Figure 2.7. Synopsis of the presented doublet code model.
2.7 Relicts of a doublet code

2.7.1 Four-fold degenerate codons: relics of prefix (ω_) doublet codons?

Each family box of codons comprises four codons with the same prefix and each of the four encodes one amino acid (leucine, valine, serine, proline, threonine, alanine, arginine, and glycine). Could this clustering into codon families be a relict of prefix doublet codons (ωA, ωB, ωC, ωD)?

This pattern of degeneracy/redundancy, which groups together codons for the same amino acid, has been previously explained in the adaptive hypothesis as the result of natural selection. This feature of the code has been reinterpreted and Crick’s (1966) wobble hypothesis explains much of the third base degeneracy (resulting in the four-fold degeneracy of the eight codon families) in terms of simple chemical considerations: a single tRNA anticodon can recognise more than one codon by non-standard base pairing. Uracil in the anticodon allows "wobbles" with U, C, A, and G at the third base position in a codon, permitting the response of a tRNA to entire codon families and purine ending pairs. With the results presented in this study I speculate that the "wobble U" may be a relict of the switch from prefix doublet codons to triplet codons. The discrimination of the third base (between ωR and ωY or even between ωA, ωB, ωC, and ωD) could be a process that arose late in genetic code evolution and is variable in some derived codes, as reflected by the presence in the code of codons encoding single amino acids, such as methionine. In the 'standard code', methionine is encoded by AUG alone. In Drosophila, bovine, and squid codes, however, methionine is encoded by AUA and AUG (Knight and Landweber 2000). These variations in third base recognition emphasise the plasticity of such an evolutionary process, possibly producing alternative genetic codes (discussed below).

Further observations, in tRNA anticodon recognition by aminoacy-tRNA synthetases, bear witness to the existence of doublet codons. aaRSs are
crucial in ensuring the fidelity of translation of the genetic code by attaching the appropriate amino acids to their corresponding tRNAs (Woese et al. 1999, 2000). Recognition and selection of the right tRNA are key in this process and this aspect of aminoacyl-tRNA synthesis has attracted a lot of attention in recent years (Ibba and Soil 2000). aaRSs interact specifically with the tRNA discriminator base, the acceptor stem, and the anticodon and it is in the interaction with the anticodon loop where I find more evidence of doublet codons. Anticodon recognition is achieved by a distinct structural module appended to the core catalytic domain of aaRSs. I have already mentioned the lack of this domain in SerRS and its possible role in the assignment of the six different codons for serine (Section 2.6). Structural rearrangements in the tRNA anticodon (Figure 2.8) appear to play an important role in optimising aaRS-tRNA recognition. Interestingly, studies on the three-dimensional structure of aaRSs in complex with their cognate tRNA revealed that ThrRS, ProRS, and ValRS (all three enzymes are involved in acylation of amino acids with four-fold degenerate codons) show interactions with anticodon bases 35 and 36 (corresponding to codon positions 1 and 2) but not with anticodon base 34 (Fukai et al 2003; Sankaranarayanan et al 1999; Sankaranarayanan and Moras 2001; Yaremchuk et al 2000). tRNA nucleotide 34, the first base of the anticodon, pairing with the third base of the codon, simply serves as a "space filler" in these aaRS-tRNA complexes and could be a vestige of a prefix doublet codon (ω_) in this model (Fig. 2.7).
Figure 2.8. Schematic representation of the three-dimensional structure of tRNA. Circles indicate the position of nucleotides, the size of which is proportional to the documented frequency with which they are involved in recognition by aaRSs. The acceptor stem refers to the helical structure formed by base pairing of bases 1 to 7 with bases 72 to 66 respectively (McClain et al., 1987).
2.7.2 The evolution of aminoacyl-tRNA synthetases: evidence for suffix(ω) doublet codon expansion?

With the codon expansion model presented I can explain the non-random pattern of half of the canonical triplet code as deriving from the use of prefix doublet codons, but can I also find relics of codons with a doublet used as a suffix?

A hint of suffix codon characteristics can be found by inspecting the patterns of initiation codons in different genetic codes in the NCBI Taxonomy Database (Wheeler et al 2000). Like the suffix codons in Figure 2.3, initiation codons can be found at different codon positions that share the second and third base, irrespective of the identity of the first base. There is further evidence, in the archaeal amino acylation system, for the existence of a 5' code expansion mechanism (ω). In most organisms tRNA aminoacylation is ensured by 20 aaRSs. Interestingly, most bacteria, archaea, chloroplasts, and mitochondria lack the aminoacyl-tRNA synthetases that are specific for asparagine (AsnRS) and glutamine (GluRS). A two-step (indirect) aminoacylation pathway is utilised in these organisms and organelles for the formation of Asn-tRNAAsn and Gln-tRNAGln (Bult et al 1996; Dennis 1997; Ibba et al 1997). In Thermus thermophilus, for instance, there are two aspartylation pathways, AspRS-1 and AspRS-2. AspRS-1 is specific for tRNAAsp, while AspRS-2 aspartylates both tRNAAsp and tRNAAsn, even though a fully functional asparaginyl-RNA synthetase (AsnRS) is present. This establishes the coexistence of two aspartylation and asparaginylation pathways where a mischarged Asp on tRNAAsn is converted into asparagine indirectly.

The third base in the anticodon (pairing with the first base in the codon) forms the identity determinant for the recognition of an amino acid-specific tRNA in aspartate/asparagine and glutamate/glutamine systems (Becker et al., 1996; Sekine et al 2001; Shimizu et al., 1992). Comparison of the X-ray crystal structures of AspRS-2 and AsnRS of Thermus thermophilus (Figure 2.9; Charron et al., 2003) provides structural evidence that non-discrimination lies
in the deviant structure of the L1 loop of the AspRS-2 anticodon binding domain and its inability to identify anticodon base C36 (Figure 2.10), the universal aspartate identity determinant. AspRS-2 and other non-discriminating aaRSs, such as GluRS-2 in *Helicobacter pylori* (Tomb et al., 1997), may be molecular fossils (Di Giulio 2002) and could therefore represent an intermediate stage in the evolution of amino acylation where, as in the suffix codons, only the second and third bases of a codon were used to encode the amino acids aspartate and glutamate (Figure 2.5).

**Figure 2.9.** (A) Comparison of anticodon-binding domains in *T. thermophilus* AspRS-2, *P. kodakaraensis* AspRS and *T. thermophilus* AsnRS. Ribbon representation of the anticodon-binding domain in AspRS-2 (left), *Pyrococcus* AspRS (middle) and *Thermus* AsnRS (right). In AspRS-2, α-helix Ha is replaced by the Lα loop. Lα and L1 loops that are specific to AspRS-2 are emphasized by red labels (Charron et al., 2003).
Figure 2.10. L1 loop regions in AspRS and AsnRS anticodon-binding domains. (A) Short L1 loops in *T.thermophilus* AspRS-2 (green), *P.kodakaraensis* AspRS (blue) and *T.thermophilus* AsnRS (pink). (B) Long L1 loops in *T.thermophilus* AspRS-1 (orange), yeast (yellow) and *E.coli* (light blue) AspRSs (for purposes of comparison, the short L1 loop region of AspRS-2 is displayed in light green). Notice in both panels the trace of the tRNA<sup>Asp</sup> backbone (orange) with identity element C36 in contact (or proximity) with L1 loops as seen in the crystal structure of the complex with AspRS-1. In AspRS-2, this loop is small and structurally homologous to that in AsnRSs, including conservation
of a proline. In discriminating *Pyrococcus* AspRS, the L1 loop, although small, lacks this proline and is not superimposable with that of AspRS-2 or AsnRS (Charron et al., 2003).

### 2.8 Genetic code variations and the emergence of a ‘true’ triplet code

The existence of non-discriminating aaRSs for aspartic and glutamic acids indicates that asparagine and glutamine are relatively recent additions to the code (Woese *et al.* 2000). Could the emergence of these ‘new’ amino acids have played a role in a possible change from a doublet to a triplet code in order to overcome doublet codon ambiguity? With most of the 24 codons occupied, the advent of new amino acids may have exhausted the doublet genetic code. Glutamine, one of the amino acids with an aaRS that is nonuniversal throughout life, is used by most bacteria as a nitrogen assimilation intermediate and may have polluted the primordial amino acid pool used for protein synthesis. Biochemically related to glutamic acid, glutamine could have competed for the same tRNA molecules as glutamic acid, causing a shift in the tRNA pool. At a given point the incorporation of glutamine in polypeptide chains could even have interfered with the polypeptide chain termination signal. Interestingly, in several protist lineages the UAA and UAG termination codons are read as glutamine (Knight *et al.* 2001), in a similar non-discriminating fashion (using the second and third bases of the codon, _AR_) with signs of first base (pyrimidine) discrimination (YAR), however, these codes are all derived from the standard code. Analogously, pressure on the genetic code to distinguish glutamine could very likely have encouraged discrimination at the first base position of the codon (i.e., between Aω, Bω, Cω, and Dω) in the suffix doublet code, thereby gradually paving the way for the foundation of the "true" triplet genetic code. Comparison of crystal structures of discriminating aaRSs in the NAN-codon group, AspRS, GlnRS (Sekine *et al.* 2001), LysRS (Cusack *et al.*, 1996), and, more recently, TyrRS (Kobayashi *et al.*, 2003), shows that the recognition of the first and the second bases of the anticodon (pairing with the third and
second base of the codon) is conserved. Specific recognition of the anticodon's third base in these aaRSs, determining the tRNA's identity, could therefore have evolved later.

The proposed mechanism for the incorporation of glutamine into the code is also applicable to the recently discovered amino acid additions to the code, pyrrolysine and selenocysteine. Pyrrolysine, for instance, shares an indirect conversion mechanism very much like the two-step aminoacylation pathway of glutamine and asparagine, described under the evolution of Aminoacyl-tRNA Synthetases (above), involving the combined action of two lysyl-tRNA synthetases (Srinivasan et al., 2002). Lysine and pyrrolysine, like aspartate/asparagine and glutamate/glutamine sets, share doublet codon affinities (_AR, in the same column of the codon table), using the first base in the triplet as an identity discriminator (lysine, RAR; pyrrolysine, YAR). More recently a second pathway has been discovered involving a natural twenty-first synthetase that directly activates pyrrolysine, both in vitro and in vivo (Blight et al., 2004; Polycarpo et al., 2004), showing that a non-canonical amino acid like pyrrolysine can be seized by the genetic code using a natural tRNA-synthetase pair.

In this model the CAR and UAR codon set for glutamine in some derived genetic codes may be reverse evolution, returning to an earlier possible code that had greater codon ambiguity. Other variations from the standard genetic code may result from similar divergences in the recognition of the first or the third base of the anticodon (corresponding to the third and the first base of the codon, respectively). AGR codons in the vertebrate mitochondrial code, for instance, encode a termination signal that may have arisen after the branching of the primary lineages when in other organisms/organelles UGR was preferred as stop codon. Discrimination between a termination signal and the incorporation of tryptophan (UGG) in these coded necessitated improved recognition of all three anticodon nucleotides at a time point after this diversion. Improved recognition of the methionine anticodon (preferring CAU over UAU) may have resulted from a similar process toward a more specific triplet code.
2.9 An ancestral doublet code

aaRSs for only 17 (or even only 16 [Ibba and Soll 2000]) of the 20 amino acids appear to be universal: as mentioned above most bacteria, archaea, chloroplasts, and mitochondria lack the aminoacyl-tRNA synthetases that are specific for asparagine (AsnRS) and glutamine (GlnRS). In addition to the lacking of AsnRS and GlnRS, at least two species of Archaea bacteria, *Methanococcus jannaschii* and *Methanobacterium thermoautotrophicum*, lack known homologues of cysteinyl-tRNA synthetases, suggesting that these amino acids may have also occurred later in evolution than others (Brown and Doolittle 1999; Woese et al. 2000). Some bacterial taxa have acquired a specific GlnRS, probably by horizontal gene transfer, from the eukaryota (Brown and Doolittle 1999; Lamour et al. 1994).

This doublet code model has a capacity of 24 codons with a vocabulary of 20 amino acids. So if the evolution of tRNA aminoacylation indeed reflects genetic code evolution, this would indicate the existence of an ancestor code with a vocabulary of amino acids (excluding amino acids with non-universal aaRSs: asparagine, glutamine, cysteine, and lysine). A possible precursor of the 24-codon model could have been the highly degenerate 16-codon genetic code represented in Figure 2.3b, in which there is no discrimination between two purines or between two pyrimidines at the first base position in the prefix doublet codons and at the third base position in suffix doublet codons. The advent of amino acids asparagine and glutamine (and perhaps cysteine) would have created evolutionary pressure on the 16 codons and may have triggered the discrimination of the first base in some prefix (ω-) codons. When this first base recognition became more specific and discriminated between two purines or pyrimidines in all prefix codons, the codon capacity of the genetic code would have increased to 24 codon possibilities.

