IMPERIAL COLLEGE LONDON

Evolutionary and Functional Aspects of Two-Component Signalling Systems

by

Xia Sheng

in the
Theoretical System Biology Group
Division of Molecular Biosciences

February 2013

IMPERIAL COLLEGE LONDON

Abstract

Theoretical System Biology Group
Division of Molecular Biosciences

by Xia Sheng

Two-component systems (TCSs) are critical for bacteria to interact with their extracellular environment. They define a type of signalling system that is composed of a transmembrane histidine kinase (HK) and a cytoplasmic response regulator (RR). In this thesis we have studied the evolutionary and functional aspects of these important signalling systems. By studying the distribution of the TCS orthologues of *E.coli* across more than 900 bacterial organisms, we have found that a pair of TCS proteins does not always coexist in one organism. The genomic localisation map of TCSs reveals a possible translocation mechanism for TCS evolution. The alignments of HKs and RRs have shown that HKs are genetically more divergent, probably due to their signal recognizing role. An analysis of the steady states of TCS dynamics has shown that the outputs of the TCSs are bistable if they have positive auto-regulation feedback loops in their transcriptional regulation. Our analysis has also shown that the phosphorylation process of a TCS is always monostable and the factors that affect steady states have been studied. For both orthodox and non-orthodox TCSs, autophosphorylation rates of the HKs are the most important factor to affect the steady states of the TCSs' outputs. To study the phosphorelay mechanism of the non-orthodox TCS ArcB/A, we constructed plasmids carrying different copy numbers of ArcB mutants with different phosphorylation sites ablated. By fitting our phosphorelay model to the data obtained from mutant ArcB constructs, we have found that ArcB most likely performs phosphorelay in an allosteric mechanism. Finally, Approximate Bayesian computation was used in order to evaluate the potential use of orthodox TCS and non-orthodox TCS architectures in synthetic biology contexts. Results show that neither of the orthodox TCS model or the non-orthodox TCS model are superior under all circumstances but that both models have advantages in some scenarios. In the appendix, we did some study on how the contact residue disorder would affect protein-ligand binding specificity.

Acknowledgements

First, I would like to thank Michael Stumpf for his supervision. I am particularly grateful for him always finding time to talk to his group members (no matter how busy he is) and always putting the interests of his group members first. Working with Michael is a great opportunity to learn not only how to do research, but also how to be a respectful leader.

I also want to thank all the previous and present members of the Theoretical Systems Biology group of Imperial College. I want to especially thank Maxime Huvet and John Pinney for their guidance. Their enthusiasm toward science has been a great inspiration. I want to thank Goran Jovanovic, Chris Barnes, Paul Kirk, Daniel Silk, Michal Komorowski, Angelique Ale and Yurong Tang for working with me on some of the projects. I want to thank Nathan Harmston for being fun and for entertaining everyone, and to thank Rodrigo Liberal, Justina Norkunalte and Juliane Liepe for all the delicious cakes. I also want to thank Sarah Fllippi, Georgia Chan, Heather Harrington, Thomas Thorne, Oleg Lenive, Adam MacLean, Siobhan McMahon, Delphine Rolando, Kamil Erguler, Isabel Holmquist, Liam Kelly, Frances Turner, Tina Toni, Waqar Ali, William Bryant, Ryan Topping and everyone else for all the help and discussion.

I also want to thank my parents (Xiaobai Sheng and Daofeng Tong) and my girlfriend (Xiaohong Jiang). Although we are not a rich family, and living in London is very expensive, my parents have still selflessly funded my living costs throughout my PhD. I started with my girlfriend before my PhD, and we have been separated in different countries for four years. I want to thank all of them for always believing in me. Their comforting and encouraging words helped me through this PhD. I owe them too much that I can only try to repay.

At last, I want to thank SHKPKF for funding by paying my tuition fees and part of my living costs. Without their funding, I would not have had the opportunity to study at Imperial College London.

DECLARATION

- This thesis is a presentation of my original research work. Wherever contributions of others are involved, their effort is mentioned at the beginning of each chapter. All else is appropriately referenced to the literature.
- This thesis is protected by copyright and other intellectual property rights. Copyright in the material it contains belongs to the author unless stated otherwise. No information or quotation from the thesis may be published or re-used without proper acknowledgement.
- I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party (*e.g.* a publisher or another researcher).

Contents

Abstract i

Acknowledgements ii

Declaration iii

List of Figures viii

List of Tables xi

Abbreviations xii

1 Introduction 1
- 1.1 Overview . . . 1
- 1.2 Signal transduction pathways . . . 2
- 1.3 General function of two-component system . . . 3
- 1.4 Structural Insights into Histidine Kinases and Response Regulators . . . 4
- 1.5 Diversity in Two-component systems . . . 7
  - 1.5.1 EnvZ/OmpR system and osmolarity response . . . 7
  - 1.5.2 CheA,B,Y in chemotaxis . . . 9
  - 1.5.3 ArcA/B responding to oxygen level and pressure change . . . 9
  - 1.5.4 PhoP/Q controlling Mg^{2+} response . . . 10
- 1.6 Specificity of TCS pairs . . . 11

2 Evolutionary study of Two-Component Systems 14
- 2.1 Background . . . 14
- 2.2 Methods . . . 16
  - 2.2.1 Searching for orthologues of TCS using `BLAST` . . . 16
  - 2.2.2 Searching for additional orthologues of TCS using `HMMER`3.0 . . . 17
  - 2.2.3 Principles of `BayesTraits` . . . 19
  - 2.2.4 Correlation test with `BayesTraits` . . . 21
  - 2.2.5 Colocalization analysis . . . 21
  - 2.2.6 Horizontal Gene Transfer test . . . 22
  - 2.2.7 Protein sequence similarity comparison . . . 23
  - 2.2.8 Hypergeometric test . . . 23
  - 2.2.9 Co-existing non-cognate TCSs graph . . . 24

2.3 The Dataset . . . 24
2.4 Results and Discussion . . . 25
2.4.1 Aligning the orthologues of TCSs in *E. coli* across all species . . . 25
2.4.2 Horizontal gene transfer test in distant organisms . . . 27
2.4.3 Colocalization analysis of the TCSs across the organisms . . . 30
2.4.4 Co-evolution test using `BayesTraits` . . . 34
2.4.5 Lifestyle test using `BayesTraits` . . . 34
2.4.6 TCS protein sequences similarity comparison . . . 35
2.4.7 Cross-talk partner inference . . . 38
2.5 Conclusions . . . 41

3 Study on the steady states of Two-component systems 46
3.1 Background . . . 46
3.2 Methods . . . 49
3.2.1 Modelling of slow reactions of two-component systems . . . 49
3.2.2 Modelling of rapid reactions of two-component systems . . . 50
3.2.3 Calculation of steady states . . . 53
3.2.4 Parameters configuration of the models . . . 54
3.2.5 Parameters randomization . . . 54
3.3 Results . . . 55
3.3.1 The model of transcriptional/translational reactions of TCS is bistable . . . 55
3.3.2 The model of post-translational reactions of TCS is monostable . . . 56
3.3.3 Numerical test for steady state in TCS post-translational reaction models . . . 57
3.3.4 Factors that could affect the steady state of orthodox and non-orthodox models . . . 57
3.3.5 Transient $[RRp]$ of a cross-talking TCS model . . . 58
3.3.6 Steady state dynamics of TCS models with cross-talk . . . 60
3.4 Conclusions . . . 62

4 The Phosphorelay Mechanism of Non-Orthodox Two-Component Systems 65
4.1 Background . . . 65
4.1.1 ArcB/A two-component system . . . 65
4.1.2 The ArcB dimer might work in a different mechanism than other TCS . . . 66
4.2 Materials and Methods . . . 67
4.2.1 Models for phosphorelay . . . 67
4.2.2 Simulation of phosphorelay models . . . 72
4.2.3 Parameter optimisation and model evaluation . . . 74
4.2.4 Sensitivity assays . . . 75
4.2.5 Bacterial strains, media and growth conditions . . . 75
4.2.6 DNA manipulations . . . 76
4.2.7 Western blot analysis . . . 76
4.2.8 In vivo bacterial two-hybrid system . . . 77
4.2.9 β-Galactosidase (β-Gal) assays . . . 77
4.3 Results . . . 78

4.3.1 Modelling of monomer and dimer histidine kinase of two-component system . . . 78
4.3.2 Modelling trans- and cis-phosphorelay of non-orthodox two-component system . . . 79
4.3.3 Activities of the constitutively active ArcB* phosphorelay mutants 80
4.3.4 WT and mutant forms of ArcB and ArcB* proteins were similarly expressed when placed on compatible plasmids pCA24N and pAPT110 . . . 82
4.3.5 Combination of ArcB mutants . . . 83
4.3.6 Fitting experimental data with systematic models. . . . 85
4.3.7 Sensitivity assay of proposed phosphorelay models . . . 87
4.4 Conclusions . . . 88

5 Bayesian design of synthetic two-component system 97
5.1 Background . . . 97
5.1.1 Engineering methods in synthetic biology . . . 97
5.1.2 Bayesian approach in system design . . . 99
5.1.3 Tow-component systems in this study . . . 101
5.2 Methods . . . 102
5.2.1 Bayesian design . . . 102
5.2.1.1 Model selection using ABC . . . 102
5.2.1.2 Prior distribution . . . 104
5.2.1.3 The distance function and output tolerance . . . 104
5.2.1.4 Deterministic models . . . 105
5.2.2 Modelling two component systems . . . 105
5.2.2.1 Models . . . 105
5.2.2.2 Distance . . . 107
5.2.2.3 Priors . . . 108
5.3 Results . . . 108
5.4 Conclusions . . . 109

6 Conclusions 114

A The roles of contact residue disorder and domain composition in characterizing protein-ligand binding specificity and promiscuity 118
A.1 background . . . 118
A.2 Methods . . . 120
A.2.1 Protein-ligand complex dataset . . . 120
A.2.2 Contact residues . . . 120
A.2.3 Protein domains . . . 120
A.2.4 Domain assignments to superfamilies . . . 121
A.2.5 Gene Ontology annotations . . . 121
A.2.6 Disorder predictions . . . 121
A.2.7 Statistical methods . . . 121
A.3 Results . . . 122

A.3.1 Analysis of disorder and domain composition of proteins with different numbers of ligands . . . 122
A.3.1.1 Analysis of contact residue disorder . . . 122
A.3.1.2 The impact of ordered multiple domains on ligand numbers 124
A.3.1.3 Gene Ontology enrichment . . . 126
A.3.2 The property of hub contact residues binding multiple ligands . . . 131
A.3.3 Disorder distribution and superfamily domain mapping in several ligand classes . . . 133
A.4 Conclusions . . . 137

Bibliography **139**

List of Figures

1.1 Schematic representation of a two-component system. 5
1.2 Diagram of domain composition of the *E. coli* two-component system ArcA/B. 6
1.3 Diagram of cross-talk in TCSs in *E. coli* summarised from Yammoto's work [194]. 13

2.1 Diagram of the strategy for `HMMER` search. 18
2.2 Histogram of the distance between each TCS pairs on the chromosome. For each TCS pair, only the ones with distance under 1000bp is drawn in this figure. 22
2.3 Orthologues of *E. coli* K-12's TCSs found by either `BLAST` or `HMMER`. . . . 26
2.4 TCS orthologues of *E. coli* K-12 ordered by their groups. 28
2.5 Orthologues of *Salmonella* TCSs. 29
2.6 Diagram of intervals within TCS pairs. 31
2.7 Sequence alignment of the domains in histidine kinases in *E. coli* K-12. . . 37
2.8 Sequence alignment of the receiver domains in respond regulators in *E. coli* K-12. 38
2.9 Sequence similarity and identity of TCS in each species. 39
2.10 Sequence similarity and similarity ratio of TCSs in each species. 40
2.11 Co-existing pairs of the candidate TCSs aligned with the phylogenetic tree. 41
2.12 Diagram of cross-talk between a TCS and orphans or other TCSs can give rise to more flexible signal transduction networks. 43

3.1 Schematic representation of slow and fast reactions involved in a typical two-component system. 48
3.2 A diagram of post-translational reactions of the three models we have built. 51
3.3 Simulation of the transcription/translation model of TCS. 55
3.4 Orthodox and non-orthodox post-translational TCS models' steady states responding to changes of different parameters. 59
3.5 Transient concentration of RRp in simple TCS and cross-talking TCS models responding to varying stimulus. 60
3.6 Simple and cross-talking post-translational TCS models' steady states responding to changes of parameters. 61

4.1 Diagram of constitutively active ArcB and mutants. 82
4.2 Phosphorelay of constitutively active ArcB variants lacking the key phosphorelay residues. 83

4.3 Phosphorelay of constitutively active ArcB variants lacking the key phosphorelay residues. . . . 84
4.4 Phosphorelay of constitutively active ArcB variants lacking the key phosphorelay residues . . . 85
4.5 Phosphorelay of constitutively active ArcB variants lacking the key phosphorelay residues. . . . 86
4.6 Phosphorelay of constitutively active ArcB variants lacking the key phosphorelay residues. . . . 87
4.7 Phosphorelay of constitutively active ArcB variants lacking the key phosphorelay residues. . . . 88
4.8 Simulation of three phosphorelay models with more constrained parameter settings. . . . 89
4.9 Simulation of three phosphorelay models with more flexible parameter settings. . . . 90
4.10 Schwarz Weight of three phosphorelay models with optimised parameters under more flexible parameter configuration. The probabilities are: trans-phosphorelay model, 0.43%; cis-phosphorelay model, 0.02%; allosteric, 99.55% . . . 91
4.11 Sensitivity of each phosphorelay model on all the parameters. The meaning of each parameter is described in Materials and Methods section. . . . 92
4.12 Contributions of individual parameters into each eigenvalue for all three models. . . . 93
4.13 FIM for all pairs of parameters for trans-phosphorelay model visualised as heat maps. . . . 94
4.14 FIM for all pairs of parameters for cis-phosphorelay model visualised as heat maps. . . . 95
4.15 FIM for all pairs of parameters for allosteric phosphorelay model visualised as heat maps. . . . 96

5.1 Bayesian approach to system design. . . . 100
5.2 Bacterial two-component systems. . . . 111
5.3 Two component systems: evolution to the noise reduction behaviour. . . . 112
5.4 Two component systems: posterior distribution for the non-orthodox system to achieve the noise reduction behaviour. . . . 112
5.5 Two component systems: evolution to the signal reproduction behaviour. 113
5.6 Two component systems: posterior distribution for the orthodox system to achieve the signal reproduction behaviour. . . . 113

A.1 Distribution of disordered contact residues of proteins with 1, 2, 3, 4 and 5 bound ligands in eukaryotes and prokaryotes. . . . 123
A.2 Distribution of disordered contact residue. Comparison of the proteins with one and more than one ligand in eukaryotes and prokaryotes. . . . 124
A.3 Pfam domain frequency. . . . 125
A.4 Association of contact residue disorder with GO annotations in ordered single domain proteins and ordered multiple domain proteins. . . . 127
A.5 Association of contact residue disorder with GO annotations in ordered single domain proteins and ordered multiple domain proteins. . . . 128
A.6 Association of contact residue disorder with GO annotations in ordered single domain proteins and ordered multiple domain proteins. . . . 129

A.7 Association of contact residue disorder with GO annotations in ordered single domain proteins and ordered multiple domain proteins. 130
A.8 Amino acid composition. Comparison of the residue composition for the whole protein, for the contact residues and for the hub contact residues. 131
A.9 Distribution of disordered contact residues of proteins binding HETAIN, HETAI and HETAIN&HETAI in eukaryotes and prokaryotes. 134
A.10 Association of protein frequency with Superfamily domain. 135
A.11 The top 58 superfamily domains with more than 20 proteins. 138

List of Tables

2.1 Example of the model of evolution for a trait that adopts three states . . 20
2.2 A table of orphan and total TCS discovered before and after using `HMMER`. 27
2.3 Table of the GC content of TCSs coding genes in *Bacillus cereus AH820*, *Bacillus anthracis Ames* and *Bacillus anthracis str Sterne* and their average GC content. 30
2.4 Correlation test of TCS pairs. In the table listed the correlation between lifestyles and TCSs. 33
2.5 Correlation test of TCS existence and lifestyles. The experiment was done using Fisher's exact test and `BayesTraits`. 34
2.6 In the table we indicate evidence for coevolution and for correlation between lifestyles and TCS proteins in *Salmonella typhi*. 36
2.7 Table of the numbers of non-cognate components that each TCS protein can cross-talk with. 42

3.1 List of parameter values used in model I and model II. Model I and Model II share most parameters while differ in the forward and reverse phosphotransfer rates of HK. 54

4.1 List of parameter values used in the phosphorelay models. 73
4.2 *E. coli* K-12 strains and plasmids used in this study 81

A.1 Kolmogorov-Smirnov test for the distributions of PDB contact residue disorder binding different numbers of ligands 122
A.2 Frequency of proteins with ordered multiple domains and its binomial proportions test. 126
A.3 Frequency of proteins with ordered multiple domains and with less than 10% disordered contact residues and its binomial proportions test 126
A.4 Frequency of amino acids on the whole proteins, the contact residues and the hub contact residues. 132
A.5 Scop domains significantly enriched in proteins with HETAIN, HETAI and HETAIN&HETAI. (P-value < 0.2) 136

Abbreviations

ABC	**A**pproximate **B**ayesian **C**omputation
ATP	**A**denosinel **Tri**ph**p**hosphate
BIC	**B**ayesian **I**nformation **C**riteria
BLAST	**B**asic **L**ocal **A**lignment **S**earch **T**ool
CA	**C**atalytic **A**ctivity
FDR	**F**alse **D**iscovery **R**ate
FIM	**F**isher **I**nformation **M**atrix
HGT	**H**orizontal **G**ene **T**ransfer
HisKA	**His**tidine **K**in**a**se (Domain)
HK	**H**istidine **K**inase
HPt	**H**istidine-containing **P**hospho**t**ransfer (Domain)
LHS	**L**atin **H**ypercube **S**ampling
LR	**L**ikelihood **R**atio
MCMC	**M**arkov **C**hain **M**onte **C**arlo
ML	**M**aximum **L**ikelihood
ODE	**O**riginal **D**ifferential **E**quation
RBH	**R**eciprocal **B**est **H**it
REC	**Rec**eiver (Domain)
RR	**R**esponse **R**egulator
SMC	**S**equential **M**onte **C**arlo
TCS	**T**wo-**C**omponent **S**ystem

Chapter 1

Introduction

1.1 Overview

The class of two-component regulatory systems (TCSs) is one of the most elementary and widespread systems for signal perception and transduction of external environmental stimuli. The concept of two-component systems was first introduced in the model bacterium *Escherichia coli* almost three decades ago [160]. Only a few TCSs have been found in eukaryotic organisms, such as the higher plant *Arabidopsis thaliana* and the yeast *Saccharomyces cerevisiae* [114]. In *Arabidopsis thaliana*, TCSs are involved in regulation of sophisticated plant development processes controlled by hormones, and the plant clock function that generates circadian rhythms. In *Saccharomyces cerevisiae*, a TCS was found to regulate an osmosensing MAP kinase cascade [102].

Approximately 50,000 TCSs have been identified from genome sequences [48], almost all of which are in prokaryotic genomes. Bacteria are single-celled organisms and each individual has the required factors for life, assimilation, growth and reproduction. Each bacterium is an independent cell (although the cells in some species remain attached to, or communicate with one another). Thus, as a way to communicate with extracellular environment, TCSs are very crucial to bacterial survival. The study of TCS in bacteria becomes very important for our understanding of how bacteria respond and interact with their environment.

1.2 Signal transduction pathways

A cell must be responsive to specific signals in its environment. Magnesium ions, for example are chemical signals for the PhoP/Q system [74]. The basic functions of signal transduction pathways are to sense stimuli and to process the signal. These molecular circuits are able to detect and sense different kinds of extracellular stimuli, and to trigger responses such as gene expression, protein post-translational modification, *etc.* In this section we summarise the most important steps of signal processing in signal transduction pathways. Diverse responses of the TCSs to signals will be reviewed in other sections.

Signal transduction processes always start with receiving of environmental signals (ion concentration changes, oxygen level changes, *etc.* [18, 74]), mostly via a cell-surface receptor. Then the signal is translated into other chemical substrates to pass on the information inside the cell.

- Transferring information into the cell from the environment by membrane receptors.

  Most chemical signal molecules are too large or too polar to pass through the membrane. Thus, the information that the signal molecules carries must be transmitted into the cell without the molecules entering the cell. Receptor proteins located across the membrane have evolved to fulfil this mission.

  Such receptor proteins are transmembrane proteins with both extracellular and intracellular domains. Binding sites on the extracellular domain can specifically recognise the signal molecule. The interaction of the signalling molecule and the receptor changes the conformation of the receptor, including the intracellular domain. These structural changes are still not enough to cause a cellular response as the changes are only restricted to a small number of molecules on the membrane. The information that the signal molecules carry must be translated into other forms to change cellular activity [161].

- Secondary messengers carrying the information from the activated receptor.

  Changes in the concentration of small molecules, also known as secondary messengers, are the next step of the signal transduction. Some very important secondary messengers include cyclic AMP and cyclic GMP, calcium ions, inositol

1,4,5-trisphosphate (IP3), and diacylglycerol (DAG) [90]. Secondary messengers are very popular in eukaryotic cells. However, the TCS pathways that we are studying in this thesis do not involve any secondary messengers.

- Protein phosphorylation is a common means of signal transfer.

  The transfer of a signal inside the cells is normally accomplished by phosphorylation of the target protein. Protein kinases are activated by the signal. These enzymes transfer phosphate groups from ATP to specific residues in proteins. Histidine kinases in TCS functions as both extracellular signal receptors and kinases. Another enzyme called protein transferase mediates the transfer of phosphate group from certain residue to another unphosphorylated specific residue on the same molecule or other specific molecules. In a TCS, the histidine kinase itself is responsible for catalysing the reaction transferring phosphate group from phosphorylated HK to unphosphorylated RR. Protein phosphorylation is reversible. Protein phosphatases are enzymes that remove specific phosphate groups through a hydrolysis reaction [18, 74, 161], and histidine kinases also exhibit phosphatase activity under the right conditions.

- Termination of signal.

  Protein phosphatases provide one important mechanism for the termination of a signal. After a signalling process has been initiated and the proper cellular response has been achieved, the signalling process must be terminated. An unstoppable signalling process could make the cells lose their responsiveness to new signals or cause a great deal of energy to be wasted [18, 74, 161].

1.3 General function of two-component system

TCSs can sense environmental stresses and transduce the information to the inside of cells for processing and adaptation. They consist of a membrane-bound histidine sensor kinase that senses a specific environmental stimulus and a corresponding response regulator that mediates the cellular response, mostly through regulating the expressions of target genes [161]. Figure 1.1 shows a general model depicting the activation of a TCS by a certain stimulus. Upon stimulation, signal transduction occurs through the transfer of phosphate groups from adenosine triphosphate (ATP) to histidine residues

in the histidine kinases through autophosphorylation. Subsequently the histidine kinase catalyses the transfer of the phosphate group on the phosphorylated histidine residues to aspartic acid residues on the response regulator. The response regulator's own conformation changes, which then typically leads to up/down regulation of target genes. The level of phosphorylation of the response regulator controls its activity [160]. In the case of non-orthodox TCSs, the HK contains HisKA (Histidine Kinase domain), REC (Receiver domain) and HPT (Histidine Phosphate Transfer domain) domains which all have one phosphorylation site.

Many two-component systems are involved in signalling systems and can sense the changes of the external environments such as temperature [30], oxygen level [110], osmolarity [26], chemo attractants [135] and pH [5]. For instance, in *E. coli* the EnvZ/OmpR TCS senses osmolarity changes and controls the differential expression of the outer membrane porin proteins OmpF and OmpC [26]. The PhoP-PhoQ system in *Edwardsiella tarda* detects changes in environmental temperature. Protein secretion is activated from 23 to 35 ℃, but it is suppressed at or below 20 ℃, or above 37 ℃ [30]. For the two component system ArcB/A, it has been shown that ArcB is activated under anaerobic and subaerobic conditions and is much less active under fully aerobic and microaerobic conditions [18].

1.4 Structural Insights into Histidine Kinases and Response Regulators

In most TCSs each component has only one phosphate-binding domain. Signal transduction starts with the autophosphorylation of the histidine residue on the histidine kinase. Subsequently, the histidine kinase catalyses the transfer of the phosphate group at the histidine residue to aspartic acid residue on the response regulator. The response regulator's own conformation changes, which then enables its DNA binding or other down-stream activities; this TCS architecture is referred to as *orthodox*. However, in some two-component systems such as ArcB/ArcA [49], TorR/TorS [70], BarA/UvrY [149] and EvgA/EvgS [174] the HKs are replaced by more complicated hybrid histidine

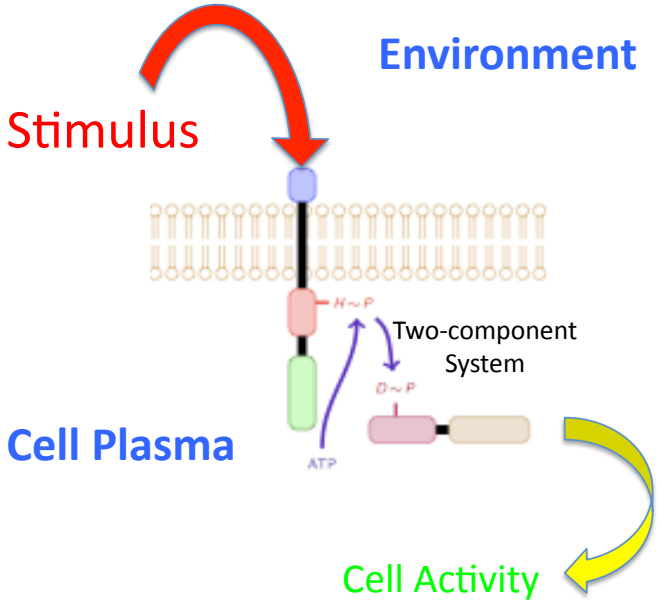


FIGURE 1.1: Schematic representation of a two-component system. Extra-cellular stimuli cause the transfer of phosphate groups from adenosine triphosphate (ATP) to histidine residues in the histidine kinases. Subsequently the histidine kinase catalyses the transfer of the phosphate group of the phosphorylated histidine residues to the aspartic acid residues on the response regulator. The response regulator's own conformation changes, and hence leads to the change in expression of target genes.

kinases, which each contains a histidine kinase (HisKA) domain, a receiver (REC) domain and a histidine-containing phosphotransfer (HPt) domain; such TCSs are called *non-orthodox* TCSs.

Compared to the one step phosphorylation of simple TCS, hybrid HKs transfer the phosphate group through a three-step phosphorelay, where the phosphate group will be transferred from HisKA to REC, then to HPt, and at last to the RR. These non-orthodox TCSs can be less sensitive to noisy inputs: only for a prolonged or sufficiently strong signal will the response be initiated [82]. The ArcB/A system is a global regulator of gene expression under microaerobic and anaerobic growth conditions. We will use it as an example to present some structural insights into non-orthodox TCS. The histidine kinase ArcB is located across the cell membrane, and the response regulator ArcA resides inside the cytoplasm. As shown in the diagram in figure 1.2, ArcB consists of a transmembrane domain, a leucine zipper, a PAS domain. ArcB is a non-orthodox sensor kinase possessing three cytosolic catalytic domains: an N-terminal transmitter domain (H1), a central receiver domain (D1), and a C-terminal phosphotransfer or secondary transmitter domain (HPt/H2) [65, 68]. The central receiver domain, about 120 amino

acid residues in length, is homologous to the ArcA receiver domain. The conserved aspartate residue at position 576 has been demonstrated to be the conserved phosphorylation group receiving site in D1. The Hpt (H2) domain transfers the phosphate group from D1 to itself and then to the REC (D2) domain on the RR. Phosphorylation activates a variable effector domain of the response regulator, which triggers the cellular response. In ArcA, aside from the phosphor-receiving domain (D2), the C-terminal effector domain contains DNA and RNA polymerase binding sites which are responsible for activation or deactivation of the gene expression of the corresponding targets.

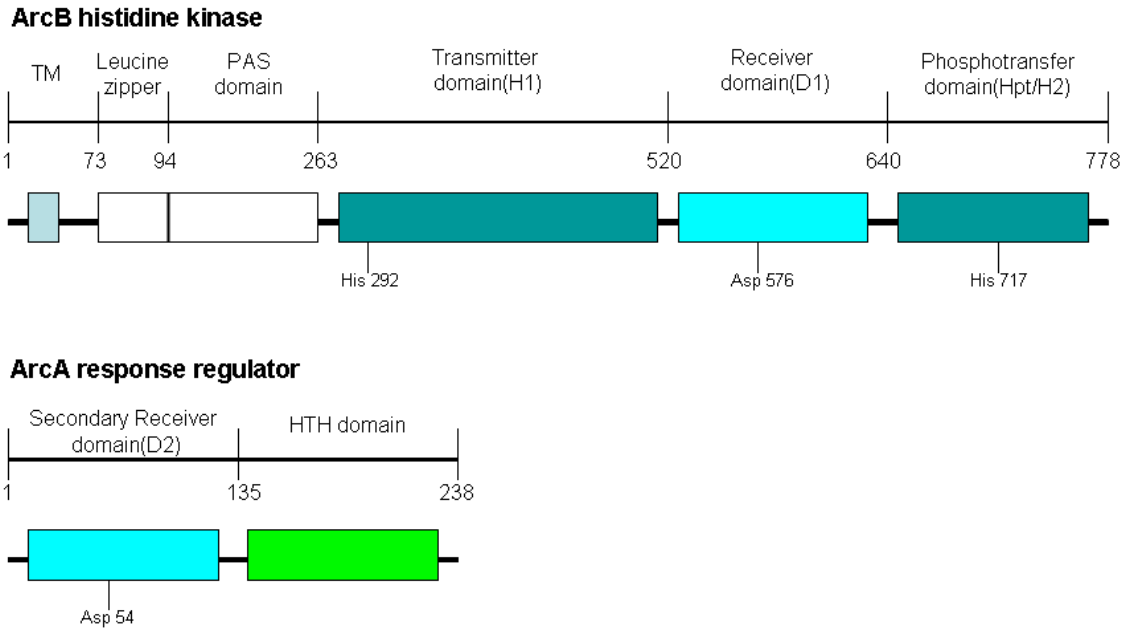


FIGURE 1.2: Diagram of domain composition of the *E. coli* two-component system ArcA/B. ArcB is attached to the plasma membrane by a transmembrane domain corresponding to residues 23 to 41 and residues 58 to 77. The linker region contains a putative leucine-zipper and a PAS domain. The primary transmitter domain (H1) contains the conserved His292, the receiver domain (D1) contains the conserved Asp576, and the histidine phosphotransfer domain (H2) contains the conserved His717. ArcA is shown with its N-terminal receiver domain containing the conserved Asp54 and its C-terminal is helixturnhelix (HTH) domain which is responsible for DNA binding.

Most sensor kinases possess a periplasmic domain which is approximately 150 amino acid residues in length, separated by two transmembrane segments, and believed to be involved in signal reception [162]. By contrast, the periplasmic sequence of ArcB is extremely small and only 16 amino acid residues long. The transmembrane domain of ArcB is immediately followed by residues which appears to have a feature characteristic of the leucine zipper motif [1, 25]. The feature of this motif is an amphipathic helix with hydrophobic residues gathered on one face and hydrophilic residues on the opposite

face, and a leucine residue at the first position in each of four contiguous heptad-repeats (LeuX6LeuX6LeuX6Leu). In general, leucine zippers are found in DNA-binding regulatory proteins [1, 25], but are also present in membrane proteins that do not bind to DNA [93, 193]. Previous studies showed that such a motif is involved in homo- or heterodimer formation through interaction of the helices from two monomers [182]. ArcB was proposed to contain a PAS domain in its linker, which is the region connecting the transmembrane domain to the catalytic domains [201]. The PAS motif has been shown to be present in proteins from all kingdoms of life, Bacteria, Archaea, and Eucarya [134, 201]. In prokaryotes, many PAS containing proteins are sensor kinases, and it has been demonstrated that they bind various small molecules such as heme, flavin, and a 4-hydroxycinnamyl chromophore to sense molecular oxygen, redox potential, and light, respectively [170]. And it was suggested that the PAS domain in ArcB is required for sensing the redox conditions [107].

In contrast to the majority of the other TCSs which perform a simple and direct one-step phosphorylation process, the non-orthodox two-component systems, such as ArcB/A, follow a four-step reaction cascade , which is also known as phosphorelay [51]. The process, take ArcB/A as an example, is initiated when anoxic conditions cause a shift in the conformation of ArcB that favours activation of its kinase activity. The active kinase then undergoes autophosphorylation at His292 by accepting the γ-phosphate of ATP [68]. The phosphate group from His292 is then transferred to Asp576, which in turn is able to direct it reversibly to His717 or irreversibly to H_2O [51]. Subsequently the phosphate group from His717 is transferred to Asp54 of ArcA [88].

1.5 Diversity in Two-component systems

The two-component systems can sense all kinds of extracellular signals and response accordingly. Here are some examples to show the diversity of TCSs.

1.5.1 EnvZ/OmpR system and osmolarity response

EnvZ/OmpR is a TCS responding to osmolarity changes in *E.coli*. It consists of the histidine kinase EnvZ and response regulator OmpR. EnvR can sense the osmolarity change and then regulates the phosphorylation state of response regulator OmpR. OmpR

is a transcription factor controlling the expression of the genes of several membrane porin proteins. The osmolarity of the cell can then be adjusted by those genes.

EnvZ is a histidine kinase composed by 450 amino acids. Same as other histidine kinases, EnvZ is located on the membrane of the cell. The main part of the molecular (residues 180-450) resides inside the cell. The cytoplasmic domain consists of the linker region at 180-220, the DHp (dimerization/histidine-containing phosphotransfer) domain at 223-289 and the CA (catalytic activity) domain at 290-450. The cytoplasmic domain of EnvZ is reported to have both kinase and phosphatase activities [128]. The periplasmic domain acts as sensor to detect signals of osmolarity change. The role of the linker region is to transduce the input signal (osmolarity change) from the periplasmic sensor domain to the cytoplasmic catalytic domain [127]. OmpR is the response regulator of this TCS. It has an N-terminal Che-Y-like receiver domain with a highly conserved Asp-55 phosphorylation site and a C-terminal DNA-binding domain [75]. A linker region connects these two domains together. Upon phosphorylation at the Asp-55 conserved site, the transcription activity of OmpR is greatly enhanced. It can bind to the regulatory sequences upstream of the promoters ompF and ompC [61].

EnvZ, a histidine kinase, undergoes autophosphorylation on the highly conserved His-243 residue [195]. The high energy phosphate group is subsequently transferred to the conserved Asp-55 residue of OmpR. Phosphorylated OmpR (OmpR-P) serves as a transcription factor modulating the expression of the major outer membrane porin genes, ompC and ompF. OmpC and OmpF form channels in the outer membrane, which allows passive diffusion of small hydrophilic molecules of less than 650 Da in size [118]. Further analysis has shown that EnvZ also has phosphatase activity, and can dephosphorylate OmpR-P [200]. In the case where the histidine kinase can both phosphorylate and dephosphorylate the cognate response regulator, we call the histidine kinase bifunctional. It has been proposed that osmotic signal regulates the ratio of the kinase-to-phosphatase activity of EnvZ to modulate the level of cellular OmpR-P. This ratio regulation is done primarily by altering the phosphatase activity [69]. The cellular OmpR-P expression levels reciprocally regulate the transcription of ompF and ompC. At low medium osmolarity a reduced level of OmpR-P, due to the decreased kinase/phosphatase activity ratio of EnvZ, favours the transcription of ompF gene. At high medium osmolarity an elevated OmpR-P level, resulting from the increased kinase/phosphatase activity ratio of EnvZ, allows the activation of ompC transcription. On the other hand, more OmpR-P

molecules bind to the ompF promoter upstream region causing repression of ompF gene expression.

1.5.2 CheA,B,Y in chemotaxis

CheA, B, Y participate in *E. coli* chemotaxis. Chemotaxis is the phenomenon whereby bacteria direct their movements according to certain chemicals in their environment. Because chemotaxis requires the cell to respond within 200 ms of stimulus application, CheA, B and Y are different from other TCSs. The response regulators of this TCS, which are CheB and CheY, do not regulate gene expression as other TCSs normally do. Instead, a signalling cascade of protein phosphorylation and dephosphorylation reactions is used to control cellular movement responding to environmental chemical changes.

The signalling cascade starts from CheA, the histidine kinase. CheA transfers its phosphate group after stimulated to two competing RRs, CheB and CheY which have reciprocally functions, to activate them. Activated CheY binds to FliM protein in the flagellar motors to augment clockwise rotation. On the other hand, activated CheB specifically demethylates MCPs (methyl-accepting chemotaxis proteins) to shift the flagellar motors to the counterclockwise state [22, 24, 129]. We only include this system in our evolutionary study but have not considered this interacting model in our other studies. Chemotaxis is a more complicated process than the TCS functions considered here and its analysis a highly active research field in its own right.

1.5.3 ArcA/B responding to oxygen level and pressure change

The general diagram showing the structure and function of this TCS has already been introduced in the previous section (Figure 1.2). So in this section we will only summarise the environmental signal and function of ArcA/B.

Life evolved on earth about 4 billion years ago when anaerobic conditions prevailed [10]. Thus life evolved from the fermentative modes of metabolism. The emergence of photosynthesis happened about 2 billion years ago, which generated the oxygen required for respiration processes [126]. Only by then, was the more effective aerobic respiratory modes of metabolism available. The fermentative modes of metabolism may waste some carbon and hydrogen compounds as excretion products. Respiratory modes of

metabolism are more efficient but can not work without oxygen [190]. Thus, facultative bacteria, *e.g. E.coli*, need to adjust their metabolism mode in response to different oxygen conditions.

When the condition is anaerobic, nitrate becomes the preferred electron acceptor. The utilisation of nitrate is regulated by the TCSs NarX/L [158] and NarQ/P [140]. Gene expression in response to O_2 concentration decrease is regulated by the cytosolic global regulator Fnr [89] and the ArcB/A two component system [67]. Under anaerobic growth condition Fnr enhances the expression of genes in the anaerobic respiratory pathways. It also represses the expression of some genes involved the aerobic respiratory pathways [181]. The ArcB/A two-component system functions mainly by repressing the expression of genes in microaerobic or aerobic conditions. The genes depressed by ArcB/A include several TCA cycle enzymes, some flavoprotein dehydrogenases, and ubiquinone oxidase [66, 67]. In a few cases, ArcA can activate expression of genes, like cytochrome *d* oxidase encoding gene [151].

ArcB is a histidine kinase which is responsible for sensing the environmental signal, in this case, O_2 concentration. However, the mechanism of signal recognition by ArcB is not known yet. It has been suggested that ArcB can indirectly sense O_2 concentration by monitoring the level of a metabolite that can exist in either an oxidized or reduced form. The quinone electron carriers located on the membrane were raised as possible compounds. *E.coli* can synthesize three different quinones, ubiquinone, menaquinone and demethylmenaquinone. They function as adapters between electron donators and electron accepters [55, 108, 180]. Ubiquinone is mainly used in aerobic respiration while menaquinone and demethylmenaquinone serve mainly in anaerobic respiration. A study has shown that the quinones could signal the redox conditions to ArcB [50]. Another study has shown that quinones do not activate a kinase in their reduced form, but pass a inhibit signal in their oxidized form [103]. Thus it has been suggested that this system is regulated by being silenced by oxidized quinones in aerobic condition [104].

1.5.4 PhoP/Q controlling Mg^{2+} response

PhoP/Q is a two-component regulatory system that controls temperature response and several virulence properties in the gram-negative bacterium *Salmonella typhimurium*.

It also controls the adaption to extracellular Mg^{2+}. The PhoQ protein is a Mg^{2+} sensor that respond to the periplasmic Mg^{2+} level by changing its conformation. PhoP is a response regulator protein that is necessary to modify approximately 3% of the *Salmonella* genes. Genes directly controlled by the PhoP protein often differ in their promoter structures. Many of the genes affected code proteins essential for growth at low Mg^{2+} concentration.

The PhoQ protein is a bi-functional sensor that has both kinase (*i.e.*, autokinase and phosphotransferase) and phosphatase activities. It is an inner membrane protein that contains two transmembrane regions and a cytoplasmic domain that includes the subdomain with the conserved histidine residue (His 277) that is the site of PhoQ phosphorylation, as well as the subdomain involved in phosphatase activity and ATP binding. PhoP protein of *Salmonella* is a member of the OmpR family response regulators. It consists of a N-terminal receiver domain which contains a conserved aspartate residue (Asp 52) which is the site for phosphorylation and C-terminal DNA binding domain with helix-turn-helix motif [153].

When the signal of low Mg^{2+} is sensed by the histidine kinase PhoQ in the periplasmic domain, the conformation of PhoQ changes, allowing it to autophosphorylate. The phosphorylated PhoQ then transfers the phosphate group to its cognate receptor, PhoP, causing related genes to be regulated [153]. It has been suggest that periplasmic domain of *Salmonella* PhoQ can also detect extracellular antimicrobial peptides [7] and pH level[137]. However, this notion has been questioned [54].

1.6 Specificity of TCS pairs

As we have described in the previous sections, TCSs share similar structure elements like the kinase activity domain, the receiver domain. These kind of similarity would enable organisms to rapidly evolve their information processing capabilities without requiring major changes. However, this similarity comes at a cost where the components must have high level of specificity for distinct pathways and avoid unwanted cross-talk. Specificity of pathways can be achieved by several mechanisms. In some cases, organisms apply spatial mechanisms such as scaffolds and subcellular localisation, or tissue-specific expression (multicellular organisms) to improve specificity. In some cases, differential

timing of gene expression is applied. In many cases, the mechanism of ensuring specificity is improving molecular recognition [117, 156, 159, 198].

TCSs have high specificity to their cognate partners. In the phosphotransfer reactions, HKs are known to exhibit a much larger preference *in vitro* for their *in vivo* cognate RRs rather than any other RRs [43, 52]. This finding suggests that a TCS maintains its specificity by a high level of molecular recognition. The regions determining specificity were proposed to be near phosphotransfer active site and active histidine site in HKs, within and near α helixes in RRs. By changing residues in these areas, Jeffrey Skerker *et al.* successfully created chimeric histidine kinases that can specifically phosphotransfer to designated non-cognate RRs [155].

Although having high specificity, reports have shown that there still remains some level of cross-talk in TCSs. The most simple cross-talking compound is phosphorylated acetyl. ArcA and some other RRs have been reported to autophosphorylate by consuming high-energy acetyl phosphate, *in vitro* and *in vivo* [37, 38, 100]. The cross-talking cases among TCS proteins include CheA phosphorylate OmpR, EnvZ phosphorylating NtrC [64], *etc.*. Cross-talking of TCSs has also been confirmed *in vivo* [184]. Ymamoto *et al.* studied the cross-talking pairings of all the TCSs of *E. coli, in vitro* [194]. Figure 1.3 summarising the cross-talking network by Ymamoto's article [194] shows how vastly cross-talking exists among all the TCSs.

These potential cross-talk options might aid relative pathways *in vivo*. When one RR is phosphorylated by non-cognate HK, the target genes under the control of this particular RR may respond to various stimuli. On the other hand, when one HK phosphorylates multiple RRs, one type of stimulus may initiate responses associate with other stimuli. We have studied the evolutionary basis for cross-talk and dynamic behaviour of cross-talk in our thesis.

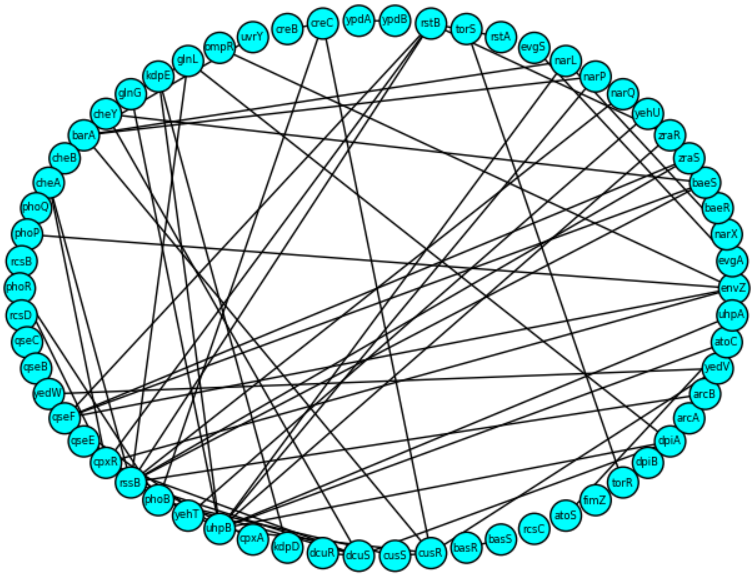


FIGURE 1.3: Diagram of cross-talk in TCSs in *E. coli* summarised from Yammoto's work [194].

Chapter 2

Evolutionary study of Two-Component Systems

Part of this work has been published as: "Evolutionary characteristics of bacterial two-component systems. "

Sheng X, Huvet M, Pinney JW, Stumpf MP. Adv Exp Med Biol. 2012;751:121-37.

2.1 Background

In most TCSs each component has only one domain with a residue that can be phosphorylated: HKs with one histidine kinase domain and RRs with one aspartate receiver domain. This TCS architecture is referred to as *orthodox*. However, in some two-component systems such as ArcB/ArcA [49], TorS/TorR [70], BarA/UvrY [149] and EvgS/EvgA [174] the HK is replaced by more complicated hybrid histidine kinases, which contain a histidine kinase (HisKA) domain, a receiver (REC) domain and a histidine-containing phosphotransfer (HPt) domain; such TCSs are called *non-orthodox* TCSs. These *non-orthodox* TCSs can exhibit ultra-sensitive behaviour and are less sensitive to noisy inputs: only for a prolonged or sufficiently strong signal will the response be initiated [82].

There are two plausible and competing models for the evolution of TCS pairs: the recruitment model and the coevolution model. In the recruitment model, novel TCSs

evolve through gene duplication of one component, which then co-opts components from heterologous systems to recruit a new cognate part. This model is supported by the structural similarity among response regulators, in which only a few residues are sufficient to determine specificity; further supporting evidence comes from the observed cross-talk of TCS within an organism. In the co-evolution model, novel TCS evolve by global duplication of both components and subsequent differentiation. This model is supported by the fact that many TCS co-occur on a chromosome [87, 147, 188].

In the past, functional arguments have been put forward in favour of the evolution of TCS architectures. Such functional reasoning can, however, be only a rough guide as we have very little solid data on *e.g.* kinetic parameters or sufficiently resolved *in vivo* data of TCS mediated signalling. In the absence of population genetic data, the best source of information on the impact of natural selection, we have to turn to the wealth of sequence information available across the prokaryotic kingdom. Here we therefore use comparative analysis of TCSs in order to investigate possible causes underlying the patterns of TCS inheritance. But in order to link to any functional information available on TCSs we consider those TCSs that have been characterised experimentally in *Escherichia coli* and other bacterial model organisms.

As more and more genome data have been obtained for bacteria, a lot of studies have been done on genome scale of two-component systems. Galperin, M *et. al.* have studied "bacterial IQ" (which reflects the abundance of signal transduction components in a given organism divided by the square of its genome size) of bacteria and archaea and found that the total number of histidine kinase increases proportionally with the square of the genome size in each organism [46, 47]. Seshasayee, A *et.al.* presented a comparative genomics survey of regulatory system proteins from about 850 prokaryotic genomes. Their study confirmed that the numbers of TCSs each scale in a quadratic fashion with genome size. Also they have suggested that cross-talking signalling network may be very prevalent across bacteria [152]. Williams, R *et. al.* have classified the gene and domain organisation of TCS gene loci from 1405 prokaryotic replicons and found that more than 90% of all TCSs locate within 200 bp to their cognate partner on the genomes [188]. Most of previous genome scale surveys of the TCSs identify the TCSs using protein domain based searching method (*i.e.* identify protein with HisKA domain as HK, protein with REC domain as RR) or are using databases based on this method (like P2CS).

In this chapter we apply a different approach to search for TCSs in the bacterial organisms by finding orthologues of *E.coli* K12 TCSs. Orthologues are defined as genes derived from a single ancestral gene in the last common ancestor of the compared species. The definition of orthologues has nothing to do with function. However, a crucial property of orthologues, which is both theoretically plausible and empirically supported, is that they usually perform equivalent functions in the respective organisms [86]. Thus, by finding orthologues of *E.coli* K12 TCSs, we can find proteins performing equivalent functions in other organisms. And to perform the equivalent functions, these proteins are very likely to communicate with the orthologues of the cognate partners in *E.coli* K12. Thus, by finding orthologues, we can also identify the partnership of these TCSs in all the bacterial organisms. Subsequently, we can study the distribution of the orphan TCSs (TCS proteins with their cognate partners absent in the same organism), and we can also study the localization of the TCS pairs in a different perspective.

