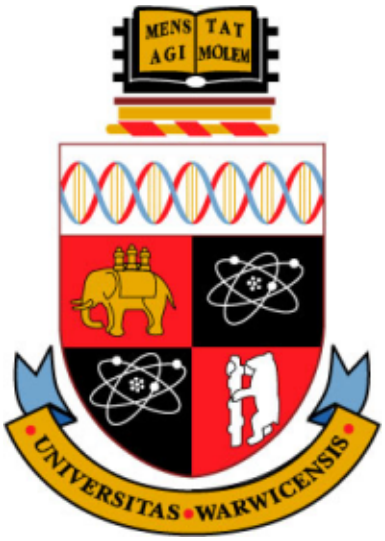


Two-component regulation: modelling, predicting & identifying protein-protein interactions & assessing signalling networks of bacteria

by

Peter J. A. Cock

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

MOAC Doctoral Training Centre

August 2008

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank the following:

- My supervisors, *Dr. David Whitworth* and *Dr. Bärbel Finkenstädt* for their help and assistance.
- *Prof. Alison Rodger* for founding and directing the MOAC Doctoral Training Centre.
- My family and in particular my parents, *Dr. Matthew Cock* and *Dr. Josephine Cock*, for their encouragement to embark on this PhD, and support during it.
- The Engineering and Physical Sciences Research Council (EPSRC) for funding.
- My PhD advisors *Prof. David Hodgson* and *Prof. David Rand* for their advice and general education from their research group meetings.
- My PhD advisory committee members *Prof. Dave Scanlan*, *Dr Hugo van den Berg* and *Dr. Andrew Mead* for keeping an eye on my progress.
- The Centre for Scientific Computing, University of Warwick, for computation resources.
- My fellow students at MOAC for friendship and company, including Yi Chan for his proof-reading, and MOAC administrators *Dr. Dorothea Mangels* and *Monica Lucena*.

And last, but not least, *H.K.* for her company, support, proof-reading, and much more . . . ♡.

Declarations

The author declares that, to the best of his knowledge, the data contained within this thesis is original and his own work under the supervision of Dr. David E. Whitworth and Dr. Bärbel Finkenstädt.

The material in this thesis is submitted for the degree of PhD to the University of Warwick only and has not been submitted to any other university. All sources of information have been specifically acknowledged in the form of references.

As noted in the main text where appropriate, several papers and a book chapter have been published based on this work:

- Cock, P. J. A. and Whitworth, D. E. (2007). Evolution of gene overlaps: relative reading frame bias in prokaryotic two-component system genes. *Journal Of Molecular Evolution*, **64**(4), 457–462.

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

- Whitworth, D. E. and Cock, P. J. A. (2008). Myxobacterial two-component systems. In *Myxobacteria: Multicellularity and Differentiation*, chapter 10. Ed. D.E. Whitworth, ASM Press, Washington DC.

- Whitworth, D. E. and Cock, P. J. A. (2008). Two-component systems of the myxobacteria: Structure, diversity and evolutionary relationships. *Microbiol.*, **154**(2), 360–372.

Abstract

Two-component signalling systems (TCSs) are found in most prokaryotic genomes. They typically comprise of two proteins, a histidine (or sensor) kinase (HK) and an associated response regulator (RR), containing transmitter and receiver domains respectively, which interact to achieve transfer of a phosphoryl group from a histidine residue (of the transmitter domain in the HK) to an aspartate residue (of the partner RR's receiver domain).

An automated analysis pipeline using the NCBI's RPS-BLAST tool was developed to identify and classify all TCS genes from completed prokaryotic genomes using the PFAM and CDD protein domain databases.

A large proportion of TCS genes were found to be simple hybrid kinases (HYs) containing both a transmitter domain and a receiver domain within a single protein, presumably the result of the fusion or combination of separate HK and RR genes. This propensity to consolidate functionality into a single protein was found to be limited in the presence of either a transmembrane sensory/input domain or a DNA binding domain – two spatially separated functions.

While HK and RR genes are usually found together in the genome, in some species a large proportion of TCS domains are found as part of complex hybrid kinases (genes containing multiple TCS domains), in isolated or orphaned genes, or in complex gene clusters. In such organisms the lack of paired HK and RR genes makes it difficult to define genome-encoded signalling networks.

Identifying paired transmitter and receiver domains from a pan-genomic survey of prokaryotes gives a database of amino acid sequences for thousands of interacting protein-protein complexes. Covariation between columns of multiple sequence alignments (MSAs) identifies particular pairs of residues representing interactions within the docked complex. Using numerical scores, these amino acids pairs were successfully used as explanatory variables in a generalised linear model (GLM) to predict the probabilities of interaction between transmitter and receiver domains.

Abbreviations

AIC	Akaike Information Criterion
bp	base pair(s)
CP	Chemical potential
GLM	Generalised linear model
H	Hpt domain (see Table 2.1)
HATPase	Histidine adenosine-triphosphatase domain
HisKA	Histidine Kinase A (phosphoacceptor) domain
HK	Histidine (or sensor) kinase (a TCS gene/protein)
Hpt	Histidine-containing phosphotransfer domain
HW	Hopp-Woods (hydrophilicity)
HY	Hybrid kinase (a TCS gene/protein)
K	HisKA domain (see Table 2.1)
KD	Kyte-Doolittle (hydrophilicity)
MI	Mutual information
MSA	Multiple sequence alignment
T_i	Class I transmitter domain
T_{ii}	Class II transmitter domain
TCS	Two component signalling (systems)
TM	Transmembrane
R	Receiver domain (see Table 2.1)
ROC	Receiver operator characteristic (curve or plot)
RR	Response regulator (a TCS gene/protein)
Y2H	Yeast two-hybrid (assay)

Species abbreviations:

- *Escherichia coli* abbreviated to *E. coli*
- *Myxococcus xanthus* abbreviated to *M. xanthus*

Chapter 1

Introducing two component signalling

1.1 Brief outline

This thesis focuses on the prediction of protein-protein interactions within or between two-component signalling (TCS) systems. TCS systems and their multistage variant, the phosphorelay, comprise the majority of prokaryotic signal pathways (Stock *et al.*, 2000; Robinson *et al.*, 2000; Hoch, 2000), and they regulate a wide variety of cellular process including motility, heat-shock responses, sporulation, antibiotic production and virulence (Hoch and Silhavy, 1995; Atkinson and Ninfa, 1999). TCS are also found in yeast (Maeda *et al.*, 1994) and plants (Mizuno, 2005), but not in animals (Thomason and Kay, 2000), which makes them a possible target of anti-microbial compounds.

A typical TCS system comprises two proteins, a histidine (or sensor) kinase (HK) and an associated response regulator (RR). Upon detecting its input signal, the HK will activate its partner RR by a protein-protein interaction which transfers a phosphoryl group. The phosphorylated RR will then elicit a response. Using two proteins in this way allows a spatial separation of the input signal's detection and the triggered response, typically linking an external environmental cue to gene regulation.

Most organisms employing TCS systems will have multiple HK and RR pairs, which could potentially interact with each other. Such signal cross-talk may be undesirable, and can be controlled by a combination of TCS interaction specificity (the focus of this study), phosphatase activity by the RR's true partner HK (Laub and Goulian, 2007), and spatial and temporal segregation of the proteins. Knowing the full TCS network for an organism will explain how the various HK input domains are connected to the RR output domains, and thus is important for understanding its adaptive behaviour.

Some HK and RR pairings can be inferred from their genome organisation, by iden-

tifying neighbouring genes in an operon. However, in some species such as *Nostoc* sp. and *Myxococcus xanthus*, a large proportion of TCS genes are complex hybrid kinases (HYs), isolated or orphaned genes, or in complex gene clusters (Whitworth and Cock, 2008a,b). In such organisms, the lack of simple paired HK and RR genes makes it difficult to define genome-encoded signalling networks. Thus, being able to predict TCS domain partnerships from genome sequences would be especially useful. This is the main aim of this thesis.

1.2 Transmitters and receivers

A typical TCS comprises two proteins, an HK and associated RR. Both types of protein are modular, with discrete domains of defined function.

HKs contain an N-terminal input domain with a C-terminal transmitter domain (Parkinson and Kofoid, 1992). Upon receiving a stimulus through its input domain, the transmitter domain hydrolyses ATP into ADP (using a histidine kinase-type ATPase (HATPase) domain within the transmitter) and auto-phosphorylates at a conserved histidine residue. Transmitter domains form dimers, and are believed to phosphorylate in trans (i.e. a protein phosphorylates its dimer partner, see Cai *et al.* (2003) and the references therein). Bilwes *et al.* (1999) classified two types of transmitter. In Class I (or orthodox) transmitter domains, the phospho-accepting histidine residue is in a HisKA motif (histidine kinase), which is also the site of the HK dimerisation. Class II (or unorthodox) transmitter domains are usually associated with motility[1]. Instead of using a HisKA motif, the phospho-accepting histidine is part of an Hpt motif (histidine-containing phosphotransfer) which is generally followed by a separate dimerisation domain. These transmitter domains are illustrated within Figures 1.1 and 1.2, respectively, and for brevity will be written as T_i and T_{ii}.

Transmitters are generally thought to form homodimers, and the HATPase of one monomer will phosphorylate the HisKA domain of the other monomer. Brencic *et al.* (2004) have demonstrated this trans-phosphorylation within the T_i dimer using VirA from *Agrobacterium tumefaciens*, part of a complex system discussed in Section 1.4.10. Current models of the Rcs system from *Escherichia coli* suggest that some transmitters could form heterodimers (see Section 1.4.7), but this has yet to be confirmed.

RRs also typically comprise two functionally separate domains, an N-terminal receiver domain and a C-terminal output domain. Upon docking with a phosphorylated HK, the phosphoryl group is transferred from the transmitter histidine residue onto a conserved aspartate residue within the receiver domain (His→Asp). Phosphorylation of the receiver domain causes

[1] Flagella biosynthesis in *Rhodospirillum centenum* is a counter example (Berleman and Bauer, 2005).

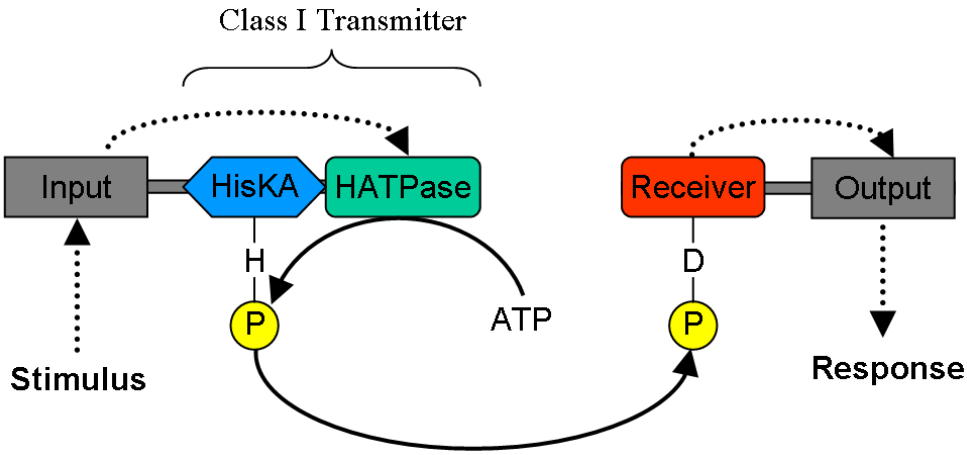


Figure 1.1: Simple two gene His→Asp TCS system consisting of an HK on the left containing a Class I (orthodox) transmitter (T_i, made up of a HisKA, in blue, and HATPase, in green), and RR on the right containing a receiver domain (in red). The blocks represent the linear domain structure of each protein, although not to scale, with the N-terminus on the left. The dimerisation of the HK is not shown. Dotted arrows show information flow, solid arrows show the phosphotransfer. When the HATPase domain uses ATP to phosphorylate the histidine residue, ADP is released, which is not shown.

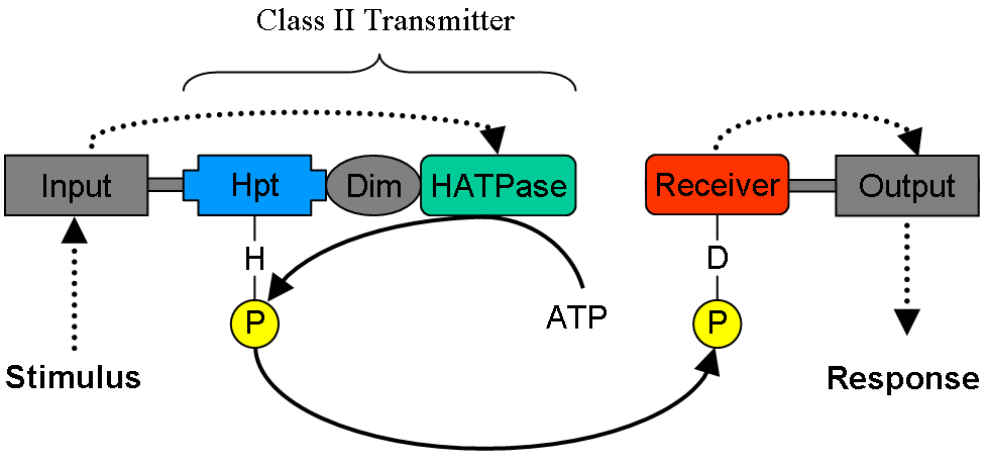


Figure 1.2: Simple two gene His→Asp TCS system consisting of an HK on the left containing a Class II (unorthodox) transmitter (T_{ii}, made up of Hpt, in blue, dimerisation region, in grey, and HATPase, in green), and RR on the right containing a receiver domain (in red). The dimerisation of the HK is not shown. *cf.* Figure 1.1

a conformational change (Lewis *et al.*, 1999), which results in an altered activity of the output domain, thus giving an output response to the initial stimulus. This output domain is frequently DNA-binding, controlling gene expression levels. In many RRs there is no separate output domain, and the protein is presumed to have a direct action based on a further interaction of the receiver domain, or to be part of a phosphorelay (discussed in the following section).

Figures 1.1 and 1.2 show typical HK and RR pairs (with Class I and Class II transmitters respectively), illustrating the normal domain arrangement within each gene. The dotted arrows show the flow of information, with solid arrows showing the phosphotransfer.

HK and RR genes do not necessarily interact in isolation, there is often additional regulation – for example from phosphatases which are enzymes that remove phosphate groups. In fact, many HK genes have a phosphatase activity from the HisKA domain (Zhu *et al.*, 2000), and can dephosphorylate their partner RR. This suggests retro-phosphorylation Asp→His from the RR receiver to the HK transmitter may be possible. The use of an Asp→His transfer is well established as the basis of the TCS phosphorelay systems discussed next.

1.3 TCS system architectures

The TCS transmitter (and its constituents) and receiver domains are modular, and are frequently found in different combinations. For example, in addition to the conventional two-gene TCS systems illustrated in Figures 1.1 and 1.2, there are numerous examples of hybrid kinases (HYs) containing both a transmitter and receiver, combining the functionality of the HK and RR into a single protein. This is illustrated in Figure 1.3, and specific examples are discussed later in this chapter. Interestingly, most eukaryotic TCS genes are hybrid kinases (Koretke *et al.*, 2000). Far more elaborate combinations of domains within a single gene have also been found and are believed to have evolved several times independently (Zhang and Shi, 2005).

The example TCS systems shown thus far (Figures 1.1, 1.2 and 1.3) all have a single phosphotransfer, His→Asp, from a transmitter to receiver. In some organisms, an Asp→His transfer from receiver to Hpt domain has been co-opted to extend the TCS system into a phosphorelay His→Asp→His→Asp (Appleby *et al.*, 1996). Figure 1.4 shows several possible His→Asp→His→Asp relays using the same basic domains. Note that in all known cases the Asp→His transfer is from a receiver to an Hpt domain (and not a HisKA domain).

The most numerous examples of phosphorelays in the literature use two separate proteins, a tripartite HY containing a Class I transmitter, receiver and Hpt domain, and a normal RR, denoted T_i-R-H + R. Some theoretical work suggests these tripartite HY genes are more

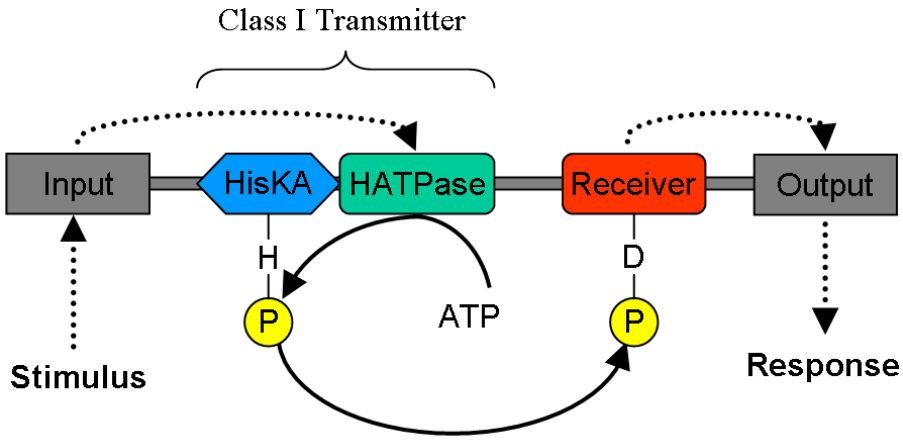


Figure 1.3: Simple one gene His→Asp TCS system consisting of an HY containing a Class I (orthodox) transmitter (T_i, made up of a HisKA, in blue, and HATPase, in green), and receiver domain (in red). Colour scheme as in Figure 1.1, where the same domains appear in two separate genes.

sensitive to their input signal, and robust to noise (Kim and Cho, 2006), which may explain their relative abundance. These, and the handful of known examples of phosphorelays with three or more genes, are discussed in the next section.

Hereafter TCS proteins will often be written in a shorthand notation. Any TCS domains within a gene are listed in the N to C-terminal order, separated with dashes. Thus for example, T_i-R denotes a simple hybrid kinase (HY) with a Class I transmitter (T_i) N-terminal to receiver domain (R), while R-T_i denotes a HY with the domain order reversed. Plus signs are used to indicate separate genes, for example T_i + R for a simple HK and RR pair, or T_i-R-H + R for a tripartite HY and RR.

1.4 Exemplar systems

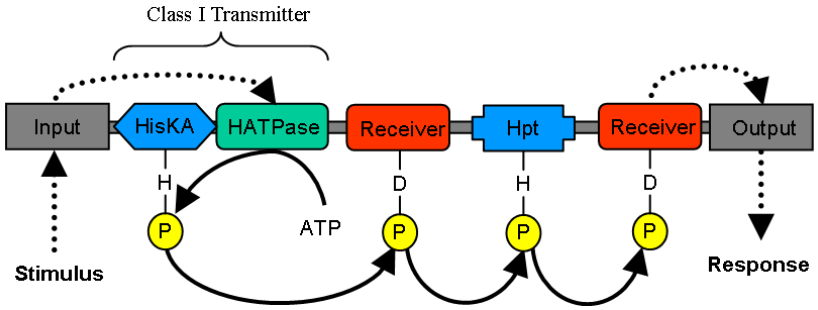
In this section a range of TCS systems from the literature are described, focusing on the canonical examples and known complex systems which illustrate particular properties of interest. Some proteins from these systems have solved 3D-structures, which are discussed in Section 1.5.

1.4.1 EnvZ/OmpR osmoregulation in *Escherichia coli*

Escherichia coli survives in both fresh water and in the gut of animals where solute concentrations (osmolarity) are much higher, and does this by adjusting the type of pores in its outer membrane (Csonka and Hanson, 1991). This osmoregulation system is controlled by the

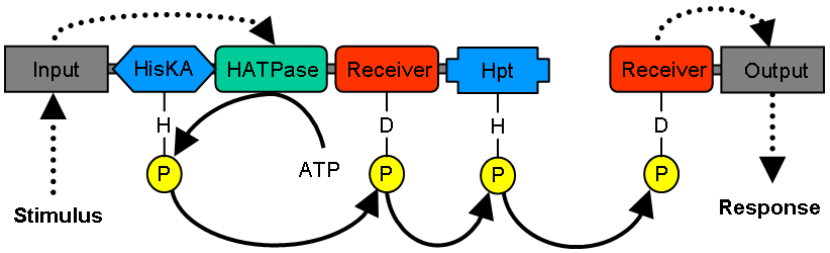
T_i-R-H-R
One gene relay

No examples documented in the literature



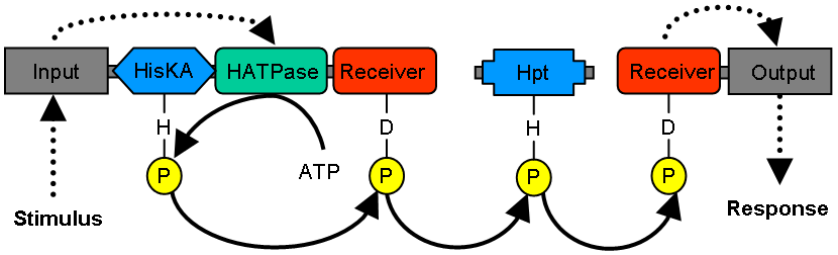
T_i-R-H + R
Two gene relay

e.g. ArcB/A in *E. coli*



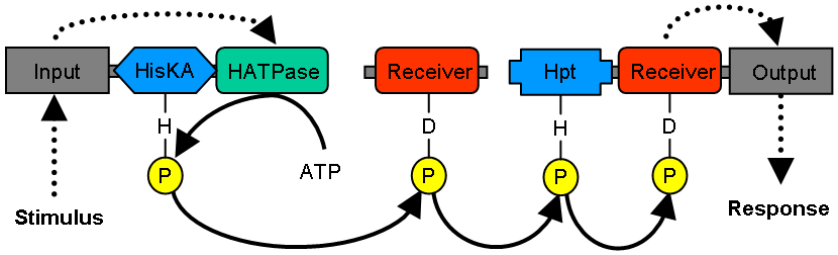
T_i-R + H + R
Three gene relay

e.g. RcsC/D/B in *E. coli*, and Sln1/Ypd1/Ssk1 in *Saccharomyces cerevisiae*.



T_i + R + H-R
Three gene relay

RcaE/F/C relay in the cyanobacteria *Fremyella diplosiphon* is similar.



T_i + R + H + R
Four gene relay

Sporulation in *Bacillus subtilis* is similar

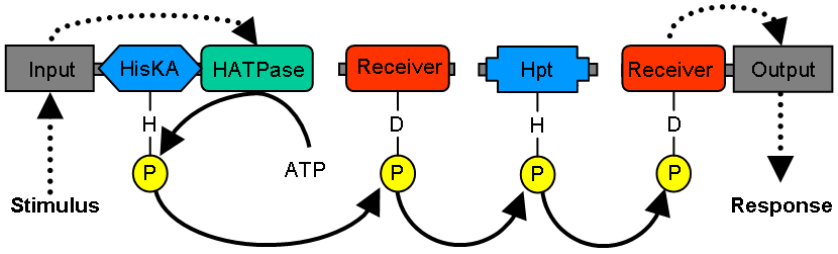


Figure 1.4: Example TCS phosphorelay systems, His→Asp→His→Asp, with one, two, three or four genes. The example systems mentioned are discussed in Section 1.4. Other variations are possible, such as the three gene relay T_i + R-H + R, but have not been reported in the literature. (*cf.* Appleby *et al.* (1996), Figure 3).

EnvZ/OmpR TCS system. This is a 'classical' TCS system with the flow of information from one HK to one RR (Figure 1.1), where the two proteins are encoded as neighbouring genes in an operon (Perraud *et al.*, 1999).

The HK EnvZ senses levels of osmolarity and phosphorylates its partner RR OmpR (i.e. His→Asp transfer), which is a transcription factor for the *omp*F and *omp*C genes which code for major outer membrane proteins (Mizuno *et al.*, 1988). OmpC is produced at high osmolarity, while OmpF is produced at low osmolarity (Dutta *et al.*, 2000). EnvZ also possesses phosphatase activity and can dephosphorylate phosphorylated OmpR (Aiba *et al.*, 1989; Zhu *et al.*, 2000).

EnvZ/OmpR is an archetypal TCS system, a simple HK and RR encoded in an operon as neighbouring genes. Furthermore, it is also typical in that it combines an N-terminal transmembrane input domain in the HK with a C-terminal DNA-binding output domain in the RR (see Chapter 4).

1.4.2 Nar regulatory TCS system of *E. coli*

The Nar system in *E. coli* regulates nitrate and nitrite metabolism, and is an example of a small TCS network (Rabin and Stewart, 1992, 1993). Encoded by neighbouring genes, NarX and NarL are an ordinary pair of HK and RR proteins. However, two further locations in the genome encode another HK, NarQ, and another RR, NarP. The TCS domains of these genes are more similar to each other than other TCS genes in *E. coli*, which helped in their identification. Both NarX and NarQ have transmembrane input domains, and will phosphorylate both NarL and NarP, which both have DNA-binding output domains (Figure 1.5). However, the interactions between these proteins are not fully symmetric, and they are also regulated differently (Darwin and Stewart, 1995).

This system illustrates that an HK and RR encoded in a single operon (such as NarX and NarL) may not be exclusive partners forming an isolated system, and furthermore that orphan TCS genes (such as NarQ or NarP) may interact with paired genes.

1.4.3 Chemotaxis in *E. coli*

Chemotaxis is directed movement in response to chemical stimuli, such as nutrient levels (see Adler (1975) for an early review). *E. coli*'s directed random walk is one of the best studied bacterial chemotaxis systems. *E. coli* cells have multiple flagella which, when rotated anti-clockwise, twist together forming a single rotating bundle, causing the bacteria to swim in a straight line. However, when rotated clockwise the bundle separates and each flagella points

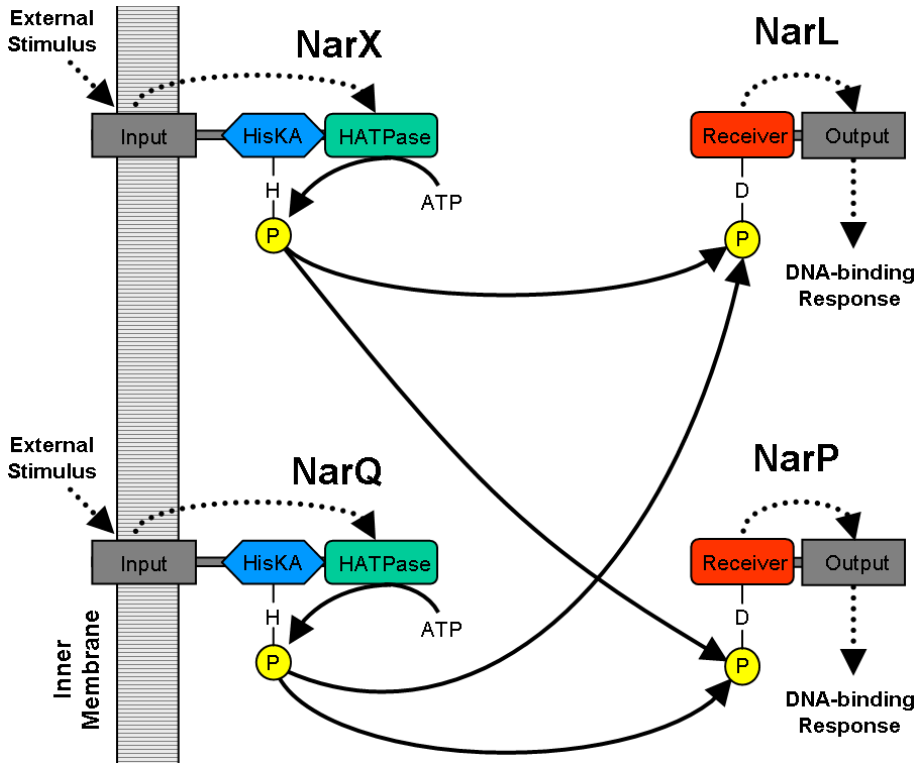


Figure 1.5: The *E. coli* Nar TCS network, discussed in Section 1.4.2, is the best studied example of cross-talk between two HKs and two RRs.

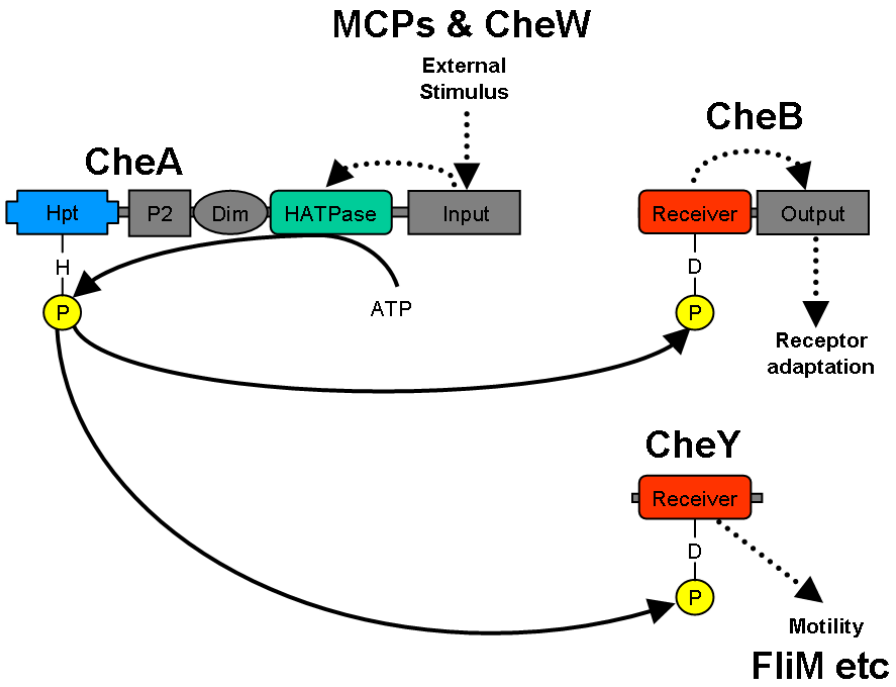


Figure 1.6: The *E. coli* chemotaxis TCS network, with one HK phosphorylating two RRs, discussed in Section 1.4.3. RR CheY is unusual in having no separate output domain.

in a random direction – resulting in an undirected tumbling at the mercy of diffusion and local currents (Macnab and Koshland, 1974). When tumbling the cell is randomly reorientated. Hence switching between anti-clockwise and clockwise rotation and back again gives a random change in direction, essentially a three dimensional random walk (Larsen *et al.*, 1974).

The mechanism *E. coli* uses to direct its random walk is very simple. If, while swimming, the levels of the attractant are increasing (or levels of a repellent are decreasing), then the cell should keep swimming. If conditions are getting worse, it should stop swimming and tumble in order to randomly select a new direction. To achieve this, the regulatory system can switch the rotation of the flagella (see RR CheY below), and detect if the attractant/repellent levels are increasing or decreasing. This "memory" works by adjusting the sensitivity of the ligand sensors, known as methyl-accepting chemotaxis proteins (MCPs), by modifying their level of methylation (Springer *et al.*, 1977; Goy *et al.*, 1977) (see RR CheB below).

The *E. coli* chemotaxis TCS system has one HK, CheA, and two RRs, CheY and CheB (Figure 1.6). CheA contains what Bilwes *et al.* (1999) called a Class II transmitter (T_{ii}), where the histidine phosphorylation site is within an Hpt domain rather than a HisKA domain (Figure 1.2). Unusually for an HK, the "input" domain of CheA is C-terminal. This region forms a complex with the CheW protein and the transmembrane MCPs, allowing CheA to indirectly receive extracellular signals.

Once stimulated, CheA self-phosphorylates its Hpt domain, which can then phosphorylate both the RRs CheB and CheY (Bourret and Stock, 2000; Wadhams and Armitage, 2004) (Figure 1.6). The second domain in HK CheY, denoted P2, has been shown to be unnecessary for this transfer, but by binding to CheB and CheY it increases the rate of phosphotransfer (Jahreis *et al.*, 2004).

The *che*A gene is unusual in having two distinct in-frame initiation sites, where the second possible start codon gives rise to a shorter protein CheAS (Smith and Parkinson, 1980). This short form CheAS lacks the N-terminal Hpt domain found in the full length CheA (or CheAL) described above, and therefore cannot be phosphorylated itself. While CheAS has been shown to be non-essential for chemotaxis (Sanatinia *et al.*, 1995), it can however phosphorylate the Hpt domain of CheA (Wolfe and Stewart, 1993; Wolfe *et al.*, 1994).

CheY has only a receiver domain and interacts directly with the flagellar proteins to control its direction of spin, achieving chemotaxis (Lowry *et al.*, 1994; Djordjevic *et al.*, 1998). This is probably the best understood RR without a separate output domain. CheB on the other hand has an output domain, a methylesterase controlling the methylation state of the MCPs, adjusting the sensitivity of the chemo-receptors (Stock and Koshland, 1978; Hayashi

et al., 1979; Yonekawa *et al.*, 1983).

The *E. coli* chemotaxis TCS system serves not only as an example of a simple one-to-many or divergent network, but also includes the exemplar T_{ii} containing HK, CheA, and output domain less RR, CheB.

1.4.4 Chemotaxis in *Rhodobacter sphaeroides*

The *E. coli* chemotaxis system, with its very simple set of TCS genes discussed above, appears to be a minimal system when compared to other bacteria – for one thing there is only a single type of flagellum which is regulated by a simple binary switch (the tumble/swim mechanism giving directed Brownian motion). Some bacteria have two flagellar systems, a polar flagellum for swimming and lateral flagella for swarming. Indeed, recent work suggests that some strains of *E. coli* have acquired a second flagellar system by horizontal gene transfer (Ren *et al.*, 2005). For recent reviews see Szurmant and Ordal (2004) and Wadhams and Armitage (2004). Given some bacteria have multiple motor systems, and may live in far less homogeneous environments than a mammalian gut, it is not surprising that the TCS networks controlling their movements can be much more complicated, for example *Sinorhizobium meliloti* (Schmitt, 2002), *Helicobacter pylori* (Jiménez-Pearson *et al.*, 2005) and *Myxococcus xanthus* (Li *et al.*, 2005).

Rhodobacter sphaeroides is a purple photosynthetic bacterium found in freshwater and marine environments. It has two different flagellar systems, a single subpolar flagellum (fla1) and multiple polar flagella (fla2) (Poggio *et al.*, 2007). In other species, dual flagellar systems allow motility in different environments (swimming and swarming) (McCarter, 2004), whereas here both systems appear to be used for swimming.

There are multiple homologues of the *E. coli* chemotaxis genes in *Rhodobacter sphaeroides*, and while many of these proteins have been shown to be involved in chemotaxis, it is possible that others are not. For example, in the related *Rhodospirillum centenum* which also has numerous *che*-homologues, Berleman and Bauer (2005) showed some controlled flagellar biosynthesis rather than chemotaxis.

Rhodobacter sphaeroides has three HKs homologous to *E. coli* CheA (CheA1, CheA2, and the special case of CheA3/CheA4), four CheW's, six RRs homologous to CheY (CheY1 through CheY6), two RRs homologous to CheB (CheB1 and CheB2), plus a more unusual protein dubbed CheBRA. Most of these genes are found in three loci. Focusing on the TCS and CheW genes only *che*Op1 contains *cheY5*, *cheY1*, *cheA1*, *cheW1* and *cheY2*; *che*Op2 contains *cheY3*, *cheA2*, *cheW2*, *cheW3* and *cheB1*; *che*Op3 contains *cheA4*, *cheB2*, *cheW4*,

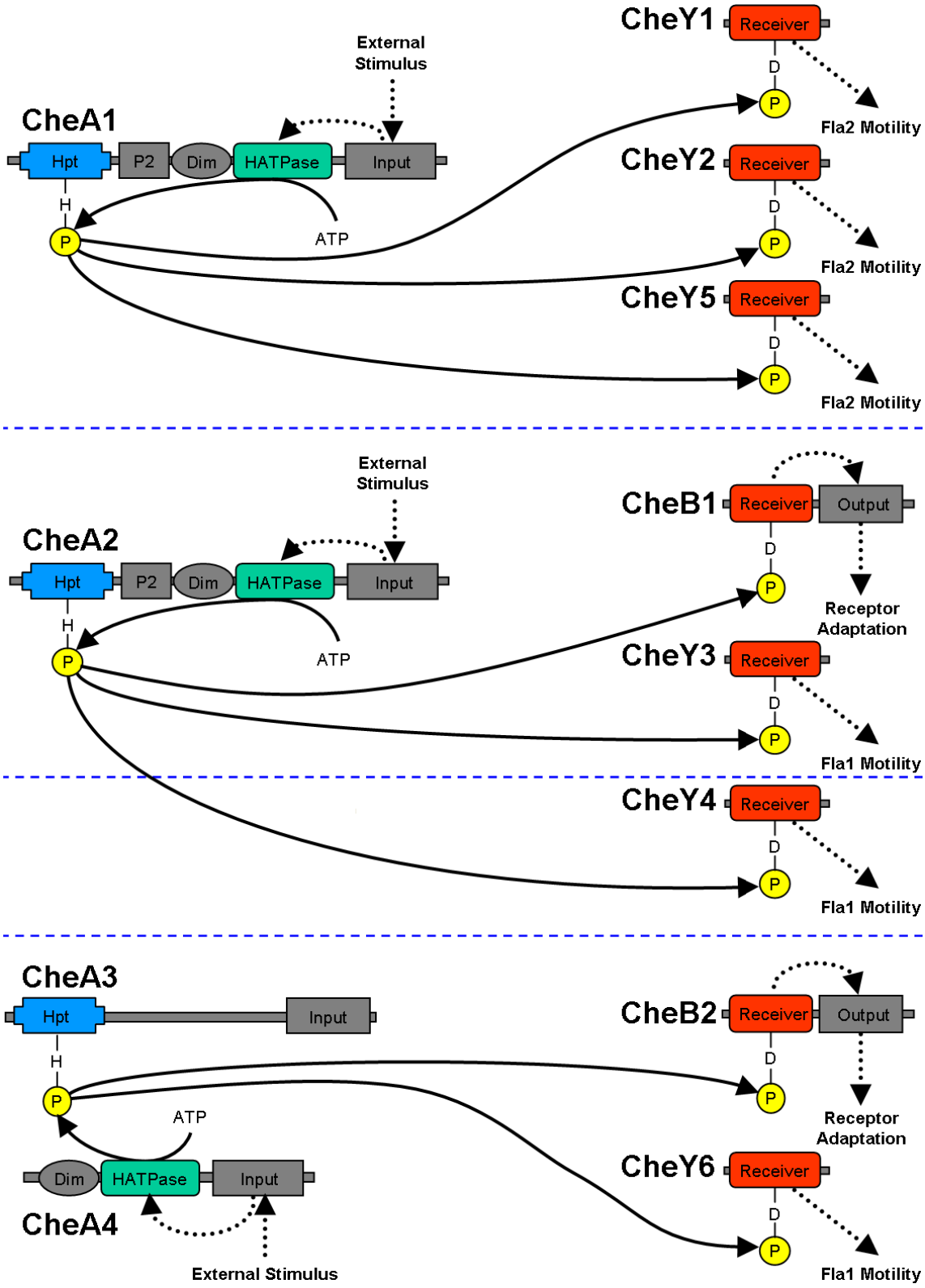


Figure 1.7: The *Rhodobacter sphaeroides* chemotaxis TCS system. Colour scheme as in other figures, but with dashed horizontal blue lines separating proteins from different operons (not spatial segregation). For figure clarity, none of the potential *in vitro* inter-operon phospho-transfers from Porter and Armitage (2004) Figure 6 are shown here (except CheA2 to CheY4, as CheA2 appears to be CheY4's only phospho-donor). *cf.* Figure 1.6 for *E. coli*.

cheY6 and *cheA3* (Porter *et al.*, 2002); with *cheY4* encoded on the small *mcpG* locus (Shah *et al.*, 2000). Note that the protein numbers do not correspond to the operon numbers (see Porter *et al.* (2002) Figure 1 for further details on the operon structure). These proteins are illustrated in Figure 1.7.

An additional small locus encodes locus tag RSP_2229, a gene dubbed cheBRA, as it resembles a CheB-CheW-CheA fusion (Mackenzie *et al.*, 2001; Porter *et al.*, 2002). However, while there are good matches to CheB and CheW proteins, the C-terminal region does not resemble a CheA T_{ii} domain, but rather a HisKA domain without an HATPase – suggesting this is a novel phosphotransfer protein. Immediately downstream of this, locus tag RSP_2230 encodes a RR which could be a seventh CheY homologue. To date, neither protein has been characterised in the literature, and any role they may have in chemotaxis remains to be explored.

One of the most interesting aspects of the *Rhodobacter sphaeroides* TCS systems is proteins CheA3 and CheA4 (both encoded in *che*Op3). These are essential for chemotaxis, and contain between them all five domains expected in CheA (as illustrated in Figure 1.6) (Porter *et al.*, 2002). These two proteins appear to function together as an HK where the HATPase of CheA4 phosphorylates the Hpt domain of CheA3 (Porter and Armitage, 2004), somewhat similar to the Rcs system in *E. coli* (Section 1.4.7). There are also even striking similarities to the trans-phosphorylation between the short and long forms of *E. coli* CheA described in Section 1.4.3.

Thus far, CheA1 appears to be non-essential for chemotaxis (Porter *et al.*, 2006). On the other hand, CheA2 is essential for aerotaxis, phototaxis, and chemotaxis (Martin *et al.*, 2001), and as noted above, CheA3/CheA4 is also essential for chemotaxis. Chemotaxis of the single subpolar flagellum (fla1) is controlled by CheY6 and either CheY3 or CheY4 (Porter *et al.*, 2006), while it was recently shown that the *che*Op1 RRs (CheY1, CheY2 and CheY5) control the multiple polar flagellar (fla2) system (del Campo *et al.*, 2007).

In vitro work has shown broad potential cross-talk between the three CheA HKs and the eight CheY or CheB like RRs (Porter and Armitage, 2004). It appears that *in vivo* there are two spatially separated chemotaxis systems controlling the polar flagellum (fla1). The *che*Op2 encoded proteins including CheA2, CheW2 and cheW3 are targeted to the cell poles, while *che*Op3 proteins CheA3/CheA4 and CheW4 are targeted to a cytoplasmic cluster, with RRs CheB1 and CheB2 throughout the cytoplasm (Martin *et al.*, 2003; Wadhams *et al.*, 2003). The RRs CheY3, CheY4 and CheY6 all appear relatively mobile but somewhat surprisingly, all three appear to associate with the cytoplasmic cluster (i.e. near CheA3/A4), but only CheY4

was shown to associate with the polar cluster (i.e. near CheA2) (Porter *et al.*, 2006). It would be interesting to see if these proteins are also segregated from the *che*Op1 system controlling fla2, or if the *in vitro* cross-talk between the systems also occurs *in vivo* which would suggest CheA2 has a key role as a master transmitter activating both flagellar systems.

Given the results of del Campo *et al.* (2007) tying the *che*Op1 RRs to the fla2 system, it now seems likely that the HK CheA1, which is also encoded on *che*Op1, does have a role in chemotaxis after all. On the basis of both the genome arrangement and the *in vitro* results reported in Porter and Armitage (2004), CheA1 might be expected to drive the fla2 system almost exclusively, and have little interaction with the fla1 system (except perhaps via CheY3).

The chemotaxis TCS genes in *Rhodobacter sphaeroides* serve as a model system where spatial segregation reduces cross-talk between otherwise compatible transmitters and receivers, which appear to be the result of both gene duplications and horizontal transfer.

1.4.5 Hybrid kinase systems in *Bacteroides*

Bacteroides thetaiotaomicron and *Bacteroides fragilis* are very unusual in having a number of HY genes (with T_i-R domains) which have both transmembrane input domains *and* DNA-binding output domains (Xu *et al.*, 2004) (discussed in more detail in Chapter 4). These systems presumably function as separate self-contained signalling pathways linking external stimuli to DNA regulatory responses (Figure 1.8).

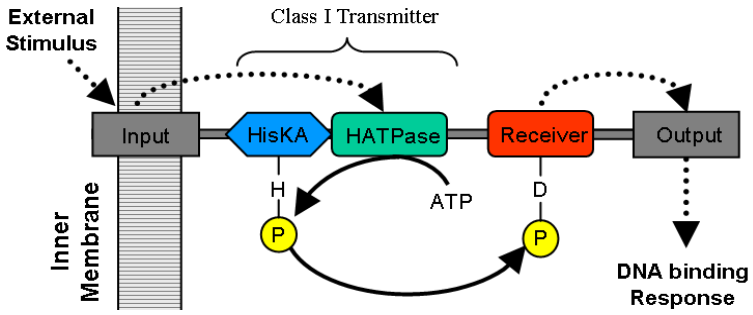


Figure 1.8: This is an example of a T_i-R hybrid kinase (HY), with a transmembrane input domain and a DNA-binding output domain. This combination is extremely rare, however multiple examples exist in *Bacteroides* species.

1.4.6 Phosphorelays with a tripartite HY and RR

All the examples so far have had single step phosphotransfers His→Asp from a transmitter to receiver domain. This section will describe examples of a His→Asp→His→Asp phosphorelay

using two genes, T_i-R-H + R, a tripartite HY containing a Class I transmitter, receiver and Hpt domain, with a normal RR containing one receiver, as illustrated in Figure 1.4.

Aerobic metabolism in *E. coli* is regulated by the ArcB/ArcA TCS system. The membrane sensor protein ArcB is a T_i-R-H tripartite HY (Iuchi *et al.*, 1990). Its partner ArcA is a cytoplamsic RR with a DNA-binding output domain and receiver domain. These genes are not located on the same operon, but it is now well established that this system is a two gene His→Asp→His→Asp phosphorelay (Tsuzuki *et al.*, 1995; Georgellis *et al.*, 1998; Matsushika and Mizuno, 1998; Kwon *et al.*, 2000).

When ArcB detects anaerobic growth conditions, the T_i auto-phosphorylates, then phosphorylates the receiver which phosphorylates the Hpt domain. Finally via an inter-protein interaction, the phosphoryl group is transfered from the ArcB Hpt to the ArcA receiver, which then regulates metabolic operons. Conversely, under aerobic conditions ArcB acts as a phosphatase and dephosphorylates ArcA, deactivating it (Kwon *et al.*, 2000; Malpica *et al.*, 2006).

The Hpt domain of ArcB is similar to that of CheA, and some work suggests that the Hpt domain of ArcB can also phosphorylate the receiver domain of CheY (Yaku *et al.*, 1997; Kato *et al.*, 1999), see Section 1.4.3. There is also some evidence that the RR ArcA is also phosphorylated by another HK, CpxA (Iuchi *et al.*, 1989), suggesting some cross-talk between the ArcB/ArcA and CpxA/CpxB TCS systems in *E. coli*.

In addition to the *E. coli* ArcB/ArcA phosphorelay described above, there are numerous other analogous T_i-R-H + R phosophorelay systems in the literature, such as the *E. coli* EvgA/EvgS system (Utsumi *et al.*, 1992, 1994; Tanabe *et al.*, 1998), and the BvgS/BvgA relay in *Bordetella pertussis* which controls virulence and is essential for the colonisation of the respiratory tract (Uhl and Miller, 1994, 1996). Another example is DorS/DorR, found in *Rhodobacter sphaeroides* encoded on neighbouring genes within chromosome II, but on opposite strands (Mouncey *et al.*, 1997; Mouncey and Kaplan, 1998).

The TosS/TorR system in *E. coli* (Jourlin *et al.*, 1999) is an interesting case as it appears that the His→Asp→His→Asp relay can run partly in reverse with HY TorS dephosphorylating RR TorR, an example of retro-phosphorylation (Ansaldi *et al.*, 2001) previously also suggested for ArcB/ArcA.

Finally, the GacS/GacA system is yet another T_i-R-H + R relay, found in Gram-negative bacteria, which is reviewed in Heeb and Haas (2001). What signal the HY GacS detects has yet to be determined, however the partner RR GacA controls the production of secondary metabolites and extracellular enzymes involved in pathogenicity to plants and animals, or antifungal activity. Interestingly, in no species were the two genes found in the same operon. GacS

was first found in *Pseudomonas syringae* with a lesion manifestation giving its original name of LemA (Hrabak and Willis, 1992), while in *Erwinia carotovora* the genes are known as ExpS (or RpfA) and ExpA. The inter-protein phosphotransfer has been demonstrated *in vitro* for BarA and UvrY, the *Escherichia coli* K12 homologues of GacS and GacA (Pernestig *et al.*, 2001).

Note that in *E. coli* alone, there are at least four T_i-R-H + R two gene His→Asp→His→Asp relays, ArcB/ArcA, EvgA/EvgS, TorS/TorR and BarA/UvrY, together making up over 10% of the TCS genes in this organism. In fact, this sort of TCS phosphorelay appears to be relatively common across the prokaryotes (see Chapter 2).

1.4.7 RcsC/RcsD/RcsB phosphorelay in *E. coli*

The RcsC/RcsD/RcsB system in *E. coli* is a three gene His→Asp→His→Asp phosphorelay, similar to the T_i-R + H + R system shown in Figure 1.4 (Takeda *et al.*, 2001; Clarke *et al.*, 2002). While the three genes are all found in the same region of the genome in *E. coli* K12, rcsC is on the opposite strand (Clarke *et al.*, 2002, Figure 2(A)). Homologous genes have been identified in other bacteria, and the system is believed to be specific to enteric pathogens/commensals (Erickson and Detweiler, 2006).

What is particularly interesting in this system is the second gene, RcsD previously known as YojN, which functions as the Hpt containing phosphotransfer protein in the phospho-cascade. It also has a non-functional T_i domain which is similar to that of RcsC but lacks the conserved histidine residue. Both RcsC and RcsD have transmembrane domains, leading Takeda *et al.* (2001) to suggest that they form a heterodimer, as illustrated in Figure 1.9. The figure shows the HATPase domain of RcsD phosphorylating the HisKA domain of RcsC, based on analogy to trans-phosphorylation in normal HK homodimers. It is not yet clear if this is the case, or indeed which input domain(s) trigger the initial phosphorylation. RcsC is understood to phosphorylate the RR RcsB, which then acts with a co-factor RcsA to bind to DNA, targeting regions known as RcsAB boxes (Wehland and Bernhard, 2000).

The Rcs system demonstrates how surprising the TCS systems can be, and shows that even transmitter domains lacking their conversed phosphorylatable histidine residue may still play an active role.

1.4.8 Sporulation in *Bacillus subtilis*

When placed under environmental stress, many bacteria will form resilient spores allowing them to wait for conditions to improve. The timing of this is extremely important, too late and there may not be enough nutritional resources to complete spore formation. On the other

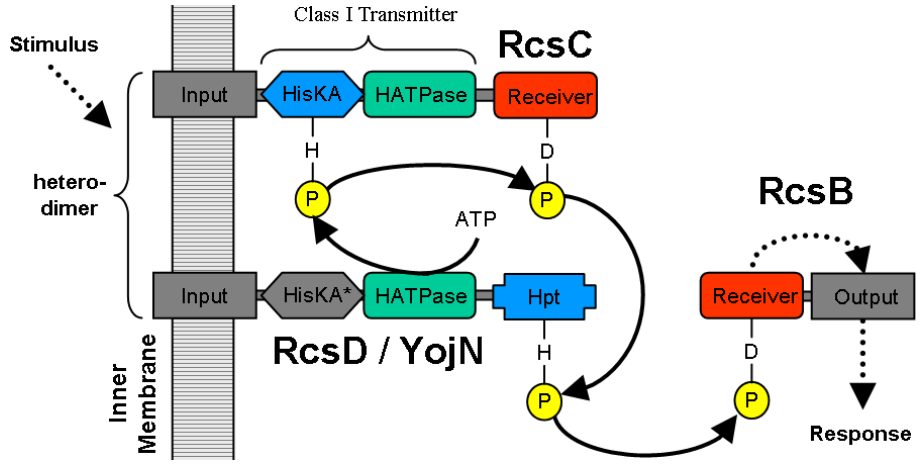


Figure 1.9: The *E. coli* Rcs TCS relay, discussed in Section 1.4.7. RcsC is presumed to form a dimer with RcsD/YojN, with the HATPase phosphorylating in trans. The colour scheme is as in previous figures, but the HisKA domain in RcsD/YojN lacking its histidine phospholaytion site is shown in grey and marked with an asterisk. Based on Takeda *et al.* (2001) Figure 8.

hand, sporulating too early does not take full advantage of limited resources, and runs the risk of being out-competed by other cells.

The TCS systems of *Bacillus subtilis* are reviewed in Fabret *et al.* (1999). The sporulation (Spo) system of *Bacillus subtilis* is one of the most studied complex TCS networks of any bacteria (Figure 1.10). There are at least five HK proteins feeding various signals into a master regulator via a phosphorelay (Hoch, 1993; Piggot and Hilbert, 2004; Barák *et al.*, 2005). The master regulator is Spo0A, a typical RR containing an N-terminal receiver domain and a C-terminal DNA-binding domain. Dependent on its phosphorylation state, Spo0A influences over 500 genes directly or indirectly (Molle *et al.*, 2003).

Early work on the *Bacillus subtilis* Spo system focused on the relatively simple His→-Asp→His→Asp phosphorelay from proteins KinA to Spo0F to Spo0B to Spo0A (Burbulys *et al.*, 1991), like that illustrated in Figure 1.4. KinA is a normal HK with a single T_i domain, and both Spo0F and Spo0A are normal RRs with a single R domain. What is particularly unusual is the histidine containing phosphotransfer protein Spo0B. This forms a dimer giving a four α-helix bundle structurally analogous to the HisKA dimer or Hpt domain, but is quite unique at the sequence level (Varughese *et al.*, 1998). The structure of the Spo0B/Spo0F complex was later solved (Zapf *et al.*, 2000), and shows the receiver residues involved are highly conserved between Spo0F and Spo0A suggesting they interact with Spo0B in the same way, consistent with earlier mutagenesis results (Tzeng and Hoch, 1997). See Section 1.5.

It is now established that at least five HK proteins will phosphorylate Spo0F, namely KinA, KinB, KinC, KinD and KinE (Jiang *et al.*, 2000), allowing the Spo system to integrate

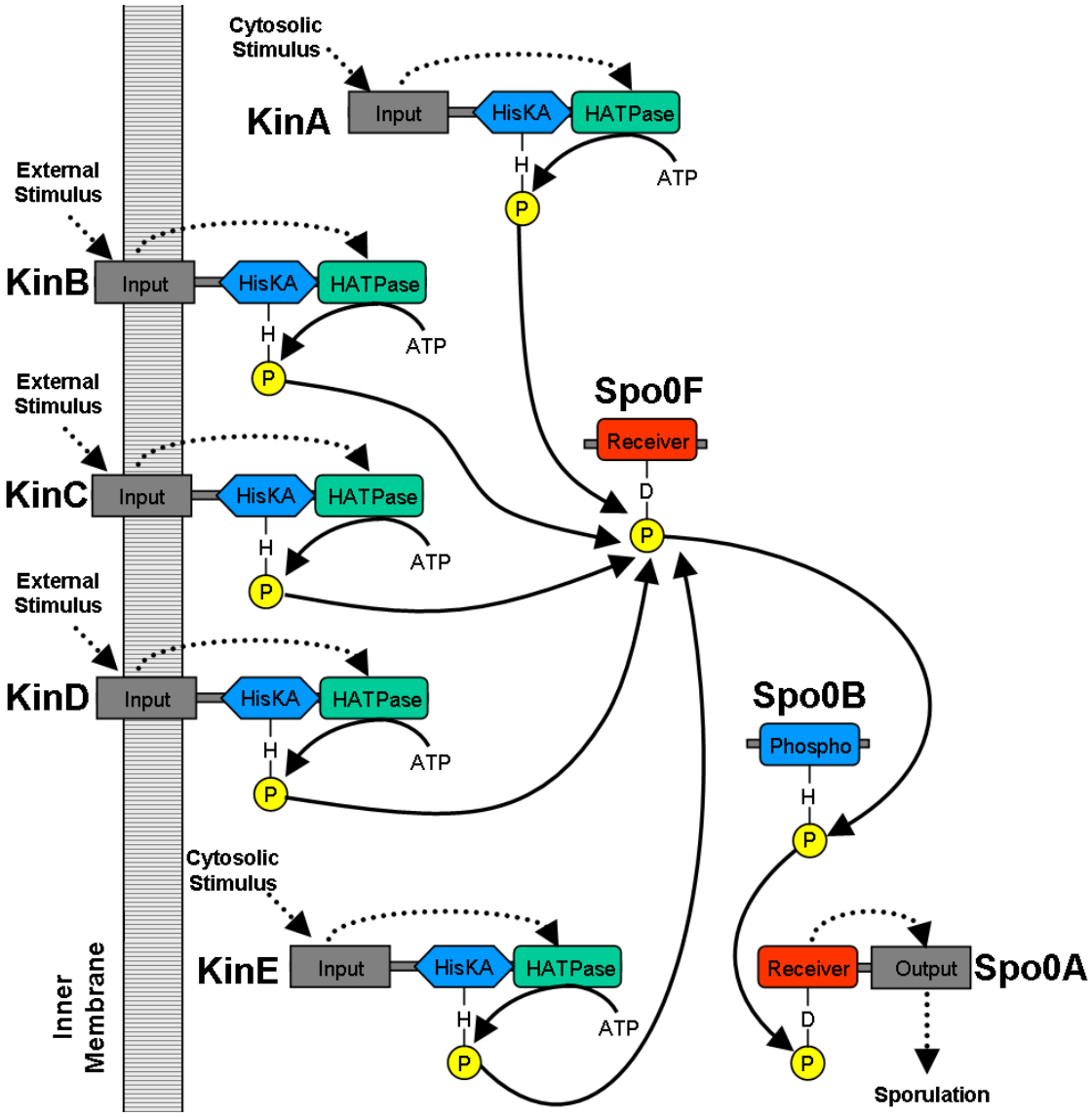


Figure 1.10: The *Bacillus subtilis* sporulation TCS network, with five HKs feeding into a phosphorelay, discussed in Section 1.4.8. The dimerisation of the HKs and Spo0B is not shown. The colour scheme is as in previous figures, with the unique phosphotransfer domain of Spo0B also shown in blue. Additional non-TCS proteins are involved in this pathway but not shown here, for example see Piggot and Hilbert (2004) Figure 2.

multiple input signals (both internal to the cell and external via transmembrane input domains). In the case of KinA, recent work has elucidated the mechanism by which the extra-cellular sensor domain controls the self-phosphoylation of the HisKA domain (Lee *et al.*, 2008)

A broad range of other proteins also intract with the system, for example Sda is known to prevent the self-phosphorylation of KinA in the event of DNA damage or replication defects (Rowland *et al.*, 2004), and Spo0E will de-phosphorylate Spo0A (Ohlsen *et al.*, 1994). For a recent overview of the full sporulation system in *Bacillus subtilis* see Piggot and Hilbert (2004) Figure 2.

Stephenson and Hoch (2002) looked at the evolution of the Spo TCS system by a comparison of *Bacillus subtilis*, *B. halodurans*, *B. anthracis* and *B. stearothermophilus*. They found the *spo*0F, *spo*0B and *spo*0A genes to be present in all species. Homologues of the *kin*A-E HK genes could only be identified by their highly conserved transmitter domains, as there were dramatic differences in the input domains of these genes. It is therefore presumed that the different *Bacillus* species use the same basic TCS network topology, but integrate different input signals according to their own very different ecological niches. It seems likely that the changes of input domains occurred by recombination events.

1.4.9 Quorum-sensing in *Vibrio harveyi*

Bacteria communicate using extracellular signal molecules termed autoinducers in a process called quorum-sensing. The higher the local population level, the higher the levels of these autoinducer signal molecules. In many cases, this communication is within species, for example collective behaviours like sporulation or bioluminescence, although interspecies communication is also possible where multiple species produce and detect the same autoinducers (Schauder and Bassler, 2001; McNab and Lamont, 2003).

A model organism in this area is the bioluminescent marine bacterium *Vibrio harveyi*, which uses a TCS network to integrate three different quorum-sensing systems (Figure 1.11). In the first system, the HY LuxN detects autoinducer AI-1 (or HAI-1 for *harveyi* autoinducer one), which is produced by LuxM. In system two, the HY LuxQ together with LuxP detects the autoinducer AI-2, whose precursor is produced by LuxS (Freeman and Bassler, 1999). Finally, the HK CqsS detects the *cholerae* autoinducer CAI-1, which is produced by CqsA (Henke and Bassler, 2004). This third system is called the *cholerae* quorum-sensing (Cqs) system as it was initially identified in the related species *Vibrio cholerae*, which has a similar TCS network sharing many of the same genes (Miller *et al.*, 2002).

The three different autoinducers AI-1, AI-2 and CAI-1 are detected independently using

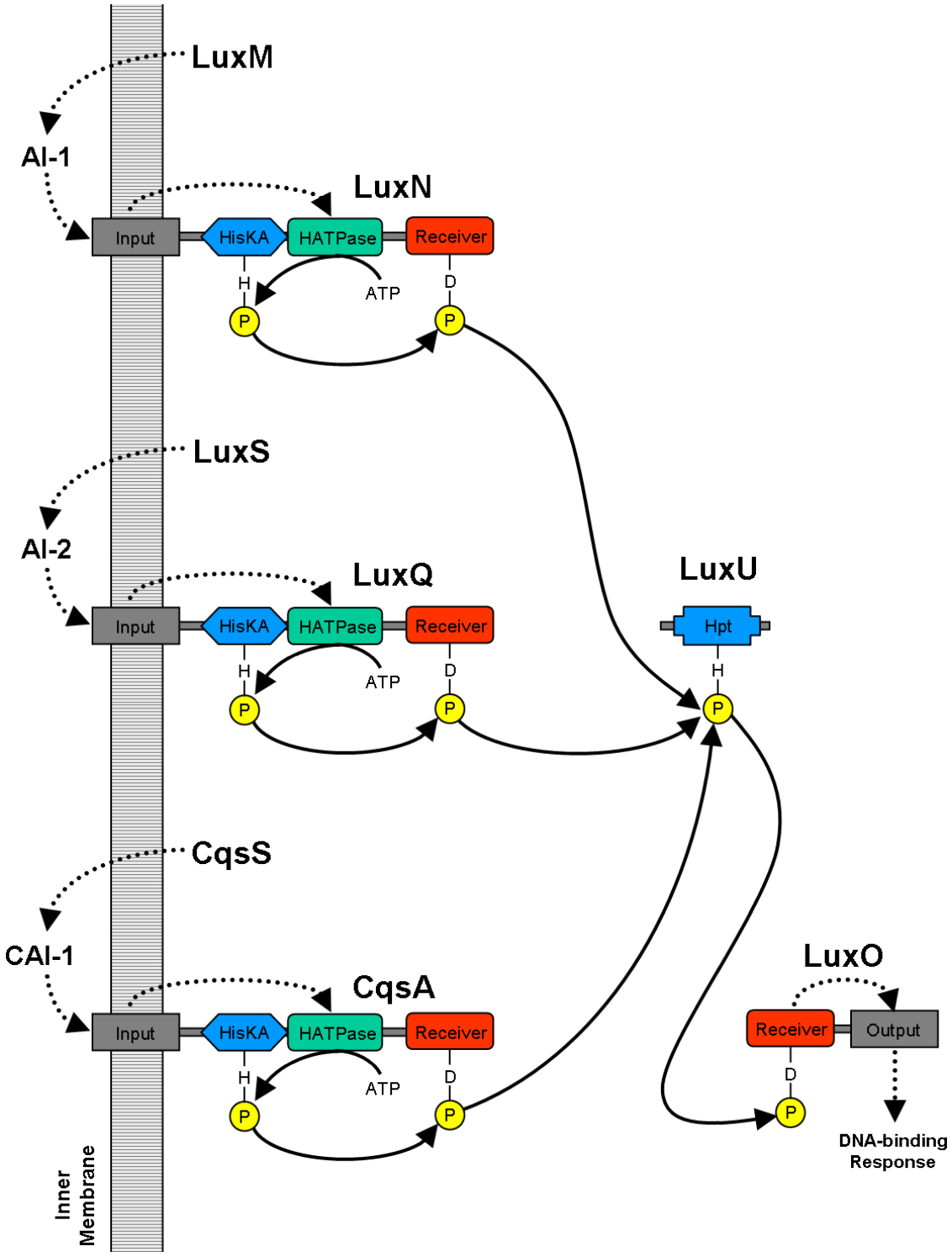


Figure 1.11: The *Vibrio harveyi* quorum-sensing TCS network, with three HYs feeding into a phosphorelay. Based on figures in Freeman and Bassler (1999); Henke and Bassler (2004); Timmen *et al.* (2006). Additional non-TCS proteins are involved in this network but not shown here. *c.f.* the *Vibrio cholerae* network shown in Miller *et al.* (2002) Figure 7.

three different transmembrane T_i-R HYs (LuxN, LuxQ and CqsS). All three HYs phosphorylate the Hpt containing protein LuxU, integrating the stimuli. LuxU will in turn phosphorylate the RR LuxO whose DNA-binding output domain regulates numerous other systems, including the bioluminescence genes (*lux*) (Freeman and Bassler, 1999; Henke and Bassler, 2004; Timmen *et al.*, 2006). The TCS network is therefore based on a T_i-R + H + R phosphorelay, but with multiple initial HYs.

Like the *Bacillus subtilis* sporulation TCS network described above, this is a many-to-one network, integrating multiple inputs into a single output response (in the form of gene regulation). Although both networks use a His→Asp→His→Asp relay, here the first phosphotransfer is intra-protein, whereas in *Bacillus subtilis* all three steps are inter-protein.

1.4.10 VirA/VigG virulence in *Agrobacterium tumefaciens*

The *Agrobacterium tumefaciens* VirA/VirG system is an interesting two gene TCS system controlling virulence (Jin *et al.*, 1990; Chang and Winans, 1992; Chang *et al.*, 1996; Jin *et al.*, 1990; Brencic *et al.*, 2004), illustrated in Figure 1.12. VirA is a transmembrane T_i-R HY which phosphorylates a RR, VirG. The receiver in VirA functions as an autoinhibitory domain. In its unphosphorylated state, this receiver domain interacts with the transmitter and prevents it from auto-phosphorylating, indirectly preventing the phosphorylation of the receiver in VirG. It would appear that the role of the VirA receiver is to prevent low level signalling from VirA to VirG, as in order for VirG to become phosphorylated, first the VirA receiver must be phosphorylated, overcoming its own inhibitory action.

1.4.11 RcaE/RcaF/RcaC phosphorelay in *Fremyella diplosiphon*

For over a century it has been known that some cyanobacteria will change colour, expressing red pigments in green light and *vice versa*. This process is called complementary chromatic adaptation (CCA), and in the filamentous cyanobacterium *Fremyella diplosiphon* it is partly controlled by the Rca TCS system (regulator for complementary chromatic adaptation, Figure 1.13). This is a complex three gene TCS phosphorelay system (Kehoe and Grossman, 1997; Stowe-Evans and Kehoe, 2004; Li and Kehoe, 2005; Kehoe and Gutu, 2006). Simplistically, it can be regarded as a variant of the T_i + R + H-R three protein relay illustrated in Figure 1.4.

The first protein in the relay is the HK RcaE, which has a chromophore-binding domain and will self-phosphorylate in red light, and then phosphorylate the receiver of RcaF. This in turn will phosphorylate the Hpt domain of HY RcaC, which has two functional receiver domains, and a third degenerate receiver. The RcaC has a DNA-binding domain which controls the

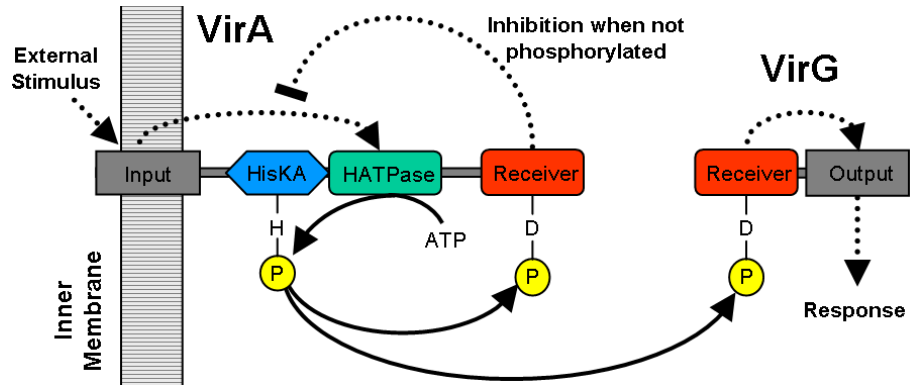


Figure 1.12: *Agrobacterium tumefaciens* VirA/VirG system, consisting of a transmembrane HY with T_i-R domains and a normal RR. This system can be thought of as a normal T_i + R two gene pair, except the VirA has a receiver domain which when unphosphorylated inhibits the self-phosphorylation of the T_i domain and/or the phosphorylation of the VirG receiver.

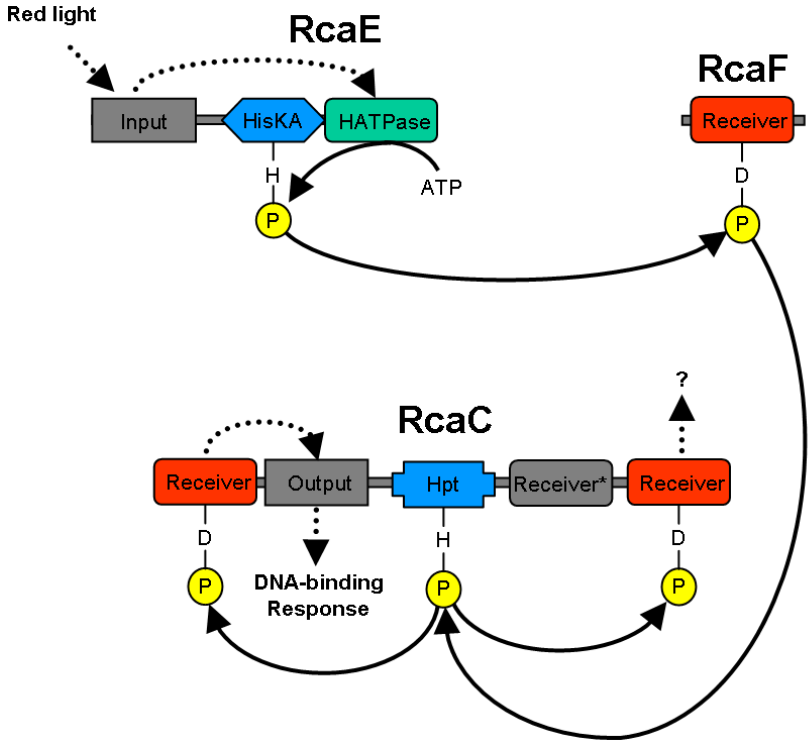


Figure 1.13: The *Fremyella diplosiphon* Rca phosphorelay, based in part on Kehoe and Gutu (2006) Figure 4. The central receiver domain of RcaC, shown in gray and marked with an asterisk, is degenerate and lacks the conserved aspartate phosphorylation site.

expression of the differently pigmented photosynthetic light-harvesting antennae. This output domain is adjacent to the N-terminal receiver, and it is presumably activated by it. The role of the second, C-terminal, receiver in RcaC is unclear, as is that of the degenerate receiver domain. Perhaps they act as in an inhibitory way, through competitive binding to the Hpt domain akin to the functionality of VirA (see Section 1.4.10).

In addition to showing yet another variation on the TCS phosphorelay, RcaC appears to show a possible competitive binding role for multiple receiver domains within a single protein. The complex HY RodK in *M. xanthus* is a similar example, containing domains T_i-R-R-R (Rasmussen *et al.*, 2006).

1.4.12 Red system in *Myxococcus xanthus*

The Red TCS system in *M. xanthus* is a complex four gene TCS network important in sporulation (Higgs *et al.*, 2005). The *red* operon encodes four consecutive TCS proteins: RedC is a conventional transmembrane HK. RedD is a complex RR containing two receiver domains (and no output domain). RedE is another HK, it appears to be cytosolic and lacking an input domain. Finally RedF is another RR, containing a receiver but no output domain.

Figure 1.14 shows the four TCS genes and the interactions suggested by a yeast two-hybrid (Y2H) assay. RedC appears to bind to the second receiver in RedD, while RedE appears to bind to both the first receiver in RedD, and the receiver of RedF (Higgs *et al.*, 2005). The fact that the RRs RedD and RedF have no output domains suggests that they could act as phosphorelay proteins, or perhaps acts directly like CheY in *E. coli* (see Section 1.4.3). To date no further clarification of this system has been published, leaving many unanswered questions, such as what controls the phosphorylation state of the HK RedE.

1.4.13 TCS systems in *Caulobacter crescentus*

The Gram-negative aquatic bacterium *Caulobacter crescentus* is a model organism for studying the cell cycle, where unusually cell division is asymmetric, yielding a swarmer cell and a non-motile stalked cell (Shapiro and Losick, 1997; McAdams and Shapiro, 2003). A recent comprehensive deletion analysis found at least 39 of its 106 TCS genes are required for cell cycle progression, growth, or morphogenesis (Skerker *et al.*, 2005).

One of the key regulators in the cell cycle is a RR CtrA which controls both polar morphogenesis and essential cell cycle processes. Another key RR in the cell cycle is DivK, which appears to act upstream of CtrA, perhaps by a phosphorelay (Wu *et al.*, 1998). The RR DivK is known to be phosphoryolated by HKs DivJ and PleC. Y2H work suggests DivK

also interacts with the sensor kinase DivL and two further uncharacterised soluble HKs, CckN and CckO (Ohta and Newton, 2003), while previous work had suggested DivL interacted with RR CtrA. DivL is of particular note as it has a phosphorylatable tyrosine residue, rather than a histidine as expected in a TCS kinase (Wu *et al.*, 1999) (making it a *tyrosine* sensor kinase rather than a *histidine* sensor kinase). Together these TCS proteins and others yet to be identified appear to form a complex and tightly regulated signalling network, reviewed in Ausmees and Jacobs-Wagner (2003).

In addition to their knock out mutant analysis, Skerker *et al.* (2005) also tested around ten HKs against a panel of 44 RRs using radioactively labelled phosphotransfer assays. This demonstrated an *in vitro* kinetic preferences for the HKs known partner RR(s), and successfully identified two orphan TCS genes as a novel system. However, even this comprehensive analysis covered only a fraction of the full interaction matrix for *Caulobacter crescentus*. This is a much more manageable task in *E. coli*, where a similar set of experiments has tested most of the possible TCS interactions (Yamamoto *et al.*, 2005), giving a fairly complete overview.

1.4.14 Phosphorelays in yeast

As mentioned in the introduction, TCS genes have also been found in yeast. A three gene osmoregulatory His→Asp→His→Asp phosphorelay was found in the budding yeast *Saccharomyces cerevisiae* (Maeda *et al.*, 1994; Posas *et al.*, 1996; Wurgler-Murphy and Saito, 1997). This is like the T_i-R + H + R example shown in Figure 1.4. The transmembrane HY Sln1 contains a transmitter domain which intra-molecularly phosphorylates its C-terminal receiver domain, which in turn phosphorylates the Hpt containing protein Ypd1, which phosphorlyates the RR Ssk1. Unusually, in the RR Ssk1 the receiver domain is C-terminal with an N-terminal output domain, the reverse of the typical domain order. This output domain interacts with a mitogen-activated protein kinase (MAPK) cascade controlling glycerol synthesis, thus modulating the composition of the cell membrane.

Later, a second RR partner for Ypd1 was identified, Skn7 (Li *et al.*, 1998; Ketela *et al.*, 1998), making this a one-to-many network (Figure 1.15). This means that Ypd1 interacts with three receiver domains, and exploration of the molecular basis of these interactions identified a common binding site on the Hpt domain of Ypd1 (Porter *et al.*, 2003; Porter and West, 2005). In addition, the crystal structure of the Sln1/Ypd1 complex has been solved (Xu *et al.*, 2003) (see Section 1.5).

The fission yeast *Schizosaccharomyces pombe* has its own TCS network (Aoyama *et al.*, 2000; Nguyen *et al.*, 2000; Buck *et al.*, 2001; Nakamichi *et al.*, 2002, 2003) which is similar,

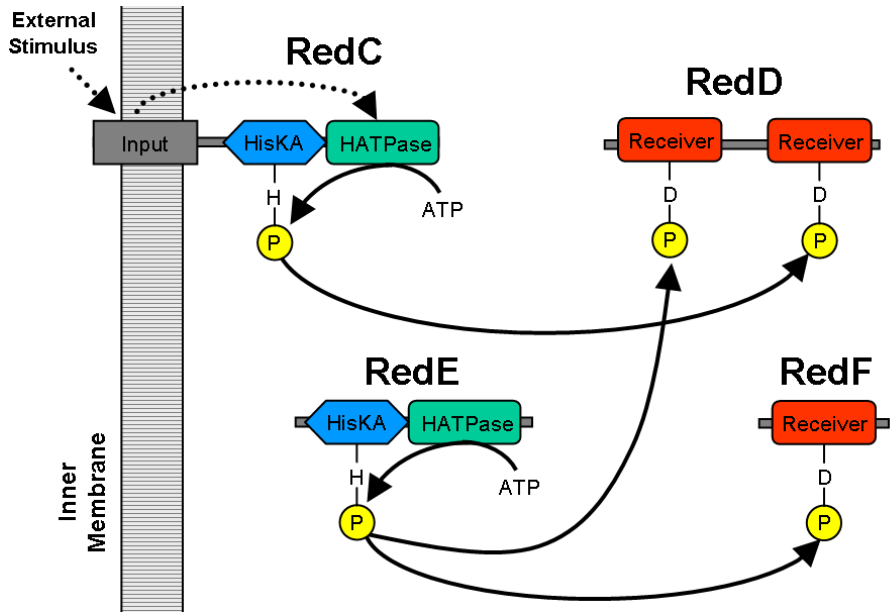


Figure 1.14: The *M. xanthus* red TCS network. Based on Y2H interactions reported in Higgs *et al.* (2005), with the directionality His→Asp by assumption.

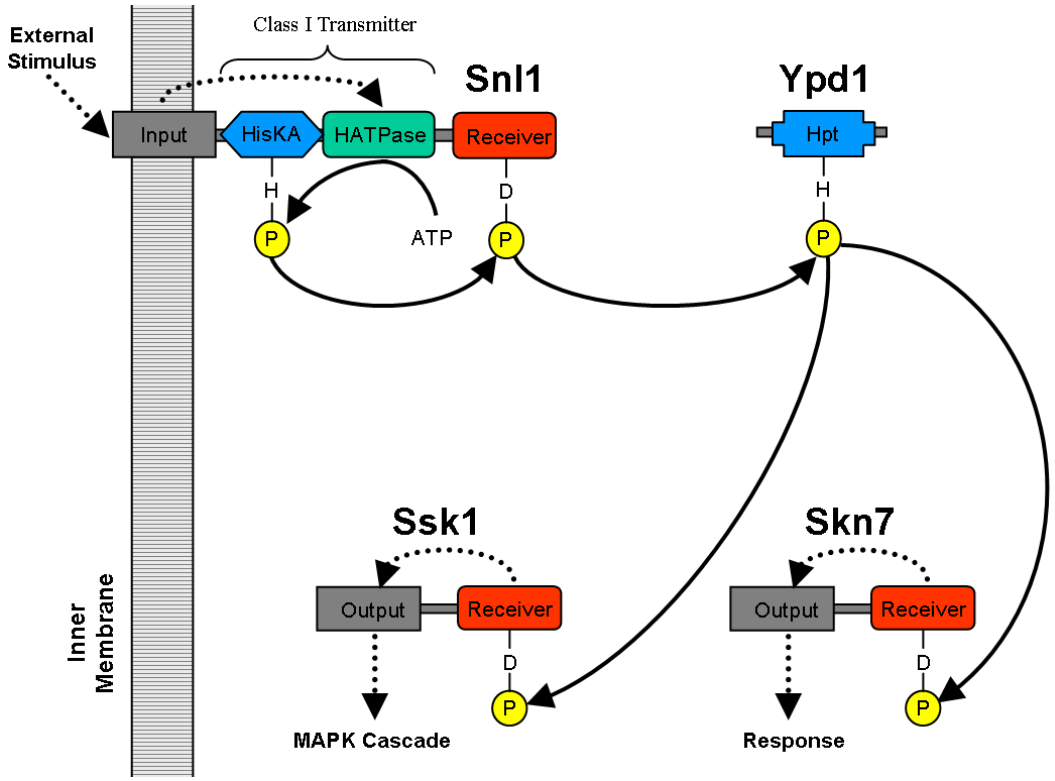


Figure 1.15: The *Saccharomyces cerevisiae* TCS network, based on a His→Asp→His→Asp phosphorelay T_i-R + H + R relay, discussed in Section 1.4.14.

but with more proteins involved (Figure 1.16). There are three known HKs (Mak1, Mak2 and Mak3 – also known as Phk3, Phk1 and Phk2) which are all cytosolic, unlike Snl1 in *Saccharomyces cerevisiae*. The three HKs appear to phosphorylate the Hpt containing Mpr1 (aka Spy1), which in turn phosphorylates two RRs (Mcs4 and Prr1). Thus this can be seen as a many-to-one-to-many network, where Mpr1 acts as a central point of regulation.

The phosphotransfer protein Mpr1 has a C-terminal Hpt, with an additional non-essential N-terminal domain which appears to facilitate binding to the RR (Tan *et al.*, 2007). The RR Mcs4 appears to be a homologue of Ssk1 in *Saccharomyces cerevisiae*, and also initiates a MAPK cascade. The TCS system responds to oxidative stress, but forms part of a larger network which also regulates mitosis and meiosis.

Both these yeast TCS networks are very similar to the quorum-sensing TCS network in *Vibrio harveyi* (see Section 1.4.9). Also, as an experimental system, yeast have one potential advantage – these appear to be their *only* TCS genes, meaning inter-system cross-talk need not be considered as a complicating factor when designing genetic modification experiments.

Interestingly the somewhat related filamentous ascomycetes encode an extensive family of HY signaling proteins, but with only one Hpt phosphotransfer protein and just two or three RRs (Catlett *et al.*, 2003). These fungi may also have just a single many-to-one-to-many TCS network, but with over ten HYs acting as inputs.

1.4.15 TCS systems in plants

Around fifty TCS genes have been identified in *Arabidopsis thaliana* (Hwang *et al.*, 2002), with a similar number found in rice (Pareek *et al.*, 2006). In some cases their role or function is somewhat understood. Many of these genes are HYs containing T_i-R domains, for example osmotic stress responses are regulated by ATHK1 which appears to act upstream of a MAPK cascade (Urao *et al.*, 1999). One unusual HY protein is WOL, containing an N-terminal transmembrane domain followed by T_i-R-R, which appears to be important in vascular morphogenesis (Mähönen *et al.*, 2000).

Plants (and cyanobacteria) also contain phytochromes, which are photosensory molecules with significant homology to histidine protein kinases (Schneider-Poetsch *et al.*, 1991). These usually lack the conserved histidine residue, and interact with G-proteins rather than RRs (NeuhausEtAl1993). Some of these proteins may feed into the circadian clock, which is a well studied system known to include several RR homologues lacking the aspartate phosphorylation site, termed pseudo response regulators (RPPs). One of these, TOC1/PRR1, is believed to be a key part of the clock (Makino *et al.*, 2000). Furthermore, the transcription of four other

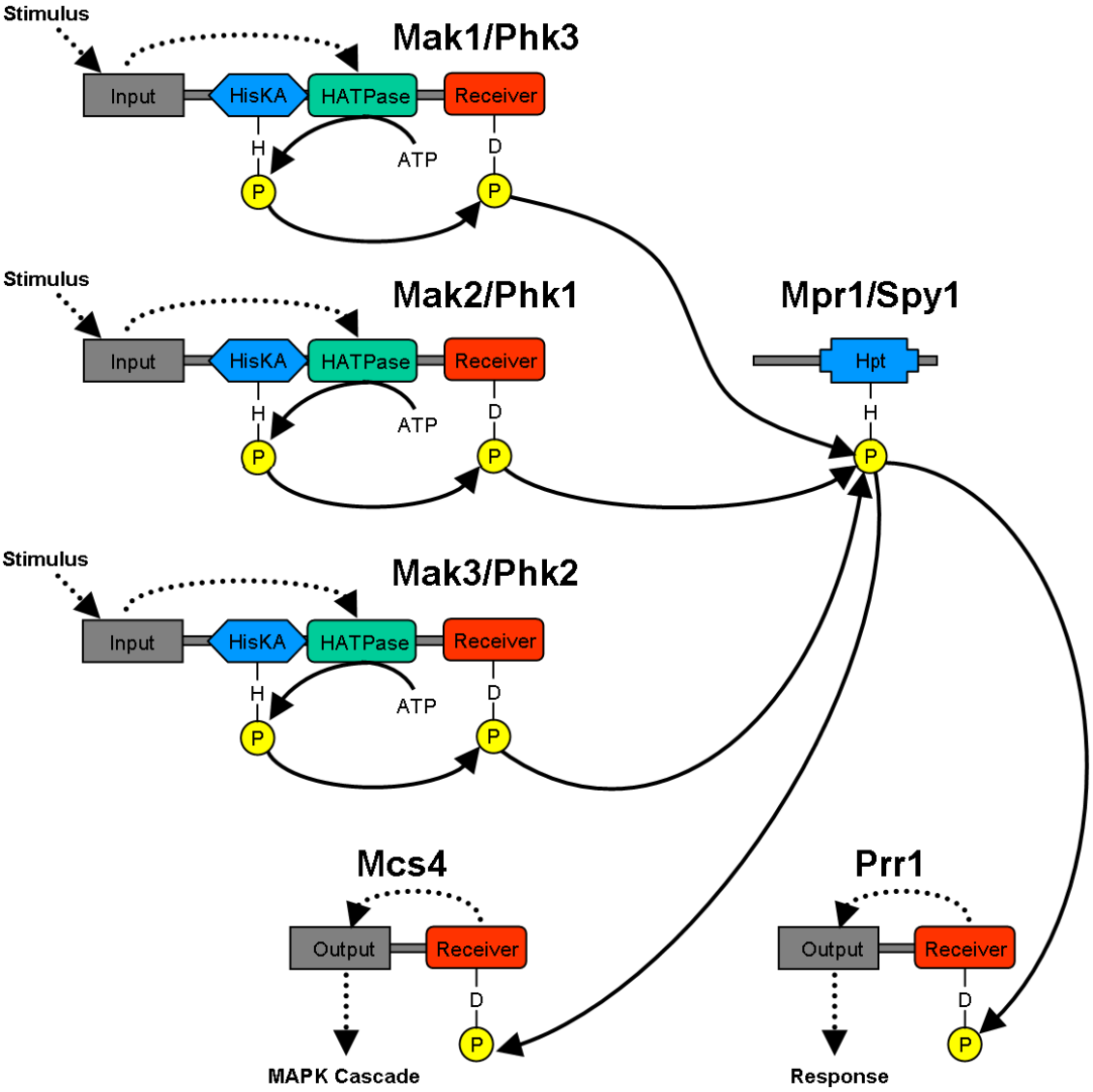


Figure 1.16: The *Schizosaccharomyces pombe* TCS network, with three HYs feeding into a phosphorelay, discussed in Section 1.4.14. Additional non-TCS proteins are involved in this pathway but not shown here, for example see Nakamichi *et al.* (2003) Figure 7.

PRR genes is also tied to the circadian rhythm (PRR3, 5, 7, and 9), peaking at different times during the day (Matsushika *et al.*, 2000; Makino *et al.*, 2002).

For recent reviews of TCS systems in plants, including the circadian clock, see Grefen and Harter (2004) and Mizuno (2005).

1.5 Three dimensional structures

Three dimensional structures have been solved for examples of the individual TCS domains, but for interacting complexes only a handful of minority cases have been resolved. In particular, a structure of the HisKA dimer in complex with a receiver domain has not yet been obtained. All the structures listed below are deposited in the Protein Data Bank (PDB) (Berman *et al.*, 2000).

CheY from *E. coli* was the first response regulator receiver domain structure to be solved (Stock *et al.*, 1989; Volz and Matsumura, 1991). Since then many others have been solved (Madhusudan *et al.*, 1996; Baikalov *et al.*, 1996; Feher *et al.*, 1997; Solà *et al.*, 1999; Birck *et al.*, 1999; Lewis *et al.*, 1999; Kern *et al.*, 1999; Guillet *et al.*, 2002; Toro-Roman *et al.*, 2005). These structures are all highly similar, with a five strand β-sheet surrounded by five α-helices in a barrel like arrangement. Figure 1.17 shows the *Bacillus subtilis* Spo0F receiver domain from PDB file 1F51 (Zapf *et al.*, 2000), while Figure 1.18 shows a simplified representation with the standard notation for the receiver secondary structure.

While all the receiver domains (with their asparate phosphorylation site) are very similar, there are several classes of histidine containing transmitter/phosphotransfer domain: HisKA, Hpt and the special case of Spo0B. These are shown schematically in Figure 1.19, and discussed below. They all share a four α-helix bundle, with the phosphorylated histidine about half-way down one helix, and are presumed to interact with their partner receiver domains in analogous fashions.

The HisKA domain, found in Class I transmitters (T_i), forms homodimers giving a four α-helix bundle. Each monomer has two α-helices, with the histidine phosphorylation site exposed about half way along the first helix. There are therefore two phosphorylation sites on opposite sides of the dimer bundle. Solved structures include *E. coli* EnvZ (Tomomori *et al.*, 1999, PDB reference 1JOY), shown in Figure 1.20.

The Hpt domain, found in Class II transmitters (T_{ii}) and as a phosphotransfer domain in TCS relays, also forms a four α-helix bundle, but this is a monomer and so there is only one histidine phosphorylation site. Solved structures include *E. coli* ArcB (Kato *et al.*, 1997; Ikegami *et al.*, 2001, PDB references 1BDJ, 1A0B, 1FR0) and RcsD/YojN (Rogova *et al.*,

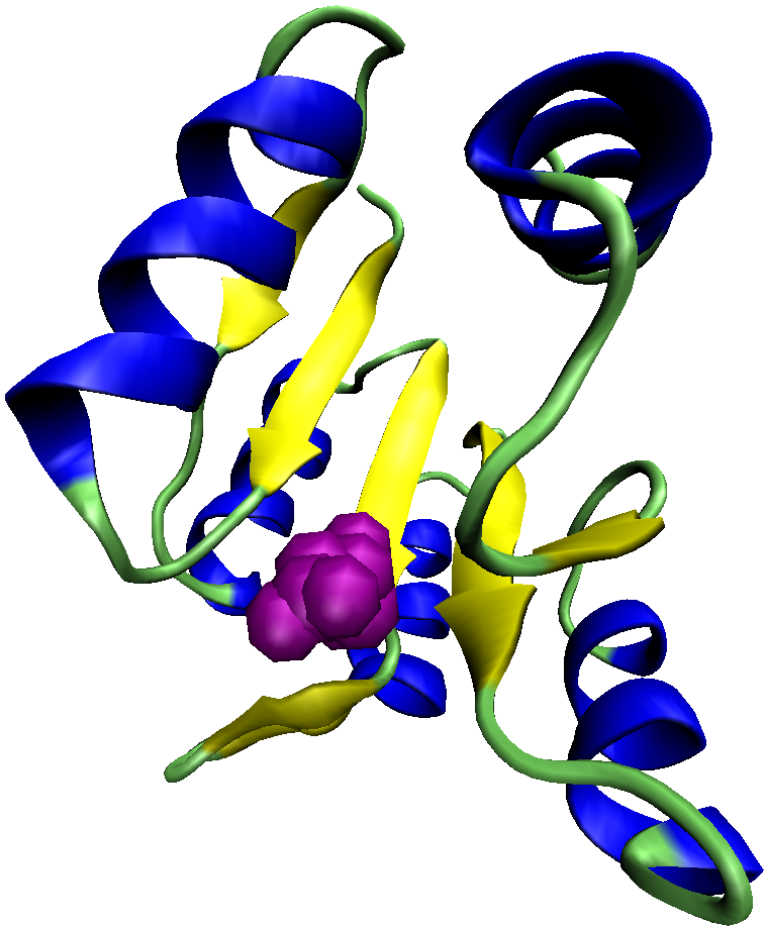


Figure 1.17: The 3D monomer structure of the receiver domain from *Bacillus subtilis* RR Spo0F (Zapf *et al.*, 2000, PDB ref. 1F51). A "cartoon" representation by the software package VMD (Humphrey *et al.*, 1996) is shown, coloured according to the secondary structure (five α-helices in blue, five β-sheets in yellow, and coils in green), with a purple space-filling representation of the phosphorylatable aspartate (D54).