With a simple primordial code, a limited number of primitive short-branched amino acids may have been utilized in peptide synthesis encoded by acceptor stem-containing mini RNA helices. Relicts of such a code may have survived
in the aminoclylation by AlaRS, SerRS and LeuRS, where specific base pairs in the acceptor stem define the identity of the tRNA and no contact is made between the enzyme and the anticodon (Giege et al 1998; Ribas de Pouplana and Schimmel 2001), a feature that is preserved in the four-fold degeneracy of these codons. With this lack of anticodon recognition, the aaRSs for alanine, serine, and leucine must predate this doublet codon model, thereby presenting a second rationale for the 24 doublet codon model encoding only 20 amino acids (plus stop codon). The codon ambiguity in this doublet code model would have required adaptations to discriminate between (isosteric) amino acids, by means of proofreading and editing as seen in ValRS and IleRS (Sankaranarayanan and Moras 2001). Alongside the establishment of amino acid discrimination, a feature that would have been greatly adaptive in limiting the ambiguous coding, the incorporation of domains involved in the recognition of the anticodon, either as prefix or as suffix, would have facilitated the expansion of the code and the diversification of living organisms. Structural arrangements in the tRNA anticodon and modification of its nucleotide also provided a general means of optimising recognition. Assignment of new amino acids required duplication and divergence of existing synthetase domains, including the improved recognition of anticodon nucleotides. From the evolution of aaRSs it appears that these processes (and the formation of a ‘true’ triplet code) were not yet fully established at the time the universal ancestor gave rise to the primary lineages of the tree of life.

2.10 Conclusion
In summary, this codon expansion model, inspired by Crick’s (1968) suggestion of a doublet code using a three-base reading frame, provides a new and simple explanation of how triplet codons evolved, or still are evolving, from two types of primordial doublet codons, without the deleterious effect of a frameshift. The gradual change from a doublet code using two kinds of highly degenerate triplet codons, providing codon space for 20 amino acids, to the current triplet code system would have limited doublet codon ambiguity and created a code pattern very similar to that of the present day canonical code.
I believe that the emergence of new variations in the ‘universal genetic code’ and the discovery of new amino acids indicate that the code structure is still subject to natural expansion, perhaps gradually bringing more of the 64 codons into use with increased diversity and specificity. The fact that many tRNAs and also many proteins require post-translational modifications and additional cofactors for function may indicate the need for such a process (Bock et al. 1991; Mehl et al 2003). A compromise between its vocabulary of natural amino acids and its error minimising efficiency, however, may have fixed the genetic code in its current format.
PART TWO

The NMR Structure of Burkholderia Pseudomallei Type III Secreted Protein BopE
3.1 Abstract

*Burkholderia pseudomallei* is a Gram-negative motile bacillus that is the causative agent of melioidosis. The molecular mechanisms of *B. pseudomallei* pathogenesis are not well understood. A number of putative type III secreted effector proteins have been identified recently by analysis of the *B. pseudomallei* genome sequence. One of these proteins, BopE, is a homologue of the potent guanine nucleotide exchange factors (GEF) SopE and SopE2 from *Salmonella enterica* (Figure 3.1). BopE is a type III secreted protein from *Burkholderia pseudomallei*. The sequence homology between SopE/E2 and BopE catalytic domains suggests the proteins share similar folds. Importantly, a number of SopE residues making contacts with Cdc42 in the SopE-Cdc42 crystal complex (Buchwald et al., 2002) are largely conserved or conservatively substituted in BopE. Therefore, it seems likely, that BopE utilizes the same mechanisms to engage and catalyse guanine nucleotide exchange in Rho GTPases as the SopE-like GEFs. However, the sequence homology between SopE/SopE2 and BopE catalytic domains (approximately 25% identity and 40% similarity) is not sufficiently high to say with certainty that the SopE/SopE2 and BopE domains adopt the same three-dimensional fold. I will report the three-dimensional structure in solution of the catalytic domain of BopE (BopE residues 78-261) in Part 2, Chapter 5, to our knowledge the first effector protein from *Burkholderia pseudomallei* to be characterised in this way.
Figure 3.1. Sequence alignment of guanine nucleotide exchange factors *Salmonella* SopE and SopE2 and BopE from *B. pseudomallei*. Conserved residues are highlighted in black and conservatively substituted residues in grey. SopE residues contacting Cdc42 in the SopE-Cdc42 crystal structure are highlighted by filled circles (Upadhyay et al., 2003).

### 3.2 Classification of *Burkholderia*

*Burkholderia* is a genus of proteobacteria, including the pathogens responsible for glanders (*Burkholderia mallei*) and for melioidosis (*Burkholderia pseudomallei*). Yabuuchi and colleagues (Yabuuchi et al., 1992) established the genus *Burkholderia* in 1992. They transferred seven species of the genus *Pseudomonas* homology group II to the new genus, with the type species designated as *Burkholderia cepacia*. The rationale for the new genus was based on the 16S rRNA sequences, DNA-DNA homology values, cellular lipid and fatty acid composition, and phenotypic characteristics. Thus, seven new species were created: *Burkholderia cepacia*, *Burkholderia mallei*, *Burkholderia pseudomallei*, *Burkholderia caryophilli*, *Burkholderia gladioli*, *Burkholderia pickettii*, and *Burkholderia solanacearum*. Based on biochemical, immunological and genetic data, *B.*
pseudomallei, B. mallei, B. cepacia and B. thailandensis are relatively closely related (Figure 3.2).

**Figure. 3.2.** The taxonomy of *Burkholderia* (Dworkin, 2001).

*B. cepacia* is an important opportunistic respiratory pathogen, particularly in patients with cystic fibrosis (CF) and chronic granulomatous disease. *B. cepacia* can have an incidence rate as high as 40% in CF patients. Twenty percent of these cases will develop acute exacerbation leading to rapid deterioration (Govan and Deretic, 1996). *B. mallei*, the causative agent of human and equine glanders, and *B. pseudomallei*, which causes melioidosis, are closely related by their nucleotide sequence similarity and by the aetiology of the diseases they cause (Woods et al., 1999).
3.3 *Burkholderia pseudomallei*

It has been reported that in more than a dozen major Gram-negative bacterial pathogens of animals and plants, virulence is largely dependent on type III secretion systems (TTSS). Type III secretion systems are key virulence determinants of *Salmonella*, *Shigella*, and other Gram-negative facultative intracellular pathogens and serve to inject bacterial proteins into target cells. A subset of type III secretion system secreted proteins (translocators) is believed to interact with the eukaryotic cell membrane and mediate the delivery of secreted effector proteins. Once inside host cells the effector proteins subvert host cell processes to the benefit of the bacteria (Cornelis, et al., 2000).

One of the Gram-negative bacteria dependent on TTSS, *B. pseudomallei*, contains a cluster of five genes with high similarity to genes of a HRP (hypersensitive response and pathogenicity) locus in the plant pathogen *Ralstonia solanacearum* and several ORFs in this locus showing significant similarity with TTSS-associated genes of plant pathogens *R. solanacearum*, *Xanthomonas campestris* and *Pseudomonas syringae* (Winstanley et al., 1999). *B. pseudomallei* is apparently a human pathogen. It is unexpected that a plant pathogen-like TTSS system would be used for human cell intoxication. Some other studies suggested a putative type III protein secretion apparatus (Bsa) in *B. pseudomallei* similar to the *Salmonella enterica* Inv/Spa/Prg systems (Stevens, et al., 2003) (Figure 3.3). Therefore *B. pseudomallei* is the first bacterial pathogen apparently armed with one TTSS system for infecting animal cells and another for plant cells. This represents a novel level of diversity and complexity in the bacterial world, and presents the challenge of elucidating the respective origins and roles of these two TTSS systems.
Figure 3.3. Gene organization of *Burkholderia pseudomallei* type III secretion system locus Bsa and comparison to *Salmonella* Typhimurium pathogenicity island-1 (Stevens, et al., 2003).

### 3.4 Melioidosis

Melioidosis, also known as pseudoglanders and Whitmore's disease (after Alfred Whitmore), is an infectious disease caused by *B. pseudomallei*. The bacteria causing melioidosis are contaminant found in water and soil and are spread to humans and animals through direct contact with the contaminated source. Melioidosis is endemic in Southeast Asia, with the greatest concentration of cases reported in Vietnam, Cambodia, Laos, Thailand, Malaysia, Myanmar, and northern Australia. It is also seen in the South Pacific, Africa, India, and the Middle East. In many of these countries, *Burkholderia pseudomallei* is so prevalent that it is a common contaminate found on laboratory cultures (White, 2003).

Besides humans, many animal species are susceptible to melioidosis. These include sheep, goats, horses, swine, cattle, dogs, and cats. Humans and animals are believed to acquire the infection by inhalation of dust, ingestion of
contaminated water, and contact with contaminated soil especially through skin abrasions, and for military personnel, by contamination of war wounds (Dance, 2002). Due to the severity of the infection, aerosol infectivity and worldwide availability, both \textit{B. pseudomallei} and \textit{B. mallei} are considered to be potential bioweapons (Wheelis, 1989; Rotz et al., 2002).

Melioidosis is diagnosed by isolating \textit{Burkholderia pseudomallei} from the blood, urine, sputum, or skin lesions. Detecting and measuring antibodies to the bacteria in the blood is another means of diagnosis. Antibiotic therapy of melioidosis is long and difficult, because of the resistance of the bacterium to many antibiotics and a tendency to relapse after recovery from clinical disease. There is no vaccine for melioidosis. Illness from melioidosis can be categorized as acute or localized infection, acute pulmonary infection, acute bloodstream infection, and chronic supplicative infection. Inapparent infections are also possible. Patients with chronic or latent melioidosis may be symptom-free for decades. The incubation period (time between exposure and appearance of clinical symptoms) is not clearly defined, but may range from 2 days to many years (Currie et al., 2000). Person-to-person transmission can occur. There is one report of transmission to a sister with diabetes who was the caretaker for her brother who had chronic melioidosis. Two cases of sexual transmission have been reported. Transmission in both cases was preceded by a clinical history of chronic prostatitis in the source patient (Brett et al., 2000). There have been three recent reports of \textit{B. pseudomallei}-infected adults with cystic fibrosis (CF) who presumably acquired the infection during extended vacations or residence in either Thailand or northern Australia (O'Carroll et al., 2003).
3.5 BopE

BopE is a type III secreted protein from *Burkholderia pseudomallei*. It has been shown that inactivation of *bopE* impairs bacterial entry into HeLa cells indicating that BopE facilitates invasion. Consistent with this notion, BopE expressed in eukaryotic cells induced rearrangements in the subcortical actin cytoskeleton, and purified BopE exhibited guanine nucleotide exchange factor activity for Cdc42 and Rac1 *in vitro* (Stevens et al., 2003).

The molecular mechanisms of *B. pseudomallei* pathogenesis are not completely understood (Stevens et al., 2004). The recently completed *B. pseudomallei* genome sequence (Holden et al., 2004) revealed a 7.3 Mb genome, unusually large for a prokaryote, comprising two chromosomes with sixteen genomic islands possibly acquired through very recent lateral transfer. One of these (Bsa) is homologous to the Inv/Spa/Prg type III secretion system (TTSS) of *Salmonella typhimurium* (Attree et al., 2001; Rainbow et al., 2002; Stevens et al., 2002).

BopE, encoded within the Bsa locus, is secreted via the Bsa apparatus and influences invasion of HeLa cells probably via its function as a guanine nucleotide exchange factor (GEF) for Rho GTPases that regulate the actin network (Stevens et al., 2003). BopE shares sequence homology with the *Salmonella* translocated effector proteins SopE (Wood et al., 1996) and SopE2 (Bakshi et al., 2000) which play an important role in *Salmonella* invasion of non-phagocytic intestinal epithelial cells. SopE is a potent guanine nucleotide exchange factor for the mammalian Rho GTPases Cdc42 and Rac1 *in vitro* and *in vivo* whereas SopE2 efficiently activates Cdc42 but not Rac1 (Hardt et al., 1998; Friebel et al., 2001).
3.6 Type III secretion system

The type III secretion system (TTSS), triggered by a close contact of the bacterium with eukaryotic host cells, involves the assembly of a dedicated secretion/translocation apparatus enabling the injection of pathogenicity (effector) proteins directly into the host cells (Table 3.1). The system components are encoded by a set of approximately twenty genes that are usually clustered in the bacterial genome, forming so-called pathogenicity islands (PIs).

The TTSS has been described in a growing number of Gram-negative pathogenic and symbiotic animal, insect and plant bacteria, including species of *Yersinia, Shigella, Pseudomonas, Burkholderia, Ralstonia, Xanthomonas, Chlamydia, Bordetella, Rhizobia, Escherichia, Photorhabdus* and *Erwinia* (Hueck, 1998, 2000; Krishnan, 2002). A major difference between animal and plant pathogens is that the latter interact with the cell cytoplasm by piercing from outside the 200 nm thick plant cell wall, while animal pathogens have to deal with only about 5 nm thick cell membranes.

