THE SPATIAL EVOLUTION OF THE CHEMOTAXIS PROTEINS OF THE *BACILLUS SUBTILIS* GROUP

A thesis submitted in fulfilment of the requirements for the degree of
MASTERS OF SCIENCE

in

BIOCHEMISTRY

of

RHODES UNIVERSITY

by

Anna Elizabeth Johanna Yssel

January 2011

ABSTRACT

The aim of this work was to study spatial evolution of the chemotaxis proteins of a group of plant-associated soil-dwelling bacteria vernacularly referred to as the *B. subtilis* group. This was achieved by creating homology models for the chemotaxis proteins if a suitable template was available, and by analysing the selective forces (positive, purifying or neutral) acting upon the chemotaxis proteins. Chemotaxis is the phenomenon in which bacteria direct their movement towards more favourable conditions, and is critical for processes such as obtaining nutrients, escaping toxic compounds, host colonization and bio-film formation. Members of the *B. subtilis* group exhibit different preferences for certain host plants, and it is therefore feasible that their chemotactic machinery are fine-tuned to respond optimally to the conditions of the various niches that the strains inhabit. Homology models were inferred for the plant growth promoting *B. amyloliquefaciens* FZB42 proteins CheB, CheC, CheD, CheR, CheW and CheY. The interactions between: CheC-CheD, the P1 and P2 domains of CheA with CheY and CheB, and the P4 and P5 domains of CheA with CheW were also modelled. The hydrophobic interactions contributing to intra- and inter-protein contacts were analysed. The models of the interactions between CheB and the various domains of CheA are of particular interest, because to date no structures have been solved that show an interaction between a histidine kinase (such as CheA) and a multidomain response regulator (such as CheB). Furthermore, evidence that phospho-CheB may inhibit the formation of phospho-CheY by competitively binding to the P2 domain of CheA is also presented. Proteins were analysed to determine if individual amino acid sites are under positive, neutral or purifying selection. The Methyl Accepting Chemotaxis Proteins (MCPs), CheA and CheV were also analyzed, but due to a lack of suitable templates, no homology models were constructed. Site-specific positive and purifying selection were estimated by comparing the ratios of non-synonymous to synonymous substitutions at each site in the sequences for the chemotaxis proteins as well as for the receptors McpA, McpB, and McpC. Homology models were coloured according to intensity of selective forces. It was found that the chemotaxis proteins of member of the *B. subtilis* group are under strong evolutionary constraints,

hence it is unlikely that positive selection in these proteins are responsible for the differences in habitat preference that these organism exhibit.

DECLARATION

I, Anna Elizabeth Johanna Yssel, declare that the work presented here is my own and that it has not been submitted for publication anywhere else.

TABLE OF CONTENTS

AKNOWLEDGEMENTS

First of all I want to express my gratitude to my parents for making my education a priority and supporting me emotionally and financially. Furthermore, I want to acknowledge my brothers and friends for their words of encouragement and Ruan for his love, patience and support. I am eternally grateful to my supervisor, Dr. Özlem Taştan Bishop, and my co-supervisor, Prof. Oleg Reva, for giving me an opportunity to work on such an exciting project, for teaching me valuable skills and for providing me with constructive criticism and much needed guidance. I want to thank Prof. Fourie Joubert from the University of Pretoria, for kindly accommodating me at the Bioinformatics and Computational Biology Unit, and the National Bioinformatics Network, the National Research Foundation and Rhodes University for providing me with funding. Prof. R. Borris (Humboldt University, Berlin) graciously provided us with the necessary information for the recently sequenced strains of *B. amyloliquefaciens*. And last but not least, I am grateful to my friend Eugene for his meticulous inspection of this work.

LIST OF FIGURES AND ILLUSTRATIONS

LIST OF TABLES

ACRONYMS

Å	Ångstrom
3D	Three-dimensional
AIC	Akaike information criterion
ATP	Adenosine Triphosphate
BLAST	Basic Local Alignment Search Tool
CASP	Critical Assessment of Techniques for Protein Structure Prediction
CDD	Conserved Domain Database
CHARMM	Chemistry at Harvard Macromolecular Mechanics
COGs	Clusters of Orthologous groups
CCW	Counter Clockwise
CW	Clockwise
DDBJ	DNA Data Bank of Japan
DNA	Deoxyribonucleic acid
EMBL	European Molecular Biology Laboratory
GDT_TS	Global Distance Test Total Score
HAMP	Histidine kinase, Adenyl cyclase, Methyl-accepting chemotaxis protein and Phosphatase
HK	Histidine Kinase

HMM	Hidden Markov Model
MEC	Mechanistic Empirical Combined Model
MCP	Methyl accepting Chemotaxis Protein
ML	Maximum Likelihood
MSA	Multiple Sequence Alignment
NCBI	National Centre for Biotechnology Information
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
Pfam	Protein Families Database
PSI-BLAST	Position Specific Iterative-BLAST
PSIPRED	Protein Structure Prediction Server
REC	Receiver domain
RMSD	Root Mean Square Deviation
RR	Response Regulator
SAH	*S*-adenosyl-L-homocysteine
SCOP	Structural Classification of Proteins
SMART	Simple Modular Architecture Research Tool
TM	Transmembrane Helix

TYPOGRAPHICAL CONVENTIONS

- Amino acids will be referred to by their three-letter abbreviations.
- Species will be referred to either by their Latin name or by the following three or four letter abbreviation:
    - R032 — *B. pumilus* SAFR032
    - 4580 — *B. licheniformis* ATCC4580
    - s168 — *B. subtilis* ssp. *subtilis* 168
    - H642 — *B. subtilis* ssp. *subtilis* JH642
    - 3610 — *B. subtilis* ssp. *subtilis* NCBI3610
    - SMY — *B. subtilis* ssp. *subtilis* SMY
    - 6633 — *B. subtilis* ssp. *spizizenii* ATCC6633
    - DSM7 — *B. amyloliquefaciens* ssp. *amyloliquefaciens* DSM7
    - ZB42 — *B. amyloliquefaciens* ssp. *plantarum* FZB42
    - B946 — *B. amyloliquefaciens* ssp. *plantarum* B946
    - 01Y2 — *B. amyloliquefaciens* ssp. *plantarum* B9601Y2
    - aoB3 — *B. amyloliquefaciens* ssp. *plantarum* GaoB3

CHAPTER 1

1. INTRODUCTION

1.1. Two component phosphorelay systems and chemotaxis in bacteria

In order to survive environments that are in a constant state of flux, bacteria must be able to respond rapidly to a myriad of changing environmental factors and to adapt to conditions such as nutrient deprivation, desiccation and presence of toxins, amongst other things. Signal transduction systems provide a link between cues (both external and internal) and the appropriate cellular responses in all forms of life (Wuichet *et al.* 2007). Prokaryotic signal transduction systems can be grouped into three major families based on domain organization and complexity: one-component systems, classical two-component systems anchored by class I histidine kinases, and multi-component systems anchored by class II histidine kinases, such as the chemotaxis system.

For the sake of simplicity, the basic characteristics of the classical two-component systems and the multi-component systems will be described with the collective term "two component phosphorelay systems" in the following text.

Two component phosphorelay systems use the transfer of phosphoryl groups to control gene transcription or protein activity in response to specific environmental cues. Such systems consist of two major components: a sensor kinase and a response regulator. The systems that control sporulation, osmoregulation, nitrogen assimilation and chemotaxis are good examples of two component phosphorelay systems that can be found in bacteria. In a typical pathway, a protein kinase will respond to a signal by trans-autophosphorylating in a dimer using the γ-phosphate of ATP (Aizawa *et al.* 2000). Thus far, studies indicate that a conserved histidine serves as the site of phosphorylation in all known protein kinases. The phosphorylated kinase transfers its phosphate group to conserved aspartate residue on the cognate response regulator protein. This regulator protein allows the cell to govern its activities according to its state of phosphorylation.

1.1.1. *Proteins and molecules involved in the chemotaxis pathway*

Chemotaxis refers to the ability of motile bacteria to sense changes in their chemical environment and to direct their movement towards more favourable conditions. Chemotaxis is controlled by a well-studied phosphorelay regulatory system which is centred on the class II histidine kinase CheA. Unlike class I histidine kinases, the class II chemotaxis system contains multiple proteins that separate input from output, and additionally, it includes several regulatory components that play a role in adaptation. Furthermore, the separation of the histidine kinase from the signal recognition proteins allows a single kinase to respond to and integrate signals from multiple stimulus receptor proteins (Aizawa *et al.* 2000). In the absence of an attractant/ repellent gradient the movement of a bacterium can be described as a series of randomized smooth runs with intermittent re-orientating tumbles. The chemotaxis system uses environmental cues to control the probability of switching between running and tumbling, thereby creating a bias that favours movement towards attractants and away from repellents. Because of their small size bacteria cannot measure attractant or repellent concentrations over the length of their cells; rather they sense concentration changes over time. Thus, chemotaxis relies on a temporal rather than a spatial sensing mechanism.

Almost all motile bacteria share the following core chemotaxis proteins: CheA, CheW, CheY, CheR and CheB. In addition to these core components, the genomes of many sequenced motile bacteria and archaea also contain proteins that catalyze CheY-p dephosphorylation (Kentner & Sourjik 2006). Chemotaxis absolutely requires the universal methyl donor *S*-adenosylmethionine (AdoMet) for methylation of membrane proteins responsible for detecting chemical cues from the environment. These methyl-accepting chemotaxis proteins (MCPs) may detect different stimuli through direct and indirect mechanisms (Kort *et al.* 1975, Stock & Baker 2009). When a chemo-effector is detected the MCP undergoes a conformation change which influences the activity of CheA (Hanlon & Ordal 1994). The methylation and demethylation of certain glutamate residues on the MCPs functions to control the adaptation of the chemotactic system to persistent stimuli (Toews *et al.* 1979, Krueger *et al.* 1992).

1.1.1.1. *MCPs*

The classical MCPs are transmembrane proteins with an amino-terminal transmembrane helix (TM1), periplasmic sensing domain, second transmembrane helix (TM2), and a large

cytoplasmic region. The cytoplasmic region contains the signalling domain (also known as the highly conserved domain) where CheA and CheW bind, and a methylation domain consisting of two helixes (MH1 and MH2) where CheR and CheB selectively methylate and demethylate conserved glutamate residues. Between the membrane and the methylation region a HAMP (histidine kinase, adenyl cyclase, methyl-accepting chemotaxis protein and phosphatase) linker is located, which propagates the signal of ligand binding to the rest of the cytoplasmic domain. Evidence suggests that MCPs form stable homodimers regardless of ligand binding. MCPS can be divided into three main classes based on the presence or absence of four original 14-residue insertion/deletion regions, found within both the signalling and methylation functional units of the cytoplasmic domain, as determined by Le Moual and Koshland (1996). It seems to be the case that the gene coding for the C-terminal cytoplasmic domain of these proteins evolved through gene duplication from a common ancestor in which the four insertion/deletion regions were deleted two by two.

The current concentration of a stimulant molecule is measured as percent receptor occupancy against its previous concentration represented by the level of adaptive methylation of the receptor (Manson *et al.* 1998). The lag time between the sub-second response to ligand binding and the slower adaptive methylation of the receptor gives the cell a rudimentary memory.

1.1.1.2. ***CheA***

The histidine kinase CheA has a complex multidomain structure (Figure 1.1) and is known to form a dimer with itself that transphosphorylates upon activation. Dimerization takes place at the P3 domain.

CheW interacts with CheA at the P5 domain to establish receptor mediated kinase regulation (Bourret *et al.* 1993, Miller *et al.* 2006). The phosphorylation reaction involves transferring of the γ-phosphate of ATP to a conserved histidine residue on the P1 domain (Shimizu *et al.* 2000, Park *et al.* 2004). A binding site for the ATP molecule is on the P4 domain, which contains several conserved regions that are essential for catalysis and for positioning ATP into the active site (Hirschman *et al.* 2001, Szurmant & Ordal 2004, Eaton & Stewart 2010); phospho-CheA donates its phosphoryl group directly to CheY which is bound to the P2 domain. CheA can also phosphorylate CheB, although no structure showing the interactions between these molecules has

been solved to date it is presumed that it takes place on the P2 interface (Park *et al.* 2004, Miller *et al.* 2006). The rate of phosphotransfer between CheA and CheY is a hundred to a thousand times faster than similar phosphotransfer reactions in other two-component systems. The P2 domain is absent in other two-component histidine kinases, therefore it is hypothesized that docking of CheY at P2 enhances the rate of the phosphotransfer (Stewart & Van Bruggen 2004).

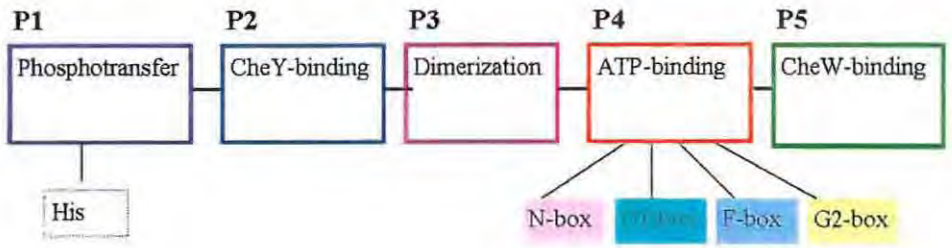


Figure 1.1. Scheme for a typical CheA. CheA forms a homo-dimer at the P3 domain. Conserved sequence motifs, N, G1, F and G2 are located in the ATP-binding domain, P4. CheA catalyzes ATP-dependent *trans*-autophosphorylation of a conserved His residue in the P1 domain. Binding of CheY to CheA is facilitated by the P2 domain which is absent in all other histidine kinases. CheW binds to the P5 domain to facilitate interaction with the MCPs.

1.1.1.3. *CheW*

CheW possesses no known regulatory or catalytic function, and acts solely as a scaffolding element (Falke *et al.* 1997). CheW interacts with the cytoplasmic domain of the MCPs as well as with CheA, thereby linking their activities (Liu & Parkinson 1989, Liu & Parkinson 1991, Gegner *et al.* 1992, Yao *et al.* 2007). CheW consists of only one domain which contains a conserved motif Asn-x-x-Gly-x-Ile-x-Pro (where x is any amino acid) which plays an important role in CheA binding (Yao *et al.* 2007). The CheW protein from *B. subtilis* has been sequenced and characterised (Hanlon *et al.* 1992), it was found that a mutation in CheW gives the bacterium a smooth swimming bias which is punctuated by brief tumbles. In contrast CheW mutants in *E.coli* are incessantly smooth swimming (Stewart *et al.* 1990). Furthermore, the consensus nucleotide binding motif present on *E.coli* CheW is absent in *B. subtilis* CheW (Robson 1984, Hanlon *et al.* 1992).

1.1.1.4. *CheY*

CheY is a single domain protein which is manifesting variety of the widespread receiver domain (REC), which is found throughout two-component response regulators (Wuichet *et al.* 2007). In most cases two-component response regulators contain, in addition to the REC domain, a second output domain. However in the chemosensory pathway CheY needs no output domain as CheY-p regulates the response by interacting directly with the flagellar motor (Falke *et al.* 1997). When CheY gets phosphorylated on its conserved aspartyl residue, it undergoes a conformational change that converts the protein from an inactive to an active state. The phosphorylation process requires presence of an Mg^{2+} ion that interacts with four conserved residues: two Asp residues, one Lys and one Thr (Lukat *et al.* 1991, Szurmant & Ordal 2004). The phosphorylation state of CheY is controlled by the opposing activities of CheA and by the rate of dephosphorylation (Cho *et al.* 2000). Dephosphorylation of CheY can either involve slow auto-dephosphorylation, or a rapid dephosphorylation by a phosphatase enzyme that functions to terminate the CheY-p signal to the flagellar motor (Guhaniyogi *et al.* 2006). The various response regulator phosphatase families required for phosphoryl group hydrolysis have different overall structures and distribution throughout archaea and bacteria, but employ similar catalytic strategies (Pazy *et al.* 2010). The concentration of CheY-p is practically the same in the adapted state as in the pre-stimulus state causing the rotational bias of the flagella to be the same under both conditions. This remarkable robust perfect adaptation to persistent stimulation is achieved trough reversible receptor methylation/demethylation (Matsuzaki *et al.* 2007).

1.1.1.5. *CheR*

The constitutively active methyl transferase CheR mediates adaptation to stimuli by the reversible methylation of glutamyl residues in the MCPs (Yi & Weis 2002). This methylation reaction involves transferring a methyl group from AdoMet to the γ-carboxyl groups of specific glutamyl residues, leading to the formation of γ-methylglutamate and *S*-adenosyl-L-homocysteine (SAH).

The elucidation of the structure of CheR revealed two domains: a small N-terminal domain which is linked to a larger α/β C-terminal domain, and a C-terminal domain with a typical nucleotide binding fold and a small antiparallel β-sheet sub domain. The AdoMet binding site is

formed by the β1/αA loop within the C-terminal domain with some residues from the N-terminal domain and linker region contributing to the interaction (Djordjevic & Stock 1997). The catalytic domains for most AdoMet dependent methyltransferases share the signature sequence (Gly/Ala)-x-(Gly/Ala/Ser)-x-Gly, where x is any amino acid (Shiomi *et al.* 2002). Studies revealed two types of interactions between MCPs and CheR (Perez *et al.* 2004). In Proteobacteria the binding site for CheR is distinct from the methylation sites and involves a small β-sub domain interacting with a NWETF pentapeptide motif on the C-terminus of the receptor (Wu *et al.* 1996, Shiomi *et al.* 2002, Yi & Weis 2002). In other classes of bacteria the NWETF motif is absent, and the role of the *β*-subdomain, outside of maintaining structural integrity, remains unclear (Perez & Stock 2007).

1.1.1.6. ***CheB***

Like CheY, CheB is known to bind to the P2 domain of CheA, although to date no structure showing this interaction has been solved. The N-terminal domain of CheB shares homology with CheY. It contains a conserved Asp site that becomes phosphorylated as well as four conserved residues that are involved in divalent cation binding. These are: two Asp residues, one Thr and one Lys (Falke *et al.* 1997). In addition to the N-terminal CheY-like REC domain, CheB contains an additional catalytic domain and a linker that joins these two domains together. The catalytic domain is responsible for methylesterase activity, which involves hydrolysis of receptor methylglutamates, formed by CheR, to release methanol and regenerate the glutamate side chains (Krueger *et al.* 1992). The methylesterase active site was identified to be a catalytic triad, consisting of Ser, His and Asp residues, which are located in a cleft formed by *cβ*1, *cβ*2 and *cβ*7 (Krueger *et al.* 1992, Djordjevic *et al.* 1998). In its unphosphorylated state CheB is inactive due to the fact that the N-terminal domain packs tightly against the C-terminal domain active site, making interaction with the methylation/demethylation sites on the receptor impossible. Upon phosphorylation a conformational change alters the juxtaposition of the regulatory and catalytic domains and CheB becomes enzymatically active (Djordjevic *et al.* 1998).

Once CheB is phosphorylated, it can catalyze its own dephosphorylation at a speed much greater than that of CheY-p. Unlike CheY-p, CheB-p does not require a phosphatase such as CheZ to hydrolyze its acyl-phosphate. It is hypothesized that the short lifetime of CheB-p is correlated

with the shorter distance that it has to travel to its targets – the MCP methylation/demethylation sites. In contrast, CheY-p has to diffuse to the more distant flagellar motor complex (Falke *et al.* 1997). CheB is not only a methyl esterase, but can also function as a deamidase, which hydrolyses certain receptor glutamines to yield ammonia and a bare glutamate side chain thereby creating sites for reversible methylation (West *et al.* 1995, Falke *et al.* 1997).

1.1.1.7. ***Spatial distribution of flagella and receptors***

In peritrichous flagellated bacteria like *E.coli and B. subtilis*, the flagella are evenly spread around the cell surface (Garrity & Ordal 1995). Unlike the flagella, the receptor complex consisting of the MCPs, CheA and CheW are located at either one or both poles of the cell (Endres *et al.* 2007). It is hypothesized that the polar localization of receptor clusters functions as a signal amplification mechanism (Shimizu & Bray 2002, Shimizu *et al.* 2003).

1.1.2. ***Chemotaxis in E. coli***

Changes in stimulus levels are sensed by the transmembrane acceptors which are connected to the histidine kinase CheA by a coupling protein CheW. When a repellent binds to the receptors (or when attractant is removed), the associated CheA kinase is activated to autophosphorylate. Phosphorylated CheA can donate the phosphor group to CheY or CheB. CheY-p diffuses to and interacts with the flagella complex to increase the likelihood of tumbly behaviour by causing the flagellum to rotate in a CW direction. This excitation signal is terminated when CheZ dephosphorylates CheY-p. It is important to note that smooth runs occur only when all four to six flagella rotate in a CCW direction, but tumbling is induced when at least one of the flagella rotate in a CW manner (Manson 2008). Adaptation to stimulus levels occurs at several conserved glutamate residues on the receptor cytoplasmic domain and involves a methylation/demethylation modification system catalyzed by CheR and CheB. The constitutively active CheR is responsible for the methylation reaction which involves AdoMet as methyl donor. Receptor methylation causes the activation of CheA kinase, while demethylation deactivates it. When CheB is phosphorylated by CheA it becomes active and hydrolyzes the methyl esters created by CheR on the receptors. This feedback loop induces a conformation change that resets CheA activity to pre-stimulus level. The methylation-demethylation process allows the bacterium to compare current stimulus concentration levels to those of the past three to four

seconds, and to respond to those changes. If the previous stimulant concentration was high, the levels of receptor methylation would be high; if it was low, receptor methylation would be low. The same principle applies for repellent concentration, but with reverse polarity (Falke *et al.* 1997). To summarize, the presence of a repellent induces the cell to tumble. After adaptation the cell is now able to receive a new signal as it moves off in a new direction. If an attractant molecule binds to the receptors (or if a repellent molecule is removed), the kinase activity of CheA is inhibited, leading to decreased CheY-p levels and an increased probability of a run (Rao *et al.* 2008). The chemotaxis system in enteric bacteria such as *E. coli* is streamlined compared to that of many other free living organisms such as *B. subtilis*. This may be due to the fact that the environments inhabited by enteric bacteria are much simpler than those of free living organisms (Manson *et al.* 1998).

1.1.3. ***Chemotaxis in B. subtilis***

Like *E. coli*, *B. subtilis* possesses a single chemotaxis operon. Furthermore, their respective pathways share five orthologous proteins: CheA, CheB, CheR, CheW and CheY. Despite the fact that the orthologs have apparent identical biochemistry, deletion studies in both organisms revealed that their individual contributions to overall function differ (Table 1.1). According to estimations, *E. coli* and *B. subtilis* diverged over 1 billion years ago (Kunst & Ogasawara 1997, Rao *et al.* 2004). Unlike *E. coli*, the genome of *B. subtilis* lacks the CheZ protein but contains additional proteins CheC, CheD and CheV (Rao *et al.* 2004). The differences between the *E.coli* and *B. subtilis* chemotaxis systems are not limited to the cytoplasmic chemotactic components, but also include the receptors and the adaptation mechanisms employed by each organism. The receptors in *E. coli* are class I receptors (Zimmer *et al.* 2000) which are methylated when attractant is added (or when repellent is removed) and demethylated when attractant is removed (or repellent is added) (Toews *et al.* 1979). McpB in *B. subtilis* is a class III receptor which release methanol in response to all stimuli (Le Moual & Koshland 1996, Zimmer *et al.* 2000). In addition to the differences in the methylation dependent adaptation processes in *B. subtilis* and *E. coli*, evidence suggests that *B. subtilis* possess additional methylation independent adaptation mechanisms which involve CheC, CheD and CheV (Rao *et al.* 2008). The flagellar motor assembly in *B. subtilis* is also different to that from *E.coli* and involves a FliY/FliN fusion protein that is not present in *E.coli* (Garrity & Ordal 1995). And the most striking difference is

an inversion of the signal conduit through the chemotaxis systems of *E. coli* and *B. subtilis*. In *E. coli* binding of an attractant to MCP inhibits phosphorylation of CheY by CheA that causes CCW rotation of flagella and a smooth run towards the higher attractant concentration, and tumbling is associated with an activated state of CheA caused by attractant removal. Contrary, *B. subtilis* tumbles by default and the CCW rotation of flagella is activated by CheY-p, which is phosphorylated by CheA, that in its turn is activated by attractant molecules bound to the MCP (Bischoff *et al.* 1993, Zhang & Phillips 2003).

Table 1.1. Results of mutational studies reveals differences between the chemotaxis mechanisms employed by *B. subtilis* and *E .coli*

Protein	E. coli	B. subtilis
CheA	*cheA* null mutant in *E. coli* (Oosawa *et al.* 1988).	*cheA* null mutant in *B. subtilis* tumbles (Fuhrer & Ordal 1991) .
CheB	*cheB* null mutants are tumbly (Parkinson & Revello 1978).	*cheB* null mutant unable to tumble in response to decreasing concentrations of asparagine (Kirby *et al.* 2000).
CheR	*cheR* mutants are exclusively smooth swimming (Springer & Koshland 1977, Parkinson & Revello 1978).	*cheR* mutants are random (Ullah & Ordal 1981).
CheW	*cheW* mutant is smooth swimming and do not excite (Liu & Parkinson 1989, Stewart *et al.* 1990)	*cheW* mutant excites and adapts but is not efficient at chemotaxis. (Hanlon *et al.* 1992).
CheY	*cheY* mutant is incessantly smooth swimming (Bischoff & Ordal 1991).	*cheY* mutant is exclusively tumbly (Bischoff & Ordal 1991).

1.1.3.1. *CheC*

CheC belongs to the CheC-like family of protein phosphatases, which function to terminate the signal of CheY-p. Members of the CheC-like family (which also includes CheX and FliY) share a highly conserved (Asp/Ser)-X_3-Glu-X_2-Asn-X_{22}-Pro motif (X represents a chain of amino acids of the length that follows the letter X). FliY and CheC, both present in *B. subtilis*, have two of these active sites, while CheX, which is not present in *B. subtilis*, has only one active site. The activity of CheC on its own is about 6% of that of the main phosphatase FliY (Szurmant *et al.* 2004). However, the activity of CheC is increased five-fold when bound to CheD (Szurmant *et al.* 2004, Chao *et al.* 2006, Muff & Ordal 2007, Rao *et al.* 2008), and furthermore the presence of CheY-p increases the affinity that CheC and CheD has for each-other (Chao *et al.* 2006, Muff & Ordal 2007). An intersection point between excitation and adaptation is created when CheD binds to the α2'-helix of CheC, which mimics the receptor substrate, thereby recruiting CheD away from the MCPs (Park *et al.* 2004, Muff & Ordal 2007).

1.1.3.2. *CheD*

CheD is a chemoreceptor glutamine deamidase that catalyzes amide hydrolysis of specific glutaminyl side chains on the chemoreceptors McpA, McpB and McpC. The active site of CheD resembles that of a cysteine hydrolase with a cysteine-histidine catalytic dyad located at the top of the $\alpha/\beta/\beta$ sandwich (Chao *et al.* 2006). The results of mutant studies suggest that CheD is required for the regulation of CheA kinase activity, and that CheC plays a role in this regulatory process (Kirby *et al.* 2001, Muff & Ordal 2007, Rao *et al.* 2008). In this model the activity of CheA is increased when CheD is bound to the receptors, and decreased when CheC recruits it away from the receptors in the presence of high CheY-p levels. Evidence for this model comes from the observation that a CheC mutant that is unable to dephosphorylate CheY-p has limited functionality. A very interesting finding is that in *Thermatoga maritima* CheD also hydrolyzes glutamyl-methyl esters at specific regulatory positions that are mostly different to the targets of CheB, although there is some overlap in specificity. It is predicted that CheD in *B. subtilis* also has the dual enzymatic activity of deamidation and demethylation (Chao *et al.* 2006), however the exact manner in which CheB and CheD are involved in tuning MCP activity is still unclear. Furthermore, CheD has been shown to interact with the HAMP domain in McpC. This interaction was not dependent on the deamidase capability of CheD, but rather its ability to somehow transmit sensory information from the N-terminal domain to the C-terminal domain (Kristich & Ordal 2004). The fact that CheD interacts with the MCPs at multiple functionally distinct sites no doubt provide impetus to further probe its role.

1.1.3.3. *CheV*

CheV is a functional homolog of CheW and also contains a C-terminal domain which is homologous to CheY (Rosario *et al.* 1994). It was shown that while the N-terminal domain of CheV can partially substitute CheW, the same cannot be said for the C-terminal domain and CheY. Based on the results of mutant studies that showed that CheV or CheW is sufficient to modulate CheA activities, it is hypothesized that they could function together as part of the same receptor-bound multi-protein complex (Rosario *et al.* 1994). In addition to its role as a scaffolding protein, it is hypothesized that CheV may play a role in the methylation independent adaptation system of *B. subtilis* in that phosphorylation of the response regulation domain of CheV leads to the inhibition of CheA by compromising its interaction with the attractant bound

MCP (Rao *et al.* 2008). According to (Kirby *et al.* 2000), CheV links CheB to CheA and this system is potentially used for a quick dephosphorylation of CheA in *Bacillus* in response to negative stimuli.

1.1.4. *Impact of understanding chemotaxis on drug-design, bioremediation, agriculture and biosensor applications, as well as the understanding of bacterial social behaviour*

Two component pathways in bacteria provide elegant mechanisms to control not only chemotaxis, but also metabolite fixation and utilization, sporulation, cell division, virulence (Harighi 2009) and antibiotic resistance. By gaining better insight into the chemotaxis pathway, many of the fundamental unanswered questions of signalling biology may be addressed. Bacteria possess the ability to monitor their own population density by sensing the concentration of autoinducer molecules that are released by the microorganisms within the population. When the concentration of these signal molecules reaches a certain threshold, quorum dependent genes are expressed (Nealson *et al.* 1970, Park *et al.* 2003). Bacterial intercellular communication allows microbial communities to determine if their population densities are sufficient or critical and to co-ordinate multi-cellular developmental processes, such as the production of nitrogen-fixing heterocysts in roughly every tenth cell in nitrogen starved *Anabaena* filaments (Aizawa *et al.* 2000). Bio-film formation or virulence responses are also quorum dependent behaviours that rely on coordinated activity of individuals. Chemotaxis is crucial in achieving population aggregates that are dense enough to establish such quorum dependent interactions (Park *et al.* 2003). The combinatorial effect of quorum sensing and chemotaxis give motile bacteria the ability to localize niches that support optimum growth.

The benefits of studying chemotaxis need not only be of academic nature, there are myriad practical applications. The industrial- and agricultural-revolution had without doubt a profoundly positive impact on the socioeconomic status of many, but not without a price to the environment. Much of our natural resources are contaminated by pesticides, acid mine drainage, sewerage, commercial solvents and other pollutants. There exists a vast array of microbes with metabolic capabilities that can be used to detoxify or degrade these pollutants (Aizawa *et al.* 2000). A very promising avenue is to use modified strains of bacteria with specific metabolic abilities and chemotactic preferences to certain stimuli in bioremediation (Miller *et al.* 2009). A better understanding of chemotaxis can also benefit medical research. Chemotaxis plays an important

role in pathogenicity and host cell invasion (Hamer *et al.* 2010) and it is also believed that the receptor proteins are related to multidrug resistance proteins (Stephenson & Hoch 2002). Due to the higher importance and abundance of two-component transduction systems in prokaryotes than in eukaryotes, these systems provide promising targets for the development of new broad spectrum antibiotics with few side effects.

Core elements of the bacterial chemotaxis system seem to be greatly conserved in all bacteria. Even in such distant organisms as *E.coli* and *B.subtilis* the chemotaxis genes *cheA, cheW, cheY* and *cheR* of one may complement the behaviour of null mutants in other (Rao *et al.* 2004). When a motile microorganism finds itself in a new environment it is likely that its chemotaxis system is the first to react to new stimuli. Although advances have been made in terms of understanding the mechanism of adaptation to stimulants, the mechanisms of adaptation of the chemotaxis system to a novel environment are absolutely obscure. With the advent of large scale genome sequencing projects, the availability of genetic information for closely related micro-organisms provided impetus to study micro-evolution of proteins and whole systems. A number of *B.subtilis* related organisms of a significant industrial importance for plant protection have been sequenced recently and will be utilized in this work to study the adaptation of the chemotaxis system of freely living soil microorganisms towards rhizosphere and plant colonization.

1.1.5. ***Members of the Bacillus subtilis group and their economic importance***

The four species *B. subtilis, B. amyloliquefaciens, B. licheniformis* and *B. pumilus* form a closely related taxonomic unit commonly referred to as the *B. subtilis* group. All members of the *B. subtilis* group are Gram-positive, motile, endospore-forming, aerobic or facultative anaerobic mesophiles (Fritze 2004). *B. subtilis* is widely used in fermentation processes, produces a large number of industrially important enzymes, and is a promising anti-fungal agent in agriculture as well as a model organism for research regarding gene regulation, cell differentiation and development (Kobayashi *et al.* 2003, Swain & Ray 2009). Some strains of this group can colonize the developing root system of plants and produce antibiotic compounds as well as plant growth promoting substances. The *Bacillus subtilis* ssp. *subtilis* strain 168 used in this study was derived from an X-ray irradiated strain, Marburg, in 1947 (Lepesant *et al.* 1975). *B. amyloliquefaciens* is a free-living soil dweller which is known to stimulate plant growth by degrading *myo*-inositol hexakisphosphate making phosphorous more available for plants. This

organism produces a large number of anti-fungal and anti-bacterial substances with pharmacological and agricultural value, such as bacillomycin-D, surfcactin, fengycin and bacillaene (Chen *et al.* 2009). Furthermore, it produces proteases and amylases which are used industrially as an ingredient in laundry detergents (Khajeh *et al.* 2006). In this study we use five *B. amyloliquefaciens* strains: plant growth promoting and rhizosphere colonizing strain FZB42, the newly sequenced type strain DSM7^T, and three other agriculturally significant strains B946, B9601Y2 and GaoB3. The strain DSM7^T appears to be non-plant associated (Borriss R., personal communication) (2010). Availability of sequences of multiple closely related strains inhabiting different ecological niches provided us with a unique opportunity to study micro-evolution at the subspecies level. Two more distant organisms, *B.pumilus* and *B.licheniformis* were used for comparison. *B. pumilus* is a widespread soil dwelling organism that colonizes the rhizosphere and produces substances that inhibit fungal and nematode related disease (Choudhary & Johri 2009). Although it is often found in food products, it is not considered to be a pathogenic threat to humans. Potential commercial applications of *B. pumilus* include the production of cellulase, which can be used to convert cellulose containing materials to soluble sugars or solvents. The *B. pumilus* SAFR-032 strain used in this study was isolated from the Spacecraft Assembly Facility at the NASA Jet Propulsion Laboratory. Spores of this strain exhibit high levels of resistance to UV radiation, γ-radiation and resistance to starvation and exposure to H_2O_2 (Gioia *et al.* 2007). *B. licheniformis* is a soil dwelling organism that is mainly associated with plant and plant materials. Its medical importance lies in the fact that it can cause food poisoning in humans (From *et al.* 2007). Its industrial applications include production of proteases and amylases (Khajeh *et al.* 2006), synthesis of an antibiotic, bacitracin, as well as a variety of organic metabolites. A potential use of different *B. licheniformis* strains as biofertilizers or biocontrol agents is under investigation (Choudhary & Johri 2009). The strain used in this study is *B. licheniformis* ATCC 14580, which is the same as the type strain DSM13^T. A prominent characteristic of the *B. subtilis* group is that its members have a strong plant-microbe interactive ability (Reva *et al.* 2004). Plant colonization is a multifaceted process which requires bacteria to respond to a variety of signals from the complex chemical environment that is the plant rhizosphere, in particular to detect and resist plant defense systems, as well as an ability to detect attractants and initiate growth on root and plant surfaces. Therefore, a finely tuned and sensitive chemotaxis system is required to ensure an evolutionary advantage for plant-associated

microorganisms. It has been shown that different strains of the *B. subtilis* group exhibit different specificity towards certain host plants (Reva *et al.* 2004). Therefore, it is reasonable to assume that the differences in the chemotaxis proteins of the members of the *B. subtilis* group may be closely correlated to differences in their habitats.

1.2. **Research aims and objectives**

The major aim of this research is to combine different bio-informatics approaches to compare and analyze the proteins involved in chemotaxis in the various members of the *B. subtilis* group to identify areas in the proteins that may confer an adaptive advantage to the organisms in their various environments. Since no structures have been solved for these proteins in any of the members of the *B. subtilis* group we have inferred homology models for some individual proteins as well as for their complexes. We have analyzed the selective forces acting upon each protein to determine which proteins are more tolerant to mutations and therefore likely to be subjected to positive selection and thus can cause a phenotype variation, in terms of habitat preference. Proteins that are less tolerant to mutations will be more conserved and have larger areas under purifying selection. We have mapped the selective forces acting on each amino acid site onto the homology models, to determine spatial proximity to known active regions.

Despite half a century of research on chemotaxis, many questions remain unanswered. This work provides a structural framework for future mutant studies that will attempt to answer some of these questions.

CHAPTER 2

2. STRUCTURAL MODELLING OF *BACILLUS AMYLOLIQUEFACIENS* FZB42 CHEMOTAXIS PROTEINS

The objective of this chapter is to give an introduction to the steps involved in homology modelling and to discuss the motivation behind creating such models for the chemotaxis proteins of *B. amyloliquefaciens* FZB42. To date, no 3D structure for any of the chemotaxis proteins from members of the *B. subtilis* group have been determined experimentally, therefore the homology models produced as part of this work provide a basis for understanding how these proteins function on a molecular level.

2.1. Introduction

Life is a characteristic that discerns objects that contain chemical processes involving thousands of different reactions occurring in an organized manner from those objects that do not, either because such functions have ceased or because the objects in question are innert (Koshland Jr. 2002). The processes of life include but are not limited to metabolism, homeostasis, growth, replication, response to stimuli and adaptation to environmental changes, which all involve proteins. The dominant presence of proteins in all the reactions that are critical for life is staggering at the least. The rapid chemical reactions occurring in a cell are dependent on the catalytic function of enzymes. The transport of molecules across otherwise impermeable membranes relies on the presence of proteins. Proteins may also have a structural function or play a role in molecular signaling. Proteins are complex polypeptide chains of linked amino acids that exhibit astonishing versatility which allows them to perform such a large array of activities (Scheeff & Fink 2003). The function of a protein is largely governed by its shape, which in turn is determined by its amino acid sequence. The distinctive 3D structures of proteins are fashioned in such a way to allow for the precise placement of particular chemical groups or other proteins, therefore any inquiry to the function of a protein must be based on an understanding of the protein's structure. The leading experimental method to obtain atomic resolution information for large macromolecules such as proteins is X-ray crystallography (Pusey *et al.* 2005). An astonishing 61 438 of the 70 695 structures in the PDB (as of 20 January,

2011) have been solved by this method which is known for the high resolution structures that it produces. In principle, when an X-ray beam is directed at an object a characteristic diffraction pattern composed of spots is formed. Due to the weak interaction of X-rays with matter, scattering is also weak, thus the presence of multiple ordered molecules is required (Oksanen & Goldman 2010). Hence it is important to purify the protein under investigation to grow a set of highly ordered crystals (Pusey *et al.* 2005, Oksanen & Goldman 2010). The diffraction pattern produced by the crystals is recorded and analyzed on the basis of Bragg's law to obtain the unit cell parameters. Each diffraction spot is characterized by a specific amplitude, wavelength and phase, which are used to infer atomic position (Schlick 2002). Challenges of X-ray crystallography include growing protein crystals, which can take anything from between a few hours to months, and trying to compute the phases of the diffracted waves in order to obtain an electron density map (Goldman & Ordal 1984, Oksanen & Goldman 2010). Furthermore, studying flexible proteins using X-ray crystallography has been proven to be problematic, since it is difficult to align molecules in an exact orientation, resulting in a smeared electron density (Lamzin *et al.* 1999).

Another popular experimental method for obtaining protein structures and the preferred method for determining flexible protein structures is multi-dimensional nuclear magnetic resonance (NMR) (Šali & Blundell 1993). This method has been used to solve 8 734 of the structures in the PDB (http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html) as available on the date of access (20 January 2011).

NMR is based on the quantum mechanical properties of atoms, in particular spin and how it is affected by applying an external magnetic field. In the absence of an external magnetic field, the spins of magnetic nuclei such as ^{1}H, ^{13}C and ^{15}N can be described as having a random orientation. When a magnetic field is applied their spins will orient either against or with the field. When probed with radio waves, the nuclei absorb energy and undergo a spin-flip from a lower to a higher energy state (McMurry 2003). The resonance frequency for different atoms is unique, but it is also affected by the local environment. Due to the effect of shielding from surrounding electrons each nucleus in a molecule produces a unique absorption signal when it comes into resonance. The observed resonances are then analyzed to obtain a list of atomic nuclei that are in close proximity to one another, and to detail the local conformation of atoms

that are bonded together (Primrose &Twyman 2006). A unique characteristic of NMR is that structure can be determined in solution, thus it is possible to control the characteristics of the fluid so that it can resemble the environment in which the protein functions. This is very important, since changes in pH and temperature can affect the structure of a protein, thus NMR enables the study of protein structure in physiological relevant conditions (Riek *et al.* 2002). Due to problems associated with overlapping peaks in NMR spectra when dealing with large proteins, structure determination by NMR is currently limited to small or medium proteins under 100 kDa (Primrose & Twyman 2006). The aforementioned experimental methods of structure determination are time consuming and not successful with all proteins (Xiang 2006), hence the available structures in the PDB represent only a fraction of the 11 636 205 protein sequences found in the UniProtKB/TrEMBL database at the time of writing this document (http://www.ebi.ac.uk/uniprot/TrEMBLstats/).

A promising alternative to experimental methods is presented by theoretical *in silico* structure determination, which can help close the gap between the growing number of sequenced protein coding genes and solved structures (Sánchez & Šali 1997, Krieger *et al.* 2003, Bujnicki 2006, Castrignanò *et al.* 2006, Ginalski 2006). There exist various methods for predicting protein structure based on sequence, these methods can be divided into two main categories: *De novo* modelling, also known as template-free modelling, and template-based modelling. *De novo* modelling is usually employed for targets for which no relationship with a protein with solved structure can be detected. Traditionally approaches that simulate the folding of a protein based on simple physics principles were the dominant methods used (Srinivasan *et al.* 2004, Moult 2005). However, newer methods for template free-prediction make use of substructure relationships on a scale ranging from a few residues, through secondary structure units to super-secondary units (Moult 2005). Template based methods employ templates that share homology with the target sequence, or non-homologous fold relationship. To evaluate current state-of-the-art structure prediction methods, a community wide event called Critical Assessment of Techniques for Protein Structure Prediction (CASP 1-9) has been performed bi-annually since 1994. The results of each CASP event have been published in an attempt to document techniques used and to highlight areas where shortcomings would benefit from further efforts (Jones 1997, Tramontano & Morea 2003, Hillisch *et al.* 2004, Moult 2005). Despite recent advances, *de novo*

The three main homology-based methods include: rigid-body assembly; segment matching and modelling by satisfaction of spatial restraints (Sánchez & Šali 1997, Fiser & Šali 2003). A fourth method, known as generalized comparative modelling, was developed more recently (Kolinski *et al.* 2001). The COMPOSER package (Sutcliffe *et al.* 1987) is a popular rigid-body assembly software tool. In rigid-body assembly the model is constructed from small rigid pieces of core regions, loops and side chains. These small rigid pieces are obtained from anatomization of related structures. Backbone construction involves averaging the C_α atom coordinates from structurally conserved residues of the template structures. Main chain atoms are superimposed from the template sequence that exhibits the highest sequence similarity to the target. A heuristic approach is used to construct the side chains and optimize their conformations. Loops are built after a library search for homologous structures (Sutcliffe *et al.* 1987).