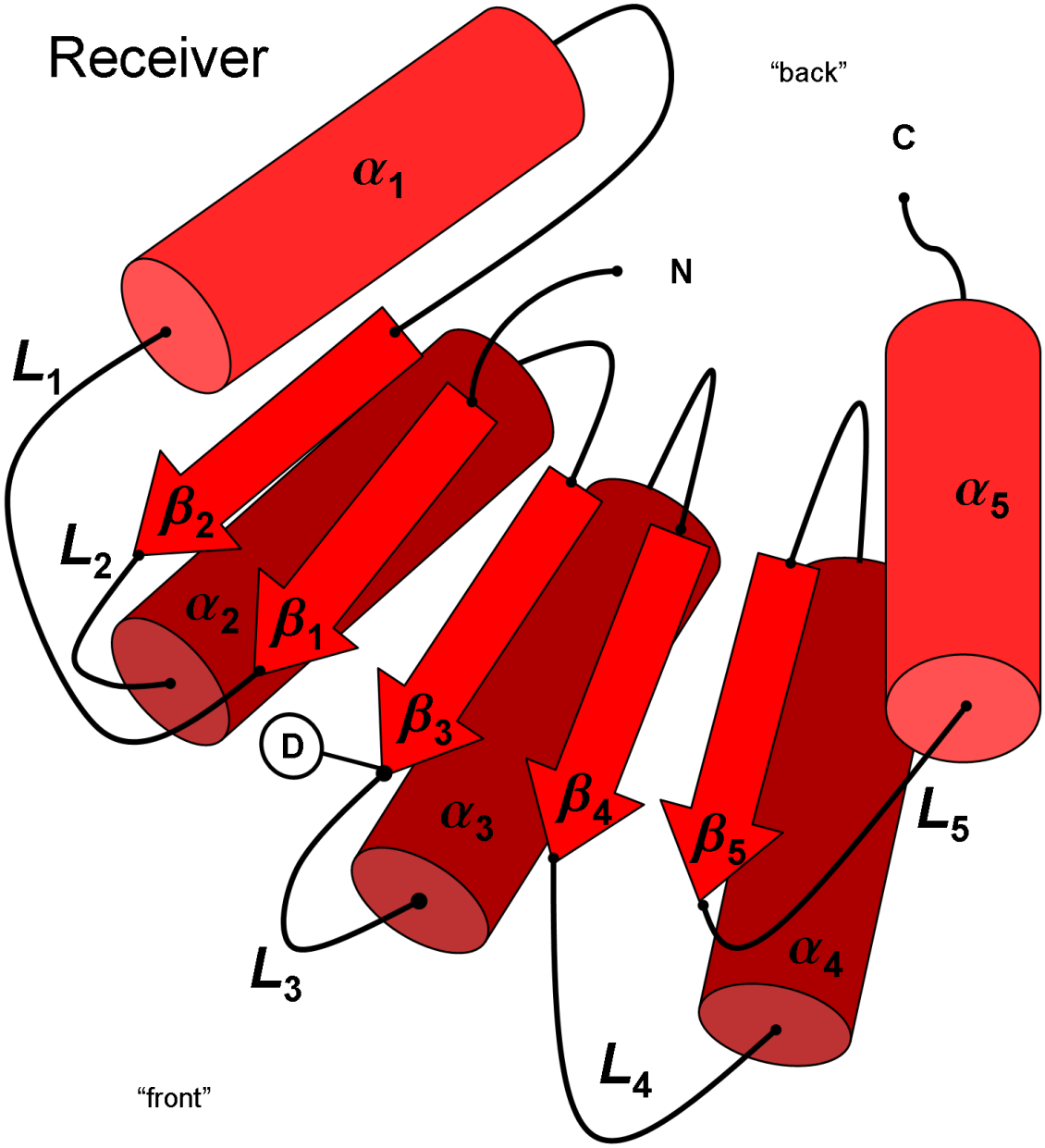


Figure 1.18: Simple diagram of the five α-helix barrel of the receiver domain, with its five strand β-sheet core. By convention, the receiver domain is considered to have a front side which forms the interaction surface (shown in the foreground) and a back side. Following the notation of Zapf *et al.* (2000) the front five loops are labeled L_1 to L_5, with the conserved phosphorylatable aspartate on the end of β_3 / start of L_3, while the four back loops have not been labeled. *cf.* the solved structure shown in Figure 1.17.

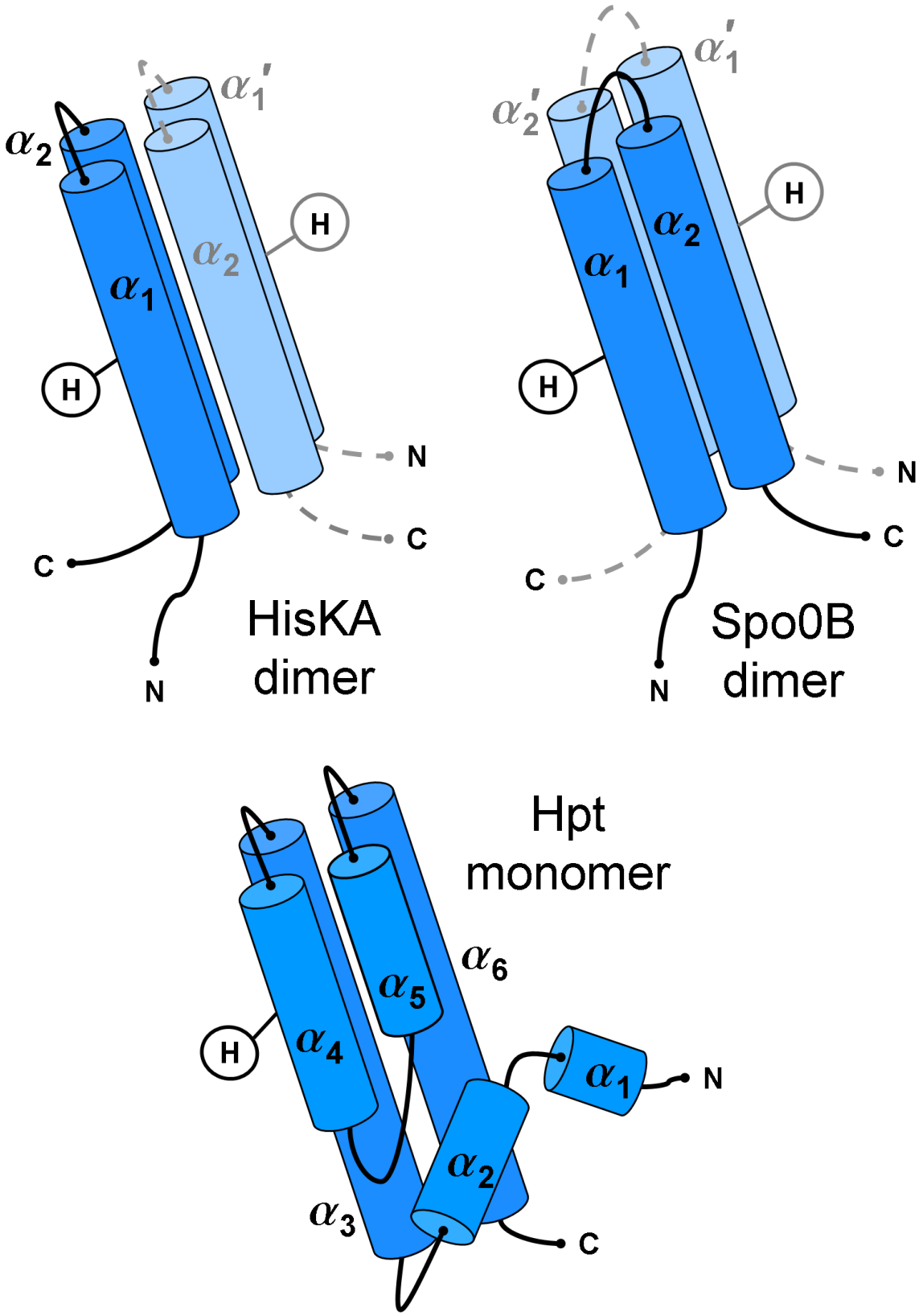


Figure 1.19: Simple diagrams of the four α-helix bundles in the HisKA and Spo0B dimers, and the Hpt monomer. In the case of the dimers, the second protein is shown in a pale blue with the backbone and labels in grey, and the α-helices numbered with a prime. *cf.* solved structures shown in Figures 1.20, 1.21 and 1.22.

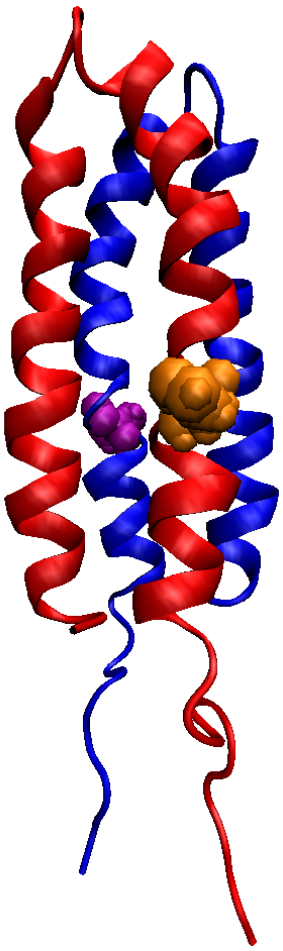


Figure 1.20: The 3D dimer structure of the HisKA domain of *E. coli* EnvZ, from PDB file 1JOY (Tomomori *et al.*, 1999). The histidine phosphorylation sites (H243) are shown using a space-filling model. One monomer is shown in red, with the phosphorylatable histidine in orange, the other is in blue with its phosphorylatable histidine in purple.

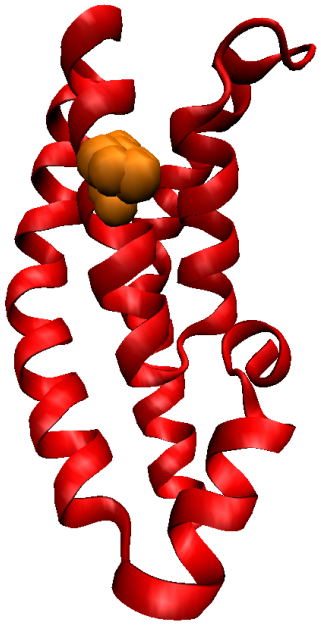


Figure 1.21: The 3D monomer structure of the Hpt domain of *E. coli* ArcB, from PDB file 1BDJ (Kato *et al.*, 1999). The protein is shown in red, with the conserved histidine phosphorylation site (H715) using a space-filling model in orange.

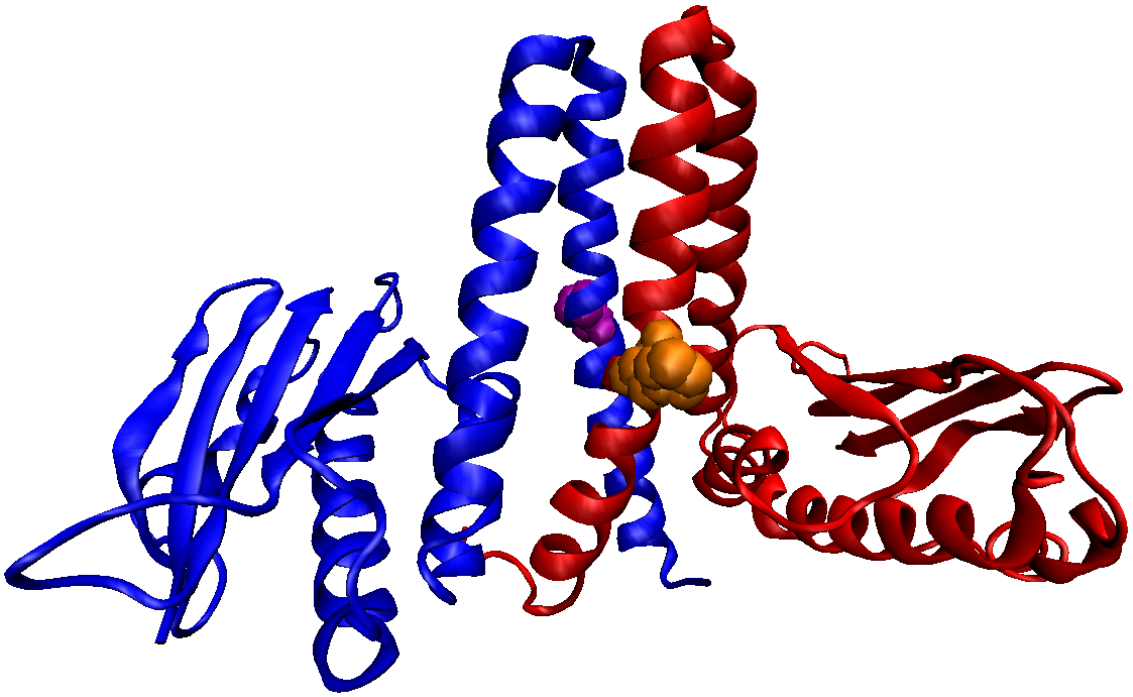


Figure 1.22: The 3D dimer structure of *Bacillus subtilis* Spo0B, from PDB file 1F51 (Zapf *et al.*, 2000). The histidine phosphorylation sites (H30) are shown using a space-filling model. One monomer is shown in red, with the phosphorylatable histidine in orange, the other dimer is in blue with its phosphorylatable histidine in purple.

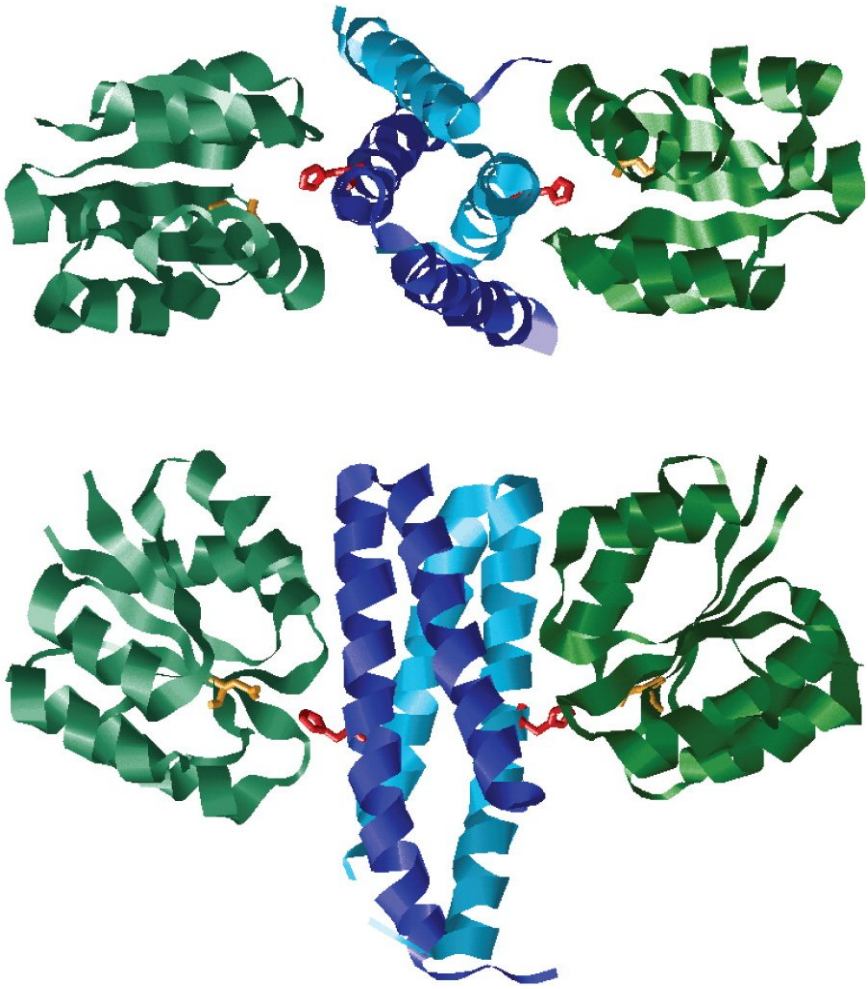


Figure 1.23: Crystal structure of the *Bacillus subtilis* dimer Spo0B in complex with the Spo0F receiver, PDB file 1F51, adapted from Laub and Goulian (2007) Figure 5. The Spo0B monomers are in light and dark blue, and two Spo0F molecules are in green. The histidine (H30) and aspartate (D54) phosphorylation sites are shown as red and yellow stick models respectively.

2004, PDB ref. 1SR2). The ArcB Hpt structure is shown in Figure 1.21. The yeast protein Ypd1 has also had its Hpt structure solved (Xu and West, 1999, PDB ref. 1QSP), and while not perfectly in agreement, does also have the four α-helix bundle structure.

The *Bacillus subtilis* protein Spo0B is a unique case. While phylogenetically distinct, it is structurally analoguous to the HisKA domain and forms a four α-helix bundle with similarly positioned histidine phosphorylation sites (Varughese *et al.*, 1998, PDB reference 1IXM). When compared to the HisKA dimer, one obvious difference is the switch of the relative positions of the second helix of the two proteins (see Figure 1.19). As discussed in Section 1.4.8, this protein acts in a phosphorelay, interacting with the receiver domains of proteins Spo0F and Spo0A. A crystallographic structure has been obtained for the Spo0B/Spo0F interaction (Zapf

et al., 2000, PDB reference 1F51), see Figure 1.23, which can be used as a basis to interpret the likely orientation of a HisKA and receiver interaction.

Structures of the Hpt domain in complex with a receiver domain have also been solved experimentally, for example the Hpt domain of ArcB with response regulator CheY from *E. coli* (Kato *et al.*, 1999, PDB reference 1BDJ). The ArcB Hpt is similar to that of CheA, which interacts with the CheY receiver as described in Section 1.4.3. The ArcB/ArcC system was described in Section 1.4.6. The yeast Ypd1/Sln1 complex (see Section 1.4.14) is another solved example of the Hpt-receiver interaction (Xu *et al.*, 2003, PDB references 1OXB and 1OXK).

1.6 TCS networks

Input and output domains in TCS systems are heterogeneous due to the variety of different stimuli they respond to, and the different responses they elicit (Mascher *et al.*, 2006). While the input and output domains define the *function*, it is the transmitter-receiver interactions which define the *topology* of the TCS pathway or network (Figure 1.24).

Skerker *et al.* (2005) and other results suggest that transmitter-receiver interactions are usually pairwise exclusive, or at least have a kinetic preference for the cognate partner. However, as the examples in the previous section show, some domains are promiscuous and have multiple partners. This means that instead of an organism being restricted to having multiple separate TCS pathways operating in parallel and in isolation, complicated networks are possible.

For example, in *E. coli* there are two Nar HK proteins which will both interact with the two Nar RR proteins (Section 1.4.2). The *Bacillus subtilis* sporulation system has five HKs all phosphorylating the same RR, a many-to-one network (Section 1.4.8). Conversely, the *E. coli* chemotaxis system has one HK phosphorylating two different RRs, a one-to-many network (Section 1.4.3).

The comparison of *Bacillus* species in Stephenson and Hoch (2002) found the same basic TCS sporulation network structure (Section 1.4.8), but with different input domains present in the HKs, presumably achieved by recombination events. More generally, TCS networks are presumably created from simple HK+RR pairs by a series of gene duplications, modifications and deletions.

The literature also shows that in addition to simple cross-talk by promiscuous phosphorylation, and TCS regulating each other at the transcriptional level (Bijlsma and Groisman, 2003), more complex interactions can also occur between systems. For example, the

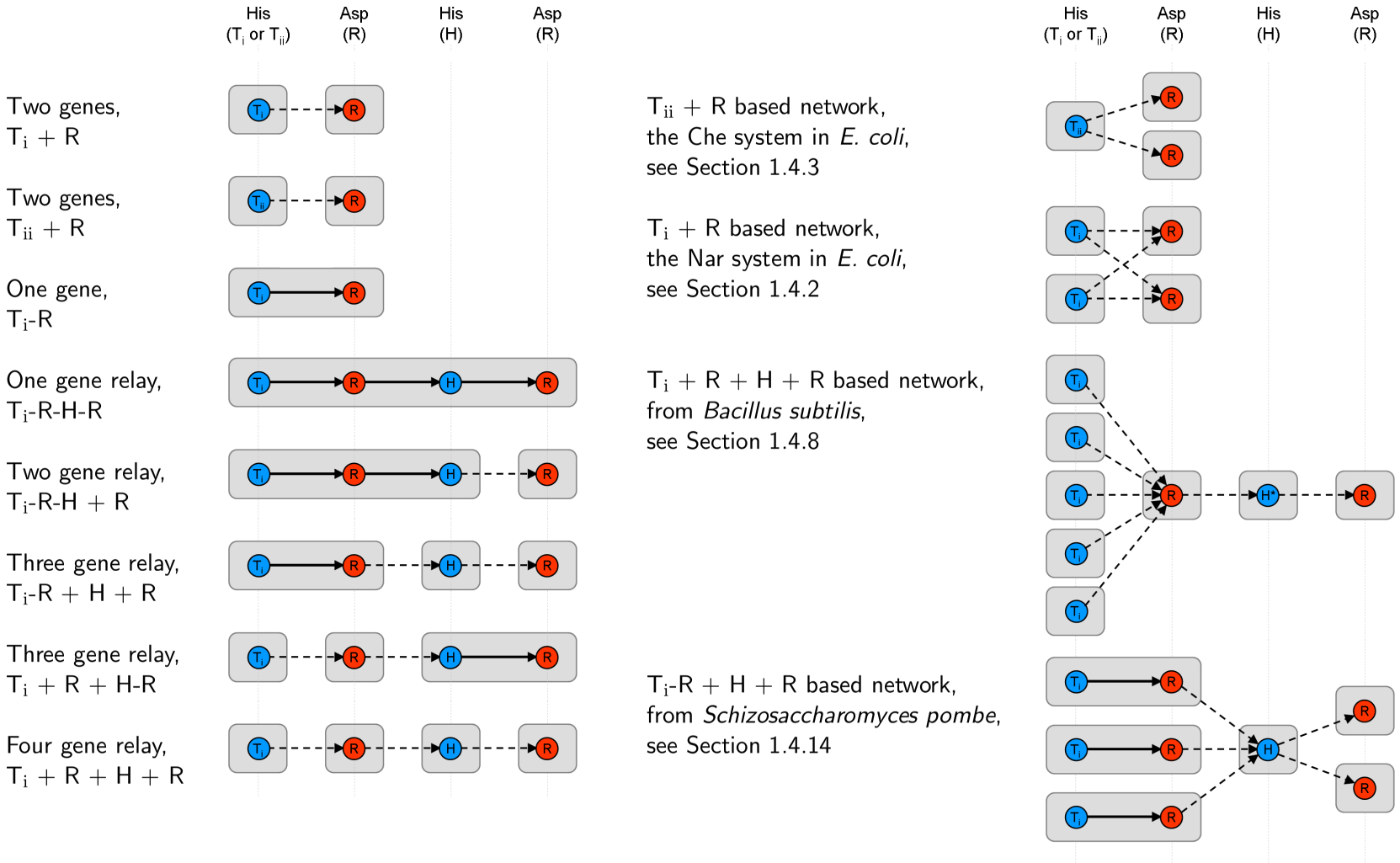


Figure 1.24: Simple schematics of His→Asp TCS systems (e.g. Figures 1.1, 1.2 and 1.3), and selected His→Asp→His→Asp phosphorelay systems (e.g. Figure 1.4) on the left, and example TCS networks on the right. Gray rounded rectangles represent proteins, with the blue and red circles representing phosphorylation sites (histidine and aspartate respectively), labeled by the domain they are found within (T_i, T_{ii}, H or R). Solid arrows are intra-protein phosphotranfers, while the dashed arrows are inter-protein.

Salmonella PhoP/PhoQ and PmrA/PmrB TCS systems can interact via the PmrD protein which is promoted by phosphorylated PhoP, and binds to phosphorylated PmrA and prevents its dephosphorylation (Kato and Groisman, 2004). The VirA/VirG system (Section 1.4.10), or HYs RcaC (Section 1.4.11) and RodK (Rasmussen *et al.*, 2006) appear to demonstrate roles for competitive phosphorylation between multiple receivers. Similarly in *Helicobacter pylori*, the CheA homologue CheAY2 has an additional receiver domain which seems to act as a phosphate sink to modulate the phosphorylation level of the partner CheY homologue, CheY1 (Jiménez-Pearson *et al.*, 2005).

More recent work on an artifical cross-talk system in genetically modified *E. coli* was reviewed in Laub and Goulian (2007). The enterococci VanS/VanR TCS system was expressed in *E. coli*, and its ability to interact with the native *E. coli* PhoR/PhoB TCS system was studied. When only one of these HKs is present (PhoR or VanS) they can phospholylate both RRs PhoB and VanR. However, when both HKs are present this cross-talk is much reduced, likely the result of the HKs having phosphatase activity on their partner RR. This suggests bifunctional HKs (with kinase and phosphatase activity) may be important in preventing unwanted cross-talk, while mono-functional HKs (kinase only) would be required in many-to-one pathways.

A complex TCS network linking multiple stimuli and responses to control the behaviour of a prokaryote can be considered analagous to a neural network in higher organisms (Hellingwerf *et al.*, 1995). The evolution of the TCS network over generations can adjust the affinity of each potential TCS interaction, and even alter the network topology, allowing prokaryotes to acquire new behaviour. This is akin to the development of neural networks over the lifetime of an individual higher organism. Hoffer *et al.* (2001) goes one step further, claiming past phosphate levels influence auto-amplification of two-component regulatory systems giving "learning" behaviour on a much faster time scale.

1.7 Predicting TCS interactions

Identifying potential interactions between HKs and RRs is the first step to building up an understanding of a prokaryote's TCS network. There are several approaches to predicting these TCS interactions, which must ultimately be verified experimentally.

1.7.1 Genome arrangement

The simplest way to deduce TCS pairings is from the arrangement of the TCS genes in the genome. In prokaryotes, co-expressed genes are normally found together in operons, and as a result genes of related or interacting function are often found next to each other. Typically the genes for HKs are found adjacent to the genes for their partner RRs in a single operon. Furthermore, a high proportion of TCS gene pairs appear to have fused into single HYs containing both a transmitter domain and a receiver domain. Such examples can easily be identified from the genome (Chapter 2).

In addition to these simple cases, there are more complicated hybrid proteins containing multiple transmitters and/or receiver domains, and also complex gene clusters containing multiple TCS genes. Conversely, some TCS genes are isolated or orphaned – they do not lie adjacent to the gene for their partner signalling protein. In general, not only do larger genomes tend to have more TCS genes, the genes and cluster also tend to be more complicated (Chapter 2). Thus in many cases the partner proteins and consequent signalling pathways for many TCS proteins are unclear, and cannot be predicted from the genome arrangement alone. In *M. xanthus* for example, a large proportion of TCS genes are complicated hybrids and/or in complex gene clusters, and their partnerships cannot be easily inferred (Whitworth and Cock, 2008a; Shi *et al.*, 2008).

1.7.2 Phylogenetics and comparative genomics

If an operon containing an HK/RR pair were to be duplicated, this would yield two initially identical TCS systems which would exhibit cross-talk if they were to be co-expressed. Over evolutionary timescales, the transmitter and receiver interaction surfaces could diverge, restricting the cross-talk, for example giving rise to two isolated HK/RR pairs. A simple interpretation of the phylogenetic relatedness of the transmitter and receiver domains within an organism can be indicative of past interactions, and thus guide predictions of the current interactions. See Figure 5 in Grebe and Stock (1999) and Figure 2 in Koretke *et al.* (2000), or for non-TCS example, Goh *et al.* (2000). Indeed, domain sequence comparison was used in the discovery of the Nar system in *E. coli* (Section 1.4.2).

Comparative genomics is also potentially very powerful for predicting TCS interactions. When a system is known in one organism, this can be used as a template to interpret the interactions of homologous genes in related species. For example, *Bacillus* sporulation (Section 1.4.8), and *Vibrio* quorum sensing (Section 1.4.9).

1.7.3 Co-expression

For any TCS phosphotransfer to take place, the two components must be co-expressed and co-located. Strict regulation can prevent cross-talk between otherwise compatible domains. The chemotaxis system in *Rhodobacter sphaeroides* appears to employ spatial segregation in this way (Section 1.4.4). Other organisms have complex temporally regulated behaviour which may also prevent unwanted TCS cross-talk. Examples include the cell cycle in *Caulobacter crescentus* (Section 1.4.13) and developmental regulation of fruiting body formation in *M. xanthus* (Whitworth and Cock, 2008a).

Co-expression data has been used to predict TCS partnerships in *Desulfovibrio vulgaris* using whole-genome microarrays to identify orphan TCS genes with similar expression patterns (Zhang *et al.*, 2006). Such studies are likely to become more common as microarray or next generation sequencing technologies become cheaper and more widely available.

1.7.4 Multiple sequence alignment based predictions

Potential interactions between TCS transmitters and receivers have begun to be predicted from their amino acid sequences. Such interactions may be valid *in vitro* but may be prevented *in vivo* if regulated to prevent co-expression.

Chapters 5 and 6 introduce a multiple sequence alignment (MSA) based approach to predicting TCS pairings between HisKA and receiver domains. During the course of this project, two independent groups published rival approaches. The work of White *et al.* (2007) is discussed in Chapter 6 as a special case of our own MSA approach to predicting TCS pairings, while Burger and van Nimwegen (2006, 2008) introduced a more complicated MSA based Bayesian model.

1.8 Research aims

To understand the behaviour of any one bacterium, an understanding of its signalling network is required. To construct a model of any network a list of parts is needed, and their interconnections. Focusing on just TCS signalling, this means a list of all transmitter, Hpt and receiver domains, *and* a list of their interactions.

The main thrust of this thesis is to predict transmitter-receiver interactions from their amino acid sequences. For any statistical problem, lots of data is usually required - in this case, lots of known transmitter-receiver pairings. Searching hundreds of fully sequenced bacterial genomes for simple TCS gene clusters (containing only one transmitter and one receiver)

provided this list (see Section 1.7.1, and Chapter 2).

This collection of paired transmitter and receiver domains was then used to identify those residues important for the pairing specificity (Chapter 5). While 3D structures of docked transmitter and receiver domains would be ideal for identifying interacting residues, as discussed above in Section 1.5, to date co-crystals have only been solved for a few atypical cases.

The related question of which residues are *essential* has been tackled experimentally for certain transmitter/receiver pairings - for example with alanine mutation studies of *Bacillus subtilis* Spo0F (Tzeng and Hoch, 1997) (see Section 1.4.8). The interactions of the Hpt domain in *Saccharomyces cerevisiae* Ypd1 with its three partner receiver domains (see Section 1.4.14) have been explored using a combination of alanine scanning mutagenesis and a Y2H assay (Porter *et al.*, 2003; Porter and West, 2005). Answering the question of which impart *partner specificity* would require mutation studies on a panel of transmitters and receivers. Nevertheless, there is some existing information regarding the role and importance of particular residues for certain known TCS pairs.

Establishing which parts of the TCS domains govern pairing specificity (Chapter 5) is the basis of the predictive model presented in Chapter 6. A selection of scoring systems are investigated to assign numerical values to amino acid pairs from the transmitter and receiver domains, for use as explanatory variables in a generalised linear model (GLM).

For bacteria with a large number of TCS genes, testing all possible domain combinations experimentally is an enormous undertaking. David Whitworth *et al.* have begun a multi-year project to systematically test all pairwise combinations of TCS transmitters and receivers in *M. xanthus* using the yeast two-hybrid assay (Y2H) (Whitworth *et al.*, 2008), a technique that has also been used in *Caulobacter crescentus* (Ohta and Newton, 2003). More recently, relatively high throughput phosphotransfer assays with radio labeling have also been reported (Yamamoto *et al.*, 2005; Skerker *et al.*, 2005). However, any sequence based prediction tool would allow biological experiments to be targeted at the most promising candidate interactions – helping to clarify the signalling pathways.

If a transmitter and receiver are predicted to interact, or even if they have been shown to interact *in vitro*, there is no guarantee they will interact *in vivo*. Both proteins must be expressed at the same time, and in the same place – a prerequisite for any such potential interaction to actually take place. For example, bacterial movement is normally controlled by TCS, and the chemotaxis machinery is known to be located at the poles of the cell in some bacteria. Temporal separation of protein expression is likely to be important in species with complicated life-cycles, such as developmental regulation in sporulation. Microarray time

course experiments have begun to be used to study this (Section 1.7.3).

As most prokaryotes contain multiple TCS systems, this is an especially numerous class of protein-protein complex. This makes it an ideal candidate system for the more general problem of predicting pairings between members of paralagous families, such as in σ-factor/anti-σ-factors (Hughes and Mathee, 1998) or G-protein coupled receptors/trimeric G-proteins (Cabrera-Vera *et al.*, 2003).

Chapter 2

Finding TCS genes and pairs

2.1 Introduction

With the long term aim of identifying thousands of HK/RR pairs to build a predictive model of TCS interactions, the first step was to list TCS genes from all available genomes. From a manual inspection it was clear that the level and quality of TCS annotation varied dramatically between species and even between strains. Any attempt to use just the genome annotation would therefore miss many genes, and also make direct comparisons between genomes problematic.

In this chapter an automated analysis pipeline for the identification and classification of TCS genes is described, using protein domain motifs from both PFAM (Bateman *et al.*, 2004) and CDD (Marchler-Bauer *et al.*, 2005) with the standalone Reversed Position Specific Blast (RPS-BLAST) tool (Marchler-Bauer and Bryant, 2004). The CDD is a composite database drawing on PFAM and other motif databases such as SMART (Schultz *et al.*, 1998) and COG (Tatusov *et al.*, 2003). The results of the analysis of hundreds of fully sequenced prokaryotic genomes are presented.

During this work the sequencing of the social-predatory bacteria *Myxococcus xanthus* DK 1622 (accession NC_008095) was completed, and this analysis contributed to the annotation of the TCS genes, acknowledged in Goldman *et al.* (2006). Additionally, publications Whitworth and Cock (2008a,b) draw directly on this work.

2.2 Source of genomes

The NCBI provides all publicly available fully sequenced prokaryotes in a variety of file formats, including as GenBank flat files which in addition to the complete DNA sequence also include

every annotated gene with its amino acid sequence. Starting from a snapshot taken in January 2004 with just a couple of hundred genomes, a local data-bank was regularly updated as and when new genomes were released.

However, for simplicity, all survey results presented herein will focus on a single collection downloaded as a compressed archive from the National Center for Biotechnology Information (NCBI) FTP site[1] on 26 February 2007. This contained all 457 of the then available fully sequenced prokaryotes, which due to multiple chromosomes or plasmids came to 807 GenBank files. These species and their accessions numbers are listed in Table A.1 on page 209.

In addition, the NCBI also provides a table of information for the sequenced prokaryotes[2] which includes information about pathogenicity, habitat, Gram staining, spores and shape. This information allows the investigation of correlations between this information and TCS usage.

2.3 Identifying TCS genes

The standalone tool RPS-BLAST was used to identify protein domains within the given amino acid translation of every annotated CDS within the GenBank files used. A single search against the PFAM database (Bateman *et al.*, 2004) was found to identify almost every TCS gene expected, but was not enough to detect or classify all the expected domains. Transmitter domains had to be defined as a small phosphotransfer sub-domain (HisKA or Hpt) followed by an HATPase. While most receiver and HATPase domains were easily detected, the phosphotransfer regions were often not recognised by the PFAM models – even with a lowered expectation threshold. Furthermore, it was difficult to exclude many non-TCS-related HATPase domains (false positives).

Both these issues were resolved satisfactorily by screening candidate TCS genes identified from PFAM domain hits against the CDD database (Marchler-Bauer *et al.*, 2005), which included a number of models for full length HKs (including the full transmitter domain) and for several non-TCS HATPase containing proteins (allowing the elimination of many HATPase false positives).

2.4 Isolated and paired genes

In addition to sequence data, the GenBank files include the location of each gene within the chromosome. Parsing this information allowed each TCS gene to be considered in the context

[1] `ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.gbk.tar.gz`

[2] `http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi` as a webpage, or as a simple tab separated text files at `ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/lproks_0.txt` and lproks_1.txt.

of its neighbouring genes, and in particular allowed a simple classification of TCS genes as either *isolated*, *paired* or *complex* (other).

A TCS gene was considered to be *isolated* or *orphaned* if there were no other TCS genes (including unassigned HATPase domains) within $5,000$ nucleotide base pairs (bp).

To be considered *paired*, two TCS genes were required to be on the same strand (i.e. same direction, forwards or reverse), no more than 100 bp apart, with an overlap no more than 100 bp. Furthermore, any pairing was rejected if another TCS gene was found within $5,000$ bp (the same cut-off distance used for an isolated gene) as this could be considered to be a complex gene cluster. These distance criteria were chosen in part based on known systems such the TCS genes in *E. coli*.

2.5 Implementation

A multi-step screening procedure was adopted. Initially, the protein sequences were searched against the PFAM database. Receiver (pfam00072) and HATPase (pfam02518) domains were accepted with expectation cut-off of 10^{-4}. A threshold of 1.0 was used for the smaller domains: HisKA pfam00512, Hpt pfam01627 and the dimerisation region pfam02895. Proteins with any matches to these five PFAM domains were then searched against the CDD database at expectation threshold 10^{-4}. The CDD includes specific models for the HisKA (smart00388, cd00082) and receiver domains (cd00156). Further HisKA and HATPase domains were found by checking CDD matches to full length HK motifs: BaeS COG0642, NtrB COG3852, NtrY COG5000, KdpD COG2205, and VicK COG5002. Some CDD models cover long motifs or entire genes (e.g. COG motifs are based on orthologous genes), thus any (partial) match was examined closely to see which sub-region of the motif had been matched (Table 2.1).

As most domains could be detected by more than one model (e.g. both pfam00072 and cd00156 for a receiver), any such hits with substantial overlap were simply merged. It was observed that some T_{ii} domains gave good matches to some of the HK COG models, potentially leading to the mis-identification of a dimerisation region in a T_{ii} domain as a HisKA sub-domain within a T_i domain. Sub-domain matches from these COG models were therefore labeled KD, denoting either HisKA (K) or dimerisation regions (D), unless there was an additional unambiguous match in the same region to either a HisKA or dimerisation region.

Small overlaps between different domain types where tolerated. A small list of less than fifty borderline overlap cases were resolved by hand.

Class I transmitter domains (T_i) were then constructed as the composite of a HisKA domain followed by a HATPase, and Class II transmitters (T_{ii}) by a Hpt domain then HATPase,

Domain	Motif	Name	Sub-region	Exp.
R, Receiver	pfam00072	Response_reg	*any*	10^{-4}
	cd00156	REC	13-103	10^{-4}
A, HATPase	pfam02518	HATPase_c	*any*	10^{-4}
	COG3852	NtrB	236-355	10^{-4}
	COG5000	NtrY	595-709	10^{-4}
	COG0642	BaeS	223-330	10^{-4}
	COG5002	VicK	339-447	10^{-4}
	COG2205	KdpD	771-880	10^{-4}
K, HisKA	pfam06580	His_kinase	*any*	1
	pfam00512	HisKA	*any*	1
	smart00388	HisKA	1-66	10^{-4}
	cd00082	HisKA	3-62	10^{-4}
D, Dimerization	pfam02895	H-kinase_dim	*any*	1
KD, HisKA or dimerization	COG3852	NtrB	134-187	10^{-4}
	COG5000	NtrY	484-556	10^{-4}
	COG0642	BaeS	113-178	10^{-4}
	COG5002	VicK	224-290	10^{-4}
	COG2205	KdpD	569-727	10^{-4}
H, Hpt	pfam01627	Hpt	*any*	1
MutL DNA mismatch repair enzyme	COG0323	MutL	19-225	10^{-5}
RsbW Anti-σ regulatory factor	COG2172	RsbW	13-125	10^{-6}
Hsp90 protein	pfam00183	HSP90	*any*	10^{-4}
DNA gyrase B carboxyl terminus	pfam00986	DNA_gyraseB_C	*any*	10^{-4}
DNA gyrase B	pfam00204	DNA_gyraseB	*any*	10^{-4}
DNA gyrase/topoisomerase IV	pfam00521	DNA_topoisoIV	*any*	10^{-4}

Table 2.1: PFAM and CDD motifs used to identify TCS domains, or exclude false positives. Motifs derived from COG tend to be for entire genes, and thus an additional restraint was used that the RPS-BLAST hit had to span the specified sub region before being considered. Domains in the bottom section of the table were used to identify non-TCS HATPases (false positives).

optionally with a dimerisation region in between. Non-TCS genes containing the HATPase domain were identified and excluded using other domain matches: DNA gyrase/topoisomerase IV (pfam02518, pfam00986, pfam00521), Hsp90 (pfam00183), MutL (COG0323) and RsbW anti-sigma factor (COG2172). A small fraction of the HATPase domains remained unassigned, and from manual inspection these were mostly transmitter domains where the phosphoacceptor domain had not been automatically identified. Genes with unassigned HATPase domains are therefore included in the gene counts reported below.

The Python programming language (www.python.org) was used to script this process, using the Biopython libraries (www.biopython.org) to manipulate sequence files and to call RPS-BLAST and parse its output. The plots in this chapter were drawn using the R programming language (R Development Core Team, 2007) based on tables of data prepared with Python scripts.

2.6 Survey results and discussion

In total $24,039$ TCS genes were identified out of $1,435,868$ annotated genes from the 457 completely sequenced prokaryotes listed in Table A.1, a number similar to that expected by extrapolation from other studies (Koretke *et al.*, 2000; Kim and Forst, 2001; Zhang and Shi, 2005). This is about two percent of all prokaryotic genes.

2.6.1 Transmitters versus receivers

Figure 2.1 shows that to a first approximation the number of transmitters and receiver domains are about equal in each species. However, in general more receivers are found than transmitters (even including the unassigned HATPase and phosphotransfer domains in the transmitter count). It seems that there is not always a simple one-to-one pairing between transmitters and receivers, and taking this at face value, on average a transmitter may have more than one partner.

One explanation for this is that the transmitter or phosphotransfer domain detection was less successful than that of receivers. While manual checks of a sample of species against published TCS lists confirmed that most domains were found, some Hpt domains were missed. However, the HATPase domain is large and not easily missed, and all unassigned HATPases (which may be part of transmitters) were included in this plot. Alternatively, perhaps there is another class of transmitter or phosphotransfer domain yet to be identified. The phosphotransfer protein Spo0B in *Bacillus subtilis* is one example of this (see Section 1.5).

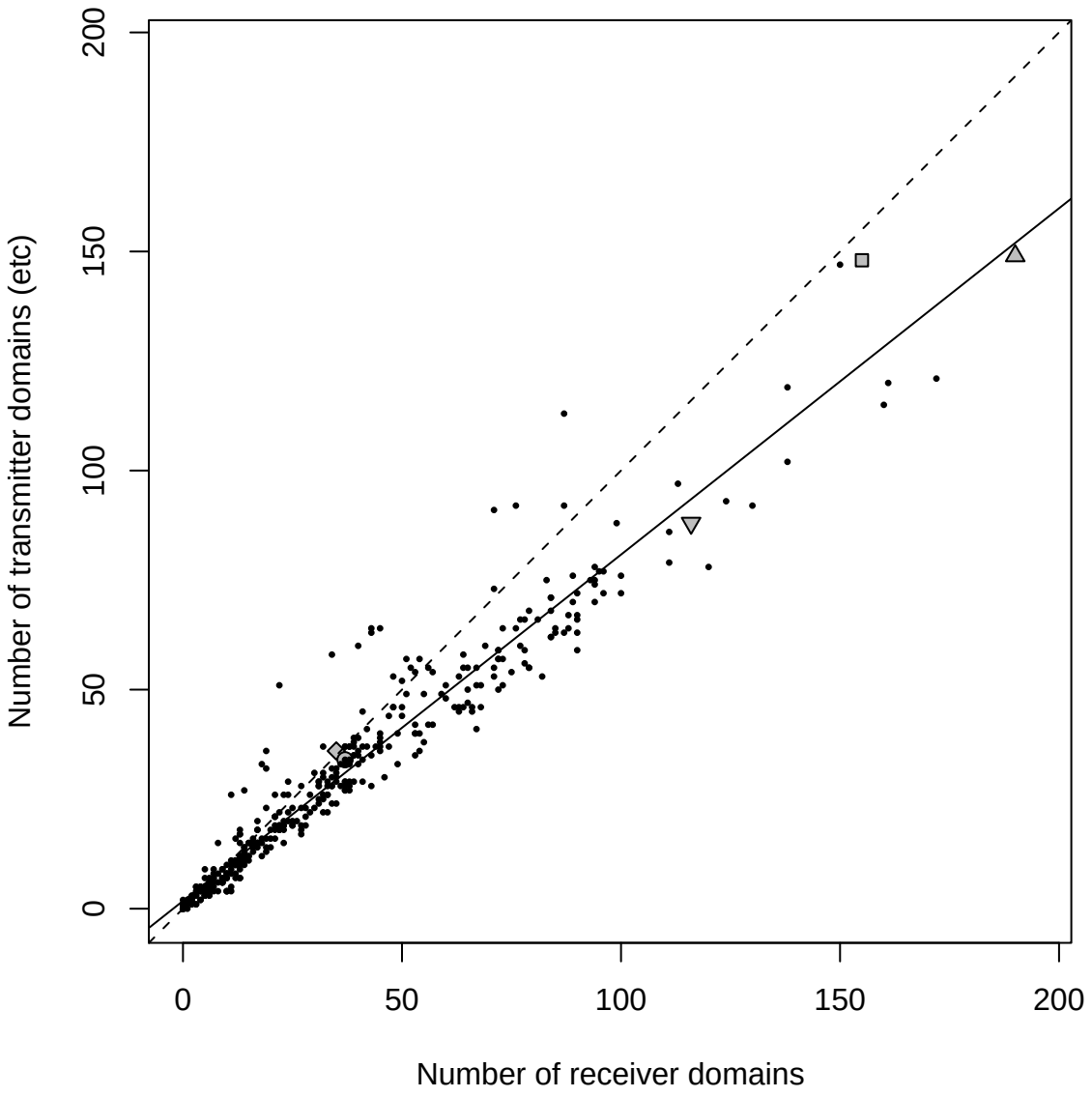


Figure 2.1: A scatter plot showing the number of receiver domains vs. the number of transmitter domains (including unassigned HATPases and phosphotransfer domains) for the 457 species listed in Table A.1. The dashed line has gradient one, as would be expected for a one-to-one pairing between transmitter and receiver domains. The solid line is a simple regression line. Certain species have been marked with a grey symbol: *Myxococcus xanthus* DK 1622 (upwards triangle), *Nostoc* sp. (square), *Anaeromyxobacter dehalogenans* 2CP-C (downwards triangle), *Bacillus subtilis* subsp. *subtilis* str. 168 (diamond) and *Escherichia coli* K12 (circle). See also Table 2.7.

Gene architecture	Isolated	Paired	Complex	Total
R	2769	5556	3493	11818
T_i	1498	4893	1807	8198
T_{ii}	17	35	232	284
K	136	120	81	337
KD	8	11	11	30
D	15	0	0	15
H	97	33	61	191
A	149	129	84	362
R-R	58	33	34	125
R-T_i	95	44	192	331
R-T_{ii}	0	0	0	0
T_i-R	764	195	467	1426
T_{ii}-R	1	18	69	88
T_i-R-H	103	64	95	262
T_i-R-R	74	18	33	125
T_i-R-R-H	17	39	32	88
R-T_i-R	16	16	32	64
Others	55	54	186	295
Total	5872	11258	6909	24039

Table 2.2: Domain architecture of all $24,039$ TCS genes identified from the 457 species listed in Table A.1, classified as isolated, paired or complex. Table 2.8 has a partial breakdown by species, while the paired genes are also listed in Table 2.3.

2.6.2 TCS architectures

Table 2.2 gives a break down of the TCS gene counts by domain architecture, with the additional classification into isolated, paired or complex based on the genomic arrangement. Notice that the typical HK (with a single T_i domain) and RR (with a single R domain) are by far the most common architectures, together making up over 80% of the TCS genes identified. Simple HK genes with a T_{ii} domain are comparatively rare, with only 284 examples found.

The third most common architecture is T_i-R, a simple hybrid kinase (HY) containing a single Class I transmitter (T_i) followed by a single receiver domain (R), with $1,424$ examples, followed by 331 cases of R-T_i where this domain order is reversed.

Table 2.2 includes 558 genes appearing to contain a phosphotransfer domain (K, KD, or H), but no HATPase or receiver domains, and 362 genes with an HATPase (A) but no other TCS associated domains. While some of these may be false positives, over half are next to or close to a TCS gene and thus may be real TCS genes - perhaps containing further unidentified domains. Further refinements to the analysis could reduce the number of these oddities, which currently make up less than 5% of the identified TCS genes. However, for the purposes of identifying most typical HK/RR pairs the results as shown are sufficient.

Table 2.2 also shows that while simple HK (with only a T domain) and RR genes (with only an R domain) dominate, there are considerable numbers of genes with more complex architectures. The "simple" hybrid kinases (HYs) with a single transmitter and single receiver are relatively common (T_i-R, R-T_i and T_{ii}-R, but interestingly no examples of R-T_{ii} were observed). These hybrid proteins can be regarded as self contained TCS pathways, as they contain a tethered transmitter-receiver pair, although being part of a phosophorelay cannot be ruled out. On the other hand, the T_i-R-H tripartite HYs are presumably part of phosophorelays, based on similar genes discussed in Section 1.4.

2.6.3 TCS gene pairs

Table 2.3 enumerates the $5,629$ TCS gene pairings for the most common TCS gene architectures, which will be described using an order-dependent plus sign notation for neighbouring genes. Given the most numerous TCS architectures are R and T_i (Table 2.2), unsurprisingly pairings between these genes are most the common, with $2,092$ cases of T_i + R and $2,689$ of R + T_i where the genes in the opposite order, making up about 85% of all TCS gene pairs. HK genes with the minority Class II transmitters are also found paired with RRs, 5 cases of T_{ii} +R and 30 cases of R+T_{ii}. Chapter 3 explores the separation or overlaps between these HK and RR gene pairs.

Chapter 4 compares the number of isolated T_i-R and R-T_i genes to the number of T_i +R and R+T_i gene pairs in an investigation of apparent gene fusion rates. These genes are also used in Chapters 5 and 6 to compile a set of known transmitter-receiver interactions.

In addition to these expected gene pairings, Table 2.3 also lists many less common pairings. 29 cases of T_i-R-H + R and 34 cases of R + T_i-R-H were identified (tripartite HY with RR), which could be two-gene phospho-relay systems like those described in Section 1.4.6. There were also 115 examples of T_i-R + R (simple HY with RR) and 53 pairs in the opposite order, R + T_i-R. These could function analogously to the VirA/G system discussed in Section 1.4.10, or alternatively they could be phospho-relays where an Hpt domain was not detected.

Other unexpected neighbours include 54 cases of R + R and 37 cases of T_i + T_i, as well as TCS gene pairing involving complicated hybrid genes. The R + R pairs are difficult to explain, given no evidence of receiver-receiver dimerization or phosophotransfer to date. On the other hand, HKs are known to dimerize so hypothetically these T_i + T_i pairs could be forming HK heterodimers, something that has yet to be demonstrated *in vivo*.

Paired	R	Ti	Tii	R-R	R-Ti	R-Tii	Ti-R	Tii-R	Ti-R-H	Ti-R-R-H	Ti-R-R	R-Ti-R	Ti-R-Ti	Other	Total
R	54	2689	30	23	4	0	53	10	34	3	5	1	0	87	2993
Ti	2092	37	0	5	8	0	2	0	0	0	0	6	0	12	2162
Tii	5	0	0	0	0	0	0	0	0	0	0	0	0	0	5
R-R	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
R-Ti	7	0	0	0	1	0	0	0	0	0	1	1	0	0	10
R-Tii	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ti-R	115	2	0	0	7	0	5	0	0	0	0	0	0	6	135
Tii-R	7	0	0	0	1	0	0	0	0	0	0	0	0	0	8
Ti-R-H	29	0	0	1	0	0	0	0	0	0	0	0	0	0	30
Ti-R-R-H	32	0	0	1	2	0	0	0	0	0	0	1	0	0	36
Ti-R-R	8	0	0	1	0	0	0	0	0	0	1	1	0	0	11
R-Ti-R	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
Ti-R-Ti	8	0	0	0	4	0	0	0	0	0	0	0	0	0	12
Other	205	3	0	1	6	0	0	0	0	0	0	5	0	5	225
Total	2563	2731	30	32	34	0	60	10	34	3	7	15	0	110	5629

Table 2.3: Domain architecture of all $5,629$ TCS gene pairs (i.e. $11,258$ genes) identified from the 457 species listed in Table A.1. The rows are the upstream gene of each pair, and the columns are the downstream gene – thus for example there are $2,092$ $\mathrm{T_i}$ + R examples of a typical HY (with a single $\mathrm{T_i}$ domain) followed by a typical RR (with a single R domain). See also Table 2.2

2.6.4 TCS associated domains

Table 2.4 lists the forty most common PFAM domains measured by the number of genes containing at least one of each domain, and how many of those were TCS genes. While the ABC transporter (pfam00005) is by far the most numerous domain, the RR receiver (pfam00072) and HATPase (pfam02518) domains are next – underlining just how ubiquitous TCS genes are. However, while most of these HATPase domains could be identified as part of transmitter domains, the HisKA domain (pfam00512) found in Class I transmitters (T_i) is only the eleventh most common domain in this table, with the Hpt domain found in Class II transmitters (T_{ii}) ranked number 618. This poor detection rate for the phosphotransfer domains led to the need for the various strategies described above in order to identify most transmitter domains.

Several of the other top forty domains are frequently associated with TCS systems, in particular the DNA binding domains trans-reg-C domain (pfam00486, 89%) and GerE domain (pfam00196, 57%), and the HAMP sensory domains (pfam00672, 53%). The second half of Table 2.4 shows other less common PFAM domains which are particularly associated with TCS genes. These include the HTH_8 (pfam02954, 51%) and LytTR (pfam04395, 71%) DNA-binding domains, and the PAS (pfam00989, 65%) and PAC (pfam00785, 54%) input associated domains, and the CHASE3 (pfam05227, 72%) sensor. The CheB_methylest domain (pfam01339, 73%) is a methylesterase output domain named after the *E. coli* RR CheB (Section 1.4.3). The KdpD domain (pfam02702, 79%) is named after the *E. coli* HK KdpD, and is believed to be an osmotic pressure sensory domain (Walderhaug *et al.*, 1992). Finally all fifty genes found with the P2 domain (pfam07194), a receiver binding domain first identified in *E. coli* CheA (see Section 1.4.3), were identified as TCS genes.

2.6.5 TCS associated input and output domains

Most of these TCS associated domains discussed above can be categorized into input or output domains. Their association with different TCS gene architectures is explored in Tables 2.5 and 2.6 (input and output domains respectively).

The HAMP domain is the most common PFAM input domain (Table 2.4), and is particularly associated with standard T_i HK genes ($1970/2221$ genes, 89%, Table 2.5). Also, the HAMP domain is the single most common input domain associated with the T_i HKs (found in $1970/8198$ T_i genes, 24%). However, in almost 90% of the T_i HK genes ($7237/8198$) there was a region of over 100 amino acids which was not recognised by any PFAM domain, suggesting that a large number of uncharacterised input domains remain to be identified.

Notice in Table 2.5 that the CheW-binding domain (pfam01584) is mostly found in

Rank	PFAM description	All genes	TCS genes	Fraction
1	pfam00005 ABC_tran	27209	14	0.00
2	pfam00072 Response_reg	14499	14499	1.00
3	pfam02518 HATPase_c	12752	11397	0.89
4	pfam00106 adh_short	10771	0	0.00
5	pfam03466 LysR_substrate	9127	7	0.00
6	pfam00126 HTH_1	9081	2	0.00
7	pfam00528 BPD_transp_1	9066	0	0.00
8	pfam00083 Sugar_tr	7195	1	0.00
9	pfam00583 Acetyltransf_1	7013	2	0.00
10	pfam00070 Pyr_redox	6912	12	0.00
11	pfam00512 HisKA	6680	6680	1.00
12	pfam04055 Radical_SAM	6315	0	0.00
13	pfam00702 Hydrolase	6123	0	0.00
14	pfam00665 rve	5699	6	0.00
15	pfam00004 AAA	5537	75	0.01
16	pfam00561 Abhydrolase_1	5458	0	0.00
17	pfam00440 TetR_N	5429	4	0.00
18	pfam00501 AMP-binding	5079	0	0.00
19	pfam00271 Helicase_C	4924	0	0.00
20	pfam01370 Epimerase	4834	0	0.00
21	pfam00270 DEAD	4716	0	0.00
22	pfam00990 GGDEF	4628	406	0.09
23	pfam00155 Aminotran_1_2	4464	1	0.00
24	pfam00486 Trans_reg_C	4439	3970	0.89
25	pfam00534 Glycos_transf_1	4436	0	0.00
26	pfam00535 Glycos_transf_2	4347	2	0.00
27	pfam00392 GntR	4298	14	0.00
28	pfam00672 HAMP	4198	2221	0.53
29	pfam00753 Lactamase_B	3954	1	0.00
30	pfam00009 GTP_EFTU	3894	0	0.00
31	pfam00171 Aldedh	3834	1	0.00
32	pfam00107 ADH_zinc_N	3804	1	0.00
33	pfam01073 3Beta_HSD	3775	0	0.00
34	pfam00593 TonB_dep_Rec	3585	0	0.00
35	pfam02653 BPD_transp_2	3499	0	0.00
36	pfam01381 HTH_3	3478	1	0.00
37	pfam02421 FeoB	3478	0	0.00
38	pfam00015 MCPsignal	3438	1	0.00
39	pfam00293 NUDIX	3416	1	0.00
40	pfam00196 GerE	3384	1920	0.57
107	pfam02954 HTH_8	1779	900	0.51
142	pfam00989 PAS	1474	965	0.65
230	pfam00785 PAC	1110	594	0.54
556	pfam04397 LytTR	541	382	0.71
618	pfam01627 Hpt	497	497	1.00
739	pfam01339 CheB_methylest	451	329	0.73
922	pfam06580 His_kinase	388	388	1.00
1153	pfam02895 H-kinase_dim	291	291	1.00
1579	pfam02702 KdpD	169	134	0.79
1602	pfam05227 CHASE3	165	118	0.72
2614	pfam07194 P2	50	50	1.00
	Total Genes	1435868	24039	0.02

Table 2.4: The top forty prokaryotic PFAM domains, measured by the number of genes containing each domain, and also selected domains found most often as part of TCS genes. This is for all annotated genes from the 457 species listed in Table A.1, using RPS-BLAST with an expectation threshold of 10^{-4}.

PFAM description	R	T_i	T_{ii}	R-R	R-T_i	R-T_{ii}	Ti-R	T_{ii}-R	T_i-R-H	T_i-R-R-H	T_i-R-R	R-T_i-R	T_i-R-T_i	Other	Total
03924 CHASE	0	28	0	0	0	0	14	0	3	11	4	0	0	4	64
05226 CHASE2	0	32	0	0	0	0	4	0	0	0	0	0	0	1	37
05227 CHASE3	0	73	0	0	0	0	21	0	1	3	3	0	0	17	118
05228 CHASE4	0	14	0	0	0	0	0	0	0	1	2	0	0	1	18
02743 Cache	0	44	1	0	0	0	17	0	0	2	7	0	0	1	72
01584 CheW	176	0	271	0	0	0	0	77	0	0	0	0	0	55	579
01590 GAF	21	335	0	3	9	0	63	0	3	10	7	6	0	34	491
00672 HAMP	1	1970	8	0	0	0	116	0	68	6	10	0	0	42	2221
02702 KdpD	0	134	0	0	0	0	0	0	0	0	0	0	0	0	134
05231 MASE1	0	27	0	0	0	0	12	0	1	0	0	0	0	1	41
00785 PAC	24	250	0	5	33	0	178	0	13	30	26	13	3	19	594
00989 PAS	49	527	0	2	51	0	213	0	26	26	26	13	0	32	965
00069 Pkinase	2	40	0	0	0	0	0	0	1	0	0	0	0	2	45
00497 SBP_bac_3	0	29	0	0	0	0	50	0	35	5	1	0	0	5	125
00474 SSF	0	24	0	0	0	0	59	0	0	0	0	0	0	0	83
00582 Usp	0	63	0	0	0	0	0	0	0	0	0	0	0	0	63
No input domain	10372	225	0	111	167	0	33	0	3	3	1	17	2	340	11274
Unidentified/other	1219	7237	25	14	119	0	1225	29	233	69	115	30	20	842	11177
Gene counts	11818	8198	284	125	331	0	1426	88	262	88	125	64	22	1208	24039

Table 2.5: Number of genes containing certain input associated PFAM domains, by TCS architecture, for all $24,039$ TCS genes identified from the 457 species listed in Table A.1. The table row "No input domain" counts genes lacking sufficient space to contain any further domains, while row "Unidentified/other" counts the number of genes containing a region of 100 amino acids or more which were not part of any TCS or PFAM domain. See Table 2.6 for the output domains.

PFAM description	R	T_i	T_{ii}	R-R	R-T_i	R-T_{ii}	Ti-R	T_{ii}-R	T_i-R-H	T_i-R-R-H	T_i-R-R	R-T_i-R	T_i-R-T_i	Other	Total
00004 AAA	3	0	0	0	0	0	0	0	0	0	0	0	0	0	3
03861 ANTAR	100	0	0	0	0	0	0	0	0	0	0	0	0	0	100
01339 CheB_methylest	306	8	0	0	0	0	14	0	0	0	1	0	0	0	329
00563 EAL	197	3	0	11	0	0	1	0	0	0	0	0	0	3	215
00990 GGDEF	304	3	0	77	0	0	5	0	0	0	0	0	0	17	406
00196 GerE	1894	0	0	0	0	0	4	0	0	0	3	0	0	2	1903
00211 Guanylate_cyc	13	0	0	0	0	0	4	0	0	0	0	3	0	4	24
01966 HD	147	2	0	0	0	0	0	0	0	0	0	0	0	1	150
00126 HTH_1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	2
01381 HTH_3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
01022 HTH_5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
02954 HTH_8	899	0	0	0	0	0	0	0	0	0	0	0	0	1	900
04967 HTH_10	4	0	0	0	0	0	0	0	0	0	0	0	0	0	4
00165 HTH_AraC	80	1	0	0	0	0	36	0	0	0	0	0	0	1	118
04397 LytTR	382	0	0	0	0	0	0	0	0	0	0	0	0	0	382
00158 Sigma54_activat	1138	0	0	0	1	0	0	0	0	0	0	0	0	1	1140
04545 Sigma70_r4	17	0	0	0	0	0	0	0	0	0	0	0	0	0	17
07228 SpoIIE	62	0	0	0	0	0	0	0	0	0	0	0	0	4	66
00440 TetR_N	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4
00486 Trans_reg_C	3962	0	0	0	0	0	0	0	0	0	0	0	0	8	3970
No output domain	2253	961	259	45	212	0	197	59	29	19	10	31	2	347	4424
Unidentified/other	1219	7237	25	14	119	0	1225	29	233	69	115	30	20	842	11177
Gene counts	11818	8198	284	125	331	0	1426	88	262	88	125	64	22	1208	24039

Table 2.6: Number of genes containing certain output associated PFAM domains, by TCS architecture, for all $24,039$ TCS genes identified from the 457 species listed in Table A.1. See Table 2.5 for the input domains.

simple HKs with T_{ii} domains (like the model system, CheA in *E. coli*, see Section 1.4.3), and never found in a simple HK with a T_i. Intriguingly CheW domains have also been found in simple RRs (with an R domain only), and in simple T_{ii}-R HYs. About 10% of the CheW-binding domains are found in genes with other more complicated TCS architectures (like the cheBRA gene discussed in Section 1.4.4). These figures also support the widely observed rule-of-thumb that T_{ii} are only associated with chemotaxis (over 95% of T_{ii} HK genes contain a CheW-binding domain).

Reassuringly, only a handful of input domains were found in the $11,818$ simple RR genes (or in more complex RR architectures such as R-R). In addition to the 196 CheW domains discussed above (pfam01584, 1.6%), there were 21 GAF domains (pfam01590, 0.2%), 24 PAC domains (pfam00785, 0.2%) and 49 PAS domains (pfam00989, 0.4%) plus a single HAMP domain (pfam00672) and two Pkinase domains (pfam00069). These numbers are too low to question the classification of these PFAM domains as input associated.

Similarly, very few output domains are found in the T_i or T_{ii} HKs (Table 2.6). Of particular note, 19% of simple RRs have no output domain at all ($2253/11818$). These may function like CheY in the *E. coli* chemotaxis system (Section 1.4.3) by directly interacting with other proteins, or some of these RRs could be part of phosphorelays like Spo0F in the *Bacillus subtilis* sporulation network (Section 1.4.8).

Interestingly no recognised output domains were found in the 262 T_i-R-H HY genes (Table 2.6), although a number of known input domains were found (Table 2.5). This is consistent with the idea that these function as the first step in a phosphorelay (see Sections 1.4.6 and 1.4.9). The same presence of some input domains but complete lack of output domains is observed for the 88 T_i-R-R-H HY genes, which may function in a similar way.

2.6.6 Species specific remarks

Figures 2.2, 2.3 and 2.4 shows how the number of TCS domains (or genes) increases with the total number of genes or genome size. These all illustrate that species with larger genomes tend to have more TCS domains and genes. As suggested in van Nimwegen (2003), these trends are well described using a squared power law distribution.

Figure 2.5 illustrates how this trend is linked to the species' lifestyle, showing that non-pathogenic prokaryotes tend to have more TCS genes for their size than pathogenic species. This can be explained on the assumption that pathogens enjoy a much less variable environment in their host than environmental organisms.

Out of the 340 representative prokaryotes in this survey, only 64 had over a hundred

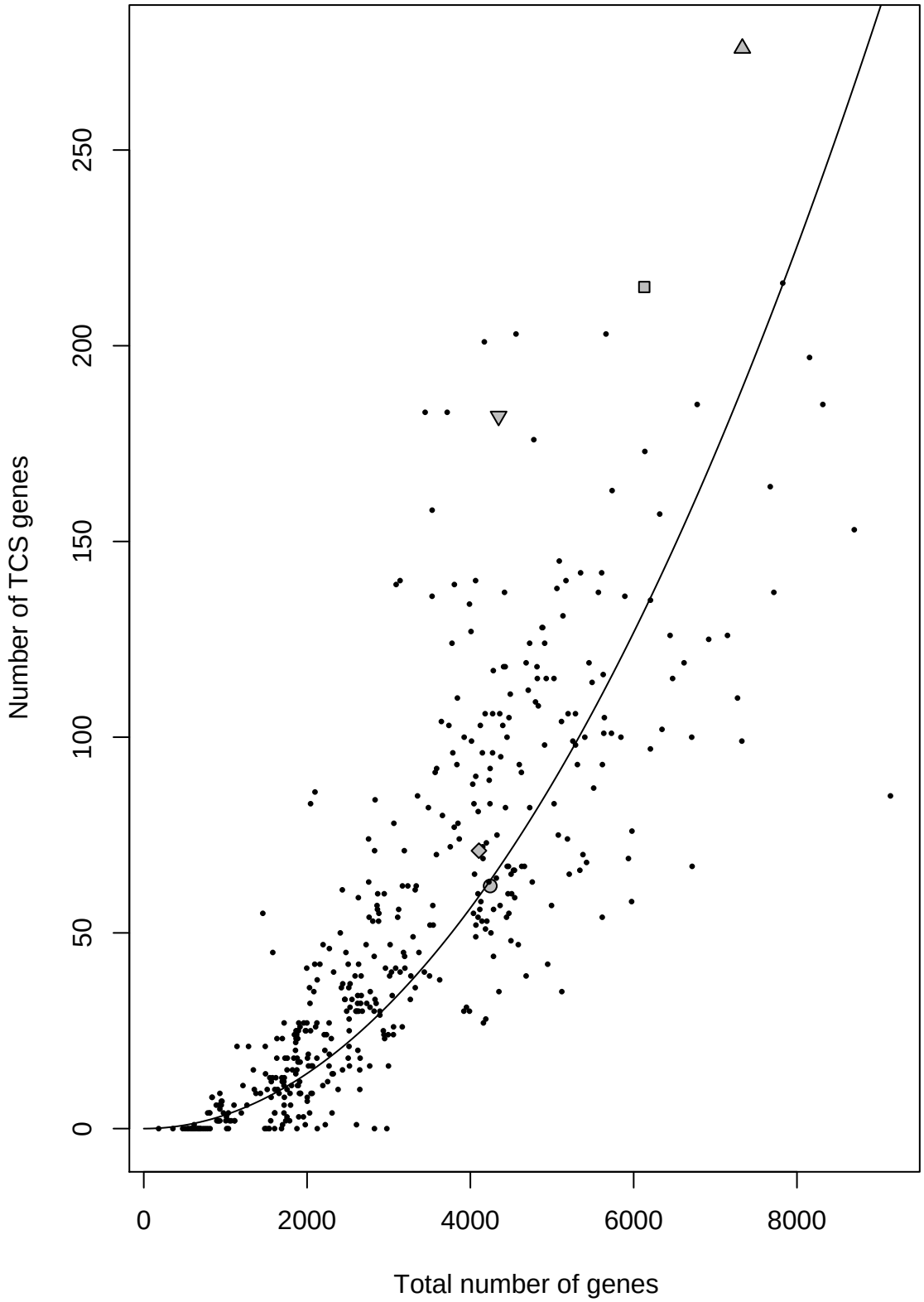


Figure 2.2: A scatter plot showing the number of TCS genes vs. the total number of genes for the 457 species listed in Table A.1. The solid line is a fitted square scaling through the origin. Symbol key as per Figure 2.1.

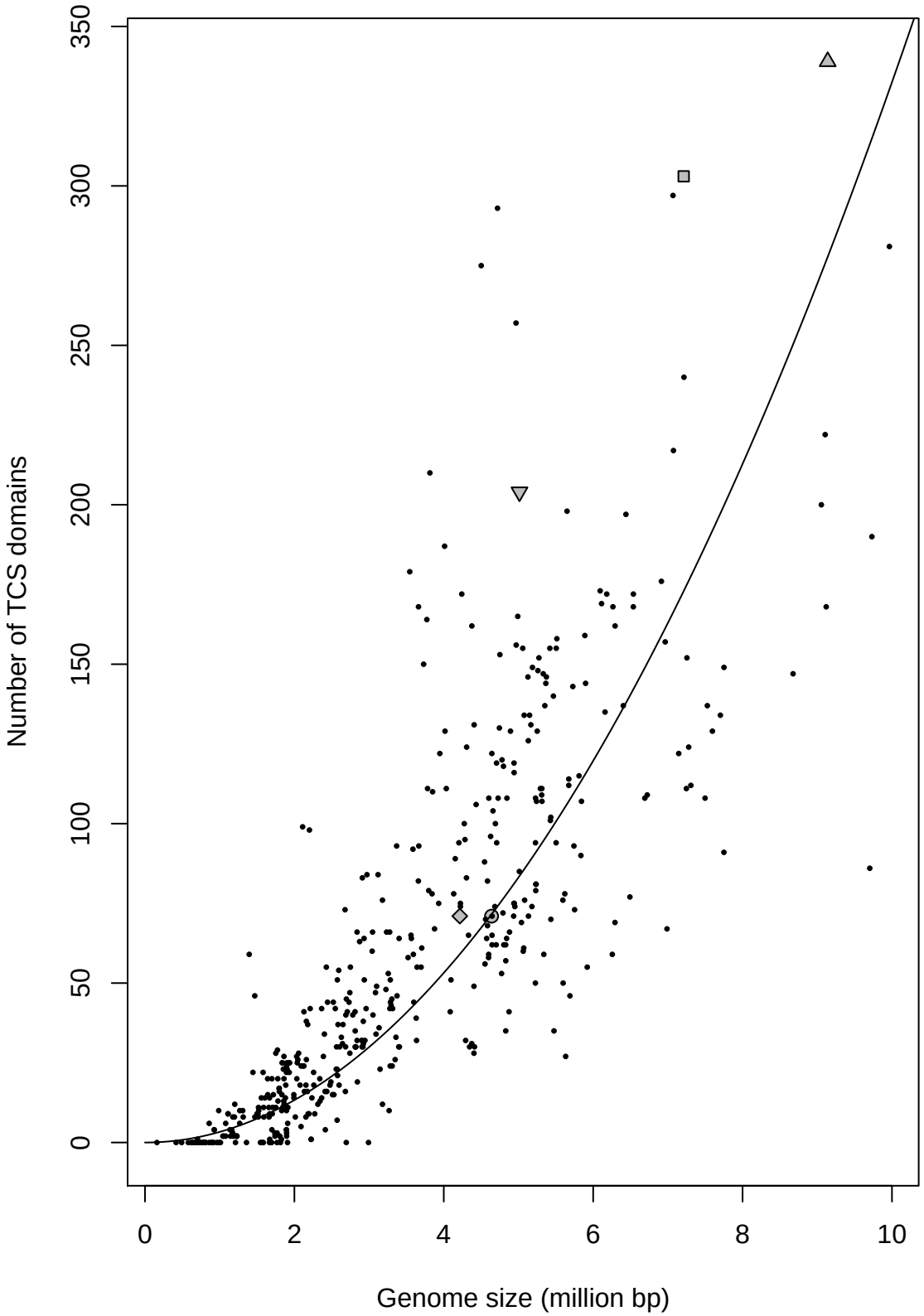


Figure 2.3: A scatter plot showing the number of TCS domains vs. the genome size for the 457 species listed in Table A.1. The solid line is a fitted square scaling through the origin. Symbol key as per Figure 2.1. See also Table 2.7.

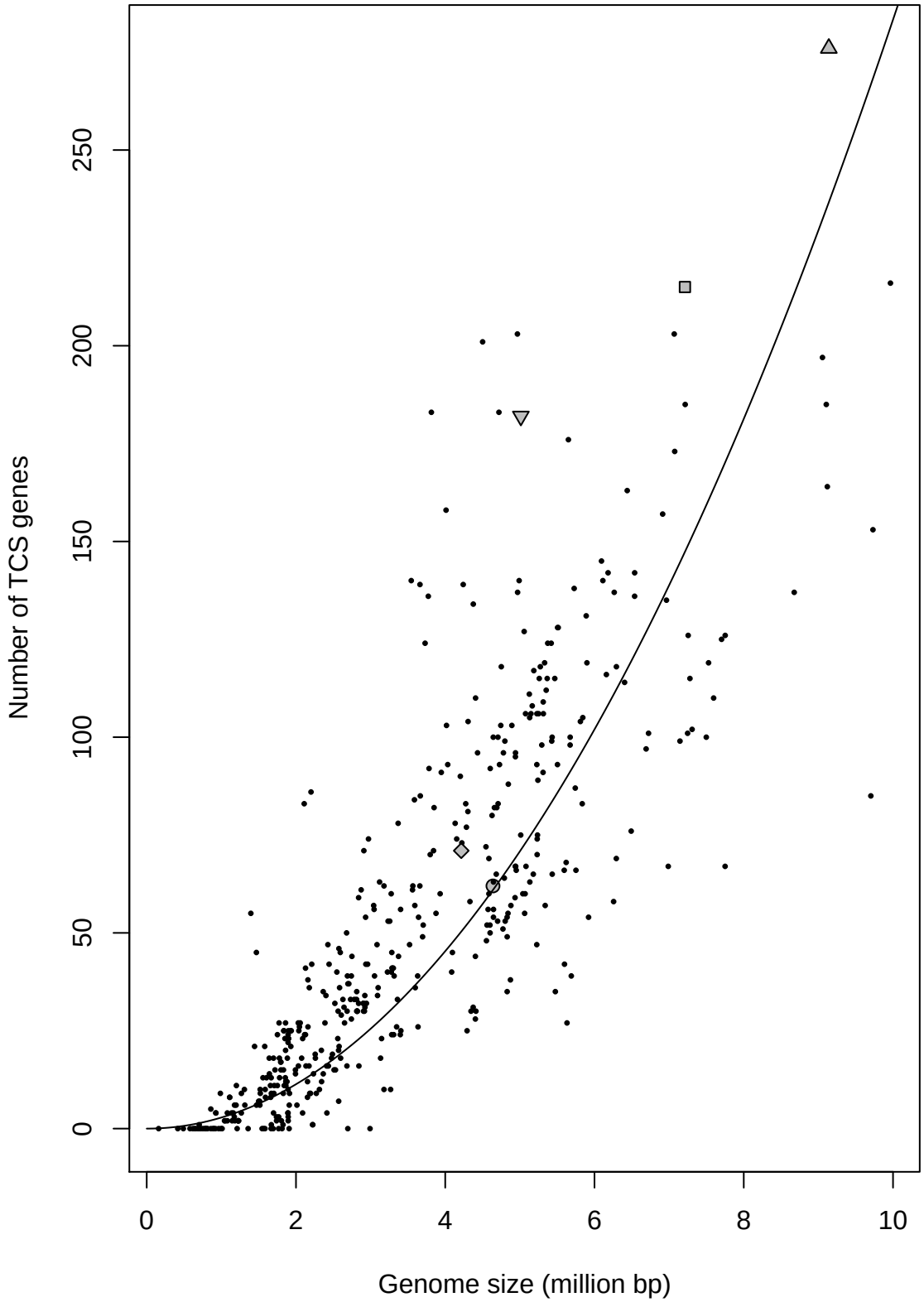


Figure 2.4: A scatter plot showing the number of TCS genes vs. the genome size for the 457 species listed in Table A.1. The solid line is a fitted square scaling through the origin. Symbol key as per Figure 2.1. See also Table 2.8.

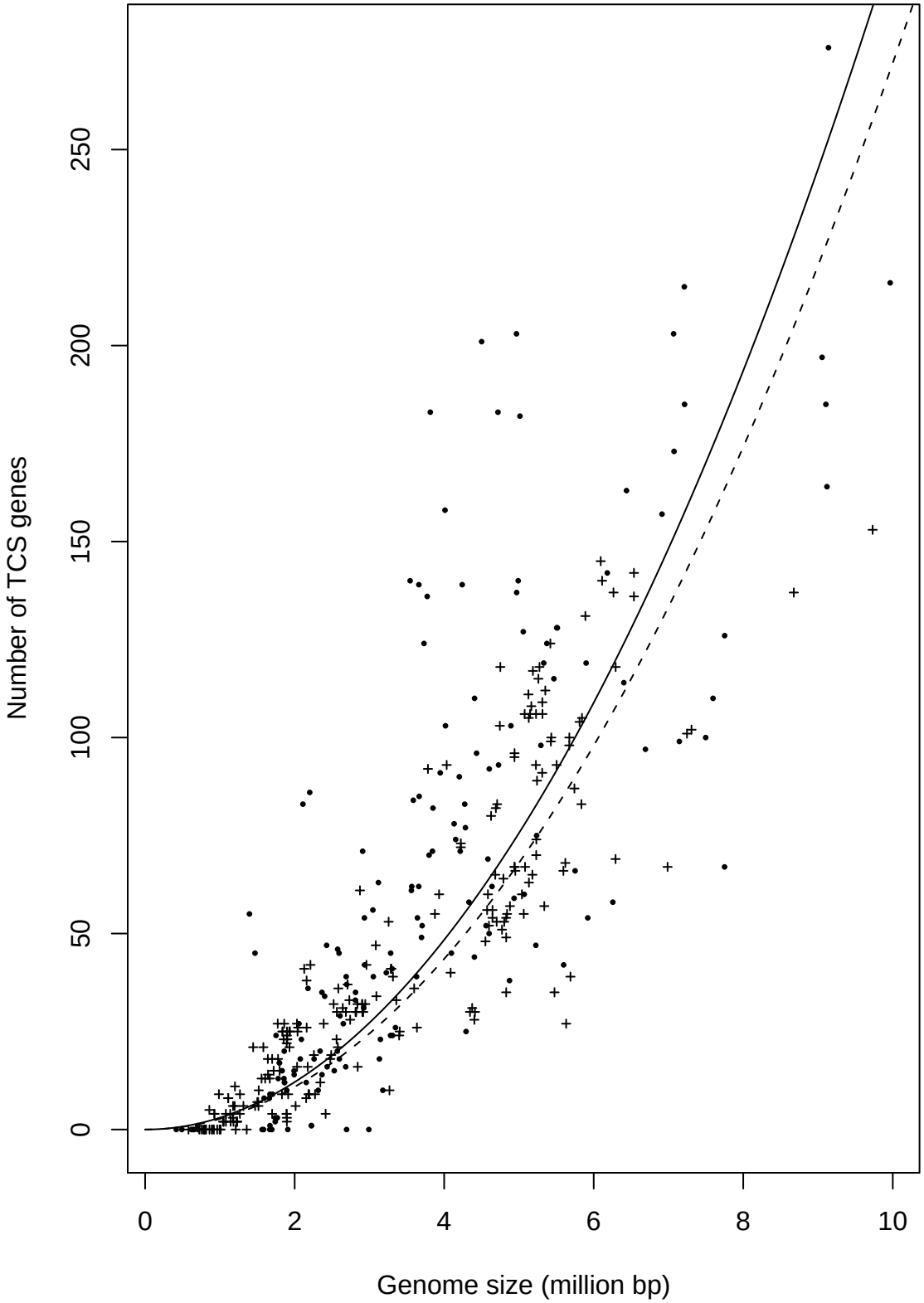


Figure 2.5: A scatter plot showing the number of TCS genes vs. the genome size for 382 of the species listed in Table A.1 which the NCBI had categorized as pathogenic (plus symbol, dashed line) or non-pathogenic (dots, solid line). See also Table 2.4.

TCS genes. These species are listed alphabetically in Tables 2.7 and 2.8 which show the number of TCS domains (R, T_i and T_{ii} etc) and the number of TCS genes (divided by domain architecture). The vast majority of these species are bacteria, with only two archaea: *Acidobacteria bacterium* Ellin345 (with 198 TCS genes) and *Methanospirillum hungatei* JF-1 (with 179 TCS genes).

Figures 2.6, 2.7 and 2.8 show scatter plots illustrating what fraction of each species' TCS genes are found together in pairs, isolated or in some other more complex arrangement. These figures attempt to represent how different species organise their TCS genes. For example, marked species *Bacillus subtilis* and *E. coli* have almost all of their TCS genes as pairs. On the other hand, larger genomes with more TCS genes tend to have less in pairs, and instead have a higher proportion of isolated (orphan) TCS genes and more complicated clusters. In particular from the marked species, *Nostoc* sp. appears to have more isolated TCS genes than normal, while the myxobacteria *M. xanthus* and *Anaeromyxobacter dehalogenans* have more TCS genes in complex clusters.

Figure 2.9 shows the number of TCS domains plotted against the number of TCS genes in each species. It is quite clear that the number of TCS domains increases faster than the number of TCS genes. That is to say, genomes with more TCS genes tend to have more complex TCS architectures (genes containing more than one TCS domain). From its position in this plot, it is not surprising that *M. xanthus* has many TCS hybrid genes with complex multi-domain architectures, such as the protein RodK (see Section 1.4.11).

From these plots (and the underlying tables) it is clear that prokaryotes with especially large numbers of TCS tend to have more complicated TCS arrangements. The myxobacteria, and *M. xanthus* in particular, are an extreme case and thus a useful model-species for the investigation of complex TCS systems (Whitworth and Cock, 2008a,b).

2.7 Potential refinements

2.7.1 More efficient searching

As implemented, every gene was searched using RPS-BLAST against the full set of PFAM domains, and then any candidate TCS genes were searched against the full set of CDD domains – each time recording the full results. This gave a rich database allowing enquiries to further refine the analysis (such as identifying HATPase associated domains), as well as identifying other TCS associated domains (potential input and output domains, e.g. Table 2.4).

However, from the point of view of identifying TCS genes, it would be more efficient

Kingdom & Group	Species	R	T_i	T_{ii}	A	K	H	Others	Total
B Acidobacteria	*Acidobacteria bacterium* Ellin345	120	64	2	2	7	3	0	198
B Betaproteobacteria	*Acidovorax avenae* subsp. *citrulli* AAC00-1	73	54	2	1	1	5	1	137
B Gammaproteobacteria	*Aeromonas hydrophila* subsp. *hydrophila* ATCC 7966	73	43	3	1	1	9	0	130
B Cyanobacteria	*Anabaena variabilis* ATCC 29413	150	125	1	1	9	9	2	297
B Deltaproteobacteria	*Anaeromyxobacter dehalogenans* 2CP-C	116	77	7	2	1	0	1	204
B Betaproteobacteria	*Azoarcus* sp. BH72	90	59	3	0	2	7	1	162
B Firmicutes	*Bacillus thuringiensis* str. *Al Hakam*	54	55	1	1	0	0	0	111
B Bacteroidetes/Chlorobi	*Bacteroides thetaiotaomicron* VPI-5482	71	77	0	3	11	0	0	162
B Alphaproteobacteria	*Bradyrhizobium japonicum* USDA 110	130	80	3	2	3	4	0	222
B Betaproteobacteria	*Burkholderia* sp. 383	81	58	3	0	3	1	1	147
B Betaproteobacteria	*Burkholderia cenocepacia* AU 1054	73	47	3	0	0	1	0	124
B Betaproteobacteria	*Burkholderia ambifaria/cepacia* AMMD	77	54	3	0	1	1	1	137
B Betaproteobacteria	*Burkholderia pseudomallei 1710b*	65	44	2	0	1	0	0	112
B Betaproteobacteria	*Burkholderia thailandensis* E264	63	43	3	0	0	0	0	109
B Betaproteobacteria	*Burkholderia fungorum/xenovorans* LB400	111	74	2	2	0	1	0	190
B Alphaproteobacteria	*Caulobacter crescentus* CB15	72	50	2	1	1	3	0	129
B Betaproteobacteria	*Chromobacterium violaceum* ATCC 12472	90	46	4	0	1	12	0	153
B Gammaproteobacteria	*Colwellia psychrerythraea* 34H	84	53	1	0	3	5	0	146
B Betaproteobacteria	*Dechloromonas aromatica* RCB	160	98	4	1	1	10	1	275
B Firmicutes	*Desulfitobacterium hafniense* Y51	77	62	1	3	0	0	0	143
B Deltaproteobacteria	*Desulfovibrio desulfuricans* G20	87	46	2	4	3	8	0	150
B Deltaproteobacteria	*Desulfovibrio vulgaris* subsp. *vulgaris* DP4	96	61	3	1	1	5	1	168
B Actinobacteria	*Frankia alni ACN14a*	59	39	0	7	1	2	0	108
B Deltaproteobacteria	*Geobacter metallireducens* GS-15	99	79	5	0	1	2	1	187
B Deltaproteobacteria	*Geobacter sulfurreducens* PCA	113	88	4	0	0	5	0	210
B Gammaproteobacteria	*Hahella chejuensis* KCTC 2396	138	72	5	1	4	19	1	240
B Other Bacteria	*Magnetococcus* sp. MC-1	172	95	5	0	1	19	1	293
B Alphaproteobacteria	*Magnetospirillum magneticum* AMB-1	138	102	2	1	3	11	0	257
B Alphaproteobacteria	*Mesorhizobium loti* MAFF303099	75	48	1	3	1	1	0	129
A Euryarchaeota	*Methanospirillum hungatei* JF-1	87	71	3	2	15	1	0	179

Continued...

Table 2.7: TCS domain counts from 64 species with more than 100 TCS genes, selected from the 340 representative species listed in bold in Table A.1. A and B denote Archaea and Bacteria as the kingdom. TCS gene architectures are listed in Table 2.8. See also Figure 2.3.

Kingdom & Group	Species	R	T_i	T_{ii}	A	K	H	Others	Total
B Betaproteobacteria	*Methylibium petroleiphilum* PM1	64	49	1	2	1	4	1	122
B Deltaproteobacteria	*Myxococcus xanthus* DK 1622	190	127	8	5	7	2	0	339
B Cyanobacteria	*Nostoc* sp. PCC 7120	155	130	1	2	6	7	2	303
B Deltaproteobacteria	*Pelobacter propionicus* DSM 2379	94	70	4	0	1	3	0	172
B Gammaproteobacteria	*Photobacterium profundum* SS9	78	46	2	3	1	7	0	137
B Betaproteobacteria	*Polaromonas* sp. JS666	78	56	2	1	1	5	1	144
B Betaproteobacteria	*Polaromonas naphthalenivorans* CJ2	71	63	1	1	1	5	2	144
B Gammaproteobacteria	*Pseudoalteromonas atlantica T6c*	85	52	1	0	4	7	0	149
B Gammaproteobacteria	*Pseudomonas aeruginosa* PAO1	93	58	4	0	1	12	0	168
B Gammaproteobacteria	*Pseudomonas entomophila* L48	89	56	3	2	0	9	0	159
B Gammaproteobacteria	*Pseudomonas fluorescens* Pf-5	124	77	3	1	0	12	0	217
B Gammaproteobacteria	*Pseudomonas putida* KT2440	100	63	3	1	0	5	0	172
B Gammaproteobacteria	*Pseudomonas syringae* pv. *phaseolicola* 1448A	94	62	4	1	1	7	0	169
B Betaproteobacteria	*Ralstonia eutropha* H16	90	59	2	1	0	4	1	157
B Betaproteobacteria	*Ralstonia metallidurans* CH34	100	63	2	2	1	7	1	176
B Betaproteobacteria	*Ralstonia solanacearum* GMI1000	65	42	1	1	0	4	2	115
B Alphaproteobacteria	*Rhizobium etli* CFN 42	82	47	2	2	1	1	0	135
B Alphaproteobacteria	*Rhizobium leguminosarum bv. viciae* 3841	90	55	2	1	0	1	0	149
B Betaproteobacteria	*Rhodoferax ferrireducens* DSM 15236, T118	90	58	2	0	2	3	1	156
B Alphaproteobacteria	*Rhodopseudomonas palustris* BisA53	84	56	5	4	2	4	0	155
B Alphaproteobacteria	*Rhodospirillum rubrum* ATCC 11170	72	49	3	1	0	6	0	131
B Gammaproteobacteria	*Saccharophagus degradans* 2-40	84	50	3	3	6	8	1	155
B Gammaproteobacteria	*Shewanella* sp. ANA-3	72	43	2	0	1	7	4	129
B Gammaproteobacteria	*Shewanella* sp. MR-4	68	42	2	1	0	6	0	119
B Gammaproteobacteria	*Shewanella amazonensis* SB2B	71	43	2	0	1	7	0	124
B Gammaproteobacteria	*Shewanella oneidensis* MR-1	71	40	2	1	3	8	1	126
B Acidobacteria	*Solibacter usitatus* Ellin6076	161	108	1	2	6	3	0	281
B Actinobacteria	*Streptomyces avermitilis* MA-4680	76	67	0	23	0	0	2	168
B Actinobacteria	*Streptomyces coelicolor* A3(2)	87	77	0	33	0	0	3	200
B Deltaproteobacteria	*Syntrophobacter fumaroxidans* MPOB	89	71	1	1	2	0	1	165
B Gammaproteobacteria	*Vibrio parahaemolyticus* RIMD 2210633	72	46	1	1	2	9	0	131
B Gammaproteobacteria	*Vibrio vulnificus* CMCP6	84	48	2	1	0	11	0	146
B Gammaproteobacteria	*Xanthomonas campestris* pv. *campestris* str. ATCC 33913	79	47	4	1	1	2	0	134
B Gammaproteobacteria	*Xanthomonas axonopodis* pv. *citri* str. 306	84	53	3	1	3	7	1	152

Kingdom & Group	Species	R	T_i	T_{ii}	R-R	R-T_{ii}	T_i-R	T_{ii}-R	T_i-R-H	Others	Total
B Acidobacteria	*Acidobacteria bacterium* Ellin345	100	48	2	1	0	9	0	0	16	176
B Betaproteobacteria	*Acidovorax avenae* subsp. *citrulli* AAC00-1	53	38	2	0	0	10	0	0	9	112
B Gammaproteobacteria	*Aeromonas hydrophila* subsp. *hydrophila* ATCC 7966	53	28	3	1	0	5	0	6	7	103
B Cyanobacteria	*Anabaena variabilis* ATCC 29413	70	69	0	0	0	21	1	0	42	203
B Deltaproteobacteria	*Anaeromyxobacter dehalogenans* 2CP-C	93	60	5	2	0	10	2	0	10	182
B Betaproteobacteria	*Azoarcus* sp. BH72	70	45	3	1	0	6	0	0	9	134
B Firmicutes	*Bacillus thuringiensis* str. *Al Hakam*	52	53	1	0	0	2	0	0	1	109
B Bacteroidetes/Chlorobi	*Bacteroides thetaiotaomicron* VPI-5482	28	34	0	0	0	42	0	0	14	118
B Alphaproteobacteria	*Bradyrhizobium japonicum* USDA 110	93	51	1	1	0	24	2	0	13	185
B Betaproteobacteria	*Burkholderia* sp. 383	71	49	2	0	0	8	1	0	6	137
B Betaproteobacteria	*Burkholderia cenocepacia* AU 1054	64	39	2	0	0	7	1	0	2	115
B Betaproteobacteria	*Burkholderia ambifaria/cepacia* AMMD	59	41	2	1	0	11	1	0	4	119
B Betaproteobacteria	*Burkholderia pseudomallei 1710b*	55	35	1	0	0	7	1	0	3	102
B Betaproteobacteria	*Burkholderia thailandensis* E264	55	36	2	0	0	5	1	0	2	101
B Betaproteobacteria	*Burkholderia fungorum/xenovorans* LB400	75	47	2	0	0	16	0	0	13	153
B Alphaproteobacteria	*Caulobacter crescentus* CB15	45	25	2	1	0	25	0	0	5	103
B Betaproteobacteria	*Chromobacterium violaceum* ATCC 12472	65	29	4	0	0	5	0	4	11	118
B Gammaproteobacteria	*Colwellia psychrerythraea* 34H	65	38	1	1	0	10	0	3	6	124
B Betaproteobacteria	*Dechloromonas aromatica* RCB	91	51	3	4	0	29	0	2	21	201
B Firmicutes	*Desulfitobacterium hafniense* Y51	74	57	1	0	0	1	0	0	5	138
B Deltaproteobacteria	*Desulfovibrio desulfuricans* G20	63	29	1	1	0	9	0	1	20	124
B Deltaproteobacteria	*Desulfovibrio vulgaris* subsp. *vulgaris* DP4	69	39	2	1	0	13	0	0	15	139
B Actinobacteria	*Frankia alni ACN14a*	52	32	0	0	0	6	0	1	9	100
B Deltaproteobacteria	*Geobacter metallireducens* GS-15	72	56	4	0	0	7	1	2	16	158
B Deltaproteobacteria	*Geobacter sulfurreducens* PCA	89	65	4	0	0	13	0	3	9	183
B Gammaproteobacteria	*Hahella chejuensis* KCTC 2396	95	48	3	4	0	11	2	4	18	185
B Other Bacteria	*Magnetococcus* sp. MC-1	75	30	2	0	0	39	3	6	28	183
B Alphaproteobacteria	*Magnetospirillum magneticum* AMB-1	92	74	0	0	0	12	2	3	20	203
B Alphaproteobacteria	*Mesorhizobium loti* MAFF303099	55	36	0	1	0	6	1	0	11	110
A Euryarchaeota	*Methanospirillum hungatei* JF-1	53	39	1	0	0	0	2	0	45	140

Continued...