Morphologically, the TTSS consists of cytoplasmic, transmembrane and extracellular domains (Figure 3.4). The cytoplasmic and transmembrane domains form the basal body, which controls protein export and provides the energy required for secretion. The basal bodies of the TTSS and the flagellum possess similar macromolecular arrangements, with several component proteins sharing either sequence or functional similarity.
<table>
<thead>
<tr>
<th>Protein</th>
<th>PDB code</th>
<th>Role in TTSS biology</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Apparatus components</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MxiM</td>
<td>1Y9L, 1Y9T</td>
<td>Pilotin responsible for guiding insertion/stabilising outer membrane ring of basal body</td>
</tr>
<tr>
<td>EscJ</td>
<td>1YJ7</td>
<td>Equivalent to FlhF in flagellar system, forms periplasmic ring closely associated with inner membrane ring</td>
</tr>
<tr>
<td>HrcQb C-terminal domain</td>
<td>1O9Y</td>
<td>Component of C-ring-like assembly on cytoplasmic side of bacterial inner membrane</td>
</tr>
<tr>
<td>LcrV</td>
<td>1R6F</td>
<td>Essential for insertion of pore components?</td>
</tr>
<tr>
<td>EspA</td>
<td>1XOU</td>
<td>Needle extension subunit</td>
</tr>
<tr>
<td><strong>Chaperones</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CesA-effector</td>
<td>1XOU</td>
<td>Specific chaperone for EspA</td>
</tr>
<tr>
<td>SycN–YscB-effector</td>
<td>1XKP</td>
<td>Specific chaperone for YopN</td>
</tr>
<tr>
<td>SycH-effector fragment</td>
<td>1TTM</td>
<td>Specific chaperone for YscM2</td>
</tr>
<tr>
<td>Spa15</td>
<td>1RY9</td>
<td>General chaperone</td>
</tr>
<tr>
<td>ShcF-effector</td>
<td>1S28</td>
<td>Specific chaperone for AvrPphF</td>
</tr>
<tr>
<td>SycT</td>
<td>2BH0, 2BSH</td>
<td>Specific chaperone for YopT</td>
</tr>
<tr>
<td><strong>Effectors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YscM2–chaperone</td>
<td>1TTW</td>
<td>Secreted regulator of secretion</td>
</tr>
<tr>
<td>AvrPto</td>
<td>1R5E</td>
<td>Interacts with host cell Pto kinase</td>
</tr>
<tr>
<td>AvrPphF–chaperone</td>
<td>1S21</td>
<td>Unknown</td>
</tr>
<tr>
<td>AvrPphB</td>
<td>1UKF</td>
<td>Cysteine protease</td>
</tr>
<tr>
<td>AvrB</td>
<td>1NH1</td>
<td>Induces phosphorylation of host RIN4</td>
</tr>
<tr>
<td>Protein</td>
<td>PDB code</td>
<td>Role in TTSS biology</td>
</tr>
<tr>
<td>------------------------</td>
<td>----------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>YopN-switch peptide</td>
<td>1XKP, 1XL3</td>
<td>Secreted regulator of secretion</td>
</tr>
<tr>
<td>YopH-inhibitor</td>
<td>1QZ0, 1XXV</td>
<td>Protein tyrosine phosphatase</td>
</tr>
<tr>
<td>SopE2-target</td>
<td>1R6E, 1R9K</td>
<td>Guanine nucleotide exchange factor</td>
</tr>
<tr>
<td>YopR</td>
<td>1Z21</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

**Table 3.1.** List of all TTSS structures published since 2004 (Johnson et al., 2005).

The basal body possesses an ~25 Å diameter channel through which all exported proteins must pass in a partially or completely unfolded state (Figure 3.4B). The supramolecular structure of the *Salmonella* TTSS apparatus resembles the flagellar basal body and comprises two parts: (i) a 7- to 8-nm-wide, 60-nm-long external needle; and (ii) a shorter cylinder formed by plates (20-40 nm in diameter) that presumably traverse both bacterial membranes and the peptidoglycan layer (Figure 3.4; Kubori et al., 1998; Kubori et al., 2000). Similar structures have also been observed in *S. flexneri* (Blocker et al., 2001; Cordes et al., 2003). The actual mechanisms of type III secretion are poorly understood.
Figure 3.4. (A) 3D reconstruction of the needle complex (NC) in the bacterial membrane. The inner and outer membrane rings made of the proteins PrgH, PrgK and InvG form large homomultimeric annular complexes, anchoring the NC in the bacterial membrane. Prgl polymerises in a helical manner to form the extra-cellular needle. (B) The insert is a cross section of the extra-cellular needle showing the helical nature of the structure and the internal pore diameter. Since the effector proteins SopE
and SopE2 are about 40 Å in width, it is thought that, in order to be secreted by the type III system, proteins must be unfolded or only partially folded (Buchwald et al., 2002). (C) Electron micrograph of purified needle complex, above right, isolated from S. Typhimurium (Cordes et al., 2003).

3.7 Rho GTPases
Rho GTPases are molecular switches that control a wide variety of signal transduction pathways in all eukaryotic cells. Rho GTPases are members of the Ras superfamily of 20- to 30-kDa GTP-binding proteins. The family numbers over sixty in mammals, and falls into five major groups: Ras, Rho, Rab, Arf and Ran (Etienne-Manneville and Hall, 2002). The proteins exist in two interconvertible forms: the GDP-bound inactive and GTP-bound active forms. Active GTPases interact with their specific downstream targets and perform their cellular functions, whereas GTP hydrolysis and liberation of phosphate inactivate the GTPases (Figure 3.5). The GTP-GDP exchange reactions are regulated by guanine nucleotide exchange factors (GEFs), guanine nucleotide dissociation inhibitors (GDIs) and GTPase-activating proteins (GAPs). Rho GTPases are key regulators for a wide range of cellular responses such as cytoskeleton reorganisation, focal adhesions, motility, cell shape change, membrane ruffling, cytokinesis, cell aggregation, cell-to-cell adhesion, gene expression, mitogenic signalling and malignant transformation (Van Aelst & D'Souza-Schorey, 1997; Ridley 2001a; Ridley, 2001b).
Figure 3.5. The GTPase molecular switch. Guanine nucleotide exchange factors (GEFs) release guanosine diphosphate (GDP) from Rho GTPases promoting the binding of guanosine triphosphate (GTP) and activation of Rho GTPases. GDP dissociation inhibitor (GDI) inhibits the dissociation of GDP from Rho GTPases and thus prevents association of GDP-GTPase to the cell membrane. GTPase activating proteins (GAPs) stimulate the intrinsic GTPase activity of Rho GTPases and convert the GTP-bound form of Rho GTPases to inactive GDP-GTPases. In resting cells Rho GTPases exist mostly in GDP-bound form and in complexes with Rho GDI in the cytosol. The GTP-bound form of Rho GTPases is associated with cell membranes.

The Rho GTPase Cdc42 is an essential component of the host cellular signalling cascades triggered by the translocated *S.typhimurium* effector
proteins during invasion. The effector proteins SopE and SopE2 bind directly to and activate host cellular Cdc42 by catalyzing guanine nucleotide exchange. Turnover numbers (k_cat) for nucleotide dissociation from Cdc42 are 5 s^{-1} for SopE and 19 s^{-1} for SopE2 respectively (Fribel et al., 2001).

3.8 Guanine nucleotide exchange factors (GEFs)
Guanine nucleotide exchange factors (GEFs) stimulate the dissociation of the GDP nucleotide from the Rho GTPases and promote the binding of GTP and activation of Rho GTPases. GEFs are varied both in structure and sequence. It has reported that the Ran GEF RCC1 adopts a 7 bladed β-propeller fold (Renault et al., 2001), while other GEFs are all α-helical proteins but with completely different arrangements of their helices (Figure 3.6). Most GEFs are multi-domain proteins. Actually all known GEFs that act on members of the Rho/Rac/Cdc42 family contain the Dbl homology (DH) domain followed by a tandem pleckstrin homology (PH) domain (Cerione & Zheng, 1996). Compared to these proteins, however the SopE/SopE2 family that also acts on members of the Rho/Rac/Cdc42 family from pathogenic bacteria comprises only a single domain (Figure 3.7).
Figure 3.6. Structures of guanine nucleotide exchange factors. (A) RCC1, the exchange factor of Ran, is a seven-bladed β-propeller. B) The Dbl and pleckstrin homology domains from HSos1, a GEF for Ras (Renault et al., 2001).

Figure 3.7. Structure of the Salmonella GEF domain (A) SopE2 (Williams, et al., 2004) and (B) SopE (Buchwald et al., 2002). Ribbon diagrams of the structures were produced with Molscript (Kraulis, 1991) and rendered with POV-RAY™.
Biochemical analyses have demonstrated that, unlike many other bacterial toxins, SopE and SopE2 activate host cellular Rho family proteins by transient interaction and not by chemical modification. Even though SopE and SopE2 do not share any recognizable sequence similarity to any known proteins from bacteria or eukaryotes, particularly the DH family proteins (Figure 3.8), they act as genuine GEFs (Stender et al., 2000).
Figure 3.8. Guanine nucleotide exchange region of Dbs in complex with Cdc42 (John Sondek and co-workers, UNC Chapel Hill). The major conformational changes are observed in the switch I/II regions of Cdc42 when α5a of Dbs inserted between the switch regions of Cdc42, where s1 is switch1 and s2 is switch 2 respectively in the above figure.
3.9 Structure of SopE-Cdc42 complex

SopE (crystal structure; Buchwald et al., 2002) and SopE2 (NMR structure; Williams et al., 2004) are composed of six α-helices arranged in two three-helix bundles. These two bundles are arranged in a Λ-shaped fashion (Figure 3.7) with size ~40 Å in length and ~30 Å in width. There is only a small two-stranded β-sheet and a peptide segment (residues 165–172) connects the two sides of the ‘Λ’. This segment harbours a GAGA169 motif (Figure 3.7A, in red), which is proposed to represent the catalytic core of SopE (Williams et al., 2004; Buchwald et al., 2002).

The SopE-Cdc42 complex interface is formed primarily by helices α2, α4 and α5 from SopE, including flanking regions that communicate mainly with the switch regions of Cdc42. The major conformational changes are observed in the switch I/II regions of Cdc42 (Figure 3.9): the catalytic GAGA169-loop of SopE is inserted between the switch regions of Cdc42 and appears to exert a push-and-pull-type movement whereby switch I is pushed aside and switch II is pulled towards the loop (Buchwald et al., 2002).
Figure 3.9. The SopE–Cdc42 complex. (A) Ribbon representation of the complex with SopE in blue and Cdc42 in yellow. The switch regions are shown in green. The catalytic loop inserted between
the switch regions is highlighted in red. (B) Schematic representation of the SopE–Cdc42 interaction with important interactions indicated (Buchwald et al., 2002).
4.1 Methods and experiments

The materials and methods used for study of BopE have been described previously (Upadhyay et al., 2004; Wu et al., 2004). The $^1$H, $^{13}$C and $^{15}$N chemical shifts of BopE$_{78-261}$ are in the BioMagResBank database (http://www.bmrb.wisc.edu) under accession number BMRB-5974.

Chemicals were from Sigma unless otherwise stated; solutions were prepared with MilliQ water (Millipore reagent water system, Milford, MA). $^{15}$NH$_4$Cl, $^{13}$C$_6$-D-glucose and $^2$H$_2$O were from Cambridge Isotope Laboratories, Inc. (Cambridge, Mass., USA) and from Spectra Stable Isotopes, Spectra Gases Inc (Branchburg, NJ, USA). Restriction grade thrombin was obtained from Calbiochem. FPLC columns and affinity resins were from Amersham Biosciences.

A DNA fragment encoding BopE residues 78–261 was amplified by PCR using _B. pseudomallei_ chromosomal DNA as a template and cloned into the pGEX-2T vector (Amersham Biosciences; cloning done by the group of Dr Edouard Galyov, BBSRC Institute for Animal Health). The resulting plasmid was used to transform _E. coli_ strain BL21(DE3). $^{15}$N-labelled and $^{15}$N/$^{13}$C-labelled BopE$_{78-261}$ were produced by expression in minimal medium with $^{15}$NH$_4$Cl and $^{15}$NH$_4$Cl/$^{13}$C$_6$-D-glucose as the sole nitrogen and nitrogen/carbon sources.

GST-BopE$_{78-261}$ was purified using glutathione sepharose 4B resin (Amersham Biosciences) according to the manufacturer's instructions. Human plasma thrombin (Calbiochem) was used to cleave BopE$_{78-261}$ from GST whilst bound to the glutathione sepharose 4B resin. BopE$_{78-261}$ was
subsequently purified further using anion exchange chromatography (Mono Q, Amersham Biosciences). The final protein contains Gly and Ser from the pGEX-2T vector at the N-terminus followed by BopE_{78-261}.

NMR samples contained 10 mg of BopE_{78-261} in 0.5 ml of 20 mM potassium phosphate, pH 5.5, 50 mM NaCl, 10 mM DTT, 1 mM EDTA, 1 mM benzamidine and 50 μM NaN₃ in 95% H₂O / 5% D₂O. The protein solution was transferred to either Wilmad 535-PP or Shigemi NMR tubes under a nitrogen atmosphere.

NMR data were recorded on a Varian Unity Inova 600 spectrometer at 25 °C. Sequence-specific backbone resonance assignments were made using 3D HNCA, HN(CO)CA, HNCACB, CBCA(CO)NH, HNCO, HN(CA)CO, (HB)CBCACO(CA)HA, HNHA and HBHA(CBCACO)NH data sets. Side chain resonance assignments were made using CCC-TOCSYNNH, HCC-TOCSY-NNH, HCCCH TOCSY and ^{15}N edited TOCSY (12.1 ms, 12.1 ms, 15.6 ms and 50 ms mixing times) spectra. Sequence-specific assignments of aromatic side chains were obtained using ^{1}H–^1H 2D TOCSY (28 ms and 50 ms mixing times), and intra-residue NOEs between the βCH2 and/or α-proton and the aromatic ring protons in ^{1}H–^1H 2D NOESY spectra (100 ms and 175 ms mixing times) and a 3D simultaneous ^{15}N/^{13}C-edited NOESY spectrum (100 ms mixing). Full details of all experiments performed on ^{15}N-labelled BopE_{78-261} are summarised in Table 4.1.