Modelling by segment matching involves breaking a target sequence into short segments, matching it to solved structural segments and then building a target structure. A database of known structures is searched to find matching segments which are fitted together based on the target sequence. Generally only the C_α atomic co-ordinates are used and conformational restraints are applied. When choosing a matching database segment the following criteria are taken into consideration: amino acid sequence similarity; conformational similarity and compatibility with the target structure based on Van der Waals' interactions. The SEGMOD is an example of a package that implements modelling by segment matching (Fain & Levitt 2001).

Modelling by satisfaction of spatial restraints is no doubt the most promising of all homology modelling techniques as various restraints can be applied to guide the modelling process. Heuristic based constraints can be obtained from secondary-structure packing, analysis of hydrophobicity and correlated mutations, empirical potentials of mean force, NMR experiments, cross-linking experiments, image reconstruction in electron microscopy, site directed mutagenesis, fluorescence spectroscopy and more (Sánchez & Šali 1997). A program like MODELLER extracts homology derived restraints on distances and dihedral angles in the target sequence from the target-template alignment; these are combined with stereochemical restraints obtained from CHARM22 force-field parameters and statistical preferences of dihedral angles and non-bonded atomic distances obtained from a representative set of all experimentally determined structures. An optimization method relying on conjugate gradients and molecular

dynamics is used to calculate the model in such a way that violations of spatial restraints are kept to a minimum (Fiser & Šali 2003).

The fourth method, known as generalized comparative modelling (GENECOMP), uses a combination of sequence comparison, threading, and lattice modelling for protein structure prediction and refinement (Kolinski *et al.* 2001). It has been shown that the four model-building methods deliver results of comparable accuracy (Fiser & Šali 2003).

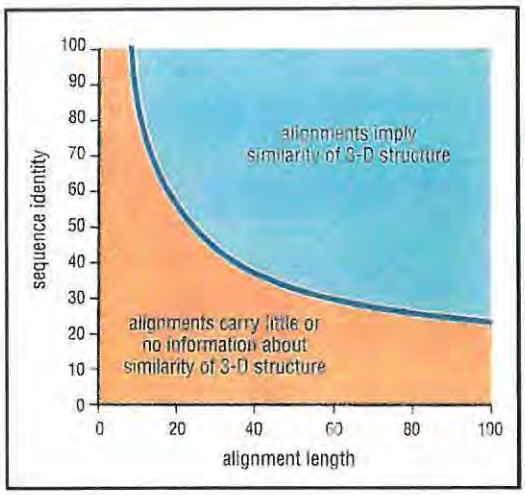


Figure 2.1. The curved line divides the graph into a safe homology modelling zone (blue) and a region known as the "twilight zone" (brown) where homology modelling becomes highly problematic. This figure was obtained from **http://swift.cmbi.kun.nl/teach/B4/drgdes_5.html** and is used here with permission of the author.

2.1.2. *Steps in the homology modelling process and the tools used to perform them*

In practice the homology modelling consists of several consecutive steps that can be iterated until a model of satisfactory quality is obtained (Krieger *et al.* 2003, Ginalski 2006). The steps are described as follows: (i) Retrieval of target sequence; (ii) Finding suitable template(s) that share homology with the target; (iii) Aligning the target and template(s) sequences; (iv) Generation of the backbone, predicting loops and modelling side chains; (v) Loop refinement; (vi) Validating the resulting model. One way to ensure that a satisfactory model is constructed is to repeat the basic model building steps until no further improvement in the model is detected (Fiser & Šali 2003).

2.1.2.1. *Sequence retrieval from publicly available databases*

The GenBank sequence database is a publicly available collection of nucleotide sequences and their protein translations that are submitted by laboratories that generate sequence data. This database is the result of collaboration between the National Center for Biotechnology Information (NCBI), the European Molecular Biology Laboratory (EMBL) Data Library from the European Bioinformatics Institute and the DNA Data Bank of Japan (DDBJ). GenBank contains sequences from over 100 000 distinct organisms and continues to grow at an exponential rate. The NCBI holds another database, RefSeq, which unlike GenBank contains only one example of each natural biological molecule for major organisms ranging from viruses to bacteria to eukaryotes. RefSeq aims to provide separate and linked records for the genomic DNA, gene transcripts and protein products of the gene transcripts. The RefSeq database is curated and is restricted to major organisms for which adequate data is available. For example in 2007 sequences for almost 4,000 distinct named organisms were available in the database. In contrast GenBank includes sequences for any organism submitted and in 2007 sequences from more than 250,000 organisms were available (2010 National Center for Biotechnology Information). Both databases continue to grow considerably, by late 2009 RefSeq contained 11,934,213 proteins from 11,536 organisms (http://www.ncbi.nlm.nih.gov/RefSeq/) and GenBank contained 108,431,692 sequence records (http://www.ncbi.nlm.nih.gov/genbank/). It is expected that this will continue increasing at a rapid pace due to recent advances in sequencing technology coupled with what seems to be a world-wide effort to sequence the genomes of as many organisms as possible.

2.1.2.2. *Template identification and selection with the HHpred server*

It is often the case that the percentage identity between the target and a possible template is high enough to allow detection with BLAST (Altschul *et al.* 1990). However, when dealing with remote homologs, more sensitive methods such as PSI-BLAST (Altschul *et al.* 1997) and methods using Hidden Markov Models (HMMs) are more successful.

The HHpred server (http://toolkit.tuebingen.mpg.de/hhpred) is a free protein function and protein structure prediction server based on the HHsearch method. This server is the first publicly available server that makes use of the pair-wise comparison of profile-HMMs. HHpred allows

the user to select from a wide choice of databases such as the PDB, SCOP, Pfam, SMART, COGs and CDD. Most other conventional search methods only search sequence databases, but HHpred can search alignment databases. Sequence-sequence comparison is inferior to profile-sequence comparison due to the presence of false positives and less than optimal alignments. In a sequence profile each column of the multiple sequence alignment contains the frequencies of each of the 20 amino acids. Thus a sequence profile holds detailed information about the conservation of each residue position, this information can be used to infer the preferred amino acids for each site and the importance of each residue position for defining members of a protein family. HMMs are a class of probabilistic models that are generally applicable to time series or linear sequences (Eddy 1998). In other words, they are statistical descriptions of the primary consensus of a group of sequences, where the state at position *i* is dependent on the state at position *i*-1. At each consensus position in the MSA, a profile HMM models the distribution of amino acid residues allowed at that position, furthermore an "insert" state and a "delete" state allow for the insertion and deletion of one or more residues between one column and the next (Eddy 1998). Simple sequence profiles contain information about the preffered amino acid sites at each position, but no information about the frequency of inserts and deletions at each column. These characteristics of profile HMMs improve the sensitivity of HHpred over other conventional methods (Soding *et al.* 2005, Hildebrand *et al.* 2009).

The three steps used by HHpred to detect homologous templates are: (i) A multiple sequence alignment (MSA) of homologs is constructed for the target sequence by multiple iterations of PSI-BLAST searches against the NCBI non-redundant database. The final alignments are annotated with the predicted secondary structure and confidence values from PSIPRED (Jones 1999, Soding *et al.* 2005) (ii) a profile HMM is created from the MSA that includes information about the predicted secondary structure; (iii) the query HMM is compared with the pre-calculated HMMs in the database. The database HMM contains secondary structure information as predicted by PSIPRED or assigned from 3D structure by DSSP (Kabsch & Sander 1983, Soding *et al.* 2005). The HHsearch software is used to search the database, and sensitivity is gained by using position-specific gap penalties. A score for the secondary structure is added to the total score. A disadvantage of this method is that marginally significant scores can be obtained for structurally analogous but non-homologous proteins. Nevertheless, during the community wide protein structure prediction benchmark CASP8, HHpred was ranked as the 2nd

best automatic structure prediction server for single domain proteins and 7th best on all targets, while being 50 times faster than the 20 top-ranked servers (Hildebrand *et al.* 2009). The results of HHpred are presented in a user friendly format quite similar to that of PSI-BLAST. The results are divided into two parts: a summary of matching templates and a list of target template-alignments. A probability score is given with each alignment, which indicates the probability in percent that the database match is a true positive. This score is a conservative which corrects for occasional high scoring false positives that result from corrupted alignments containing non-homologous sequences. Secondary structure information is visible as part of the target-template alignments, making it easy for the user to see if there are gaps that would make homology modelling difficult to ascertain. The sequence identity between target and template is listed in each case, as well as the resolution, making it easy for the user to select the most suitable template from a list of candidates.

2.1.2.3. *Aligning target and templates using Promals3D*

The key determinant of model quality is the construction of an accurate target template alignment, for an inaccurate alignment will result in an inaccurate model (Venclovas *et al.* 2003, Soding *et al.* 2005, Bujnicki 2006). The PROMALS3D web-server employs a progressive method that integrates state of the art alignment techniques such as probabilistic consistency of profile-profile comparisons and additional information from database homologs and predicted secondary structures (Figure 2.2). The user can submit as input multiple sequences as well as 3D structures. Similar sequences are clustered together for fast and easy alignment, and more advanced techniques are applied to align divergent clusters. The initial step involves aligning similar sequences using a scoring function of weighted sum-of-pairs of BLOSUM62 scores. The resulting pre-aligned groups are relatively distant from each other, and a representative sequence from each is selected. The representatives from each group are subjected to PSI-BLAST searches for the retrieval of additional homologs from the UNIREF database and PSIPRED secondary structure prediction. Thereafter a HMM model of profile-profile alignments with predicted secondary structures is applied to pairs of representative sequences to obtain posterior probabilities of residue matches. These probabilities are combined with constraints derived from homologs with 3D structures to arrive at a probability-consistency scoring function. To form the final multiple alignment, the pre-aligned groups that were constructed in the first stage are

merged with the alignment of representatives. The advantage of PROMALS3D over more conventional methods is that it produces high quality alignments that are consistent with sequences and structures of proteins (Pei *et al.* 2008). The PROMALS3D web-server is available at http://prodata.swmed.edu/promals3d/promals3d.php.

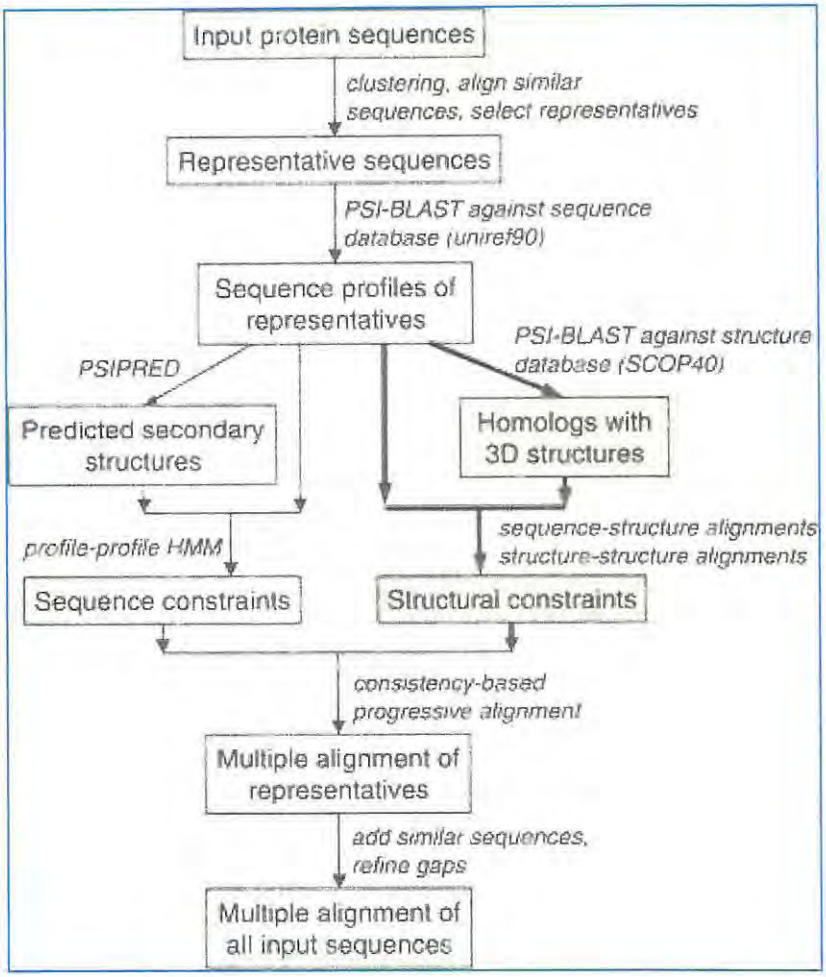


Figure 2.2. A PROMALS3D flow diagram showing the steps followed to create a multiple sequence alignment. The stages indicated in bold contain additional steps to remove redundancy when more than one structure is available. This image was obtained from Pei *et al.* (2008) with permission from the author and Oxford University Press.

2.1.2.4. *Modelling by satisfaction of spatial restraints*

MODELLER is no doubt one of the most popular homology modelling programs due to high accuracy (Dalton & Jackson 2007). The program is used with a scripting language and does not include a graphical user interface. The user provides an alignment between the target sequence and the template(s) as well as a Python script containing MODELLER commands. The output is a 3D model for the target sequence that encompasses all main-chain and side-chain non-hydrogen atoms. Given an alignment the model is obtained without any user intervention as MODELLER automatically derives the constraints from the alignment with the known structure (Šali 2009), which are then expressed as probability density functions (pdfs) for the various restraint types. The pdfs restrain C_α-C_α and backbone N-O distances as well as backbone and side-chain dihedral angles for different residue types to produce a model with minimum violations of these restraints (Wallner & Elofsson 2003). A noteworthy feature of the method is that the spatial restraints are obtained empirically from a database of protein structure alignments (Šali & Blundell 1993). An objective function is obtained by combining the spatial restraints with CHARMM energy terms enforcing proper stereochemistry (Šali & Blundell 1993, Brooks *et al.* 2009). The objective function is optimized in Cartesian space to produce the model. A variable target function method is used to achieve model optimization (Braun & Gō 1985) by employing methods of conjugate gradients and molecular dynamics with simulated annealing (Šali 2009). A number of slightly different models can be calculated by varying the initial structure. The inconsistencies between these models can then be applied to assess inaccuracies in the corresponding regions of the fold (Šali 2009).

2.1.2.5. *Model evaluation with MetaMQAPii, Verify3D PROCHECK and ProQ*

One of the persistent problems in determining 3D structure is how to determine the correctness of the final model. The number of errors in a homology model depends on the percentage sequence identity between the target and template and also the number of errors in the template itself. Methods for estimating errors in structures can roughly be divided into two categories: statistical potential based methods or physics based energy calculation methods. The first method compiles statistical profiles of spatial features and interaction energies of experimentally determined structures and then assigns a score that indicates how well a given characteristic of the constructed model resembles the same characteristic in experimental structures. The second

category evaluates the stereo-chemistry of the model to ensure consistency with the physiochemical rules by checking for anomalies in φ-ψ angles and bond lengths amongst other symmetry and geometry checks (Primrose & Twyman 2006). Although various methods have been developed to detect local inaccuracies in unrefined crystallographic models, their usefulness in evaluating computational models is debatable.

MetaMQAPii (https://genesilico.pl/toolkit/unimod?method=MetaMQAPii) is a meta-server that was developed specifically for the quality assessment of computational protein models. It is based on a multivariate regression model, which uses scores from VERIFY3D, PROSA, BALA, ANOLEA, PROVE, TUNE, REFINER and PROQRES, but in which trivial parameters are controlled. Trivial features can be calculated directly from each atom and includes parameters such as: solvent accessibility, depth in the structure and the number of local and non-local neighbours. MetaMQAPii predicts the Root Mean Square Deviations (RMSD) of individual C_α atoms between the computational model and the unknown native structure, as well as global deviations expressed as a Global Distance Test total score (GDT_TS). The GDT_TS is a measure of similarity between two protein structures with identical sequences but different 3D structures. Additionally MetaMQAPii also produces a PDB file of the model in which the temperature fields have been replaced with the MetaMQAPii scores, making it easy to visualize areas of inaccuracy (Pawlowski *et al.* 2008). According to current opinion in available literature a sound model typically has a GDT_TS score of >75 and RMSD < 2Å (Soares *et al.* 2009, Wang *et al.* 2010). It is imperative to note that MetaMQAPii merely predicts the deviation of a homology model from the native structure; a true deviation can only be calculated if the native structure is known. Therefore scores must be interpreted as estimations of model quality and not as ultimate validations (Kaminska *et al.* 2010).

Verify3D (http://nihserver.mbi.ucla.edu/Verify_3D/) (Luthy *et al.* 1992) was initially designed as a tool to help in the refinement of crystallographic structures, but has become a popular method for the evaluation of homology models (Katiyar *et al.* 2009, Makkar *et al.* 2009). Verify3D uses a statistical approach to measure the accuracy of a structure by comparing its compatibility to its own amino acid sequence using a 3D profile scoring function. This method is known as 3D-1D comparison. Every residue in the structures is allocated a structural class based on its position and environment (α-helix, β-sheet, loop, polar, non-polar, etc.). To obtain a score

for each of the 20 amino acids in a particular class a collection of correct structures is used as a reference. An incorrectly modelled segment in an otherwise correct structure can be identified by adding and plotting the scores of a sliding 21-residue window for each residue (Luthy *et al.* 1992). A compatibility score greater than 0 corresponds to an acceptable side chain environment, while negative scores are considered problematic. If more than 80% of the residues in a structure score are greater than 0.2, the structures are considered to be of experimental quality (Kativar *et al.* 2009. Makkar *et al.* 2009).

PROCHECK (http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/) assesses the stereochemical quality of a protein by determining the degree that residue geometry resembles or deviates from the norm, as derived from stereochemical parameters of well-refined high-resolution structures. PROCHECK produces a number of plots describing and evaluating the overall residue-by-residue geometry of the structure. Regions highlighted by PROCHECK as unusual are not always necessarily erroneous, but may contain distortions caused, for example, by ligand binding (Laskowski *et al.* 1993). Among the plots generated by PROCHECK, the Ramachandran plot indicates φ-ψ torsion angles for all residues in the structure, excluding those at the chain termini. The Ramachandran plot colouring divides it into most favourable, generously allowed, additionally allowed and disallowed regions. Based on analysis of 118 structures with a resolution of at least 2.0 Ångstrom (Å), and an R factor no greater than 20%, a good quality structure should have more than 90% residues in the most favored region (Katiyar *et al.* 2009).

ProQ (http://www.sbc.su.se/~bjornw/ProQ/) is a neural network that predicts the quality of computational models by extracting structural features such as frequency of atom-atom contacts (Wallner & Elofsson 2003). The ProQ-LG neural network was trained to evaluate models using the Levitt and Gerstein (LG) score, which represents the significance of a score associated with the best subsection of a structural alignment between a model and a correct structure (Cristobal *et al.* 2001, Wallner & Elofsson 2003).

2.1.2.6. *Loop refinement*

Loops represent the most variable regions of a structure, and insertions and deletions often occur here. Often homology restraints such as Cα-Cα distance restraints cannot be applied in these regions. Given a good energy function, loop refinement can relax the backbone closer to the native state (Xiang 2006). MODELLER provides several loop optimization methods, all of which rely on scoring functions and optimization protocols that were altered for loop modelling (Šali 2009). Loop regions can be refined automatically after standards model building or manually on an existing PDB file. For automatic loop refinement the *loopmodel* class must be used instead of the standard *automodel* class. When an existing PDB file is modified it is necessary to redefine the *loopmodel.select loops atoms()* routine as no alignment is available for automatic loop detection. The user can also define the range of atoms to be refined in the script, otherwise, all atoms are selected by default. The first step of the loop modelling method involves the generation of an initial loop conformation by positioning the atoms of the loops with uniform spacing on the line connecting the main-chain carbonyl oxygen and amide nitrogen atoms of the N- and C- terminal anchor regions respectively. Subsequently a number of models (as specified by the user) is inferred by randomizing the initial loop conformations by ±5Å in each of the Cartesian directions. Model optimization is performed twice: first only the loop atoms are taken into consideration, and thereafter the rest of the system is taken into account. An atomistic distance-dependent mean force for non-bond interactions is used as a basis for the optimization (Melo & Feytmans 1997). All amino acids are categorized according to 40 atom classes and a potential as MODELLER cubic spline restraints are applied. This procedure does not involve any homology-derived restraints (Šali 2009).

2.1.3. *The chemotaxis proteins of the B. subtilis group*

Although various structures for chemotaxis proteins have been solved experimentally, none of these structures belong to members of the economically important *B. subtilis* group. As stated in Chapter 1, the bacterial chemotaxis system is very diverse and there are marked differences between the mechanisms employed by different taxons. Since structural information is crucial for a holistic understanding of a protein's function, the main objective as outlined in this chapter was to construct homology models for the individual chemotaxis proteins of *B. amyloliquefaciens* ssp. *plantarum* FZB42 (referred to in remainder of text as *B. amyloquefaciens*

FZB42) as well as for some interacting partners. *B. amyloliquefaciens* FZB42 was selected as target organism because future work will aim to address the observation that certain strains of *B. amyloliquefaciens* are plant associated while others are not by investigating the effect that variations in the chemotaxis proteins have on habitat preferences. The chapter is concluded with a discussion of the fitness of the models. The motivation behind building the homology models was to map the selective forces acting on individual sites of the target proteins onto the 3D structures to identify regions that are under strong evolutionary constraints as well as regions where mutations may give rise to novel functions. The complexes were analyzed to determine which residues are important for protein-protein recognition. These objectives are covered in subsequent chapters. Future work will attempt to reconstruct the network of interacting partners in the chemotaxis pathway by using these models in docking studies.

2.1.4. *Prediction of residue interactions important for maintaining tertiary and quaternary structures*

In order to understand the molecular basis of stability and function in a protein or a protein-protein complex it is essential to study the various weak and strong interactions that maintain structural integrity. These interactions include: hydrophobic interactions, ionic interactions, disulphide bonds, hydrogen bonds, aromatic-aromatic interactions, aromatic-sulphur interactions and cation-π interactions (Tina *et al.* 2007).

2.2. Methods

The application of the steps, described previously, to modelling various chemotaxis proteins from *B. amyloliquefaciens* FZB42 will be discussed in the following sections. Models were visualized in PyMOL (Schrödinger 2010).

2.2.1. *Sequence retrieval*

Amino acid sequences for the proteins CheA, CheB, CheC, CheD, CheR, CheW and CheY from *B. amyloliquefaciens* FZB42 (NC_009725) were retrieved from the NCBI database. These sequences served as the targets for homology modelling. To create a dataset of orthologs for the *B. subtilis* group to be used in further analysis, orthologs for the target proteins and McpA, McpB and McpC were obtained for *B. subtilis* 168 (NC_000964), *B. subtilis* SMY (ABQN01000001-ABQN01000009), *B. subtilis* JH642 (ABQM01000001- ABQM01000009) *B.*

subtilis NCBI 3610 (ABQL01000001-ABQL01000005), *B. licheniformis* str. ATCC 14580 (NC_006270.3) and *B. pumilus* SAFR-032 (NC_009848) which were retrieved from the NCBI database. Newly sequenced *B. amyloliquefaciens* strains DSM7^{T}, B946, B9601Y2 and GaoB3 were provided by Prof. R. Borriss (Humboldt University, Berlin) and added to the datasets for each of the proteins under study. These datasets together with template structures were used as input for the multiple sequence alignment program PROMALS3D (Pei *et al.* 2008). Datasets were also built for CheV, McpA, McpB and McpC, although no homology models were built for these proteins due to a lack of suitable templates.

2.2.2. *Template identification*

To identify suitable templates for modelling the chemotaxis proteins of *B. amyloliquefaciens* FZB42 the HHpred server was used. The *pdb70* was selected as the HMM database. The regularly updated *pdb70* is an alignment database built around sequences of known structure and uses full-length sequences from the PDB as seeds. The number of PSI-BLAST iterations was set at 8 which is the default value. The minimum coverage of the query by the PSI-BLAST matches were set at 20 which means that at least 20% of the residues of the query must align with residues from the matched sequence for it to be included in the profile. Benchmark tests have shown that this value improves sensitivity without negatively affecting selectivity. The alignment mode was set to "local" to increase sensitivity for remote homolog detection (Soding *et al.* 2005). Suitable templates were chosen based on their coverage of the target, their sequence identity to the target, secondary structure features and probability of being a true homolog. Although template structures were verified before submission to the PDB and can therefore be assumed to be of acceptable quality, they were analyzed with PROCHECK to identify residues in disallowed regions which may lead to local inaccuracies in the models based on them. Template structures and sequences were downloaded from the PDB following their identification by HHpred.

2.2.3. *Target-template alignments*

Correct alignment between target and template is critical in homology modelling, as misaligned sections will give rise to erroneous regions in the model (Xiang 2006). To improve alignment quality over that of a simple pairwise alignment, PROMALS3D was used to align orthologs from the *B. subtilis* group with templates. Results from PROMALS3D were compared to those of

HHpred. If needed, the alignments were edited by removing N- terminal and C-terminal sections where the alignment quality was poor due to the presence of gaps. The modelling procedure was repeated until models of acceptable quality were obtained.

2.2.4. *Modelling by satisfaction of spatial restraints*

We have used the MODELLER9v7 package to build homology models for CheB, CheC, CheD, CheR, CheW, CheY and the P1 and P2 domains of CheA from *B. amyloliquefaciens* FZB42 as well as for interactions between: CheC and CheD; CheY and the two phosphatase active sites of CheC; CheY with the P1 domain of CheA; CheY with the P2 domain of CheA; CheB and the P1 domain of CheA; The N-terminal domain of CheB with the P2 domain of CheA and for CheW in complex with the P4 and P5 domains of CheA.

In the case of the single models the target sequences were aligned with only one template that covered most or all of the target sequence length. In the case of the complexes single or multiple templates were used as described in following sections. Initially only one model was built for each target, which was then assessed based on its Discrete Optimized Protein Energy (DOPE) score. DOPE uses an enhanced reference state that correlates with non-interacting atoms in a homogenous sphere. The radius of the sphere is dependent on a sample true structure and accounts for the finite and spherical shape of native structures. Because the DOPE score has an arbitrary scale, it must be normalized before different proteins can be compared; this normalized version is indicated as DOPE Z (Šali 2009). Generally a DOPE Z score <-1 is considered an indication of a native like structure. In this study, if the model quality was not satisfactory, the alignment was adjusted and another model was built until sufficient improvement, in terms of DOPE Z score, was made. Thereafter, 100 models with slight variations in structure were built. Again, models were assessed using the DOPE Z scores. No refinement was performed during the modelling process because energy based refinement usually leads to degradation instead of improvement of the model, unless the sample space is sufficiently restricted to avoid false attractors (Qian *et al.* 2004). For each target, the model with the lowest DOPE Z score was selected for further steps such as loop refinement (if needed).

2.2.5. *Model validation*

For this study MetaMQAPii and ProQ were used to assess the quality of individual proteins. MetaMQAPii can only evaluate single chain structures, thus complexes were evaluated using ProQ, PROCHECK and Verify3D.

2.2.6. *Loop refinement*

Structures with inaccurately modelled regions, as identified by model quality assessment programs, were subjected to loop refinement until a sufficient model, based on DOPE Z score and model quality assessment results, was obtained. The region to be refined was specified and initially 5 models were generated. Based on the DOPE Z score the best model was then selected and another 100 models were built with the same range of residues selected for refinement. The models were then evaluated based on DOPE Z score. The "refine.very_slow" option was selected for all refinement steps to produce models of the highest possible quality. The process was repeated for various loop regions until an acceptable DOPE Z score was obtained. Models needing loop refinement were: CheB, CheC, CheD, CheR, CheW, CheY, the interaction between CheW and CheAP3P4, CheC and CheD, CheB and CheAP1, $CheB_N$ and CheAP2

2.2.7. *Renumbering models*

When MODELLER creates a .pdb file it arbitrarily numbers the first atom as "1", in single chain models no chain identifier is assigned, in multi chain models, chains are named alphabetically and residues are numbered in a continuous fashion. Therefore the residue numbers often do not correspond to the actual position in the target amino acid sequence. To rectify this problem ModifyPDB, which is a part of the ResDe package (Hintze & Johnson 2010), was used to renumber the residues and rename the chains in the .pdb files of all modeled complexes as well as the models for CheD, and CheAP2.

2.2.8. *Prediction of residue interactions important for maintaining tertiary and quaternary structure using Protein Interactions Calculator*

Protein Interactions Calculator (PIC) (http://crick.mbu.iisc.ernet.in/~PIC/) is a server which accepts a 3D structure of a protein or a multiple protein complex as input, and then calculates various interactions such as: hydrophobic interactions within 5Å, ionic interactions, disulphide bonds, main-chain to main-chain hydrogen bonds, main-chain to side-chain hydrogen bonds,

side-chain to side-chain hydrogen bonds, aromatic-aromatic interactions, aromatic-sulphur interactions and cation-π interactions. Standard published criteria are used as the basis for all interaction calculations (Tina *et al.* 2007).

2.2.9. *Identification of active sites*

The active sites for all the templates have been determined experimentally. Active sites in the homology models were located by aligning the models with their respective templates and mapping the active sites to structure. The analysis of specific residues that play a role in domain-domain contacts for multidomain chemotaxis proteins and complexes within the *B. subtilis* group will be discussed in Chapter 3 and Chapter 4 respectively.

2.3. Results

The results obtained in an attempt to infer homology models for the chemotaxis proteins of *B. amyloliquefaciens* FZB42 are described in the following sections.

2.3.1. *Sequence retrieval*

Amino acid sequences for the proteins CheA, CheB, CheC, CheD, CheR, CheW and CheY from *B. amyloliquefaciens* FZB42 were retrieved from the NCBI database. These sequences served as the targets for homology modelling.

2.3.2. *Template identification*

Target sequences were submitted to the HHpred server to identify possible templates for homology modelling. Templates were selected based on their percentage sequence identity to the target, the number of gaps in the alignment, correlation between the predicted secondary structure of the template and target secondary structure, as well as the probability that the template is a homolog of the target. Template and target information are summarized in Table 2.1 and Table 2.2.

Table 2.1. A summary of target and template information for single protein models.

Target	NCBI accession number	Template PDB ID	Template organism	Identity	Residue range of target that aligns with template (sequence length of full protein)
CheB	YP_001421220.1	1A2O_A	*S. typhimurium*	39%	2 –354 (355)
CheC	YP_001421223.1	1XKR_A	*T. maritima*	30%	4 –207 (209)
CheD	YP_001421224.1	2F9Z_C	*T. maritima*	42%	11 –158 (166)
CheR	YP_001421681.1	1AF7_A	*Salmonella typhimurium*	28%	2 –254 (256)
CheW	YP_001421222.1	2QDL_A	*T. tengcongensis*	29%	9 –156 (157)
CheY	YP_001421211.1	1TMY_A	*T. maritima*	71%	1 -119 (120)
CheAP1	YP_001421221.1	3KYJ_A	*R. sphaeroides*	21%	1 -128 (670)
CheAP2	YP_001421221.1	1U0S_A	*T. maritima*	35%	163 -247 (670)

Table 2.2. A summary of target and template information for complexes. ID refers to the percentage identity between target and template

Protein complex	Template(s) for 1st chain	Template organism(s)	ID	Template(s) for 2nd chain	Template organism(s)	ID
CheC CheD	2F9Z_A	*T. maritima*	31%	2F9Z_C	*T. maritima*	42%
CheAP4P5 CheW	2CH4_A	*T. maritima*	47%	2CH4_W	*T. maritima*	29%
CheAP1 CheY	3KYJ_A	*R. sphaeroides*	21%	3KYJ_B	*R. sphaeroides*	29%
CheAP1 CheB	3KYJ_A	*R. sphaeroides*	21%	1A2O_A. 3KYJ_B	*S. typhimurium* *R. sphaeroides*	39% 34%
CheAP2 CheY	1U0S_A	*T. maritima*	35%	1U0S_Y	*T. maritima*	71%
CheAP2 CheB	1U0S_A	*T. maritima*	35%	1A2O_A 1U0S_Y	*S. typhimurium* *T. maritima*	39% 43%

2.3.2.1. *CheB*

The HHpred search results returned two possible templates that had a 100% probability of being a true homolog of *B. amyloliquefaciens* FZB42 CheB. The two possible templates were PDB ID: 1A2O chain A (Djordjevic *et al.* 1998) and 1CHD chain A (West *et al.* 1995). Both of these templates come from *S. typhimurium* CheB, but only 1A2O chain A covered the almost entire sequence length of the target (Table 2.1 and Table 2.2). The alignment between the target and the template 1A2O chain A revealed a gap of about 9 residues in a loop region, which is known to connect the N-terminal regulatory domain with the C-terminal methylesterase domain (Figure

2.3). The templates and target share ~39% sequence identity. The stereochemical quality of the template was assessed with PROCHECK and it was found that 88.6% residues were in most favoured regions, 9.4% were in additionally allowed regions, 1% were in generously allowed regions and 1% were in disallowed regions (Figure 2.8). The N-terminal domain of CheB shares homology to CheY, and various other bacterial response regulator proteins. Suitable templates for the interactions between CheB and the P1 and P2 domains of CheY were also identified, and will be discussed in the following section.

2.3.2.2. ***CheC***

The HHpred search algorithm returned only one possible template with a 100% probability of being a true homolog of the target sequence. The template comes from *T. maritima* CheC, two entries in the PDB are representative of this protein: 1XKR chain A and 2F9Z chain A. The structure 1XKR had a better resolution of the two and was selected as template (Table 2.1 and Table 2.2). The HHpred alignment between target and template revealed high similarity in secondary structure with 2 gaps of 4 residues in total in the alignment (Figure 2.4). The template shares ~30% sequence identity with the target. Other possible templates as predicted by HHpred had a probability score between 20.2- 99.7%. CheC shares homology with CheX (Park *et al.* 2004), which is a member of the CheC family of CheY-phosphatases, but is absent in the *B. subtilis* group. Other homologs include the bacterial flagellar motor switch protein FliM (Park *et al.* 2006), Cholesteryl ester transferase inhibitor protein found in baboons (Buchko *et al.* 2000) and the antimicrobial peptide Carnocin CP52 (Sprules *et al.* 2004). Suitable templates for the interaction between CheC and CheY, and CheC and CheD were also identified, which will be discussed later on in this chapter. A Ramachandran plot of the template was generated with PROCHECK (Figure 2.8) to determine the stereochemical quality of the structure. 95% of residues were found in most favoured regions, 4.4% in additional allowed regions and 0.6% in generously allowed regions.

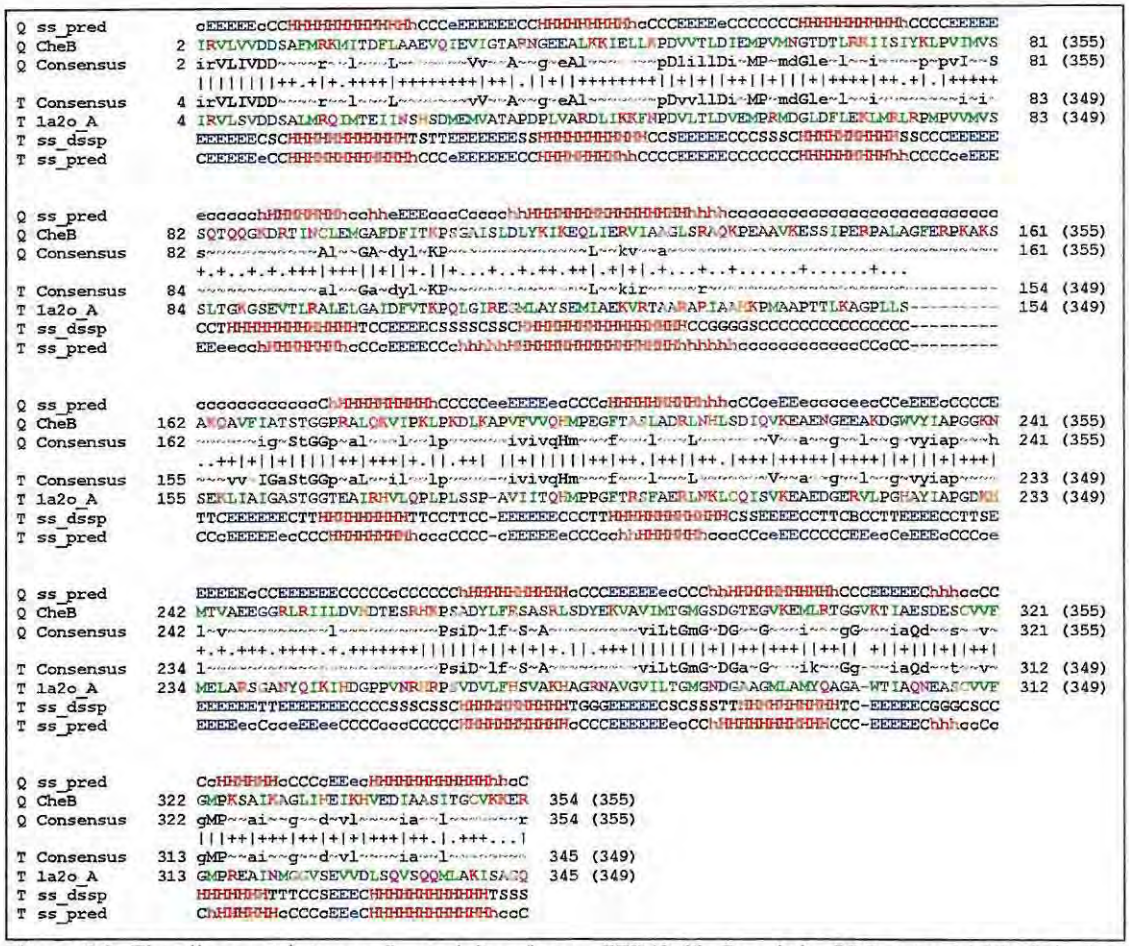


Figure 2.3. The alignment between *B. amyloliquefaciens* FZB42 CheB and the PDB entry 1A2O chain A as produced by HHpred. The predicted secondary structure for the target (Q ss_pred) and template (T ss_pred) as well as the actual secondary structure of the template as predicted by DSSP (T ss_dssp) are shown. Upper and lower case amino acids in the consensus sequences indicate high (60%) and moderate (40%) conservation, respectively. Symbols indicating the quality of the column–column match: '|' very good, '+' good, '.' neutral, '–' bad and '=' very bad.

2.3.2.3. *CheD*

The HHpred search results for possible templates *of B. amyloliquefaciens* FZB42 protein CheD returned only one result with a 100% probability of being a homolog: The structure for *T. maritima* CheD (PDB ID: 2F9Z chain C) (Table 2.1. and Table 2.2. The HHpred alignment between the target and the selected template revealed quite similar secondary structures, with only 3 gaps of 12 residues in total (Figure 2.5); furthermore the target and template share ~42% sequence identity. PROCHECK analysis of the template revealed that 82% of the residues fell into most favoured regions while 18% were in additional allowed regions (Figure 2.8.). It should

be noted that the template comes from a complex between CheC and CheD (from the organism *T. maritima*) and this interaction may have caused some distortion.

2.3.2.4. *CheR*

Only one possible template with a 100% probability of being a homolog to *B. amyloliquefaciens* FZB42 CheR was revealed by an HHpred search (Table 2.1 and Table 2.2). The template sequence comes from *S. typhimurium* CheR and is represented by two entries in the PDB: 1AF7 chain A (Djordjevic & Stock 1997) and 1BC5 chain A (Djordjevic & Stock 1998), the structure 1AF7 had the better resolution of the two. The alignment between the target and template reveals 6 gaps with a total of 13 missing residues (Figure 2.6). The target and template share a sequence identity of 29%. The target sequence also shares homology with a variety of other methyltransferase proteins. PROCHECK analysis revealed that 91.5% of the residues are in most favoured regions, 7.7% are in additional allowed regions and 0.8% is in generously allowed regions (Figure 2.8).

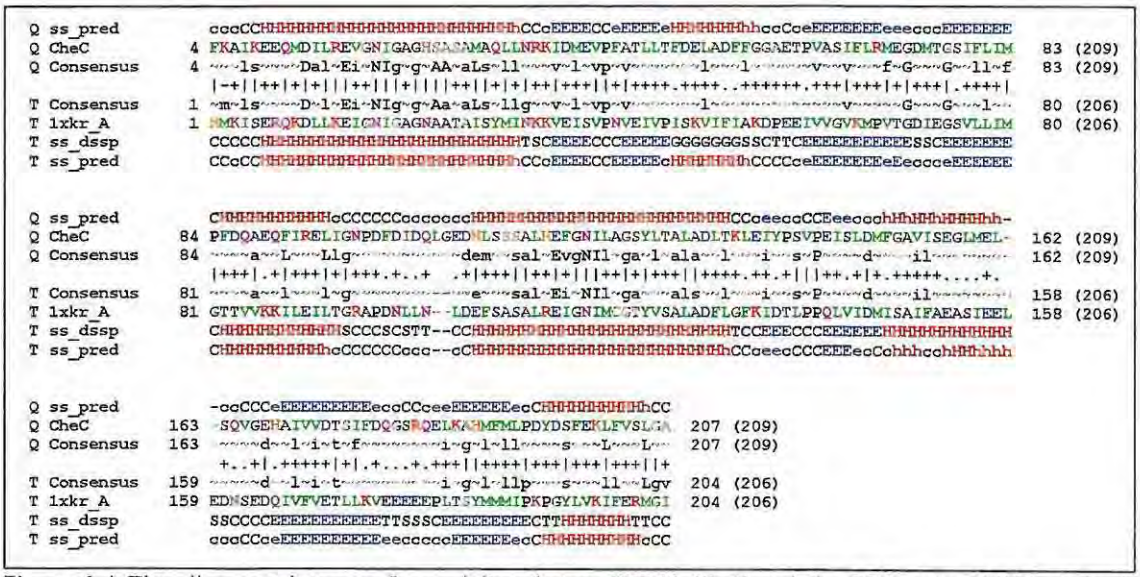


Figure 2.4 The alignment between *B. amyloliquefaciens* FZB42 CheC and the PDB entry 1XKR chain A as produced by HHpred. The predicted secondary structure for the target (Q ss_pred) and template (T ss_pred) as well as the actual secondary structure of the template as predicted by DSSP (T ss_dssp) are shown. Upper and lower case amino acids in the consensus sequences indicate high (60%) and moderate (40%) conservation, respectively. Symbols indicating the quality of the column-column match: '' very good, '+' good, '.' neutral, '-' bad and '=' very bad.