Table 2.8: Domain architecture of TCS genes from 64 species with more than 100 TCS genes, selected from the 340 representative species listed in bold in Table A.1. A and B denote Archaea and Bacteria as the kingdom. TCS domain counts are listed in Table 2.7. See also Figure 2.4

Kingdom & Group	Species	R	T_i	T_{ii}	R-R	R-T_{ii}	T_i-R	T_{ii}-R	T_i-R-H	Others	Total
B Betaproteobacteria	*Methylibium petroleiphilum* PM1	48	33	1	0	0	11	0	0	7	100
B Deltaproteobacteria	*Myxococcus xanthus* DK 1622	126	92	3	3	0	14	5	0	33	276
B Cyanobacteria	*Nostoc* sp. PCC 7120	79	73	0	0	0	20	1	0	42	215
B Deltaproteobacteria	*Pelobacter propionicus* DSM 2379	62	43	4	1	0	20	0	1	8	139
B Gammaproteobacteria	*Photobacterium profundum* SS9	58	31	2	2	0	8	0	5	8	114
B Betaproteobacteria	*Polaromonas* sp. JS666	58	41	1	0	0	4	1	0	14	119
B Betaproteobacteria	*Polaromonas naphthalenivorans* CJ2	49	44	1	0	0	9	0	0	12	115
B Gammaproteobacteria	*Pseudoalteromonas atlantica T6c*	58	32	1	1	0	9	0	5	11	117
B Gammaproteobacteria	*Pseudomonas aeruginosa* PAO1	70	41	2	1	0	13	1	2	7	137
B Gammaproteobacteria	*Pseudomonas entomophila* L48	68	41	1	0	0	9	1	2	9	131
B Gammaproteobacteria	*Pseudomonas fluorescens* Pf-5	90	48	1	0	0	20	1	5	8	173
B Gammaproteobacteria	*Pseudomonas putida* KT2440	73	44	1	1	0	14	1	0	8	142
B Gammaproteobacteria	*Pseudomonas syringae* pv. *phaseolicola* 1448A	70	43	2	0	0	15	1	1	8	140
B Betaproteobacteria	*Ralstonia eutropha* H16	72	43	1	0	0	13	1	0	5	135
B Betaproteobacteria	*Ralstonia metallidurans* CH34	86	50	1	0	0	10	1	1	8	157
B Betaproteobacteria	*Ralstonia solanacearum* GMI1000	58	35	1	0	0	5	0	0	5	104
B Alphaproteobacteria	*Rhizobium etli* CFN 42	61	35	2	2	0	6	0	0	10	116
B Alphaproteobacteria	*Rhizobium leguminosarum bv. viciae* 3841	65	41	2	2	0	9	0	0	7	126
B Betaproteobacteria	*Rhodoferax ferrireducens* DSM 15236, T118	74	46	2	1	0	6	0	0	8	137
B Alphaproteobacteria	*Rhodopseudomonas palustris* BisA53	59	34	3	1	0	17	2	2	10	128
B Alphaproteobacteria	*Rhodospirillum rubrum* ATCC 11170	52	36	2	1	0	8	1	1	9	110
B Gammaproteobacteria	*Saccharophagus degradans* 2-40	62	37	2	1	0	8	1	0	16	127
B Gammaproteobacteria	*Shewanella* sp. ANA-3	54	31	2	0	0	4	0	4	11	106
B Gammaproteobacteria	*Shewanella* sp. MR-4	52	33	2	1	0	3	0	3	6	100
B Gammaproteobacteria	*Shewanella amazonensis* SB2B	54	33	2	2	0	3	0	4	6	104
B Gammaproteobacteria	*Shewanella oneidensis* MR-1	54	32	2	2	0	1	0	5	9	105
B Acidobacteria	*Solibacter usitatus* Ellin6076	98	59	1	1	0	34	0	0	23	216
B Actinobacteria	*Streptomyces avermitilis* MA-4680	72	63	0	0	0	4	0	0	25	164
B Actinobacteria	*Streptomyces coelicolor* A3(2)	84	75	0	0	0	2	0	0	36	197
B Deltaproteobacteria	*Syntrophobacter fumaroxidans* MPOB	65	49	1	0	0	16	0	0	9	140
B Gammaproteobacteria	*Vibrio parahaemolyticus* RIMD 2210633	54	29	1	0	0	11	0	5	8	108
B Gammaproteobacteria	*Vibrio vulnificus* CMCP6	56	28	2	2	0	10	0	7	6	111
B Gammaproteobacteria	*Xanthomonas campestris* pv. *campestris* str. ATCC 33913	52	28	3	1	0	13	0	1	8	106
B Gammaproteobacteria	*Xanthomonas axonopodis* pv. *citri* str. 306	56	32	3	1	0	12	0	2	12	118

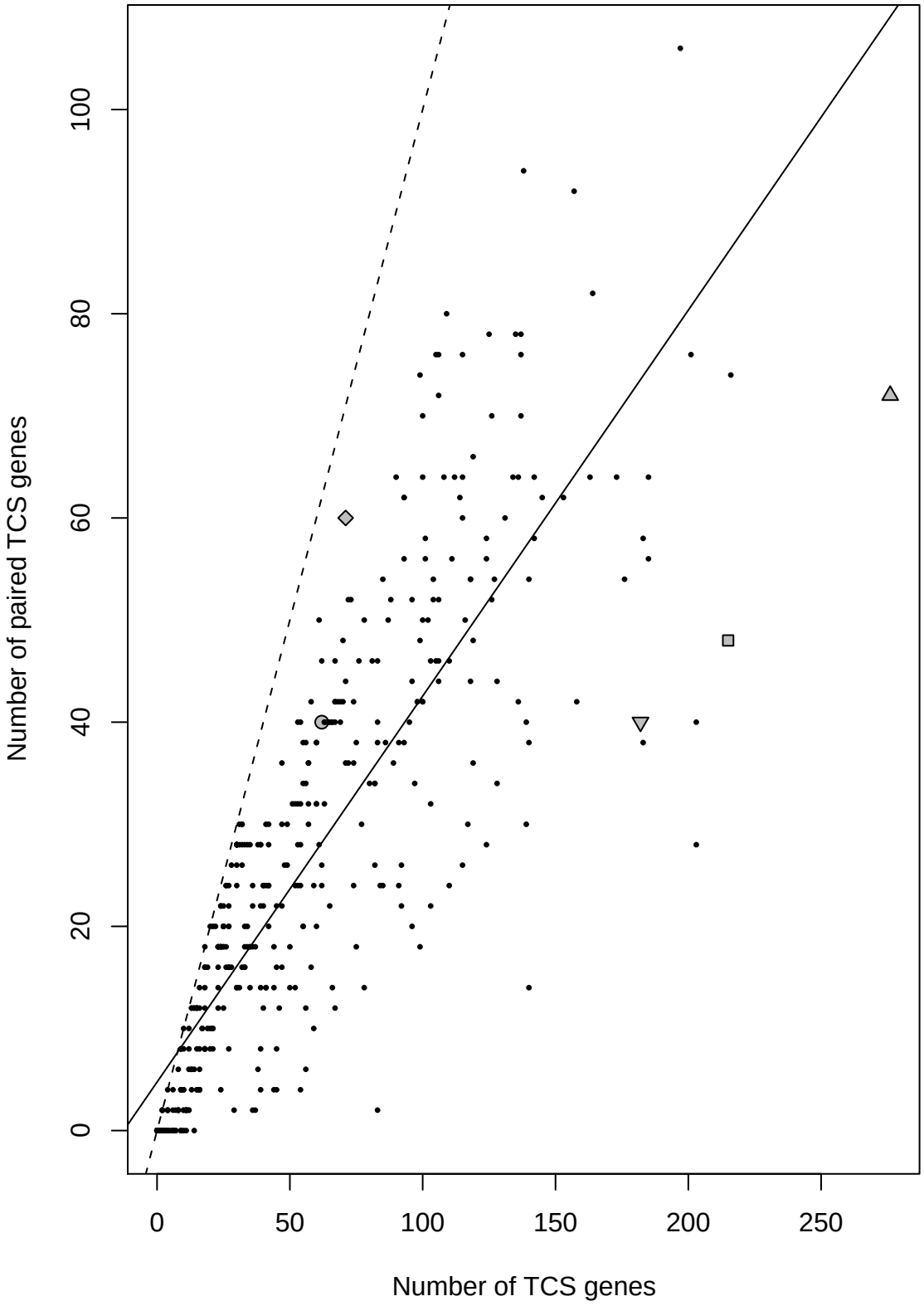


Figure 2.6: A scatter plot showing the number of TCS genes vs. the number of paired TCS genes for the 457 species listed in Table A.1. The dotted line has gradient one, and thus any point on or close to this line has all or more of its TCS genes in pairs. The solid line is a simple trend line. Symbol key as per Figure 2.1. See also Figures 2.7 and 2.8.

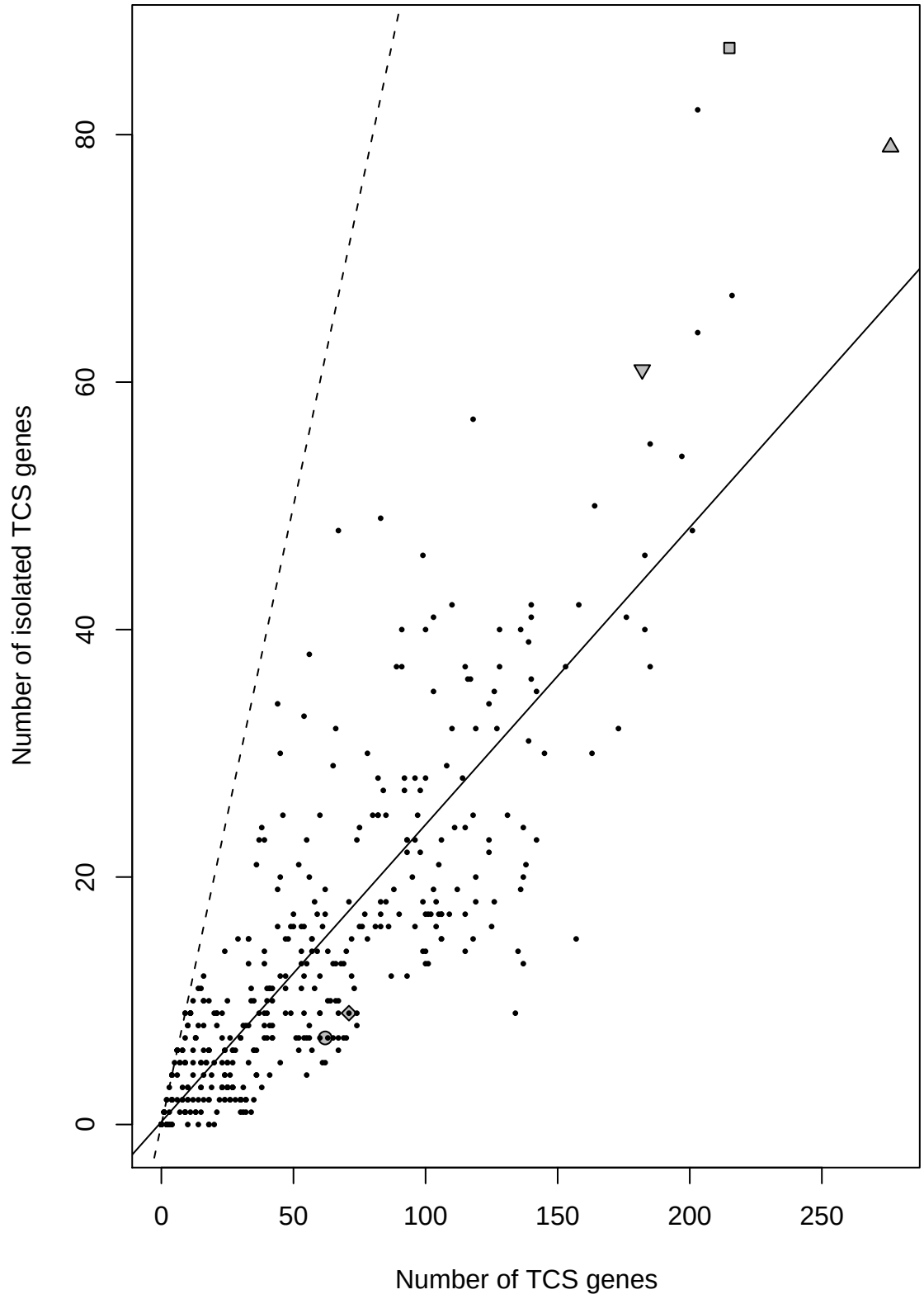


Figure 2.7: A scatter plot showing the number of TCS genes vs. the number of isolated TCS genes for the 457 species listed in Table A.1. The dotted line has gradient one, and thus any point on or close to this line has all or more of its TCS genes isolated from each other. The solid line is a simple trend line. Symbol key as per Figure 2.1. See also Figures 2.6 and 2.8.

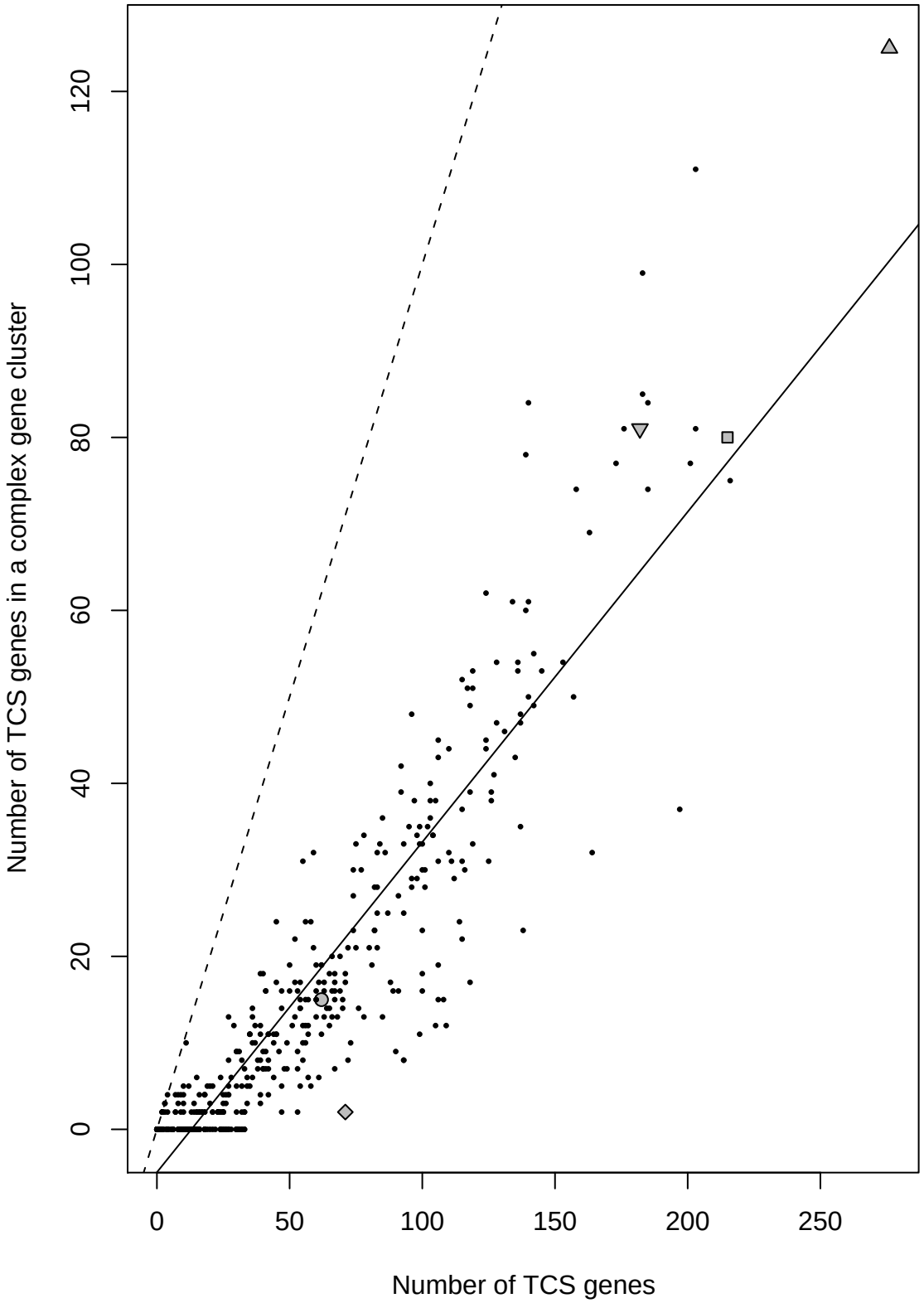


Figure 2.8: A scatter plot showing the number of TCS genes vs. the number of TCS genes in a complex gene cluster (not paired or isolated) for the 457 species listed in Table A.1. The dotted line has gradient one, and thus any point on or close to this line has all or more of its TCS genes in complex gene clusters. The solid line is a simple trend line. Symbol key as per Figure 2.1. See also Figures 2.6 and 2.7.

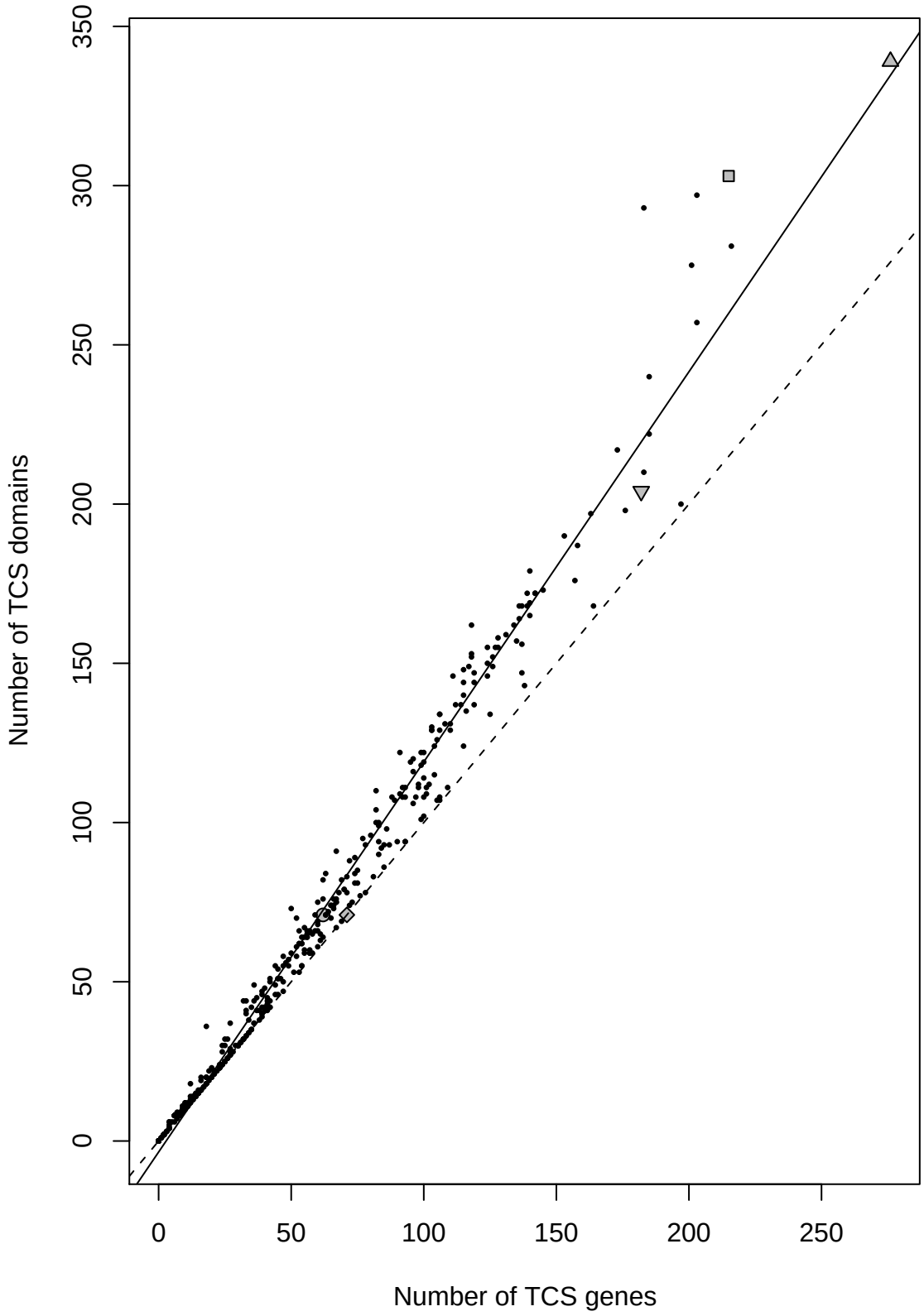


Figure 2.9: A scatter plot showing the number of TCS domains vs. the number of TCS genes for the 457 species listed in Table A.1. The dotted line has gradient one, and shows the expected number of domains if each TCS gene contained only one TCS domain. The solid line is a simple trend line. Symbol key as per Figure 2.1.

to first compile a minimal RPS-BLAST database containing only the relatively few domains of interest. This would require knowing *a priori* which domains would be required, but running RPS-BLAST would then require far less RAM (with only tens of motifs to hold in memory) and correspondingly less run time (fewer models to search).

2.7.2 Updates to PFAM

This survey was conducted using a set of PFAM models from *circa* 2004. Since then these protein models have been updated, and now give slightly different results. For instance, the current Hpt domain pfam01627 model appears to be more sensitive, and detects domains that previously had been ignored. In addition the PFAM database now includes three additional models for HisKA domains, HisKA_2 (pfam07568) and HisKA_3 (pfam07730) based on Grebe and Stock (1999) and HWE_HK (pfam07536) based on Karniol and Vierstra (2004). Using these new HisKA models may reduce the apparent surplus of receivers seen in Figure 2.1.

Also, it may now be possible to have an equally efficient screen based purely on the latest PFAM domains, without using the CDD or its subsidiary motif collections. RPS-BLAST was used for both PFAM and CDD domains in order to keep the analysis pipeline simple, although the PFAM database uses the tool HMMER (Eddy, 1998) internally. If only PFAM domains were to be used, switching from RPS-BLAST to HMMER should be considered.

2.8 Conclusion

The results of this survey demonstrate just how common TCS genes are in prokaryotes, and that over 80% of these genes are simple HKs and RRs (with only a T_i and an R TCS domain each). Furthermore, about half of these simple genes are found as neighbours, confirming that the simplistic TCS scheme introduced in Figure 1.1 on page 3 really is typical, and that the more exotic examples discussed in Section 1.4 are the exceptions rather than the norm.

In automatically identifying HK and RR gene pairs (Section 2.4), it was necessary to determine how far apart neighbouring TCS genes were, or to what extent they overlapped. Plotting a histogram of the gene overlap/separation presented several interesting features worthy of explanation (Chapter 3, Cock and Whitworth (2007a)).

The rest of this thesis will concentrate on the T_i with R pairings found here as either T_i + R or R + T_i gene pairs, or as simple hybrids, T_i-R or R-T_i. Chapter 4 will explore factors affecting the relative numbers of simple pairs and simple hybrids. The remainder of the thesis will explore these T_i/R pairs as a training dataset to predict the partnerships from the amino acid sequences.

Chapter 3

Phase preference in gene overlaps

3.1 Introduction

Overlapping genes can be found in all domains of life, but are particularly common in viruses and prokaryotes (Normark *et al.*, 1983; Rogozin *et al.*, 2002; Makalowska *et al.*, 2005). Any gene overlap allows the same number of genes to be encoded in a smaller genome, potentially advantageous to organisms under selective pressure to minimize the size of their genome, and enables mechanisms of co-regulation to operate, including translational coupling (Normark *et al.*, 1983; Oppenheim and Yanofsky, 1980; McCarthy, 1990).

When discussing two adjacent or overlapping genes, there are three distinct geometries to consider. Using the terminology of Rogozin *et al.* (2002), these are convergent (tail-to-tail, $\rightarrow\leftarrow$), divergent (head-to-head, $\leftarrow\rightarrow$), and unidirectional (head-to-tail, or tandem, $\rightarrow\rightarrow$). In the first two cases the genes are on opposite strands of the DNA, and the labels "gene one" and "gene two" are interchangeable. However, for unidirectional overlaps this symmetry is broken, with an upstream gene ("gene one") and a downstream gene ("gene two") on the same strand.

In general, there are three separate phases to consider, which, based on the gene *separation* length modulo three, Kingsford *et al.* (2007) labeled as phases $+0$, $+1$ and $+2$. *i.e.* separations of $3i$, $3i+1$ and $3i+2$ (for $i \in \mathbb{Z}$)[1]. By including the stop codons when calculating separation or overlap lengths, an overlap of length n corresponds to a separation of length $-n$, which means the three phases $+0$, $+1$ and $+2$ can also be described in terms of overlaps of $3i$, $3i-1$ and $3i-2$ (for $i \in \mathbb{Z}$).

For unidirectional gene overlaps, both genes are encoded on the same DNA strand, and the three phases can be considered as relative reading frames. Only phases $+1$ and $+2$ will be

[1]The set of integers, $\{\ldots,-3,-2,-1,0,1,2,3,\ldots\}$, is denoted by $\mathbb{Z}$. Similarly $\mathbb{N}$ will denote the set of natural numbers, $\{1,2,3,\ldots\}$.

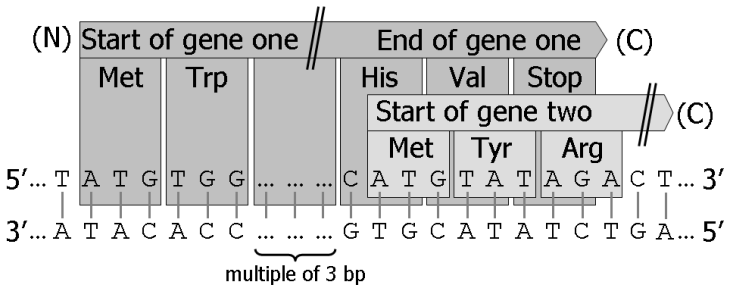


Figure 3.1: An example unidirectional gene overlap in phase $+1$, where the overlap length takes the form $3i - 1$ for $i \in \mathbb{N}$, in this case 8 bp. Note that the reading frame of gene two is 1 bp advanced from that of gene one.

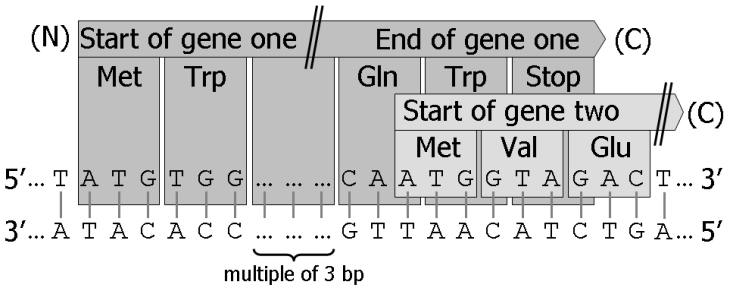


Figure 3.2: An example unidirectional gene overlap in phase $+2$, where the overlap length takes the form $3i - 2$ for $i \in \mathbb{N}$, in this case 7 bp. Note that the reading frame of gene two is 2 bp advanced from that of gene one.

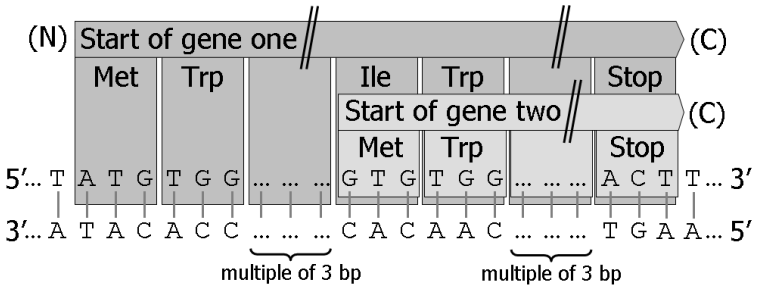


Figure 3.3: An example unidirectional gene overlap in phase $+0$, where the overlap is a multiple of three bp (i.e. overlap length $3i$ for $i \in \mathbb{N}$). The two genes' reading frames are in-phase, thus they both terminate at the same stop codon. Such overlaps can also be considered as one gene with alternative initiation sites giving two possible protein products with a common C-terminal region, and are excluded from the analysis.

considered, for example as illustrated in Figures 3.1 and 3.2 for overlaps of 8 and 7 nucleotide base pairs (bp). Phase $+0$ (or in-phase) unidirectional overlaps are a special case as both genes share the same reading frame. As a consequence, they must share the same stop codon, and therefore this is equivalent to alternative initiation sites for a single gene (Figure 3.3). Such cases are rare, and have been excluded from the following analysis.

Short overlaps of less than six bp are extremely restricted due to the amino acid genetic code (Tables 3.1 and 3.2), as such cases require the nucleotides in the overlap to be dual coding for a start/stop codon in two different genes. In the case of divergent genes, the two start codons would overlap. Similarly for convergent genes, the two stop codons would overlap. Unidirectionally overlapping genes are on the same strand, and here the stop codon of the upstream gene overlaps with the start codon of the downstream gene. In many cases, overlaps of particular lengths are rendered impossible by the genetic code. For overlaps of six or more nucleotides, the start/stop codons merely share nucleotides with ordinary amino acid coding codons in the other gene. These start/stop codon restrictions lead to very different distributions for short overlaps (less than 6 bp), compared to longer overlaps (6 or more bp).

This chapter starts with a motivational observation about unidirectional gene overlaps in TCS systems (Cock and Whitworth, 2007a), then a survey of all gene overlaps from the sequenced prokaryotes. Existing work looking at gene overlaps is summarized, and the remainder of the chapter focuses on explaining unidirectional gene overlaps.

3.2 Observed separations or overlaps from TCS genes

Figure 3.4 shows a bar chart of gene separation/overlap for unidirectional gene pairs. As in Cock and Whitworth (2007a), this is restricted to neighbouring HK and RR genes (in either order), where the HK has a single transmitter (T_i or T_{ii}) and the RR a single receiver domain (see Section 2.4). Using TCS partners allows us to be confident that each overlap is between biologically linked genes, which we would expect to be co-expressed.

In addition to peaks at overlaps of lengths one and four, there is a clear phase bias in the gene overlaps which is emphasized by the colour scheme. Related Figures 3.5 and 3.6 show this same dataset divided according to the order of the two genes, RR then HK and HK then RR respectively. Both figures show very similar distributions.

Extending the analysis to all neighbouring gene pairs shows a smoother version of the same pattern (Figure 3.13, described in following section), making this a global phenomenon, and not some quirk of TCS operons. The remainder of this chapter will focus on this general case, rather than on only the TCS gene pairs.

TTT	F	Phe		TCT	S	Ser	TAT	Y	Tyr	TGT	C	Cys
TTC	F	Phe		TCC	S	Ser	TAC	Y	Tyr	TGC	C	Cys
TTA	L	Leu		TCA	S	Ser	TAA	*	Stop	TGA	*	Stop
TTG	L	Leu	(i)	TCG	S	Ser	TAG	*	Stop	TGG	W	Trp
CTT	L	Leu		CCT	P	Pro	CAT	H	His	CGT	R	Arg
CTC	L	Leu		CCC	P	Pro	CAC	H	His	CGC	R	Arg
CTA	L	Leu		CCA	P	Pro	CAA	Q	Gln	CGA	R	Arg
CTG	L	Leu	(i)	CCG	P	Pro	CAG	Q	Gln	CGG	R	Arg
ATT	I	Ile	(i)	ACT	T	Thr	AAT	N	Asn	AGT	S	Ser
ATC	I	Ile	(i)	ACC	T	Thr	AAC	N	Asn	AGC	S	Ser
ATA	I	Ile	(i)	ACA	T	Thr	AAA	K	Lys	AGA	R	Arg
ATG	M	Met	(i)	ACG	T	Thr	AAG	K	Lys	AGG	R	Arg
GTT	V	Val		GCT	A	Ala	GAT	D	Asp	GGT	G	Gly
GTC	V	Val		GCC	A	Ala	GAC	D	Asp	GGC	G	Gly
GTA	V	Val		GCA	A	Ala	GAA	E	Glu	GGA	G	Gly
GTG	V	Val	(i)	GCG	A	Ala	GAG	E	Glu	GGG	G	Gly

Table 3.1: This table shows what the NCBI refers to as Translation Table 11, the genetic code used for bacteria, archaea, prokaryotic viruses and chloroplast proteins. Recognised initiation start codons are marked with (i).

TTT	F	Phe		TCT	S	Ser	TAT	Y	Tyr	TGT	C	Cys
TTC	F	Phe		TCC	S	Ser	TAC	Y	Tyr	TGC	C	Cys
TTA	L	Leu	(i)	TCA	S	Ser	TAA	*	Stop	TGA	W	Trp
TTG	L	Leu	(i)	TCG	S	Ser	TAG	*	Stop	TGG	W	Trp
CTT	L	Leu		CCT	P	Pro	CAT	H	His	CGT	R	Arg
CTC	L	Leu		CCC	P	Pro	CAC	H	His	CGC	R	Arg
CTA	L	Leu		CCA	P	Pro	CAA	Q	Gln	CGA	R	Arg
CTG	L	Leu	(i)	CCG	P	Pro	CAG	Q	Gln	CGG	R	Arg
ATT	I	Ile	(i)	ACT	T	Thr	AAT	N	Asn	AGT	S	Ser
ATC	I	Ile	(i)	ACC	T	Thr	AAC	N	Asn	AGC	S	Ser
ATA	I	Ile	(i)	ACA	T	Thr	AAA	K	Lys	AGA	R	Arg
ATG	M	Met	(i)	ACG	T	Thr	AAG	K	Lys	AGG	R	Arg
GTT	V	Val		GCT	A	Ala	GAT	D	Asp	GGT	G	Gly
GTC	V	Val		GCC	A	Ala	GAC	D	Asp	GGC	G	Gly
GTA	V	Val		GCA	A	Ala	GAA	E	Glu	GGA	G	Gly
GTG	V	Val	(i)	GCG	A	Ala	GAG	E	Glu	GGG	G	Gly

Table 3.2: This table shows what the NCBI refers to as Translation Table 4, the genetic code used for mould, protozoan, coelenterate mitochondria and mycoplasma/spiroplasma. Recognised initiation start codons are marked with (i). *cf.* Table 3.1 where TGA is a stop codon, and TTA is not recognised as an initiation start codon.

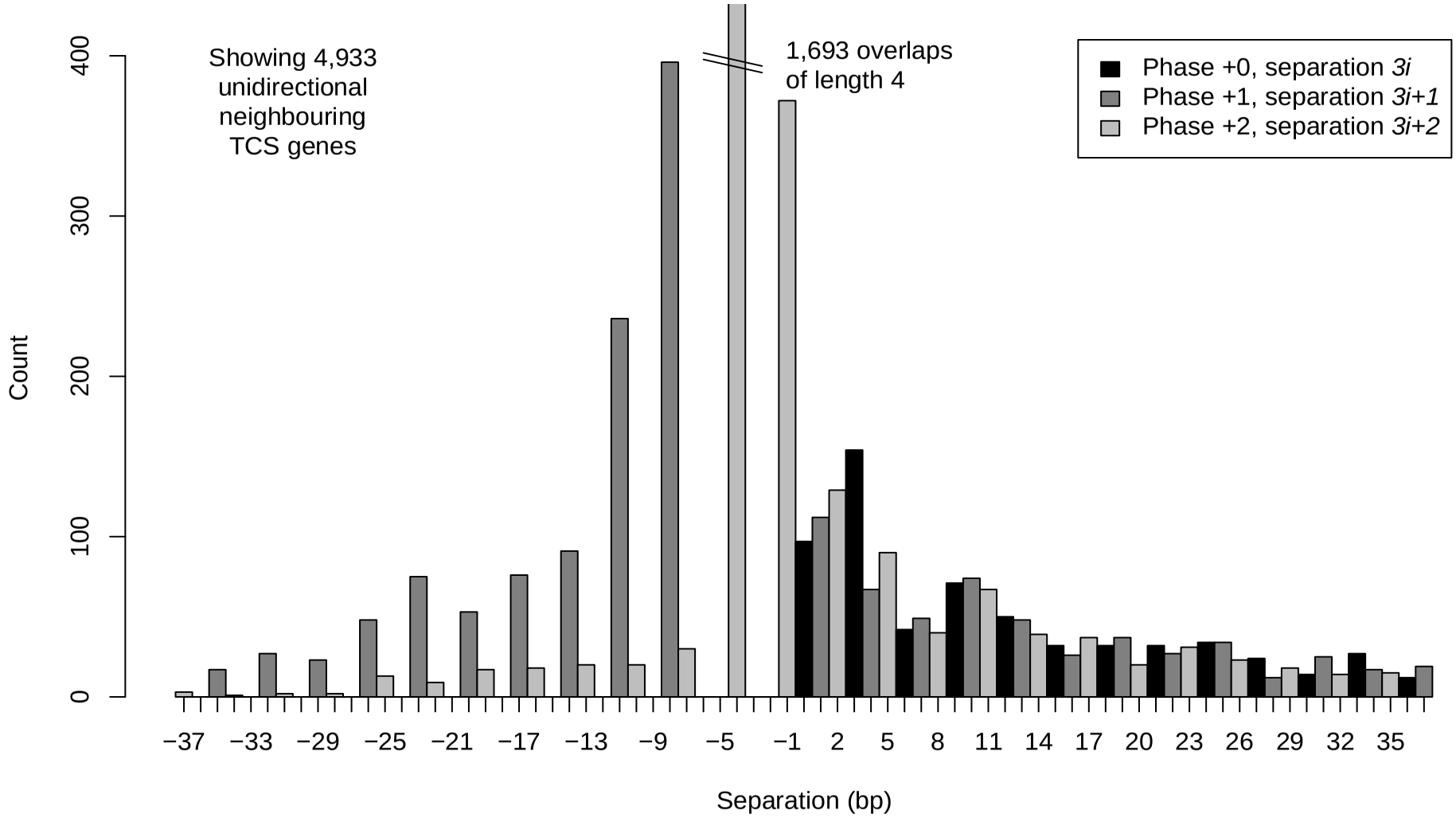


Figure 3.4: Unidirectional (→→) TCS gene pair separation or overlap, from the 457 species listed in Table A.1. Unlike Figure 3.13, this only shows simple HK and RR neighbours (in either order). Figures 3.5 and 3.6 show this dataset divided according to the order of the two genes.

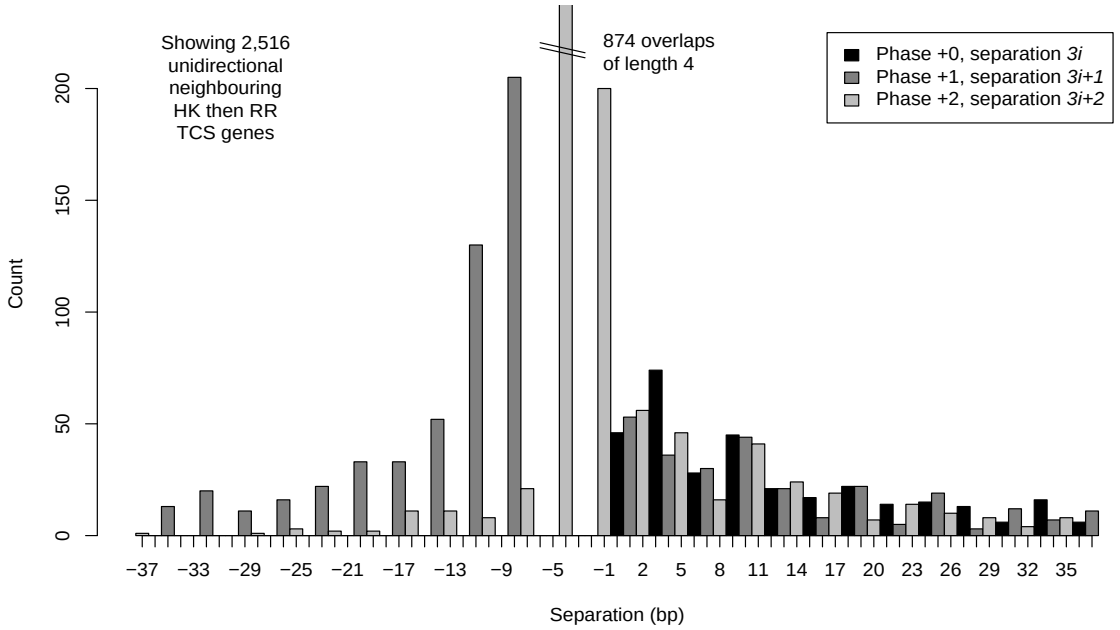


Figure 3.5: Unidirectional (→→) HK then RR gene pair separation or overlap (in that order, i.e. T + R pairs), from the 457 species listed in Table A.1.

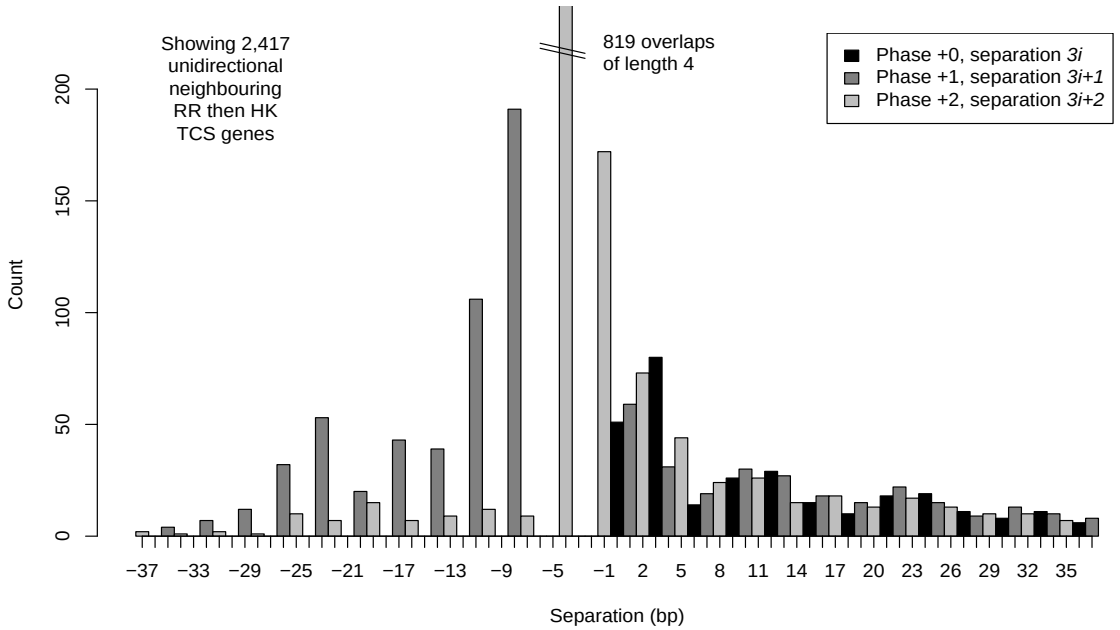


Figure 3.6: Unidirectional (→→) RR then HK gene pair separation or overlap (in that order, i.e. R + T pairs), from the 457 species listed in Table A.1.

3.3 Observed separations or overlaps from all genes

Separation/overlap frequencies were tabulated for divergent, convergent and unidirectional adjacent genes based on the annotation in the GenBank files of the 457 species listed in Table A.1. For simplicity, any genes with non-exact locations, ambiguous sequences, internal stop codons, invalid start or stop codons (as verified using the declared genetic code, Table 3.1 or 3.2), or special cases with non-continuous coding sequences (e.g. from ribosomal slippages) were excluded, as were cases where one gene was entirely within another (fully overlapped). Separated gene pairs with any ambiguous sequence between them were also excluded.

In total $1,428,039$ gene pairs were considered: $210,209$ divergent, $210,494$ convergent and $1,007,336$ unidirectional gene pairs; a split of 14.7%, 14.7% and 70.5% respectively (3 sf). Within each orientation, the proportion of overlapping gene pairs varies considerably. Only 3.6% of divergent genes are annotated as overlapping, compared with 13% of convergent pairs and 21% (2 sf) of unidirectional pairs. These ratios agree within one percent with published results (Fukuda *et al.*, 2003; Kingsford *et al.*, 2007).

In the following bar charts, for all three orientations there is a clear division between short overlaps ($n < 6$), and longer overlaps ($n \geq 6$) where there are periodic patterns in the distributions, which have been emphasized by using three alternating colours for the three phases.

3.3.1 Divergent gene pairs

Figure 3.7 illustrates divergent gene pairs, which make up about 15% of neigbouring genes. Of these, only 3.6% or about $7,600$ pairs (2 sf) are overlapped. Divergent overlaps require any promoter or translation initiation sites to be dual-encoded in the complementary strand of the other gene's coding region - which may go some way to explaining their rarity.

This bar chart can be divided into three regions which exhibit different behaviours: separated genes, short overlaps, and long overlaps. There is a periodic behaviour in the separated genes, with phase $+2$ apparently most common. Also noteworthy is a drop in the observed counts around six bp, which could be explained by requirements of the ribosomal binding site having to be dual coded with the other gene's start codon.

Short divergent overlaps ($n < 6$) are restricted by the limited set of possible start codons. Table 3.3 shows these overlaps tabulated according to the start codons used. In cases where a pair of start codons cannot give a particular overlap length, a dash has been shown. Otherwise the observed count is given, which is zero in some cases. Because divergent overlaps are symmetric with respect to the strands, these tables are symmetric.

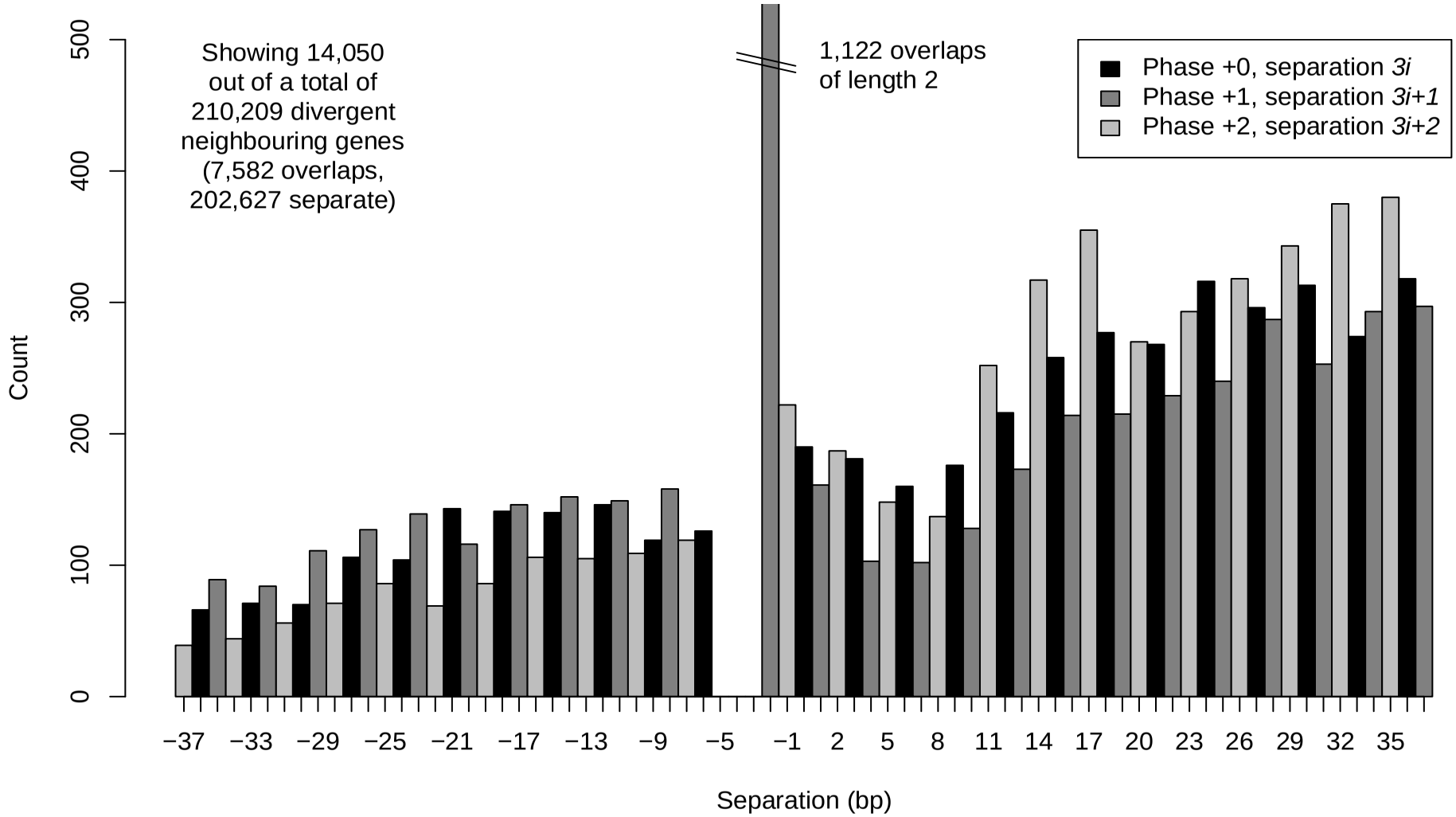


Figure 3.7: Divergent (←→) gene separation or overlap, from all unambiguously annotated genes from the 457 species listed in Table A.1.

$n = 1$

			gene two start codon							
			ATA	**A**TC	**A**TG	**A**TT	**C**TG	**G**TG	**T**TA	**T**TG
	ATA	(TA**T**)	-	-	-	-	-	-	0	0
	ATC	(GA**T**)	-	-	-	-	-	-	0	0
gene one	ATG	(CA**T**)	-	-	-	-	-	-	0	219
start codon	ATT	(AA**T**)	-	-	-	-	-	-	0	0
(and reverse	CTG	(CA**G**)	-	-	-	-	-	3	-	-
complement)	GTG	(CA**C**)	-	-	-	-	3	-	-	-
	TTA	(TA**A**)	0	0	0	0	-	-	-	-
	TTG	(CA**A**)	0	0	219	0	-	-	-	-

222 overlaps

$n = 2$

			gene two start codon							
			ATA	**AT**C	**AT**G	**AT**T	**CT**G	**GT**G	**TT**A	**TT**G
	ATA	(T**AT**)	0	1	1	0	-	-	-	-
	ATC	(G**AT**)	1	0	0	0	-	-	-	-
gene one	ATG	(C**AT**)	1	0	1120	0	-	-	-	-
start codon	ATT	(A**AT**)	0	0	0	0	-	-	-	-
(and reverse	CTG	(C**AG**)	-	-	-	-	-	-	-	-
complement)	GTG	(C**AC**)	-	-	-	-	-	-	-	-
	TTA	(T**AA**)	-	-	-	-	-	-	-	-
	TTG	(C**AA**)	-	-	-	-	-	-	-	-

1122 overlaps

$n = 4$

			gene two start codon							
			A**TA**	A**TC**	A**TG**	A**TT**	C**TG**	G**TG**	T**TA**	T**TG**
	ATA	(**TA**T)	0	-	-	-	-	-	0	-
	ATC	(**GA**T)	-	-	-	-	-	-	-	-
gene one	ATG	(**CA**T)	-	-	-	-	-	-	-	-
start codon	ATT	(**AA**T)	-	-	-	-	-	-	-	-
(and reverse	CTG	(**CA**G)	-	-	-	-	-	-	-	-
complement)	GTG	(**CA**C)	-	-	-	-	-	-	-	-
	TTA	(**TA**A)	0	-	-	-	-	-	0	-
	TTG	(**CA**A)	-	-	-	-	-	-	-	-

0 overlaps

$n = 5$

			gene two start codon							
			AT**A**	AT**C**	AT**G**	AT**T**	CT**G**	GT**G**	TT**A**	TT**G**
	ATA	(**T**AT)	-	-	-	0	-	-	-	-
	ATC	(**G**AT)	-	-	0	-	0	0	-	0
gene one	ATG	(**C**AT)	-	0	-	-	-	-	-	-
start codon	ATT	(**A**AT)	0	-	-	-	-	-	0	-
(and reverse	CTG	(**C**AG)	-	0	-	-	-	-	-	-
complement)	GTG	(**C**AC)	-	0	-	-	-	-	-	-
	TTA	(**T**AA)	-	-	-	0	-	-	-	-
	TTG	(**C**AA)	-	0	-	-	-	-	-	-

0 overlaps

Table 3.3: Divergent overlap sequences of varying lengths, $n < 6$, tabulated by start codon. Bold indicates which parts of the two start codons would coincide, with dashes for impossible combinations. Note that these tables are symmetric. Same dataset as Figure 3.7.

$n \geq 6$, phase $+0$

			gene two start codon							
			ATA	ATC	ATG	ATT	CTG	GTG	TTA	TTG
	ATA	(TAT)	0	0	2	0	0	1	0	0
	ATC	(GAT)	0	0	0	0	0	0	0	0
gene one	ATG	(CAT)	2	0	641	0	8	651	0	598
start codon	ATT	(AAT)	0	0	0	0	1	0	0	0
(and reverse	CTG	(CAG)	0	0	8	1	1	1	0	1
complement)	GTG	(CAC)	1	0	651	0	1	152	0	197
	TTA	(TAA)	0	0	0	0	0	0	0	0
	TTG	(CAA)	0	0	598	0	1	197	0	103

2357 overlaps

$n \geq 6$, phase $+1$

			gene two start codon							
			ATA	ATC	ATG	ATT	CTG	GTG	TTA	TTG
	ATA	(TAT)	0	0	0	0	0	0	0	0
	ATC	(GAT)	0	0	0	0	0	0	0	0
gene one	ATG	(CAT)	0	0	838	0	4	627	0	610
start codon	ATT	(AAT)	0	0	0	0	0	0	0	1
(and reverse	CTG	(CAG)	0	0	4	0	0	2	0	1
complement)	GTG	(CAC)	0	0	627	0	2	170	0	198
	TTA	(TAA)	0	0	0	0	0	0	0	0
	TTG	(CAA)	0	0	610	1	1	198	0	124

2575 overlaps

$n \geq 6$, phase $+2$

			gene two start codon							
			ATA	ATC	ATG	ATT	CTG	GTG	TTA	TTG
	ATA	(TAT)	1	0	0	0	0	0	0	0
	ATC	(GAT)	0	0	0	0	0	0	0	0
gene one	ATG	(CAT)	0	0	275	1	2	381	0	300
start codon	ATT	(AAT)	0	0	1	0	1	0	0	0
(and reverse	CTG	(CAG)	0	0	2	1	1	0	0	1
complement)	GTG	(CAC)	0	0	381	0	0	105	0	143
	TTA	(TAA)	0	0	0	0	0	0	2	0
	TTG	(CAA)	0	0	300	0	1	143	0	93

1306 overlaps

Table 3.4: Divergent overlap nucleotide sequences of length $n \geq 6$, tabulated by phase according to the start codons used. This is a continuation of Table 3.3.

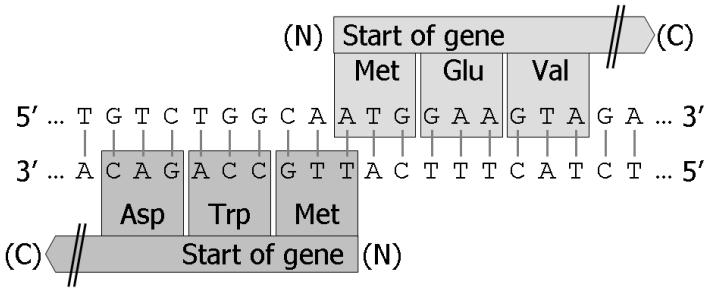


Figure 3.8: An example divergent gene overlap of length $n = 1$, ...CA**A**TG... (or ...CA**T**TG... on the reverse complement strand) where the two genes commence **A**TG... and **T**TG... (using the standard start codon ATG, and an atypical start codon, TTG).

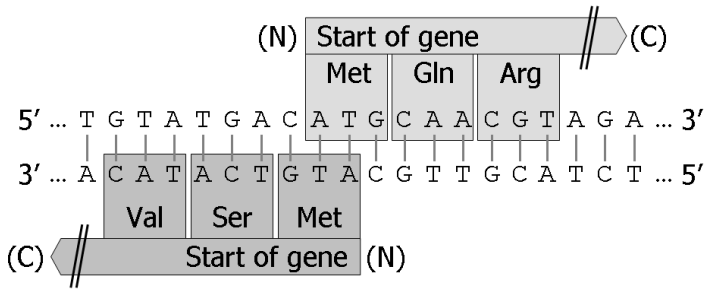


Figure 3.9: An example divergent gene overlap of length $n = 2$, ...C**AT**G... (which is a palindromic sequence) where both genes commence **AT**G... using the standard start codon, ATG.

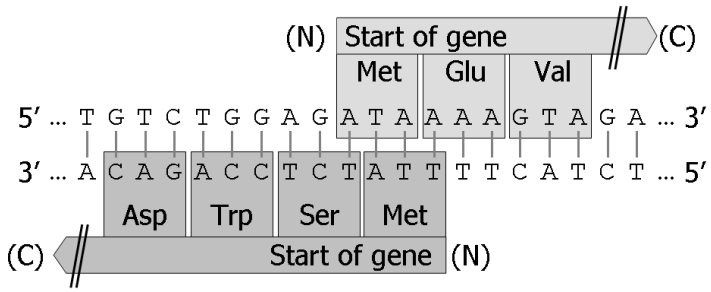


Figure 3.10: An example divergent gene overlap of length $n = 4$, ...ATAA... (or ...TTAT... on the reverse complement strand) where the two genes commence ATAA... and TTAT... (using atypical start codons ATA and TTA).

There are over two hundred divergent gene overlaps of length $n = 1$, almost exclusively ...CA**A**TG... or equivalently ...CA**T**TG... (its reverse complement), which was observed 219 times. In this particular overlap the two genes commence with start codons **A**TG... and **T**TG..., as illustrated in Figure 3.8. The $n = 1$ overlap ...CA**C**TG... or ...CA**G**TG... was found three times, while the remaining theoretical possibilities were not observed (Table 3.3).

Over a thousand divergent overlaps of length $n = 2$ were found, almost exclusively where both genes commence with the standard start codon **ATG**... giving an overlap region ...C**ATG**... (a palindromic sequence, 1120 cases). This is illustrated in Figure 3.9. Table 3.3 shows several other possible overlaps of length $n = 2$ using different start codons pairings, which were observed only once or not at all.

A divergent overlap of length $n = 3$ would require a start codon whose reverse complement is also a start codon, and this is not possible in the known genetic codes. Divergent overlaps of lengths $n = 4$ and 5 bp are possible using atypical start codons (e.g. Figure 3.10), but were not observed (Table 3.3)

For longer divergent overlaps ($n \geq 6$), phase $+2$ is clearly least common, while the number of overlaps in phases $+0$ and $+1$ are similar, with slightly more in phase $+1$ (Figure 3.7, Table 3.4). Note that while long overlaps in phase $+2$ are least common, separations in phase $+2$ are most common. This symmetry may be due in part to misannotated start codons.

3.3.2 Convergent gene pairs

Around 15% of all neighbouring genes are convergent, and of these 13% (2 sf) are overlapping. Figure 3.11 shows separations/overlap lengths in convergent gene pairs.

Short convergent overlaps ($n < 6$) are restricted by the limited set of three possible stop codons. Table 3.5 shows these overlaps tabulated according to the stop codons used. In cases where a pair of stop codons cannot give a particular overlap length, a dash has been shown. Since divergent overlaps are symmetric with respect to the strands, these tables are symmetric.

Figure 3.11 shows about a third of the convergent overlaps are of length four, a distribution spike previously reported (Fukuda *et al.*, 1999). In fact, overlaps of length $n = 4$ are the only possible short convergent overlaps ($n < 6$), as overlaps of lengths 1, 2, 3 and 5 bp cannot appear due to stop codon limitations (Table 3.5).

Convergent overlaps of length $n = 4$ can occur in three ways (Table 3.5), with both genes using stop codon TAA (2359 cases), both using TAG (2729 cases), or one using TAA and the other using TAG (3329 cases, illustrated in Figure 3.12). Convergent overlaps of length

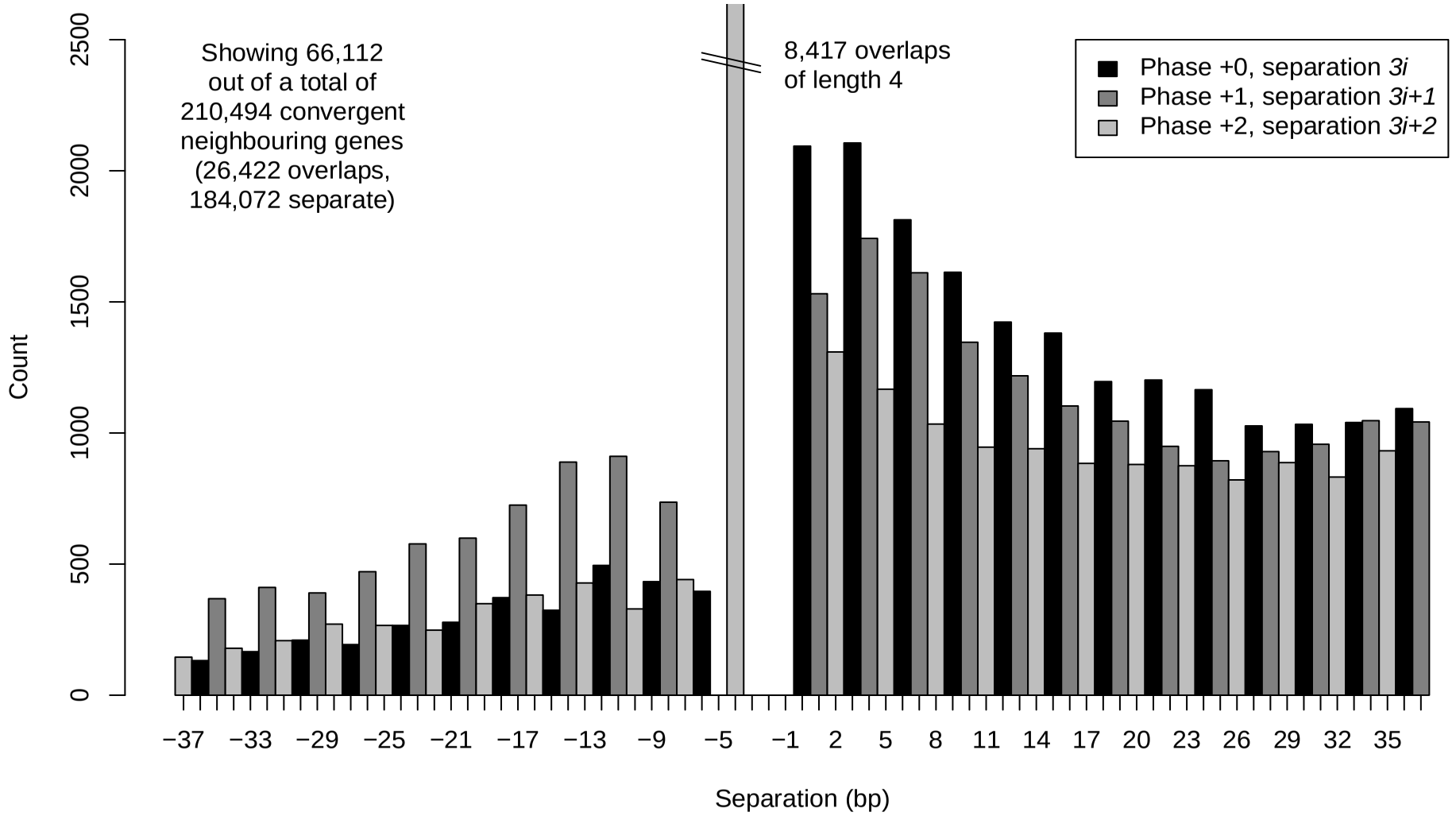


Figure 3.11: Convergent (→←) gene separation or overlap, from all unambiguously annotated genes from the 457 species listed in Table A.1. *cf.* Kingsford *et al.* (2007) Figure 1.

$n = 1$		*gene two stop (reverse complement)*		
		TAA (**T**TA)	TAG (**C**TA)	TGA (**T**CA)
gene	TA**A**	-	-	-
one	TA**G**	-	-	-
stop	TG**A**	-	-	-
				0 overlaps

$n = 2$		*gene two stop (reverse complement)*		
		TAA (**TT**A)	TAG (**CT**A)	TGA (**TC**A)
gene	T**AA**	-	-	-
one	T**AG**	-	-	-
stop	T**GA**	-	-	-
				0 overlaps

$n = 4$		*gene two stop (reverse complement)*		
		TAA (T**TA**)	TAG (C**TA**)	TGA (T**CA**)
gene	**TA**A	2359	3329	-
one	**TA**G	3329	2729	-
stop	**TG**A	-	-	-
				8417 overlaps

$n = 5$		*gene two stop (reverse complement)*		
		TAA (TT**A**)	TAG (CT**A**)	TGA (TC**A**)
gene	**T**AA	-	-	-
one	**T**AG	-	-	-
stop	**T**GA	-	-	-
				0 overlaps

$n \geq 6$, phase $+0$		*gene two stop (reverse complement)*		
		TAA (TTA)	TAG (CTA)	TGA (TCA)
gene	TAA	892	713	1020
one	TAG	713	460	912
stop	TGA	1020	912	753
				4750 overlaps

$n \geq 6$, phase $+1$		*gene two stop (reverse complement)*		
		TAA (TTA)	TAG (CTA)	TGA (TCA)
gene	TAA	1051	1203	1508
one	TAG	1203	372	1519
stop	TGA	1508	1519	2753
				8406 overlaps

$n \geq 6$, phase $+2$		*gene two stop (reverse complement)*		
		TAA (TTA)	TAG (CTA)	TGA (TCA)
gene	TAA	217	307	1065
one	TAG	307	251	1614
stop	TGA	1065	1614	1395
				4849 overlaps

Table 3.5: Convergent overlap sequences of varying lengths, n, tabulated by stop codons. Bold indicates which parts of the two stop codons would coincide, with dashes for impossible combinations. Note that these tables are symmetric. From the same dataset as Figure 3.11.

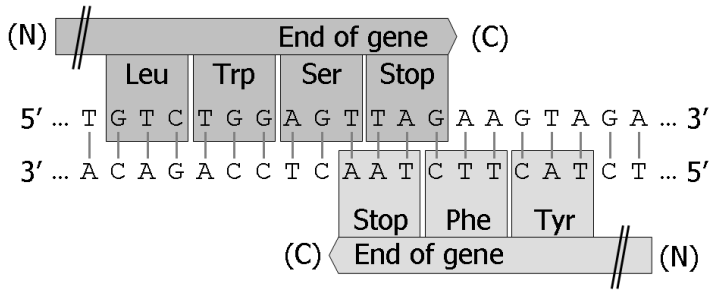


Figure 3.12: An example convergent gene overlap of length $n = 4$, `...TTAG...` (or `...CTAA...` on the reverse complement strand) where the two genes end `...TTAG` and `...CTAA` using stop codons `TAG` and `TAA` respectively.

$n = 4$ using the stop codon `TGA` are impossible. In this situation, since one gene would finish `...NTGA`, giving an overlap of `NTGA` with reverse complement `TCAN`, the other gene would end `...TCAN` requiring a stop codon of the form `CAN`.

For the longer convergent overlaps ($n \geq 6$), phase $+1$ overlaps are clearly the most common (Figure 3.11, Table 3.5), as reported in Rogozin *et al.* (2002) (where phase $+1$ is referred to as C2) and Kingsford *et al.* (2007). On the other hand, for separate convergent pairs, phase $+0$ is slightly more common than phase $+1$, than phase $+2$, also reported in Kingsford *et al.* (2007). However, the overlap and separated gene spectra do not appear to be mirror images of each other, as claimed in Kingsford *et al.* (2007).

3.3.3 Unidirectional gene pairs

The most common gene pair orientation is unidirectional, with over 70% of cases. This orientation also has the highest overlap rate (21%, 2 sf).

For unidirectional overlaps, the three phases can be viewed in terms of relative reading frames. As mentioned in the introduction, the special case of "in phase" overlaps (of length $3i$ for $i \in \mathbb{N}$, phase $+0$) reduces to alternative start codons for a single gene which were excluded (Figure 3.3). This leaves two possible overlap phases, phase $+1$ (overlaps of length $3i - 1$, Figure 3.1) and phase $+2$ (overlaps of length $3i - 2$ for $i \in \mathbb{N}$, Figure 3.2).

Figure 3.13 shows unidirectional gene pairs. Overlaps of length $n = 1$ make up about one sixth of unidirectional overlaps, while overlaps of length $n = 4$ constitute just under half. The high number of length 1 and 4 bp overlaps has long been recognised (Eyre-Walker, 1996). Overlaps of 2, 3 and 6 bp are prevented by the genetic code, however there are a handful of overlaps of 5 bp made possible by an atypical start codon, `ATT` (Table 3.6).

The simplest possible unidirectional overlap is by one bp. For example, Figure 3.14

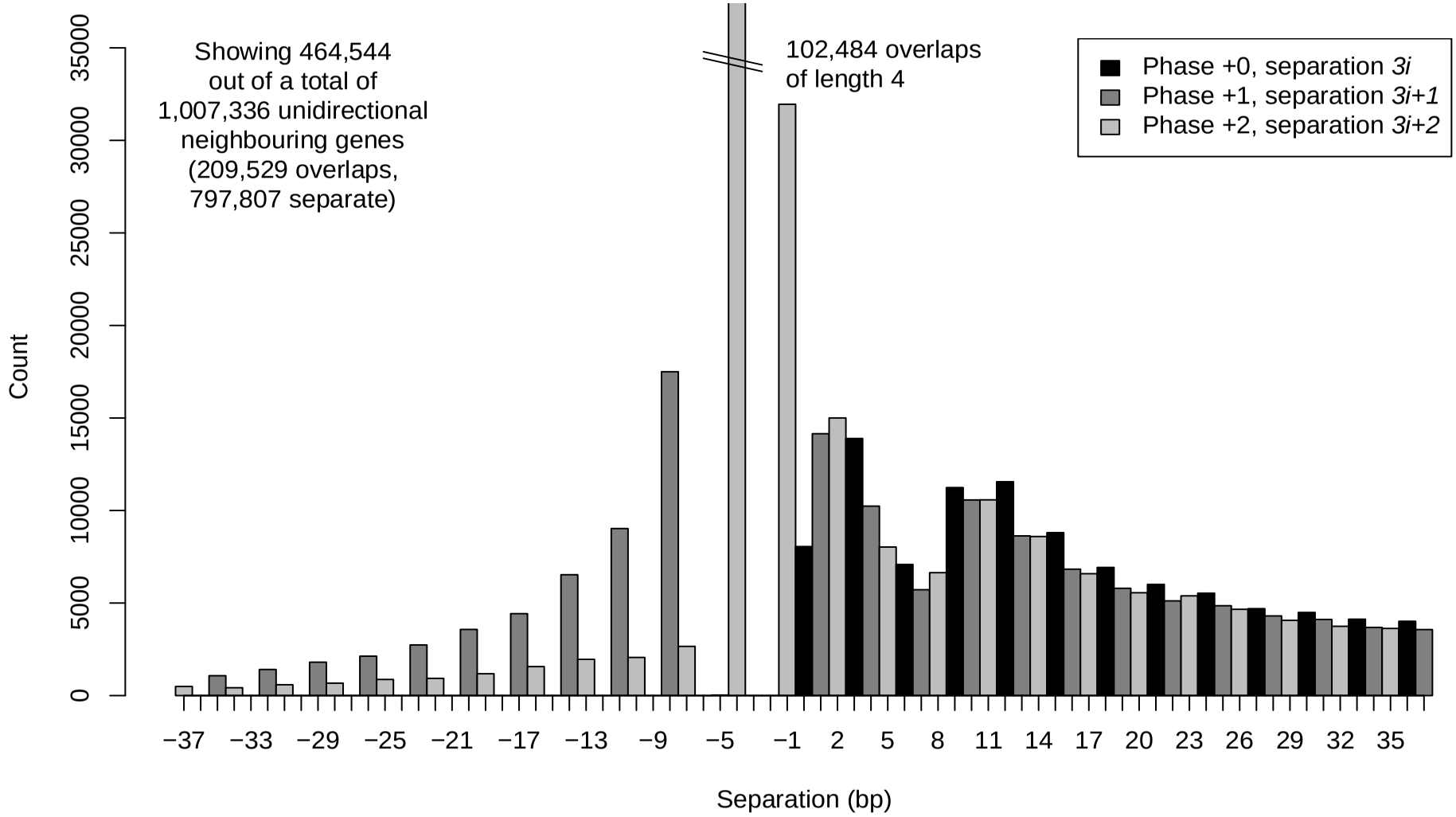


Figure 3.13: Unidirectional (→→) gene separation or overlap, from all unambiguously annotated genes from the 457 species listed in Table A.1. See also Table 3.6. *cf.* Figure 3.4

$n = 1$		*gene two start codon*							
		ATA	**A**TC	**A**TG	**A**TT	**C**TG	**G**TG	**T**TA	**T**TG
gene	TA**A**	2	2	15618	15	-	-	-	-
one	TA**G**	-	-	-	-	-	1331	-	-
stop	TG**A**	0	1	14982	1	-	-	-	-
									31952 overlaps

$n = 2$		*gene two start codon*							
		ATA	**AT**C	**AT**G	**AT**T	**CT**G	**GT**G	**TT**A	**TT**G
gene	T**AA**	-	-	-	-	-	-	-	-
one	T**AG**	-	-	-	-	-	-	-	-
stop	T**GA**	-	-	-	-	-	-	-	-
									0 overlaps

$n = 4$		*gene two start codon*							
		A**TA**	A**TC**	A**TG**	A**TT**	C**TG**	G**TG**	T**TA**	T**TG**
gene	**TA**A	73	-	-	-	-	-	14	-
one	**TA**G	5	-	-	-	-	-	4	-
stop	**TG**A	-	-	78840	-	116	18487	-	4945
									102484 overlaps

$n = 5$		*gene two start codon*							
		AT**A**	AT**C**	AT**G**	AT**T**	CT**G**	GT**G**	TT**A**	TT**G**
gene	**T**AA	-	-	-	22	-	-	-	-
one	**T**AG	-	-	-	4	-	-	-	-
stop	**T**GA	-	-	-	2	-	-	-	-
									28 overlaps

$n \geq 6$, phase $+1$		*gene two start codon*							
		ATA	ATC	ATG	ATT	CTG	GTG	TTA	TTG
gene	TAA	25	9	18514	14	47	3073	28	3073
one	TAG	2	3	5490	4	6	1166	8	1072
stop	TGA	0	3	18768	1	24	3873	0	2579
									57782 overlaps

$n \geq 6$, phase $+2$		*gene two start codon*							
		ATA	ATC	ATG	ATT	CTG	GTG	TTA	TTG
gene	TAA	11	21	3730	56	16	1003	6	826
one	TAG	1	3	2163	5	6	724	0	469
stop	TGA	4	0	4440	3	25	2611	0	1160
									17283 overlaps

Table 3.6: Unidirectional overlap nucleotide sequences of varying lengths, n, tabulated according to the upstream gene stop codon (gene one) and the downstream gene start codon (gene two). Bold nucleotides indicate which parts of the start and stop codons would have to coincide, impossible combinations are shown as a dash. Note that the combination of start codon TTA with stop codon TGA is not possible for overlaps of length $n < 6$ in either of the genetic codes used (Tables 3.1 and 3.2). From the same dataset as Figure 3.13.

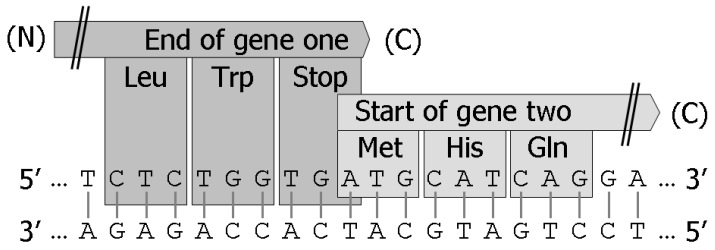


Figure 3.14: An example unidirectional gene overlap of length $n = 1$, `...TG`**`A`**`TG...` where the upstream gene one ends `...TG`**`A`** (stop codon `TGA`) and the downstream gene two starts **`A`**`TG...` (standard start codon `ATG`).

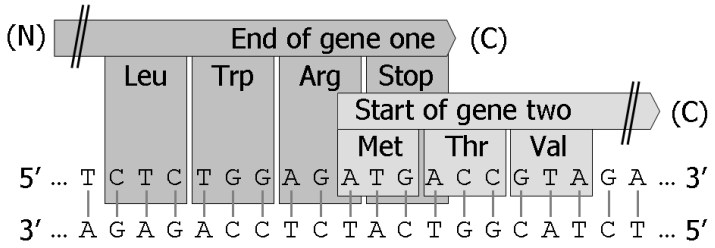


Figure 3.15: An example unidirectional gene overlap of length $n = 4$, `...ATGA...` where the upstream gene one ends `...ATGA` (stop codon `TGA`) and the downstream gene two starts `ATGA...` (standard start codon `ATG`).

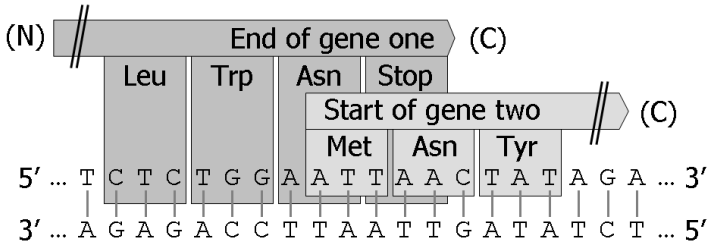


Figure 3.16: An example unidirectional gene overlap of length $n = 5$, `...ATTAA...` where the upstream gene one ends `...ATTAA` (stop codon `TAA`) and the downstream gene two starts `ATTAA...` (atypical start codon `ATT`).

shows an overlap ...TG**A**TG... where the overlapping region (A) is the last nucleotide of a stop codon in gene one (...TG**A**) and the first nucleotide of a start codon in gene two (**A**TG...). Referring to Table 3.6, overlaps of ...TG**A**TG... and ...TA**A**TG... make up the bulk of $n = 1$ overlaps, with a handful of others using different start codons.

The most numerous unidirectional overlap is length $n = 4$, principally ...ATGA... (gene one ends ...TGA and gene two starts ATG..., illustrated in Figure 3.15). Because prokaryotes employ multiple possible start codons, other overlaps of four are also possible, with GTGA and TTGA being the next most common (Table 3.6).

An overlap of five nucleotides is rare, requiring the atypical start codon ATT (Table 3.6). One example of this is shown in Figure 3.16. A similar analysis based on the codon tables excludes overlaps of two nucleotides. The in-phase cases of $n = 3$ and $n = 6$ can also be specifically ruled out as these would require a codon to serve double duty as both a stop and a start site, or reduce the downstream gene to a trivial six nucleotide start-stop sequence.

Thus for $n \leq 6$ the major overlaps are $n = 1$ and 4, plus the slight possibility of $n = 5$ with an atypical start codon. These extremely common short overlaps almost certainly cause translational coupling (Oppenheim and Yanofsky, 1980; McCarthy, 1990), resulting in both genes being expressed at a similar level. Eyre-Walker (1996) discusses the use of particular alternative stop codons in this context.

For overlaps of more than six nucleotides, it is striking that overlaps in phase $+1$ are far more common than in those in phase $+2$. Plots like this have previously been published although the phase bias for longer overlaps was not apparent due to lack of data (Eyre-Walker, 1996), or not stressed (Borodovsky *et al.*, 1999; Johnson and Chisholm, 2004), until this work (Cock and Whitworth, 2007a).

3.4 Current understanding of gene overlaps

The three different overlap orientations share certain features. For short overlaps ($n < 6$), the absence of certain overlap lengths is trivially explained by the codon table. For the longer overlaps ($n \geq 6$) there appears to be an exponential drop off in the observed counts, but with an additional effect with a periodicity equal to the codon length of three bp. There are also some visible phase effects in the separation distribution for neighbouring non-overlapping genes.

One important question to address when considering potentially overlapping genes is how they were predicted or annotated. For stop codons, there is little chance of error (barring abnormalities like ribosomal slippages), but for start codons the situation is less clear cut –

indeed multiple start codons can be used *in vitro* resulting in different gene products from the same gene (e.g. the genes *inf*B (Nyengaard *et al.*, 1991; Laursen *et al.*, 2002) or *che*A (Smith and Parkinson, 1980) in *E. coli* and other species) as illustrated in Figure 3.3.

How each genome was annotated, and any biases in start codon selection regarding potential gene overlaps will complicate this analysis. To avoid concerns about mis-annotated start codons Rogozin *et al.* (2002) restricted their analysis to evolutionarily conserved overlapping gene pairs, but this does have the downside of restriction to a much smaller dataset. In Cock and Whitworth (2007a) our analysis was restricted to TCS genes only, where because these genes contain characterized domains, they are more likely to be genuine than predicted genes of unknown function.

There has been little work in the literature looking at the phase patterns in divergent genes, presumably handicapped by both the relative scarcity of divergent gene overlaps, and uncertainty over the reliability of the start codon annotation which is doubly crucial in this head-to-head orientation.

Several groups have looked at the more common case of convergent overlaps, where being tail-to-tail any ambiguity of start codons is not so important. Rogozin *et al.* (2002) concluded the phase $+1$ preference in convergent gene overlaps could be explained because this offered the least mutual constraint on nonconservative amino acid replacements in both overlapping coding sequences (Krakauer, 2000), and overlaps in this phase were therefore more likely to be retained by positive selection.

To explain this reasoning, the three nucleotides within a codon will be referred to as $c1$, $c2$ and $c3$. In this notation $c3$ corresponds to the "wobble position" (Crick, 1966). When two convergent genes overlap in phase $+1$, the $c1$ positions of each gene coincide, while the $c2$ position of one gene matches a $c3$ position in the other, and *vice versa*. This arrangement allows non-synonymous changes in one gene by point mutation of a nucleotide in the $c2$ position, with minimal disruption to the other gene where this change is in the $c3$ or wobble position, and the change is therefore likely to be synonymous. This imposes minimal mutual constraints on the co-evolution of the two genes, compared to the other two phases. In particular, in phase $+2$, the wobble positions in both genes coincide, which means any non-synonymous change in one gene will likely be non-synonymous in the other gene.

Kingsford *et al.* (2007) later observed that due to the likelihood of finding alternative stop codons in the reverse complement of a non-overlapping gene coding sequence, the simple loss of a stop codon will produce a similar three bp periodic pattern of overlapping gene lengths. They concluded convergent gene overlaps arose by random extension of genes to the

next in-frame stop codon, followed by selection against longer overlaps, which was modeled using an exponential fitness function. This data-driven explanation is much simpler than the mutual constraint arguments of Rogozin *et al.* (2002), although as the authors observed, the two are not mutually contradictory.

Our own work focused on the phase bias in unidirectional overlaps, initially from a mutual restraint perspective (Cock and Whitworth, 2007a), similar to that of Rogozin *et al.* (2002). This chapter presents an alternative start/stop codon analysis following the method of Kingsford *et al.* (2007). While there is one start codon to worry about in unidirectional gene pairs, this orientation is by far the most common, providing a wealth of data to analyse.

3.5 Long unidirectional gene overlaps

Short overlaps of length $n < 6$ have been discussed above (Section 3.3.3), while multiples of three are excluded ($n = 3i$ for $i \in \mathbb{N}$, phase $+0$). This leaves the longer overlaps ($n \geq 6$) where there are two distinct phases to consider. Based on shared information content alone, it has been predicted there would be no phase preference for unidirectional gene overlaps (Krakauer, 2000), but this is not the case (Figure 3.4, Table 3.6).

For the overlapping region, each nucleotide is encoding two different codons - and thus controls the amino acid sequence of two different proteins. Due to the nature of the amino acid translation table, mutations in the wobble position of a codon ($c3$) generally make least difference to the resulting amino acid, and mutations in this position are most likely to be tolerated. Looking at the amino acid nature, mutations in codon position one ($c1$) generally are less damaging than mutations in position two ($c2$).

With a phase difference of $+1$, the wobble position ($c3$) in gene one corresponds to a fragile $c2$ nucleotide in gene two. With a phase difference of $+2$, the wobble position ($c3$) in gene *two* corresponds to a $c2$ nucleotide in gene *one*. If it is assumed that there is no difference in the evolutionary pressures on two overlapping genes, then by a symmetry argument, one would predict there should be no preference for the $+1$ or $+2$ phase differences (for overlaps of more than six bp). However, any simple argument based on either of the upstream or downstream gene being "more important" is somewhat undermined by our analysis of TCS gene pairs, HK + RR pairs versus RR + HK pairs, where the genes are presumably of equal importance to the organism, and yet the same phase bias was observed (Figures 3.5 and 3.6).

On the basis of this codon based mutual constraint argument, the observed phase $+1$ bias in longer overlaps (Figures 3.4 and 3.13) suggests that in general the tail end of gene one is "more important" than the start of gene two. That is to say, a phase $+1$ overlap allows

maximal non-synonymous changes in the start of gene two while minimizing changes in gene one. This argument can be assessed directly by considering hypothetical point mutations in the overlap region (Cock and Whitworth, 2007a).

On the basis of this mutual constraint argument, in Cock and Whitworth (2007a) we proposed that unidirectional gene overlaps tend to arise from non-overlapping unidirectional genes by the N-terminal extension of the downstream gene by the (gradual) adoption of an alternative start codon. Under this model, the overlap region was originally single coding for the C-terminal region of the upstream gene, and becomes dual coding giving an initially random N-terminal prefix to the original downstream gene. There would then be evolutionary pressure to optimize the amino acid composition of this new stretch of protein, which can occur more easily *without* disruption to the upstream gene in phase $+1$. Assuming generation of overlaps in either phase happens at similar frequencies, this mutual constraint suggests that phase $+1$ overlaps are more likely to be retained by purifying selection, resulting in a phase $+1$ bias.

Alternatively, unidirectional overlaps could arise by the C-terminal extension of the upstream gene - for example a (point) mutation in the original stop codon leading to an extension to the next in-frame stop codon. Applying the same argument as above suggests a phase $+2$ bias would be expected. In Cock and Whitworth (2007a) a simplistic codon frequency analysis suggested the codon usage in non-overlapped regions was more similar to that of the overlapped region of the upstream gene than the downstream gene, taken as support for the creation of these overlaps by extension of the downstream gene via an alternative start codon.

In an alternative explanation, the following section will explore unidirectional overlap generation by the adoption of alternative start or stop codons, based on the analysis of Kingsford *et al.* (2007) for convergent gene overlaps.

3.6 Generating overlaps from alternative start/stop codons

Kingsford *et al.* (2007) introduced a model explaining the convergent gene overlap phase bias based on the likelihood of finding alternative stop codons in the reverse complement of non-overlapping gene coding sequences. They concluded convergent gene overlaps arose by the loss of a stop codon leading to random extension of a gene to the next in-frame stop codon, followed by a selection against longer overlaps which was modelled as an exponential fitness function.

This analysis is repeated here and generalised to cover the other gene orientations. The extension to look for alternative stop codons in non-overlapping unidirectional gene pairs

is straightforward. Alternative start codons can also be searched for to generate potential unidirectional or divergent gene overlaps. This analysis does not explicitly look for Shine-Dalgarno binding sites (Shine and Dalgarno, 1975), nor any other translation initiation regions, which would have to be dual coding with the protein sequence of the other gene. However, the function of these regions is not known to require a precise distance from the start codon, and thus should be irrelevant to the phase effects we are interested in.

3.6.1 Generating convergent overlaps from alternative stop codons

Kingsford *et al.* (2007) looked at non-overlapping convergent gene pairs, and considered how these could give rise to overlaps by the adoption of an alternative stop codon. They searched each non-overlapped gene for the first reverse complement stop codon. A histogram of the convergent overlap which would result from a neighbouring gene adopting this stop codon revealed the same phase bias observed in Figure 3.11. Such an overlap could be created by a simple point mutation of the existing stop codon (assuming the gene separation was in the correct phase) or by a more radical mutation.

This analysis is repeated in Figure 3.17, where the horizontal axis is labeled by the resulting overlap length. As in Kingsford *et al.* (2007), there is a striking phase bias, with phase $+1$ most common. Each phase appears to show an exponential-like decay, but with different rates. As a result, for overlaps up to about 12 bp, phase $+0$ is more common than phase $+2$, but this ranking switches for the longer overlaps. Kingsford *et al.* (2007) Figure 3 shows similar behaviour, however in their dataset phase $+2$ overtakes phase $+0$ to become the second most common phase at around forty bp.

3.6.2 Generating divergent overlaps from alternative start codons

Figure 3.18 shows a similar analysis, searching non-overlapping divergent genes for the first reverse complement start codon (looking for any valid start codon in the relevant codon table). This appears to show a similar overlap phase bias to that observed in Figure 3.7, with phase $+1$ more common than phase $+0$ than phase $+2$.

A large number of potential overlaps of lengths one and two are found, with $n = 2$ about ten times more common, as expected. However, these short overlaps make up a far higher fraction of the generated overlaps than observed in annotated overlaps. In addition, there are a substantial number of overlaps of length $n = 5$ predicted, which does not agree with the overserved data where none have been annotated. These differences are likely due to the fact that any potentially valid start codon has been accepted when generating the hypo-

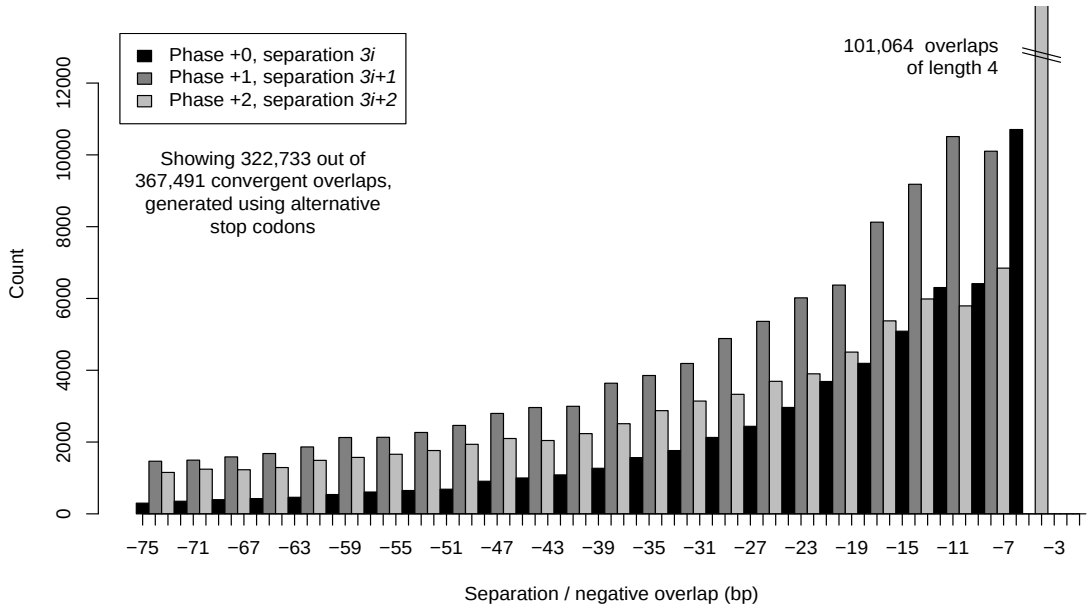


Figure 3.17: Hypothetical convergent overlaps generated by an alternative stop codon, from the 457 species listed in Table A.1.

thetical overlaps, which does not reflect the start codon usage observed in annotated overlaps (Table 3.3). One particular problem with this simplistic analysis is that some alternative start codons defined for the bacterial genetic code may only be used by a subset of species. More generally, different start codons have different initiation rates and thus different usage rates.

This start codon bias is dealt with very simply in Figure 3.19, where only overlaps generated using a commonly observed start codon (`ATG`, `GTG` or `TTG`) were accepted (these three start codons together make up the bulk of all annotated start codons observed). The number of $n = 1$ and $n = 2$ overlaps are much reduced compared to Figure 3.18 (using any start codon), although still over-represented. There are only a handful of overlaps of length $n = 5$ predicted now, which does better match the observed distribution. In reducing the number of short overlaps, there are correspondingly more longer overlaps ($n \geq 6$), which now show a very clear phase bias (with phase $+1$ most common), much more pronounced than in the observed distribution (Figure 3.7), although phases $+0$ and $+2$ are now almost equally common. There is a smooth exponential decay, bar the low number of overlaps of length $n = 6$, although that is observed in Figure 3.7. This exponential decay appears to be faster than in the oberserved data (Figure 3.7).

Given the comparatively small sample of divergent gene overlaps observed, and the relatively indistinct phase patterns therein (Figure 3.7), further comparison with these predictions (Figures 3.18 and 3.19) has not been pursued.

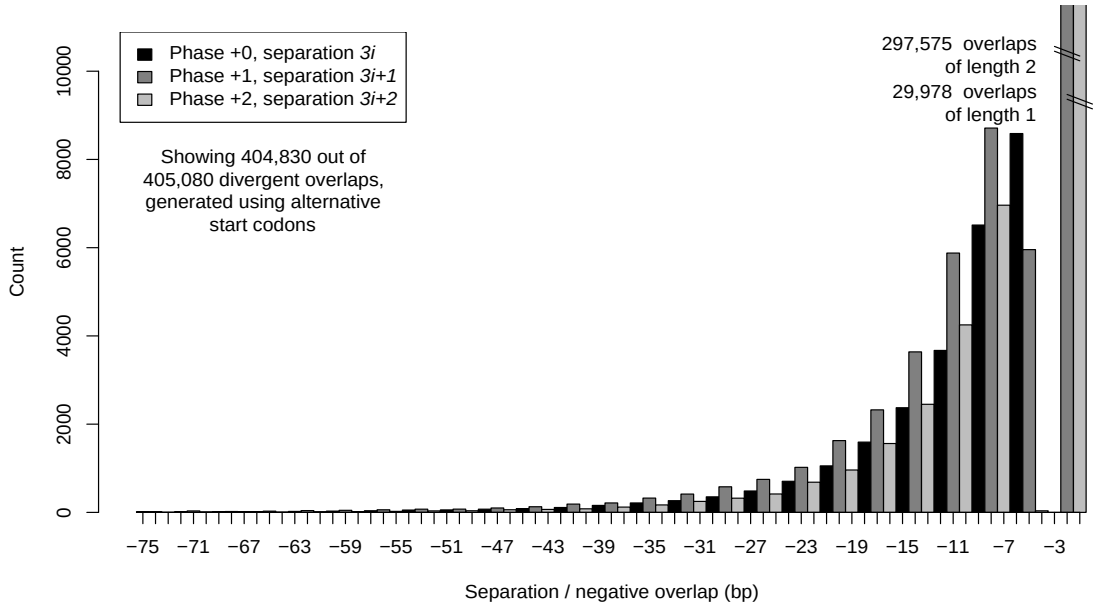


Figure 3.18: Hypothetical divergent overlaps generated by any potential valid start codon, from the 457 species listed in Table A.1. Any valid start codon in the relevant codon table was accepted.