<table>
<thead>
<tr>
<th>Spectral Width (Hz)</th>
<th>Experiments</th>
<th>F₁</th>
<th>F₂</th>
<th>F₃</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>8000</td>
<td>8000</td>
<td></td>
<td>Macura et al., 1980</td>
</tr>
<tr>
<td></td>
<td>^{1}H–^{1}H Nosey</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>^{1}H–^{1}H Tocsy</td>
<td>8000</td>
<td>8000</td>
<td></td>
<td>Bax et al., 1985</td>
</tr>
<tr>
<td></td>
<td>^{1}H–^{15}N–HSQC</td>
<td>8000</td>
<td>2000</td>
<td></td>
<td>Kay et al., 1992</td>
</tr>
<tr>
<td></td>
<td>^{1}H–^{15}N NH₂</td>
<td>8000</td>
<td>2000</td>
<td></td>
<td>Kay et al., 1992</td>
</tr>
<tr>
<td></td>
<td>selective HSQC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence</td>
<td>T1 (ms)</td>
<td>T2 (ms)</td>
<td>T1' (ms)</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>---------</td>
<td>---------</td>
<td>----------</td>
<td>----------------------------------</td>
<td></td>
</tr>
<tr>
<td>HNCO</td>
<td>8000</td>
<td>2112</td>
<td>2000</td>
<td>Grzesiek et al., 1992a</td>
<td></td>
</tr>
<tr>
<td>HN(CA)CO</td>
<td>8000</td>
<td>2112</td>
<td>2000</td>
<td>Kay et al., 1992</td>
<td></td>
</tr>
<tr>
<td>HNCA</td>
<td>8000</td>
<td>4200</td>
<td>2000</td>
<td>Muhandiram et al., 1994</td>
<td></td>
</tr>
<tr>
<td>HN(CO)CA</td>
<td>8000</td>
<td>2112</td>
<td>2000</td>
<td>Bax et al., 1991</td>
<td></td>
</tr>
<tr>
<td>CBCA(CO)NH</td>
<td>8000</td>
<td>9300</td>
<td>2000</td>
<td>Muhandiram et al., 1994</td>
<td></td>
</tr>
<tr>
<td>CBCANH</td>
<td>8000</td>
<td>9300</td>
<td>2000</td>
<td>Muhandiram et al., 1994</td>
<td></td>
</tr>
<tr>
<td>CC(CO)NH</td>
<td>8000</td>
<td>11461.6</td>
<td>2000</td>
<td>Grzesiek et al., 1993</td>
<td></td>
</tr>
<tr>
<td>HC(CO)NH</td>
<td>8000</td>
<td>3000</td>
<td>2000</td>
<td>Sattler et al., 1995</td>
<td></td>
</tr>
<tr>
<td>HBHA(CBCACO)NH</td>
<td>8000</td>
<td>3600</td>
<td>2000</td>
<td>Grzesiek et al., 1992b</td>
<td></td>
</tr>
<tr>
<td>(HB)CBCACO(CA)HA</td>
<td>6000.6</td>
<td>9300.2</td>
<td>1810.6</td>
<td>Kay et al., 1993</td>
<td></td>
</tr>
<tr>
<td>$^{13}$C HCCH-TOSCY</td>
<td>8000</td>
<td>4500</td>
<td>3600</td>
<td>Kay et al., 1990</td>
<td></td>
</tr>
<tr>
<td>$^{13}$C HCCH-COSY</td>
<td>8000</td>
<td>3600</td>
<td>4500</td>
<td>Bax et al., 1990</td>
<td></td>
</tr>
<tr>
<td>$^{15}$N-edited NOESY-HSQC</td>
<td>8000</td>
<td>6600</td>
<td>2000</td>
<td>Zhang et al., 1994</td>
<td></td>
</tr>
<tr>
<td>$^{15}$N-edited TOCSY-HSQC</td>
<td>8000</td>
<td>6599.6</td>
<td>2000</td>
<td>Marion et al., 1989</td>
<td></td>
</tr>
<tr>
<td>HNHA</td>
<td>8000</td>
<td>6599.6</td>
<td>2000</td>
<td>Kuboniwa et al., 1994</td>
<td></td>
</tr>
<tr>
<td>HNHB</td>
<td>8000</td>
<td>6599.6</td>
<td>2000</td>
<td>Archer et al., 1991</td>
<td></td>
</tr>
<tr>
<td>$^{13}$C$^{15}$N-NOSEY</td>
<td>8000</td>
<td>2112</td>
<td>2000</td>
<td>Pascal et al., 1994</td>
<td></td>
</tr>
<tr>
<td>HACAN_SQ</td>
<td>8000</td>
<td>1400</td>
<td>3620.1</td>
<td>Kanelis et al., 2000</td>
<td></td>
</tr>
<tr>
<td>(HB)CBCA(CO)N(CA)HA</td>
<td>8000</td>
<td>2112</td>
<td>2000</td>
<td>Kanelis et al., 2000</td>
<td></td>
</tr>
<tr>
<td>$^{15}$N-COSY</td>
<td>6599.6</td>
<td>2000</td>
<td></td>
<td>Marion et al., 1989</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1. Acquisition Parameters for MNR experiments performed on *Burkholderia pseudomallei* BopE<sub>78-261</sub>.
All data were processed using NMRPipe (Delaglio et al., 1995) and analysed using Sparky (T. D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco). $^1$H chemical shifts were referenced to DSS. $^{15}$N and $^{13}$C chemical shifts were referenced indirectly to DSS (Wishart et al., 1995).

The $^1$H-$^{15}$N HSQC spectrum of the $^{15}$N-labelled GEF domain of BopE is shown in Figure 4.1. Backbone N, NH, Ca, Cβ, CO, Hα and Hβ resonances have been assigned for all but one of the residues in BopE$_{78-261}$. Complete assignment of side chain $^{13}$C and $^1$H resonances has been made for approximately 90% of these residues.

**Figure 4.1.** $^1$H-$^{15}$N HSQC spectrum of uniformly $^{15}$N-enriched BopE catalytic domain (residues 78-261). Assignments are indicated alongside the corresponding signals. Side chain NH$_2$ groups of Asn and Gln are indicated by lines parallel to the $^1$H axis.
4.2. Methods for resonance assignment

The strategy for determination of a protein's 3D structure by NMR first requires that signals in NMR spectra be assigned to individual atoms or functional groups of the protein. Backbone resonances (those of amide N, amide H, Cα, Hα, C(O), Cβ and Hβ) are assigned first. The resonances of aliphatic and aromatic side chain C's and H's are then assigned. Various methods have been created to complete this requirement. The following spectra have been used to complete BopE resonance assignment.

1. Backbone assignment

$^{15}$N HSQC. HSQC (Figure 4.1) is a 2D spectrum. It is used as a template of $^1$H and $^{15}$N resonances to aid the analysis of other 3D spectra that have $^1$H and $^{15}$N dimensions.

$^{15}$N HSQC_NH$_2$. N15HSQC_NH$_2$ is used to select for side chain NH$_2$ groups of asparagines and glutamine residues.

HN(CO)CA. HN(CO)CA correlates each backbone $^1$H and $^{15}$N amide resonance with the Cα resonance of the previous amino acid in the protein sequence.

HNCA. HNCA correlates each backbone $^1$H and $^{15}$N amide resonance with the Cα resonance of the current and also often the previous amino acid in the protein sequence.

CBCA(CO)NH. CBCA(CO)NH correlates each backbone $^1$H and $^{15}$N amide resonance with the Cα and Cβ resonance of the previous amino acid in the protein sequence.
**HNCACB.** HNCACB correlates each backbone $^1$H and $^{15}$N amide resonance with the Cα and Cβ resonance of the current amino acid and also often the previous amino acid in the protein sequence.

**HNCO.** HNCO correlates each backbone $^1$H and $^{15}$N amide resonance with the C(O) resonance of the previous amino acid in the protein sequence.

**HN(CA)CO.** HN(CA)CO correlates each backbone $^1$H and $^{15}$N amide resonance with the C(O) resonance of the current amino acid in the protein sequence.

**HCA(CO)CANH.** HCA(CO)CANH correlates each backbone $^1$H and $^{15}$N amide resonance with the C(O) resonance of the current amino acid in the protein sequence, the same as HN(CA)CO does.

**HBHACBCA(CO)NH.** HBHACBCA(CO)NH correlate each backbone $^1$H and $^{15}$N amide resonance with the Hα and Hβ resonances of the previous amino acid in the protein sequence.

**$^{15}$N TOCSY.** $^{15}$N TOCSY correlates each backbone $^1$H and $^{15}$N amide resonance with the Hα resonance of the current amino acid in the protein sequence. (in smaller proteins, e.g. less than 10kDa, this experiment will also correlate each backbone $^1$H and $^{15}$N with side chain Hβ, Hγ resonances etc.).

**HNHA.** HNHA correlates each backbone $^1$H and $^{15}$N amide resonance with the Hα resonance of the current amino acid in the protein sequence.

**HNHB.** HNHB correlates each backbone $^1$H and $^{15}$N amide resonance with the Hβ resonance(s) of the current amino acid in the protein sequence.

**(HB)CBCA(CO)CAHA.** (HB)CBCA(CO)CAHA can be used to check existing assignments as it correlates Hα, Cα, Cβ and C(O) resonances within an amino acid, independently of the backbone $^1$H and $^{15}$N amide resonance.
Using the above spectra, 100% backbone assignment of BopE$_{78-261}$ was completed.

2. Side chain assignment
The following spectra were used to determine the chemical shift assignments of side chain nuclei.

**C(CO)NH.** C(CO)NH correlates the backbone $^{1}\text{H}$ and $^{15}\text{N}$ resonance of one amino acid with the side chain aliphatic $^{13}\text{C}$ resonances of the previous amino acid.

**H(CCO)NH.** H(CCO)NH correlates the backbone $^{1}\text{H}$ and $^{15}\text{N}$ resonances of one amino acid with the side chain aliphatic $^{1}\text{H}$ resonances of the previous amino acid.

**HCCH TOCSY.** HCCH TOCSY is used to provide those aliphatic $^{1}\text{H}$ and $^{13}\text{C}$ chemical shifts not assigned in C(CO)NH and H(CCO)NH.

**HCCH COSY.** HCCH COSY does as HCCH TOCSY except that in HCCH COSY, correlations are only observed between $^{1}\text{H}$ nuclei that are three or fewer covalent bounds apart. This means that HCCH COSY can be used for spin system identification.

**C13arom_CBHD.** C13arom_CBHD is used to find chemical shifts of Hδ in aromatic rings.

**C13arom_CBHE.** C13arom_CBHE is used to find chemical shifts of He in aromatic rings.

**WG TOCSY.** WG TOCSY is used to find $^{1}\text{H}$ chemical shifts of aromatic rings.
TN NOESY. TN NOESY is used to find $^1$H chemical shifts of aromatic rings and to assign NOEs involving aromatic ring protons.

4.3. Structural restraints

In order to determine the 3D structure of BopE$_{78-261}$, the known chemical shifts were used to find distance restraints. Nuclear Overhauser enhancement (NOE) NMR spectra are used to provide this information. NOE distance restraints were obtained by analysis of $^1$H-$^1$H 2D NOESY (100ms and 175ms mixing times) (Macura et al., 1980), $^{15}$N-NOESY HSQC (50ms, 100ms, and 150ms mixing times) (Zhang et al., 1994) and simultaneous 3D $^{15}$N/$^{13}$C-edited NOESY (100ms mixing time) spectra (Pascal et al., 1994).

$^{15}$N NOESY. $^{15}$N NOESY is a 3D spectrum used to find NOEs between the backbone and side chain amide $^1$H’s and other types of $^1$H.

$^{13}$C NOESY. $^{13}$C NOESY is a 3D spectrum used to find aliphatic-aliphatic, aliphatic-aromatic and aromatic-aromatic NOEs as well as NOEs involving amide protons.

TN NOESY. TN NOESY is a 2D $^1$H-$^1$H spectrum that in smaller proteins can be used for assignment of aliphatic and aromatic NOEs, but in larger proteins (12-15 kDa and above) the 2D $^1$H-$^1$H NOESY is typically only useful for assignment of aromatic-aromatic and aromatic-aliphatic NOEs.

4.4 Structure calculation

Each NOE was assigned to one of four restraint distances based on the peak intensity: 1.8-2.8 Å, 1.8-3.3 Å, 1.8-5.0 Å and 1.8-6.0 Å, corresponding to strong, medium, weak and very weak NOEs. Distances involving methyl groups, aromatic ring protons, and non-stereospecifically assigned methylene protons were represented as a $(\sum r^{-6})^{-1/6}$ sum (Nilges et al., 1993). Backbone dihedral angles $\phi$ and $\psi$ were predicted from $^{13}$C$_{\alpha}$, $^{13}$C$_{\beta}$, $^{13}$C', $^1$H$_{\alpha}$ and
backbone $^{15}$N chemical shifts using TALOS (Cornilescu et al., 1999). Dihedral
angles were restrained to TALOS-predicted values ± 30° and ± 50° for α-
helices and β-strands, respectively. Hydrogen bond restraints were obtained
from hydrogen-deuterium exchange experiments: uniformly $^{15}$N-labelled
BopE$_{78-261}$ in NMR buffer was lyophilized and resuspended in 99.96% D$_2$O. A
series of $^1$H-$^{15}$N HSQC spectra was then recorded to determine amide
protons protected from exchange with the solvent. NH resonances remaining
after two hours were assigned as having hydrogen bonds. For hydrogen
bond distance constraints, the NH-O distance was assigned lower and upper
distance bounds of 1.5 and 2.8 Å, and the N-O distance assigned lower and
upper distance bounds of 2.4 and 3.5 Å.