We have also compared the orthologues data and life style data of the bacteria. And we have compared the sequence difference among the HKs and the RRs respectively. We have found results either support the coevolution model or the recruitment model.

2.2 Methods

2.2.1 Searching for orthologues of TCS using `BLAST`

We used 23s rRNA sequences as benchmarks to calculate each organism's distance to *E. coli* K-12 (or *Salmonella typhi*); the 23s rRNA evolves slowly and in a clock-like fashion and is therefore better suited for gauging evolutionary distances between species than any single gene or protein sequence [63]. The sequences were aligned using the multiple sequence alignment tool `MAFFT` (*http://mafft.cbrc.jp/alignment/software/*) [77] using default parameters. The function `dist.dna` in the `R` package `ape` (*http://cran.r-project.org/web/packages/ape/index.html*) was used to calculate the pairwise distance between every two organisms. A symmetric matrix of pairwise distances was generated, from which each organism's evolutionary distance to *E. coli* K-12 (or *Salmonella typhi*) was obtained [9]. The sequences of *E. coli* K-12 (or *Salmonella typhi*) TCSs were retrieved from P2CS [9] (MIST returns the same result [179]).

To identify the orthologues of all the 30 two-component systems (TCSs) of *E. coli* K-12 (as well as 19 TCSs in *Salmonella typhi*) in all the other 950 bacteria species, a reciprocal best hit (RBH) approach [62, 115] was applied. The Basic Local Alignment Search Tool (`BLAST`) is used as the software tool for RBH. `BLAST` can find local similarity between sequences. It uses the query sequences to compare with the database and score for all the possible results. The higher the score, the more reliable the result is.

For each TCS protein in *E. coli* (or *Salmonella typhi*), we used `BLAST` to search sequence databases for homologues. The identified sequences were ordered by their `BLAST` alignment scores, from high to low. We used the top hit in the result to search back in *E. coli* (or *Salmonella typhi*) protein database using `BLAST`. If the original queried *E. coli* (or *Salmonella typhi*) TCS sequence was the top hit in the second search result, then these two proteins were identified as orthologues. A problem for this approach is that it might occasionally catch sequences which matches very well, but are in fact very short in length, which is obviously not the result we want to obtain. To avoid this problem, we set the threshold of the proportion of the alignment length compared to the whole sequence length to be 50%. And we set the threshold of the identity between the aligned fragments to be 30%, which means the identical part between the query sequences and result sequences should be longer than 30% of the whole aligned fragment length. Only those sequences that fulfil all the above restrictions were identified as orthologues of their TCS counterparts in *E. coli* K-12 (or *Salmonella typhi*).

2.2.2 Searching for additional orthologues of TCS using `HMMER`3.0

`BLAST` is reasonably accurate and fast at identifying homologues and orthologues. However, as it uses protein sequences directly to search for proteins (perhaps processing similar cellular function), it might miss protein orthologues in distant organisms where the sequences have changed a lot, while still maintaining the same function. To make sure we miss as few true positive orthologues as possible we need a more sensitive tool. Thus we chose `HMMER`, which applies probabilistic model profiles to search for closely related proteins. `HMMER` is supposed to be more accurate and better at detecting remote homologs compared to `BLAST` and other sequence alignment and database search tools based on older scoring methodologies [191].

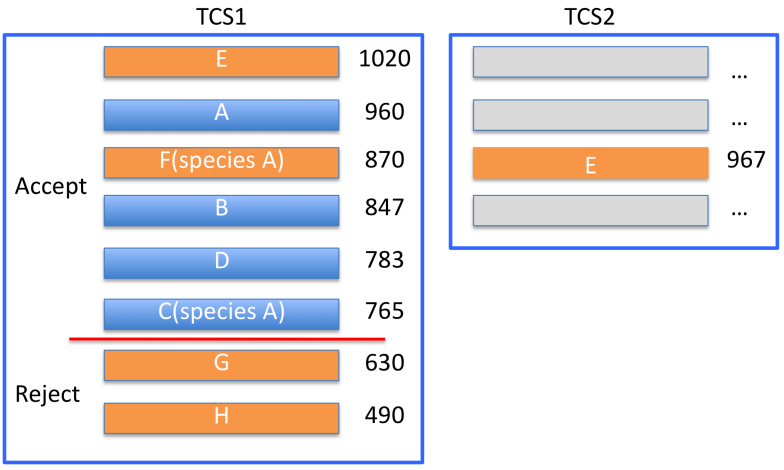


FIGURE 2.1: Diagram of the strategy for HMMER search. It shows the HMMER result of protein TCS1 and TCS2. Blue boxes represent the orthologue sequences obtained by BLAST search. Orange boxes represent the results newly obtained by HMMER search. Grey boxes represent the sequences obtained by either method. The numbers on the right side of the boxes indicate the score of corresponding sequences. Sequence F and G are from the same species, species A. For the search result of TCS1, sequence C of BLAST result has the lowest score 765. So the threshold is set to 765, while any result with lower score is rejected. Sequence E appears in both TCS1 and TCS2 results. As it has higher score in the TCS1 result, we consider it as an orthologue of TCS1. Sequence F and sequence C are both from species A. As F scores higher than C, sequences is considered as the orthologue of TCS in species A.

The software version of HMMER we used is HMMER3.0. As HMMER needs to first build a probabilistic profile based on the alignment of known orthologues, we used all the predicted orthologue sequences identified by BLAST to build the HMM model profile, which was then used to search in the whole bacterial protein sequence database. The result contains as many as 50, 000 entries, which is obviously too many. Thus, we performed some further pruning to decrease the size of the result. For each protein, we used the lowest score of the identified orthologues (obtained by BLAST) as a threshold. All the hits with scores less than the threshold were discarded. For each sequence that appears in more than one proteins' result, we categorised it as the orthologue of the protein for which it had the highest score, considering them as more related. If there was more than one candidate hit for a certain TCS protein in a certain species, the hit with the highest score was taken. Figure 2.1 gives a example of how the results were processed. The 50% alignment length threshold was also applied here, which means the alignment length was required to be longer than 50% of the query sequence length or 50% of the profile length.

2.2.3 Principles of `BayesTraits`

`BayesTraits` is a software used to analyse trait evolution among groups of species where a phylogeny is available. It can be applied to the analysis of traits of discrete states, or continuously varying traits. Hypotheses about pairs of traits related to evolution information can be tested with this software.

`BayesTraits` uses Markov chain Monte Carlo (MCMC) methods to derive posterior distributions and maximum likelihood (ML) methods to derive point estimates of the parameters of statistical models, and the values of traits at ancestral nodes of phylogenies.

`BayesTraits` can be used with a single phylogenetic tree. In this case, only uncertainty about model parameters is tested. It can also be applied to samples of trees when in that case phylogenetic uncertainty is taken into account.

In this software package, BayesMultiState is used to test the correlation of traits of discrete states evolving on phylogenetic trees. This is useful for testing models of trait evolution. It can be applied to traits that adopt two or more discrete states (as in our research) [124].

Multistate method and Discrete method in the package fit continuous-time Markov models to the discrete character data. This model allows the trait to change from the state it is in at any given moment to any other state over infinitesimal intervals of time. The rate parameters of the model estimate these transition rates. The model traverses the tree estimating transition rates and the likelihood associated with different states at each node. Table 2.1 shows an example of the model of evolution for a trait that can adopt three states, 0,1, and 2. The q_{ij} are the transition rates among the three states, and these are what the method estimates on the tree, based on the distribution of states among the species. If these rates differ from zero, this indicates that they are a significant component of the model. The main diagonal elements are not estimated but are a function of the other values in their row. For a trait that adopts four states, the matrix would have twelve entries, whereas for a binary trait the matrix would have just two entries. `BayesTraits` does not test hypotheses but provides the information needed to make hypotheses tests. The likelihoods ratio (LR) test is often used to compare two likelihoods derived from nested models (models that can be expressed such that one is

State	0	1	2
0	-	q_{01}	q_{02}
1	q_{10}	-	q_{12}
2	q_{20}	q_{21}	-

TABLE 2.1: Example of the model of evolution for a trait that adopts three states

a special or general case of the other). The likelihood ratio statistic is calculated as

$$\text{LR} = 2[\text{log-likelihood(Better Fitting Model)} - \text{log-likelihood(Worse Fitting Model)}]$$

The likelihood ratio statistic is distributed as a χ^2 random variable with degrees of freedom equal to the difference in numbers of parameters between the two models. However, in some circumstances [124] the test may follow a χ^2 with fewer degrees of freedom. Variants of the LR test include the Akaike Information Criterion and the Bayesian Information Criterion. These tests will not be described here. They are discussed in many works on phylogenetic inference and statistical model choice [124]. The LR, Akaike and Bayesian Information Criterion tests presume that the likelihood is at or near its maximum likelihood value. In a MCMC framework tests of likelihood often rely on Bayes factors. The logic is similar to that for the likelihood ratio test, except here we compare the marginal likelihoods of two models rather than their maximum likelihoods. The marginal likelihood of a model is the integral of the model likelihoods over all values of the models parameters and over possible trees. In practice this marginal likelihood is difficult to estimate but it can be well approximated by the harmonic mean of the likelihoods allowing the Markov chain to run for a very large number of iterations (typically millions).

`BayesTraits` calculates the logarithm of the harmonic mean of the likelihoods as the program runs, having ignored the likelihoods during the burn-in period when the model is moving towards convergence. The running tally of harmonic means is read from the final iteration of the chain and the values for the independent and dependent models are then compared. The test statistic is just

$$2(Log[\text{Harmonic Mean (Better Model)}] - Log[\text{Harmonic Mean (Worse Model)}])$$

Any positive value favours the dependent model, but conventionally a ratio greater than

2 is taken as positive evidence, greater than 5 is strong and greater than 10 is very strong evidence.

2.2.4 Correlation test with `BayesTraits`

The software `BayesTraits` was downloaded from *http://www.evolution.reading.ac.uk/BayesTraits.html*. The traits *TCS gene existence* or *oxygen requirement*, *pathogenic behaviour*, *taxon group*, *habitat place*, *living temperature etc.* of 635 bacteria organisms with at least one known lifestyle were taken as traits for the analysis. For example, the trait of oxygen requirement contains four states: aerobic, anaerobic, microaerobic, facultative. The likelihood is calculated as:

$$\text{Likelihood ratio} = -2[Log(Lh_{independent}) - Log(Lh_{dependent})]$$

where "independent" and "dependent" denote the mode of evolution of two traits with respect to each other. From the asymptotic chi-square distribution we obtain the p value of each lifestyle and gene combination. The null hypothesis was that the two traits were independent. If $p < 0.05$, the null hypothesis would be rejected, and the tested gene and lifestyle would be correlated. The False Discovery Rate (FDR) was controlled at the level of 5%, in order to correct for the number of tests performed.

2.2.5 Colocalization analysis

The orthologues identified by `BLAST` and `HMMER` were used to screen the organisms that contain both components for all TCS pairs. We then tracked the genomic locations of the genes that are translated into these proteins using the gene data obtained from the NCBI genome database. The sequences located further than 1000bp from each other or are located on different chromosomes can clearly not be considered to be colocalized. As shown in figure 2.2, for the sequences shorter than 1000bp, there are only a limited number of TCS pairs (38 pairs out of 5007 pairs in all) separated by between 200bp and 400bp. So 300bp was selected as a criteria to determine colocalization. If both components of the pair were on the same chromosome, and the distance between them was less than 300bp, we considered them as colocalized.

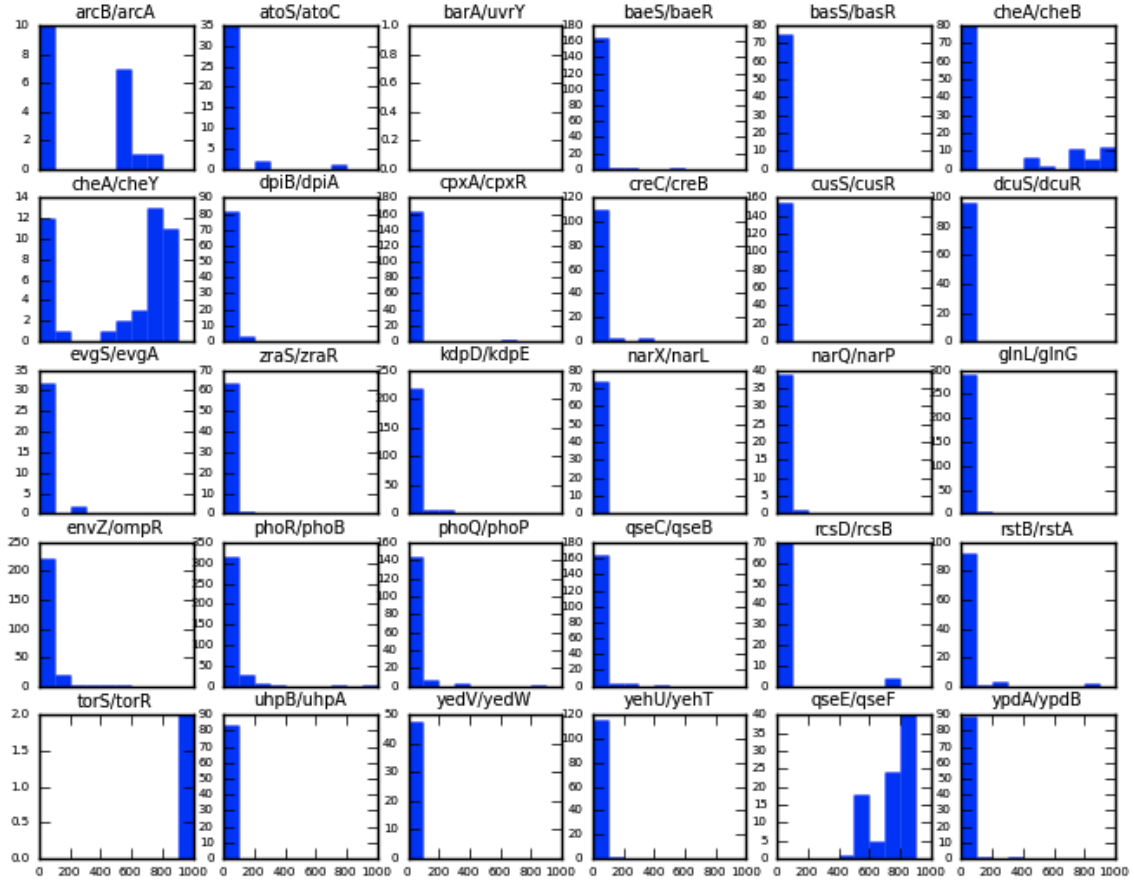


FIGURE 2.2: Histogram of the distance between each TCS pairs on the chromosome. For each TCS pair, only the ones with distance under 1000bp is drawn in this figure.

2.2.6 Horizontal Gene Transfer test

When the organisms are encoding proteins, they may have different preferences for certain codons combinations. One of the consequences is that they might have quite different GC content. Here we are using GC content as a benchmark to detect horizontal gene transfer in a number of candidate genomes. We downloaded cDNA sequence files of all the available bacterial organisms from the NCBI cDNA database (*http://www.ncbi.nlm.nih.gov*). The `Python` module `Biopython` (*http://biopython.org/wiki/Biopython*) was used to parse the sequences and to calculate GC content of all the cDNA sequences [33]. GC content of orthologues of dpiB/A, dcuS/R, baeS/R, ypdA/B coding cDNAs were analysed in the organisms *Bacillus cereus AH820, Bacillus anthracis Ames* and *Bacillus anthracis str. Sterne*. The average GC content in the above three organisms and *E. coli* K-12 was calculated. Sequences shorter than 300bp are discarded.

2.2.7 Protein sequence similarity comparison

Protein sequences information was retrieved from the NCBI protein database (*http://www.ncbi.nlm.nih.gov*). HisKA, HPT and REC regions of the proteins were extracted from the whole length using the `Python` module `Biopython` (*http://biopython.org/wiki/Biopython*). The sequences were then aligned using the multiple sequence alignment tool `MAFFT` (*http://mafft.cbrc.jp/alignment/software/*) with default parameter setting [76, 78]. The alignment files were viewed using `Jalview` (*http://www.jalview.org/*) with all the parameters again set to default values, except for the colour which is set to ZAPPO. The alignment result was analysed by the `R` package `seqinr` (*http://cran.r-project.org/web/packages/seqinr/index.html*) using the `dist.alignment` function, which calculates similarity and identity of aligned sequences (with 'method' parameter set to 'similarity' or 'identity').

2.2.8 Hypergeometric test

To analyse the probability of the orphan TCS components being cross-talking with non-cognate components, we used the hypergeometric test to test the significance of the correlation between every HK/RR and their possible non-cognate counterparts (29 RR/HK for each HK/RR). Briefly, if we were testing for a candidate interacting non-cognate RR for certain HK, we set the total number of species to be g, which is 950 in our case. The number of species with this HK as an orphan is k; the number of species with the RR's orthologues present is m; the number of species with both HK orphans and certain RR's orthologues is x (overlapping number). Then the probability of observing no less than x overlaps by chance is given by

$$\text{P = phyper (x-1 , m , g-m , k , lower.tail = FALSE)}$$

We used the upper tail of the distribution where orphan HK/RR coexist with their proposed cross-talking partners. The false discovery rate was taken care of by using `R` function p.adjust with the "method" parameter set to "fdr".

2.2.9 Co-existing non-cognate TCSs graph

After the above hypergeometric test, we had got 363 pairs of non-cognate co-inherited pairs (p value < 0.05). For these possible cross-talking partners, we chose the 78 pairs among them with a p value of less than 1×10^{-10}, together with 5 randomly chosen HK/RR pairs as negative control. For each chosen pair, if both components exist in certain organism, the value was 1. For all the other situation, the value was 0. A graph for coexisting non-cognate TCSs is drawn, and the species were aligned according to the 23S rRNA phylogenetic tree. The numbers of non-cognate components that each TCS protein could cross-talk with were extracted from the list of 363 non-cognate pairs we had obtained.

2.3 The Dataset

The protein amino acid sequences, genome sequences and 23s rRNA sequences of 950 bacterial organisms were downloaded from NCBI (protein, RNA and complete genome databases) in fasta format in May of 2009. The TCS sequences and annotations of *E. coli* K-12 were downloaded from the TCS database P2CS (`www.p2cs.org/`) [9] in May of 2009. The sequences were manipulated using `Python` module `Biopython` (`http://biopython.org/wiki/Biopython`)

For lifestyle traits, the feature *habitat* included *aquatic*, *host-associated*, *multiple*, *specialized* and *terrestrial*. *Gram stain* contained traits *positive* and *negative*. *Oxygen requirement* included *aerobic*, *anaerobic*, *facultative* and *microaerophilic*. *Temperature range* included *hyperthermophilic*, *mesophilic*, *psychrophilic* and *thermophilic*.

We downloaded lifestyle information of 635 organisms with known lifestyle from NCBI in June of 2009; the lifestyles of the bacteria include habitat, gram stain, oxygen requirement and temperature range. Among our species we have habitat information for 517 organisms, we know the oxygen requirements for 462 organisms, for 477 organisms we know gram stain information, and for 503 we have information about the temperature ranges at which they exist.

2.4 Results and Discussion

2.4.1 Aligning the orthologues of TCSs in *E. coli* across all species

We searched for orthologues of 62 *E. coli* K-12 TCSs in 950 bacteria organisms with `BLAST` applying reciprocal best hit approach. There are several observations worth noting in the result obtained `BLAST` (figure 2.3). The result is ordered by species' distances to *E.coli* K-12.

- In some organisms (in the columns close to position 300), although they are not too far from *E. coli* in phylogenetic distance, there are rarely any orthologues of TCSs, while the organisms near them all have abundant TCS orthologues.
- On the other hand, there are some organisms (around position 450 in figure 2.3) having orthologue "islands" for dpiA/B and dcuS/R, while the organisms nearby have almost no orthologues for these genes.
- A TCS is a system which functions through two parts. However, in the figure there are many orphan HKs and RRs which do not have their cognate partners in a number of organisms.

For the first point, we found all the organisms involved are endosymbiotic bacteria in certain insects. It is reasonable to infer that their living environment is very homogeneous and predictable, and thus they have lost some of the abilities of responding to outside stimulations.

For the second point, we suspected that this was caused by horizontal gene transfer, and performed suitable tests described later in this chapter.

For the last point, generally histidine kinases phosphorylate their cognate response regulator with higher efficiency. But it has also been reported that some TCS can cross-talk with non-cognate counterparts [194]. Cross-talk could be an effective mean to integrate signals to control multiple outputs. In figure 2.3, we can see that many TCS proteins are missing their cognate HKs or RRs. To function they have to cooperate with component from other TCSs. We still have not found solid evidence supporting that cross-talk exists in the organisms missing one of the TCS partners. However, cross-talk seems to

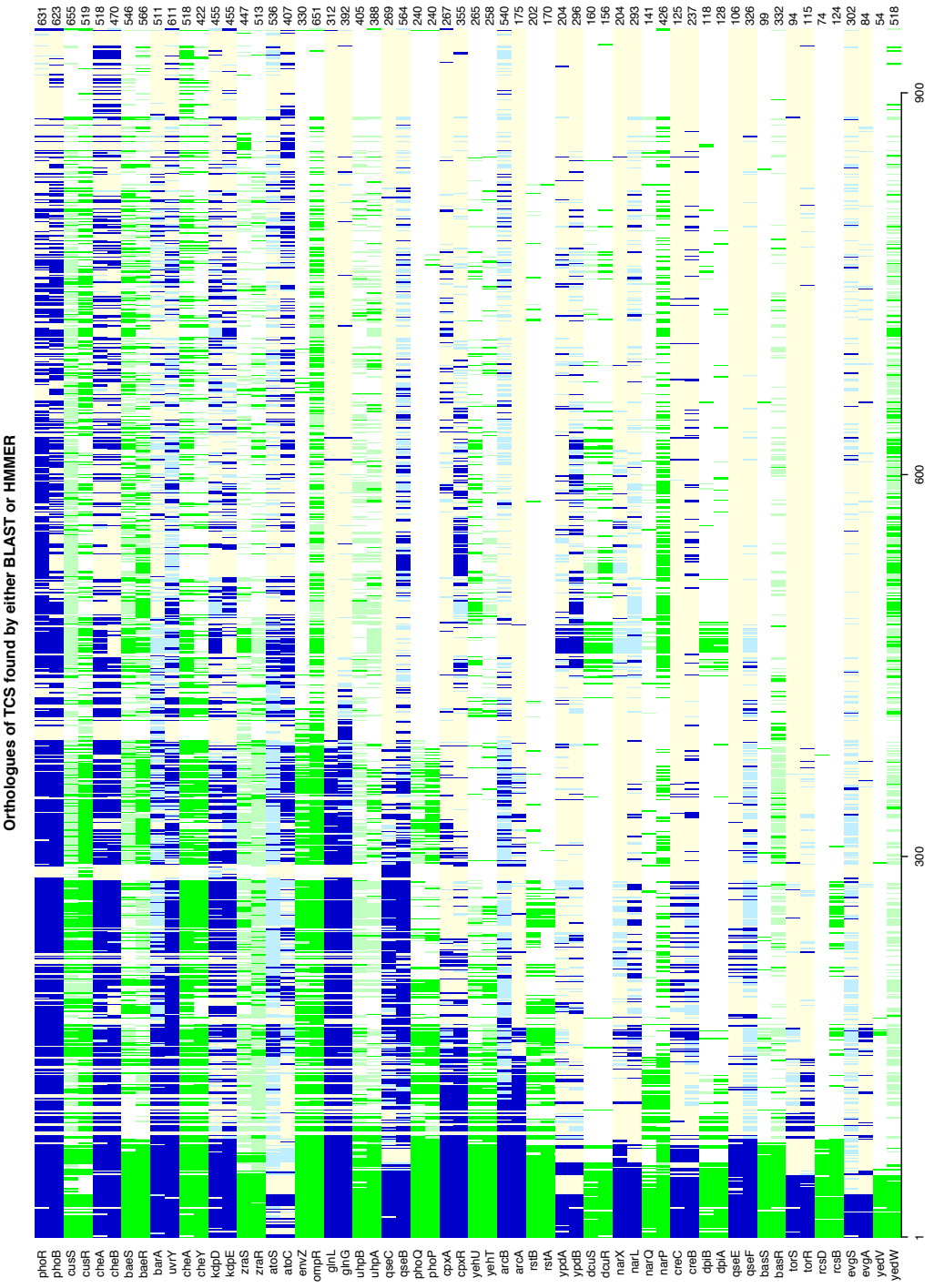


FIGURE 2.3: Orthologues of *E. coli* K-12's TCSs found by either `BLAST` or `HMMER`. The X axis lists the orthologues ordered by the distance of the organism to *E. coli* K-12, from left to right, from close to distant. On the left side of Y axis are the names of the TCS proteins, while on the right side listed the total numbers of orthologues discovered. Green and blue stripes/areas indicate orthologues discovered by `BLAST` in that certain species, where white and light yellow stripes/areas mean there is no orthologue in the corresponding species. Light green and light blue stripes/areas indicate the orthologues that were identified by `HMMER`, while missed by `BLAST`. The neighbouring proteins with the same colour set are cognate histidine kinases and response regulators. HKs are drawn on the upper side for each pair of TCS.

be a natural and promising way to explain why so many organisms lack their cognate partners.

However, we have two ways in which we can understand the orphan HKs and RRs in figure 2.3. First, as stated above, the orphans might be able to cross-talk with some non-cognate TCS proteins in the organism. Second, it is also possible that our method is not sensitive enough, and the orphans arose just because we failed to discover their cognate parts. To test the later possibility, we applied a more sensitive programme, `HMMER` 3.0, to identify more orthologues.

Method	Data type	HK	RR
BLAST	total	6112	7777
	orphan	1118	2783
	orphan/total	18.29%	35.79%
BLAST&HMMER	total	9328	10814
	orphan	2101	3587
	orphan/total	22.52%	33.17%

TABLE 2.2: A table of orphan and total TCS discovered before and after using `HMMER`.

In the `HMMER` result (figure 2.3, Y axis, right side), new TCS orthologues were found across all species(the numbers of newly found orthologues ranges from 0 to 426 for different organisms). These new orthologues filled some of the "holes" in the `BLAST` results. But still there were lots of TCS proteins in the figure missing their cognate parts, as shown in table 2.2, which leads us back to the conclusion that some TCSs apply cross-talk as an alternative pathway to implement missing cognate parts.

One may argue that ordering by distance to *E. coli* as shown in figure 2.3 would not reflect their evolutionary relationship. Thus we have ordered the same result (without separating the `BLAST` and `HMMER` results) by their genetic group. The result shown in figure 2.4 is similar to the result ordered by organism distance (figure 2.3).

We have done the similar procedure on the TCSs of *Salmonella typhi*. As shown in figure 2.5, the result has similar pattern as that of *E. coli* K-12.

2.4.2 Horizontal gene transfer test in distant organisms

As shown in figure 2.3 organisms *Bacillus cereus AH820*, *Bacillus anthracis Ames* and *Bacillus anthracis str Sterne* have orthologues of dpiB/A, dcuS/R, baeS/R, ypdA/B.



FIGURE 2.4: TCS orthologues of *E. coli* K-12 ordered by their groups. The X axis lists the orthologues ordered by the distance of the groups to *E. coli* K-12, from left to right, from closely to distantly related groups (as measured by the 23s rRNA). On the left side of Y axis are the names of the TCS proteins, while on the right side listed the total numbers of orthologues discovered. ArcB, TorS, BarA and EvgS are hybrid histidine kinase in *non-orthodox* TCS (marked by grey background). Light green stripes/areas indicate orthologues discovered by `BLAST` or HMM in that certain species, whereas light yellow stripes mean that no orthologue in the corresponding species could be identified. For each pair of TCS proteins we indicate the presence of the HK above the RR. Red squares indicate the "islands" of species in the phylogeny that showed higher levels of conservation of TCS pairs that evolutionary far from *E.coli* K-12.

However, in the organisms closer to or further from *E. coli*, we rarely found any orthologues for the TCS pairs mentioned above. This suggests that those *Bacillus* organisms might have got these genes from *E. coli* many generations ago. In that case, features of the TCS genes in those *Bacillus* should be different from other genes in the same organism, while more similar to that of *E. coli*.

We tested our horizontal transfer hypotheses by determining the GC content of the genes.

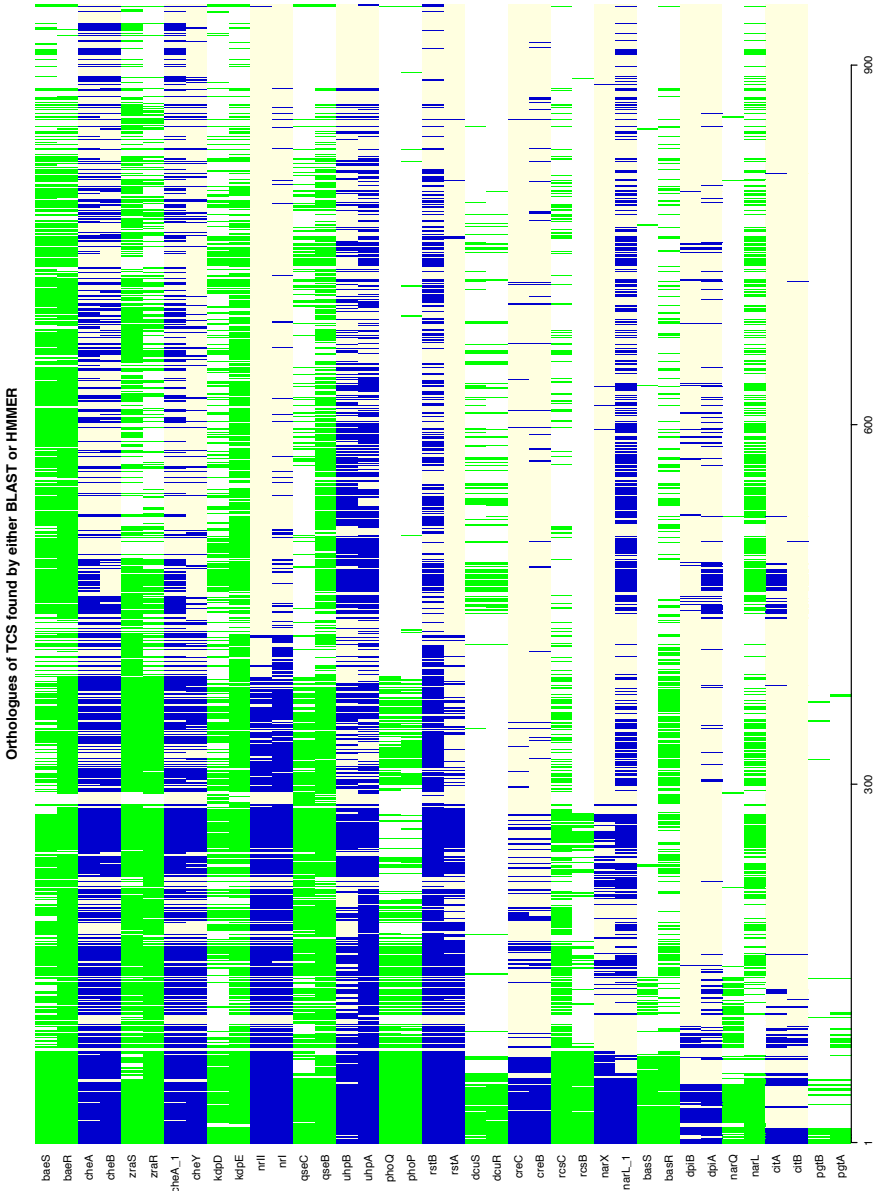


FIGURE 2.5: Orthologues of *Salmonella* TCSs. The X axis lists the 19 pairs of TCS orthologues ordered by the distance of the group to *Salmonella typhi*, from left to right, from closely to distantly related groups (as measured by the 23s rRNA). On the left side of Y axis are the names of the TCS proteins, while on the right side listed the total numbers of orthologues discovered. Green and blue stripes/areas indicate orthologues discovered by `BLAST` or HMM in that certain species, whereas light yellow and white stripes mean that no orthologue in the corresponding species could be identified. For each pair of TCS proteins we indicate the presence of the HK above the RR.

One of the hallmarks of gene transfer is different GC content between the transferred gene and the remaining genome. And our test showed no deviation between these two quantities. At the same time we found that the GC content of the candidate genes was significantly lower than that in *E. coli* (approximately 35% in those genes compared to

	Bacillus_cereus_AH820	Bacillus_anthracis_Ames	Bacillus_anthrcacis_str_Sterne
dpiB(%)	33.71	34.70	34.70
dpiA(%)	37.23	36.58	36.59
dcuS(%)	33.89	33.64	33.64
dcuR(%)	32.49	32.15	32.06
baeS(%)	35.05	34.90	34.90
baeR(%)	37.72	37.24	37.24
ypdA(%)	36.50	36.55	36.55
ypdB(%)	33.47	33.33	33.33
Average(%)	36.07±3.05	36.18±2.97	36.13±2.99

TABLE 2.3: Table of the GC content of TCSs coding genes in *Bacillus cereus AH820*, *Bacillus anthracis Ames* and *Bacillus anthracis str Sterne* and their average GC content.

average of 51% in *E. coli* K-12). This further excluded the possibility that these genes were transferred from a recent ancestor of *E. coli*. Thus, it was concluded that recent horizontal gene transfer is not responsible for the dpiB/A, dcuR/S, baeS/R and ypdA/B "islands" in figure 2.3.

Our GC content test result is shown in table 2.3. The average GC content of the three organisms are $36.07 \pm 3.05\%$, $36.18 \pm 2.97\%$ and $16.13 \pm 2.99\%$. The average GC content of *E. coli* K-12 is $51.90 \pm 4.41\%$. If the TCS "islands" were coded by genes transferred from the organisms close to *E. coli* K-12 through plasmid injection or other means we would expect that the nucleotide composition of the genes coding for TCSs dpiB/A, dcuS/R, baeS/R, ypdA/B should be more similar to the nucleotide composition of the genes in *E. coli* K-12 instead of the genes from there original organisms. However, as we can see in table 2.3, GC content of the above mentioned TCSs coding genes are all in the range of $35 \pm 3\%$. As the GC content of the TCS "islands" is significantly different (p value < 0.05) from that of the *E. coli* genes, while none of them is significantly different from the GC content of the genes in their original organisms. We concluded that the TCS "islands" in those organisms are not obtained by horizontal gene transfer.

2.4.3 Colocalization analysis of the TCSs across the organisms

In the previous studies it was found that in most of TCS pairs the two components located within the same operon in the genome, which might result in a positive feedback

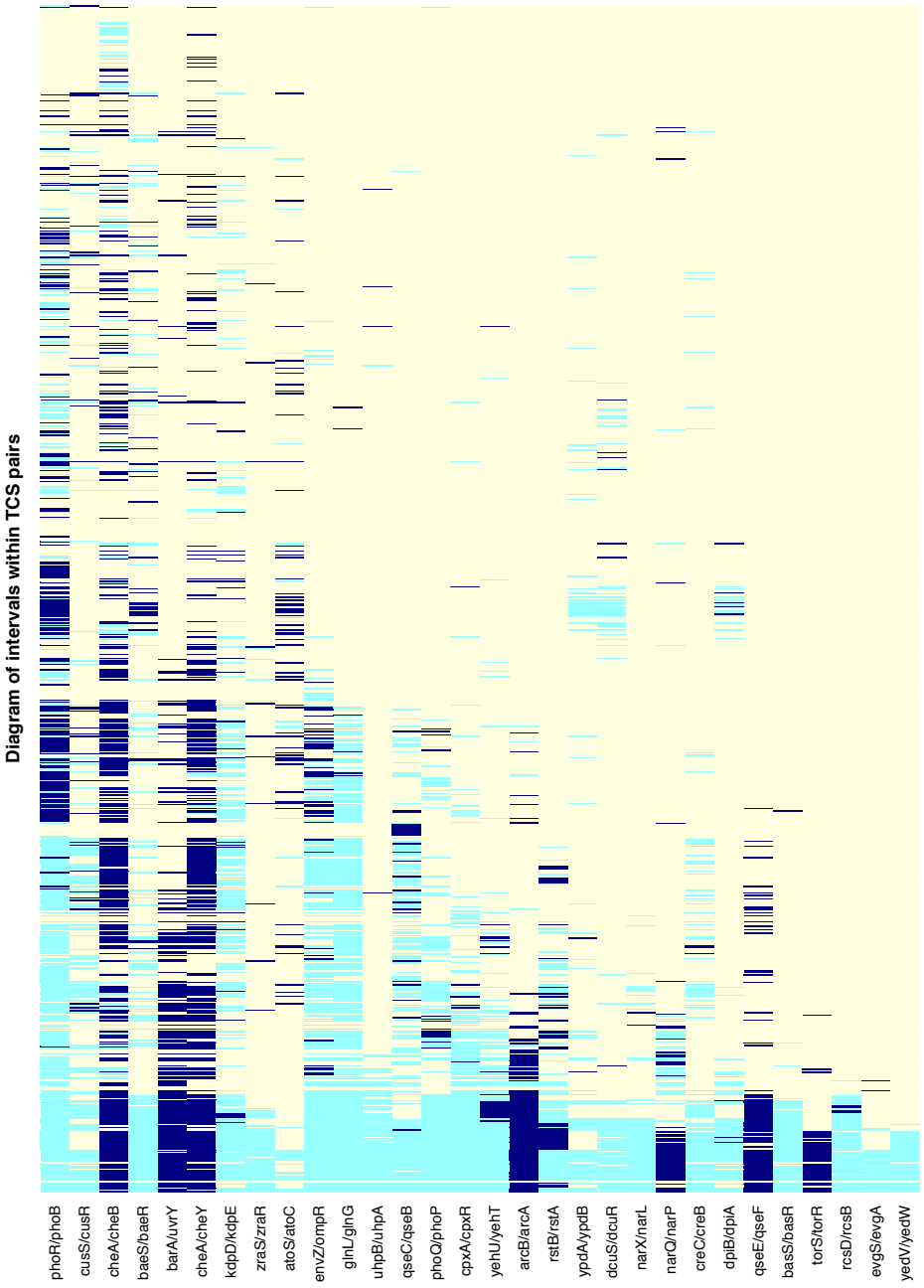


FIGURE 2.6: Diagram of intervals within TCS pairs. In this diagram, on the X axis listed the orthologues ordered by their organism's distance to *E. coli* K-12. From right to left of this figure listed the organisms with distance from close to distant to *E. coli* K-12. Listed on Y axis are the names of the TCS protein pairs. Light blue stripe indicates that TCS pair located within 300bp to each other in the genome. Navy blue stripe indicates that TCS pair located further than 300bp to each other, or was it presented in different genomes. Grey colour means that either or both of the components of the TCS pair were absent in that organism.

loop while the signal pathway is activated [113, 172]. However, some TCSs, *e.g.* ArcA/B in *E. coli*, are located quite far from each other on the chromosome. In the ArcA/B system, ArcA, the RR, is located in the arcA operon, while ArcB, the HK, is located in the arcB operon more than 1,000,000bp away from its cognate partner. To investigate how colocalized and separate TCS pairs occur across all the species, we performed the colocalization analysis for all the TCSs pairs we had discovered from our previous analyses. In the cases while the distance of the two TCS components is below 300bp on the chromosome, we defined them as colocalizing genes.

Figure 2.6 is a diagram of TCS gene intervals. It shows that most (23 out of 30 in *E. coli* K-12) of the TCSs are colocalizing on the genome. One thing worth noting is that some TCSs, like phoR/B and baeS/R, colocalize in some organisms but are widely dispersed in others. This difference in localization may suggest some gene recombination taking place in the ancestor when the break point happens.

As shown in figure 2.6, most of the TCS pairs colocalize in the genome. If a pair of TCS genes colocalize in one species, it is most likely that they also colocalize in all the other species where orthologues of this pair exist. The opposite cases also exist, *e.g.* the numbers of colocalized and separate phoR/B orthologues are similar to each other, neither type dominating. This could be explained by multiple copies of phoR/B orthologues in the bacteria. Suppose one species has one copy of a phoB orthologue, and two copies of phoR; while one of the copies (copy A) is in the same operon with the phoB orthologue, the other one (copy B) locates in a different operon which located far away. If the two orthologues have slight difference in their sequences. We would have similar probability of finding each copy as the orthologue of phoR in this species. However, this hypothesis still needs further experiments to decide whether it is acceptable.

In total 23 out of 30 TCS pairs colocalize on the chromosome in *E. coli* K-12. For the remaining 7 TCS pairs which are not colocalized (in *E. coli* K-12), two of them are chemotaxis related, which are involved in the chemotaxis system involving more than three TCS pairs, which are therefore not typical two-component systems; three out of the other 5 pairs contain hybrid histidine kinases. As there are only 4 hybrid histidine kinase among the 30 TCS pairs, we can conclude that a very high proportion of hybrid HK-containing TCSs are not colocalized, which suggests that TCSs with hybrid histidine kinases might evolve through a different route than the simple TCSs. One possible

mechanism for the evolution of hybrid HKs is that it might be a consequence of gene fusion of normal TCSs [34]. As the gene fusion could cause the gene to translocate to a different position, this hypothesis would support the fact that non-orthodox TCSs have higher chance of not being colocalized.

On the other hand, localizing in different operons could have some functional advantages. The particular gene distribution may give non-orthodox TCS the advantage of more flexible gene regulation. As HK and RR are not in the same operon, they could be activated in different dynamic patterns, which may help them to be competent in more complicated scenarios.

	Habitat	Gram Stain	Oxygen Requirement	Temperature range
arcA/arcB	√	√ √	√ √	
atoC/atoS	√	√ √	√ √	√
baeR/baeS		√ √	√ √	√ √
basR/basS		√	√	√
cpxR/cpxA		√ √	√ √	√
creB/creC		√ √	√	√
cusR/cusS		√ √	√ √	√ √
dcuR/dcuS		√ √		
dpiA/dpiB		√	√ √	
evgA/evgS	√	√	√	√
glnG/glnL		√ √	√ √	
kdpE/kdpD	√	√ √	√ √	√ √
narL/narX	√	√ √	√ √	√
qseB/qseC		√ √	√ √	
qseF/qseE	√	√	√	
rstA/rstB		√ √	√ √	
torR/torS		√ √	√	
uhpA/uhpB		√ √	√ √	√ √
yedW/yedV		√	√	√
yehT/yehU		√ √	√ √	
ypdB/ypdA		√ √	√ √	
zraR/zraS	√	√ √	√ √	√ √
ompR/envZ	√ √	√ √	√ √	
uvrY/barA	√	√ √	√ √	√
narP/narQ		√ √	√	√
phoP/phoQ		√ √	√ √	
phoB/phoR	√	√ √	√ √	
cheB/cheA	√	√ √	√ √	
cheY/fimZ		√ √	√ √	√
rcsB/rcsC	√	√ √	√ √	√
rcsD/rssB		√	√	

TABLE 2.4: Correlation test of TCS pairs. In the table listed the correlation between lifestyles and TCSs. They have at most two ticks in a cell which indicated positive for correlation between lifestyle and either one or both two components. P value lesser than 0.05 were designated to be significant.False Discovery Rate (FDR) was controlled at the level of 5%.

2.4.4 Co-evolution test using `BayesTraits`

We then tested for correlated evolution of the two components in each TCS studied here. The `BayesTraits` software [123] can be applied to the analysis of traits that adopt a finite number of discrete states, or to the analysis of continuously varying traits. Hypotheses about models of evolution, about ancestral states and about correlations among pairs of traits across phylogenies can be tested.

Even after correcting stringently for multiple testing we found that the components of all TCS (in *E. coli* K-12) appear to evolve in a correlated, dependent manner, as may perhaps be expected naively. However, for some TCSs which have very unbalanced HK/RR orthologues ratios the evidence for dependent evolution is reduced. The fact that all TCSs appear to show evidence for correlated evolutionary history of their respective component parts is a piece of further evidence in favour of the co-evolution model (result for *Salmonella typhi* is shown in table 2.6, last column).

Method	Life style	PspA	PspB	PspC	PspD	PspE	PspF	PspG	ArcA	ArcB
Fisher's exact test	Oxygen_Req	1.20E-03	1.74E-08	7.85E-10	7.67E-04	0.053	1.00E-05	1.96E-07	2.55E-10	3.06E-05
	Gram_Stain	1.77E-18	1.08E-13	0.189	9.96E-04	7.66E-03	1.46E-15	5.02E-06	2.07E-11	3.51E-10
	Habitat	2.82E-06	3.19E-05	8.70E-04	1.04E-11	0.010	2.58E-06	6.13E-07	1.20E-04	1.52E-07
	Temp._range	0.678	0.131	0.186	0.151	0.812	0.186	0.151	0.131	0.233
Pagel's Bayes-Traits	Oxygen_req	0.850	0.940	0.980	1	0.940	0.970	1	0.980	1
	Gram_Stain	0.010	0.400	0.178	0.915	0.205	0.237	0.373	0.282	0.019
	Habitat	1	1	1	1	1	1	1	1	1
	Temp._range	0.9962	0.819	0.921	0.9997	0.9968	0.991	0.983	0.994	0.996

TABLE 2.5: Correlation test of TCS existence and lifestyles. The experiment was done using Fisher's exact test and `BayesTraits`.

2.4.5 Lifestyle test using `BayesTraits`

Fisher's exact test is a statistical method to test for correlation between different traits. In order to assess potential functional factors affecting patterns of presence/absence of HKs and RRs in TCSs we also tested for correlation with the organisms' lifestyle (downloaded from *http://www.ncbi.nlm.nih.gov*) using `BayesTraits`. The phage shock protein (Psp) stress response system is an important part of the stress response machinery in many bacteria, including *Escherichia coli* K12. Previous study showed that it has close connection with the two component system ArcA/B [63]. In this chapter

we used bacterial lifestyles and information of this system and ArcA/B to compare the correlation test methods Fisher's exact test and `BayesTraits`. As shown in table 2.5, we first test for correlation between the existence of Psp system proteins and lifestyles using Fisher's exact test. It turns out that except for temperature range, all the Psp system genes are correlated with all the lifestyles (p value< 0.05). But this correlation might be spurious, as the lifestyles are related to certain phylogenetic clusters, while gene presence is also related to some phylogenetic clusters. Thus Psp genes and lifestyles may not be correlated with each other, but all are simply correlated with the phylogeny. And when we use BayesTraits, which applies the phylogenetic tree as a reference parameter, almost all the gene/lifestyle pairs lost the correlation. Thus, we conclude that Fisher's exact test might not be an ideal method for testing the correlation between two phylogeny related traits because of being too lenient.

In table 2.4 we show the detailed results which paint a very nuanced picture: almost all of the TCSs show significant correlation (even after applying an FDR correction) to the lifestyles *oxygen requirement* and *gram stain*, as may well be expected given the importance of these factors for the metabolic and physiological processes of bacteria. But in several cases, only one of the TCS partners shows such correlation with *habitat* and *temperature range*, which may suggest presence of cross-talk and could provide tentative evidence in favour of the recruitment model (the result for *Salmonella typhi* is shown in table 2.6).

2.4.6 TCS protein sequences similarity comparison

Histidine and aspartate phosphorylation domains show high levels of similarity and this may act as a further confounding factor. To address this, in a final step of our analysis (also aimed at detecting any remaining orphan HKs and RRs in the species considered here) we extracted the protein sequences of all the TCSs orthologues that we have identified and tested for their similarity.

As a starting point, we first examined the sequence alignment of HKs and RRs in *E. coli* K-12. To make the sequences more comparable, we aligned them by their domains. HKs may have 3 phosphate group binding domains, and all have HisKA domains. And if they are hybrid HKs, they may also have HPT domains and REC domains. In the meantime, RRs only have one phosphate group binding domain, a REC domain. As shown in figure

	Habitat	Gram_stain	Oxy_req	Temp_range	Co_evo
citA/citB			√		√
pgtB/pgtA			√		√
rcsC/rcsB	√	√ √	√ √	√	√
baeS/baeR	√	√ √	√ √	√	√
cheA/cheB	√	√ √	√ √	√	√
cheA/cheY		√ √	√ √	√	√
rstB/rstA	√	√ √	√ √		√
narX/narL		√ √	√ √	√	√
phoQ/phoP		√ √	√ √		√
kdpD/kdpE		√ √	√ √	√	√
dpiB/dpiA		√	√	√	√
qseC/qseB		√ √	√ √	√ √	√
zraS/zraR	√ √	√ √	√ √		√
nrII/nrI	√	√ √	√ √		√
uhpB/uhpA	√	√ √	√ √	√	√
basS/basR		√ √	√ √	√	√
dcuS/dcuR		√ √	√		√
creC/creB		√ √	√		√

TABLE 2.6: In the table we indicate evidence for coevolution and for correlation between lifestyles and TCS proteins in *Salmonella typhi*. Each tick indicates correlation of existence of one TCS pair, or correlation between lifestyle and either one of the two components. P values smaller than 0.05 were designated to be significant after controlling the False Discovery Rate (FDR) at the 5% level.