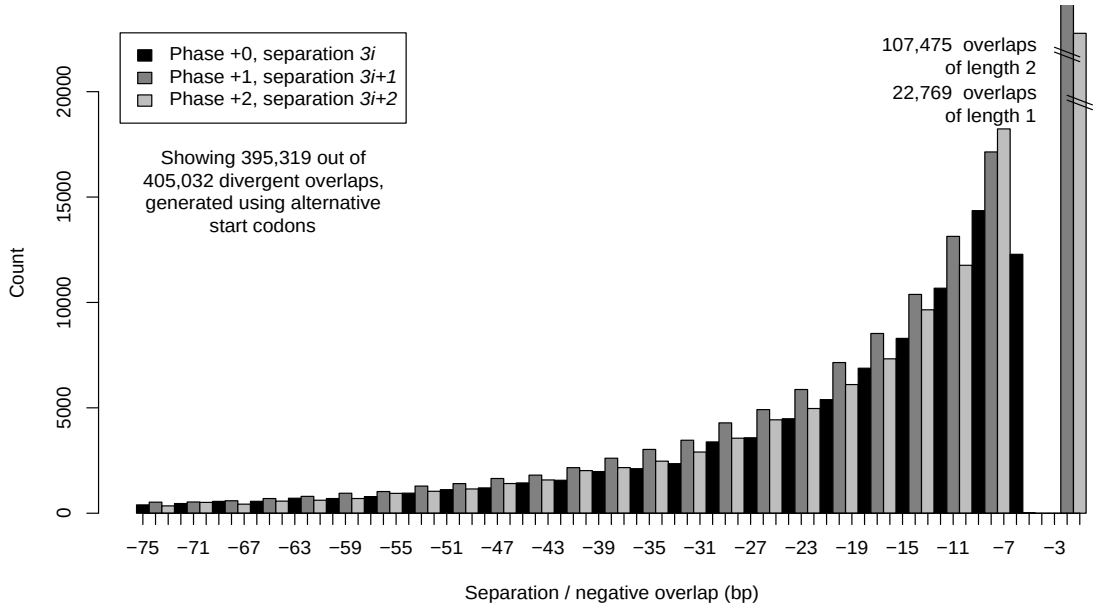


Figure 3.19: Hypothetical divergent overlaps generated by any common start codon (`ATG`, `GTG` or `TTG`), from the 457 species listed in Table A.1.

3.6.3 Generating unidirectional overlaps from alternative start/stop codons

Analysis of non-overlapping unidirectional genes is slightly more complicated, as both N-terminal extension of the downstream gene via a new start codon, and C-terminal extension of the upstream gene via a new stop codon must be considered.

Figure 3.20 shows the resultant overlaps from the most C-terminal out-of-phase start codon in the upstream gene, using any valid start codon in the genetic code assigned to each organism. For the longer overlaps ($n \geq 6$) there is a smooth exponential decay, with phase $+1$ about twice as common as phase $+2$. This is the phase bias expected, but it is not as pronounced as observed in annotated overlaps (Figure 3.13).

Large numbers of overlaps of lengths $n = 1$ and 4 are predicted, with about twice as many of length $n = 4$. This does not match the oberved ratio where overlaps of $n = 4$ are about three times as common (Table 3.6). Another noteable difference between Figure 3.20 and the observed distribution is the high peak at overlap length $n = 5$, which is almost completely absent in Figure 3.13. As discussed in Section 3.3.3, unidirectional overlaps of length $n = 5$ are only possible using the alternative start codon `ATT`. As in the case of divergent gene overlaps above, this analysis is simplistic in that even "rare" start codons like `ATT` are given equal weighting. This peak at $n = 5$ is absent in Figure 3.21 where only the commonly observed three start codons were considered.

Figure 3.21 shows other differences. Overlaps of $n = 1$ and 4 are also much reduced, which is also to be expected as only the standard three start codons were considered, restricting the set of possible short overlaps (see Table 3.6). Also their ratio is closer to that observed. For the longer overlaps ($n \geq 6$), the phase bias is now much stronger, with phase $+1$ about three times as common as phase $+2$, in much closer agreement with the observed bias (Figure 3.13, Table 3.6). Also of note, in Figure 3.21 the decay rate is much slower than in Figure 3.20.

Figure 3.22 shows the resultant overlaps from the most N-terminal out-of-phase stop codon in the downstream gene. The breakdown of short overlaps ($n < 6$) is in reasonable agreement, with a only a small fraction of $n = 5$ overlaps generated, although the ratio of overlaps of length $n = 1$ and 4 is skewed. The longer overlaps ($n \geq 6$) show exponential decay as expected, except that the number of overlaps of length $n = 6$ is comparatively low. There is a slight phase bias, with phase $+2$ somewhat more common, which is the opposite of the bias observed (Figure 3.13).

Notice that the phase bias in the hypothetical overlaps from alternative start codons (Figures 3.20 and 3.21) matches that observed in annotated overlaps (Figure 3.13), but those generated from alternative stop codons do not (Figure 3.22). This would appear to support

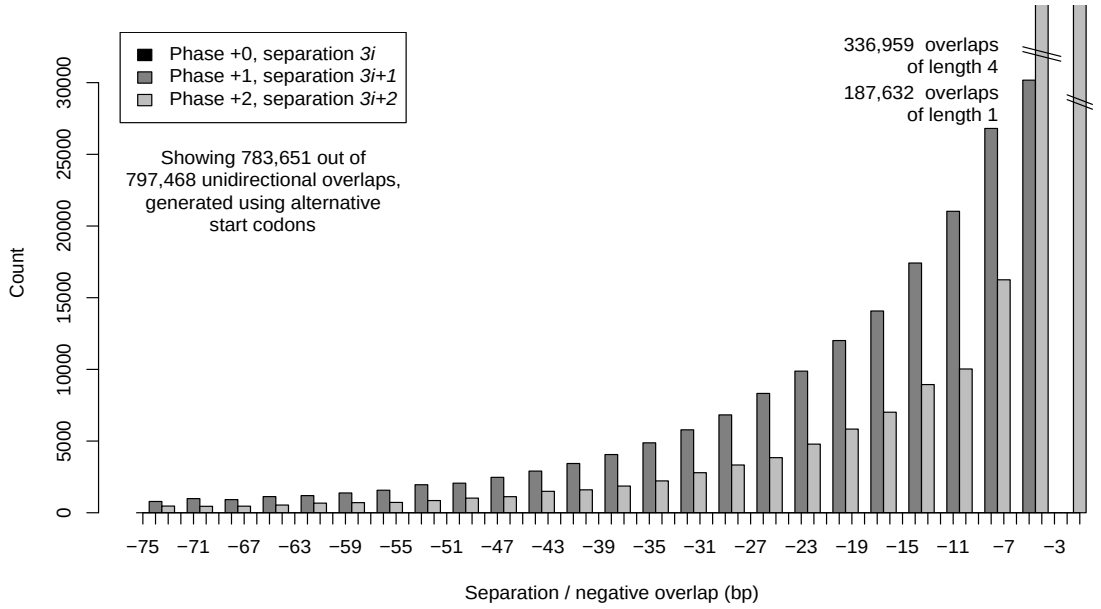


Figure 3.20: Hypothetical unidirectional overlaps generated by any valid start codon within the upstream gene of non-overlapping unidirectional gene pairs, selected from the 457 species listed in Table A.1. Any valid start codon in the relevant codon table was accepted.

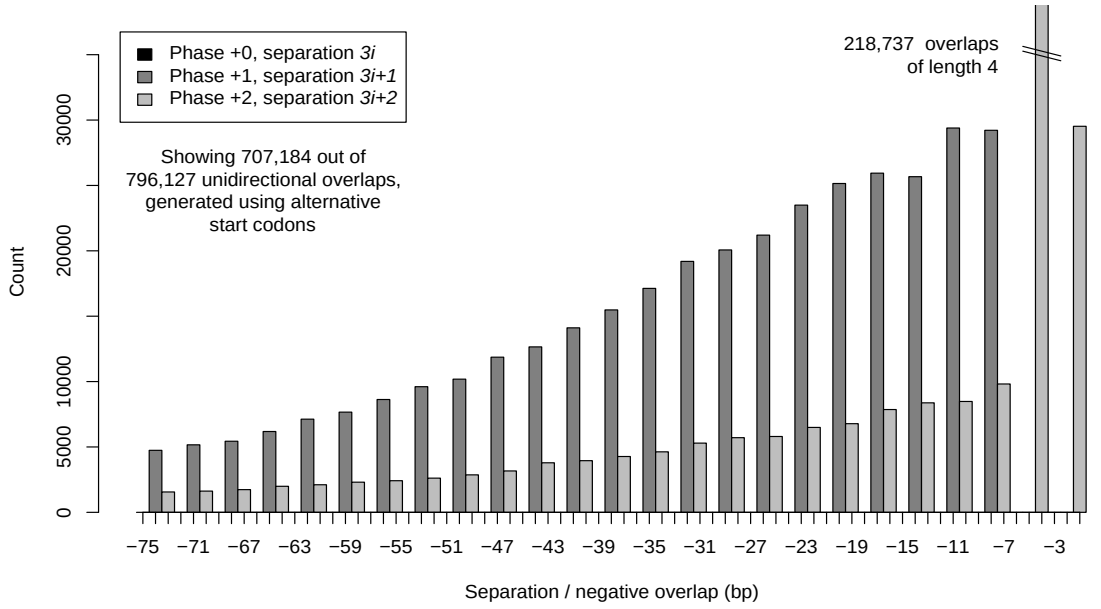


Figure 3.21: Hypothetical unidirectional overlaps generated by any common start codon (`ATG`, `GTG` or `TTG`) within the upstream gene of non-overlapping unidirectional gene pairs, selected from the 457 species listed in Table A.1.

the hypothesis that unidirectional gene overlaps tend to be generated from non-overlapping unidirectional gene pairs by the N-terminal extension of the downstream gene by the adoption of an alternative start codon within the upstream gene, rather than C-terminal extension of the upstream gene by a adoption of new stop codon within the downstream gene.

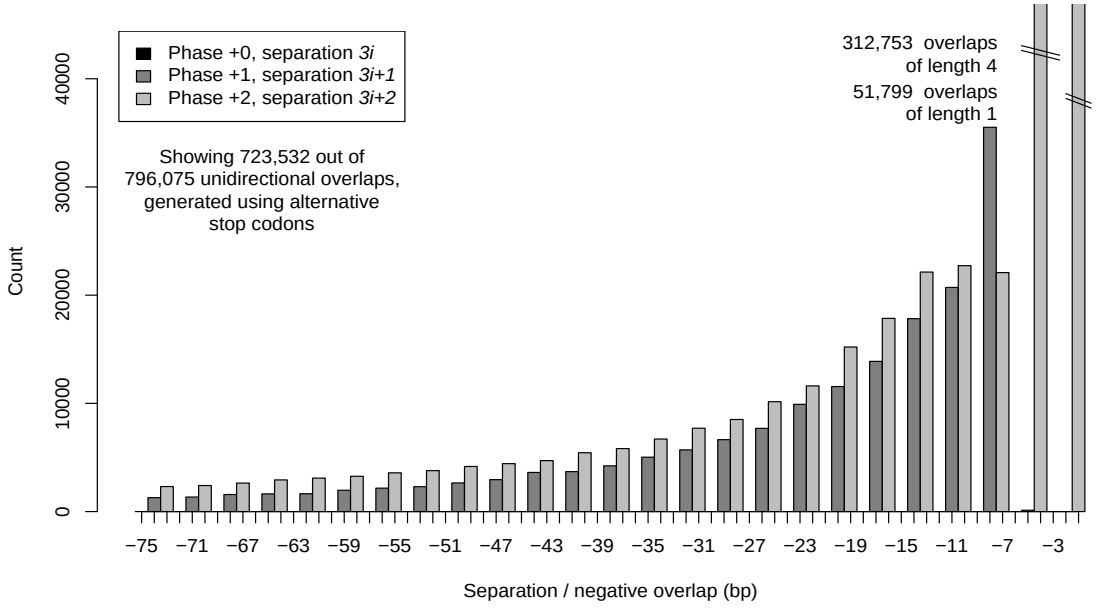


Figure 3.22: Hypothetical unidirectional overlaps generated by any potential alternative stop codon within the downstream gene of non-overlapping unidirectional gene pairs, selected from the 457 species listed in Table A.1.

3.7 Predicting overlap length spectra

The results of the previous section can be used to generate overlap spectra showing the expected phase biases for convergent overlaps (using alternative stop codons) and unidirectional overlaps (using alternative start codons). However the exponential decay rates do not quite match, which in their model for convergent overlaps Kingsford *et al.* (2007) resolved using a two-step model. Firstly, a new convergent overlap is generated by the adoption of a new stop codon. Secondly, the fitness of this new overlap is limited by its length, modelled with an exponential fitness function. This idea could be applied to unidirectional overlaps (and potentially divergent overlaps), modelling their generation by the adoption of a new start codon, followed by a length-based selection.

3.8 Discussion

Given the relative scarcity of divergent overlap examples, we have not examined the case of divergent overlaps in any detail. However, it appears that the overlap phase biases here may also be explained by the relative frequencies of start codons within the other reading frames of normal coding sequences.

Both the mutual constraint selection model (Cock and Whitworth, 2007a), and the alternative start codon model presented above based on Kingsford *et al.* (2007), explain the observed unidirectional overlap phase bias. The models are compatible, both assuming that overlaps tend to happen by the N-terminal extension of a downstream gene. The alternative start codon model suggests a strong phase bias at the creation of the overlap, while the mutual constraint model suggests a strong phase selection bias after the overlaps is created. Either or both may be valid.

Similarly, for convergent overlaps, the alternative stop codon (Kingsford *et al.*, 2007) and mutual constraint selection (Rogozin *et al.*, 2002) models are also compatible. I am inclined to follow Kingsford *et al.* (2007) in concluding that the inherent phase bias in the observed frequency of start and stop codons in alternative reading frames of coding sequences gives the more elegant and straightforward explanation of the convergent overlap patterns observed, and also favour this explanation for unidirectional overlaps.

The phase patterns in overlapping genes have no immediately apparent evolutionary function, but rather are inherently linked to the genetic code itself. The genetic code will have evolved under various pressures. Itzkovitz and Alon (2007) explored a range of hypothetical genetic codes, and concluded that the genetic codes observed in nature are near optimal for allowing additional information to be encoded within a protein sequence. This work did not specifically mention nucleotide sequences simultaneously encoding two proteins, but rather arbitrary (short) sequences representing possible DNA-binding regions or other motifs. Furthermore, the presence of alternative out of frame stop codons within a gene (hidden stop codons) was looked at from the point of view of robustness to translational frameshift errors (Seligmann and Pollock, 2004). The standard genetic code was found to terminate erroneous reads sooner than the hypothetical genetic codes, which is advantageous as less resources are wasted constructing and degrading non-functional proteins. It seems reasonable that functions like double-coding and hidden stop codons may have shaped the genetic code, and thus indirectly contribute to the overlap phase patterns observed.

One assumption in the work presented in this chapter, has been taking annotated start codons at face value. From the perspective of phase biases, as long as two genes still overlap,

the chosen start codon does not actually matter. Also, mis-annotations would be expected to cause marked phase bias in the distribution of non-overlapping genes. This may indeed be the case in divergent overlaps (Figure 3.7), but does not appear to be an issue in the unidirectional gene pairs (Figure 3.13). In Cock and Whitworth (2007a), filtering the dataset to exclude any gene pairs with possible alternative start codons actually made the overlap phase bias slightly stronger, however this dramatically reduces the size and therefore the reliability of the dataset.

While the genetic code itself appears to induce these phase biases in longer overlaps, without searching for translation initiation sites it is not clear how many of these long unidirectional (or divergent) overlaps are biologically relevant. Translational coupling provides a biological reason for short unidirectional gene overlaps, but may not apply to the longer overlaps reported. Indeed, a recent analysis by Pallejá *et al.* (2008) concluded that most long overlaps are mis-annotations. The phase biases of these longer overlaps may be an artefact - our simple model of alternative start codon selection with an exponential length-based fitness criterion can be applied equally well to the genome annotation process!

Finally, the high number of unidirectional overlaps of length $n = 5$ generated using "rare" alternative start codons (Figure 3.20) may be of biological relevance, with the handful of cases annotated being just the tip of the iceberg. Eyre-Walker (1996) noted skewed ratios of alternative stop codons in short overlaps ($n = 1$ or 4), so a similar atypical start codon usage in this context is not unreasonable. Perhaps future work will reveal that overlaps of $n = 5$ are in fact far more common than current annotation indicates.

Chapter 4

TCS gene fusion

4.1 Introduction

The prototypical TCS system consists of two proteins, a histidine kinase (HK) containing a transmitter domain (T), and its partner response regulator (RR) containing a receiver domain (R) (e.g. Figure 1.1). The domain motif based search described in Chapter 2 identified a large number of more complicated genes containing multiple TCS domains (see Table 2.2 on page 47), the most common of which are *simple* hybrid kinases (HYs) containing a single T and a single R domain (e.g. Figure 1.3).

One explanation for such HY genes is they are the result of a gene fusion event, merging neighbouring HK and RR genes into a single composite. However these genes are created, for so many HYs to be preserved across multiple species, these systems must retain some functionality, so it is presumed that there is intra-protein phosphorylation between the T and R domains in a simple HY.

If these simple HY proteins do form self-contained TCS systems, then the presence of an output domain could be expected. In the absence of a separate output domain, a response could be elicited directly by the receiver domain itself, similarly to CheY (Section 1.4.3). Alternatively, rather than constituting a self-contained TCS, a simple HY protein could be part of a phosphorelay, as in the *Vibrio harveyi* quorum-sensing system (Section 1.4.9). However, relatively few phosphotransfer proteins have been identified to date (see Section 1.4).

In this chapter the lists of paired T and R domains generated in Section 2.4 are used to infer apparent net TCS gene fusion rates, and to investigate factors influencing this, in particular the presence of transmembrane helices in HKs, and the presence of DNA-binding domains in RRs. In the absence of such features there is a relative abundance of fused genes. These results were published in Cock and Whitworth (2007b).

Two gene pairs		Isolated hybrids			
Domains	Count	Domains	Count	Fused	Total
$T_i + R$	2092	T_i-R	764	27%	2856
$T_{ii} + R$	5	T_{ii}-R	1	17%	6
$R + T_i$	2689	R-T_i	95	3%	2784
$R + T_{ii}$	30	R-T_{ii}	0	0%	30
Total	4816	Total	860	15%	5676

Table 4.1: Minimal TCS systems from 457 species (Table A.1), data points from Tables 2.2 and 2.3. The fused column shows the fraction in isolated hybrids, an apparent net fusion rate.

4.2 Minimal TCS systems

A *minimal* TCS system will be taken as one containing a single T and a single R domain, either as neighbouring HK and RR genes, or as isolated HYs. Only isolated HYs are used (where there are no other TCS genes within $5,000$ bp) allowing direct comparison to paired genes (where again, there are no other TCS genes within $5,000$ bp, see Section 2.4) to infer apparent net fusion rates.

These systems will be denoted as T + R or R + T for the two gene pairs, and T-R or R-T for the hybrids, with plus signs denoting two separate neighbouring genes, and minuses indicating multiple domains within one gene. Similarly, TR and RT denote minimal TCS systems as either two gene pairs, or as single gene hybrids.

Table 4.1 shows minimal TCS systems identified as part of the survey described in Chapter 2. Given the small number of T_{ii} systems, this chapter will focus on the T_i systems only. There are similar numbers of TR and RT systems ($2,856$ and $2,784$ respectively, excluding T_{ii} systems), however, the proportion of single-gene systems (and therefore the apparent fusion rate), was found to be markedly different for TR and RT geometries: 27% and 3% respectively were hybrid kinases (independence rejected with chi-squared p-value < 0.001).

It is clear that the apparent net fusion rate depends on the domain order, with the TR fusion rate an order of magnitude higher than that for RT systems.

4.3 Transmembrane and DNA-binding domains

Two factors important in TCS function that might affect TCS gene fusion are transmembrane (TM) helices in input domains, and the presence of DNA-binding domains, as these domains require separate spatial localisation for function. TCSs were therefore assessed for the presence of TM helices and DNA-binding domains.

Transmembrane (TM) predictions were made for genes containing T domains using the online web-interface to TMHMM v2.0 (Krogh *et al.*, 2001) (as recommended in an independent

comparison (Moller *et al.*, 2001)). Note that the TMHMM software is now available to download as a standalone tool, which would allow its complete integration into the analysis pipeline, rather than the semi-automatic procedure adopted which prepared batches of queries for the online tool.

The presence of DNA-binding output domains was based on matches to any of the following PFAM domains with an expectation threshold of 10^{-4} using RPS-BLAST (see Section 2.5): pfam00486, pfam00196, pfam02954, pfam04397, pfam00165, pfam04545, pfam00447, pfam00440, pfam01381, pfam04967, pfam00249, pfam00126 or pfam01022.

Table 4.2 shows the breakdown of the minimal TCS systems (with Class I transmitters) by both domain architecture and the presence of TM and DNA-binding domains. Overall, 83% of minimal TCSs possessed TM helices and 26% contained DNA-binding domains, with 20% containing both. However, these breakdowns are very dependent on the TCS system architecture.

Most HKs possessed TM helices ($4,223$ out of $4,781$ [88%]), as did a large number of hybrid kinases (456 out of 859 [53%]). The proportion of TM HKs was found to be much higher in R + T_i systems than in T_i + R systems (97% and 77% respectively), and additionally, was much lower in R-T systems than in T-R systems (1% and 60% respectively).

There also appears to be a relationship between TCS geometry and the presence of DNA-binding domains (Table 4.2). Of $4,739$ RRs within minimal TCSs, 29% possess DNA-binding output domains. Small numbers of HYs were found to have DNA-binding domains (5% of T_i-R proteins, but no R-T_i cases), while in two gene pairs, DNA-binding domains were more common (59% of T_i + R pairs, and 6% of R + T_i pairs). Thus it seems that hybrid kinases possessing DNA-binding domains are selected against, and this selection is stronger for R-T_i hybrids than for T_i-R hybrids (0% of DNA-binding RT geometry TCSs are hybrid kinases, compared to 3% of TR geometry TCSs).

The chi-squared test rejects the independence of gene order and the proportion of fused genes with p-values of 6.1×10^{-3} and 1.6×10^{-237} when TM helices and DNA-binding domains respectively are excluded. However, considering only those TCSs that lack both TM helices

	T_i + R	T_i-R	Fused	R + T_i	R-T_i	Fused	Total	Fraction
TM only	697	422	38%	2454	1	0%	3574	63%
DNA only	311	6	2%	29	0	0%	346	6%
TM+DNA	927	33	3%	145	0	0%	1105	20%
Neither	157	303	66%	61	94	61%	615	11%
Total	2092	764	27%	2689	95	3%	5640	100%

Table 4.2: Occurrence of minimal TCSs, sub-divided by the presence/absence of transmembrane helices (TM), DNA-binding domains (DNA) and TCS geometry.

and DNA-binding domains, we see similar proportions of fused genes in TR and RT geometries (61% RT, 66% TR, p-value for independence 0.416), implying that the presence of TM helices and DNA-binding domains are the main factors affecting the propensity for gene fusion. It also suggests that in the absence of domains requiring specific spatial localisation, evolution generates (and/or retains) fused gene products at a higher frequency than separated gene pairs, which agrees with more general studies (Snel *et al.*, 2000; Kummerfeld and Teichmann, 2005).

4.4 TCS domain location in HK and RR genes

The results of the previous section clearly show domain order (TR versus RT) is linked to the apparent fusion rate. A related question is the location of the T and R domains within HK and RR genes. Figure 4.1 uses bar-charts to show the position of the T within each HK, and the R within each RR, for the paired systems discussed in this chapter.

It is apparent that the transmitter domain is typically within 25 amino acids of the C-terminus of the HK, and there are normally hundreds of amino acids preceding it at the N-terminus (which presumably includes an input domain). Figure 4.2 shows the same data divided according to whether or not the HK is TM, and the same patterns persist.

For the RRs, the opposite holds. The receiver is usually within 25 amino acids of the N-terminus, and there is typically a C-terminal region of 50 to 400 amino acids (presumably containing an output domain). Figure 4.3 shows that DNA-binding domains are all within a C-terminal region of around 100 amino acids, while for non-DNA-binding output domains there is a wider range of lengths.

4.5 TCS domain location in HY genes

Figures 4.4 and 4.5 show similar plots for the T_i-R and R-T_i isolated hybrids, with the addition of separation of the TCS domains. For the T_i-R hybrids, there is generally an N-terminal region of up to 1000 aa, presumably containing an input domain. The region between the TCS domains is typically around 25 aa, with a second peak at about 150 aa perhaps indicative of an intermediate domain. Most of these HYs have a C-terminal region less than 50 aa, however, some are longer suggesting the presence of an output domain.

For the R-T_i hybrids (Figure 4.5), the N and C-terminal regions are generally short (under 50 aa), but there is considerable variation in the intra-TCS domain spacing, where input and output domains might be expected.

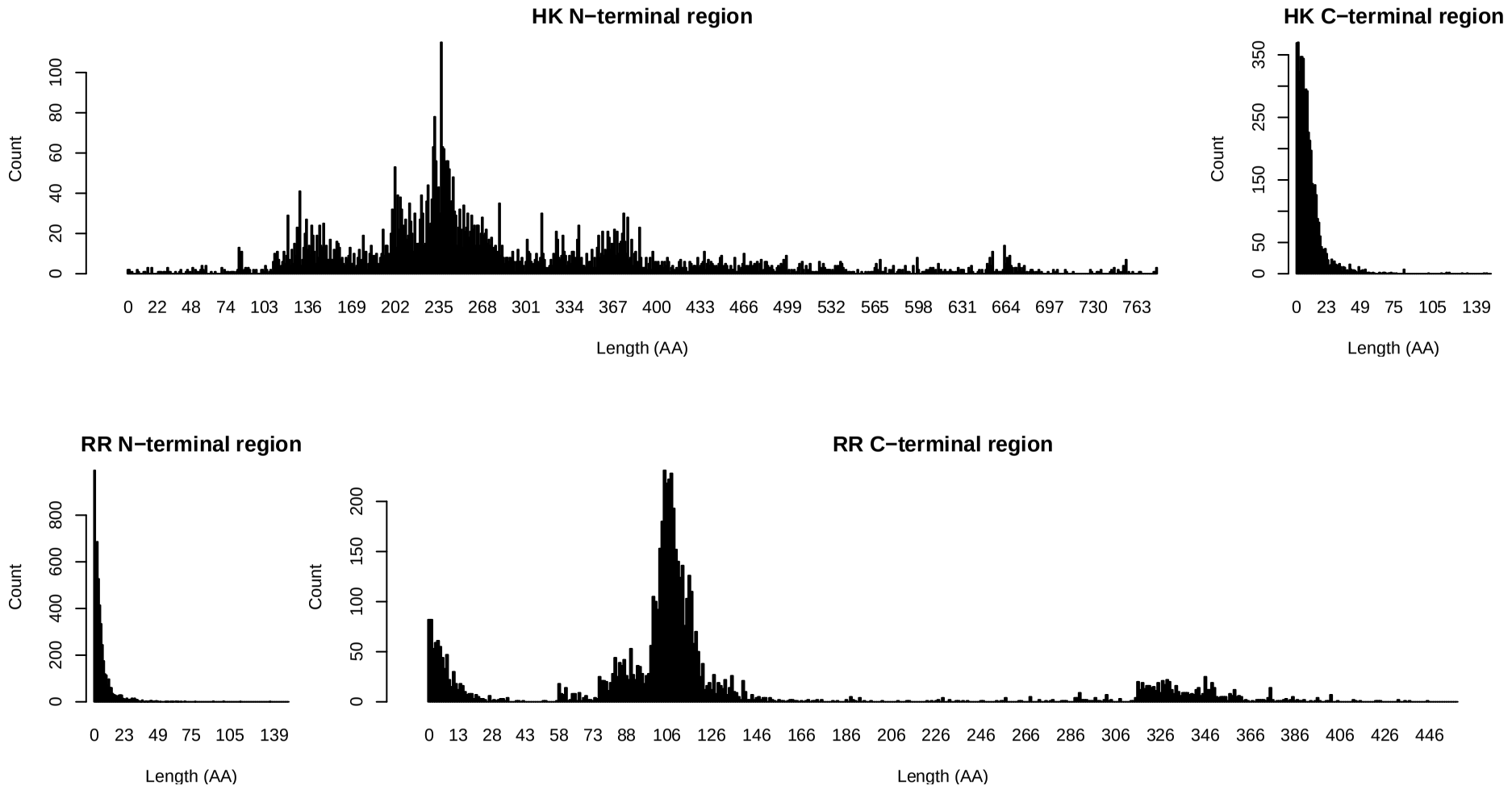


Figure 4.1: Lengths of the regions N and C-terminal of the transmitter and receiver domains in HK and RR proteins respectively. These bar-charts show only paired HK and RR proteins, the top two are for HKs and the bottom two for RRs, with N-terminal lengths on the left, and C-terminal lengths on the right.

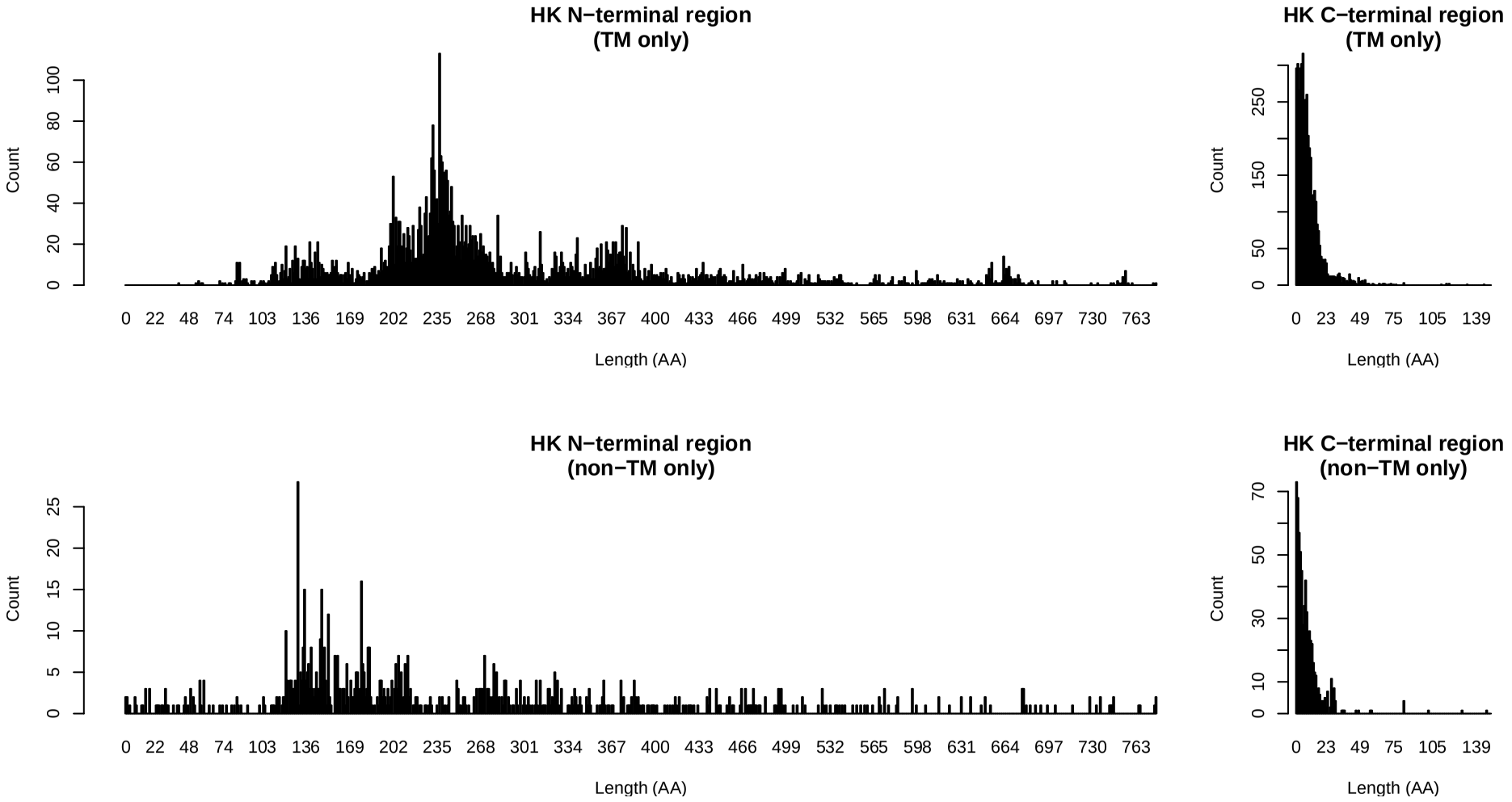


Figure 4.2: Lengths of the regions N and C-terminal of the transmitter domain in HKs. These bar-charts show only HKs paired with RR proteins, the top two are for TM HKs and the bottom two for non-TM HKs – a sub-division of the HK data shown in Figure 4.1.

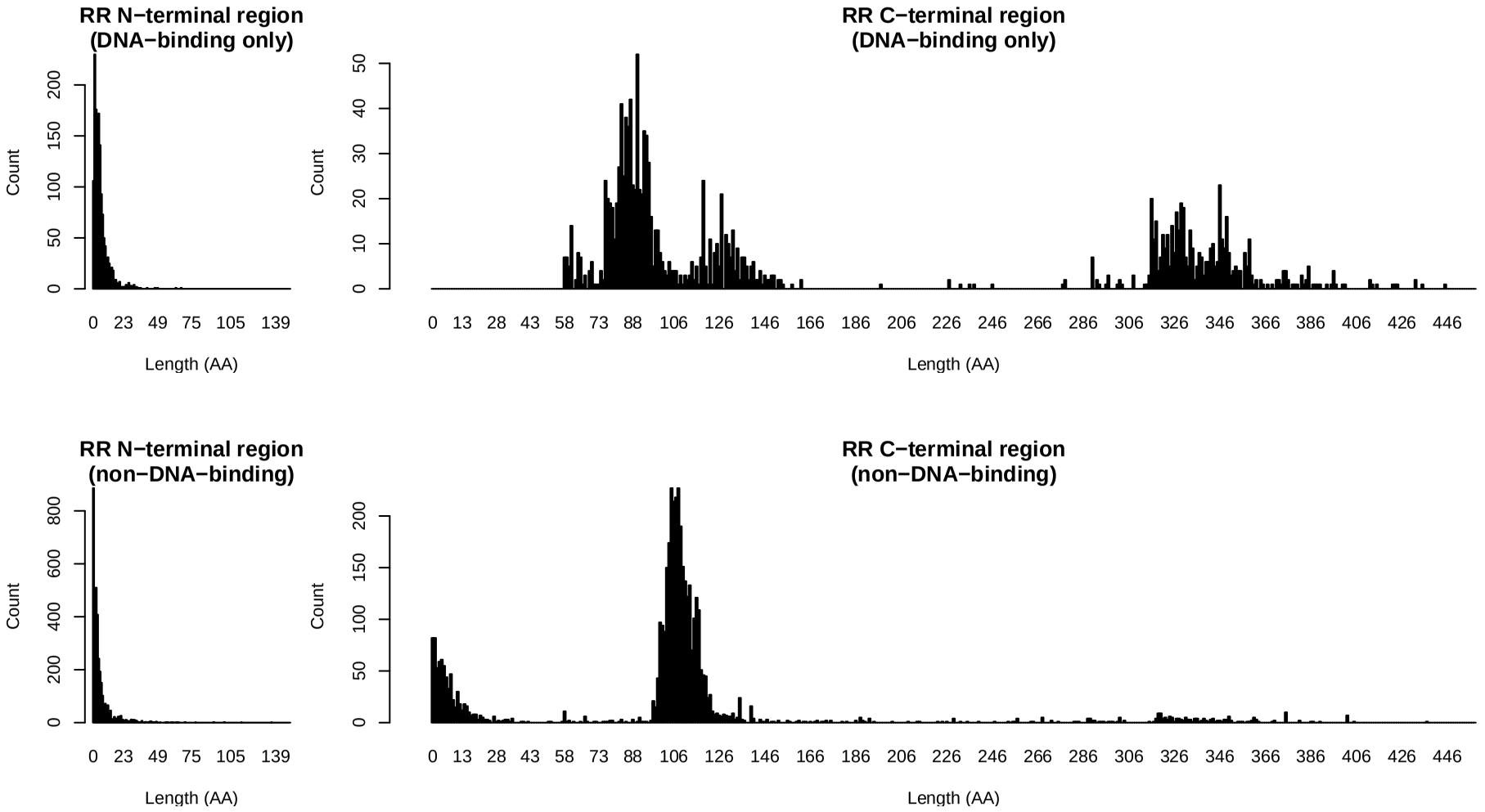


Figure 4.3: Lengths of the regions N and C-terminal of the receiver domain in RRs. These bar-charts show only RRs paired with HK proteins, the top two are for RR containing a DNA-binding domain and the bottom two for RRs lacking a DNA-binding domain – a sub-division of the RR data shown in Figure 4.1.

T_i-R (isolated hybrids)

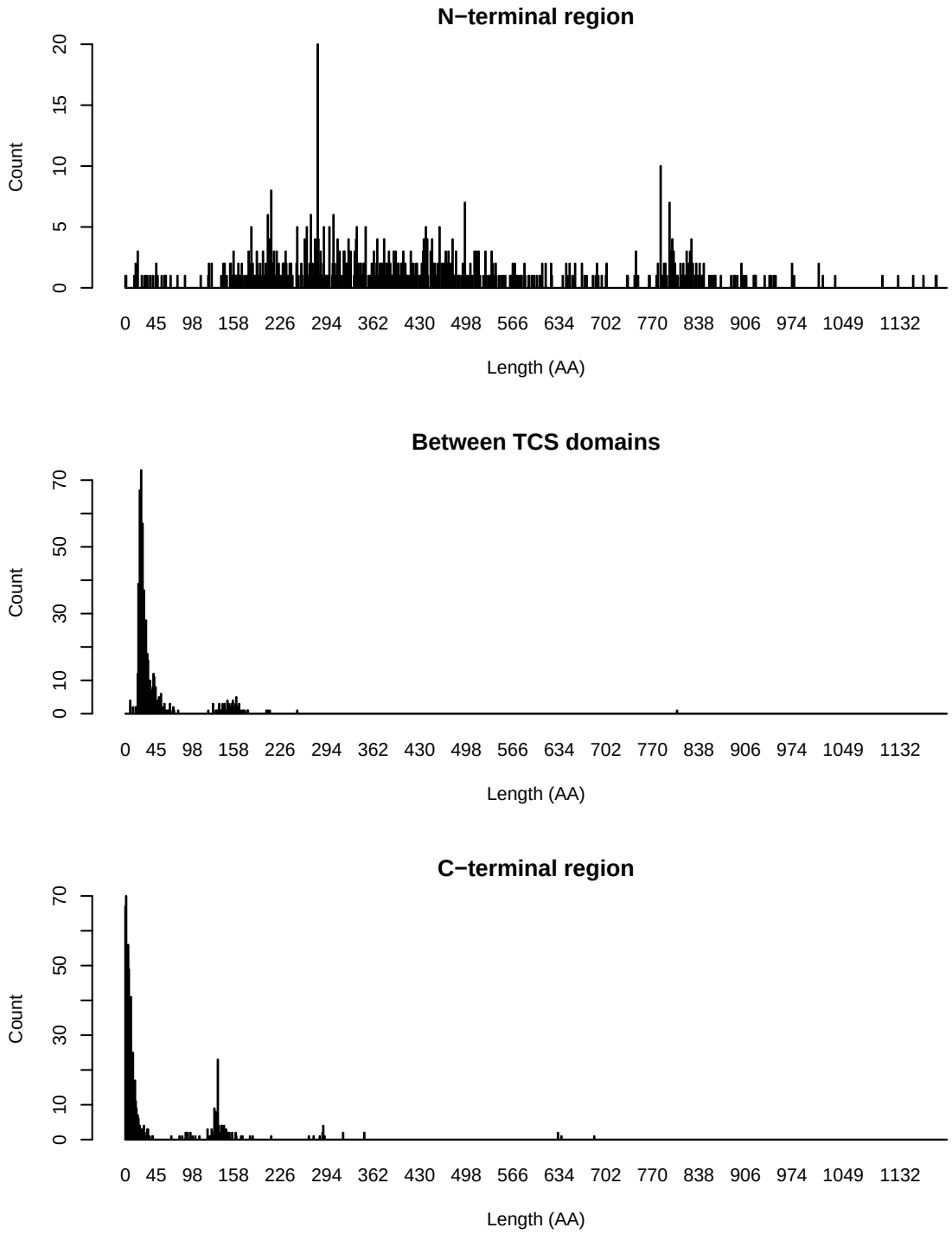


Figure 4.4: Bar-charts of the lengths of the regions N-terminal to the transmitter, between the receiver and transmitter, and C-terminal to the receiver domain, in isolated T_i-R HYs.

R-T_i (isolated hybrids)

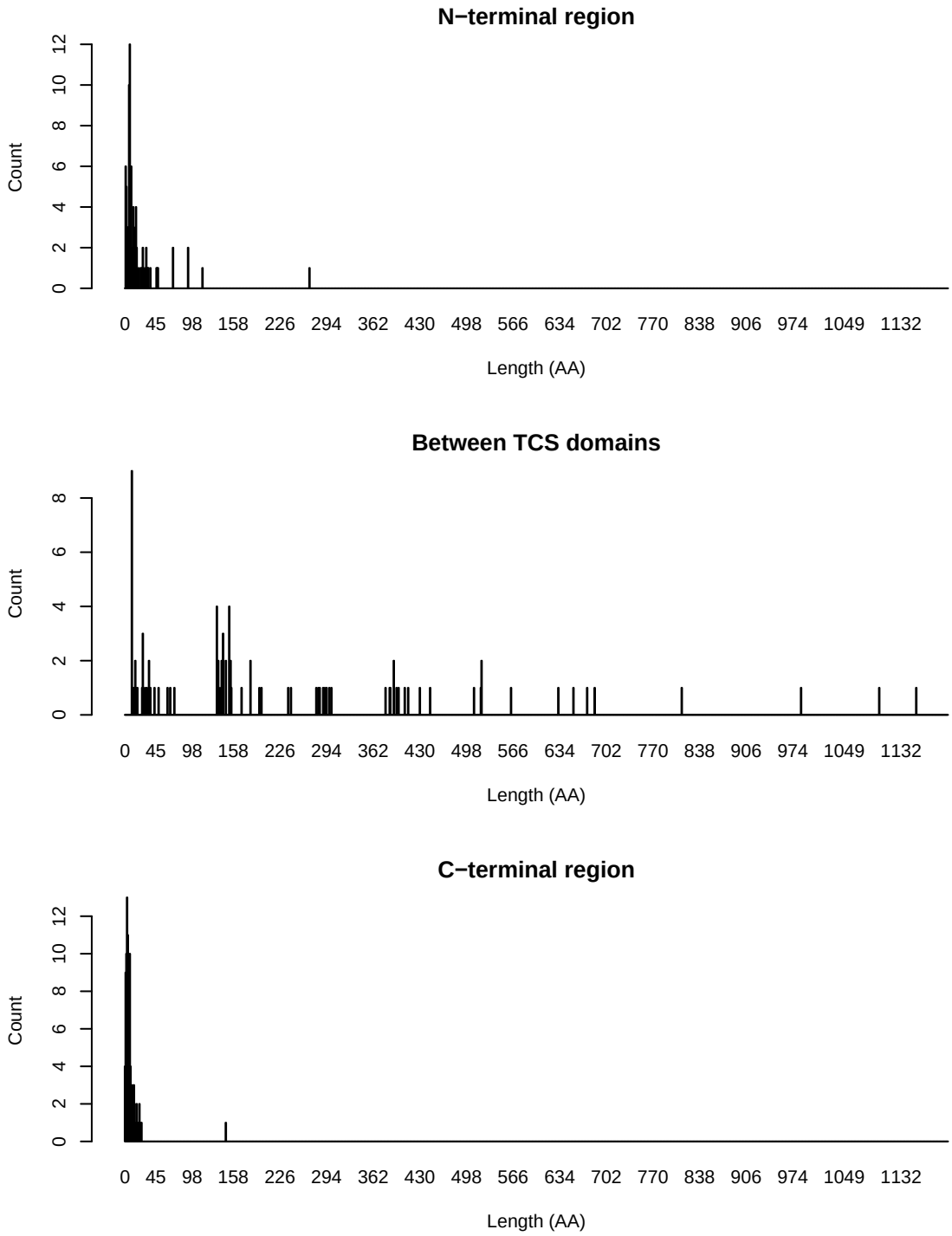


Figure 4.5: Bar-charts of the lengths of the regions N-terminal to the receiver, between the receiver and transmitter, and C-terminal to the transmitter domain, in isolated R-T_i HYs.

4.6 Domain separation

Figure 4.6 shows the separation (nucleotide bps) between encoded transmitter and receiver domains for the four minimal TCS geometries. Domains tend to be much closer to each other in TR than RT systems, which is consistent with the observation that most RRs possess C-terminal output domains, and typical HKs have N-terminal input domains (Section 4.4).

A minimum separation between the TCS domains is apparent in the HY genes. These TCS domain separations are also shown measured in amino acids in Figures 4.4 and 4.5.

A sub-population of T_i-R systems is apparent with an average domain separation of ~ 500 bp, ~ 400 bp larger than the main population, suggesting the presence of an additional domain between the transmitter and receiver domains. A periodicity of ~ 400 bp may also be present for the inter-domain distances of RT systems, presumably reflecting integer values of intervening domains between the transmitter and receiver domains.

4.7 Discussion

If HYs arise from the fusion of HK and RR genes, then many of the trends observed can be explained due to the typical domain arrangements. HKs tend to have an N-terminal input domain, and a C-terminal transmitter (INPUT-T_i), while RRs tend to have an N-terminal receiver domain and C-terminal output domain (R-OUTPUT) (Section 4.4). Thus a TR fusion, starting with an INPUT-T_i HK upstream of an R-OUTPUT RR, would give a domain layout of INPUT-T_i-R-OUTPUT, while an RT fusion would give R-OUTPUT-INPUT-T_i instead.

DNA-binding output domains require some steric freedom in order to function, and hypothetically this would be hampered by being a central domain within a HY. Thus an RT fusion (expected to give R-OUTPUT-INPUT-T_i) would be less functional than a TR fusion (which would be INPUT-T_i-R-OUTPUT), which matches the observed data (39 T_i-R with DNA-binding, but no R-T_i cases). However, this alone does not seem to explain the scarcity of HYs with DNA-binding domains.

TM proteins are known to have an N-terminal signal peptide which marks them for membrane export. In a TR fusion event, this marker would be preserved at the N-terminal of the resulting hybrid gene. However, in RT fusions the marker would be lost or rendered non-functional by virtue of being in the central core of the new fused gene. Another important order dependent difference between hypothetical fusions of TM HK with RR genes is the location of the TM input domain. If as is normally the case, the HK domain layout is INPUT-T_i, then in a TR fusion the TM input domain remains at the N-terminus of the protein, and thus

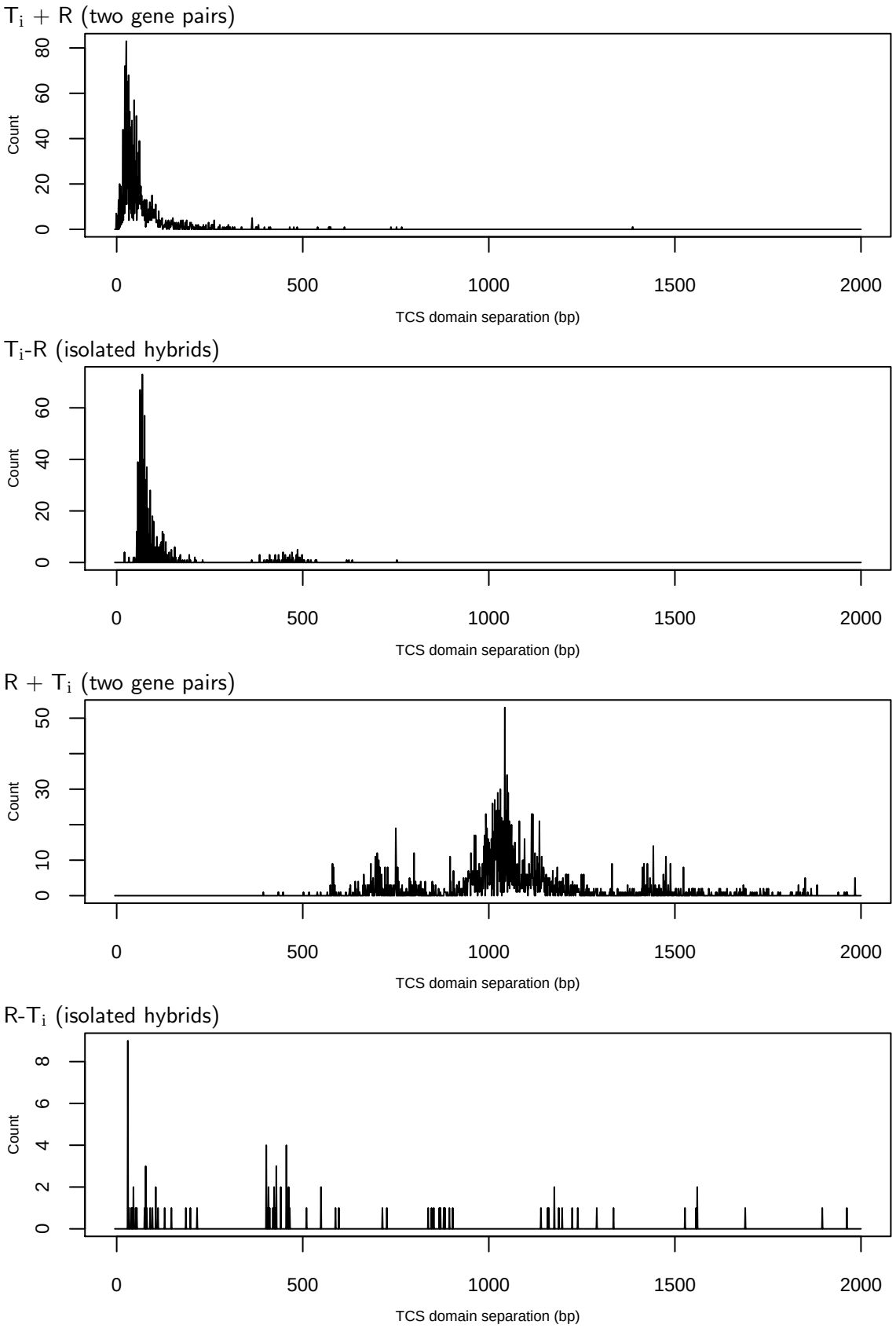


Figure 4.6: TCS domain separation in minimal systems, for the four system architectures.

should remain functional. In this case, the previously cytosolic RR domains become membrane associated, which may impact their functionality, depending on the nature of the output signal (a particular concern for DNA-binding). However, for an RT fusion this would place the TM region in the middle of the resulting HY protein, which may complicate or prevent the membrane insertion (even if given a suitable membrane export signaling marker). Furthermore, if there are an odd number of TM helices, this would put the transmitter and receiver domains on opposite sides of the membrane, preventing the phosphotransfer.

On the basis of these arguments, while TM T_i-R HYs should be possible (and indeed 455 were identified), no TM R-T_i HYs would be expected. In fact, a single TM R-T_i HY was found, StyS of *Xanthomonas axonopodis pv. citri* (GI:21244368), with two predicted TM helices in the middle of the protein. This may simply be a false positive TM prediction.

20% of minimal TCS systems contain both a TM input domain and a DNA-binding output domain. From the arguments above, here an RT fusion event would be non-functional (and no R-T_i HYs were observed with both TM and DNA-binding domains), while a TR fusion of a TM HK and DNA-binding RR resulting in an INPUT-T_i-R-OUTPUT domain structure does appear to be *possible*. Such proteins would be membrane bound, thus in order for the output domain to bind to the DNA, the DNA itself would have to move into proximity with the HY, which seems energetically unfavourable. On the contrary, for normal RRs or cytosolic HYs, the protein itself is mobile and can diffuse to the DNA. It is therefore somewhat surprising that a total of 33 HYs containing both a transmembrane domain and a DNA-binding output domain were observed (Table 4.2). These were from *Bacteroides thetaiotaomicron*, *Bacteroides fragilis* (multiple occurrences each, also reported in Xu *et al.* (2004), see Section 1.4.5) and three strains of *Bacillus cereus* (one example each). If these proteins truly have transmembrane helices, and are therefore membrane anchored, perhaps their DNA-binding domains act to spatially organize the DNA in a more complex regulatory process than the simple modulation of gene expression?

TMHMM was trained on a set of only 160 known TM proteins, of which 76 are from Eukaryota (64 from Mammalia), 3 from Archaea (all Halobacterium), 76 from Bacteria (of which 48 are Escherichia) and 5 from viruses (Krogh *et al.*, 2001, suplementary information). Thus only half are from prokaryotes, and half of these are from Escherichia. It is possible that there is some species bias in the transmembrane predictions.

Finally, 11% of minimal TCS systems contain neither a TM input domain nor a DNA-binding output domain. With no further information about the nature of the domains in these proteins, there is no reason *a priori* to expect the proportion of HYs to be different for the TR

and RT geometries, which indeed are very similar (61% RT, 66% TR, p-value for independence 0.416). Moreover, the fact that the apparent net fusion rate is so high in the absence of a TM input and a DNA-binding output domain (with about two thirds of these systems being HYs), suggests that when otherwise prevented, TCS genes will readily fuse into single proteins – effectively becoming one-component systems.

Considering the domain separations of T_i-R and R-T_i systems (Figures 4.4, 4.5 and 4.6), it is apparent that very few of the hybrid systems have a TCS domain separation less than ten amino acids (i.e. 30 bp). This suggests that a linker region of at least ten amino acids is required for appropriate interaction between the receiver and transmitter domains in both T_i-R and R-T_i HYs.

Providing there is a sufficiently flexible linker region between the T and R domains, fusion allows the diffusion limited step in the TCS signalling cascade to be eliminated, resulting in a quicker/stronger response to the stimuli. It would also reduce the chance of cross-talk with other receivers. TCS gene fusion could therefore be described as a heavy handed approach to tuning the bacteria's decision making network.

Taken together, this data suggests that the presence of TM helices and DNA-binding domains are the main factors affecting the propensity for TCS gene fusion. It also suggests that in the absence of domains requiring specific spatial localisation, evolution generates (and/or retains) fused gene products at a higher frequency than separated gene pairs, which agrees with more general studies (Snel *et al.*, 2000; Kummerfeld and Teichmann, 2005). This data provides numerical support for the TCS dogma that the HK and RR are two separate components in order to be able to couple spatially separated stimuli and responses.

4.8 Conclusion

In summary, these findings suggest that the presence of TM helices and DNA-binding domains appear to be the primary factors correlating with observed rates of apparent TCS gene fusion. In the absence of such domains there appears to be a general tendency to formation (and/or retention) of fused TCS systems. A further consideration is the relative genetic distance between encoded transmitter and receiver domains. This appears to be related to apparent TCS gene fusion rates in a geometry-specific manner, presumably a reflection of the typical domain ordering found in HK and RR genes. Additionally, it appears that any minimal linker random-coil between T_i-R or R-T_i domains must be at least ten amino acids long.

HYs enforce a single cellular location upon the entire TCS. 83% of minimal TCSs contain at least one TM helix, suggesting that the role of the majority of TCSs is to couple

extracellular sensing to internal responses. 20% of minimal TCSs contain both TM and DNA-binding domains, coupling extracellular sensing with transcriptional responses, and of these only 3% are HYs. However, in addition to enforcing a single cellular location upon the entire TCS, the gene fusion of a HK with RR also removes a diffusion-limited step in signal transduction (formation of the transmitter-receiver complex), such that the resulting hybrid kinase would, if functional, exhibit an increase in signalling speed and efficiency. Such fused TCSs could be regarded as a step backwards towards one-component systems (OCS), which consist of single proteins directly coupling an input and output domain (Ulrich *et al.*, 2005). However, the newly-formed hybrid kinase would retain its phosphotransfer signalling mechanism, providing additional opportunities for modulation of signal transduction by extrinsic kinases and phosphatases.

Chapter 5

Identifying amino acid residues for TCS partner specificity

5.1 Introduction

This chapter analyses paired multiple sequence alignments (MSAs), with the aim of identifying amino acid positions (MSA columns) important in the specificity of TCS HisKA-receiver protein-protein interactions. The residues thus identified are used in the following chapter to predict TCS protein pairings.

In general, any highly conserved MSA columns have been preserved during evolution and are interpreted as critically important for protein function. In the case of TCS domains, such residues are presumably essential for the correct structural folding, the formation of the protein-protein complex, or are perhaps catalytic residues required to permit phosphotransfer. However, highly conserved MSA columns by their nature cannot restrict partner specificity, where some variability is required. The aim of the analysis presented is to identify co-varying MSA column pairs from the two proteins, with the expectation that these will represent interacting residues in the protein-protein complex governing partner specificity.

By means of an analogy, think of the protein-protein complex as a matching key and lock (Figure 5.1), and consider a collection of keys and locks, where all the keys fit in all the locks, but only unlock their specific partner. Comparing the keys to each other, some parts are perfectly conserved, in particular the length and cross section of the blade. There are no such constraints on the the key bow (or handle) where many designs are possible and equally functional. Similarly some parts of the lock such as the key hole and tumbler mechanism will be conserved, while the casing or any handle need not be. The lock-key pairings are governed by variations in the key teeth and compensating variations to the pin lengths in the lock cylinder

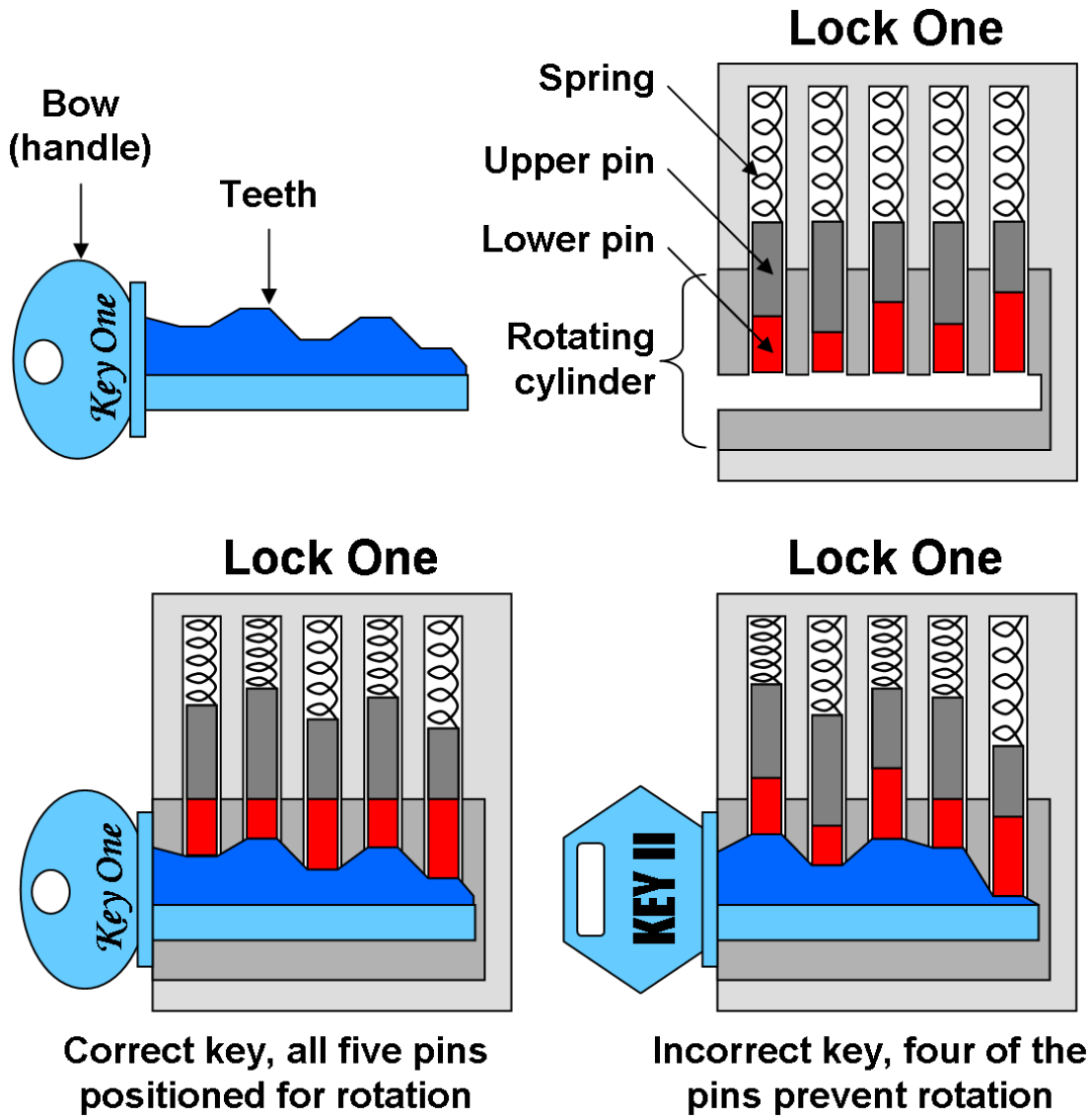


Figure 5.1: Simple pin tumbler cylinder lock and keys. In this type of lock, when the key is inserted the pins within the lock are raised up to different heights by the key teeth. Each pin is divided into two parts (coloured grey and red here), and can therefore sheer at a particular height, allowing the cylindrical lock-barrel to turn. The important parts for specificity are the key teeth (in dark blue) and the lock pins (in red). The two keys have been drawn in two contrasting styles, emphasising that the bow (handle) is unimportant.

tumblers. To identify key-lock parings within this set, we need only consider the key teeth and the lock pins - everything else is either perfectly conserved or irrelevant.

Returning to the world of proteins, we need to identify which parts of the two proteins are important for pair specificity (corresponding to the teeth and pins in the key-lock example), and in particular how they interact (which key tooth matches which lock pin). A naive analysis of the properties of a set of key-lock pairs should be able to identify these tooth/pin pairings as correlations - but may also select spurious matches, for example in the choice of metal, or the branding. Similarly we might expect some false positives in our MSA analysis from phylogenetic effects.

Various approaches have been used in the literature to tackle similar problems such as identifying protein-protein interaction sites from correlated mutations in MSAs (Göbel *et al.*, 1993; Pazos *et al.*, 1997). One particular measure is mutual entropy or mutual information (MI) (Shannon and Weaver, 1949), a measure of co-variation which can be calculated directly from the sequence letters. It has previously been used to identify potential residue interactions within proteins or between interacting proteins (Korber *et al.*, 1993; Giraud *et al.*, 1998; Atchley *et al.*, 1999, 2000; Buck and Atchley, 2005; Martin *et al.*, 2005; Gouveia-Oliveira and Pedersen, 2007; Dunn *et al.*, 2008), having earlier been successfully applied to RNA structure analysis (Chiu and Kolodziejczak, 1991; Gutell *et al.*, 1992; Gorodkin *et al.*, 1997; Adami, 2004).

In this chapter several approaches, including the use of the MI (Section 5.5.4), are discussed as methods to identify which amino acid residues are important for the specificity of TCS interactions using paired MSAs. As there is only limited and indirect information available about the spatial orientation of the HisKA-receiver complex, inter-residue distance information has not been used except to interpret the results. If the specificity residues can be identified and interactions predicted *without* the use of three-dimensional data, this makes the approach applicable to a wider range of protein-protein interactions where there is no three-dimensional data at all. In the following chapter a predictive model is considered, based on these particular amino acids selected from the sequences of the HisKA and receiver domains.

5.2 TCS protein complexes

The survey of TCS genes generated thousands of paired transmitter and receiver domains (Chapter 2), including both Class I transmitters (with HisKA domains) and the minority Class II transmitters (with Hpt domains, see Section 1.2). This and the following chapter focus exclusively on the more numerous Class I transmitters, where the HisKA-receiver complex is important for phosphotransfer.

3D structures of docked transmitter and receiver domains would be ideal for identifying interacting residues, based on physical proximity. However, to date co-crystals have only been solved for a few atypical cases (Section 1.5), and even predominantly buried residues have been shown to play a role in recognition (McLaughlin *et al.*, 2007). A related question of which residues are *essential* can be tackled experimentally for a given transmitter/receiver pairing - for example with alanine mutation studies, e.g. Tzeng and Hoch (1997); Jiang *et al.* (1999); McLaughlin *et al.* (2007). Doing this to answer the question of which impart *partner specificity* would require mutation studies on a panel of transmitter/receivers. Nevertheless, there is existing information regarding the role and importance of particular residues for certain known TCS pairs, and this will be useful to validate the results of the analysis.

5.3 Approach

If one accepts that all HisKA dimers (or receiver domains) share broadly the same three dimensional structure, and that any two examples could be superimposed on each other with a rigid body motion, then columns in a MSA correspond to specific locations on the average domain structure. Furthermore, assuming that all interacting HisKA and receiver complexes adopt the same orientation, if the structures for these protein complexes were available they too could be superimposed on each other. With this image in mind, any pair of amino acid residues from this generic HisKA-receiver complex can be represented by the corresponding pair of MSA columns.

Only a minority of MSA column pairs will represent amino acids in close contact, forming part of the interaction surface. Amino acids making up the interaction site between the two protein families would be expected to exert mutual constraints, and thus to have co-evolved, manifesting as a correlation or co-variation between columns of paired MSAs. Figure 5.2 shows a schematic of the HisKA and receiver MSAs, sorted according to known domain pairings, and represents the possible amino acid interactions as a grid of inter-protein MSA column pairs.

Hydrophilicity is believed to play a role in TCS interactions (Kojetin *et al.*, 2003), and therefore correlations in hydrophilicity between MSA columns were calculated. Two alternative hydrophilicity scorings were used, those of Kyte-Doolittle (KD) (Kyte and Doolittle, 1982) and Hopp-Woods (HW) (Hopp and Woods, 1981). Although these give numerical values and therefore a standard Pearson correlation could be calculated, the scores are not normally distributed as they come from a small discrete set of amino acids. Instead two rank based correlations were used with tie corrections, Spearman's ρ and Kendall's τ (Kendall and Gibbons,

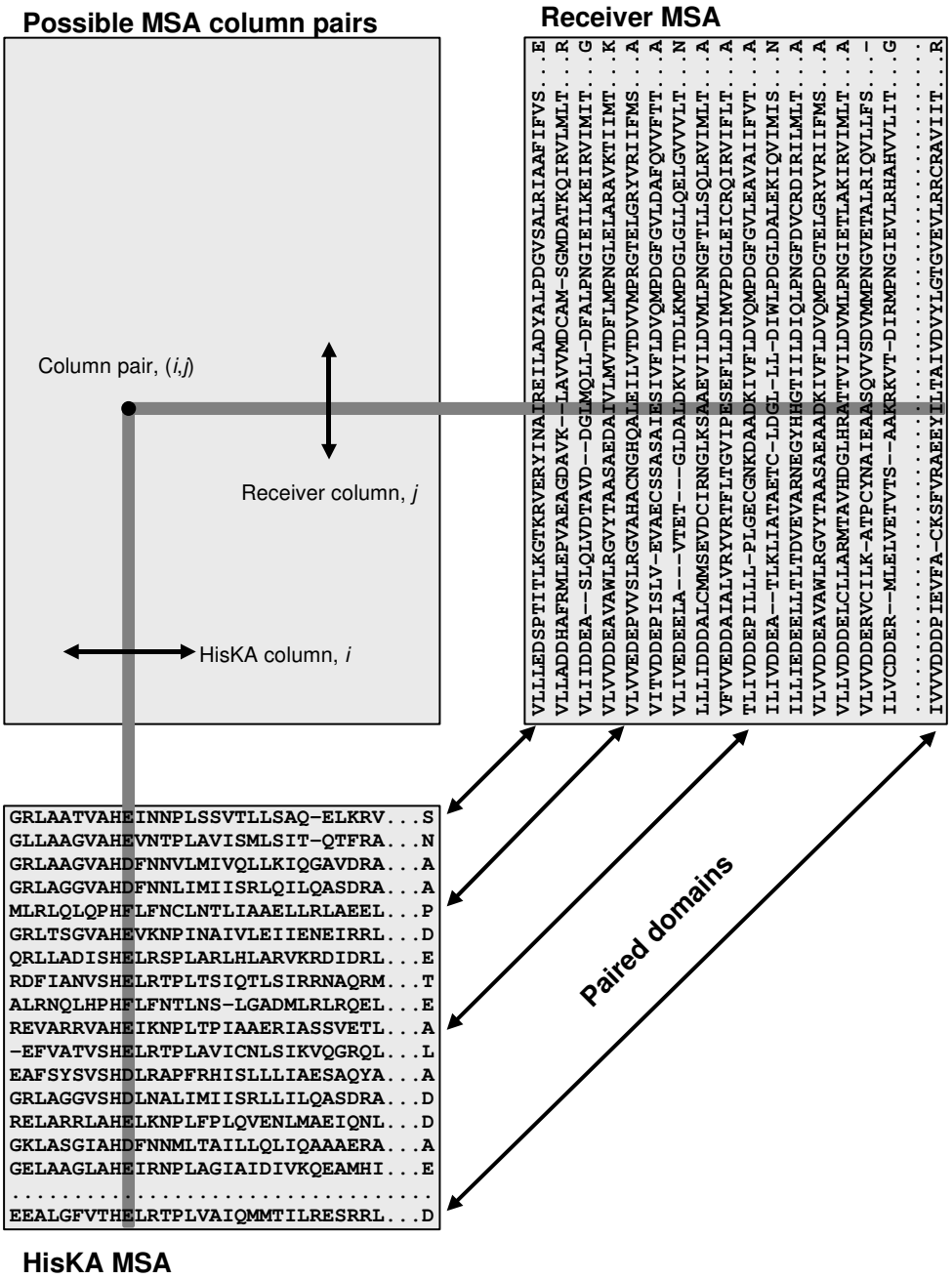


Figure 5.2: Schematic figure showing a possible pair of MSA columns (i, j) from the paired HisKA MSA column i and receiver domain MSA column j. In the example shown, this column pair defines a list of paired amino acids (E,I), (E,-), (D,L), (D,I), ..., (E,I).

1990).

Another MSA column pair score evaluated, dubbed CP, was calculated as the sum of the chemical potentials between each amino acid pair. This used an existing statistical chemical potential derived empirically from known protein complexes (Lu *et al.*, 2003), an approach building on earlier work (Tanaka and Scheraga, 1976; Miyazawa and Jernigan, 1985; Sippl, 1990; Moont *et al.*, 1999). This chemical potential, which is illustrated in Figure 5.3, assigns values ranging from -4.4 for the amino acid pair (Cys,Cys) to $+1.9$ for (Asp,Asp), with the mapping largely explained by hydrophilicity. Low potentials between close amino acids should indicate energetically favourable complexes, and thus may identify interacting complexes.

In addition to these scores based on the physical properties of the amino acids themselves, two scores based on the letters themselves were considered. The first of these was the χ^2 statistic, which measures the difference in distribution of amino acid pairs from that expected if the two MSA columns were independent. Finally, the mutual information (MI) between each MSA column pair was also calculated.

MSA column pairs with an extreme score (e.g. a high MI) may be indicative of co-evolution due to mutual constraint posited for amino acid pairs playing a role in TCS partner specificity. Grids of the MSA column pairs were plotted, with each square coloured according to the score (i.e. a coloured matrix, also termed a level-plot or heatmap), with the rows and columns labelled by the amino acid sequences of a reference HisKA and receiver domain. This allows any "hot-spots" to be interpreted in the context of known 3D structures and literature regarding the protein-protein interaction surface. These column pairs with extreme value scores were also visualised as lines between the HisKA and receiver reference sequences shown running across the top and bottom of the figure.

One key validation step is to perform the same calculations using randomised protein pairings, to give an estimate of the typical range of scores which could be expected from the natural variation of the protein sequence motifs concerned. Should the distribution of scores from the random pairings show little difference to those using the pairings inferred from the genome arrangement, that scoring system is clearly unsuitable for the intended purpose.

As there is not yet a solved 3D structure for the HisKA-receiver complex which could be used as a template (Section 1.5), the analogous Spo0B-Spo0F complex has been used as a model (e.g. Laub and Goulian (2007)). By superimposing a known HisKA structure onto the Spo0B dimer, a rough approximation of the expected HisKA-receiver complex was created. This allowed crude pairwise distances to be calculated, providing another way to evaluate the MSA column pairs selected by the different scoring systems.

5.4 Implementation

To reduce potential sampling bias, HisKA and receiver domain pairs were taken from the 340 representative species listed in bold in Table A.1. The alignments were generated using three different datasets, two-gene pairs ($3,473$ neighbouring HK and RR genes), HYs ($1,434$ single isolated genes containing both a T_i and R domain), and a combined dataset ($4,907$ pairs). These sequences were not filtered for redundancy based on sequence similarity.

Additional sequences consisting of the HisKA domain from *E. coli* EnvZ (PDB ref. 1JOY), or the receiver domain of *Bacillus subtilis* Spo0F (PDB ref. 1F51), listed below, were included in each MSA to provide a convenient reference point for describing the MSA columns, and for inferring distances as described later. However, for the calculation of the MSA column pair correlation scores, these two reference sequences were excluded.

```
>REF HisKA domain from 1JOY EnvZ
DRTLLMAGVSHDLRTPLTRIRLATEMMSEQDGYLAESINKDIEECNAIIEQFIDYLR

>REF Receiver domain from 1F51 Spo0F
KILIVDDQSGIRILLNEVFNKEGYQTFQAANGLQALDIVTKERPDLVLLDMKIPGMDGI
EILKRMKVIDENIRVIIMTAYGELDMIQESKELGALTHFAKPFDIDEIRDAVKKYLPL
```

The quality of the MSAs themselves was expected to play some role, and this was assessed by using two different alignment programs, CLUSTAL W (Thompson *et al.*, 1994) and MUSCLE (Edgar, 2004). Specifically, CLUSTAL W version 1.83 was used with its default settings, and MUSCLE version 3.7 with a maximum of three iterations (`-maxiter 3`), and the alignments output using the CLUSTAL W file format (`-clwstrict`).

When calculating scores using randomised pairings, a same sized set of HisKA-receiver pairings was generated from the alignments using sampling with replacement. This is equivalent to generating two new "paired" MSAs by copying rows (domain sequences) from the rows of the original MSAs selected at random. A bootstrapping procedure based on this procedure is discussed at the end of this chapter.

MSA columns where the fraction of gap characters exceeded a given threshold (50% unless otherwise stated) were excluded from the analysis, unless the column happened to include an amino acid from the reference sequence. The chemical potential between a gap character and an ordinary amino acid or another gap was taken as zero. Similarly, gaps were assigned a hydrophilicity score of zero before calculating the ρ or τ correlations. For the χ^2 and MI scores gap characters were treated as another letter.

A crude 3D structure for the HisKA-receiver complex was created by a three-dimensional alignment of the *E. coli* EnvZ HisKA dimer (PDB 1JOY, Figure 1.20) onto the Spo0B dimer in the *Bacillus subtilis* Spo0B-Spo0F complex (PDB 1F51, Figure 1.22), based on the structural analogy of their four α-helix bundles shown in Figure 1.19. Specifically, a rigid body motion mapping one dimer onto the other was selected using singular value decomposition to minimise the least squares Euclidean distance between the C_α atoms[1] of the conserved hisitidine phosphorylation sites and the three amino acids either side of them (within the same α-helices). This ensured the histidine phosporoylation sites of EnvZ and Spo0B were co-located and that the orientation of the α-helix bundles was consistent. Note that structural alignment is in general much more complicated, see for example Taylor and Orengo (1989) and Levitt and Gerstein (1998).

Since the *E. coli* EnvZ and *Bacillus subtilis* Spo0F domain sequences were included in the HisKA and receiver MSAs, using this crude HisKA-receiver complex C_α distancescould straightforwardly be assigned to most MSA column pairs. For columns in the MSA where a gap had been introduced in the reference sequences, a simple linear interpolation between the C_α atoms of the residues either side of the gap was used to assign a notional set of coordinates to this MSA column, and thus pairwise distances.

The whole analysis was scripted in python (www.python.org). MSA files were loaded using the Biopython libraries (www.biopython.org). The `Bio.PDB` module in Biopython (Hamelryck and Manderick, 2003) was used to load the PDB files and calculate the structural alignment. Spearman's ρ and Kendall's τ correlations were calculated using the `Bio.Cluster` module in Biopython (de Hoon *et al.*, 2004), while the χ^2 and MI calculations were implemented by hand in python.

The figures were drawn using the python library ReportLab (www.reportlab.org), or R (R Development Core Team, 2007) invoked from python via RPy (Moreira and Warnes, 2003). In particular, the smoothed scatter plots were drawn using the R function `smoothScatter` from the Bioconductor package (Gentleman *et al.*, 2004). This uses a kernel density estimate, where each scatter point is replaced with a "smoothed out" kernel using a standard two-dimensional normal or Gaussian distribution. The plot area is divided using a fine grid, and each grid square is coloured according to the sum of these values, giving an image resembling a contour plot.

[1]The C_α atom is the protein back bone carbon atom to which the amino acid residue side chain is attached.

5.5 Column pair correlations and results

In this section, for each scoring system discussed, a similar set of figures is presented. Firstly, the distribution of the observed scores is shown together with that given by a random set of protein pairings. This also serves as a colour key for a following figure, where the score of each MSA column pair is shown on a coloured grid, with the receiver residues shown horizontally and the HisKA residues vertically, with the axes labelled with the reference structures (including their secondary structure using dark and light greys for the α-helices and β-sheets). A further figure (or pair of figures) shows the MSA column pairs giving extreme scores as lines connecting those residues on the two reference sequences, drawn horizontally across the top and bottom (again with their secondary structure indicated). Finally smoothed scatter plots have been used to show how the correlation scores compare to the inferred distance from the crude protein complex.

Unless otherwise stated, these figures are from CLUSTAL W alignments from HisKA and receiver domains found in neighbouring HK and RR genes (i.e. not HYs).

5.5.1 Chemical potential summations

The simplest MSA column pair score considered was the summation of the Lu *et al.* (2003) chemical potential of each amino acid pair (illustrated in Figure 5.3). The resulting distribution of scores is shown in Figure 5.4, where it is clear that there is no difference in pattern between those from the identified domain pairs, and a control set of random domain pairings. This indicates that this approach has not identified any MSA column pairs linked to the domain pairings. Figure 5.5 confirms this by showing these scores as a grid, where there is no spatial patterning of interest. Analysis of the MUSCLE alignments or the alternative datasets showed no difference (data not shown).

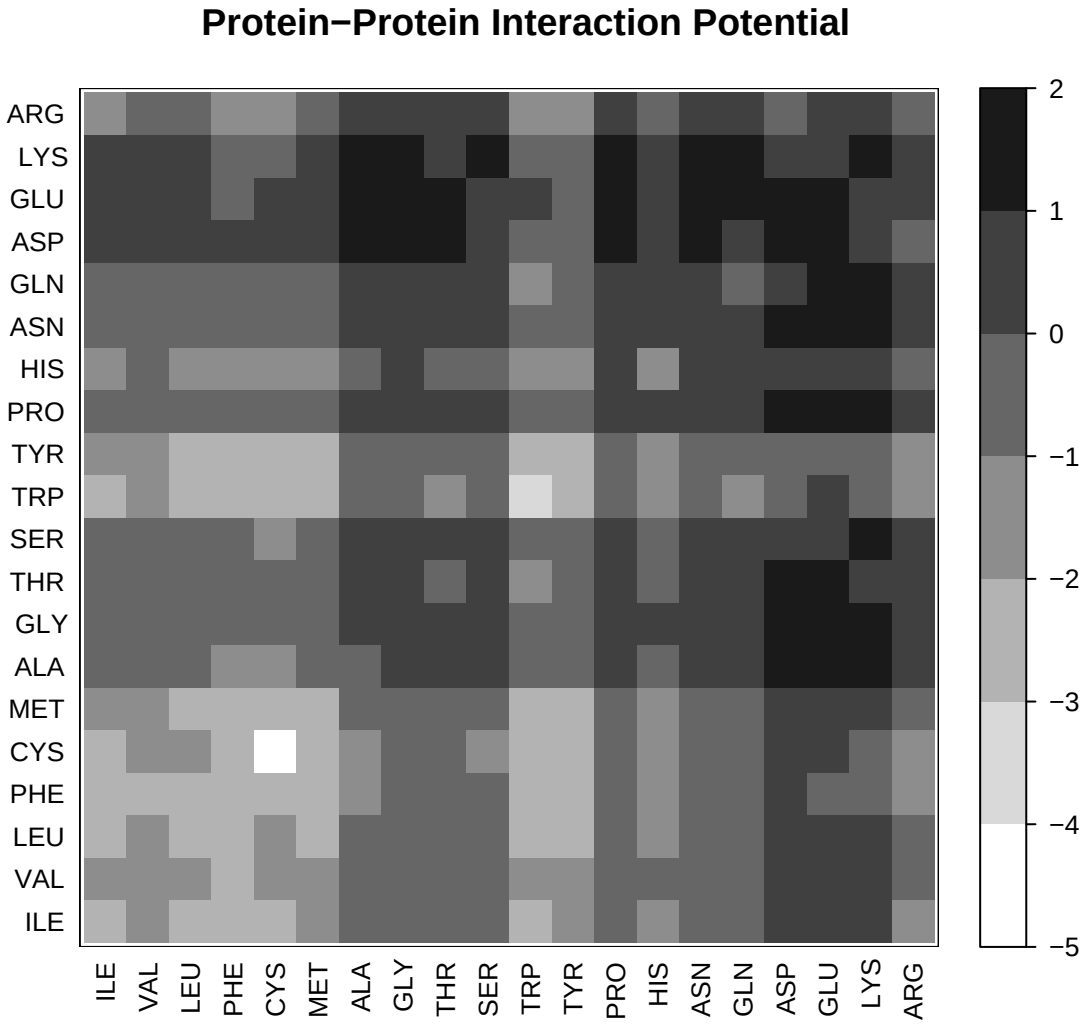


Figure 5.3: The chemical potential from Lu *et al.* (2003), shown as a coloured grid where the potential for each amino acid combination is indicated by the brightness (key on right). The potential assigned to each amino acid pair ranges from -4.4 for (Cys,Cys) to $+1.9$ for (Asp,Asp), with the mapping largely explained by hydrophilicity.

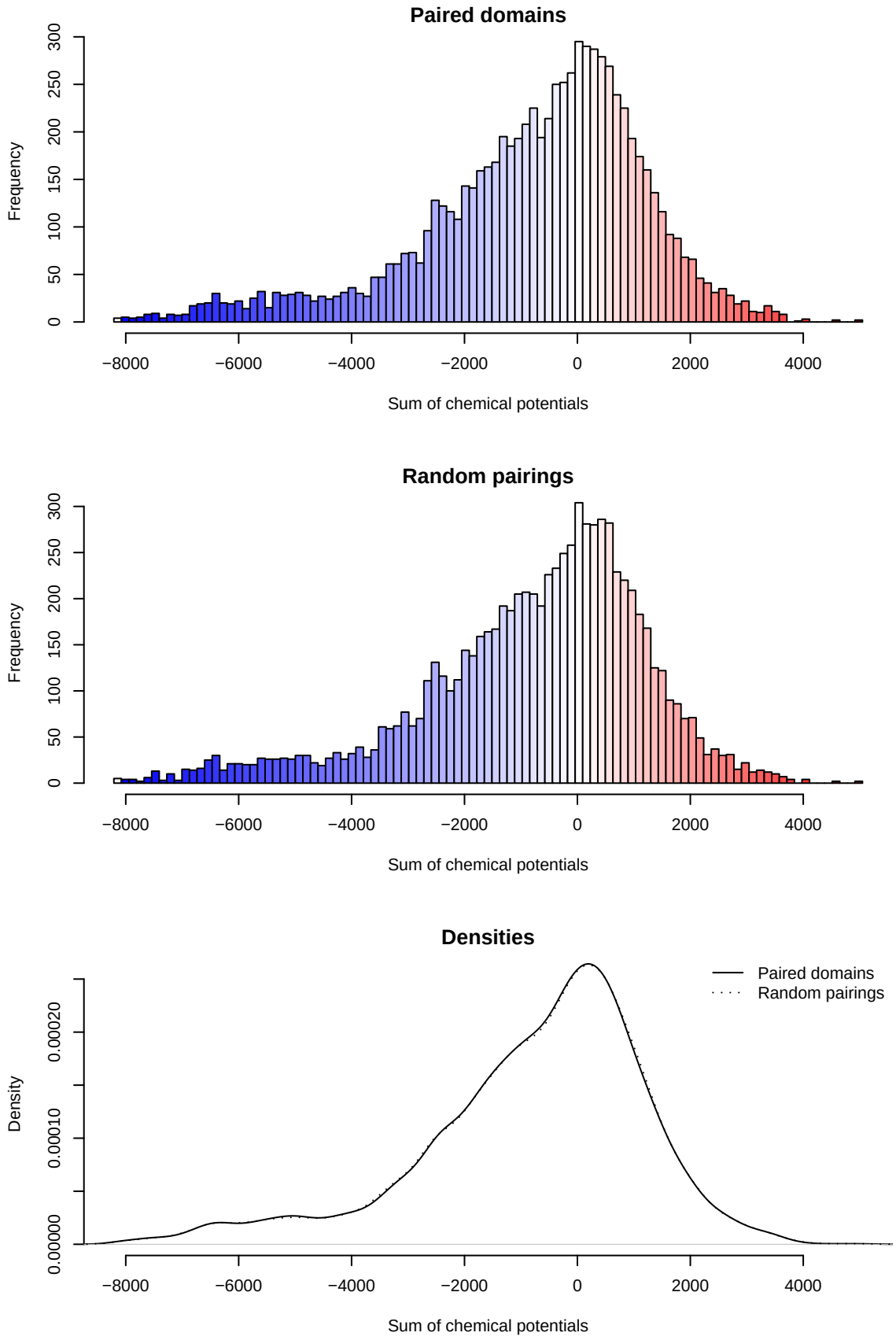


Figure 5.4: Histogram of chemical potential summations (CP score) from HisKA and receiver pairs (top), and randomised pairings (middle). These are shown as two almost identical density curves (bottom).

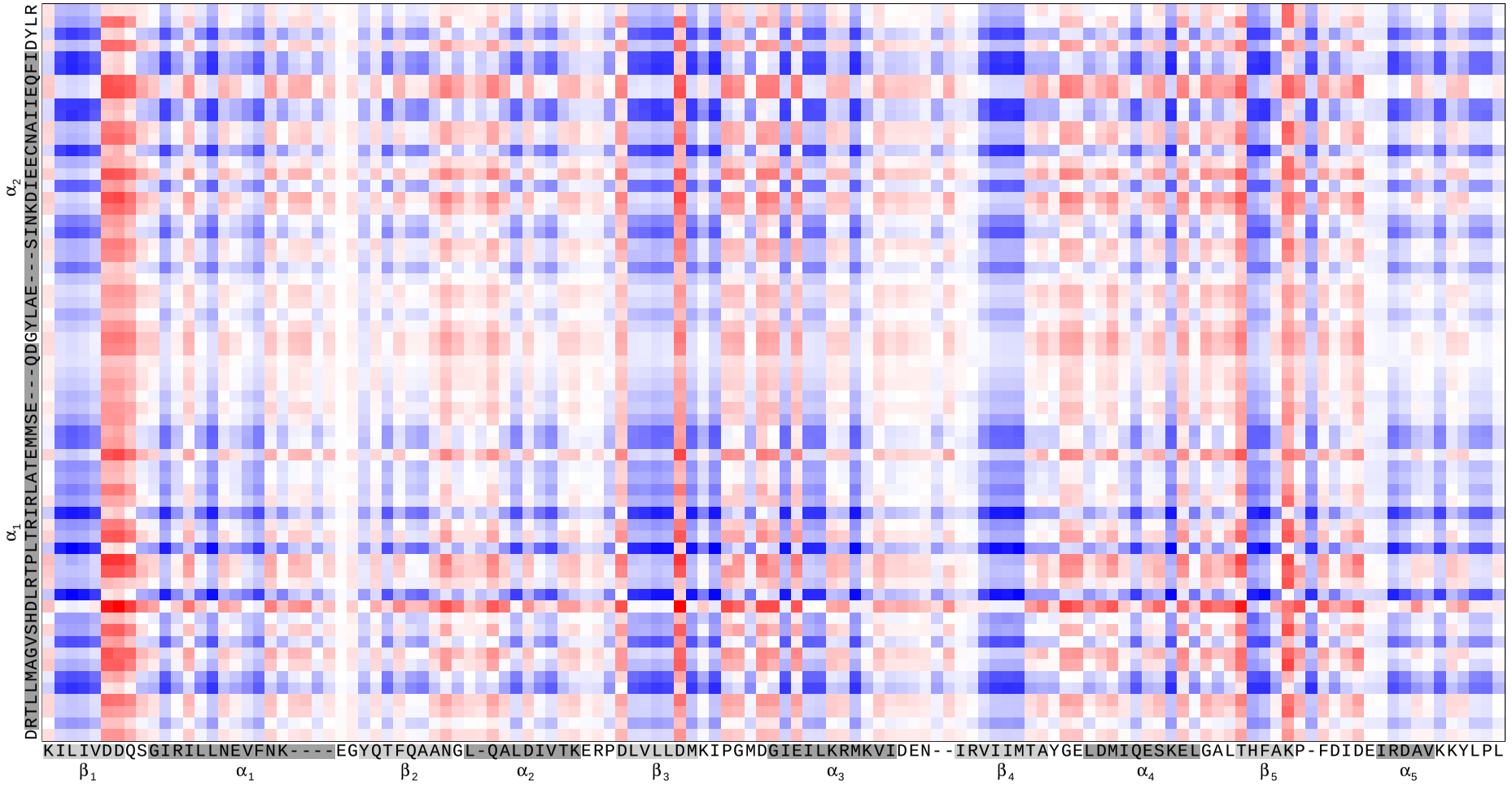


Figure 5.5: Grid of co-varying amino acid positions of HisKA (vertical) and receiver (horizontal) domains, for two gene domain pairs. The vertical axis corresponds to columns of the HisKA MSA, and the horizontal to columns of the receiver MSA. Each axis is labelled with the associated reference protein and its secondary structure. Each grid square corresponds to a column pairing, coloured according to the sum of the chemical potentials (CP score). Negative totals are blue, positive totals are red. See Figure 5.4 for the colour key.

5.5.2 Hydrophilicity correlations

Using a mapping such as the KD or HW hydrophilicity (Table 5.1), the list of amino acids in any MSA column can be translated into a list of numbers. For a pair of columns from the HisKA and receiver MSAs, this gives a set of paired hydrophilicity scores for which a correlation can be calculated. This could be a simple linear regression (treating the hydrophilicities as a continuous variable), such as a Pearson correlation. However, as there are only twenty amino acids, and some of them have the same hydrophilicity, this is actually a discrete problem. It is therefore more appropriate to use a rank based correlation such as Spearman's ρ or Kendall's τ, both of which include a tie correction.

Amino acid		KD	HW	Amino acid		KD	HW
A	Ala	-1.80	-0.50	N	Asn	3.50	0.20
C	Cys	-2.50	-1.00	P	Pro	1.60	0.00
D	Asp	3.50	3.00	Q	Gln	3.50	0.20
E	Glu	3.50	3.00	R	Arg	4.50	3.00
F	Phe	-2.80	-2.50	S	Ser	0.80	0.30
G	Gly	0.40	0.00	T	Thr	0.70	-0.40
H	His	3.20	-0.50	V	Val	-4.20	-1.50
I	Ile	-4.50	-1.80	W	Trp	0.90	-3.40
K	Lys	3.90	3.00	Y	Tyr	1.30	-2.30
L	Leu	-3.80	-1.80	X	Xxx	0.00	0.00
M	Met	-1.90	-1.30	-	Gap	0.00	0.00

Table 5.1: KD and HW hydrophilicity/hydrophobicity scores. The final rows show the unknown amino acid X and gap character, for which an arbitrary value of zero was typically used.

Given lists of n paired values (here hydrophilicities from two MSA columns), denote their ranks by x_i and y_i (ranging from 1 to n) for $i = 1, \ldots, n$. Tied elements are assigned the mean of the ranks they would otherwise be given. Spearman's ρ is defined as follows where it is useful to introduce $N' := n(n^2-1)/6$,

$$\rho := 1 - \frac{6\sum_i \left[(x_i - y_i)^2\right]}{n(n^2-1)} = \frac{N' - \sum_i \left[(x_i - y_i)^2\right]}{N'}. \tag{5.1}$$

Kendall and Gibbons (1990) then gives a tie corrected form of Spearman's ρ,

$$\rho := \frac{N' - \sum_i \left[(x_i - y_i)^2\right] - U' - V'}{\sqrt{N' - 2U'}\sqrt{N' - 2V'}}, \tag{5.2}$$

where $U' := \sum(u^3 - u)/12$ and $V' := \sum(v^3 - v)/12$ are sums over the observed ranks with u and v the number of elements with each rank in lists x_i and y_i respectively. In the absence of ties, $U' = V' = 0$ and this reduces to Equation (5.1).