Structures were calculated using XPLOR-NIH 2.06 (Schwieters et al., 2003),
using the standard simulated annealing protocol starting from random
extended structures. Default values were used for all force constants and
molecular parameters. The ensemble of NMR structures was analysed for
violated restraints using the VMD-XPLOR visualization package (Schwieters
et al., 2001). The structure determination was carried out iteratively whereby
consistently violated restraints were reassigned, wherever possible, using
existing structures or removed until a consistent set of constraints was
obtained with few violations in the ensemble. The ensemble of structures was
further refined with XPLOR-NIH standard refinement protocols using the final
set of restraints. The quality of the structures was checked using
PROCHECK-NMR (Laskowski et al., 1996).
5.1 Abstract

BopE is a type III secreted protein from *Burkholderia pseudomallei*, the aetiological agent of a severe emerging infection of humans and animals called melioidosis. BopE is a guanine nucleotide exchange factor for the Rho GTPases Cdc42 and Rac1. The three-dimensional structure of the catalytic domain of BopE (amino acids 78-261) by NMR spectroscopy is reported here. It shows that BopE$_{78-261}$ comprises two three-helix bundles ($\alpha_1\alpha_4\alpha_5$ and $\alpha_2\alpha_3\alpha_6$) with the nucleotide exchange catalytic motif located in the loop connecting helices $\alpha_3$ and $\alpha_4$. This fold is similar to that adopted by the catalytic domains of BopE homologues SopE and SopE2, *Salmonella* guanine nucleotide exchange factors for Rho GTPases. Whereas the two three-helix bundles of SopE$_{78-240}$ and SopE2$_{69-240}$ form the arms of a ‘A’ shape, however, BopE$_{78-261}$ adopts a closed conformation with substantial interactions between the two three-helix bundles. I propose that prolines play a significant role in these conformational differences between BopE and SopE/E2: three prolines cause the $\alpha_5$-$\alpha_6$ loop of SopE$_{78-240}$ and SopE2$_{69-240}$ to bulge out and keep the three-helix bundles apart, whereas BopE$_{78-261}$ contains only one proline in the $\alpha_5$-$\alpha_6$ connecting region; and BopE proline 204 disrupts $\alpha_5$, positioning part of $\alpha_5$ as a bridge between the two three-helix bundles. Analysis of the molecular interface in the SopE$_{78-240}$-Cdc42 complex crystal structure indicates that, in a BopE-Cdc42 interaction, the closed conformation of BopE catalytic domain would engender steric clashes with the switch regions of Cdc42. This implies that BopE$_{78-261}$ must undergo a conformational change from closed to open forms in order to catalyse guanine nucleotide exchange. The requirement for such a conformational change may explain the lower turnover number of BopE compared to SopE and SopE2.
5.2 Structure analysis
Analytical ultracentrifugation experiments in conjunction with analytical size exclusion chromatography show that BopE$_{78-261}$ is monomeric in aqueous solution. CD spectroscopy indicates that the protein is predominantly $\alpha$-helical, with predicted secondary structure composition of 59% $\alpha$-helix and 7% $\beta$-strand (Upadhyay et al., 2004). Overall, BopE has approximately 16 % and 17 % sequence identity with the *Salmonella* guanine nucleotide exchange factors SopE and SopE2. Within the catalytic domain (comparing residues 78-240 of SopE and SopE2 with residues 78-240 of BopE), the sequence identity/similarity with SopE and SopE2 is approximately 25 %/40 % and 24 %/39 %. Moreover, BopE is predicted to have similar secondary structure to SopE/E2 and residues involved in the intermolecular interface of the SopE-Cdc42 complex (Buchwald et al., 2002) are largely conserved or conservatively substituted in all three proteins, suggesting the proteins have similar three dimensional structures. Furthermore, SopE residues (Asp103, Gln109, Asp124 and Gly168) identified by mutational studies to be critical for catalytic efficiency are conserved in all three proteins (Schlumberger et al., 2003). It seems likely, therefore, that BopE utilises the same mechanism as SopE in catalysing guanine nucleotide exchange in Rho GTPases.

5.3 Structure determination
As described in previous chapters, NMR data have been collected and analysed in order to determine the three-dimensional structure of BopE catalytic domain. A semi-automated procedure for iterative NOE assignment was used to generate the structure of BopE$_{78-261}$. The final structures were generated using 2546 NOE-derived distance restraints (comprising 853 intra-residue, 1312 sequential and medium range and 381 long range NOEs), 202 hydrogen bond restraints and 176 $\psi$ and $\phi$ dihedral angle restraints. Twenty structures were chosen to represent the ensemble of NMR structures on the basis of lowest energy and minimal distance and torsional angle restraint
violations. All of these structures were consistent with both experimental data and standard covalent geometry, displaying no violations greater than 0.5 Å for distance restraints or 5° for dihedral angles. Details of the final set of structural restraints and their violations within the final ensemble are listed in Table 5.1.

Table 5.1. Structural Statistics on the Final Set of NMR-Derived Structures of the BopE GEF Domain

<table>
<thead>
<tr>
<th>Total number of NOE restraints</th>
<th>2546</th>
</tr>
</thead>
<tbody>
<tr>
<td>intraresidue</td>
<td>853</td>
</tr>
<tr>
<td>sequential/med. range (i to i+1-4)</td>
<td>1312</td>
</tr>
<tr>
<td>long range</td>
<td>381</td>
</tr>
<tr>
<td>Number of dihedral angle restraints</td>
<td>176</td>
</tr>
<tr>
<td>Number of hydrogen bond restraints</td>
<td>202</td>
</tr>
<tr>
<td>Rmsd for backbone atoms</td>
<td>1.12^a, 1.02^b</td>
</tr>
</tbody>
</table>
| Average numbers of NOE violations
  >0.3Å (per structure)          | 10   |
| >0.5Å (per structure)          | 0    |
| Average number of dihedral angle violations
  >5° (per structure)           | 0    |

<table>
<thead>
<tr>
<th>Ramachandran plot regions^c</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Most favoured (%)</td>
<td>78.9</td>
<td>76.6</td>
</tr>
<tr>
<td>Additional allowed (%)</td>
<td>15.5</td>
<td>15.9</td>
</tr>
<tr>
<td>Generously allowed (%)</td>
<td>3.7</td>
<td>5.8</td>
</tr>
</tbody>
</table>

^a The rmsd from the mean structure calculated over residues 84-97, 111-131, 144-156, 160-169, 178-188, 206-215, and 223-247. ^b The r.m.s.d from the mean structure calculated over residues 84-97, 111-131, 144-156, 178-188, 206-215, and 223-247. ^c Calculated with PROCHECK-NMR.

The ensemble of structures and the structure closest to the mean are shown in Figure 5.1.
Figure 5.1. Structure of the BopE GEF domain (residues 78-261). (A) Backbone (N, Cα and C') trace of the 20 lowest energy structures coloured as a continuum from blue at N-terminus to red at C-terminus. Backbone trace was generated with MOLMOL (Koradi et al., 1996)
Figure 5.1. Structure of the BopE GEF domain (residues 78-261). (B) Ribbon diagram of the structure closest to the mean coloured as in part A. The α-helices and β-hairpin are labelled. The $^{171}$GAGT$^{174}$ catalytic motif lies between the β-hairpin and α4. Ribbon diagram was generated with Molscript (Kraulis et al., 1991).

The ensemble of structures has a backbone r.m.s.d. from the mean over the regular secondary structure elements of 1.12 Å. A Ramachandran plot (Figure 5.2) of the structures with PROCHECK-NMR (Laskowski et al., 1996) indicates that 94.4 % of the residues (excluding Gly and Pro residues) lie in the most favoured or additionally allowed regions. The few non-glycine residues to fall into the generously allowed regions and disallowed regions correspond to residues located at the termini or loop regions where the NMR restraint density is low. The structure closest to the mean of the ensemble was chosen for illustrative purposes and for further discussion.
Figure 5.2. An example Ramachandran plot for the NMR structures of BopE generated with PROCHECK-NMR (Laskowski et al., 1996).
5.4 Three-dimensional structure of BopE$_{78-261}$

BopE$_{78-261}$ consists of six $\alpha$-helices arranged in two three-helix bundles, $\alpha_1\alpha_4\alpha_5$ and $\alpha_2\alpha_3\alpha_6$. The two bundles interact extensively to form a compact overall structure (Figure 5.1B) with the relative orientation of the bundles defined by 49 inter-bundle NOEs (compared with 20 inter-bundle NOEs for SopE$_{269-240}$; Williams et al., 2004). This fold is very similar to that adopted by the catalytic domains of BopE homologues SopE and SopE2, *Salmonella* guanine nucleotide exchange factors for Rho GTPases. Whereas the two three-helix bundles of SopE$_{78-240}$ and SopE$_{269-240}$ form the arms of a ‘Λ’ shape, BopE$_{78-261}$ adopts a closed conformation with substantial interactions between the two three-helix bundles (Figure 5.3). The three-helix bundles are connected by a loop between helices $\alpha_1$ and $\alpha_2$, a $\beta$-hairpin (residues 162 to 169) followed by a loop that contains the $^{171}$GAGT$^{174}$ catalytic motif between helices $\alpha_3$ and $\alpha_4$, and a type II turn between helices $\alpha_5$ and $\alpha_6$. The $\beta$-hairpin displays a higher r.m.s.d. than other secondary structure elements and omission of the $\beta$-hairpin from the calculation lowers the backbone RMSD from 1.12 Å to 1.02 Å.
Figure 5.3. Backbone superposition of the solution structure of SopE_{69-240} (blue, Williams et al., 2004) with the solution structure of BopE_{78-240} (yellow), demonstrating the similarities and differences in the two structures. Both structures consist of two three-helix bundles with linkers that include a $\beta$-hairpin followed by a loop that contains the GAGA (GAGT in BopE) catalytic motif (shown in red). Backbone superposition was generated with MOLMOL (Koradi et al., 1996).
5.5 Sequence and secondary structure comparison of *B. pseudomallei* BopE\(_{78-261}\) with *Salmonella* SopE\(_{78-240}\) and SopE\(_{26-240}\)

As mentioned previously, BopE has been identified (Stevens et al., 2002) as a homologue of the *Salmonella* effector proteins SopE and SopE2 (Figure 5.4). Overall, BopE has approximately 16% and 17% sequence identity with SopE and SopE2. Within the catalytic domain (comparing residues 78-240 of SopE and SopE2 with residues 78-240 of BopE), the sequence identity/similarity with SopE and SopE2 is approximately 25%/40% and 24%/39%.

The secondary structures of BopE\(_{78-261}\), SopE\(_{78-240}\) and SopE\(_{26-240}\) are generally similar (Figure 5.4 and Figure 5.5).
Figure 5.4. BopE, SopE, SopE2 and CopE sequence and secondary structure comparison. Sequence alignment (performed using ClustalW) of guanine nucleotide exchange factors BopE from *Burkholderia pseudomallei* (BopE B.psm, GenBank accession number CAH38998), SopE from *Salmonella Typhimurium* (SopE.tm, GenBank accession number AF043239), SopE2 from *Salmonella Dublin* (SopE2.dbi, GenBank accession number Q9K1Z2) and *Chromobacterium violaceum* CopE (CopE C.vtm, GenBank accession number AAQ57975). Residues conserved in at least three of the sequences are highlighted in green and conservatively substituted residues in blue. SopE residues that interact with Cdc42 in the SopE78-240-Cdc42 crystal structure are indicated by black filled circles, except interacting...
residues from the catalytic GAGA motif (residues 166-169 in SopE and SopE2) that are indicated by red filled circles (Buchwald, 2002). Secondary structure elements are indicated as red blocks (α-helices) and cyan arrows (β-strands). The secondary structure for *Chromobacterium violaceum* CopE was predicted using the PredictProtein server (http://www.predictprotein.org/).
Figure 5.5. (A) Comparison of the solution structure of BopE\textsubscript{78-261} (red; PDB code 2AIC) with the solution structure of SopE\textsubscript{26-240} (blue; PDB code 1R6E and 1R9K) and crystal structure of Cdc42-bound SopE\textsubscript{78-240} (green; PDB code 1GZS) demonstrating the similarities and differences in the SopE\textsubscript{78-240}, SopE\textsubscript{26-240} and BopE\textsubscript{78-261} structures. All three structures consist of two three-helix bundles connected by a $\beta$-hairpin followed by a loop that contains the $^{168}$GAGA$^{169}$ (SopE/E2) / $^{171}$GAGT$^{174}$ (BopE) catalytic
motif. Ribbon diagrams of the structures were produced with Molscript (Kraulis, 1991) and rendered with POV-Ray™.