2.7, HisKA domains have a very low level of conservation. HPT domains and REC domains have higher levels of conservation. REC domains of the RRs (figure 2.8) have very high level of conservation, and share similar patterns with REC domains in hybrid HKs. As HisKA is present in all the HKs, and HisKA has much lower conservation level than REC in RR, we can conclude that HKs in *E. coli* K-12 have much lower similarity to each other than RRs.

To investigate whether the above observation holds for other species, we studied the similarity level among all the HKs or RRs in each species. We separated the TCSs into two groups, HKs and RRs, aligned the sequences of each group and calculated the similarity levels inside each group. As shown in figure 2.9 RRs exhibit higher levels of similarity than HKs ($p = 7.6 \times 10^{-4}$, Kolmogorov-Smirnov test), which is in accordance with the case in *E. coli* K-12.

All the species we studied have different distances to *E. coli* K-12. The histograms treat them, however, identically. It would be interesting to see if there is anything different in sequence similarity between the species close to *E. coli* and the species far from it.

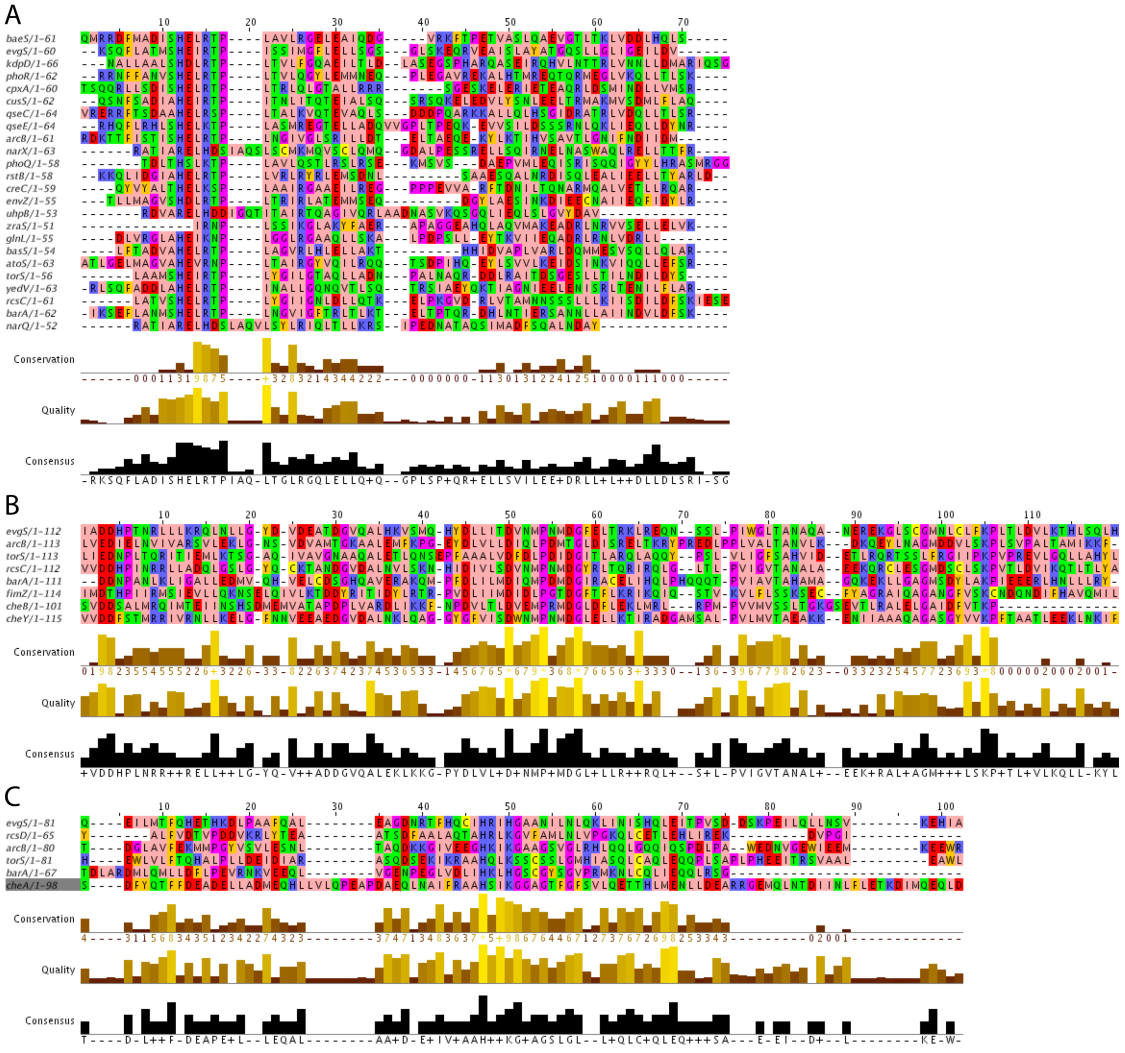


FIGURE 2.7: Sequence alignment of the domains in histidine kinases in *E. coli* K-12. A, alignment of HisKA domains in all the HKs. B, alignment of the REC domains in hybrid HKs. C, alignment of the HPT domains in hybrid HKs.

As shown in figure 2.10, the similarity of both HKs and RRs seems to be stable to the variation of their distance to *E. coli*. However, in the species far from *E. coli*, there are more species with less than four HK or RR sequences, and these species frequently have higher similarity levels. This is natural: fewer identified orthologues objects will lead to higher similarity rate. The similarity ratio of HKs and RRs shows a similar pattern, the ratio is always lower than 1 for the species with more than 4 HK and RR orthologues. Thus, it can be concluded that HKs are more diverse in the majority of bacteria.

This phenomenon might be a consequence of their different roles in the cell, or the different constraints under which they operate. HKs, for example, are responsible for

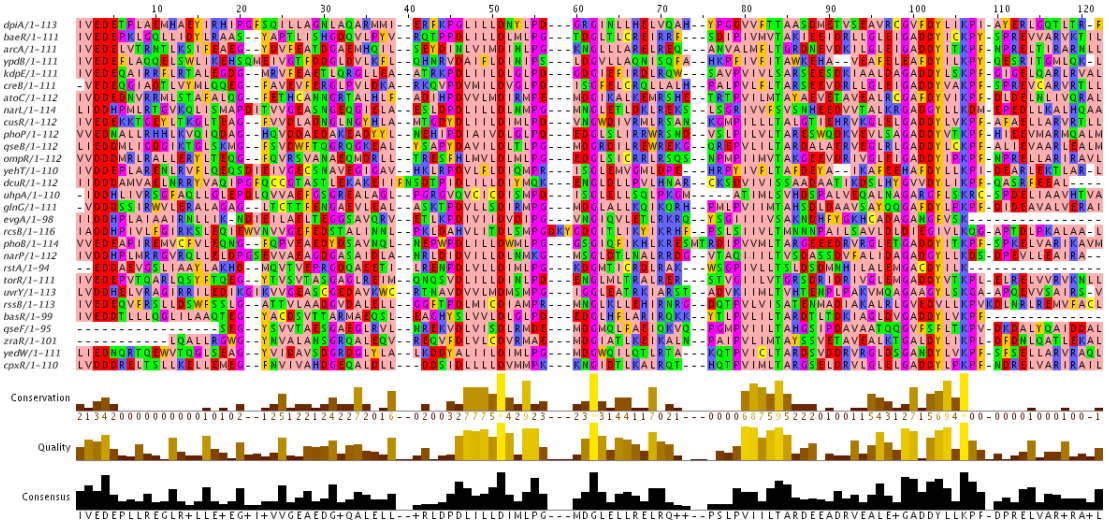


FIGURE 2.8: Sequence alignment of the receiver domains in respond regulators in *E. coli* K-12.

responding to a broad range of often complicated extra-cellular signals, which may have caused them to diverge more, even though they must generally embed into a cell membrane which constrains the evolution of parts of their sequence. RRs on the other hand must be able to interact with their corresponding HK, receive phosphate groups from them and then bind to DNA. The latter is known to already impose some constrains on the sequence [133], and so a larger proportion of the RR sequence may be under the influence of selection than is the case for their corresponding HKs. Nevertheless the high degree of similarity of proteins within the same species is remarkable, for it also potentially opens up more scope for cross-talk and co-opting parts from other TCSs: if the different HKs and RRs within a given organism show such similarity at the sequence level, then some overlap in functional similarity or the ability to interact with each others' respective counterparts is plausible. Needless to say this will require further experimental testing.

2.4.7 Cross-talk partner inference

To investigate which pairs of non-cognate TCSs could be possibly cross-talking, we performed a hypergeometric test on all the pairs of non-cognate partners. We expected to see all the TCSs pairs (near $30 \times 29 \times 2 \approx 1800$) to fail such a test expect for a few with p value lower than 0.05, which will be our candidates for cross-talking partners.

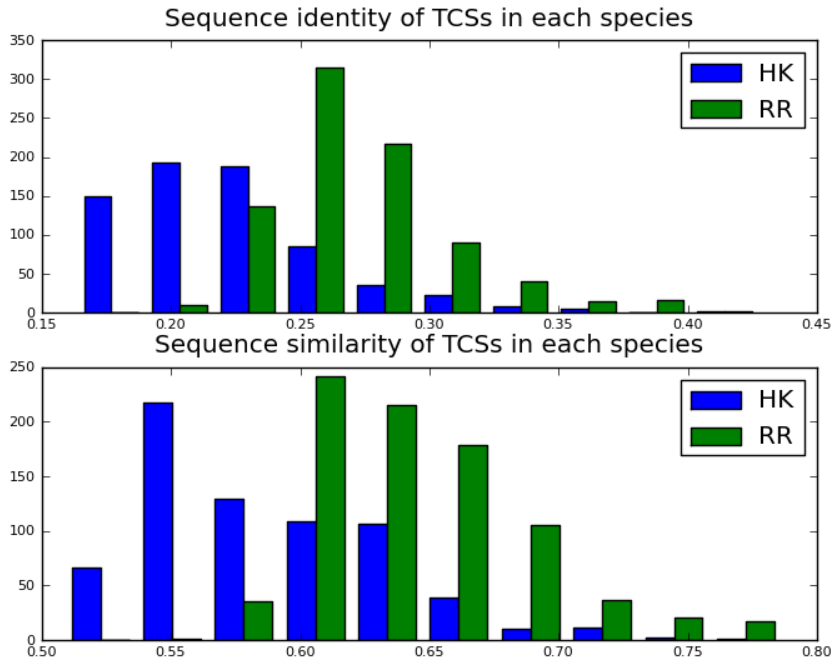


FIGURE 2.9: Sequence similarity and identity of TCS in each species. A comparison of histidine kinases and response regulators inside each bacteria species. Numbers of each similarity/identity range were counted and displayed in either blue or green. X axis lists similarity range. Y axis lists the counts of each category.

Surprisingly, we found that by this criterion more than 300 candidates could be possible cross-talking pairs. To narrow down the results, we chose 78 of these pairs with p values less than 1×10^{-10}, and evaluated the pairs' coexistence map aligned with the phylogenetic tree. If the pairs are cross-talking, then we may expect them to coexist in many species. And their distribution should potentially form clusters. We also chose 5 randomly picked pairs as negative controls. As shown in figure 2.11, candidate cross-talking partners are indeed forming clusters. However, some of the negative controls also form clusters in the figure. In table 2.7, we listed the number of candidate cross-talking partners each TCS has based on the hypergeometric result with p value less than 0.05. When we compare the number of each TCS with their cognate partners, we found that they seem to be anti-correlated, *e.g.* atoS has 21 cross-talking partners while atoC has only 7; qseE has only 2 cross-talking partners, while qseF has 26. The Pearson's correlation test supported our hypothesis. The correlation coefficient of the numbers is -0.527 suggesting that the HKs' and RRs' cross-talking numbers are anti-correlated with p value of 0.0014. This unexpected relationship suggests that TCSs may have some

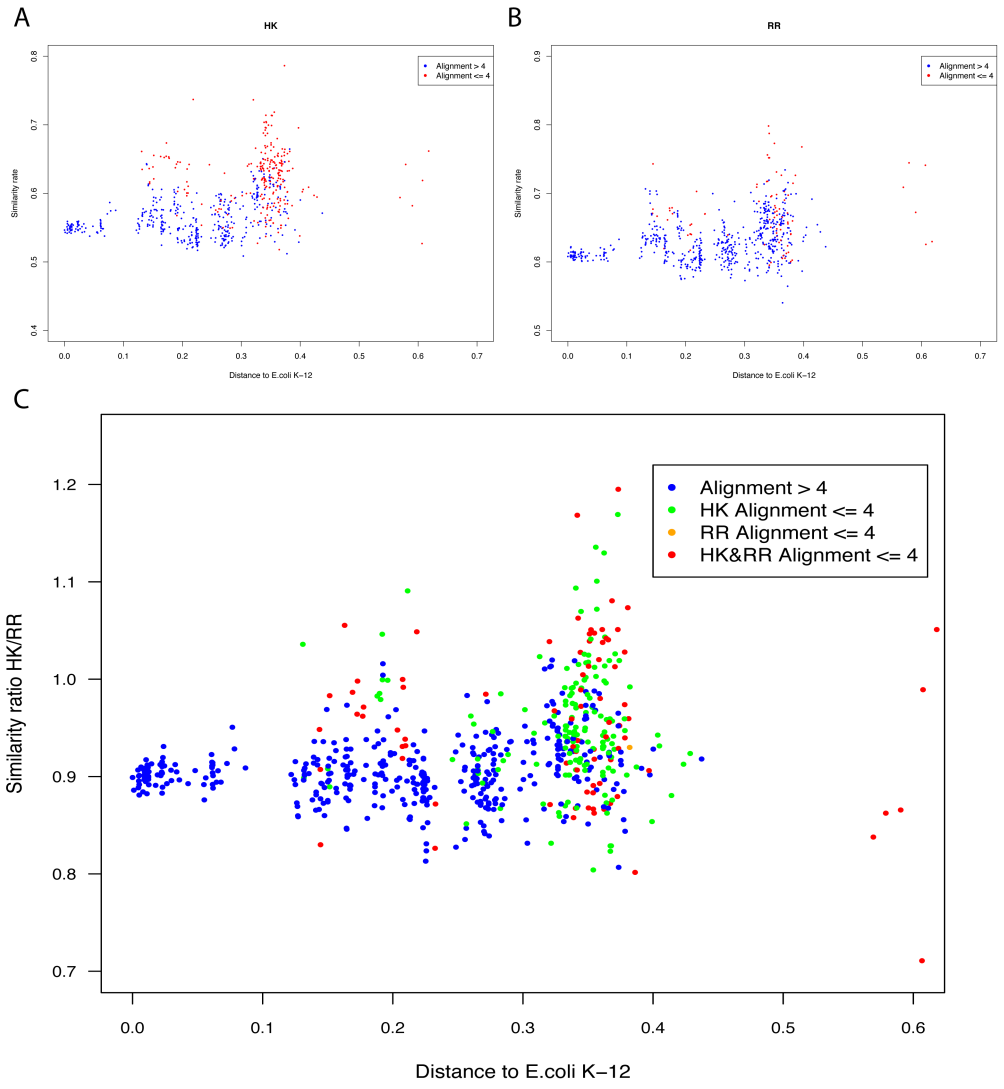


FIGURE 2.10: Sequence similarity and similarity ratio of TCSs in each species. A, similarity rate of HKs in each species arranged by the distance from that species to *E. coli* K-12. Species with more than 4 HK sequences are shown in blue colour. Red colour indicates species with no more than 4 HK sequences. B, similarity rate of RRs in each species arranged by the distance from that species to *E. coli* K-12. Species with more than 4 RR sequences are shown in blue colour. Red colour indicates species with no more than 4 RR sequences. C, similarity ratio of HK/RR in each species arranged by the distance from that species to *E. coli* K-12. Species with more than 4 RR sequences of both HKs and RRs are shown in blue colour. Red colour indicates species with no more than neither 4 RR nor HK sequences. Green dots are the species with no more than 4 HKs but more than 4 RRs, while orange dots shows the opposite situation

complementary mechanism involving cross-talking.

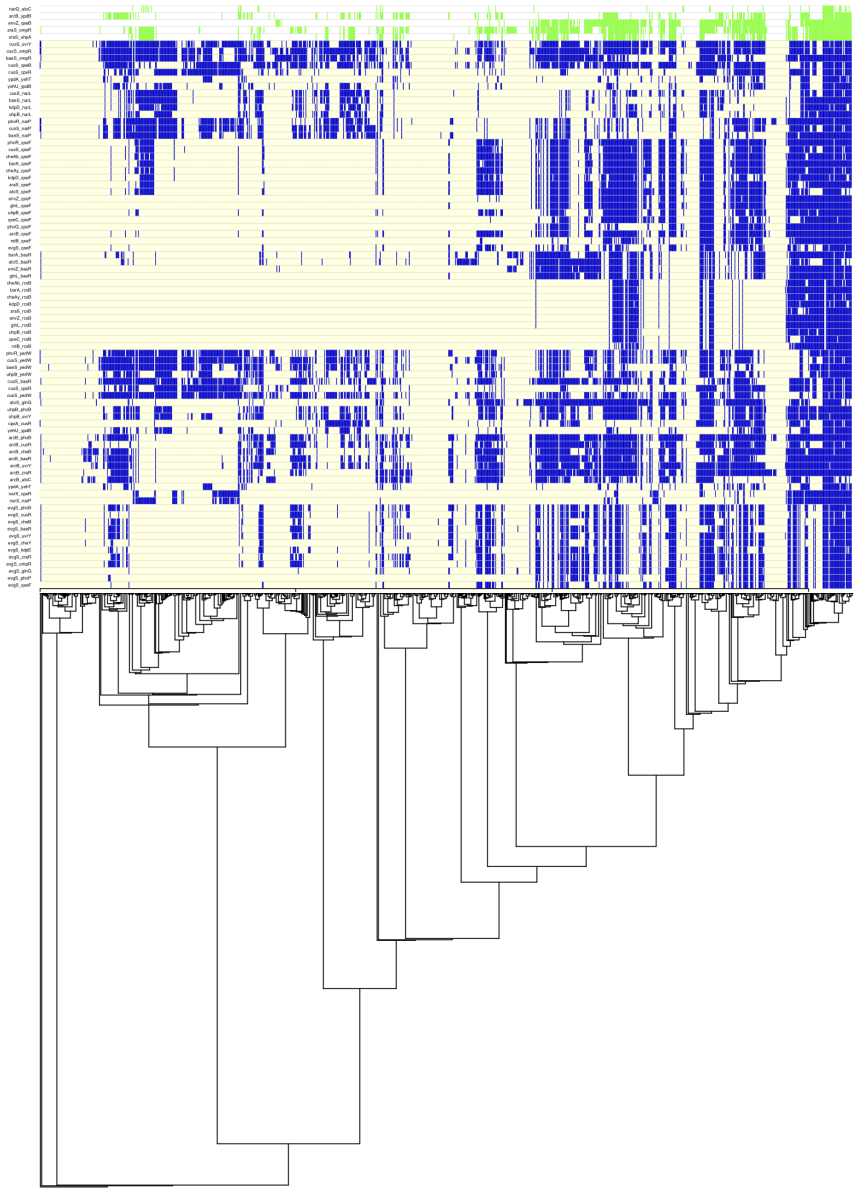


FIGURE 2.11: Co-existing pairs of the candidate TCSs aligned with the phylogenetic tree. Blue stripes are species with coexisting candidate cross-talking pairs. Green stripes are species with coexisting randomly picked pairs. *E. coli* and its neighbours are are drawn along the right edge of the figure.

2.5 Conclusions

In this chapter, we have characterised the patterns of evolutionary conservation for all TCSs present in *E. coli*; we identified orthologues in some 950 bacterial species; tested for correlated evolution and the genetic neighbourhood of the components of

TCS Pairs	No. of cross-talk for HK	No. of cross-talk for RR
torS/torR	24	10
baeS/baeR	16	13
phoR/phoB	16	13
atoS/atoC	21	7
qseE/qseF	2	26
uhpB/uhpA	19	8
cusS/cusR	18	8
yehU/yehT	12	13
evgS/evgA	20	4
arcB/arcA	21	2
barA/uvrY	8	15
basS/basR	8	15
narX/narL	9	12
rcsD/rcsB	0	21
envZ/ompR	5	15
kdpD/kdpE	12	8
creC/creB	3	16
cheAb/cheB	9	9
ypdA/ypdB	9	9
cheAy/cheY	10	7
dcuS/dcuR	8	9
phoQ/phoP	11	6
rstB/rstA	13	4
glnL/glnG	7	9
qseC/qseB	5	11
cpxA/cpxR	9	6
yedV/yedW	1	14
narQ/narP	5	9
zraS/zraR	6	8
dpiB/dpiA	4	4

TABLE 2.7: Table of the numbers of non-cognate components that each TCS protein can cross-talk with. The number for each HK/RR is shown together with their cognate partners. Then the correlation coefficient between HKs' and their cognate RRs' cross-talking protein number was calculated using Pearson's correlation test using `R`. The correlation coefficient of these two groups is -0.527. The two groups are anti correlated with p value of 0.0014.

each TCS; tested for evidence between presence or absence of TCSs and their respective components and environmental as well as life-style factors; and finally assessed levels of sequence similarity of TCSs in each of the 950 species. Taken together, as we will show below, the results from these lines of investigation are broadly in line with what would be expected under a mixture of both evolutionary models. The discussion of these findings and further implications for the evolution of simple prokaryotic signalling systems will conclude this chapter.

As we have stated in the background section, most researches on genome scale survey of TCSs are based on the TCSs found by domain searching methods. The idea of

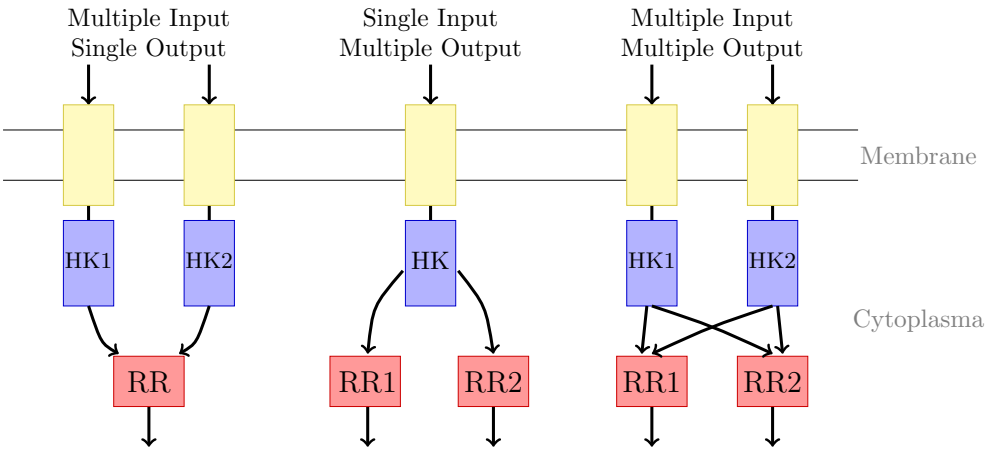


FIGURE 2.12: Diagram of cross-talk between a TCS and orphans or other TCSs can give rise to more flexible signal transduction networks.

this approach is to screen the proteins of interest for the presence of a TCS-defining domain. The domain sequence profiles can be obtained from the database such as Pfam [42]. HisKA, Hpt and REC domains are usually used to identify TCS proteins. Other domains like HATPase are also considered in the search. Comparing to the domain searching methods, the orthologue searching method that we have applied is able to provide a view of the distribution of the paired and orphan TCSs across all the organisms. However, the search result with our method can not be considered as a whole set of TCS proteins in each species. Suppose we have some organism expressing more TCSs than *E.coli* K12 or expressing TCSs which are not orthologues of any *E.coli* K12 TCSs, our method will definitely miss these "additional" TCSs. However, we have to use this orthologue searching method because most of our studies in this chapter is on the partnership of the TCSs.

We had initially set out to search for evidence in favour of either the recruitment or the co-evolution model of TCS evolution. Our evidence suggests that a more nuanced description is called for: there is tentative evidence in favour of either model but in summary both recruitment and co-evolutionary processes appear to have shaped the evolutionary history of TCSs.

- We found statistical evidence against independent evolution of the two components making up the 30 systems extant in *E. coli*.

- TCSs appear to be colocalized in general.
- Many, but not all TCSs, are inherited in a highly coordinated fashion, where either both or none are present in a species.
- We found evidence that is in line with evidence of ancient horizontal gene transfer of complete TCSs.
- We found significant evolutionary association between each component of the TCS pairs.
- For a number of TCSs we observed asymmetries in the relative presence of orphan HKs and RRs, *e.g.* some TCSs appear to have preferentially lost RRs (or HKs).
- We also found high levels of sequence similarity, in particular for RRs, which may suggest the facility of cross-talk and the possibility different, more complex signalling interactions involving components from different TCSs.
- For the non-orthodox TCSs we found a noticeable decrease in colocalization compared to the orthodox TCSs.

We have performed the tests not only in *E.coli* K-12, but also in *Salmonella typhi*. Taken together we propose that the two previously proposed models apply to different extents to different TCS cases, but are neither mutually exclusive nor incompatible. Shared functionality (trans-membrane, phospho-transfer, phospho-reception and DNA binding) may even have exerted comparable levels of selection on the complements of RRs and HKs within the same species.

We conclude on a cautionary note: cross-talk as shown in figure 2.12 may occur in different forms leading to multiple architectures, which in turn may affect signal processing and an organism's response to its environment in ways that are more subtle than the simple TCS picture might suggest. Such added flexibility, will of course, further confound our attempts at explaining TCS evolution in terms of simple models. Thus our understanding of the evolution of more complex signal transduction systems is unlikely to be furthered by simplified models. In population genetics we have a framework in which to study the interplay between the different forces acting on single genes (or small sets of genes): mutation, recombination, selection and drift. These forces will also act on the genes considered here (or in other systems) but be additionally modulated by

interactions among genes or their protein products as well as functional considerations. In particular we may expect many interactions — and network architectures such as those depicted in figure 2.12 — to be specific to particular species or species clusters and generally not be detectable by orthology arguments such as the ones used here. Rather direct experimental observation will be required.

Chapter 3

Study on the steady states of Two-component systems

3.1 Background

Two component systems are the most prevalent signal transduction systems in bacteria. TCSs are used to sense environmental stimuli and to transduce the information into the cells. This will lead to a range of downstream processes and responses, primarily through transcriptional regulation of target genes.

Sensing and transmission of a signal into the cytoplasm is performed by a histidine kinase (HK, sensor) which typically resides in the cellular membrane of bacteria; the HK then transmits the information by phosphorylating a cognate response regulator (RR, receptor). Histidine kinases have a transmembrane region near the N-terminus of the protein, which anchors HK in the membrane. The outer membrane region is usually relatively short, and is believed to be responsible for sensing the extracellular stimulus. The detailed molecular mechanisms for sensing signals are mostly unknown. The vast majority of RRs are DNA binding proteins that control the expression of genes required in the response. The targets of RRs include genes responding to changes in temperature [30], oxygen level [110], osmolarity [26], pH [5], *etc.*. In most TCSs, RRs can also bind upstream of its own operon to regulate its own gene transcription. In some cases an RR's expression is downregulated by itself, like in the case of the ArcB/A system [18]. In some cases it is positive feedback where the RR's expression is upregulated by itself, like

for the MprB/A system [172]. In this chapter, we are specifically studying the positive feedback loop transcription model, because positive feedback loops have been reported to be one of the requirements for bistability in dynamic systems [186].

In contrast to a simple two-component system, where the histidine kinase has only one site to be phosphorylated, a non-orthodox TCS has a hybrid histidine kinase instead. The structure and activation of orthodox and non-orthodox TCSs have been discussed in chapter 1.

In bacteria, it is easy to see the advantage of bistability. it allows a few cells to enter a state that would be better adapted to one circumstance or another should that circumstance arise. This type of bimodal distribution may allow the cells to specialize in advance of changed circumstances. As a sensing and signal transduction system, some types of TCSs might also acquire the property of bistability to respond distinctly to varying levels of extracellular signals. Tiwari *et al.* have discussed different bistability phenomena in bacteria, and two-component systems have also been discussed in their review [173]. However, it has not been pointed out what elements in two-component system might affect the potential bistability of the system. It has been reported that auto-activation can lead to bistability [80]. Some research has also shown that a positive feedback is a necessary, but not sufficient, condition for a chemical system to be bistable [35, 36]. Thus we have only modelled auto-activated TCSs with positive transcriptional feedback loops to test for bistability. Something worth mentioning at this stage is that bistability is not always beneficial for the system. If a signalling system is meant to provide gradient responding to changing input signals, bistability which coexist with hysteresis should not be a good choice for the system. Thus, in this chapter, we are not claiming that a bistable system is better than a monostable system. Instead, we are trying to identify systematically the conditions that may lead to bistability.

In a typical two-component system model, there are reactions for dimerization, binding, transcription, translation, degradation, autophosphorylation, phosphotransfer and dephosphorylation, as shown in figure 3.1. If they are all modelled for calculation of bistability by ordinary differential equations (ODE), the system will have 13 variables and 28 parameters. Steady state of these equations cannot be determined analytically. Even when we approach this problem numerically, the process is quite slow (data not shown). Thus we divide the system into two parts, and solve them individually. The first

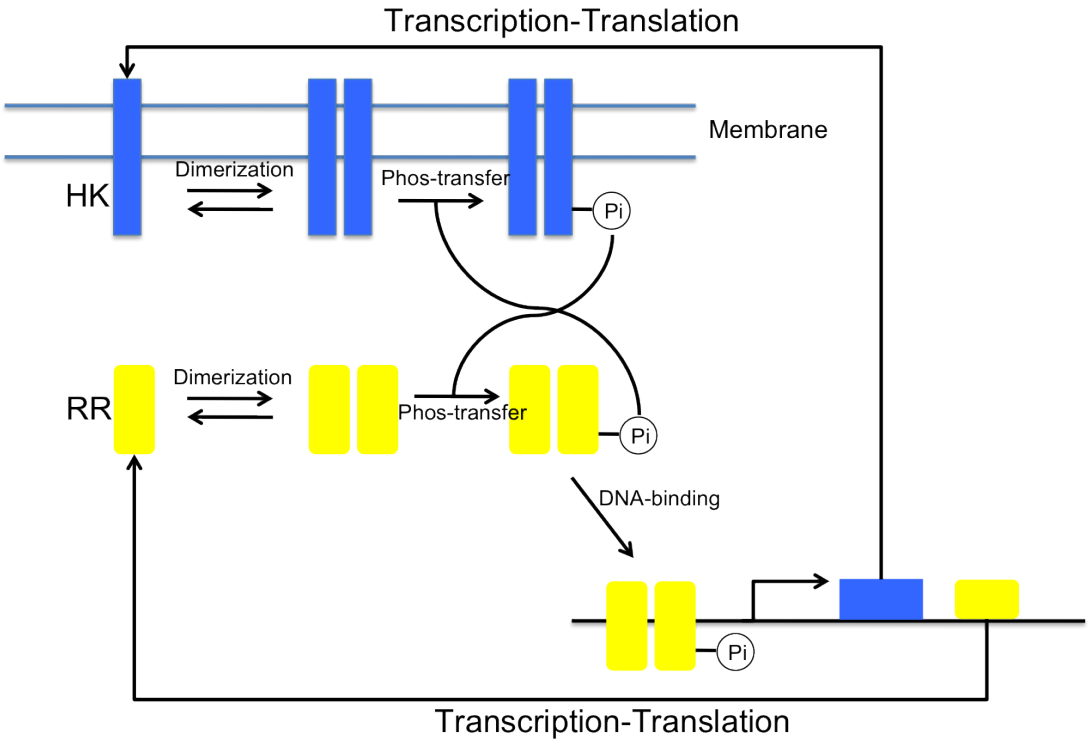


FIGURE 3.1: Schematic representation of slow and fast reactions involved in a typical two-component system. HK stands for histidine kinase. RR stands for response regulator. Upon stimulation, the HK changes its conformation which is then autophosphorylated. The phosphate group will then transfer to the unphosphorylated RR. The phosphorylated RR will then bind to the promoter of its own transcription operon and initiate transcription and translation of itself and its cognate partner. The freshly translated HK and RR protein monomer will bind to form a homodimer to be active required for the reaction [161]. The degradation reactions are not included in this diagram.

part involves transcription, translation and degradation reactions. These reactions are relatively slow. The second set of reactions includes all the other molecular processes. This type of reaction tends to be faster. When we are studying the slow reactions, we consider the fast reactions to be always at steady state. On the other hand, when we are studying the fast reactions, we always consider the slow reactions to be halted. Then we can combine these two parts of results together to get a full picture of the system. The method was also applied by Tiwari *et. al.* in their study [172].

Two-component system proteins have very high specificity in a one to one pattern with their cognate partner. However, it has been reported that they can also cross-talk with proteins from other TCS with lower affinity [154, 194]. For example, CheA can phosphorylate OmpR [64], and EnvZ can phosphorylate NtrC [64] and our analysis in the previous chapter has also pointed towards widespread ability of cross-talk in

the TCSs. The sequence motif determining specificity is believed to be located in the contact region of cognate HK/RR pairs [155]. Skerker *et al.* [155] showed that by replacing amino acid sequences in this region by the sequence from the same region from a second TCS can rewire its specificity to bind to the second TCS. Although the region that controls specificity has thus been found, the underlying mechanism still remains unknown. There are an increasing number of studies of how cross-talking might affect the dynamic features of the whole system compared to a single TCS [53, 112, 130]. We start from the simplest case, the interaction between two parallel TCSs. There are three types of cross-talking arrangements between two TCSs. In the first case, one HK can react with its cognate RR and another RR from a different TCS, which we call SIMO (single input multiple output) motif. In the second case, the RR can react with a different HK in addition to its cognate HK. We call this the MISO (multiple input single output). In the third case, the two TCSs' HK and RR can interact with each other's respective partners, which we call MIMO (multiple input multiple output). Because MIMO has the reactions of both SIMO and MISO motif, and can be transformed into either MISO or SIMO motif by simply removing certain interaction reactions. In this chapter, we have studied the steady state characteristics of cross-talking model with MIMO motif.

3.2 Methods

3.2.1 Modelling of slow reactions of two-component systems

We use chemical reactions to describe the TCSs that we have modelled. These models follow mass-action kinetics. The reactions we have used to describe the transcription/-translation/degradation reactions are:

$$(RR_2)p + \text{promoter} \underset{k_r}{\overset{k_f}{\rightleftharpoons}} (RR_2)p - \text{promoter}$$

$$(RR_2)p - \text{promoter} \xrightarrow{k_t} (RR_2)p - \text{promoter} + RR$$

$$\emptyset \xrightarrow{k_b} RR$$

$$RR \xrightarrow{k_{deg}} \emptyset$$

In our models, the active phosphorylated response regulator dimer binds to the upstream promoter of any target gene, including its own coding gene. When binding to the upstream region of $(RR_2)p$, the response regulator gene is transcribed at a higher rate. Without binding of the transcription factor, transcription only occurs at the basal rate k_b. Translation of mRNA into protein is not explicitly modelled. It has instead been integrated into the transcription reactions. The concentration of phosphorylated RR dimer is assumed to be proportional to the total RR dimer concentration for simplicity [157]. The invert dissociation constant of the dimerized phosphorylation form $(RR_2)p$ to RR was set to be K_d. The ordinary differential equation for monomer $[RR]$ is:

$$d[RR]/dt = \frac{Vmax[RR]^2}{[RR]^2 + K} - k_{deg}[RR] + k_b$$

In the above equation, $K = k_r/(k_f K_d), Vmax = k_t[\text{promoter}]_{total}$. The response regulator needs to form homodimers to be active. And only the phosphorylated dimer is able to bind to the target promoter sequence. These reactions are relatively much more rapid. Thus we have considered these reactions to be always at steady state. And we have assumed $[(RR_2)p]$ is proportional to $[RR]$ for simplicity. Slow reactions are the same for both orthodox TCSs and non-orthodox TCSs.

3.2.2 Modelling of rapid reactions of two-component systems

Post-translational reactions of two-component system, including phosphorylation, dephosphorylation, dimerization and binding, are significantly more rapid than the above mentioned slow reactions. We consider them to be always in steady state when we are studying the slow reactions. On the other hand, we consider the slow reactions to be approximately static when studying rapid reactions. In the dimerization reactions, HK and RR both form dimers, 2 × monomer ⇌ homodimer. As the dimerization reaction constant is independent of the phosphorylation state [132]. We have $(HK_2)_{total} = K_{d1}HK_{total}^2, (RR_2)_{total} = K_{d2}RR_{total}^2$. Because we consider all the transcription-translation reaction to be static when calculating steady states of rapid reactions, there will be no changes in the total concentrations of HK and RR. Thus the total concentrations of the HK dimer and the RR dimer are also assumed, and because only dimerized HK and RR are active, we will only focus on dimerized molecules. For simplicity, we will just denote them as HK and RR instead of HK_2, RR_2.

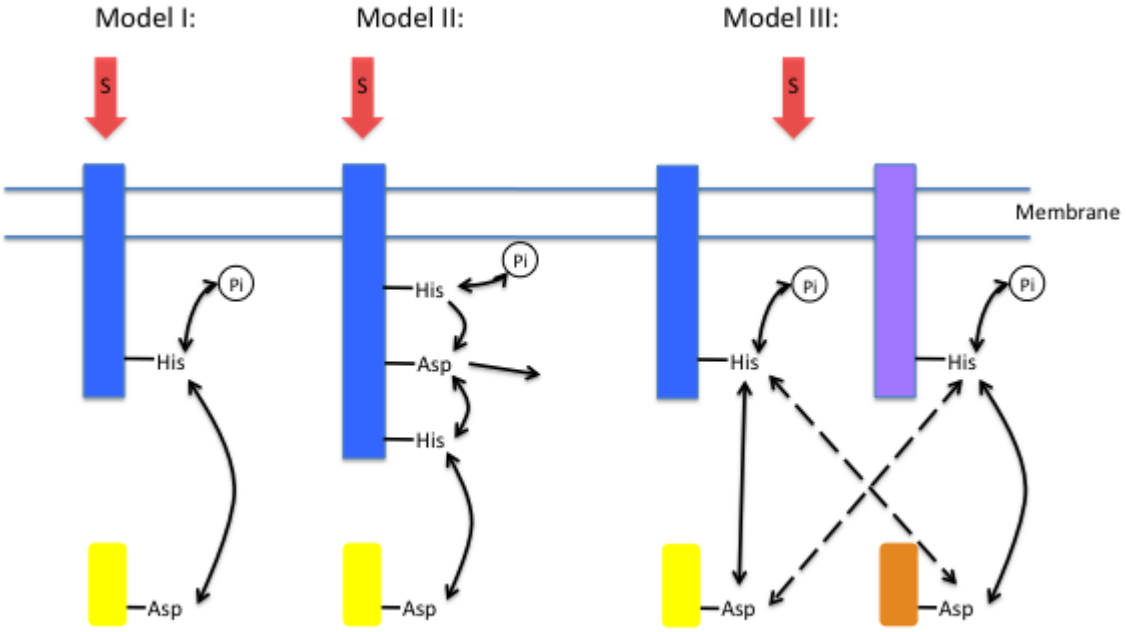


FIGURE 3.2: A diagram of post-translational reactions of the three models we have built. Red arrows with letter S stand for the extracellular stimulus. Blue and purple squares denotes HKs. Yellow and orange squares denotes RRs. Because no dimerization reaction is involved in these models. We didn't draw them as dimer for simplification reason. Black lines with arrow indicate the directions of phosphorylation or dephosphorylation. Dashed lines indicates cross-talking reactions between different TCS which have much lower reaction rates.

We have built three models of the fast reactions based on their different phosphorylation styles. The three models are the orthodox model with a simple histidine kinase, the non-orthodox model with a hybrid histidine kinase, and a general cross-talking model with two interacting orthodox TCSs. Diagrams of the three models of post-translational reactions are shown in figure 3.2 The features and chemical reactions for the models are listed below:

Model I, orthodox model. In this model, the HK can be autophosphorylated and dephosphorylated. Phosphorylated HK can bind to unphosphorylated RR, and the phosphate group is then transferred to RR. This reaction is reversible. Chemical reactions are

shown below:

$$HK \underset{k_2}{\overset{k_1}{\rightleftharpoons}} HKp$$
$$HKp + RR \underset{kd_1}{\overset{kb_1}{\rightleftharpoons}} HKpRR \xrightarrow{kt_1} HK + RRp$$
$$HK + RRp \underset{kd_2}{\overset{kb_2}{\rightleftharpoons}} HKRRp \xrightarrow{kt_2} HKp + RR$$

Model II, non-orthodox model. The hybrid HK participates in this model. HK1, 2 and 3 respectively denote the hybrid HK with site H1, D1, H2 phosphorylated. $HK0$ denotes the unphosphorylated hybrid HK. Model II differs from model I greatly in terms of the phosphorylation process, where the phosphate group is transferred from H1 to D1, then to H2 and at last to RR. During dephosphorylation, the phosphate group is transferred from RR to H2, then to D1 and is then lost from D1. Chemical reactions are shown below:

$$HK0 \xrightarrow{k_1} HK1$$
$$HK1 \xrightarrow{k_2} HK2$$
$$HK2 \xrightarrow{k_3} HK3$$
$$HK3 \xrightarrow{k_4} HK2$$
$$HK2 \xrightarrow{k_5} HK0$$
$$HK3 + RR \underset{kd_1}{\overset{kb_1}{\rightleftharpoons}} HKpRR \xrightarrow{kt_1} HK0 + RRp$$
$$HK0 + RRp \underset{kd_2}{\overset{kb_2}{\rightleftharpoons}} HKRRp \xrightarrow{kt_2} HK3 + RR$$

Model III, cross-talking orthodox model. Model III is composed of two interacting models of the same type as model I. In this model, HK can react with RR' while HK' can react with RR. The cross-talking reaction rate is much lower than the reaction rate with the

cognate partner. Chemical reactions are shown below:

$$HK \underset{k_2}{\overset{k_1}{\rightleftharpoons}} HKp$$

$$HKp + RR \underset{k_4}{\overset{k_3}{\rightleftharpoons}} HKpRR \xrightarrow{k_5} HK + RRp$$

$$HK + RRp \underset{k_7}{\overset{k_6}{\rightleftharpoons}} HKRRp \xrightarrow{k_8} HKp + RR$$

$$HK' \underset{k_{22}}{\overset{k_{21}}{\rightleftharpoons}} HK'p$$

$$HK'p + RR' \underset{k_{24}}{\overset{k_{23}}{\rightleftharpoons}} HK'pRR' \xrightarrow{k_{25}} HK' + RR'p$$

$$HK' + RR'p \underset{k_{27}}{\overset{k_{26}}{\rightleftharpoons}} HK'RR'p \xrightarrow{k_{28}} HK'p + RR'$$

$$HKp + RR' \underset{k_{10}}{\overset{k_9}{\rightleftharpoons}} HKpRR' \xrightarrow{k_{11}} HK + RR'p$$

$$HK + RR'p \underset{k_{13}}{\overset{k_{12}}{\rightleftharpoons}} HKRR'p \xrightarrow{k_{14}} HKp + RR'$$

$$HK'p + RR \underset{k_{16}}{\overset{k_{15}}{\rightleftharpoons}} HK'pRR \xrightarrow{k_{17}} HK' + RRp$$

$$HK' + RRp \underset{k_{19}}{\overset{k_{18}}{\rightleftharpoons}} HK'RRp \xrightarrow{k_{20}} HK'p + RR$$

3.2.3 Calculation of steady states

Phosphorylated response regulator dimer can bind to the promoter upstream of target genes and initiate subsequent reactions. Thus we use the concentration of the phosphorylated response regulator dimer $[(RR_2)p]$ as output of the models. To analytically solve the equations, we first simplify the steady states equations by hand. Then we used the `Solve` function in Mathematica® 7.0 to solve the simplified equations analytically. To numerically solve the equations we use the `Python` programming language. The autophosphorylation rate k_1 of HK or H1 in the hybrid HK, is the input of the system. Upon stimulation, HK has increased autophosphorylation rate. We let k_1 increase from 10% of the maximum speed to its maximum speed gradually. At each point, we first use the numerical integrator to simulate the system for sufficiently long time (usually 100s, which has shown to suffice to reach close enough to the steady state for the next step) and record the final state. Then we apply the `fsolve` function with the final state in `scipy` to obtain the numerical solution for steady states. For each autophosphorylation rate, we obtain two steady states, one with lower initial state and one with higher initial state. If either one of the two steady states is more than 10% higher than the other one,

the system is defined to behave in a bistable manner. 10 000 simulations with different parameter sets are run for each model, and we record the percentage of bistable results.

3.2.4 Parameters configuration of the models

The parameters of model I and model II were set to the values shown in table 3.1. The values were used previously by Tiwari *et al.* [172].

Parameter		Value	Description
Model I	**Model II**		
k_1	k_1, k_2, k_3	0.00132 s^{-1}	Auto-phosphorylation and forward phosphotransfer in HK
k_2	k_4, k_5	0.001 s^{-1}	Dephosphorylation and reverse phosphotransfer in HK
k_{t1}	k_{t1}	0.5 s^{-1}	Phosphotransfer rate of HKp~RR complex
k_{t2}	k_{t2}	0.05 s^{-1}	Phosphotransfer rate of HK~RRp complex
$(k_{d1}+k_{t1})/k_{b1}$	$(k_{d1}+k_{t1})/k_{b1}$	50 μM	Michaelis-Menten constant for HKp~RR phosphotransfer
$(k_{d2}+k_{t2})/k_{b2}$	$(k_{d2}+k_{t2})/k_{b2}$	50 μM	Michaelis-Menten constant for HK~RRp phosphotransfer

TABLE 3.1: List of parameter values used in model I and model II. Model I and Model II share most parameters while differ in the forward and reverse phosphotransfer rates of HK.

3.2.5 Parameters randomization

In numerical simulations, each model is simulated with randomized parameters using Latin Hypercube Sampling [168]. Most of the parameters, if not specifically mentioned, range from 1 to 100. Cross-talking reaction rates in model III are set to be 0 to 20% (uniformly randomized) of the corresponding cognate partner reaction rates. 10 000 hypercubes are sampled for each model.

3.3 Results

3.3.1 The model of transcriptional/translational reactions of TCS is bistable

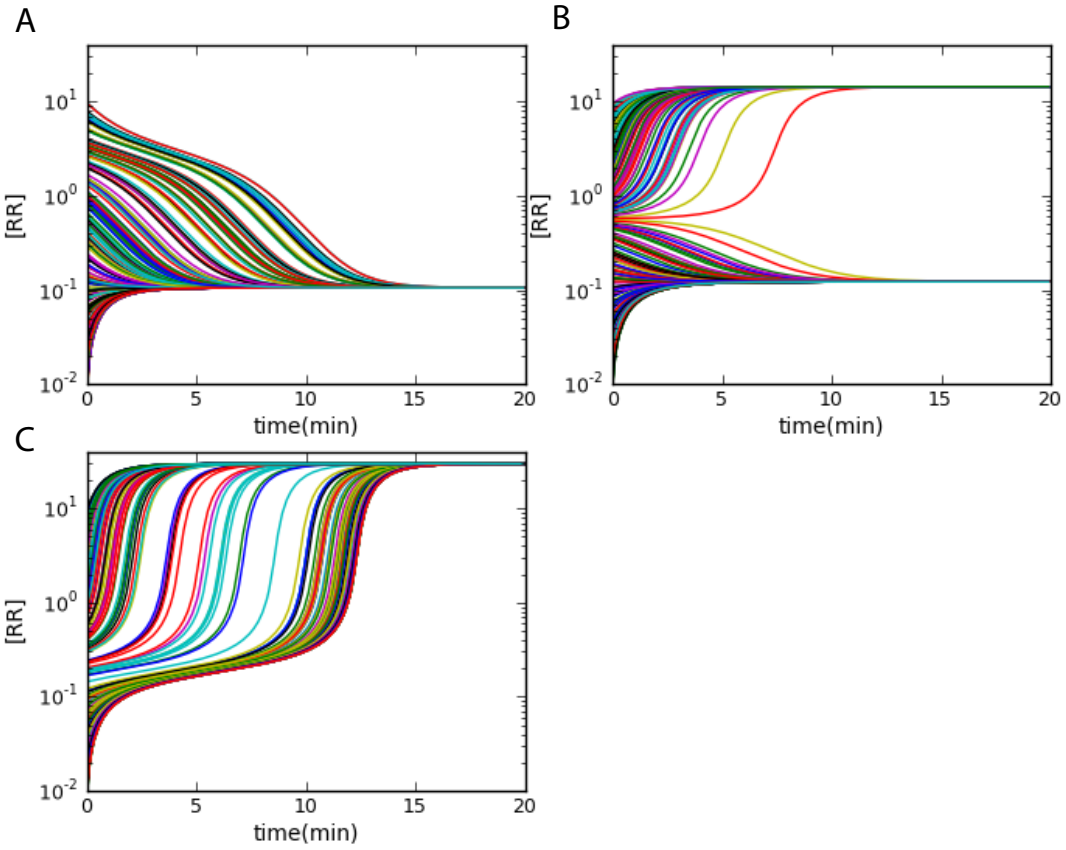


FIGURE 3.3: Simulation of the transcription/translation model of TCS. X axis is the simulated time. Y axis is the transit RR concentration. The parameters of basal rate of RR synthesis $k_b = 0.1min^{-1}$, degradation rate $k_{deg} = 1min^{-1}$, and the rate $K = 10$. The maximal Rate of RR synthesis V_{max} was set to different values in each simulation. A, $V_{max} = 5min^{-1}$, the system has monostability with a lower level of RR steady state; B, $V_{max} = 15min^{-1}$, the system is bistable; C, $V_{max} = 30min^{-1}$, the system has monostability with a higher level of RR steady state.