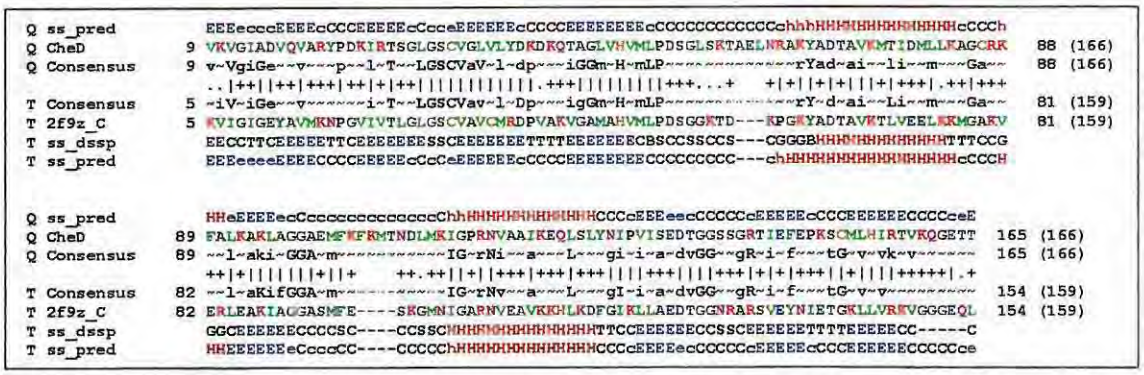


Figure 2.5. The alignment between *B. amyloliquefaciens* FZB42 CheD and the PDB entry 2F9Z chain C as produced by HHpred. The predicted secondary structure for the target (Q ss_pred) and template (T ss_pred) as well as the actual secondary structure of the template as predicted by DSSP (T ss_dssp) are shown. Upper and lower case amino acids in the consensus sequences indicate high (60%) and moderate (40%) conservation, respectively. Symbols indicating the quality of the column-column match: '|' very good, '+' good, '.' neutral, '-' bad and '=' very bad

2.3.2.5. *CheW*

Three possible templates with a 100% probability of being homologs of *B. amyloliquefaciens* FZB42 were identified by an HHpred search. These templates are all CheW proteins from the organisms *E. coli* (Li *et al.* 2007), *Thermoanaerobacter tengcongensis* (Yao *et al.* 2007) and *T. maritima* (Griswold *et al.* 2002, Park *et al.* 2006). The *T. tengcongensis* structure (PDB ID: 2QDL chain A) shares ~30% sequence identity with the target (Table 2.1. and Table 2.2), and the HHpred alignment between the two sequences revealed only one gap (Figure 2.7). The geometry of the selected template was assessed with PROCHECK and it was found that 85.6% of the residues fell within most favoured regions, 11.4% in additional allowed regions and 3.0% in generously allowed regions. No residues were located in disallowed regions (Figure 2.12). An appropriate template for the interaction between CheW and the P4 and P5 domains of CheA was identified and will be discussed in a later section of this chapter.

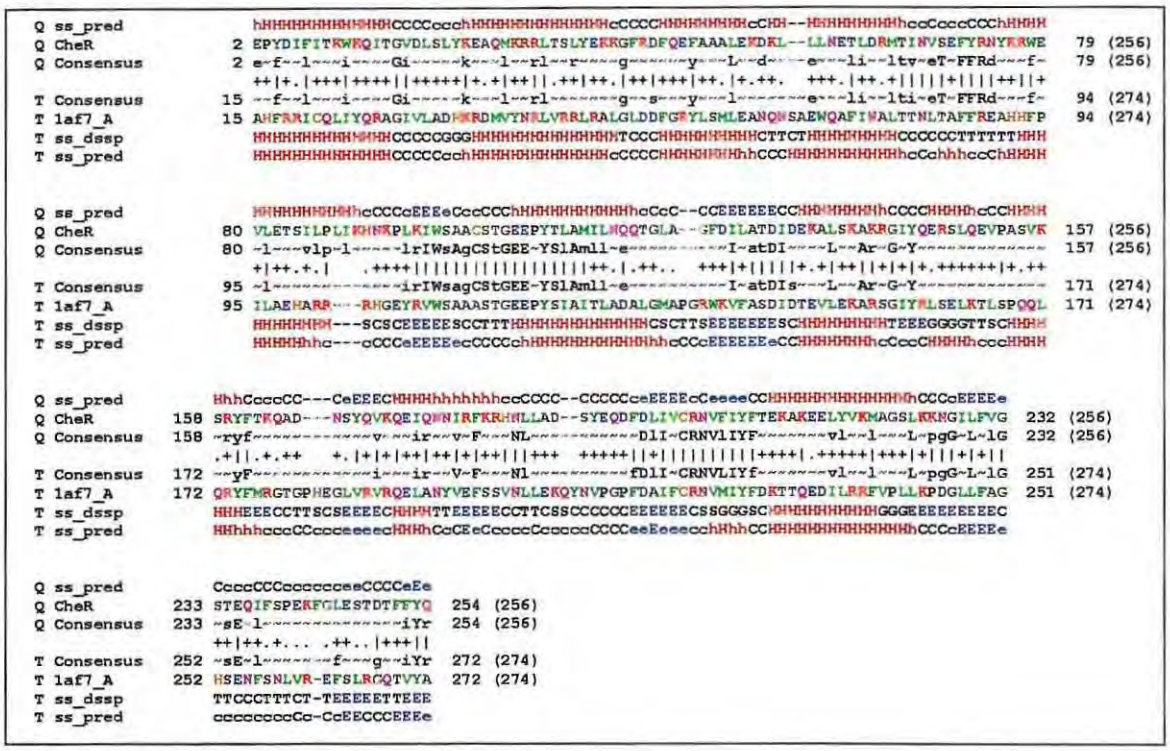


Figure 2.6. The alignment between *B. amyloliquefaciens* FZB42 CheR and the PDB entry 1AF7 chain A as produced by HHpred. The predicted secondary structure for the target (Q ss_pred) and template (T ss_pred) as well as the actual secondary structure of the template as predicted by DSSP (T ss_dssp) are shown. Upper and lower case amino acids in the consensus sequences indicate high (60%) and moderate (40%) conservation, respectively. Symbols indicating the quality of the column–column match: '|' very good, '+' good, '.' neutral, '–' bad and '=' very bad

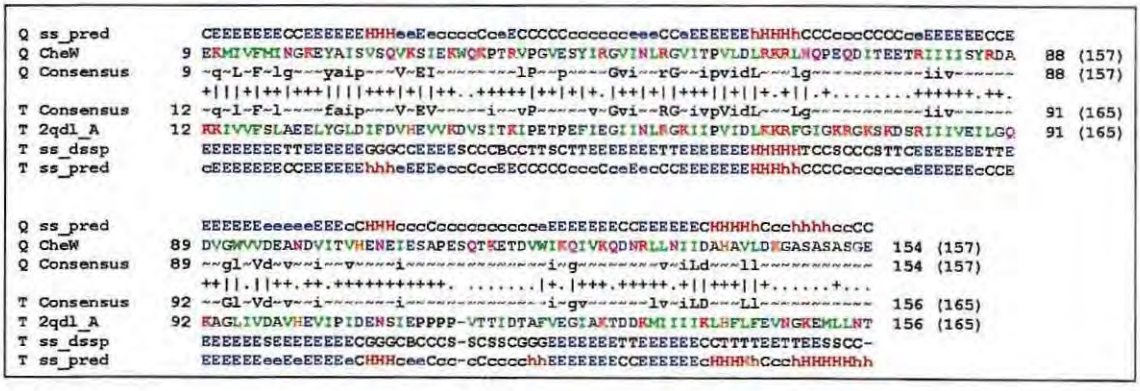


Figure 2.7. The alignment between *B. amyloliquefaciens* FZB42 CheW and the PDB entry 2QDL chain A as produced by HHpred. The predicted secondary structure for the target (Q ss_pred) and template (T ss_pred) as well as the actual secondary structure of the template as predicted by DSSP (T ss_dssp) are shown. Upper and lower case amino acids in the consensus sequences indicate high (60%) and moderate (40%) conservation, respectively. Symbols indicating the quality of the column–column match: '|' very good, '+' good, '.' neutral, '–' bad and '=' very bad.

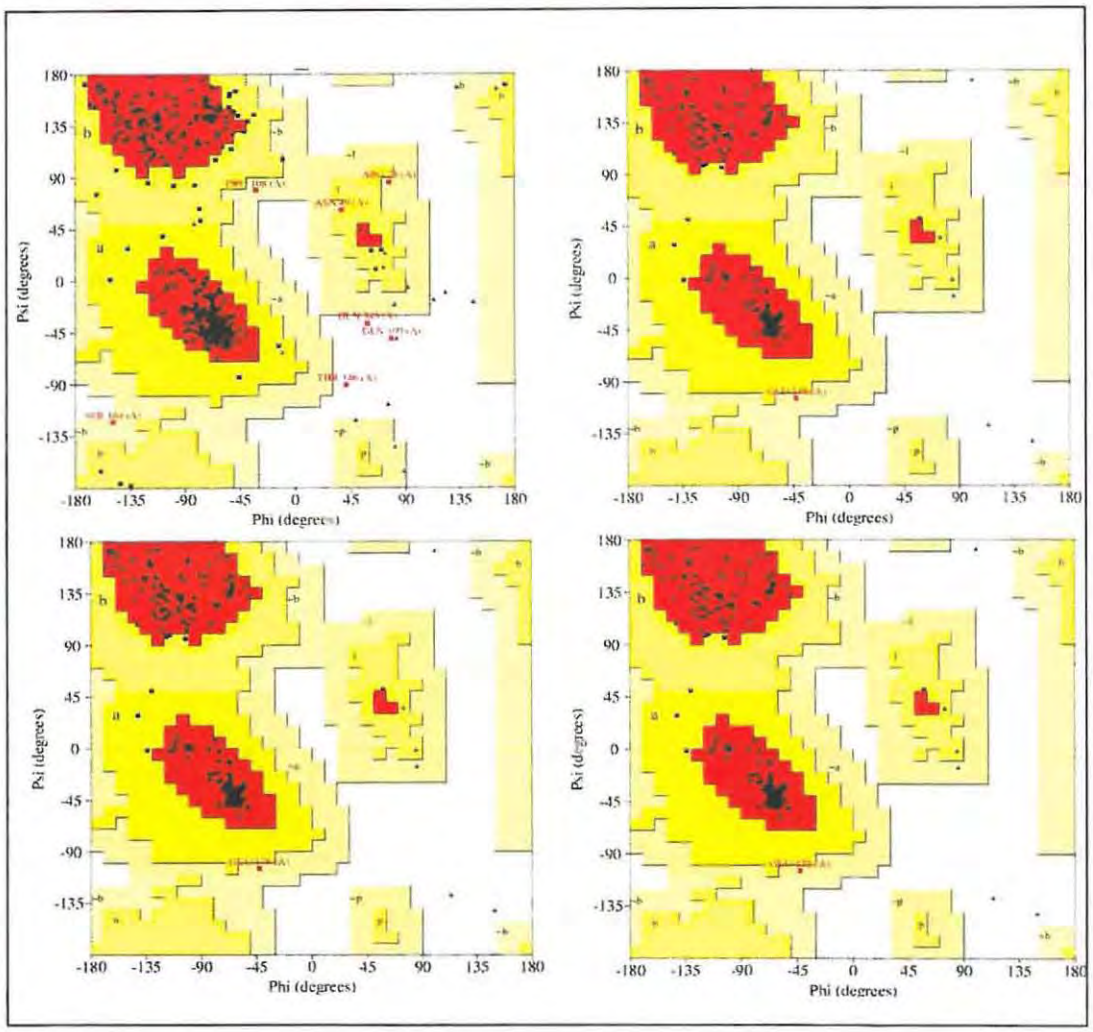


Figure 2.8. A Ramachandran plot of the structures: 1A20 chain A (top left); 1XKR chain A (top right) 2F9Z chain C (bottom left) and 1AF7 (bottom right). Most sterically favored regions (red), additional allowed regions (dark yellow), generously allowed regions (light yellow), disallowed regions (white) α-helix (A), β-sheet (B), left-handed-helix (L).

2.3.2.6. *CheY*

The HHpred search for possible templates for *B. amyloliquefaciens* FZB42 CheY produced almost ninety different structures with a 100% probability score. The top ranked template was the *T. maritima* CheY protein with five representatives in the PDB. Four of these five structures (1TMY -4TMY) represent different apo-forms of Mg^{2+} bound and free CheY (Usher *et al.* 1998). The fifth structure (PDB ID 1U0S chain Y) is part of an interaction between CheY and the P2 domain of CheA (Park *et al.* 2004). The template sequence shares a 71% sequence identity with

the target (Table 2.1 and Table 2.2). The alignment between the target and template showed a high degree of similarity between the predicted secondary structure of the target and the actual and predicted secondary structure of the template, with no gaps and high sequence conservation (Figure 2.9). The template was evaluated with PROCHECK to determine if any regions exhibit unusual geometry (Figure 2.12). No residues were found to be in disallowed or generously allowed regions, 92.4% were found in most favoured regions and 7.6% in additional allowed regions.

2.3.2.7. *CheA domains P1 and P2*

The HHpred search did not return any templates that could cover the entire length of the large multimeric CheA protein. The PDB entries 3KYJ chain A (Bell *et al.* 2010) and 1U0S chain A, were selected as templates for modelling the P1 and P2 domains respectively (Table 2.1 and Table 2.2). A model of both domains, connected by a loop, was based on these two templates. The structure 3KYJ, which shares ~21% sequence identity with the target, is representative of the interaction between the $CheA_3P1$ domain with $CheY_6$ from *R. sphaeroides* and was also used as a template to model CheAP1 interacting with CheY and CheB in *B. amyloliquefaciens* FZB42. The template 1U0S is representative of an interaction between CheY and CheAP2 from *T. maritima* and was also used as a template to model the interaction between the homologous proteins from the target organism. Chain A of 1U0S shares ~35% sequence identity with the target. The alignments between the two templates and the appropriate sections of the target protein reveal some inconsistencies between the predicted secondary structure characteristics of the target and the predicted as well as actual secondary structure of the templates (Figure 2.10 and Figure 2.11). Therefore these regions may be problematic for homology modelling. A suitable template for the interaction between CheAP4P5 and CheW was also identified, and will be discussed later on in this chapter. The Ramachandran dihedral statistics for both templates were good: 1U0S chain A: 91.4% most favoured, 7.4% additionally allowed, 1.2% generously allowed; and 3KYJ chain A: 95.7% most favoured, 4.3% generously allowed (Figure 2.12).

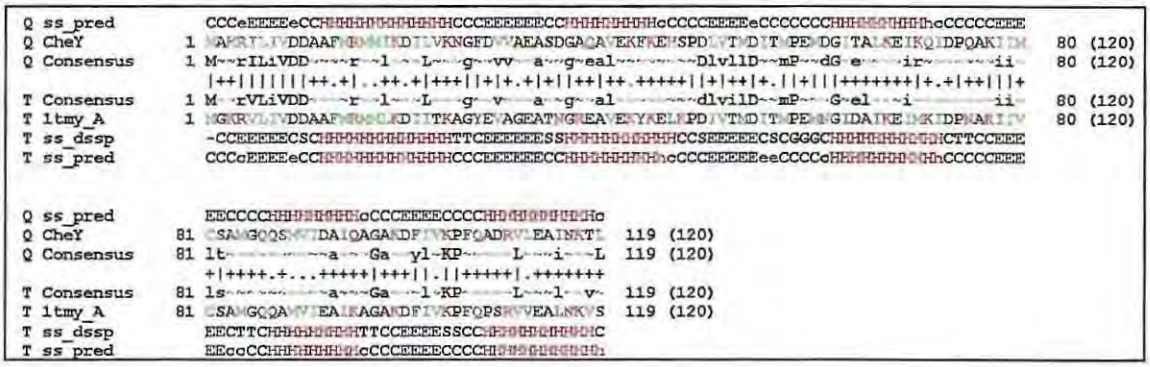


Figure 2.9. The alignment between *B. amyloliquefaciens* FZB42 CheY and the PDB entry 1TMY chain A as produced by HHpred. The predicted secondary structure for the target (Q ss_pred) and template (T ss_pred) as well as the actual secondary structure of the template as predicted by DSSP (T ss_dssp) are shown. Upper and lower case amino acids in the consensus sequences indicate high (60%) and moderate (40%) conservation, respectively. Symbols indicating the quality of the column–column match: '|' very good, '+' good, '.' neutral, '–' bad and '=' very bad.

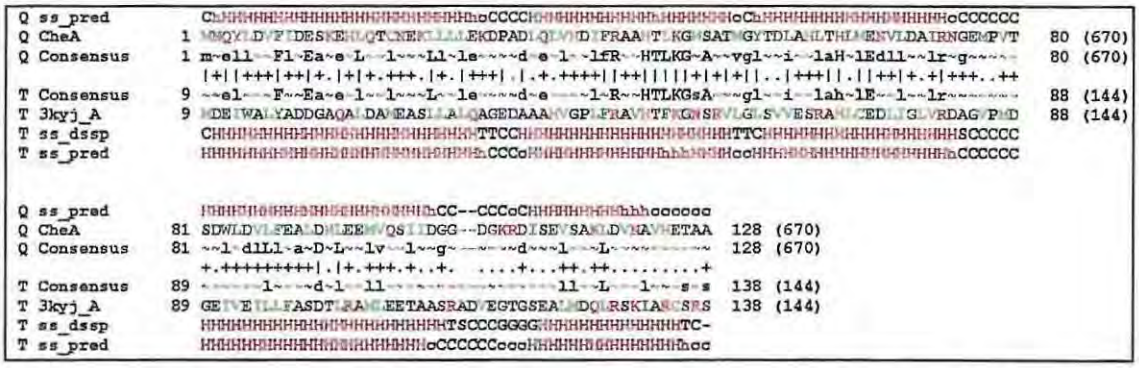


Figure 2.10. The alignment between B. amyloliquefaciens FZB42 CheAP1 and the PDB entry 1B3Q chain A as produced by HHpred. The predicted secondary structure for the target (Q ss_pred) and template (T ss_pred) as well as the actual secondary structure of the template as predicted by DSSP (T ss_dssp) are shown. Upper and lower case amino acids in the consensus sequences indicate high (60%) and moderate (40%) conservation, respectively. Symbols indicating the quality of the column–column match: '|' very good, '+' good, '.' neutral, '–' bad and '=' very bad.

2.3.2.8. *CheV*

The CheV protein has two domains: a C-terminal CheY like domain that can accept a phosphate group from CheA, and a CheW like N-terminal domain. The HHpred search results returned 4 matches for the N-terminal domain of CheV, four of which came from solved structures of CheW and one experimental representative of the regulatory domain of CheA. About 95 matches for the C-terminal domain were returned, representing a variety of response regulators found in

bacteria. However none of them had a sequence identity of more than 30%, therefore no attempt was made to build a homology model for CheV.

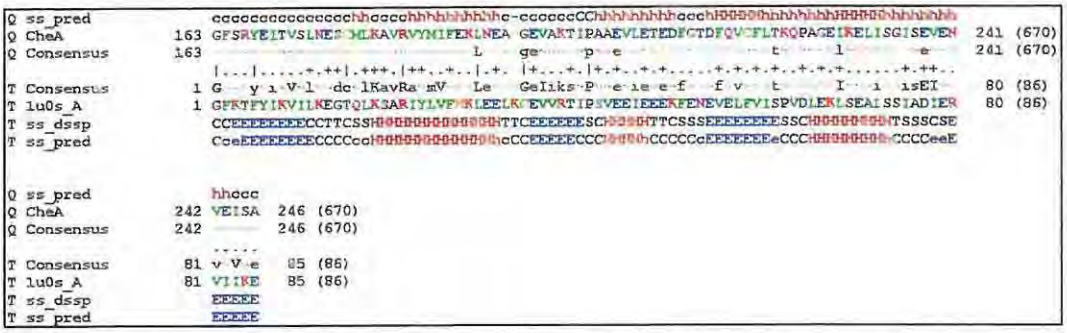


Figure 2.11. The alignment between *B. amyloliquefaciens* FZB42 CheAP2 and the PDB entry 1U0S chain A as produced by HHpred. The predicted secondary structure for the target (Q ss_pred) and template (T ss_pred) as well as the actual secondary structure of the template as predicted by DSSP (T ss_dssp) are shown. Upper and lower case amino acids in the consensus sequences indicate high (60%) and moderate (40%) conservation, respectively. Symbols indicating the quality of the column-column match: '|' very good, '+' good, '.' neutral, '-' bad and '=' very bad.

2.3.2.9. *Interaction between CheC and CheD*

The complex between the chemotaxis deamidase CheD and the chemotaxis phosphatase CheC from *T. maritima* (PDB ID 2F9Z) (Chao *et al.* 2006) was chosen as a template to model the similar interaction between the homologous *B. amyloliquefaciens* FZB42 proteins (Table2.1 and Table2.2). The alignment between CheD and 2F9Z chain C was discussed in a previous section (Figure 2.5). The alignment between CheC and 2F9Z chain A is the same as that of CheC with 1XKR chain A, since they represent the same protein from *T. maritima*, except that three N-terminal residues are absent in 2F9Z chain A. When more than one structure of a protein from the same organism is present, HHpred only shows the alignment with the structure that has the best resolution. Thus, in the case of the interaction between CheC-CheD a target-template alignment generated by Promals3D was used. The stereochemical quality of the template was checked with PROCHECK and it was found that 82.7% of residues are in most favoured regions and 17.3% are in additional allowed regions (Figure 2.17). According to the HHpred results chain A of the template shares ~30% sequence identity with CheC, and chain C shares ~40% identity with CheD. The template structure was determined with X-ray diffraction at 2.4Å.

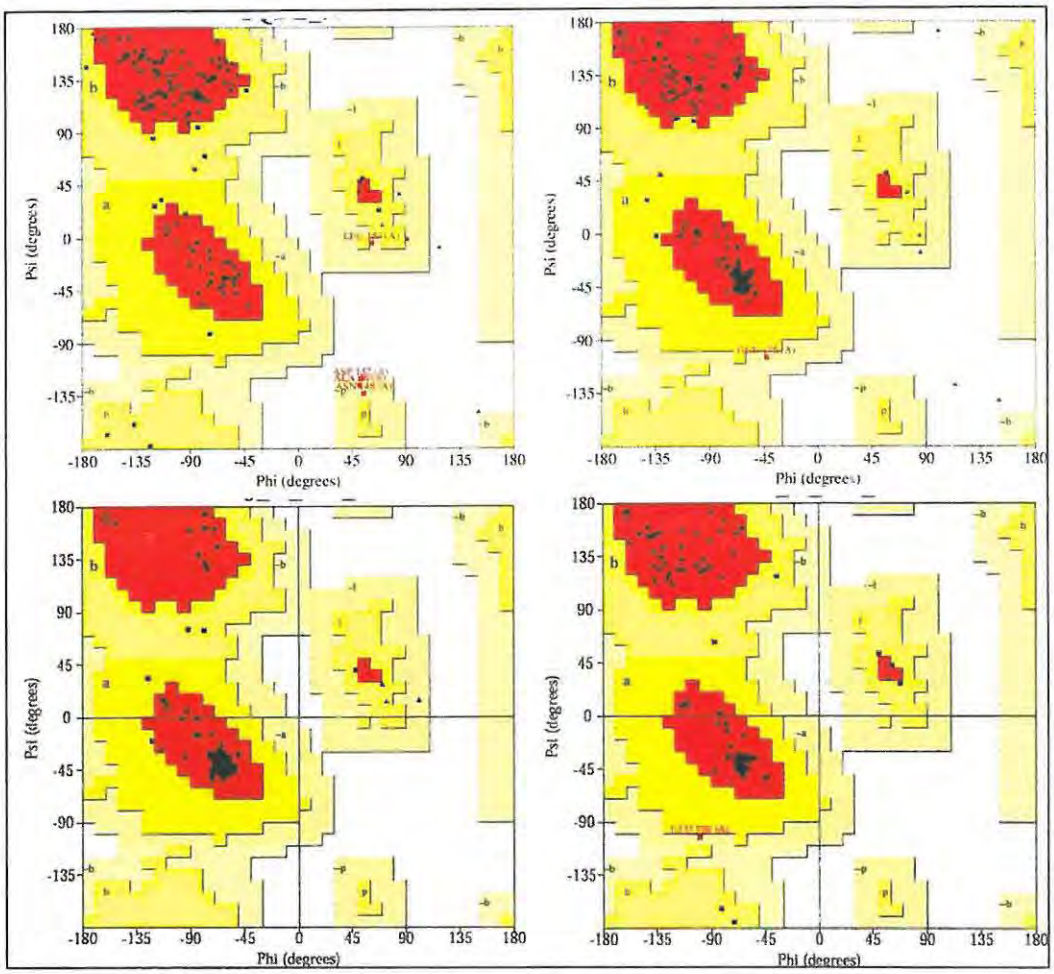


Figure 2.12. A Ramachandran plot of the structures: 2QDL chain A (top left); 1TMY (top right); 3KYJ chain A (bottom left) and 1U0S chain A(bottom right) . Most sterically favored regions (red), additional allowed regions (dark yellow), generously allowed regions (light yellow), disallowed regions (white) α-helix (A), β-sheet (B), left-handed-helix (L).

2.3.2.10. *Interaction between CheA domains P4 and P5 with CheW*

A suitable template for the interaction between CheA domains P4 and P5 with CheW was downloaded from the PDB (Table 2.1 and Table 2.2). The template (PDB ID: 2CH4 chains A and W) (Park *et al.* 2006) from the organism *T. maritima* was determined at 3.5Å. Promals3D was used to generate the target-template alignments. Sequence identity between the target and template sequences were calculated using BioEdit (Hall 1999). The sequence identity between chain A of the template and CheA was ~47%; and ~29% between chain W of the template and the target CheW. The Ramachandran dihedral statistics of template, as determined by PROCHECK, revealed the following about the location of the residues: 71.2% in most favoured

regions, 27.3% in additional allowed regions, 1.0% in generously allowed regions and 0.5% in disallowed regions (Figure 2.17). According to these results the stereochemistry of the structure was less than ideal, but due to the lack of availability of better templates, homology modelling of CheW was performed using the structure 2CH4 as template.

2.3.2.11. *Interaction between CheY and the P1 domain of CheA*

The 1.40Å resolution crystal structure of the histidine containing phosphotransfer domain (P1) of $CheA_3$ in complex with $CheY_6$ from *R. sphaeroides* (PDB ID: 3KYJ) (Bell *et al.* 2010) was selected as a template to model the similar interaction between *B. amyloliquefaciens* FZB42 CheAP1 and CheY (Table 2.1). The chains in the template that are homologous to the target sequences share ~21% and ~29% sequence identity to CheA and CheY respectively. The target template alignments produced by HHpred revealed a gap area larger than 10 residues between CheY and 3KYJ chain B, but only 2 gaps between CheAP1 and 3KYJ chain A (Figure 2.14 and Figure 2.15). The PROCHECK evaluation revealed that the template contains 95.1% of residues in most favoured regions with the remaining 4.9% in additional allowed regions (Figure 2.17). These results affirm that the template has valid stereochemistry. Due to the low sequence identity between the template and CheAP1 as well as the gap in the alignment with CheY some difficulties with homology modelling were anticipated.

2.3.2.12. *Interaction between CheB and the P1 domain of CheA*

A model for the interaction between *B. amyloliquefaciens* FZB42 CheB and the P1 domain of CheA was based on the templates 3KYJ chains A and B, and 1A20 chain A. 3KYJ chain B shares a ~34% sequence identity with CheB. Since 3KYJ chain B only covers the N-terminal domain of CheB (Figure 2.16), the structure 1A2O was also included in the alignment to enable derivation of homology constraints for the full length CheB protein. Other template properties and PROCHECK evaluation results have been discussed in previous sections.

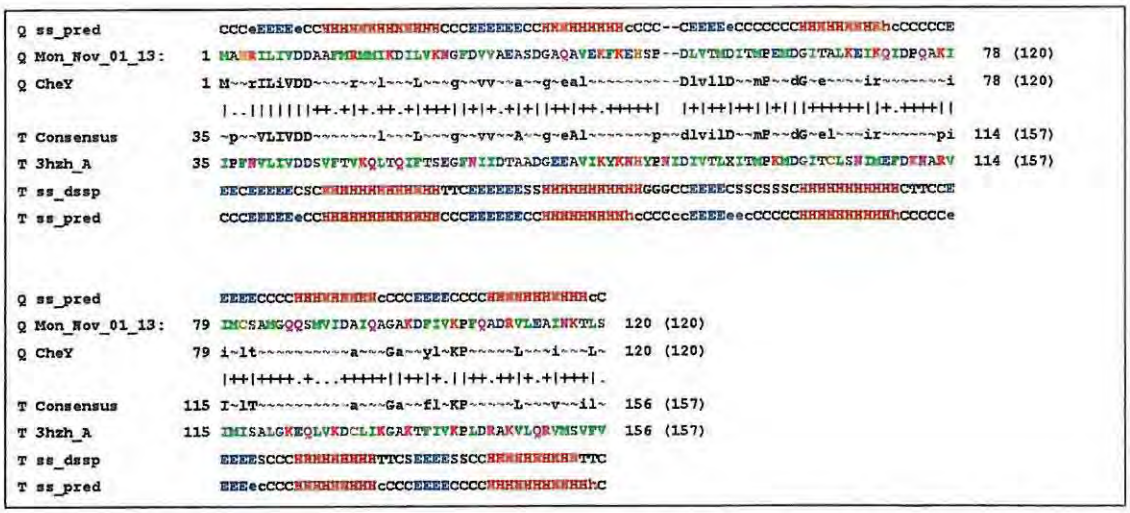


Figure 2.13. The alignment between *B. amyloliquefaciens* FZB42 CheY and the PDB entry 3HZH chain A as produced by HHpred. The predicted secondary structure for the target (Q ss_pred) and template (T ss_pred) as well as the actual secondary structure of the template as predicted by DSSP (T ss_dssp) are shown. Upper and lower case amino acids in the consensus sequences indicate high (60%) and moderate (40%) conservation, respectively. Symbols indicating the quality of the column–column match: '|' very good, '+' good, '.' neutral, '–' bad and '=' very bad.

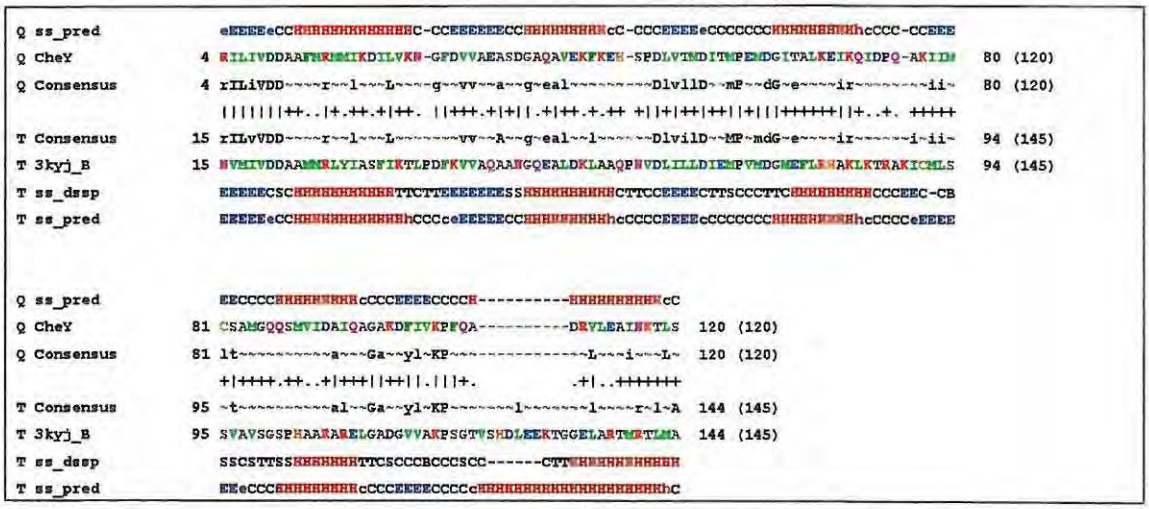


Figure 2.14. The alignment between *B. amyloliquefaciens* FZB42 CheY and the PDB entry 3KYJ chain B as produced by HHpred. The predicted secondary structure for the target (Q ss_pred) and template (T ss_pred) as well as the actual secondary structure of the template as predicted by DSSP (T ss_dssp) are shown. Upper and lower case amino acids in the consensus sequences indicate high (60%) and moderate (40%) conservation, respectively. Symbols indicating the quality of the column–column match: '|' very good, '+' good, '.' neutral, '–' bad and '=' very bad.

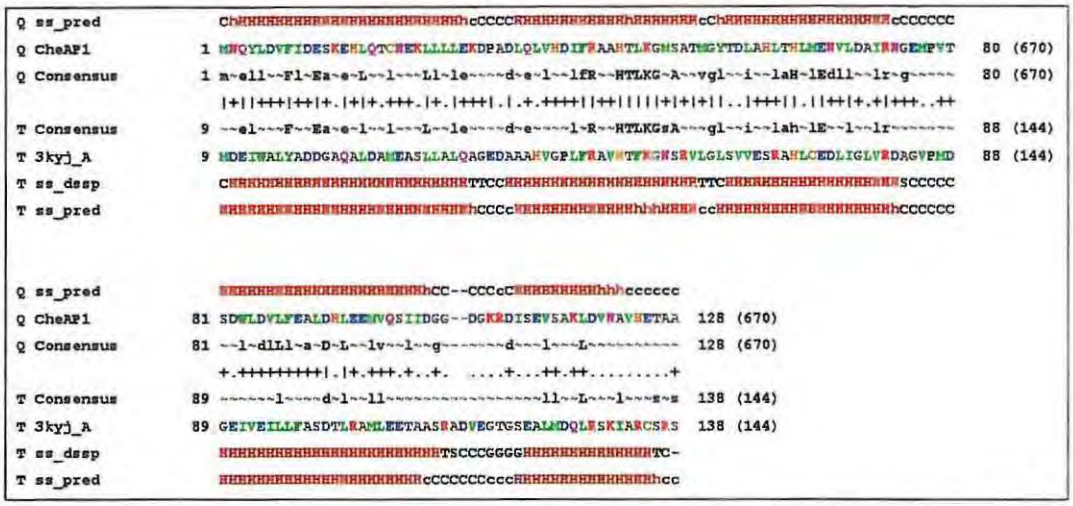


Figure 2.15. The alignment between *B. amyloliquefaciens* FZB42 CheA and the PDB entry 3KYJ chain A as produced by HHpred. The predicted secondary structure for the target (Q ss_pred) and template (T ss_pred) as well as the actual secondary structure of the template as predicted by DSSP (T ss_dssp) are shown. Upper and lower case amino acids in the consensus sequences indicate high (60%) and moderate (40%) conservation, respectively. Symbols indicating the quality of the column–column match: '|' very good, '+' good, '.' neutral, '–' bad and '=' very bad.

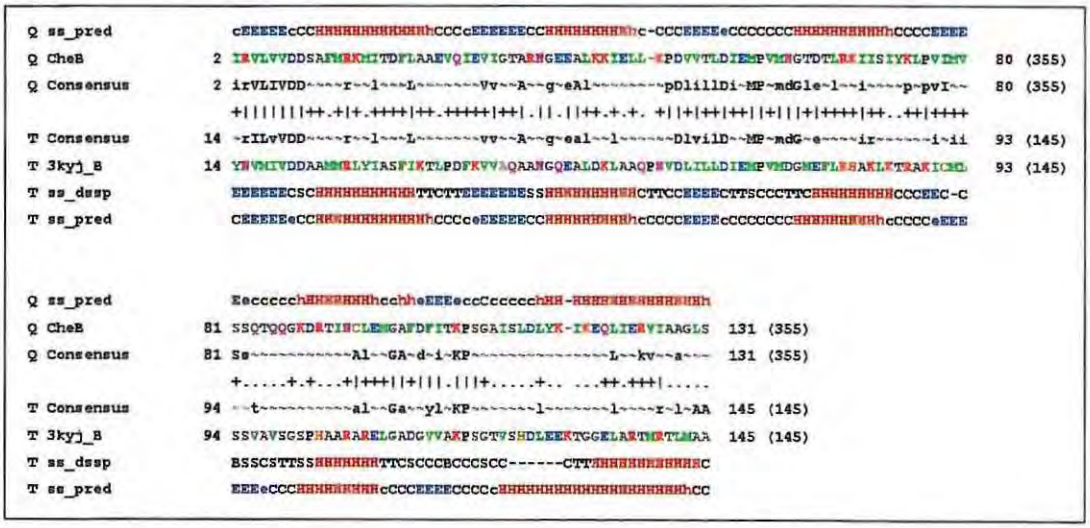


Figure 2.16. The alignment between *B. amyloliquefaciens* FZB42 CheB and the PDB entry 3KYJ chain B as produced by HHpred. The predicted secondary structure for the target (Q ss_pred) and template (T ss_pred) as well as the actual secondary structure of the template as predicted by DSSP (T ss_dssp) are shown. Upper and lower case amino acids in the consensus sequences indicate high (60%) and moderate (40%) conservation, respectively. Symbols indicating the quality of the column–column match: '|' very good, '+' good, '.' neutral, '–' bad and '=' very bad.

2.3.2.13. *Interaction between CheY and the P2 domain of CheA*

The HHpred searches for a suitable template for CheY also returned a possible template for modelling the interaction between CheY and the P2 domain of CheA. The template structure (PDB ID: 1U0S) comes from the organism *T. maritima* and was determined at a resolution of 1.9Å (Park *et al.* 2004). The amino acid sequence of the CheY chain of 1U0S is identical to the structure 1TMY, except that the first residue is missing in the structure 1U0S. Since the template-target alignments (Figure 2.11 and Figure 2.9) were already discussed in previous sections, they will be omitted here. There is a 71% identity between the target and template sequences for the CheY protein, and a 35% identity between those for the CheAP2 domain. The template showed good stereochemistry with 92.5% residues in most favoured regions, 7% in additional allowed regions and 0.5% in generously allowed regions. These results confirm the good geometry of the template (Figure 2.17)

2.3.2.14. *The interaction between $CheB_N$ and CheAP2*

The model for the interaction between the N-terminal domain of CheB and the P2 domain of CheY was based on the templates 1U0S (chains Y and A), and 1A2O (chain A), which have been discussed elsewhere. The Y chain of 1U0S comes from the same protein represented by the structure 1TMY. It should be noted that when there is more than one representative structure of the same protein from the same organism, HHpred only shows a target-template alignment with the template that has the best resolution (Figure 2.18). The sequence identities between target and templates were: ~43% between the $CheB_N$ and 1U0S chain Y, ~42% between $CheB_N$ and ~34% between 1U0S chain A and CheA. Although sequence identity was somewhat low, it fell well within the "safe zone" for homology modelling (Figure 2.1).

2.3.3. *Target-template alignment*

PROMALS3D was used to construct multiple sequence alignments between representative sequences from the *B. subtilis* group and appropriate templates. In each case the target and template sequences were copied from these MSAs and used as input for the modelling process. These alignments were compared to those produced by HHpred and in most cases were found to be almost identical. If an alignment did not result in a satisfactory model the alignment was improved by removing N-terminal and C-terminal sections. These loop regions often contain a

large number of gaps which makes modelling them problematic. The ".pir" alignments used as input for the models can be found on the disk containing the data of this research in the sub-directory Chapter2_data/pir_files. Each file-name contains the name of the protein or protein complex, followed by the names and chains of the templates used. For example the file "CheAP1_CheB_1A2O_A_3KYJ_AB.pir" contains the alignment between the targets CheA domain P1 and CheB and the templates 1A2O chain A and 3KYJ chains A and B.

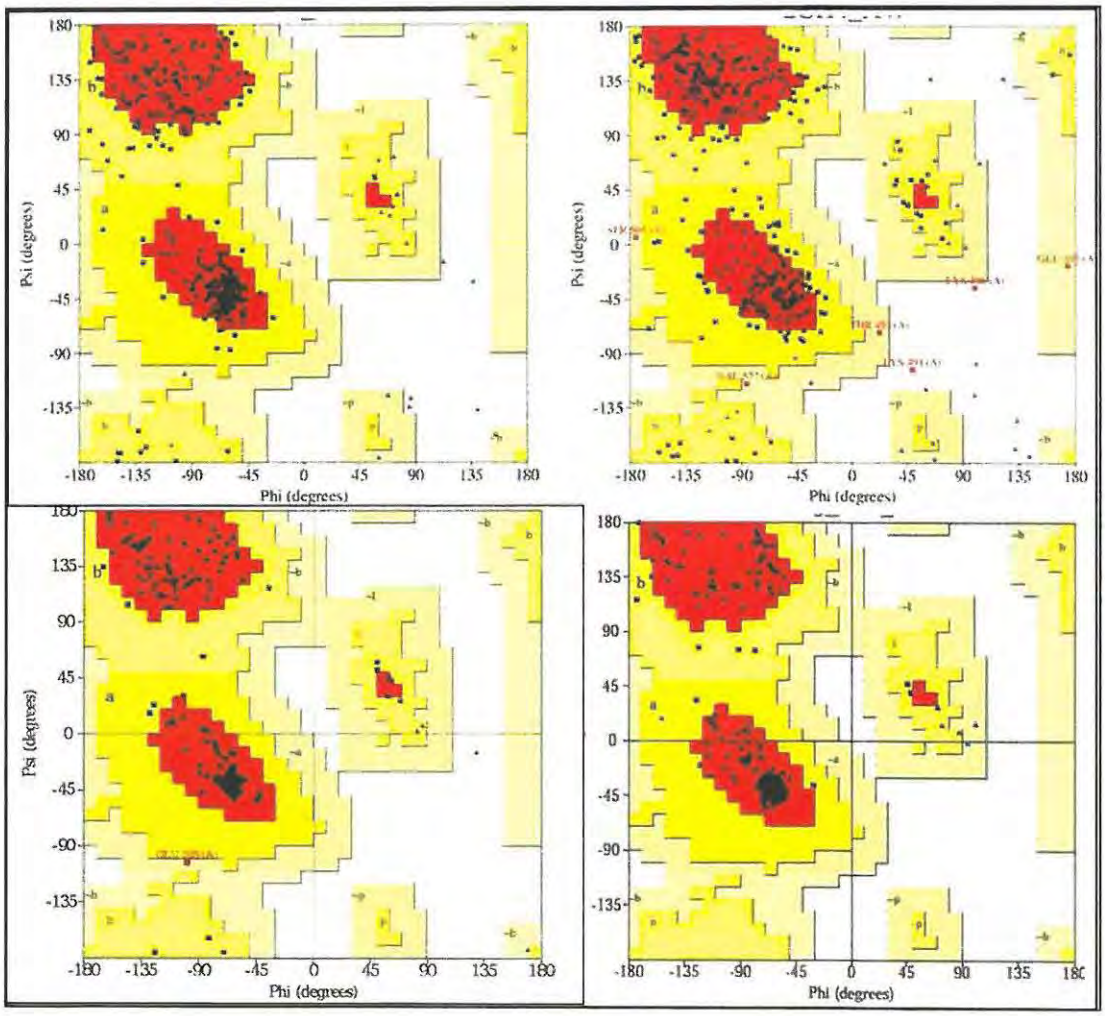


Figure 2.17. A Ramachandran plot of the structure 2F9Z chains A and C (top left), 2CH4 chains A and W (top right), 1U0S chains A and Y (bottom left) and 3KYJ chain A and B (bottom right). Most sterically favored regions (red), additional allowed regions (dark yellow), generously allowed regions (light yellow), disallowed regions (white). α-helix (A), β-sheet (B), left-handed-helix (L).