Figure 5.5. (B) Molecular surface images representing the Cdc42-interacting face of the structures of SopE_{78-240} (left above), SopE_{69-240} (right above) and BopE_{78-261} (bottom), oriented as in Figure 5.5A. Surfaces are coloured according to electro-potential: blue represents a positive charge and red a negative charge. The pattern of charge distribution over the active site-containing surfaces of BopE_{78-261}, SopE_{78-240} and SopE_{78-240} is grossly similar: all three proteins contain a central belt of positive charge with a concentration of negative charge near the apex of the Λ. The most significant charge distribution differences shown by
BopE\textsubscript{78-261} may be removed if, as anticipated, BopE\textsubscript{78-261} undergoes a conformational change upon interaction with target Rho GTPases (see Section 5.7). Molecular surface images were produced using GRASP (Nicholls et al., 1991).

One major difference is that BopE helix \( \alpha 6 \) is considerably longer than its SopE and SopE2 counterparts; this is partly due to the fact that BopE has sixteen extra amino acids at its C-terminus. The functional significance of the greater length of \( \alpha 6 \) and C-terminal extension of BopE is unknown, but it might be speculated for example that the BopE C-terminal extension is involved in nucleotide exchange by interaction with host cell Rho GTPases to facilitate the push-pull mechanism of nucleotide release observed in the SopE\textsubscript{78-240-Cdc42} crystal structure (Buchwald et al., 2002) and presumably adopted by other members of the SopE/SopE2/BopE family of bacterial guanine nucleotide exchange factors. It is worth noting here that the NOE density in the region of the BopE\textsubscript{78-261} C-terminal tail is very low (see Figure 5.1A), perhaps reflecting the flexibility required for this part of BopE to influence the catalysis of nucleotide exchange. Detailed structural analysis of a BopE-GTPase complex is needed to confirm or rule out the possibility of a functional role for the extended C-terminal region of BopE.

A second major difference between BopE and SopE/E2 catalytic domains is that BopE\textsubscript{78-261} helix \( \alpha 5 \) is disrupted by residue Pro204 (Figures 5.1 and 5.6). As well as a secondary structure difference, this has important consequences for the tertiary structure of BopE\textsubscript{78-261}; these consequences are described below. For ease of comparison of BopE\textsubscript{78-261} with its SopE/E2 counterparts, I have retained the numerical order of the helices by terming the resulting two parts of \( \alpha 5 \) produced by this disruption as \( \alpha 5' \) and \( \alpha 5'' \).
Figure 5.6. SopE$^{78-240}$ and BopE$^{78-261}$ fold comparison. In order to highlight the major secondary structure and conformational differences between SopE$^{78-240}$ (green) and BopE$^{78-261}$ (purple), the $\alpha$-helices of the two structures are shown and the locations of relevant proline residues are highlighted. Note the contrast between the protuberance of the $\alpha5-\alpha6$ loop in SopE$^{78-240}$ and the compactness of the corresponding turn in BopE$^{78-261}$. Also note the disruption of BopE$^{78-261}$ helix $\alpha5$ into two parts, labelled $\alpha5'$ and $\alpha5''$, that permits $\alpha5'$ in particular to interact with the $\alpha2\alpha3\alpha6$ bundle. Both characteristics appear to be due to the presence or absence of proline residues and result in the greater compactness of BopE$^{78-261}$ relative to SopE$^{78-240}$ and SopE$^{26-240}$. 

99
5.6. Tertiary structure comparison of *B. pseudomallei* BopE<sub>78-261</sub> with *Salmonella* SopE<sub>78-240</sub> and SopE<sub>69-240</sub>

BopE<sub>78-261</sub> adopts a similar fold to its *Salmonella* counterparts SopE<sub>78-240</sub> and SopE<sub>69-240</sub> but is more closed with extensive interaction between the two three-helix bundles (Figures 5.5 and 5.6). As an illustration of the more extensive association between the bundles in BopE<sub>78-261</sub>, 49 inter-bundle NOEs were assigned in BopE<sub>78-261</sub> compared with 20 such NOEs in the previous structure determination of SopE<sub>69-240</sub> by this laboratory (Williams et al., 2004). Further illustration of the different relative arrangements of the three-helix bundles comes from comparison of inter-helical angles, several of which are very different between BopE<sub>78-261</sub> and the two *Salmonella* GEFs (Table 5.2).

**Table 5.2:** Comparison of Helix Crossing Angles in the Solution Structures of BopE<sub>78-261</sub> and SopE<sub>69-240</sub> and the Crystal Structure of Cdc42-Bound SopE<sub>78-240</sub>

<table>
<thead>
<tr>
<th>helix pair</th>
<th>BopE&lt;sub&gt;78-261&lt;/sub&gt;</th>
<th>SopE&lt;sub&gt;69-240&lt;/sub&gt;&lt;sup&gt;b&lt;/sup&gt;</th>
<th>SopE&lt;sub&gt;78-240&lt;/sub&gt;&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Δ1&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Δ2&lt;sup&gt;e&lt;/sup&gt;</th>
<th>Δ3&lt;sup&gt;f&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>-158.37</td>
<td>-149.36</td>
<td>-141.08</td>
<td>-9.01</td>
<td>-17.29</td>
<td>-8.28</td>
</tr>
<tr>
<td>1-3</td>
<td>-26.25</td>
<td>-32.32</td>
<td>-43.49</td>
<td>6.07</td>
<td>17.24</td>
<td>11.17</td>
</tr>
<tr>
<td>1-4</td>
<td>147.08</td>
<td>131.71</td>
<td>132.28</td>
<td>15.37</td>
<td>14.80</td>
<td>-0.56</td>
</tr>
<tr>
<td>1-5</td>
<td>-18.60</td>
<td>-31.76</td>
<td>-35.21</td>
<td>13.16</td>
<td>16.61</td>
<td>3.45</td>
</tr>
<tr>
<td>1-6</td>
<td>161.15</td>
<td>134.07</td>
<td>123.33</td>
<td>27.08</td>
<td>37.82</td>
<td>-7.42</td>
</tr>
<tr>
<td>2-3</td>
<td>166.50</td>
<td>161.34</td>
<td>165.24</td>
<td>5.16</td>
<td>1.26</td>
<td>-3.89</td>
</tr>
<tr>
<td>2-4</td>
<td>-39.80</td>
<td>-76.45</td>
<td>-79.58</td>
<td>36.65</td>
<td>39.78</td>
<td>3.13</td>
</tr>
<tr>
<td>2-5</td>
<td>-150.74</td>
<td>-134.40</td>
<td>-133.28</td>
<td>16.34</td>
<td>17.46</td>
<td>-1.12</td>
</tr>
<tr>
<td>2-6</td>
<td>-6.60</td>
<td>-19.65</td>
<td>-17.78</td>
<td>13.05</td>
<td>11.18</td>
<td>-1.87</td>
</tr>
<tr>
<td>3-4</td>
<td>150.04</td>
<td>112.53</td>
<td>105.35</td>
<td>37.51</td>
<td>44.69</td>
<td>7.18</td>
</tr>
<tr>
<td>3-5</td>
<td>23.79</td>
<td>31.15</td>
<td>38.08</td>
<td>-7.36</td>
<td>-14.29</td>
<td>-6.93</td>
</tr>
<tr>
<td>3-6</td>
<td>170.87</td>
<td>163.25</td>
<td>159.99</td>
<td>7.62</td>
<td>10.88</td>
<td>3.26</td>
</tr>
<tr>
<td>4-5</td>
<td>165.68</td>
<td>141.83</td>
<td>142.39</td>
<td>23.85</td>
<td>23.29</td>
<td>-0.55</td>
</tr>
<tr>
<td>4-6</td>
<td>-33.17</td>
<td>-94.23</td>
<td>-94.23</td>
<td>51.01</td>
<td>61.06</td>
<td>10.05</td>
</tr>
<tr>
<td>5-6</td>
<td>157.04</td>
<td>132.82</td>
<td>121.91</td>
<td>24.22</td>
<td>35.13</td>
<td>10.91</td>
</tr>
</tbody>
</table>

<sup>a</sup> Calculated with Interlhx (K. Yap, University of Toronto). <sup>b</sup>SopE2 GEF domain NMR structure (PDB entry 1R9K). <sup>c</sup>Sop E GEF domain crystal structure (PDB entry 1GZS). <sup>d</sup>The BopE<sub>78-261</sub> helix crossing angle minus the SopE<sub>69-240</sub> helix crossing angle. <sup>e</sup>The BopE<sub>78-261</sub> helix crossing angle minus the SopE<sub>78-240</sub> helix crossing angle. <sup>f</sup>The SopE<sub>69-240</sub> helix crossing angle minus the SopE<sub>78-240</sub> helix crossing angle.
The tight interactions between the two three-helix bundles of BopE<sub>78-261</sub> involve an intricate network of charge and hydrophobic interactions. Prominent among the residues involved are five arginine residues at positions 100, 182, 200, 207 and 230 that are almost unique to BopE: SopE and SopE2 do not possess arginines in any of the corresponding positions but the putative bacterial GEF family member CopE has arginines corresponding to BopE Arg200 and Arg207 (Figure 5.4). Three of the BopE arginines, Arg200, Arg207 and Arg230, form an association between helices α5 and α6 while Arg182 links helices α4 and α2 at the Cdc42-binding face of BopE through a salt bridge with Glu125. Arg100 (in helix α1) occupies a hydrophobic pocket between helix α2 and helix α5. The conformational change from closed to open BopE that would seem to be required to allow binding with Cdc42 (described in the next section) requires that the salt bridge between Arg182 and Glu125 (connecting helices α4 and α2) is disrupted, perhaps triggering the release of some of the other arginine-mediated interactions.

BopE residue Pro204 (corresponding to Ile199 in SopE and Ala199 in SopE2) promotes these inter-bundle interactions by disrupting helix α5 into two parts termed α5' and α5''. As a consequence, α5' is positioned to bridge the α1α4α5 and α2α3α6 bundles and its residues are able to interact with residues in α2 and α6 of the α2α3α6 bundle (Figure 5.6).

In contrast to BopE proline 204, three SopE/E2 prolines appear to impede inter-bundle interaction and therefore contribute to the more open conformation adopted by SopE<sub>69-240</sub> in solution relative to BopE<sub>78-261</sub>. Near to the apex of the Λ formed by the two three-helix bundles, the loop connecting SopE<sub>69-240</sub> and SopE<sub>78-240</sub> helices α5 and α6 bulges, presumably due to the presence of Pro211, Pro219 and Pro221. Due to the lack of prolines at positions corresponding to 219 (Asn224) and 221 (Leu226), BopE<sub>78-261</sub> helix α6 begins earlier in the amino acid sequence than SopE/E2 α6 (Figure 5.4 and Figure 5.7.) and the BopE<sub>78-261</sub> α5-α6 connecting element is a four-residue type II turn rather than the seven-residue loop observed in SopE<sub>69-240</sub> and SopE<sub>78-240</sub>. I reason that this protrusion of the polypeptide
chain in the α5-α6 loop at the apex of the Λ, not observed in BopE78-261 due to key amino acid differences, counteracts extensive inter-bundle interaction in SopE78-240 and SopE263-240 (Figure 5.7).
Figure 5.7. Structure Comparison in α5-α6 of BopE$_{78-261}$ (red) and SopE$_{78-240}$ (green) shows the difference in number and location of proline residues. Three prolines cause the α5-α6 loop of SopE$_{78-240}$ and SopE$_{269-240}$ to bulge out and keep the three-helix bundles apart, whereas BopE$_{78-261}$ contains only one proline in the α5-α6 connecting region; and BopE proline 204 disrupts α5, positioning part of α5 as a bridge between the two three-helix bundles.
5.7 Implications of tertiary structure differences between BopE_{78-261} and SopE_{78-240} / SopE2_{69-240} for interaction with Rho GTPases

The question then arises as to whether the conformational difference between the catalytic domain of BopE and those of SopE and SopE2 has implications for interaction with Rho GTPases. Analysis of the interface between SopE_{78-240} and Cdc42 in the SopE_{78-240}-Cdc42 complex crystal structure (Buchwald et al., 2002) reveals that the interaction can be broken down into two major components (Figure 5.8): a ridge on Cdc42 formed by residues 35 to 41 (switch region I) fits into a groove on SopE_{78-240} and residues Val36 and Asp38 of Cdc42 wedge into the gap between the two three-helix bundles of SopE_{78-240}. The latter interaction in particular indicates that, in its closed conformation, BopE_{78-261} would experience steric clashes with Cdc42. The resulting implication is that BopE catalytic domain must undergo a change from its closed conformation to a more open conformation like those of SopE and SopE2 catalytic domains in order to carry out its guanine nucleotide exchange function. A requirement for such a large scale conformational change is consistent with, and may at least partially explain, the observed differences in turnover number for guanine nucleotide exchange between BopE_{78-261} and its Salmonella counterparts: a $k_{cat}$ of 0.48 s$^{-1}$ was measured for BopE_{78-261} - induced guanine nucleotide exchange in Rac1 (a similar rate was measured for Cdc42) (Stevens et al., 2003), whereas the $k_{cat}$ values for guanine nucleotide exchange in Cdc42 are $5 \pm 1$ s$^{-1}$ and $19 \pm 3$ s$^{-1}$ for SopE_{78-240} and SopE2_{69-240} respectively (Friebel et al., 2001).
Figure 5.8. Molecular surface representation of the structure of the SopE$_{78-240}$-Cdc42 complex (Buchwald et al., 2002). The molecules have been separated at the interface of the complex with the intention to demonstrate their shape complementarity. The position of the interface is denoted by the dotted line areas and is shaded in grey on both molecules, with the switch region I contact area shaded in yellow. A ridge on Cdc42 (right) formed by switch region I residues 35 to 41 fits into a groove on SopE$_{78-240}$ (left) with residues Val36 and Asp38 of Cdc42 wedging into the gap between the two three-helix bundles of SopE$_{78-240}$. Surfaces are coloured according to electro-potential: blue represents a positive charge and red a negative charge. Molecular surface images were produced using GRASP (Nicholls et al., 1991).