Kendall's τ works by looking at the $\frac{1}{2}n(n-1)$ pairs of rank entries (i and j), counting concordant pairs which have the same rank orders (i.e. $x_i < y_i$ and $x_j < y_j$, or $x_i > y_i$ and

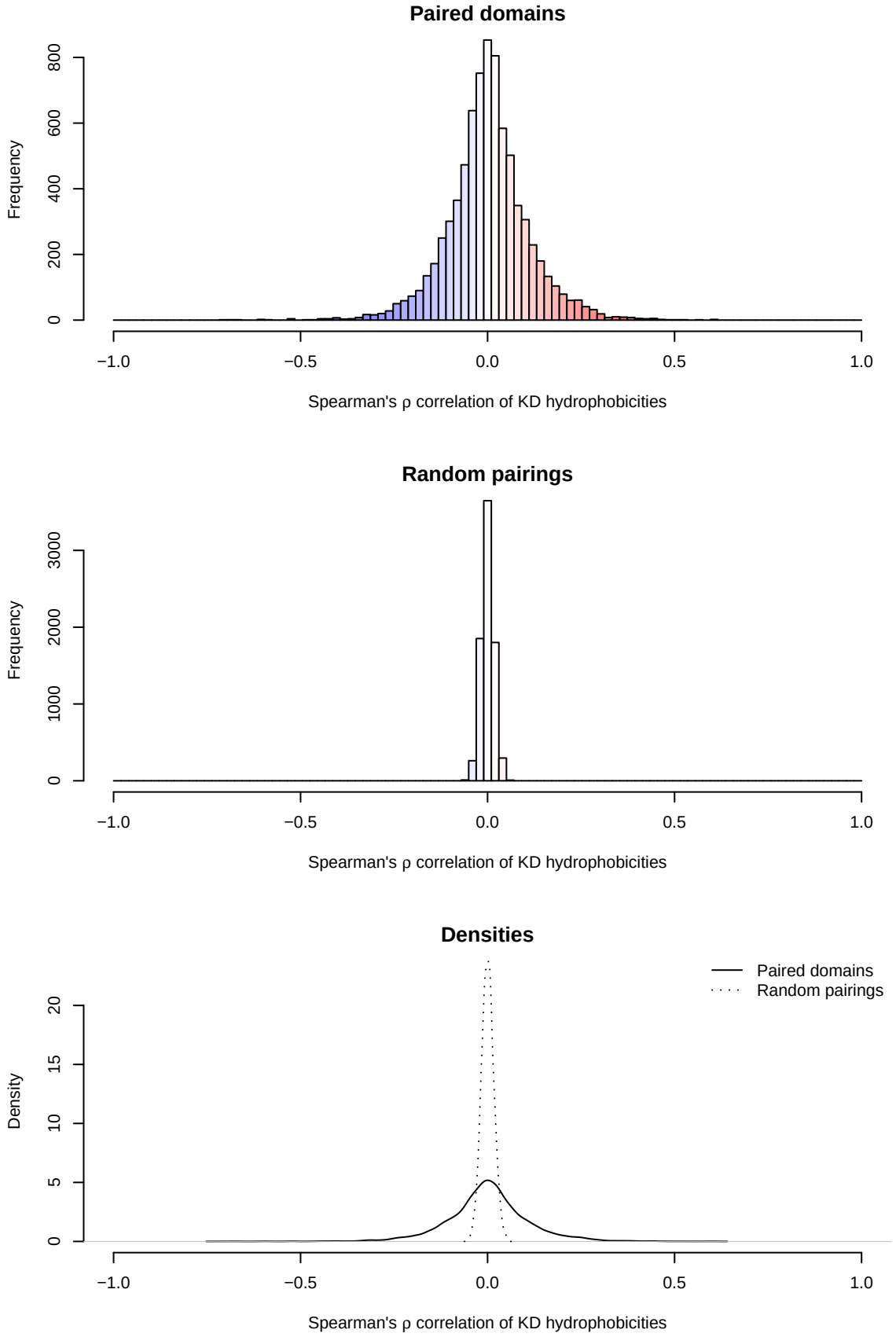


Figure 5.6: Histogram of KD Spearman's ρ from HisKA and receiver pairs (top), and randomised pairings (middle). These are shown as density curves (bottom).

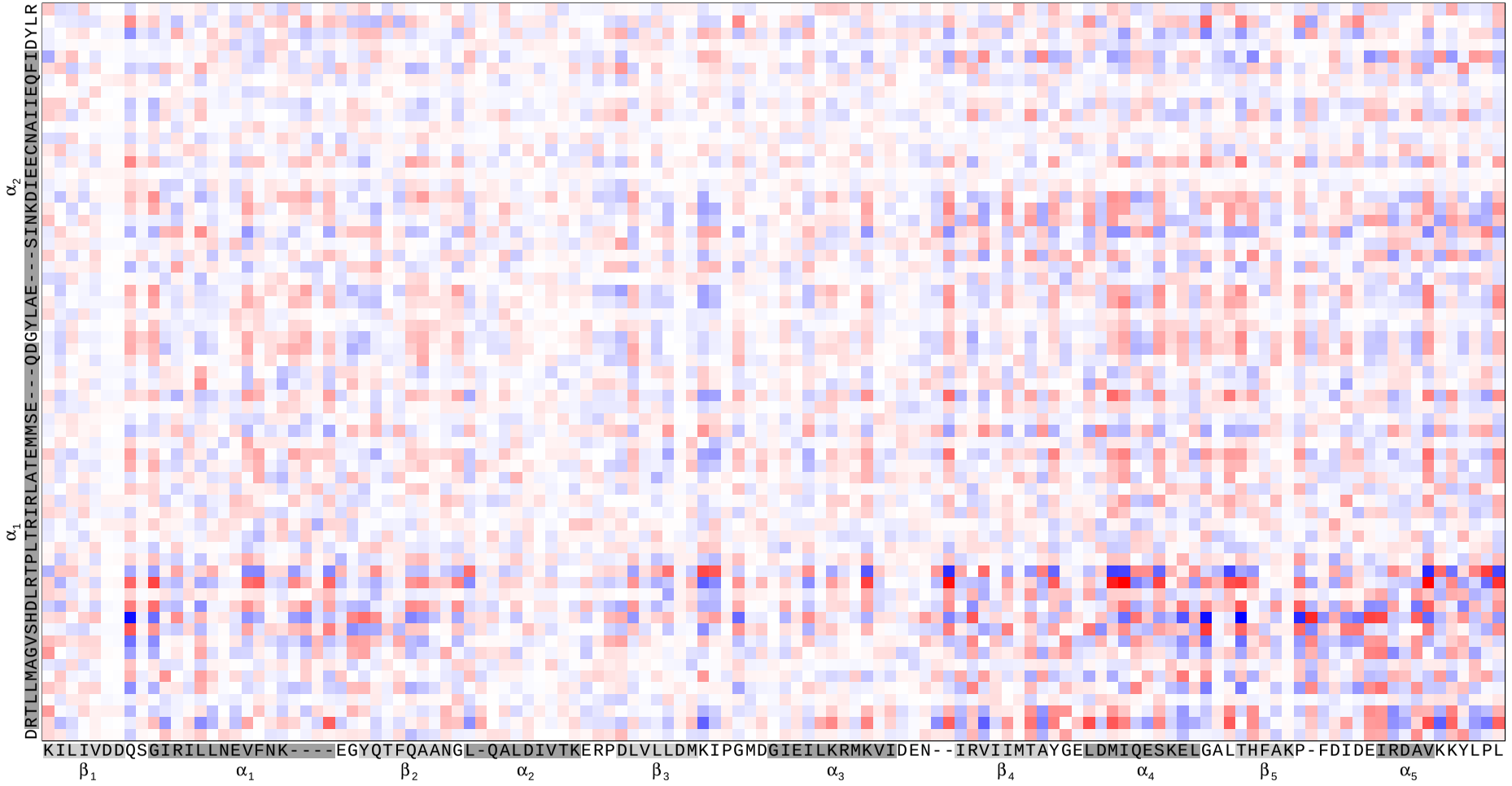


Figure 5.7: Grid of co-varying amino acid positions of HisKA (vertical) and receiver (horizontal) domains, for two gene domain pairs. Each grid square corresponds to a column pairing between the HisKA MSA and the receiver MSA, coloured according to the KD Spearman's ρ correlation. Negative correlations are blue, positive correlations are red. See Figure 5.6 for the colour key.

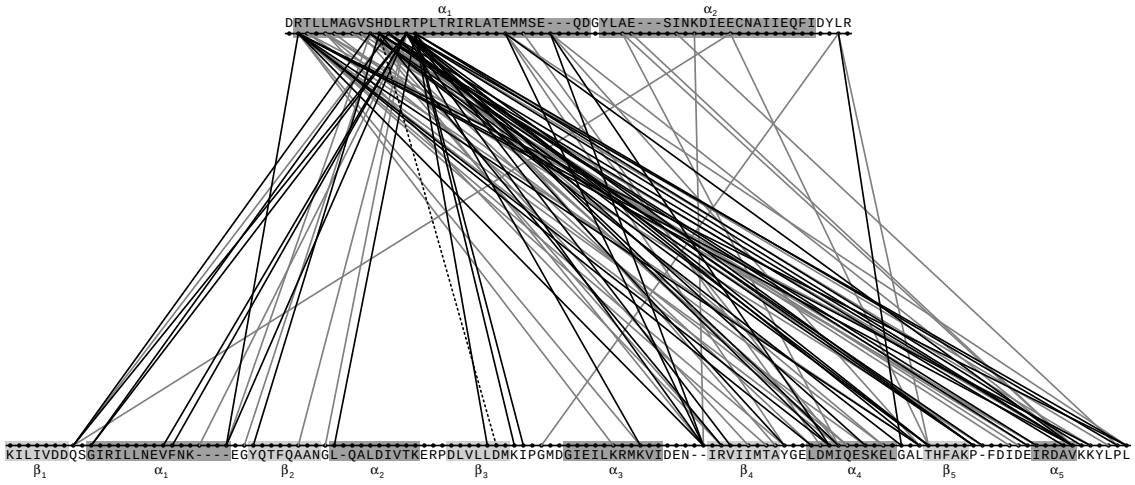


Figure 5.8: Top 100 MSA column pairs with positive KD Spearman's ρ correlations.

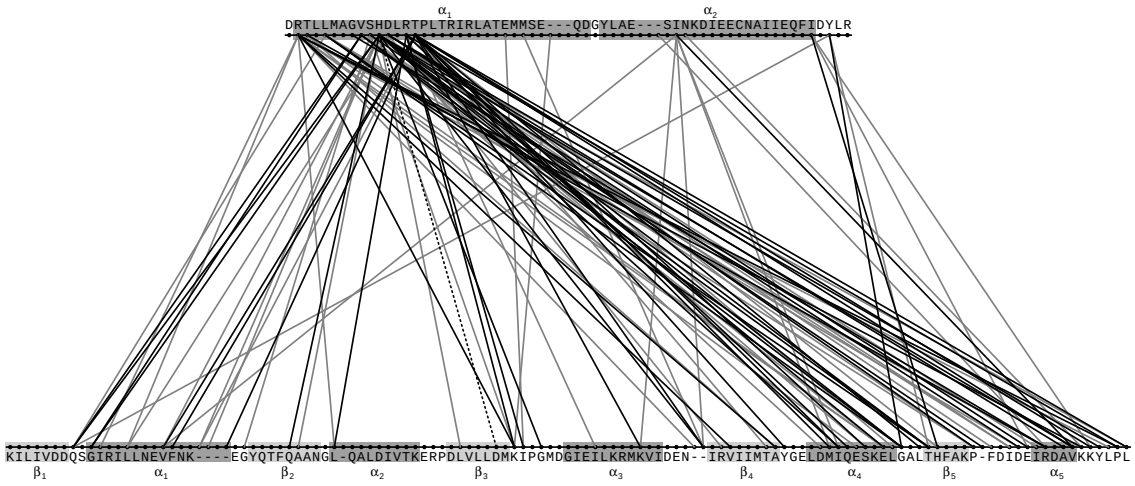


Figure 5.9: Top 100 MSA column pairs with negative KD Spearman's ρ correlations.

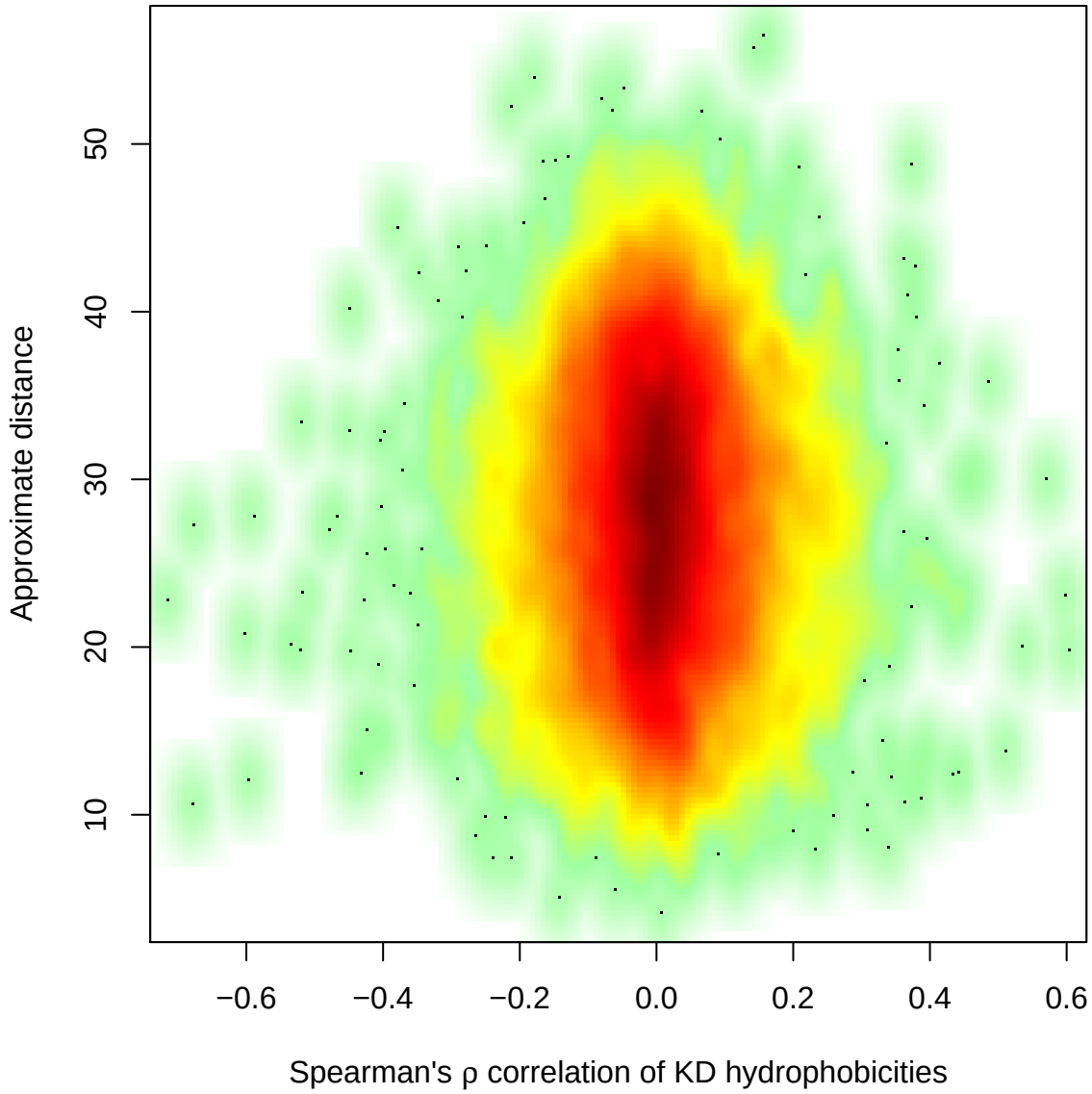


Figure 5.10: Smoothed scatter plot of KD Spearman's ρ correlations against estimated distances from a crude protein-protein complex. Rather than attempting to show a scatter plot with tens of thousands of points, this figure (and later similar plots) indicate the local smoothed kernel density estimate by color, ranging from white to green to yellow to red to brown. Only one hundred outliers are shown as points (in the green regions).

$x_j > y_j$), and discordant pairs with opposite rank orders (e.g. $x_i < y_i$ and $x_j > y_j$). Writing P for the number of concordant pairs and Q for the number of discordant pairs, Kendall and Gibbons (1990) defines

$$\tau := \frac{2(P-Q)}{n(n-1)}. \tag{5.3}$$

Correction factors $U := \sum u(u-1)$ and $V := \sum v(v-1)$, summed over the observed ranks, are included in the tie corrected variant,

$$\tau := \frac{2(P-Q)}{\sqrt{n(n-1)-U}\sqrt{n(n-1)-V}}. \tag{5.4}$$

As before, in the absence of ties the correction factors vanish ($U = V = 0$) giving the uncorrected Equation (5.3). In this chapter the tie corrected forms of ρ and τ are used exclusively.

Figure 5.6 shows the distribution of KD Spearman's ρ correlations for the paired proteins, and as a control for randomised domain pairings. The distributions have similar bell shaped distributions centred at the origin. The paired domains show a much broader range of ρ correlations, suggesting those column pairs with extreme ρ correlation scores may be important for the protein-protein interaction specificity.

Inspection of a grid of these correlation scores (Figure 5.7) shows many of these column pairs with extreme ρ correlations are associated with the first α-helix of the HisKA (bright red or blue against a white background where $\rho \sim 0$). This is also apparent in Figures 5.8 and 5.9 which show the column pairs with the highest or lowest 100 KD ρ correlations.

Figure 5.10 shows these KD ρ correlations plotted against the estimated separation of the associated amino acids in the reference protein complex. Overall there is no correlation, although by eye one might argue that the column pairs with the most extreme correlation scores are slightly closer together than average (in that the top left and top right corners of the plot are empty). Although not shown in full, plots using the HW hydrophilicity values with Spearman's ρ yield much the same results as those discussed above using the KD hydrophilicity scale (see Figure 5.16, described below).

Kendall's τ is an alternative rank-based correlation. Figure 5.11 shows the distribution of KD τ correlations for the paired proteins and the randomised pairings. Both distributions are symmetric about the origin, however the paired data clearly shows more extreme correlations, both positive and negative. As with the ρ results, the fact that there is a noticeable difference between these two distributions is encouraging if these correlations are capturing something of the interaction information.

Figure 5.12 shows the KD τ correlations as a grid, where the MSA column pairs with extreme positive (or negative) correlations are shown in bright red (or blue). Since most

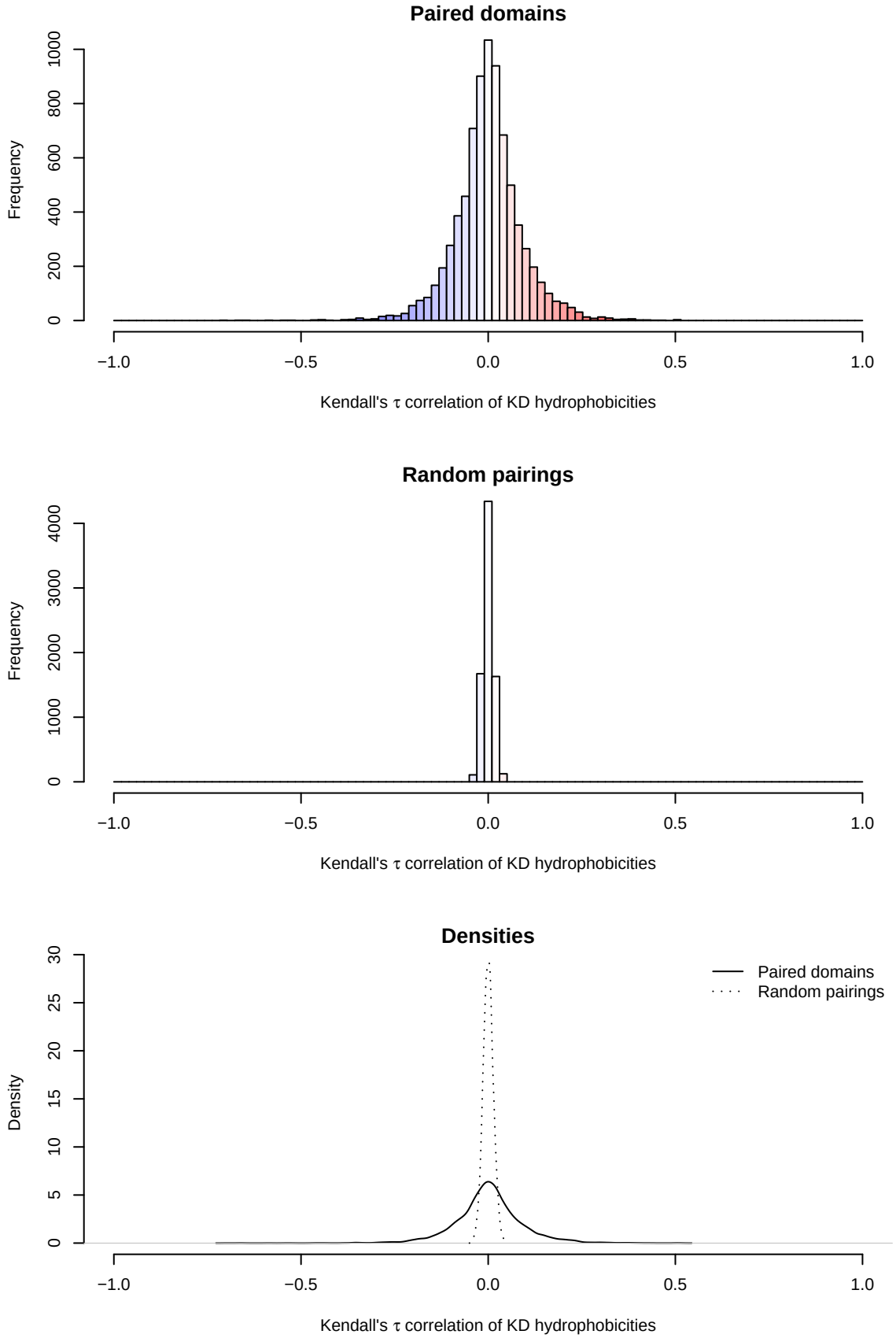


Figure 5.11: Histogram of KD Kendall's τ from HisKA and receiver pairs (top), and randomised pairings (middle). These are shown as density curves (bottom).

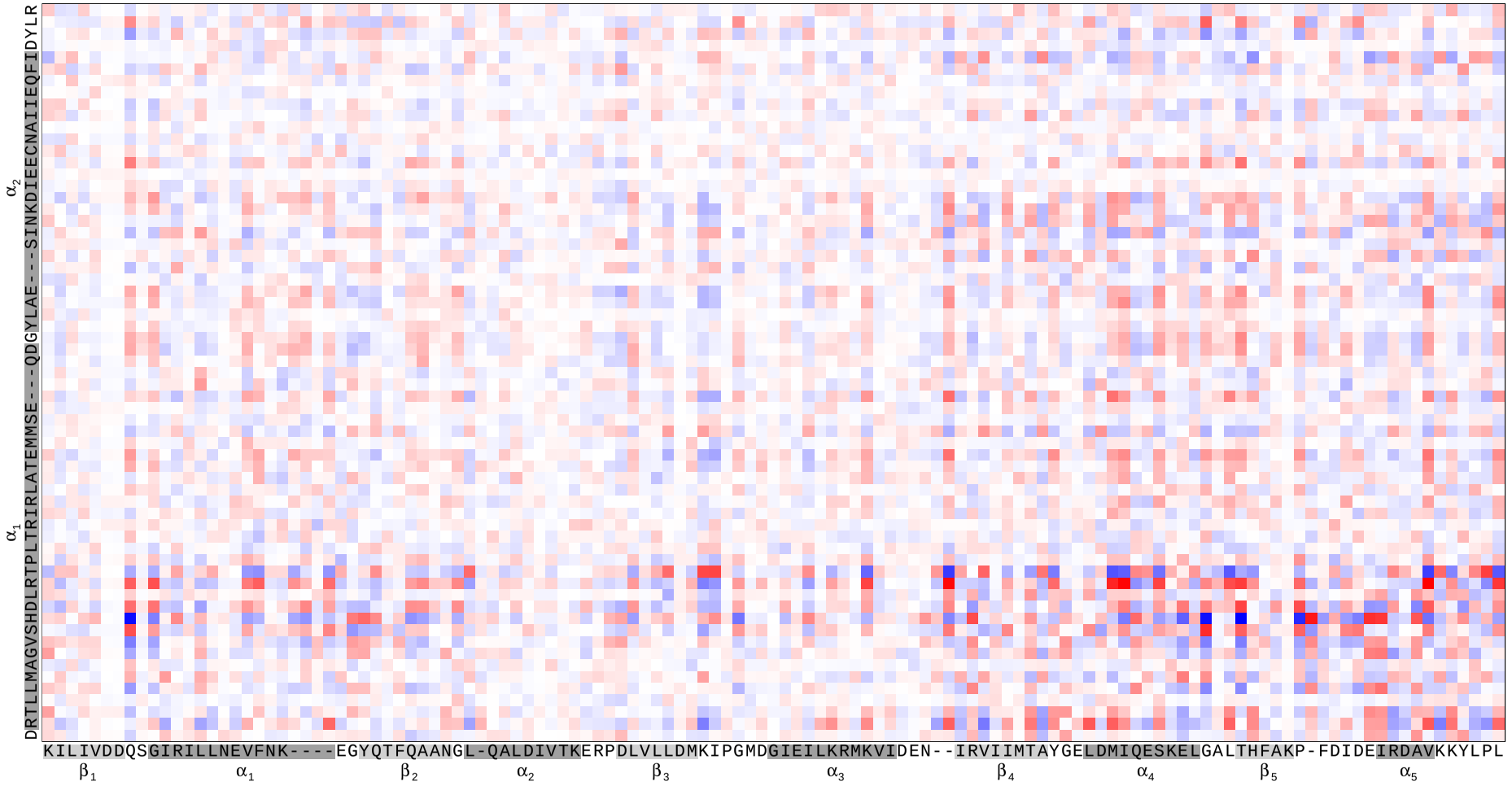


Figure 5.12: Grid of co-varying amino acid positions of HisKA (vertical) and receiver (horizontal) domains, for two gene domain pairs. Each grid square corresponds to a column pairing between the HisKA MSA and the receiver MSA, coloured according to the KD Kendall's τ correlation. Negative correlations are blue, positive correlations are red. See Figure 5.11 for the colour key.

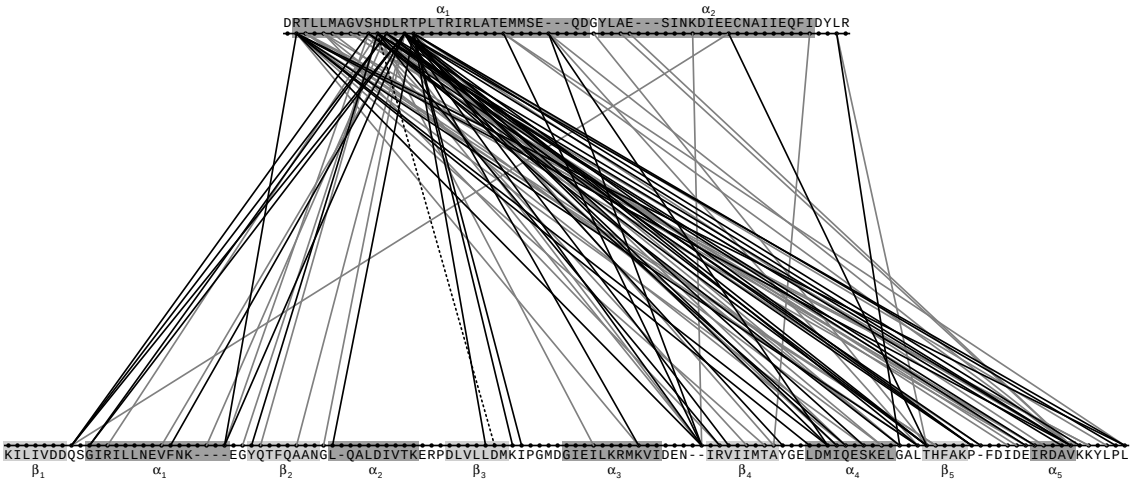


Figure 5.13: Top 100 MSA column pairs with positive KD Kendall's τ correlations.

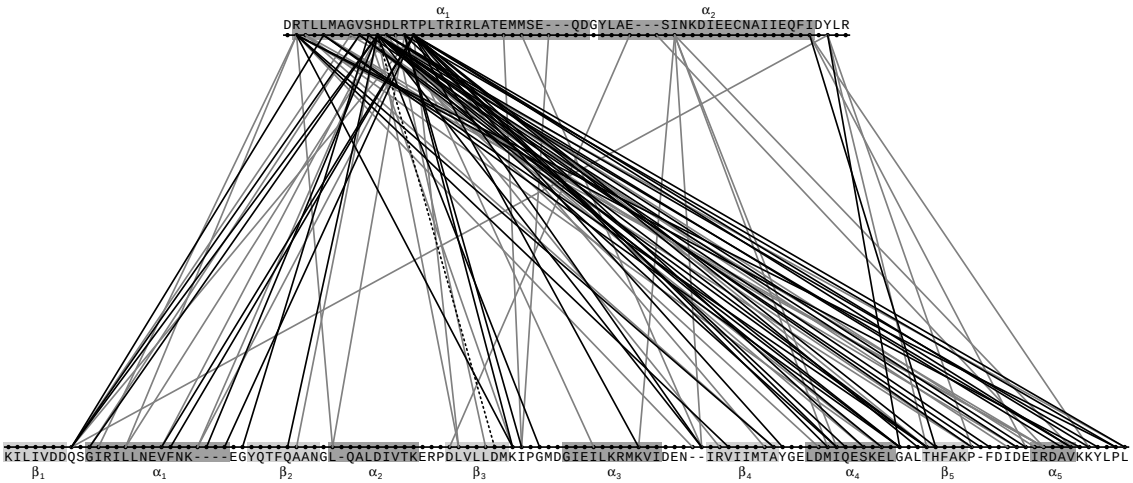


Figure 5.14: Top 100 MSA column pairs with negative KD Kendall's τ correlations.

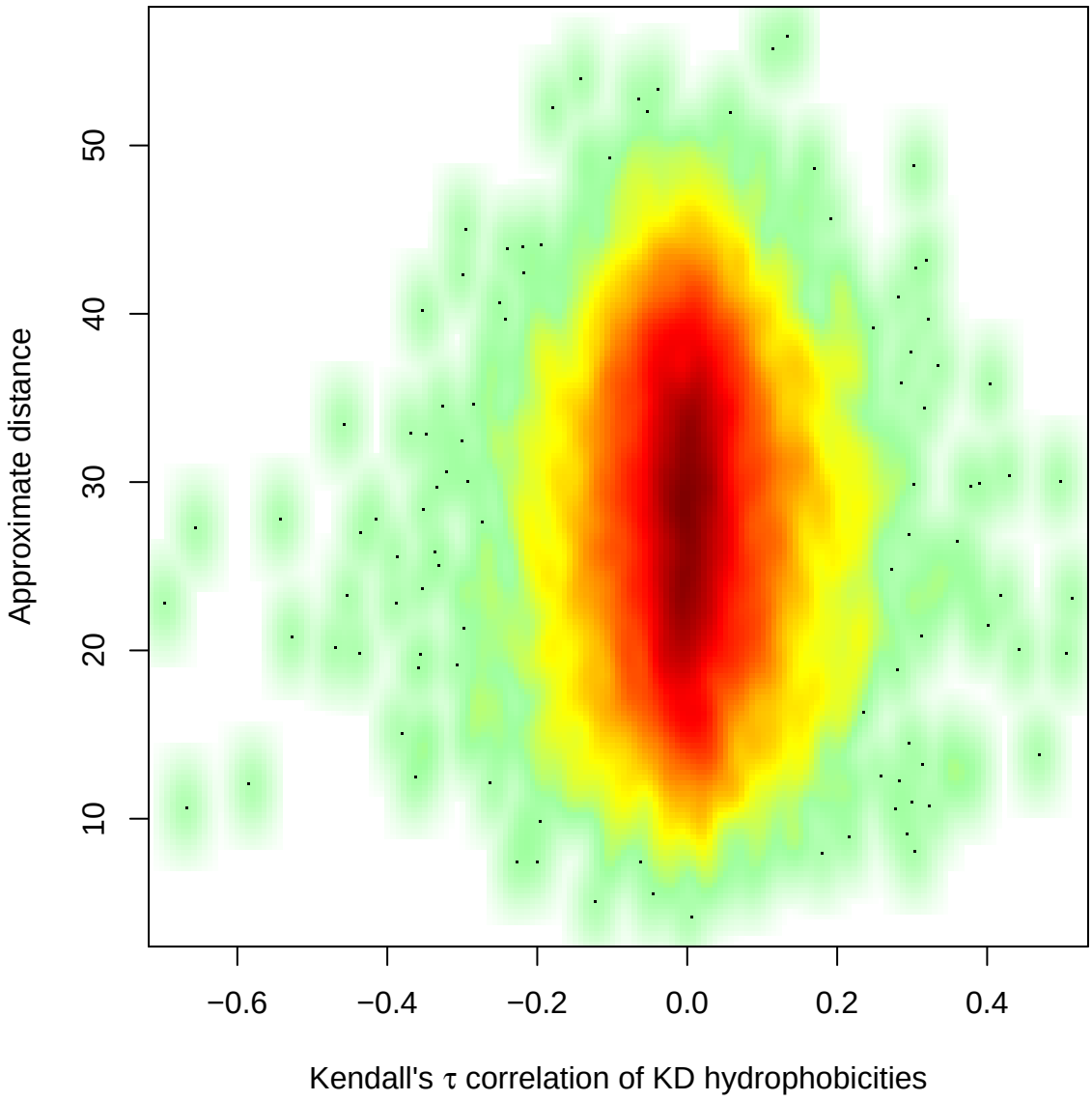


Figure 5.15: Smoothed scatter plot of KD Kendall's τ correlations against estimated distances from a crude protein-protein complex. Those column pairs with an extreme Kendall's τ are slightly closer together than average based on the estimate distances.

column pairs have $\tau \sim 0$, the majority of this grid is white. As with KD ρ, there are a number of extreme values, particularly associated with the first α-helix of the HisKA. These extreme values are also illustrated in Figures 5.13 and 5.14, where it can be observed that many of the receiver residues selected are on the interaction face of the protein.

Figure 5.15 plots the KD τ correlation scores against the residue separation from the inferred protein-protein structure. Again one might argue that those column pairs with extreme τ values do appear to be slightly closer together than average, but even so, at best this is only a slight improvement over KD ρ (Figure 5.10).

As with Spearman's ρ, a repeat analysis using the HD hydrophilicity rather than the KD hydrophilicity gave very similar distributions of Kendall's τ correlations, although the precise MSA column pairs highlighted do differ (data not shown in full). These alternative results are summarised in Figure 5.16, which shows a number of smoothed scatter plots comparing the KD and HW hydrophilicity scores, Spearman's ρ and Kendall's τ, against the estimated residue separation.

Kendall's τ is much more computationally expensive than Spearman's ρ, but under the following circumstances it proved more robust. In calculating the results described above, any gap characters in the MSA (and the few unknown amino acids recorded as X) were given a zero score. One alternative explored was to assign a null or NA value, and exclude such pairs from the correlations. This lead to the score for gap-rich MSA column pairs being determined by the minority of non-gapped residues, which often gave a spuriously high ρ correlation, as shown in Figure 5.17. Interestingly, this artifact was not seen when this gap handling approach was used with the Kendall's τ correlation, where the distribution remained symmetrical about zero.

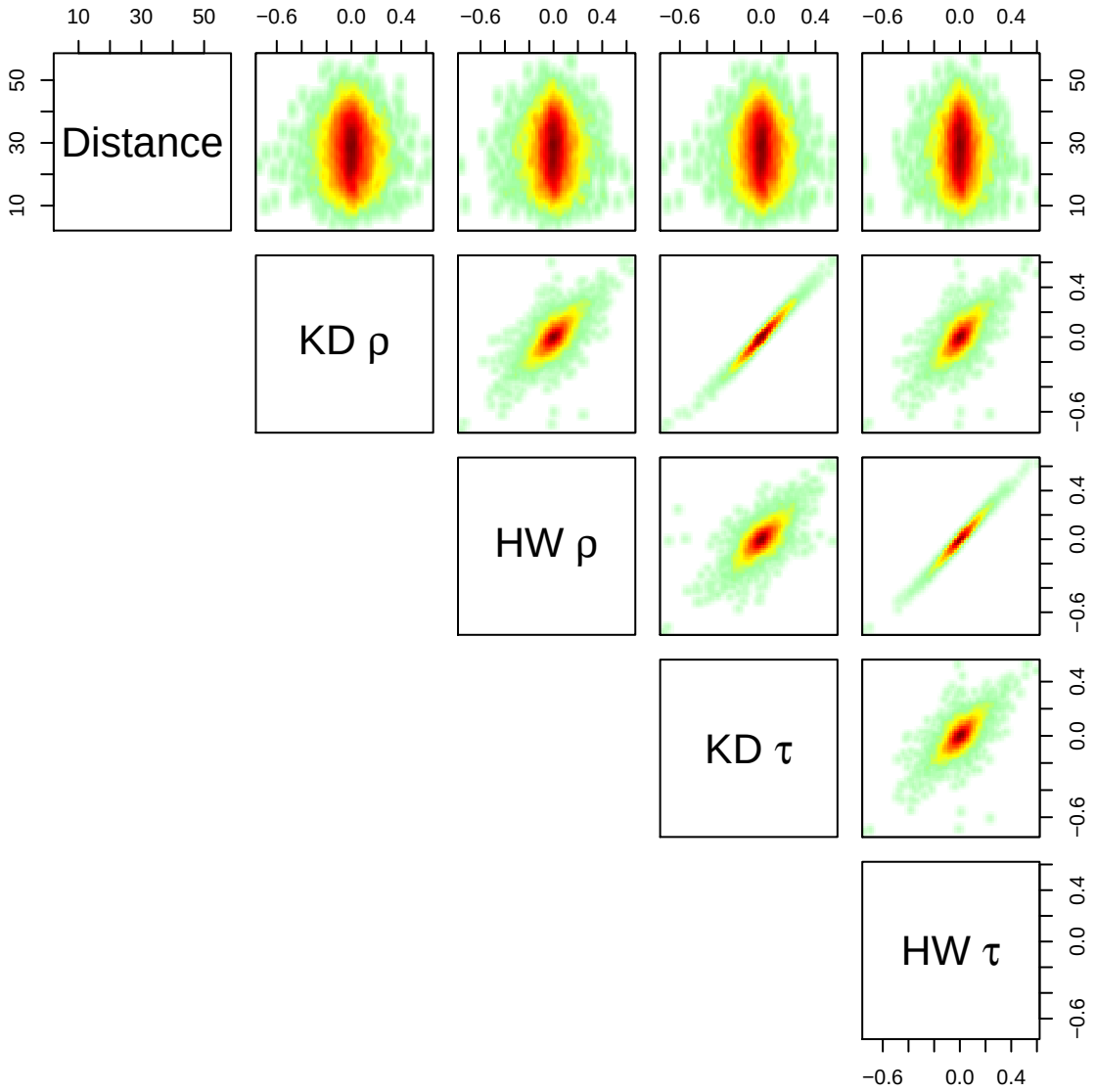


Figure 5.16: Assorted smoothed scatter plots for the hydrophilicity based scores and the estimated distances from a crude protein-protein complex. The captions along the diagonal indicate which scoring method is shown on the associated row/column. For example, the top row shows smoothed scatter plots of the estimated distances against the KD ρ, HW ρ, KD τ and HW τ correlations (cf. Figures 5.15 and 5.10). The KD ρ and KD τ scores correlate extremely well with each other, as do the HW ρ and HW τ scores.

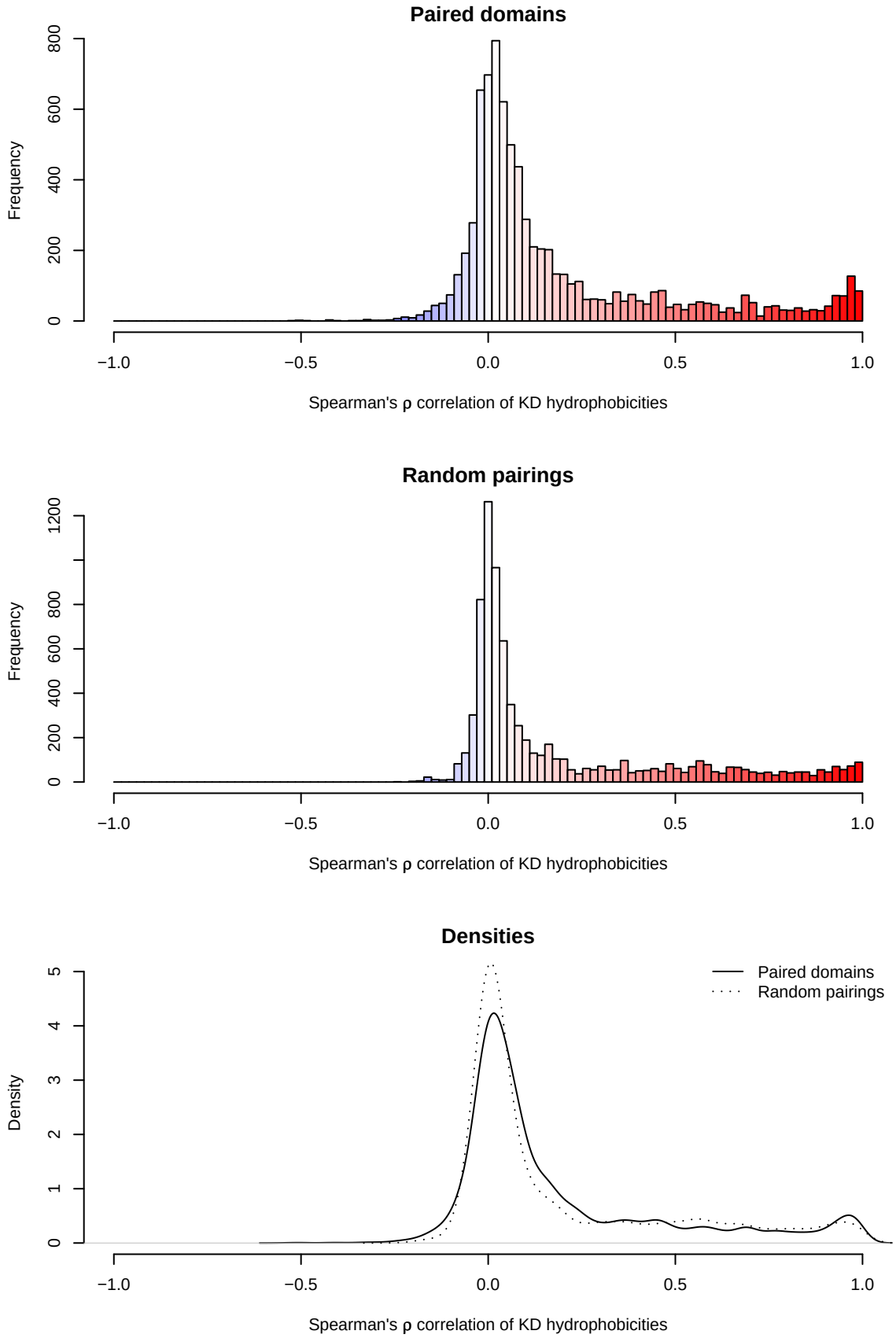


Figure 5.17: Histogram of KD Spearman's ρ from HisKA and receiver pairs (top), and randomised pairings (middle). These are shown as density curves (bottom). In Figure 5.6, gap and unknown residues were given a score of zero. Here, they are instead excluded from the correlation, which is then calculated from the remaining residue pairs. This gives rise to a spurious set of gap-rich column pairs with $\rho \approx +1$.

5.5.3 Chi-squared score

A given MSA column pair can be viewed as a list of paired amino acids (one from the HisKA, and one from the partner receiver). A simple contingency table can tabulate the observed amino acid pairings. The χ^2 (chi-squared) statistic is a measure of how the observed amino acid pairs compare to the expected distributions if the HisKA and receiver columns were independent.

Let vectors $\vec{A}$ and $\vec{B}$ of length l represent the HisKA and receiver MSA columns under consideration. For any letters α in $\vec{A}$ and β in $\vec{B}$, define the observed and expected distributions of amino acid pairs as follows:

$$\begin{aligned} \text{Obsv}_{\vec{A},\vec{B}}(\alpha,\beta) &:= \text{Count of } (\alpha,\beta) \text{ in } (\vec{A},\vec{B}) \\ \text{Obsv}_{\vec{A}}(\alpha) &:= \text{Count of } \alpha \text{ in } \vec{A} \\ \text{Obsv}_{\vec{B}}(\beta) &:= \text{Count of } \beta \text{ in } \vec{B} \\ \text{Expt}_{\vec{A},\vec{B}}(\alpha,\beta) &:= \tfrac{1}{l} \times \text{Obsv}_{\vec{A}}(\alpha) \times \text{Obsv}_{\vec{B}}(\beta) \end{aligned} \tag{5.5}$$

The chi-squared statistic χ^2 is then given by:

$$\chi^2(\vec{A},\vec{B}) := \sum_{\alpha,\beta} \frac{[\text{Expt}_{\vec{A},\vec{B}}(\alpha,\beta) - \text{Obsv}_{\vec{A},\vec{B}}(\alpha,\beta)]^2}{\text{Expt}_{\vec{A},\vec{B}}(\alpha,\beta)} \tag{5.6}$$

Note that when χ^2 is calculated for each column pair, the set possible amino acids summed over in Equation (5.6) changes, and thus the number of degrees of freedom also changes.

For a contingency table the χ^2 statistic is typically used in (Pearson's) χ^2 test for independence. This gives a p-value with the null hypothesis that the marginal distributions (here the two amino acid distributions in $\vec{A}$ and $\vec{B}$) are independent. Such a p-value could be used to rank MSA column pairs, however there are good computational reasons to work directly with the χ^2 value. In addition to the additional computation time required, sorting and comparing small p-values can cause computational problems due to limited floating point number resolution. The χ^2 test is also best avoided when dealing with a large number of samples (as here, with thousands of protein-protein pairs).

Figures 5.18 shows a histogram of observed χ^2 values for MSA column pairs from the paired proteins, and the matching distribution for randomly paired proteins. It is very clear that paired proteins give much higher χ^2 values, well outside the random distribution. These scores are shown on a grid in Figure 5.19 with the 100 highest scoring column pairs in red.

Figure 5.20 shows the χ^2 scores plotted against the estimated distance between the residues in the protein-protein complex. Column pairs with the highest χ^2 scores are closer together than average. While this approach does seem to have merit, there is much more literature and precedent for the next method discussed, MI.

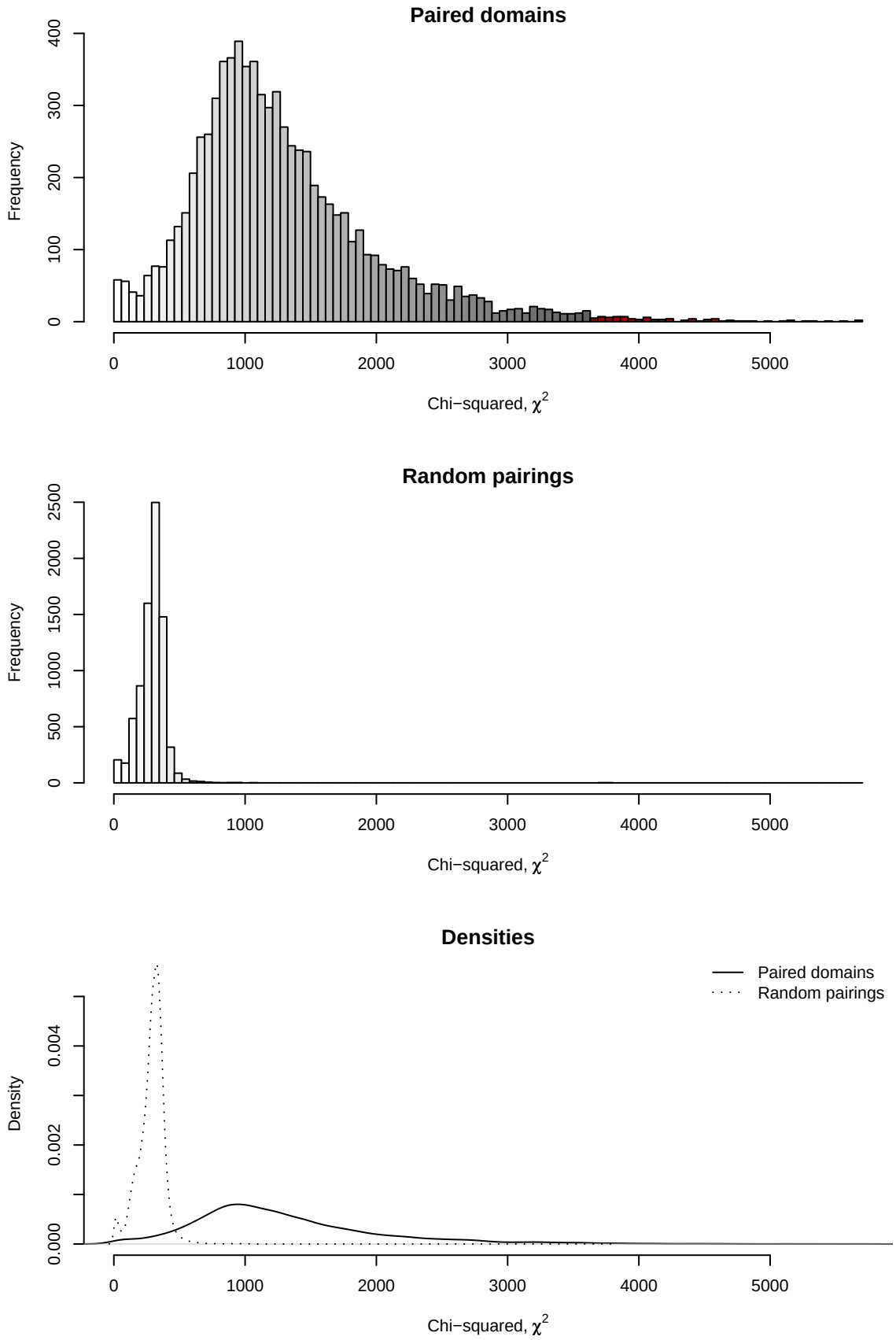


Figure 5.18: Histogram of χ^2 scores from HisKA and receiver pairs (top), and randomised pairings (middle). These are shown as density curves (bottom). Based on MSAs generated by CLUSTAL W.

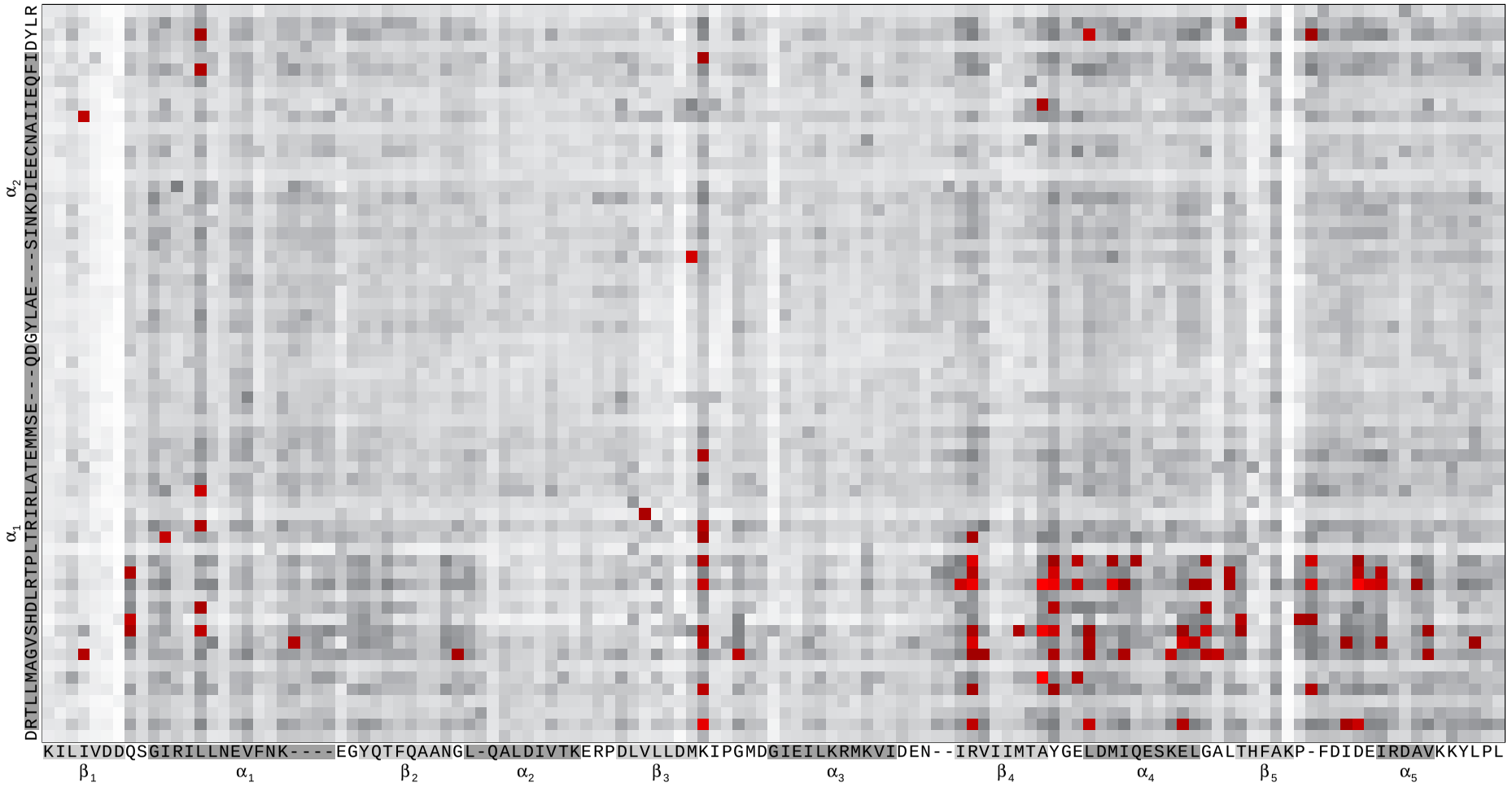


Figure 5.19: Grid of co-varying amino acid positions of HisKA (vertical) and receiver (horizontal) domains, as determined by χ^2, for two gene domain pairs. Each grid square corresponds to a column pairing between the HisKA MSA and the receiver MSA, coloured according to the χ^2 value. Low scores are white, with higher scores in grey except for the top 100 scores which are in red. See Figure 5.18 for the colour key. Based on MSAs generated by CLUSTAL W.

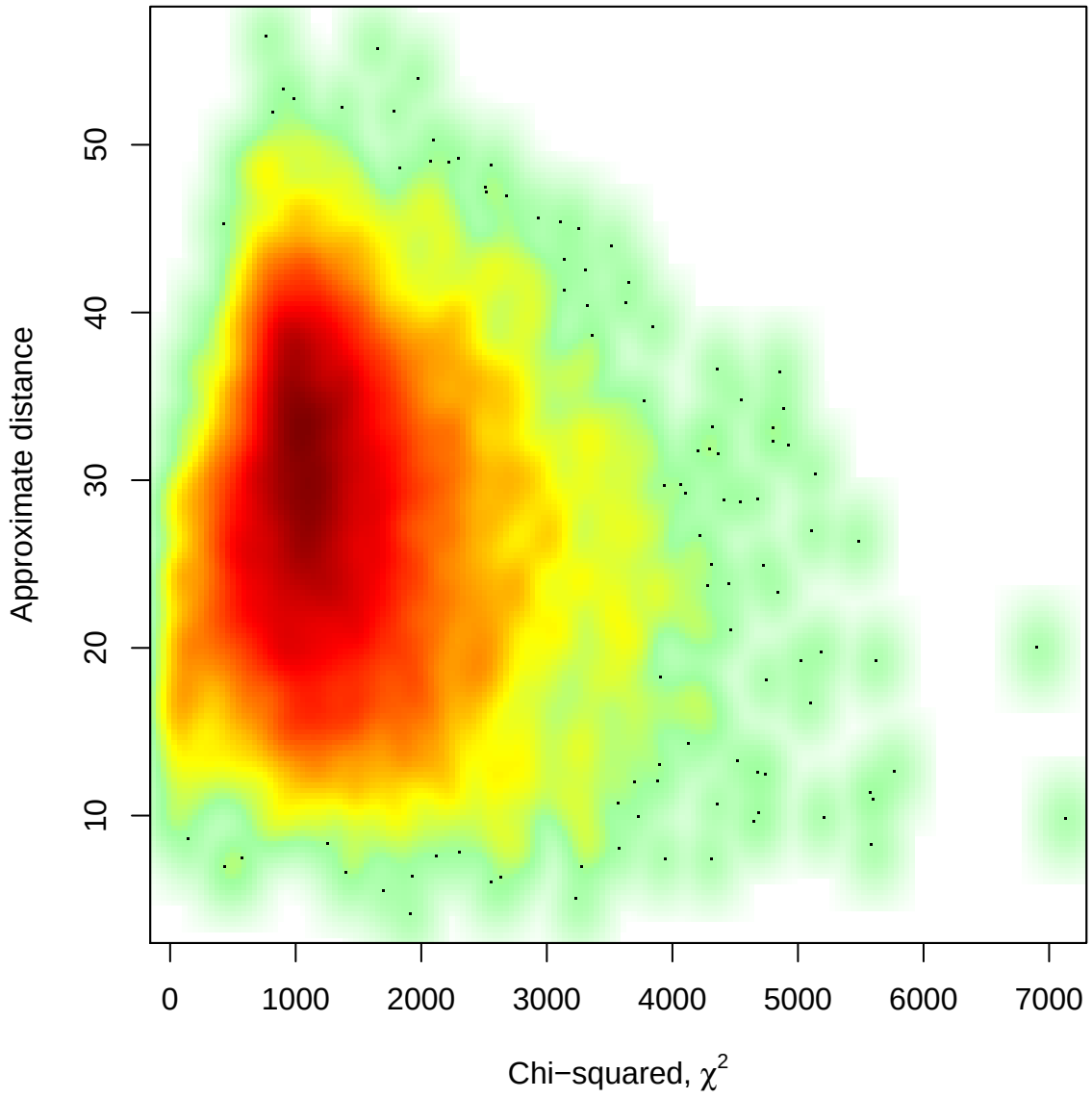


Figure 5.20: Smoothed scatter plot of χ^2 scores against estimated distances from a crude protein-protein complex. Those column pairs with high χ^2 are closer together than average, based on the estimate distances.

5.5.4 Mutual information

The final scoring system considered for MSA column pairs was MI. Unlike the chemical potential (CP) and hydrophilicity scores used above, this does not map the amino acids onto a numerical system, but, like the χ^2 statistic, works directly with the data itself.

As before, let vectors $\vec{A}$ and $\vec{B}$ of length l represent the HisKA and receiver MSA columns under consideration. Let A and B represent random variables describing the amino acids found in the two MSA columns $\vec{A}$ and $\vec{B}$. As per Shannon and Weaver (1949), MI is then defined by the following sum over the observed letters α in $\vec{A}$ and β in $\vec{B}$, as follows[2]:

$$\mathrm{MI}(\vec{A},\vec{B}) := \sum_{\alpha,\beta}\left\{P[A=\alpha,B=\beta]\cdot\log\left(\frac{P[A=\alpha,B=\beta]}{P[A=\alpha]P[B=\beta]}\right)\right\} \tag{5.7}$$

Using this definition it is clear that MI is commutative, that is $\mathrm{MI}(\vec{A},\vec{B}) = \mathrm{MI}(\vec{B},\vec{A})$. Furthermore, it is zero when the random variables A and B are independent (i.e. when $P[A=\alpha,B=\beta] = P[A=\alpha]P[B=\beta]$). In particular, if either column $\vec{A}$ or $\vec{B}$ is perfectly conserved, this implies the MI is zero. Although not immediately apparent from this equation, MI is always positive, shown later. Statisticians may note that MI can also be defined as the Kullback-Leibler divergence between the joint distribution $P[A=\alpha,B=\beta]$ and the product $P[A=\alpha]P[B=\beta]$ (Kullback and Leibler, 1951).

The probabilities in this equation are calculated as the observed frequencies of the amino acids in MSA columns $\vec{A}$ and $\vec{B}$ (which are both of length l). Using the notation introduced in Equation (5.5) we can now express Equation (5.7) in a form suitable for direct calculation:

$$\mathrm{MI}(\vec{A},\vec{B}) = \frac{1}{l}\sum_{\alpha,\beta}\left\{\mathsf{Obsv}_{\vec{A},\vec{B}}(\alpha,\beta)\cdot\log\left(\frac{l\cdot\mathsf{Obsv}_{\vec{A},\vec{B}}(\alpha,\beta)}{\mathsf{Obsv}_{\vec{A}}(\alpha)\cdot\mathsf{Obsv}_{\vec{B}}(\beta)}\right)\right\} \tag{5.8}$$

These equations can be generalised to a summation over all possible amino acid pairs, provided the summand is taken as zero when an amino acid has not been observed, avoiding the undefined term $\log(0/0)$.

As an alternative to Equation (5.7), MI can also be defined in terms of entropies (Shannon and Weaver, 1949). Summing over the observed values, the (marginal) entropies are defined as:

$$\begin{aligned} H(\vec{A}) &:= -\sum_{\alpha}\Big\{P[A=\alpha]\cdot\log(P[A=\alpha])\Big\} \\ H(\vec{B}) &:= -\sum_{\beta}\Big\{P[B=\beta]\cdot\log(P[B=\beta])\Big\} \end{aligned} \tag{5.9}$$

[2] Gap characters are treated as letters.

Note that these are positive quantities, zero only when the sequence is perfectly conserved. They measure our uncertainty about the value of the random variables A or B. Similarly the joint entropy is:

$$\begin{aligned} H(\vec{A},\vec{B}) &:= -\sum_{\alpha,\beta}\Big\{P[A=\alpha,B=\beta]\cdot\log(P[A=\alpha,B=\beta])\Big\} \\ &= H(\vec{B},\vec{A}) \end{aligned} \tag{5.10}$$

Again, this is a positive quantity. The joint entropy is always at least the individual entropy, $H(\vec{A}) \leq H(\vec{A},\vec{B})$ and $H(\vec{B}) \leq H(\vec{A},\vec{B})$. Similarly, it is bounded by the sum of the individual entropies, $H(\vec{A},\vec{B}) \leq H(\vec{A}) + H(\vec{B})$, with equality only when A and B are independent. These inequalities are intuitive if the entropy is thought of as the information content of the amino acid sequences. The MI can be expressed as the difference between the joint entropy and the sum of the two marginal entropies:

$$\mathrm{MI}(\vec{A},\vec{B}) = H(\vec{A}) + H(\vec{B}) - H(\vec{A},\vec{B}) \tag{5.11}$$

Proof. This result follows from the observation that $P[A=\alpha] = \sum_\beta P[A=\alpha,B=\beta]$ and similarly for $P[B=\beta]$, thus:

$$\begin{aligned} H(\vec{A}) + H(\vec{B}) - H(\vec{A},\vec{B}) &= -\sum_\alpha\Big\{\Big(\sum_\beta P[A=\alpha,B=\beta]\Big)\log(P[A=\alpha])\Big\} \\ &\quad -\sum_\beta\Big\{\Big(\sum_\alpha P[A=\alpha,B=\beta]\Big)\log(P[B=\beta])\Big\} \\ &\quad +\sum_{\alpha,\beta}\Big\{P[A=\alpha,B=\beta]\cdot\log(P[A=\alpha,B=\beta])\Big\} \\ &= -\sum_{\alpha,\beta}\left\{P[A=\alpha,B=\beta]\cdot\log\left(\frac{P[A=\alpha,B=\beta]}{P[A=\alpha]P[B=\beta]}\right)\right\} \\ &= \mathrm{MI}(\vec{A},\vec{B}) \end{aligned}$$

□

From Equation (5.11), it follows $0 \leq \mathrm{MI}(\vec{A},\vec{B}) \leq \max(H(\vec{A}),H(\vec{B}))$, that is to say the MI is postive and limited by the variability of both sequences. A high MI value is only possible when $\vec{A}$ and $\vec{B}$ are correlated with a high variability. If either $\vec{A}$ or $\vec{B}$ is very conserved (or perfectly conserved) then the MI will be small (or zero).

MI was originally calculated with a base two logarithm, giving an information measure in bits, which is natural for binary codes. The choice of base is essentially a scaling issue. When dealing with amino acids, some authors have continued to use base two (Atchley *et al.*, 1999, 2000), but others such as Dunn *et al.* (2008) have adopted base 20 (the number of amino

acids, ignoring potential gap characters). Here, the natural logarithm has been used (base e), meaning the MI scores are in *nats* rather than bits (Comley and Dowe, 2005). This choice was in part for consistency with link functions in the following chapter. Gouveia-Oliveira and Pedersen (2007) explores other choices, including a weighted scaling according to the number of amino acids present in the two columns.

As with the previous scoring systems discussed above, MI was calculated between the columns of the paired HisKA and receiver MSAs. Figure 5.21 shows the distribution of MI values for the protein pairs, and that given by a randomised pairing of the proteins. There is a very clear difference in distribution, with the paired proteins giving much higher MI values. As with the χ^2 plot (Figure 5.18), this difference is much more pronounced than that seen in the equivalent plots for the earlier column pair correlation scores, for example KD Kendall's τ (Figure 5.11).

Figure 5.22 shows these MI scores as a grid, with the column pairs giving the top 100 MI scores highlighted in red. These tend to be associated with the first α-helix in the HisKA, and cover a range of points in the receiver.

Figures 5.21 and 5.22 are both based on MSAs generated with CLUSTAL W. Figures 5.23 and 5.24 show the same information using MSAs generated by MUSCLE. The MI analysis using the output from the two alignment programs identifies similar but non-identical sets of column pairings, as illustrated in Figures 5.25 and 5.26 for the CLUSTAL W and MUSCLE alignments, respectively.

A visual inspection suggests the column pairs with MI may be interesting in terms of the known interaction surfaces. For more concrete support, Figure 5.27 shows the MI scores plotted against the inferred separation (for the CLUSTAL W MSAs, the results for MUSCLE are similar). Column pairs with a high MI score are generally closer together than average.

5.6 Mapping scores onto protein structures

When the alignments used to calculate the correlation and mutual information/entropy scores were created, a reference sequence was included for which a known 3D structure was available. By mapping alignment positions to these reference sequences, most alignment columns can be mapped to a 3D position. For alignment positions corresponding to a gap in the reference sequence, this is not so straightforward.

Taking the HisKA and receiver structures in isolation, any MSA column score can be displayed visually by colouring the protein model of the reference sequence. For example, each position in the HisKA could be assigned the maximum of all the MSA column pair scores for

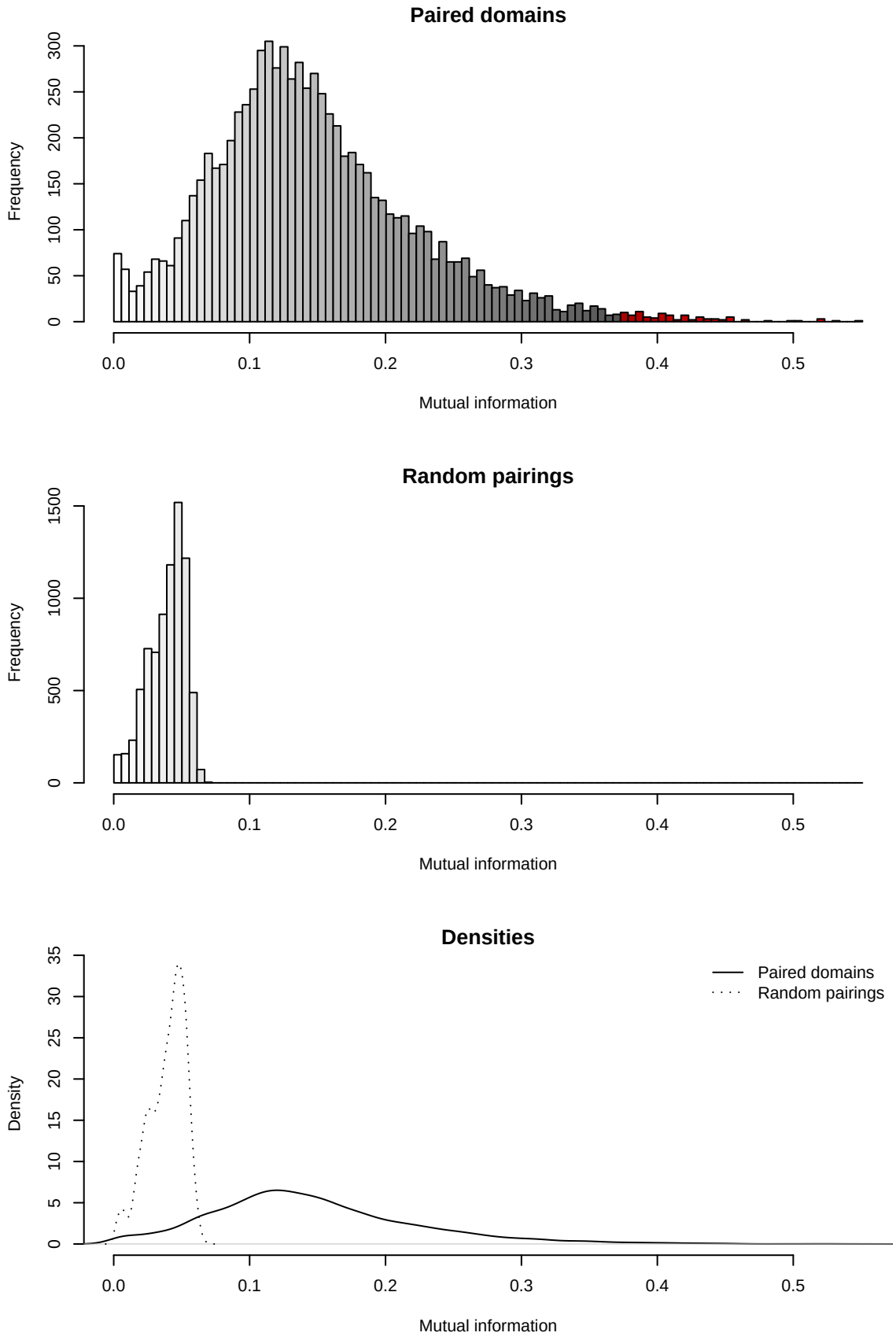


Figure 5.21: Histogram of MI scores from HisKA and receiver pairs (top), and randomised pairings (middle). These are shown as density curves (bottom). Based on MSAs generated by CLUSTAL W.

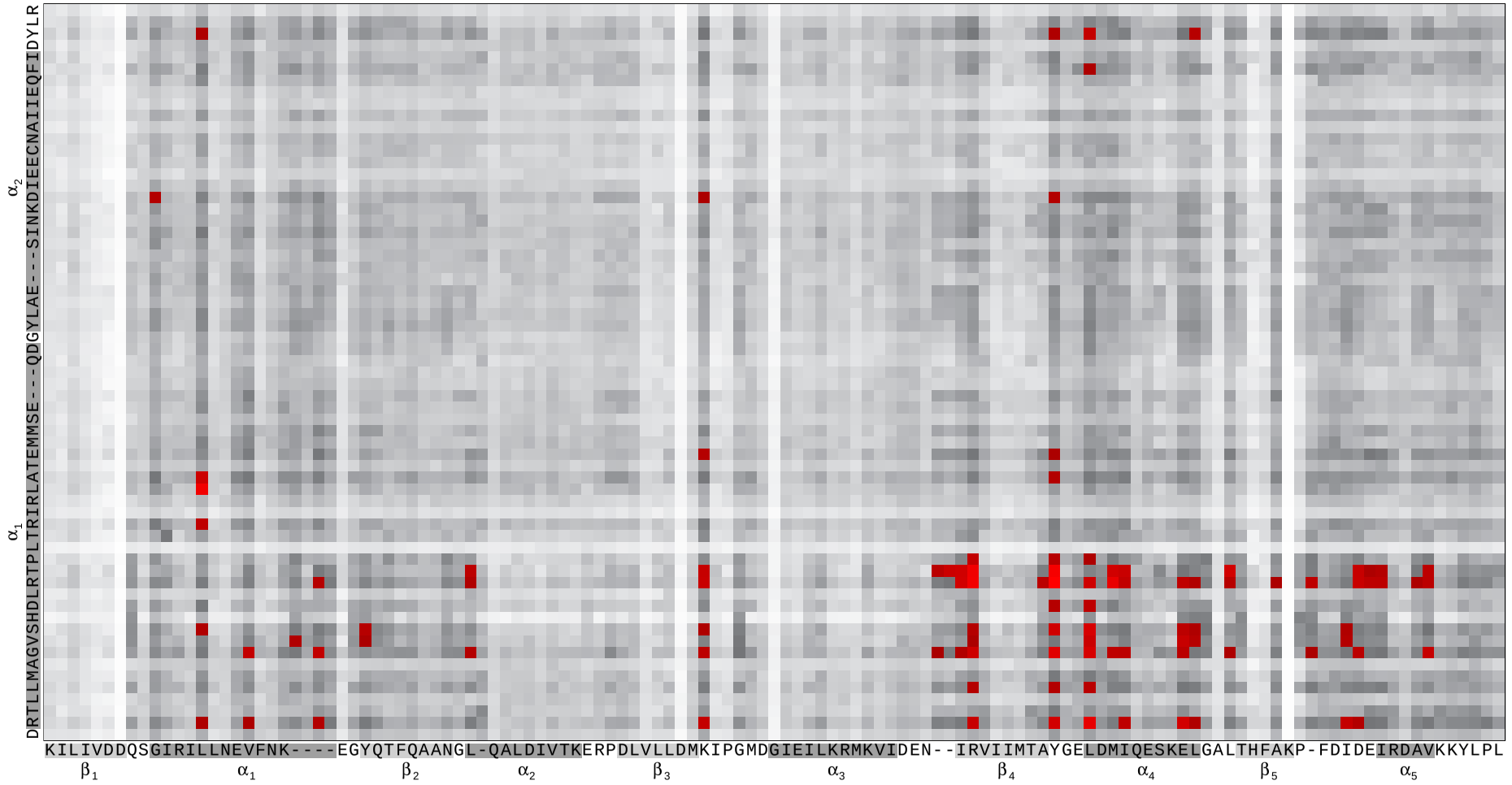


Figure 5.22: Grid of co-varying amino acid positions of HisKA (vertical) and receiver (horizontal) domains, as determined by MI, for two gene domain pairs. Each grid square corresponds to a column pairing between the HisKA MSA and the receiver MSA, coloured according to the MI. Low scores are white, with higher scores in grey except for the top 100 scores which are in red. See Figure 5.21 for the colour key. Based on MSAs generated by CLUSTAL W.

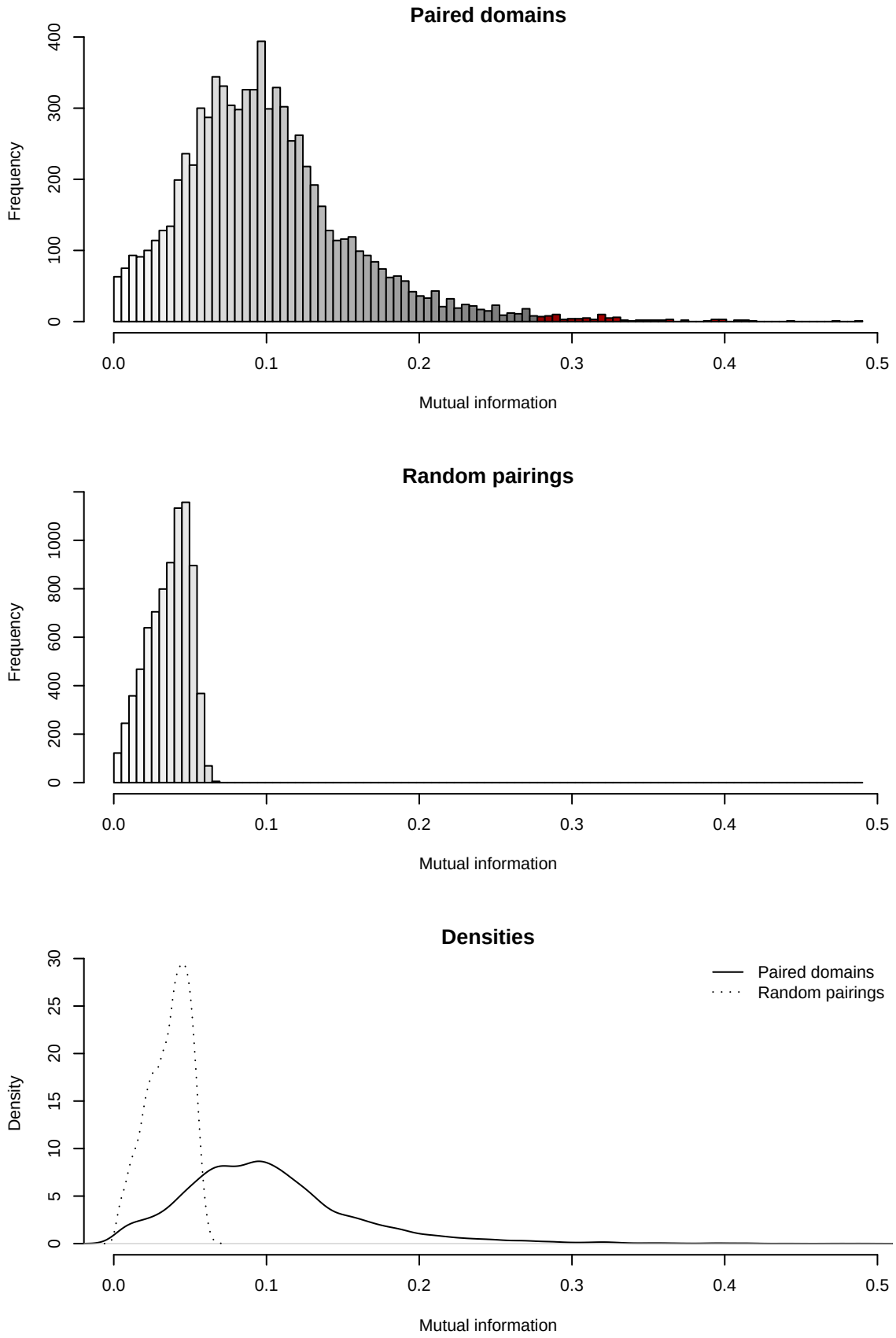


Figure 5.23: Histogram of MI scores from HisKA and receiver pairs (top), and randomised pairings (middle). These are shown as density curves (bottom). Based on MSAs generated by MUSCLE.

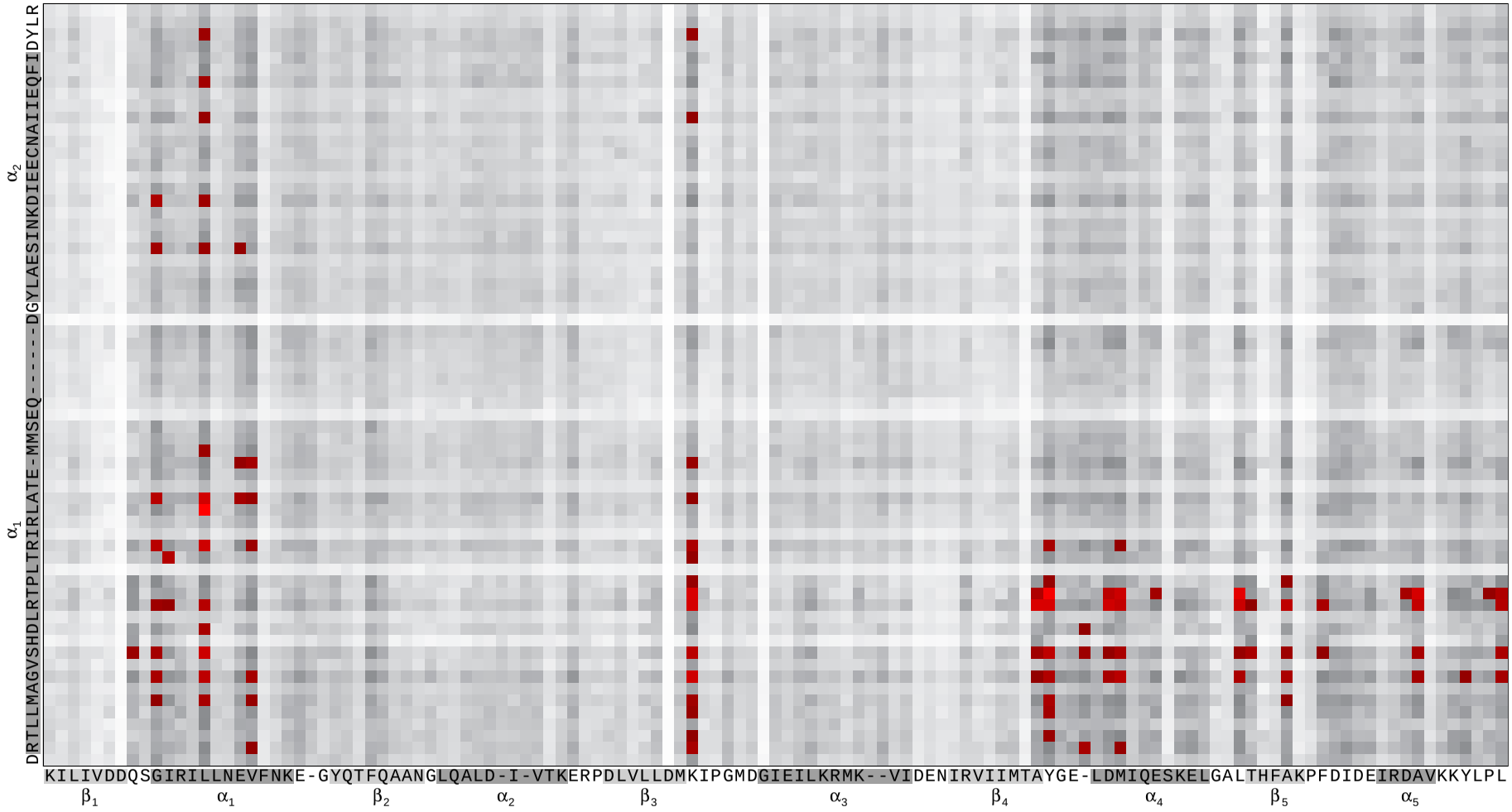


Figure 5.24: Grid of co-varying amino acid positions of HisKA (vertical) and receiver (horizontal) domains, as determined by MI, for two gene domain pairs. Each grid square corresponds to a column pairing between the HisKA MSA and the receiver MSA, coloured according to the MI. Low scores are white, with higher scores in grey except for the top 100 scores which are in red. See Figure 5.23 for the colour key. Based on MSAs generated by MUSCLE.

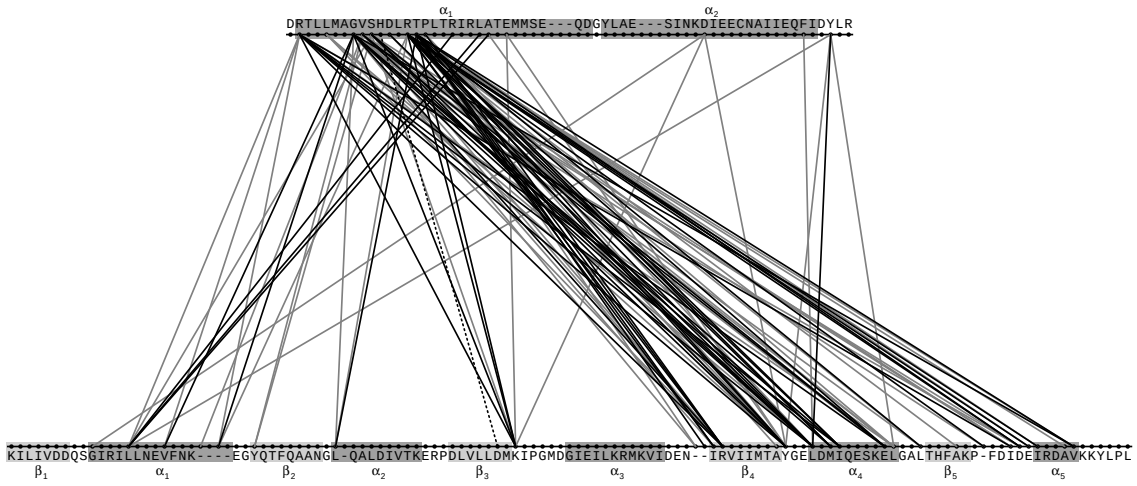


Figure 5.25: Top 100 co-varying amino acid positions of HisKA and receiver domains, as determined by MI, for two gene domain pairs. Based on MSAs generated by CLUSTAL W.

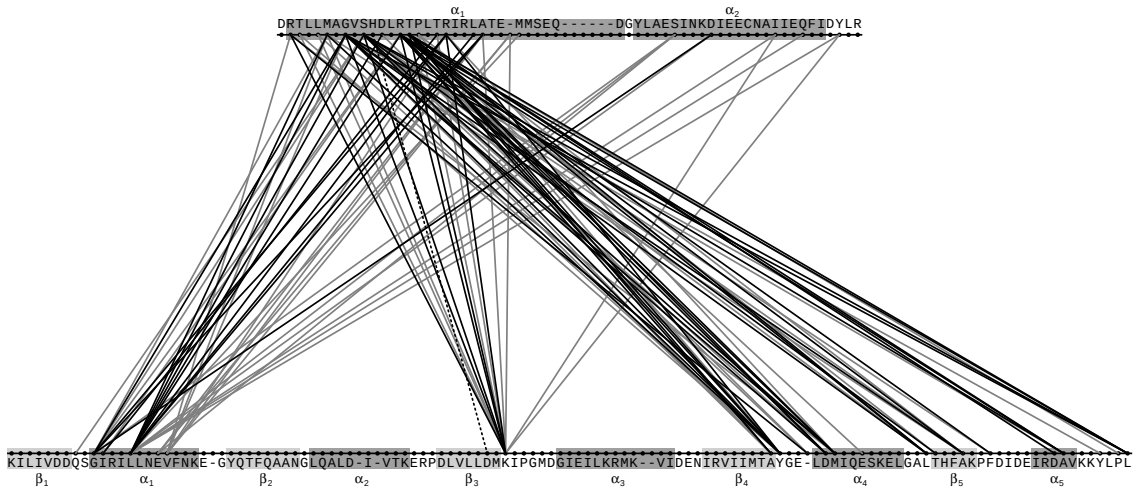


Figure 5.26: Top 100 co-varying amino acid positions of HisKA and receiver domains, as determined by MI, for two gene domain pairs. Based on MSAs generated by MUSCLE.

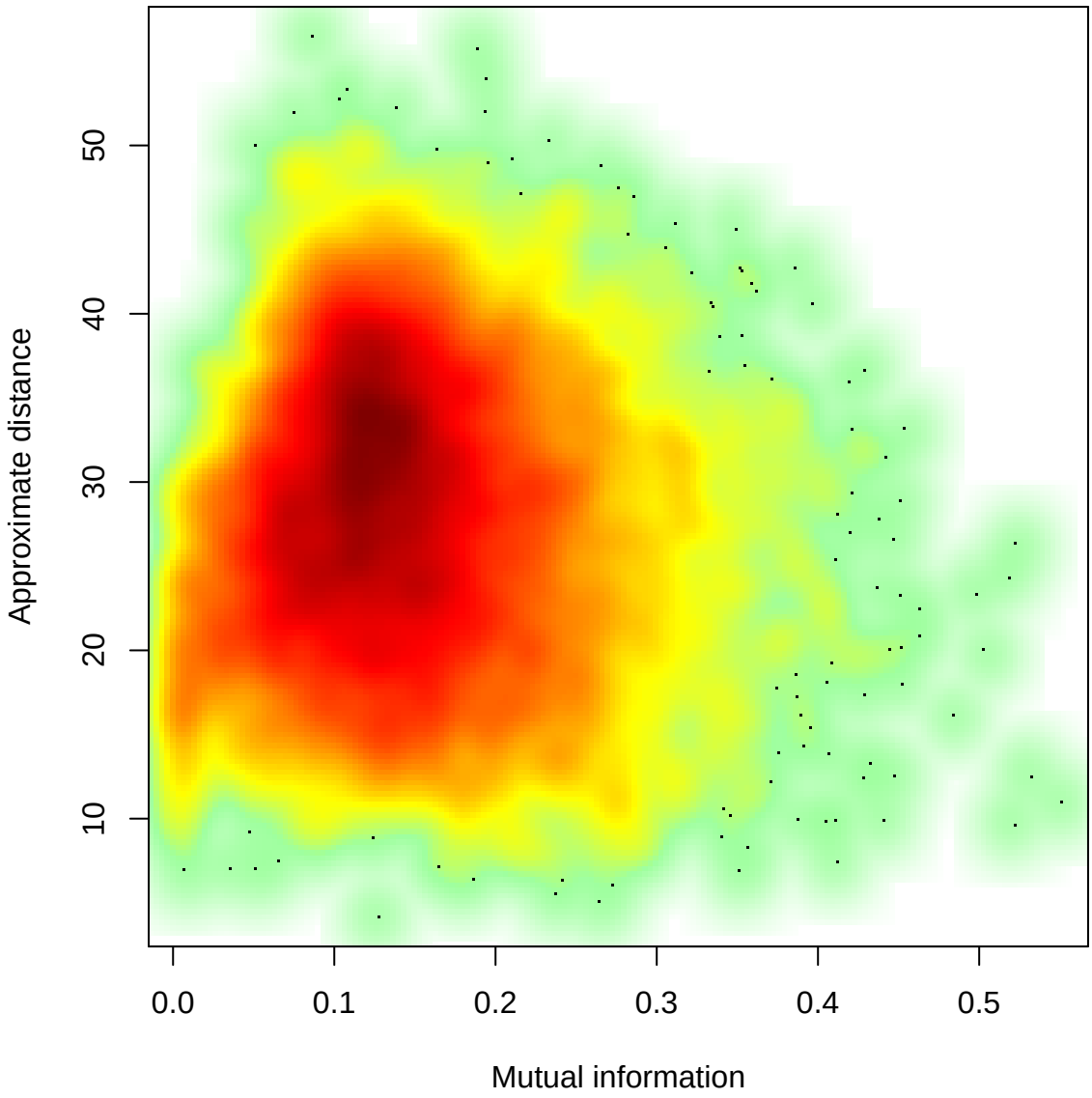


Figure 5.27: Smoothed scatter plot of MI against estimated distances from a crude protein-protein complex. Those column pairs with a high MI are closer together than average based on the estimate distances. Based on MSAs generated by CLUSTAL W.

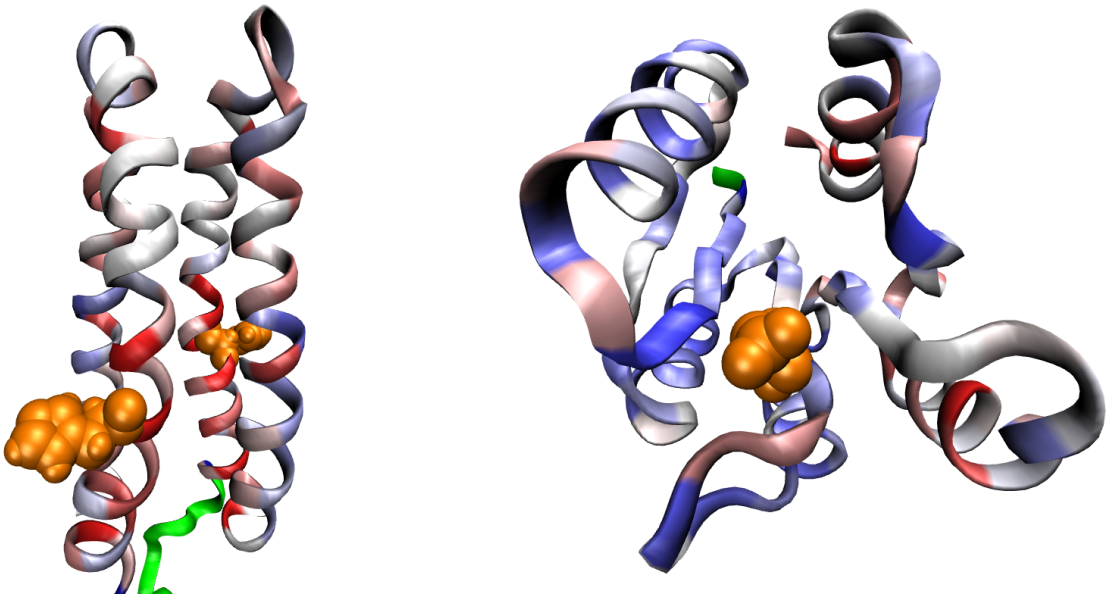


Figure 5.28: This shows the KD τ correlation score with the largest absolute value for each column in the HisKA and receiver alignment, mapped onto reference 3D structures. On the left is the HisKA dimer structure, EnvZ, orientated to show the conserved histidine (shown by a space filling model in orange). On the right is the receiver structure, Spo0F, orientated to show the conserved aspartate (shown by a space filling model in orange). The KD τ colour ranges from blue (negative) through white to red (positive), although the colours do not exactly match those used in previous figures. Residues which were not in the alignments are in green.