As mentioned in the methods section, the ODE for the monomer RR is

$$d[RR]/dt = \frac{V_{max}[RR]^2}{[RR]^2 + K} - k_{deg}[RR] + k_b,$$

where $K = k_r/(k_f K_d), Vmax = k_t[promoter]_{total}$. At steady state, the concentration of RR stays in a dynamic balanced state, where $d[RR]/dt = 0$. By substituting this into

the above equation, we obtain

$$k_{deg}[RR]^3 - (k_b + V_{max})[RR]^2 - k_{deg}K[RR] + k_bK = 0.$$

This equation has three positive roots. When they are different, the middle root corresponds to an unstable state, the other two roots correspond to two stable states. Thus we can conclude that this model is bistable when the above equation has three distinct positive roots for certain parameter sets. It has to be mentioned that if the dimerization reaction is removed from this system, the system is going to be monostable. Thus, besides positive feedback, dimerization is another necessary condition for this system to achieve bistability.

3.3.2 The model of post-translational reactions of TCS is monostable

In calculating the steady states of post-transcriptional models, we have transcriptional/translational reactions remaining static, which means that the total concentrations of the HK dimers and the RR dimers remain constant. Combining this with the other steady state equations, we can get an equation for our output $[RRp]$. For model I and model II, we can obtain analytical results, which are equations each with one negative root and one positive root. Only the positive root is relevant and the roots for model I is:

$$[RRp] = \frac{-X + \sqrt{X^2 + 4a(1+a)(1+b)cd(ac+bd)[RR_{tot}]}}{2(1+b)d(ac+bd)}, \tag{3.1}$$

where $X = (1+a)(ac+bd) + a(1+b)cd([HK_{tot}] - [RR_{tot}]), a = k_1/k_2, b = kt_2/kt_1, c = kb_1/(kd_1 + kt_1), d = kb_2/(kd_2 + kt_2)$.

Similarly, for model II we obtain:

$$[RRp] = \frac{-Y + \sqrt{Y^2 + 4a_3(1+a_0)(1+b)cd(a_3c+bd)[RR_{tot}]}}{2(1+b)d(a_3c+bd)} \tag{3.2}$$

where $Y = (1+a_0)(a_3c+bd)+a_3(1+b)cd([HK_{tot}]-[RR_{tot}]), a_1 = k_1/k_2, a_2 = k_1/k_5, a_3 = (k_1k_3)/(k_4k_5), a_0 = a_1 + a_2 + a_3, b = kt_2/kt_1, c = kb_1/(kd_1 + kt_1), d = kb_2/(kd_2 + kt_2)$.

We can see that the steady state of model I and model II share a similar form. They only differ in some of the coefficients (a and $a3$).

3.3.3 Numerical test for steady state in TCS post-translational reaction models

Because the post-translational cross-talking model for simple TCSs (model III) can only be solved with considerable numerical effort, we choose to use a computationally more efficient route. The shape of the simplified equations share some similarities with the equations of model I. Thus we can infer that they might also have only one steady state.

To test our assumption, we have performed numerical tests to solve the steady state of model III, as well as the other two models, for monostability. Results of 10 000 replicated simulations with different parameters showed that all of the three models are monostable, which is consistent with our analytical result.

3.3.4 Factors that could affect the steady state of orthodox and non-orthodox models

We have obtained analytical solutions for orthodox and non-orthodox models. With these solutions we can analyse how the phosphorelay of non-orthodox systems might affect their steady states. We can also analyse how different parameters will affect the system's steady state.

Under the parameters we have chosen $[RRp]$ ends up at a steady state of $9.22\mu M$ for the orthodox model. The steady state for $[RRp]$ is at $9.42\mu M$ for the non-orthodox model. Steady state of $[RRp]$ is 2% higher in the non-orthodox model than in the orthodox model. Our non-orthodox models always have higher values than the orthodox models even for the varying parameters. The reason for a higher steady state could be that non-orthodox systems have more sites to be phosphorylated. Thus it has more phosphorylated states, which means that the whole system can hold more phosphate groups.

We have also varied some important parameters in both models to see how they can affect the steady state. Because model I and model II always have different steady states of $[RRp]$ we only depict the fold-change of their $[RRp]$ steady states compared to the baseline level (the parameters shown in table 3.1). The stimulus level, which is reflected in the autophosphorylation rate, changes the level of the steady state most noticeably.

When the stimulus is increased 100 folds, the $[RRp]$ increases to 6 (non-orthodox) or 8 (orthodox) times of the original concentration (figure 3.4A). This is in consistent with the function of the TCSs, which is to sense changing levels of a stimulus and to response accordingly. However,the orthodox model changes more than the non-orthodox model, which is in agreement with earlier analyses [82].

In figure 3.4B, when $[HK_{total}]$ is increased from 1% to 100% of $[RR_{total}]$, $[RRp]$ has decreased 3% (non-orthodox) or 7% (orthodox). This fold change is not very significant. The Michaelis-Menten constant is a parameter controlling the dissociation tendency of an intermediate substrate. The larger this parameter, the more unstable is the intermediate substrate. In figure 3.4C, when the Michaelis-Menten constant is increased from $1\mu M$ to 100 μM, $[RRp]$ has increased 7% (non-orthodox) or 8% (orthodox). This change is not very pronounced, either. When the constant is less than 10 the non-orthodox model changes more than the orthodox model. When the constant is larger than 10 the non-orthodox model appears to become more robust to parameter changes.

These results show that the levels of the steady states of TCSs are most sensitive to the input stimuli. They are robust to the proportion of $[HK]$ to $[RR]$. Finally steady states of non-orthodox models seems to be more robust to parameter changes in most cases.

3.3.5 Transient $[RRp]$ of a cross-talking TCS model

To study how the dynamic behaviour of the system can be affected by cross-talk, we have compared the transient $[RRp]$ of simple TCSs and cross-talking TCSs upon various stimuli. In this study, we have modelled two orthodox TCSs named A and B, responding to different stimuli. In the "single" scenario, TCSs A and B are functioning independently without cross-talk. While in the "cross-talk" scenario, the A and B TCSs (each TCS has the same parameters as in the "single" scenario) have cross-talk to each other at a rate which corresponds to 10% of their normal reaction rate. The models are simulated with different stimuli sets where TCS A and TCS B receive different input stimuli. In this series of simulations we only check their transient levels instead of their steady states. The results are shown in figure 3.5.

In figure 3.5A we see that, when TCS A and TCS B are treated with the same stimuli, their transient levels are close to each other. This means that the systems are sensitive to

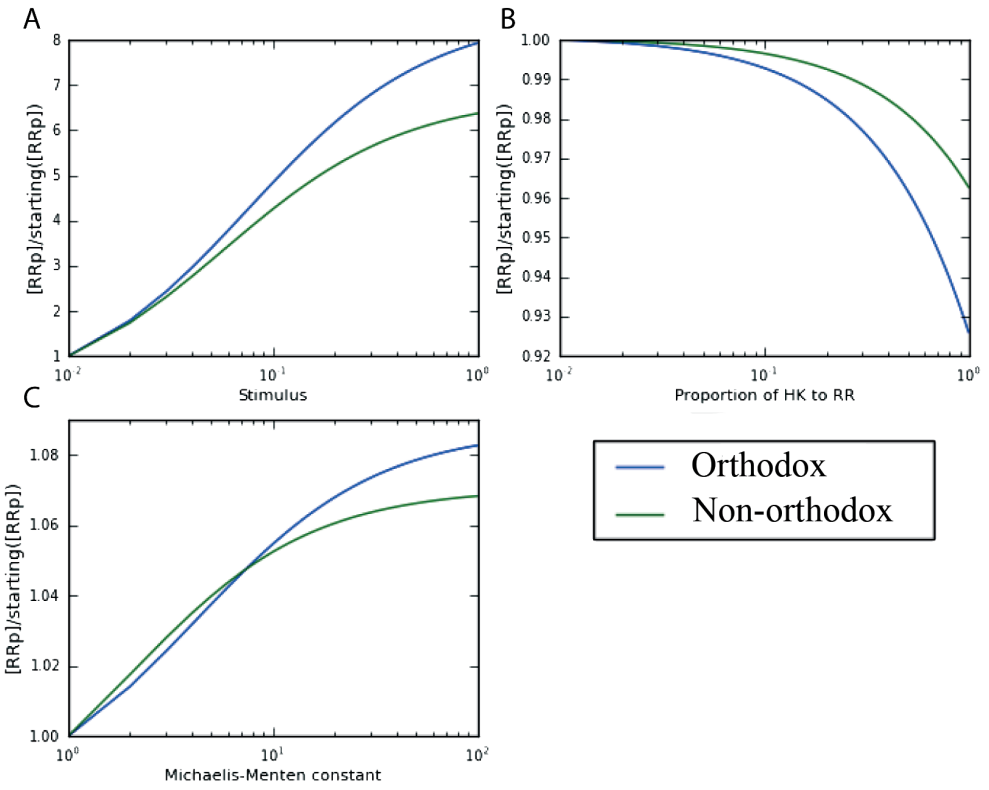


FIGURE 3.4: Orthodox and non-orthodox post-translational TCS models' steady states responding to changes of different parameters. The parameters of the models are set according to the list in table 3.1, except one of the parameters is set to be variable. Along the X axis is the varying parameter; the Y axis is the the level of steady state [RRp] of given changing parameter comparing to [RRp]'s steady state at baseline parameters. A, the stimulus, which is k_1 in both model, is changing from 1% to 100% of the original stimulus level. B, the concentration of total $[HK]$ is increasing from 1% to 100% of total $[RR]$. C, the Michaelis-Menten constant, which is $(k_{d1} + k_{t1})/k_{b1}$ and $(k_{d2} + k_{t2})/k_{b2}$ in model I and II, is increasing from 1 to 100 μM.

stimuli. However, each single TCS in a cross-talking system does not seem to change as much as the corresponding isolated system. When TCS A and TCS B are treated with opposite stimuli (shown in figure 3.5B), the cross-talking systems show higher mean output levels, which suggests that a cross-talking system would have higher response levels to the same signal, and that the cross-talking system is more stable to changing stimuli. This observation also holds for figure 3.5C, D where TCS A and TCS B are treated with yet more different stimuli.

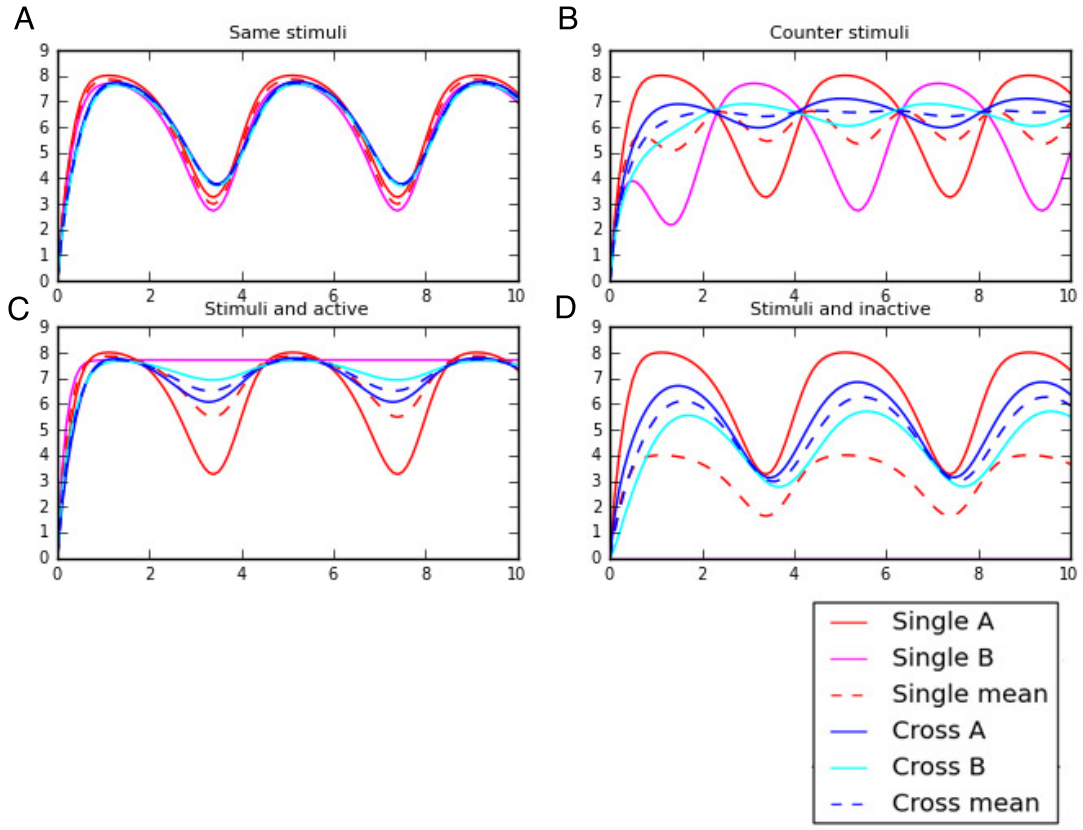


FIGURE 3.5: Transient concentration of [RRp] in "single" TCSs (TCS A and TCS B) and "cross-talk" TCSs (TCS A and TCS B) models responding to varying stimulus. The "single" TCSs and "cross-talk" TCSs are simulated with varying stimulus, reflected by varying autophosphorylation rate. TCSs A and B share the same parameters except that the phosphotransfer rate k_{t1} of model A is 1.5 times larger than the rate in model B. Along the X axis we show time measured by seconds; the Y axis is the fold change of the $[RRp]$ in each model. This figure shows transient $[RRp]$ of TCS A and TCS B respectively and also shows the mean value of TCS A and B in each scenario. (A) TCS A and B are treated by the same sinusoid shaped stimuli; (B)TCS A and TCS B are treated by opposite sinusoid shaped stimuli; (C) TCS A has sinusoid shaped stimulus and TCS B is constitutively active; (D) TCS A has sinusoid shaped stimulus and TCS B is constitutively inactive.

3.3.6 Steady state dynamics of TCS models with cross-talk

We next simulate two simple TCSs, A and B, working separately as well as interacting. The parameters we have used are the same as those used for figure 3.5. But here we are recording their steady states instead of their transient concentrations.

In figure 3.6A, when TCS A and TCS B are treated with the same stimulus, their steady states are close to each other. The level of $[RRp]$ has increased 10 folds for both simple

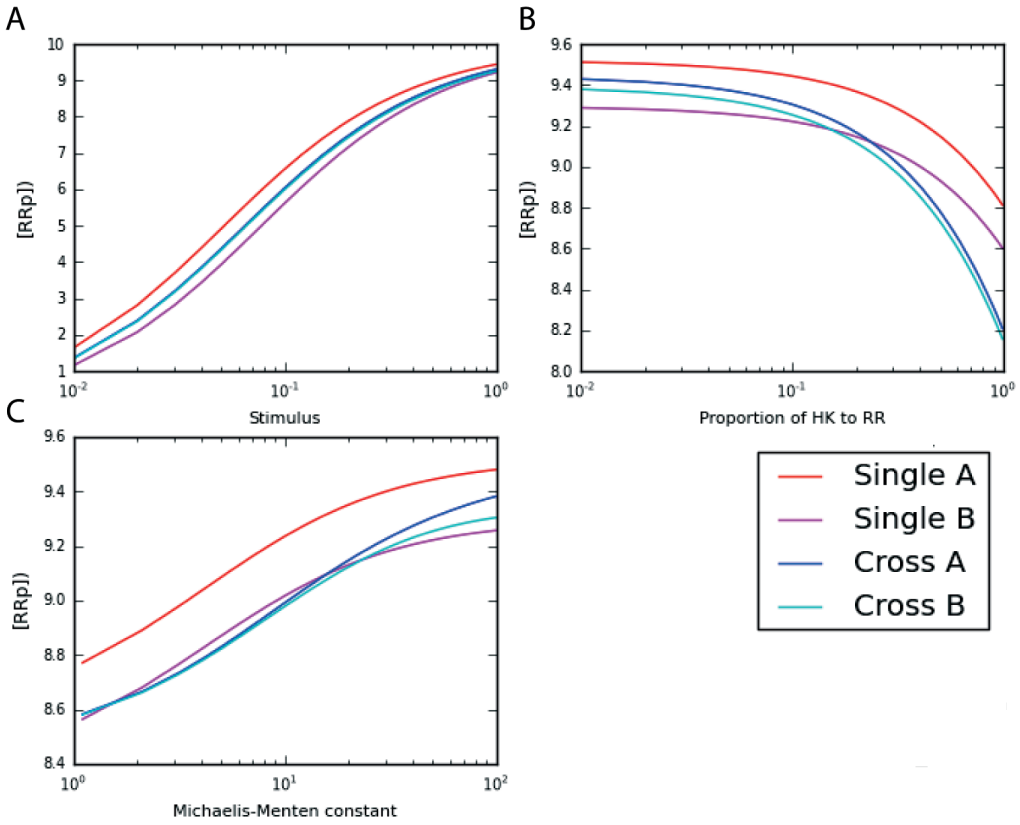


FIGURE 3.6: Simple and cross-talking post-translational TCS models' steady states responding to changes of parameters. Two simple TCSs, A and B, are set to share the parameters except that the phosphotransfer rate of HK RRp in TCS A is 1.5 folds increased compared to that of TCS B. TCS A and B are simulated assuming they are operating separately (single) or are interacting with each other (cross). The other parameters of the models are set according to the list in table 3.1The X axis is the varying parameter; the Y axis is the the level of steady state [RRp] of the given changing parameter compared to [RRp]'s steady state at the baseline parameters. (A) the stimulus, which is k_1 in both model, is changing from 1% to 100% of the original stimulus. (B) the concentration of total $[HK]$ is increasing from 1% to 100% of total $[RR]$. (C) the Michaelis-Menten constant, which is $(k_{d1} + k_{t1})/k_{b1}$ and $(k_{d2} + k_{t2})/k_{b2}$ in model I and II, is increasing from 1 to 100 μM.

and cross-talking models. In figures 3.6B, C, the changes are much less significant. This result means that the stimulus is the most important factor that affects the signal for both simple and cross-talking models. In figure 3.6 the differences of the levels of cross-talking A and B are not as high as for single A and B, which could mean that our cross-talking model can integrate the signal of the two constituent TCSs to achieve a more balanced signal. In real signal pathways, this may help to induce a less intense but more broad-spectrum response to a single input stimulus.

3.4 Conclusions

In this chapter we have studied the properties of the steady state behaviour in two component systems. As TCSs are responding to diverse stimuli, their function might require them to be bistable or monostable. In this chapter we have studied which elements might affect the steady states of TCSs.

We have built models of TCS transcription and post-translational reactions. These are large models with more than 10 variables and more than 25 rate parameters. Because we want to know the detailed relationship between TCS's bistability and its parameters numerical result alone would not fulfill our aim. Thus we have applied an approximation to simplify the calculation for steady states. In this method, we separate all the reactions into two parts: slow reactions (degradation, transcription, translation) and fast reactions (dimerization, phosphorylation, dephosphorylation). These two groups of reaction are thought to be approximately separable.

Because in TCS the RR is the molecule that promotes target gene expression or other cell activities we use RR or phosphorylated RR levels as the output signal strength of the whole system. When a stimulus arrives, the conformation of the HK molecule changes and thereby leads to an increase (or decrease) of the HK's autophosphorylation rate. Thus we use changing autophosphorylation rate of the HK to simulate the input signal of each TCS. Under these assumptions and conditions, we have obtained some valuable results and we have come up with the following conclusions.

Because phosphorylation reactions are not involved in slow reaction we monitor the monomer RR concentration $[RR]$ as the output of the slow-reaction models. The steady state equation of $[RR]$ can be simplified to a cubic equation:

$$k_{deg}[RR]^3 - (k_b + V_{max})[RR]^2 - k_{deg}K[RR] + k_bK = 0$$

Because the equation describes a biological system, all the parameters in the equation should be positive. Under this condition, the three roots of this equation would be either all positive real, or two complex roots and one positive real root. In the former condition, the system has three steady states. Two of them are stable states, the medium steady state is an unstable state. The system is bistable in this case. In the other case, the system will have only one steady state which is also a stable state. The system

is monostable in this case. By changing the initial $[RR]$ while fixing all the other parameters and variables, the system can switch between monostability and bistability as illustrated in figure 3.3.

When calculating steady states for the fast reactions, we have constructed 3 models that differ in their phosphorylation reactions. We first studied the steady state of the orthodox TCS and non-orthodox TCS models. We found that in order to calculate the steady state concentration of the phosphorylated RR dimer, which is the output of TCS models, we need to focus only on the reactions of the HK/RR dimers. These two models are both monostable, and the stable states are as the expressions shown in equation (3.1) and equation (3.2).

From the two expressions, we can see the stable states of both orthodox and non-orthodox models depend only on the total HK and RR dimer concentrations $[HK_{tot}]$ and $[RR_{tot}]$. We also found that the orthodox TCS and the non-orthodox TCS's steady state values are very similar. The differences between these two expressions reflect the differences in the phosphorylation processes of the two models.

To study how the parameters might change the steady states of these models, we simulated both models with the same parameter sets, and kept forward phosphorylation rate and reverse phosphorylation rate the same in both models. We then studied how steady states change as functions of the parameters. The parameters we have selected to vary are the stimulus of the system, HK/RR concentration ratio, and Michalis-Menten constant of phospho transfer from HK to RR. The absolute value of steady states differ for these two models (data not shown). The result in figure 3.4 shows that stimulus is the factor that affects the steady state most. When the stimulus increases 100 fold, the steady state increases 6 to 8 folds. The same fold-change in the HK/RR ratio and the reaction constants only lead to a small change in the steady state (approx. 0.1 fold). We also found that the non-orthodox model's steady state always changes less. This can be attributed to the buffering effect of the multi-layer phosphorylation in non-orthodox TCSs [82]. The steady state of the non-orthodox TCS thus becomes more robust to changes in environmental signal, transcription/translation rate, or HK/RR binding efficiency.

The steady state of the cross-talking TCS proves to be more difficult to analyse. However, the simplified cross-talking equations have symmetrical forms to that of the simplified orthodox steady state equations. We thus infer that they are also monostable. The numerical simulations confirm our reasoning.

By stimulating the cross-talking TCSs with different combinations of stimuli (figure 3.5), we found that compared to the separate TCSs, cross-talk can decrease the difference between the cross-talking TCSs. But the mean response of cross-talking TCSs is higher than that of the separate single TCSs.

We have also tested the influences of the stimulus, the HK/RR ratio and the Michaelis-Menten constant on the cross-talking TCS model by performing numerical simulations (figure 3.6). Results show that just like the transient concentrations, the cross-talking models' steady states are closer to each other than to the separate TCSs.

Chapter 4

The Phosphorelay Mechanism of Non-Orthodox Two-Component Systems

This work was carried out in collaboration with Dr. Goran Jovanovic and Professor Martin Buck. All simulations and statistical analyses were performed by me; but for consistency the work is discussed together.

4.1 Background

4.1.1 ArcB/A two-component system

As we have described in chapter 1, judging by the phosphorylation mechanisms, the TCSs can be categorized into the orthodox TCS and the non-orthodox TCS. The non-orthodox TCS performs a three-step phosphorelay instead of simple phosphorylation like the orthodox TCS performs.

The Arc (Anoxic redox control) system is a non-orthodox two-component system that in general mediates regulation of operons implicated in respiratory metabolism and enables facultatively aerobic bacteria to sense and respond to different respiratory conditions [104]. The Arc TCS comprises the membrane-bound tripartite hybrid HK ArcB and the cognate response regulator ArcA. The transcription of arcB is constitutive under all

growth conditions. Under anoxic conditions arcA transcription is significantly increased in a Fnr (global regulator that together with Arc system regulates transcription in response to O_2) dependent manner; ArcB autophosphorylates and through His292-Asp576-His717-Asp54 phosphorelay phosphorylates ArcA. Phosphorylated ArcA then represses many operons involved in aerobic respiration and up-regulates genes involved in anaerobic respiration (*e.g.* cydAB) and fermentation (*e.g.* pfl). It has been shown that the ArcB/A system activation is of particular importance under microaerobic growth conditions [2]. Under aerobic growth conditions ArcB kinase activity is inhibited by quinone, electron carriers that promote the oxidation of redox-active cysteine residues (C180, C241), which have been implicated in formation of an inactive ArcB dimer through intermolecular disulfide bond formation [18, 50, 103]. ArcA is dephosphorylated by ArcB-dependent reverse phosphorelay Asp54-His717-Asp576-Pi. It is likely that ArcB-Asp576 get dephosphorylated because the high-energy phospho-aspartyl bond is unstable [104]. The non-specific acetyl-P-dependent phosphorylation of ArcA can be dephosphorylated by ArcB's phosphatase activity [104].

4.1.2 The ArcB dimer might work in a different mechanism than other TCS

It is established that in TCSs the HKs form dimers after being translated from their coding mRNA. Dimerization is reported to be necessary for kinase activity in most TCSs [165, 193]. The reason for dimerization is probably because of the molecular mechanism of autophosphorylation in these TCSs. Autophosphorylation is the process where an activated HK homodimer changes its conformation and its His site gets phosphorylated by the catalyst (CA) domain of the monomer in the same HK. In most HKs this reaction follows an intermolecular pattern where one HK monomer is phosphorylated by the catalyst domain of the other HK monomer in the same dimer. TCSs that autophosphorylate in this manner include CheA, VirA, EnvZ, NtrB [119, 125, 165, 166, 192, 196] *etc.*. This pattern of autophosphorylation requires both monomers.

Intramolecular autophosphorylation, where one HK monomer is phosphorylated by its own catalyst (CA) domain, is still possible, although this does not appear to be widespread [131]. The hybrid HK ArcB is also reported to autophosphorylate in this

pattern, which could suggest that ArcB may not need the other monomer to get phosphorylated.

Although HKs function as dimers, it has also been reported that the HKs may phosphorylate their cognate RRs while in a special monomeric form (an HK monomer is covalently bound to a truncated HK monomer which only have dimerization and phosphotransfer domain) [139]. This fact suggests that as an HK has the potential to work with one of the monomers missing or defunct, they always have the ability of forming dimers.

Combining with the special autophosphorylation pattern of ArcB, we found several questions intriguing about both ArcB and non-orthodox HK. We already know ArcB autophosphorylates in a intramolecular pattern. Could ArcB be able to work with one of the monomer being complete dysfunction? Which pattern of phosphorelay does ArcB perform? How does HK transmitter module acts in the context of reverse phosphorelay? Here, we use computational simulations as well as the constitutively active ArcB (ArcB*) and their phosphorelay mutants as theoretical and in vivo experimental model system in an attempt to answer some of those questions.

4.2 Materials and Methods

4.2.1 Models for phosphorelay

Models of phosphorelay of TCSs were built following the descriptions of phosphorelay of two-component systems in the review written by Malpica *et al.* [104]. We have also investigated some other papers on simple TCSs [81, 82, 164]. The ODE models were built using the programme `COPASI` [60, 109]. Models were stored in SBML3.0 format as .xml files to be better shared with other programmes.

We have also build models of hybrid histidine kinase that use different mechanisms of phosphorelay. In the first model which we call trans-phosphorelay model, the phosphate group jumps from one monomer to the other during phosphorelay. In an alternative cis-phosphorelay model, the phosphate group stays in the original monomer during phosphorelay. The dimers are not allowed to bind more than one phosphate group at the same time. In our ODEs of the phosphorelay models, HK0 denotes the unphosphorylated HK

dimer; Xi {X in (R,L); i in (1,2,3)} denotes HKs with a specific site phosphorylated. For example, $R2$ ($HKHK - R2$) denotes a HK dimer phosphorylated on the second phosphorylation site (in D1 domain) of the right monomer. In a dimer, we define each monomer as *left* or *right* monomer for illustrating purpose, making them different. kab denotes the rate of the reaction which has a as a substrate, b as a product. For example, $kr1l2$, which only exists in trans-phosphorelay model, denotes the reaction rate of the phosphate group transferring from H1 of the right monomer to D1 of the left monomer. Another example is that $kl30$ denotes the reaction rate of HK transferring its phosphate group from H2 domain of the left monomer to RR (So HK changes from L3 to HK0). The models we built are as follows:

Reactions of trans-phosphorelay model:

$$HKHK \rightarrow HKHK\text{–}L1$$
$$HKHK\text{–}L1 \rightarrow HKHK\text{–}R2$$
$$HKHK\text{–}R2 \rightleftharpoons HKHK\text{–}L3$$
$$HKHK\text{–}R2 \rightarrow HKHK$$
$$HKHK\text{–}L3 + RR \rightleftharpoons HKHK + RRp$$

$$HKHK \rightarrow HKHK\text{–}R1$$
$$HKHK\text{–}R1 \rightarrow HKHK\text{–}L2$$
$$HKHK\text{–}L2 \rightleftharpoons HKHK\text{–}R3$$
$$HKHK\text{–}L2 \rightarrow HKHK$$
$$HKHK\text{–}R3 + RR \rightleftharpoons HKHK + RRp$$

The ODEs of trans-phosphorelay model are then,

$$
\begin{aligned}
d(HK0)/dt = &- k0l1 \times HK0 - k0r1 \times HK0 + kl20 \times L2 + kr20 \times R2 \\
&+ kl30 \times L3 \times rr + kr30 \times R3 \times rr - k0l3 \times HK0 \times rrp \\
&- k0r3 \times HK0 \times rrp + kl10 \times L1 + kr10 \times R1 \\
d(L1)/dt = &k0l1 \times HK0 - kl10 \times L1 - kl1r2 \times L1 \\
d(R1)/dt = &k0r1 \times HK0 - kr10 \times R1 - kr1l2 \times R1 \\
d(L2)/dt = &kr1l2 \times R1 - kl2r3 \times L2 + kr3l2 \times R3 - kl20 \times L2 \\
d(R2)/dt = &kl1r2 \times L1 - kr2l3 \times R2 + kl3r2 \times L3 - kr20 \times R2 \\
d(L3)/dt = &kr2l3 \times R2 - kl3r2 \times L3 - kl30 \times L3 \times rr + k0l3 \times HK0 \times rrp \\
d(R3)/dt = &kl2r3 \times L2 - kr3l2 \times R3 - kr30 \times R3 \times rr + k0r3 \times HK0 \times rrp \\
d(rr)/dt = &- kl30 \times L3 \times rr - kr30 \times R3 \times rr + k0l3 \times HK0 \times rrp + k0r3 \times HK0 \times rrp \\
d(rrp)/dt = &kl30 \times L3 \times rr + kr30 \times R3 \times rr - k0l3 \times HK0 \times rrp - k0r3 \times HK0 \times rrp
\end{aligned}
$$

For the cis-phosphorelay model we have:

$$
\begin{aligned}
HKHK &\rightarrow HKHK\text{–}L1 \\
HKHK\text{–}L1 &\rightarrow HKHK\text{–}L2 \\
HKHK\text{–}L2 &\rightleftharpoons HKHK\text{–}L3 \\
HKHK\text{–}L2 &\rightarrow HKHK \\
HKHK\text{–}L3 + RR &\rightleftharpoons HKHK + RRp
\end{aligned}
$$

$$
\begin{aligned}
HKHK &\rightarrow HKHK\text{–}R1 \\
HKHK\text{–}R1 &\rightarrow HKHK\text{–}R2 \\
HKHK\text{–}R2 &\rightleftharpoons HKHK\text{–}R3 \\
HKHK\text{–}R2 &\rightarrow HKHK \\
HKHK\text{–}R3 + RR &\rightleftharpoons HKHK + RRp
\end{aligned}
$$

while the corresponding ODEs are

$$
\begin{aligned}
d(HK0)/dt = &- k0l1 \times HK0 - k0r1 \times HK0 + kl20 \times L2 + kr20 \times R2 \\
&+ kl30 \times L3 \times rr + kr30 \times R3 \times rr - k0l3 \times HK0 \times rrp \\
&- k0r3 \times HK0 \times rrp + kl10 \times L1 + kr10 \times R1 \\
d(L1)/dt = &k0l1 \times HK0 - kl10 \times L1 - kl1l2 \times L1 \\
d(R1)/dt = &k0r1 \times HK0 - kr10 \times R1 - kr1r2 \times R1 \\
d(L2)/dt = &kl1l2 \times L1 - kl2l3 \times L2 + kl3l2 \times L3 - kl20 \times L2 \\
d(R2)/dt = &kr1r2 \times R1 - kr2r3 \times R2 + kr3r2 \times R3 - kr20 \times R2 \\
d(L3)/dt = &kl2l3 \times L2 - kl3l2 \times L3 - kl30 \times L3 \times rr + k0l3 \times HK0 \times rrp \\
d(R3)/dt = &kr2r3 \times R2 - kr3r2 \times R3 - kr30 \times R3 \times rr + k0r3 \times HK0 \times rrp \\
d(rr)/dt = &- kl30 \times L3 \times rr - kr30 \times R3 \times rr + k0l3 \times HK0 \times rrp + k0r3 \times HK0 \times rrp \\
d(rrp)/dt = &kl30 \times L3 \times rr + kr30 \times R3 \times rr - k0l3 \times HK0 \times rrp - k0r3 \times HK0 \times rrp
\end{aligned}
$$

As a final model, we also consider a model which requires full functionality on both monomers at each site in order to complete the phosphorelay. We refer to this models

as the allosteric phosphorelay model and write it as

$$
\begin{aligned}
HKHK &\rightarrow HKHK{-}L1 \\
HKHK{-}L1 &\rightarrow HKHK{-}L2 \\
HKHK{-}L1 &\rightarrow HKHK{-}R2 \\
HKHK{-}L2 &\rightleftharpoons HKHK{-}L3 \\
HKHK{-}R2 &\rightleftharpoons HKHK{-}L3 \\
HKHK{-}L2 &\rightarrow HKHK \\
HKHK{-}L3 + RR &\rightleftharpoons HKHK + RRp
\end{aligned}
$$

$$
\begin{aligned}
HKHK &\rightarrow HKHK{-}R1 \\
HKHK{-}R1 &\rightarrow HKHK{-}R2 \\
HKHK{-}R1 &\rightarrow HKHK{-}L2 \\
HKHK{-}R2 &\rightleftharpoons HKHK{-}R3 \\
HKHK{-}L2 &\rightleftharpoons HKHK{-}R3 \\
HKHK{-}R2 &\rightarrow HKHK \\
HKHK{-}R3 + RR &\rightleftharpoons HKHK + RRp
\end{aligned}
$$

with the corresponding ODEs given by,

$$
\begin{aligned}
d(HK0)/dt =& -k0l1 \times HK0 - k0r1 \times HK0 + kl20 \times L2 + kr20 \times R2 \\
&+ kl30 \times L3 \times rr + kr30 \times R3 \times rr - k0l3 \times HK0 \times rrp \\
&- k0r3 \times HK0 \times rrp + kl10 \times L1 + kr10 \times R1 \\
d(L1)/dt =& k0l1 \times HK0 - kl10 \times L1 - kl1r2 \times L1 - kl1l2 \times L1 \\
d(R1)/dt =& k0r1 \times HK0 - kr10 \times R1 - kr1l2 \times R1 - kr1r2 \times R1 \\
d(L2)/dt =& kr1l2 \times R1 - kl2r3 \times L2 + kr3l2 \times R3 - kl20 \times L2 + kl1l2 \times L1 \\
&- kl2l3 \times L2 + kl3l2 \times L3 \\
d(R2)/dt =& kl1r2 \times L1 - kr2l3 \times R2 + kl3r2 \times L3 - kr20 \times R2 + kr1r2 \times R1 \\
&- kr2r3 \times R2 + kr3r2 \times R3 \\
d(L3)/dt =& kr2l3 \times R2 - kl3r2 \times L3 - kl30 \times L3 \times rr + k0l3 \times HK0 \times rrp \\
&+ kl2l3 \times L2 - kl3l2 \times L3 \\
d(R3)/dt =& kl2r3 \times L2 - kr3l2 \times R3 - kr30 \times R3 \times rr + k0r3 \times HK0 \times rrp \\
&+ kr2r3 \times R2 - kr3r2 \times R3 \\
d(rr)/dt =& -kl30 \times L3 \times rr - kr30 \times R3 \times rr + k0l3 \times HK0 \times rrp + k0r3 \times HK0 \times rrp \\
d(rrp)/dt =& kl30 \times L3 \times rr + kr30 \times R3 \times rr - k0l3 \times HK0 \times rrp - k0r3 \times HK0 \times rrp
\end{aligned}
$$

In this model kinase activity depends on both phosphorylation sites to be present, while the phosphatase activity (dephosphorylation always occurs from the D1 domain) is more lenient.

4.2.2 Simulation of phosphorelay models

The ODE models are simulated using the `scipy` module in `Python` [122]. The parameters we used are as described in table 4.1 , but we also explore parameter space more globally by sampling parameters using Latin Hypercube sampling as described in the previous chapters. We have modelled HK dimers composed of wildtype or mutated monomers. In a dimer, we define each monomer as *left* or *right* monomer for illustrating purpose. The *left* monomer is arbitrarily defined as the monomer with higher expression level.

In models of trans-phosphorelay or cis-phosphorelay, reaction rates are set to 0 if either of the substrate or product can not exist because of the site-directed mutation. For

Parameter	Value	Description
k_{0y1},k_{x1y2},k_{x2y3}	0.00132 s^{-1}	Auto-phosphorylation and forward phosphotransfer in HK
k_{x20},k_{x3y2}	0.001 s^{-1}	Dephosphorylation and reverse phosphotransfer in HK
k_{x30}	0.5 s^{-1}	Phosphotransfer rate from HKp to RR
k_{0x3}	0.05 s^{-1}	Phosphotransfer rate from RRp to HK
k_{x10}	0.0001 s^{-1}	Spontaneous dephosphorylation rate of HK1
Initial [HK]	1μM	Initial unphosphorylated HK concentration
Initial [phosphorylated HK]	0μM	Initial concentration of all phosphorylated HK
Initial [RR]	8μM	Initial unphosphorylated RR concentration
Initial [RRp]	2μM	Initial concentration phosphorylated HK(this is also the value for negative control)

*: x,y could be either ″L″ or ″R″.

TABLE 4.1: List of parameter values used in the phosphorelay models. Forward phosphorelay includes the reactions the transfer the phosphate group toward response regulator. Reverse phosphorelay includes the reactions the transfer the phosphate group away from response regulator until dephosphorylation.

instance, if site 2 (D1) of the right dimer is mutated (D576A), $kl1r2$, $kr2l3$, $kl3r2$, $kr20$ in the trans-phosphorelay model, and $kr1r2$, $kr2r3$, $kr3r2$, $kr20$ in the cis-phosphorelay model will all be set to 0. In the allosteric model the forward phosphotransfer reaction rate of site X is set to 0 if either of the monomers involves a mutation at site X. For instance, if site 2 of the right dimer is mutated, $kl1r2$, $kr2l3$, $kr1l2$, $kl2r3$, $kr2r3$, $kl2l3$, $kr1r2$, $kl1l2$ *etc.* will all be set to 0. However, reverse phosphotransfer is blocked only when both phosphorylation sites are mutated.

In the simulations, as in the experiments, the system will have plasmids producing different monomers with different copy numbers. Suppose the system contains nl copies of HK U and nr copies of HK V (U, V can be wildtype ArcB, H292A, D576A or H717A). Then the proportion of UU, VV homodimer and UV heterodimer in total HK dimers would be $[nl/(nl+nr)]^2$, $[nr/(nl+nr)]^2$ and $nl\,nr/(nl+nr)^2$. The output of the system is calculated as:

$$\sum \text{proportion of a dimer} \times \text{output of this pure dimer}$$

To simplify the simulation of the models, we assume that similar reactions share the same reaction rate. In this more constrained configuration there are only 4 parameters for each of the three models, $kf, kr, kt1$ and $kt2$. They denote the reaction rates of the forward phosphorylation, reverse phosphorylation, phosphotransfer from HK to RR and phosphotransfer from RR to HK, with $kf \in \{k0x1, kx1y2, kx2y3\}, kr \in \{kx20, kx3y2\}, kt1 = kx30, kt2 = k0x3, x \in \{l, r\}, y \in \{l, r\}$. In another model where the parameters are more flexible, and thus more realistic, we have 7 parameters in this model: $kf0 = k0x1, kf1 = kx1y2, kf2 = kx2y3, kr3 = kx3y2, kr2 = kx20, kt1 = kx30, kt2 = k0x3, x \in \{l, r\}, y \in \{l, r\}$. In both configurations, the spontaneous dephosphorylation rates of HK1 are set to be 10% of the dephosphorylation rate of D1 (kr or $kr2$).

4.2.3 Parameter optimisation and model evaluation

We apply the parameters listed in table 4.1 to the models, and optimise the parameters to obtain the minimum difference between the models' output and the observed experimental data. We use the `fmin_l_bfgs_b` function in the `python` module `scipy` to optimise the parameters. To keep the parameters in a reasonable biological range, we set the boundaries for the parameters. For each parameter, the lower boundary is set to be 0.1% of the unoptimised parameter, while the upper boundary is set to be 1000 folds of the original parameter. The output of each optimised model is compared with experimental data and the Bayesian Information Criteria (BIC) is calculated for each model as

$$\text{BIC} = n \cdot \ln(\widehat{\sigma_e^2}) + k \cdot \ln(n).$$

Here n is the number of datapoints, k is the number of free parameters, $\widehat{\sigma_e^2}$ is the error variance, which is defined as $\frac{1}{n}\sum_{i=1}^{n}\text{error}_i$.

Subsequently, the Schwarz weight of each model is calculated as [163]:

$$\begin{aligned}\Delta_i &= \text{BIC}_i - \text{BIC}_{min} \\ w_i &= \frac{\exp(\frac{-\Delta_i}{2})}{\sum_{j=1}\exp(\frac{-\Delta_j}{2})}\end{aligned}$$

In the above equations, $i, j \in \{\text{trans-, cis-, allosteric}\}$. The Schwarz weight equals the probability that, among the three models, a certain model is the one that describes the real biological system judging by their differences to the experimental data.

4.2.4 Sensitivity assays

We performed sensitivity analysis on the model parameters using the `matlab` package `StochSens`. This package calculates the Fisher Information Matrix (FIM) for each model and calculates their sensitivities from the FIMs [85] (FIMs for deterministic models are used). The models used are the same as described previously.

4.2.5 Bacterial strains, media and growth conditions

The experimental work was carried out by Dr. Goran Jovanovic and is here discussed for the sake of completeness. The bacterial strains used in this study are shown in Table 4.2. Strains were constructed by transduction using the P1vir bacteriophage [111] and were grown in Luria-Bertani (LB) broth or on LB agar plates at 37℃ [111]. To eliminate the Kan cassette from a $\Delta arcA$::Kan mutant and construct the marker-less $\Delta arcA$ variant (strain MVA113), we used plasmid pCP20 and the method described by Cherepanov and Wackernagel [32] (see Table 4.2). The colony PCR and creD/ yjjY pair of primers was used to verify that $\Delta arcA$ mutant in MVA113 strain had the correct structure. For in vivo bacterial two-hybrid system assays, strains were grown in LB at 30℃. For aerobic growth, overnight cultures of cells were diluted 100-fold into 5 ml of LB in a universal tube with loose fitting caps and shaken at 200 rpm. For microaerobic growth, overnight cultures of cells were diluted 100-fold into 5 ml of LB and shaken at 100 rpm. For induction of the T5/lacUV5 promoters, 0.1 mM isopropyl-β-D-galactopyranoside

(IPTG) was added for 1 hour. For scoring the lacZ$^+$ colonies, indicator plates containing 40 μl of 20 mg/ml stock solution of 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) and 0.5 mM IPTG were used. Antibiotics were routinely used at the following concentrations: ampicillin (Amp; 100 $\mu g\, ml^{-1}$), kanamycin (Kan; 25 (or 50 for BACTH) $\mu g ml^{-1}$), and chloramphenicol (Cam; 30 $\mu g ml^{-1}$), and spectinomycin (Spc; 50 $\mu g ml^{-1}$).

4.2.6 DNA manipulations

Plasmids used in this study are listed in Table 4.2. Using PCR-based site-specific mutagenesis (Quickchange mutagenesis kit, Stratagene) of the plasmid templates pGJ23 and pGJ73 and pGJ74, we constructed plasmids pGJ72, pGJ74 and pGJ74/78/80, respectively (see Table 4.2). All constructs were verified by DNA sequencing. For the in vivo bacterial two-hybrid system experiments, the ArcB* and its phosphorelay mutants were fused to either T25 (plasmid pKT25) or T18 (pUT18C) Cya domains as described in Jovanovic *et al.* [71]. The genes were amplified from corresponding plasmids (see Table 4.2) using primers that introduce XbaI-KpnI restriction sites, cloned in pGEM-T Easy (Promega), and than subcloned in-frame into multiple cloning sites (MCS) of pKT25 and pUT18C, respectively. All constructs were verified by DNA sequencing. Transformation of bacteria was performed as described by Miller [111].

4.2.7 Western blot analysis

For the Western blot analysis, bacterial cells were harvested at mid-log phase (OD600$\sim$0.5) and re-suspended in a mix of 30 μl 4% SDS and 30 μl Laemelli buffer (Sigma). Samples used for Western blotting were normalized according to cell growth measured at OD600. Samples were separated on 12.5% (SDS)-PAGE and transferred onto PVDF membrane using a semidry transblot system (Bio-Rad). Western blotting was performed as described [40] using antibodies to ArcB (a gift from D. Georgellis) (1:15000 with anti-rabbit). The proteins were detected using the ECL plus Western Blotting Detection Kit according to manufacturers guidelines (GE Healthcare). Images were captured in a FujiFilm intelligent Dark Box by an image analyser with a charge-coupled device camera (LAS-3000). Densitometry analysis was performed with MultiGauge 3.0 software (FujiFilm USA Inc., Valhalla, NY) and quantification (results expressed in arbitrary units) performed using the AIDA software.

4.2.8 In vivo bacterial two-hybrid system

The adenylate cyclase Cya^- based bacterial two-hybrid system (BACTH) allows detection of protein-protein interactions in vivo and is particularly appropriate for studying interactions among membrane proteins [72, 73]. The BACTH assay was used here to study in vivo protein-protein interactions of ArcB* and its phosphorelay mutants. Protein fusions (see Figure 4.4) were assayed as described [71]. As a negative control we used pKT25 and pUT18C vectors in the absence of fusion proteins, as a positive control we used pKT25-zip and pUT18C-zip plasmids carrying fused GCN4 leucine-zipper sequence [72]. After co-transformation of the BTH101 strain with the two plasmids expressing the fusion proteins, selection plates (Kan, Amp, X-gal and IPTG) were incubated at room temperature for 72 h. The levels of the interactions were quantified by β-Galactosidase activity in liquid cultures (see below). Chromosomal LacZ expression 3-fold above the negative control (vectors alone) value was scored as a positive interaction signal.

4.2.9 β-Galactosidase (β-Gal) assays

Activity from a single copy chromosomal ϕ ($'cydA-lacZ$) transcriptional fusion exclusively regulated by ArcA [18] was assayed to gauge the level of cydA promoter activation level. Cells were grown overnight at 37°C in LB broth containing the appropriate antibiotic (s) and then diluted 100-fold (initial OD600nm$\sim$0.025) into the same medium (5 ml). Following incubation to OD600nm 0.2-0.3, cultures were induced with IPTG for 1 h and then assayed for β-Gal activity as described by Miller [111]. The β-Gal activity from a chromosomal lacZ gene in BTH101 strain was assayed to estimate the protein-protein interactions in BACTH assay. Bacteria were grown in LB medium containing 100 $\mu g\, ml^{-1}$ Amp and 50 $\mu g\, ml^{-1}$ Kan at 30°C for 16 h, then cultures were diluted 1:25 and grown until the OD600 nm$\sim$0.3, then 0.5 mM IPTG was added and the cells incubated for a further 1 h at 30°C. For all β-Gal assays, mean values of six independent assays taken from technical duplicates of three independently grown cultures of each strain were used to calculate activity. The data are shown as a mean values with SD error bars.