It might then be asked whether BopE catalytic domain exists in equilibrium in solution between closed and open forms, or does it undergo a conformational change upon interaction with target protein? These two possibilities are not
necessarily mutually exclusive – there may be an equilibrium in solution but one that lies strongly towards the closed conformation. It is noted here that only one set of peaks was observed in \(^1\)H-\(^{15}\)N HSQC spectra recorded over a period of months, and there is no obvious exchange broadening of the peaks, indicating that any exchange between closed and open conformations is in the fast exchange regime. Finally, given the sequence and conformational differences between BopE and SopE/E2 catalytic domains, there remains the possibility that there are as yet unknown differences in specificity among the members of this family of bacterial guanine nucleotide exchange factors, with the potential for modulation of the activities of small G-proteins other than Cdc42 and Rac1.

5.8 Conclusions
The molecular mechanisms of \textit{B. pseudomallei} pathogenesis are not well understood. A number of putative type III-secreted effector proteins have been identified recently by analysis of the \textit{B. pseudomallei} genome sequence (Stevens et al., 2002). One of these proteins, BopE, is a homologue of the potent guanine nucleotide exchange factors SopE (Wood et al., 1996) and SopE2 (Bakshi et al., 2000) from \textit{Salmonella enterica} (Figure 5.4). SopE and SopE2 catalyse nucleotide exchange in mammalian Rho GTPases, contributing to disruption of the host cell membrane and invasion of the host cell (Hardt et al., 1998; Friebel et al., 2000 and 2001). BopE likewise acts as a GEF for the Rho GTPases Cdc42 and Rac1 \textit{in vitro} and may play a role in the invasion of non-phagocytic epithelial cells (Stevens et al., 2003). This work shows that BopE and SopE/SopE2 catalytic domains adopt similar three-dimensional folds comprising two three-helix bundles but also shows that BopE has a more compact conformation, involving significant inter-bundle interaction, than its \textit{Salmonella} counterparts. It is worth noting, however, that SopE residues involved in contacting Cdc42 in the SopE\textsubscript{78-240}-Cdc42 complex crystal structure (Buchwald et al., 2002) are largely conserved or conservatively substituted in BopE (Figure 5.4). SopE residues (Asp103, Gln109, Asp124 and Gly168) shown by mutation to be functionally important
(Schlumberger et al., 2003) are, moreover, conserved in BopE. It seems likely, therefore, that despite its more closed conformation, BopE ultimately utilises the same mechanism as SopE in catalyzing guanine nucleotide exchange in Rho GTPases. The possibility of large scale change in BopE from closed to open conformations prior to or upon binding Rho GTPase targets and a functional role for the C-terminal extension, including α6, will require further investigation.
PART THREE

Mathematical Biology
6.1 Abstract

The human genome comprises 3000 Mb of DNA, more than 98% of which does not encode proteins (Collins et al., 2004). Cross-species vertebrate genome comparison has lead to the identification of a considerable number of conserved non-genic sequences (CNGs) (also referred to as conserved non-coding sequences (CNSs or CNCs)) that are not homologous to any other sequences or to each other and are not repetitive (Margulies et al., 2003; Bejerano et al., 2004 and Dermitzakis et al., 2005). Human-mouse genome comparison indicates that there are approximately 327,000 CNGs (with at least 100 bp of ungapped alignment with at least 70% identical nucleotides that do not correspond to known coding genes or non-coding RNAs) (Dermitzakis et al., 2005). Four hundred and eighty-one segments longer than 200 bp are absolutely conserved between orthologous regions of the human, rat and mouse genomes, suggesting that these ultra-conserved elements in many species are more than 300 million years old (Bejerano et al., 2004). CNG function is generally poorly characterised. Using Ramsey Theory (Graham et al., 1980) and assuming neutral substitution rates, I estimate that a region of only several hundred bases in length is required to give rise to a human-mouse CNG of at least 200 bp. Hence it is likely that many CNGs have arisen by chance and do not necessarily have a function. This result does not rule out the possibility that some CNGs have a function but does help to explain why it has been difficult to assign a function to many CNGs and why, for example, deletion of gene deserts containing a total of 1243 CNGs in mice resulted in no detectable phenotypic changes and only small changes in the expression of genes that are adjacent to the deletions (Nobrega et al., 2004). The existence of many CNGs may be explained according to Ramsey Theory whereby vertebrate genomes are large enough
to contain patterns such as CNGs just as patterns of meaningful words have been found in large texts such as the Bible (Drosnin et al., 1997).

6.2 Statistical evidence
Cross-species genome comparison is a fundamental method for identifying biologically essential elements. Using this method, a considerable number of conserved non-genic sequences (CNGs) have been identified in vertebrates (Margulies et al., 2003; Bejerano et al., 2004 and Dermitzakis et al., 2005; Table 6.1) and also insects (Glazov et al., 2005). Some CNGs are even more conserved than coding sequences (Bejerano et al., 2004; Dermitzakis et al., 2003), which might indicate that at least some CNGs are functionally important. Experimental attempts to test the hypothesis that CNGs are previously unidentified genes showed that this is unlikely (Dermitzakis et al., 2002; Mural et al., 2002; Frazer et al., 2001 and Waterston et al., 2002). Several other potential CNG functions have been suggested, including a role as regulatory regions (Duret et al., 1993; Hardison et al., 1997 and Hardison, 2000) and matrix attachment regions (Glazko et al., 2003). There is evidence, moreover, that CNGs are important in vertebrate development (Woolfe et al., 2005). Deletion of 1817 kb and 983 kb gene deserts (gene-poor regions greater than 500 kb) containing a total of 1243 CNGs in mice, however, resulted in no detectable phenotypic changes and only small changes in the expression of genes that are adjacent to the deletions (Nobrega et al., 2004).
<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Number of Conserved exons</th>
<th>Average conserved exon length</th>
<th>Number of CNGs</th>
<th>Average CNG length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20713</td>
<td>292.22</td>
<td>27890</td>
<td>151.45</td>
</tr>
<tr>
<td>2</td>
<td>18029</td>
<td>264.50</td>
<td>32427</td>
<td>160.67</td>
</tr>
<tr>
<td>3</td>
<td>13241</td>
<td>281.05</td>
<td>26257</td>
<td>154.33</td>
</tr>
<tr>
<td>4</td>
<td>9090</td>
<td>289.36</td>
<td>22184</td>
<td>154.62</td>
</tr>
<tr>
<td>5</td>
<td>10675</td>
<td>297.70</td>
<td>24521</td>
<td>155.13</td>
</tr>
<tr>
<td>6</td>
<td>10519</td>
<td>303.66</td>
<td>19813</td>
<td>151.25</td>
</tr>
<tr>
<td>7</td>
<td>10616</td>
<td>284.14</td>
<td>18895</td>
<td>153.64</td>
</tr>
<tr>
<td>8</td>
<td>7742</td>
<td>297.22</td>
<td>16300</td>
<td>153.46</td>
</tr>
<tr>
<td>9</td>
<td>8893</td>
<td>293.69</td>
<td>14679</td>
<td>153.78</td>
</tr>
<tr>
<td>10</td>
<td>9456</td>
<td>280.01</td>
<td>16123</td>
<td>153.08</td>
</tr>
<tr>
<td>11</td>
<td>12086</td>
<td>298.70</td>
<td>16471</td>
<td>151.84</td>
</tr>
<tr>
<td>12</td>
<td>10890</td>
<td>282.46</td>
<td>13541</td>
<td>148.81</td>
</tr>
<tr>
<td>13</td>
<td>4466</td>
<td>295.70</td>
<td>11696</td>
<td>157.87</td>
</tr>
<tr>
<td>14</td>
<td>7015</td>
<td>301.80</td>
<td>11859</td>
<td>153.48</td>
</tr>
<tr>
<td>15</td>
<td>8261</td>
<td>282.79</td>
<td>10620</td>
<td>153.66</td>
</tr>
<tr>
<td>16</td>
<td>9032</td>
<td>283.00</td>
<td>9670</td>
<td>159.58</td>
</tr>
<tr>
<td>17</td>
<td>11557</td>
<td>285.36</td>
<td>9127</td>
<td>147.63</td>
</tr>
<tr>
<td>18</td>
<td>3921</td>
<td>274.32</td>
<td>9331</td>
<td>156.12</td>
</tr>
<tr>
<td>19</td>
<td>8540</td>
<td>334.70</td>
<td>2264</td>
<td>156.11</td>
</tr>
<tr>
<td>20</td>
<td>4904</td>
<td>314.30</td>
<td>7084</td>
<td>151.13</td>
</tr>
<tr>
<td>21</td>
<td>1837</td>
<td>326.60</td>
<td>2978</td>
<td>150.02</td>
</tr>
<tr>
<td>22</td>
<td>3894</td>
<td>331.30</td>
<td>2159</td>
<td>142.65</td>
</tr>
<tr>
<td>X</td>
<td>8244</td>
<td>298.63</td>
<td>23295</td>
<td>153.20</td>
</tr>
<tr>
<td>Y</td>
<td>602</td>
<td>284.00</td>
<td>993</td>
<td>220.50</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td><strong>294.88</strong></td>
<td></td>
<td><strong>156.00</strong></td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td><strong>295.36</strong></td>
<td></td>
<td><strong>153.20</strong></td>
</tr>
</tbody>
</table>

**Table 6.1. Numbers of conserved exons and CNGs for each chromosome (Dermitzakis et al., 2005).**

I hypothesise that most CNGs have never been under selective pressure and that substitution rate is the only dynamic force on them. This hypothesis is supported by the results presented in Figures 6.1 and 6.2 (using data in Table 6.1): average CNG length is essentially constant across all human chromosomes except the Y chromosome whereas average conserved exon length is much more variable (Figure 6.1); and average CNG density (chromosome length divided by number of CNGs) varies little across chromosomes 1-18 and 20 (the range of values obtained is 7382 to 9663, where 7382 indicates that on average there is a CNG every 7382 bases), whereas the density of conserved exons is more variable (range 7471 to 111...
25308) across chromosomes 1-18 and 20 (Figure 6.2). As an interesting aside, the results of these simple calculations show that most of the smaller chromosomes, including the Y chromosome, tend to have distinct properties (Figures 6.1 and 6.2). The numbers of conserved exons and CNGs on the Y chromosome, for example, are 602 and 993 (the other chromosomes have between 1837 and 20713 exons each and between 2159 and 32427 CNGs each), so there may be a bias in the data for the Y chromosome due to its small size. It is possible that the numbers of conserved exons and CNGs on the Y chromosome are low at least partly because there is only one copy of the Y chromosome.

Figure 6.1. Comparison of the average lengths of conserved exons and CNGs for each human chromosome.
Figure 6.2. Comparison of the densities of conserved exons and CNGs for each human chromosome.
The densities were obtained by dividing the number of bases in a particular chromosome by the number of exons/CNGs on that chromosome.

Superficial examination of the data in Figures 6.1 and 6.2 indicates that exon evolution has been somewhat chromosome-dependent whereas CNGs have evolved largely independently of which chromosome they are on. Therefore I consider further whether CNGs and exons have been shaped by the same dynamic force. In order to try to answer this question, I checked the correlation between average exon length and average CNG length. (The data used are the same as those used for Figure 6.1).

The Pearson correlation coefficient is written:

\[ r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \]

where \( \bar{x} \) and \( \bar{y} \) are the sample means of \( x_i \) (the average length of conserved exons) and \( y_i \) (the average length of conserved CNGs), \( s_x \) and \( s_y \) are the
sample standard deviations of $x_i$ and $y_i$ and the sum is from $i = 1$ to $n$ ($n=24$ with chromosome Y and $n=23$ without chromosome Y).

Using this formula and data from Table 6.1, the correlation coefficient is $-0.13$ without the Y chromosome and $-0.25$ with the Y chromosome included. The average lengths of conserved exons and CNGs for each human chromosome therefore have a negative and low correlation.

These statistical analyses suggest that even if CNGs have been shaped by selection, then selection pressure has been exerted in a different way on CNGs than on exons. Logically, however, it is more likely that most CNGs have not been shaped by selection. Therefore I have built a simple mathematical model using Ramsey Theory to test whether long-term conservation of numerous stretches of non-coding DNA could have occurred under no selective pressure.

6.3 Ramsey Theory model

Humans are predisposed to finding patterns but some patterns can arise by chance from randomness. Ramsey Theory is used to explain those patterns that arise from randomness; it concerns the conditions under which patterns must appear and, indeed, implies that complete disorder is impossible. Problems in Ramsey Theory typically ask a question of the form: how many elements of some structure must there be to guarantee that a particular property will hold? Perhaps the most famous case explained by Ramsey Theory is The Bible Code (Figure 6.3). It has been claimed that the Hebrew text of Genesis contains intentional coincidences of words or phrases that appear as letters with equal spacing (Drosnin, 1997 and 2003). However McKay et al. subsequently disproved this claim by proving that the same code can be found in the first 78,064 letters (the same length as Genesis) of the Hebrew translation of War and Peace (Figure 6.4; Bar-Hillel et al., 1998; McKay et al., 1999). There are about 3 billion letters in the 'life book', the human genome. Is this sufficiently large to find patterns such as CNGs in the same way that patterns can be found in the Bible or any other sufficiently
I use Ramsey Theory and some previous results concerning substitution rates to indicate that yes, it is likely that many CNGs have arisen by chance.