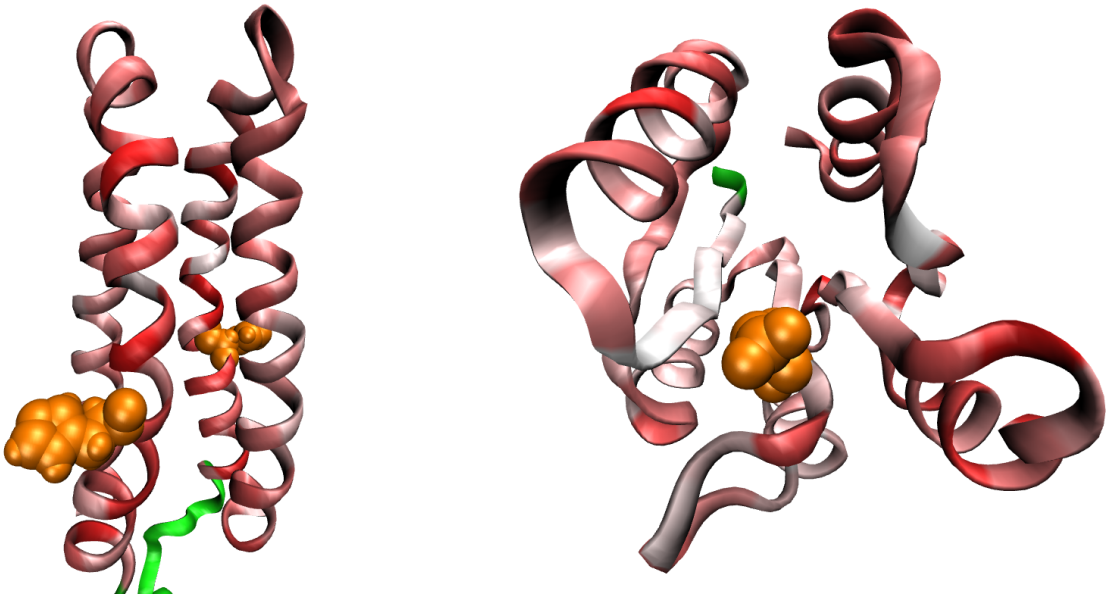


Figure 5.29: Maximum MI score for each column in the HisKA and receiver alignment mapped onto reference 3D structures (EnvZ and Spo0F as in Figure 5.28). MI colour ranges from white (zero) to red (maximum). Residues which were not in the alignments are in green.

that position, and similarly for the receiver. Figure 5.28 uses this approach to show the KD Kendall's τ correlation with the largest absolute value, while Figure 5.29 shows the maximum MI score.

Simplified representations of the reference protein structures, which are not dependent on the amino acid side chains, are most suitable for representing a generic HisKA or receiver. The "new ribbon" representation in VMD was used (Humphrey *et al.*, 1996), rather than a "cartoon" representation which reduces α-helices to simple cylinders discarding the orientation of the individual residues. It should be noted that these representations do not attempt to show the amino acid *pairings* identified (between the proteins), which is a much more complicated problem to visualise.

5.7 Summary and comparison of results

Figures 5.30 and 5.31 (for the CLUSTAL W and MUSCLE MSAs respectively, using the HK and RR domain pairs) summarise the correlations between the scores considered, and the approximate distance in the protein-protein complex. Those column pairs with the highest MI are by definition highly variable, and from these figures show a broad range of hydrophobicity correlations (represented in the figures by the KD Kendall's τ), while those with a low MI have KD $\tau \sim 0$. These figures also show the χ^2 and MI scores correlate well, and these show the most convincing link to the distances. However, distance isn't everything – McLaughlin *et al.* (2007) shows predominantly buried residues can play a role in recognition.

Although Figures 5.30 and 5.31 look very similar, the precise MSA column pairs selected do differ for the two alignment methods. Using the reference sequences EnvZ and Spo0F, it is possible to cross-reference most of the column pairs (but not those where the patten of gaps is different) allowing a direct comparison of the scores from the CLUSTAL W MSAs to those from the MUSCLE MSAs.

Figures 5.32, 5.33 and 5.34 show this for MI. The results from the two alignment methods correlate well, but overall the CLUSTAL W MSAs give somewhat higher MI scores. The spread of the points can be significantly reduced by excluding column pairs which are "gap rich" leaving a much clearer diagonal trend (comparing the top and bottom sub-figures, which exclude columns with more than 1% gaps, or only columns with more than 50% gaps). This suggests that most differences between the two alignment tools are down to the placement of insertions.

As a consequence of this discrepancy due to the different gap placements, many of the column pairs selected by MI using the CLUSTAL W MSAs have a much lower MI from

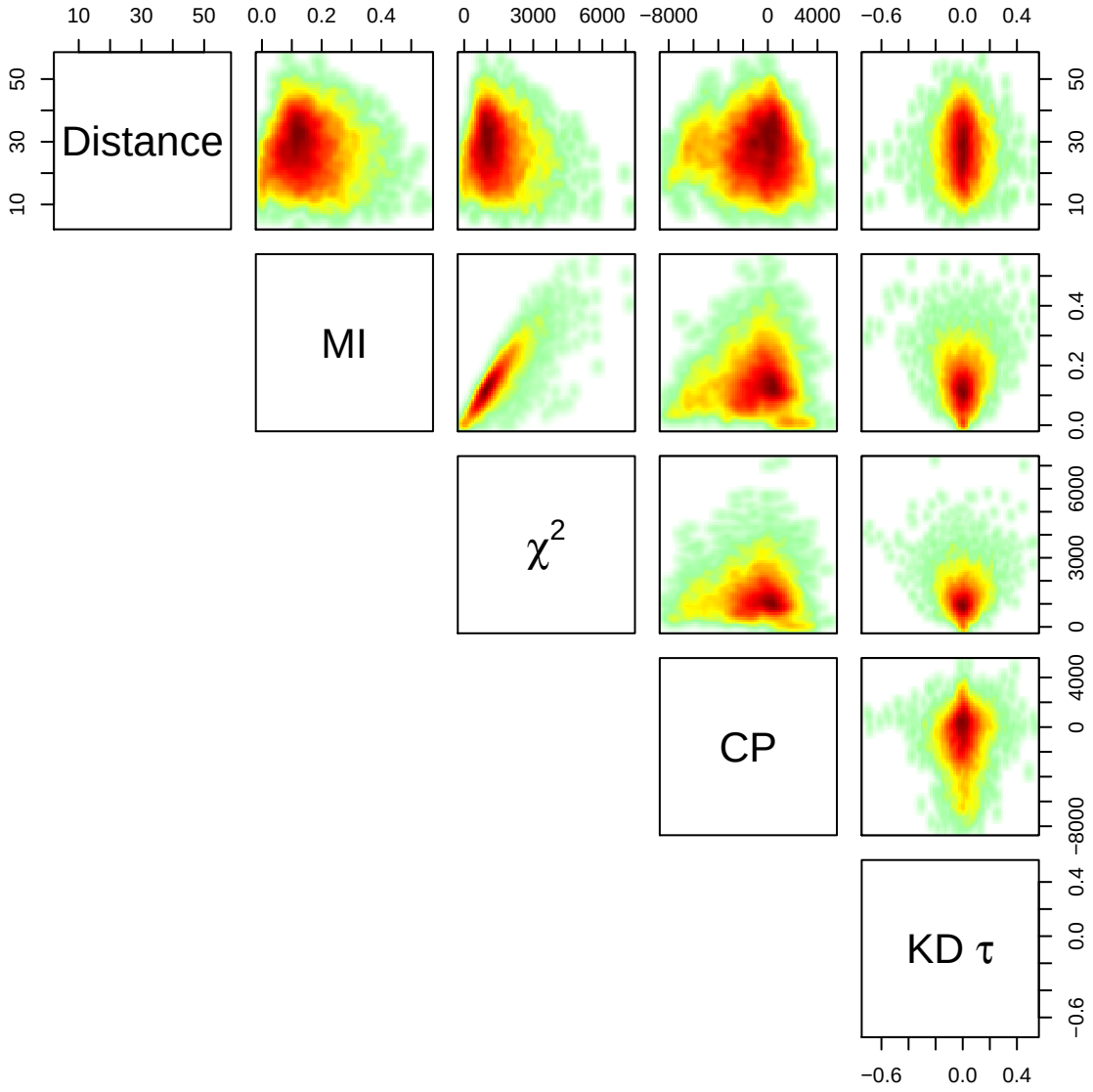


Figure 5.30: Assorted smoothed scatter plots for selected scores and the estimated distances from a crude protein-protein complex. As in Figure 5.16, the captions along the diagonal indicate which scoring method is shown on the associated row/column. These plots are based on MSAs created with CLUSTAL W.

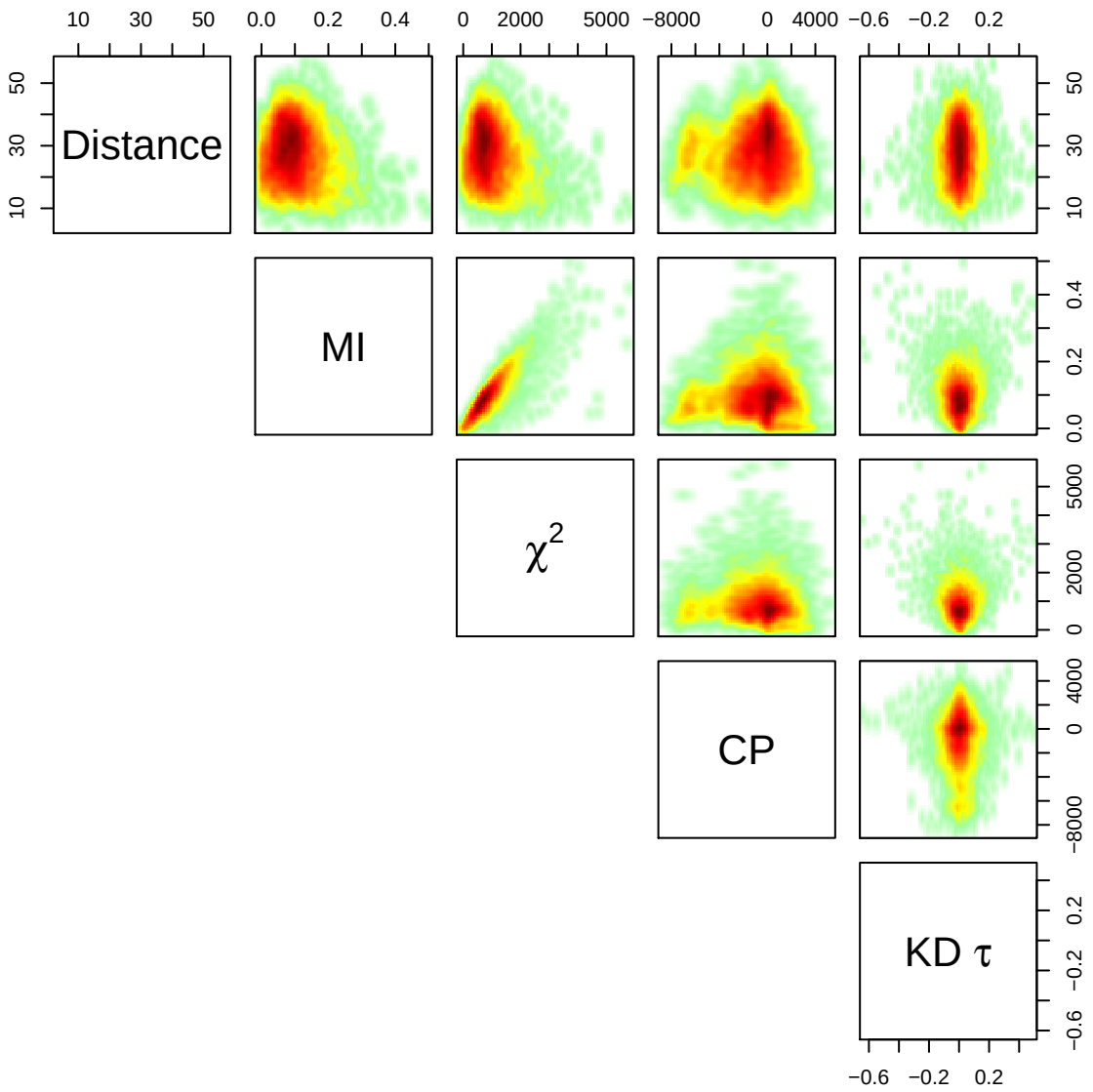


Figure 5.31: Assorted smoothed scatter plots for selected scores and the estimated distances from a crude protein-protein complex, based on MSAs created with MUSCLE.

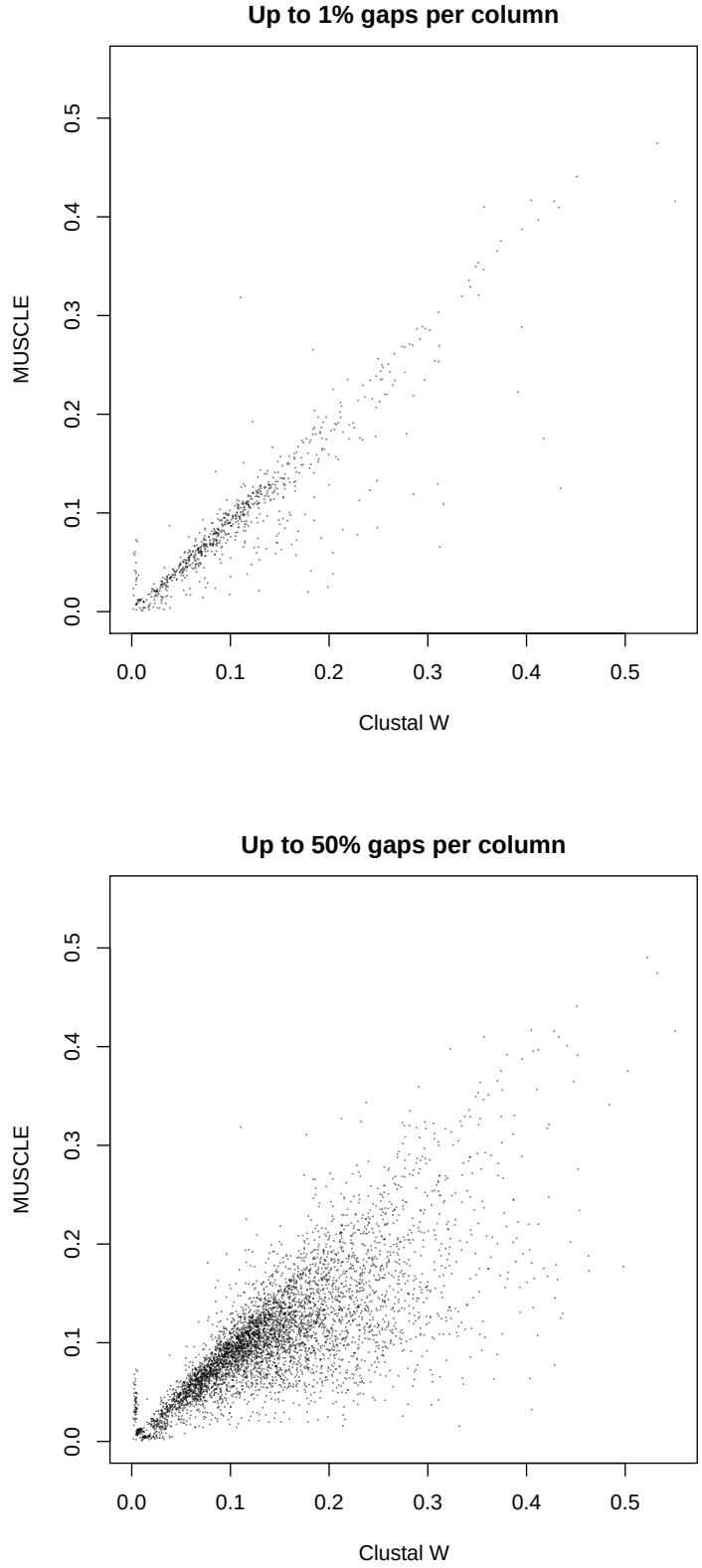


Figure 5.32: Scatter plot of the MI scores from the CLUSTAL W and MUSCLE paired MSAs using domains from HK and RR gene pairs. Column pairs were cross-referenced using the EnvZ and Spo0F reference sequences.

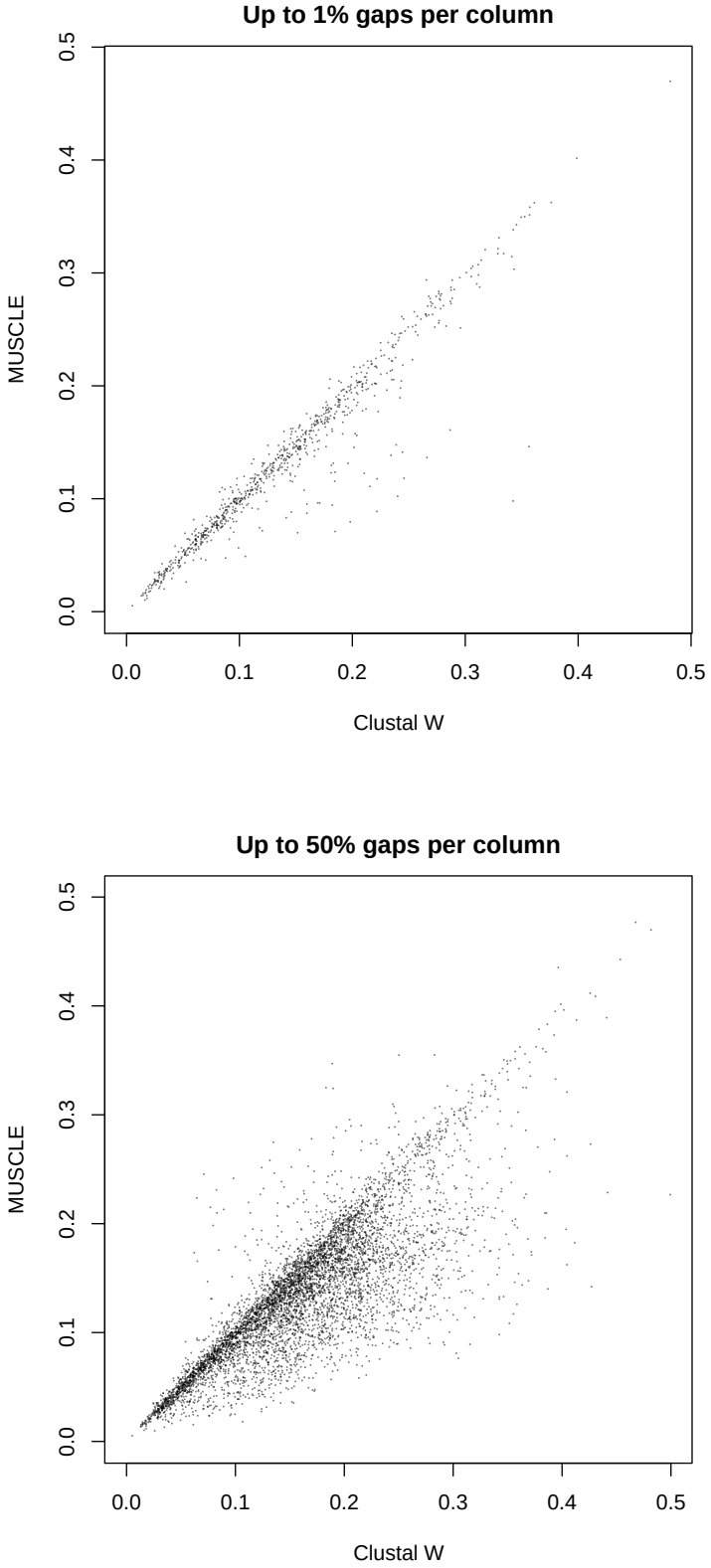


Figure 5.33: Scatter plot of the MI scores from the CLUSTAL W and MUSCLE paired MSAs using domains from HY genes. Column pairs were cross-referenced using the EnvZ and Spo0F reference sequences.

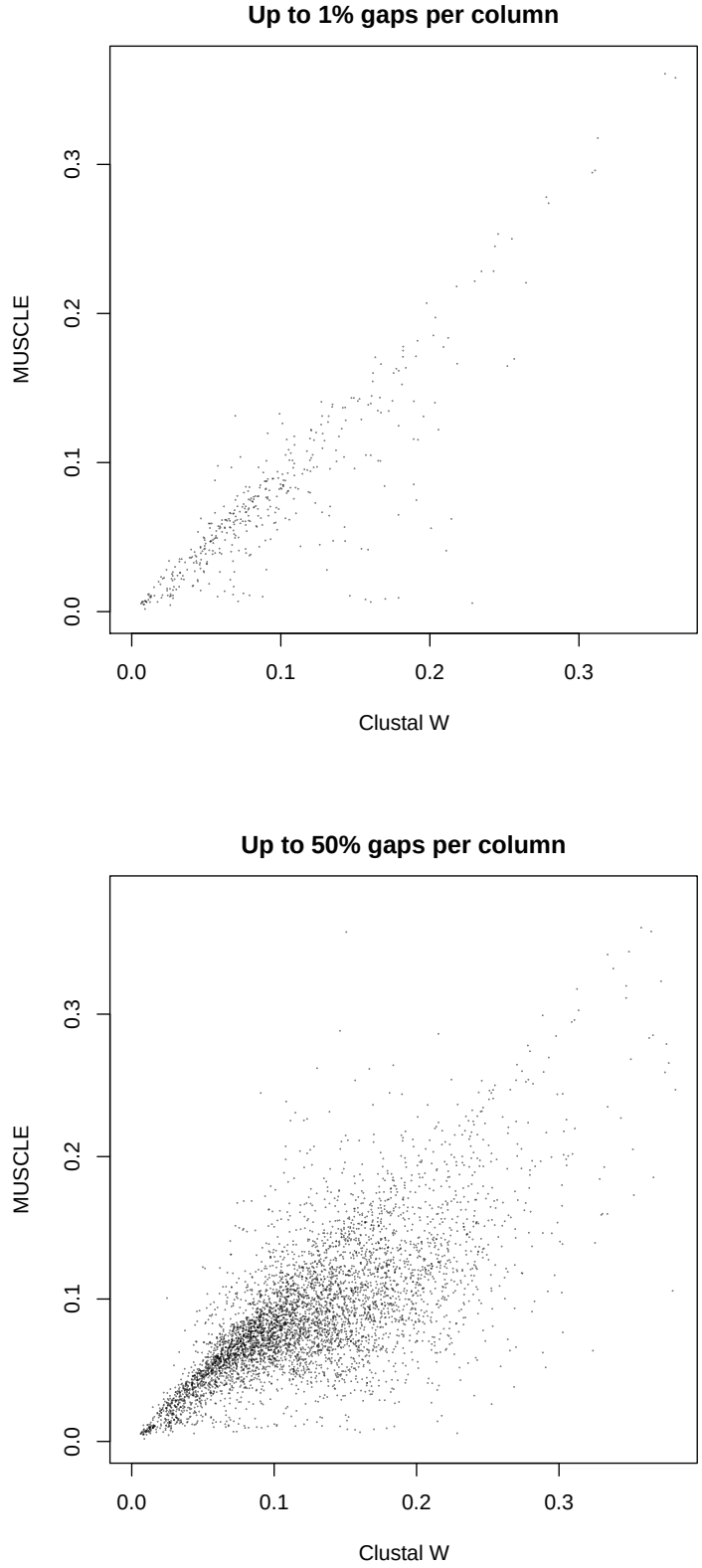


Figure 5.34: Scatter plot of the MI scores from the CLUSTAL W and MUSCLE paired MSAs using domains from both HK and RR gene pairs and HY genes. Column pairs were cross-referenced using the EnvZ and Spo0F reference sequences.

the MUSCLE MSA, and thus are not ranked as highly. However, there is still a large overlap between the top column pairs by MI given by the two different alignment sets (particularly if gap rich columns are explicitly excluded).

Additionally, Figure 5.32 shows a small population of column pairs with a CLUSTAL W MI ~ 0 but MUSCLE MI ~ 0.05 for the HK and RR paired domains. Interestingly this grouping does not occur on the HY dataset, nor the combined dataset (Figures 5.33 and 5.34). These column pairs all include the RR Spo0F residues D10 and K104 in the MUSCLE MSA. The two RR aspartate residues (DD) corresponding to Spo0F D9 and D10 at the end of β_1 are highly conserved (see Figure 1.18 for the secondary structure naming). However in some RRs there are three aspartate residues (DDD), and in others a glutamic acid and two aspartate (EDD, chemically similar). This variability complicates the alignment and is likely to explain the difference between the two methods in this region. K104 corresponds to another highly conserved RR residue, at the end of β_5. In the CLUSTAL W MSA this is conserved in all but 7 sequences. In the MUSCLE MSA, however, the K is missing or mis-aligned in 126 sequences, meaning that this column is much more variable and therefore can by chance reach a higher but still small MI score, ~ 0.05.

Although the results from the Clustal W and MUSCLE MSAs do not differ substantially, there are other ways to build MSAs (see Edgar and Batzoglou (2006) for a recent review). In particular, given a number of solved 3D structures are available for both the HisKA and receiver domains (see Section 1.5), it would be possible to build MSAs taking advantage of this spatial information, for example using the tool 3DCoffee (O'Sullivan *et al.*, 2004), and repeat the analyses here.

5.8 Discussion

Based on these results, in the following chapter potentially informative column pairs for the HisKA-receiver specificity are selected using MI. Column pairs with high MI are assumed to represent positions on the two proteins which have co-evolved and interact in some way. When either column is (almost) perfectly conserved, the column pair is uninformative, and the MI is (almost) zero. In particular, this automatically down-weights any gap-rich columns.

One limitation of MI is that it can be drowned out by background noise for small MSAs. Given Martin *et al.* (2005) suggested at least 125 sequences be used to avoid this issue, as the MSAs here contain thousands of sequences this isn't expected to be a problem.

Two other groups recently published work using MI to identify the residues controlling the HisKA-receiver interaction specificity (White *et al.*, 2007; Skerker *et al.*, 2008). The

later paper included experimental verification that disrupting the residues identified negatively impacted the phosphotransfer, and more interestingly also demonstrated *in vivo* switching of specificity of the HisKA-receiver complex. Even without this external verification, the results above are highly encouraging in that MI appears to select column pairs are important for the HisKA-receiver specificity. In the following chapter, these column pairs are used as the basis of a predictive model for determining HisKA-receiver pairings from their amino acid sequences.

One open question at the close of this chapter is how many of the highly scoring column pairs identified are actually biologically important. In Figure 5.21 (and similar plots), the observed scores are compared to those generated using a randomisation of the TCS domain pairings. Repeating this procedure multiple times would allow the variability of these scores to be estimated. Such a bootstrap proedure would allow a p-value based cut off (incorpoating a multiple testing correction) to select only those column pairings with a statistically significant score. In Chapter 6 however, the number of column pairs to include is considered from an alternative perspective.

Chapter 6

Predictions using a generalised linear model (GLM)

6.1 Introduction

While interacting TCS domains are usually found together as pairs in the genome, in some species such as *Nostoc* sp. or *M. xanthus*, a large proportion of TCS domains are found in more complicated arrangements (Whitworth and Cock, 2008a,b, and Chapter 2). In such organisms, being able to predict TCS domain partnerships from genome sequences would be especially useful, as the lack of paired HK and RR genes makes it difficult to define genome-encoded signalling networks. Even without consideration of co-expression and co-localisation (see Section 1.7), the ability to predict potential pairings solely from amino acid sequences would be a useful tool to guide experimental investigation, especially in cases where a large number of possible combinations makes exhaustive testing a daunting prospect.

In addition to the TCS systems, there are many other classes of protein-protein interactions where the two interacting proteins are encoded as paralogous sets within genomes, and it is therefore not immediately clear which of the possible combinations are biologically relevant. Examples of such systems include G-protein coupled receptors/trimeric G-proteins (Cabrera-Vera *et al.*, 2003) and σ factors/anti-σ factors (Hughes and Mathee, 1998). Any methodology demonstrated for predicting TCS interactions may therefore prove to have wide utility.

6.2 Modelling approach

GLMs are used to explain a dependent variable Y_i (with samples indexed by i) using a weighted linear combination of K explanatory variables X_{ik} (for $k = 1, 2, \ldots, K$). The weights are model parameters which must be estimated by fitting the model to a training dataset. The aim here is to predict if two proteins interact, given only their amino acid sequences (and a training dataset). In the GLM framework, this is a binary classification problem (Y is a binary random variable, $Y_i = 1$ if the proteins interact, $Y_i = 0$ otherwise). This is handled as a binomial logistic regression,

$$\begin{aligned} \text{logit}(P[Y_i = 1 | X_{ik} \text{ for all } k]) &= \log \frac{P[Y_i = 1 | X_{ik} \text{ for all } k]}{P[Y_i = 0 | X_{ik} \text{ for all } k]} \\ &= w_0 + \sum_{k=1}^{K} w_k X_{ik} + \epsilon_i \end{aligned} \tag{6.1}$$

where $P[Y_i = 1]$ denotes the probability of interaction, w_0 an intercept parameter, w_k the K weights, and ϵ_i the error.

The sample index i is an integer used to index different combinations of the two protein domains (arranged in a simple list rather than as a grid). The explanatory variables will be based on the amino acids in the MSA entries of the relevant domains (indicated by i). However, rather than attempting to include all the amino acids in these protein domains, only K MSA column pairs will be selected, indexed by the integer $k = 1, 2, \ldots, K$. Any single value of k identifies a column of the HisKA MSA *and* a column of the receiver MSA, and can therefore be considered as a pair of indices (see Figure 6.1). Each value X_{ik} is calculated from the amino acid pair given by the MSA columns indicated k and the MSA rows indicated by i.

Estimating the w_0 and the K weights, w_k, requires samples (training data) of both known interactions ($Y_i = 1$) *and* non-interactions ($Y_i = 0$). The weights are then selected to best fit the training data. To generate suitable training data, members of the two paralogous protein families under consideration (HisKA and receiver domains) must first be identified (for example using domain based searches, or sequence similarity to known exemplars), and interactions and non-interactions between these protein family members identified. The TCS training sets used in this chapter were compiled from multiple prokaryotes, as described in Chapter 2, assuming paired domains interact exclusively with their partner, and not with any other paralogues encoded in that genome.

Given a test set of proteins for which we wish to make predictions (such as the TCS domain complement of a genome of interest), an amino acid MSA is generated for each domain type (HisKA and receiver domains), combining the training and test datasets. Using the MI of the interacting domains in the training data, K column pairs are selected, and then the chosen

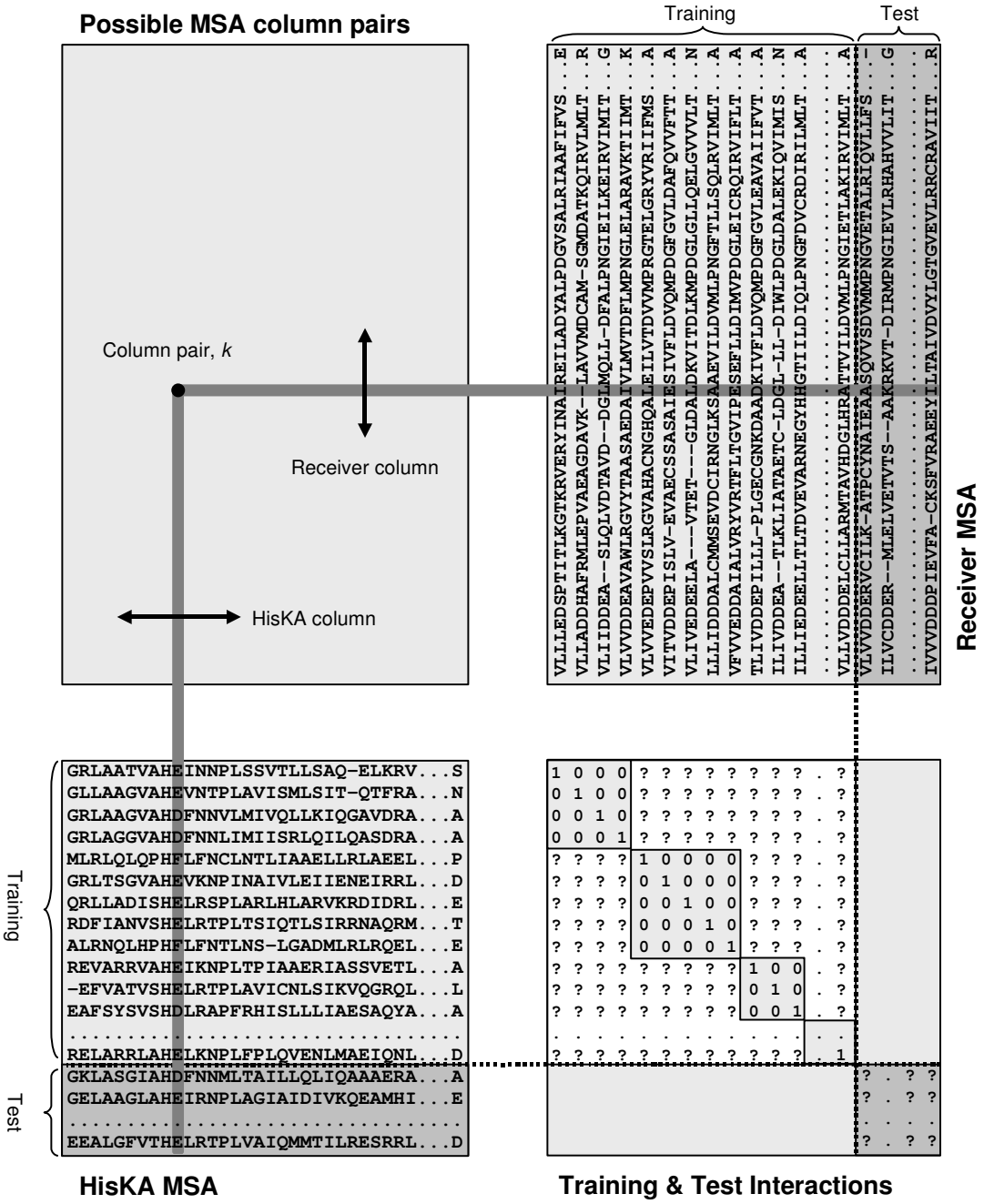


Figure 6.1: Overview of the MSA data and the indexing used (cf. Figure 5.2). The bottom left and top right panels show abbreviated representative HisKA and receiver domain protein sequence MSAs. Column pairs are indexed by k, an example column pair is shown (dark grey bars), represented as a point in the possible MSA column pair space (top left panel). Dotted lines separate the training and test data. The bottom right panel shows the known interactions for the training sequences (upper left of dotted lines), and the interactions we wish to predict for the test sequences (bottom right of dotted lines, "?" on grey). The training interactions are shown with a block structure reflecting inter- and intra-species combinations, where the intra-species interactions are known ("0" or "1" on pale grey) while the hypothetical inter-species interactions are unknown ("?" on white).

scoring system is applied to calculate the explanatory variables X_{ik}. Next, the GLM is fitted to the training data, estimating the model parameters w_k. A prediction for Y_i can then be made for any two sequences from the two MSAs (e.g. all combinations of the test proteins), by extracting the relevant amino acids from the MSAs, applying the chosen scoring system to give K numerical scores X_{ik}, and substituting these values together with the estimated w_k into Equation (6.1).

6.2.1 Selecting column pairs

In the context of a GLM, there are established methods for variable selection. One typical method employes the Akaike Information Criterion (AIC) (Akaike, 1974), which requires multiple models to be fitted testing different possible combinations of input variables. However, since it is not computationally feasible to assess the fit of all possible combinations of MSA column pairs as explanatory variables, we must use a strategy to tease out suitable subsets.

As argued in Chapter 5, assuming all consorting paralogues follow the same docking orientation, any pair of columns from these two MSAs represents a potential inter-residue interaction between the corresponding positions of the protein-protein complex. The vast majority of these residue pairs will represent spatially separated amino acids, which will make no contribution to the interaction specificity. However, a minority of column-couplets will be informative, representing variable amino acids which do interact (possibly indirectly) in the protein-protein complex. Our first question is how to identify these explanatory column pairs, and then how to interpret them numerically.

Any solved structures of the protein-protein complex of interest could be used to assign typical inter-residue distances, and thus short list pertinent MSA column pairs. Unfortunately, no generic TCS HisKA-receiver complex has been solved to date, the closest available being a sequence dissimilar but structurally analogous complex, PDB reference 1F51 (Zapf *et al.*, 2000). Chapter 5 instead explored several automated data driven techniques to select informative column pairs, concluding with the selection of Mutual information (MI) as the most promising candidate.

Column pairs with high MI are assumed to represent positions on the two proteins which have co-evolved and interact in some way. When either column is (almost) perfectly conserved, the column pair is uninformative, and the MI is (almost) zero. Thus taking the MI ranking gives an ordering allowing the simple selection of the top K column pairs, leaving us only with the choice of K (how many terms to include).

6.2.2 Column pair scores

Several ways to assign a numerical score to a pair of amino acids were investigated for use as explanatory variables in a GLM to predict protein domain interactions. These included two scoring functions, S^V and S^H, which are independent of the MSA column pair under consideration (and symmetric with respect to the two protein families). In addition a series of scores, S_k^P, S_k^M, S_k^ζ, S_k^η and S_k^Ω were constructed, specific to each column pair k, based on the observed amino acid frequencies within the training data.

The integer $k = 1, 2, \ldots, K$ indexes the K column pairs selected by MI. Recall any single value of k can be considered as a pair of indices identifying a column of the HisKA MSA *and* a column of the receiver MSA (see Figure 6.1). Let A_k and B_k represent random variables describing the amino acids found in the training data for the two MSA columns of column pair k, and let a_{ik} and b_{ik} be the specific amino acids from column pair k in sample i. For each score, a separate GLM was constructed setting $X_{ik} = S^V(a_{ik}, b_{ik})$, $X_{ik} = S^H(a_{ik}, b_{ik})$, ..., or $X_{ik} = S_k^\Omega(a_{ik}, b_{ik})$ for substitution into Equation (6.1).

S^V denotes an existing statistical chemical potential (Lu *et al.*, 2003), introduced in Section 5.5.1. Low potentials between close amino acids should indicate energetically favourable complexes, and thus may identify interacting complexes. A major component of this chemical potential is hydrophilicity, which is believed to play a role in TCS interactions (Kojetin *et al.*, 2003). Therefore a hydrophilicity compatibility score S^H is introduced, defined as the product of the Kyte-Doolittle (KD) hydrophilicities of the two amino acids (Kyte and Doolittle, 1982) (see Section 5.5.2). Pairs of hydrophobic or hydrophilic residues have a positive score, while a miss-matched combination is assigned a negative score. For example, $S^H(\text{Asn}, \text{Asp}) = -1.8 \times 3.5 = -6.3$. For any gap characters or ambiguous amino acids, S^V and S^H were taken as zero.

The other scores are probabilistic, based on observed amino acid frequencies in the MSA columns for the column pair under consideration, treating gap characters as another letter. Taking different scores for each column pair allows for different physical interactions between different parts of the protein complex. Given the selection of column pairs using MI, it was natural to consider these kinds of scores.

First S_k^P is defined as the probability of an amino acid pair occurring in MSA column pair k given a positive interaction ($Y = 1$) in the training data,

$$S_k^P(\alpha, \beta) := P\left[A_k = \alpha, B_k = \beta | Y = 1\right]. \tag{6.2}$$

Then the MI contribution score S_k^M is defined as

$$S_k^M(\alpha,\beta) := S_k^P(\alpha,\beta)S_k^\zeta(\alpha,\beta) \tag{6.3}$$

where S_k^ζ is another scoring function,

$$S_k^\zeta(\alpha,\beta) := \log\left(\frac{P\left[A_k=\alpha, B_k=\beta\,|\,Y=1\right]}{P[A_k=\alpha|Y=1]\,P[B_k=\beta|Y=1]}\right). \tag{6.4}$$

Of the preceding scores, the best predictions were given by S_k^ζ, and therefore two further logged probability ratios were evaluated,

$$S_k^\eta(\alpha,\beta) := \log\left(\frac{P\left[A_k=\alpha, B_k=\beta\,|\,Y=1\right]}{P\left[A_k=\alpha, B_k=\beta\right]}\right) \tag{6.5}$$

and

$$S_k^\Omega(\alpha,\beta) := \log\left(\frac{P\left[A_k=\alpha, B_k=\beta\,|\,Y=1\right]}{P\left[A_k=\alpha, B_k=\beta|\,Y=0\right]}\right). \tag{6.6}$$

All the probabilistic scores are estimated from the training data using the observed amino acid frequency counts in the relevant MSA column pair. For example, for column pair k and HisKA/receiver sample i where the amino acid pair is a_{ik}, b_{ik}, score S_k^P is estimated as

$$\hat{S}_k^P(a_{ik},b_{ik}) = \frac{\sum_j I(a_{jk}=a_{ik})I(b_{jk}=b_{ik})I(Y_j=1)}{\sum_j I(Y_j=1)}, \tag{6.7}$$

where $I(\cdot)$ is the indicator function,

$$I(\text{expression}) := \begin{cases} 1 & \text{if expression is true,} \\ 0 & \text{if expression is false.} \end{cases} \tag{6.8}$$

Estimates $\hat{S}_k^M$, $\hat{S}_k^\zeta$, $\hat{S}_k^\eta$ and $\hat{S}_k^\Omega$ are constructed analogously.

6.2.3 Related column pair scores

The last three scores defined above, S_k^ζ, S_k^η and S_k^Ω, are all closely related, differing only by their denominator. Under a couple of reasonable assumptions, these can be shown to be approximately equal.

When considering the estimation of $P\left[A_k=\alpha|\,Y=1\right]$, $P\left[A_k=\alpha|\,Y=0\right]$ or $P\left[A_k=\alpha\right]$ the same MSA column entries are simply counted with different weightings. By construction the training data herein is structured into blocks where only within species interactions are considered, and each HisKA and receiver form a single exclusive partnership (see Figure 6.1). Ignoring the random sampling for cross-validation, this means that $P\left[A_k=\alpha|\,Y=1\right]$ counts each HisKA residue once, while $P\left[A_k=\alpha|\,Y=0\right]$ counts each once per non-partner receiver in the same species, and $P\left[A_k=\alpha\right]$ counts each once per receiver in the same species. It is

therefore not unreasonable to expect these three distributions are similar (and likewise for the receiver domains),

$$\begin{aligned} P\left[A_k{=}\alpha \,|\, Y{=}1\right] &\approx P\left[A_k{=}\alpha \,|\, Y{=}0\right] \approx P\left[A_k{=}\alpha\right], \\ P\left[B_k{=}\beta \,|\, Y{=}1\right] &\approx P\left[B_k{=}\beta \,|\, Y{=}0\right] \approx P\left[B_k{=}\beta\right]. \end{aligned} \tag{6.9}$$

In selecting the K column pairs, we expect the amino acids in those columns to be mutually dependant for interacting domain pairs,

$$P\left[A_k = \alpha, B_k = \beta | \, Y{=}1\right] \neq P\left[A_k = \alpha | \, Y{=}1\right] P\left[B_k = \beta | \, Y{=}1\right]. \tag{6.10}$$

However, for non-interacting domains the distributions could be independent,

$$P\left[A_k = \alpha, B_k = \beta | \, Y{=}0\right] \approx P\left[A_k = \alpha | \, Y{=}0\right] P\left[B_k = \beta | \, Y{=}0\right]. \tag{6.11}$$

Given that the training data contains many more non-interactions than interactions, this can be stretched further,

$$P\left[A_k = \alpha, B_k = \beta\right] \approx P\left[A_k = \alpha\right] P\left[B_k = \beta\right]. \tag{6.12}$$

Starting from the definition of S_k^Ω in Equation (6.6), and substituting the results from Equations (6.11) and (6.9), leads to the definition of S_k^ζ in Equation (6.4),

$$\begin{aligned} S_k^\Omega(\alpha,\beta) &= \log\left(\frac{P\left[A_k = \alpha, B_k = \beta \,|\, Y{=}1\right]}{P\left[A_k = \alpha, B_k = \beta | \, Y{=}0\right]}\right) \\ &\approx \log\left(\frac{P\left[A_k = \alpha, B_k = \beta \,|\, Y{=}1\right]}{P\left[A_k = \alpha | \, Y{=}0\right] P\left[B_k = \beta | \, Y{=}0\right]}\right) \\ &\approx \log\left(\frac{P\left[A_k = \alpha, B_k = \beta \,|\, Y{=}1\right]}{P\left[A_k = \alpha | \, Y{=}1\right] P\left[B_k = \beta | \, Y{=}1\right]}\right) \\ &= S_k^\zeta(\alpha,\beta). \end{aligned} \tag{6.13}$$

Thus $S_k^\Omega(\alpha,\beta) \approx S_k^\zeta(\alpha,\beta)$.

Similarly, from the definition of S_k^η in Equation (6.5), substituting the results from Equations (6.12) and (6.9), also leads to S_k^ζ,

$$\begin{aligned} S_k^\eta(\alpha,\beta) &= \log\left(\frac{P\left[A_k = \alpha, B_k = \beta \,|\, Y{=}1\right]}{P\left[A_k = \alpha, B_k = \beta\right]}\right) \\ &\approx \log\left(\frac{P\left[A_k = \alpha, B_k = \beta \,|\, Y{=}1\right]}{P\left[A_k = \alpha\right] P\left[B_k = \beta\right]}\right) \\ &\approx \log\left(\frac{P\left[A_k = \alpha, B_k = \beta \,|\, Y{=}1\right]}{P\left[A_k = \alpha | \, Y{=}1\right] P\left[B_k = \beta | \, Y{=}1\right]}\right) \\ &= S_k^\zeta(\alpha,\beta) \end{aligned} \tag{6.14}$$

giving $S_k^\eta(\alpha,\beta) \approx S_k^\zeta(\alpha,\beta)$. Thus S_k^ζ, S_k^η and S_k^Ω could be expected to be approximately equal for a large training dataset.

6.2.4 Restricted models

Under the assumption that the K interaction pair scores are independent, Bayes Theorem provides an elegant interpretation of $\sum S_k^\eta$ and $\sum S_k^\Omega$ in terms of the probability of interaction Y given the K amino acid pairs (α_k and β_k for $k = 1, 2, \ldots, K$):

$$\log\left(P\left[Y=1|A_k=\alpha_k, B_k=\beta_k \text{ for all } k\right]\right) = \log\left(P\left[Y=1\right]\right) + \sum_{k=1}^{K} S_k^\eta(\alpha_k, \beta_k), \tag{6.15}$$

$$= \log\left(\frac{P\left[Y=1\right]}{P\left[Y=0\right]}\right) + \sum_{k=1}^{K} S_k^\Omega(\alpha_k, \beta_k). \tag{6.16}$$

Proof. Starting from Bayes Theorem, and applying the independence assumption,

$$P\left[Y=1|A_k=\alpha_k, B_k=\beta_k \text{ for all } k\right] = \frac{P\left[Y=1\right] P\left[A_k=\alpha_k, B_k=\beta_k \text{ for all } k|Y=1\right]}{P\left[A_k=\alpha_k, B_k=\beta_k \text{ for all } k\right]}$$

$$= P\left[Y=1\right] \prod_{k=1}^{K} \frac{P\left[A_k=\alpha_k, B_k=\beta_k|Y=1\right]}{P\left[A_k=\alpha_k, B_k=\beta_k\right]}. \tag{6.17}$$

Taking logarithms and using the definition of S_k^η in Equation (6.5), gives

$$\log\left(P\left[Y=1|A_k=\alpha_k, B_k=\beta_k \text{ for all } k\right]\right) = \log\left(P\left[Y=1\right]\right) + \sum_{k=1}^{K} \log\left(\frac{P\left[A_k=\alpha_k, B_k=\beta_k|Y=1\right]}{P\left[A_k=\alpha_k, B_k=\beta_k\right]}\right)$$

$$= \log\left(P\left[Y=1\right]\right) + \sum_{k=1}^{K} S_k^\eta(\alpha_k, \beta_k). \tag{6.18}$$

as claimed in Equation (6.15).

A derivation similar to that of Equation (6.17) gives

$$P\left[Y=0|A_k=\alpha_k, B_k=\beta_k \text{ for all } k\right] = P\left[Y=0\right] \prod_{k=1}^{K} \frac{P\left[A_k=\alpha_k, B_k=\beta_k|Y=0\right]}{P\left[A_k=\alpha_k, B_k=\beta_k\right]}. \tag{6.19}$$

Using Equations (6.17) and (6.19), together with the definition of S_k^Ω in Equation (6.6) gives

$$\text{logit}\left(P\left[Y=1|A_k=\alpha_k, B_k=\beta_k \text{ for all } k\right]\right) = \log\left(\frac{P\left[Y=1|A_k=\alpha_k, B_k=\beta_k \text{ for all } k\right]}{P\left[Y=0|A_k=\alpha_k, B_k=\beta_k \text{ for all } k\right]}\right)$$

$$= \log\left(\frac{P\left[Y=1\right]}{P\left[Y=0\right]}\right) + \sum_{k=1}^{K} \log\left(\frac{P\left[A_k=\alpha_k, B_k=\beta_k|Y=1\right]}{P\left[A_k=\alpha_k, B_k=\beta_k|Y=0\right]}\right)$$

$$= \log\left(\frac{P\left[Y=1\right]}{P\left[Y=0\right]}\right) + \sum_{k=1}^{K} S_k^\Omega(\alpha_k, \beta_k) \tag{6.20}$$

as claimed in Equation (6.16). □

Equations (6.15) and (6.16) have the form of a GLM with equal weights, with log and logit link-functions respectively. Hence, the following restricted GLMs were also formulated for these scores (and the closely related S_k^ζ score), with only two parameters w_0 and $w_1 = w_k$ for $k = 2, 3, \ldots, K$,

$$\text{log}(P[Y_i = 1 | X_{ik} \text{ for all } k]) = w_0 + w_1 \sum_{k=1}^{K} X_{ik} + \epsilon_i \tag{6.21}$$

where for $X_{ik} = \hat{S}_k^\eta(\alpha_k, \beta_k)$ it would be expected $w_0 \approx \log(P[Y=1])$ and $w_k \approx 1$, and

$$\text{logit}(P[Y_i = 1 | X_{ik} \text{ for all } k]) = w_0 + w_1 \sum_{k=1}^{K} X_{ik} + \epsilon_i \tag{6.22}$$

where for $X_{ik} = \hat{S}_k^\Omega(\alpha_k, \beta_k)$ it would be expected that $w_0 \approx \log\left(\frac{P[Y=1]}{P[Y=0]}\right)$ and $w_k \approx 1$.

The equal weight models in Equations (6.21) and (6.22) will be referred to as as *restricted GLMs*, while Equation (6.1) is an *unrestricted GLM*. The particular sub-case $\sum S_k^\zeta$ is used directly with a threshold in White *et al.* (2007) to rank protein-protein interactions.

6.2.5 Model assessment

Baldi *et al.* (2000) reviews a range of model assessment criteria, including receiver operator characteristic (ROC) curves Fawcett (2003). These plot the true positive rate against the false positive rate, which when done with a moving threshold gives a line ranging from the bottom left corner $(0, 0)$ (model predicts everything is false) to the top right corner $(1, 1)$ (model predicts everything is true). In good models, the curve will be well above the diagonal, ideally reaching close to the top left corner $(0, 1)$ giving every true positive prediction with no false positives. Calculating the area under an ROC curve gives a simple assessment of the model performance, typically in the range 0.5 (random) to 1 (perfect). A model with an ROC area less than 0.5 is worse than random, with an area of 0 possible for inverted perfect model (always wrong).

To compare the predictive performance of the different models, the list of known interactions and non-interactions was divided randomly into a *training set* (typically 80%) used for fitting and a *test set* (20%) used for out-of-sample prediction. The area under the ROC curve was then calculated, and the procedure repeated five times, using a different random split each time.

In addition, predictions were made for the interactions among the TCS proteins of model organisms, *E. coli* K-12 (NC_000913), *Bacillus subtilis* (NC_000964) and *Caulobacter crescentus* (NC_002696).

6.3 Implementation

Paired TCS HisKA and receiver domains were compiled from 340 prokaryote species (excluding multiple strains, Table A.1) as described in Chapter 2, comprising $3,473$ cases from neighbouring HK and RR genes, and $1,434$ hybrid kinases (single genes containing both a HisKA and receiver). The $4,907$ paired domains were taken as positive interactions ($Y = 1$) while the $114,763$ other possible inter-species combinations of these domains were taken as non-interactions ($Y = 0$). Many of the figures shown in this chapter draw on a subset containing just the $3,473$ two gene pairs and their $55,046$ presumed non-interactions (inter-species combinations). Similarly, a third dataset was given by considering only the $1,434$ hybrid gene domain pairs, and their $19,726$ other inter-species combinations. For the initial model cross-validation, 80% of the presumed interactions and 80% of the presumed non-interactions were taken as the training dataset, with the remaining 20% held out as a test dataset.

Note that only the HisKA domain from the $\mathrm{T_i}$ was used, and not the HATPase, as it does not seem to be important for the kinase-receiver interaction (Ohta and Newton, 2003). This was supported by an initial investigation of the MI scores where a MSA for the full $\mathrm{T_i}$ domain was used (data not shown), also observed in Skerker *et al.* (2008).

As in Chapter 5, two different alignment programs were used to generate the MSAs, CLUSTAL W version 1.83 with its default settings, and MUSCLE version 3.7 with a maximum of three iterations (`-maxiter 3`), and the alignments output using the CLUSTAL W file format (`-clwstrict`). MI was calculated between columns of the two MSAs by pairing rows for the known interactions in the training data, treating any gap characters as another amino acid. The columns pairs were then ranked by their MI, and the top K highest scoring couplets were selected as input to the model via one of the described scoring functions.

One numerical complication for some of the probabilistic scores is the logarithm of zero (and potentially also the undefined ratio $\frac{0}{0}$) can occur when the amino acid pair (A_{ik}, B_{ik}) has not been observed in an interacting pair. In this situation, for $\hat{S}_k^M$ the natural limit value zero was used, while for $\hat{S}_k^\zeta$, $\hat{S}_k^\eta$ and $\hat{S}_k^\Omega$ the minimum observed score of any amino acid pair was taken. In the contrary situation where (A_{ik}, B_{ik}) has *only* been observed in interacting pairs, for $\hat{S}_k^\Omega$ we have a non-zero numerator with a zero denominator, and instead take the *maximum* observed score of any amino acid pair from that column pair.

The use of an alternative power transformation in S_k^ζ, square root in place of the natural logarithm, made minimal difference, as did choosing an ad-hoc value of zero (square root only) or one for undefined ratios where amino acids had not been observed in the training data (data not shown). The use of a log link function, as suggested by Equation (6.15), proved

numerically intractable.

GLM parameters (weights w_k and intercept term w_0) were estimated using the `glm` function in the statistical programming language R (R Development Core Team, 2007). The ROCR library (Sing *et al.*, 2005) was used to generate ROC curves and their areas. The process was scripted in Python (www.python.org) using RPy (Moreira and Warnes, 2003) and Biopython (www.biopython.org). Figures were drawn using ReportLab (www.reportlab.org) or R.

6.4 Results

Figure 6.2 shows plots of the AIC, and the performance of the models assessed using the area under the ROC curve, against K for five random divisions of the full dataset using CLUSTAL W MSAs. AIC is a measure of the trade-off between the model fit and the number of parameters, with the aim of minimising the AIC. The area under a ROC curve is a measure of predictive performance – a perfect model would give an area of 1, a random model 0.5. In-sample predictions are naturally more successful (have a higher ROC area) than out-of-sample (previously unseen test data), but even here the predictions are substantially better than random.

As expected, for all scoring functions the performance of the unrestricted GLM, Equation (6.1), increases with the number of terms K. This begins to plateau at $K \geq 30$, but has still not quite saturated at $K = 100$. Increasing K further yields only marginal improvements at increased computational cost (data not shown). While model assessments such as the distribution of residuals suggest the fit could be improved, nevertheless there is good predictive power when assessed on the previously unseen test data (out-of-sample predictions). The S_k^{Ω} and S_k^{η} scores perform best (test ROC area ≥ 0.87 (2sf) for $K \geq 30$), with little to choose between them, followed by S_k^{ζ} (≥ 0.81), S_k^{M} (≥ 0.75), S_k^{P} (≥ 0.72), S^{H} (≥ 0.72) and S^{V} (≥ 0.67) scores in decreasing order – all better than a random model (area ≈ 0.5). The same ordering of the scores is seen using the AIC, where again the S_k^{Ω} and S_k^{η} scores are almost indistinguishable.

The restricted GLM (where the K scores are summed with equal weighting) with a log link-function, Equation (6.21), proved numerically intractable. However, with a logit-link function the restricted GLMs for the S_k^{Ω}, S_k^{η} and S_k^{ζ} scores, Equation (6.22), performed almost indistinguishably from the unrestricted GLM (Equation 6.1) for $K < 20$, but show a noticeably lower ROC area with more terms. Impressively for such simple models, these still give a test ROC area ≥ 0.85 for S_k^{Ω} and S_k^{η}, and ≥ 0.73 for S_k^{ζ} for $K \geq 30$ (Figure 6.2). The AIC also

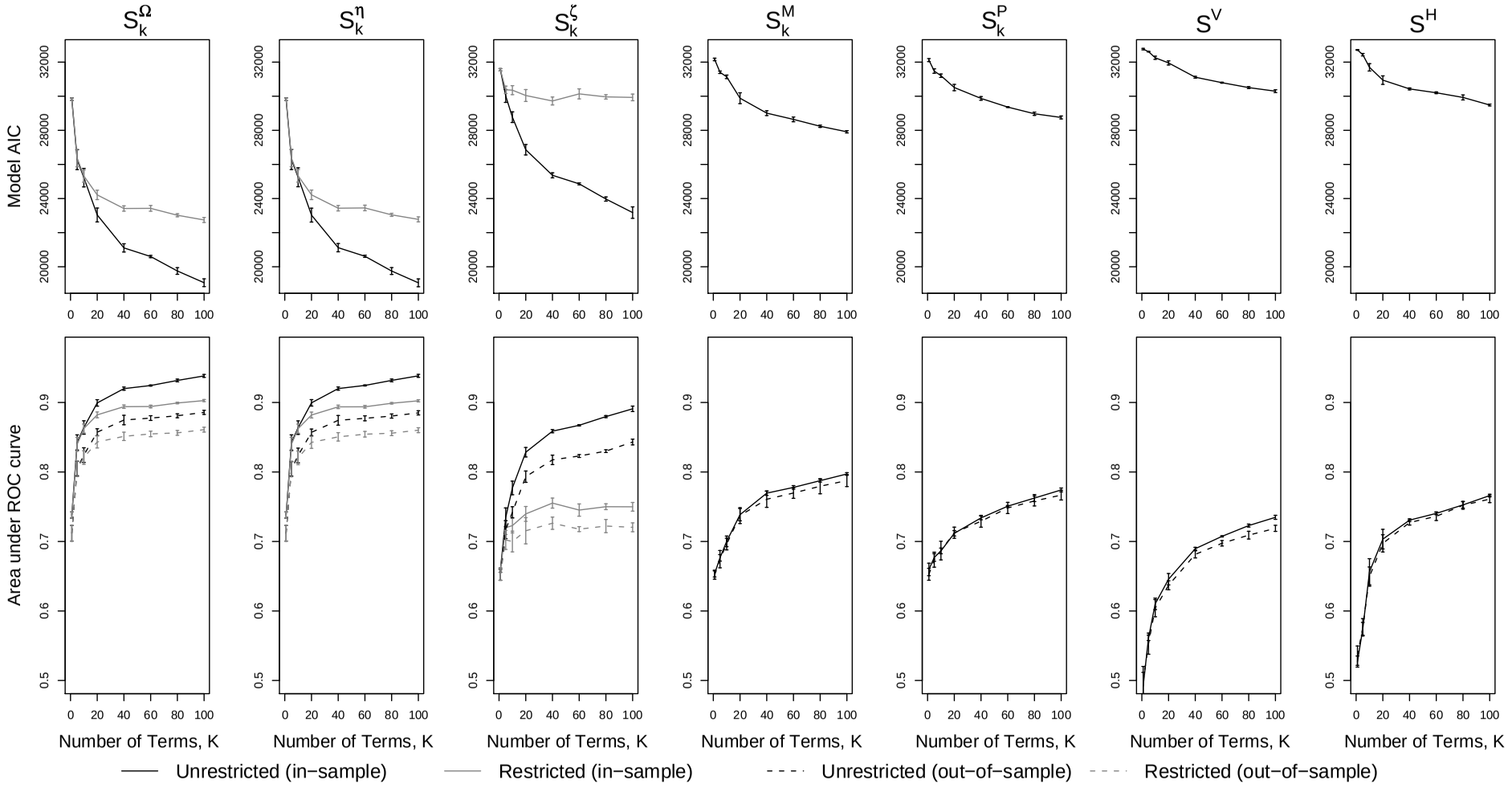


Figure 6.2: Model performance on full dataset, with 80% for training, using CLUSTAL W MSAs. The points are median values, while the error bars show one standard deviation, based on five randomised splits of 80% training ($3,925$ positive and $91,788$ non-interacting samples) and 20% test (982 and $22,948$ samples). First row of figures shows model AIC (fitting the training data), second row shows area under ROC curve for predictions on both the training data (in-sample, solid lines) and unseen test data (out-of-sample, dashed lines). Unrestricted GLM shown in black, Equation (6.1), restricted equal weight GLM in grey, Equation (6.22) (S_k^Ω, S_k^η and S_k^ζ only).

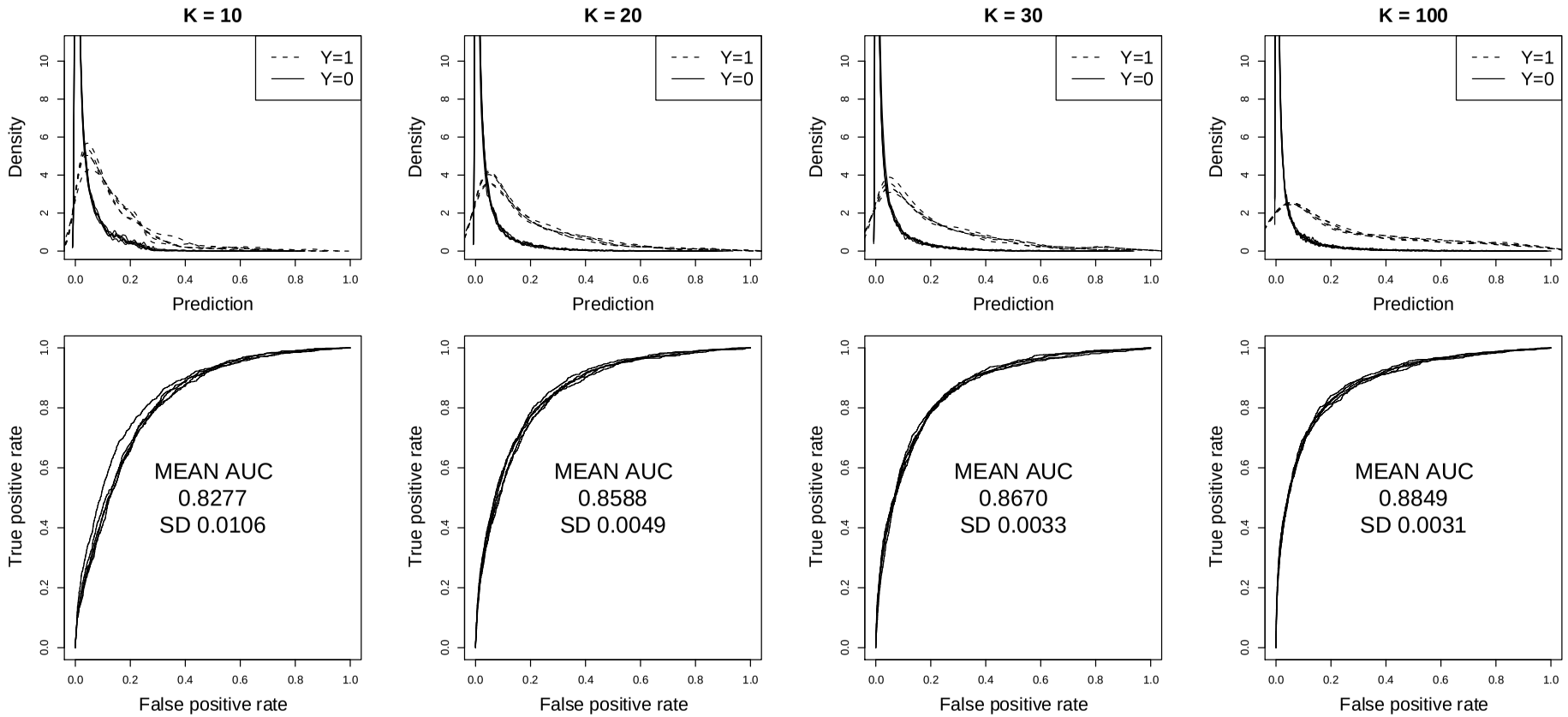


Figure 6.3: Unrestricted model performance on full dataset, with 80% for training, using CLUSTAL W MSAs and S_k^{Ω} for $K = 10$, 20, 30 and 100 (from left to right). Five random splits of the data into training and test are shown. The top row of figures show the predicted interaction probabilities for the expected interactions ($Y = 1$) and non-interactions ($Y = 0$), showing the class separation improves with higher K. The bottom row of figures are ROC plots, labelled with the mean area under the curve, where the area increases as K is increased.

suggests that the unrestricted GLM gives a better fit. As might be expected, the equal-weight restriction was not useful on the remaining scores (data not shown).

Of the seven scores explored, S_k^{Ω} and S_k^{η} give the best predictive performance, both in the restricted and unrestricted GLM, with little to chose between them. Given the choice of a logit link function, S_k^{Ω} was selected as the preferred score because of the simple Bayesian interpretation of $\sum S_k^{\Omega}$ (Section 6.2.4).

Figure 6.3 explores the performance of the S_k^{Ω} score in more detail, for the five random divisions of the full dataset. This figure is divided into four columns for $K = 10, 20, 30$ and 100. The top row of plots shows the range of predictions for known interactions ($Y = 1$) and non-interactions ($Y = 0$), showing the class separation improves with higher K but that even at $K = 100$ there are a number of false positives. The bottom row of plots shows the corresponding ROC curves.

Figure B.1 shows the same set of results as Figure 6.2 but from MUSCLE MSAs rather than CLUSTALW W MSAs. By eye, the two figures are practically identical, showing that the details of the alignment algorithm play a much smaller role that the choice of scoring system and the number of terms, K.

The performance of the models trained and evaluated on two-gene data (Figures 6.4 and 6.5), and the trends therein, is broadly similar to that using the full dataset (Figures 6.2 and 6.3), and marginally more successful. By contrast, the models trained and evaluated on hybrid kinase data (Figures 6.6 and 6.7) show markedly less predictive power, particularly on the test data (out-of-sample predictions). These differences are not simply due to the small training sample effect, as shown by Figures B.2 and B.4 where the combined and two-gene datasets were split to give a similarly sized training set to that of the HY dataset (Figure 6.6) and still show better predictive performance. Otherwise comparing the individual scores, the same general trends persist in the HY dataset predictions.

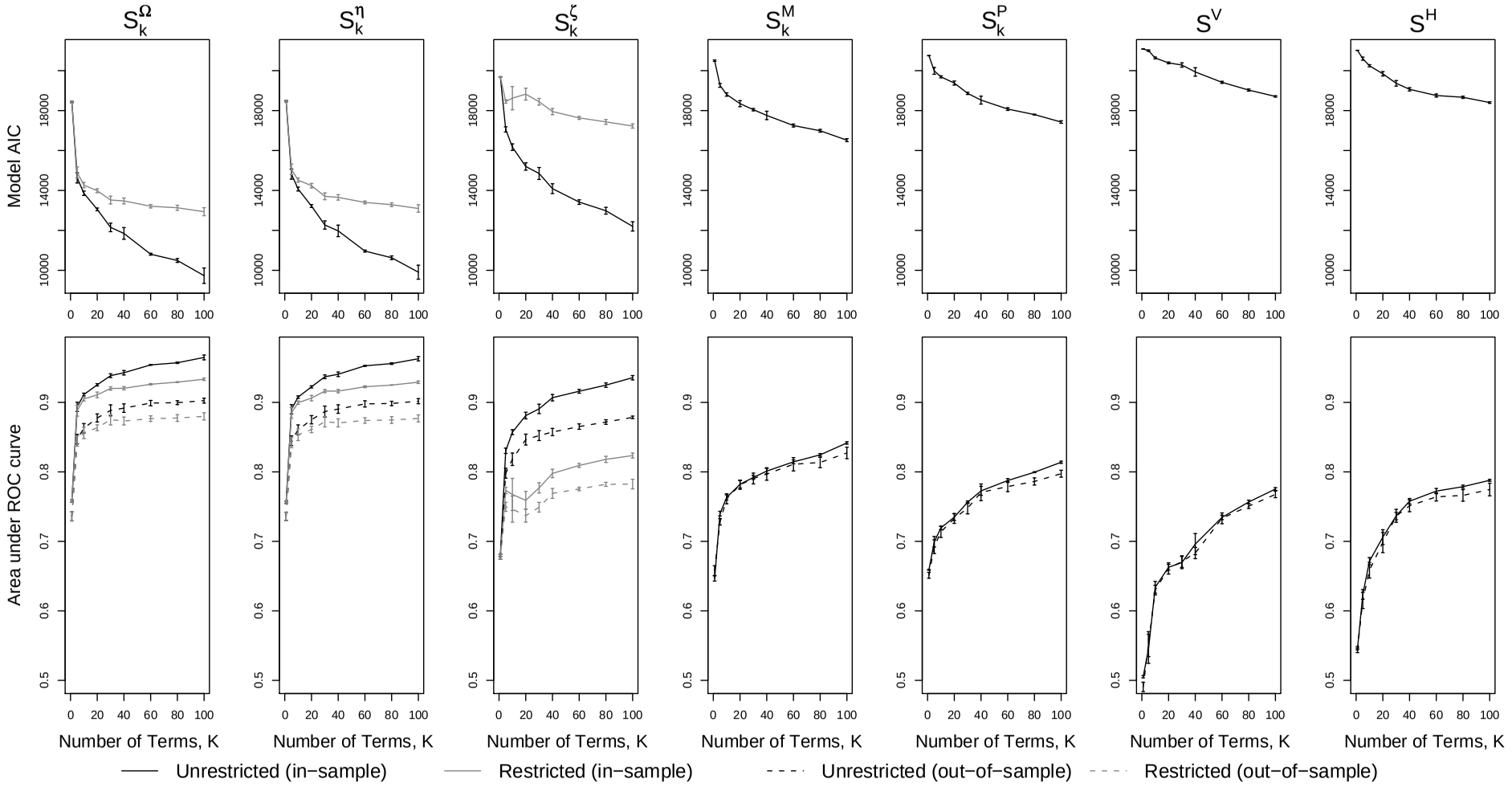


Figure 6.4: Model performance using only conventional two gene TCS systems, 80% for training, CLUSTAL W MSAs. Five randomised splits, trained on 80% ($2,778$ positive interactions and $44,036$ non-interactions), tested on 20% (695 and $11,010$). Key as per Figure 6.2, *cf.* Figure 6.6.

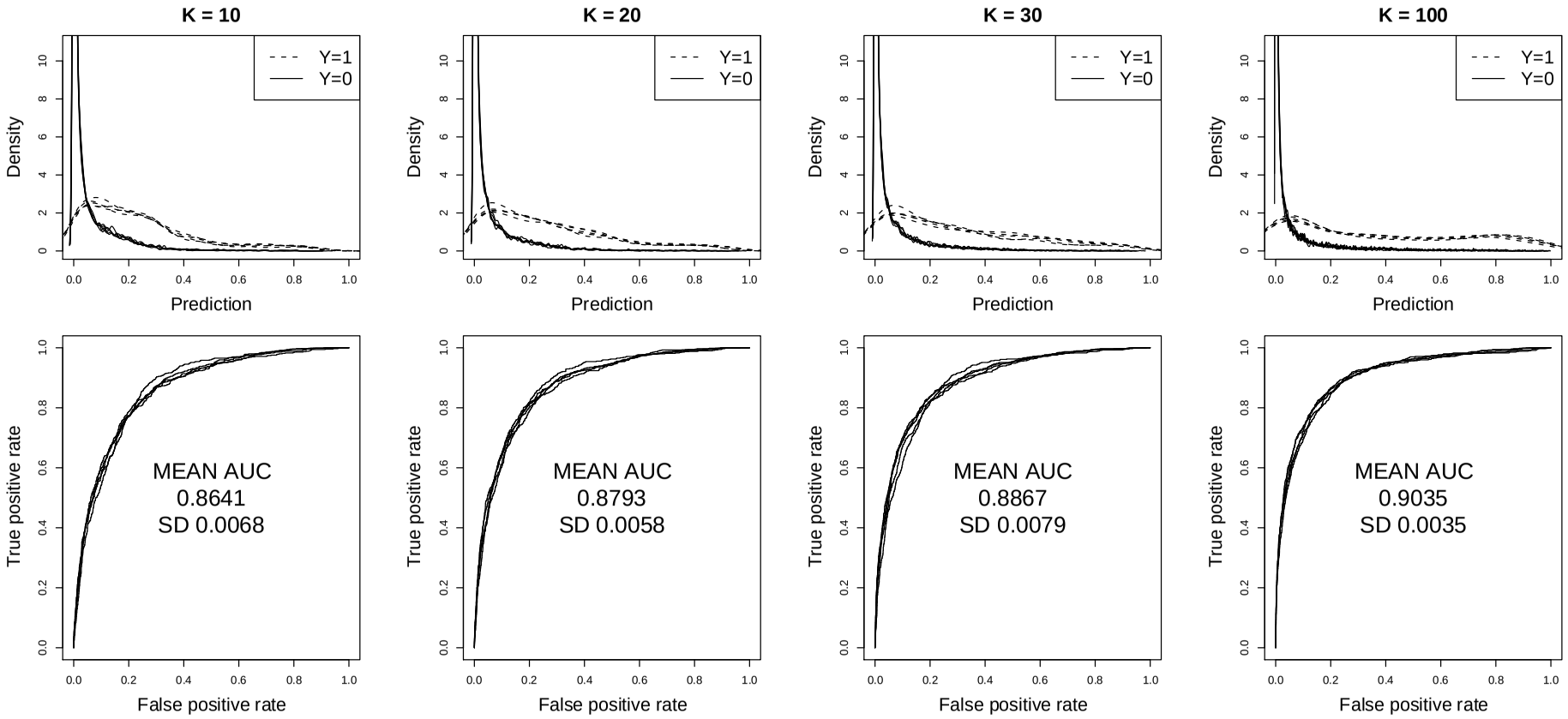


Figure 6.5: Unrestricted model performance using only conventional two gene TCS systems, with 80% for training, using CLUSTAL W MSAs and S_k^{Ω} for $K = 10$, 20, 30 and 100 (from left to right). Five random splits of the data into training and test are shown. The top row of figures show the predicted interaction probabilities for the expected interactions ($Y = 1$) and non-interactions ($Y = 0$), showing the class separation improves with higher K. The bottom row of figures are ROC plots, labelled with the mean area under the curve, where the area increases as K is increased.

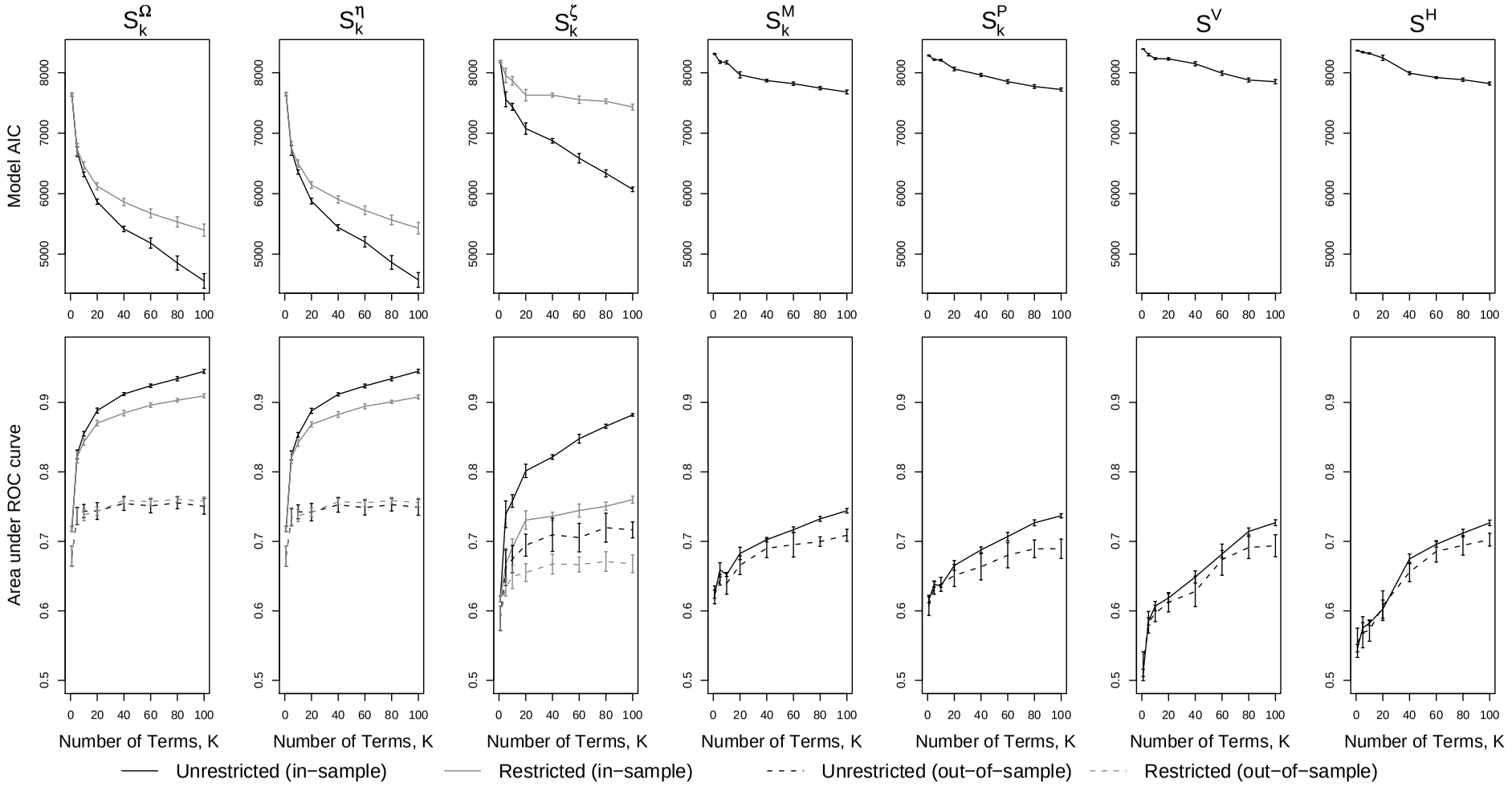


Figure 6.6: Model performance using only HYs (TCS genes with both HisKA and receiver), CLUSTAL W MSAs. Five randomised splits, trained on 80% ($1,147$ positive interactions and $15,780$ non-interactions), tested on 20% (287 and $3,946$). Key as per Figure 6.2, *cf.* Figure B.4.

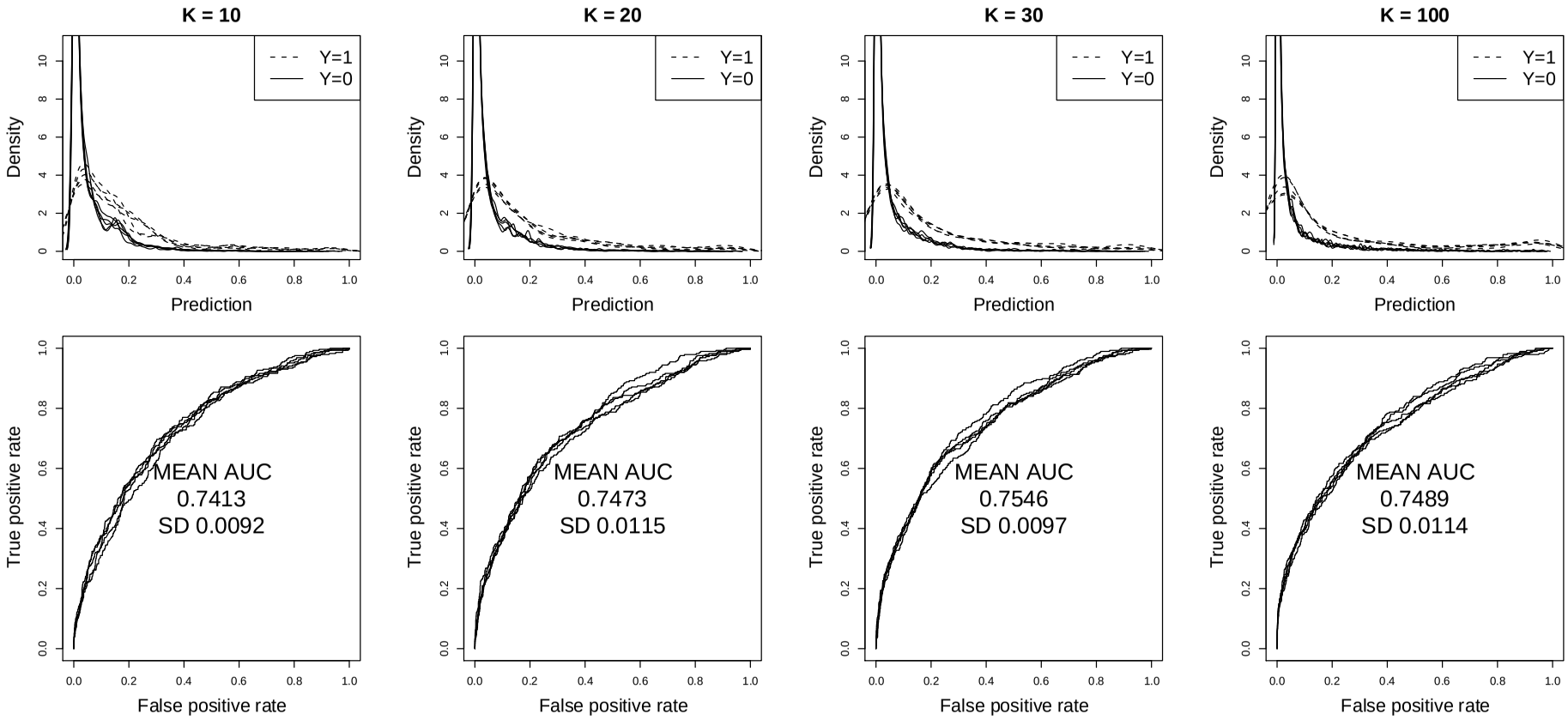


Figure 6.7: Unrestricted model performance using only only HYs (TCS genes with both HisKA and receiver), with 80% for training, using CLUSTAL W MSAs and S_k^{Ω} for $K = 10$, 20, 30 and 100 (from left to right). Five random splits of the data into training and test are shown. The top row of figures show the predicted interaction probabilities for the expected interactions ($Y = 1$) and non-interactions ($Y = 0$), showing the class separation improves with higher K. The bottom row of figures are ROC plots, labelled with the mean area under the curve, where the area increases as K is increased.

6.5 Application to *Escherichia coli*

Figure 6.8 shows the prediction performance for *E. coli* K-12, where the training data was the full dataset excluding excluding *Escherichia*, *Salmonellae*, *Shigellas* and *Yersinias*. The in-sample performance is weaker than that in Figure 6.2, perhaps due to certain *E. coli* proteins disrupting the MSAs. More strikingly, the out-of-sample predictions for the *E. coli* interactions are better, and plateau at only $K = 20$ terms for S_k^Ω, S_k^η and S_k^ζ. This suggests *E. coli* is very similar to the training data average. Perhaps this is a reflection of sampling bias towards laboratory-cultured bacteria.

Figure 6.9 explores the performance of the unrestricted model using the S_k^Ω in more detail, showing the class separation and the actual ROC curves for $K = 10$, 20, 30 and 100. From left to right (increasing the number of terms, K), the class separation and ROC area both improve. These figures show that with $K \geq 20$, almost all the true positives can be identified with the false negative rate at only around 20%.

Figure 6.10 shows the predictions for *E. coli* using S_k^Ω and $K = 100$ with the unrestricted GLM (Equation 6.1). To answer the question "What is the partner for a given protein?", we can focus on a single row or column (where the top score is marked with a vertical or horizontal bar). The top scoring receiver is a known partner for $16/28$ HisKA domains, and an established HisKA partner is identified for $16/28$ receiver domains.

Alternatively, looking at this matrix with a global threshold allows us to identify potential crosstalk between systems. For instance, a threshold of 0.25 (red squares) identifies 5 interactions in addition to $17/30$ known pairings. Of these unexpected interactions, two suggest potential coupling of the YpdA-YpdB system with YehU-YehT system, neither of which has an apparent phenotype, raising the possibility that these are redundant through crosstalk. There are other noticeable groupings of HisKAs and receivers with similar predicted interaction profiles, in particular NarQ-NarP and NarX-NarL (plus perhaps UhbB-UhpA). It is well established that the Nar systems intercommunicate (Rabin and Stewart, 1992), while some interactions with the Uhp system have also been demonstrated *in vitro* (Yamamoto *et al.*, 2005). The remaining strong false positives ($P > 0.5$) are CusS/YedW, RstB/OmpR and QseC/BasR.

Unorthodox TCS HKs which contain an Hpt domain rather than a HisKA, such as CheA, were excluded from this analysis. Thus the lack of predicted partners for CheB and CheY fits with expectations. Similarly, the phosphotransfer Hpt domains in the tripartite T_i-R-H + R systems (Section 1.4.6) were also excluded, thus no interactions are expected for RRs ArcA, EvgS, TorR and UvrY. Other than an apparent false positive for NarQ/UvrY, there

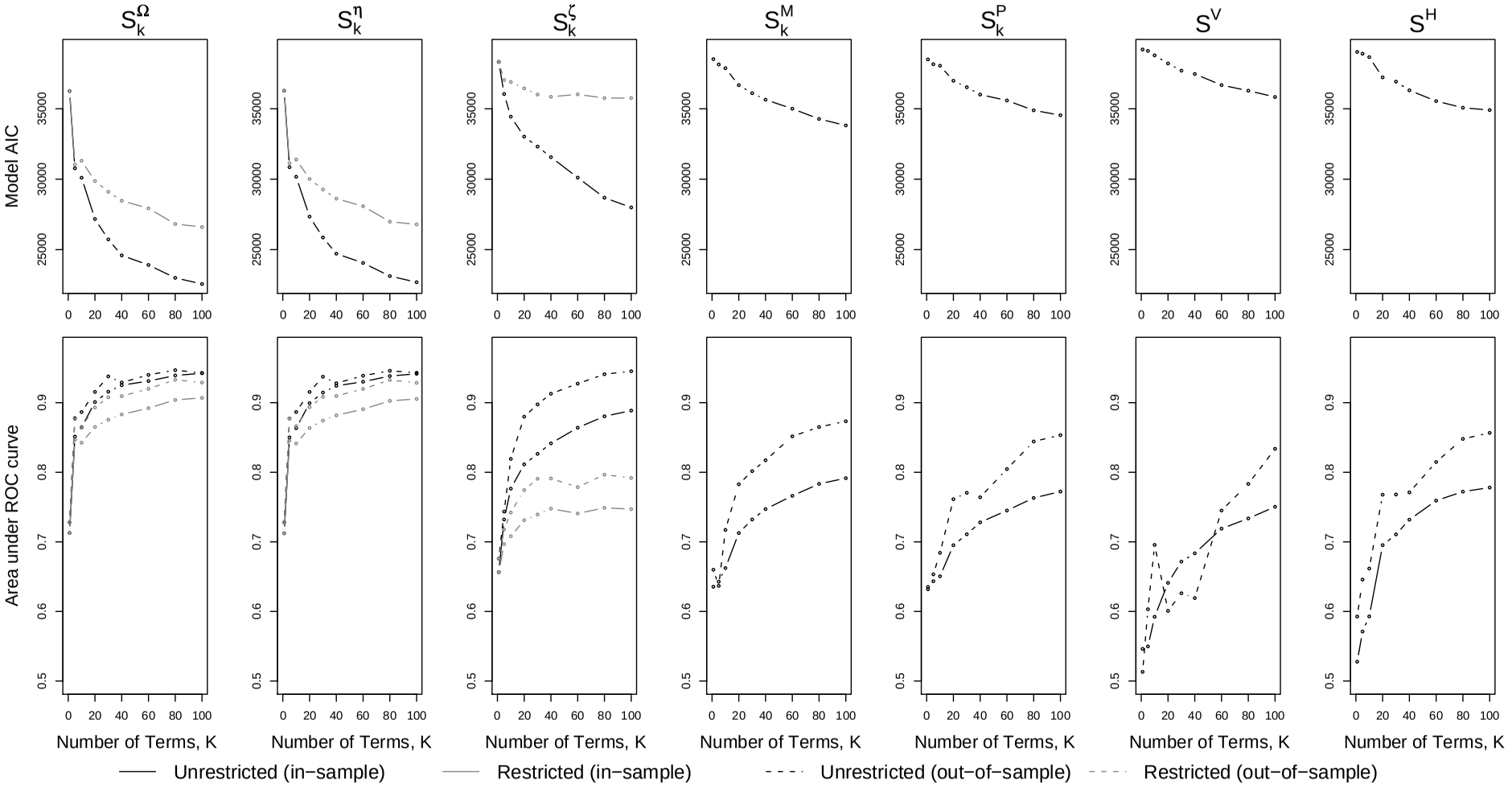


Figure 6.8: Model performance predicting the TCS interactions of *E. coli* (dashed lines, expected interactions shown in Figure 6.10), trained on full dataset excluding *Escherichia*, and the closely related *Salmonellae*, *Shigellas* and *Yersinias* (solid lines). Key as per Figure 6.2.

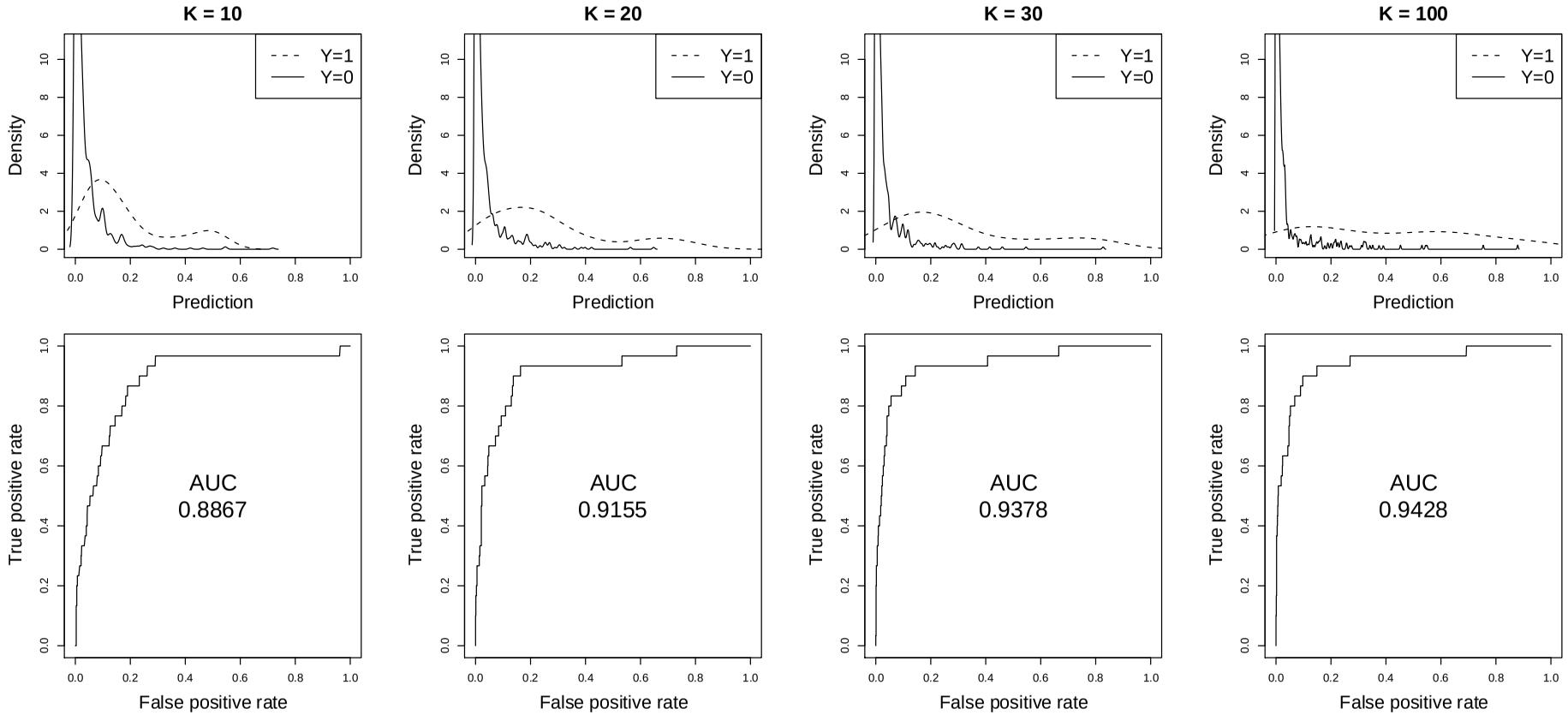


Figure 6.9: Performance of out-of-sample predictions for *E. coli* using S_k^{Ω} for $K = 10$, 20, 30 and 100 (from left to right). Unrestricted model performance is shown predicting the TCS interactions of *E. coli*, trained on the full dataset excluding *Escherichia*, *Salmonellae*, *Shigellas* and *Yersinias*. The top row of figures show the predicted interaction probabilities for the expected interactions ($Y = 1$) and non-interactions ($Y = 0$), showing the class separation improves with higher K. The bottom row of figures are ROC plots, labelled with the area under the curve, where the area increases as K is increased.