4.3 Results

4.3.1 Modelling of monomer and dimer histidine kinase of two-component system

Our aim in this research is to study the dynamics of phosphorylation in two component systems. We have first investigated two models of the phosphorelay process of non-orthodox TCSs. In the first model, the histidine kinase is modelled to function as a monomer, while in the second model, the histidine kinase is modelled to be functioning as a dimer. The sequence of phosphorelay was described before. We simulate this phosphorelay process both in the monomer model and the dimer model. When a HK is not phosphorylated, it is denoted as HK_0. If the molecule is phosphorylated at site H1, D1,H2, the molecule is denoted as HK_1, HK_2 and HK_3 respectively. The chemical equations of the TCS models are as follows: Model I, monomer model:

$$HK_0 \rightarrow HK_1$$

$$HK_1 \rightarrow HK_2$$

$$HK_2 \rightleftharpoons HK_3$$

$$HK_2 \rightarrow HK_0$$

$$HK_3 + RR \rightleftharpoons HK_0 + RRp$$

Model II, dimer model:

$$HK + HK \rightleftharpoons HKHK_{total}$$

$$HKHK_0 \rightarrow HKHK_1$$

$$HKHK_1 \rightarrow HKHK_2$$

$$HKHK_2 \rightleftharpoons HKHK_3$$

$$HKHK_2 \rightarrow HKHK_0$$

$$HKHK_3 + RR \rightleftharpoons HKHK_0 + RRp$$

Because HKs almost always form dimers, we are trying to analyse if there is any benefit in dynamic behaviour for HKs to function in this way compared to a monomer. In model II, the dimer model, we assume the system has already reached a steady state. Because the dimerization reaction constant is independent of the phosphorylation state [132], as we have already stated in chapter 3, the total HK dimer concentration will not change during the phosphorelay. Thus the dimerization reaction does not need to be considered in model II. If we ignore the dimerization reaction of model II, we can easily see that model I and model II are in fact the same model but with different variable names. Our simulations also show that if given equivalent sets of parameters, these two models will have the same temporal behaviour (data not shown). Thus we can confirm that the bacterial TCSs do not evolve dimer HKs because they have a dynamic advantage. It might just be that in a dimer the 3D structure has other advantages, or indeed enables the phosphorelay mechanism in the first place.

4.3.2 Modelling trans- and cis-phosphorelay of non-orthodox two-component system

Non-orthodox two-component systems have three sites able to bind to phosphate groups. If we are modelling this system, we have to specify whether we allow one HK dimer to simultaneously bind more than one phosphate group, or if we allow one HK dimer to bind only one phosphate group at a time. By comparing these two mechanisms we found no difference when comparing their output signals at transient time points and at steady state (data not shown). Because the multiple-sites-phosphorylation model is computationally less convenient and it is also not supported by the previous studies, we use the one-phosphate-group-at-a-time assumption in our models.

We first only proposed two phosphorelay models, the trans-phosphorelay and cis-phosphorelay models. We can see that the wild type trans-phosphorelay model and cis-phosphorelay model are in fact identical except for the choice of variable names. Their kinetic behaviour is thus identical and we have to conclude that the phosphorelay mechanism is thus non-identifiable from time-resolved data.

In order to induce difference in these two models we therefore introduce point mutations into the models. We define that if one of the phosphorylation sites on HK (H1, D1 or H2) is mutated, then all of the reactions involving that site will be hindered and their

reaction rates are set to 0. However, in the system with mutant homodimer it is still not possible to separate the cis-phosphorelay model from the trans-phosphorelay model. Output of both systems would be fully blocked. Thus we instead consider the models for mixed HKs. In this configuration we model systems that contain two different HKs which are either wild type HK or H1, D1, H2 mutant. They can bind to each other freely, as was experimentally verified using bacterial two-hybrid assays. Suppose we have mixed wild type and H1 mutant; we can then expect to have different proportions of wild type homodimer, H1 mutant homodimer, and wild type $\times$ H1 mutant heterodimer in the system. As we have stated before, homodimer models of both phosphorelay patterns will produce the same result. However, H1 mutant $\times$ D1 mutant heterodimer and D1 mutant $\times$ H2 mutant heterodimer could be expected to produce different results for trans- and cis-phosphorelay models.

From the arguments above we can conclude that:

- The dimer and monomer models are different. However, if we only consider the process after the dimerization, these two models are the same.
- Multi-phosphorylation is not different from the single-phosphorylation model.
- The trans- and cis-phosphorelay model are indistinguishable using the wild type model.
- By introducing mutants into the system, we are able to generate difference in these two phosphorelay patterns.
- To tell if an TCS is using trans- or cis-phosphorelay mechanism, we still have to use experiments.

Based on these results artificial ArcB constructs were generated with phosphorylation sites at H1, D1 and H2 ablated. These are then used to compare the predictions resulting from the different phosphorelay models.

4.3.3 Activities of the constitutively active ArcB* phosphorelay mutants

To assess the activities of the hybrid sensor in phosphorelay we constructed mutants of constitutively active ArcB* [71] (see figure 4.1 and Table 4.2) and used a $\Delta arcB$ strain

TABLE 4.2: *E. coli* K-12 strains and plasmids used in this study

Strain/plasmid	**Relevant characteristics**	**Reference**
Strain		
ASA12	MC4100 *recA*, λRSS2[ϕ ('*cydA-lacZ*)] (kan^r)	Bekker *et al.*, 2010 [18]
MVA92	Δ*arcB59*	Jovanovic *et al.*, 2009 [71]
MVA104	MVA92[ϕ ('*cydA-lacZ*)] (kan^r)	This work MVA92 × ASA12
JWK4364	BW25113 Δ*arcA*::Kan (kan^r)	Baba *et al.*, 2006 [6]
MVA113	Δ*arcB59* Δ*arcA*	This work MVA92 × JWK44364 × pCP20
MVA114	MVA113 [ϕ ('*cydA-lacZ*)] (kan^r)	This work MVA113 × ASA12
BTH101	cya^-, lac^+	A gift from D. Ladant
XL1-Blue	tet^r	Laboratory collection
Plasmid		
pCA24N	Expression vector, $P_{T5/lac}$ promoter, $lacI^q$, *ori* pMB1, (cam^r)	Kitagawa et al., 2005 [83]
pJW5536 (-)	$P_{T5/lac}$-6*xhis-arcB* (arcB cloned into pCA24N, encoding ArcB wild type), $lacI^q$ (cam^r)	Kitagawa et al., 2005 [83]
pGJ23	pJW5536 (-) encodes ArcB $_{C180A/C241A}$ (ArcB*) (cam^r)	Jovanovic et al., 2009 [71]
pGJ30	pGJ23 encodes ArcB* $_{H292A}$ (cam^r)	Jovanovic et al., 2009 [71]
pGJ72	pGJ23 encodes ArcB* $_{D576A}$ (cam^r)	This work
pGJ27	pGJ23 encodes ArcB* $_{H717A}$ (cam^r)	Jovanovic et al., 2009 [71]
pAPT110	Expression vector, P_{lacUV5} promoter, $lacI^q$, *ori* p15A, (spc^r)	A gift from J. Beckwith
pGJ73	P_{lacUV5}-6*xhis-arcB* (*arcB* cloned into pAPT110 *Xba*I-*Kpn*I, encoding ArcB wild type (spc^r)	This work
pGJ74	pGJ73 encodes ArcB* (spc^r)	This work
pGJ80	pGJ74 encodes ArcB* $_{H292A}$ (spc^r)	This work
pGJ78	pGJ74 encodes ArcB* $_{D576A}$ (spc^r)	This work
pGJ76	pGJ74 encodes ArcB* $_{H717A}$ (spc^r)	This work
pKT25	IPTG-inducible vector containing the T25 domain of Cya upstream of the MCS (kan^r)	A gift from D. Ladant
pUT18C	IPTG-inducible vector containing the T18 domain of Cya upstream of the MCS (amp^r)	A gift from D. Ladant
pKT25-zip	GCN4 leucine zipper fusion to the C-terminus of the T25 domain of Cya in pKT25 (kan^r)	A gift from D. Ladant
pUT18C-zip	GCN4 leucine zipper fusion to the C-terminus of the T18 domain of Cya in pUT18C (amp^r)	A gift from D. Ladant
pCP20	FLP^+, λcI857$^+$, λp_R Rep^ts, (amp^r, cam^r)	Cherepanov and Wackernagel, 1995 [32]
pGEM-T Easy	Cloning vector (amp^r)	Promega

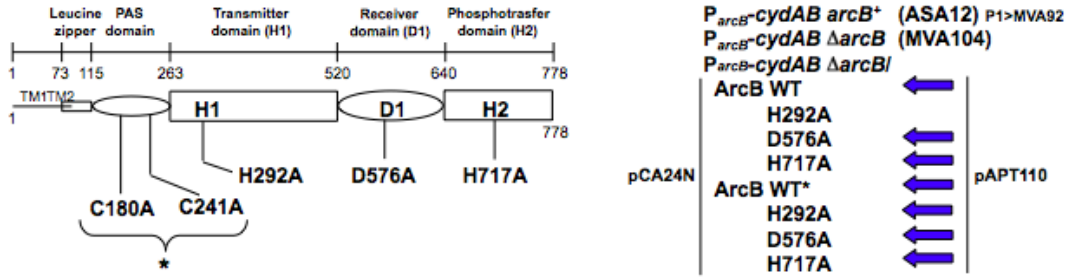


FIGURE 4.1: Diagram of constitutively active ArcB and mutants. The left subfigure shows the digram of the mutation site. The right subfigure shows the plasmids used to carry the constitutively active ArcB* and mutants constructs.

carrying the chromosomal ϕ ('$cydA-lacZ$) transcriptional fusion specifically regulated only by ArcA with the Fnr regulatory sequences deleted [18] (see Table 4.2). To avoid ArcA expression level to be the limiting factor, the experiments were carried out in microaerobiosis where expression of ArcA is induced by Fnr.

4.3.4 WT and mutant forms of ArcB and ArcB* proteins were similarly expressed when placed on compatible plasmids pCA24N and pAPT110

Initially we showed that constitutively active form of ArcB, ArcB*, activates transcription of ϕ ('$cydA-lacZ$) fusion independent of growth conditions, bypassing all upstream signalling needed for the activation of the kinase activity (Figure 4.2). We then determined that WT and mutant forms of ArcB and ArcB* proteins were similarly expressed when placed on compatible plasmids pCA24N and pAPT110, respectively (Fig. 4.3). Since dimerization of sensor is prerequisite for its kinase activity, the ArcB* and its phosphorelay mutants were tested for the pairwise interaction using bacterial two-hybrid (BACTH) system (see Materials and Methods section and Figure 4.4); the results showed that mutations introduced into ArcB* do not interfere with formation of dimers. As a control experiment, the ArcB* or its separate phosphorelay mutant was expressed from either pCA24N or pAPT110 vector in a $\Delta arcB$ strain, and efficiency of ArcA phosphorylation was assessed by measuring β-Gal activity of ϕ ('$cydA-lacZ$) fusion following ArcA-dependent activation of transcription (Figure 4.5). As expected, the results showed that any disruption of phosphorelay present in both monomers of sensor kinase homodimer abolishes activation of ArcA. Notably, some residual activation above

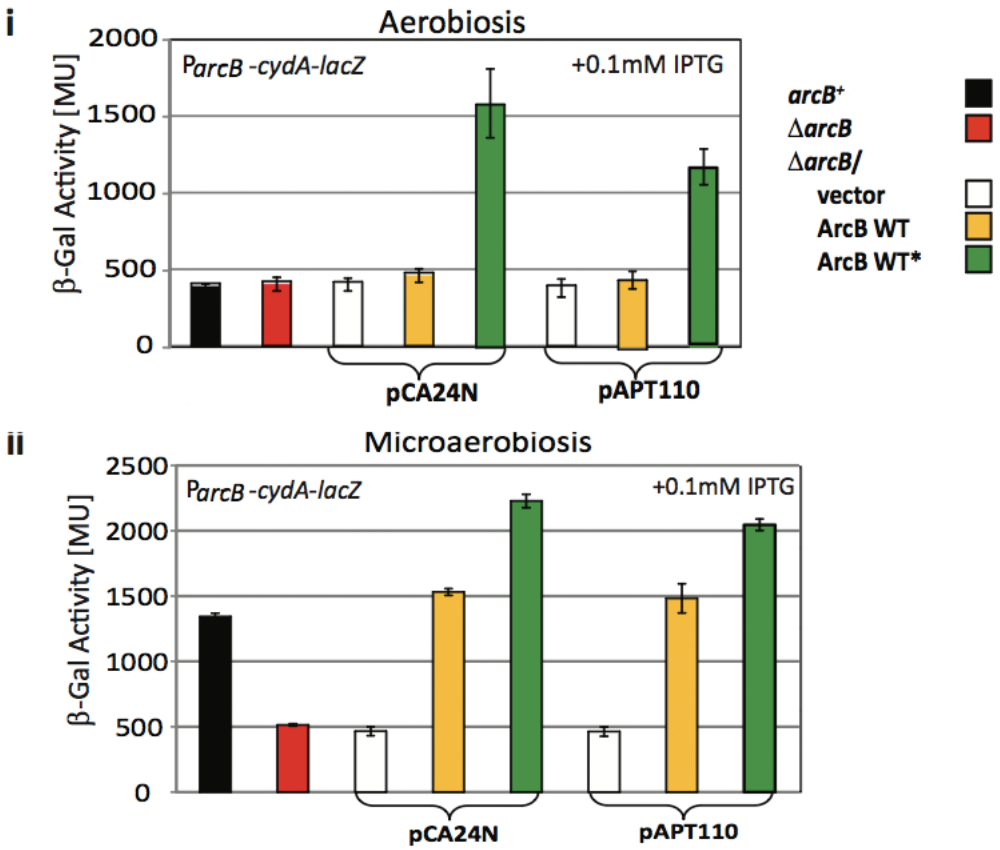


FIGURE 4.2: Phosphorelay of constitutively active ArcB variants lacking the key phosphorelay residues. Expression of a $cydA-lacZ$ chromosomal transcription fusion under unique control of the ArcB/A system was measured using a β-Gal assay (see Materials and Methods) in cells grown in aerobiosis (i) or microaerobiosis (ii) in the presence of 0.1 mM IPTG. $arcB^+$ (MG1655 PcydA-lacZ); $\Delta arcB$ (MVA104); $\Delta arcB$ (MVA104) carried vector pCA24N or pAPT110 alone, expressed plasmid (pCA24N- or pAPT110-based) borne ArcB wild type [pKW5536(-) or pGJ73] or constitutively active ArcB* (pGJ23 or pGJ74). For all β-Gal assays, mean values of six independent assays taken from technical duplicates of three independently grown cultures of each strain were used to calculate activity. The data are shown as a mean values with SD error bars.

the control basal level (vector alone) was detected when ArcB^{*D576A} or ArcB^{*H717A} mutants were expressed from a low copy number vector (Figure 4.5).

4.3.5 Combination of ArcB mutants

Finally, the combinations of either ArcB* or separate phosphorelay mutant of ArcB* with different ArcB* mutants were co-expressed in a $\Delta arcB\phi$ ('$cydA-lacZ$) strain and phosphorylation efficiency of ArcA is determined as above (Figure 4.6). The results showed that:

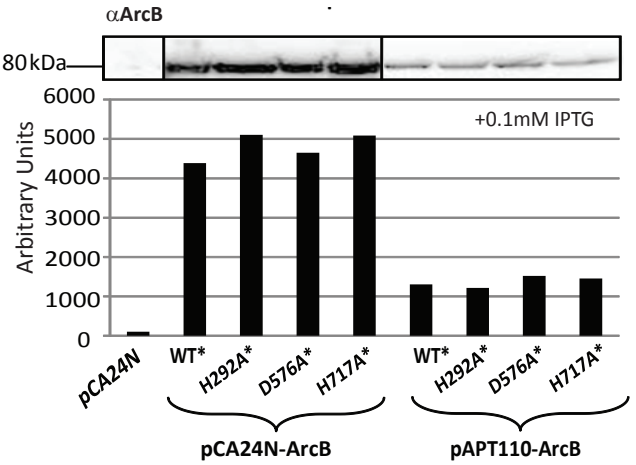


FIGURE 4.3: Phosphorelay of constitutively active ArcB variants lacking the key phosphorelay residues. The level of ArcB* and its variant proteins (see Figure 4.1) expression from either pCA24N or pAPT110 based plasmids in a Δ*arcB* (MVA104) strain grown in microaerobiosis (control was vector pCA24N alone) was assessed using Western blotting and antibodies against ArcB (α-ArcB) (see Materials and Methods). The level of expression was presented in arbitrary units after quantification using `ImageJ` software. For all β-Gal assays, mean values of six independent assays taken from technical duplicates of three independently grown cultures of each strain were used to calculate activity. The data are shown as a mean values with SD error bars.

- Co-expression of ArcB* with any of the sensor kinase phosphorelay mutants significantly diminished activation of ArcA, ArcB*^{H292A} mutation having the most detrimental effect;

- Co-expression of different ArcB* mutants that disrupt the sensor kinase phosphorelay abolished activation of ArcA, with the most pronounced effect of ArcB*H292A mutation in combinations;

- Below basal level activation of ϕ ('$cydA - lacZ$) when at least one monomer of ArcB* carried intact D576 and H717 and inactivated H292 residues, suggesting the reverse phosphorelay activity of corresponding ArcB* mutant monomer and so phosphatase activity of ArcB* mutant sensor kinase and dephosphorylation of ArcA (that might be activated by cross-talk with *e.g.*, acetyl-P).

To answer whether non-specifically activated ArcA might contribute to the basal level expression (only vectors) of ϕ ('$cydA - lacZ$) fusion in Δ*arcB* strain, we compared this result to the basal level activity obtained in a Δ*arcB*Δ*arcA* strain lacking ArcA response regulator (Figure 4.7). Apparently, the presence of ArcA contributes to the basal level expression seen in Δ*arcB* strain co-expressing both vectors alone.

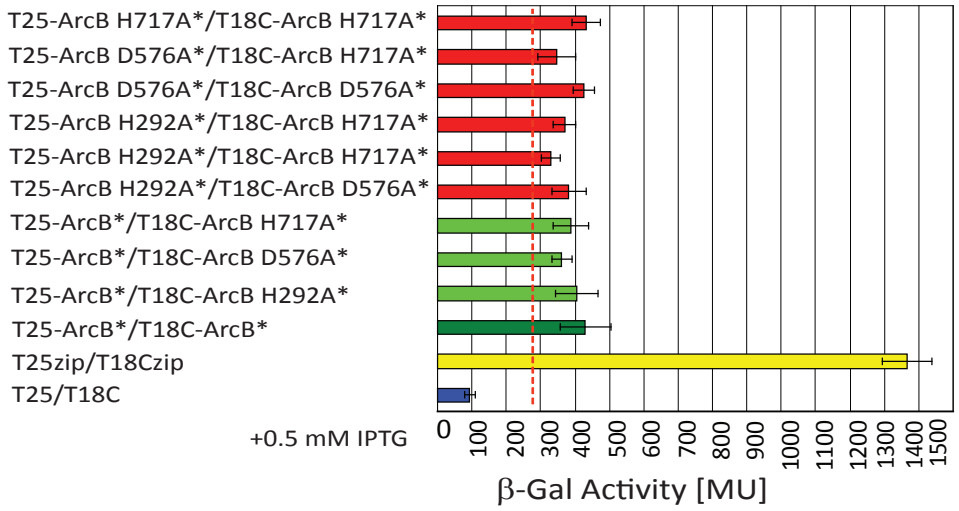


FIGURE 4.4: Phosphorelay of constitutively active ArcB variants lacking the key phosphorelay residues. The in vivo BACTH system was used to detect protein-protein interactions between the ArcB* and its variant proteins fused to T25 or T18 subunit of adenylate cyclase and expressed in the presence of 0.5 mM IPTG in a *cya*$^{-}$ BTH101 strain. For construction of fusion proteins, growth conditions and calculation of mean values with S.D. see Materials and Methods. Negative control: BTH101/pKT25+pUT18C vectors alone; Positive control: BTH101/pKT25-zip+pUT18C-zip; the levels of the interactions were quantified by β-Gal activity in liquid cultures grown micro-aerobically. Chromosomal LacZ expression 3-fold above the negative control (vectors alone) value was scored as a positive interaction signal. For all β-Gal assays, mean values of six independent assays taken from technical duplicates of three independently grown cultures of each strain were used to calculate activity. The data are shown as a mean values with SD error bars.

4.3.6 Fitting experimental data with systematic models.

We had first hypothesized that the phosphorelay process of non-orthodox TCSs should be one of either the trans-phosphorelay or cis-phosphorelay model. However, simulation results show that even after optimisation under a constrained parameter configuration, the models disagree with the experimental data at several places (figure 4.8A, B). In both the trans- and cis-phosphorelay models, the H1 $\times$ WT produces higher level of output than that of the D1 $\times$ WT or the H2 $\times$ WT, while the experimental data is the opposite. Also in trans-phosphorelay model, the H1 $\times$ H2 double mutant produces higher level of output than that of D1 $\times$ H2 and H1 $\times$ H2 double mutants, which is also opposite to the experimental data.

Thus we created a new model which we called the allosteric model. In this model, phosphorelay is dependent on the integrity of both molecules in a dimer. Details of this model were already described above. Simulation with optimised parameters shows that

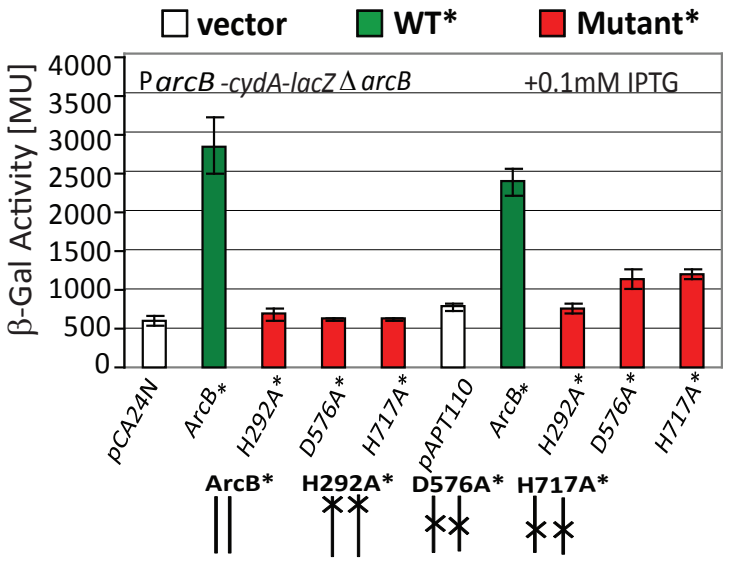


FIGURE 4.5: Phosphorelay of constitutively active ArcB variants lacking the key phosphorelay residues. Expression of a $cydA-lacZ$ chromosomal transcription fusion was measured using a β-Gal assay in $\Delta arcB$ (MVA104) cells grown in microaerobiosis in the presence of 0.1 mM IPTG co-expressing ArcB* or its variants from either pCA24N- (left hand part) or pAPT110-based (right hand part) plasmids (see schematic presentation below graph). For all β-Gal assays, mean values of six independent assays taken from technical duplicates of three independently grown cultures of each strain were used to calculate activity. The data are shown as a mean values with SD error bars.

the allosteric model fits the experimental data much better than the other two models, although it still differs from the experimental data in some ways (figure 4.8C). The values of the BIC (Schwarz criterion) of trans-, cis- and allosteric phosphorelay models are -25.63, -28.53 and -40.28, which shows that the allosteric model is the one that fits best. The resulting Schwarz weight for the allosteric model is also as high as 99% (figure not shown), which means compared to the other two models, the allosteric model has very high probability to be the model that produces the experimental data.

We also considered the model with a more flexible parameter configuration (7 parameters, as described in the Materials and Methods section) to get a better fit to the experimental data (figure 4.9). The BICs of trans-, cis- and allosteric phosphorelay models changes to -20.25, -14.04 and -31.14 (BICs do not improve because it has strict punishment for parameter numbers). The Schwarz weight of allosteric model is again much higher than for the other two models, which is 99.55% (as shown in figure 4.10).

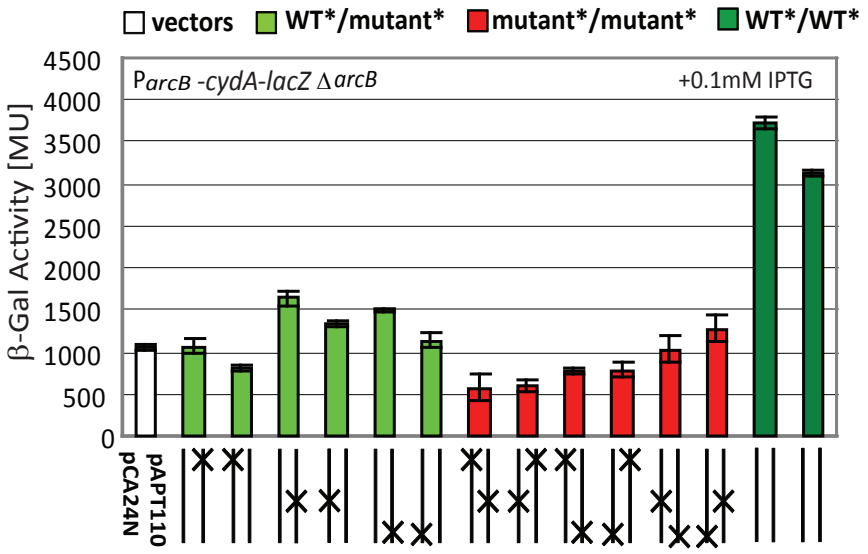


FIGURE 4.6: Phosphorelay of constitutively active ArcB variants lacking the key phosphorelay residues. As in figure 4.5 except cells expressed different combinations of the ArcB* and/or its variants from co-transformed pCA24N- and pAPT100-based plasmids (see schematic presentation below graph). For all β-Gal assays, mean values of six independent assays taken from technical duplicates of three independently grown cultures of each strain were used to calculate activity. The data are shown as a mean values with SD error bars.

4.3.7 Sensitivity assay of proposed phosphorelay models

The three phosphorelay models are subjected to the `Matlab` package `StochSens` using optimised parameters. The sensitivities of the deterministic models calculated from the FIM is shown in figure 4.11. The figure shows that all the models are much more sensitive to variation in the parameter $kf0$, which is the autophosphorylation rate. This observation is in agreement with the results we have observed in chapter 3: the steady states of the TCSs show higher sensitivity to autophosphorylation rate changes. The `StochSens` assays are performed on wild type TCS models with no autophosphorylation deficiency. In figure 4.11 we can see that the sensitivities of the three models are very similar, especially trans- and cis-phosphorelay models are almost indistinguishable. Combined with the figures 4.12, 4.13, 4.14 and 4.15, we can see that the dynamic properties of the wild type models, especially trans- and cis-phosphorelay models, are generally similar, which further confirms that the wild type trans- and cis-phosphorelay models are very hard to be distinguished.

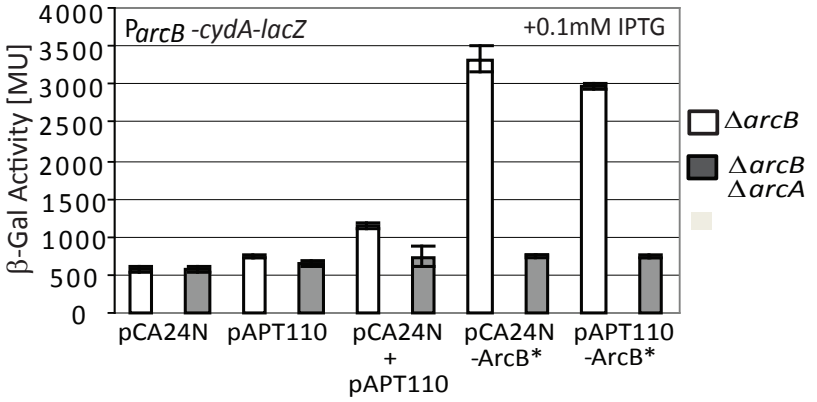


FIGURE 4.7: Phosphorelay of constitutively active ArcB variants lacking the key phosphorelay residues. Expression of a $cydA-lacZ$ chromosomal transcription fusion was measured using a β-Gal assay in either $\Delta arcB$ (MVA104) or $\Delta arcB\Delta arcA$ (MVA114) cells grown in microaerobiosis in the presence of 0.1 mM IPTG carrying either pCA24N or pAPT110 alone, both vectors alone, or expressing ArcB* from either pCA24N- or pAPT110-based plasmids. For all β-Gal assays, mean values of six independent assays taken from technical duplicates of three independently grown cultures of each strain were used to calculate activity. The data are shown as a mean values with SD error bars.

4.4 Conclusions

In the TCSs the HKs require dimerization to be active as a functioning signalling unit [165, 193]. In a HK signalling pathway, successfully sensing a specific signal is followed by autophosphorylation of the conserved histidine residue in the HisKA domain. Numerous studies on autophosphorylation have shown that most HKs autophosphorylate in an intermolecular manner, such as CheA, VirA, EnvZ and NtrB [119, 125, 165, 166, 192, 196]. The intramolecular autophosphorylation mechanism appears to occur less frequently but has been reported in some TCSs, including ArcB [131].

ArcB belongs to a relatively rare type of HK, the hybrid HK, which takes a 3-step phosphorelay instead of a one-step phosphorylation as the simple HKs. Given the fact that HKs are active in dimer form, the same question for autophosphorylation is also of interest to phosphorelay: is the phosphorelay of a certain hybrid HK happening in a intramolecular way or a intermolecular way?

We want to answer this question by fitting mechanistic models to experimental data. First we had to design experiments that can make these models to produce distinguishable results. Then deterministic models of phosphorelay of wild type ArcB were

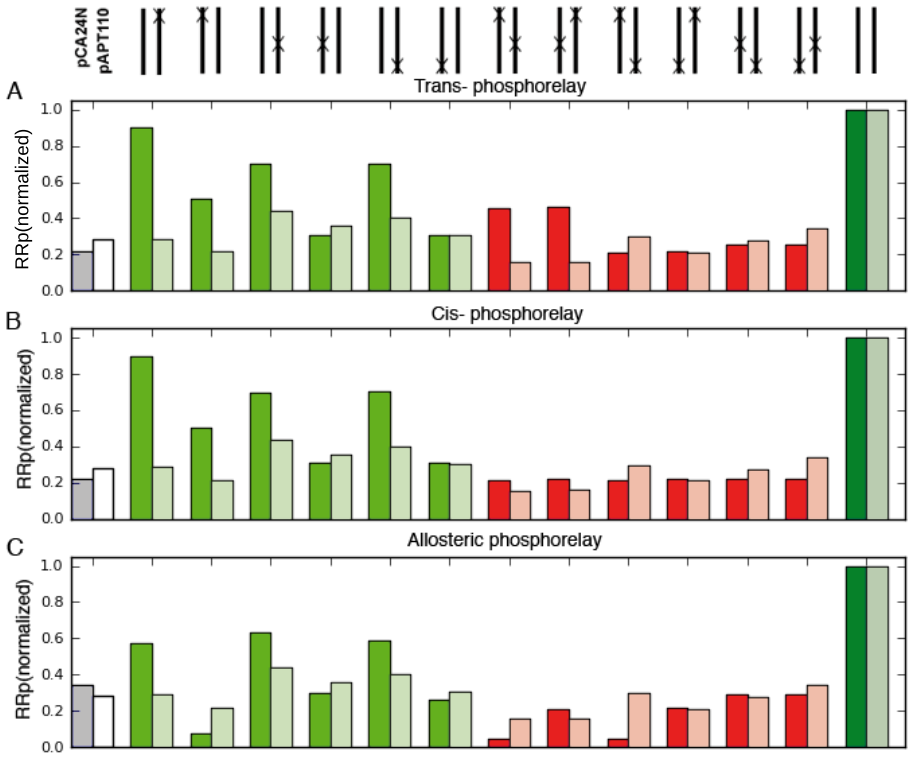


FIGURE 4.8: Simulation of three phosphorelay models with more constrained parameter settings. In these simulations, there are only 4 parameters for each of the three models, $kf, kr, kt1$ and $kt2$, as described in Materials and Methods section of this chapter. Normalised [RRp] is used as output for each model. The top legend shows the contents of the transfected plasmids in each lane. The most left lane is the negative control where only empty plasmid are transfected. In all the other lanes, the left bar stands for the constructs based on the higher copy number plasmids pCA24N, while the right bar denotes the constructs based on the lower copy number plasmids pAPT110. In each subfigures, the bars of the darker colours show the levels of the normalised [RRp] from model simulation. The bars of the lighter colours show the levels of normalised β- Gal activities of experimental data (as shown in 4.6). A, trans-phosphorelay model; B, cis-phosphorelay model; D, allosteric phosphorelay model.

constructed. The intermolecular mechanism was named trans-phosphorelay, while the intramolecular mechanism was named cis-phosphorelay. The $[RRp]$ was used as output of the models. However, by comparing the ODEs of these two deterministic models, we can see that these two models are actually all the same, except for different variable names. The `StochSensS` analysis has also confirmed that the trans- and cis-phosphorelay models are too similar and cannot be distinguished by their dynamic behaviour (figures 4.11, 4.12, 4.13, 4.14). Thus we have introduced "mutations" into the models. In our configuration, a hybrid HK is allowed to have a point mutation in one of its three conserved phosphorylation sites (His in HisKA and HPt domain, Asp in REC domain. They

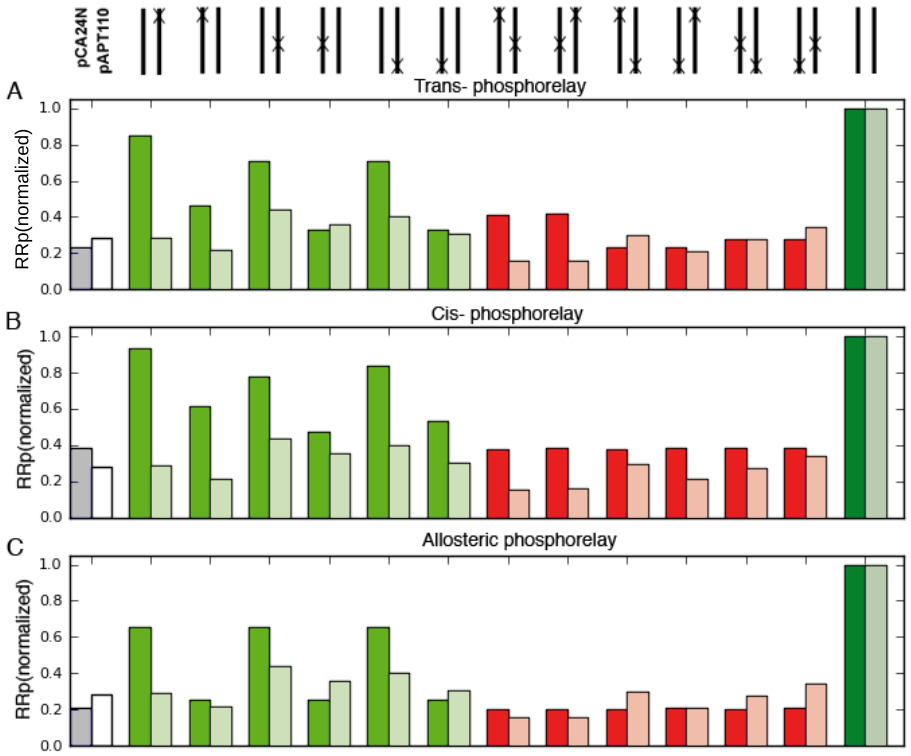


FIGURE 4.9: Simulation of three phosphorelay models with more flexible parameter settings. In these simulations, there are 7 parameters for each of the three models, $kf0, kf1, kf2, kr3, kr2, kt1$ and $kt2$, as described in Materials and Methods section of this chapter. Normalised [RRp] is used as output for each model. The top legend shows the contents of the transfected plasmids in each lane. The most left lane is the negative control where only empty plasmid are transfected. In all the other lanes, the left bar stands for the constructs based on the higher copy number plasmids pCA24N, while the right bar denotes the constructs based on the lower copy number plasmids pAPT110. In each subfigures, the bars of the darker colours show the levels of the normalised [RRp] from model simulation. The bars of the lighter colours show the levels of normalised β- Gal activities of experimental data (as shown in 4.6). A, trans-phosphorelay model; B, cis-phosphorelay model; D, allosteric phosphorelay model.

will be referred to as H1, H2 and D1 in the following text). If a site is mutated, all the reaction rates involving phosphorylation at this site will be set to 0. Three types of dimers are possible when a mutated HK is added to the wild type cells: wild type $\times$ wild type dimer, mutation $\times$ mutation dimer, and wild type $\times$ mutation dimer. The first type of dimer is the same one as in the models we have just mentioned. As for the second type of dimer, the phosphorelay reactions are completely blocked considering either trans- or cis-phosphorelay model, meaning they are still indistinguishable. As for the last type, the phosphorelay reactions are both partly blocked, thus still indistinguishable. We have come up with a more complex type of system, where the cells

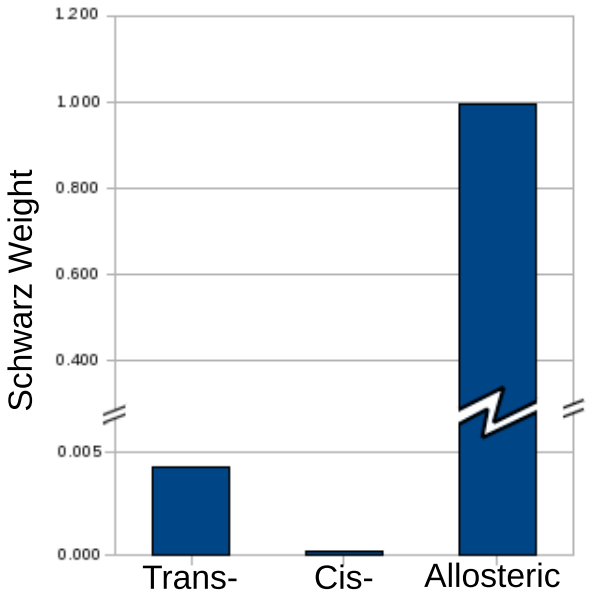


FIGURE 4.10: Schwarz Weight of three phosphorelay models with optimised parameters under more flexible parameter configuration. The probabilities are: trans-phosphorelay model, 0.43%; cis-phosphorelay model, 0.02%; allosteric, 99.55%

produce no wild type HKs but only two types of mutants. In this case, we find that the two models can finally be separated. Suppose that we have one H1 mutant mixed with a D1 mutant, homodimers will still be indistinguishable as we have just pointed out. However, the H1 × D1 heterodimer will have half of the phosphorylation reactions blocked in the trans-phosphorelay model, but are fully blocked in the cis-phosphorelay model. We can see that if the two mutant monomers in a heterodimer contains mutations that are at neighbouring positions in the phosphorelay chain (H1-D1-H2), trans- and cis-phosphorelay models will produce different output. The other candidate mutant combination is D1 × H2. Thus we have decided to design experiments with pairs of ArcB mutants.

The constitutively active ArcB mutants were constructed (shown in figure 4.2) so that the system will have a constant input signal. The plasmids that carry the mutants have different copy numbers (shown in figure 4.3) in order to bring more details to the output data. Figure 4.4 proves that the binding affinity of the mutants to each other or to the wild type proteins are rarely affected by their mutations. Thus in a system with mixed ArcBs (wild type and/or mutants), a monomer is able form a dimer with any other

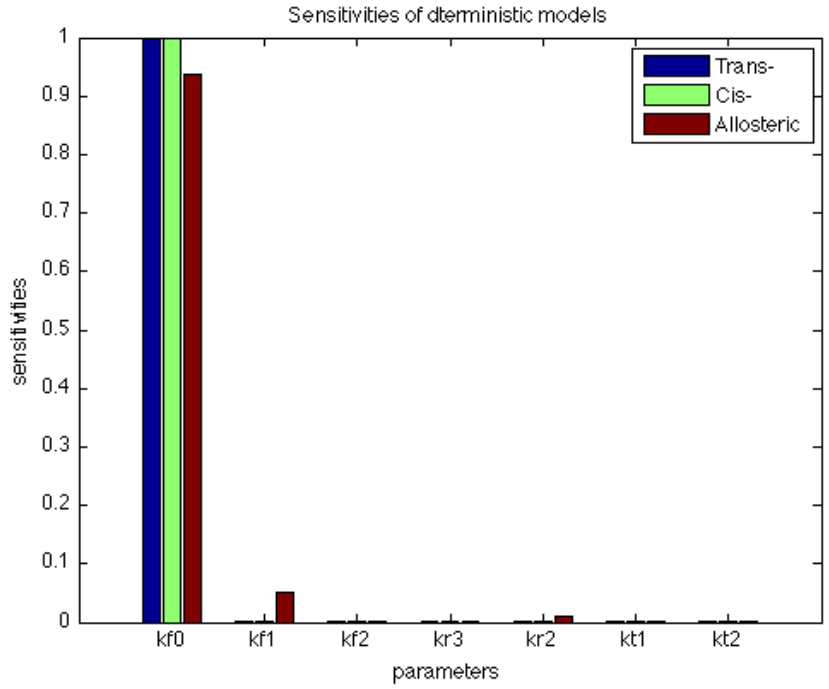


FIGURE 4.11: Sensitivity of each phosphorelay model on all the parameters. The meaning of each parameter is described in Materials and Methods section.

monomers with nearly equal probabilities.

The data used for model fitting we generated from the experiment performed with constitutively active ArcB mutants transfected into cells using plasmids with different copy numbers as shown in figure 4.6. The reason that some mutants combinations have output lower than the negative control is probably due to the phosphatase activity and the lack of phosphorylation activity of ArcB. However, we have found some inconsistency in figure 4.6 compared to our models. In our models, all the wild type $\times$ mutant combinations should have the same output level, while in the data wild type $\times$ H1 group has significantly lower output than the other two groups. This can be explained because the wild type $\times$ H1 has no mutations in the sites involved in dephosphorylation (D1, H2) which implies higher dephosphorylation activity. Another inconsistency is that the D1 $\times$ H2 group have higher output levels than the other double mutants groups. In our models this group should have the same output level as H1 $\times$ D1 in the trans-phosphorelay model or the same output level as the other two double mutants groups in the cis-phosphorelay model. This can not be explained even considering dephosphorylation (dephosphorylation would bring other contradictions). Finally, the wild type $\times$ H2

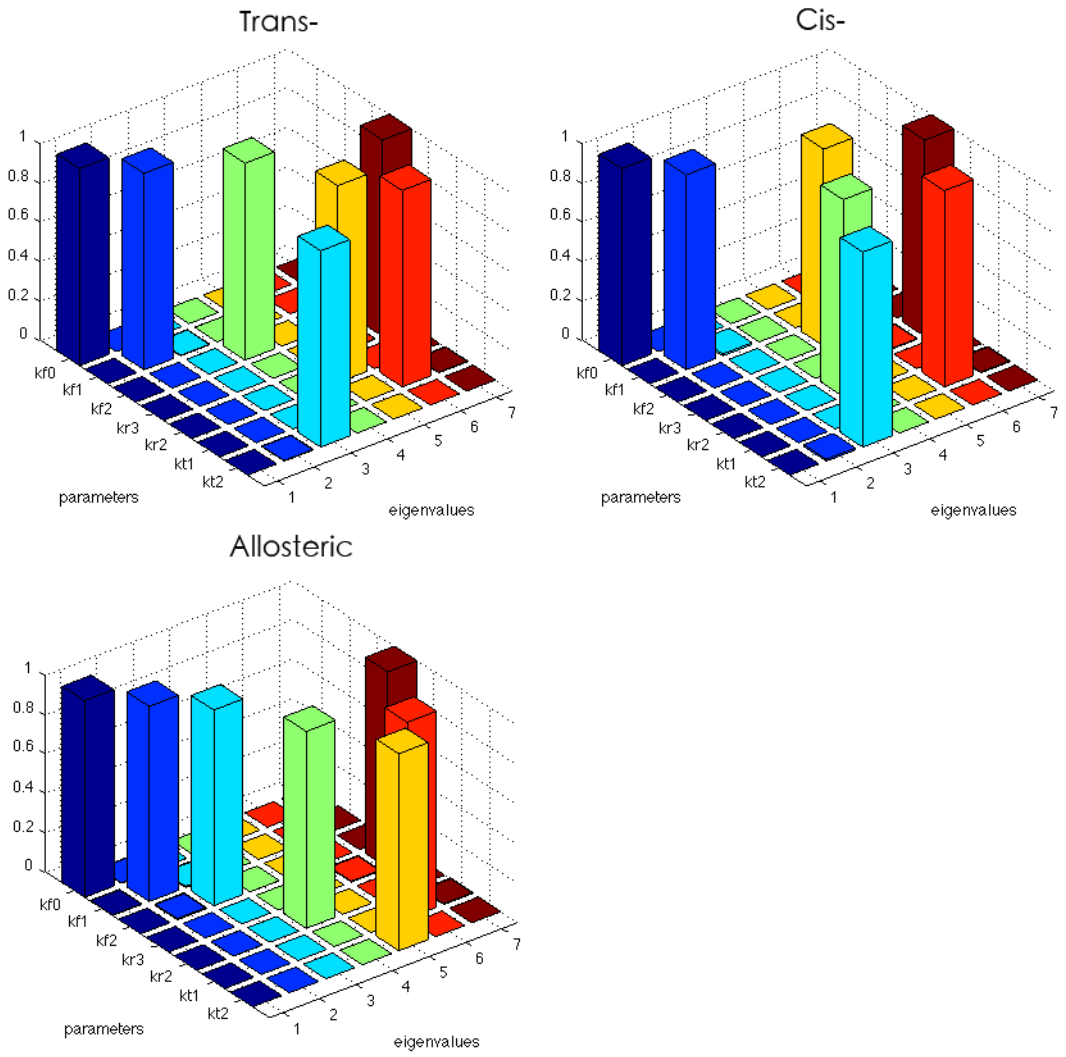


FIGURE 4.12: Contributions of individual parameters into each eigenvalue for all three models.

group shows comparable level of output as D1 × H2, while the later one should be much lower in our models.

Thus we think our hypothesis might need some modifications as the trans- and cis-phosphorelay models both fail to describe the experimental data. We therefore came up with a new model called allosteric model where for every pair of sites (*e.g.* both H1 sites in a dimer HK), both sites need to be functioning to enable the phosphorylation to take place at one of the sites. The simulations with optimised parameters (figures 4.8, 4.9) show that this leads to improved (even though not perfect) agreement between the model and the data. Schwarz weight has also shown that among the three models, the probability that the allosteric phosphorelay model describes the real ArcB/A system is

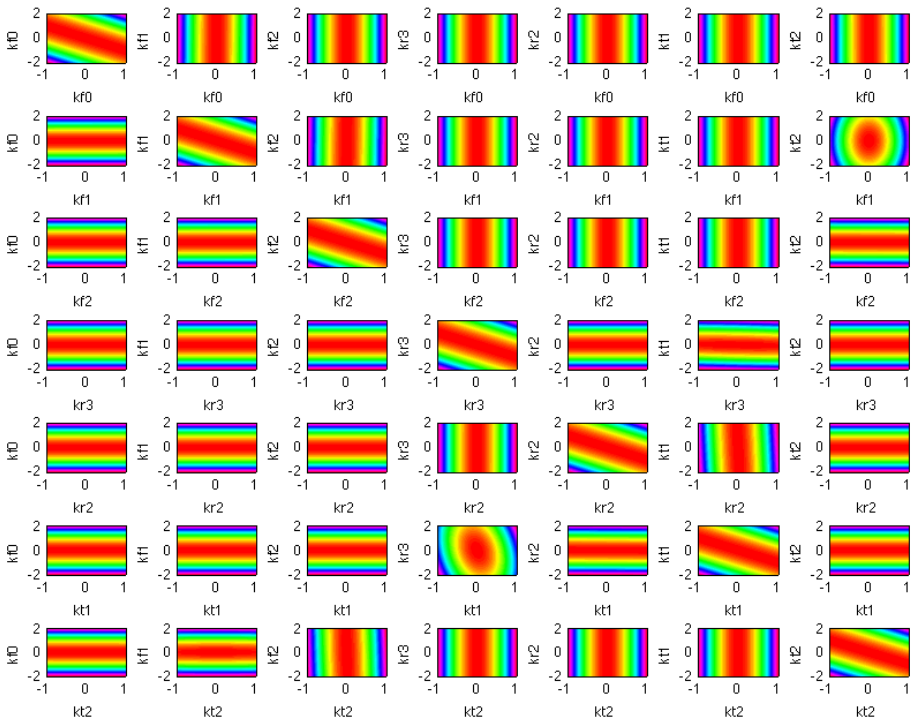


FIGURE 4.13: FIM for all pairs of parameters for trans-phosphorelay model visualised as heat maps.

over 99%. Considering that these three models describe all plausible simple phosphorelay mechanisms, we think it is reasonable to conclude that ArcB/A is employing an allosteric phosphorelay mechanism, maybe with some modifications.