![Figure 6.3](image)

Figure 6.3. One of the examples mentioned in Drosnin's book, The Bible Code.

Some studies (Smith et al., 2002; Hardison et al., 2003) suggest that substitution rates vary deterministically across the non-coding non-repetitive
regions of the human genome. Properties of substitution rates from the non-coding part of the human genome have been identified using human-chimpanzee pair-wise comparisons over 4.7 Mb of sequence. These properties are (1) regional variation, with significant differences between alignments of tens of kilobases on the same chromosome; (2) local similarity, with adjacent one kb blocks within alignments tending to have similar substitution rates; (3) repeatable variation, with the substitution rate variation down the human and chimpanzee lineages being positively correlated; (4) no evidence to indicate that the observed substitution rate variation is owing to selection. I think that regional variation gives rise to characteristics such as CNGs, with regions with slower substitution rates being the source of most CNGs.

The probability of finding even one ultra-conserved element, let alone 481 or even 327,000, has been calculated as vanishingly small (less than $10^{-22}$) (Bejerano et al., 2004); a mean of the Bernoulli variable for substitution rate from a 1 Mb window was used to determine this probability. I argue that the local substitution rates in this analysis were overestimated because an average value was used, ignoring regional variations, and that the implications of Ramsey Theory were ignored. Assuming the four results of Smith and co-workers (Smith et al., 2002), I can show that most of the 327,000 CNGs actually have a high probability of occurring. Employing Ramsey Theory and assuming neutral substitution rates, I calculate that a region of a 300 million year old ancestor sequence of only 573 bases in length is required for humans and mice to retain a CNG of at least 200 bp with 100% identity and 95% identity to chicken (see Section 6.4). This analysis indicates, therefore, that many CNGs have remained by chance and are non-functional. Identification of functional CNGs may be facilitated by comparison of genomes from evolutionarily distant species such as human and fugu (Woolfe et al., 2005; Goode et al., 2005), since the numbers of CNGs shared between species decreases with evolutionary distance.
6.4 Assessment of likelihood that CNGs have arisen by chance

Suppose a genomic region with substitution rate $r$ and length $N$ (in nucleotides) existed 300,000,000 years ago (the approximate time of a common ancestor of birds and mammals - Figure 6.5). From the time that the human and mouse lineages split (80,000,000 years ago), what is the value of the length $N$ required, according to Ramsey theory, in order that human and mouse inherit an identical 200 bp segment (CNG) and that chicken inherits this segment with 95% identity to human and mouse?

![Diagram](image)

**Figure 6.5.** Illustration of the hypothetical situation used to assess the likelihood that a CNG can arise by chance. A 300 million year (myr) old genomic region is represented by the long yellow line at the top of the diagram. The human and mouse lineages split approximately 80,000,000 years ago and the avian and mammalian lineages split approximately 300,000,000 years ago. A 200 bp segment that has been inherited by human and mouse with 100% identity and by chicken with 95% identity to human and mouse is represented by the short yellow line towards the bottom of the diagram.
Firstly, the substitution rate, defined as the number of substitutions per site per year, is calculated as follows:

\[ K = 2Tr, \]

where \( K \) is the number of substitutions between two homologous sequences (species), \( T \) is the time of divergence between the two sequences (species), and \( r \) is defined as the number of substitutions per site per year (Patthy, 1999). The substitution rate is therefore given by

\[ 2 \times 3 \times 10^8 \times r \times N = 0.05 \times N \]

where \( 3 \times 10^8 \) is the number of years and 0.05 is the total change due to substitutions (for retention of 95% identity with chicken).

This gives a substitution rate in substitutions per site per year of

\[ r = 8.3333 \times 10^{-11} \]

With this substitution rate estimate, the chance of a nucleotide change since the time when the co-ancestor of the human and mouse lineages lived (80,000,000 years ago) is:

\[ 8.3333 \times 10^{-11} \text{ substitutions per site per year} \times 8 \times 10^7 \text{ years} = 6.6666 \times 10^9. \]

The probability of a nucleotide base remaining the same is then \( (1 - 6.6666 \times 10^{-3}) = 0.9933 \), and therefore the probability of a region of 200 nucleotides remaining absolutely conserved is:

\[ (0.9933)^{200} \times (0.9933)^{200} = 0.0679. \]
As the size of the genomic region \( (N) \) increases, the probability of finding an ultra conserved region of 200 nucleotides increases. If I pick the first base to 200\(^{th} \) base in this genomic region, the second base to 201\(^{st} \) base and so on, then there are \((N - 200 + 1)\) events in total (Figure 6.6). Hence it will be expected there are \((0.9933)^{400} \times (N - 200 + 1)=0.0679 \times (N - 200 + 1)\) CNGs in region with length \( N \).

![Figure 6.6](image)

**Figure 6.6** The conserved genomic region (indicated in yellow) may be selected from the first base to 200\(^{th} \) base, the second base to 201\(^{st} \) base, and so on.

We find that for a region with \( N \geq 214 \) nucleotides in the 3\( \times \)10\(^8 \) year old ancestor sequence, the number of finding a segment of 200 nucleotides that has remained 100\% conserved between human and mouse and 95\% conserved with chicken is greater than 1:

\[
0.0679 \times (N - 200 + 1) > 1
\]

Mathematically, therefore, CNGs must occur. This result is perhaps counter-intuitive, but human intuition focuses on a specific outcome, the 200 bp conserved segment, ignoring the fact that we are calculating the probability of this outcome from a huge sample space.

I note that the substitution rate for CNG sequences estimated above (8.3333\( \times \)10\(^{-11} \) substitutions per site per year) is 25-44 times slower than a recent estimate of the average rate of substitution in mammalian DNA of 2.1-3.7 \( \times \)10\(^{-9} \) substitutions per year (Hardison et al., 2003). Bejerano *et al.* (2004)
calculate that CNG elements are changing at a rate that is roughly 20 times slower than the average for the genome. Taking these figures into account, we estimate that an upper value of the substitution rate in CNGs might be something like

\[ r = 8.3333 \times 10^{-11} \times \frac{44}{20} = 1.8333 \times 10^{-10} \]

With this substitution rate, the chance of a substitution is:

\[ 1.8333 \times 10^{-10} \times 8 \times 10^7 = 1.4666 \times 10^{-2} \]

The probability of a nucleotide base remaining the same is 0.9853 and, therefore, the probability of 200 bp that are absolutely conserved between human and mouse is

\[ (0.9853)^{200} \times (0.9853)^{200} = 2.6754 \times 10^{-3}. \]

From the relationship \( 2.6754 \times 10^{-3} \times (N - 200 + 1) \geq 1 \), we find that even with the faster substitution rate, a region of the ancestor sequence with length greater than or equal to only 573 bp is needed for occurrence of a 200 bp conserved element with 100% identity between human and mouse and 95% identity between human/mouse and chicken.
PART THREE
Chapter Seven
Evolution Is Unpredictable

7.1 Abstract
Many arguments on whether evolution is predictable have been posed since Darwin proposed the concept of evolution by natural selection. This big question was again posed at Kavli’s theoretical physics conference (Collins, 2004). Among 25 forward-looking questions (Section 7.5) suggested by the meeting’s attendees, this nagging question was more precisely rephrased: can the theory of evolution be quantitative and predictive? Here I demonstrate that the evolution of genomes cannot be properly predicted no matter how genomes are quantified. If it is assumed that the prediction of evolution of other systems must be traced back to the molecular level, then their accurate prediction will be impossible, too.

7.2 The limitation of methodology
From daily experiences, humans are predisposed to thinking that a particular system cannot be predicted accurately because the exact initial conditions and some factors that affect the system are not yet known. However some systems do not fall into this category. The impossibility of predicting these systems is not because of the limitation of knowledge but the limitation of methodology: practicable methods for accurate prediction do not exist.

There are some examples wherein accurate predictions are impossible even if all initial conditions and factors that affect events are known.

The Uncertainty Principle (Heisenberg, 1929; Hey et al., 1987) states that one cannot assign exact simultaneous values to the position and momentum of a quantum mechanical system. Varied philosophic controversies have been
continuing since then. In the spectacular debate on quantum mechanics between Bohr and Einstein, actually the primary argument was whether there existed methods that could be used to accurately measure the position and the momentum of a particle simultaneously.

Similarly, some complex systems that are sensitive to initial conditions such as the weather or the stock market are also difficult to predict over any useful time range. This is attributed to the butterfly effect and the finite model usage. The butterfly effect states that small variations in the initial conditions of a dynamical system produce large variations in the long-term behaviour of the system. When simulating the weather, for example, only finite models such as manual or computer models can be used to analyse data. This limitation of finite model usage means that one or more tiny factors must be omitted. Owing to the butterfly effect, weather predictions are useless after a certain amount of time (Lorenz, 1963).

7.3 The restrictions of biology
Can the theory of evolution be quantitative and predictive? It is obvious that this question is not clear because there may exist different solutions when varied assumptions are given. As to genomic evolution, at least people would expect to know whether a new evolving sequence is functional, and then what exact functions the new evolving sequence has. In addition, as the philosophical arguments used in uncertainty principle and butterfly effect (see Section 7.2), if one wants to predict evolution of some specific systems, the tools and methods used to measure or predict evolution must be stressed. That is, the method for prediction should be addressed.

Therefore, what restrictions may a possible method have when trying to detect the function of DNA sequences? Biologists have known for decades that heritable changes in gene expression can occur without accompanying changes in DNA sequence. Position effect variegation (PEV) in Drosophila was first described by Hermann Muller and others in the 1930s (Figure 7.1). In their experiments, for example, placing the white eye gene close to
heterochromatin, a tightly packaged and transcriptionally silent chromosomal region, would often give rise to flies with patches of red and white eye color (Dillon, 2003 and Wakimoto, 1998).

Figure 7.1. Eye-opener: studies of Drosophila showed that a gene's position can affect its function (Dillon, 2003).

Moreover, genes are also affected by long-range repressors and enhancers, which could reside megabasepairs away (Nobrega et al., 2003 and Cai et al., 1996). These studies all suggest that a gene's exact function is affected by other DNA sequences and its relative position in the genome.

Next, what methods could be used to detect functions of a DNA sequence? All possible methods can only be divided into two categories logically (Snyder et al., 2003):

1. Direct method, analysing the sequence features: One may find a sequence's functions by checking or comparing in databases for similar sequence features.
2. Indirect methods, analyzing the sequence's products: Via over-activating or inactivating a given sequence to create 'abnormal' life, the function of the
sequence may be detected by comparing the abnormal life form to the original one.

However, methods based on these two categoric methods cannot be used to predict a sequence's exact function. The proof is as follows.

7.4 The evolution of genomes is unpredictable

Assume that there is a new genome and an arbitrary sequence is chosen from this genome (Figures 7.2a, 7.2b). People would like to know whether the sequence is functional, and if it is, what functions the sequence has. According to the methodology mentioned above, one may check the sequence features. However its exact function is affected by genomic configuration. For finding the exact function, other sequence features should be checked (Figures 7.2c, 7.2d). Again and again, the procedure falls into an infinite self-recurrence. Alternatively, one can analyse the sequence's products. However the possible functions from the 'abnormal' form cannot express the exact function of the original sequence because the exact functions are still relative to other DNA sequences in the original genome. Therefore the procedure still falls into an infinite self-recurrence.
Figure 7.2. The exact function of a DNA sequence is affected by other DNA sequences. a) A new genome. b) An arbitrary sequence (red) is chosen from this genome. c) The (red) sequence's exact function is affected by the genomic configuration. Hence other sequences' function (for example those shown in yellow) should be checked. d) Again, in order to establish the exact function(s) of the yellow sequences, the purple sequence should be checked, and so on.

These logical arguments not only fit the evolutionary 'future life', but also fit the 'existing life'. Current methodologies are unable to provide a DNA sequence the exact measurement in a genome in the same way that the Uncertainty Principle applies to particles are in the quantum world. Uncertainty could be the nature of life. Therefore the best mathematical language, perhaps, for describing a specific gene is a probability distribution. This may partially answer another question presented at the Kavli conference: to understand biology, is new mathematics required?
7.5 Appendix: Kavli’s theoretical physics conference

More than 150 leading physicists gathered at the University of California at Santa Barbara to commemorate the 25th anniversary of the school’s Kavli Institute for Theoretical Physics in October 2004. They looked back at the past quarter of a century of physics and ahead to what will happen in the next. Kavli’s theoretical physics conference covered many other areas, such as nanophysics, astrophysics (particularly gravitational waves), and quantum computing and gravity. It also served as an impromptu celebration of the Nobel Prize awards, one of the 2004 winners being the Institute’s director, David Gross. In wrapping things up, Gross presented 25 forward-looking questions that ran the gamut of topics, culled from suggestions by the meeting’s attendees. Several concerned the application of physics and mathematics to life sciences (Collins, 2004):

1. How can one tell the shape of an organism by looking at its genome?
2. Can the theory of evolution be quantitative and predictive?
3. To understand biology, is new mathematics required, the way that string theory requires new mathematics?
References