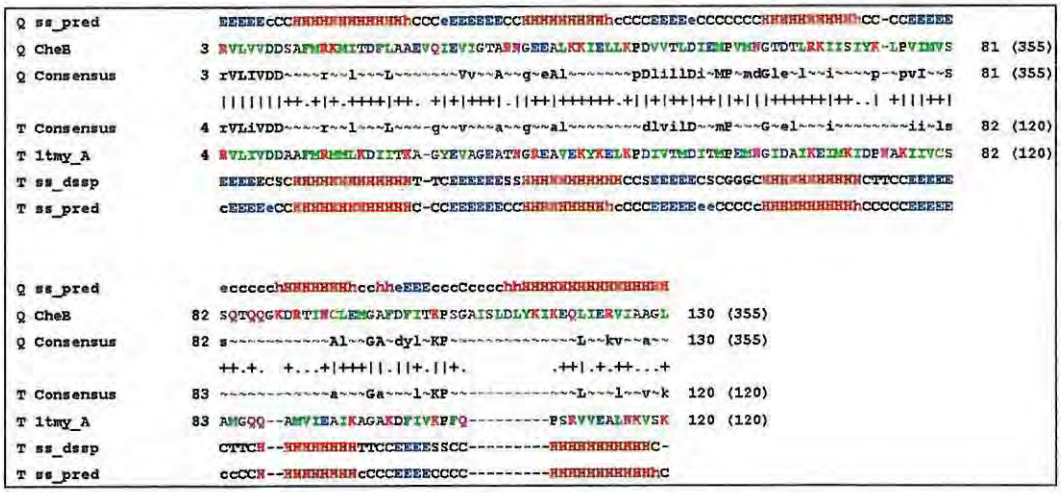


Figure 2.18. The alignment between *B. amyloliquefaciens* FZB42 CheB and the PDB entry 1TMY (which represents the same protein as 1U0S chain Y) as produced by HHpred. The predicted secondary structure for the target (Q ss_pred) and template (T ss_pred) as well as the actual secondary structure of the template as predicted by DSSP (T ss_dssp) are shown. Upper and lower case amino acids in the consensus sequences indicate high (60%) and moderate (40%) conservation, respectively. Symbols indicating the quality of the column–column match: '|' very good, '+' good, '.' neutral, '–' bad and '=' very bad.

2.3.4. *Homology modelling, loop refinement, model validation and model properties results*

The individual results for each single model as well as complex obtained by the protocol described in Section 2.3.3 are described in the sections that follow. The input and output for the various model quality assessment programs, as well as the final renumbered models, can be found on the accompanying disk under the folder Chapter 2, in the appropriately named subdirectories. For each template MODELLER generated 100 models which were evaluated on DOPE Z score. Models of individual proteins were evaluated with MetaMQAPii and ProQ, while complexes were evaluated with Verify3D, PROCHECK and ProQ. Loop refinement was performed when needed. Results from PIC can be found on the disk containing supplementary material.

2.3.4.1. *CheB*

The top ranked model had a DOPE Z score of -0.832. The model was submitted to MetaMQAPii to identify local areas of inaccuracy (Figure 2.19). MetaMQAPii returned a GDT_TS of 72.958. The predicted deviation from the unknown native structure had an RMSD of 2.453Å. The large flexible linker region (residues 130-163) that connects the N-terminal and C-terminal domains proved to be the single largest source of error, and moved substantially after loop refinement.

Loop refinement was performed on the following residue ranges: 131-164; 160-166; 255-266; 56-61; 279-281; and 225-229. After loop refinement the DOPE Z score improved to -1.079, which is indicative of a native-like structure. The final model was also submitted to MetaMQAPii, which returned a GDT_TS of 80.493 and an RMSD of 1.578Å. The evaluation results showed an improvement in overall model quality following loop-refinement (Figure 2.19). ProQ was also used to evaluate the model before and after refinement. The predicted LGscore before refinement was 5.293, which placed the model in the "extremely good model" category. After refinement the score improved further to 6.827. Based on these results, the model for CheB can be considered a good representative of the actual structure that can be used in docking studies, and can also serve as an adequate reference structure for studying positive and purifying selection operating on the protein in question. The active sites of CheB are listed in Table 2.3.

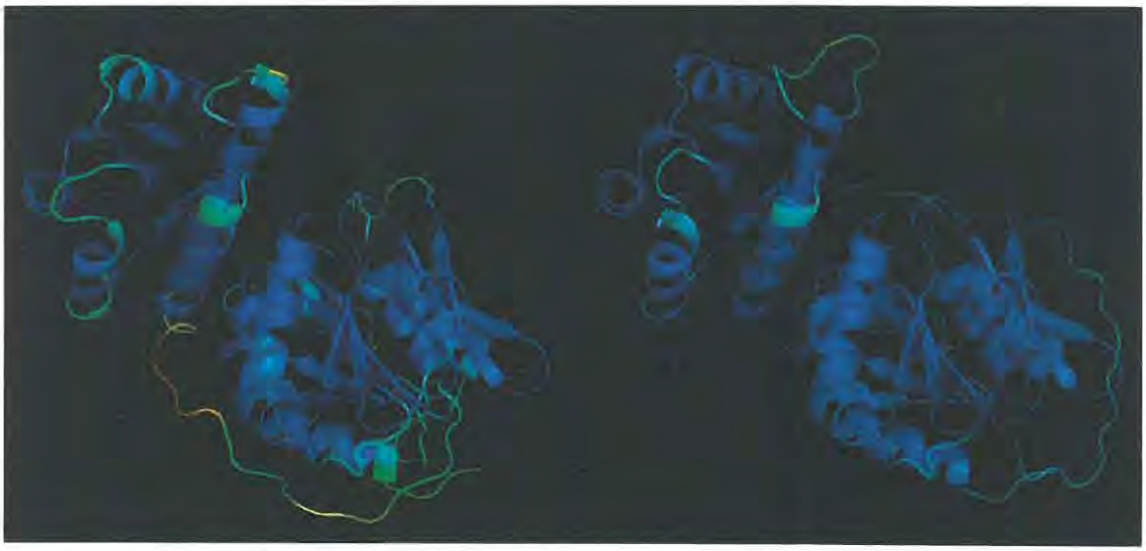


Figure 2.19. The initial (left) and final (right) models for CheB coloured according to MetaMQAPii scores. The spectrum of colours from blue to red indicates the spectrum of residues predicted to be correct to those predicted to be incorrect.

Table 2.3. A list of the active sites of CheB from the template organism *S. typhimurium* and the corresponding positions in *B. amyloliquefaciens* FZB42 and *B. subtilis subtilis* 168.

S. typhimurium	**Function**	***B. amyloliquefaciens* FZB42**	***B. subtilis subtilis* 168**
Ser-164, His-190, Asp-286	Catalytic triad with methylesterase activity (Krueger *et al.* 1992, West *et al.* 1995).	Ser-171, His-198, Asp-294	Ser-173, His-200 Asp-296
Asp-56	Acceptor site for Mg^{2} dependent phosphoryl transfer from the histidine kinase CheA. (Djordjevic *et al.* 1998).	Asp-54	Asp-54
Asp-10, Asp-11, and Asp-56, Glu-58	Acidic cluster where Mg^{2} binds (Djordjevic *et al.* 1998).	Asp-8, Asp-9, Asp-54, Glu-56	Asp-8, Asp-9, Asp-54, Glu-56

Residues indicated in **bold** and *italic* are not conserved throughout the members of the *B. subtilis* group.

2.3.4.2. *CheC*

The top ranked model had a DOPE Z score of -1.085 and a ProQ LGscore of 4.799 ("extremely good model"). Despite the good DOPE Z score and predicted LGscore, the MetaMQAPii analysis revealed inaccurately modelled loop regions. The most problematic loops were those of the N- and C-termini. The loop regions that were refined were: 1-8; 100-110; 180-184 and 204-209. After loop refinement the DOPE Z score improved to -1.490 and predicted LGscore improved to 6.446. The MetaMQAPii GDT_TS scores before and after refinement were 78.469 and 83.732 respectively. The predicted RMSD from the unknown native structure before loop refinement was 2.252Å, improving to 1.489Å after loop refinement. These results indicate that overall model quality showed a marked improvement after loop refinement. An inaccurately modelled region in the final model stretches from residues 99 – 110. The high variability in residue composition of the target and the template in this region made homology modelling difficult to achieve. Nevertheless the evaluation statistics are in line with those of high quality models that can be used in ligand docking studies. The model for CheC can be considered a reliable foundation for studying positive and purifying selection in terms of protein structure in accordance with the objectives of this work (Figure 2.20). The active sites of CheC are listed in Table 2.4.

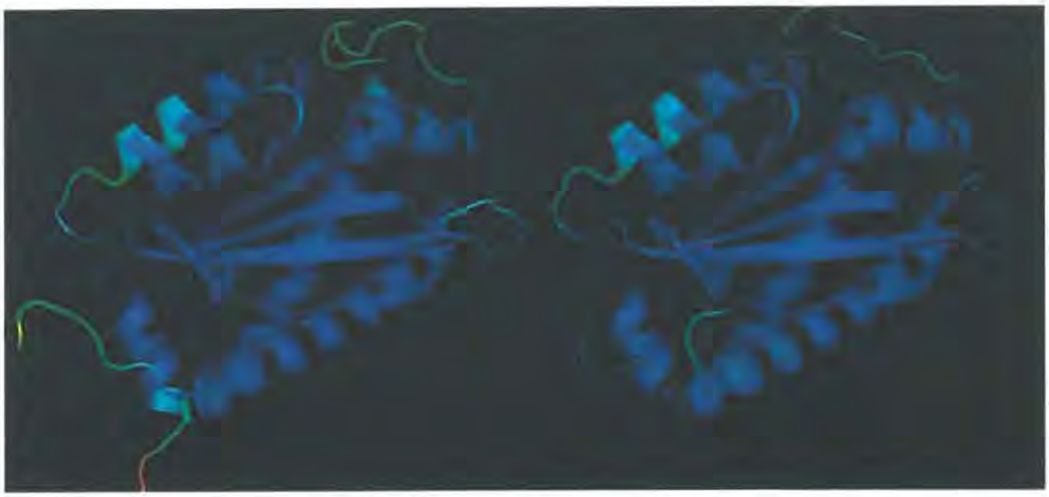


Figure 2.20. The initial (left) and final (right) models for CheC coloured according to MetaMQAPii scores. The spectrum of colours from blue to red indicates the spectrum of residues predicted to be correct to those predicted to be incorrect.

Table 2.4. A list of the active sites of CheC from the template organism *T. maritima* and the corresponding positions in *B. amyloliquefaciens* FZB42 and *B. subtilis subtilis* 168.

T. maritima	Function	*B. amyloliquefaciens* FZB42	*B. subtilis subtilis* 168
Glu-13 and Asn-16	First phosphatase active site (Park *et al.* 2004).	Glu-17 and Asn-20	Glu-17 and Asn-20 (Muff & Ordal 2007).
Glu-112 and Asn-115	Second phosphatase active site (Park *et al.* 2004).	Glu-118 and Asn-121	Glu-118 and Asn-121 (Muff & Ordal 2007).

2.3.4.3. *CheD*

The top ranked structure had a Z DOPE score of -1.002 and a ProQ predicted LGscore of 4.175 ("extremely good model"). MetaMQAPii identified several inaccurately modelled residues therefore, loop refinement was performed. Regions that were subjected to loop refinement were residues: 50-53; 93-101; 147-149; 75-77 and 49-52. After loop refinement the DOPE Z score improved to -1.281 and the ProQ predicted LGscore increased to 5.237. The MetaMQAPii GDT_TS scores before and after loop-refinement were 68.289 and 76.342 respectively. The RMSD from the predicted unknown native structure changed from 3.061Å to 2.044Å after loop refinement. The presence of gaps in the loop regions that stretch from residues 46-57 and 91-103 affected the derivation of homology constraints from the template in a negative manner. The final model (Figure 2.21) was renumbered using the ModifyPDB program from the ResDe4

package. The model for CheD can be considered as sufficiently reliable to provide a structural framework to study sequence-function relationships and site directed selection. The active sites of CheD are listed in Table 2.5.

Table 2.5. A list of the active sites of CheD from the template organism *T. maritima* and the corresponding positions in *B. amyloliquefaciens* FZB42 and *B. subtilis subtilis* 168.

T. maritima	Function	*B. amyloliquefaciens* FZB42	*B. subtilis subtilis* 168
Cys-27, His-44, Thr-21	Cysteine hydrolase active site cluster (Chao *et al.* 2006).	Cys-33, His-50, Thr-27	Cys-33, His-50, Thr-27

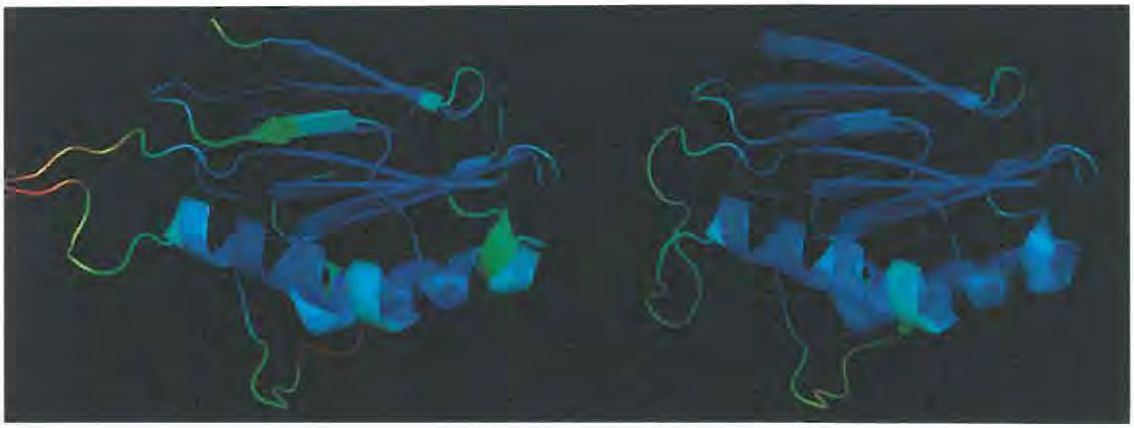


Figure 2.21. The initial (left) and final (right) models for CheD coloured according to MetaMQAPii scores. The spectrum of colours from blue to red indicates the spectrum of residues predicted to be correct to those predicted to be incorrect.

2.3.4.4. *CheR*

The top ranked model had a DOPE Z score of -1.150. The LGscore as predicted by ProQ was 4.536, which placed the model in the "extremely good" category. The MetaMQAPii evaluation of this structure returned a GDT_TS of 69.238, and the RMSD from the predicted unknown native structure was estimated to be 2.125Å. Areas with the most notable inaccuracies corresponded to regions with gaps between the target and template alignment. Regions subjected to loop refinement were residues: 87-95; 233-245; 39-41; 185-195 and 119-124. Loop refinement did somewhat relax the backbone conformation of these regions, and a slight improvement in overall model quality was achieved (Figure 2.22). The final model had a DOPE Z score of -1.470, a predicted LGscore of 5.665 and a GDT_TS of 78.613; and the RMSD from the native structure was predicted to be 1.559Å. These results suggest that the model of CheR is

of adequate quality to be used in future ligand docking studies, and more than adequate to serve as a 3D framework to study positive and purifying selection. Active site positions were determined by overlaying the model with the template structure and mapping experimentally determined active sites to the model (Table 2.6).

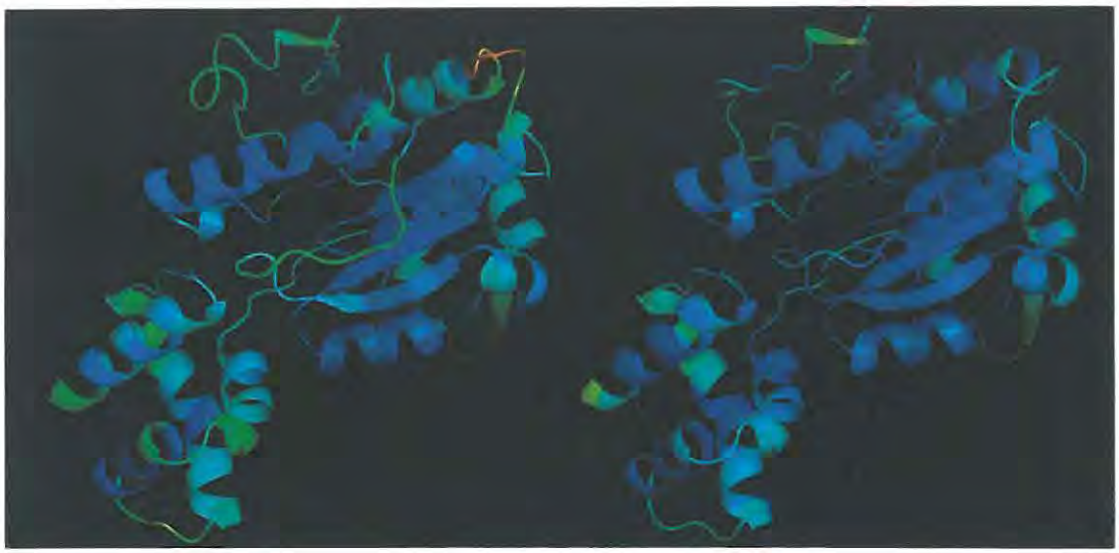


Figure 2.22. The initial (left) and final (right) models for CheR coloured according to MetaMQAPii scores. The spectrum of colours from blue to red indicates the spectrum of residues predicted to be correct to those predicted to be incorrect.

Table 2.6. A list of the active sites of CheR from the template organism *S. typhimurium* and the corresponding positions in *B. amyloliquefaciens* FZB42 and *B. subtilis subtilis* 168.

S. typhimurium	**Function**	***B. amyloliquefaciens* FZB42**	***B. subtilis subtilis* 168**
Ile-155 and Val-232	Aliphatic residues surrounding adenine ring.	Ile-131 and Val-203	Ile-131 and Val-203
Asp-154	Hydrogen bonds with the hydroxyl groups of the ribose ring.	Asp-130	Asp-130
Ala-38	Backbone carbonyl oxygen hydrogen bonds to amino group from adenine ring.	Thr-15	Thr-15
Asn-212	Side-chain carbonyl oxygen hydrogen bonds to amino group from adenine ring. (Djordjevic & Stock 1997)	Asn-185	Asn-185

2.3.4.5. *CheW*

The top ranked model had a DOPE Z score of -1.347. ProQ returned a predicted LGscore of 4.013 ("extremely good model"). However, evaluation of the model by MetaMQAPii showed

several inaccurately modelled regions. The MetaMQAPii predicted GDT_TS and RMSD for the model were 65.287 and 3.418Å, respectively. From the MetaMQAPii results it was evident that the core of the model was of good quality, but that the surface was problematic, thus further refinement was needed (Figure 2.23). The following residues were selected for loop refinement: 108-119; 69-76; 38-43; 1-8 and 147-149. Following loop refinement, the DOPE Z score of the final model improved to -1.668, the GDT_TS increased to 73.885, and the predicted RMDS from the unknown native structure was predicted to be 2.395Å. The predicted LGscore increased to 5.173. These results suggest that the predicted accuracy of the model may be too low for atomic-level detail analysis, but that the model is sufficient to function as a reliable framework for studying positive and purifying selection in terms of 3D structure. The sites important for CheW coupling to CheA are listed in Table 2.7.

Table 2.7. A list of the active sites of CheW from the template organism *T. tengcongensis* and the corresponding positions in *B. amyloliquefaciens* FZB42 and *B. subtilis subtilis* 168.

T. tengcongensis	**Function**	***B. amyloliquefaciens* FZB42**	***B. subtilis subtilis* 168**
Asn-55, Gly-58, Ile-60, Pro-62	CheA binding (Yao *et al.* 2007).	Asn-52, Gly55, ***Ile-57***, Pro-59	Asn-52, Gly55, ***Val-57***, Pro-59

Residues indicated in **bold** and *italic* are not conserved throughout the members of the *B. subtilis* group.

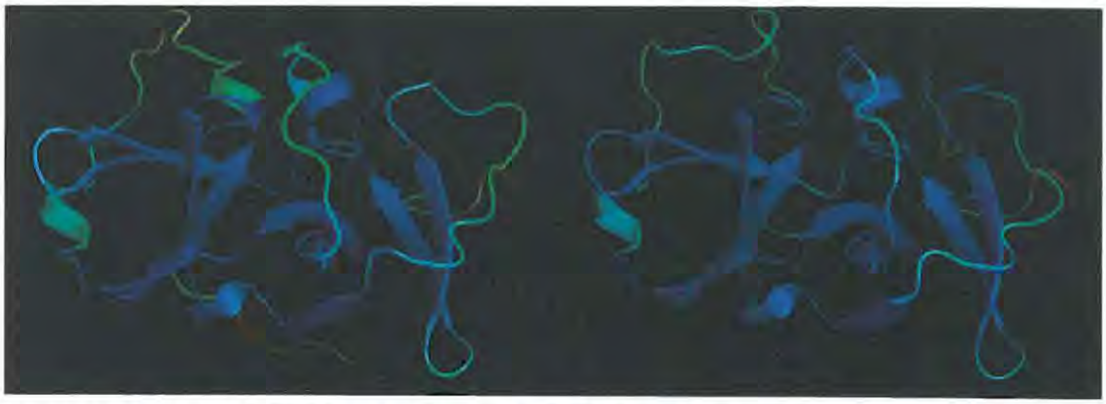


Figure 2.23. The initial (left) and final (right) models for CheW colour according to MetaMQAPii scores. The spectrum of colours from blue to red indicates the spectrum of residues predicted to be correct to those predicted to be incorrect.

2.3.4.6. ***CheY***

The top ranked model out of a hundred (with a DOPE Z score -1.836) was selected. ProQ predicted an LGscore of 4.299 ("extremely good model"). Evaluation by MetaMQAPii showed that most errors were located in loops and in a short helix-like structure (Figure 2.24). Loop refinement was performed on residues: 83-87; 57 and 61. The GDT_TS score before and after refinement was 82.917 and 84.792, respectively. The predicted RMSD from the unknown native structure improved from 1.294Å to 1.190Å. The predicted LGscore after refinement was 4.584. DOPE Z score after refinement was -1.879. These results suggest that our model of CheY is sufficiently reliable to be used in *ab initio* ligand docking studies in the future. Furthermore, it is a more than adequate framework to interpret sequence-function relationships and to study site directed positive and purifying selection in terms of 3D structure, in accordance to the objective of this study. The residues involved in phosphorylation of CheY are listed in (Table 2.8).

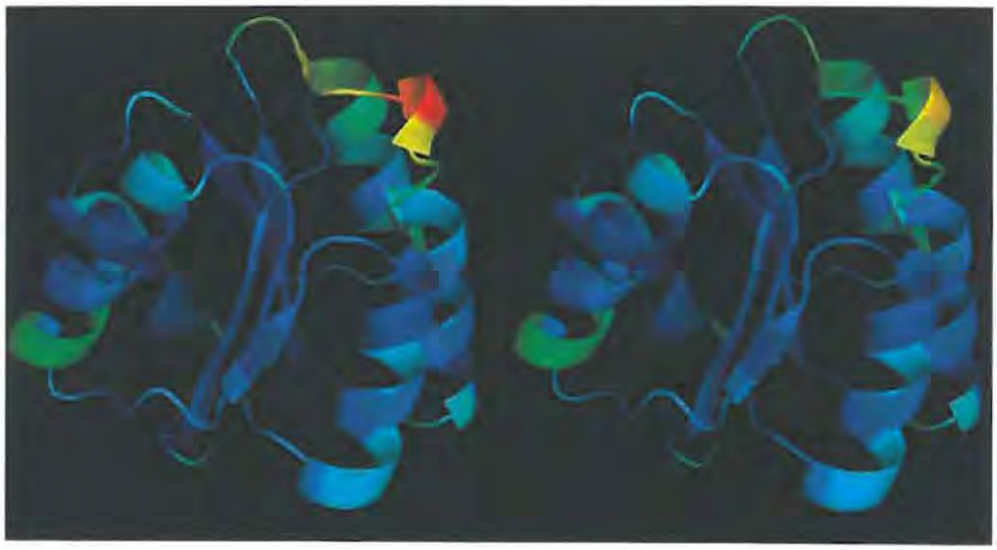


Figure 2.24. The initial (left) and final (right) models for CheY colour according to MetaMQAPii scores. The spectrum of colours from blue to red indicates the spectrum of residues predicted to be correct to those predicted to be incorrect.

Table 2.8. A list of the active sites of CheY from the template organism *T. maritima* and the corresponding positions in *B. amyloliquefaciens* FZB42 and *B. subtilis subtilis* 168.

T. maritima	Function	*B. amyloliquefaciens* FZB42	*B. subtilis subtilis* 168
Asp-54	Acceptor site for Mg^{2+} dependent phosphoryl transfer from the histidine kinase CheA	Asp-54	Asp-54
Asp-9, Asp-10, Asp-54	Acidic cluster where Mg^{2+} binds	Asp-9, Asp-10, Asp-54	Asp-9, Asp-10, Asp-54
Lys-104	Forms salt bridge with active site	Lys-104 (Usher *et al.* 1998)	Lys-104 (Volz 1993)

2.3.4.7. *CheAP1 and CheAP2*

The target template alignments between CheAP1 with 3KYJ chain A and between CheAP2 with 1U0S chain A were individually used as inputs for MODELLER. One hundred models were generated, and in each case a representative model with the lowest DOPE Z score was selected. The selected model for CheAP1 had a DOPE Z score of -2.301 while the model for CheAP2 had a DOPE Z score of -1.778. Evaluation of the models (Figure 2.25) by MetaMQAPii and ProQ suggested that the quality of both models approximate that of experimentally determined structures: CheAP1: RMSD by MetaMQAP 1.704Å, GDT_TS by MetaMQAP 78.373, LGscore by ProQ 5.166 ("extremely good model"); and CheAP2: RMSD by MetaMQAP 1.606Å, GDT_TS by MetaMQAP 77.647, LGscore by ProQ 3.279 ("very good model"). No loop refinement was performed on either of these models. The position of the important P1-domain His residue which accepts the γ-phosphate of ATP is listed in Table 2.9.

Table 2.9. The position of the conserved P1 domain His residue from the template organism *R. sphaeroides* and the corresponding positions in *B. amyloliquefaciens* FZB42 and *B. subtilis subtilis* 168.

R. sphaeroides	Function	*B. amyloliquefaciens* FZB42	*B. subtilis subtilis* 168
His-51	Site of phosphorylation (Bell *et al.* 2010).	His-44	His-46 (Fuhrer & Ordal 1991).

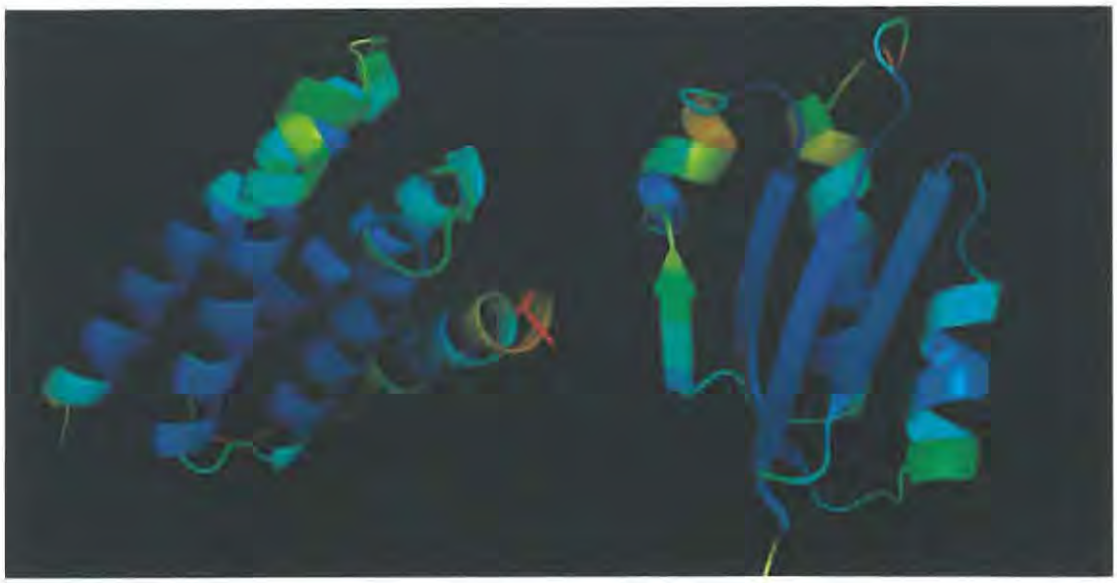


Figure 2.25. The models for CheAP1 (left) and CheAP2 (right) coloured according to MetaMQAPii scores. The spectrum of colours from blue to red indicates the spectrum of residues predicted to be correct to those predicted to be incorrect.

2.3.4.8. *CheC interacting with CheD*

The best model out of a hundred, based on DOPE Z score, was evaluated with ProQ PROCHECK and Verify3D. MetaMQAPii was not used to evaluate any of the complexes because it can only accept single chain structures. Loop refinement was performed on the following residues: 91-101, 172-179, 202-206 and 254-266. The Ramachandran dihedral statistics as well as the Verify3D energy profile showed slight improvement after loop refinement. It should be noted that the presence of residues in disallowed regions (as highlighted by PROCHECK) is not necessarily erroneous, but could be caused by a spatial distortion of the structure due to interaction between the two proteins. The chains and residue numbers of the final model were renumbered using the ResDe package so that the residue numbers correspond to that of the target and the chain name corresponds to that of the target protein. Homology models with similar PROCHECK scores were used in another study to identify key residues, and served as a structural basis to study the interaction between SMAD proteins and DNA (Makkar *et al.* 2009). Therefore, although the quality of the CheC-CheD model is not quite comparable to

that of an experimentally determined structure, it can still be used to identify characteristic features, and combined with information on site-directed positive and purifying selection they can be useful for identifying functionally important residues. Also the model could possibly be used in the reconstruction of the assembly that is formed by when phospho-CheY interacts with CheC bound to CheD. The model quality assessment scores are summarized in Table 2.10.

Table 2.10. Model quality assessment scores for the structure representing interaction between CheC and CheD

Evaluation Method	Assessment scores before refinement	Assessment scores after refinement
DOPE Z score:	-0.608	-0.776
ProQ LGscore :	5.504	5.998
PROCHECK: Most favourable regions Additionally allowed regions Generously allowed regions Disallowed regions	 87.5% 10% 1.9% 0.6%	 88.4% 10% 1.2% 0.3%
Verify3D: % residues with averaged 3D–1D score >= 0.2	~77%	~83%

2.3.4.9. *Interaction between CheA domains P4 and P5 with CheW*

A hundred models were built and the best one was selected based on DOPE Z score. The model was evaluated and loop refinement was performed. Residues that were selected for loop refinement are the following: 322-324. 1-3, 462-462,187-192, 73-78,233-238, 384-393, 249-258, 222-226, 98-105, 117-118, 175-177, 36-39, 63-66,418-422, 438-446, 329-333, 268-271,125-128, 315-317. After loop refinement, the overall quality of the model improved such that it approaches that of an experimentally determined structure. Some secondary structural features of CheW such as two α helices are present in the CheW structure, but absent in the single CheW structure that was based on the template 2QDL. Model quality assessment scores are summarized in Table 2.11.

Table 2.11. Model quality assessment scores for the structure representing interaction between CheA domains P4 and P5 with CheW.

Evaluation Method	Assessment scores before refinement	Assessment scores after refinement
DOPE Z score:	-0.677	-1.002
ProQ LGscore :	5.328	6.451
PROCHECK: Most favourable regions Additionally allowed regions Generously allowed regions Disallowed regions	 88.4% 9.9% 1.4% 0.2%	 90.1% 8.5% 1.2% 0.2%
Verify3D: % residues with averaged 3D–1D score > 0.2	~89%	~96%

2.3.4.10. *Interaction between CheY and the P1 domain of CheA*

The Ramachandran dihedral statistics, the Verify3D scoring function and ProQ LGscore for the model were all good and no refinement was performed. Assessment scores are listed in Table 2.12.

Table 2.12. Model quality assessment scores for the structure representing interaction between CheY and the P1 domain of CheA.

Evaluation Method	Assessment scores
DOPE Z score:	-1.065
ProQ LGscore :	4.225
PROCHECK: Most favourable regions Additionally allowed regions Generously allowed regions Disallowed regions	 93.8% 5.3% 0.4% 0.4%
Verify3D: % residues with averaged 3D–1D score > 0.2	~78%

2.3.4.11. *Interaction between CheB and the P1 domain of CheA*

MODELLER was used to generate a hundred models with slight variations and these models were ranked according to their normalized objective function scores, with the best model selected for further analysis. Model evaluation results indicated a model of good quality, but with some local errors. Loop refinement was performed on the following residues: 260-270; 271-281; and 282. Again, the best model out of a hundred was selected as representative. Following loop refinement the model showed marked improvement in overall quality (Table 2.13).

Table 2.13. Model quality assessment scores for the structure representing interaction between CheB and the P1 domain of CheA.

Evaluation Method	Assessment scores before refinement	Assessment scores after refinement
DOPE Z score:	-0.901	-1.005
ProQ LGscore :	5.793	6.095
PROCHECK: Most favourable regions Additionally allowed regions Generously allowed regions Disallowed regions	 88.4% 9.9% 1.4% 0.2%	 92.5% 6.1% 1.2% 0.2%
Verify3D: % residues with averaged 3D–1D score > 0.2	~80%	~84%

2.3.4.12. *The interaction between CheY and the P2 domain of CheA*

MODELLER was here used to generate a hundred models with slight variations. From these models the best ranking structure was selected as being representative. The representative structure was evaluated with the model quality assessment tools described in previous sections (Table 2.14). The evaluation statistics indicated a model of very high quality, which can be considered on par with an experimentally determined structure, thus it was deemed that no loop refinement was necessary.

Table 2.14. Model quality assessment scores for the structure representing interaction between CheY and the P2 domain of CheA.

Evaluation Method	Assessment scores
DOPE Z score:	-1.552
ProQ LGscore :	4.932
PROCHECK: Most favourable regions Additionally allowed regions Generously allowed regions Disallowed regions	 92.1% 7.90% 0% 0%
Verify3D: % residues with averaged 3D–1D score > 0.2	~89%

2.3.4.13. *Interaction between CheA P2 domain with the N-terminal domain of CheB*

A hundred models representing slight variations of the interaction between the N-terminal-domain of CheB with the P2 domain of CheA were generated using MODELLER. The model with the lowest DOPE Z score was evaluated with the tools described above and deemed of high sufficient quality to justify not performing any loop refinement (Table 2.15).

Table 2.15. Model quality assessment scores for the structure representing interaction between the N-terminal domain of CheB and the P2 domain of CheA.

Evaluation Method	Assessment scores
DOPE Z score:	-1.160
ProQ LGscore :	3.992
PROCHECK: Most favourable regions Additionally allowed regions Generously allowed regions Disallowed regions	 93.8% 5.1% 1.0% 0%
Verify3D: % residues with averaged 3D–1D score > 0.2	 ~95%

2.4. **Discussion**

The gap between the number of experimentally solved structures and known proteins sequences is growing and it is likely that the vast majority of protein structures will never be determined experimentally. Homology modelling has become a very important tool in bridging this gap by making it possible to deduce structural properties from sequences. In this study target sequences from *B. amyloliquefaciens* FZB42 were used to infer representative 3D models of the chemotaxis proteins present in members of the *B. subtilis* group. The quality of the final models was assessed using MetaMQAPii, and loop refinement was performed when needed. Models were assessed using several measures of similarity and model accuracy, including DOPE Z scores, single models were evaluated by MetaMQAPii and ProQ while complexes were evaluated with ProQ, PROCHECK and VERIFY3D. The evaluation statistics for the single models showed that the models for CheB, CheC, CheR, CheY, CheAP1 and CheAP2 approach the quality of experimentally determined structures, and therefore can be used in future to study protein-protein and protein-ligand interactions. Models for CheD and CheW are of somewhat lower quality and would not be suitable for high resolution docking studies. Nevertheless, all models built for the

individual proteins were suitable for the prime objective of this research, which was to study site directed positive and purifying selection operating on each of the proteins in terms of 3D structure. The availability of suitable templates for the interactions between CheC and CheD, CheAP3P4 and CheW, CheY with CheAP1, CheY with P2, CheB with CheAP1 and $CheB_N$ with CheAP2 provided motivation to build computational models for these interactions. The protein-protein contacts of the complexes were identified using the Protein Interactions Calculator. Evaluation statistics of the complexes are in line with that of high quality models.

The use of state of the art techniques, strict modelling protocols and careful selection of templates has contributed to the generation of high quality computational models. These models provide the structural basis towards understanding functional similarities and differences among the chemotaxis proteins of the members of the *B. subtilis* group. A noteworthy outcome of the work described in this chapter is the construction of homology models that show CheB interacting with CheA. To date, information on the interaction between these two proteins in motile bacteria is limited to mutational studies, and no structure showing such an interaction has been solved. This highlights the importance of the role that computational modelling can play in bridging the knowledge gap that exists between available sequence information and experimentally determined structures.

CHAPTER 3

3. ANALYSIS OF SITE DIRECTED POSITIVE- AND PURIFYING SELECTION

The objective of this chapter is to describe site directed positive and purifying selection operating on the chemotaxis proteins and MCPs of the *B. subtilis* group, using the sequences from *B. amyoliquefaciens* FZB42 as reference. The spatial distribution of the sites under purifying/ positive selection is analysed using the homology models presented in Chapter2 as structural references. The pitfalls of using sequences from organisms that are very closely related, thereby biasing results towards purifying selection, are also discussed. The work presented in this chapter is unique, since to date no other studies on site directed evolutionary forces operating on the chemotaxis proteins of a closely related group of bacteria have been undertaken.

3.1. Introduction

The protein sequences of extant organisms have been shaped by a lengthy evolutionary history. As mutations, insertions and deletions occur, changes in the amino acid sequence of a protein accumulate, ultimately affecting the function and 3D structure of the protein (Cothia & Lesk 1986, Krissinel 2007, Rausell *et al.* 2010). There are three broad categories of mutations: Synonymous or silent mutations; non-synonymous mutations and non-sense mutations. Synonymous mutations occur due to redundancy in the genetic code, when such a mutation occurs one codon is changed to another that codes for the same amino acid and causes no change in the protein product. When a codon is changed to one that encodes for a different amino acid it is called a non-synonymous mutation, and, based on the physicochemical properties of the changed amino acid may or may not alter the protein structure and/or function. When a non-synonymous mutation is neutral, one amino acid is replaced by another with similar physicochemical properties. Non-sense mutations come about when an amino acid is changed to a stop codon, resulting in an incomplete protein, normally with deleterious effects. Mutations can also be classified according to the type of nucleic acid substitution, when a purine replaces a purine or a pyrimidine replaces a pyrimidine it is referred to as a transition, but when a pyrimidine replaces a purine or vice versa it is called a transversion. During evolution the

structure of a protein is more stable and changes at a considerably slower rate than the associated DNA sequence (Cothia & Lesk 1986, Krieger *et al.* 2003). Furthermore, areas that are biologically important such as active sites, ligand binding sites or protein-protein interaction areas are usually more conserved than sites that are not functionally important. Mutations in functionally important areas may reduce the fitness of the organism and will be selected against. Sites that exhibit high variability in amino acid composition are considered to be tolerant to functional constraints, but such sites may also be under positive Darwinian selection, giving an adaptational advantage to the organism (Doron-Faigenboim *et al.* 2005). Comparison of normalized synonymous mutation rates (K_a) and non-synonymous mutation (K_s) rates in genes that code for proteins provides an essential means for understanding molecular evolution (Yang *et al.* 2000a, Yang *et al.* 2000b). If non-synonymous mutations do not affect the fitness of the protein, they become fixed within the population at the same rate as synonymous mutations, in this case the rate (ω) of $K_a/K_s = 1$ (Massingham & Goldman 2005). If non-synonymous mutations have a deleterious effect on protein fitness it would be fixed at a lower rate than synonymous mutations. This process is known as purifying selection, making ω smaller than 1. Beneficial non-synonymous mutations will be retained in a population and ω would be larger than 1 (Yang *et al.* 2000a). Positive selection is a rare event as most protein domains are under purifying selection to conserve the functional integrity of the structure. In situations where an organism finds itself in a new environment or when its environment is undergoing significant changes positive selection can be expected, and in some cases it can play a role in speciation. Positive selection may lead to an alteration of the biochemical or enzymatic properties of the protein (Creevey & McInerney 2002).

Many different evolutionary models for inferring K_a/K_s are available and new ones are continuously being developed (Li *et al.* 1985, Doron-Faigenboim & Pupko 2007, Hershberg *et al.* 2007). The most popular methods take into consideration the different transition and transversion probabilities, codon bias, and among-site rate variation. The first methods that were developed inferred a global K_a/K_s value for the entire sequence, or used a sliding window approach to infer K_a/K_s for sub-sequences (Fares *et al.* 2002). Modern methods that estimate K_a/K_s for every amino acid site in a sequence enable the detection of individual sites undergoing positive selection despite a low global K_a/K_s.

Mechanistic codon-based evolutionary models, such as those developed by Goldman and Yang (1994) as well as Muse and Gaut (1994) include parameters for the transition-transversion bias and codon frequencies. The model by Goldman and Yang (1994) uses the Grantham physicochemical distance matrix (Grantham 1974, Doron-Faigenboim & Pupko 2007) to account for the different replacement probabilities between amino acids.

Mechanistic Bayesian models such as those developed by (Yang *et al.* 2000a) assume a prior distribution of K_a/K_s ratios. A drawback to these methods is that the different replacement rates of amino acids are ignored, and therefore they cannot give an accurate description of biological reality.

A more recent development is the estimation of an empirical codon-substitution matrix, based on a large number of coding datasets as opposed to parameterized models. The drawback of using a non-parameterized model is that it may lead to under-fitting of the data being studied.(Doron-Faigenboim & Pupko 2007) suggested a combined codon model that integrates information on empirical amino acid replacement probabilities with theoretical assumptions about transition-transversion bias, codon frequencies and different selection forces acting within and among genes. Context-dependent empirical amino acid replacement probability matrices (Naylor *et al.* 1995, Adachi & Hasegawa 1996, Koshi & Goldstein 1996) can be integrated into the combined codon model to produce a realistic, context-dependent codon model, which allows for improved accuracy when estimating K_a/K_s and phylogeny (Doron-Faigenboim & Pupko 2007).

3.1.1. *Selecton version 2.4: a web-based tool for the identification of site-specific selection*

To estimate site-specific selection of amino acids the Selecton server implements various evolutionary codon models and favours an empirical Bayesian approach (Yang *et al.* 2000a, Swanson *et al.* 2003, Doron-Faigenboim & Pupko 2007, Stern *et al.* 2007). The Mechanistic Empirical Combined (MEC) model (Doron-Faigenboim & Pupko 2007), the M8 (Yang & Bielawski 2000, Yang *et al.* 2000a) and the M5 (Yang *et al.* 2000a) models all allow for positive selection, while the M8a and M7 models do not. To verify that positive selection is significant

the likelihoods of a model that allows for positive selection must be compared to that of a nested null model that does not (Swanson *et al.* 2003) Selecton enables comparison between the following models: M8 vs. M8a; MEC vs. M8a; and M5 vs. M8a. Neither the MEC nor the M5 models have nested null models, therefore statistical significance testing requires comparing Akaike Information Criterion scores (Akaike 1974) with that of the null model. A common theme of the evolutionary models available on the Selecton server is that the codon evolution along a phylogenetic tree is described in probabilistic terms by assuming a statistical distribution to account for heterogeneous K_a/K_s values among sites. For this work the distributions were approximated with 14 discreet categories, which is the maximum allowed by Selecton and give the most accurate results.