In this chapter we have shown that the *in vivo* phosphorelay mechanism of ArcB/A is very likely to involve the type of allosteric phosphorelay mechanism discussed above. The phosphorelay mechanisms of the other non-orthodox TCSs need further experiments to confirm.

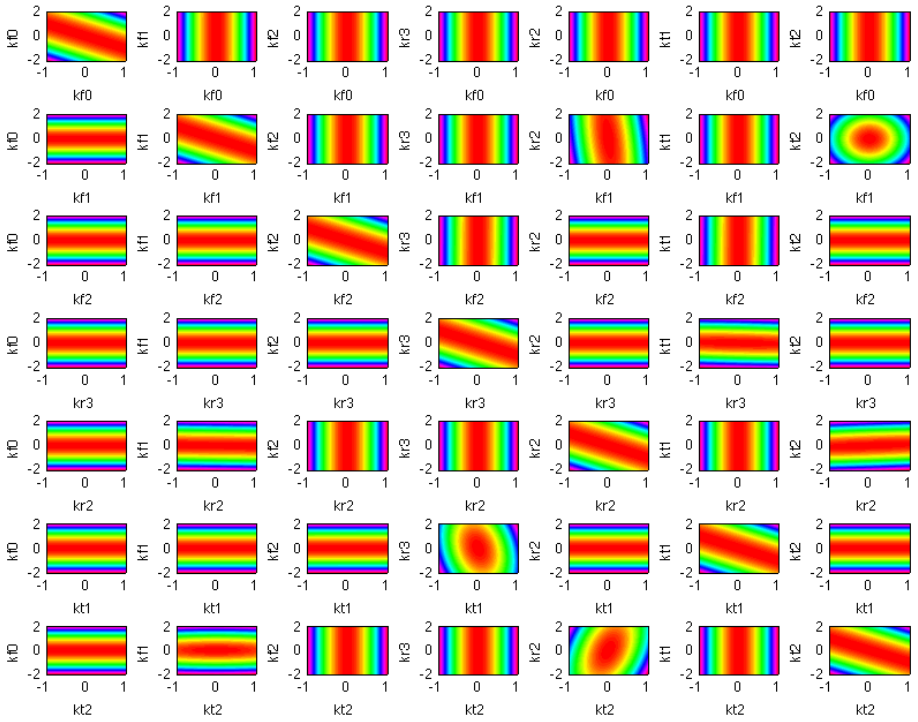


FIGURE 4.14: FIM for all pairs of parameters for cis-phosphorelay model visualised as heat maps.

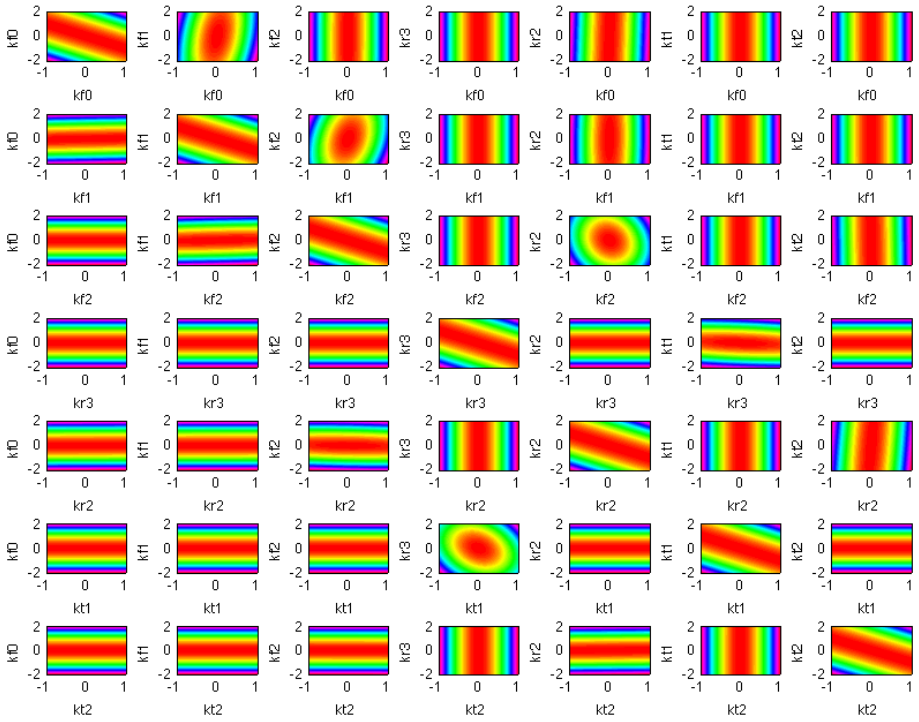


FIGURE 4.15: FIM for all pairs of parameters for allosteric phosphorelay model visualised as heat maps.

Chapter 5

Bayesian design of synthetic two-component system

The work presented here has been published in: "Bayesian design of synthetic biological systems. "

Barnes CP, Silk D, Sheng X, Stumpf MP. Proc Natl Acad Sci U S A. 2011 Sep 13;108(37):15190-5.

5.1 Background

5.1.1 Engineering methods in synthetic biology

As we are beginning to understand the mechanisms governing biological systems we are starting to identify potential ways of guiding or controlling the behaviour of cellular and molecular systems. Rationally reengineering organisms for biomedical or biotechnological purposes has become the central aim of the synthetic biology. By redirecting regulatory and physical interactions or by altering molecular binding affinities we may, for example, control metabolic processes [106, 143] or alter intra- and intercellular communication and decision making processes [84, 197]. The range of potential applications of such engineered systems is vast: designing microbes for biofuel production [44, 150] and bioremediation [29]; developing control strategies which drive stem cells through the various decisions to become terminally differentiated (or back) [57, 167], with the aim

of developing novel therapeutics [99, 101]; construction of new drug-delivery systems with homing microbes delivering molecular medicines directly to the site where they are needed [3]; use of bacteria or bacterial populations (employing swarming and quorum sensing) as biosensors [141]; and gaining better understanding of all manner of biological systems by systematically probing their underlying molecular machinery.

A range of tools and building blocks for such engineered biological systems are now available which allow us to build such systems from simple and reusable biological components [27]. In electronic systems, such modularity has been crucial and has allowed the cost-effective production of reliable components that can be combined to produce desired outputs. Biology, however, poses different and novel challenges that are intimately linked to the biophysical and biochemical properties of biomolecules and the media in which they are suspended. Especially in crowded environments such as found inside living cells the lack of insulation between different components, *i.e.*, the very real possibility of undesired cross talk, can create problems;with increasing miniaturization similar, albeit quantum effects, are now also surfacing in electronic circuits [95].

As synthetic biology gears up to bring engineering methods and tools to bear on biological problems the way in which we manipulate biological systems and processes is likely to change. Historically, each new branch of engineering has gone through a phase of what can be described as tinkering before rationally planned and executed designs became common place. Arguably, this practice is the current state of synthetic biology and it has indeed been suggested that the complexity of synthetic biological systems over the past decade has reached a plateau [138]. From the earliest days, explicit quantitative modelling of systems has been integral to the vision and practice of synthetic biology and it will become increasingly important in the future. The ability to model how a natural or synthetic system will perform under controlled conditions must in fact be seen as one of the hallmarks of success of the integrative approaches taken by systems and synthetic biology.

Here we present a statistical approach to the design of synthetic biological systems that utilizes methods from Bayesian statistics to train models according to specified input-output characteristics. It incorporates modelling and automated design and is general in the sense that it can be applied to any system that can be described by a mathematical

model which can be simulated. Because of the statistical nature of this approach, previously challenging problems such as handling stochastic models, accounting for kinetic parameter uncertainty, and incorporating environmental stochasticity can all be handled in a straightforward and consistent manner.

5.1.2 Bayesian approach in system design

The question of how to design a system to perform a specified task can be viewed as an analogue to reverse engineering. In design we want to elucidate the most appropriate system to achieve our design objectives; in reverse engineering we aim to infer the most probable system structure and dynamics that can give rise to some observational data. In this respect, the design question can be viewed as statistical inference on data we wish to observe.

In the Bayesian approach to statistical inference the posterior distribution is the quantity of interest and is given by the normalized product of the likelihood and the prior. In most practical applications the posterior distribution cannot be derived analytically, but if the likelihood (and prior) can be expressed mathematically we can use Monte Carlo methods to sample from the posterior. In many cases where the model structure is complex the likelihood cannot be written in closed form and traditional Monte Carlo techniques cannot be applied. These include inference for the types of stochastic processes encountered in systems and synthetic biology. In these cases a family of techniques known collectively as approximate Bayesian computation (ABC) can be applied: ABC uses model simulations to approximate the posterior distribution directly. Here we use a sequential Monte Carlo ABC algorithm known as ABC SMC to move from the prior to the approximate posterior via a series of intermediate distributions [177]. This framework can also be used to perform Bayesian model selection [175] and has been implemented in the software package ABC-SysBio [96].

Figure 5.1 depicts the approach presented here. The design objectives are first specified through input-output characteristics. Here these have been depicted as a single time series, though the method can be applied in a much broader sense with multiple inputs and outputs. A set of competing designs is then specified through deterministic or stochastic models, each containing a set of kinetic parameters and associated prior distributions. The distance function measures the discrepancy between the model output

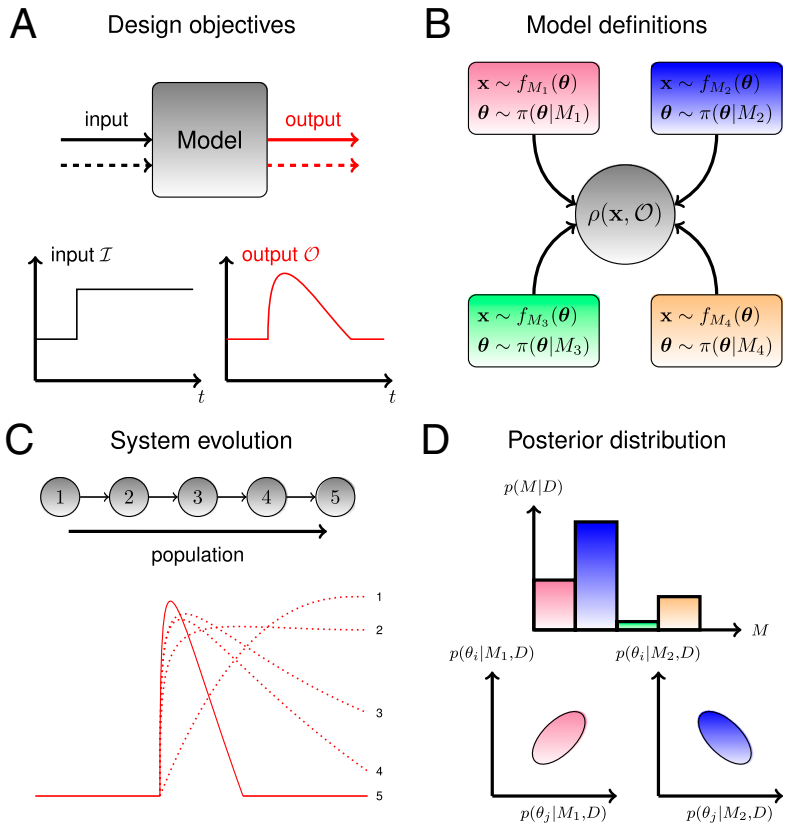


FIGURE 5.1: Bayesian approach to system design. A, The design objectives are encoded by the specification of input and output characteristics; (B) One or more competing designs for the system are specified together with priors on the parameters. A distance function, $\rho(\mathrm{x}, \mathcal{O})$, relates model output, x, to the desired output characteristic, $\mathcal{O}$; C, The system is evolved using sequential Monte Carlo. Each population more accurately approximates the desired behaviour; D, The model posterior probability encodes the ability of each design to achieve the desired behaviour. The parameter posterior shows para- meters that are sensitive or insensitive to the input-output specification.

and the objective. In principal it is possible to specify a distribution over the objective and each model could also contain experimental error. The ABC SMC algorithm then automatically evolves the set of models toward the desired design objectives. The results are a set of posterior probabilities representing the probability for each design to achieve the specified design objectives in addition to the posterior probability distribution of the associated kinetic parameters. This approach is similar in spirit to some existing methods for the automated design of genetic networks such as those adopting evolutionary algorithms [23, 45], Monte Carlo methods [17, 41], or optimization [16, 39, 145] but the advantages of our method over traditional ones are that we can utilize powerful concepts from Bayesian statistics in the design of complex biological systems, including

- the rational comparison of models under parameter uncertainty using Bayesian model selection, which automatically accounts for model complexity (number of parameters) and robustness to parameter uncertainty;
- a posterior distribution over possible design parameter values that can be analysed for parameter sensitivity and robustness and provide credible limits on design parameters;
- the treatment of stochastic systems at the design stage including the design of systems with required probability distributions on system components; and
- methods for the efficient exploration of high-dimensional parameter space.

In the following we demonstrate the power of this approach by examining, from this unique perspective, systems that have been of interest in the recent literature. We look at the ability of two bacterial TCS topologies to achieve four types of particular input-output behaviours.

5.1.3 Tow-component systems in this study

TCSs allow bacteria to sense external environmental stimuli and relay information into the cell, *e.g.*, to the gene expression apparatus as already discussed at length in the previous chapters. The TCSs that we are studying in this chapter are the orthodox TCS which applies simple phosphorylation process and non-orthodox TCS which performs a phosphorelay upon stimulation. The detailed structures and phosphorylation processes of these two TCSs are described in the previous chapter (chapter 1). The reactions involved are depicted in figure 5.2A and B. Here we have applied the Bayesian approach to directly compare the ability of orthodox and non-orthodox designs to achieve various input-output behaviours, using ODE models similar to ones described previously [82] (see Methods section). Below we discuss the approximate Bayesian computation (ABC) framework in the necessary detail to understand how we can employ Bayesian ideas in the design of synthetic TCSs.

5.2 Methods

5.2.1 Bayesian design

5.2.1.1 Model selection using ABC

In Bayesian inference comparison of a discrete set of models can be be performed using the marginal posterior. Consider the joint space defined by $(M, \theta) \in \mathcal{M} \times \Theta_{\mathcal{M}}$; Bayes theorem can then be written

$$\pi(M \mid y) = \frac{f(y \mid M)\pi(M)}{\int_{\mathcal{M}} f(y \mid M')\pi(M')dM'} = \frac{f(y \mid M)\pi(M)}{\sum_{\mathcal{M}} f(y \mid M')\pi(M')dM'},$$

where $f(y \mid M)$, the marginal likelihood, can be written

$$f(y \mid M) = \int_{\Theta_{\mathcal{M}}} \pi(\theta \mid M) f(y \mid \theta, M) d\theta.$$

Therefore the posterior probability of a model is given by the normalized marginal likelihood which may or may not be weighted depending on whether the prior over models is informative or uniform respectively. It has recently been noted that model selection using summary statistics can be problematic because the summary statistic must be sufficient for the joint space, $\{M, \theta\}$, rather than just θ [144]. This is not a concern here since in all our examples we use the full data set with no summary or we define our posterior distributions through the summary statistics.

Model selection can be incorporated into the ABC framework by introducing the model indicator M and proceeding with inference on the joint space. For example, the ABC rejection algorithm with model selection [11] proceeds as follows

MR1 Sample M^* from $\pi(M)$.

MR2 Sample θ^* from $\pi(\theta \mid M^*)$.

MR3 Simulate a dataset x^* from $f(x \mid \theta^*, M^*)$.

MR4 If $\rho(x^*, y) \leq \epsilon$ accept (M^*, θ^*), otherwise reject.

MR5 Return to R1.

Once N samples have been accepted, an approximation to the marginal posterior, $\pi(M = m \mid y)$, is given by

$$\pi(M = m \mid y) = \frac{\#\text{accepted}\, m}{N}.$$

Model selection can also be incorporated into the ABC SMC algorithm [175]. To obtain N samples $\{(M,\theta)^1,(M,\theta)^2,(M,\theta)^3...,(M,\theta)^N\}$ from the posterior, defined as, $\pi(M,\theta \mid \rho(x^*,y) \leq \epsilon_T)$, proceed as follows

MS1 Initialize $\epsilon = \infty$

Set the population indicator $t = 0$

MS2.0 Set the particle indicator $i = 1$

MS2.1 If $t = 0$, sample (M^{**},θ^{**}) from the prior $\pi(M,\theta) = \pi(M)\pi(\theta \mid M)$.

If $t > 0$, sample M^* with probability $P_{t-1}(M^*)$ and perturb $M^{**} \sim KM_t(M \mid M^*)$.

Sample θ^* from the previous population $\{\theta(M^{**})_{t-1}\}$ with weights w_{t-1}.

Perturb the particle, $\theta^{**} \sim K_{t,M^{**}}(\theta \mid \theta^*)$ where $K_{t,M}$ is the perturbation kernel.

If $\pi(M^{**},\theta^{**}) = 0$, return to MS2.1

Simulate a candidate dataset $x^* \sim f(x \mid M^{**},\theta^{**})$.

If $\rho(x^*,y) > \epsilon$ return to MS2.1

MS2.2 Set $(M,\theta)_t^i = (M^{**},\theta^{**})$ and $d_t^i = \rho(x^*,y)$, calculate the weight as

$$w_t^i(M_t^i,\theta_t^i) = \begin{cases} 1 & \text{if } t = 0 \\ \frac{\pi(M_t^i,\theta_t^i)}{S_1 S_2} & \text{if } t > 0 \end{cases}$$

where

$S1 = \sum_{j\in\mathcal{M}} P_{t-1}(M_{t-1}^j) KM_t(M_t^i \mid M_{t-1}^j)$

and

$S_2 = \sum_{k\in M_t^i = M_{t-1}} \frac{w_{t-1}^k K_{t,M^i}(\theta_t^i \mid \theta_{t-1}^k)}{P_{t-1}(M_t^i = M_{t-1})}$

If $i < N$, set $i = i + 1$, go to MS2.1

S3 Normalize the weights.

Obtain the marginal model probabilities given by

$P_t(M_t = m) = \sum_{k\in M_t^i = M_{t-1}} w_t^i(M_t^i,\theta_t^i)$

Determine ϵ such that $Pr(d_t \leq \epsilon) = 0.9$.

If $\epsilon > \epsilon_T$, set $t = t + 1$, go to MS2.0.

There are two obvious additions to the algorithm when compared to parameter inference. The model kernel, KM_t, perturbs the resampled models using a multinomial distribution, and the additional term in the weight denominator accounts for the probability of observing the current model given the previous population. This is described in detail in

Toni and Stumpf's paper [175] as well as in the paper which resulted from the research carried out and presented here [11].

5.2.1.2 Prior distribution

The prior distribution encodes our knowledge of the system and should be set according to known biochemical properties. However, often the kinetic parameters are not well known and can be very difficult or even impossible to measure *in vivo*. In these cases we make the prior distribution non informative by specifying a large range over possible, biophysically and biochemically plausible values. As more information becomes available, through experimental studies or otherwise, the prior can be updated to reflect our increased knowledge of the system. Interestingly, for some systems, our design method could help to constrain kinematic parameters where experimental data are unavailable.

5.2.1.3 The distance function and output tolerance

In system design we would rarely insist on achieving the true posterior distribution corresponding to $\epsilon = 0$, but would like to reach the objective within some tolerance. A theorem due to Wilkinson (2008) [187] states that if we assume that the data can be considered as

$$y = \eta(\hat{\theta}) + e,$$

where $\eta(\hat{\theta})$ is a draw from the model at the 'best' input and e is an additive, independent error, then the approximate posterior distribution, $\pi(\theta \mid \rho(x^*, y) \leq \epsilon)$ can be interpreted as the 'true posterior $\pi(\hat{\theta} \mid y)$. While the independence assumption is not always true, this theorem provides some insight into the relationship between the final ϵ value and the tolerance on our specified behaviour. For example when using uniform kernels, as in this study, if our desired output behaviour is a constant of 0.5 and we finish the inference at $\epsilon = 0.05$ our final trajectories will be distributed $U(0.45, 0.55)$ giving a tolerance of $\pm 10\%$ on the output behaviour. This can be used when considering our desired output objectives. To achieve other error distributions, such as Gaussian errors, we can always explicitly specify the error model in the design objectives.

5.2.1.4 Deterministic models

Inference for deterministic models such as ordinary differential equations can be problematic since there is a one to one relationship between the parameter vector θ and the data set x. Therefore, in the absence of observational error, the posterior distribution resembles a delta function, $\delta(\theta - \hat{\theta})$ where $\hat{x} = f(\hat{\theta})$ is data 'closest to y. An additional problem for ABC methods is that the minimum distance, $\rho(\hat{x}, y)$, is greater than zero [176]. However, in practice, observational data have associated experimental errors and when this is included explicitly in the model, the problem is resolved. In the case of systems design, we omit the explicit error model for clarity, but note that it could be included with assumptions on the form of the distribution.

5.2.2 Modelling two component systems

5.2.2.1 Models

The models we used were based on the ones found in [82]. S is the input stimulus of the system. The concentration of phosphorylated RR is used as the output of the system. All simulations were performed in the range $0 \leq t \leq 10$.

Orthodox system

We modelled the following reactions

$$\begin{aligned} HK + S &\xrightarrow{k_1} HKp + S \\ HKp + RR &\xrightarrow{k_2} HK + RRp \\ HKp &\xrightarrow{k_3} HK \\ HK + RRp &\xrightarrow{k_4} HKp + RR \\ RRp &\xrightarrow{k_5} RR \end{aligned}$$

Additionally we assumed that the total concentration of $HK_{tot} = HK + HKp$ and $RR_{tot} = RR + RRp$ were equal to one. This resulted in the following ordinary differential

equations

$$\begin{aligned}\frac{d[HK]}{dt} &= k_2[HKp][RR] + k_3[HKp] - k_4[HK][RRp] - k_1[HK][S] \\ \frac{d[RRp]}{dt} &= k_2[HKp][RR] - k_4[HK][RRp] - k_5[RRp].\end{aligned}$$

Non-orthodox system

We labelled the occupied states of the phosphorelay as

	H1	D1	H2
HK_1	x	x	x
HK_2	o	x	x
HK_3	x	o	x
HK_4	x	x	o
HK_5	o	o	x
HK_6	o	x	o
HK_7	x	o	o
HK_8	o	o	o

where H1, D1 and H2 are the binding domains on the Histidine Kinase and x, o represent an empty, occupied domain respectively. We modelled the following reactions

$$HK + S \xrightarrow{k_1} HKp + S$$

Again we assumed that the total concentration of $HK_{tot} = \sum HKi$ and $RR_{tot} = RR + RRp$ were equal to one. This resulted in the following ordinary differential equations

$$\begin{aligned}
\frac{d[HK_1]}{dt} &= k_4[HK_4][RR] + k_6[HK_3] - k_7[HK_1][RRp] + k_8[HK_2] - k_1[HK_1][S] \\
\frac{d[HK_2]}{dt} &= k_4[HK_6][RR] + k_6[HK_5] - k_7[HK_2][RRp] - k_8[HK_2] + k_1[HK_1][S] - k_2[HK_2] \\
\frac{d[HK_3]}{dt} &= -k_3[HK_3] + k_4[HK_7][RR] + k_5[HK_4] - k_6[HK_3] - k_7[HK_3][RRp] + k_8[HK_5] \\
&\quad -k_1[HK_3][S] + k_2[HK_2] \\
\frac{d[HK_4]}{dt} &= k_3[HK_3] - k_4[HK_4][RR] - k_5[HK_4] + k_6[HK_7] + k_7[HK_1][RRp] + k_8[HK_6] \\
&\quad -k_1[HK_4][S] \\
\frac{d[HK_5]}{dt} &= -k_3[HK_3] + k_4[HK_8][RR] + k_5[HK_6] - k_6[HK_5] - k_7[HK_5][RRp] - k_8[HK_5] \\
&\quad +k_1[HK_3][S] \\
\frac{d[HK_6]}{dt} &= k_3[HK_5] - k_2[HK_6] - k_4[HK_6][RR] - k_5[HK_6] + k_6[HK_8] + k_7[HK_2][RRp] \\
&\quad -k_8[HK_6] + k_1[HK_4][S] \\
\frac{d[HK_7]}{dt} &= k_2[HK_6] - k_4[HK_7][RR] - k_6[HK_7] + k_7[HK_3][RRp] + k_8[HK_8] - k_1[HK_7][S] \\
\frac{d[RRp]}{dt} &= k_4[RR]([HK_4] + [HK_6] + [HK_7] + [HK_8]) - k_7[RRp]([HK_1] + [HK_2] + [HK_3] \\
&\quad +[HK_5]) - k_9[RRp]
\end{aligned}$$

5.2.2.2 Distance

The distance functions for input-output behaviours $\rho(x, \mathcal{O})_{1-4}$ were defined to be

$$\begin{aligned}
\rho(x, \mathcal{O})_1 &= \{\mathbb{H}(0)(\text{argmax}_t x_t - 2.0), \mathbb{H}(0)(\text{argmin}_t x_t - 4.0)\} \\
\rho(x, \mathcal{O})_2 &= \sqrt{\sum_t (x_t - 1.0)^2} \\
\rho(x, \mathcal{O})_3 &= \sqrt{\sum_t (x_t - 0.5)^2} \\
\rho(x, \mathcal{O})_4 &= \{\rho(x, \mathcal{O})_1, \mathbb{H}(0)(1 - \text{max}x_t - 0.2), \mathbb{H}(0)(\text{min}x_t - 0.2)\}
\end{aligned}$$

where $\mathbb{H}(0)$ is the Heaviside function ensuring the distance is positive.

5.2.2.3 Priors

The priors on all variables were distributed as $U(0, 1000)$.

5.3 Results

Figures 5.2 C-F show four types of behaviour that may be desired in synthetic TCS systems (*e.g.*, for bioremediation or biopharmaceutical applications), and the corresponding posterior probabilities of the orthodox and non-orthodox models to achieve them. In Figure 5.2C the specified behaviour (output, $[RRp]$, blue dashed lines) is that of a fast response to a square pulse input signal (S, red solid line). That is the output should show a maximum within 0.1 s after the pulse starts and should return to its minimum within 0.1 s after the pulse ends. Different from the experiment in figure 5.2F, there is no constraint for the value of the output level here. As can be seen from the posterior probabilities, both models achieve this behaviour easily, as one would expect from a signalling system, with the orthodox system slightly outperforming the non-orthodox system. In Figure 5.2D the ability of the two systems to achieve a steady output state at $t > 2$ s under a constant input signal is examined. The output level is desired to stabilize at high level after stimulated for 2 s. And again both systems perform comparably with the orthodox system appearing slightly more favourable.

In figure 5.2E the input signal is a high-frequency sinusoid with a mean of 0.6, and the desired output is a constant signal with the same mean and root mean square < 0.3 (which means we want the output to be as stable as possible); this behaviour would mimic a system that is robust to high-frequency noise. The output trajectories at some intermediate and final populations are shown in figure 5.3. The system takes more than 100 intermediate populations to evolve parameters posterior distribution producing satisfying outputs. In this example the non-orthodox system clearly outperforms the orthodox system (figure 5.2E), which indicates the increased robustness to noisy signals that comes with the relay architecture [82]. But the direct comparison of the two models' ability to cope with noise, which is becoming possible in this approach, also reveals some unexpected characteristics: Inspection of the posterior distribution, which is the distribution of parameters that produce desired outputs (figure 5.4), shows that all the dephosphorylation reaction rates, $k_6; k_8; k_9$, are minimized whereas the rate of

the signal induced autophosphorylation (k_1) is large. We can see that in figure 5.4, in all the contour maps involving $k_6; k_8; k_9$, the distributions are compressed at the lower end of the axises; while for the contour maps involving k_1, the distributions are near the upper side of the axises. The distribution plots in the diagonal subfigures shows the same pattern. Thus the noise reduction mechanism in the non-orthodox system works by saturating the system.

In figure 5.2F the input is again a step function but the output is more specific; it must reach to > 0.8 within 0.5 s of the pulse start and drop to < 0.2 within 0.5 s of the pulse end, thus approximately reproducing the input (Figure 5.5 shows the evolution of the system in this case). Here the orthodox model clearly outperforms the non-orthodox model. Inspection of the posterior distribution (figure 5.6) shows that both the rate of the signal induced autophosphorylation and the rate of phosphorylation of the response regulator by the histidine kinase, $k_1; k_2$, are large whereas the rate of dephosphorylation of the response regulator, k_5, is small. We can see in figure 5.6 that for the contour maps involving k_5, the distributions are compressed at the lower end of the axises; while for the contour maps involving $k_1; k_2$, the distributions are near the upper side of the axises. Diagonal distribution plots show the same behaviour. These properties ensure that the shape of the signal is transferred faithfully through the system.

5.4 Conclusions

In this chapter we have presented a novel method for the design of synthetic biological systems employing ideas from Bayesian statistics. We have demonstrated its utility in bacteria two-component system. Its application in more biological systems can be seen in the associated published paper [11]. This method has advantages over traditional design approaches in that the modelling and model evaluation/characterization is incorporated directly into the design stage. The statistical nature of the method has many attractive features including the handling of stochastic systems, the ability to perform model selection and the handling of parameter uncertainty in a well-defined manner. Model selection using summary statistics can be problematic in other settings [144] but is not a concern here because we either use the full dataset with no summary or we define our posterior distributions through the desired system outputs. We used the `ABC-SysBio` and `cudasim` softwares [96, 199], which takes as input a set of SBML files and as such can

be used by bioengineers and experimentalists to rationally com- pare their competing designs for a system. By using this method we hope that the implementation time of synthetic systems can be reduced by defining a program of experimental work based on the posterior probabilities of each design.

Monte Carlo sampling of parameter spaces has been used to assess the robustness of engineered and biological systems in the past [56, 79, 202]. But like in the statistical case, simple Monte Carlo sampling tends to waste too much effort and time on those regions which are of no real interest for reverse engineering or design purposes. Our statistically based sequential approach homes in onto those regions where the probability of observing the desired behaviour is appreciable, which allows us a more nuanced comparative assessment of different design proposals, especially when dynamics are expected (or indeed desired) to exhibit elements of stochasticity. And the Bayesian model selection approach automatically strikes a balance between the systems abilities to generate the desired behaviour effectively but also robustly.

Further developments will include the incorporation of methods for model abstraction to reduce computation time [116] and to handle a database of standard parts as in other existing design software systems [59, 105]. Moreover, it is also possible to include the generation of novel structures (by, *e.g.*, using stochastic context-free grammars [8] to propose alterations to a reaction/ interaction network) as part of the design process. Just like in the case where ideas from control engineering and statistics can gainfully be combined to reverse-engineer the structure of naturally evolved biological systems, we feel that in the design of synthetic systems such a union will also be fruitful.

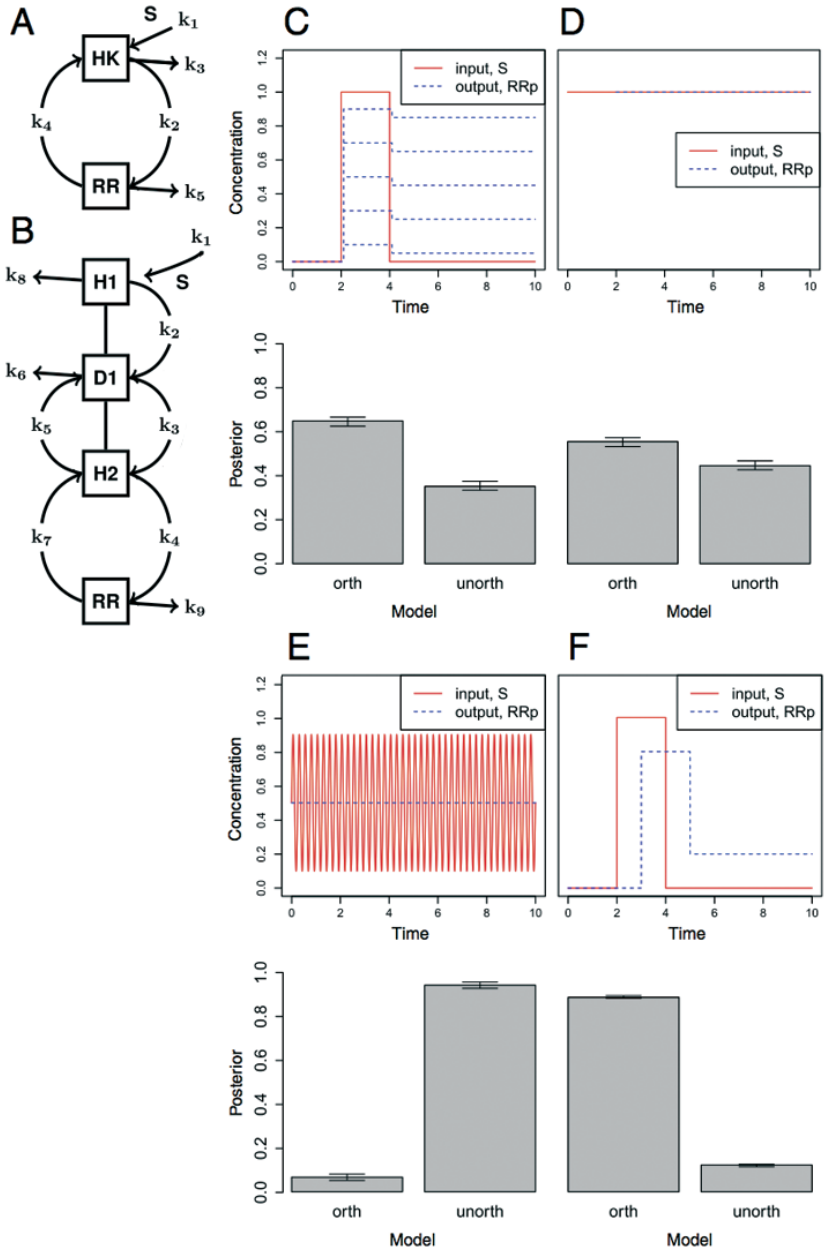


FIGURE 5.2: Bacterial two-component systems. A, Orthodox system where HK denotes the histidine kinase and RR the response regulator, both of which have phosphorylated forms, HKp and RRp. Arrows represent reactions involving phosphate groups and the k_i represent the rate parameters. S is the input stimulus signal that causes autophosphorylation of the histidine kinase; B, Phosphorelay system with three phosphate-binding domains, where H, D refer to histidine and aspartate domains, respectively. (C-F) Specified input-output behaviour (upper) and posterior probabilities for the two designs to achieve it (lower). The input signal corresponds to the stimulus, S, and the output signal is represented by the concentration of phosphorylated response regulator, RRp. The error bars indicate the variability in the marginal model posteriors over three separate runs.

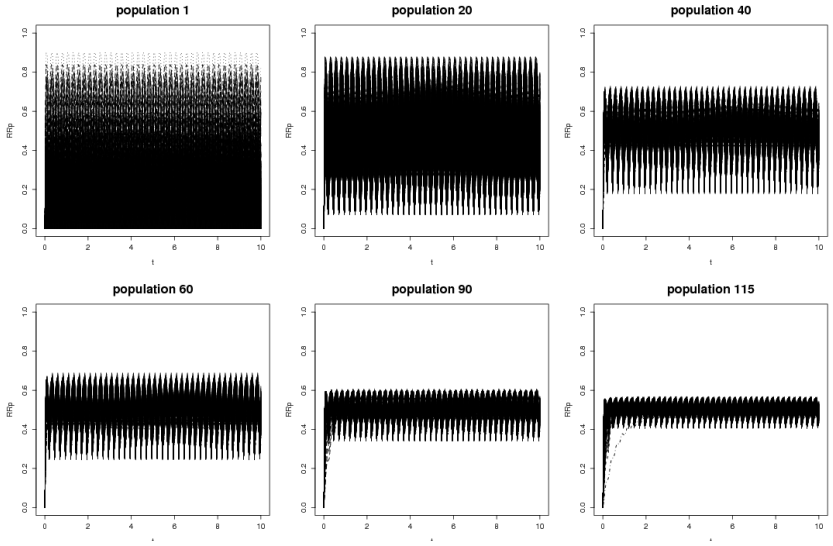


FIGURE 5.3: Two component systems: evolution to the noise reduction behaviour.

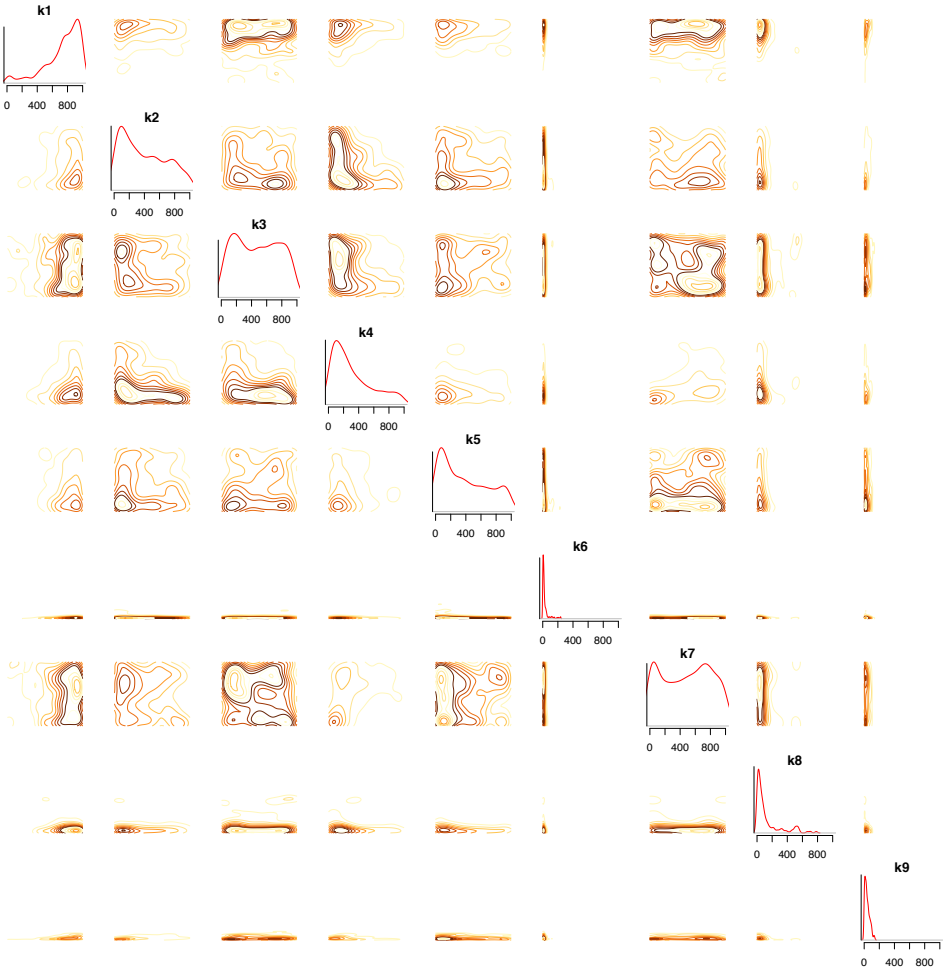


FIGURE 5.4: Two component systems: posterior distribution for the non-orthodox system to achieve the noise reduction behaviour.

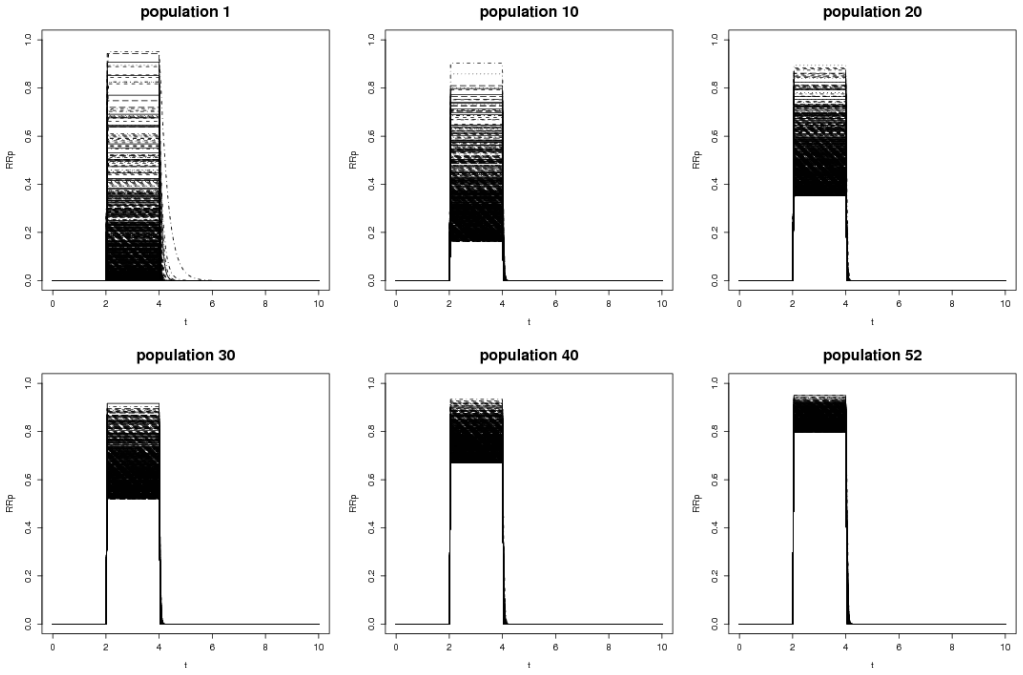


FIGURE 5.5: Two component systems: evolution to the signal reproduction behaviour.

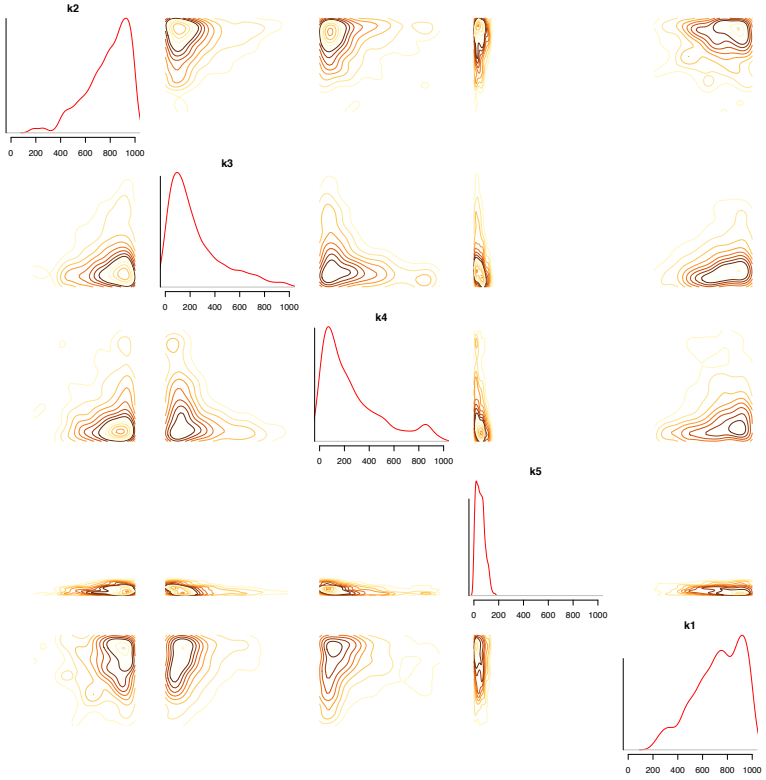


FIGURE 5.6: Two component systems: posterior distribution for the orthodox system to achieve the signal reproduction behaviour.

Chapter 6

Conclusions

In this thesis, we have studied the evolutionary and functional aspects of two-component systems. Some of the results have confirmed the findings of other scholars, while others of the results have improved our knowledge of the TCSs. We have also applied some novel methods to the analysis of the TCSs.

Although the TCS proteins mostly function in a one-to-one pattern and have very high specificity to their cognate partners, cross-talk with non-cognate TCS still exists. The histidine kinase CheA can react with either response regulator CheB or CheY [64]. And a number of TCSs in *E. coli* are found to be able to react with non-cognate partners *in vitro* [194]. In chapter 2, by comparing the orthologues of *E. coli*'s TCSs across all bacterial species, we found that the composition of TCSs in each species is quite flexible. And in many species a pair of TCS do not even coexist (nor are they both absent, figure 2.3, 2.4, 2.5). This lead us to the conclusion that TCSs might be able to cross-talk with non-cognate partners in the species where their partners are missing. In figure 2.6, a large proportion of the TCSs are colocalizing in some species but are separated in others. This suggests that TCS genes are able to change position on the chromosome during evolution. When this happens, the regulation pattern changes for this TCS. And hybrid HKs might also have evolved in this way.

From figure 2.6, we have also found that the non-orthodox TCSs' coding genes tend to localize separately on the chromosome. However, in orthodox TCSs, the HKs' and RRs' coding genes tends to locate close to their cognate partners' genes. One possible

mechanism for the evolution of hybrid HKs is that they might be a consequence of gene fusion of simple HKs [34]. As the gene fusion could cause the gene to translocate to a different position on the chromosome, this hypothesis is supported by the fact that non-orthodox TCSs have a higher chance of not being colocalized (figure 2.6).

Because TCS pairs are not always located in the same operon on the chromosome, the pairs are not always regulated together. In chapter 3, we have come to the conclusion that the post-translational reactions of TCSs are monostable. The TCS will be bistable if it has a positive feedback loop in its transcriptional regulatory circus. The auto regulation of RR seems to be the key to the TCS's bistability, if it exists. There are scenarios where this may be advantageous and our analysis has mapped out the scenarios when this can occur.

The analysis of steady states shows that just like the orthodox TCS, the non-orthodox TCS has only one stable steady state (monostable) in its post-translational signalling process. Figure 3.4 shows that the steady states of orthodox and non-orthodox systems change differently with the parameter. The steady states of orthodox and non-orthodox are both more sensitive to stimuli than any of the other parameters. And the steady state of the non-orthodox system appears to be more robust to all the parameter perturbations.

Our results in figures 2.3, 2.4 and 2.5 have further confirmed the possibility of cross-talk. In chapter 3, we have studied the steady states and dynamic behaviour of cross-talking TCSs. Numerical simulation shows that just like a sole TCS, the cross-talking TCS is also monostable when we consider post-translational chemical reactions. By comparing the steady states of the simple TCS and the cross-talking TCSs, we found that the cross-talking model can integrate the signal of the two constituent TCSs to achieve a much balanced steady state (figure 3.6) regarding to the parameter changes. Study of the transient [RRp] level shows that cross-talking TCSs can be more robust in scenarios with contradictory stimuli. And it can achieve higher levels of response to co-activating stimuli (figure 3.5).

In chapter 4, we have built several ODE models to understand the phosphorelay mechanism of the non-orthodox TCSs. The phosphorelay defines the phosphorylation process taking place in the non-orthodox TCS where the phosphate group transfers from H1 to D1 to H2 domains after autophosphorylation. What we were interested to know is the mechanism of phosphorelay in a HK dimer of the non-orthodox TCS system (ArcB/A in this thesis). Based on previous studies, we thought that there might most probably be two mechanisms of the phosphorelay. In the cis-phosphorelay model, the phosphate group transfers straight along one monomer, while in the trans-phosphorelay model, the phosphate group follows a zigzag way, jumping between the two monomers.

However, in the deterministic models, we found that trans- and cis-phosphorelay can not be distinguished analytically or numerically. The models showed that only through double mutant heterodimers we will be able to observe differences between these two models. Thus we used molecular biology approaches to generate constitutively active double mutant ArcB molecules. These mutants have different defective phosphorylation sites, which prohibit normal phosphorelay processes if combined appropriately. After transfecting these mutants into cells using two plasmids with different copy numbers, we monitored the cells' β-gal activity which reflects the activity of [RRp]. However, neither of the trans- or cis-phosphorelay model's simulation result fits the data (figures 4.8A, B and 4.9A, B).

Thus we have come up with a third model, called the allosteric model, where both of the phosphorylation sites in a dimer need to be wild type to allow the conformation change and the subsequent phosphor-transfer. This new model proves to be much better in fitting the data than the other models. It has 99% probability to be the true model compared to the other two models in light of the data.