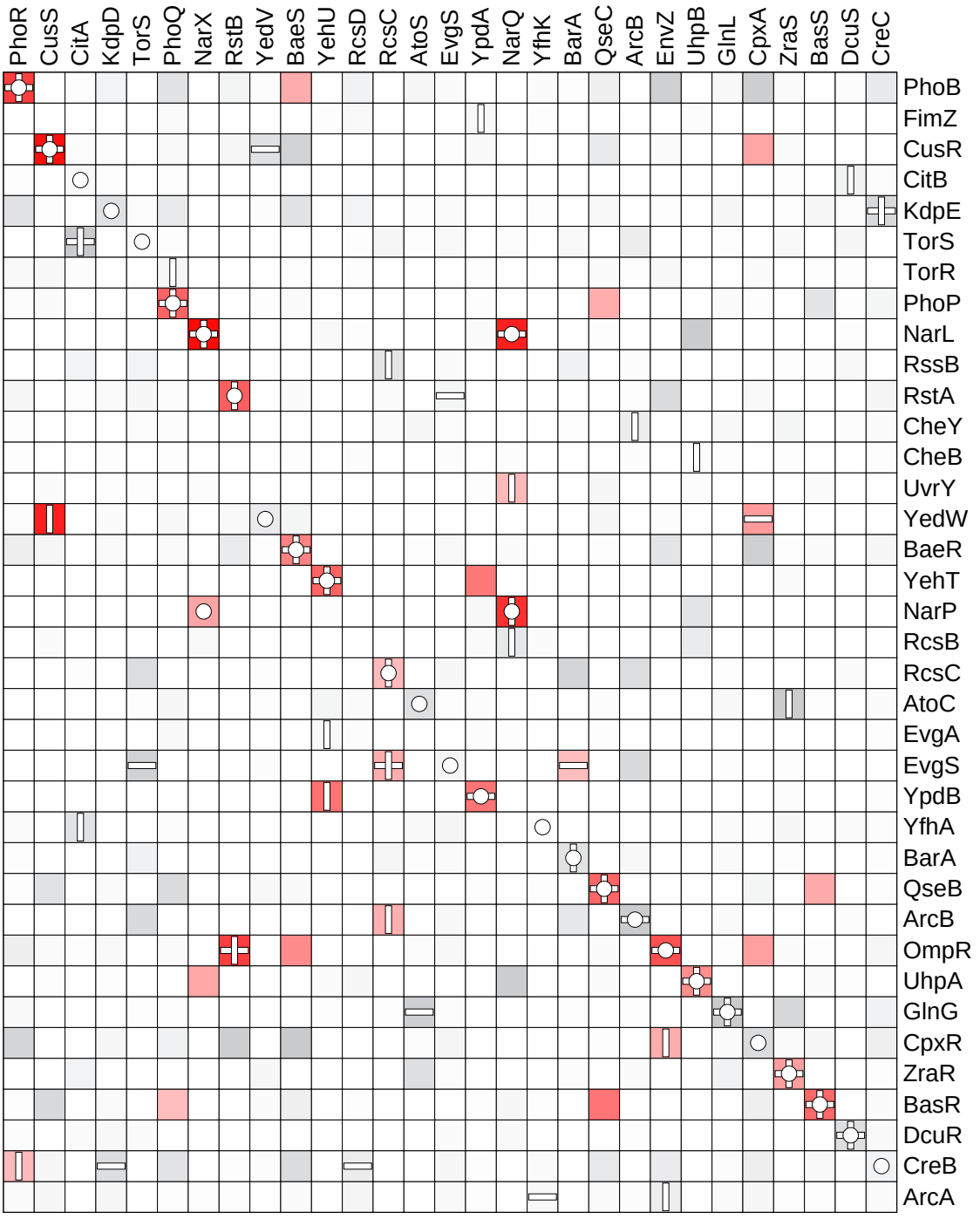


Figure 6.10: Out-of-sample predictions for *E. coli* using S_k^{Ω} and $K = 100$ with an unrestricted GLM. Rows are HisKA domains, columns are receivers, sorted by gene location. Predicted probability of interaction shown by colour (linear scaling from 0.0 as white to 0.25 in pale grey, to 1.0 as red). In each row and column, the highest score is indicated with a vertical or horizontal bar. Cells where the score is the highest in that row and column therefore have a cross-hair shown. White circles show interactions expected from the genome arrangement (which are therefore roughly on the diagonal) and/or the literature (the *Nar* system, see Section 1.4.2).

are no interactions predicted for these RRs. Likewise the only expected interaction from the RcsC/RcsD/RcsB system relay (Section 1.4.7) is between the T_i and R domains of RcsC, which scores highly. However, there are strong predictions for the RcsC transmitter with the receivers of EvgS and ArcB (presumably false positives). On a positive note, there are no predicted interactions with the non-function T_i domain in RcsD.

Finally, there are no strong partnerships predicted for the orphan receivers FimZ and RssB, suggesting their partners (if they have any) are not HisKA domains.

6.6 Application to *Bacillus subtilis*

Figure 6.11 show the various models applied to *Bacillus subtilis*, where here S_k^η and S_k^Ω reach an ROC area of almost 0.9. Figure 6.12 explores the performance of the unrestricted model using the S_k^Ω in more detail, showing the class separation and ROC curves for $K = 10$, 20, 30 and 100. From left to right (increasing the number of terms, K), the class separation and ROC area both improve.

Figure 6.13 shows a grid of the *Bacillus subtilis* predictions for S_k^Ω and $K = 100$. For the HisKA domains, the known partner is the top scoring receiver for $12/28$ cases. Similarly, a known partner is selected for $9/24$ receivers. Using a global threshold of 0.25 (red squares) selects only $11/28$ expected interactions. However, while these interactions score less than 0.25, for all five of KinA to KinE, the top scoring receiver is correctly identified as Spo0F (see Section 1.4.8, *cf.* predictions in supplementary Table 2 of Burger and van Nimwegen (2006) where Spo0F is selected as the most likely partner only for KinA and KinC). Note that there are no strong predictions for the receiver Spo0A, consistent with its only known phosphorylation route via the atypical phosphotransfer protein Spo0B. Additionally, as there are no T_{ii} HK proteins in this dataset, the lack of any predicted partners for RRs CheB, CheY and CheV is also as expected.

This grid also shows a number of HisKA and receiver domains with similar predicted interaction profiles, for example HisKA domains YdfH, YocF, YvqE and YwpD and receiver domains YfiK, YhcZ, YocG, YvqC and DegU could form a multiply connected network. The LytS/LytT system also shows a number of possible cross talk interactions.

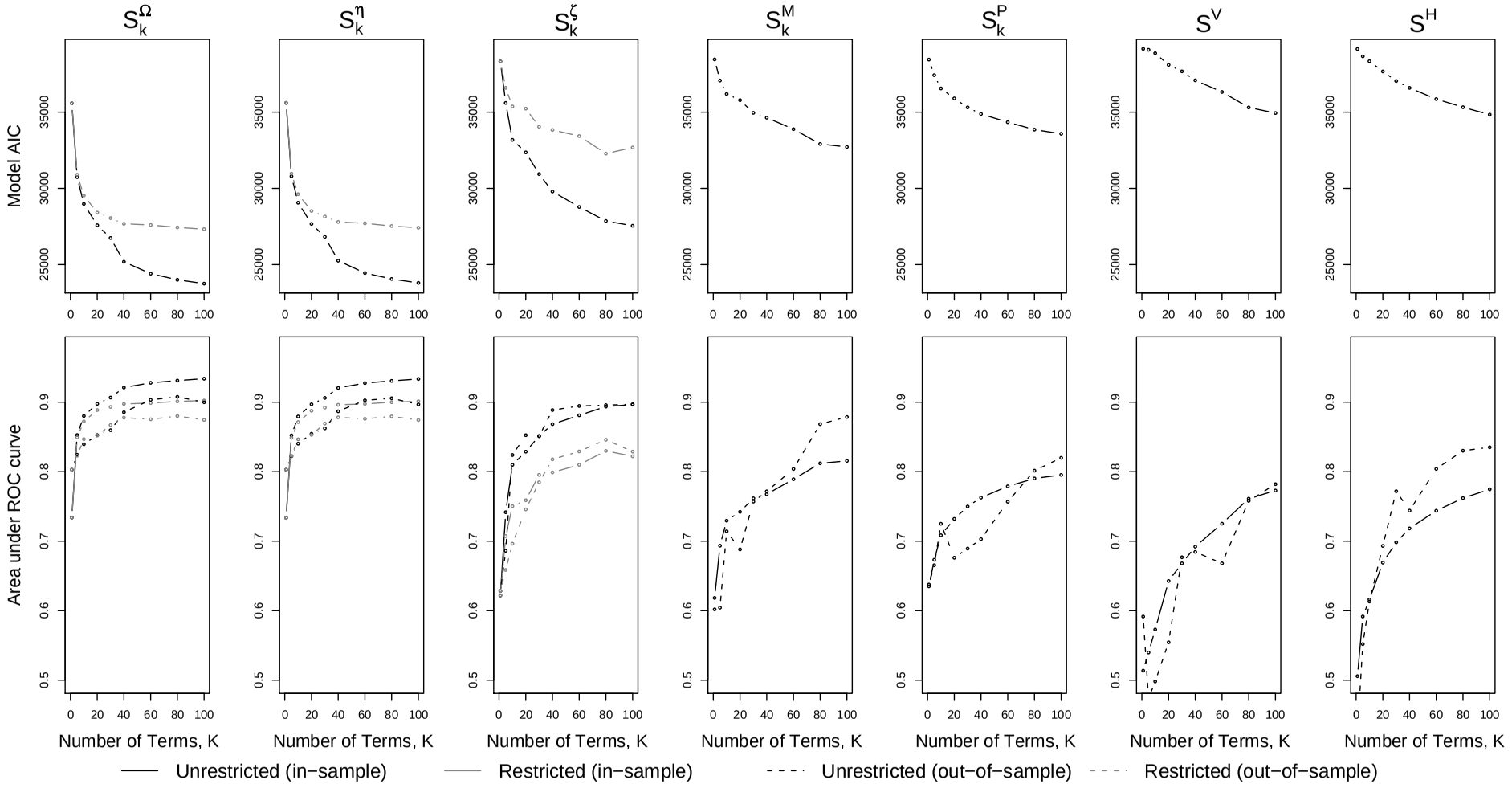


Figure 6.11: Model performance predicting the TCS interactions of *Bacillus subtilis* (dashed lines, expected interactions shown in Figure 6.13), trained on full dataset excluding *Bacillus* (solid lines). Using CLUSTAL W MSAs. Key as per Figure 6.2.

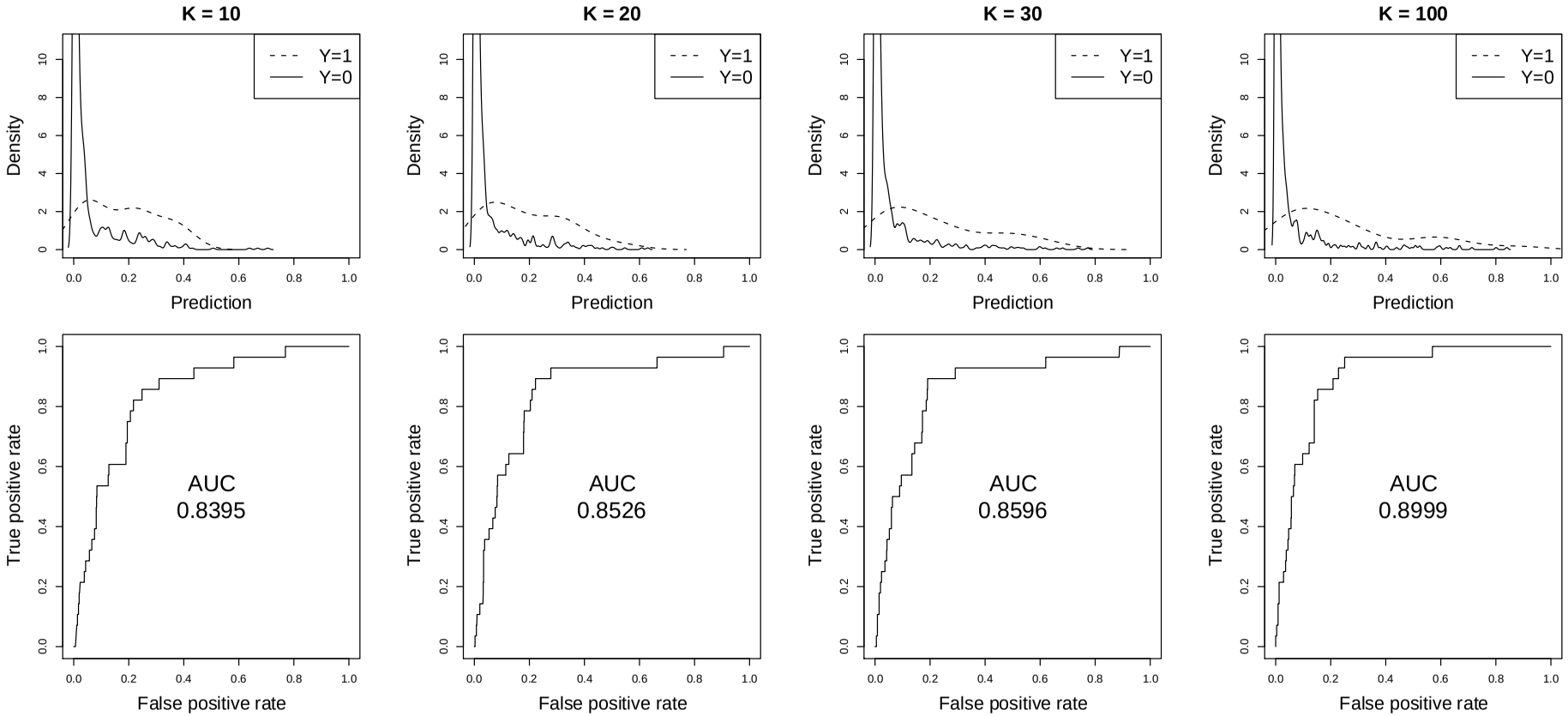


Figure 6.12: Performance of out-of-sample predictions for *Bacillus subtilis* using S_k^{Ω} for $K = 10$, 20, 30 and 100 (from left to right). Unrestricted model performance is shown predicting the TCS interactions of *Bacillus subtilis*, trained on the full dataset excluding *Bacilli*. The top row of figures show the predicted interaction probabilities for the expected interactions ($Y = 1$) and non-interactions ($Y = 0$), showing the class separation improves with higher K. The bottom row of figures are ROC plots, labelled with the area under the curve, where the area increases as K is increased.

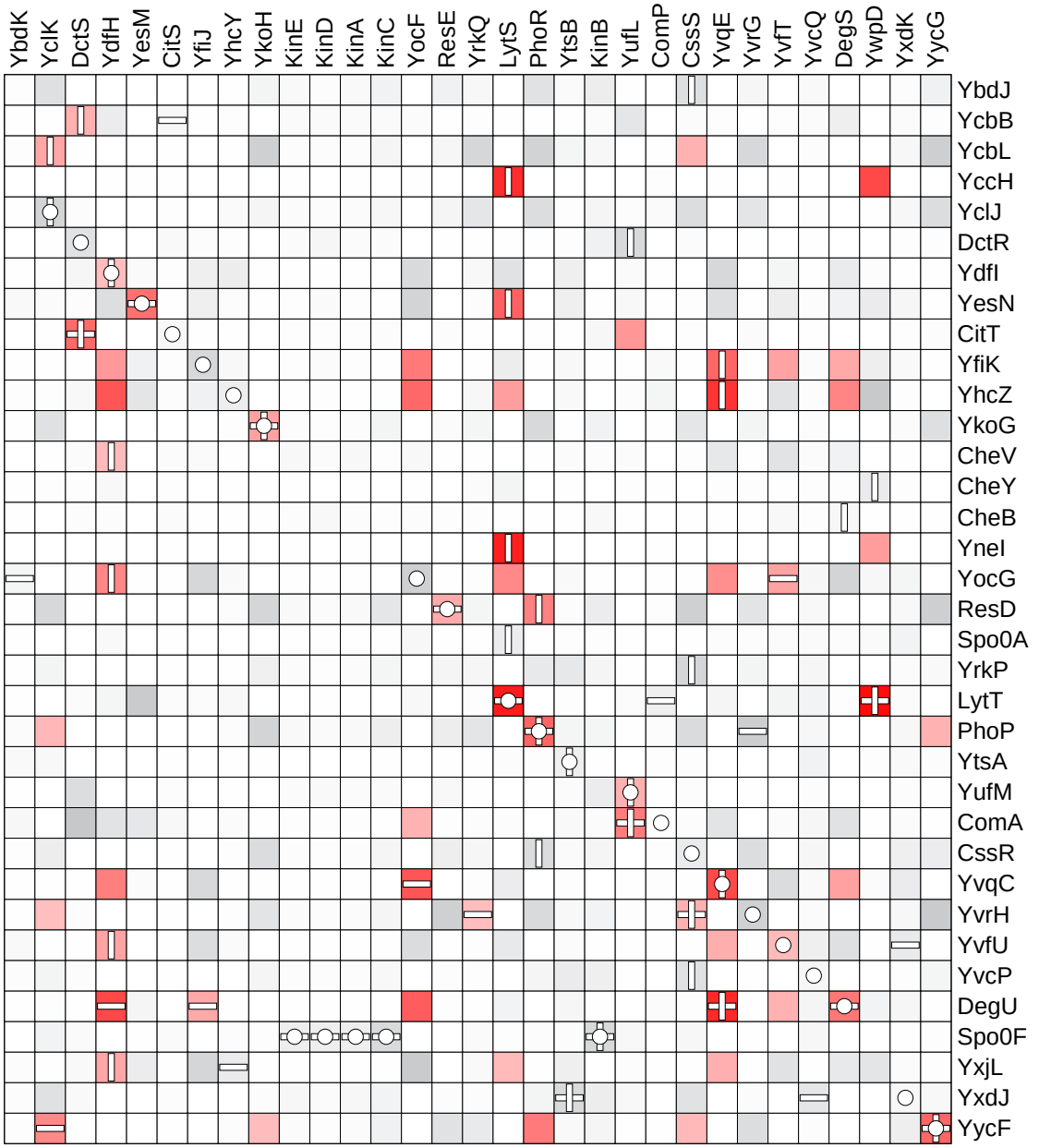


Figure 6.13: Out-of-sample predictions for *Bacillus subtilis* using S_k^{Ω} and $K = 100$ with an unrestricted GLM, using CLUSTAL W MSAs. Rows are HisKA domains, columns are receivers, sorted by gene location. Predicted probability of interaction shown by colour (linear scaling from 0.0 as white to 0.25 in pale grey, to 1.0 as red). In each row and column, the highest score is indicated with a vertical or horizontal bar. Cells where the score is the highest in that row and column therefore have a cross-hair shown. White circles show interactions expected from the genome arrangement (which are therefore roughly on the diagonal) and/or the literature (the *Spo* system, see Section 1.4.8).

6.7 Application to *Caulobacter crescentus*

Based on predictions for other bacteria, the outstanding model performance seen for *E. coli* and *Bacillus subtilis* discussed above with ROC areas around 0.9 for S_k^{Ω} and S_k^{η} is not universal. Figure 6.14 shows predictions for *Caulobacter crescentus* compared to those expected from the genome pairs or Skerker *et al.* (2005). Here the ROC areas for S_k^{Ω} and S_k^{η} reach only around 0.8, shown in more detail in Figure 6.15.

A grid of the unrestricted GLM predictions for S_k^{Ω} with $K = 100$ is shown in Figure 6.16. Using a global threshold of 0.25 (red squares) only $13/41$ expected interactions are identified. However, a number of the "false positives" include domain pairs on the diagonal which did not pass the stringent criteria used to automatically identify domain pairs: CC0238/CC0237, CC0248/CC0247 and NtrY/NtrX (CC1742/CC1743).

Several domains show similar predicted interaction profiles (which could be shown visually with a clustering method) perhaps suggesting they inter-phosphorylate. In particular there is broad cross talk predicted between the HisKA domains from proteins CC0026, CC0921, CC0934, CC2521, CC2670, CC2852, CC2971, CC2988, CC3075, CC3102, CC3191 and CC3219 with the receivers from proteins CC0921, CC0934, CC2852 and CC3102. There is also the possibility of cross talk between the CC2932/PetR (CC2931) and CC1181/CC1182 systems, and between CC0248/CC0247 and CC1768/CC1767.

Remaining unexpected possible predicted interactions include CC3198/CC1150 and CC2755/CC1293, plus nearby genes CC3474/CC3477 which while scoring less that 0.25 are nevertheless each other's predicted top partner.

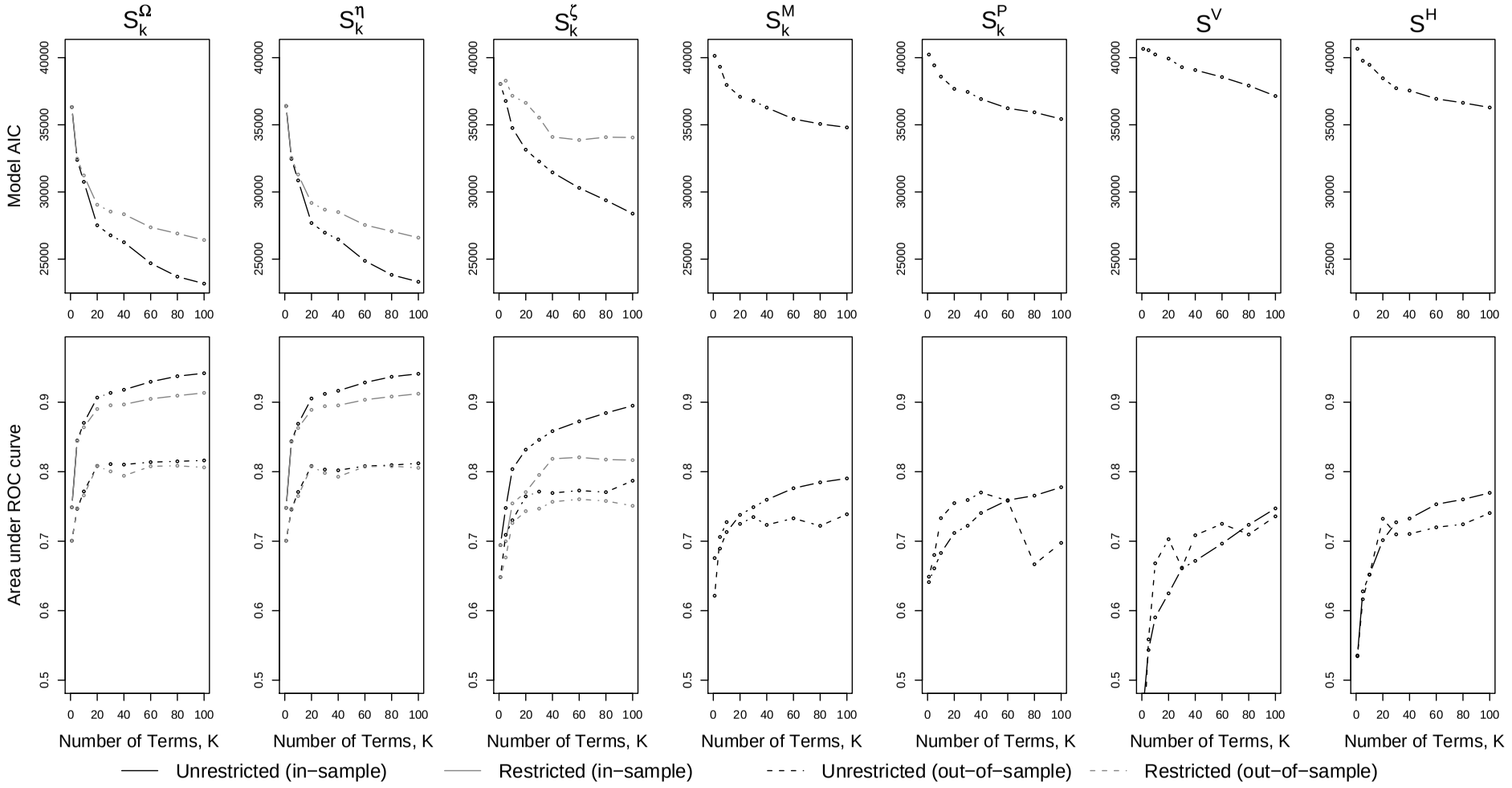


Figure 6.14: Model performance predicting the TCS interactions of *Caulobacter crescentus* (dashed lines, expected interactions shown in Figure 6.16), trained on full dataset excluding *Caulobacteria* (solid lines). Using CLUSTAL W MSAs. Key as per Figure 6.2.

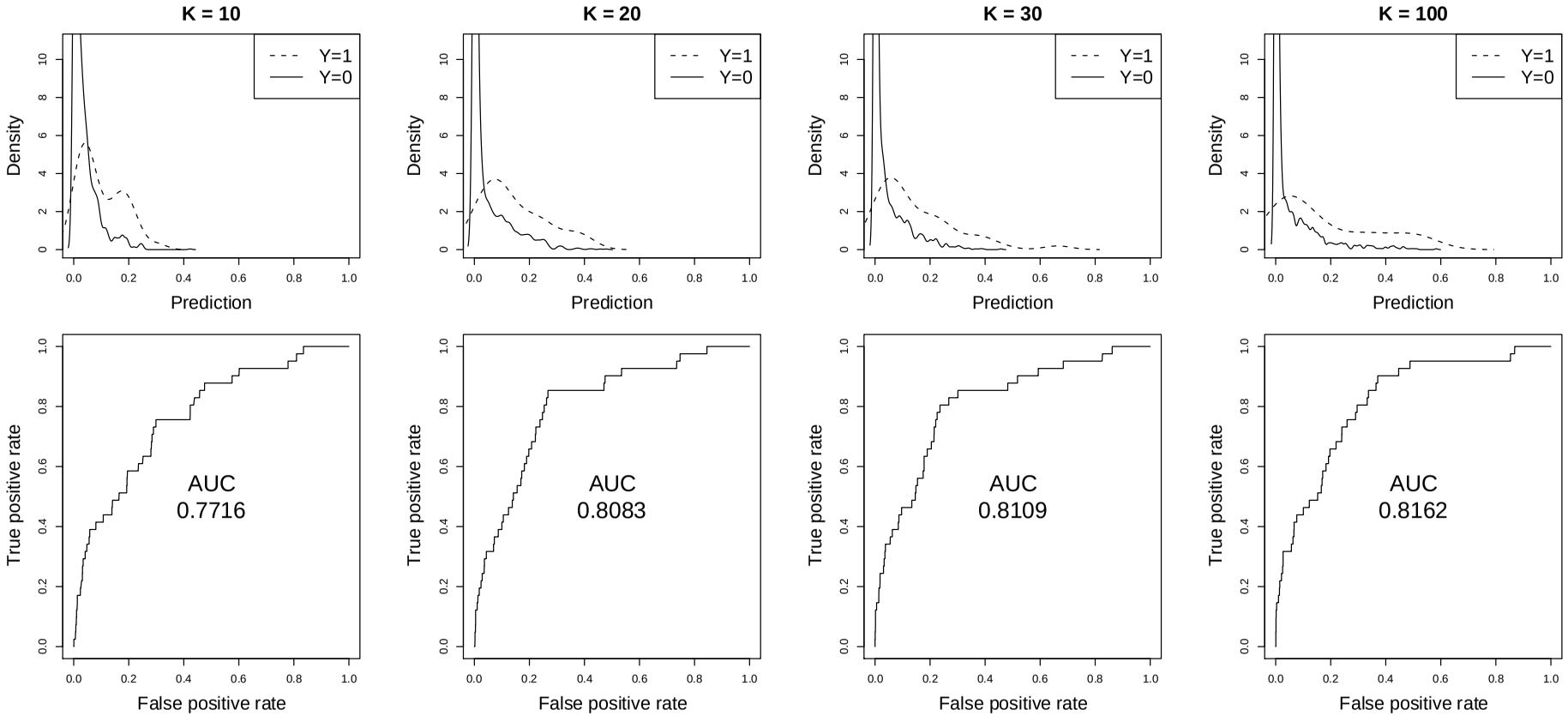


Figure 6.15: Performance of out-of-sample predictions for *Caulobacter crescentus* using S_k^{Ω} for $K = 10$, 20, 30 and 100 (from left to right). Unrestricted model performance is shown predicting the TCS interactions of *Caulobacter crescentus*, trained on the full dataset excluding *Caulobacteria*. The top row of figures show the predicted interaction probabilities for the expected interactions ($Y = 1$) and non-interactions ($Y = 0$), showing the class separation improves with higher K. The bottom row of figures are ROC plots, labelled with the area under the curve, where the area increases as K is increased.

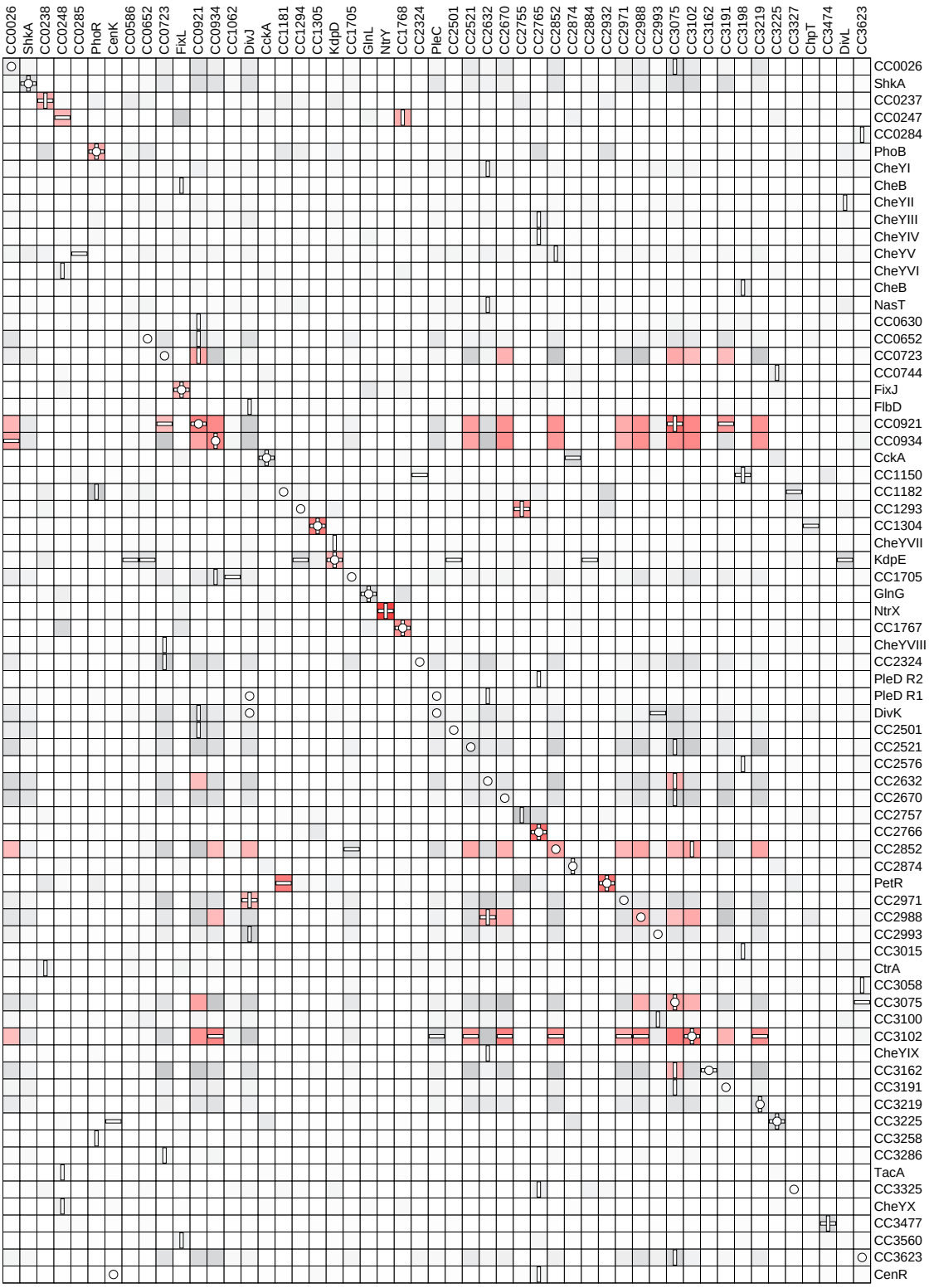


Figure 6.16: Out-of-sample predictions for *Caulobacter crescentus* using S_k^{Ω} and $K = 100$ with an unrestricted GLM, using CLUSTAL W MSAs. Rows are HisKA domains, columns are receivers, sorted by gene location. Most gene names have been taken from `http://caulo.stanford.edu/GeneList.htm` (Shapiro group) as the GenBank annotation predates many of these assignations. As in Figure 6.10, predicted probability of interaction is shown by colour, with the highest score in each row and column indicated with a vertical or horizontal bar. White circles show interactions expected from the genome arrangement (roughly on the diagonal) or Skerker *et al.* (2005).

6.8 Application to *Nostoc* and *M. xanthus*

Figures 6.17 and 6.18 for *Nostoc* sp, and Figure 6.20 and 6.21 for *M. xanthus* shows the model performance predicting those interactions in these species that were inferred from their genome arrangements. Because both organisms have highly complex TCS arrangements, with a high proportion of orphaned and complex HYs, this prediction assessment is only based on part of the full interaction grid.

Figure 6.19 shows the prediction grid for *Nostoc* using S_k^{Ω} with $K = 100$ in an unrestricted GLM, with those combinations with $P \geq 0.25$ tabulated in Table 6.1, which include both non-neighbouring HY and RR pairs, and predictions for some complex HY proteins as well. The equivalent results for *M. xanthus* are shown in Figure 6.22 and Table 6.2.

6.9 Discussion

When trying to assign TCS domain partnerships, a biologist would typically start from the genomic organisation before even looking at the actual sequences. Two simple pieces of such information are the separation of the two domains in nucleotides, and whether or not the domains are from the same gene (a hybrid kinase). The GLM framework allows model extension by the addition of more explanatory terms, thus including this information would be straightforward, but it would trivially explain our automatically compiled dataset. However, for a large experimentally determined dataset of interactions and non-interactions, it would be intriguing to explore a composite model using both sequence data *and* genomic organisation.

The use of MUSCLE rather than CLUSTAL W for the construction of the MSAs made minimal difference to the predictive performance of the models, while the number of explanatory variables (K) and the scoring function function used to generate them had marked effects. In the results shown up to $K = 100$ terms have be used, more explanatory variables gave only marginal improvement and becomes increasingly computationally expensive. Selecting terms with MI led to reasonable predictions, while an alternative scheme selecting columns with low MI was found to give predictions little better than random (data not shown).

The hydrophilicity and chemical potential scores (S^H and S^V) give similar prediction performance, which is expected as they are highly correlated. Using the GLM framework to assign different weights (with different signs) allows some residue pairs to be labelled as having compatible physicochemical properties and others to be repulsive, and this gives a model with measurable predictive power. In contrast, the S_k^P, S_k^M, S_k^{ζ}, S_k^{η} and S_k^{Ω} scoring systems are specific to the observed frequencies for each amino acid column pair, and all give

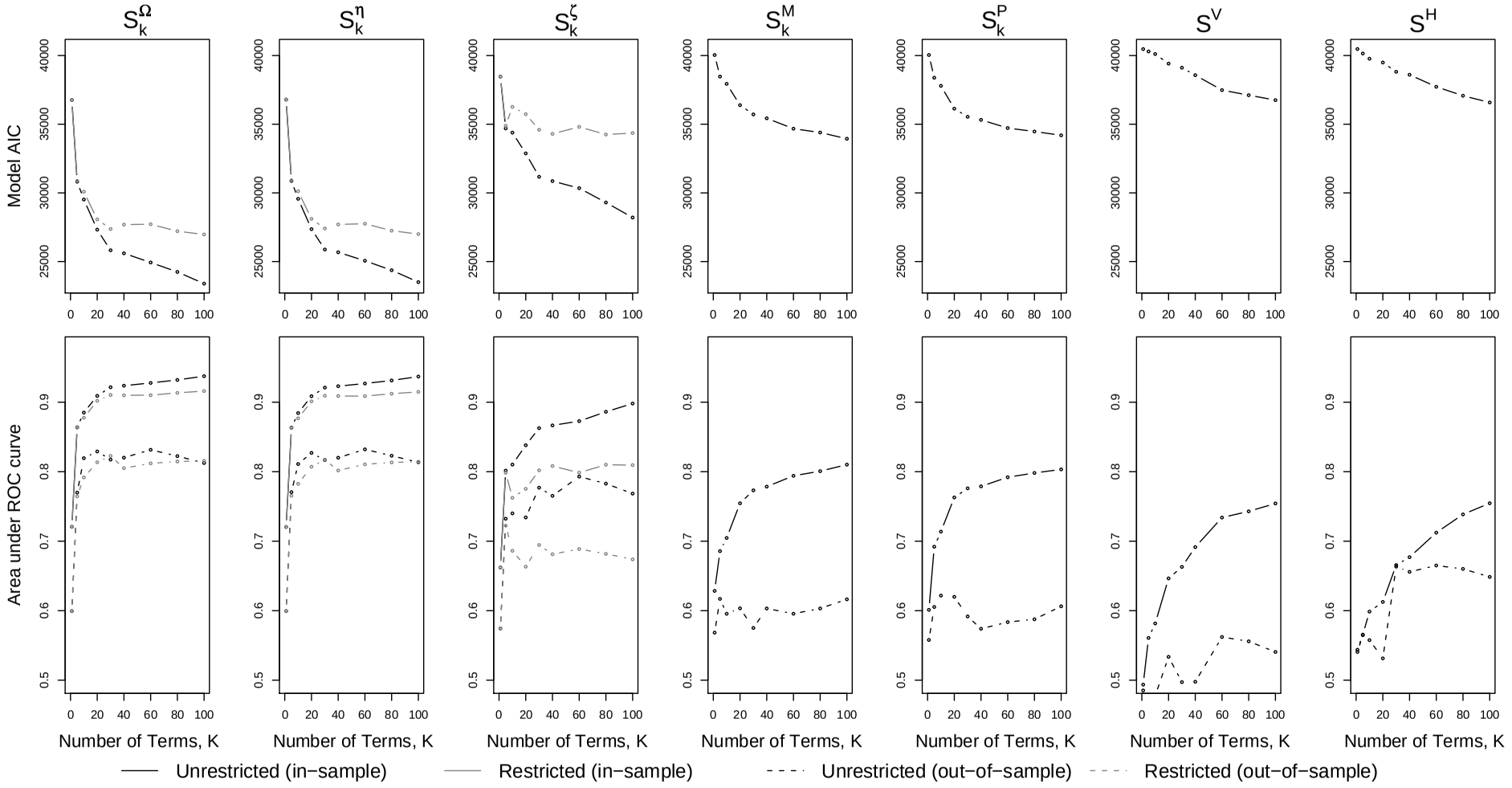


Figure 6.17: Model performance predicting the TCS interactions of *Nostoc* sp. (dashed lines, expected interactions from genome arrangement), trained on full dataset excluding *Nostoc* (solid lines). Using CLUSTAL W MSAs. Key as per Figure 6.2.

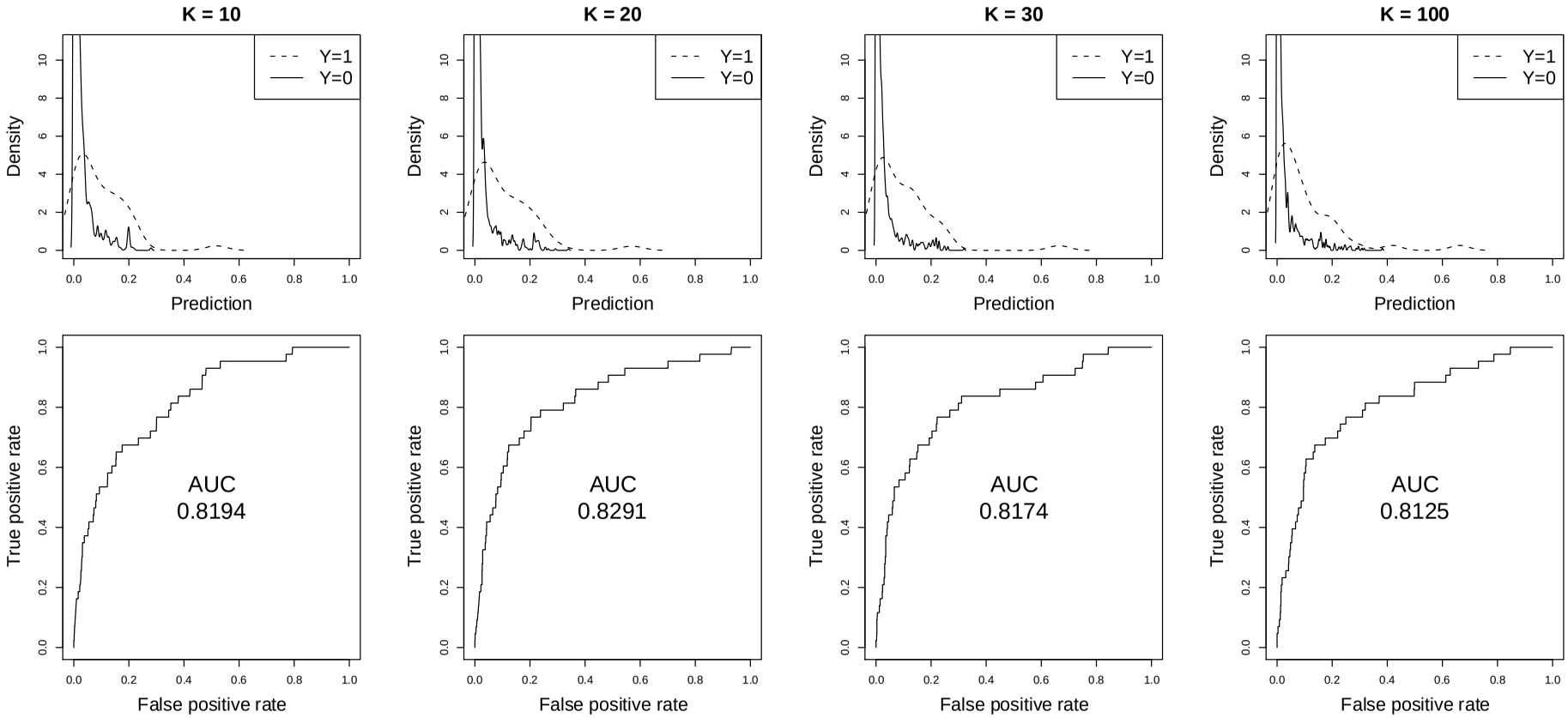


Figure 6.18: Performance of out-of-sample predictions for *Nostoc* sp. using S_k^{Ω} for $K = 10$, 20, 30 and 100 (from left to right). Unrestricted model performance is shown predicting the TCS interactions of *Nostoc* sp. expected from the genome arrangement, trained on the full dataset excluding *Nostoc*. The top row of figures show the predicted interaction probabilities for the expected interactions ($Y = 1$) and non-interactions ($Y = 0$), showing the class separation improves with higher K. The bottom row of figures are ROC plots, labelled with the area under the curve, where the area increases as K is increased.

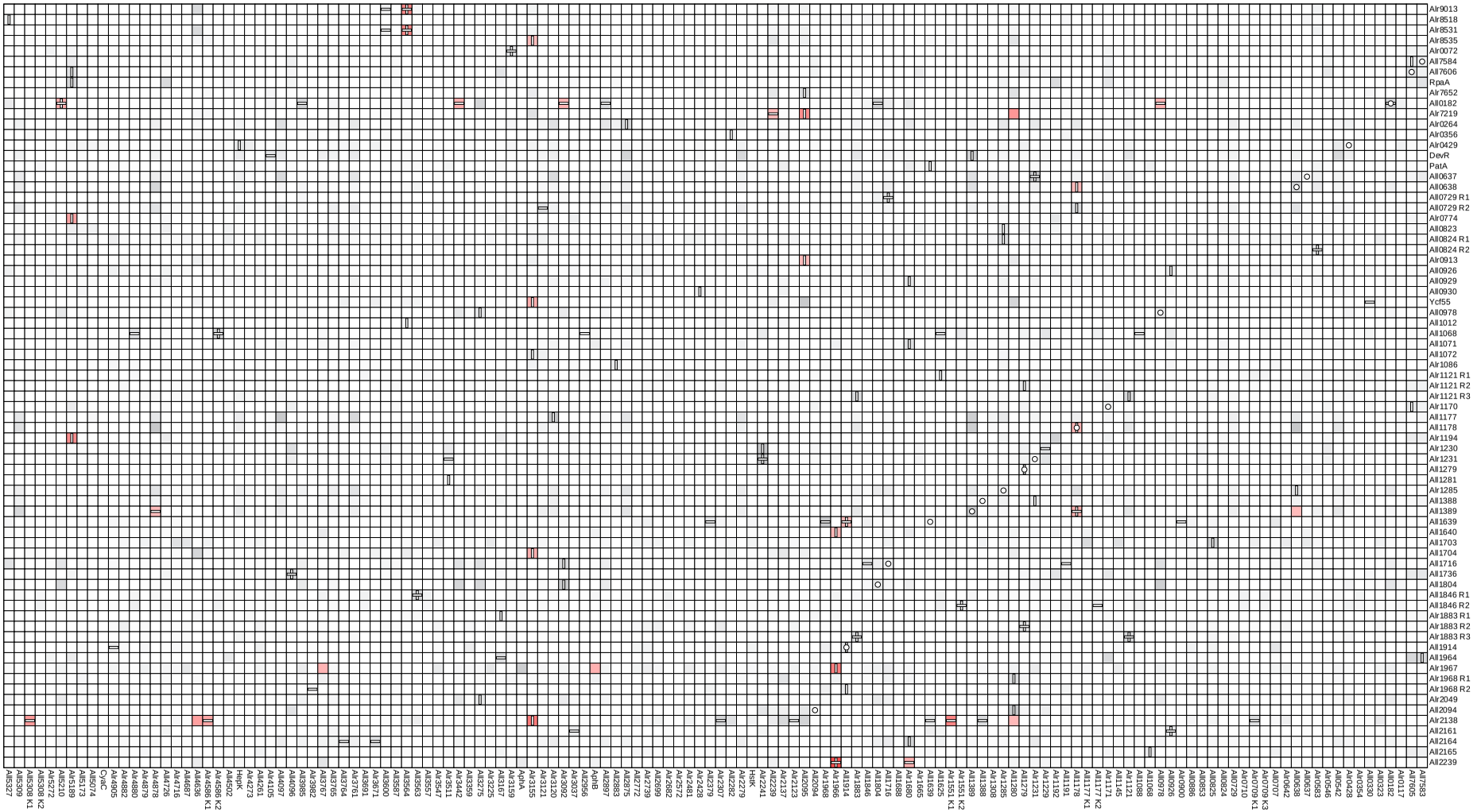
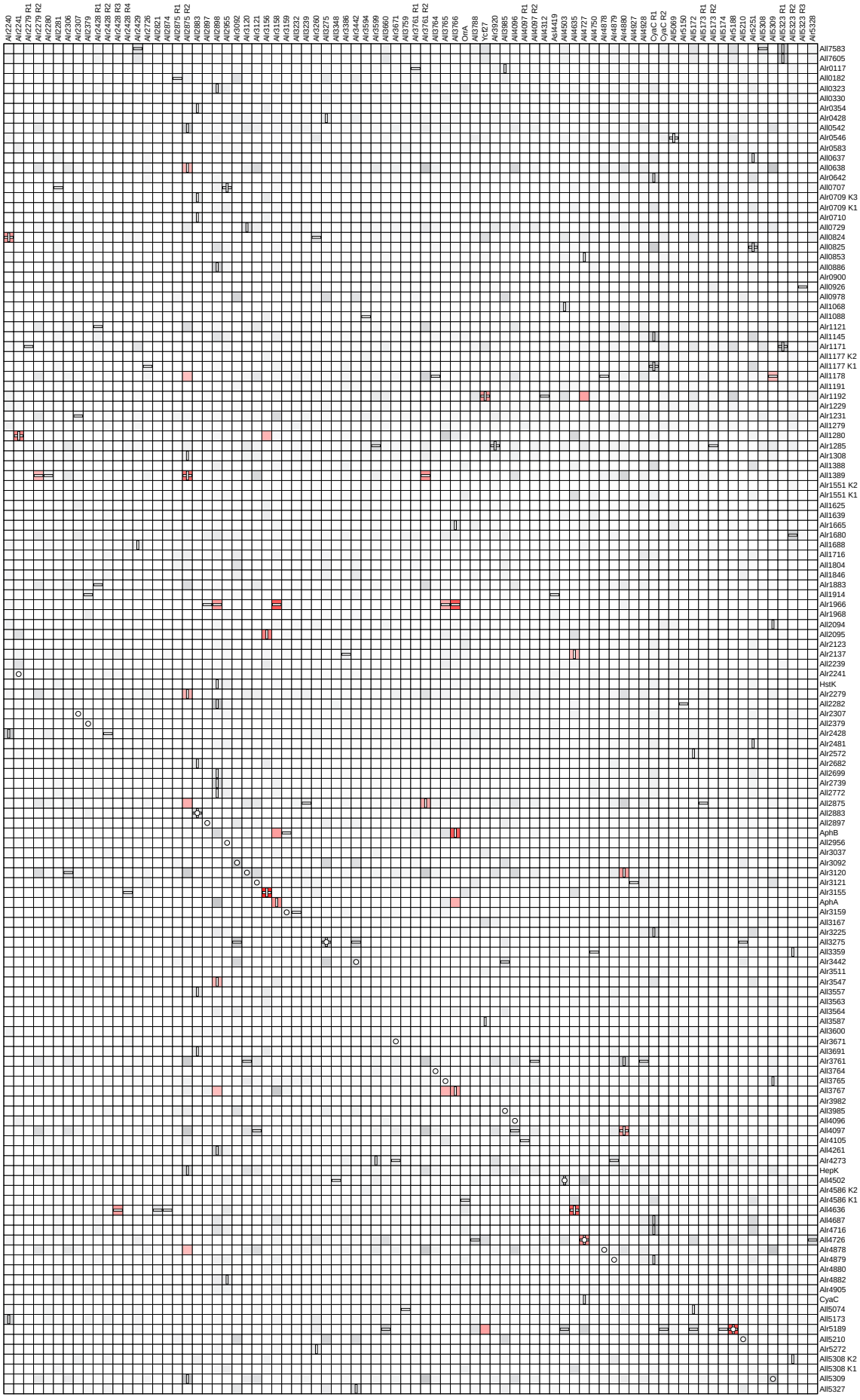


Figure 6.19: Out-of-sample predictions for *Nostoc* sp. using S_k^{Ω} and $K = 100$ with an unrestricted GLM. Rows are HisKA domains, columns are receivers, sorted by gene location. Predicted probability of interaction shown by colour (linear scaling from 0.0 as white to 0.25 in pale grey, to 1.0 as red). White circles show interactions expected from the genome arrangement (and are therefore roughly on the diagonal). Continued on following page.


Alr2240
Alr2241
Alr2279 R1
Alr2279 R2
Alr2280
All2281
Alr2306
Alr2307
All2379
Alr2428 R1
Alr2428 R2
Alr2428 R3
Alr2428 R4
Alr2429
Alr2726
All2821
All2874
All2875 R1
All2875 R2
All2883
All2887
All2898
All2955
Alr3092
Alr3120
Alr3121
Alr3155
Alr3158
Alr3159
All3232
All3239
All3260
All3275
All3348
Alr3386
Alr3442
Alr3594
All3599
All3660
All3671
All3759
Alr3761 R1
Alr3761 R2
All3764
All3765
All3766
All3788
Yc627
Alr3920
All3985
All4096
All4097 R1
All4097 R2
All4312
As14419
All4503
All4635
All4727
All4750
All4878
All4879
All4880
All4927
All4928
CyaC R1
CyaC R2
All5069
Alr5150
Alr5172
All5173 R1
All5173 R2
All5174
All5188
All5210
All5251
All5308
All5309
All5323 R1
All5323 R2
All5323 R3
All5328
All7583
All7605
Alr0117
All0182
All0323
All0330
Alr0354
Alr0428
All0542
Alr0546
Alr0583
All0637
All0638
Alr0642
All0707
Alr0709 K3
Alr0709 K1
Alr0710
All0729
All0824
All0825
All0853
All0886
Alr0900
All0926
All0978
All1068
All1088
Alr1121
All1145
Alr1171
All1177 K2
All1177 K1
All1178
All1191
Alr1192
Alr1229
Alr1231
All1279
All1280
Alr1285
Alr1308
All1388
All1389
Alr1551 K2
Alr1551 K1
All1625
All1639
Alr1665
Alr1680
All1688
All1716
All1804
All1846
Alr1883
All1914
Alr1966
Alr1968
All2094
All2095
Alr2123
Alr2137
All2239
Alr2241
HstK
Alr2279
All2282
Alr2307
All2379
Alr2428
Alr2481
Alr2572
Alr2682
All2699
Alr2739
All2772
All2875
All2883
All2897
AphB
All2956
Alr3037
Alr3092
Alr3120
Alr3121
Alr3155
AphA
Alr3159
All3167
Alr3225
All3275
All3359
Alr3442
Alr3511
Alr3547
All3557
All3563
All3564
All3587
All3600
Alr3671
All3691
Alr3761
All3764
All3765
All3767
Alr3982
All3985
All4096
All4097
Alr4105
All4261
Alr4273
HepK
All4502
Alr4586 K2
Alr4586 K1
All4636
All4687
Alr4716
All4726
Alr4878
Alr4879
Alr4880
Alr4882
Alr4905
CyaC
All5074
All5173
Alr5189
All5210
Alr5272
All5308 K2
All5308 K1
All5309
All5327

Name	TCS domains	Name	TCS domains	Prediction	Expected
All0638	T_i-R	All2875	T_i-R-R-H (R2)	0.275	
All0638	T_i-R	All1389	T_i-R	0.266	
All0824	T_i-R-R	Alr2240	R	0.373	
All0978	T_i-R	All0182	T_i-R	0.295	
All1178	T_i-R	All1389	T_i-R	0.296	
All1178	T_i-R	All1178	T_i-R	0.287	Yes
All1178	T_i-R	All5309	T_i-R	0.269	
All1178	T_i-R	All0638	T_i-R	0.261	
All1178	T_i-R	All2875	T_i-R-R-H (R2)	0.250	
Alr1192	T_i	Ycf27	R	0.379	
Alr1192	T_i	All4727	R	0.362	
All1280	T_i	Alr2241	R-T_i	0.472	
All1280	T_i	Alr7219	R	0.413	
All1280	T_i	Alr3156	R	0.294	
All1280	T_i	Alr2138	R	0.275	
All1389	T_i-R	All2875	T_i-R-R-H (R2)	0.536	
All1389	T_i-R	Alr3761	T_i-R-R (R2)	0.418	
All1389	T_i-R	Alr2279	T_i-R-R-H (R2)	0.279	
Alr1551	K-T_i (K1)	Alr2138	R	0.482	
Alr1680	T_i	All2239	R-T_i	0.306	
All1914	T_i-R	All1639	T_i-R	0.258	
Alr1966	T_i	All2239	R-T_i	0.679	
Alr1966	T_i	All3766	R	0.651	
Alr1966	T_i	Alr3158	R	0.639	
Alr1966	T_i	Alr1967	R	0.511	
Alr1966	T_i	All2898	R	0.388	
Alr1966	T_i	All3765	R-T_i	0.310	
Alr1966	T_i	All1640	R	0.280	
All2095	T_i	Alr3156	R	0.500	
All2095	T_i	Alr7219	R	0.449	
All2095	T_i	Alr0913	R	0.265	
Alr2137	T_i	All4635	R	0.255	
All2239	R-T_i	Alr7219	R	0.265	
Alr2279	T_i-R-R-H	All2875	T_i-R-R-H (R2)	0.297	
All2875	T_i-R-R-H	Alr3761	T_i-R-R (R2)	0.330	
All2875	T_i-R-R-H	All2875	T_i-R-R-H (R2)	0.315	
AphB	T_i	All3766	R	0.646	
AphB	T_i	Alr3158	R	0.381	
AphB	T_i	Alr1967	R	0.304	
Alr3092	T_i-R	All0182	T_i-R	0.288	
Alr3120	T_i-R	Alr4880	R-T_i	0.382	
Alr3155	T_i	Alr3156	R	0.812	
Alr3155	T_i	Alr2138	R	0.521	
Alr3155	T_i	Ycf55	R	0.322	
Alr3155	T_i	All1704	R	0.312	
Alr3155	T_i	Alr8535	R	0.259	
AphA	T_i	Alr3158	R	0.398	
AphA	T_i	All3766	R	0.308	
Alr3442	T_i-R	All0182	T_i-R	0.308	
Alr3547	T_i	All2898	R	0.293	
All3564	T_i	Alr9013	R	0.460	
All3564	T_i	Alr8531	R	0.460	
All3767	T_i	All3766	R	0.363	
All3767	T_i	All3765	R-T_i	0.300	
All3767	T_i	Alr1967	R	0.294	
All3767	T_i	All2898	R	0.256	
All4097	R-T_i-R	Alr4880	R-T_i	0.384	
Alr4586	K-T_i (K1)	Alr2138	R	0.346	
All4636	T_i	All4635	R	0.624	
All4636	T_i	Alr2428	R-H-R-T_i-R-R (R3)	0.376	
All4636	T_i	Alr2138	R	0.368	
All4726	T_i	All4727	R	0.417	Yes
Alr4878	T_i-R	All1389	T_i-R	0.264	
Alr4878	T_i-R	All2875	T_i-R-R-H (R2)	0.258	
Alr5189	T_i	Alr5188	R	0.662	Yes
Alr5189	T_i	Alr1194	R	0.444	
Alr5189	T_i	Ycf27	R	0.365	
Alr5189	T_i	Alr0774	R	0.360	
All5210	T_i-R	All0182	T_i-R	0.316	
All5308	R-K-T_i (K1)	Alr2138	R	0.390	

Table 6.1: Unrestricted model predictions $P \geq 0.25$ for *Nostoc* sp. using S_k^{Ω} and $K = 100$, corresponding to the red squares in Figure 6.19. The interactions are grouped by the HisKA domain, with possible receiver interactions listed according the predicted probability of interaction. The final column shows if the interaction was expected from the genome arrangement. Where a protein contains more than one HisKA or receiver domain, the relevant domain is indicated in brackets.

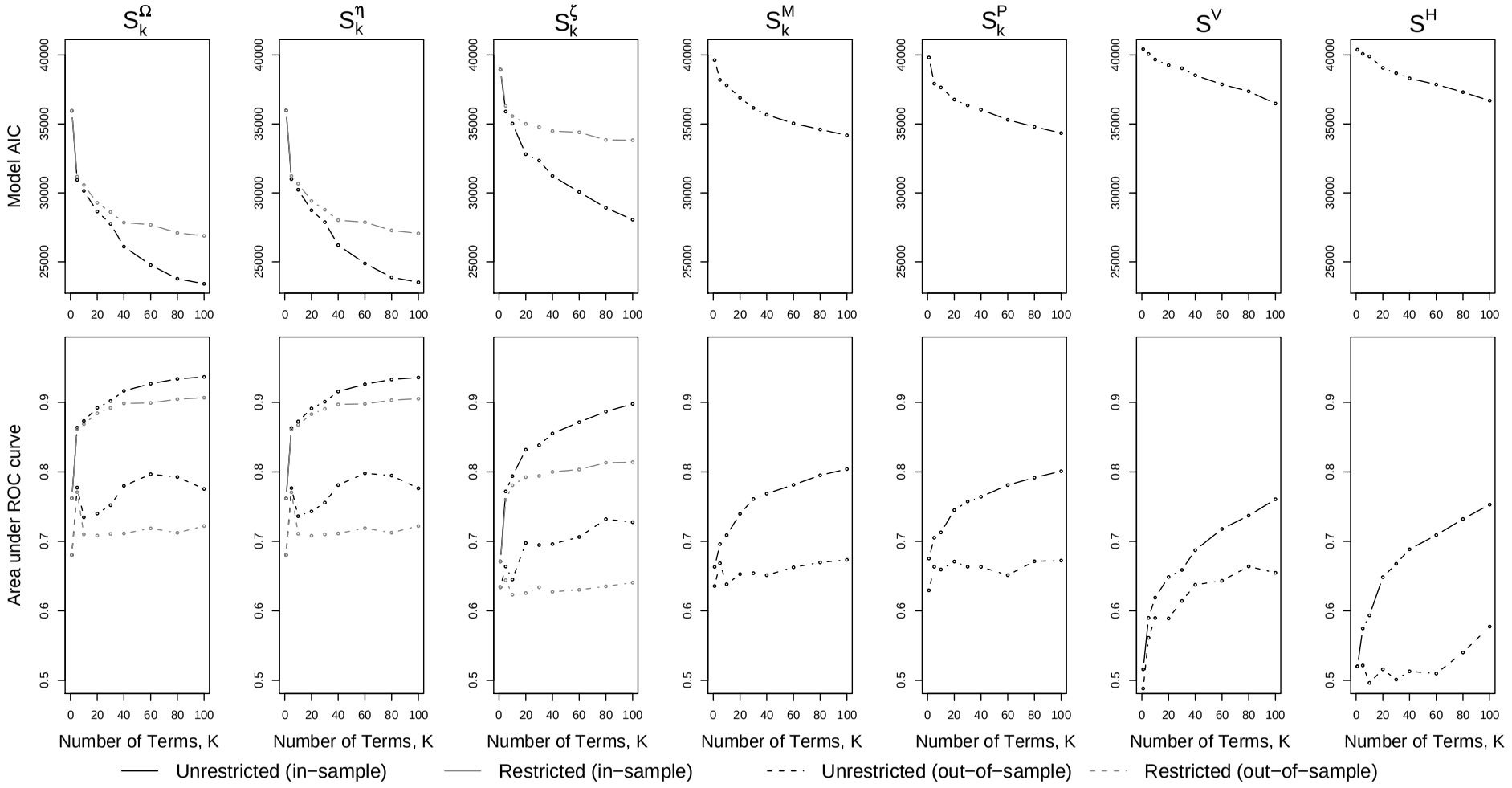


Figure 6.20: Model performance predicting the TCS interactions of *M. xanthus* (dashed lines, expected interactions from genome arrangement), trained on full dataset excluding *Myxobacteria* (solid lines). Using Clustal W MSAs. Key as per Figure 6.2.

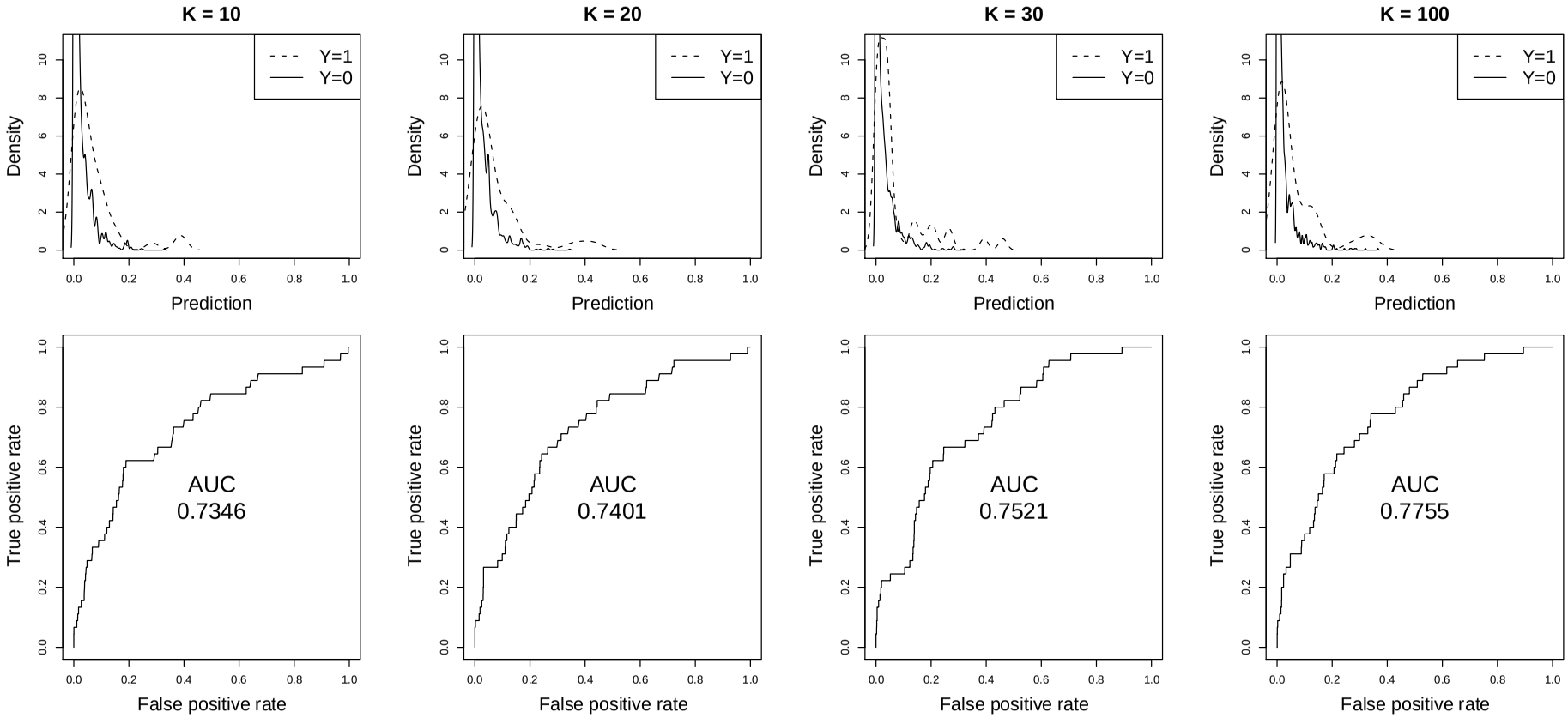


Figure 6.21: Performance of out-of-sample predictions for *M. xanthus* using S_k^{Ω} for $K = 10$, 20, 30 and 100 (from left to right). Unrestricted model performance is shown predicting the TCS interactions of *M. xanthus* expected from the genome arrangement, trained on the full dataset excluding *Myxobacteria*. The top row of figures show the predicted interaction probabilities for the expected interactions ($Y = 1$) and non-interactions ($Y = 0$), showing the class separation improves with higher K. The bottom row of figures are ROC plots, labelled with the area under the curve, where the area increases as K is increased.

Figure 6.22: Out-of-sample predictions for *M. xanthus* using S_k^{Ω} and $K = 100$ with an unrestricted GLM. Rows are HisKA domains, columns are receivers, sorted by gene location. Predicted probability of interaction shown by colour (linear scaling from 0.0 as white to 0.25 in pale grey, to 1.0 as red). In each row and column, the highest score is indicated with a vertical or horizontal bar. Cells where the score is the highest in that row and column therefore have a cross-hair shown. White circles show interactions expected from the genome arrangement (and are therefore roughly on the diagonal). Continued on following two pages.

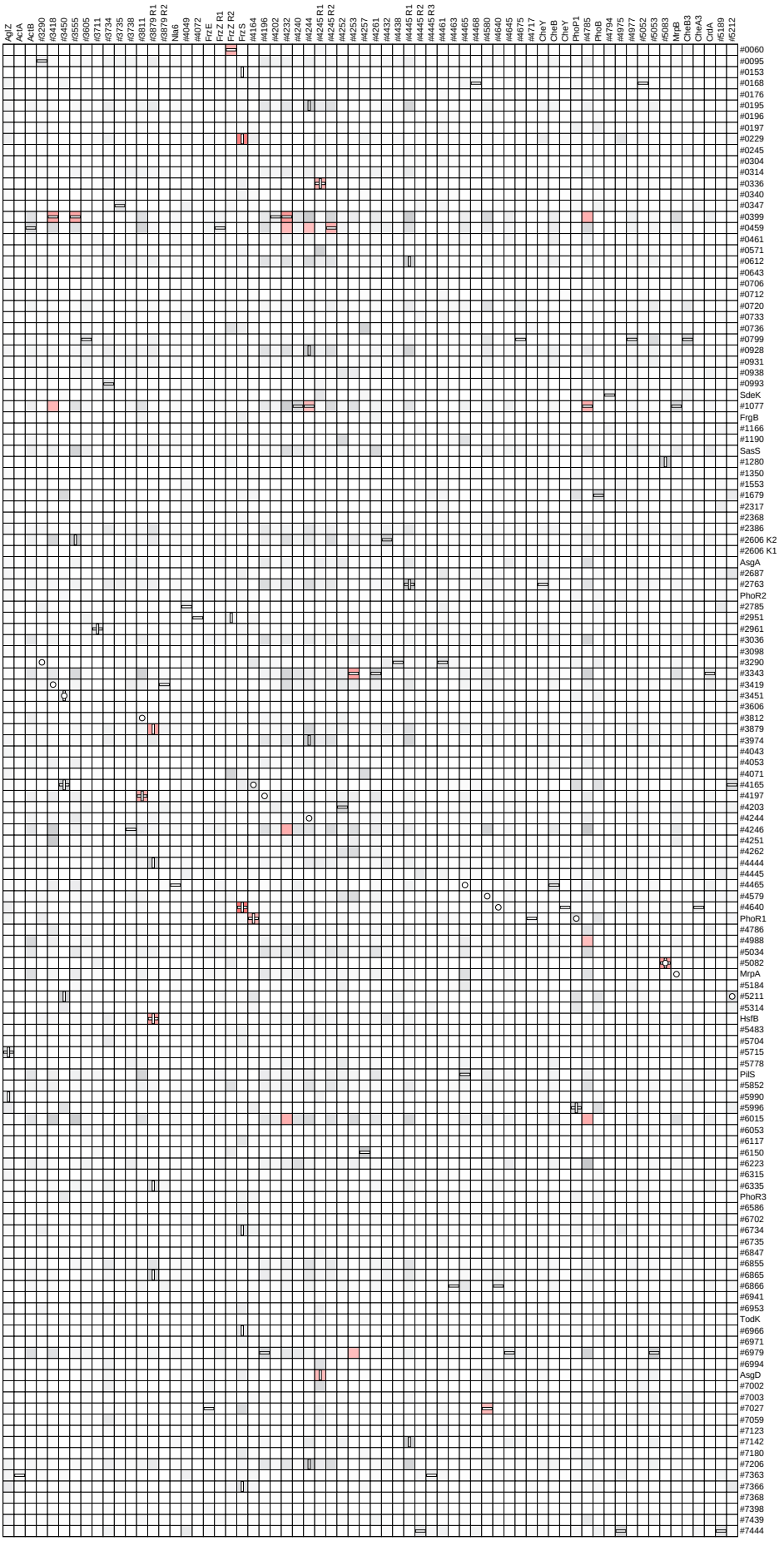


Figure 6.22 continued.

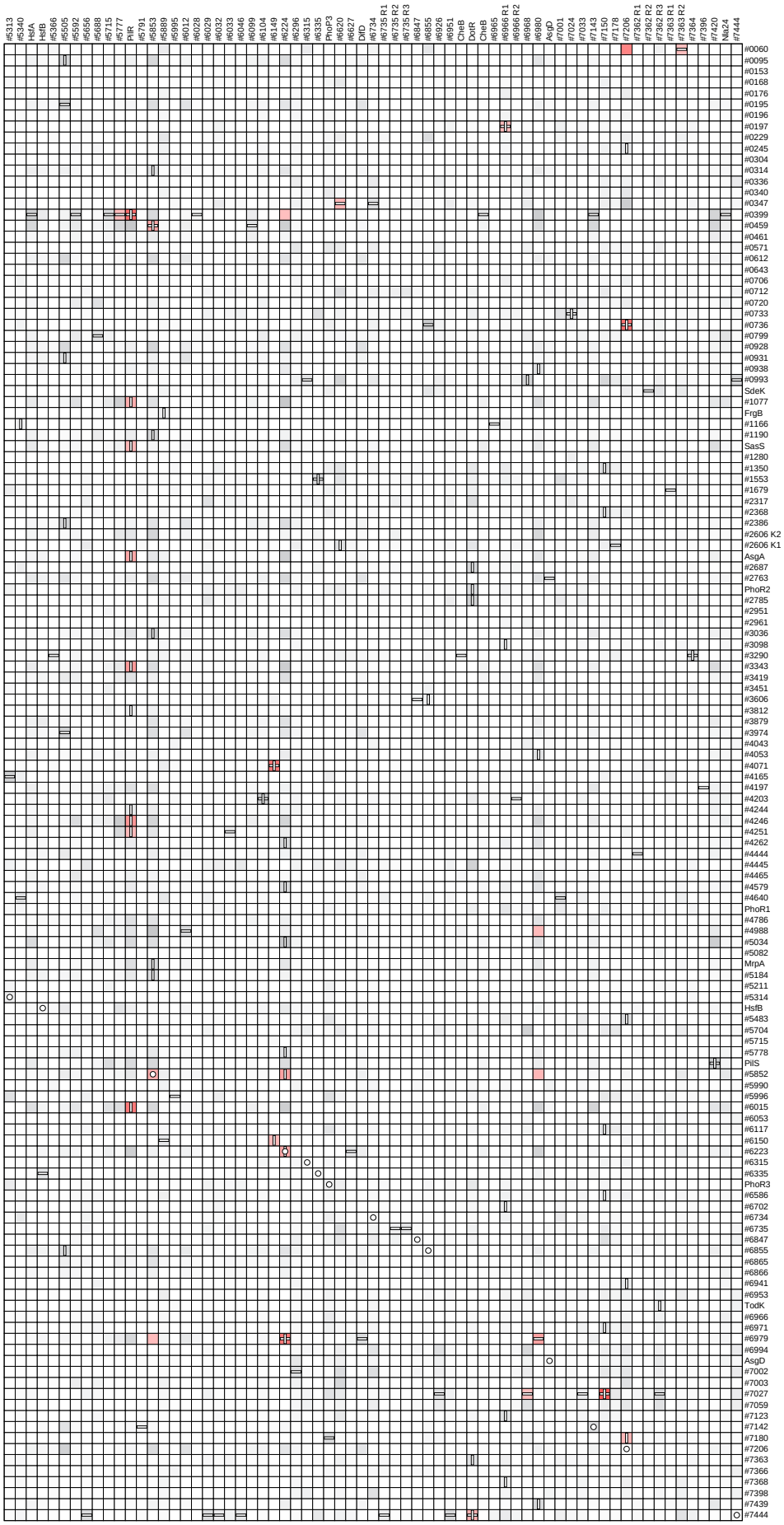


Figure 6.22 continued.

Name	TCS domains	Name	TCS domains	Prediction	Expected
MXAN_0060	T_i	MXAN_0931	T_i-R	0.514	
MXAN_0060	T_i	MXAN_7206	T_i-R	0.481	
MXAN_0060	T_i	MXAN_2763	T_i-R	0.328	
MXAN_0060	T_i	FrzZ	R-R (R2)	0.312	
MXAN_0060	T_i	MXAN_2386	T_i-R	0.289	
MXAN_0060	T_i	MXAN_1093	R	0.283	
MXAN_0060	T_i	MXAN_7363	T_i-R-R (R2)	0.256	
MXAN_0197	T_i	MXAN_6966	R-R-T_i (R1)	0.266	
MXAN_0229	R-T_i-R	FrzS	R	0.496	
MXAN_0336	R-T_i	MXAN_4245	R-R (R1)	0.305	
MXAN_0347	T_i	MXAN_1093	R	0.338	
MXAN_0347	T_i	MXAN_6620	R	0.285	
MXAN_0399	T_i	PilR	R	0.577	
MXAN_0399	T_i	SasR	R	0.439	
MXAN_0399	T_i	MXAN_4232	R	0.390	
MXAN_0399	T_i	MXAN_1378	R	0.342	
MXAN_0399	T_i	MXAN_3418	R	0.341	
MXAN_0399	T_i	MXAN_3555	R	0.337	
MXAN_0399	T_i	MXAN_2671	R	0.327	
MXAN_0399	T_i	Nla28	R	0.314	
MXAN_0399	T_i	MXAN_4785	R	0.311	
MXAN_0399	T_i	MXAN_5777	R	0.260	
MXAN_0399	T_i	MXAN_6224	R	0.253	
MXAN_0459	T_i	MXAN_5853	R	0.343	
MXAN_0459	T_i	MXAN_2671	R	0.315	
MXAN_0459	T_i	MXAN_4245	R-R (R2)	0.309	
MXAN_0459	T_i	MXAN_4232	R	0.269	
MXAN_0459	T_i	MXAN_4244	R-T_i	0.252	
MXAN_0736	T_i	MXAN_7206	T_i-R	0.544	
MXAN_0736	T_i	MXAN_0931	T_i-R	0.520	
MXAN_0736	T_i	MXAN_2386	T_i-R	0.441	
MXAN_0736	T_i	MXAN_2763	T_i-R	0.351	
MXAN_0736	T_i	MXAN_1093	R	0.344	
MXAN_0799	K	MXAN_0732	R	0.323	
MXAN_1077	T_i	PilR	R	0.323	
MXAN_1077	T_i	MXAN_4785	R	0.318	
MXAN_1077	T_i	Nla28	R	0.287	
MXAN_1077	T_i	MXAN_3418	R	0.282	
MXAN_1077	T_i	MXAN_1378	R	0.275	
MXAN_1077	T_i	MXAN_4244	R-T_i	0.265	
SasS	T_i	PilR	R	0.281	
MXAN_1679	T_i	MXAN_1680	R	0.267	Yes
AsgA	R-T_i	PilR	R	0.340	
MXAN_3343	T_i	PilR	R	0.395	
MXAN_3343	T_i	MXAN_4253	R	0.362	
MXAN_3343	T_i	MXAN_1378	R	0.296	
MXAN_3343	T_i	SasR	R	0.263	
MXAN_3343	T_i	MXAN_0460	R-R (R1)	0.258	
MXAN_3879	T_i-R-R	MXAN_3879	T_i-R-R (R1)	0.349	
MXAN_4043	K	MXAN_1093	R	0.425	
MXAN_4071	T_i	MXAN_6149	R	0.526	
MXAN_4197	T_i	MXAN_3811	R	0.321	
MXAN_4246	T_i	PilR	R	0.413	
MXAN_4246	T_i	MXAN_4232	R	0.316	
MXAN_4246	T_i	SasR	R	0.263	
MXAN_4251	T_i	PilR	R	0.271	
MXAN_4465	T_i-R	SasR	R	0.264	
MXAN_4640	T_i-R	FrzS	R	0.540	
PhoR1	T_i	MXAN_4164	R	0.270	
MXAN_4988	T_i	SasR	R	0.319	
MXAN_4988	T_i	MXAN_6980	R	0.265	
MXAN_4988	T_i	MXAN_4785	R	0.256	
MXAN_5082	T_i	MXAN_5083	R	0.362	Yes
HsfB	R-T_i	MXAN_3879	T_i-R-R (R1)	0.367	
MXAN_5852	T_i	MXAN_6224	R	0.366	
MXAN_5852	T_i	MXAN_5853	R	0.314	Yes
MXAN_5852	T_i	MXAN_6980	R	0.282	
MXAN_6015	T_i	PilR	R	0.526	
MXAN_6015	T_i	SasR	R	0.376	
MXAN_6015	T_i	MXAN_1378	R	0.347	
MXAN_6015	T_i	MXAN_4785	R	0.315	
MXAN_6015	T_i	MXAN_4232	R	0.310	
MXAN_6015	T_i	MXAN_2671	R	0.254	
MXAN_6150	T_i	MXAN_6149	R	0.266	
MXAN_6223	T_i	MXAN_6224	R	0.330	Yes
MXAN_6979	T_i	MXAN_6224	R	0.523	
MXAN_6979	T_i	MXAN_6980	R	0.364	
MXAN_6979	T_i	MXAN_5853	R	0.263	
MXAN_6979	T_i	MXAN_4253	R	0.259	
AsgD	R-T_i	MXAN_4245	R-R (R1)	0.266	
MXAN_7002	T_i	MXAN_1093	R	0.366	
MXAN_7027	T_i	MXAN_7150	R	0.634	
MXAN_7027	T_i	MXAN_4580	R	0.268	
MXAN_7027	T_i	MXAN_6968	R	0.255	
MXAN_7180	T_i	MXAN_7206	T_i-R	0.279	
MXAN_7444	R-T_i	DotR	R	0.299	

Table 6.2: Unrestricted model predictions $P \geq 0.25$ for *M. xanthus* using S_k^{Ω} and $K = 100$, corresponding to the red squares in Figure 6.22. The interactions are grouped by the HisKA domain, with possible receiver interactions listed according the predicted probability of interaction. The final column shows if the interaction was expected from the genome arrangement. Where a protein contains more than one HisKA or receiver domain, the relevant domain is indicated in brackets.

better performance than the S^H and S^V scores. Given a suitably large and uniformly sampled training dataset, these scores themselves could capture useful information about the *nature* of each interaction, especially if coupled with spatial information about the protein-protein complex.

By their construction, S_k^ζ, S_k^η and S_k^Ω are closely related (Section 6.2.3). In terms of predictive power there is little to choose betweenS_k^η and S_k^Ω, but S_k^ζ was found to be less effective. One difference between these scores is the calculation of S_k^ζ only looks at the amino acid frequencies in the training set's positive interactions ($Y = 1$), while S_k^η and S_k^Ω also consider the non-interactions ($Y = 0$), and thus S_k^ζ may be less robust on smaller datasets.

The simple Bayesian interpretations given in Equations (6.15) and (6.16) allow $\sum S_k^\eta$ and $\sum S_k^\Omega$ to be linked to the probability of interaction given the amino acids residues. The argument in Section 6.2.3 extends this to give some justification for the use of $\sum S_k^\zeta$ as an interaction indicator in White *et al.* (2007). Similarly, $\sum S_k^\eta$ and $\sum S_k^\Omega$ can also be used directly with a sliding threshold to rank interactions (indeed, this would give the same ROC curve as the restricted GLM used here). However, this simple Bayesian interpretation assumes all K column pair scores to be independent. As a MSA column can be found in multiple column pairs (e.g. Figures 5.25 and 5.26), this assumption may be problematic for large K. Handling column interdependence explicitly in a Bayesian framework, perhaps along the lines of Burger and van Nimwegen (2006, 2008), may be an interesting alternative to the GLM approach herein.

The relatively poor out-of-sample prediction performance of all the models trained on the hybrid dataset is probably of biological significance. We posit that the specificity between HisKA and receiver domains in hybrid kinase proteins is less selective than in two-gene partners, because in these proteins the two domains are tethered to each other and thus the phosphotransfer suffers less competition from other receivers. i.e. domains in hybrid kinases impose weaker co-evolutionary dependence on each other, than in two-gene TCS pairs. See also Cock and Whitworth (2007b). Following this rationale, the assumption of exclusive interaction between domains from hybrid kinases (used to compile our dataset) is perhaps unjustified.

In parameter selection, the AIC is intended to penalise over-fitting. Typically plotting the AIC against the number of terms (here K) will show a clear minimum for some value of K. However, in these results the AIC was never minimised - even for large $K > 100$ where the predictive performance assessed by the ROC area had clearly plateaued. Also, less direct evidence comes from the better than expected predictive performance on the *E. coli*

and *Bacillus subtilis* datasets (Sections 6.5 and 6.6). Taken together, this may indicate that some of these models are over fitting the data, and that K should be limited based on some other criteria. Biological knowledge could be used to suggest a sensible cut off, or a human judgement call based on the ROC area curves. One more rigourous alternative suggested in Chapter 5 would be to perform a bootstrap analysis on the protein pairs in order to assign a p-value to each MI score. A threshold (say $p = 0.01$) would translate into a limit on K.

6.10 Conclusion

A GLM framework for predicting protein-protein interactions has been presented, and its predictive performance assessed for a variety of scoring functions therein using the AIC and the area under the ROC curve. All scoring functions considered provide predictive power when used as explanatory variables. There is most predictive power from the probabilistic scores S_k^{Ω} and S_k^{η}, with S_k^{ζ} a close third best, all based on observed amino acid frequencies in the training dataset. However, the more general fully defined S^H and S^V scores also have some predictive power and may warrant consideration in smaller datasets where amino acid frequencies cannot be calculated reliably.

The predictive power of the models was found to be lower when trained on and applied to domains from hybrid kinases, and this is not simply an effect of a small training set. This may be a biological phenomena, namely that these hybrid gene domain pairs are not as monogamous as those in two-gene pairs, on that basis that being physically tethered to each other the gives the "correct" pairing a thermodynamic advantage. Hypothetically this results in less evolutionary pressure to maintain exclusive specificity at the protein-protein interface for hybrid kinases. Thus there are two handicaps to predicting these pairings, firstly the reliability of the interaction calls in the training dataset is undermined, and secondly the amino acid "signal" would be expected to be weaker.

The unweighted sum of the S_k^{ζ}, S_k^{η} or S_k^{Ω} logged probability ratio scores has been shown to have merit as an single explanatory variable in the GLM framework. These summations can also be used directly with a threshold as a simple predictive guide for protein-protein interactions, with the virtue of being comparatively simple to understand and implement. However, using the unrestricted GLM improves on the predictive power.

Finally, the predictions shown for *E. coli*, *Bacillus subtilis* and *Caulobacter crescentus* show broad agreement with current knowledge. The predicted interactions for more complex organisms such as *Nostoc* sp. and *M. xanthus* should provide a useful list of candidate interactions for experimental verification.

Chapter 7

Conclusions and future work

Chapter 1 introduced TCS systems, and some of the known systems and networks. Chapter 2 described a protein domain motif based survey of fully sequenced prokaryotes, which identified a number of trends. In particular, it was found that prokaryotes with a large TCS gene complement tend to have less simple TCS transmitter-receiver pairs, and instead have more complex hybrid genes and TCS gene clusters. *Nostoc* sp. and *M. xanthus* are particular examples of this, where the genome arrangement precludes simple deduction of TCS interactions.

Patterns in the phase of gene overlaps were considered in Chapter 3, a general property of prokaryotic genes (not specific to TCS gene pairs). The observed phase bias in longer unidirectional gene overlaps could be explained by the genetic code itself, provided these most such overlaps arose from the selection of a new start codon for the downstream gene. However, with the biological validity of these long overlaps somewhat in question (Pallejá *et al.*, 2008), the same model could equally well describe the annotation process.

Typical two gene TCS systems, consisting of neighbouring HK and RR genes, and simple HY genes, containing one transmitter and one receiver, were the focus of Chapter 4. While many TCS systems have evolved using these single proteins, it was found that in most cases these lacked a TM input domain or DNA binding output domain. These domains impose specific spatial constraints on the mobility of the protein, which would generally be impaired by the merger of separate HK and RR genes into a HY. Further work would be required using a phylogenetic analysis to determine what proportion of HY genes have evolved from the *in situ* fusion of neighbouring HK and RR genes, and how many can be best explained by recombination events. Additionally, refinements in TCS PFAM domain models should also allow better detection of potentially missed phosphotransfer domains which could indicate that some of these HY systems are in fact part of larger phosphorelays (Section 1.4.6).

Chapters 5 and 6 described a model developed to predict TCS interactions between

HisKA and receiver domains from their amino acid sequences. Firstly, important amino acid positions in the two domains were identified using MI, and then these were scored numerically as explanatory variables in a GLM. The best predictive power came from two probabilistic scores S_k^{Ω} and S_k^{η}, constructed as log odds ratios of observed amino acid frequencies in the training data. After model validation, predictions were made for a number of model organisms, including *Nostoc* sp. and *M. xanthus* which have a particularly complex set of TCS genes. After exploring sensible upper limits in the number of model terms to avoid over-fitting, the next logical step would be to attempt experimental verification of the TCS predictions made in Chapter 6.

Predicted interactions between specific TCS proteins can be verified experimentally, both *in vivo* (for example, examining the phenotypes of knock out mutants), or *in vitro* (for example, the Y2H assay, e.g. Whitworth *et al.* (2008)). For more robust proof demonstration of actual phosphotransfer is usually confirmed using a radioisotope assay (e.g. Yamamoto *et al.* (2005) and Skerker *et al.* (2005)). For most organisms where few if any TCS interactions have been confirmed, predictions can usefully guide experiments to target particular combinations. In model organisms such as *E. coli*, most native TCS interactions are already reasonably well characterised, and new insights from these predictions are likely to be limited. However, the models here could be applied to mutagenized or engineered variants of the proteins (Skerker *et al.*, 2008), to clarify the mechanisms involved.

If this modelling approach does prove useful, there are a number of relatively straightforward improvements that could be made. Firstly, the set of published sequenced genomes keeps expanding, allowing for ever larger training sets to be compiled. It may now be feasible to apply this method to other less common protein-protein interactions. Secondly, once the training data has been compiled, the construction of new MSAs combining the training data and a given test dataset is a major bottleneck. An obvious step would be to build alignments of the test data once, and train the GLM on this data. After this upfront cost, test sequences could be mapped onto these alignments one at a time in order to match up their amino acids to those deemed to impart interaction specificity.

The simplistic Bayesian interpretation of $\sum S_k^{\Omega}$ or $\sum S_k^{\eta}$ (Section 6.2.4) suggests that as an alternative to using a GLM to take a weighted sum of these terms, explicitly modelling the inter-column dependencies could yield better predictions. Using MI identifies many columns which show correlation to multiple positions in the other domain - perhaps this information can be used explicitly to bundle MSA column pairs into larger groups describing the interaction interface as a number of separate modules, each of which could be scored individually.

In addition to helping to explain the biology of existing TCS systems, any successful interaction prediction scheme will also have potential practical applications. The re-wiring of HK interaction preference has already been demonstrated in *E. coli* by genetic manipulation (Skerker *et al.*, 2008). Bespoke TCS domains could have an important role in synthetic biology, or perhaps even therapeutically as a means of manipulating the signalling pathways within prokaryotic pathogens.

Appendix A

Species List

Table A.1: List of all 457 sequenced species downloaded from the NCBI's FTP site on 26 Feb 2007 (807 accessions/GenBank files), of which the 340 in bold are taken as representative species (ignoring multiple strains, 615 accessions/GenBank files).

Species name	Accessions
Acidobacteria bacterium Ellin345	**NC_008009**
Acidothermus cellulolyticus 11B	**NC_008578**
Acidovorax JS42	**NC_008765, NC_008766, NC_008782**
Acidovorax avenae citrulli AAC00-1	**NC_008752**
Acinetobacter sp ADP1	**NC_005966**
Aeromonas hydrophila ATCC 7966	**NC_008570**
Aeropyrum pernix	**NC_000854**
Agrobacterium tumefaciens C58 Cereon	**NC_003062, NC_003063, NC_003064, NC_003065**
Agrobacterium tumefaciens C58 UWash	NC_003304, NC_003305, NC_003306, NC_003308
Alcanivorax borkumensis SK2	**NC_008260**
Alkalilimnicola ehrlichei MLHE-1	**NC_008340**
Anabaena variabilis ATCC 29413	**NC_007410, NC_007411, NC_007412, NC_007413**
Anaeromyxobacter dehalogenans 2CP-C	**NC_007760**
Anaplasma marginale St Maries	**NC_004842**
Anaplasma phagocytophilum HZ	**NC_007797**
Aquifex aeolicus	**NC_000918, NC_001880**
Archaeoglobus fulgidus	**NC_000917**
Arthrobacter FB24	**NC_008537, NC_008538, NC_008539, NC_008541**
Arthrobacter aurescens TC1	**NC_008711, NC_008712, NC_008713**
Aster yellows witches-broom phytoplasma AYWB	**NC_007716, NC_007717, NC_007718, NC_007719, NC_007720**
Azoarcus BH72	**NC_008702**
Azoarcus sp EbN1	**NC_006513, NC_006823, NC_006824**
Bacillus anthracis Ames	**NC_003997**
Bacillus anthracis Ames 0581	NC_007322, NC_007323, NC_007530
Bacillus anthracis str Sterne	NC_005945
Bacillus cereus ATCC14579	**NC_004721, NC_004722**
Bacillus cereus ATCC 10987	NC_003909, NC_005707
Bacillus cereus ZK	NC_006274, NC_007103, NC_007104, NC_007105, NC_007106, NC_007107
Bacillus clausii KSM-K16	**NC_006582**
Bacillus halodurans	**NC_002570**
Bacillus licheniformis ATCC 14580	**NC_006270**
Bacillus licheniformis DSM 13	NC_006322
Bacillus subtilis	**NC_000964**
Bacillus thuringiensis Al Hakam	**NC_008598, NC_008600**
Bacillus thuringiensis konkukian	NC_005957, NC_006578
Bacteroides fragilis NCTC 9434	**NC_003228, NC_006873**
Bacteroides fragilis YCH46	NC_006297, NC_006347
Bacteroides thetaiotaomicron VPI-5482	**NC_004663, NC_004703**

Continued...