The M5 (gamma) model assumes a gamma distribution of K_a/K_s among sites. Unlike the beta distribution, the gamma distribution is not constrained and thus allows for all types of selection (positive, neutral and purifying). A drawback of the model is that under discrete categorization of the gamma distribution it is sometimes the case that no category of positive selection is obtained, despite positive selection operating on the protein. This is especially likely to happen when a large proportion of the protein is subject to strong purifying selection with only a few sites under positive selection (Yang *et al.* 2000a, Stern *et al.* 2007).

The M8 (beta & ω) model allows for all types of selection and enables nesting of null models. The parameters for the M8 model are estimated using Maximum Likelihood: p_0 is the proportion of sites drawn from a beta distribution defined by the interval [0, 1]; p_1 (=1-p_0) is the proportion of sites drawn from an additional category ω_s (which is constrained to be ≥ 1). Sites drawn from the ω_s category are under neutral or positive selection while those drawn from the beta distribution are under purifying selection (Swanson *et al.* 2003). The biggest drawback of the M8 model is that it weighs all amino-acid replacement probabilities more or less equally. Hence the same probability for a Leu codon mutating to a Trp codon (UUG → UGG) will be given to a Leu codon changing to a Phe codon (UUG → UUU) since they both require 1 transversion. However, according to amino-acid similarity matrices, the latter mutation is 5 times more likely than the former (Doron-Faigenboim & Pupko 2007, Stern *et al.* 2007)

The M8a model is a variation of the M8 model, but because the ω_s category is constrained to be 1, only purifying and neutral selection is allowed (Stern *et al.* 2007).

The M7 model, which is nested in the M8 model, does not include the ω_s category. The beta distribution is defined solely on the interval [0, 1]; therefore positive selection is not allowed under this model.

The MEC model differs from all the other models employed by Selecton, in that it is the only method that takes into account the differences between amino acid replacement probabilities. The MEC model expands the 20 by 20 JTT amino acid replacement probability matrix (Jones *et al.* 1992) into a 61 by 61 sense-codon probability matrix. Parameters for the rate of transition (tr) and transversion (tv) are integrated into the matrix. Intensities of different selective forces are taken into account by multiplying all non-synonymous substitutions with ω (for which a prior distribution is assumed). In the MEC model ω does not stand for positive selection, but only for the intensity of positive or purifying selection. K_a/K_s can be computed from ω (Doron-Faigenboim & Pupko 2007, Stern *et al.* 2007). To compensate for the fact that the JTT matrix inherently assumes selection, MEC introduces another parameter (f) for the proportion of sites undergoing no selection. The MEC model is more sensitive compared to other models, because different amino-acid replacement probabilities are treated differently, K_a is computed differently and therefore a position with radical replacements will obtain a higher K_a value than a position with more moderate replacements. Because there is no null model nested within the MEC model the statistical analysis is performed by comparing Akaike Information Content scores between the MEC and the M8a model.

When using the Selecton server it is recommended to use sequences in which the distance is in the order of magnitude of the distances in mammalian sequences to prevent a distortion of K_a/K_s ratio computations.

3.2. Methodology

The methods used to analyse site directed positive and purifying selection operating on each of the proteins under investigation, as well as the spatial distribution of purifying selection, is discussed here.

3.2.1. *Analysis of site directed positive and purifying selection*

The Selecton server was used to study site specific positive and purifying selection operating on the proteins CheA, CheB, CheC, CheD, CheR, CheV, CheW, CheY, McpA, McpB and McpC from the closely related taxonomic unit commonly referred to as the *B. subtilis* group. Nucleotide sequences from the organisms *B. pumilus* SAFR032, *B. licheniformis* ATCC4580, *B. subtilis* ssp. *subtilis* 168, *B. subtilis* ssp. *subtilis* JH642, *B. subtilis* ssp. *subtilis* NCBI3610, *B. subtilis* ssp. *subtilis* SMY, *B. subtilis* ssp. *spizizenii* ATCC6633, *B. amyloliquefaciens* ssp. *amyloliquefaciens* DSM7, *B. amyloliquefaciens* ssp. *plantarum* FZB42, *B. amyloliquefaciens* ssp. *plantarum* B946, *B. amyloliquefaciens* ssp. *plantarum* B9601Y2 and *B. amyloliquefaciens* ssp. *plantarum* GaoB3 were used to study the selective forces acting upon the chemotaxis proteins and receptor proteins. Annotated sequences for CheA from *B. amyloliquefaciens* ssp. *plantarum* B9601Y2 and *B. amyloliquefaciens* ssp. *plantarum* GaoB3 and for CheR, CheV, McpA, McpB or McpC from *B. subtilis* ssp. *spizizenii* ATCC6633 were not available at the time of study. Sequences for the housekeeping gene *gyrA* were used as a control. For a full list of accession numbers of sequences used, please refer to the folder "Chapter3_data/Accesion_numbers.txt" on the accompanying disk. Note that the newly determined sequences for the *B. amyloliquefaciens* strains, DSM7^T, B946, B9601Y2 and GaoB3.were not publicly available at the time of print. A Python script was used to colour code the results of the M7 model onto the homology models built in Chapter 2 to get a visual representation of the spatial distribution of sites under purifying selection. This was particularly useful for comparison with the results obtained from PIC and allowed for the identification of residues that are under strong functional constraints in the *B. subtilis* group, due to their importance in inter- and intra- protein interactions. Structures were visualized in PyMOL.

3.2.2. *Intensity of purifying selection on residues with known functional importance*

To determine which of the sites under strong purifying selection (according to the M7 model) also have a functional role within the closely related taxonomic unit, active sites were identified by overlaying the homology models with the templates that they were based on, for which the active sites have been determined experimentally (see Chapter 2). In each case the structures and sequences from *B. amyloliquefaciens* FZB42 were used as references. Unless otherwise indicated all residues are numbered according to their position in *B. amyloliquefaciens* FZB42.

3.2.3. *Spatial distribution of residues under purifying selection*

To analyse the spatial distribution of sites under purifying selection (according to the M7 model), sites that play a role in intra- and inter-protein interactions were predicted with the Protein Interaction Calculator (PIC) server (see Chapter 2).

3.3. Results and Discussion

The results obtained after analysing the chemotaxis proteins, as well as the receptors McpA, McpB and McpC for site directed positive and purifying selection are presented in this section. Furthermore an attempt is made to compare the various evolutionary models available on the Selecton server with eachother in terms of their accuracy in predicting functionally important residues. The spatial distributions of purifying selected sites on each protein, according to the M7 model, are also presented here.

3.3.1. *Analysis of site directed positive and purifying selection*

To analyze the possibility that positive selection acts on the various chemotaxis proteins and MCPs found in members of the *B. subtilis* group we used various codon models that are available on the Selecton server to determine the K_a/K_s ratio at each amino acid site. The multiple sequence alignments that were analyzed as well as the results (K_a/K_s scores, phylogenetic trees, likelihood and parameters) obtained can be found on the disk containing supplementary data under a folder titled "Chapter3_data". Likelihood values and parameters, as well as residues detected to be under positive selection, can be found in Table 3.1, which also contains results of significance tests for proteins that were assumed to be under positive selection. According to the MEC model only McpB and McpC were predicted to contain sites

under positive selection, and comparison with the M8a model verified the significance of this prediction. According to the M8 model McpA and McpC contained sites under positive selection, but the likelihoods ratio test comparison with M8a revealed that these predictions were not significant, therefore the hypotheses that positive selection is operating on McpA and McpC were rejected.

Table 3.1. Likelihood values under the five models that allow for hypothesis testing as well as sites under positive selection as inferred by the three models allowing for positive selection, applied to each of the chemotaxis proteins studied. * Amino acid sites are numbered according to their position in the reference sequence from *B. amyloliquefaciens* FZB 42. ** The M8a null model was not run when no positive selection was detected by an alternative model. *** Statistical significance test passed, positive selection is significant. Parameters of the models are: shape parameters of beta distribution (α, β); transition/transversion ratio of M8 model (κ), ω_s (additional category representing positive selection), p_1 (proportion of ω_s), rate of transition for MEC model (tr), rate of transversion for MEC model (tv), sites under no selection for MEC model (f)

Protein	Model	Selecton optimized parameters	Selected sites * (outcome of significance test)
CheA	MEC	log-likelihood: -7467.2 tr: 4.3907 tv: 4.09729 f=prob (selection): 1 α: 0.156847 β: 2.65081	None
	M8	log-likelihood: -7530.31 κ: 2 α: 0.117898 β: 2.74028 additional omega category: 1.00757 prob (additional omega category): 0.100099	None
	M5	log-likelihood: -7533.71 κ: 1.45556 α: 0.211658 β: 2.57373	None
	M8a	Not applicable**	Not applicable**
	M7	log-likelihood: -7532.05 κ: 1.46878 α: 0.216503 β: 2.74325 additional omega category: 1.5 prob(additional omega category): 0	None allowed
CheB	MEC	log-likelihood: -4571.99 tr: 4.3956 tv: 3.90708 f=prob (selection): 1 α: 0.244708 β: 2.65081	None
	M8	log-likelihood: -4650.24 κ: 2 α: 0.117898	25, 90, 135, 151, 247, 250, 277, 279, 348 (Reject model)

		β: 2.23607 additional omega category: 1.1848 prob (additional omega category): 0.157799	
	M5	log-likelihood: -4628.44 κ: 1.82298 α: 0.332689 β: 2.70564	None
	M7	log-likelihood: -4625.9 κ: 1.85743 α: 0.320017 β: 2.57373 additional omega category: 1.5 prob(additional omega category): 0	None allowed
	M8a (M8a vs. M8a)	log-likelihood: -4628.1 κ: 1.8 α: 0.343743 β: 2.70564 additional omega category: 1 prob(additional omega category): 0.0172209	None allowed (Reject MEC model)
CheC	MEC	log-likelihood: -2273.07 tr: 4.3907 tv: 4.20128 f=prob (selection): 1 α: 0.198811 β: 2.65081	None
	M8	log-likelihood: -2283.86 κ: 1.49318 α: 0.288281 β: 2.70564 additional omega category: 1.00757 prob (additional omega category): 0.0170192	None
	M5	log-likelihood: -2284.46 κ: 1.48069 α: 0.279987 β: 2.70564	None
	M7	log-likelihood: -2283.24 κ: 1.48069 α: 0.260047 β: 2.62382 additional omega category: 1.5 prob(additional omega category): 0	None allowed
	M8a	Not applicable**	Not applicable**
CheD	MEC	log-likelihood: -1903.39 tr: 4.3907 tv: 2.67717 f=prob (selection): 1 α: 0.167313 β: 2.65081	None
	M8	log-likelihood: -1925.98 κ: 2 α: 0.117898 β: 2.70564 additional omega category: 1.06441 prob (additional omega category): 0.102638	None
	M5	log-likelihood: -1922.42 κ: 1.64789	None

		α: 0.264621	
		β: 2.59195	
	M7	log-likelihood: -1920.83	None allowed
		κ: 1.66357	
		α: 0.260905	
		β: 2.53391	
		additional omega category: 1.5	
		prob (additional omega category): 0	
	M8a	Not applicable**	Not applicable**
CheR	MEC	log-likelihood: -3037.08	None
		tr: 4.3907	
		tv: 4.18024	
		f=prob (selection): 1	
		α: 0.177004	
		β: 2.65081	
	M8	log-likelihood: -3081.11	None
		κ: 1.64929	
		α: 0.229952	
		β: 2.72133	
		additional omega category: 1.00757	
		prob (additional omega category): 0.0180317	
	M5	log-likelihood: -3081.6	None
		κ: 1.49564	
		α: 0.231403	
		β: 2.6234	
	M7	log-likelihood: -3081.51	None allowed
		κ: 1.49564	
		α: 0.210318	
		β: 2.36144	
		additional omega category: 1.5	
		prob (additional omega category): 0	
	M8a	Not applicable**	Not applicable**
CheV	MEC	log-likelihood: -3255.31	None
		tr: 4.18272	
		tv: 4.27032	
		f=prob (selection): 1	
		α: 0.189516	
		β: 2.65081	
	M8	log-likelihood: -3276.13	None
		κ: 2	
		α: 0.117898	
		β: 2.70564	
		additional omega category: 1.00757	
		prob (additional omega category): 0.149106	
	M5	log-likelihood: -3281.69	None
		κ: 1.4326	
		α: 0.247749	
		β: 2.59195	
	M7	log-likelihood: -3278.69	None allowed
		κ: 1.44818	
		α: 0.217637	
		β: 2.53516	
		additional omega category: 1.5	
		prob (additional omega category): 0	
	M8a	Not applicable**	Not applicable**
CheW	MEC	log-likelihood: -1871.32	None

		tr: 4.41861 tv: 4.35509 f=prob (selection): 1 α: 0.230077 β: 2.65081	
	M8	log-likelihood: -1890.71 κ: 2 α: 0.117898 β: 2.45967 additional omega category: 1.00757 prob (additional omega category): 0.229044	None
	M5	log-likelihood: -1891.46 κ: 1.64521 α: 0.309511 β: 2.70564	None
	M7	log-likelihood: -1890.04 κ: 1.49762 α: 0.275644 β: 2.6131 additional omega category: 1.5 prob (additional omega category): 0	None allowed
	M8a	Not applicable**	Not applicable**
CheY	MEC	log-likelihood: -1080.5 tr: 4.27051 tv: 2.34761 f=prob (selection): 1 α: 0.122291 β: 2.53622	None
	M8	log-likelihood: -1088.24 κ: 2.54818 α: 0.117898 β: 2.70564 additional omega category: 1.00757 prob (additional omega category): 0.0172209	None
	M5	log-likelihood: -1088.19 κ: 2.20317 α: 0.117898 β: 2.70564	None
	M7	log-likelihood: -1088.11 κ: 1.8 α: 0.117898 β: 2.70564 additional omega category: 1.5 prob (additional omega category): 0	None allowed
	M8a	Not applicable**	Not applicable**
McpA	MEC	log-likelihood: -7920.95 tr: 4.41777 tv: 4.35509 f=prob (selection): 1 α: 0.256004 β: 2.65081	None
	M8	log-likelihood: -8060.46 κ: 2 α: 0.117898 β: 2.23607 additional omega category: 1.5	7, 30, 50, 69, 72, 74, 88, 91, 93, 95, 99, 106, 126, 127, 157, 172, 186, 200, 218, 219,225 234, 235, 244, 247, 252, 257, 264, 269, 289, 299, 314, 318, 348, 432, 457, 471, 534, 541, 559, 569, 606, 619,

		prob (additional omega category): 0.215639	646, 657 (Reject model)
	M7	log-likelihood: -8025.68 κ: 1.47403 α: 0.361188 β: 2.61653 additional omega category: 1.5 prob (additional omega category): 0	None allowed
	M8a vs. M8	log-likelihood: -8031.34 κ: 1.458 α: 0.376805 β: 2.70564 additional omega category set to 1 prob(additional omega category): 0.0172209	None allowed
	M5	log-likelihood: -8031.5 κ: 1.60254 α: 0.359466 β: 2.70564	None
McpB	MEC***	log-likelihood: -8322.41 tr: 4.40759 tv: 4.03245 f=prob (selection): 1 α: 0.281649 β: 2.65081	91 (Accept model)
	M8	log-likelihood: -8409.39 κ: 2 α: 0.13514 β: 1.76393 additional omega category: 1.01889 prob (additional omega category): 0.227528	None
	M5	log-likelihood: -8392.47 κ: 1.72919 α: 0.453528 β: 2.70564	None
	M7	log-likelihood: -8381.32 κ: 1.5846 α: 0.411978 β: 2.67839 additional omega category: 1.5 prob (additional omega category): 0	None allowed
	M8a vs. MEC	log-likelihood: -8387.73 κ: 1.8 α: 0.414054 β: 2.5114 additional omega category set to 1 prob (additional omega category): 0.0172209	None allowed Accept MEC model
McpC	MEC***	log-likelihood: -8417.07 tr: 4.37014 tv: 4.22558 f=prob (selection): 1 α: 0.30723 β: 2.65081	67, 71, 123 (Accept model)
	M8	log-likelihood: -8591.86 κ: 2 α: 0.181153 β: 1.58754	67, 71, 122, 123, 242, 354, 536, 598, 651 (Reject model)

		additional omega category: 1.18286 prob (additional omega category): 0.126688	
	M5	log-likelihood: -8543.93 κ: 1.62946 α: 0.464575 β: 2.70564	None
	M7	log-likelihood: -8537.64 κ: 1.66629 α: 0.492547 β: 2.62297 additional omega category: 1.5 prob (additional omega category): 0	None allowed
	M8a vs. MEC	log-likelihood: -8540.1 κ: 1.62 α: 0.469229 β: 2.57959 additional omega category set to 1 prob(additional omega category): 0.0172209	None allowed (Accept MEC model)
	M8a vs. M8	log-likelihood: -8540.08 κ: 1.62 α: 0.469524 β: 2.57959 additional omega category set to 1 prob (additional omega category): 0.0172209	None allowed (Reject M8 model)
gyrA	MEC	log-likelihood: -7891.51 tr: 4.20163 tv: 3.12492 f=prob(selection): 1 alpha: 0.122291 beta: 2.53622	None
	M8	log-likelihood: -7909 κ:2 α: 0.117898 β: 2.70564 additional omega category: 1.00757 prob (additional omega category): 0.0365025	None
	M5	log-likelihood: -7913.85 κ: 1.443 α: 0.117898 β: 2.70564	None
	M7	log-likelihood: -7908.23 κ: 1.44379 α: 0.117898 β: 2.70564 additional omega category: 1.5 prob (additional omega category): 0	None allowed
	M8a	Not applicable**	Not applicable**

* Amino acid sites are numbered according to their position in the reference sequence from *B. amyloliquefaciens* FZB 42

** The M8a null model was not run when no positive selection was detected by an alternative model

*** Statistical significance test passed, positive selection is significant.

Parameters of the models are: shape parameters of beta distribution(α, β); transition/transversion ratio of M8 model (κ) ,ω_s (additional category representing positive selection), p_1 (proportion of ω_s), rate of transition for MEC model(tr), rate of transversion for MEC model (tv), sites under no selection for MEC model (f)

3.3.2. *Intensity of purifying selection on residues with known functional importance*

In order to assess the various evolutionary models for their ability to detect functionally important sites, K_a/K_s values obtained at known active sites (that are conserved throughout the genomes of the organisms under study) were compared with each other. The M5, M7 and M8 models were comparable in accuracy (Table 3.2).

Table 3.2. K_a/K_s values at known active sites as determined under various models. Note that the values are not normalized.

Protein	Function	Ka/Ks: M5	Ka/Ks: M7	Ka/Ks: M8	Ka/Ks: MEC
CheA:					
His-44	Site of phosphorylation on P1 domain	0.014	0.015	0.0082	0.012
Asn-406	ATP binding on P4 domain, N-box	*0.013*	*0.015*	*0.0074*	0.012
Asp-446	ATP binding on P4 domain G1-box	*0.011*	*0.012*	*0.0062*	0.011
Gly-448	ATP binding on P4 domain G1-box	0.018	0.02	0.0062	0.019
Gly450	ATP binding on P4 domain G1-box	0.018	0.02	0.0099	0.019
Phe-484	ATP binding on P4 domain F-box	0.015	0.016	0.0086	0.018
Phe-488	ATP binding on P4 domain F-box	0.018	0.02	0.011	0.021
Gly-499	ATP binding on P4 domain G2-box	0.017	0.2	0.0095	0.019
Gly-501	ATP binding on P4 domain G2-box	0.017	0.017	0.0084	0.016
Gly-503	ATP binding on P4 domain G2-box	0.018	0.02	0.0098	0.019
CheB:					
Ser-171	Catalytic triad	0.0300	0.0330	0.0130	*0.0130*
His-198	Catalytic triad	0.0230	0.0250	0.0098	0.0200
Asp-294	Catalytic triad	*0.0160*	*0.0180*	*0.0068*	0.0190
Asp-54	Phosphoryl transfer	*0.0150*	*0.0170*	*0.0063*	0.0180
Asp-8	Mg^{2+} binding	*0.0160*	*0.0170*	*0.0064*	0.0180
Asp-9	Mg^{2+} binding	*0.0160*	*0.0170*	*0.0067*	0.0190
Glu-56	Mg^{2+} binding	*0.0160*	*0.0170*	*0.0065*	0.0200
CheC:					
Glu-17	First phosphatase active site	*0.0180*	*0.0170*	*0.0190*	0.0180
Asn-20	First phosphatase active site	*0.0170*	*0.0160*	*0.0180*	*0.0130*
Glu-118	Second phosphatase active site	*0.0180*	*0.0170*	*0.0190*	0.0180
Asn-121	Second phosphatase active site	*0.0170*	*0.0160*	*0.0180*	*0.0130*

Protein	Function	Ka/Ks: M5	Ka/Ks: M7	Ka/Ks: M8	Ka/Ks: MEC
CheD:					
Cys-33	Cysteine hydrolase activity	0.0280	0.0300	0.0130	0.0220
His-50	Cysteine hydrolase activity	0.0210	0.0220	0.0096	0.0140
Thr-27	Cysteine hydrolase activity	0.0200	0.0210	0.0087	***0.0100***
CheR:					
Ile-131	AdoMet interaction	***0.0120***	***0.0110***	***0.0120***	***0.0080***
Val-203	AdoMet interaction	0.0140	0.0140	0.0140	***0.0086***
Asp-130	AdoMet interaction	***0.0094***	***0.0088***	***0.0092***	0.0090
Thr-15	AdoMet interaction	0.0160	0.0150	0.0160	***0.0080***
Asn-185	AdoMet interaction	***0.0094***	***0.0089***	***0.0093***	***0.0082***
CheW:					
Asn-52	CheA binding	***0.0130***	***0.0120***	***0.0061***	***0.0110***
Gly55	CheA binding	0.0190	0.0170	0.0084	0.0200
Ile-57	CheA binding	0.0960	0.0950	0.1600	0.0900
Pro-59	CheA binding	0.0260	0.0240	0.0130	0.0200
CheY:					
Asp-54	Phosphoryl transfer	***0.0081***	***0.0082***	***0.0085***	0.0150
Asp-9	Mg^{2+} binding	***0.0086***	***0.008***	***0.0090***	0.0160
Asp-10	Mg^{2+} binding	***0.0085***	***0.008***	***0.0090***	0.0160

K_a/K_s scores indicated in **bold** and *italic* falls under the category of most intense purifying selection

3.3.3. *Spatial distribution of residues under purifying selection*

The results of the M7 model were colour coded onto the reference sequences by Selecton (Figure 3.1).When a homology model was available for a protein the Selecton results were also colour coded onto the structure using a PyMOL script. The results for each protein will be discussed here in detail.

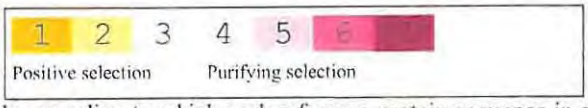


Figure 3.1 The selection scale according to which each reference protein sequence in a MSA is coloured. K_a/K_s scores are projected onto the primary sequence of the protein, using a 7-colou r scale. Shades of yellow (colours 1 and 2) indicate a normalized K_a/K_s ratio > 1, and shades of bordeaux (colours 3 through 7) indicate a normalized K_a/K_s ratio < 1:

3.3.3.1. *CheA*

The functional organization of CheA was inferred from the homology models constructed in Chapter 2 (Figure 3.2). The P1, P4 and P5 domain, as well as the region between the C-terminus of the P2 domain and the N-terminus of the P4 (i.e. the P3 domain and flanking linker regions) were predicted to be under predominantly purifying selection, while the P2 domain was predicted to be under neutral selection.

Table 3.3. Selecton colour-coded results for CheA. K_a/K_s scores under the M7 model are projected onto the primary sequence of the protein. The number in the top left of each column indicates sequence position.

1	11	21	31	41
M S Q Y L D V E I D	E S K E H L Q T C S	E E L L L L E K D P	A D L Q L N H D I F	R N A H T L K G M S
51	61	71	81	91
N T M G Y T E L A H	L T H L M E N V L D	N I R N G E M P V T	S D W L D V L I E A	L D H L E E M N Q S
101	111	121	131	141
I I D G G D G N R D	I S E V S A K L E N	N A V H E T A A S A	E T A E P P A S E Q	Q A S T E W N N D K
151	161	171	181	191
E E R T N I E E A E	E Q G E S R Y E I T	V S L N E S C M L K	A N R N Y M I E E K	L N K A G H N A R T
201	211	221	231	241
I P A A E V L E T E	D F G T D E Q V C F	L T K Q P A G E I K	E L I S G I S E V E	N V E I S A G A P L
251	261	271	281	291
K T N E K P Q A A E	P V K E A P V K K T	E K Q P K P Q A K T	E E Q P K H H S G S	K T I R V S I E R L
301	311	321	331	341
D S S M N L E E E L	V I D R G R L E Q I	A K E L D H N K L T	E T N E R L T R I S	G D L Q S I I L N E
351	361	371	381	391
R N M P N E T N E N	R F P R M I R Q L Q	N E L N S N I E L S	I I G N E T E L E R	T N I D E I G E P L
401	411	421	431	441
V H L I R N S I D H	G I E S P E V R V N	K G K P L S G H N V	L K A Y H S G N H V	F I E N E E D G A G
451	461	471	481	491
L N R K K I L E N A	L E R S N I T E R D	N E T L E D N E I Y	E L I I N P G F S T	N D Q I S D I S G R
501	511	521	531	541
G N G L D N M K N K	L E S L G G S V S V	K S A E G Q G S L F	S I Q L P L T L S I	I S N L L I K L E E
551	561	571	581	591
E T E A I P I S S I	I E T N N I D K S D	I L Q T H E R E N I	D E R G Q I N P I N	Y L K K E E G I T D
601	611	621	631	641
S K Q D A D Q F H V	I V N K K G D K S T	N E N V K S E I G Q	Q E N N L K S L G D	Y L T N V E N I S G
651	661			
N T I L G N G E N N	L I I D C N N L I			

Table 3.4. Interacting residues that play a role in maintaining hydrophobic contact between the N-terminal and C-terminal domains of CheB. K_a/K_s values were obtained under the M7 model.

N-terminal residue	K_a/K_s value	C-terminal residue	K_a/K_s value
Leu-95	0.032	Pro-175	0.029
Phe-100	0.1	Val-319	0.023
Phe-100	0.1	Val-320	0.022
Phe-102	0.024	Met-199	0.014
Phe-102	0.024	Pro-200	0.029
Phe-102	0.024	Phe-203	0.024

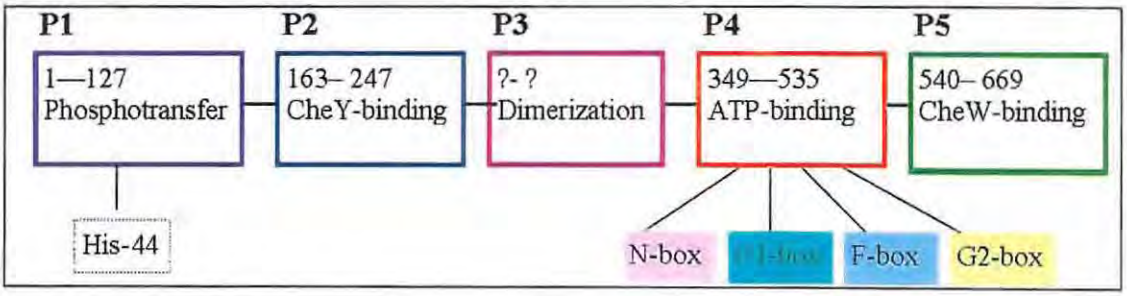


Figure 3.2. Functional organization of CheA. Domain P1 comprise of residues 1 to 127; domain P2 of residues 163 to 247; domain P4 of residues 349 to 535 and domain P5 of residues 540 – 669. The exact borders of domain P3 remain unknown. The His-44 residue in the P1 is the site of phosphorylation. The N- G1-,F- and G2-boxes in domain P4 are crucial for ATP binding and catalysis (Lupas & Stock 1989, Parkinson & Kofoid 1992, Hirschman *et al.* 2001).

3.3.3.2. *CheB*

The CheB protein contained patches of localized purifying selection. These patches surround the N-terminal domain acidic cluster where Mg^{2+} binding and phosphotransfer take place. Another obvious patch under purifying selection was in the interface between the N-terminal and C-terminal domains (Figure 3.3 and Figure 3.4). A comparison of the Selecton results with the results from PIC revealed that the residues that are important for maintaining hydrophobic contact between the N- and C-terminal domains are under varying degrees of purifying selection (Table 3.4). All the residues involved in these hydrophobic interactions are conserved throughout members of the *B. subtilis* group, except for Phe-100, which is encoded as a Val in *B. licheniformis* ATCC14580. Response regulators such as CheB are predicted to be dynamic in terms of their structural conformation, and are thought to oscillate between the active and inactive conformations. Phosphorylation is thought to shift the oscillation equilibrium towards

the active form (Anand *et al.* 2000). In its unphosphorylated form the N-terminal blocks the methylesterase active site on the C-terminal domain. Hughes *et al.* (2001) showed that phosphorylation of CheB caused an increase in solvent accessibility of the active site.

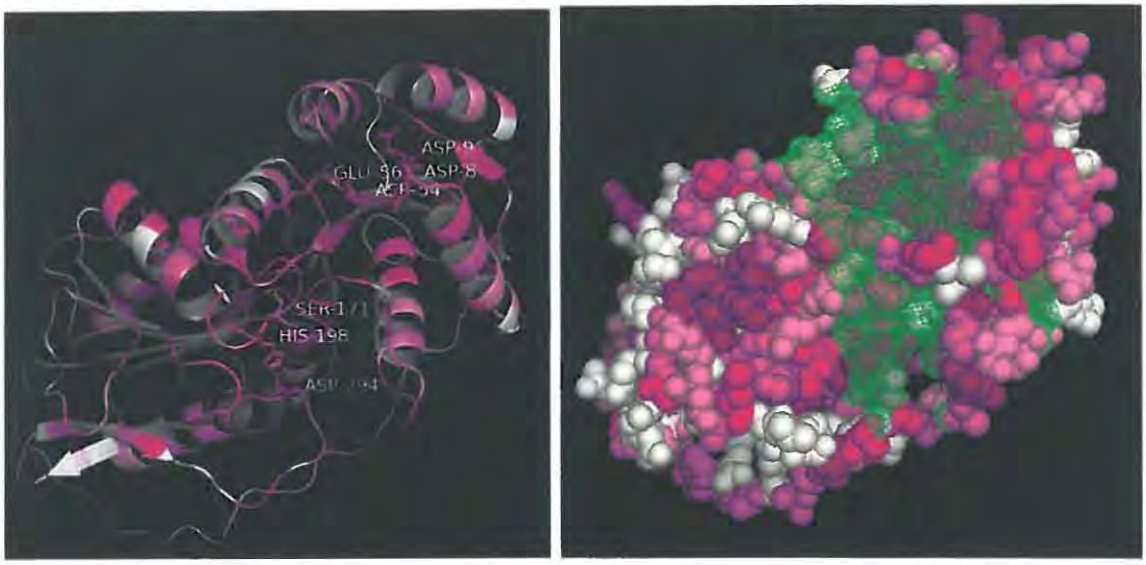


Figure 3.3 (left). The homology model for CheB from *B. amyloliquefaciens* FZB42 shown as a cartoon and coloured according to K_a/K_s values obtained under the M7 model. The rangeof colours represents the selective forces from neutral (white) to intense purifying selection (purple). Known active sites are indicated as sticks and labelled.

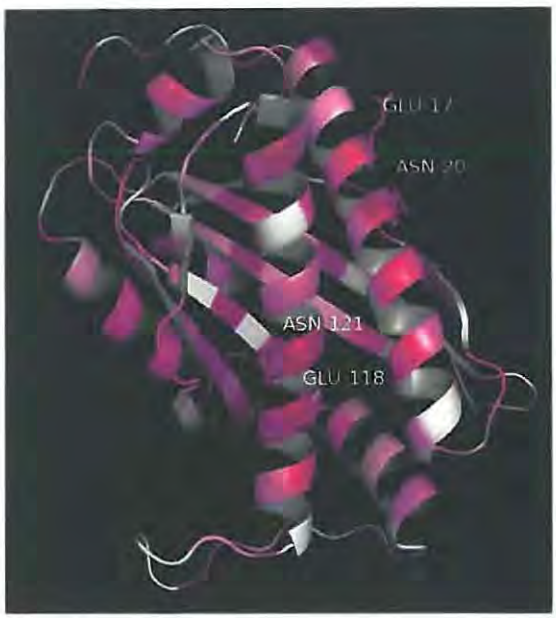
Figure 3.4 (right). The same structure as represented in Figure 3.3, but shown with space filling spheres, the green mesh highlights patches that are under purifying selection that play a role in phosphor binding and methylesterase catalysis.

3.3.3.3. *CheC*

The active sites of CheC were inferred to be under intense purifying selection. The patterns of purifying selection seem more or less evenly dispersed throughout the protein, with the active sites under intensive purifying selection as well as some core residues that could possibly play a role in maintaining tertiary structure (Figure 3.5). A comparison of Selecton and PIC results revealed that the following residues under intensive purifying selection do indeed play a role in hydrophobic interactions within the protein: Phe-50, Ile-67, Ile-68, Met-71, Phe-80, ILe-82, and Met-83.

3.3.3.4. *CheD*

Although the amino acids that comprise the catalytic triad of CheD (Thr-27, Cys-33 and His-50) are conserved throughout members of the *B. subtilis* group used in this study (refer to MSA), they were predicted to be under varying degrees of purifying selection (Figure 3.6). Residues in the core region that are under intensive purifying selection which also play a role in internal hydrophobic interactions are: Ile-13, Ile-25, Met-52, Ile-130, Ile-145 and Ile-156.



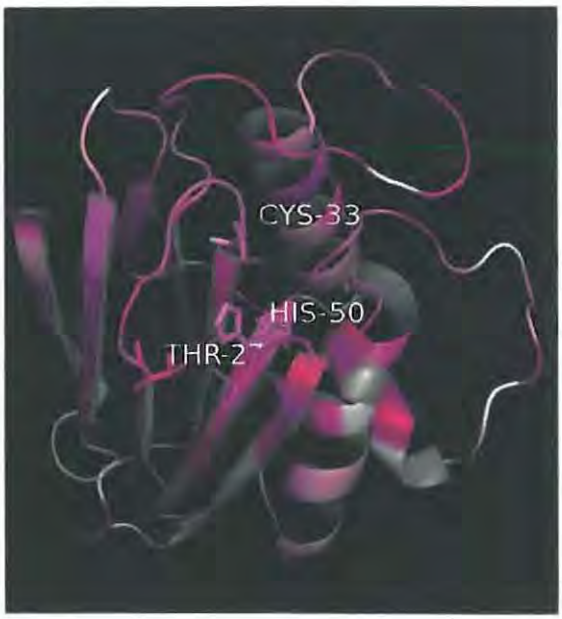


Figure 3.5 (left). The homology model for CheC from *B. amyloliquefaciens* FZB42 shown as a cartoon and coloured according to K_a/K_s values obtained under the M7 model. The rangeof colours represents the selective forces from neutral (white) to intense purifying selection (purple). Known active sites are indicated as sticks and labelled

Figure 3.6 (right). The homology model for CheD from *B. amyloliquefaciens* FZB42 shown as a cartoon and coloured according to K_a/K_s values obtained under the M7 model. The rangeof colours represents the selective forces from neutral (white) to intense purifying selection (purple). Known active sites are indicated as sticks and labelled

3.3.3.5. *CheR*

Purifying selection operating on CheR was predicted to be more intensive in the inter-domain interface and on the face of the protein containing the small β-sub domain, whilst the opposite side of the protein contains large areas under neutral selection (Figure 3.7 and Figure 3.8). A comparison between the Selecton and PIC results showed that the results predicted to play a part in the hydrophobic interactions between the N-terminal and C-terminal domains are under

varying degrees of purifying selection (Table 3.5). As far as the β-sub domain is concerned, in Proteobacteria the binding site for CheR is distinct from the methylation sites and involves the small β-sub domain interacting with a NWETF pentapeptide motif on the C-terminus of the receptor (Wu *et al.* 1996, Shiomi *et al.* 2002). In other classes of bacteria the NWETF motif is absent from the receptors, and the role of the β-subdomain, outside of maintaining structural integrity, remains a mystery (Perez & Stock 2007).

.**Table 3.5.** Interacting residues that play a role in maintaining hydrophobic contact between the N-terminal and C-terminal domains of CheR.

N-terminal residue	K_a/K_s value	C-terminal residue	K_a/K_s value
11-Trp	0.021	131-Ile	0.011
17-Val	0.014	206-Tyr	0.014
17-Val	0.014	207-Phe	0.012

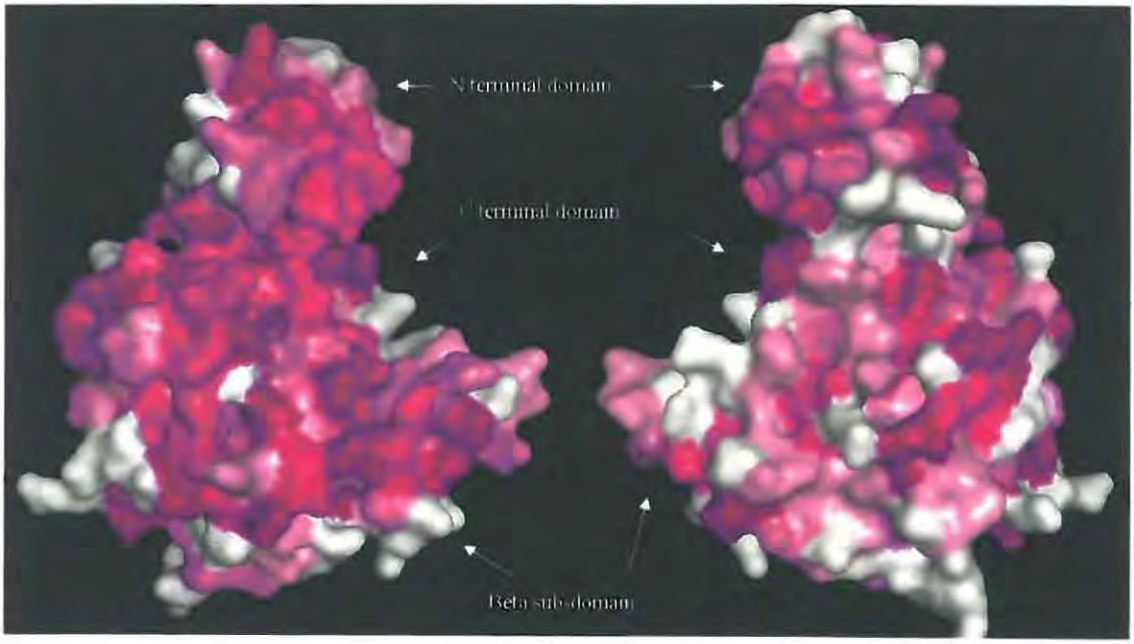


Figure 3.7. Two images of the homology model for CheR from *B. amyloliquefaciens* FZB42 indicating the protein surface and coloured according to K_a/K_s values obtained under the M7 model. The rangeof colours represents the selective forces from neutral (white) to intense purifying selection (purple). The image on the left shows the surface that is under intensive purifying selection, the image on the right is a 180 ° rotation on the Y-axis, showing the opposite side that is under less constraining purifying selection

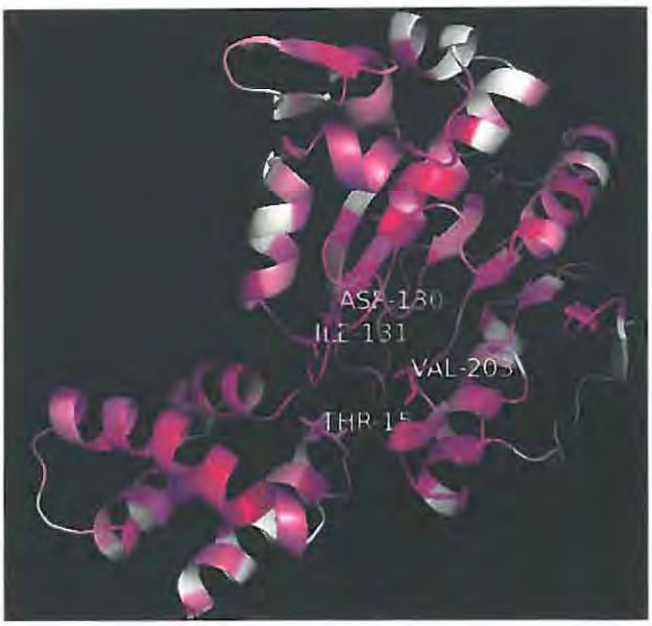


Figure 3.8. The homology model for CheR from *B. amyloliquefaciens* FZB42 shown as a cartoon and coloured according to Selecton scores as described above. The structure is shown from a different angle as that of Figure 3.7. Active site residues are indicated as sticks and labelled.

Table 3.6. Selecton colour-coded results for CheV. K_a/K_s scores under the M7 model are projected onto the primary sequence of the protein. The number in the top left of each column indicates sequence position.

1	11	21	31	41
LSLQQYDLL	DSGTSLLIN	NEEVGVSTG	INNMSNREII	QPNDVTSNPQ
51	61	71	81	91
SHKDVIGMIK	LRGEILPNIS	LYTEEGNDAE	EAKDDSNINI	EISKRSINEH
101	111	121	131	141
VGSNSQIHRN	SWEAIFNPTS	LNQGMERELT	GIISLESTMI	HLPDVESIIY
151	161	171	181	191
DIESASGVET	MHVHQMGFME	RRAGKSLIIN	EHSPLLMRLL	TDELTEAGYS
201	211	221	231	241
EIVTEENGSE	AYDYVVELTD	NGANLSDHIE	MISDIIMPK	SDGHRLTSLL
251	261	271	281	291
NDNPQSSDVP	IMIESSLITD	NLRHRGEVVG	VEEQISNPII	SELIKSVETY
301				
VIE				

3.3.3.6. *CheV*

The CheY-like domain of CheV contains all of the most highly conserved residues found in response regulators: two acidic residues where Mg^{2+} binds (Glu-181 and Asp-182), the phosphorylated residue (Asp-235) and the conserved lysine (Lys-287) which forms a salt-bridge with the active site. These residues were all detected to be under intensive purifying selection. The CheW-like domain of CheV contains a deviation from the Asn-x-x-Gly-x-Ile-x-Pro

consensus. In members of *B. subtilis* the Asn is replaced with Lys/Thr/Ser. None of the residues in the consensus is under purifying selection.