In chapter 5, we have applied a novel method to study orthodox and non-orthodox systems' performance under different scenarios from the perspective of synthetic biology. The difference between inference and design is that in the former we try to reconstruct the system that has given rise to the data that we observe, whereas in the latter, we seek to construct the system that produces the data that we would like to observe. In chapter 5, we apply an approximate Bayesian computation scheme which required us to simulate from different competing models to arrive at rational criteria.

Jeong-Rae Kim *et. al.* argued that non-orthodox TCSs were superior to orthodox system as they could exhibit ultra-sensitive behaviour and were less sensitive to noisy inputs[82]. When using this method to compare the orthodox TCS and the non-orthodox TCS, we have found that each of these two systems has its own advantage in cellular signalling scenarios (figure 5.2). The orthodox system performs slightly better when rapid activation is required (figure 5.2C) or when a reliably maintained active state is required (figure 5.2D). Thanks to its short phosphorylation process, the system can receive and lose phosphate group quickly. As for the non-orthodox TCS, this system is much more robust in noisy environments. The phosphorelay can provide the system with a buffer for all kinds of noise. Neither the orthodox system nor the non-orthodox system can be said to be better than the other. Their dynamic features are suitable for the role they play in signal transduction process.

In summary, by combining arguments from comparative genomics, systems biology and synthetic biology, we were able to shed new light on the functional and evolutionary study of TCSs. In particular we were able to provide new insights into the phosphorelay mechanism through a combination of modelling, experiment and statistical model selection.

Appendix A

The roles of contact residue disorder and domain composition in characterizing protein-ligand binding specificity and promiscuity

The work presented here has been published as: "The roles of contact residue disorder and domain composition in characterizing protein-ligand binding specificity and promiscuity"

Tang Y, Sheng X, Stumpf MP. Mol Biosyst. 2011 Dec;7(12):3280-6.

A.1 background

Interactions between proteins and small molecular ligands play a crucial role in many biological processes, such as cell signalling and enzymatic reactions, which are regulated by these interactions. Analysing these interactions is of fundamental biological interest and critical for drug discovery and understanding the structure and dynamics of biological networks more generally. To understand the function and basic mechanisms behind ligand-binding interactions, accurate annotation and analysis of the target proteins are required to improve our understanding of interaction.

Interactions between proteins and ligands have been analysed from different perspectives. Bordner [20] and Lopez *et al.* [98] have focused on identification of ligand-binding sites and functionally important residues on proteins, which can give useful clues for predicting the biochemical function of proteins. Andersson *et al.* [4] characterized and

mapped the ligand-binding cavities of proteins that are formed by solvent accessible surfaces, which can provide valuable information on protein similarity obtained from sequence comparisons. Rausell *et al.* [142] explored the role of protein-ligand interactions in different protein subfamilies, as well as the structural distribution of specificity determining positions at protein interfaces. Protein-ligand binding affinity prediction presented by Cheng *et al.* [31] is also of considerable current interest, in order to elucidate interactions between proteins and ligands, including the position and orientation of a flexible ligand in relation to its protein receptor. The above studies principally looked at presence/absence information and the locations of ligand-binding sites. They did not address the problems of specificity and promiscuity of ligand-binding. Bashton *et al.* [13–15] presented an automatic assignment of potential ligands to specific domain; however, they focus on cognate ligand mapping for enzymes and offer little scope for interpreting promiscuity of ligand-binding. Carbonell and Faulon [28] addressed protein promiscuity directly; here a large number of enzymes binding promiscuously were studied in order to generate suitable predictors. However, this study dealt exclusively with the (important) case of enzymes and does not extrapolate to other proteins.

From the above studies the reasons why some proteins interact with many ligands and others interact with only a single ligand are still not sufficiently understood. It has been suggested that disorder plays an important functional role in binding multiple interaction partners [58] and functional diversity [146]. Also, functional properties of individual domains may need to be considered in order to understand their potential roles in interaction promiscuity [120, 178]. Therefore, we study here if disorder and domain composition of proteins in their unbound form play a functional role in multiple ligand binding. Disorder and domain properties of proteins (in their unbound form) binding different classes of ligands have so far not been explored comprehensively. In the present study, we focus our efforts on characterizing the functional and structural properties of proteins binding different numbers of ligands and different ligand types. First, we study the impact of protein contact residue disorder and domain composition on numbers of bound ligands. Based on this analysis, we then investigate whether the difference of disorder and domain architecture is reflected in GO annotations. Second, the properties of hub contact residues are assessed; here hub contact residues refer to contact residues that bind several ligands. Last, the three most common classes of ligands are mapped onto superfamily domains.

A.2 Methods

A.2.1 Protein-ligand complex dataset

The protein-ligand complexes from the March 2010 release of the Protein Data Bank (PDB) [19] were obtained by the following procedure: first, the protein complex entries with ligand molecules were selected; second, to obtain high-resolution structural information where the binding conformation is more reliable, entries with X-ray diffraction studies with a resolution better than 2 Åwere selected; third, the dataset was divided into two subsets, corresponding to eukaryotes and prokaryotes, respectively; fourth, identical and similar PDB sequences were extracted using PDB ID and protein chain information from PDBsum [91, 92]. Because the main differences in ligand binding characteristics appear at 97% of sequence identity in PDB, we used the program CD-HIT [94] to cluster all the PDB sequences at 97% identity to reduce sequence redundancy (while maintaining statistical power), although that only removed some mutants, isoforms, etc. After CD-HIT clustering, 4324 eukaryotic clusters and 5017 prokaryotic clusters were identified, and we selected representative sequences from all the clusters. The bound ligands of the representative sequences were extracted from PDBsum by the corresponding PDB ID and sequence. The ligands are given in the mmCIF-format file [21] and we filtered out those ligands that showed no apparent biological significance.

A.2.2 Contact residues

Contact residues refer to amino acid residues which interact with ligands through their side chains. The contact residues of protein chains with ligands were obtained from PDBsum [91, 92].

A.2.3 Protein domains

The domain information corresponding to a PDB structure was obtained from Pfam [42]. When a protein sequence contains only one domain, the protein is defined as a single domain protein. When a protein has more than one domain, the protein is denoted as a multiple domain protein. Pfam domain frequency was calculated as the fraction of the proteins with a single domain or multiple domains. Here ordered domains are defined

as those that contain less than 40% of disordered residues; changing the value of this threshold does not affect the results.

A.2.4 Domain assignments to superfamilies

Domain assignments at the superfamily level are provided by the SUPERFAMILY resource [189], in which domain assignments are generated using an expert-curated set of profile hidden Markov models. For a PDB entry domain, there are five SCOP classifications in SUPERFAMILY: class, fold, superfamily, family and protein. We attached the superfamily classification to each PDB entry considered below.

A.2.5 Gene Ontology annotations

To assign protein function to PDB chains, firstly we used SIFTS [183] which provides residue level annotations for PDB sequences from the Gene Ontology (GO) Consortium resource [12]. Then we used the GO database to classify these PDB chains by biological process, cellular component and molecular function.

A.2.6 Disorder predictions

For each PDB sequence disordered amino acids were obtained from FoldIndex [136], which predicts if a given protein sequence is intrinsically unfolded using an algorithm based on the average residue hydrophobicity and net charge of the sequence. Then the percentage of disordered contact residues was estimated as disordered contact residue counts divided by contact residue counts on a PDB sequence.

A.2.7 Statistical methods

Kolmogorov-Smirnov tests and binomial proportion tests were used to compare distributions and proportions, respectively, in this study. To investigate the correlation between contact residue disorder and GO annotations, we calculated Z-scores, which had previously been used to examine the association of protein disorder predictions and GO annotations [58, 185]. We also calculated Z-scores to determine the significance of the correlation between superfamily domains and the frequency of proteins binding some

class (es) of ligand (s). All of the above statistical methods were performed in the R language; where appropriate, we controlled the false-discovery rate using the approach of Benjamini and Hochberg (with q=0.05), and only the adjusted *P*-values are reported here.

A.3 Results

A.3.1 Analysis of disorder and domain composition of proteins with different numbers of ligands

A.3.1.1 Analysis of contact residue disorder

In the protein-ligand complex dataset we find that almost half of the proteins are only known to bind a single ligand, and only 25% of proteins bind more than two ligands. The maximum known numbers of bound ligands are 11 and 13 in eukaryotic and prokaryotic datasets, respectively. Because the number of proteins with more than 5 ligands is not large enough to meet our statistical requirements, we only analysed the fraction of contact amino acid residues in disordered regions for proteins with up to five bound ligands. Figure A.1 shows the distributions of the prevalence of disorder among the contact residues in eukaryotes and prokaryotes when the number of bound ligands is 1, 2, 3, 4 and 5 (after removing proteins with less than 3 contact residues). Eukaryotic and prokaryotic proteins both show a strong preference for ordered contact residues, but contact residues of prokaryotic proteins are more ordered than those in eukaryotes across all ligand categories. Furthermore, as ligand counts increase the contact residues tend to become more disordered in both domains of life.

Pairs of ligand number	*P*-values in eukaryotes	*P*-values in prokaryotes
1 and 2	4.499000e−04	1.665400e−14
1 and 3	5.358000e−10	<8.800000e−16
1 and 4	5.358000e−10	1.893333e−10
1 and 5	8.681333e−05	3.786000e−06

TABLE A.1: Kolmogorov-Smirnov test for the distributions of PDB contact residue disorder binding different numbers of ligands

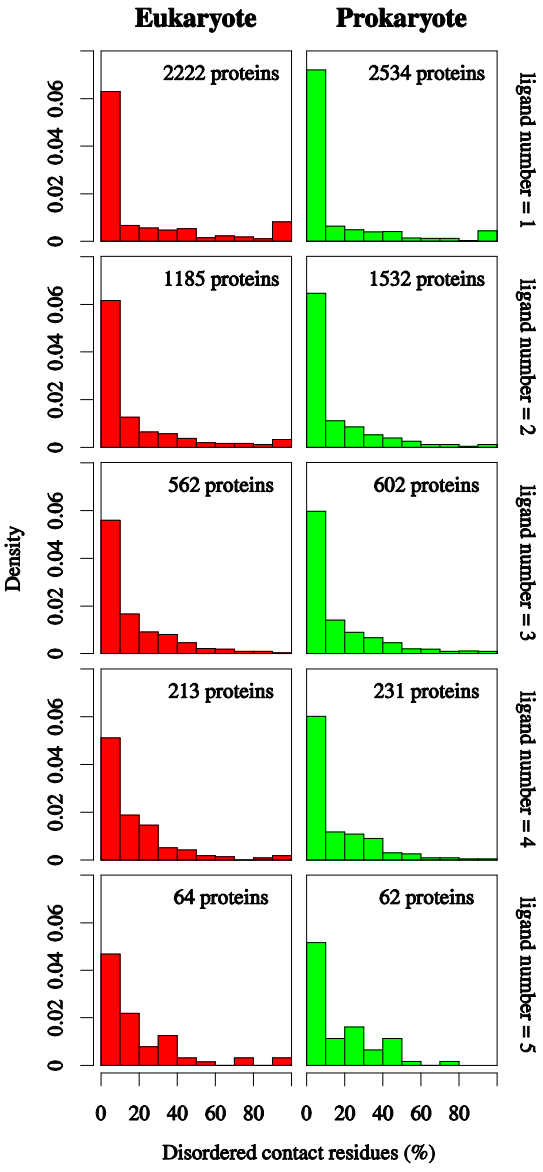


FIGURE A.1: Distribution of disordered contact residues of proteins with 1, 2, 3, 4 and 5 bound ligands in eukaryotes and prokaryotes.

Table A.1 contains the P-values (after correcting for multiple testing) of the Kolmogorov-Smirnov test that we used to test if two distributions are different from each other. It shows that the distribution of proteins binding one ligand is significantly different from those binding 2, 3, 4 and 5 ligands. Furthermore, the contact residue disorder percentages in proteins with one ligand and more than one ligand (see Figure A.2) are both significantly different in eukaryotes (P-value $= 4.996\text{e}-15$) and in prokaryotes (P-value $< 2.2\text{e}-16$). Thus contact residues of proteins with more than one ligand are more disordered compared to those with only a single ligand, which could suggest, quite

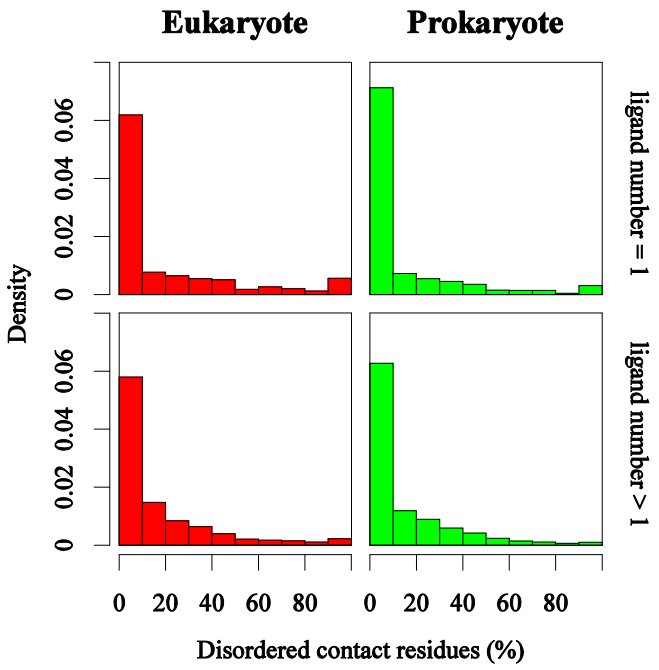


FIGURE A.2: Distribution of disordered contact residue. Comparison of the proteins with one and more than one ligand in eukaryotes and prokaryotes.

reasonably, that more disordered contact residues can bind more ligands.

Generally, disordered proteins have more charged residues. Therefore, we investigated the charged contact residue percentages in proteins with one ligand (25.04% in eukaryotes; 13.21% in prokaryotes) and more than one ligand (48.26% in eukaryotes; 25.33% in prokaryotes). We found that they are both significantly different in eukaryotes (P-value $< 2.2e-16$) and in prokaryotes (P-value $< 2.2e-16$), and contact residues of proteins with more than one ligand tend to have more charged residues than those with only a single ligand.

A.3.1.2 The impact of ordered multiple domains on ligand numbers

In addition to binding site disorder, the number of bound ligands may also depend on the domain composition of proteins. Therefore, we assigned Pfam domains to all proteins (see Methods) and investigated the correlation between multiple domains and the number of ligands. The correlation between the domain count and ligand number is 0.129 (P-value $< 2.2e{-}16$) in eukaryotes and 0.097 (P-value $= 4.68e{-}12$) in prokaryotes, respectively but the prevalence of multiple domain proteins does not increase simply with

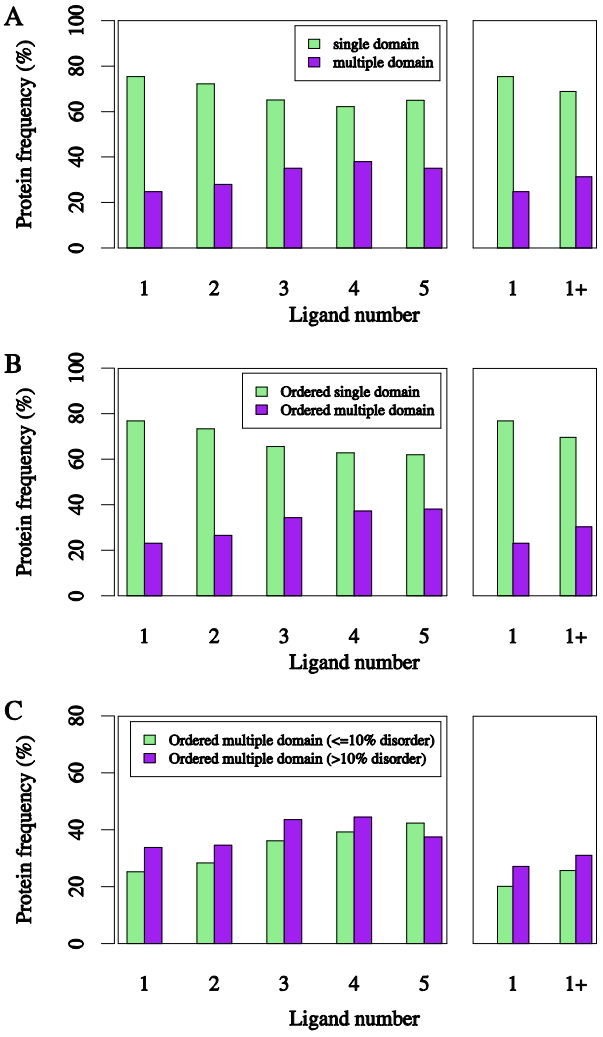


FIGURE A.3: Pfam domain frequency. The frequency of proteins are compared with ligand number increasing between A, single domain proteins and multiple domain proteins; B, ordered single domain proteins and ordered multiple domain proteins and C, ordered multiple domain proteins when contact residue disorder is less than 10% and more than 10%.

ligand count (see Figure A.3A). Further, we find that there appears to be a slight increase in the frequency of multiple domain architectures (31.24%) among proteins binding more than one ligand (compared to 24.70% for single domain proteins, P-value $= 1.265\text{e}-15$).

Figure A.3B shows that the number of ordered multiple domains rises with ligand number, and Table A.2 shows that proteins with more than one ligand tend to have multiple domains more often than those with only a single ligand. Similarly, proteins with multiple (ordered) domains with more than one known ligand (30.37%) are more abundant than those with only a single ligand (23.08%, P-value $= 5.118\text{e}-12$).

Pairs of ligand number	Frequency (%)	*P*-values
1 and 2	23.08 and 26.54	3.857000e−03
1 and 3	23.08 and 34.36	1.824800e−11
1 and 4	23.08 and 37.18	1.558000e−08
1 and 5	23.08 and 38.04	1.718667e−03

TABLE A.2: Frequency of proteins with ordered multiple domains and its binomial proportions test.

Pairs of ligand number	Frequency (%)	*P*-values
1 and 2	25.19 and 28.33	3.4070e−02
1 and 3	25.19 and 36.15	7.1680e−07
1 and 4	25.19 and 39.30	3.4380e−05
1 and 5	25.19 and 42.31	1.0996e−02

TABLE A.3: Frequency of proteins with ordered multiple domains and with less than 10% disordered contact residues and its binomial proportions test

As shown in Figure A.1, the fraction of disordered contact residues is less than 10 percent in most proteins. Figure A.3C shows that multiple ordered domains increase with ligand number in proteins with less than 10% contact residue disorder, but not in proteins with more than 10% contact residue disorder. The relative frequencies of proteins with ordered multiple domains binding different ligand numbers are shown in Table A.3. Ordered multiple domains are more common among proteins with more than one ligand (32.15%) than those with one ligand (25.19%) when contact residue disorder is less than 10% (P-value = 9.214e−08). However, when contact residue disorder is more than 10%, the percentages are at best barely significantly different at 33.85% (proteins with one ligand) and 38.73% (protein with more than one ligand, P-value = 0.02066). This suggests that in the absence of disordered contact residues, having more domains may help these proteins to bind multiple ligands.

A.3.1.3 Gene Ontology enrichment

The differences in disorder and domain architecture are reflected by the GO annotations of proteins with one or more ligands. We investigated the biological process, cellular component and molecular function terms and, in the initial scan, selected GO terms that had P-value lower than 0.2 and that contained more than 50 proteins (we use P-value as a

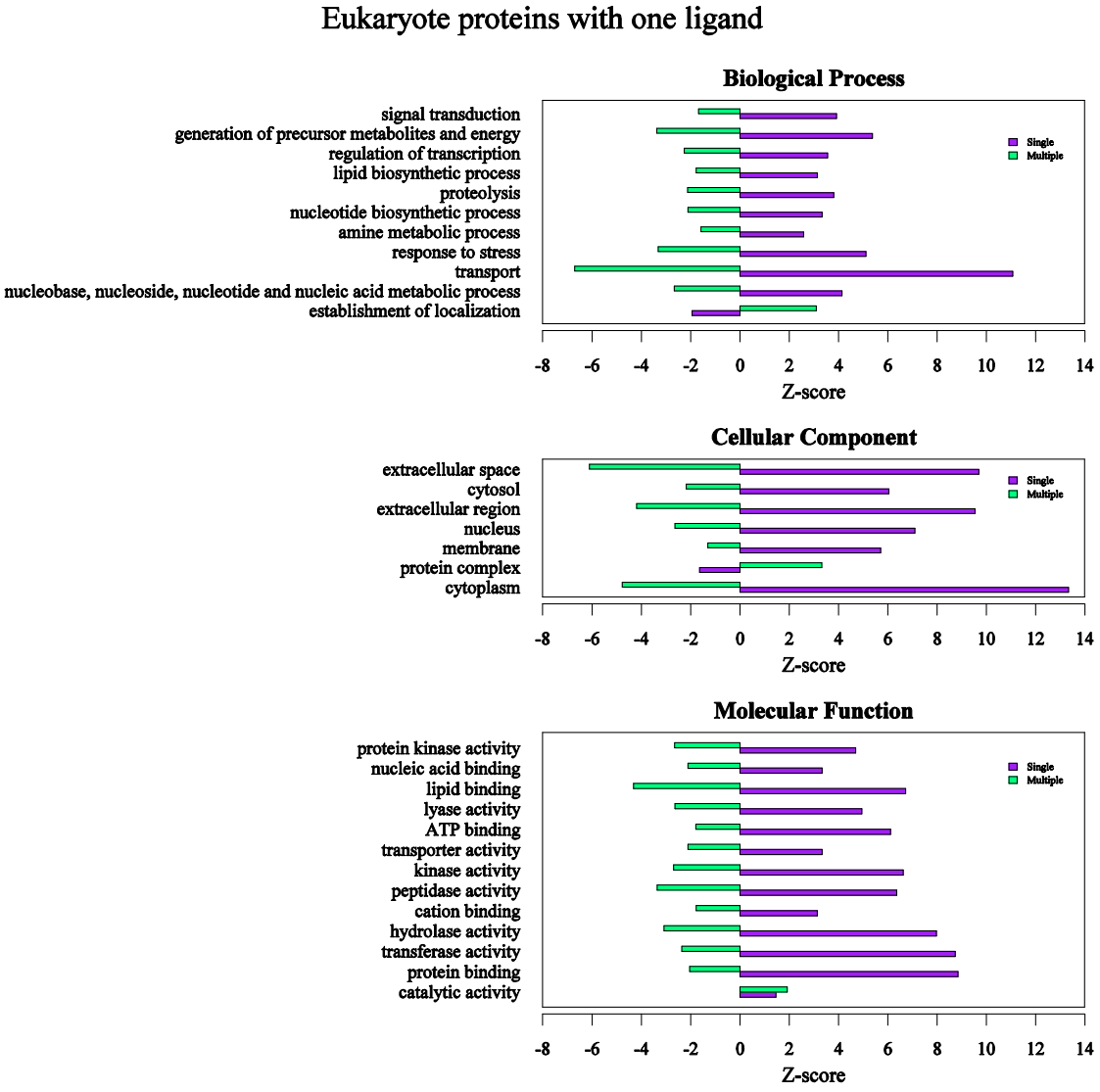


FIGURE A.4: Association of contact residue disorder with GO annotations in ordered single domain proteins and ordered multiple domain proteins. Legend Single is ordered single domain proteins and legend"Multiple" is ordered multiple domain proteins. Eukaryote proteins with one ligand.

coarse tool to scan for plausible hypotheses). The Z-score (see Methods) for the number of disordered contact residues associated with each annotation in ordered single domain and ordered multiple domains is plotted in Figure A.4, A.5, A.6and A.4. Z-score values and *P*-values are showed in Table S1 of the supplementary information in the paper of Yurong and me *et al.* [169] (this table is too large to be shown here). A positive (negative) Z-score shows that more (less) contact residue disorder is associated with the corresponding annotation than would be expected by chance. For proteins binding one ligand the contact residue disorder is found to be enriched in ordered single domain

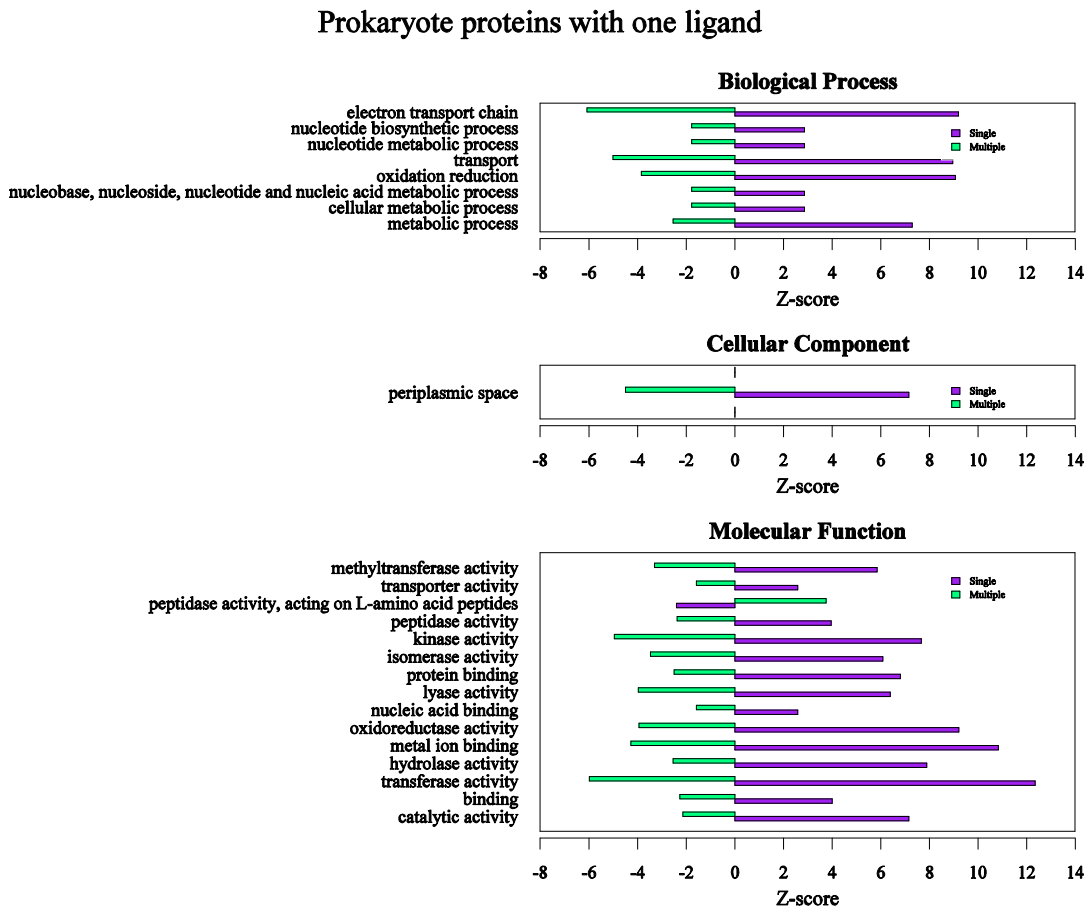


FIGURE A.5: Association of contact residue disorder with GO annotations in ordered single domain proteins and ordered multiple domain proteins. Legend Single is ordered single domain proteins and legend Multiple is ordered multiple domain proteins. Prokaryote proteins with one ligand.

proteins and depleted in ordered multiple domain proteins for almost all annotations related to biological process, cellular component and molecular function (see Figure A.4 and Figure A.5). This suggests that the proteins binding one ligand do not contain multiple domains or have low disorder percentage, which is consistent with the result in section A.3.1.2. For the ordered multiple domain proteins binding several ligands, contact residue disorder is found to be enriched for some annotations but depleted for other others (see Figure A.6 and Figure A.7). This, too, indicates that multiple domains might help proteins to bind more ligands in the absence of disordered contacts. Disorder appears to be a critical factor that enables proteins to bind more ligands, especially in some annotation classes. For example, in Figure A.6, eukaryotic single domain and multiple domain proteins with more ligands involved in two molecular function GO annotations, kinase activity and protein binding, are strongly associated with disorder, in

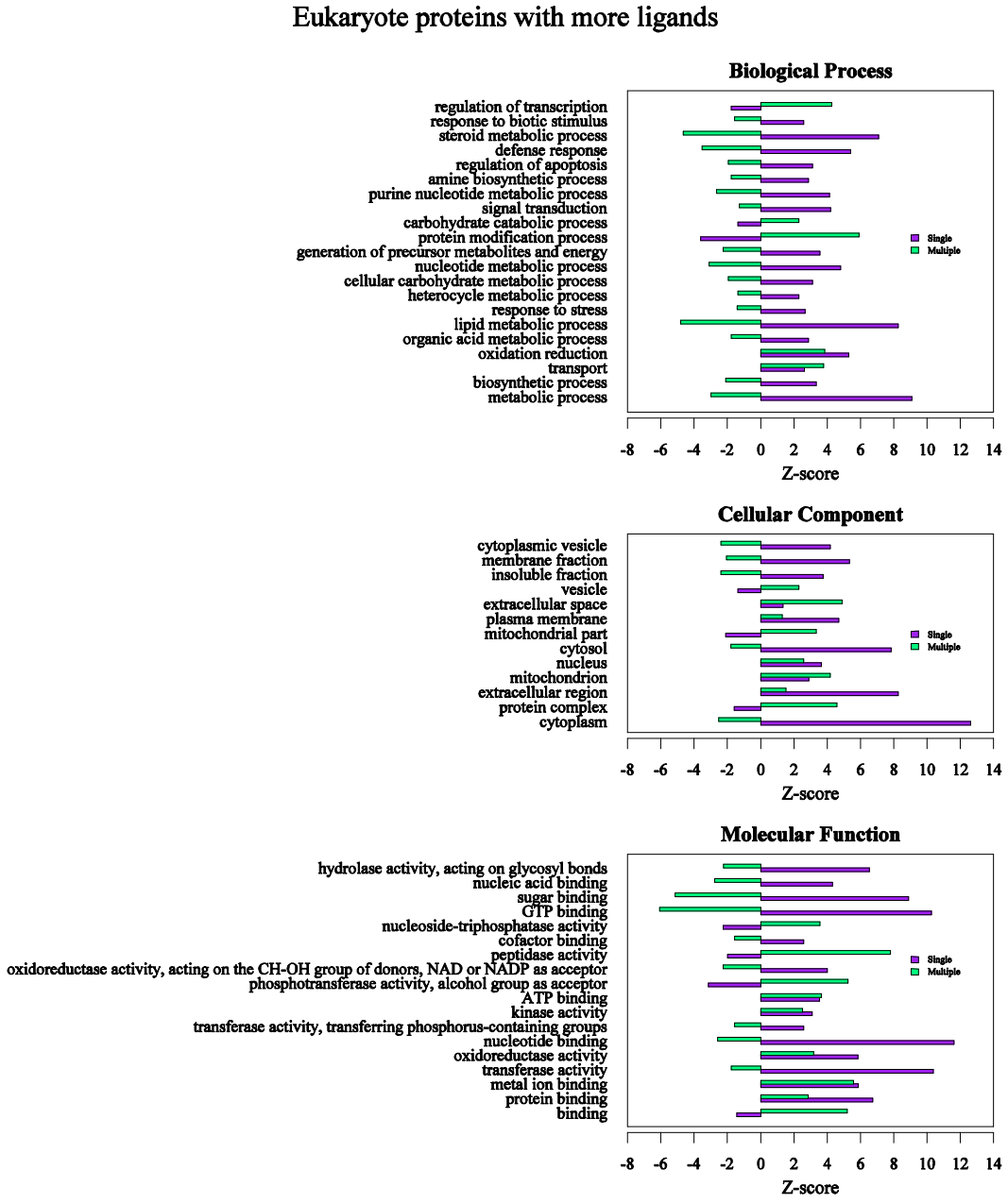


FIGURE A.6: Association of contact residue disorder with GO annotations in ordered single domain proteins and ordered multiple domain proteins. Legend Single is ordered single domain proteins and legend Multiple is ordered multiple domain proteins. Eukaryote proteins with more ligands.

support of previous observations of high disorder content in proteins involved in these two molecular functions [185]. Proteins with protein binding functional annotation are also enriched for disorder, which agrees with earlier results that hub proteins tend to exhibit disorder [58]. Furthermore eukaryotic single domain and multiple domain proteins that are annotated as localized in the nucleus and that bind several ligands are enriched for

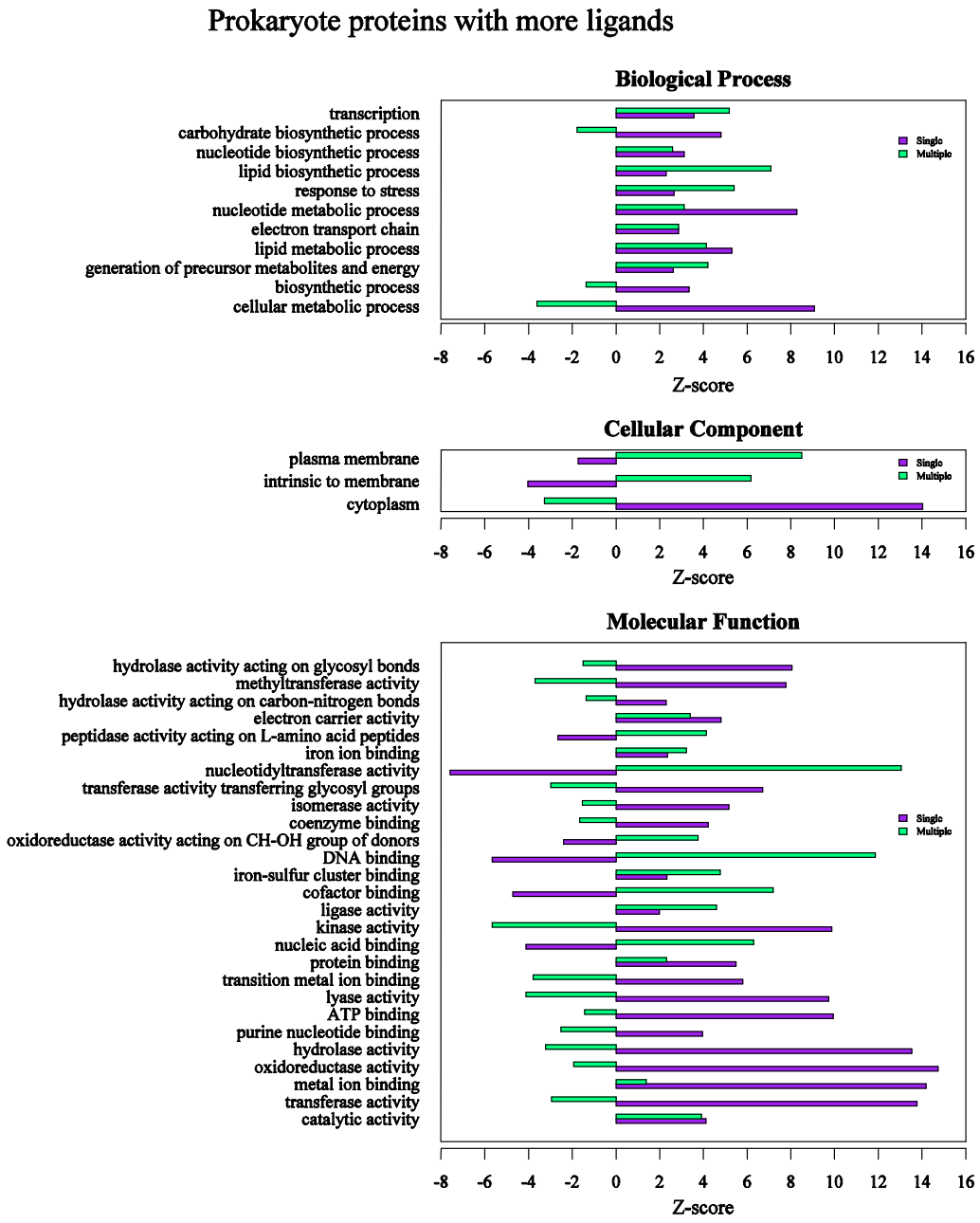


FIGURE A.7: Association of contact residue disorder with GO annotations in ordered single domain proteins and ordered multiple domain proteins. Legend Single is ordered single domain proteins and legend Multiple is ordered multiple domain proteins. Prokaryote proteins with more ligands.

disorder. A previous study has also demonstrated that nuclear proteins are typically enriched for disorder [185]. Interestingly, however, for only few annotations may single domain proteins show low levels of disorder while still binding multiple ligands. This further reflects the complexity of the problematic relationship between protein disorder and binding specificity versus promiscuity.

A.3.2 The property of hub contact residues binding multiple ligands

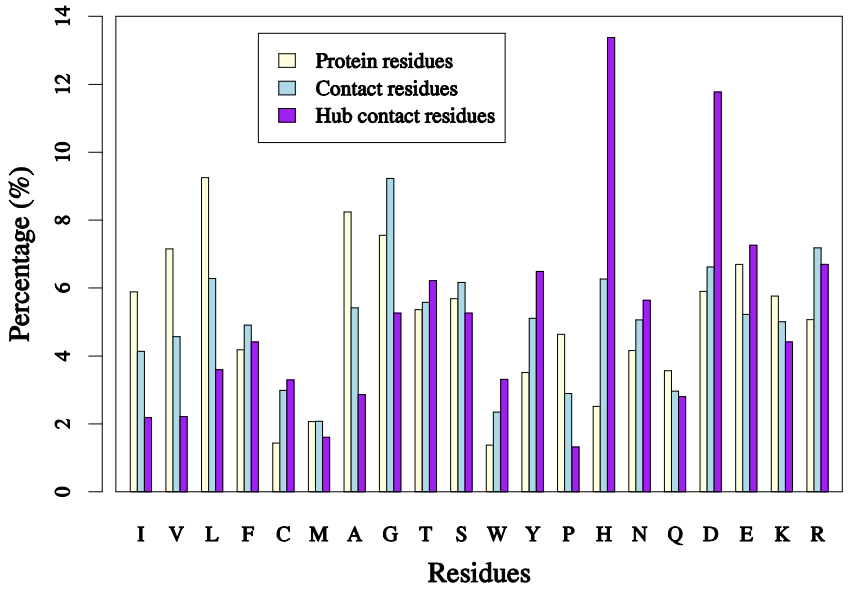


FIGURE A.8: Amino acid composition. Comparison of the residue composition for the whole protein, for the contact residues and for the hub contact residues.

Some contact residues bind multiple ligands while others only bind single ligands. We refer to contact residues with multiple ligands as hub contact residues. To determine the properties of hub contact residues, we compared the percentage of hub contact residues with contact residues in disordered regions, multiple domains and ordered multiple domains, but failed to detect any significant differences. However, we found that 75.86% of contact residues are within domains; compared to normal contact residues (74.62%), hub contact residues (80.47%) show a preference (P-value $< 2.2e-16$) for domains, which might help hub contact residues to bind more ligands. For in-depth analysis of hub contact residues, we computed the amino acid composition for the whole protein, all contact residues and hub contact residues (see Table A.4 in the supplementary information). Figure A.8 shows the proportion of different residues for the whole protein, the contact residues and hub contact residues. The 20 amino acids are listed on the x-axis according to their hydrophobicity. The results show that the whole protein residue composition, contact residue composition and hub contact residue composition are significantly different in important physicochemical respects.

First, we investigated if the contact residues and hub contact residues have a preference for hydrophobic residues (IVLFCMA) or hydrophilic (GTSWYPHNQDEKR) residues.

Residues	Frequency in protein residues	Frequency in contact residues	Frequency in hub contact residues
I	5.886	4.133	2.183
V	7.149	4.565	2.217
L	9.247	6.279	3.596
F	4.181	4.905	4.412
C	1.431	2.986	3.297
M	2.070	2.078	1.608
A	8.242	5.416	2.861
G	7.554	9.225	5.262
T	5.362	5.580	6.216
S	5.688	6.162	5.262
W	1.373	2.349	3.309
Y	3.506	5.108	6.491
P	4.639	2.895	1.321
H	2.517	6.263	13.373
N	4.157	5.061	5.641
Q	3.564	2.965	2.803
D	5.899	6.620	11.776
E	6.697	5.221	7.261
K	5.760	5.007	4.412
R	5.072	7.186	6.698

TABLE A.4: Frequency of amino acids on the whole proteins, the contact residues and the hub contact residues.

Contact residues are less likely to be hydrophobic (P-value $< 2.2e-16$) and contain a higher percentage of hydrophilic residues (P-value $< 2.2e-16$) compared to the whole protein. The hub contact residues are further depleted in hydrophobic residues (P-value $< 2.2e-16$) and enriched in hydrophilic residues (P-value $< 2.2e-16$) compared to the general contact residues. Second, we considered differences in charged residue (DEKR) and polar residue (TSYHNQDEKR) composition. The result shows that the contact residues prefer charged residues and polar residues (P-value $= 0.001045$ and P-value $< 2.2e-16$). Compared to the contact residues, the hub contact residues are further enriched for charged residues and polar residues (P-value $< 2.2e-16$ and P-value $< 2.2e-16$). Last, we found that histidine (H) and aspartate (D) residues are significantly enriched among the hub contact residues compared to other contact residues (P-value $< 2.2e-16$). It is known that two-component signalling systems in bacteria have H and D residues in the phosphorylation sites of their histidine kinase and receiver domains [171]; H and D residues are also found as catalytic triads of serine protease. Thus we conjecture that the peak of H and D might due to the prevalence of

TCSs and serine proteases in PDB. Further study needs to be done to test our hypothesis. From the above results we conclude that the hub contact residues are more likely to have more hydrophilic, charged, polar and His-Asp catalytic triad residues than other contact residues, suggesting perhaps a basic underlying mechanism for ligand interaction promiscuity.

A.3.3 Disorder distribution and superfamily domain mapping in several ligand classes

There are more than 10,000 kinds of ligands in the protein-ligand complex database. These ligands can be classified into ten classes according to the conventional mmCIF-format [21]. Among these ten classes, HETAIN (predominantly inhibitors and non-canonical biological molecules) and HETAI (small ions) are the most common ligand classes [97]. We classified all proteins as binding HETAIN, binding HETAI and binding others, if they bind any of the other eight classes. We found that around 90% of the annotated proteins bind either HETAIN or HETAI or both classes simultaneously. Therefore, the following comparisons are based on proteins binding HETAIN (and not HETAI), HETAI (but not HETAIN) and HETAIN&HETAI simultaneously.

We studied contact residue disorder of proteins binding these three ligand classes in eukaryotes and prokaryotes. Figure A.9 in the supplementary information shows the distribution of contact residue disorder percentage of proteins binding the three classes of ligands. The figure indicates that proteins binding HETAI have the highest contact residue disorder percentage and those with HETAIN have a lower contact residue disorder percentage. The Kolmogorov-Smirnov test shows that the distribution of proteins binding HETAIN and HETAI are significantly different in both eukaryotes (P-value $< 2.2\text{e}-16$) and in prokaryotes (P-value $= 9.05\text{e}-07$). And HETAIN and HETAI-binding proteins are also significantly different from those, which bind both classes of ligands (P-value < 0.05 in eukaryotes and prokaryotes).

To further elucidate the differences between proteins binding HETAI, HETAIN and HETAIN&HETAI, we assigned the superfamily domains and analysed the mapping between superfamily domains and their ligands (see the Methods section). We chose the top 58 superfamily domains with more than 20 proteins (see Figure A.10 and Figure A.11). From Figure A.11, we see that one of these common domains bind exclusively

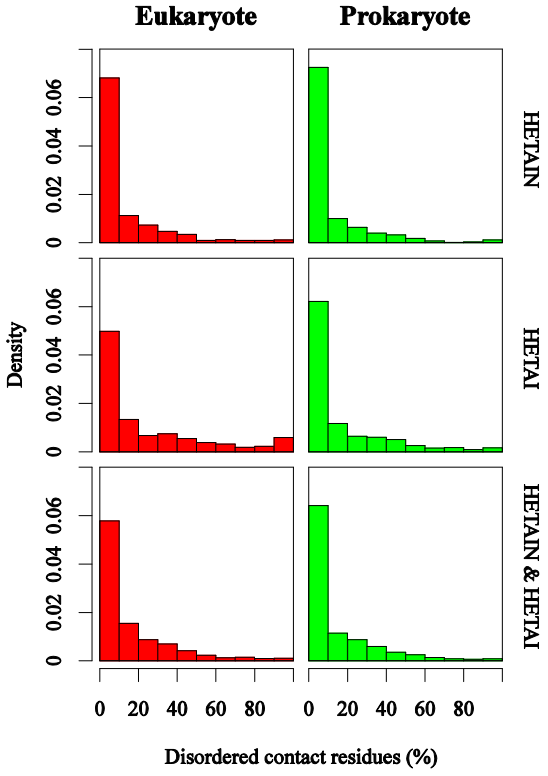


FIGURE A.9: Distribution of disordered contact residues of proteins binding HETAIN, HETAI and HETAIN&HETAI in eukaryotes and prokaryotes.

only one ligand type. To discriminate between proteins with HETAIN, HETAI and HETAIN&HETAI ligands, we used a permutation test to obtain a null model distribution and computed Z-score (see Methods). Superfamily domains were selected that show at least weak association with any of the three classes; the Z-score for the observed number of proteins associated with each selected superfamily domain is given in Table A.5 of the supplementary information. Figure A.10 plots the Z-score values for these superfamily classes. We observe, for example, that there are more proteins binding HETAIN&HETAI associated with THDP-binding (Thiamin diphosphate-binding fold) and GroES-like domains. Some of the superfamily domains examined here, EF-hand, Cupredoxins and Zn-dependent exopeptidases are associated with HETAI-type ligands (see Figure A.10). There are more proteins binding HETAIN associated with Globin-like domains, FAD/NAD-linked domains, Nuclear receptor ligand-binding domains and NAD (P)-binding Rossmann-fold domains. While previous studies [13, 121, 148] have found that NAD (P)-binding Rossmann-fold domain bind NAD (nicotinamide adenine dinucleotide), our results provide strong evidence in favour of Rossmann-folds to bind

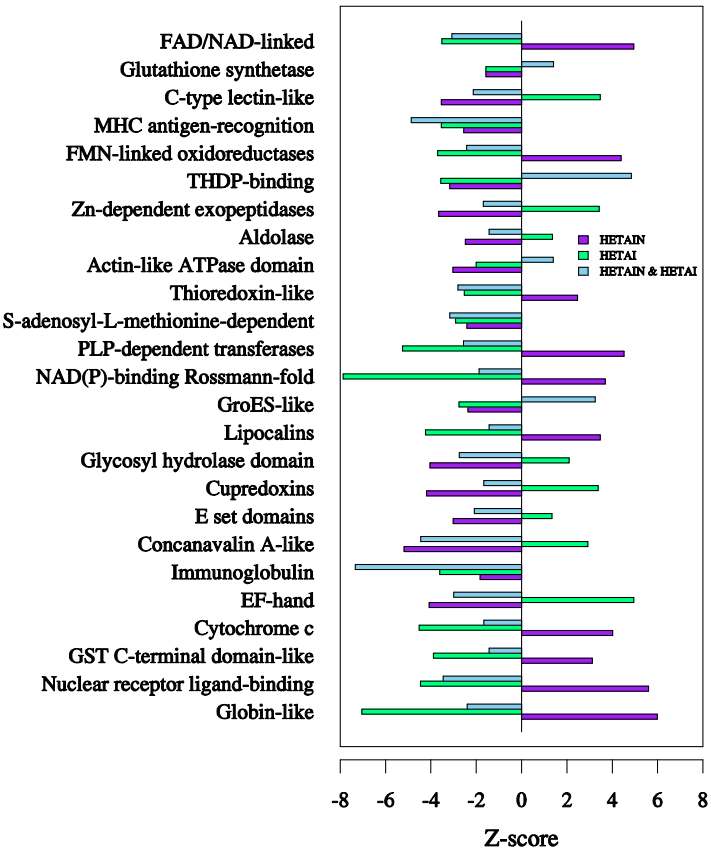


FIGURE A.10: Association of protein frequency with Superfamily domain. Z-score values are computed for the observed number of proteins binding three ligand classes (HETAI, HETAIN and HETAIN&HETAI) associated with each selected superfamily domain. Positive (negative) Z-score indicates that more (less) proteins are associated with the indicated superfamily domain than would be expected by chance.

more generally to ligands that (like NAD) belong to the HETAIN class. FAD/NAD-like association of HETAIN agrees well with the ligand-domain mapping tool PROCOGNATE7, in which the domain is mapped to some HETAIN type ligands like 3AA, ACM and AJ3. Our observations for the Globin-like and Nuclear receptor ligand-binding domains also agree with PROCOGNATEs in silico predictions. The immunoglobulin domain and MHC antigen-recognition domains are not associated with any of the three common classes of ligands, which may well reflect their immunological role.