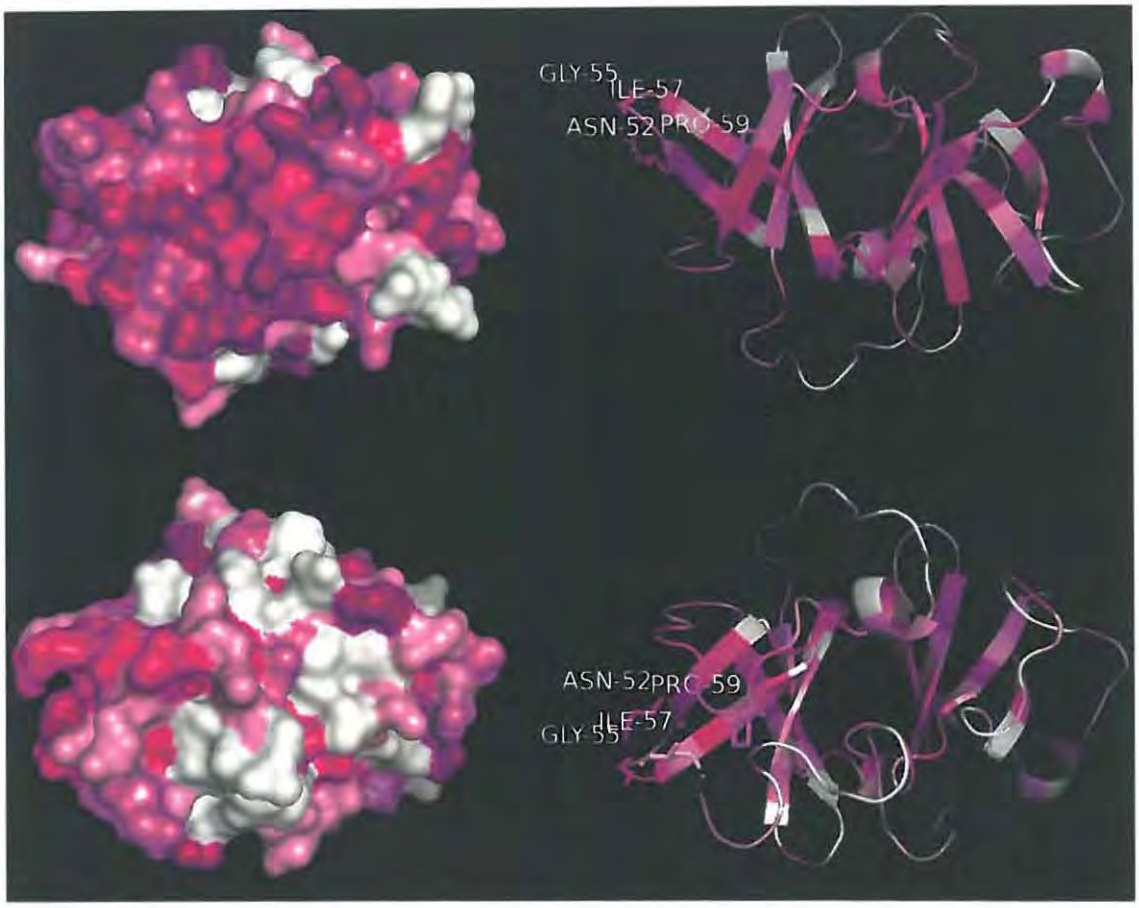


Figure 3.9. Four images of the homology model for CheW from *B. amyloliquefaciens* FZB42 indicating the proteins surface and coloured according to K_a/K_s values obtained under the M7 model. The rangeof colours represents the selective forces from neutral (white) to intense purifying selection (purple). The image on the top left shows the surface that is under intensive purifying selection. The image on the top right shows the molecule in the same orientation as in the top left but represented as a cartoon with CheA binding sites indicated as sticks and labelled. The images on the bottom left and right are a 180 ° rotation on the X-axis, showing the opposite side that is under less constraining purifying selection. The bottom left image shows the surface of the molecule while the bottom right image shows the molecule as a cartoon with CheA binding sites indicated as sticks and labelled

3.3.3.7. *CheW*

Purifying selection was predicted to be more intensive on the one face of the protein that is comprised mostly of β-sheets, with the opposite face containing mostly loops under predominantly neutral selection (Figure 3.9). The hydrophobic residues Met-11, Ile-23, Ile-47,

Ile-81, Ile-82 and Ile-137 which are under intense purifying selection were detected by PIC as involved in hydrophobic reactions, and probably play key roles in internal core formation.

3.3.3.8. *CheY*

Purifying selection operating on the surface of CheY seemed to be patchy. One of these localized patches that were predicted to be under strong purifying selection contained the residues involved in phosphoryl binding. A comparison between Selecton and PIC results revealed that the region residues Met-53 and Met-80 that are involved in maintaining hydrophobic interactions within the protein are under intensive purifying selection (Figure 3.10).

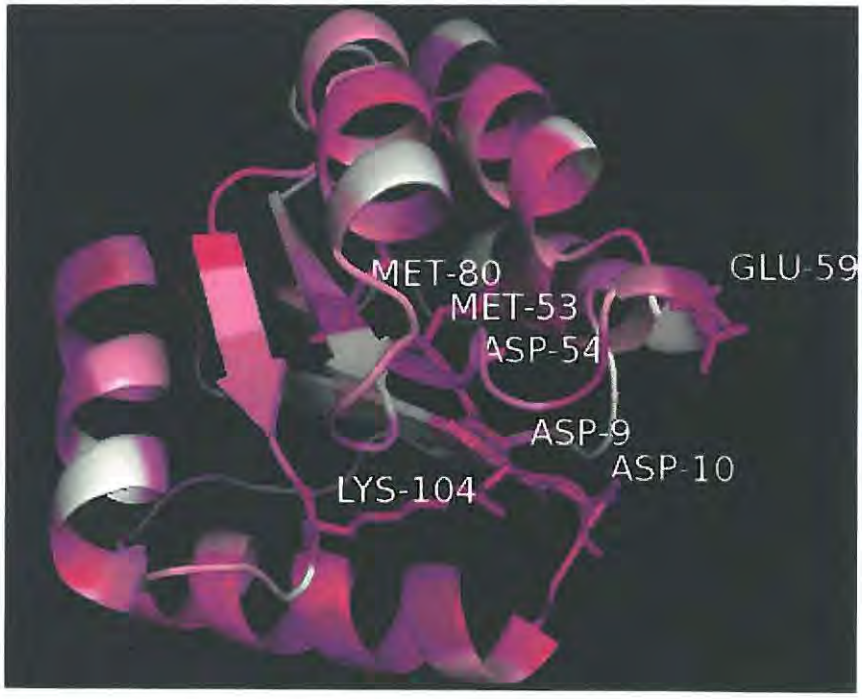


Figure 3.10. The homology model for CheY from *B. amyloliquefaciens* FZB42 shown as a cartoon and coloured according to K_a/K_s values obtained under the M7 model. The rangeof colours represents the selective forces from neutral (white) to intense purifying selection (purple). Known active sites (Asp-9, Asp-10, Asp-54 and Lys-104), core residues involved in hydrophobic interactions (Met-53 and Met-80) and a amino acid (Glu-59) in a conserved structural feature are all indicated as sticks and labelled.

3.3.3.9. *MCPs*

Under the M7 model it was predicted that the C-terminal domains of the MCPs are predominantly under purifying selection while the N-terminal domains are under neutral selection. The methylation and signalling domains are located on the C-terminal region of the MCP's, therefore it would be expected that evolutionary pressure would favour retention of function and thus mutations would not accumulate at a high rate. On the other hand, The N-terminal domains contain the ligand binding- and transmembrane regions; therefore, due to exposure to changing environmental factors, mutations in these regions would be tolerated more easily than in regions where function must be conserved to ensure the survival of the organism. Under the MEC model it was predicted that McpB site 91 and McpC sites 67, 71 and 123 may be under positive selection (Table 3.1). These sites all fall within the extracellular sensing domains of the two proteins.

Table 3.7. Selecton colour-coded results for McpA. K_a/K_s scores under the M7 model, projected onto the primary sequence of the protein. The number in the top left of each colom indicates sequence position.

McpA				
1	11	21	31	41
MKKIVQIIKQ	RSLTRSLLIS	FLAIFILFIS	TLAFIANQSA	VSTLERQMMG
51	61	71	81	91
SNLDNVQQIN	DMINTSISQK	EDGTAYFSDW	LIKDRYKPEN	QSQITDKITE
101	111	121	131	141
YMKINKEVES	IYTSETEGHF	TRVPDLQMPK	GVNPIGRDWS	EKAVENKGKV
151	161	171	181	191
VVTDPNRTAS	INTMNNDVVQ	QTKDGSGMVN	INMKIDELLK	TTKKVNIGKS
201	211	221	231	241
GMAFILTKDK	KVNAIPNRTS	GIVLNGEWAE	QLFSSKKGDF	QYKYDGQEKK
251	261	271	281	291
MAYDINGLTG	WKITGTMYLD	DIHENALPIL	HLSLLVLVLA	IVIGTIVMVL
301	311	321	331	341
IIRSITKPLK	QLNTSSAQIS	EGNLIETIEI	QSKDILGELG	KSFANMAASL
351	361	371	381	391
RSLIHAIQDS	VNNNASSSIE	LTNSAEQTSL	NDIDILLAIE	QFSNGNISQS
401	411	421	431	441
INIISAAEHI	YQMNSGLKDM	AKASAVITES	SATSAEVANS	GGKLVHQTVG
451	461	471	481	491
QMNVIDRSNK	EAIQNVRQLI	TNSKHINIL	RNINGIADQT	NLLALSANIE
501	511	521	531	541
NARAGEYGRG	FSNNAIENRK	LNVQSADSAR	EIESLIIEIN	NEINTSLGMF
551	561	571	581	591
QSVNKIVESG	LDIINQTITS	EKRISEMTNQ	INGEIQNMNA	TVEQLTAGSQ
601	611	621	631	641
QISGASEQIA	SIAKESSVSI	QDIANSNIEQ	LNSMIEISSS	SETISQSAEI
651				
LRDMTSRESI				

Table 3.8. Selecton colour-coded results for MepB. K_a/K_s scores under the M7 model, projected onto the primary sequence of the protein. The number in the top left of each colom indicates sequence position.

MepB				
1	11	21	31	41
MKKFINWCTK	ASISRKLIIS	FIVILIIPIL	VLEFSSYHTA	SNSLDDQISG
51	61	71	81	91
NAKNNIESFN	TTITSDIGAK	AKVIEFFSET	LKGSSFSKSN	ISALEEKFGQ
101	111	121	131	141
YTSIHKDVAR	IYGGTEDGGY	AQAPKENTPA	DYDPRTRTWY	KDAVKAGGTL
151	161	171	181	191
NVIDPNTANS	DGSMNVIVAK	QMQDGTGVVA	MDITIDQLLK	QMQGIKIGQK
201	211	221	231	241
GFVFITSKNK	TYNAHKDHKP	GDKVSTPWLN	EVNSKDSGII	SMTLDDQNKK
251	261	271	281	291
MAFTINKLTG	WKIEGSMELN	IIKDASQPNL	TMGMIVLAAS	IIIGGILILL
301	311	321	331	341
IVRSITSPLK	RLVRSSSQIS	GGDLTETIDI	RSKDELGELG	ASFNEMGESL
351	361	371	381	391
RGLISAIQNS	VDNNAASSEE	LTNSASQMSK	AIEHIIMAIF	QFSNGNIEQS
401	411	421	431	441
EKVDSSSVKL	NHINDGLAAV	SRTSSSITEA	SKQSKEAAGT	GEEYNQQIVG
451	461	471	481	491
QMNLINQSMQ	QAEANVKGLE	AKSKDITHIL	RNINGIADQI	SLIALNANIE
501	511	521	531	541
ANRAGEYGRG	FSMENEFNRK	LAVQSAGSAK	EIEKLIQEIT	AEINTSLHMI
551	561	571	581	591
TSMNEIMQSG	LAVIDRIKES	IQNIFGMIND	IAEKLQSMNG	TVEQLSDSSQ
601	611	621	631	641
HVSANVTDIS	DMSRESAASI	QDINASSNEFQ	LNSMLEISTS	ADTLAQMNEE
651				
LRELTKQFKI				

Table 3.9. Selecton colour-coded results for MepC. K_a/K_s scores under the M7 model, projected onto the primary sequence of the protein. The The number in the top left of each colom indicates sequence position.

MepC				
1	11	21	31	41
MFKKMHIKIA	VFVSAMLIVI	NVFLTVSSYL	TIKPMMTDEA	KRTIENVTNS
51	61	71	81	91
LGQNIILQLK	NDETVLLRLA	GGELSRSFLS	DGNKETARLF	NDELKQIGQN
101	111	121	131	141
DKYVALNNVG	TANKQMFTYP	KADFAKDYDP	TNRTWYKMAA	EKPDKNVWID
151	161	171	181	191
PNKDANTGDM	IVTASKAIQN	SGTVVCVASL	FLKLSSIQSM	VNEQKVPNKG
201	211	221	231	241
FNFLADENGS	LLAMPEKQCK	NISEDQTLKD	ITTDKEGIKE	LQGNMNVYQT
251	261	271	281	291
VKETGWKNST	QFEKDQIMGV	ADKMNTVSIF	MSLIALVITI	GLSYFLAKTI
301	311	321	331	341
IGPIQQLIAK	TNSNSAGDLI	VRAQTKSKDE	VGELIKDINQ	MVENMKIMVE
351	361	371	381	391
QMKASSENMS	DTSDQLTVIS	QFINETSGQI	NKAIEDNAAG	ATEQNSEVNI
401	411	421	431	441
INIKSENLSV	KIKGIADRVG	SIKQLSKSST	DNSNKGLDAI	GQLIMKSNEN
451	461	471	481	491
NSGTKKNLAM	LNDLFEQIKS	IFEMNTNISA	ISDQINLLNI	NASIESNEVG
501	511	521	531	541
ISGRGFAVVA	EFNRNLNQQS	ALSSGHISFT	VRLIQSFTKE	ASHNMIEISR
551	561	571	581	591
MSDFDNSAIH	ETGFVLNLII	TFMQSLNQGI	DHIYNDIQKM	SEFQAAIQEA
601	611	621	631	641

3.4. **Conclusion**

The search for positive selection operating on protein coding nucleotide sequences is fraught with difficulty. Furthermore, erroneous statistical methods and unjustifiable assumptions hinder attempts to understand evolutionary processes at the macromolecular level. It is also difficult to distinguish cases where positive selection is operating at an amino acid site from cases where purifying selection is simply relaxed (Hughes 2007). Nevertheless, this study attempted to identify individual loci in the chemotaxis proteins and MCPs where mutations may confer an adaptive advantage to various members of the *B. subtilis* group. We have employed various evolutionary models using an empirical Bayesian framework that compares alternative models that allow for positive selection to a null model that does not. Under the M8 model all amino-acid replacements are weighted almost equally, which is not an accurate depiction of biological reality. The M5 model fails to detect positively selected sites when majority of the protein is under strong purifying selection. The M7 and M8 models are alike, except for the fact that the M7 model assumes only a beta distribution with no additional categories and therefore mainly allows for purifying selection. The MEC model differs from the other models, in that it accounts for differences in amino acid replacement probabilities. The MEC model will give a higher K_a value to a position with a radical replacement than to a position with a moderate replacement, making it model more sensitive for detecting positive selection than the other models described here. The results of the MEC model suggest that the chemotaxis proteins in the cytoplasm are more selectively constrained, whilst the receptors McpB and McpC evolve more quickly. Perhaps due to increased selective pressure, as resulting from ligand-binding in the extracellular environment. It is important to note that despite the fact that the MEC model is sensitive when it comes to the detection of sites under positive selection, our results indicate that it is less reliable in detecting sites under purifying selection.

The fact that the organisms analysed in this study are so closely related could have made the detection of positive selection more difficult. A case in point is the protein CheY. Five strains of

B. amyloliquefaciens were included in a set of twelve organisms, and as the amino acid sequence for CheY is 100% identical for all the *B. amyloliquefaciens* strains, this would bias the K_a/K_s results to favour purifying selection and mask positive selection.

There are reasonable arguments against the use of K_a/K_s ratios as a test for adaptive evolution. One such argument is that if a phenotypic adaptation required a series of amino acid changing mutations in a gene, natural selection alone is insufficient to produce it unless every mutation in the series is at least somewhat advantageous (Hughes 2007). Based on the results presented in this chapter it seems unlikely that positive selection of some of the proteins involved in chemotaxis is responsible for different host preferences of the members of the *B. subtilis* group. A reasonable alternative is that variation in the methylation sites of the receptors can play a role, since these sites are important for tuning receptor sensitivity. Reversible methylation/demethylation takes place on Glu residues; however, some of the methylation sites are encoded as Gln and must be deamidated prior to methylation. There are variations in these sites between the various members of the *B. subtilis* group. It is highly unlikely that such mutations would be detected as under positive selection by the methods used here.

To test which evolutionary model most accurately predicts purifying selection the results of MEC, M5, M7 and M8, when applied to the proteins CheA. CheB, CheC, CheD, CheR, CheW and CheY, were compared to each other. According to Doron-Faigenboim *et al.* (2005) and Gertow *et al.* (2004), it would be reasonable to expect that active sites and interaction epitopes would be under the strongest purifying selection. The M5, M7 and M8 models compared quite well in their results, although there were cases where conserved active site residues were not detected to be under the most intensive degree of purifying selection as indicated by the Selecton colour codes. However the numerical results indicate that these sites were still detected to be under some degree of purifying selection. In cases where small datasets are used, the signal of purifying selection may be too weak to pick up reliably.

When attempting to detect possible active sites the MEC model performed the worst of all models employed in this study, and the conclusion is drawn that the MEC model is not suitable for inferring which residues in a protein may have a functional role, especially not in cases where

the dataset is relatively small and consist of closely related sequences. It seems to be the case that, when homologs are available for which the active sites are known, simply looking for conserved sites in the MSA is more reliable than looking for sites under purifying selection. Nevertheless, when dealing with a new protein, the analysis of site directed purifying selection using the M5, M7 or M8 models can be somewhat useful for identifying putative active sites. Since not much is known about the specific residues that are involved in the interactions between the chemotaxis proteins of the *B. subtilis* group, the structures were coloured according to Selecton results (for M7 model), and compared with the results from PIC. This will be discussed in Chapter 4.

By comparing the results from PIC with that of Selecton we were able to determine which residues that are involved in intra- and inter-protein interactions are also under evolutionary constraints within members of the *B. subtilis* group. The fact that there was a large overlap between the residues identified by PIC as important for intra- and inter protein interaction and those identified by the M7 models as being under strong purifying selection shows that the M7 model is sufficiently reliable for the prediction of functionally important residues. The identified residues could prove to be critical for retention of the protein fold as well as for protein-protein interactions. However mutational experiments will be needed to confirm this hypothesis. The biologically important sites identified in this study can serve as guidelines in designing site-directed mutagenesis experiments and aid in the understanding of the basis for residue conservation in homologous proteins.

CHAPTER 4

4. ANALYSIS OF MODEL PROPERTIES REPRESENTING PROTEIN-PROTEIN INTERACTIONS

Since not much is known about the specific residues that are involved in the interactions between the chemotaxis proteins of the *B. subtilis* group, the objective of this chapter is to describe the steps taken to identify interacting residues using an approach that combines the results of PIC with those of Selecton. PyMol session files showing the interactions between protein partners, with residues coloured according to the M7 model scores can be found on the disk containing supplementary information in the folder titled "Pymol_session_files".

4.1. Introduction

Purifying selection acts to eliminate selectively deleterious mutations, such as those that alter residues that play a role in protein activity and structure. To identify sites under varying degrees of purifying and neutral selection in the chemotaxis proteins and MCPs from members of the *B. subtilis* group, we analysed the K_a/K_s ratios under the M7 model. For more detail on this analysis see Chapter 3. Residues that play a role in protein-protein interactions were determined with the Protein Interaction Calculator (PIC) server (see Chapter 2). In each case the structures and sequences from *B. amyloliquefaciens* FZB42 were used as references. Unless otherwise indicated all residues are numbered according to their position in *B. amyloliquefaciens* FZB42.

4.2. Methodology

To obtain an idea of the spatial distribution of sites under purifying selection, the results from the M7 model described in Chapter 3 were mapped onto the homology models of protein complexes built in Chapter2.

4.3. Results

Residues predicted to play an important role in protein-protein interactions for some of the interaction pairs described in Chapter 2 are listed in the following tables: Table 4.1, Table 4.2 Table 4.3, Table 4.4, Table 4.5 and Table 4.6. A subset of these residues that are also under purifying selection was mapped onto homology models in order to visualise their spatial distributions (Figure 4.1, Figure 4.2, Figure 4.3, Figure 4.4, Figure 4.5 and Figure 4.6).

Table 4.1. Residues involved in protein-protein interactions between the P4 and P5 domains of CheA and CheW. Residues that are highlighted are under strong purifying selection.

CheA Residue	CheW Residue	Comment
Hydrophobic interactions within 5Å		
Ile-497	Val-100	Both residues conserved in all members
Ile-497	Ile-12	Both residues conserved in all members. CheW Ile-12 under intensive strong purifying selection.
Ile-497	Val-25	Both residues conserved in all members
Ile-497	Val-28	Both residues conserved in all members
Phe-553	Pro-41	Both residues conserved in all members
Phe-596	Pro-41	Both residues conserved in all members
Ile-598	Pro-41	CheW Pro-41 conserved in all members. CheA Ile-598 replaced by Val in organisms: R032; 4580; 6633; H642; 3610; SMY and S168
Leu-638	Val-40	Both residues conserved in all members
Leu-638	Val-50	Both residues conserved in all members
Leu-638	Ile-57	CheA Leu-638 conserved in all members. CheW Ile-57 replaced by Val in organisms: R032; 4580; 6633; H642; 3610; SMY and S168
Tyr-641	Val-43	Both residues conserved in all members
Tyr-641	Tyr-46	Both residues conserved in all members
Tyr-641	Ile-47	Both residues conserved in all members. CheW Ile-47 under intensive strong purifying selection.
Tyr-641	Pro-59	Both residues conserved in all members
Tyr-641	Leu-61	CheA Tyr-641 conserved in all members. CheW Leu-61 replaced by an Ile in: R032; 4580; 6633; H642; 3610; SMY and S168
Leu-642	Val-40	Both residues conserved in all members
Leu-642	Val-43	Both residues conserved in all members
Leu-642	Ile-47	Both residues conserved in all members. CheW Ile-47 under intensive strong purifying selection.
Leu-642	Val-50	Both residues conserved in all members
Leu-642	Pro-59	Both residues conserved in all members
Val-645	Val-40	Both residues conserved in all members
Val-645	Val-43	Both residues conserved in all members
Ile-648	Val-40	Both residues conserved in all members
Ala-651	Ile-57	CheA Ala-651 conserved in all members. CheW Ile-57 replaced by Val in organisms: R032; 4580; 6633; H642; 3610; SMY and S168

CheA Residue	CheW Residue	Comment
Val-659	Val-50	Both residues conserved in all members
Val-659	Ile-57	CheA Ala-651 conserved in all members. CheW Ile-57 replaced by Val in organisms: R032; 4580; 6633; 11642; 3610; SMY and S168
Leu-661	Val-40	Both residues conserved in all members
Leu-661	Pro-41	Both residues conserved in all members
Protein-Protein Main Chain-Side Chain Hydrogen Bonds		
Tyr-641	Glu-44	CheA Tyr-641 is conserved in all members. CheW Glu-44 is replaced by Ala in organism: R032
Asp-492	Lys-10	Both residues are under strong purifying selection.
Tyr-641	Ile-47	Both residues conserved in all members. CheW Ile-47 is under strong purifying selection.
Protein-Protein Side Chain-Side Chain Hydrogen Bonds		
Asp-398	Arg-54	Both residues conserved in all members. CheA Asp-398 is under strong purifying selection.
Ionic interactions within 6Å		
Asp-398	Arg-54	Both residues conserved in all members. CheA Asp-398 is under strong purifying selection.
Asp-492	Lys-10	Both residues are under strong purifying selection.
Asp-496	Lys-29	CheA Asp-496 is under strong purifying selection. CheW Lys-29 is replaced by Met in organism: 4580.
Aromatic-Aromatic Interactions within 4.5 and 7 Ångstroms		
Tyr-641	Tyr-46	Both residues conserved in all members.

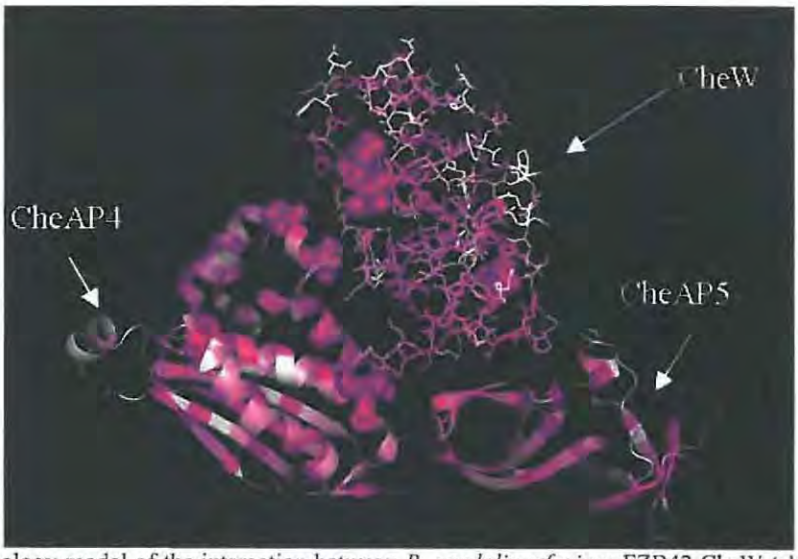


Figure 4.1 A homology model of the interaction between *B. amyloliquefaciens* FZB42 CheW (shown as sticks) and CheA domains P4 and P5 (shown as cartoons). The proteins are coloured according to K_a/K_s values obtained under the M7 model. The rangeof colours represents the selective forces from neutral (white) to intense purifying selection (purple). The residues indicated as spheres are under intensive purifying selection and were determined by PIC as important for protein-protein interactions.

Table 4.2. Residues involved in protein-protein interactions between CheC and CheD. Residues that are highlighted are under strong purifying selection.

CheC Residue	CheD residue	Comment
Hydrophobic interactions within 5Å		
Phe-50	Ile-13	Both residues under strong purifying selection.
Pro-63	Val-159	Both residues conserved in all members.
Phe-68	Phe-102	Both residues conserved in all members. CheC Phe-68 under strong purifying selection.
Phe-80	Phe-102	Both residues conserved in all members. CheC Phe-80 under strong purifying selection.
Phe-101	Val-159	Both residues conserved in all members. CheC Phe-101 under strong purifying selection.
Met-150	Val-159	Both residues conserved in all members. CheC Met-150 under strong purifying selection.
Ala-153	Met-101	Both residues conserved in all members. CheD Met-101 under strong purifying selection.
Val-154	Met-101	Both residues conserved in all members. CheD Met-101 under strong purifying selection.
Val-154	Phe-102	Both residues conserved in all members.
Leu-159	Ile-13	Both residues conserved in all members. CheD Ile-13 under strong purifying selection.
Met-160	Ile-113	CheC Met-160 replaced by Ile in organism: 4580. CheD Ile-113 replaced by Val in organism: 4580.
Met-160	Met-52	CheC Met-160 replaced by Ile in organism: 4580. CheD Met-52 under strong purifying selection.
Met-160	Leu-53	CheC Met-160 replaced by Ile in organism: 4580. CheD Leu-53 conserved in all members.
Val-165	Ile-13	CheC Val-165 replaced by Leu in organism: R032. CheD Ile-13 under strong purifying selection.
Val-165	Met-52	CheC Val-165 replaced by Leu in organism: R032. CheD Met-53 under strong purifying selection.
Protein-Protein Main Chain-Side Chain Hydrogen Bonds		
Gln-164	Lys-60	CheD Lys-60 is under strong purifying selection. CheC Gln-165 is replaced by Pro in organism: R032
Glu-157	Ser-32	Both residues conserved in all members
Glu-145	Lys-103	CheD Lys-103 is under strong purifying selection. CheC Glu-145 is replaced by Asp in organism: R032.
Protein-Protein Side Chain-Side Chain Hydrogen Bonds		
Asp-149	Met-101	Both residues under strong purifying selection.
Met-160	Ser-32	CheC Met-160 replaced by Ile in organism: 4580. CheD Ser-32 conserved in all members.
Tyr-195	Ser-59	CheC Tyr-195 replaced by Phe in organism: R032. CheD Ser-59 replaced by Ala in organism: R032.
Ionic interactions within 6Å		
Glu-61	Arg-157	CheC Glu-61 replaced by Asp in organisms: R032; 4580; 6633; 11642; 3610; SMY and S168. CheD Arg-157 conserved in all members.
Aromatic-Aromatic Interactions within 4.5 and 7 Ångstroms		
Phe-68	Phe-102	Both residues conserved in all members. CheC Phe-68 under strong purifying selection.
Phe-80	Phe-102	Both residues conserved in all members. CheC Phe-80 under strong purifying selection.

CheC Residue	CheD residue	Comment
Cation-Pi Interactions within 6 Ångstrom.		
Tyr-195	Lys-60	CheC Tyr-195 replaced by Phe in organism: R032. CheD lys-60 under strong purifying selection.

Figure 4.2. A homology model of the interaction between *B. amyloliquefaciens* FZB42 CheC and CheD. The proteins are coloured according to K_a/K_s values obtained under the M7 model. The rangeof colours represents the selective forces from neutral (white) to intense purifying selection (purple). The residues indicated as spheres are under intensive purifying selection and were determined by PIC as important for protein-protein interactions.

Table 4.3. Residues involved in protein-protein interactions between the P1 domain of CheA and CheY. Residues that are highlighted are under strong purifying selection.

CheA residue	CheY residue	Comment
Hydrophobic interactions within 5Å		
Tyr-4	Phe-13	Both residues conserved in all members.
Tyr-4	Met-1	Both residues conserved in all members. CheY Met-17 under strong purifying selection.
Val-7	Ala-12	CheA Val-7 is replaced with Ile in organism R302. CheY Ala-12 is conserved in all members.
Val-7	Phe-13	CheA Val-7 is replaced with Ile in organism R302. CheY Phe-13 is conserved in all members.
Val-7	Met-16	CheA Val-7 is replaced with Ile in organism R302.CheY Met-16 is under strong purifying selection.
Phe-8	Phe-13	Both residues conserved in all members.

CheA residue	CheY residue	Comment
Met-49	Phe-13	Both residues conserved in all members. CheA Met-49 is under strong purifying selection.
Ala-51	Phe-106	Both residues conserved in all members.
Met-53	Phe-13	Both residues conserved in all members. CheA Met-53 is under strong purifying selection.
Protein-Protein Main Chain-Main Chain Hydrogen Bonds		
Thr-52	Gln-107	Both residues conserved in all members.
Protein Main Chain-Side Chain Hydrogen Bonds.		
Met-53	Gln-107	Both residue conserved in all members. CheA Met-53 under strong purifying selection.
Protein Side Chain-Side Chain Hydrogen Bonds.		
Gln-3	Met-16	Both residues conserved in all members. CheY met-16 under strong purifying selection
Aromatic-Aromatic Interactions within 4.5 and 7 Ångstroms		
Phe-8	Phe-13	Both residues conserved in all members.
Aromatic-Sulphur Interactions within 5.3 Ångstroms.		
Tyr-4	Met-17	Both residues conserved in all members. CheY Met-17 under strong purifying selection.
Met-53	Phe-13	Both residues conserved in all members. CheA Met-53 under strong purifying selection.

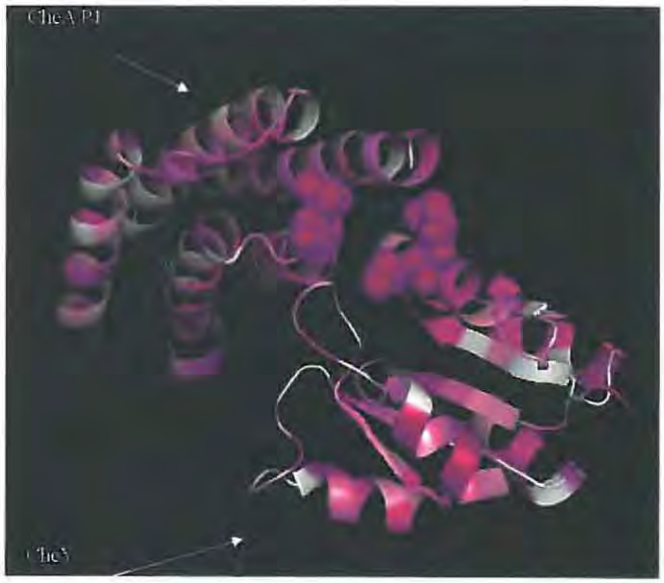


Figure 4.3. A homology model of the interaction between *B. amyloliquefaciens* FZB42 CheY and P1 domain of CheA. The proteins are coloured according to K_a/K_s values obtained under the M7 model. The rangeof colours represents the selective forces from neutral (white) to intense purifying selection (purple). The residues indicated as spheres are under intensive purifying selection and were determined by PIC as important for protein-protein interactions.

Table 4.4. Residues involved in protein-protein interactions between the P1 domain of CheA and CheB. Residues that are highlighted are under strong purifying selection.

CheA residues	CheB residues	Comment
Hydrophobic Interactions within 5 Ångstroms		
Tyr-4	Phe-12	Both residues conserved in all members.
Val-7	Ala-11	CheA Val-7 is replaced with Ile in organism R302.
Val-7	Phe-12	CheA Val-7 is replaced with Ile in organism R302.
Phe-8	Phe-12	Both residues conserved in all members.
Met-49	Phe-12	Both residues conserved in all members. CheA Met-49 is under strong purifying selection.
Ala-51	Ile-110	Both residues conserved in all members. CheB Ile-110 is under strong purifying selection.
Met-53	Phe-12	Both residues conserved in all members. CheA Met-53 is under strong purifying selection.
Ala-59	Ile-110	Both residues conserved in all members. CheB Ile-110 is under strong purifying selection.
Protein-Protein Main Chain-Main Chain Hydrogen Bonds		
Ala-51	Ala-109	CheA Ala-51 is conserved in all members. CheB Ala-109 replaced by Ser in organisms: R032; 4580; 6633; 11642; 3610; SMY and S168.
Ala-51	Ile-110	Both residues conserved in all members. CheB Ile-110 is under strong purifying selection.
Protein-Protein Main Chain-Side Chain Hydrogen Bonds		
Gly-48	Ser-107	Both residues conserved in all members.
Glu-11	Ala-11	Both residues conserved in all members. CheA Glu-11 is under strong purifying selection.
Glu-11	Phe-12	Both residues conserved in all members. CheA Glu-11 is under strong purifying selection.
Protein-Protein Side Chain-Side Chain Hydrogen Bonds		
His44	Glu-56	Both residues conserved in all members. CheB Glu-56 is under strong purifying selection.
Thr-52	Met-16	CheA Thr-52 conserved in all members. CheB Met-16 replaced by a Leu in organism: R032.
Ionic interactions within 6Å		
His-44	Glu-56	Both residues conserved in all members. CheB Glu-56 is under strong purifying selection.
Aromatic-Aromatic Interactions within 4.5 and 7 Ångstroms		
Tyr-4	Phe-12	Both residues conserved in all members.
Phe-8	Phe-12	Both residues conserved in all members.
Aromatic-Sulphur Interactions within 5.3 Ångstroms		
Met-49	Phe-12	Both residues conserved in all members. CheA Met-49 is under strong purifying selection
Met-53	Phe-12	Both residues conserved in all members. CheA Met-53 is under strong purifying selection.

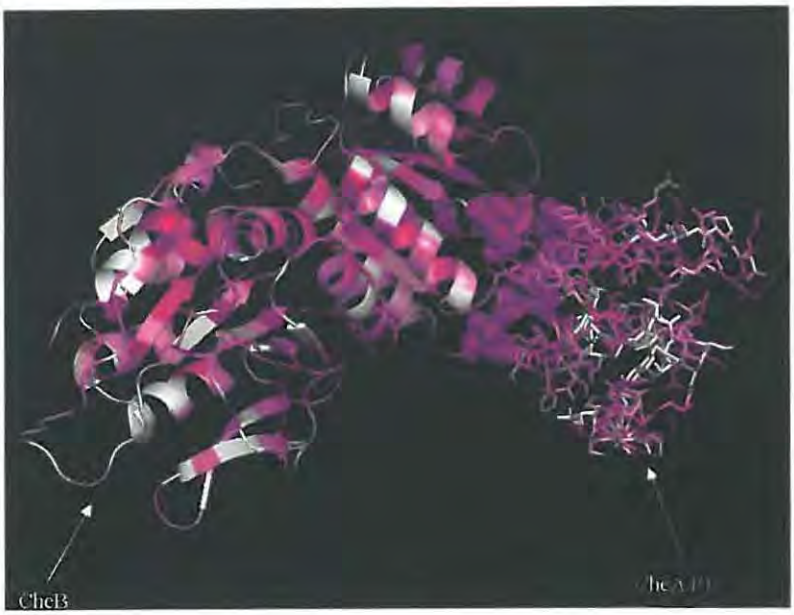


Figure 4.4. A homology model of the interaction between *B. amyloliquefaciens* FZB42 CheB and P1 domain of CheA. The proteins are coloured according to K_a/K_s values obtained under the M7 model. The rangeof colours represents the selective forces from neutral (white) to intense purifying selection (purple). The residues indicated as spheres are under intensive purifying selection and were determined by PIC as important for protein-protein interactions.

Table 4.5. Residues involved in protein-protein interactions between the P2 domain of CheA and CheY. Residues that are highlighted are under strong purifying selection.

CheY Residue	CheA Residue	Comment
Hydrophobic Interactions within 5 Ångstroms		
Met-84	Met-178	CheY Met-84 is under strong purifying selection. CheA Met-178 is replaced by Leu in organisms: R032 and 4580.
Met-84	Leu- 179	Both residues are under strong purifying selection.
Val-90	Leu-179	Both residues conserved in all members.
Val-90	Val-182	Both residues conserved in all members.
Ile-91	Val-182	Both residues conserved in all members.
Ile-91	Tyr-185	Both residues conserved in all members.
Ile-94	Val-182	Both residues conserved in all members.
Ile-94	Tyr-185	Both residues conserved in all members.
Ile-94	Met-186	Both residues conserved in all members. CheA Met-186 under strong purifying selection.
Phe-101	Leu-179	Both residues conserved in all members
Phe-101	Val-182	Both residues conserved in all members.
Phe-101	Met-186	Both residues conserved in all members. CheA Met-186 under strong purifying selection.

CheY Residue	CheA Residue	Comment
Val-103	Met-178	CheY Val-103 conserved in all members. CheA Met-178 is replaced by Leu in organisms: R032 and 4580.
Val-103	Leu-179	Both residues conserved in all members
Pro-105	Met-178	CheY Pro-105 conserved in all members. CheA Met-178 is replaced by Leu in organisms: R032 and 4580.
Protein-Protein Main Chain-Side Chain Hydrogen Bonds		
Lys-117	Ile-236	CheY Lys-117 under strong purifying selection. CheA Ile-236 replaced with Val in organisms: 4580; 6633; H642; 3610; SMY and S168.
Phe-101	Arg-183	Both residues conserved in all members
Phe-101	Arg-183	Both residues conserved in all members
Glu-113	Glu-238	Both residues under strong purifying selection.
Protein-Protein Side Chain-Side Chain Hydrogen Bonds		
Asp-100	Met-186	Both residues conserved in all members. CheA Met-186 under strong purifying selection.
Arg-110	Glu-238	Both residues conserved in all members. CheA Glu-238 under strong purifying selection.
Asp-100	Arg-183	Both residues conserved in all members
Gln-87	Glu-210	Both residues conserved in all members. CheA Glu-210 under strong purifying selection.
Glu-113	Ser-237	Both residues conserved in all members. CheY Glu-113 under strong purifying selection.
Ionic Interactions within 6 Ångstroms		
Lys-99	Glu-189	Both residues conserved in all members. CheA Glu-189 under strong purifying selection.
Asp-100	Arg-183	Both residues conserved in all members.
Arg-110	Glu-238	Both residues conserved in all members. CheA Glu-238 under strong purifying selection.
Glu-113	Arg-183	Both residues conserved in all members. CheY Glu-113 under strong purifying selection.
Protein-Protein Aromatic-Sulphur Interactions		
Phe-101	Met-186	Both residues conserved in all members. CheA Met-186 under strong purifying selection.
Cation-Pi Interactions within 6 Ångstroms		
Phe-101	Arg-183	Both residues conserved in all members.

Figure 4.5 A homology model of the interaction between *B. amyloliquefaciens* FZB42 CheY and P2 domain of CheA. The proteins are coloured according to K_a/K_s values obtained under the M7 model. The rangeof colours represents the selective forces from neutral (white) to intense purifying selection (purple). The residues indicated as spheres are under intensive purifying selection and were determined by PIC as important for protein-protein interactions.

Table 4.6. Residues involved in protein-protein interactions between the P2 domain of CheA and the N-terminal domain of CheB. Residues that are highlighted are under strong purifying selection.

CheB residue	CheA residue	Comment
Hydrophobic Interactions within 5 Ångstroms		
Ile-92	Val-182	Both residues conserved in all members. CheB Ile-92 is under strong purifying selection.
Ile-92	Tyr-185	Both residues conserved in all members.
Leu-95	Val-182	Both residues conserved in all members.
Leu-95	Tyr-185	Both residues conserved in all members.
Leu-95	Met-186	Both residues conserved in all members. Met-186 is under strong purifying selection
Phe-102	Leu-179	Both residues conserved in all members
Phe-102	Val-182	Both residues conserved in all members
Phe-102	Met-186	Both residues conserved in all members. Met-186 is under strong purifying selection
Pro-106	Met-178	CheB Pro-106 is conserved in all members. CheA Met-178 is replaced by Leu in organisms: R032 and 4580.
Protein-Protein Main Chain-Side Chain Hydrogen Bonds		
Phe-102	Arg-183	Both residues conserved in all members
Glu-123	Ser-237	CheB Glu-123 is under strong purifying selection. Both residues conserved in all members.
Gln-120	Glu-238	CheB Gln-120 is replaced by a Met in organism: 4580. CheA Glu-238 is under strong purifying selection.

CheB residue	CheA residue	Comment
Protein-Protein Side Chain-Side Chain Hydrogen Bonds		
Asp-101	Met-186	Both residues are under strong purifying selection.
Thr-104	Met-178	CheB Thr-104 is replaced by an Ala in organism: 4580. CheA Met-178 is replaced by Leu in organisms: R032 and 4580.
Gln-120	Glu-238	CheB Gln-120 is replaced by a Met in organism: 4580. CheA Glu-238 is under strong purifying selection.
Gln-83	Met-178	CheB Gln-83 is replaced by a Leu in organism: 4580. CheA Met-178 is replaced by Leu in organisms: R032 and 4580.
Asp-101	Arg-183	Both residues conserved in all members
Ionic Interactions within 6 Ångstroms		
Lys-88	Glu-210	CheB Lys-88 is replaced by an Ala in organism: 4580. CheA Glu-210 is under strong purifying selection.
Asp-101	Arg-183	Both residues conserved in all members. CheB Asp-101 is under strong purifying selection.
Arg-124	Glu-238	Both residues are conserved in all members. CheA Glu-238 is under strong purifying selection.

Figure 4.6. A homology model of the interaction between *B. amyloliquefaciens* FZB42 CheB N terminal domain and P2 domain of CheA. The proteins are coloured according to K_a/K_s values obtained under the M7 model. The rangeof colours represents the selective forces from neutral (white) to intense purifying selection (purple). The residues indicated as spheres are under intensive purifying selection and were determined by PIC as important for protein-protein interactions.

4.4. **Conclusion**

PIC identified residues that may be involved in protein-protein interactions. These residues were then compared with the results from the M7 model employed by the Selecton server. Residues that were both under purifying selection and also involved in inter protein interactions were mapped to the homology models. The majority of putative interaction residues as determined by PIC are conserved throughout members of the *B. subtilis* group, yet they were not predicted to be under strong purifying selection. Nevertheless, our results show that by combining structural information with information about selection operating at each amino acid site, it is possible to make a conservative prediction of which amino acids are important for forming quaternary structures between specific proteins within a given group of organism.

CHAPTER 5

5. A DETAILED ANALYSIS OF THE INTERACTIONS BETWEEN THE HISTIDINE KINASE CHEA AND ITS COGNATE RESPONSE REGULATORS CHEY AND CHEB

The aim of this chapter is to provide an overview of the interaction between CheB and the P1 domain of CheA. In addition, evidence for phospho-CheB competing with CheY to bind to the P2 domain of CheA is also presented. For practical reasons the focus will mainly be on residues involved in hydrophobic interactions. However, detailed results of hydrogen bonds, ionic interactions, aromatic interactions and so forth can be found on the disk containing supplementary material.

5.1. Introduction

Classical two component systems found in bacteria typically consist of a histidine protein kinase (HK) and a response regulator (RR). The operational status of these two components is controlled by three phosphotransfer reactions: (i) the presence of an environmental stimulus triggers the ATP-dependent autophosphorylation of a conserved histidine residue on the histidine containing phosphotransfer (Hpt) domain of the HK. (ii) The transfer of the phosphor group to an aspartate residue within the receiver domain of the RR. (iii) Dephosphorylation of the RR to return the system to prestimulus state. When the RR is phosphorylated it interacts with genes or proteins to trigger the appropriate cellular responses. Dephosphorylation of the RR can be accomplished by autophosphatase activity of the RR, phosphatase activity of the HK or by an external phosphatase (Parkinson & Kofoid 1992, Parkinson 1993, Volz 1993, Casino *et al.* 2009, Bell *et al.* 2010). The lifetime of a phospho-RR is correlated with its function. The sporulation transcription regulator Spo0A needs to be active long enough to influence the genes under its control, and therefore have a half-life of the order of several hours. On the other hand, CheY must be activated and deactivated rapidly to influence flagellar rotation and has a half-life of less than two minutes (Muff & Ordal 2007). A single microorganism may contain as many as 200 – 300 HH-RR pairs. However, each HK only interacts with one or two RRs, indicating extraordinary partner specificity (Casino *et al.* 2009). The highly specific nature of

phosphotransfer reactions between HH-RR pairs provides a control mechanism to prevent unwanted crosstalk (Bell *et al.* 2010) Mistaken binding between an activated HK and the wrong RR can have disastrous effects such as incorrect gene regulation (Hoch & Varughese 2001). A unique feature of the chemotaxis system HK, CheA, is the presence of the P2 domain, which is absent in all other known HKs. It is believed that this domain is responsible for the exceptionally fast phosphotransfer reaction that occurs between the Hpt domain (also known as the P1 domain) of CheA and CheY (Stewart & Van Bruggen 2004).

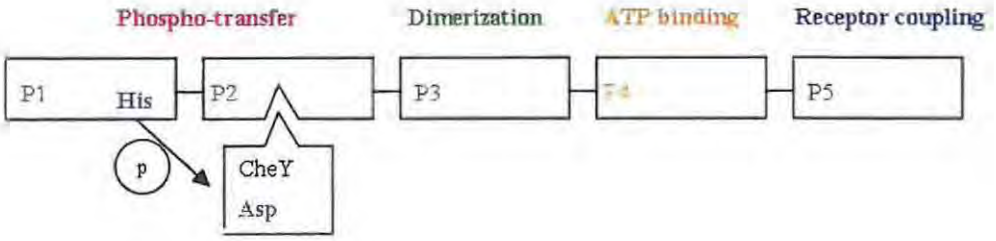


Figure 5.1 Schematic diagram of the domains of CheA The functional roles of each domain were determined experimentally using various methods. The Histidine in P1 that transfers the phosphoryl group to CheY is indicated, as well as the accepting Aspartate in CheY. CheY docking at P2 enhances the rate of phosphoryl transfer from P1.

Despite the availability of large quantities of biochemical and mutational information on two-component signal transduction, the signalling process is not well understood on a structural level The majority of HK-RR structures that have been solved until now are representatives of the class 1 HKs (Casino *et al.* 2009). The chemotaxis system is centred on a class II HK (Dutta *et al.* 1999), therefore high resolution structures of functional complexes of typical class II HK-RR pairs must be solved to shed light on the process of signal transduction in the chemotaxis system. Structures that show CheY interacting with the P1 and P2 domains of CheA have been solved for *R. sphaeroides* and *T. maritima* respectively. Nevertheless, no structure showing CheB interacting with CheA has been solved to date. As part of this work, homology models were inferred for the interactions between CheY and the P1 domain of CheA, CheY and the P2 domain of CheA, and for CheB and the P1 domain of CheA. Furthermore, the possibility that phospho-CheB can interact with the P2 domain of CheA was also explored by creating a homology model for the N-terminal domain ($CheB_N$) interacting with P2. The objective was to determine the similarities and differences in the HK-RR intermolecular recognition interactions to better understand the exquisite partner specificity that exists between these proteins. Please

note that, unless otherwise indicated, all residues are numbered according to their position in the *B. amyloliquefaciens* ssp. *plantarum* FZB42.

5.1.1. *General topology of response regulators*

A protein belonging to the class of response regulators usually have a C-terminal output domain, which is most often a DNA-binding domain, the activity of which is regulated by the phosphorylation of an N-terminal receiver domain (Parkinson & Kofoid 1992, Tzeng & Hoch 1997). The C-terminal effector domain of CheB functions as a methyl-esterase and glutamine-deamidase. In its unphosphorylated state the N-terminal domain blocks the catalytic active site on the C-terminal domain, but upon phosphorylation the protein undergoes a conformational change that increases the solvent accessibility of this area (West *et al.* 1995, Jurica & Stoddard 1998, Anand *et al.* 2000, Hughes *et al.* 2001). Unlike CheB, CheY is a single domain receiver protein. The N-terminal domain of CheB ($CheB_N$) and CheY display a (β/α) 5 scaffold topology, where five β-strands are arranged in a parallel sheet with the topology $\beta2$-$\beta1$-$\beta3$-$\beta4$-$\beta5$. Five amphiphilic α-helices are clustered into two groups on opposing sides of the central β sheet. This arrangement is believed to be typical of all receiver domains (Volz 1993, Tzeng & Hoch 1997, Szurmant & Hoch 2010). The active site is comprised of the phosphor accepting aspartate on $\beta3$. An acidic pocket essential for Mg^{2+} binding is formed by four invariant residues: an aspartate pair on the same face of the turn following β1, the strand $\beta4$ holding a Thr/Ser in place and a Lys on β5 that forms a hydrogen bond with the phosphor binding site (Sanders *et al.* 1989, Lukat *et al.* 1990, Volz 1993, Bellsolell *et al.* 1996, Hoch & Varughese 2001,). $CheB_N$ shares a high sequence identity with CheY (37.7%, as calculated with Jalview); therefore it is reasonable to assume that similar residues are involved in the interactions between these two proteins and CheA.

5.1.2. *Interaction of CheY and CheB with the P1 domain in B. amyloliquefaciens*

Up to the present, the only determined structure of an interaction between the histidine phospho transfer-domain of a class II HK and a cognate response regulator comes from the *R sphaeroides* interacting partners $CheA_3P1$-$CheY_6$ (PDB ID: 3KYJ). Comparison with solved structures centred on class I HK revealed several structural similarities. Based on the structures of solved HK-RR interaction pairs a typical interaction comprises an alpha-1 helix of the RR contacting with a C-terminal helical region containing the phosphorus-binding histidine of the HK

(Szurmant *et al.* 2008, Casino *et al.* 2009, Bell *et al.* 2010). The interface between $CheA_3P1$ and $CheY_6$ was shown to be dominated by hydrophobic interactions, with only one hydrogen bond and no salt bridges formed between the interacting partners. This weak interaction explains the ephemeral nature of the complexes. The largest contribution to the binding interfaces comes from a protruding methionine finger on the N-terminal end of $CheY_6$ helix α1, which is accommodated by the hydrophobic pocket formed by helices αA and αB of $CheA_3P1$ (Bell *et al.* 2010). Homology models for *B. amyloliquefaciens* FZB42 CheY-CheAP1 and CheB-CheAP1 (Figure 5.2) were built based on the template 3KYJ to determine which residues are involved in HK-RR recognition in this organism.

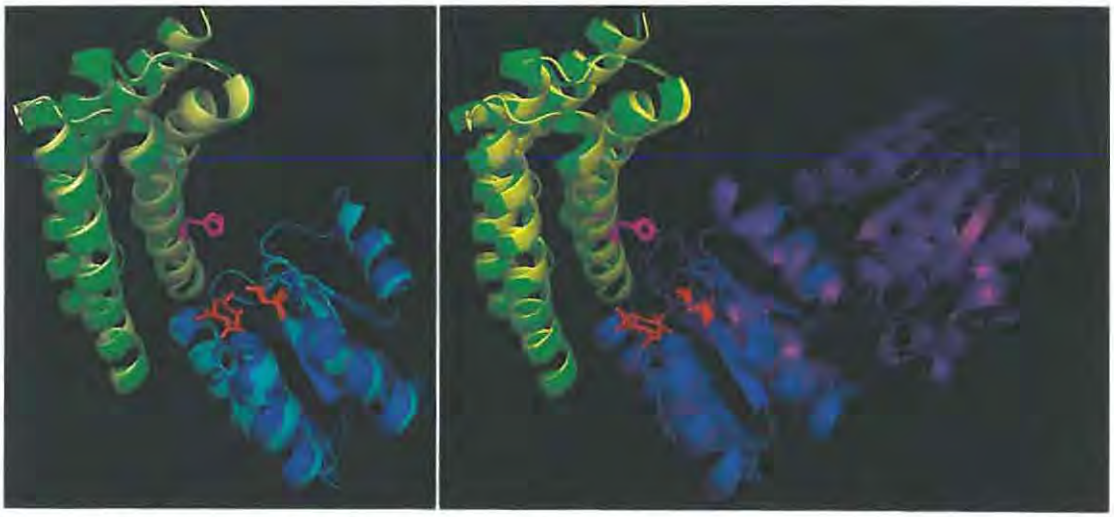


Figure 5.2. Left: Superposition of the homology model for *B. amyloliquefaciens* FZB42 CheY-CheAP1 and the *R. sphaeroides* structure $CheY_6$-CheA (PDB ID: 3KYJ). Colour coding for 3KYJ is as follows: $CheY_6$ is shown in blue and $CheA_3P1$ is shown in green. For the homology model CheY is shown in teal and CheAP1 in pale yellow. The phospho-accepting histidine residue is indicated as a magenta stick on both structures the three aspartate residues that form part of the active site of CheY are indicated as red sticks on both structures. The RMSD between the experimental structure and the homology model is 1.388Å. **Right:** Superposition of the homology model CheB-CheAP1 and 3KYJ. Homology model of CheB is shown in purple and CheAP1 in yellow. Experimental structure is coloured the same as before. Active sites are indicated in the same manner as before. The RMSD between the experimental structure and the homology model is 2.185Å.

Sequence and structural analysis reveal that an alanine on α1 on $CheY_6$ and a glycine on the loop region following αB of $CheA_3P1$ from *R. sphaeroides* are conserved in the *B. subtilis* group. Other interaction residues are conserved within the *B. subtilis* group, but are of a different amino acid type than in *R. sphaeroides* (Figure 5.5, Figure 5.6 and Figure 5.7). Based on the homology and PIC results it is clear that the largest contribution to the binding interface comes from a phenyl ring on the N-terminal part of α1 (Phe-12 in CheB and Phe-13 in CheY) that inserts into a hydrophobic pocket formed by Tyr-4, Val-7 and Phe-8 on αA, and Met-49 and Met-53 on αB (Figure 5. 3 and Figure 5. 4). The α1 Phenyl is conserved in both RRs for all members of the *B. subtilis* group analysed in this study. The CheA residues involved in this interaction are all conserved with the exception of Val-7, which is replaced by an Ile in *B. pumilus* SAFR-032. According to the PIC results, CheB binds to P1 more strongly than CheY does, as the CheB interaction involves more hydrogen bonds, and additional ionic and aromatic interactions (See Chapter 4).

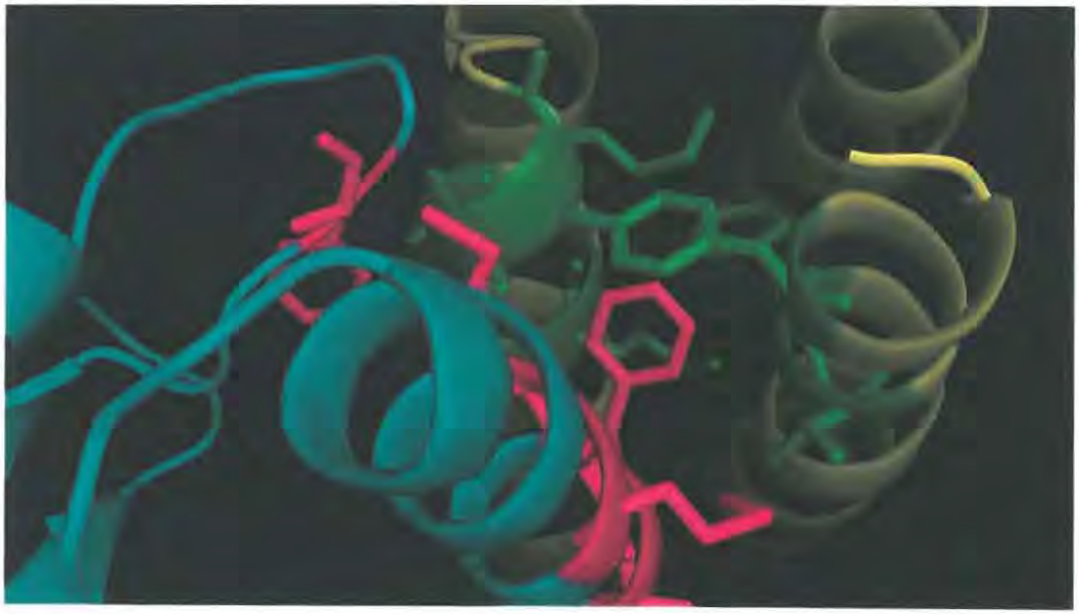


Figure 5.3. A close-up of the interaction between CheY (teal) and CheA P1domain (pale yellow). Residues that are indicated as sticks are involved in the hydrophobic interaction. Pink residues of CheY interact with green residues of CheA

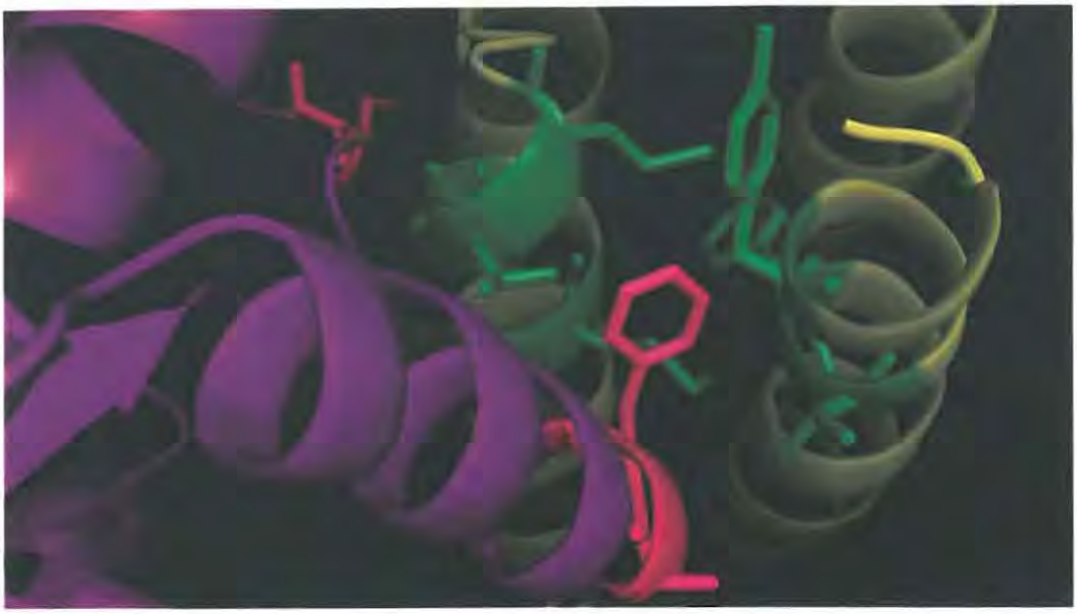


Figure 5.4 A close-up of the interaction between CheB (purple) and the CheA P1 domain (pale yellow). Residues that are indicated as sticks are involved in the hydrophobic interaction. Pink residues of CheB interact with green residues of CheA.

Figure 5.5. Multiple sequence alignment of a section of CheAP1 from the *B. subtilis* group and CheA₁P1 from *R. sphaeroides*. The sequence for *B. amyloliquefaciens* are in the red dashed-line box, and its abbreviation "ZB42" is highlighted in grey. Residues in blue were identified in the *R. sphaeroides* structure 3KYJ to play an important role in the hydrophobic interaction with CheY. Residues coloured in green play a role in inter-protein hydrophobic interactions with CheY and CheB. The alignment was generated by Promals3D and viewed in Jalview.

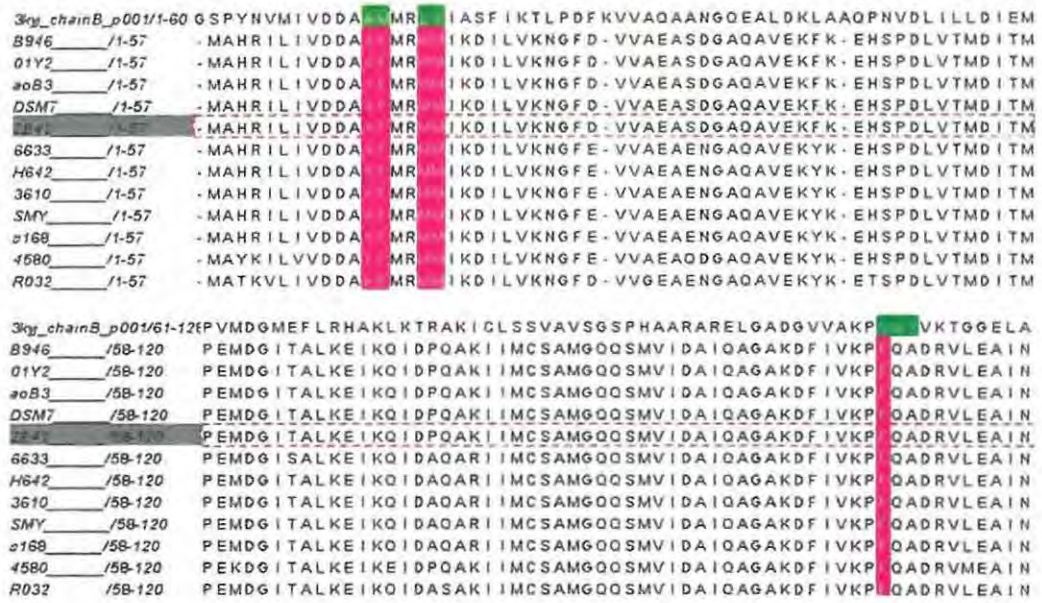


Figure 5.6. Multiple sequence alignment of a section of CheY from the *B. subtilis* group and $CheY_6$ from *R. sphaeroides*. The sequence for *B. amyloliquefaciens* are in the red dashed-line box, and its abbreviation "ZB42" is highlighted in grey Green residues were identified in the *R. sphaeroides* structure 3KYJ to play an important role in the hydrophobic interaction with $CheA_3P1$. Pink residues are conserved in CheY within the *B. subtilis* group and were identified by PIC as important for inter-protein hydrophobic interactions with the P1 domain of CheA.

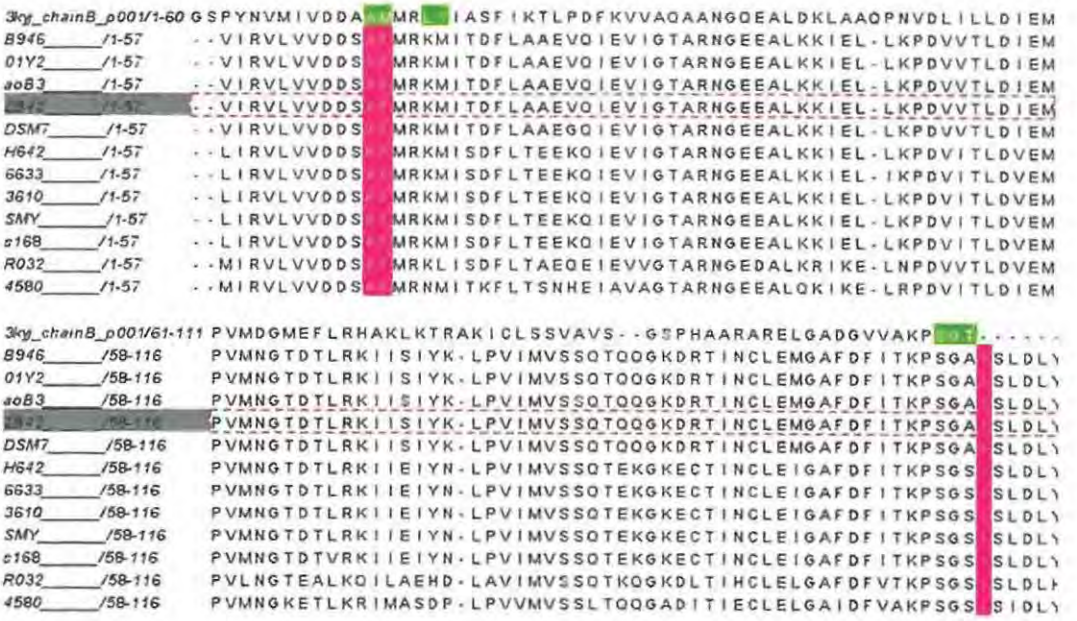


Figure 5.7 Multiple sequence alignment of a section of CheB from the *B. subtilis* group and CheY6 from *R. sphaeroides*. Colouring is the same as for Figure 5.6, except that pink residues are conserved in CheB within the *B. subtilis* group and were identified by PIC as important for inter-protein hydrophobic interactions with the P1 domain of CheA.

5.2. Interactions of CheY and CheB with P2 in the *B. subtilis* group

Interaction of CheY with P2 accelerates the phospho-transfer reaction by bringing CheY in close spatial proximity to the phospho-histidine located in P1 (Stewart & Van Bruggen 2004). It was shown that, when the P2 domain is absent, the P1 domain has a low affinity for CheY and phospho-transfer to CheY is markedly slower compared to when P2 is present (Stewart *et al.* 2000). In its inactive state the methylesterase CheB cannot bind to P2 using the surface similar to that of CheY, because of a steric clash involving the C-terminal domain. However, once activated by phosphorylation, the molecule changes from a closed to an open conformation and the domains are repositioned to allow access to this surface (Djordjevic *et al.* 1998) Therefore it is theoretically possible for the N-terminal domain of phospho-CheB to interact with the P2 domain of CheA.

The CheY-CheAP2 and $CheB_N$-CheAP2 models were analysed with PIC to identify residues involved in surface recognition and inter-protein interactions (Figure 5.8, Figure 5.9 and Figure 5.10). The hydrophobic interactions involve a phenyl ring (CheY: Phe-102 and CheB: Phe-101) that stacks closely with P2 Val-182, Leu-179 and Met-186. This interaction is flanked by an Ile- (CheY position: 91, CheB position 92) and a Pro-residue (CheY position: 105, CheB position: 106). The results from PIC indicated that reactions involving hydrogen bonds, ionic interactions, and cation-pi interactions between CheY and P2 are also conserved between $CheB_N$ and P2 (Figure 5.11 and Figure 5.12).

These results show that, despite similarities with the *T. maritima* CheY-CheAP2 interaction, the *B. subtilis* group interactions involve different amino acids. For instance, in *T. maritima* it is the alanine-arginine pair that stacks with the phenyl ring, while two leucine residues enclose the ring (Park *et al.* 2004). More importantly, these results also indicate that it is possible for the N-terminal domain of phospho-CheB to interact with the P2 domain of CheA and thereby directly antagonize the formation of phospho-CheY. Such a model has not been proposed before. The presence of a phosphatase protein to terminate the phospho-CheY signal is not critical for chemotaxis, thus it is possible that in the more ancestral system the antagonistic binding of phospho-CheB to P2 had the purpose of preventing new phospho-CheY being formed, while the transient nature of the already-formed phospho-CheY molecule would result in lower levels of

phospho-CheY and a return to default flagellar rotation. CheB plays a role in adaptation by selective demethylation of specific conserved residues on the MCPs, thus it is possible that CheB provides an intersecting point between adaptation and signal termination. This model is consistent with the observation that a *cheB* null-mutant is random, and not smooth swimming as would be expected from the fact that overmethylation of the receptors activates CheA (Rosario *et al.* 1995). In a *cheB* null-mutant the levels of phospho-CheY would be controlled by the phosphatase action of CheC and FliY thereby giving rise to intermittent swimming and tumbling. Furthermore *cheB* null-mutants adapt normally to repellents but cannot adapt to high attractant concentrations (Kirsch *et al.* 1993). This suggests that when a cell is incubated in the presence of attractant, phospho-CheB serves to lower phospho-CheY levels to reduce smooth swimming which may take the bacterium away from the favourable environment.

It should be noted that in *E.coli* it would not be possible for phospho-CheB to directly antagonize the formation of phospho-CheY by interacting with P2 as the CheY residues involved in binding P2 are not conserved in CheB (McEvoy *et al.* 1999).

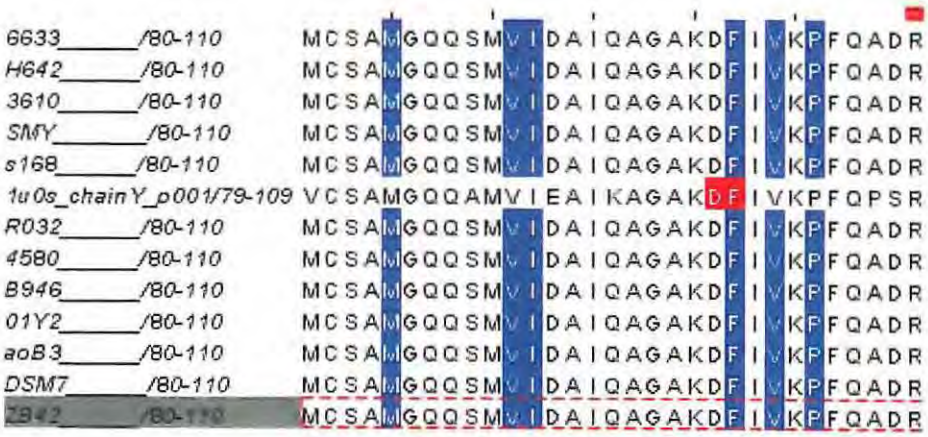


Figure 5.8. Multiple sequence alignment of CheY from *T. maritima* (PDB ID: 1U0S) and CheY from members of the *B. subtilis* group showing conserved residues which are important for CheY-CheAP2 interaction. Residues that are highlighted red were previously identified as important for P2 binding in *T. maritima*. Residues highlighted in blue were identified by PIC as important for interaction with CheA and are conserved in members of the *B. subtilis* group.

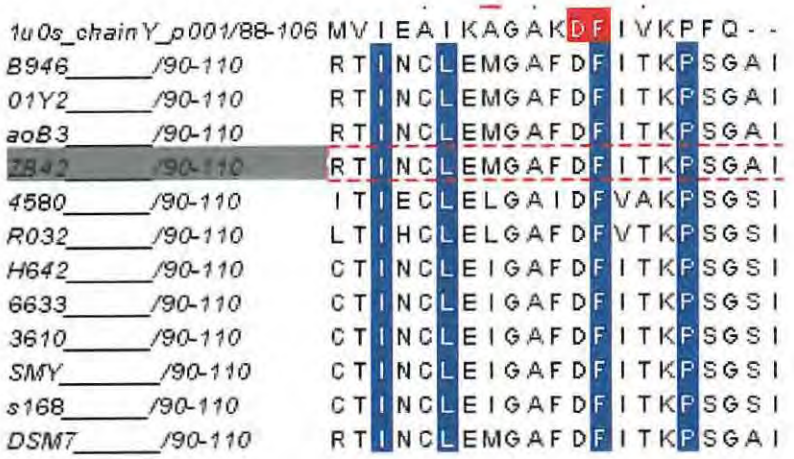


Figure 5.9. Multiple sequence alignment of CheY from *T. maritima* (PDB ID: 1U0S) and CheB from members of the *B. subtilis* group showing conserved residues which are important for CheY-CheAP2 interaction. Residues that are highlighted red were previously identified as important for P2 binding in *T. maritima*. Residues highlighted in blue were identified by PIC as important for interaction with CheA and are conserved in members of the *B. subtilis* group.

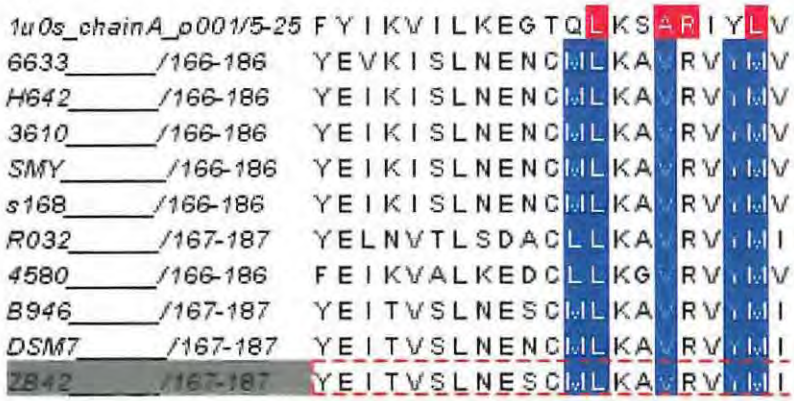


Figure 5.10. Multiple sequence alignment of CheA from *T. maritima* (PDB ID: 1U0S) and members of the *B. subtilis* group showing conserved residues which are important for CheY-CheAP2 interaction. Residues that are highlighted red were previously identified as important for CheY binding in *T. maritima*. Residues highlighted in blue were identified by PIC as important for interaction with CheY and $CheB_x$ and are conserved in members of the *B. subtilis* group.

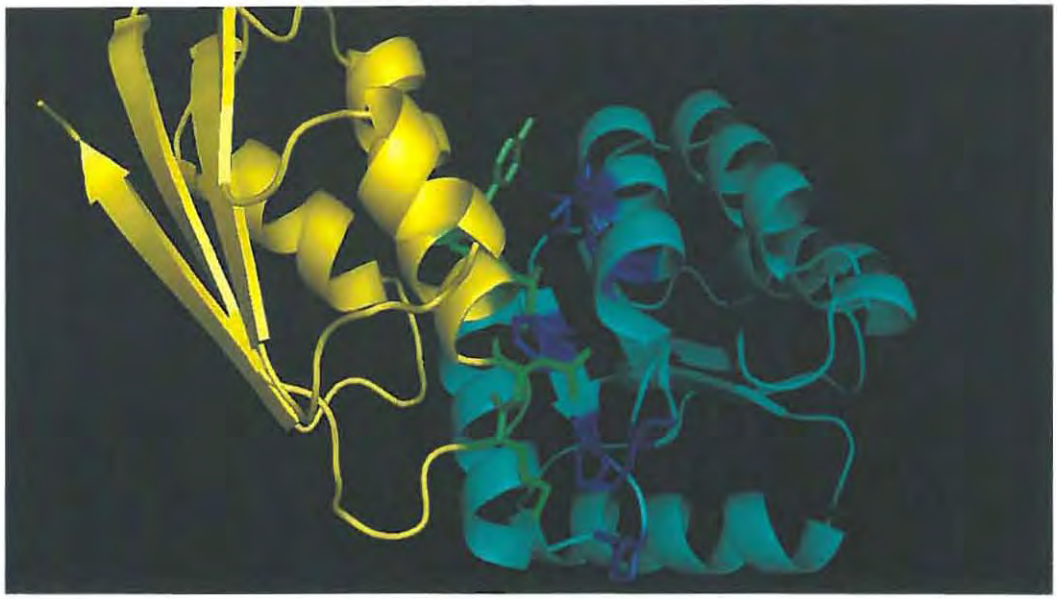


Figure 5.11. A close-up of the interaction between CheY (teal) and CheA P2 domain (yellow). Residues that are indicated as sticks are involved in the hydrophobic interaction. Blue residues from CheY interact with green residues from P2

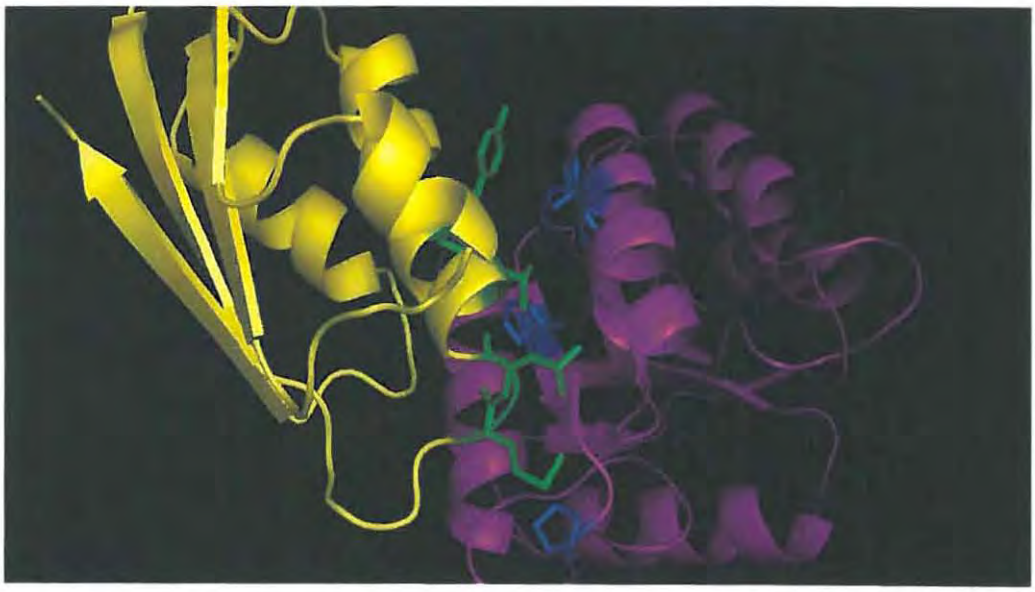


Figure 5.12. A close-up of the interaction between CheB (purple) and CheA P2 domain (yellow). Residues that are indicated as sticks are involved in the hydrophobic interaction. Blue residues from CheY interact with green residues from P2

5.3. **Conclusion**

Insight into the mechanisms involved in the partner specificity that exists between HK and RRs will not only aid the prediction of interaction partners, but could also be applied to rewiring bacterial sensory pathways in synthetic biology (Bell *et al.* 2010). Furthermore, two component signal transduction pathways are very widespread in bacteria, but absent in animals (Casino *et al.* 2009), making it an ideal target for broad spectrum antibiotics. Up to now structural information on the interaction between class II HKs and associated RRs is very sparse. Available structures are restricted to the single domain chemotaxis response regulator CheY, and no experimental structures showing the multidomain methylesterase CheB interacting with CheA exist. The N-terminal domain of CheB exhibits high similarity to the CheY protein, thus a model of $CheB_N$ based on CheY is most likely acceptable for the purposes of drawing structure based conclusions. In line with the above, the $CheB_N$ residues predicted to be important for complex formation correspond to residues identified in other studies as taking part in interacting with CheAp1 and P2 (Tzeng & Hoch 1997, Appleby & Bourret 1998, Hoch & Varughese 2001, Park *et al.* 2004, Casino *et al.* 2009). As far as the interaction between P1 with CheB and CheY is concerned, homology models of reasonable quality suggest that unphosphorylated CheB and CheY interact with P1 in a manner that is characteristic of other HK-RR interactions. It seems to be the case that there is a molecular "code" of amino acids that is used by HKs to recognize their cognate RR, and that this interaction is mostly of the weak hydrophobic type

As for the interaction between CheY with P2, the residues involved in this interaction are also conserved in the homologous surface of $CheB_N$. This raises some interesting questions with regards to the possible role that phospho-CheB may play as an inhibitor of CheY binding to P2. Furthermore, the fact that unphosphorylated CheB cannot make use of P2 to facilitate interaction with P1, as is the case for CheY, indicates that there must be some inherent feature of CheB which enhances its affinity for CheAP1. Nevertheless, experimental studies must be performed to confirm these hypotheses.

CHAPTER 6

6. CONCLUDING REMARKS

Chemotaxis is pivotal in the interactions between plants and bacteria. Plants secrete a plethora of chemicals which may be chemo-attractants or -repellents for plant-associated bacteria. Research has shown that there is no standard set of chemo-attractants for bacteria living in the rhizosphere (Yao & Allen 2006) and even at a subspecies level differences in host preferences can be observed (Reva *et al.* 2004, Yao & Allen 2006). With this in mind it is reasonable to hypothesize that variations in the chemotactic responses of different bacteria may be due to evolutionary selection conferring adaptation to divergent habitats. However, results suggested that it is unlikely that positive selection operates on the proteins involved in chemotaxis of the *B. subtilis* group. It is also possible that the results of the site specific positive and purifying selection analyses were biased to favour purifying selection because the sequences analyzed were very closely related. For example, the nucleotide sequence for the CheY proteins was more than 96% identical between *B. amyloliquefaciens* strains.

Homology models of the chemotaxis proteins of *B. amyloliquefaciens* spp. *plantarum* FZB42 were constructed. Additionally, the interactions between some of the proteins were also modelled. Models were evaluated using various model quality assessment programs. If present, inaccurately modelled loop regions were refined to relieve unlikely conformations. Evaluation results indicated that the quality of the homology models built as part of this work approached that of experimentally determined structures. The modelling of CheR, CheW, the P1 and P2 domains of CheA as well as CheB and CheY interacting with the P1 and P2 domains of CheA highlighted the difficulties in constructing homology based models for targets with low sequence identity to selected templates.

Interacting residues in complexes were identified using the Proteins Interaction Calculator server. Following a computational approach it was possible to gain further insight into the structures of the various chemotaxis proteins found in members of the *B. subtilis* group, as well as for some of the interaction partners. Of particular interest are the homology models of the

interaction between CheB and the P1 and P2 domains of CheA, since no experimental structural information for these interactions is available. The homology model of the N-terminal domain of CheB with the P2 domain of CheA, together with sequence information for CheY and CheB from members of the *B. subtilis* group suggested that it is possible for phospho-CheB to compete with CheY, thereby directly inhibiting the formation of phospo-CheY. Such a mechanism has not been proposed before, hence this hypothesis would benefit from experimental investigation.

The models built for the purpose of the present study can be used in future to investigate ligand-protein and protein-protein interactions in an attempt to model an *in silico* reconstruction of the chemotaxis pathway. The work presented here should contribute to a structure-based understanding of chemotaxis, and in particular to aid future research on the spatial evolution of the chemotaxis pathway in the *B. subtilis* group.

REFERENCES

Adachi, J. & Hasegawa, M. 1996, "Model of amino acid substitution in proteins encoded by mitochondrial DNA", *Journal Of Molecular Evolution,* vol. 24, no. 2, pp 459-468

Aizawa, S., Harwood, C.S. & Kadner, R.J. 2000, "Signaling components in bacterial locomotion and sensory reception", *The Journal Of Bacteriology,* vol. 182, no. 6, pp. 1459-1471.

Akaike, H. 1974, "A new look at the statistical model identification", *IEEE Transactions on Automatic Control,* vol. 19, no. 6, pp. 716-723.

Altschul, S.F., Gish, W., Miller, W., Meyers, E.W. & Lipman, D.J. 1990, "Basic local alignment search tool", *Journal Of Molecular Biology,* vol. 215, no. 3, pp. 403-410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. 1997, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Research,* vol. 25, no. 17, pp. 3389-3402.

Anand, G.S., Goudreau, P.N., Lewis, J.K. & Stock, A.M. 2000, "Evidence for phosphorylation-dependent conformational changes in methylesterase CheB", *Protein Science,* vol. 9, no. 5, pp. 898-906.

Appleby, J.L. & Bourret, R.B. 1998, "Proposed signal transduction role for conserved CheY residue Thr87, a member Of the response regulator active-Site Quintet", *The Journal Of Bacteriology,* vol. 180, no. 14, pp. 3563-3569.

Bell, C.H., Porter, S.L., Strawson, A., Stuart, D.I. & Armitage, J.P. 2010, "Using structural information to change the phosphotransfer specificity of a two-component chemotaxis signalling complex", *PLoS Biology,* vol. 8, no. 2, pp. e1000306.

Bellsolell, L., Cronet, P., Majolero, M., Serrano, L. & Coll, M. 1996, "The three-dimensional structure of two mutants Of the signal transduction protein CheY suggest its molecular activation mechanism", *Journal Of Molecular Biology,* vol. 257, no. 1, pp. 116-128.

Bischoff, D.S. & Ordal, G.W. 1991, "Sequence and characterization of *Bacillus subtilis* CheB, a homolog of Escherichia coli CheY, and its role in a different mechanism of chemotaxis", *Journal Of Biological Chemistry,* vol. 266, no. 19, pp. 12301-12305.

Bischoff, D.S., Bourret, R.B., Kirsch, M.L. & Ordal, G.W. 1993, "Purification and characterization of *Bacillus subtilis* CheY", *Biochemistry,* vol. 32, no. 35, pp. 9256-9261.

Bourret, R.B., Davagnino, J. & Simon, M.I. 1993, "The carboxy-terminal portion of the CheA kinase mediates regulation of autophosphorylation by transducer and CheW", *The Journal Of Bacteriology,* vol. 175, no. 7, pp. 2097-2101.

Braun, W. & Gō, N. 1985, "Calculation of protein conformations by proton-proton distance constraints : A new efficient algorithm", *Journal Of Molecular Biology*, vol. 186, no. 3, pp. 611-626.

Brooks, B.R., Brooks, C.L., Mackerell, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A.R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R.W., Post, C.B., Pu, J.Z., Schaefer, M., Tidor, B., Venable, R.M., Woodcock, H.L., Wu, X., Yang, W., York, D.M. & Karplus, M. 2009, *CHARMM: The Biomolecular Simulation Program*, Wiley Subscription Services, Inc., A Wiley Company.

Browne, W.J., North, A.C.T., Phillips, D.C., Brew, K., Vanaman, T.C. & Hill, R.L. 1969, "A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme", *Journal Of Molecular Biology*, vol. 42, no. 1, pp. 65-70.

Buchko, G.W., Cushley, R.J., Rozek, A., Kennedy, M.A. & Kanda, P. 2000, "Structural studies of a baboon (Papio sp.) plasma protein inhibitor of cholesteryl ester transferase", *Protein Science*, vol. 9, no. 8, pp. 1548-1558.

Bujnicki, J. 2006, "Protein structure prediction: concepts and applications. Anna Tramontano.", *ChemBioChem*, vol. 7, no. 6, pp. 990-991.