Scop domain	Name	HN_zscore	HI_zscore	HT&HI_zscore	HN_pvalue	HI_pvalue	HT&HI_pvalue
46458	Globin-like	5.978	-7.048	-2.396	2.254E-09	1.810E-12	1.658E-02
48508	Nuclear receptor ligand-binding domain	5.604	-4.463	-3.456	2.093E-08	8.075E-06	5.473E-04
47616	GST C-terminal domain-like	3.116	-3.886	-1.435	1.836E-03	1.020E-04	1.512E-01
46626	Cytochrome c	4.015	-4.513	-1.670	5.938E-05	6.399E-06	9.491E-02
47473	EF-hand	-4.078	4.944	-2.995	4.552E-05	7.671E-07	2.744E-03
48726	Immunoglobulin	-1.830	-3.607	-7.338	6.725E-02	8.587E-03	8.587E-03
49899	Concanavalin A-like lectins/glucanases	-5.185	2.924	-4.448	2.160E-07	3.453E-03	8.674E-06
81296	E set domains	-3.019	1.339	-2.086	2.532E-03	1.807E-01	3.701E-02
49503	Cupredoxins	-4.189	3.370	-1.669	2.801E-05	7.512E-04	9.505E-02
51011	Glycosyl hydrolase domain	-4.034	2.098	-2.743	5.475E-05	3.592E-02	6.081E-03
50814	Lipocalins	3.466	-4.236	-1.435	5.289E-04	2.277E-05	1.512E-01
50129	GroES-like	-2.362	-2.762	3.243	1.818E-02	5.741E-03	1.184E-03
51735	NAD(P)-binding Rossmann-fold domains	3.689	-7.865	-1.879	2.250E-04	3.678E-15	6.020E-02
53383	PLP-dependent transferases	4.514	-5.245	-2.562	6.376E-06	1.559E-07	1.041E-02
53335	S-adenosyl-L-methionine-dependent methyltransferases	-2.405	-2.912	-3.165	1.616E-02	3.596E-03	1.552E-03
52833	Thioredoxin-like	2.472	-2.528	-2.806	1.343E-02	1.148E-02	5.023E-03
53067	Actin-like ATPase domain	-3.028	-2.007	1.395	2.463E-03	4.473E-02	1.631E-01
51569	Aldolase	-2.485	1.365	-1.435	1.294E-02	1.722E-01	1.512E-01
53187	Zn-dependent exopeptidases	-3.653	3.418	-1.689	2.588E-04	6.316E-04	9.118E-02
52518	Thiamin diphosphate-binding fold (THDP-binding)	-3.163	-3.563	4.844	1.564E-03	3.668E-04	1.273E-06
51395	FMN-linked oxidoreductases	4.384	-3.703	-2.426	1.165E-05	2.131E-04	1.526E-02
54452	MHC antigen-recognition domain	-2.550	-3.543	-4.868	1.078E-02	3.954E-04	1.129E-06
56436	C-type lectin-like	-3.536	3.466	-2.135	4.067E-04	5.289E-04	3.273E-02
56059	Glutathione synthetase ATP-binding domain-like	-1.575	-1.575	1.405	1.153E-01	1.153E-01	1.601E-01
55424	FAD/NAD-linked reductases, dimerisation (C-terminal) domain	4.944	-3.519	-3.074	7.641E-07	4.333E-04	2.116E-03

TABLE A.5: Scop domains significantly enriched in proteins with HETAIN, HETAI and HETAIN&HETAI. (P-value < 0.2)

A.4 Conclusions

The results of this study indicate that protein-ligand binding specificity and diversity are functions of protein (contact residue) disorder, domain composition and the physicochemical properties of the contact residues. Hub contact residues can bind more ligands compared to conventional contact residues because they contain more hydrophilic, charged, polar and his-asp catalytic triad residues. Furthermore, specific superfamily domains prefer certain classes of ligands. In order to qualify any redundancy in our dataset and its potential effects on the analysis, we investigated the fold and superfamily composition of the dataset. The result shows that the dataset has a bias for some folds, for example, "TIM beta/alpha-barrel", "P-loop containing nucleoside triphosphate hydrolases", "Immunoglobulin-like beta-sandwich", "NAD (P)-binding Rossmann-fold domains", "Ferredoxin-like" and "Flavodoxin-like". Those proteins with the above folds make up some 13.65% of the proteins considered here. In terms of superfamily composition, the dataset exhibits a preference for the "P-loop containing nucleoside triphosphate hydrolases", "NAD (P)-binding Rossmann-fold domains" and "Immunoglobulin" superfamilies. The proteins with the above superfamilies account for 6.57% of the number of all the proteins. Furthermore, we investigated the ligand composition of the dataset. The result shows that the dataset has a preference for some ligands, such as GOL (Glycerol), NAG (N-acetyl-d-glucosamine), HEM (Protoporphyrin ix containing fe), FLC (Citrate anion), SPM (Spermine), EDO (1,2-ethanediol) and ADP (Adenosine-5'-diphosphate), which together make up 20.73%. Even with this caveat, the overall picture remains clear and biophysically appealing: disorder and multiple domain architecture are complementary factors that affect protein-ligand binding specificity, affinity and promiscuity jointly. These new and more nuanced insights hold potentially valuable clues for rationally interfering with protein-activity by designing appropriate ligands. Obviously, here we have taken only a first step by searching systematically for factors contributing to promiscuity and specificity. But by investigating the extent of contact residue disorder and domain architecture of potential drug targets we may be able to triage the vast space of potential ligands in a suitable manner, identifying promising classes of ligands. Once more structural insights have been gained; we can also do it by homing in onto manageable sets of particularly appropriate or promising ligands.

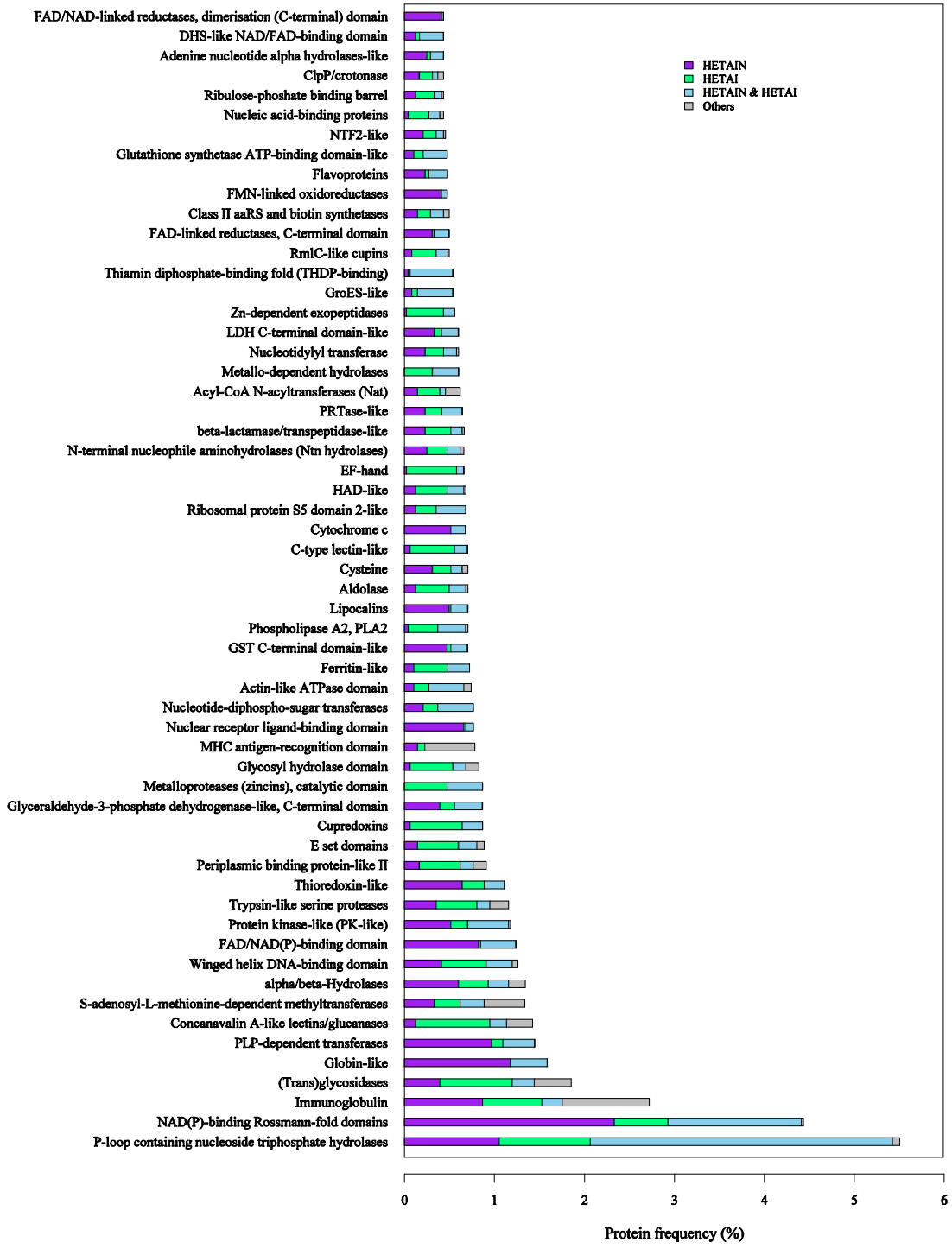


FIGURE A.11: The top 58 superfamily domains with more than 20 proteins.

Bibliography

[1] Abel, T. Gene regulation. Action of leucine zippers. *Nature*, 1989.

[2] Alexeeva, S., Hellingwerf, K.J. and Teixeira de Mattos, M.J. Requirement of arca for redox regulation in escherichia coli under microaerobic but not anaerobic or aerobic conditions. *J Bacteriol*, 185(1):204–9, 2003.

[3] Anderson, J.C., Clarke, E.J., Arkin, A.P. and Voigt, C.A. Environmentally controlled invasion of cancer cells by engineered bacteria. *J Mol Biol*, 355(4):619–27, 2006.

[4] Andersson, C.D., Chen, B.Y. and Linusson, A. Mapping of ligand-binding cavities in proteins. *Proteins*, 78(6):1408–22, 2010.

[5] Attwood, P.V., Piggott, M.J., Zu, X.L. and Besant, P.G. Focus on phosphohistidine. *Amino acids*, 32(1):145–156, 2007.

[6] Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M. *et al.* Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection. *Mol Syst Biol*, 2:2006.0008, 2006.

[7] Bader, M.W., Sanowar, S., Daley, M.E., Schneider, A.R., Cho, U., Xu, W., Klevit, R.E., Le Moual, H. *et al.* Recognition of antimicrobial peptides by a bacterial sensor kinase. *Cell*, 122(3):461–72, 2005.

[8] Baldi, P. and Brunak, S. *Bioinformatics: the machine learning approach.* MIT Press, Cambridge, Mass., 1998.

[9] Barakat, M., Ortet, P., Jourlin-Castelli, C., Ansaldi, M., Méjean, V. and Whitworth, D.E. P2CS: a two-component system resource for prokaryotic signal transduction research. *BMC genomics*, 10:315, 2009.

[10] Barnabas, J., Schwartz, R.M. and Dayhoff, M.O. Evolution of major metabolic innovations in the precambrian. *Orig Life*, 12(1):81–91, 1982.

[11] Barnes, C.P., Silk, D., Sheng, X. and Stumpf, M.P.H. Bayesian design of synthetic biological systems. *Proc Natl Acad Sci U S A*, 108(37):15,190–5, 2011.

[12] Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C. and Apweiler, R. The goa database in 2009–an integrated gene ontology annotation resource. *Nucleic Acids Res*, 37(Database issue):D396–403, 2009.

[13] Bashton, M., Nobeli, I. and Thornton, J.M. Cognate ligand domain mapping for enzymes. *J Mol Biol*, 364(4):836–52, 2006.

[14] Bashton, M., Nobeli, I. and Thornton, J.M. Procognate: a cognate ligand domain mapping for enzymes. *Nucleic Acids Res*, 36(Database issue):D618–22, 2008.

[15] Bashton, M. and Thornton, J.M. Domain-ligand mapping for enzymes. *J Mol Recognit*, 23(2):194–208, 2010.

[16] Batt, G., Yordanov, B., Weiss, R. and Belta, C. Robustness analysis and tuning of synthetic gene networks. *Bioinformatics*, 23(18):2415–22, 2007.

[17] Battogtokh, D., Asch, D.K., Case, M.E., Arnold, J. and Schuttler, H.B. An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of neurospora crassa. *Proc Natl Acad Sci U S A*, 99(26):16,904–9, 2002.

[18] Bekker, M., Alexeeva, S., Laan, W., Sawers, G., Teixeira de Mattos, J. and Hellingwerf, K. The ArcBA two-component system of Escherichia coli is regulated by the redox state of both the ubiquinone and the menaquinone pool. *Journal of bacteriology*, 192(3):746–754, 2010.

[19] Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L. *et al.* The protein data bank. *Acta Crystallogr D Biol Crystallogr*, 58(Pt 6 No 1):899–907, 2002.

[20] Bordner, A.J. Predicting small ligand binding sites in proteins using backbone structure. *Bioinformatics*, 24(24):2865–71, 2008.

[21] Bourne, P.E., Berman, H.M., McMahon, B., Watenpaugh, K.D., Westbrook, J.D. and Fitzgerald, P.M. Macromolecular crystallographic information file. *Methods Enzymol*, 277:571–90, 1997.

[22] Bourret, R.B. and Stock, A.M. Molecular information processing: lessons from bacterial chemotaxis. *J Biol Chem*, 277(12):9625–8, 2002.

[23] Bray, D. and Lay, S. Computer simulated evolution of a network of cell-signaling molecules. *Biophys J*, 66(4):972–7, 1994.

[24] Bren, A. and Eisenbach, M. How signals are heard during bacterial chemotaxis: protein-protein interactions in sensory signal propagation. *J Bacteriol*, 182(24):6865–73, 2000.

[25] Buckland, R. Leucine zipper motif extends. *Nature*, 1989.

[26] Buckler, D.R., Anand, G.S. and Stock, A.M. Response-regulator phosphorylation and activation: a two-way street? *Trends in microbiology*, 8(4):153–156, 2000.

[27] Canton, B., Labno, A. and Endy, D. Refinement and standardization of synthetic biological parts and devices. *Nat Biotechnol*, 26(7):787–93, 2008.

[28] Carbonell, P. and Faulon, J.L. Molecular signatures-based prediction of enzyme promiscuity. *Bioinformatics*, 26(16):2012–9, 2010.

[29] Cases, I. and de Lorenzo, V. Genetically modified organisms for the environment: stories of success and failure and what we have learned from them. *Int Microbiol*, 8(3):213–22, 2005.

[30] Chakraborty, S., Li, M., Chatterjee, C., Sivaraman, J., Leung, K.Y. and Mok, Y.K. Temperature and mg2+ sensing by a novel phop-phoq two-component system for regulation of virulence in edwardsiella tarda. *J Biol Chem*, 285(50):38,876–88, 2010.

[31] Cheng, T., Liu, Z. and Wang, R. A knowledge-guided strategy for improving the accuracy of scoring functions in binding affinity prediction. *BMC Bioinformatics*, 11:193, 2010.

[32] Cherepanov, P.P. and Wackernagel, W. Gene disruption in escherichia coli: Tcr and kmr cassettes with the option of flp-catalyzed excision of the antibiotic-resistance determinant. *Gene*, 158(1):9–14, 1995.

[33] Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25(11):1422–1423, 2009.

[34] Cock, P.J.A. and Whitworth, D.E. Evolution of prokaryotic two-component system signaling pathways: gene fusions and fissions. *Molecular Biology and Evolution*, 24(11):2355–2357, 2007.

[35] Conradi, C., Flockerzi, D. and Raisch, J. Multistationarity in the activation of a mapk: parametrizing the relevant region in parameter space. *Math Biosci*, 211(1):105–31, 2008.

[36] Conradi, C., Flockerzi, D., Raisch, J. and Stelling, J. Subnetwork analysis reveals dynamic features of complex (bio)chemical networks. *Proc Natl Acad Sci U S A*, 104(49):19,175–80, 2007.

[37] Dailey, F.E. and Berg, H.C. Change in direction of flagellar rotation in escherichia coli mediated by acetate kinase. *J Bacteriol*, 175(10):3236–9, 1993.

[38] Danese, P.N., Snyder, W.B., Cosma, C.L., Davis, L.J. and Silhavy, T.J. The cpx two-component signal transduction pathway of escherichia coli regulates transcription of the gene specifying the stress-inducible periplasmic protease, degp. *Genes Dev*, 9(4):387–98, 1995.

[39] Dasika, M.S. and Maranas, C.D. Optcircuit: an optimization based method for computational design of genetic circuits. *BMC Syst Biol*, 2:24, 2008.

[40] Elderkin, S., Jones, S., Schumacher, J., Studholme, D. and Buck, M. Mechanism of action of the escherichia coli phage shock protein pspa in repression of the aaa family transcription factor pspf. *J Mol Biol*, 320(1):23–37, 2002.

[41] Feng, X.J., Hooshangi, S., Chen, D., Li, G., Weiss, R. and Rabitz, H. Optimizing genetic circuits by global sensitivity analysis. *Biophys J*, 87(4):2195–202, 2004.

[42] Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P. *et al.* The pfam protein families database. *Nucleic Acids Res*, 38(Database issue):D211–22, 2010.

[43] Fisher, S.L., Kim, S.K., Wanner, B.L. and Walsh, C.T. Kinetic comparison of the specificity of the vancomycin resistance vansfor two response regulators, vanr and phob. *Biochemistry*, 35(15):4732–40, 1996.

[44] Fortman, J.L., Chhabra, S., Mukhopadhyay, A., Chou, H., Lee, T.S., Steen, E. and Keasling, J.D. Biofuel alternatives to ethanol: pumping the microbial well. *Trends Biotechnol*, 26(7):375–81, 2008.

[45] François, P. and Hakim, V. Design of genetic networks with specified functions by evolution in silico. *Proc Natl Acad Sci U S A*, 101(2):580–5, 2004.

[46] Galperin, M.Y. A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial iq, extroverts and introverts. *BMC Microbiol*, 5:35, 2005.

[47] Galperin, M.Y., Higdon, R. and Kolker, E. Interplay of heritage and habitat in the distribution of bacterial signal transduction systems. *Mol Biosyst*, 6(4):721–8, 2010.

[48] Gao, R. and Stock, A.M. Biological insights from structures of two-component proteins. *Annual Review of Microbiology*, 63:133–154, 2009.

[49] Georgellis, D., Kwon, O., De Wulf, P. and Lin, E.C. Signal decay through a reverse phosphorelay in the Arc two-component signal transduction system. *The Journal of biological chemistry*, 273(49):32,864–32,869, 1998.

[50] Georgellis, D., Kwon, O. and Lin, E.C. Quinones as the redox signal for the arc two-component system of bacteria. *Science*, 292(5525):2314–6, 2001.

[51] Georgellis, D., Lynch, A.S. and Lin, E.C. In vitro phosphorylation study of the arc two-component signal transduction system of Escherichia coli. *Journal of bacteriology*, 179(17):5429–5435, 1997.

[52] Grimshaw, C.E., Huang, S., Hanstein, C.G., Strauch, M.A., Burbulys, D., Wang, L., Hoch, J.A. and Whiteley, J.M. Synergistic kinetic interactions between components of the phosphorelay controlling sporulation in bacillus subtilis. *Biochemistry*, 37(5):1365–75, 1998.

[53] Groban, E.S., Clarke, E.J., Salis, H.M., Miller, S.M. and Voigt, C.A. Kinetic buffering of cross talk between bacterial two-component sensors. *J Mol Biol*, 390(3):380–93, 2009.

[54] Groisman, E.A. and Mouslim, C. Sensing by bacterial regulatory systems in host and non-host environments. *Nat Rev Microbiol*, 4(9):705–9, 2006.

[55] Guest, J.R. and Shaw, D.J. Molecular cloning of menaquinone biosynthetic genes of escherichia coli k12. *Mol Gen Genet*, 181(3):379–83, 1981.

[56] Hafner, M., Koeppl, H., Hasler, M. and Wagner, A. 'glocal' robustness analysis and model discrimination for circadian oscillators. *PLoS Comput Biol*, 5(10):e1000,534, 2009.

[57] Hanna, J.H., Saha, K. and Jaenisch, R. Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. *Cell*, 143(4):508–25, 2010.

[58] Haynes, C., Oldfield, C.J., Ji, F., Klitgord, N., Cusick, M.E., Radivojac, P., Uversky, V.N., Vidal, M. *et al.* Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol*, 2(8):e100, 2006.

[59] Hill, A.D., Tomshine, J.R., Weeding, E.M.B., Sotiropoulos, V. and Kaznessis, Y.N. Synbioss: the synthetic biology modeling suite. *Bioinformatics*, 24(21):2551–3, 2008.

[60] Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L. *et al.* Copasi–a complex pathway simulator. *Bioinformatics*, 22(24):3067–74, 2006.

[61] Huang, K.J. and Igo, M.M. Identification of the bases in the ompf regulatory region, which interact with the transcription factor ompr. *J Mol Biol*, 262(5):615–28, 1996.

[62] Huvet, M., Toni, T., Sheng, X., Thorne, T., Jovanovic, G., Engl, C., Buck, M., Pinney, J.W. *et al.* The evolution of the phage shock protein response system: interplay between protein function, genomic organization, and system function. *Mol Biol Evol*, 28(3):1141–55, 2011.

[63] Huvet, M., Toni, T., Jovanovic, G., Engl, C., Buck, M. and Stumpf, M.P.H. Model-based evolutionary analysis: the natural history of phage-shock stress response. *Biochemical Society Transactions*, 37(Pt 4):762–767, 2009.

[64] Igo, M.M., Ninfa, A.J., Stock, J.B. and Silhavy, T.J. Phosphorylation and dephosphorylation of a bacterial transcriptional activator by a transmembrane receptor. *Genes Dev*, 3(11):1725–34, 1989.

[65] Ishige, K., Nagasawa, S., Tokishita, S. and Mizuno, T. A novel device of bacterial signal transducers. *The EMBO journal*, 13(21):5195–5202, 1994.

[66] Iuchi, S. arcA (dye), a global regulatory gene in Escherichia coli mediating repression of enzymes in aerobic pathways. *Proceedings of the National Academy of . . .*, 1988.

[67] Iuchi, S., Cameron, D.C. and Lin, E.C. A second global regulator gene (arcb) mediating repression of enzymes in aerobic pathways of escherichia coli. *J Bacteriol*, 171(2):868–73, 1989.

[68] Iuchi, S. and Lin, E.C. Mutational analysis of signal transduction by ArcB, a membrane sensor protein responsible for anaerobic repression of operons involved in the central aerobic pathways in Escherichia coli. *Journal of bacteriology*, 174(12):3972–3980, 1992.

[69] Jin, T. and Inouye, M. Ligand binding to the receptor domain regulates the ratio of kinase to phosphatase activities of the signaling domain of the hybrid escherichia coli transmembrane receptor, taz1. *J Mol Biol*, 232(2):484–92, 1993.

[70] Jourlin, C., Ansaldi, M. and Méjean, V. Transphosphorylation of the TorR response regulator requires the three phosphorylation sites of the TorS unorthodox sensor in Escherichia coli. *Journal of molecular biology*, 267(4):770–777, 1997.

[71] Jovanovic, G., Engl, C. and Buck, M. Physical, functional and conditional interactions between ArcAB and phage shock proteins upon secretin-induced stress in Escherichia coli. *Molecular microbiology*, 74(1):16–28, 2009.

[72] Karimova, G., Pidoux, J., Ullmann, A. and Ladant, D. A bacterial two-hybrid system based on a reconstituted signal transduction pathway. *Proc Natl Acad Sci U S A*, 95(10):5752–6, 1998.

[73] Karimova, G., Dautin, N. and Ladant, D. Interaction network among escherichia coli membrane proteins involved in cell division as revealed by bacterial two-hybrid analysis. *J Bacteriol*, 187(7):2233–43, 2005.

[74] Kato, A., Tanabe, H. and Utsumi, R. Molecular characterization of the phop-phoq two-component system in escherichia coli k-12: identification of extracellular mg2+-responsive promoters. *J Bacteriol*, 181(17):5516–20, 1999.

[75] Kato, M., Aiba, H., Tate, S., Nishimura, Y. and Mizuno, T. Location of phosphorylation site and dna-binding site of a positive regulator, ompr, involved in activation of the osmoregulatory genes of escherichia coli. *FEBS Lett*, 249(2):168–72, 1989.

[76] Katoh, K., Asimenos, G. and Toh, H. Multiple alignment of dna sequences with mafft. *Methods Mol Biol*, 537:39–64, 2009.

[77] Katoh, K. and Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics*, 9(4):286–298, 2008.

[78] Katoh, K. and Toh, H. Parallelization of the mafft multiple sequence alignment program. *Bioinformatics*, 26(15):1899–900, 2010.

[79] Keenan, K.G. and Valero-Cuevas, F.J. Experimentally valid predictions of muscle force and emg in models of motor-unit function are most sensitive to neural properties. *J Neurophysiol*, 98(3):1581–90, 2007.

[80] Keller, A.D. Specifying epigenetic states with autoregulatory transcription factors. *J Theor Biol*, 170(2):175–81, 1994.

[81] Kierzek, A.M., Zhou, L. and Wanner, B.L. Stochastic kinetic model of two component system signalling reveals all-or-none, graded and mixed mode stochastic switching responses. *Molecular BioSystems*, 6(3):531–542, 2009.

[82] Kim, J.R. and Cho, K.H. The multi-step phosphorelay mechanism of unorthodox two-component systems in E. coli realizes ultrasensitivity to stimuli while maintaining robustness to noises. *Computational biology and chemistry*, 30(6):438–444, 2006.

[83] Kitagawa, M., Ara, T., Arifuzzaman, M., Ioka-Nakamichi, T., Inamoto, E., Toyonaga, H. and Mori, H. Complete set of orf clones of escherichia coli aska library

(a complete set of e. coli k-12 orf archive): unique resources for biological research. *DNA Res*, 12(5):291–9, 2005.

[84] Kobayashi, H., Kaern, M., Araki, M., Chung, K., Gardner, T.S., Cantor, C.R. and Collins, J.J. Programmable cells: interfacing natural and engineered gene networks. *Proc Natl Acad Sci U S A*, 101(22):8414–9, 2004.

[85] Komorowski, M., Costa, M.J., Rand, D.A. and Stumpf, M.P.H. Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proc Natl Acad Sci U S A*, 108(21):8645–50, 2011.

[86] Koonin, E.V. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, 39:309–38, 2005.

[87] Koretke, K., Lupas, A., Warren, P., Rosenberg, M. and Brown, J. Evolution of Two-Component Signal Transduction. *Molecular Biology and Evolution*, 17(12):1956, 2000.

[88] Kwon, O., Georgellis, D. and Lin, E. Phosphorelay as the Sole Physiological Route of Signal Transmission by the Arc Two-Component System of Escherichia coli. *Journal of bacteriology*, 182(13):3858, 2000.

[89] Lambden, P.R. and Guest, J.R. Mutants of escherichia coli k12 unable to use fumarate as an anaerobic electron acceptor. *J Gen Microbiol*, 97(2):145–60, 1976.

[90] Lambert, D.G. Signal transduction: G proteins and second messengers. *Br J Anaesth*, 71(1):86–95, 1993.

[91] Laskowski, R.A., Hutchinson, E.G., Michie, A.D., Wallace, A.C., Jones, M.L. and Thornton, J.M. Pdbsum: a web-based database of summaries and analyses of all pdb structures. *Trends Biochem Sci*, 22(12):488–90, 1997.

[92] Laskowski, R.A. Pdbsum new things. *Nucleic Acids Res*, 37(Database issue):D355–9, 2009.

[93] Lau, P.C., Wang, Y., Patel, A., Labbé, D., Bergeron, H., Brousseau, R., Konishi, Y. and Rawlings, M. A bacterial basic region leucine zipper histidine kinase regulating toluene degradation. *Proceedings of the National Academy of Sciences of the United States of America*, 94(4):1453–1458, 1997.

[94] Li, W. and Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–9, 2006.

[95] Liang, W., Shores, M., Bockrath, M., Long, J. and Park, H. Kondo resonance in a single-molecule transistor. *Nature*, 417(6890):725–729, 2002.

[96] Liepe, J., Barnes, C., Cule, E., Erguler, K., Kirk, P., Toni, T. and Stumpf, M.P.H. Abc-sysbio–approximate bayesian computation in python with gpu support. *Bioinformatics*, 26(14):1797–9, 2010.

[97] Lopez, G., Valencia, A. and Tress, M. Firedb–a database of functionally important residues from proteins of known structure. *Nucleic Acids Res*, 35(Database issue):D219–23, 2007.

[98] López, G., Valencia, A. and Tress, M.L. firestar–prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res*, 35(Web Server issue):W573–7, 2007.

[99] Lu, T.K., Khalil, A.S. and Collins, J.J. Next-generation synthetic gene networks. *Nat Biotechnol*, 27(12):1139–50, 2009.

[100] Lynch, A.S. and Lin, E.C. Transcriptional control mediated by the arca two-component response regulator protein of escherichia coli: characterization of dna binding at target promoters. *J Bacteriol*, 178(21):6238–49, 1996.

[101] Macarthur, B.D., Ma'ayan, A. and Lemischka, I.R. Systems biology of stem cell fate and cellular reprogramming. *Nat Rev Mol Cell Biol*, 10(10):672–81, 2009.

[102] Maeda, T., Wurgler-Murphy, S.M. and Saito, H. A two-component system that regulates an osmosensing MAP kinase cascade in yeast. *Nature*, 369(6477):242–245, 1994.

[103] Malpica, R., Franco, B., Rodriguez, C., Kwon, O. and Georgellis, D. Identification of a quinone-sensitive redox switch in the arcb sensor kinase. *Proceedings of the National Academy of Sciences of the United States of America*, 101(36):13,318–13,323, 2004.

[104] Malpica, R., Sandoval, G.R.P., Rodríguez, C., Franco, B. and Georgellis, D. Signaling by the arc two-component system provides a link between the redox state of

the quinone pool and gene expression. *Antioxidants & redox signaling*, 8(5-6):781–795, 2006.

[105] Marchisio, M.A. and Stelling, J. Computational design of synthetic gene circuits with composable parts. *Bioinformatics*, 24(17):1903–10, 2008.

[106] Martin, V.J.J., Pitera, D.J., Withers, S.T., Newman, J.D. and Keasling, J.D. Engineering a mevalonate pathway in escherichia coli for production of terpenoids. *Nat Biotechnol*, 21(7):796–802, 2003.

[107] Matsushika, A. and Mizuno, T. Characterization of three putative sub-domains in the signal-input domain of the ArcB hybrid sensor in Escherichia coli(1). *Journal of Biochemistry*, 127(5):855–860, 2000.

[108] Meganathan, R. Biosynthesis of menaquinone (vitamin k2) and ubiquinone (coenzyme q): a perspective on enzymatic mechanisms. *Vitam Horm*, 61:173–218, 2001.

[109] Mendes, P., Hoops, S., Sahle, S., Gauges, R., Dada, J. and Kummer, U. Computational modeling of biochemical networks using copasi. *Methods Mol Biol*, 500:17–59, 2009.

[110] Mika, F. and Hengge, R. A two-component phosphotransfer network involving ArcB, ArcA. *genesdev.cshlp.org*, 2009.

[111] Miller, J.H. *A short course in bacterial genetics: a laboratory manual and handbook for Escherichia coli and related bacteria*. Cold Spring Harbor Laboratory Press, Plainview, N.Y., 1992.

[112] Mitrophanov, A.Y. and Groisman, E.A. Signal integration in bacterial two-component regulatory systems. *Genes Dev*, 22(19):2601–11, 2008.

[113] Mitrophanov, A.Y., Hadley, T.J. and Groisman, E.A. Positive autoregulation shapes response timing and intensity in two-component signal transduction systems. *Journal of Molecular Biology*, 401(4):671–680, 2010.

[114] Mizuno, T. Two-component phosphorelay signal transduction systems in plants: from hormone responses to circadian rhythms. *Bioscience, biotechnology, and biochemistry*, 69(12):2263–2276, 2005.

[115] Moreno-Hagelsieb, G. and Latimer, K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics (Oxford, England)*, 24(3):319–324, 2008.

[116] Myers, C.J., Barker, N., Jones, K., Kuwahara, H., Madsen, C. and Nguyen, N.P.D. iBioSim: a tool for the analysis and design of genetic circuits. *Bioinformatics (Oxford, England)*, 25(21):2848–2849, 2009.

[117] Newman, J.R.S. and Keating, A.E. Comprehensive identification of human bzip interactions with coiled-coil arrays. *Science*, 300(5628):2097–101, 2003.

[118] Nikaido, H. and Vaara, M. Molecular basis of bacterial outer membrane permeability. *Microbiol Rev*, 49(1):1–32, 1985.

[119] Ninfa, E.G., Atkinson, M.R., Kamberov, E.S. and Ninfa, A.J. Mechanism of autophosphorylation of escherichia coli nitrogen regulator ii (nrii or ntrb): trans-phosphorylation between subunits. *J Bacteriol*, 175(21):7024–32, 1993.

[120] Nobeli, I., Favia, A.D. and Thornton, J.M. Protein promiscuity and its implications for biotechnology. *Nat Biotechnol*, 27(2):157–67, 2009.

[121] Ohlsson, I., Nordström, B. and Brändén, C.I. Structural and functional similarities within the coenzyme binding domains of dehydrogenases. *J Mol Biol*, 89(2):339–54, 1974.

[122] Olivier, B.G., Rohwer, J.M. and Hofmeyr, J.H.S. Modelling cellular processes with python and scipy. *Mol Biol Rep*, 29(1-2):249–54, 2002.

[123] Pagel, M. and Meade, A. Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo. *The American naturalist*, 167(6), 2006.

[124] Pagel, M., Meade, A. and Barker, D. Bayesian estimation of ancestral character states on phylogenies. *Systematic biology*, 53(5):673–684, 2004.

[125] Pan, S.Q., Charles, T., Jin, S., Wu, Z.L. and Nester, E.W. Preformed dimeric state of the sensor protein vira is involved in plant–agrobacterium signal transduction. *Proc Natl Acad Sci U S A*, 90(21):9939–43, 1993.

[126] Papagiannis, M.D. Life-related aspects of stellar evolution. *Orig Life*, 14(1-4):43–50, 1984.

[127] Park, H. and Inouye, M. Mutational analysis of the linker region of envz, an osmosensor in escherichia coli. *J Bacteriol*, 179(13):4382–90, 1997.

[128] Park, H., Saha, S.K. and Inouye, M. Two-domain reconstitution of a functional protein histidine kinase. *Proc Natl Acad Sci U S A*, 95(12):6728–32, 1998.

[129] Parkinson, J.S. Bacterial chemotaxis: a new player in response regulator dephosphorylation. *J Bacteriol*, 185(5):1492–4, 2003.

[130] Pawelczyk, S., Scott, K.A., Hamer, R., Blades, G., Deane, C.M. and Wadhams, G.H. Predicting inter-species cross-talk in two-component signalling systems. *PLoS One*, 7(5):e37,737, 2012.

[131] Pena-Sandoval, G. and Georgellis, D. The ArcB Sensor Kinase of Escherichia coli Autophosphorylates by an Intramolecular Reaction. *Journal of bacteriology*, 192(6):1735, 2010.

[132] Perron-Savard, P., De Crescenzo, G. and Le Moual, H. Dimerization and dna binding of the salmonella enterica phop response regulator are phosphorylation independent. *Microbiology*, 151(Pt 12):3979–87, 2005.

[133] Pinney, J.W., Amoutzias, G.D., Rattray, M. and Robertson, D.L. Reconstruction of ancestral protein interaction networks for the bZIP transcription factors. *Proceedings of the National Academy of Sciences of the United States of America*, 104(51):20,449–20,453, 2007.

[134] Ponting, C.P. and Aravind, L. PAS: a multifunctional domain family comes to light. *Current biology : CB*, 7(11):R674–7, 1997.

[135] Porter, S.L., Wadhams, G.H. and Armitage, J.P. Signal processing in complex chemotaxis pathways. *Nat Rev Microbiol*, 9(3):153–65, 2011.

[136] Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I. and Sussman, J.L. Foldindex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, 21(16):3435–8, 2005.

[137] Prost, L.R., Daley, M.E., Le Sage, V., Bader, M.W., Le Moual, H., Klevit, R.E. and Miller, S.I. Activation of the bacterial sensor kinase phoq by acidic ph. *Mol Cell*, 26(2):165–74, 2007.

[138] Purnick, P.E.M. and Weiss, R. The second wave of synthetic biology: from modules to systems. *Nat Rev Mol Cell Biol*, 10(6):410–22, 2009.

[139] Qin, L., Dutta, R., Kurokawa, H., Ikura, M. and Inouye, M. A monomeric histidine kinase derived from envz, an escherichia coli osmosensor. *Mol Microbiol*, 36(1):24–32, 2000.

[140] Rabin, R.S., Collins, L.A. and Stewart, V. In vivo requirement of integration host factor for nar (nitrate reductase) operon expression in escherichia coli k-12. *Proc Natl Acad Sci U S A*, 89(18):8701–5, 1992.

[141] Rajendran, M. and Ellington, A.D. Selection of fluorescent aptamer beacons that light up in the presence of zinc. *Anal Bioanal Chem*, 390(4):1067–75, 2008.

[142] Rausell, A., Juan, D., Pazos, F. and Valencia, A. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc Natl Acad Sci U S A*, 107(5):1995–2000, 2010.

[143] Ro, D.K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A., Eachus, R.A. *et al.* Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440(7086):940–3, 2006.

[144] Robert, C.P., Cornuet, J.M., Marin, J.M. and Pillai, N.S. Lack of confidence in approximate bayesian computation model choice. *Proc Natl Acad Sci U S A*, 108(37):15,112–7, 2011.

[145] Rodrigo, G., Carrera, J. and Jaramillo, A. Genetdes: automatic design of transcriptional networks. *Bioinformatics*, 23(14):1857–8, 2007.

[146] Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M. *et al.* Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci U S A*, 103(22):8390–5, 2006.

[147] Root-Bernstein, R.S. Peptide self-aggregation and peptide complementarity as bases for the evolution of peptide receptors: a review. *Journal of molecular recognition : JMR*, 18(1):40–49, 2005.

[148] Rossmann, M.G., Moras, D. and Olsen, K.W. Chemical and biological evolution of nucleotide-binding protein. *Nature*, 250(463):194–9, 1974.

[149] Sahu, S.N., Acharya, S., Tuminaro, H., Patel, I., Dudley, K., LeClerc, J.E., Cebula, T.A. and Mukhopadhyay, S. The bacterial adaptive response gene, barA, encodes a novel conserved histidine kinase regulatory switch for adaptation and modulation of metabolism in Escherichia coli. *Molecular and cellular biochemistry*, 253(1-2):167–177, 2003.

[150] Savage, D.F., Way, J. and Silver, P.A. Defossiling fuel: how synthetic biology can transform biofuel production. *ACS Chem Biol*, 3(1):13–6, 2008.

[151] Sawers, G. and Suppmann, B. Anaerobic induction of pyruvate formate-lyase gene expression is mediated by the arca and fnr proteins. *J Bacteriol*, 174(11):3474–8, 1992.

[152] Seshasayee, A.S.N. and Luscombe, N.M. Comparative genomics suggests differential deployment of linear and branched signaling across bacteria. *Mol Biosyst*, 7(11):3042–9, 2011.

[153] Shin, D. and Groisman, E.A. Signal-dependent binding of the response regulators phop and pmra to their target promoters in vivo. *J Biol Chem*, 280(6):4089–94, 2005.

[154] Siryaporn, A. and Goulian, M. Cross-talk suppression between the cpxa-cpxr and envz-ompr two-component systems in e. coli. *Mol Microbiol*, 70(2):494–506, 2008.

[155] Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M. and Laub, M.T. Rewiring the specificity of two-component signal transduction systems. *Cell*, 133(6):1043–1054, 2008.

[156] Skerker, J.M., Prasol, M.S., Perchuk, B.S., Biondi, E.G. and Laub, M.T. Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. *PLoS biology*, 3(10):e334, 2005.

[157] Smolen, P., Baxter, D.A. and Byrne, J.H. Frequency selectivity, multistability, and oscillations emerge from models of genetic regulatory systems. *Am J Physiol*, 274(2 Pt 1):C531–42, 1998.

[158] Stewart, V., Parales, Jr, J. and Merkel, S.M. Structure of genes narl and narx of the nar (nitrate reductase) locus in escherichia coli k-12. *J Bacteriol*, 171(4):2229–34, 1989.

[159] Stiffler, M.A., Chen, J.R., Grantcharova, V.P., Lei, Y., Fuchs, D., Allen, J.E., Zaslavskaia, L.A. and MacBeath, G. Pdz domain binding selectivity is optimized across the mouse proteome. *Science*, 317(5836):364–9, 2007.

[160] Stock, A.M. Protein phosphorylation and regulation of adaptive responses in bacteria. *Microbiological reviews*, 53(4):450–490, 1989.

[161] Stock, A.M., Robinson, V.L. and Goudreau, P.N. Two-component signal transduction. *Annual review of biochemistry*, 69:183–215, 2000.

[162] Stock, J., Surette, M. and Levit, M. Two-component signal transduction systems: structure-function relationships and mechanisms of catalysis. *Two-component signal ...*, 1995.

[163] Stumpf, M.P.H. and Thorne, T. Multi-model inference of network properties from incomplete data. *J. Integrative Bioinformatics*, 3(2), 2006.

[164] Sureka, K., Ghosh, B., Dasgupta, A., Basu, J., Kundu, M. and Bose, I. Positive feedback and noise activate the stringent response regulator rel in mycobacteria. *PloS one*, 3(3):e1771, 2008.

[165] Surette, M.G., Levit, M., Liu, Y., Lukat, G., Ninfa, E.G., Ninfa, A. and Stock, J.B. Dimerization is required for the activity of the protein histidine kinase chea that mediates signal transduction in bacterial chemotaxis. *J Biol Chem*, 271(2):939–45, 1996.

[166] Swanson, R.V., Bourret, R.B. and Simon, M.I. Intermolecular complementation of the kinase activity of chea. *Mol Microbiol*, 8(3):435–41, 1993.

[167] Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K. and Yamanaka, S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5):861–72, 2007.

[168] TANG, B. Orthogonal array-based latin hypercubes. *Journal of the American Statistical Association*, 88(424):1392–1397, 1993.

[169] Tang, Y., Sheng, X. and Stumpf, M.P.H. The roles of contact residue disorder and domain composition in characterizing protein-ligand binding specificity and promiscuity. *Mol Biosyst*, 7(12):3280–6, 2011.

[170] Taylor, B.L. and Zhulin, I.B. PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiology and molecular biology reviews : MMBR*, 63(2):479–506, 1999.

[171] Thomason, P. and Kay, R. Eukaryotic signal transduction via histidine-aspartate phosphorelay. *J Cell Sci*, 113 (Pt 18):3141–50, 2000.

[172] Tiwari, A., Balázsi, G., Gennaro, M.L. and Igoshin, O.A. The interplay of multiple feedback loops with post-translational kinetics results in bistability of mycobacterial stress response. *Physical biology*, 7(3):036,005, 2010.

[173] Tiwari, A., Ray, J.C.J., Narula, J. and Igoshin, O.A. Bistable responses in bacterial genetic networks: designs and dynamical consequences. *Mathematical biosciences*, 231(1):76–89, 2011.

[174] Tomenius, H., Pernestig, A.K., Méndez-Catalá, C.F., Georgellis, D., Normark, S. and Melefors, O. Genetic and functional characterization of the Escherichia coli BarA-UvrY two-component system: point mutations in the HAMP linker of the BarA sensor give a dominant-negative phenotype. *Journal of bacteriology*, 187(21):7317–7324, 2005.

[175] Toni, T. and Stumpf, M.P.H. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics (Oxford, England)*, 26(1):104–110, 2010.

[176] Toni, T. Approximate bayesian computation for parameter inference and model selection in systems biology. *PhD thesis Imperial College London .*, 2010.

[177] Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M.P.H. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187, 2009.

[178] Tsai, C.J., Ma, B. and Nussinov, R. Protein-protein interaction networks: how can a hub protein bind so many different partners? *Trends Biochem Sci*, 34(12):594–600, 2009.

[179] Ulrich, L.E. and Zhulin, I.B. Mist: a microbial signal transduction database. *Nucleic Acids Res*, 35(Database issue):D386–90, 2007.

[180] Unden, G. Differential roles for menaquinone and demethylmenaquinone in anaerobic electron transport of e. coli and their fnr-independent expression. *Arch Microbiol*, 150(5):499–503, 1988.

[181] Unden, G., Achebach, S., Holighaus, G., Tran, H.G., Wackwitz, B. and Zeuner, Y. Control of fnr function of escherichia coli by o2 and reducing conditions. *J Mol Microbiol Biotechnol*, 4(3):263–8, 2002.

[182] van Heeckeren, W.J., Sellers, J.W. and Struhl, K. Role of the conserved leucines in the leucine zipper dimerization motif of yeast GCN4. *Nucleic acids research*, 20(14):3721–3724, 1992.

[183] Velankar, S., McNeil, P., Mittard-Runte, V., Suarez, A., Barrell, D., Apweiler, R. and Henrick, K. E-msd: an integrated data resource for bioinformatics. *Nucleic Acids Res*, 33(Database issue):D262–5, 2005.

[184] Verhamme, D.T., Arents, J.C., Postma, P.W., Crielaard, W. and Hellingwerf, K.J. Investigation of in vivo cross-talk between key two-component systems of escherichia coli. *Microbiology*, 148(Pt 1):69–78, 2002.

[185] Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. and Jones, D.T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*, 337(3):635–45, 2004.

[186] Wilhelm, T. The smallest chemical reaction system with bistability. *BMC Systems Biology*, 3:90, 2009.

[187] Wilkinson, R.D. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Biometrika*, 2008.

[188] Williams, R.H. and Whitworth, D.E. The genetic organisation of prokaryotic two-component system signalling pathways. *BMC genomics*, 11(1):720, 2010.

[189] Wilson, D., Madera, M., Vogel, C., Chothia, C. and Gough, J. The superfamily database in 2007: families and functions. *Nucleic Acids Res*, 35(Database issue):D308–13, 2007.

[190] Wilson, T.H. and Lin, E.C. Evolution of membrane bioenergetics. *J Supramol Struct*, 13(4):421–46, 1980.

[191] Wistrand, M. and Sonnhammer, E.L.L. Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. *BMC bioinformatics*, 6:99, 2005.

[192] Wolfe, A.J. and Stewart, R.C. The short form of the chea protein restores kinase activity and chemotactic ability to kinase-deficient mutants. *Proc Natl Acad Sci U S A*, 90(4):1518–22, 1993.

[193] Yaku, H. and Mizuno, T. The membrane-located osmosensory kinase, EnvZ, that contains a leucine zipper-like motif functions as a dimer in Escherichia coli. *FEBS letters*, 417(3):409–413, 1997.

[194] Yamamoto, K., Hirao, K., Oshima, T., Aiba, H., Utsumi, R. and Ishihama, A. Functional characterization in vitro of all two-component signal transduction systems from Escherichia coli. *The Journal of biological chemistry*, 280(2):1448–1456, 2005.

[195] Yang, Y. and Inouye, M. Requirement of both kinase and phosphatase activities of an escherichia coli receptor (taz1) for ligand-dependent signal transduction. *J Mol Biol*, 231(2):335–42, 1993.

[196] Yang, Y., Park, H. and Inouye, M. Ligand binding induces an asymmetrical transmembrane signal through a receptor dimer. *J Mol Biol*, 232(2):493–8, 1993.

[197] You, L., Cox, 3rd, R.S., Weiss, R. and Arnold, F.H. Programmed population control by cell-cell communication and regulated killing. *Nature*, 428(6985):868–71, 2004.

[198] Zarrinpar, A., Park, S.H. and Lim, W.A. Optimization of specificity in a cellular protein interaction network by negative selection. *Nature*, 426(6967):676–80, 2003.

[199] Zhou, Y., Liepe, J., Sheng, X., Stumpf, M.P.H. and Barnes, C. GPU accelerated biochemical network simulation. *Bioinformatics (Oxford, England)*, 27(6):874–876, 2011.

[200] Zhu, Y., Qin, L., Yoshida, T. and Inouye, M. Phosphatase activity of histidine kinase envz without kinase catalytic domain. *Proc Natl Acad Sci U S A*, 97(14):7808–13, 2000.

[201] Zhulin, I.B., Taylor, B.L. and Dixon, R. PAS domain S-boxes in Archaea, Bacteria and sensors for oxygen and redox. *Trends in Biochemical Sciences*, 22(9):331–333, 1997.

[202] Zio, E. *The Monte Carlo simulation method for system reliability and risk analysis.* Springer, New York, 2012.