Casino, P., Rubio, V. & Marina, A. 2009, "Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction", *Cell*, vol. 139, no. 2, pp. 325-336.

Castrignanò, T., De Meo, P.D., Cozzetto, D., Talamo, I.G. & Tramontano, A. 2006, "The PMDB protein model database", *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D306-D309.

Chao, X., Muff, T.J., Park, S., Zhang, S., Pollard, A.M., Ordal, G.W., Bilwes, A.M. & Crane, B.R. 2006, "A receptor-modifying deamidase in complex with a signaling phosphatase reveals reciprocal regulation", *Cell*, vol. 124, no. 3, pp. 561-571.

Chen, X.H., Koumoutsi, A., Scholz, R., Schneider, K., Vater, J., Süssmuth, R., Piel, J. & Borriss, R. 2009, "Genome analysis of *Bacillus amyloliquefaciens* FZB42 reveals its potential for biocontrol of plant pathogens", *Journal Of Biotechnology*, vol. 140, no. 1-2, pp. 27-37.

Cho, H.S., Lee, S., Yan, D., Pan, X., Parkinson, J.S., Kustu, S., Wemmer, D.E. & Pelton, J.G. 2000, "NMR structure of activated CheY", *Journal Of Molecular Biology*, vol. 297, no. 3, pp. 543-551.

Choudhary, D.K. & Johri, B.N. 2009, "Interactions of *Bacillus spp.* and plants – with special reference to induced systemic resistance (ISR)", *Microbiological Research*, vol. 164, no. 5, pp. 493-513.

Cothia, C. & Lesk, A.M. 1986, "The relation between the divergence of sequence and structure in proteins", *The EMBO Journal,* vol. 5, no. 4, pp. 823-826.

Creevey, C.J. & McInerney, J.O. 2002, "An algorithm for detecting directional and non-directional positive selection, neutrality and negative selection in protein coding DNA sequences", *Gene,* vol. 300, no. 1-2, pp. 43-51.

Cristobal, S., Zemla, A., Fischer, D., Rychlewski, L. & Elofsson, A. 2001, "A study of quality measures for protein threading models", *BMC Bioinformatics,* vol. 2, no. 5.

Dalton, J.A.R. & Jackson, R.M. 2007, "An evaluation of automated homology modelling methods at low target–template sequence similarity", *Bioinformatics,* vol. 23, no. 15, pp. 1901-1908.

Djordjevic, S., Goudreau, P.N., Xu, Q., Stock, A.M. & West, A.H. 1998, "Structural basis for methylesterase CheB regulation by a phosphorylation-activated domain", *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 95, no. 4, pp. 1381-1386.

Djordjevic, S. & Stock, A.M. 1998, "Chemotaxis receptor recognition by protein methyltransferase CheR", *Nature,* vol. 5, no. 6, pp. 446-450.

Djordjevic, S. & Stock, A.M. 1997, "Crystal structure of the chemotaxis receptor methyltransferase CheR suggests a conserved structural motif for binding S-adenosylmethionine", *Structure,* vol. 5, no. 4, pp. 545-558.

Doron-Faigenboim, A. & Pupko, T. 2007, "A Combined empirical and mechanistic codon model", *Molecular Biology And Evolution,* vol. 24, no. 2, pp. 388-397.

Doron-Faigenboim, A., Stern, A., Mayrose, I., Bacharach, E. & Pupko, T. 2005, "Selecton: a server for detecting evolutionary forces at a single amino-acid site", *Bioinformatics,* vol. 21, no. 9, pp. 2101-2103.

Dutta, R., Qin, L. & Inouye, M. 1999, "Histidine kinases: diversity of domain organization", *Molecular Microbiology,* vol. 34, no. 4, pp. 633-640.

Eaton, A.K. & Stewart, R.C. 2010, "Kinetics of ATP and TNP-ATP binding to the active site of CheA from *Thermotoga maritima*", *Biochemistry,* vol. 49, no. 27, pp. 5799-5809.

Eddy, S.R. 1998, "Profile hidden Markov models", *Bioinformatics,* vol. 14, no. 9, pp. 755-763.

Endres, R.G., Falke, J.J. & Wingreen, N.S. 2007, "Chemotaxis receptor complexes: from signaling to assembly", *PLoS Computational Biology,* vol. 3, no. 7, pp. e150.

Fain, B. & Levitt, M. 2001, "A novel method for sampling alpha-helical protein backbones", *Journal Of Molecular Biology,* vol. 305, no. 2, pp. 191-201.

Falke, J.J., Bass, R.B., Butler, S.L., Chervitz, S.A. & Danielson, M.A. 1997, "The two-component signalling pathway of bacterialchemotaxis: a molecular view of signal transduction by receptors, kinases, and adaptation enzymes", *Annual Review Of Cell And Developmental Biology*, vol. 13, no. 1, pp. 457-512.

Fares, M.A., Elena, S.F., Ortiz, J., Moya, A. & Barrio, E. 2002, "A Sliding Window-Based Method to Detect Selective Constraints in Protein-Coding Genes and Its Application to RNA Viruses", *Journal Of Molecular Evolution*, vol. 55, no. 5, pp. 509-521.

Fernandez-Fuentes, N., Rai, B.K., Madrid-Aliste, C.J., Eduardo Fajardo, J. & Fiser, A. 2007, "Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments", *Bioinformatics*, vol. 23, no. 19, pp. 2558-2565.

Fiser, A. & Šali, A. 2003, "Modeller: generation and refinement of homology-based protein structure models" in *Methods In Enzymology*, ed. Charles W. Carter,Jr.and Robert M.Sweet, Academic Press, , pp. 461-491.

Fritze, D. 2004, "Taxonomy of the genus *Bacillus* and related genera: the aerobic endospore-forming bacteria", *Phytopathology*, vol. 94, no. 11, pp. 1245-1248.

From, C., Hormazabal, V. & Granum, P.E. 2007, "Food poisoning associated with pumilacidin-producing *Bacillus pumilus* in rice", *International Journal Of Food Microbiology*, vol. 115, no. 3, pp. 319-324.

Fuhrer, D.K. & Ordal, G.W. 1991, "*Bacillus subtilis* CheN, a homolog of CheA, the central regulator of chemotaxis in *Escherichia coli.*", *The Journal Of Bacteriology*, vol. 173, no. 23, pp. 7443-7448.

Garrity, L.F. & Ordal, G.W. 1995, "Chemotaxis in *Bacillus subtilis*: how bacteria monitor environmental signals", *Pharmacology And Therapeutics*, vol. 68, no. 1, pp. 87-104.

Gegner, J.A., Graham, D.R., Roth, A.F. & Dahlquist, F.W. 1992, "Assembly of an MCP receptor, CheW, and kinase CheA complex in the bacterial chemotaxis signal transduction pathway", *Cell*, vol. 70, no. 6, pp. 975-982.

Gertow, K., Bellanda, M., Eriksson, P., Boquist, S., Hamsten, A., Sunnerhagen, M. & Fisher, R.M. 2004, "Genetic and structural evaluation of fatty acid transport protein-4 in relation to markers of the insulin resistance syndrome", *Journal Of Clinical Endocrinology Metabolism*, vol. 89, no. 1, pp. 392-399.

Ginalski, K. 2006, "Comparative modeling for protein structure prediction", *Current Opinion In Structural Biology*, vol. 16, no. 2, pp. 172-177.

Gioia, J., Yerrapragada, S., Qin, X., Jiang, H., Igboeli, O.C., Muzny, D., Dugan-Rocha, S., Ding, Y., Hawes, A., Liu, W., Perez, L., Kovar, C., Dinh, H., Lee, S., Nazareth, L., Blyth, P., Holder, M., Buhay, C., Tirumalai, M.R., Liu, Y., Dasgupta, I., Bokhetache, L., Fujita, M., Karouia, F., Eswara Moorthy, P., Siefert, J., Uzman, A., Buzumbo, P., Verma, A., Zwiya, H., McWilliams, B.D., Olowu, A., Clinkenbeard, K.D., Newcombe, D., Golebiewski, L., Petrosino, J.F., Nicholson, W.L., Fox, G.E., Venkateswaran, K., Highlander, S.K. & Weinstock, G.M. 2007, "Paradoxical DNA Repair and Peroxide Resistance Gene Conservation in *Bacillus pumilus* SAFR-032", *PLoS ONE*, vol. 2, no. 9, pp. e928.

Goldman, D.J. & Ordal, G.W. 1984, "In vitro methylation and demethylation of methyl-accepting chemotaxis proteins in *Bacillus subtilis*", *Biochemistry*, vol. 23, no. 12, pp. 2600-2606.

Goldman, N. & Yang, Z. 1994, "A codon-based model of nucleotide substitution for protein-coding DNA sequences", *Molecular Biology And Evolution*, vol. 11, no. 5, pp. 725-736.

Grantham, R. 1974, "Amino acid difference formula to help explain protein evolution", *Science*, vol. 185, no. 4154, pp. 862-864.

Griswold, I.J., Zhou, H., Matison, M., Swanson, R.V., McIntosh, L.P., Simon, M.I. & Dahlquist, F.W. 2002, "The solution structure and interactions of CheW from *Thermotoga maritima*", *Nature Structural And Molecular Biology*, vol. 9, no. 2, pp. 121-125.

Guhaniyogi, J., Robinson, V.L. & Stock, A.M. 2006, "Crystal structures of beryllium fluoride-free and beryllium fluoride-bound CheY in complex with the conserved C-terminal peptide of CheZ reveal dual binding modes specific to CheY conformation", *Journal Of Molecular Biology*, vol. 359, no. 3, pp. 624-645.

Hall, T.A. 1999, "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT", *Nucleic Acids Symposium Series*, vol. 41, pp. 95-98.

Hamer, R., Chen, P., Armitage, J., Reinert, G. & Deane, C. 2010, "Deciphering chemotaxis pathways using cross species comparisons", *BMC Systems Biology*, vol. 4, no. 1, pp. 3.

Hanlon, D.W., Márquez-Magaña, L.M., Carpenter, P.B., Chamberlin, M.J. & Ordal, G.W. 1992, "Sequence and characterization of *Bacillus subtilis* CheW.", *Journal Of Biological Chemistry*, vol. 267, no. 17, pp. 12055-12060.

Hanlon, D.W. & Ordal, G.W. 1994, "Cloning and characterization of genes encoding methyl-accepting chemotaxis proteins in *Bacillus subtilis*", *Journal Of Biological Chemistry*, vol. 269, no. 19, pp. 14038-14046.

Harighi, B. 2009, "Genetic evidence for CheB- and CheR-dependent chemotaxis system in *A. tumefaciens* toward acetosyringone", *Microbiological Research*, vol. 164, no. 6, pp. 634-641.

Hershberg, R., Tang, H. & Petrov, D. 2007, "Reduced selection leads to accelerated gene loss in Shigella", *Genome Biology,* vol. 8, no. 8, pp. R164.

Hildebrand, A., Remmert, M., Biegert, A. & Söding, J. 2009, "Fast and accurate automatic structure prediction with HHpred", *Proteins: Structure, Function And Bioinformatics,* vol. 77, no. S9, pp. 128-132.

Hillisch, A., Pineda, L.F. & Hilgenfeld, R. 2004, "Utility of homology models in the drug discovery process", *Drug Discovery Today,* vol. 9, no. 15, pp. 659-669.

Hintze, B.J. & Johnson, S.J. 2010, "ResDe: a new tool for visual definition of distance restraints for crystallographic refinement", *Journal Of Applied Crystallography,* vol. 43, no. 6, pp. 1540-1542.

Hirschman, A., Boukhvalova, M., VanBruggen, R., Wolfe, A.J. & Stewart, R.C. 2001, "Active site mutations in CheA, the signal-transducing protein kinase of the chemotaxis system in *Escherichia coli*", *Biochemistry,* vol. 40, no. 46, pp. 13876-13887.

Hoch, J.A. & Varughese, K.I. 2001, "Keeping signals straight in phosphorelay signal transduction", *The Journal Of Bacteriology,* vol. 183, no. 17, pp. 4941-4949.

Hughes, A.L. 2007, "Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level", *Heredity,* vol. 99, no. 4, pp. 364-373.

Hughes, C.A., Mandell, J.G., Anand, G.S., Stock, A.M. & Komives, E.A. 2001, "Phosphorylation causes subtle changes in solvent accessibility at the interdomain interface of methylesterase CheB", *Journal Of Molecular Biology,* vol. 307, no. 4, pp. 967-976.

Jones, D.T. 1999, "Protein secondary structure prediction based on position-specific scoring matrices", *Journal Of Molecular Biology,* vol. 292, no. 2, pp. 195-202.

Jones, D.T. 1997, "Progress in protein structure prediction", *Current Opinion In Structural Biology,* vol. 7, no. 3, pp. 377-387.

Jones, D.T., Taylor, W.R. & Thornton, J.M. 1992, "The rapid generation of mutation data matrices from protein sequences", *Computer Applications In The Biosciences : CABIOS,* vol. 8, no. 3, pp. 275-282.

Jurica, M.S. & Stoddard, B.L. 1998, "Mind your B's and R's: bacterial chemotaxis, signal transduction and protein recognition", *Structure,* vol. 6, no. 7, pp. 809-813.

Kabsch, W. & Sander, C. 1983, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features", *Biopolymers,* vol. 22, no. 12, pp. 2577-2637.

Kaminska, K.H., Purta, E., Hansen, L.H., Bujnicki, J.M., Vester, B. & Long, K.S. 2010, "Insights into the structure, function and evolution of the radical-SAM 23S rRNA

methyltransferase Cfr that confers antibiotic resistance in bacteria", *Nucleic Acids Research,* vol. 38, no. 5, pp. 1652-1663.

Katiyar, A., Lenka, S.K., Lakshmi, K., Chinnusamy, V. & Bansal, K.C. 2009, "In silico characterization and homology modeling of thylakoid-bound ascorbate peroxidase from a drought tolerant wheat cultivar", *Genomics, Proteomics And Bioinformatics,* vol. 7, no. 4, pp. 185-193.

Kentner, D. & Sourjik, V. 2006, "Spatial organization of the bacterial chemotaxis system", *Current Opinion In Microbiology,* vol. 9, no. 6, pp. 619-624.

Khajeh, K., Shokri, M.M., Asghari, S.M., Moradian, F., Ghasemi, A., Sadeghi, M., Ranjbar, B., Hosseinkhani, S., Gharavi, S. & Naderi-Manesh, H. 2006, "Acidic and proteolytic digestion of α-amylases from *Bacillus licheniformis* and *Bacillus amyloliquefaciens*: Stability and flexibility analysis", *Enzyme And Microbial Technology,* vol. 38, no. 3-4, pp. 422-428.

Kirby, J.R., Kristich, C.J., Saulmon, M., Zimmer, M.A., Garrity, L., Zhulin, I.B. & Ordal, G. 2001, "CheC is related to the family of flagellar switch proteins and acts independently from CheD to control chemotaxis in *Bacillus subtilis*", *Molecular Microbiology,* vol. 42, no. 3, pp. 573-585.

Kirby, J.R., Niewold, T.B., Maloy, S. & Ordal, G.W. 2000, "CheB is required for behavioural responses to negative stimuli during chemotaxis in *Bacillus subtilis*", *Molecular Microbiology,* vol. 35, no. 1, pp. 44-57.

Kirsch, M.L., Peters, P.D., Hanlon, D.W., Kirby, J.R. & Ordal, G.W. 1993, "Chemotactic methylesterase promotes adaptation to high concentrations of attractant in *Bacillus subtilis.*", *Journal Of Biological Chemistry,* vol. 268, no. 25, pp. 18610-18616.

Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., Andersen, K.K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P., Boland, F., Brignell, S.C., Bron, S., Bunai, K., Chapuis, J., Christiansen, L.C., Danchin, A., Dabarbouilla, M., Dervyn, E., Deuerling, E., Devine, K., Devine, S.K., Dreesen, O., Errington, J., Fillinger, S., Foster, S.J., Fujita, Y., Galizzi, A., Gardan, R., Eschevins, C., Fukushima, T., Haga, K., Harwood, C.R., Hecker, M., Hosoya, D., Hullo, M.F., Kakeshita, H., Karamata, D., Kasahara, Y., Kawamura, F., Koga, K., Koski, P., Kuwana, R., Imamura, D., Ishimaru, M., Ishikawa, S., Ishio, I., Le Coq, D., Masson, A., Maual, C., Meima, R., Mellado, R.P., Moir, A., Moriya, S., Nagakawa, E., Nanamiya, H., Nakai, S., Nygaard, P., Ogura, M., Ohanan, T., O'Reilly, M., O'Rourke, M., Pragai, Z., Pooley, H.M., Rapoport, G., Rawlins, J.P., Rivas, L.A., Rivolta, C., Sadaie, A., Sadaie, Y., Sarvas, M., Sato, T., Saxild, H.H., Scanlan, E., Schumann, W., Seegers, J.F.M.L., Sekiguchi, J., Sekowska, A., Saror, S.J., Simon, M., Stragier, P., Studer, R., Takamatsu, H., Tanaka, T., Takeuchi, M., Thomaides, H.B., Vagner, V., van Dijl, J.M., Watabe, K., Wipat, A., Yamamoto, H., Yamamoto, M., Yamamoto, Y., Yamane, K., Yata, K., Yoshida, K., Yoshikawa, H., Zuber, U. & Ogasawara, N. 2003, "Essential *Bacillus subtilis* genes", *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 100, no. 8, pp. 4678-4683.

Kolinski, A., Betancourt, M., Kihara, D., Rotkiewicz, P. & Skolnick, J. 2001, "Generalized comparative modeling (GENECOMP): A combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement", *Proteins: Structure, Function And Bioinformatics,* vol. 44, no. 2, pp. 133-149.

Kort, E.N., Goy, M.F., Larsen, S.H. & Adler, J. 1975, "Methylation of a membrane protein involved in bacterial chemotaxis", *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 72, no. 10, pp. 3939-3943.

Koshi, J. & Goldstein, R. 1996, "Probabilistic reconstruction of ancestral protein sequences", *Journal Of Molecular Evolution,* vol. 42, no. 2, pp. 313-320.

Koshland Jr., D.E. 2002, "Special essay: The seven pillars of life", *Science,* vol. 295, no. 5563, pp. 2215-2216.

Krieger, E., Nabuurs, S.B. & Vriend, G. 2003, "Homology Modeling" in *Structural Bioinformatics*, vol. 44, ed. Bourne P. and Weissig H, John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/0471721204.ch25.

Krissinel, E. 2007, "On the relationship between sequence and structure similarities in proteomics", *Bioinformatics,* vol. 23, no. 6, pp. 717-723.

Kristich, C.J. & Ordal, G.W. 2004, "Analysis of chimeric chemoreceptors in *Bacillus subtilis* reveals a role for CheD in the function of the McpC HAMP domain", *The Journal Of Bacteriology,* vol. 186, no. 17, pp. 5950-5955.

Krueger, J.K., Stock, J. & Schutt, C.E. 1992, "Evidence that the methylesterase of bacterial chemotaxis may be a serine hydrolase", *Biochimica et Biophysica Acta (BBA) - Protein Structure And Molecular Enzymology,* vol. 1119, no. 3, pp. 322-326.

Kunst, F. & Ogasawara, N. 1997, "The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*", *Nature,* vol. 390, no. 6657, pp. 249.

Lamzin, V.S., Morris, R.J., Dauter, Z., Wilson, K.S. & Teeter, M.M. 1999, "Experimental Observation of Bonding Electrons in Proteins", *Journal Of Biological Chemistry,* vol. 274, no. 30, pp. 20753-20755.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., & Thornton, J.M., 1993, "PROCHECK: a program to check the stereochemical quality of protein structures", *Journal Of Applied Crystallography,* vol. 26, no. 2, pp. 283-291.

Le Moual, H. & Koshland, J.,Daniel E. 1996, "Molecular evolution of the C-terminal cytoplasmic domain of a superfamily of bacterial receptors involved in taxis", *Journal Of Molecular Biology,* vol. 261, no. 4, pp. 568-585.

Lepesant, J.A., Billault, A., Kejzlarová-Lepesant, J., Pascal, M., Kunst, F. & Dedonder, R. 1975, "Identification of the structural gene for sucrase in *Bacillus subtilis marburg*", *Biochimie*, vol. 56, no. 11-12, pp. 1465-1470.

Li, W.H., Wu, C.I. & Luo, C.C. 1985, "A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes", *Molecular Biology And Evolution*, vol. 2, no. 2, pp. 150-174.

Li, Y., Hu, Y., Fu, W., Xia, B. & Jin, C. 2007, "Solution structure of the bacterial chemotaxis adaptor protein CheW from *Escherichia coli*", *Biochemical And Biophysical Research Communications*, vol. 360, no. 4, pp. 863-867.

Liu, J.D. & Parkinson, J.S. 1991, "Genetic evidence for interaction between the CheW and Tsr proteins during chemoreceptor signaling by *Escherichia coli*.", *The Journal Of Bacteriology*, vol. 173, no. 16, pp. 4941-4951.

Liu, J.D. & Parkinson, J.S. 1989, "Role of CheW protein in coupling membrane receptors to the intracellular signaling system of bacterial chemotaxis", *Proceedings Of The National Academy Of Sciences Of The United States Of America*, vol. 86, no. 22, pp. 8703-8707.

Lukat, G.S., Lee, B.H., Mottonen, J.M., Stock, A.M. & Stock, J.B. 1991, "Roles of the highly conserved aspartate and lysine residues in the response regulator of bacterial chemotaxis.", *Journal Of Biological Chemistry*, vol. 266, no. 13, pp. 8348-8354.

Lukat, G.S., Stock, A.M. & Stock, J.B. 1990, "Divalent metal ion binding to the CheY protein and its significance to phosphotransfer in bacterial chemotaxis", *Biochemistry*, vol. 29, no. 23, pp. 5436-5442.

Lupas, A. & Stock, J. 1989, "Phosphorylation of an N-terminal regulatory domain activates the CheB methylesterase in bacterial chemotaxis.", *Journal Of Biological Chemistry*, vol. 264, no. 29, pp. 17337-17342.

Luthy, R., Bowie, J.U. & Eisenberg, D. 1992, "Assessment of protein models with three-dimensional profiles", *Nature*, vol. 356, no. 6364, pp. 83-85.

Makkar, P., Metpally, R.P.R., Sangadala, S. & Reddy, B.V.B. 2009, "Modeling and analysis of MH1 domain of Smads and their interaction with promoter DNA sequence motif", *Journal Of Molecular Graphics And Modelling*, vol. 27, no. 7, pp. 803-812.

Manson, M.D. 2008, "The tie that binds the dynamic duo: the connector between AS1 and AS2 in the HAMP domain of the *Escherichia coli* Tsr chemoreceptor", *The Journal Of Bacteriology*, vol. 190, no. 20, pp. 6544-6547.

Manson, M.D., Armitage, J.P., Hoch, J.A. & Macnab, R.M. 1998, "Bacterial locomotion and signal transduction", *The Journal Of Bacteriology*, vol. 180, no. 5, pp. 1009-1022.

Massingham, T. & Goldman, N. 2005, "Detecting amino acid sites under positive selection and purifying selection", *Genetics*, vol. 169, no. 3, pp. 1753-1762.

Matsuzaki, Y., Kikuchi, S. & Tomita, M. 2007, "Robust effects of Tsr–CheBp and CheA–CheYp affinity in bacterial chemotaxis", *Artificial Intelligence In Medicine*, vol. 41, no. 2, pp. 145-150.

McEvoy, M.M., Bren, A., Eisenbach, M. & Dahlquist, F.W. 1999, "Identification of the binding interfaces on CheY for two of its targets the phosphatase CheZ and the flagellar switch protein FliM", *Journal Of Molecular Biology*, vol. 289, no. 5, pp. 1423-1433.

McMurry, J.C. 2003, *Organic Chemistry*, 6th edn, Brooks/Cole, CA.

Melo, F. & Feytmans, E. 1997, "Novel knowledge-based mean force potential at atomic level", *Journal Of Molecular Biology*, vol. 267, no. 1, pp. 207-222.

Miller, A.S., Kohout, S.C., Gilman, K.A. & Falke, J.J. 2006, "CheA kinase of bacterial chemotaxis: chemical mapping of four essential docking sites", *Biochemistry*, vol. 45, no. 29, pp. 8699-8711.

Miller, L.D., Russell, M.H. & Alexandre, G. 2009, "Chapter 3 diversity in bacterial chemotactic responses and niche adaptation" in *Advances in Applied Microbiology*, ed. Allen I. Laskin,Sima Sariaslani and Geoffrey M.Gadd, Academic Press, , pp. 53-75.

Moult, J. 2005, "A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction", *Current Opinion In Structural Biology*, vol. 15, no. 3, pp. 285-289.

Muff, T.J. & Ordal, G.W. 2007, "The CheC phosphatase regulates chemotactic adaptation through CheD", *Journal Of Biological Chemistry*, vol. 282, no. 47, pp. 34120-34128.

Muse, S.V. & Gaut, B.S. 1994, "A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome.", *Molecular Biology and Evolution*, vol. 11, no. 5, pp. 715-724.

Naylor, G.J.P., Collins, T.M. & Brown, M.W. 1995, "Hydrophobicity and phylogeny", *Nature*, vol. 373, pp. 565-566.

Nealson, K.H., Platt, T. & Hastings, J.W. 1970, "Cellular control of the synthesis and activity of the bacterial luminescent system", *The Journal Of Bacteriology*, vol. 104, no. 1, pp. 313-322.

Oksanen, E. & Goldman, A. 2010, "Introduction to macromolecular X-Raycrystallography" in *Comprehensive Natural Products II*, eds. Lew Mander & Hung-Wen Liu, Elsevier, Oxford, pp. 51-89.

Oosawa, K., Hess, J.F. & Simon, M.I. 1988, "Mutants defective in bacterial chemotaxis show modified protein phosphorylation", *Cell*, vol. 53, no. 1, pp. 89-96.

Park, S., Beel, B.D., Simon, M.I., Bilwes, A.M. & Crane, B.R. 2004, "In different organisms, the mode of interaction between two signaling proteins is not necessarily conserved", *Proceedings Of The National Academy Of Sciences Of The United States Of America*, vol. 101, no. 32, pp. 11646-11651.

Park, S., Borbat, P.P., Gonzalez-Bonet, G., Bhatnagar, J., Pollard, A.M., Freed, J.H., Bilwes, A.M. & Crane, B.R. 2006, "Reconstruction of the chemotaxis receptor-kinase assembly", *Nature Structural And Molecular Biology*, vol. 13, no. 5, pp. 400-407.

Park, S., Chao, X., Gonzalez-Bonet, G., Beel, B.D., Bilwes, A.M. & Crane, B.R. 2004, "Structure and function of an unusual family of protein phosphatases: the bacterial chemotaxis proteins CheC and CheX", *Molecular Cell*, vol. 16, no. 4, pp. 563-574.

Park, S., Lowder, B., Bilwes, A.M., Blair, D.F. & Crane, B.R. 2006, "Structure of FliM provides insight into assembly of the switch complex in the bacterial flagella motor", *Proceedings Of The National Academy Of Sciences*, vol. 103, no. 32, pp. 11886-11891.

Park, S., Wolanin, P.M., Yuzbashyan, E.A., Silberzan, P., Stock, J.B. & Austin, R.H. 2003, "Motion to Form a Quorum", *Science*, vol. 301, no. 5630, pp. 188.

Parkinson, J.S. & Kofoid, E.C. 1992, "Communication modules in bacterial signaling proteins", *Annual Review Of Genetics*, vol. 26, no. 1, pp. 71-112.

Parkinson, J.S. 1993, "Signal transduction schemes of bacteria", *Cell*, vol. 73, no. 5, pp. 857-871.

Parkinson, J.S. & Revello, P.T. 1978, "Sensory adaptation mutants of *E. coli*", *Cell*, vol. 15, no. 4, pp. 1221-1230.

Pawlowski, M., Gajda, M., Matlak, R. & Bujnicki, J. 2008, "MetaMQAP: A meta-server for the quality assessment of protein models", *BMC Bioinformatics*, vol. 9, no. 1, pp. 403.

Pazy, Y., Motaleb, M.A., Guarnieri, M.T., Charon, N.W., Zhao, R. & Silversmith, R.E. 2010, "Identical phosphatase mechanisms achieved through distinct modes of binding phosphoprotein substrate", *Proceedings Of The National Academy Of Sciences*, vol. 107, no. 5, pp. 1924-1929.

Pei, J., Kim, B. & Grishin, N.V. 2008, "PROMALS3D: a tool for multiple protein sequence and structure alignments", *Nucleic Acids Research*, vol. 36, no. 7, pp. 2295-2300.

Perez, E. & Stock, A.M. 2007, "Characterization of the *Thermotoga maritima* chemotaxis methylation system that lacks pentapeptide-dependent methyltransferase CheR:MCP tethering", *Molecular Microbiology*, vol. 63, no. 2, pp. 363-378.

Perez, E., West, A.H., Stock, A.M. & Djordjevic, S. 2004, "Discrimination between different methylation states of chemotaxis receptor Tar by receptor methyltransferase CheR", *Biochemistry,* vol. 43, no. 4, pp. 953-961.

Primrose, S.B. & Twyman, R.M. 2006, *Principles of Gene Manipulations and Genomics,* 7th edn, Blackwell Publishing, United Kingdom.

Pusey, M.L., Liu, Z., Tempel, W., Praissman, J., Lin, D., Wang, B., Gavira, J.A. & Ng, J.D. 2005, "Life in the fast lane for protein crystallization and X-ray crystallography", *Progress In Biophysics And Molecular Biology,* vol. 88, no. 3, pp. 359-386.

Qian, B., Ortiz, A.R. & Baker, D. 2004, "Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation", *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 101, no. 43, pp. 15346-15351.

Rao, C.V., Glekas, G.D. & Ordal, G.W. 2008, "The three adaptation systems of *Bacillus subtilis* chemotaxis", *Trends In Microbiology,* vol. 16, no. 10, pp. 480-487.

Rao, C.V., Kirby, J.R. & Arkin, A.P. 2004, "Design and diversity in bacterialchemotaxis: a comparative study in *Escherichia coli* and *Bacillus subtilis*", *PLoS Biology,* vol. 2, no. 2, pp. e49.

Rausell, A., Juan, D., Pazos, F. & Valencia, A. 2010, "Protein interactions and ligand binding: from protein subfamilies to functional specificity", *Proceedings Of The National Academy Of Sciences,* vol. 107, no. 5, pp. 1995-2000.

Reva, O.N., Dixelius, C., Meijer, J. & Priest, F.G. 2004, "Taxonomic characterization and plant colonizing abilities of some bacteria related to *Bacillus amyloliquefaciens* and *Bacillus subtilis*", *FEMS Microbiology Ecology,* vol. 48, no. 2, pp. 249-259.

Riek, R., Fiaux, J., Bertelsen, E.B., Horwich, A.L. & Wüthrich, K. 2002, "Solution NMR techniques for large molecular and supramolecular structures", *Journal Of The American Chemical Society,* vol. 124, no. 41, pp. 12144-12153.

Robson, R.L. 1984, "Identification of possible adenine nucleotide-binding sites in nitrogenase Fe- and MoFe-proteins by amino acid sequence comparison", *FEBS letters,* vol. 173, no. 2, pp. 394-398.

Rosario, M.M., Fredrick, K.L., Ordal, G.W. & Helmann, J.D. 1994, "Chemotaxis in *Bacillus subtilis* requires either of two functionally redundant CheW homologs", *The Journal Of Bacteriology,* vol. 176, no. 9, pp. 2736-2739.

Rosario, M.M.L., Kirby, J.R., Bochar, D.A. & Ordal, G.W. 1995, "Chemotactic methylation and behavior in *Bacillus subtilis*: role of two unique proteins, CheC and CheD", *Biochemistry,* vol. 34, no. 11, pp. 3823-3831.

Rost, B. 1999, "Twilight zone of protein sequence alignments", *Protein Engineering,* vol. 12, no. 2, pp. 85-94.

Šali, A. 2009, *MODELLER: A Program For Protein Structure Modeling,* University of California.

Šali, A. & Blundell, T.L. 1993, "Comparative protein modelling by satisfaction of spatial restraints", *Journal Of Molecular Biology,* vol. 234, no. 3, pp. 779-815.

Sánchez, R. & Šali, A. 1997, "Advances in comparative protein-structure modelling", *Current Opinion In Structural Biology,* vol. 7, no. 2, pp. 206-214.

Sanders, D.A., Gillece-Castro, B.L., Stock, A.M., Burlingame, A.L. & Koshland, D.E. 1989, "Identification of the site of phosphorylation of the chemotaxis response regulator protein, CheY", *Journal Of Biological Chemistry,* vol. 264, no. 36, pp. 21770-21778.

Scheeff, E.D. & Fink, J.L. 2003, *Fundamentals of Protein Structure*, John Wiley & Sons, Inc.

Schlick, T. 2002, *Molecular Modeling and Simulation,* Springer-Verlag New York, LLC.

Schrödinger, L.L.C. 2010, *The PyMOL Molecular Graphics System, Version 1.3.*

Shimizu, T.S., Le Novere, N., Levin, D.M., Beavil, A.J., Sutton, B.J. & Bray, D. 2000, "Molecular model of a lattice of signalling proteins involved in bacterial chemotaxis", *Nature Cell Biology,* vol. 2, no. 11, pp. 292-296.

Shimizu, T.S., Aksenov, S.V. & Bray, D. 2003, "A spatially extended stochastic model of the bacterial chemotaxis signalling pathway", *Journal Of Molecular Biology,* vol. 329, no. 2, pp. 291-309.

Shimizu, T.S. & Bray, D. 2002, "Modelling the bacterial chemotaxis receptor complex",*Novartis Foundation Symposium* 247:162-77; discussion 177-81.

Shiomi, D., Zhulin, I.B., Homma, M. & Kawagishi, I. 2002, "Dual recognition of the bacterial chemoreceptor by chemotaxis-specific domains of the CheR methyltransferase", *Journal Of Biological Chemistry,* vol. 277, no. 44, pp. 42325-42333.

Soares, D.C., Barlow, P.N., Newbery, H.J., Porteous, D.J. & Abbott, C.M. 2009, "Structural models of human eEF1A1 and eEF1A2 reveal two distinct surface clusters of sequence variation and potential differences in phosphorylation", *PLoS ONE,* vol. 4, no. 7, pp. e6315.

Soding, J., Biegert, A. & Lupas, A.N. 2005, "The HHpred interactive server for protein homology detection and structure prediction", *Nucleic Acids Research,* vol. 33, no. suppl_2, pp. W244-248.

Springer, W.R. & Koshland, D.E. 1977, "Identification of a protein methyltransferase as the cheR gene product in the bacterial sensing system", *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 74, no. 2, pp. 533-537.

Sprules, T., Kawulka, K.E., Gibbs, A.C., Wishart, D.S. & Vederas, J.C. 2004, "NMR solution structure of the precursor for carnobacteriocin B2, an antimicrobial peptide from *Carnobacterium piscicola*", *European Journal Of Biochemistry,* vol. 271, no. 9, pp. 1745-1756.

Srinivasan, R., Fleming, P.J. & Rose, G.D. 2004, "Ab initio protein folding using LINUS" in *Methods In Enzymology*, ed. Ludwig Brand and Michael L. Johnson, Academic Press, pp. 48-66.

Stephenson, K. & Hoch, J.A. 2002, "Virulence- and antibiotic resistance-associated two-component signal transduction systems of Gram-positive pathogenic bacteria as targets for antimicrobial therapy", *Pharmacology And Therapeutics,* vol. 93, no. 2-3, pp. 293-305.

Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E. & Pupko, T. 2007, "Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach", *Nucleic Acids Research,* vol. 35, no. suppl 2, pp. W506-W511.

Stewart, R.C., Roth, A.F. & Dahlquist, F.W. 1990, "Mutations that affect control of the methylesterase activity of CheB, a component of the chemotaxis adaptation system in *Escherichia coli*", *The Journal Of Bacteriology,* vol. 172, no. 6, pp. 3388-3399.

Stewart, R.C., Jahreis, K. & Parkinson, J.S. 2000, "Rapid phosphotransfer to CheY from a CheA protein lacking the CheY-binding domain", *Biochemistry,* vol. 39, no. 43, pp. 13157-13165.

Stewart, R.C. & Van Bruggen, R. 2004, "Association and dissociation kinetics for CheY interacting with the P2 domain of CheA", *Journal Of Molecular Biology,* vol. 336, no. 1, pp. 287-301.

Stock, J.B. & Baker, M.D. 2009, "Chemotaxis" in *Encyclopedia Of Microbiology*, ed. Moselio Schaechter, Academic Press, Oxford, pp. 71-78.

Sutcliffe, M.J., Hayes, F.R.F. & Blundell, T.L. 1987, "Knowledge based modelling of homologous proteins, part II: rules for the conformations of substituted sidechains", *Protein Engineering,* vol. 1, no. 5, pp. 385-392.

Swain, M.R. & Ray, R.C. 2009, "Biocontrol and other beneficial activities of *Bacillus subtilis* isolated from cowdung microflora", *Microbiological Research,* vol. 164, no. 2, pp. 121-130.

Swanson, W.J., Nielsen, R. & Yang, Q. 2003, "Pervasive adaptive evolution in mammalian fertilization proteins", *Molecular Biology And Evolution,* vol. 20, no. 1, pp. 18-20.

Szurmant, H., Bobay, B.G., White, R.A., Sullivan, D.M., Thompson, R.J., Hwa, T., Hoch, J.A. & Cavanagh, J. 2008, "Co-evolving motions at protein-protein interfaces of two-component signaling systems identified by covariance analysis ", *Biochemistry,* vol. 47, no. 30, pp. 7782-7784.

Szurmant, H. & Hoch, J.A. 2010, "Interaction fidelity in two-component signaling", *Current Opinion In Microbiology,* vol. 13, no. 2, pp. 190-197.

Szurmant, H., Muff, T.J. & Ordal, G.W. 2004, "*Bacillus subtilis* CheC and FliY are members of a novel class of CheY-P-hydrolyzing proteins in the chemotactic signal transduction cascade", *Journal Of Biological Chemistry,* vol. 279, no. 21, pp. 21787-21792.

Szurmant, H. & Ordal, G.W. 2004, "Diversity in chemotaxis mechanisms among the Bacteria and Archaea", *Microbiology And Molecular Biology Reviews,* vol. 68, no. 2, pp. 301-319.

Tina, K.G., Bhadra, R. & Srinivasan, N. 2007, "PIC: protein interactions calculator", *Nucleic Acids Research,* vol. 35, no. suppl 2, pp. W473-W476.

Toews, M.L Goy, M.F., Springer, M.S. & Adler, J. 1979, "Attractants and repellents control demethylation of methylated chemotaxis proteins in *Escherichia coli*", *Proceedings Of The National Academy Of Sciences Of The United States Of America,* vol. 76, no. 11, pp. 5544-5548.

Tramontano, A. & Morea, V. 2003, "Assessment of homology-based predictions in CASP5", *Proteins: Structure, Function, And Bioinformatics,* vol. 53, no. S6, pp. 352-368.

Tzeng, Y. & Hoch, J.A. 1997, "Molecular recognition in signal transduction: the interaction surfaces of the Spo0F response regulator with its cognate phosphorelay proteins revealed by alanine scanning mutagenesis", *Journal Of Molecular Biology,* vol. 272, no. 2, pp. 200-212.

Ullah, A.H. & Ordal, G.W. 1981, "In vivo and in vitro chemotactic methylation in Bacillus subtilis.", *The Journal Of Bacteriology,* vol. 145, no. 2, pp. 958-965.

Usher, K.C., De La Cruz, A.F.A., Dahlquist, F.W., James Remington, S., Swanson, R.V. & Simon, M.I. 1998, "Crystal structures of CheY from *Thermotoga maritima* do not support conventional explanations for the structural basis of enhanced thermostability", *Protein Science,* vol. 7, no. 2, pp. 403-412.

Venclovas, C., Zemla, A., Fidelis, K. & Moult, J. 2003, "Assessment of progress over the CASP experiments", *Proteins: Structure, Function And Bioinformatics,* vol. 53, no. S6, pp. 585-595.

Volz, K. 1993, "Structural conservation in the CheY superfamily", *Biochemistry,* vol. 32, no. 44, pp. 11741-11753.

Wallner, B. & Elofsson, A. 2003, "Can correct protein models be identified?", *Protein Sceince,* vol. 12, no. 5, pp. 1073-1086.

Wang, W., Baker, P. & Seah, S.Y.K. 2010, "Comparison of two metal-dependent pyruvate aldolases related by convergent evolution: substrate specificity, kinetic mechanism, and substrate channeling", *Biochemistry,* vol. 49, no. 17, pp. 3774-3782.

West, A.H., Martinez-Hackert, E. & Stock, A.M. 1995, "Crystal structure of the catalytic domain of the chemotaxis receptor methylesterase, CheB", *Journal Of Molecular Biology,* vol. 250, no. 2, pp. 276-290.

Wu, J., Li, J., Li, G., Long, D.G. & Weis, R.M. 1996, "The receptor binding site for the methyltransferase of bacterial chemotaxis is distinct from the sites of methylation", *Biochemistry,* vol. 35, no. 15, pp. 4984-4993.

Wuichet, K., Alexander, R.P. & Zhulin, I.B. 2007, "Comparative genomic and protein sequence analyses of a complex system controlling bacterial chemotaxis" in *Methods In Enzymology,* ed. Melvin I. Simon,Brian R.Crane and Alexandrine Crane, Academic Press, , pp. 1, 3-31.

Xiang, Z. 2006, "Advances in homology protein structure modeling", *Current Protein And Peptide Science,* vol. 7, pp. 217-227.

Yang, Z. & Bielawski, J.P. 2000, "Statistical methods for detecting molecular adaptation", *Trends In Ecology & Evolution,* vol. 15, no. 12, pp. 496-503.

Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A.K. 2000a, "Codon-substitution models for heterogeneous selection pressure at amino acid sites", *Genetics,* vol. 155, no. 1, pp. 431-449.

Yang, Z., Swanson, W.J. & Vacquier, V.D. 2000b, "Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites", *Molecular Biology And Evolution,* vol. 17, no. 10, pp. 1446-1455.

Yao, J. & Allen, C. 2006, "Chemotaxis is required for virulence and competitive fitness of the bacterial wilt pathogen *Ralstonia solanacearum*", *The Journal Of Bacteriology,* vol. 188, no. 10, pp. 3697-3708.

Yao, W., Shi, L. & Liang, D. 2007, "Crystal structure of scaffolding protein CheW from *Thermoanaerobacter tengcongensis*", *Biochemical And Biophysical Research Communications,* vol. 361, no. 4, pp. 1027-1032.

Yi, X. & Weis, R.M. 2002, "The receptor docking segment and S-adenosyl-homocysteine bind independently to the methyltransferase of bacterial chemotaxis", *Biochimica et Biophysica Acta (BBA) - Protein Structure And Molecular Enzymology,* vol. 1596, no. 1, pp. 28-35.

Zhang, W. & Phillips, G.N. 2003, "Structure of the oxygen sensor in *Bacillus subtilis*: signal transduction of chemotaxis by control of symmetry", *Structure,* vol. 11, no. 9, pp. 1097-1110.

Zimmer, M.A., Tiu, J., Collins, M.A. & Ordal, G.W. 2000, "Selective methylation changes on the *Bacillus subtilis* chemotaxis receptor McpB promote adaptation", *Journal Of Biological Chemistry,* vol. 275, no. 32, pp. 24264-24272.