Species name	Accessions
Bartonella bacilliformis KC583	**NC_008783**
Bartonella henselae Houston-1	**NC_005956**
Bartonella quintana Toulouse	**NC_005955**
Baumannia cicadellinicola Homalodisca coagulata	**NC_007984**
Bdellovibrio bacteriovorus	**NC_005363**
Bifidobacterium adolescentis ATCC 15703	**NC_008618**
Bifidobacterium longum	**NC_004307, NC_004943**
Bordetella bronchiseptica	**NC_002927**
Bordetella parapertussis	**NC_002928**
Bordetella pertussis	**NC_002929**
Borrelia afzelii PKo	**NC_008273, NC_008274, NC_008277, NC_008564, NC_008565, NC_008566, NC_008567, NC_008568, NC_008569**
Borrelia burgdorferi	**NC_000948, NC_000949, NC_000950, NC_000951, NC_000952, NC_000953, NC_000954, NC_000955, NC_000956, NC_000957, NC_001318, NC_001849, NC_001850, NC_001851, NC_001852, NC_001853, NC_001854, NC_001855, NC_001856, NC_001857, NC_001903, NC_001904**
Borrelia garinii PBi	**NC_006128, NC_006129, NC_006156**
Bradyrhizobium japonicum	**NC_004463**
Brucella abortus 9-941	**NC_006932, NC_006933**
Brucella melitensis	**NC_003317, NC_003318**
Brucella melitensis biovar Abortus	NC_007618, NC_007624
Brucella suis 1330	**NC_004310, NC_004311**
Buchnera aphidicola	**NC_004545, NC_004555**
Buchnera aphidicola Cc Cinara cedri	NC_008513
Buchnera aphidicola Sg	NC_004061
Buchnera sp	**NC_002252, NC_002253, NC_002528**
Burkholderia 383	**NC_007509, NC_007510, NC_007511**
Burkholderia cenocepacia AU 1054	**NC_008060, NC_008061, NC_008062**
Burkholderia cenocepacia HI2424	NC_008542, NC_008543, NC_008544, NC_008545
Burkholderia cepacia AMMD	**NC_008385, NC_008390, NC_008391, NC_008392**
Burkholderia mallei ATCC 23344	**NC_006348, NC_006349**
Burkholderia mallei NCTC 10229	NC_008835, NC_008836
Burkholderia mallei SAVP1	NC_008784, NC_008785
Burkholderia pseudomallei 1710b	**NC_007434, NC_007435**
Burkholderia pseudomallei K96243	NC_006350, NC_006351
Burkholderia thailandensis E264	**NC_007650, NC_007651**
Burkholderia xenovorans LB400	**NC_007951, NC_007952, NC_007953**
Campylobacter fetus 82-40	**NC_008599**
Campylobacter jejuni	**NC_002163**
Campylobacter jejuni 81-176	NC_008770, NC_008787, NC_008790
Campylobacter jejuni RM1221	NC_003912
Candidatus Blochmannia floridanus	**NC_005061**
Candidatus Blochmannia pennsylvanicus BPEN	NC_007292
Candidatus Carsonella ruddii PV	**NC_008512**
Candidatus Pelagibacter ubique HTCC1062	**NC_007205**
Candidatus Ruthia magnifica Cm Calyptogena magnifica	**NC_008610**
Carboxydothermus hydrogenoformans Z-2901	**NC_007503**
Caulobacter crescentus	**NC_002696**
Chlamydia muridarum	**NC_002182, NC_002620**
Chlamydia trachomatis	**NC_000117**
Chlamydia trachomatis A HAR-13	NC_007429, NC_007430
Chlamydophila abortus S26 3	**NC_004552**
Chlamydophila caviae	**NC_003361, NC_004720**
Chlamydophila felis Fe C-56	**NC_007899, NC_007900**
Chlamydophila pneumoniae AR39	**NC_002179**
Chlamydophila pneumoniae CWL029	NC_000922
Chlamydophila pneumoniae J138	NC_002491
Chlamydophila pneumoniae TW 183	NC_005043
Chlorobium chlorochromatii CaD3	**NC_007514**
Chlorobium phaeobacteroides DSM 266	**NC_008639**
Chlorobium tepidum TLS	**NC_002932**
Chromobacterium violaceum	**NC_005085**
Chromohalobacter salexigens DSM 3043	**NC_007963**

Continued. . .

Species name	Accessions
Clostridium acetobutylicum	**NC_001988, NC_003030**
Clostridium novyi NT	**NC_008593**
Clostridium perfringens	**NC_003042, NC_003366**
Clostridium perfringens ATCC 13124	NC_008261
Clostridium perfringens SM101	NC_008262, NC_008263, NC_008264, NC_008265
Clostridium tetani E88	**NC_004557, NC_004565**
Clostridium thermocellum ATCC 27405	**NC_009012**
Colwellia psychrerythraea 34H	**NC_003910**
Corynebacterium diphtheriae	**NC_002935**
Corynebacterium efficiens YS-314	**NC_004369**
Corynebacterium glutamicum ATCC 13032 Bielefeld	**NC_006958**
Corynebacterium glutamicum ATCC 13032 Kitasato	NC_003450
Corynebacterium jeikeium K411	**NC_003080, NC_007164**
Coxiella burnetii	**NC_002971, NC_004704**
Cyanobacteria bacterium Yellowstone A-Prime	**NC_007775**
Cyanobacteria bacterium Yellowstone B-Prime	NC_007776
Cytophaga hutchinsonii ATCC 33406	**NC_008255**
Dechloromonas aromatica RCB	**NC_007298**
Dehalococcoides CBDB1	**NC_007356**
Dehalococcoides ethenogenes 195	**NC_002936**
Deinococcus geothermalis DSM 11300	**NC_008010, NC_008025**
Deinococcus radiodurans	**NC_000958, NC_000959, NC_001263, NC_001264**
Desulfitobacterium hafniense Y51	**NC_007907**
Desulfotalea psychrophila LSv54	**NC_006138, NC_006139, NC_006140**
Desulfovibrio desulfuricans G20	**NC_007519**
Desulfovibrio vulgaris DP4	**NC_008741, NC_008751**
Desulfovibrio vulgaris Hildenborough	NC_002937, NC_005863
Ehrlichia canis Jake	**NC_007354**
Ehrlichia chaffeensis Arkansas	**NC_007799**
Ehrlichia ruminantium Gardel	**NC_006831**
Ehrlichia ruminantium Welgevonden	NC_005295
Ehrlichia ruminantium str. Welgevonden	NC_006832
Enterococcus faecalis V583	**NC_004668, NC_004669, NC_004670, NC_004671**
Erwinia carotovora atroseptica SCRI1043	**NC_004547**
Erythrobacter litoralis HTCC2594	**NC_007722**
Escherichia coli 536	NC_008253
Escherichia coli APEC O1	NC_008563
Escherichia coli CFT073	NC_004431
Escherichia coli K12	**NC_000913**
Escherichia coli O157H7	NC_002127, NC_002128, NC_002695
Escherichia coli O157H7 EDL933	NC_002655, NC_007414
Escherichia coli UTI89	NC_007941, NC_007946
Escherichia coli W3110	AC_000091
Francisella tularensis FSC 198	**NC_008245**
Francisella tularensis holarctica	NC_007880
Francisella tularensis holarctica OSU18	NC_008369
Francisella tularensis novicida U112	NC_008601
Francisella tularensis tularensis	NC_006570
Frankia CcI3	**NC_007777**
Frankia alni ACN14a	**NC_008278**
Fusobacterium nucleatum	**NC_003454**
Geobacillus kaustophilus HTA426	**NC_006509, NC_006510**
Geobacter metallireducens GS-15	**NC_007515, NC_007517**
Geobacter sulfurreducens	**NC_002939**
Gloeobacter violaceus	**NC_005125**
Gluconobacter oxydans 621H	**NC_006672, NC_006673, NC_006674, NC_006675, NC_006676, NC_006677**
Gramella forsetii KT0803	**NC_008571**
Granulobacter bethesdensis CGDNIH1	**NC_008343**
Haemophilus ducreyi 35000HP	**NC_002940**
Haemophilus influenzae	**NC_000907**
Haemophilus influenzae 86 028NP	NC_007146
Haemophilus somnus 129PT	**NC_006298, NC_008309**
Hahella chejuensis KCTC 2396	**NC_007645**
Haloarcula marismortui ATCC 43049	**NC_006389, NC_006390, NC_006391, NC_006392, NC_006393, NC_006394, NC_006395, NC_006396, NC_006397**

Continued. . .

Species name	Accessions
Halobacterium sp	**NC_001869, NC_002607, NC_002608**
Haloquadratum walsbyi	**NC_008212, NC_008213**
Halorhodospira halophila SL1	**NC_008789**
Helicobacter acinonychis Sheeba	**NC_008229, NC_008230**
Helicobacter hepaticus	**NC_004917**
Helicobacter pylori 26695	**NC_000915**
Helicobacter pylori HPAG1	NC_008086, NC_008087
Helicobacter pylori J99	NC_000921
Hyperthermus butylicus	**NC_008818**
Hyphomonas neptunium ATCC 15444	**NC_008358**
Idiomarina loihiensis L2TR	**NC_006512**
Jannaschia CCS1	**NC_007801, NC_007802**
Lactobacillus acidophilus NCFM	**NC_006814**
Lactobacillus brevis ATCC 367	**NC_008497, NC_008498, NC_008499**
Lactobacillus casei ATCC 334	**NC_008502, NC_008526**
Lactobacillus delbrueckii bulgaricus	**NC_008054**
Lactobacillus delbrueckii bulgaricus ATCC BAA-365	NC_008529
Lactobacillus gasseri ATCC 33323	**NC_008530**
Lactobacillus johnsonii NCC 533	**NC_005362**
Lactobacillus plantarum	**NC_004567, NC_006375, NC_006376, NC_006377**
Lactobacillus sakei 23K	**NC_007576**
Lactobacillus salivarius UCC118	**NC_006529, NC_006530, NC_007929, NC_007930**
Lactococcus lactis	**NC_002662**
Lactococcus lactis cremoris MG1363	NC_009004
Lactococcus lactis cremoris SK11	NC_008503, NC_008504, NC_008505, NC_008506, NC_008507, NC_008527
Lawsonia intracellularis PHE MN1-00	**NC_008011, NC_008012, NC_008013, NC_008014**
Legionella pneumophila Lens	**NC_006366, NC_006369**
Legionella pneumophila Paris	NC_006365, NC_006368
Legionella pneumophila Philadelphia 1	NC_002942
Leifsonia xyli xyli CTCB0	**NC_006087**
Leptospira borgpetersenii serovar Hardjo-bovis JB197	**NC_008510, NC_008511**
Leptospira borgpetersenii serovar Hardjo-bovis L550	NC_008508, NC_008509
Leptospira interrogans serovar Copenhageni	**NC_005823, NC_005824**
Leptospira interrogans serovar Lai	NC_004342, NC_004343
Leuconostoc mesenteroides ATCC 8293	**NC_008496, NC_008531**
Listeria innocua	**NC_003212, NC_003383**
Listeria monocytogenes	**NC_003210**
Listeria monocytogenes 4b F2365	NC_002973
Listeria welshimeri serovar 6b SLCC5334	**NC_008555**
Magnetococcus MC-1	**NC_008576**
Magnetospirillum magneticum AMB-1	**NC_007626**
Mannheimia succiniciproducens MBEL55E	**NC_006300**
Maricaulis maris MCS10	**NC_008347**
Marinobacter aquaeolei VT8	**NC_008738, NC_008739, NC_008740**
Mesoplasma florum L1	**NC_006055**
Mesorhizobium BNC1	**NC_008242, NC_008243, NC_008244, NC_008254**
Mesorhizobium loti	**NC_002678, NC_002679, NC_002682**
Methanobacterium thermoautotrophicum	**NC_000916**
Methanococcoides burtonii DSM 6242	**NC_007955**
Methanococcus jannaschii	**NC_000909, NC_001732, NC_001733**
Methanococcus maripaludis S2	**NC_005791**
Methanocorpusculum labreanum Z	**NC_008942**
Methanopyrus kandleri	**NC_003551**
Methanosaeta thermophila PT	**NC_008553**
Methanosarcina acetivorans	**NC_003552**
Methanosarcina barkeri fusaro	**NC_007349, NC_007355**
Methanosarcina mazei	**NC_003901**
Methanosphaera stadtmanae	**NC_007681**
Methanospirillum hungatei JF-1	**NC_007796**
Methylibium petroleiphilum PM1	**NC_008825, NC_008826**
Methylobacillus flagellatus KT	**NC_007947**
Methylococcus capsulatus Bath	**NC_002977**
Moorella thermoacetica ATCC 39073	**NC_007644**

Continued...

Species name	Accessions
Mycobacterium KMS	**NC_008703, NC_008704, NC_008705**
Mycobacterium MCS	**NC_008146, NC_008147**
Mycobacterium avium 104	**NC_008595**
Mycobacterium avium paratuberculosis	NC_002944
Mycobacterium bovis	**NC_002945**
Mycobacterium bovis BCG Pasteur 1173P2	NC_008769
Mycobacterium leprae	**NC_002677**
Mycobacterium smegmatis MC2 155	**NC_008596**
Mycobacterium tuberculosis CDC1551	**NC_002755**
Mycobacterium tuberculosis H37Rv	NC_000962
Mycobacterium ulcerans Agy99	**NC_008611**
Mycobacterium vanbaalenii PYR-1	**NC_008726**
Mycoplasma capricolum ATCC 27343	**NC_007633**
Mycoplasma gallisepticum	**NC_004829**
Mycoplasma genitalium	**NC_000908**
Mycoplasma hyopneumoniae 232	**NC_006360**
Mycoplasma hyopneumoniae 7448	NC_007332
Mycoplasma hyopneumoniae J	NC_007295
Mycoplasma mobile 163K	**NC_006908**
Mycoplasma mycoides	**NC_005364**
Mycoplasma penetrans	**NC_004432**
Mycoplasma pneumoniae	**NC_000912**
Mycoplasma pulmonis	**NC_002771**
Mycoplasma synoviae 53	**NC_007294**
Myxococcus xanthus DK 1622	**NC_008095**
Nanoarchaeum equitans	**NC_005213**
Natronomonas pharaonis	**NC_007426, NC_007427, NC_007428**
Neisseria gonorrhoeae FA 1090	**NC_002946**
Neisseria meningitidis FAM18	**NC_008767**
Neisseria meningitidis MC58	NC_003112
Neisseria meningitidis Z2491	NC_003116
Neorickettsia sennetsu Miyayama	**NC_007798**
Nitrobacter hamburgensis X14	**NC_007959, NC_007960, NC_007961, NC_007964**
Nitrobacter winogradskyi Nb-255	**NC_007406**
Nitrosococcus oceani ATCC 19707	**NC_007483, NC_007484**
Nitrosomonas europaea	**NC_004757**
Nitrosomonas eutropha C71	**NC_008341, NC_008342, NC_008344**
Nitrosospira multiformis ATCC 25196	**NC_007614, NC_007615, NC_007616, NC_007617**
Nocardia farcinica IFM10152	**NC_006361, NC_006362, NC_006363**
Nocardioides JS614	**NC_008697, NC_008699**
Nostoc sp	**NC_003240, NC_003241, NC_003267, NC_003270, NC_003272, NC_003273, NC_003276**
Novosphingobium aromaticivorans DSM 12444	**NC_007794**
Oceanobacillus iheyensis	**NC_004193**
Oenococcus oeni PSU-1	**NC_008528**
Onion yellows phytoplasma	**NC_005303**
Parachlamydia sp UWE25	**NC_005861**
Paracoccus denitrificans PD1222	**NC_008686, NC_008687, NC_008688**
Pasteurella multocida	**NC_002663**
Pediococcus pentosaceus ATCC 25745	**NC_008525**
Pelobacter carbinolicus	**NC_007498**
Pelobacter propionicus DSM 2379	**NC_008607, NC_008608, NC_008609**
Pelodictyon luteolum DSM 273	**NC_007512**
Photobacterium profundum SS9	**NC_005871, NC_006370, NC_006371**
Photorhabdus luminescens	**NC_005126**
Picrophilus torridus DSM 9790	**NC_005877**
Pirellula sp	**NC_005027**
Polaromonas JS666	**NC_007948, NC_007949, NC_007950**
Polaromonas naphthalenivorans CJ2	**NC_008757, NC_008758, NC_008759, NC_008760, NC_008761, NC_008762, NC_008763, NC_008764, NC_008781**
Porphyromonas gingivalis W83	**NC_002950**

Continued. . .

Species name	Accessions
Prochlorococcus marinus AS9601	**NC_008816**
Prochlorococcus marinus CCMP1375	NC_005042
Prochlorococcus marinus MED4	NC_005072
Prochlorococcus marinus MIT9313	NC_005071
Prochlorococcus marinus MIT 9303	NC_008820
Prochlorococcus marinus MIT 9312	NC_007577
Prochlorococcus marinus MIT 9515	NC_008817
Prochlorococcus marinus NATL1A	NC_008819
Prochlorococcus marinus NATL2A	NC_007335
Propionibacterium acnes KPA171202	**NC_006085**
Pseudoalteromonas atlantica T6c	**NC_008228**
Pseudoalteromonas haloplanktis TAC125	**NC_007481, NC_007482**
Pseudomonas aeruginosa	**NC_002516**
Pseudomonas aeruginosa UCBPP-PA14	NC_008463
Pseudomonas entomophila L48	**NC_008027**
Pseudomonas fluorescens Pf-5	**NC_004129**
Pseudomonas fluorescens PfO-1	NC_007492
Pseudomonas putida KT2440	**NC_002947**
Pseudomonas syringae phaseolicola 1448A	**NC_005773, NC_007274, NC_007275**
Pseudomonas syringae pv B728a	NC_007005
Pseudomonas syringae tomato DC3000	NC_004578, NC_004632, NC_004633
Psychrobacter arcticum 273-4	**NC_007204**
Psychrobacter cryohalolentis K5	**NC_007968, NC_007969**
Psychromonas ingrahamii 37	**NC_008709**
Pyrobaculum aerophilum	**NC_003364**
Pyrobaculum islandicum DSM 4184	**NC_008701**
Pyrococcus abyssi	**NC_000868, NC_001773**
Pyrococcus furiosus	**NC_003413**
Pyrococcus horikoshii	**NC_000961**
Ralstonia eutropha H16	**NC_008313, NC_008314**
Ralstonia eutropha JMP134	NC_007336, NC_007337, NC_007347, NC_007348
Ralstonia metallidurans CH34	**NC_007971, NC_007972, NC_007973, NC_007974**
Ralstonia solanacearum	**NC_003295, NC_003296**
Rhizobium etli CFN 42	**NC_007761, NC_007762, NC_007763, NC_007764, NC_007765, NC_007766**
Rhizobium leguminosarum bv viciae 3841	**NC_008378, NC_008379, NC_008380, NC_008381, NC_008382, NC_008383, NC_008384**
Rhodobacter sphaeroides 2 4 1	**NC_007488, NC_007489, NC_007490, NC_007493, NC_007494, NC_009007, NC_009008**
Rhodococcus RHA1	**NC_008268, NC_008269, NC_008270, NC_008271**
Rhodoferax ferrireducens T118	**NC_007901, NC_007908**
Rhodopseudomonas palustris BisA53	**NC_008435**
Rhodopseudomonas palustris BisB18	NC_007925
Rhodopseudomonas palustris BisB5	NC_007958
Rhodopseudomonas palustris CGA009	NC_005296, NC_005297
Rhodopseudomonas palustris HaA2	NC_007778
Rhodospirillum rubrum ATCC 11170	**NC_007641, NC_007643**
Rickettsia bellii RML369-C	**NC_007940**
Rickettsia conorii	**NC_003103**
Rickettsia felis URRWXCal2	**NC_007109, NC_007110, NC_007111**
Rickettsia prowazekii	**NC_000963**
Rickettsia typhi wilmington	**NC_006142**
Roseobacter denitrificans OCh 114	**NC_008209, NC_008386, NC_008387, NC_008388, NC_008389**
Rubrobacter xylanophilus DSM 9941	**NC_008148**
Saccharophagus degradans 2-40	**NC_007912**
Salinibacter ruber DSM 13855	**NC_007677, NC_007678**
Salmonella enterica Choleraesuis	**NC_006855, NC_006856, NC_006905**
Salmonella enterica Paratypi ATCC 9150	NC_006511
Salmonella typhi	**NC_003198, NC_003384, NC_003385**
Salmonella typhi Ty2	NC_004631
Salmonella typhimurium LT2	**NC_003197, NC_003277**

Continued...

Species name	Accessions
Shewanella ANA-3	**NC_008573, NC_008577**
Shewanella MR-4	**NC_008321**
Shewanella MR-7	**NC_008320, NC_008322**
Shewanella W3-18-1	**NC_008750**
Shewanella amazonensis SB2B	**NC_008700**
Shewanella denitrificans OS217	**NC_007954**
Shewanella frigidimarina NCIMB 400	**NC_008345**
Shewanella oneidensis	**NC_004347, NC_004349**
Shigella boydii Sb227	**NC_007608, NC_007613**
Shigella dysenteriae	**NC_007606, NC_007607**
Shigella flexneri 2a	**NC_004337, NC_004851**
Shigella flexneri 2a 2457T	NC_004741
Shigella flexneri 5 8401	NC_008258
Shigella sonnei Ss046	**NC_007384, NC_007385**
Silicibacter TM1040	**NC_008042, NC_008043, NC_008044**
Silicibacter pomeroyi DSS-3	**NC_003911, NC_006569**
Sinorhizobium meliloti	**NC_003037, NC_003047, NC_003078**
Sodalis glossinidius morsitans	**NC_007712, NC_007713, NC_007714, NC_007715**
Solibacter usitatus Ellin6076	**NC_008536**
Sphingopyxis alaskensis RB2256	**NC_008036, NC_008048**
Staphylococcus aureus COL	**NC_002951, NC_006629**
Staphylococcus aureus MW2	NC_003923
Staphylococcus aureus Mu50	NC_002758, NC_002774
Staphylococcus aureus N315	NC_002745, NC_003140
Staphylococcus aureus NCTC 8325	NC_007795
Staphylococcus aureus RF122	NC_007622
Staphylococcus aureus USA300	NC_007790, NC_007791, NC_007792, NC_007793
Staphylococcus aureus aureus MRSA252	NC_002952
Staphylococcus aureus aureus MSSA476	NC_002953, NC_005951
Staphylococcus epidermidis ATCC 12228	**NC_004461, NC_005003, NC_005004, NC_005005, NC_005006, NC_005007, NC_005008**
Staphylococcus epidermidis RP62A	NC_002976, NC_006663
Staphylococcus haemolyticus	**NC_007168**
Staphylococcus saprophyticus	**NC_007350, NC_007351, NC_007352**
Streptococcus agalactiae 2603	**NC_004116**
Streptococcus agalactiae A909	NC_007432
Streptococcus agalactiae NEM316	NC_004368
Streptococcus mutans	**NC_004350**
Streptococcus pneumoniae D39	**NC_008533**
Streptococcus pneumoniae R6	NC_003098
Streptococcus pneumoniae TIGR4	NC_003028
Streptococcus pyogenes M1 GAS	**NC_002737**
Streptococcus pyogenes MGAS10270	NC_008022
Streptococcus pyogenes MGAS10394	NC_006086
Streptococcus pyogenes MGAS10750	NC_008024
Streptococcus pyogenes MGAS2096	NC_008023
Streptococcus pyogenes MGAS315	NC_004070
Streptococcus pyogenes MGAS5005	NC_007297
Streptococcus pyogenes MGAS6180	NC_007296
Streptococcus pyogenes MGAS8232	NC_003485
Streptococcus pyogenes MGAS9429	NC_008021
Streptococcus pyogenes SSI-1	NC_004606
Streptococcus sanguinis SK36	**NC_009009**
Streptococcus thermophilus CNRZ1066	**NC_006449**
Streptococcus thermophilus LMD-9	NC_008500, NC_008501, NC_008532
Streptococcus thermophilus LMG 18311	NC_006448
Streptomyces avermitilis	**NC_003155, NC_004719**
Streptomyces coelicolor	**NC_003888, NC_003903, NC_003904**
Sulfolobus acidocaldarius DSM 639	**NC_007181**
Sulfolobus solfataricus	**NC_002754**
Sulfolobus tokodaii	**NC_003106**
Symbiobacterium thermophilum IAM14863	**NC_006177**

Continued...

Species name	Accessions
Synechococcus CC9311	**NC_008319**
Synechococcus CC9605	**NC_007516**
Synechococcus CC9902	**NC_007513**
Synechococcus elongatus PCC 6301	**NC_006576**
Synechococcus elongatus PCC 7942	NC_007595, NC_007604
Synechococcus sp WH8102	**NC_005070**
Synechocystis PCC6803	**NC_000911, NC_005229, NC_005230, NC_005231, NC_005232**
Syntrophobacter fumaroxidans MPOB	**NC_008554**
Syntrophomonas wolfei Goettingen	**NC_008346**
Syntrophus aciditrophicus SB	**NC_007759**
Thermoanaerobacter tengcongensis	**NC_003869**
Thermobifida fusca YX	**NC_007333**
Thermococcus kodakaraensis KOD1	**NC_006624**
Thermofilum pendens Hrk 5	**NC_008696, NC_008698**
Thermoplasma acidophilum	**NC_002578**
Thermoplasma volcanium	**NC_002689**
Thermosynechococcus elongatus	**NC_004113**
Thermotoga maritima	**NC_000853**
Thermus thermophilus HB27	**NC_005835, NC_005838**
Thermus thermophilus HB8	NC_006461, NC_006462, NC_006463
Thiobacillus denitrificans ATCC 25259	**NC_007404**
Thiomicrospira crunogena XCL-2	**NC_007520**
Thiomicrospira denitrificans ATCC 33889	**NC_007575**
Treponema denticola ATCC 35405	**NC_002967**
Treponema pallidum	**NC_000919**
Trichodesmium erythraeum IMS101	**NC_008312**
Tropheryma whipplei TW08 27	**NC_004551**
Tropheryma whipplei Twist	NC_004572
Ureaplasma urealyticum	**NC_002162**
Verminephrobacter eiseniae EF01-2	**NC_008771, NC_008786**
Vibrio cholerae	**NC_002505, NC_002506**
Vibrio fischeri ES114	**NC_006840, NC_006841, NC_006842**
Vibrio parahaemolyticus	**NC_004603, NC_004605**
Vibrio vulnificus CMCP6	**NC_004459, NC_004460**
Vibrio vulnificus YJ016	NC_005128, NC_005139, NC_005140
Wigglesworthia brevipalpis	**NC_003425, NC_004344**
Wolbachia endosymbiont of Brugia malayi TRS	**NC_006833**
Wolbachia endosymbiont of Drosophila melanogaster	NC_002978
Wolinella succinogenes	**NC_005090**
Xanthomonas campestris	**NC_003902**
Xanthomonas campestris 8004	NC_007086
Xanthomonas campestris vesicatoria 85-10	NC_007504, NC_007505, NC_007506, NC_007507, NC_007508
Xanthomonas citri	**NC_003919, NC_003921, NC_003922**
Xanthomonas oryzae KACC10331	**NC_006834**
Xanthomonas oryzae MAFF 311018	NC_007705
Xylella fastidiosa	**NC_002488, NC_002489, NC_002490**
Xylella fastidiosa Temecula1	NC_004554, NC_004556
Yersinia enterocolitica 8081	**NC_008791, NC_008800**
Yersinia pestis Antiqua	**NC_008120, NC_008121, NC_008122, NC_008150**
Yersinia pestis CO92	NC_003131, NC_003132, NC_003134, NC_003143
Yersinia pestis KIM	NC_004088, NC_004838
Yersinia pestis Nepal516	NC_008118, NC_008119, NC_008149
Yersinia pestis biovar Mediaevails	NC_005810, NC_005813, NC_005814, NC_005815, NC_005816
Yersinia pseudotuberculosis IP32953	**NC_006153, NC_006154, NC_006155**
Zymomonas mobilis ZM4	**NC_006526**

Appendix B

GLM predictions

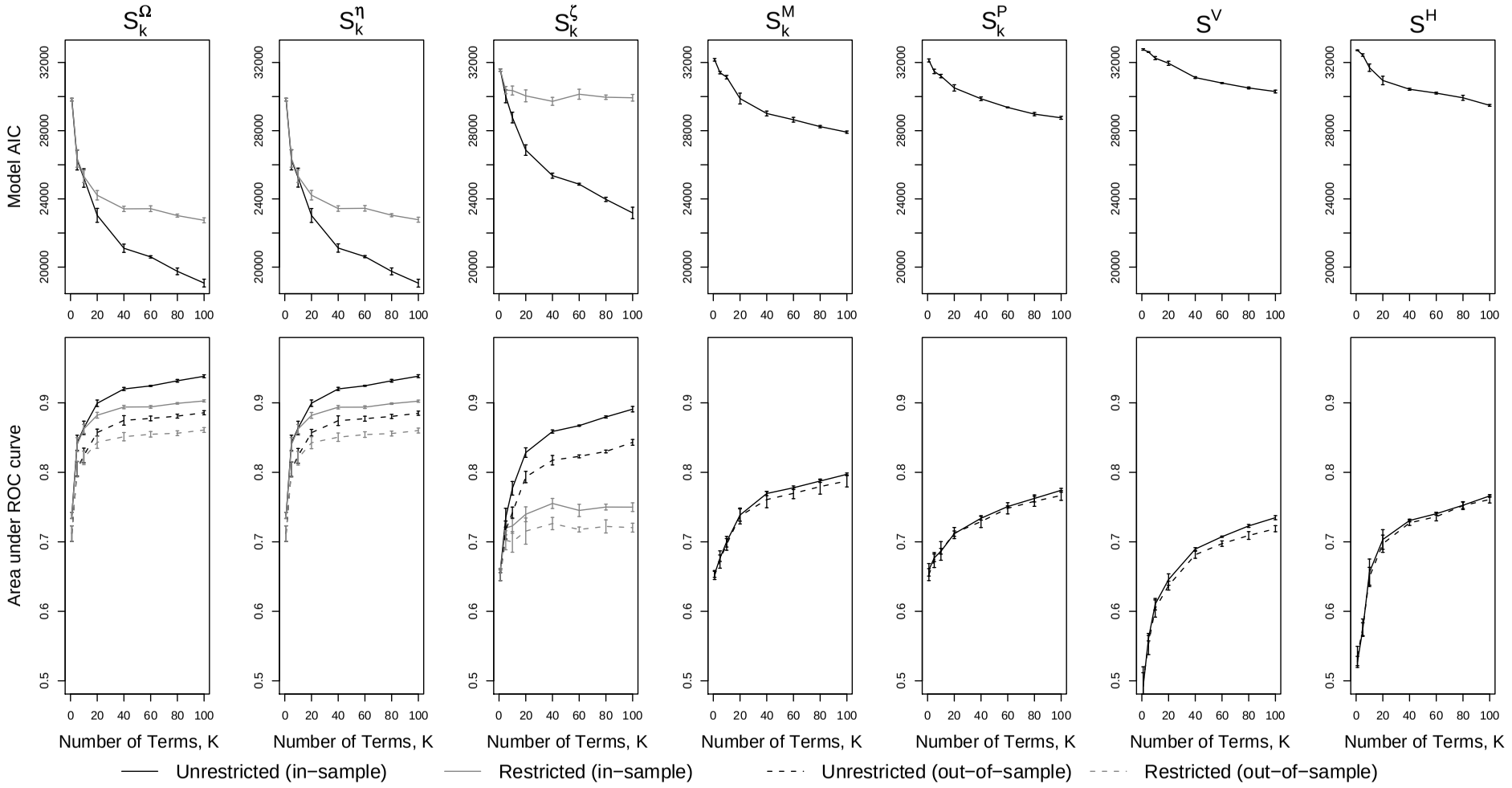


Figure B.1: Model performance on full dataset, with 80% for training, using MUSCLE MSAs. *cf.* Figure 6.2 which uses the same dataset but with CLUSTAL W MSAs, and gives very similar prediction performance.

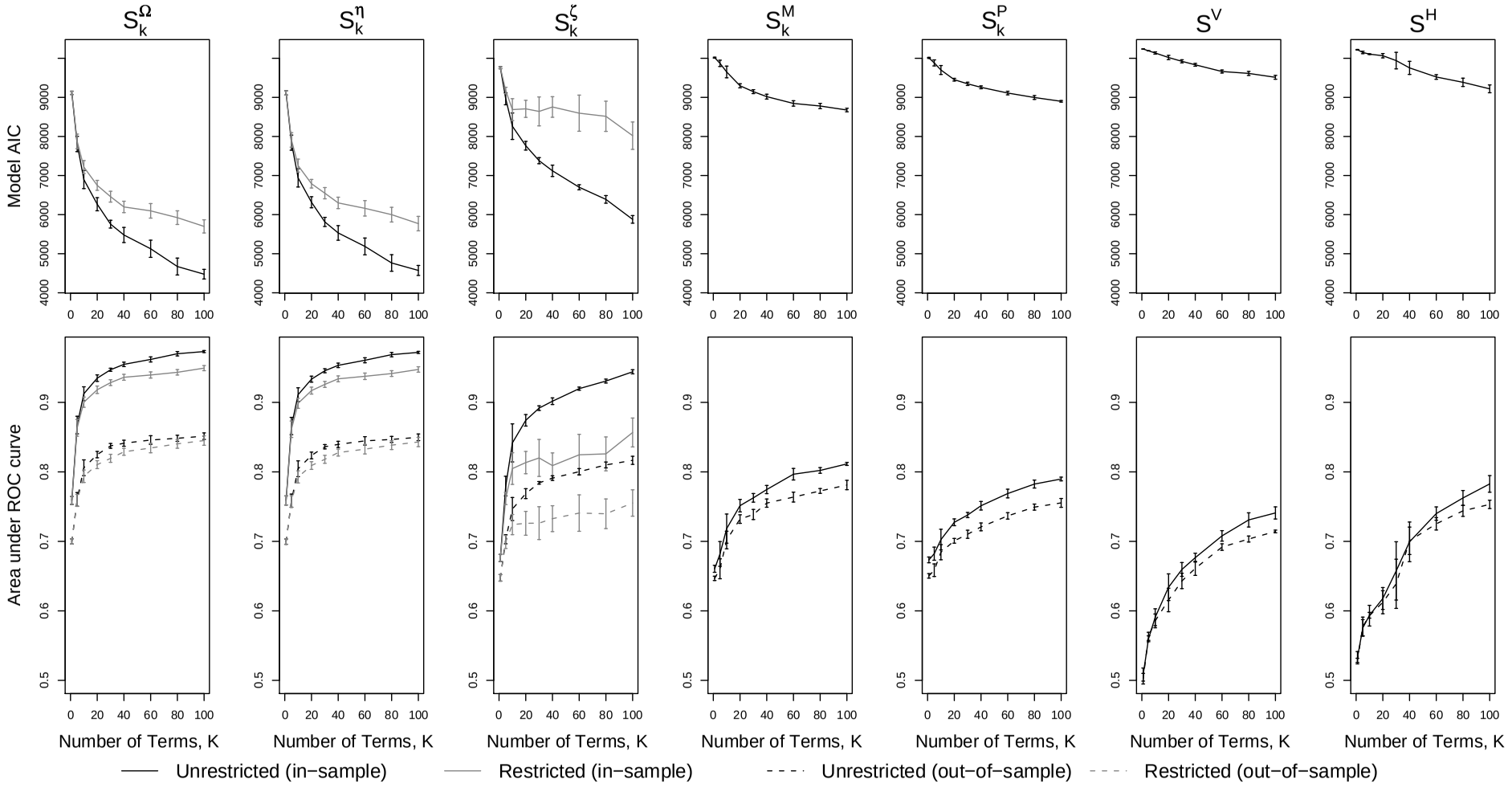


Figure B.2: Model performance on full dataset, using CLUSTAL W MSAs. Five randomised splits of 25% training and 75% test. Key as per Figure 6.2 (the same dataset but using 80% for training) and later figures using a similarly sized training set.

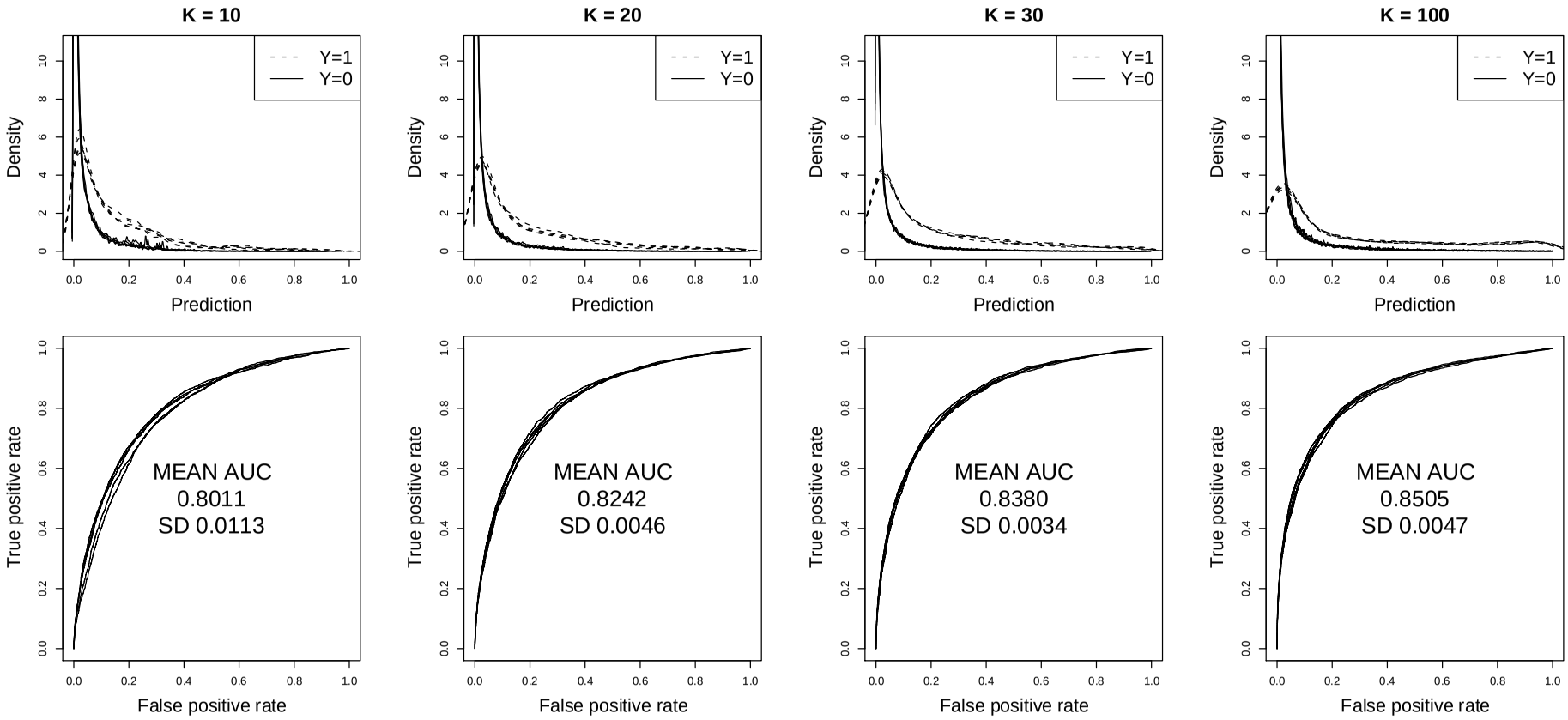


Figure B.3: Unrestricted model performance on full dataset, with 25% for training, using CLUSTAL W MSAs and S_k^Ω for $K = 10$, 20, 30 and 100 (from left to right). Five random splits of the data into training and test are shown. The top row of figures show the predicted interaction probabilities for the expected interactions ($Y = 1$) and non-interactions ($Y = 0$), showing the class separation improves with higher K. The bottom row of figures are ROC plots, labelled with the mean area under the curve, where the area increases as K is increased. *cf.* Figure 6.3 where 80% of the data is used for training..

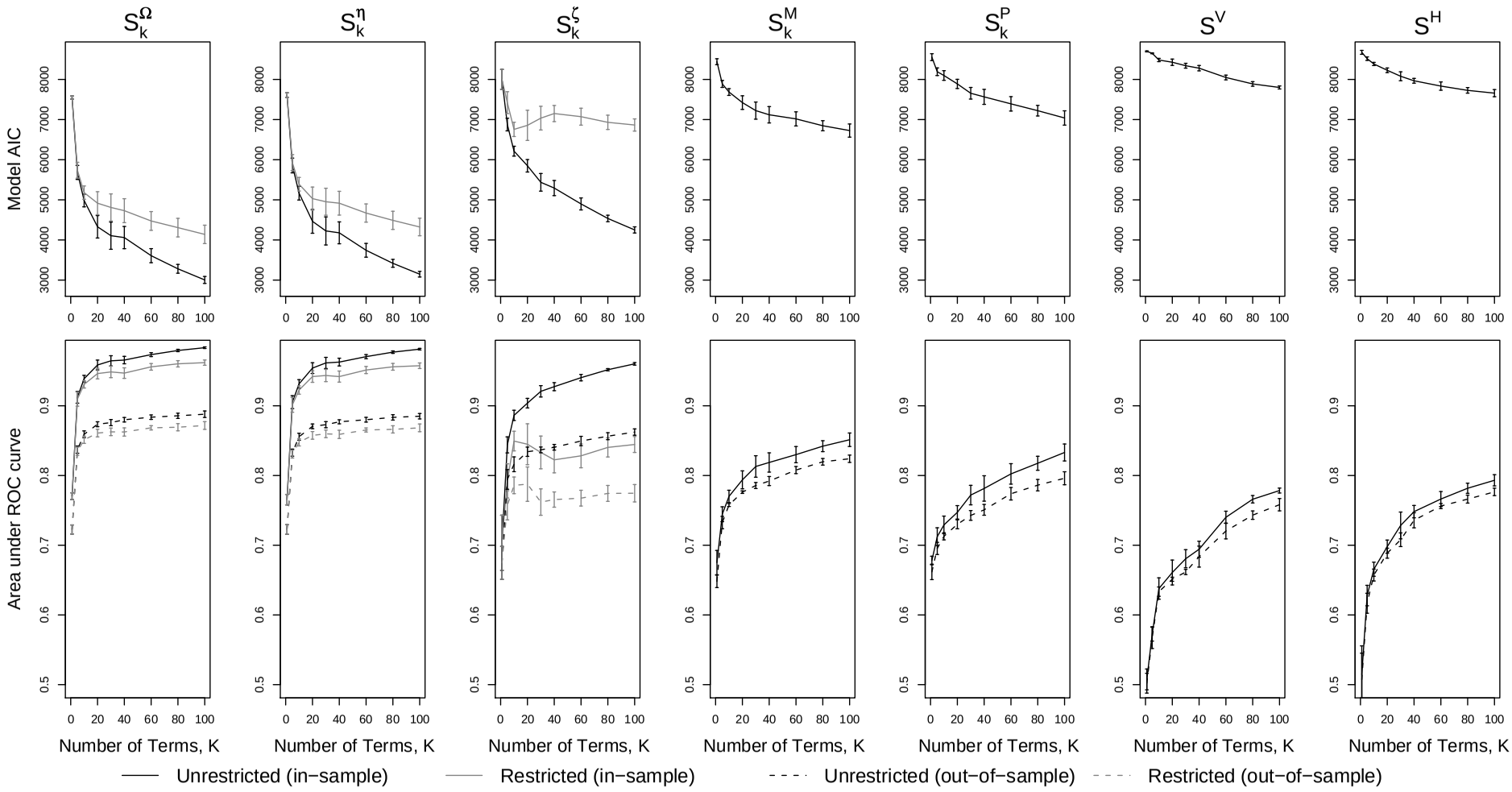


Figure B.4: Model performance using only conventional two gene TCS systems, using CLUSTAL W MSAs. Five randomised splits, trained on 33% ($1,146$ positive interactions and $18,165$ non-interactions), tested on 67% ($2,327$ and $36,881$). Key as per Figure 6.2, *cf.* Figure 6.4 where 80% of the data is used for training.

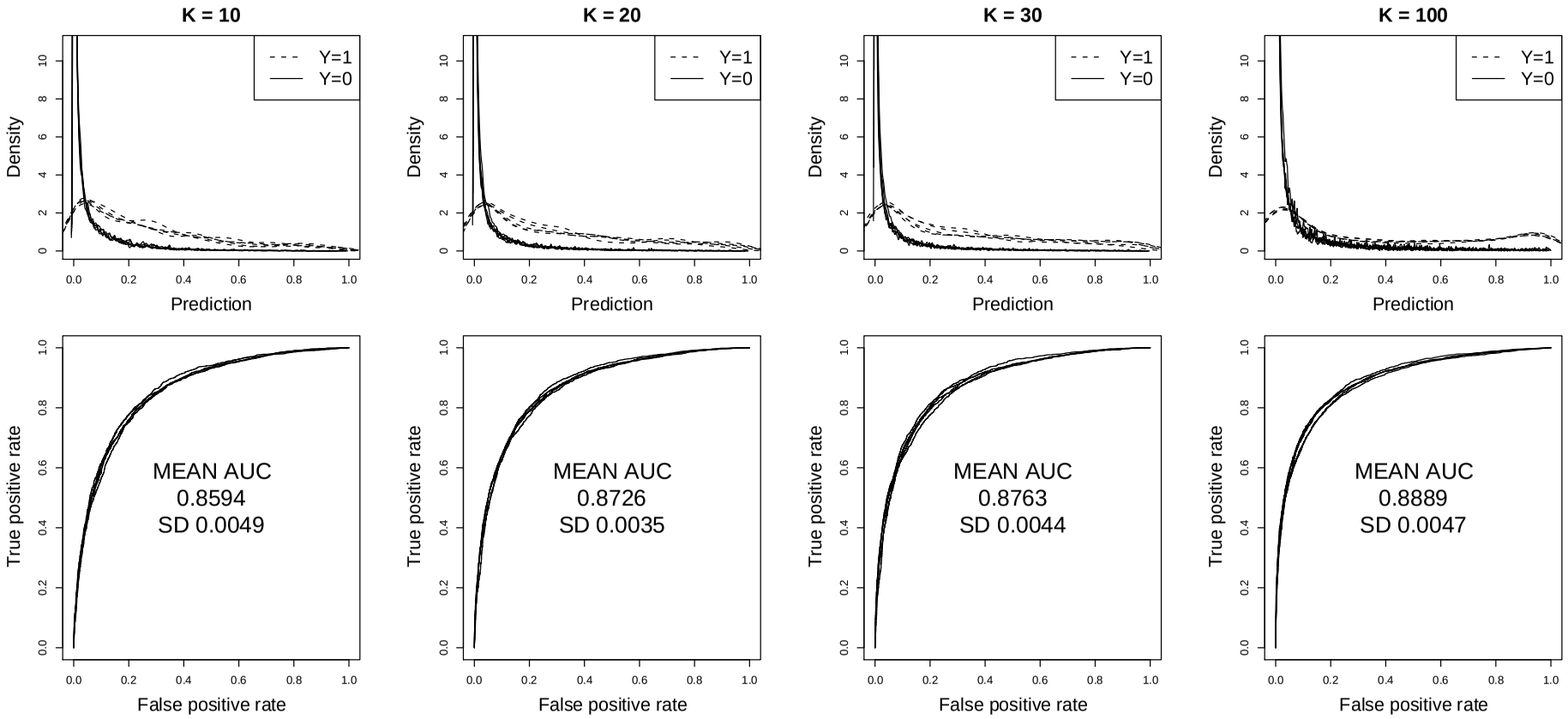


Figure B.5: Unrestricted model performance using only conventional two gene TCS systems, with 33% for training, using CLUSTAL W MSAs and S_k^{Ω} for $K = 10$, 20, 30 and 100 (from left to right). Five random splits of the data into training and test are shown. The top row of figures show the predicted interaction probabilities for the expected interactions ($Y = 1$) and non-interactions ($Y = 0$), showing the class separation improves with higher K. The bottom row of figures are ROC plots, labelled with the mean area under the curve, where the area increases as K is increased. *cf.* Figure 6.5

Bibliography

Adami, C. (2004). Information theory in molecular biology. *Physics of Life Reviews*, **1**(1), 3–22.

Adler, J. (1975). Chemotaxis in bacteria. *Annu. Rev. Biochem.*, **44**, 341–356.

Aiba, H., Nakasai, F., Mizushima, S., and Mizuno, T. (1989). Evidence for the physiological importance of the phosphotransfer between the two regulatory components, EnvZ and OmpR, in osmoregulation in *Escherichia coli*. *J. Biol. Chem.*, **264**, 3973–3977.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Ansaldi, M., Jourlin-Castelli, C., Lepelletier, M., Théraulaz, L., and Méjean, V. (2001). Rapid dephosphorylation of the TorR response regulator by the TorS unorthodox sensor in *Escherichia coli*. *J. Bact.*, **183**(8), 2691–2695.

Aoyama, K., Mitsubayashi, Y., Aiba, H., and Mizuno, T. (2000). Spy1, a histidine-containing phosphotransfer signaling protein, regulates the fission yeast cell cycle through the Mcs4 response regulator. *J. Bact.*, **182**(17), 4868–4874.

Appleby, J. L., Parkinson, J. S., and Bourret, R. B. (1996). Signal transduction via the multi-step phosphorelay: not necessarily a road less traveled. *Cell*, **86**, 845–848.

Atchley, W. R., Terhalle, W., and Dress, A. W. (1999). Positional dependence, cliques, and predictive motifs in the bHLH protein domain. *J. Mol. Evol.*, **48**(5), 501–516.

Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W., and Dress, A. W. (2000). Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol. Biol. Evol.*, **17**, 164–178.

Atkinson, M. R. and Ninfa, A. J. (1999). Two-component systems. In S. Baumberg, editor, *Prokaryotic Gene Expression*, pages 194–228. Oxford University Press, Oxford, UK.

Ausmees, N. and Jacobs-Wagner, C. (2003). Spatial and temporal control of differentiation and cell cycle progression in *Caulobacter crescentus*. *Annu. Rev. Biobiol.*, **57**, 225–247.

Baikalov, I., Schroder, I., Kaczor-Grzeskowiak, M., Grzeskowiak, K., Gunsalus, R. P., and Dickerson, R. E. (1996). Structure of the *Escherichia coli* response regulator NarL. *Biochemistry*, **35**(34), 11053–11061.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**(5), 412–424.

Barák, I., Ricca, E., and Cutting, S. M. (2005). Micromeeting - from fundamental studies of sporulation to applied spore research. *Mol. Micro.*, **55**(2), 330–338.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Res.*, **32**(Database issue), D138–D141.

Berleman, J. E. and Bauer, C. E. (2005). A che-like signal transduction cascade involved in controlling flagella biosynthesis in *Rhodospirillum centenum*. *Mol. Micro.*, **55**(5), 1390–1402.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Structure*, **28**(Nucleic Acids Res.), 235–242.

Bijlsma, J. J. E. and Groisman, E. A. (2003). Making informed decisions: regulatory interactions between two-component systems. *Trends Microbiol.*, **11**(8), 359–366.

Bilwes, A., Alex, L., Crane, B., and Simon, M. (1999). Structure of CheA, a signal-transducing histidine kinase. *Cell*, **96**, 131–141.

Birck, C., Mourey, L., Gouet, P., Fabry, B., Schumacher, J., Rousseau, P., Kahn, D., and Samama, J.-P. (1999). Conformational changes induced by phosphorylation of the FixJ receiver domain. *Structure*, **7**(12), 1505–1515.

Borodovsky, M., Hayes, W. S., and Lukashin, A. V. (1999). Statistical predictions of coding regions in prokaryotic genomes by using inhomogeneous markov models. In R. Charlebois, editor, *Organisation of the prokaryotic genome*. ASM Press, Washington, DC, USA.

Bourret, R. B. and Stock, A. M. (2000). Molecular information processing: Lessons from bacterial chemotaxis. *J. Biol. Chem.*, **277**(12), 9625–9628.

Brencic, A., Xia, Q., and Winans, S. C. (2004). VirA of *Agrobacterium tumefaciens* is an intradimer transphosphorylase and can actively block *vir* gene expression in the absence of phenolic signals. *Mol. Micro.*, **52**(5), 1349–1362.

Buck, M. J. and Atchley, W. R. (2005). Networks of coevolving sites in structural and functional domains of serpin proteins. *Mol. Biol. Evol.*, **22**(7), 1627–1634.

Buck, V., Quinn, J., Soto Pino, T., Martin, H., Saldanha, J., Makino, K., Morgan, B. A., and Millar, J. B. (2001). Peroxide sensors for the fission yeast stress-activated mitogen-activated protein kinase pathway. *Mol. Biol. Cell.*, **12**(2), 407–419.

Burbulys, D., Trach, K., and Hoch, J. (1991). The initiation of sporulation in *Bacillus subtilis* is controlled by a multicomponent phosphorelay. *Cell*, **64**, 545–552.

Burger, L. and van Nimwegen, E. (2006). A Bayesian algorithm for reconstructing two-component signaling networks. *Lecture Notes in Bioinformatics, Proceedings 6th international workshop algorithms in bioinformatics (WABI)*.

Burger, L. and van Nimwegen, E. (2008). Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Molecular Systems Biology*, **12**.

Cabrera-Vera, T. M., Vanhauwe, J., Thomas, T. O., Medkova, M., Preininger, A., Mazzoni, M. R., and Hamm, H. E. (2003). Insights into G protein structure, function, and regulation. *Endocrine Reviews*, **24**(6), 765–781.

Cai, S., Khorchid, A., Ikura, M., and Inouye, M. (2003). Probing catalytically essential domain orientation in histidine kinase EnvZ by targeted disulfide crosslinking. *J. Mol. Biol.*, **328**, 409–418.

del Campo, A. M., Ballado, T., de la Mora, J., Poggio, S., Camarena, L., and Dreyfus, G. (2007). Chemotactic control of the two flagellar systems of *Rhodobacter sphaeroides* is mediated by different sets of CheY and FliM proteins. *J. Bact.*, **189**(22), 8397–8401.

Catlett, N. L., Yoder, O. C., and Turgeon, B. G. (2003). Whole-genome analysis of two-component signal transduction genes in fungal pathogens. *Eukaryotic Cell*, **2**(6), 1151–1161.

Chang, C. H. and Winans, S. C. (1992). Functional roles assigned to the periplasmic linker and receiver domains of the *Agrobacterium tumefaciens* VirA protein. *J. Bact.*, **174**(21), 7033–7039.

Chang, C. H., Zhu, J., and Winans, S. C. (1996). Pleiotropic phenotypes caused by genetic ablation of the receiver module of the *Agrobacterium tumefaciens* VirA protein. *J. Bact.*, **178**(15), 4710–4716.

Chiu, D. K. Y. and Kolodziejczak, T. (1991). Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci.*, **7**(3), 347–352.

Clarke, D. J., Joyce, S. A., Toutain, C. M., Jacq, A., and Holland, I. B. (2002). Genetic analysis of the RcsC sensor kinase from *Escherichia coli* K-12. *J. Bact.*, **184**(4), 1204–1208.

Cock, P. J. A. and Whitworth, D. E. (2007a). Evolution of gene overlaps: relative reading frame bias in prokaryotic two-component system genes. *J. Mol. Evol.*, **64**(4), 457–462.

Cock, P. J. A. and Whitworth, D. E. (2007b). Evolution of prokaryotic two-component system signalling pathways: gene fusions and fissions. *Mol. Biol. Evol.*, **24**(11), 2355–2357.

Comley, J. W. and Dowe, D. (2005). Minimum message length and generalized Bayesian nets with asymmetric languages. In P. Grünwald, I. J. Myung, and M. A. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*, chapter 11, pages 265–294. MIT Press, Cambridge, MA, USA.

Crick, F. H. C. (1966). Codon-anticodon pairing: the wobble hypothesis. *J. Mol. Biol.*, **19**(2), 548–555.

Csonka, L. and Hanson, A. (1991). Prokaryotic osmoregulation: genetics and physiology. *Annu. Rev. Microbiol.*, **45**, 569–606.

Darwin, A. J. and Stewart, V. (1995). Expression of the *narX*, *narL*, *narP*, and *narQ* genes of *Escherichia coli* K-12: regulation of the regulators. *J Bact.*, **177**(13), 3865–3869.

Djordjevic, S., Goudreau, P., Xu, Q., Stock, A. M., and West, A. H. (1998). Structural basis for methylesterase CheB regulation by a phosphorylation-activated domain. *PNAS*, **95**(4), 1381–1386.

Dunn, S. D., Wahl, L. M., and Gloor, G. B. (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*.

Dutta, R., Yoshida, T., and Inouye, M. (2000). The critical role of the conserved Thr247 residue in the functioning of the osmosensor EnvZ, a histidine kinase/phosphatase, in *Escherichia coli*. *J. Biol. Chem.*, **275**(49), 38645–38653.

Eddy, S. (1998). Profile hidden markov models. *Bioinformatics*, **14**, 755763.

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

Edgar, R. C. and Batzoglou, S. (2006). Multiple sequence alignment. *Curr. Opin. Struct. Biol.*, **16**(3), 368–373.

Erickson, K. D. and Detweiler, C. S. (2006). The Rcs phosphorelay system is specific to enteric pathogens/commensals and activates *ydeI*, a gene important for persistent *Salmonella* infection of mice. *Mol. Micro.*, **62**(3), 883–894.

Eyre-Walker, A. (1996). The close proximity of *Escherichia coli* genes: consequences for stop codon and synonymous codon use. *J. Mol. Evol.*, **42**(2), 73–78.

Fabret, C., Feher, V. A., and Hoch, J. A. (1999). Two-component signal transduction in *Bacillus subtilis*: How one organism sees its world. *J. Bacteriol.*, **181**(7), 1975–1983.

Fawcett, T. (2003). ROC graphs: Notes and practical considerations for data mining researchers. *Technical Report HPL-2003-4, HP Labs*.

Feher, V. A., Zapf, J. W., Hoch, J. A., Whiteley, J. M., McIntosh, L. P., Rance, M., Skelton, N. J., Dahlquist, F. W., and Cavanagh, J. (1997). High-resolution NMR structure and backbone dynamics of the *Bacillus subtilis* response regulator, Spo0F: implications for phosphorylation and molecular recognition. *Biochemistry*, **36**(33), 10015–10025.

Freeman, J. A. and Bassler, B. L. (1999). Sequence and function of LuxU: a two-component phosphorelay protein that regulates quorum sensing in *Vibrio harveyi*. *J. Bact.*, **181**(3), 899–906.

Fukuda, Y., Washio, T., and Tomita, M. (1999). Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. *Nucleic Acids Res.*, **27**(8), 1847–1853.

Fukuda, Y., Nakayama, Y., and Tomita, M. (2003). On dynamics of overlapping genes in bacterial genomes. *Gene.*, **323**, 181–187.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, **5**, R80.

Georgellis, D., Kwon, O., Wulf, P. D., and Lin, E. C. C. (1998). Signal decay through a reverse phosphorelay in the Arc two-component signal transduction system. *J. Biol. Chem.*, **273**(49), 32864–32869.

Giraud, B. G., Lapedes, A., and Liu, L. C. (1998). Analysis of correlations between sites in models of protein sequences. *Phys. Rev. E*, **58**(5), 6312–6322.

Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1993). Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, **18**(4), 309–317.

Goh, C.-S., Bogan, A. A., Joachimiak, M., Walther, D., and Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. *J. Mol. Biol.*, **299**, 283–293.

Goldman, B. S., Nierman, W. C., Kaiser, D., Slater, S. C., Durkin, A. S., Eisen, J., Ronning, C. M., Barbazuk, W., Blanchard, M., Field, C., Halling, C., Hinkle, G., Iartchuk, O., Kim, H. S., Mackenzie, C., Madupu, R., Miller, N., Shvartsbeyn, A., Sullivan, S. A., Vaudin, M., Wiegand, R., and Kaplan, H. B. (2006). Evolution of sensory complexity recorded in a myxobacterial genome. *PNAS*, **103**(41), 2355–2357.

Gorodkin, J., Heyer, L., Brunak, S., and Storomo, G. (1997). Displaying the information contents of structural RNA alignments: the structure logos. *Bioinformatics*, **13**(6), 583–586.

Gouveia-Oliveira, R. and Pedersen, A. G. (2007). Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms for molecular biology*, **2**.

Goy, M. F., Springer, M. S., and Adler, J. (1977). Sensory transduction in *Escherichia coli*: role of a protein methylation reaction in sensory adaptation. *PNAS*, **74**(11), 4964–8.

Grebe, T. W. and Stock, J. B. (1999). The histidine protein kinase superfamily. *Adv. Microb. Physiol.*, **41**, 139–227.

Grefen, C. and Harter, K. (2004). Plant two-component systems: principles, functions, complexity and crosstalk. *Planta*, **219**(5), 733–742.

Guillet, V., Ohta, N., Cabantous, S., Newton, A., and Samama, J.-P. (2002). Crystallographic and biochemical studies of DivK reveal novel features of an essential response regulator in *Caulobacter crescentus*. *J. Biol. Chem.*, **277**(44), 42003–42010.

Gutell, R. R., Power, A., Hertz, G. Z., Putz, E. J., and Stormo, G. D. (1992). Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucl. Acids Res.*, **20**(21), 5785–5795.

Hamelryck, T. and Manderick, B. (2003). PDB file parser and structure class implemented in Python. *Bioinformatics*, **19**(17), 2308–2310.

Hayashi, H., Koiwai, O., and Kozuka, M. (1979). Studies on bacterial chemotaxis II. Effect of *cheB* and *cheZ* mutations on the methylation of methyl-accepting ghemotaxis protein of *Escherichia coli*. *J. Biochem.*, **85**(5), 1213–1223.

Heeb, S. and Haas, D. (2001). Regulatory roles of the GacS/GacA two-component system in plant-associated and other gram-negative bacteria. *Mol. Plant Microbe Interact.*, **14**(12), 1351–1363.

Hellingwerf, K. J., Postma, P. W., Tommassen, J., and Westerhoff, H. V. (1995). Signal transduction in bacteria: phospho-neural network(s) in *Escherichia coli* ? *FEMS Microbiol. Rev.*, **16**(4), 309–321.

Henke, J. M. and Bassler, B. L. (2004). Three parallel quorum-sensing systems regulate gene expression in *Vibrio harveyi*. *J. Bact.*, **186**(20), 6902–6914.

Higgs, P. I., Cho, K., Whitworth, D. E., Evans, L. S., and Zusman, D. R. (2005). Four unusual two-component signal transduction homologs, RedC to RedF, are necessary for timely development in *Myxococcus xanthus*. *J. Bact.*, **187**(23), 8191–8195.

Hoch, J. A. (1993). The phosphorelay signal transduction pathway in the initiation of *Bacillus subtilis* sporulation. *J. Cell. Biochem.*, **51**(1), 55–61.

Hoch, J. A. (2000). Two-component and phosphorelay signal transduction. *Curr. Opin. Microbiol.*, **3**(2), 165–170.

Hoch, J. A. and Silhavy, T. J. (1995). *Two-Component Signal Transduction*. American Society for Microbiology, Washington, DC, USA.

Hoffer, S. M., Westerhoff, H. V., Hellingwerf, K. J., Postma, P. W., and Tommassen, J. (2001). Autoamplification of a two-component regulatory system results in "learning" behavior. *J. Bact.*, **183**(16), 4914–4917.

de Hoon, M. J. L., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. *Bioinformatics*, **20**(9), 1453–1454.

Hopp, T. P. and Woods, K. R. (1981). Prediction of protein antigenic determinants from amino acid sequences. *PNAS*, **78**(6), 3824–3828.

Hrabak, E. M. and Willis, D. K. (1992). The *lemA* gene required for pathogenicity of *Pseudomonas syringae* pv. *syringae* on bean is a member of a family of two-component regulators. *J. Bact.*, **174**(9), 3011–3020.

Hughes, K. T. and Mathee, K. (1998). The anti-sigma factors. *Annu. Rev. Microbiol.*, **53**, 231–286.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, **14**, 33–38.

Hwang, I., Chen, H., and Sheen, J. (2002). Two-component signal transduction pathways in *Arabidopsis*. *Plant Physiol.*, **129**(2), 500–515.

Ikegami, T., Okada, T., Ohki, I., Hirayuma, J., Mizuno, T., and Shirakawa, M. (2001). Solution structure and dynamic character of the histidine-containing phosphotransfer domain of anaerobic sensor kinase ArcB from *Escherichia coli*. *Biochemistry*, **40**, 375–386.

Itzkovitz, S. and Alon, U. (2007). The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res.*, **17**, 405–412.

Iuchi, S., Furlong, D., and Lin, E. C. (1989). Differentiation of *arcA*, *arcB*, and *cpxA* mutant phenotypes of *Escherichia coli* by sex pilus formation and enzyme regulation. *J. Bact.*, **171**(5), 2889–2893.

Iuchi, S., Matsuda, Z., Fujiwara, T., and Lin, E. C. C. (1990). The *arcB* gene of *Escherichia coli* encodes a sensor-regulator protein for anaerobic repression of the *arc* modulon. *Mol. Micro.*, **4**(5), 715–727.

Jahreis, K., Morrison, T. B., Garzon, A., and Parkinson, J. S. (2004). Chemotactic signaling by an *Escherichia coli* CheA mutant that lacks the binding domain for phosphoacceptor partners. *J. Bact.*, **186**(9), 2664–2672.

Jiang, M., Tzeng, Y., Feher, V. A., Perego, M., and Hoch, J. A. (1999). Alanine mutants of the Spo0F response regulator modifying specificity for sensor kinases in sporulation initiation. *Mol. Micro.*, **33**(2), 389–395.

Jiang, M., Shao, W., Perego, M., and Hoch, J. A. (2000). Multiple histidine kinases regulate entry into stationary phase and sporulation in *Bacillus subtilis*. *Mol. Micro.*, **38**(3), 535–542.

Jiménez-Pearson, M., Delany, I., Scarlato, V., and Beier, D. (2005). Phosphate flow in the chemotactic response system of *Helicobacter pylori*. *Microbiol.*, **151**, 3299–3311.

Jin, S., Prusti, R. K., Roitsch, T., Ankenbauer, R. G., and Nester, E. W. (1990). Phosphorylation of the VirG protein of *Agrobacterium tumefaciens* by the autophosphorylated VirA protein: essential role in biological activity of VirG. *J. Bact.*, **172**(2), 4945–4950.

Johnson, Z. I. and Chisholm, S. W. (2004). Properties of overlapping genes are conserved across microbial genomes. *Genome Res.*, **14**, 2268–2272.

Jourlin, C., Bengrine, A., Chippaux, M., and Méjean, V. (1999). An unorthodox sensor protein (TorS) mediates the induction of the *tor* structural genes in response to trimethylamine N-oxide in *Escherichia coli*. *Mol. Micro.*, **20**(6), 1297–1306.

Karniol, B. and Vierstra, R. D. (2004). The HWE histidine kinases, a new family of bacterial two-component sensor kinases with potentially diverse roles in environmental signaling. *J. Bacteriol.*, **186**(2), 267–269.

Kato, A. and Groisman, E. A. (2004). Connecting two-component regulatory systems by a protein that protects a response regulator from dephosphorylation by its cognate sensor. *Genes & Development*, **18**(18), 2302–2313.

Kato, M., Mizuno, T., Shimizu, T., and Hakoshima, T. (1997). Insights into multistep phosphorelay from the crystal structure of the C-terminal HPt domain of ArcB. *Cell*, **88**, 717–723.

Kato, M., Shimizu, T., Mizuno, T., and Hakoshima, T. (1999). Structure of the histidine-containing phosphotransfer (HPt) domain of the anaerobic sensor protein ArcB complexed with the chemotaxis response regulator CheY. *Acta. Crystallogr., Sect D: Biol. Crystallogr.*, **55**, 1257–1263.

Kehoe, D. M. and Grossman, A. R. (1997). New classes of mutants in complementary chromatic adaptation provide evidence for a novel four-step phosphorelay system. *J. Bact.*, **179**(12), 3914–3921.

Kehoe, D. M. and Gutu, A. (2006). Responding to color: The regulation of complementary chromatic adaptation. *Annu. Rev. Plant Biol.*, **57**, 127–150.

Kendall, M. and Gibbons, J. D. (1990). *Rank Correlation Methods*. Edward Arnold, UK, fifth edition.

Kern, D., Volkman, B. F., Luginbhl, P., Nohaile, M. J., Kustu, S., and Wemmer, D. E. (1999). Structure of a transiently phosphorylated switch in bacterial signal transduction. *Nature*, **402**, 894–898.

Ketela, T., Brown, J. L., Stewart, R. C., and Bussey, H. (1998). Yeast Skn7p activity is modulated by the Sln1p-Ypd1p osmosensor and contributes to regulation of the HOG1 pathway. *Mol. Gen. Genet.*, **259**(4), 372–378.

Kim, D. and Forst, S. (2001). Genomic analysis of the histidine kinase family in bacteria and archaea. *Microbiol.*, **147**, 1197–1212.

Kim, J. and Cho, K. (2006). 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.

Kingsford, C., Delcher, A. L., and Salzberg, S. L. (2007). A unified model explaining the offsets of overlapping and near-overlapping prokaryotic genes. *Mol. Biol. Evol.*, **24**(9), 2091–2098.

Kojetin, D. J., Thompson, R. J., and Cavanagh, J. (2003). Sub-classification of response regulators using the surface characteristics of their receiver domains. *FEBS Letters*, **554**(3), 231–236.

Korber, B. T., Farber, R. M., Wolpert, D. H., and Lapedes, A. S. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *PNAS*, **90**(15), 7176–7180.

Koretke, K. K., Lupas, A. N., Warren, P. V., Rosenberg, M., and Brown, J. R. (2000). Evolution of two-component signal transduction. *Mol. Biol. Evol.*, **17**(2), 1956–1970.

Krakauer, D. (2000). Stability and evolution of overlapping genes. *Evolution*, **54**(3), 731–739.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer., E. L. L. (2001). Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes. *J. Mol. Biol.*, **305**(3), 567–580.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86.

Kummerfeld, S. K. and Teichmann, S. A. (2005). Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet.*, **21**, 25–30.

Kwon, O., Georgellis, D., and Lin, E. C. C. (2000). Phosphorelay as the sole physiological route of signal transmission by the Arc two-component system of *Escherichia coli*. *J. Bact.*, **182**(13), 3858–3862.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Biophysical Journal*, **157**, 105–132.

Larsen, S. H., Reader, R. W., Kort, E. N., Tso, W. W., and Adler, J. (1974). Change in direction of flagellar rotation is the basis of the chemotactic response in *Escherichia coli*. *Nature*, **249**(4552), 74–77.

Laub, M. T. and Goulian, M. (2007). Specificity in two-component signal transduction pathways. *Annu. Rev. Genet.*, **41**, 121–45.

Laursen, B. S., de A Steffensen, S. A., Hedegaard, J., Moreno, J. M., Mortensen, K. K., and Sperling-Petersen, H. U. (2002). Structural requirements of the mRNA for intracistronic translation initiation of the enterobacterial *infB* gene. *Genes to Cells*, **7**(9), 901–10.

Lee, J., Tomchick, D. R., Brautigam, C. A., Machius, M., Kort, R., Hellingwerf, K. J., and Gardner, K. H. (2008). Changes at the KinA PAS-A dimerization interface influence histidine kinase function. *Biochemistry*, **47**(13), 4051–4064.

Levitt, M. and Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *PNAS*, **95**(11), 59135920.

Lewis, R. J., Brannigan, J. A., Muchov, K., Bark, I., and Wilkinson, A. J. (1999). Phosphorylated aspartate in the structure of a response regulator protein. *J. Mol. Biol.*, **294**(1), 9–15.

Li, L. and Kehoe, D. M. (2005). In vivo analysis of the roles of conserved aspartate and histidine residues within a complex response regulator. *Mol. Micro.*, **55**(5), 1538–1552.

Li, S., Ault, A., Malone, C. L., Raitt, D., Dean, S., Johnston, L. H., Deschenes, R. J., and Fassler, J. S. (1998). The yeast histidine protein kinase, Sln1p, mediates phosphotransfer to two response regulators, Ssk1p and Skn7p. *EMBO J.*, **17**, 6952–6962.

Li, Y., Bustamante, V. H., Lux, R., Zusman, D., and Shi, W. (2005). Divergent regulatory pathways control A and S motility in *Myxococcus xanthus* through FrzE, a CheA-CheY fusion protein. *J. Bact.*, **187**(5), 1716–1723.

Lowry, D. F., Roth, A. F., Rupert, P. B., Dahlquist, F. W., Moy, F. J., Domaille, P. J., and Matsumura, P. (1994). Signal transduction in chemotaxis. a propagating conformation change upon phosphorylation of CheY. *J. Biol. Chem.*, **269**(42), 26358–26362.

Lu, H., Lu, L., and Skolnick, J. (2003). Development of unified statistical potentials describing protein-protein interactions. *Biophysical Journal*, **84**(3), 1895–1901.

Mackenzie, C., Choudhary, M., Larimer, F. W., Predki, P. F., Stilwagen, S., Armitage, J. P., *et al.* (2001). The home stretch, a first analysis of the nearly completed genome of *Rhodobacter sphaeroides* 2.4.1. *Photosynth. Res.*, **70**(1), 19–41.

Macnab, R. and Koshland, Jr., D. E. (1974). Bacterial motility and chemotaxis: Light-induced tumbling response and visualization of individual flagella. *J. Mol. Biol.*, **84**(3), 399–406.

Madhusudan, Zapf, J. W., Whiteley, J. M., Hoch, J. A., Xuong, N., and Varughese, K. I. (1996). Crystal structure of a phosphatase-resistant mutant of sporulation response regulator Spo0F from *Bacillus subtilis*. *Structure*, **4**(6), 679–690.

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

Mähönen, A. P., Bonke, M., Kauppinen, L., Riikonen, M., Benfey, P. N., and Helariutta, Y. (2000). A novel two-component hybrid molecule regulates vascular morphogenesis of the *Arabidopsis* root. *Genes Dev.*, **14**(23), 2938–2943.

Makalowska, I., Lin, C. F., and Makalowski, W. (2005). Overlapping genes in vertebrate genomes. *Comput. Biol. Chem.*, **29**(1), 1–12.

Makino, S., Kiba, T., Imamura, A., Hanaki, N., Nakamura, A., Suzuki, T., Taniguchi, M., Ueguchi, C., Sugiyama, T., and Mizuno, T. (2000). Genes encoding pseudo-response regulators: Insights into His-to-Asp phosphorelay and circadian rhythm in *Arabidopsis thaliana*. *Plant Cell Physiol.*, **41**, 791–803.

Makino, S., Matsushika, A., Kojima, M., Yamashino, T., and Mizuno, T. (2002). The APRR1/TOC1 quintet implicated in circadian rhythms of *Arabidopsis thaliana*: I. characterization with APRR1-overexpressing plants. *Plant and Cell Physiology*, **43**(1), 58–69.

Malpica, R., Sandoval, G. R. P., Rodriguez, C., Franco, B., and Georgellis, D. (2006). Signaling by the Arc two-component system provides a link between the redox state of the quinone pool and gene expression. *Antioxid. Redox. Signal.*, **8**, 781–795.

Marchler-Bauer, A. and Bryant, S. H. (2004). CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.*, **32**(Web Server issue), W327–W331.

Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Liebert, C. A., Liu, C., Lu, F., Marchler, G. H., Mullokandov, M., Shoemaker, B. A., Simonyan, V., Song, J. S., Thiessen, P. A., Yamashita, R. A., Yin, J. J., Zhang, D., and Bryant, S. H. (2005). CDD: a conserved domain database for protein classification. *Nucleic Acids Res.*, **33**(Database issue), D192–D196.

Martin, A. C., Wadhams, G. H., Shah, D. S. H., Porter, S. L., Mantotta, J. C., Craig, T. J., Verdult, P. H., Jones, H., and Armitage, J. P. (2001). CheR- and CheB-dependent chemosensory adaptation system of *Rhodobacter sphaeroides*. *J. Bact*, **183**(24), 7135–7144.

Martin, A. C., Nair, U., Armitage, J. P., and Maddock, J. R. (2003). Polar localization of CheA2 in *Rhodobacter sphaeroides* requires specific Che homologs. *J. Bact*, **185**(16), 4667–4671.

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

Mascher, T., Helmann, J. D., and Unden, G. (2006). Molecular information processing: Lessons from bacterial chemotaxis. *Microbiol. Mol. Biol. Rev.*, **70**(5), 910–938.

Matsushika, A. and Mizuno, T. (1998). A dual-signaling mechanism mediated by the ArcB hybrid sensor kinase containing the histidine-containing phosphotransfer domain in *Escherichia coli*. *J. Bact.*, **180**(15), 3973–3977.

Matsushika, A., Makino, S., Kojima, M., and Mizuno, T. (2000). Circadian waves of expression of the APRR1/TOC1 family of pseudo-response regulators in *Arabidopsis thaliana*: Insight into the plant circadian clock. *Plant Cell Physiol.*, **41**, 1002–1012.

McAdams, H. H. and Shapiro, L. (2003). A bacterial cell-cycle regulatory network operating in time and space. *Science*, **301**(5641), 1874–1877.

McCarter, L. L. (2004). Dual flagellar systems enable motility under different circumstances. *J. Mol. Microbiol. Biotechnol.*, **7**, 18–29.

McCarthy, J. E. (1990). Post-transcriptional control in the polycistronic operon environment: studies of the *atp* operon of *Escherichia coli*. *Mol. Micro.*, **4**(8), 1233–1240.

McLaughlin, P. D., Bobay, B. G., Regel, E. J., Thompson, R. J., Hoch, J. A., and Cavanagh, J. (2007). Predominantly buried residues in the response regulator Spo0F influence specific sensor kinase recognition. *FEBS Letters*, **581**(7), 1425–1429.

McNab, R. and Lamont, R. J. (2003). Microbial dinner-party conversations: the role of LuxS in interspecies communication. *J. Med. Microbiol.*, **52**, 541–545.

Miller, M. B., Skorupski, K., Lenz, D., Taylor, R. K., and Bassler, B. L. (2002). Parallel quorum sensing systems converge to regulate virulence in *Vibrio cholerae*. *Cell*, **110**(3), 303–314.

Miyazawa, S. and Jernigan, R. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**(3), 534–552.

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

Mizuno, T., Kato, M., Jo, Y.-L., and Mizushim, S. (1988). Interaction of OmpR, a positive regulator, with osmoregulators *ompC* and *ompF* genes of *Escherichia coli*. studies with wild-type and mutant OmpR proteins. *J. Biol. Chem.*, **263**(2), 1008–1012.

Molle, V., Fujita, M., Jensen, S., Eichenberger, P., Gonzalez-Pastor, J., Liu, J., and Losick, R. (2003). The Spo0A regulon of *Bacillus subtilis*. *Mol. Micro.*, **50**(5), 1683–1701.

Moller, S., Croning, M. D. R., and Apweiler, R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**(7), 646–653.

Moont, G., Gabb, H., and Sternberg, M. J. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*, **35**(3), 364–73.

Moreira, W. and Warnes, G. R. (2003). RPy, a robust Python interface to the R programming language.

Mouncey, N. J. and Kaplan, S. (1998). Redox-dependent gene regulation in *Rhodobacter sphaeroides* 2.4.1T: Effects on dimethyl sulfoxide reductase (*dor*) gene expression. *J. Bact.*, **180**(21), 5612–5618.

Mouncey, N. J., Choudhary, M., and Kaplan, S. (1997). Characterization of genes encoding dimethylsulfoxide reductase of *Rhodobacter sphaeroides* 2.4.1T: an essential metabolic gene function encoded on chromosome II. *J. Bact.*, **179**(24), 7617–7624.

Nakamichi, N., Yamada, H., Aoyama, K., Ohmiya, R., Aiba, H., and Mizuno, T. (2002). His-to-Asp phosphorelay circuitry for regulation of sexual development in *Schizosaccharomyces pombe*. *Biosci. Biotechnol. Biochem.*, **66**(12), 2663–2672.

Nakamichi, N., Yanada, H., Aiba, H., Aoyama, K., Ohmiya, R., and Mizuno, T. (2003). Characterization of the Prr1 response regulator with special reference to sexual development in *Schizosaccharomyces pombe*. *Biosci. Biotechnol. Biochem.*, **67**(3), 547–555.

Nguyen, A. N., Lee, A., Place, W., and Shiozaki, K. (2000). Multistep phosphorelay proteins transmit oxidative stress signals to the fission yeast stress-activated protein kinase. *Molecular Biology of the Cell*, **11**(4), 1169–1181.

Normark, S., Bergstrom, S., Edlund, T., Grundstrom, T., Jaurin, B., Lindberg, F. P., and Olsson, O. (1983). Overlapping genes. *Annu. Rev. Genet.*, **17**, 499–525.

Nyengaard, N. R., Mortensen, K. K., Lassen, S. F., Hershey, J. W., and Sperling-Petersen, H. U. (1991). Tandem translation of *E. coli* initiation factor IF2 beta: purification and characterization in vitro of two active forms. *Biochem. Biophys. Res. Commun.*, **181**(3), 1572–9.

Ohlsen, K. L., Grimsley, J. K., and Hoch, J. A. (1994). Deactivation of the sporulation transcription factor Spo0A by the Spo0E protein phosphatase. *PNAS*, **91**(5), 1756–1760.

Ohta, N. and Newton, A. (2003). The core dimerization domains of histidine kinases contain recognition specificity for the cognate response regulator. *J. Bact.*, **185**(15), 4424–4431.

Oppenheim, D. S. and Yanofsky, C. (1980). Translational coupling during expression of the tryptophan operon of *Escherichia coli*. *Genetics*, **95**, 785–795.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., and Notredame, C. (2004). 3DCoffee: Combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.*, **340**(2), 385–395.

Pallejá, A., Harrington, E. D., and Bork, P. (2008). Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? *BMC Genomics*, **9**, 335.

Pareek, A., Singh, A., Kumar, M., Kushwaha, H. R., Lynn, A. M., and Singla-Pareek, S. L. (2006). Whole-genome analysis of *Oryza sativa* reveals similar architecture of two-component signaling machinery with *Arabidopsis*. *Plant Physiol.*, **142**, 380–397.

Parkinson, J. S. and Kofoid, E. C. (1992). Communication modules in bacterial signalling proteins. *Ann. Rev. Genet.*, **26**, 71–112.

Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.*, **271**(4), 511–523.

Pernestig, A. K., Melefors, O., and Georgellis, D. (2001). Identification of UvrY as the cognate response regulator for the BarA sensor kinase in *Escherichia coli*. *J. Biol. Chem.*, **276**(1), 225–231.

Perraud, A., Weiss, V., and Gross, R. (1999). Signalling pathways in two-component phosphorelay systems. *Trends Microbiol.*, **7**(3), 115–120.

Piggot, P. J. and Hilbert, D. W. (2004). Sporulation of *Bacillus subtilis*. *Curr. Opin. Microbiol.*, **7**(6), 579–586.

Poggio, S., Abreu-Goodger, C., Fabela, S., Osorio, A., Dreyfus, G., Vinuesa, P., and Camarena, L. (2007). A complete set of flagellar genes acquired by horizontal transfer coexists with the endogenous flagellar system in *Rhodobacter sphaeroides*. *J. Bact.*, **189**(8), 3208–3216.

Porter, S. L. and Armitage, J. P. (2004). Chemotaxis in *Rhodobacter sphaeroides* requires an atypical histidine protein kinase. *J. Biol. Chem.*, **279**(52), 54573–54580.

Porter, S. L., Warren, A. V., Martin, A. C., and Armitage, J. P. (2002). The third chemotaxis locus of *Rhodobacter sphaeroides* is essential for chemotaxis. *Mol. Micro.*, **46**(4), 1081–1094.

Porter, S. L., Wadhams, G. H., Martin, A. C., Byles, E. D., Lancaster, D. E., and Armitage, J. P. (2006). The CheYs of *Rhodobacter sphaeroides*. *J. Biol. Chem.*, **281**(43), 32694–32704.

Porter, S. W. and West, A. H. (2005). A common docking site for response regulators on the yeast phosphorelay protein YPD1. *Biochim. Biophys. Acta., Proteins & Proteomics*, **1748**(2), 138–145.

Porter, S. W., Xu, Q., and West, A. H. (2003). Ssk1p response regulator binding surface on histidine- containing phosphotransfer protein ypd1p. *Eukaryotic Cell*, **2**(1), 27–33.

Posas, F., Wurgler-Murphy, S. M., Maeda, T., Witten, E. A., Thai, T. C., and Saito, H. (1996). Yeast HOG1 MAP kinase cascade is regulated by a multistep phosphorelay mechanism in the SLN1-YPD1-SSK1 "two-component" osmosensor. *Cell*, **86**(6), 865–875.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rabin, R. S. and Stewart, V. (1992). Either of two functionally redundant sensor proteins, NarX and NarQ, is sufficient for nitrate regulation in *Escherichia coli* K-12. *PNAS*, **89**, 8419–8423.

Rabin, R. S. and Stewart, V. (1993). Dual response regulators (NarL and NarP) interact with dual sensors (NarX and NarQ) to control nitrate and nitrite-regulated gene expression in *Escherichia coli* K-12. *J. Bact.*, **175**(11), 3259–3268.

Rasmussen, A. A., Wegener-Feldbrügge, S., Porter, S. L., Armitage, J. P., and Søgaard-Andersen, L. (2006). Four signalling domains in the hybrid histidine protein kinase RodK of *Myxococcus xanthus* are required for activity. *Mol. Micro.*, **60**(2), 525–534.

Ren, C., Beatson, S. A., Parkhill, J., and Pallen, M. J. (2005). The Flag-2 locus, an ancestral gene cluster, is potentially associated with a novel flagellar system from *Escherichia coli*. *J. Bact.*, **187**(4), 1430–1440.

Robinson, V. L., Buckler, D. R., and Stock, A. M. (2000). A tale of two components: a novel kinase and a regulatory switch. *Nat. Struct. Biol.*, **7**(8), 626–633.

Rogova, V. V., Bernharda, F., Löhra, F., and Dötsch, V. (2004). Solution structure of the *Escherichia coli* YojN histidine-phosphotransferase domain and its interaction with cognate phosphoryl receiver domains. *J. Mol. Biol.*, **343**(4), 1035–1048.

Rogozin, I. B., Spiridonov, A. N., Sorokin, A. V., Wolf, Y. I., Jordan, I. K., Tatusov, R. L., and Koonin, E. V. (2002). Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet.*, **18**(5), 228–232.

Rowland, S., Burkholder, W., Cunningham, K., Maciejewski, M., Grossman, A., and King, G. (2004). Structure and mechanism of action of Sda, an inhibitor of the histidine kinases that regulate initiation of sporulation in *Bacillus subtilis*. *Mol. Cell.*, **13**(5), 689–701.

Sanatinia, H., Kofoid, E. C., Morrison, T. B., and Parkinson, J. S. (1995). The smaller of two overlapping *cheA* gene products is not essential for chemotaxis in *Escherichia coli*. *J. Bact.*, **177**(10), 2713–20.

Schauder, S. and Bassler, B. L. (2001). The languages of bacteria. *Genes Dev.*, **15**(12), 1468–1480.

Schmitt, R. (2002). *Sinorhizobial* chemotaxis: a departure from the enterobacterial paradigm. *Microbiol.*, **148**(3), 627–631.

Schneider-Poetsch, H. A., Braun, B., Marx, S., and Schaumburg, A. (1991). Phytochromes and bacterial sensor proteins are related by structural and functional homologies. hypothesis on phytochrome-mediated signal-transduction. *FEBS Lett.*, **281**, 245–9.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. *PNAS*, **95**(11), 5857–5864.

Seligmann, H. and Pollock, D. D. (2004). The ambush hypothesis: Hidden stop codons prevent off-frame gene reading. *DNA and Cell Biology*, **23**(10), 701–705.

Shah, D. S. H., Porter, S. L., Harris, D. C., Wadhams, G. H., Hamblin, P. A., and Armitage, J. P. (2000). Identification of a fourth cheY gene in *Rhodobacter sphaeroides* and interspecies interaction within the bacterial chemotaxis signal transduction pathway. *Mol. Micro.*, **35**(1), 101–112.

Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois, USA.

Shapiro, L. and Losick, R. (1997). Protein localization and cell fate in bacteria. *Science*, **276**, 712–718.

Shi, X., Wegener-Feldbrügge, S., Huntley, S., Hamann, N., Hedderich, R., and Søgaard-Andersen, L. (2008). Bioinformatics and experimental analysis of proteins of two-component systems in *Myxococcus xanthus*. *J. Bact.*, **190**(2), 613–624.

Shine, J. and Dalgarno, L. (1975). Determinant of cistron specificity in bacterial ribosomes. *Nature*, **254**(5495), 34–38.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: Visualizing classifier performance in R. *Bioinformatics*, **21**(20), 3940–3941.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, **213**(4), 859–883.

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

Skerker, J. M., Perchuk, B. S., Siryaporn, A., Lubin, E. A., Ashenberg, O., Goulian, M., and Laub, M. T. (2008). Rewiring the specificity of two-component signal transduction systems. *Cell*, **133**(6), 1043–54.

Smith, R. A. and Parkinson, J. S. (1980). Overlapping genes at the cheA locus of *Escherichia coli*. *PNAS*, **77**(9), 5370–4.

Snel, B., Bork, P., and Huynen, M. (2000). Genome evolution: gene fusion versus gene fission. *Trends Genet.*, **16**, 9–11.

Solà, M., Gomis-Rüth, F. X., Serrano, L., González, A., and Coll, M. (1999). Three-dimensional crystal structure of the transcription factor PhoB receiver domain. *J. Mol. Biol.*, **285**(2), 675–687.

Springer, M. S., Goy, M. F., and Adler, J. (1977). Sensory transduction in *Escherichia coli*: a requirement for methionine in sensory adaptation. *PNAS*, **74**(1), 183–7.

Stephenson, K. and Hoch, J. A. (2002). Evolution of signalling in the sporulation phosphorelay. *Mol. Micro.*, **46**(2), 297–304.

Stock, A. M., Mottonen, J. M., Stock, J. B., and Schutt, C. E. (1989). Three dimensional structure of CheY, the response regulator of bacterial chemotaxis. *Nature*, **337**, 745–749.

Stock, A. M., Robinson, V. L., and Goudreau, P. N. (2000). Two-component signal transduction. *Annu. Rev. Biochem.*, **69**, 183–215.

Stock, J. B. and Koshland, D. E. (1978). A protein methylesterase involved in bacterial sensing. *PNAS*, **75**(8), 3659–3663.

Stowe-Evans, E. L. and Kehoe, D. M. (2004). Signal transduction during light-quality acclimation in cyanobacteria: a model system for understanding phytochrome-response pathways in prokaryotes. *Photochem. Photobiol. Sci.*, **3**, 495–502.

Szurmant, H. and Ordal, G. W. (2004). Diversity in chemotaxis mechanisms among the bacteria and archaea. *Microbiol. Mol. Biol. Rev.*, **68**(2), 301–319.

Takeda, S., Fujisawa, Y., Matsubara, M., Aiba, H., and Mizuno, T. (2001). A novel feature of the multistep phosphorelay in *Escherichia coli*: a revised model of the RcsC → YojN → RcsB signalling pathway implicated in capsular synthesis and swarming behaviour. *Mol. Micro.*, **40**(2), 440–450.

Tan, H., Janiak-Spens, F., and West, A. H. (2007). Functional characterization of the phosphorelay protein Mpr1p from *Schizosaccharomyces pombe*. *FEMS Yeast Res.*, **7**(6), 912–921.

Tanabe, H., Masuda, T., Yamasaki, K., Katoh, A., Yoshioka, S., and Utsumi, R. (1998). Molecular interaction between proteins involved in EvgAS signal transduction of *Escherichia coli*. *Biosci. Biotechnol. Biochem.*, **62**(1), 78–82.

Tanaka, S. and Scheraga, H. A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, **9**(6), 945–950.

Tatusov, R., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**(41).

Taylor, W. R. and Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.*, **208**(1), 1–22.

Thomason, P. and Kay, R. (2000). Eukaryotic signal transduction via histidine-aspartate phosphorelay. *J. Cell Sci.*, **113**, 3141–3150.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Timmen, M., Bassler, B. L., and Jung, K. (2006). AI-1 influences the kinase activity but not the phosphatase activity of LuxN of *Vibrio harveyi*. *J. Biol. Chem.*, **281**(34), 24398–24404.

Tomomori, C., Tanaka, T., Dutta, R., Park, H., Saha, S. K., Zhu, Y., Ishima, R., Liu, D., Tong, K. I., Kurokawa, H., Qian, H., Inouye, M., and Ikura, M. (1999). Solution structure of the homodimeric core domain of *Escherichia coli* histidine kinase EnvZ. *Nat. Struct. Biol.*, **6**(8), 729–734.

Toro-Roman, A., Wu, T., and Stock, A. M. (2005). A common dimerization interface in bacterial response regulators KdpE and TorR. *Protein Science*, **14**, 3077–3088.

Tsuzuki, M., Ishige, K., and Mizun0, T. (1995). Phosphotransfer circuitry of the putative multi-signal transducer, ArcB, of *Escherichia coli*: in vitro studies with mutants. *Mol. Micro.*, **18**(5), 953–962.

Tzeng, Y. and Hoch, J. A. (1997). Molecular recognition in signal transduction: The interaction surfaces of the Spo0F response regulator with its cognate phosphorelay proteins revealed by alanine scanning mutagenesis. *J. Mol. Biol.*, **272**, 200–212.

Uhl, M. and Miller, J. (1994). Autophosphorylation and phosphotransfer in the bordetella pertussis BvgAS signal transduction cascade. *PNAS*, **91**(3), 1163–1167.

Uhl, M. and Miller, J. (1996). Integration of multiple domains in a two-component sensor protein: the *Bordetella pertussis* BvgAS phosphorelay. *EMBO J.*, **15**(5), 1028–1036.

Ulrich, L. E., Koonin, E. V., and Zhulin, I. B. (2005). One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol.*, **13**(2), 52–56.

Urao, T., Yakubov, B., Satoh, R., Yamaguchi-Shinozaki, K., Seki, M., Hirayama, T., and Shinozaki, K. (1999). A transmembrane hybrid-type histidine kinase in *Arabidopsis* functions as an osmosensor. *Plant Cell*, **11**, 1743–1754.

Utsumi, R., Katayama, S., Ikeda, M., Igaki, S., Nakagawa, H., Miwa, A., Taniguchi, M., and Noda, M. (1992). Cloning and sequence analysis of the evgAS genes involved in signal transduction of *Escherichia coli* K-12. *Nucleic Acids Symp Ser.*, **27**, 149–150.

Utsumi, R., Katayama, S., Taniguchi, M., Horie, T., Ikeda, M., Igaki, S., Nakagawa, H., Miwa, A., Tanabe, H., and Noda, M. (1994). Newly identified genes involved in the signal transduction of *Escherichia coli* K-12. *Gene*, **140**(1), 73–77.

van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. *Trends Genet.*, **19**, 479–484.

Varughese, K., Madhusudan, Zhou, X., Whiteley, J., and Hoch, J. (1998). Formation of a novel four-helix bundle and molecular recognition sites by dimerization of a response regulator phosphotransferase. *Molecular Cell*, **2**(4), 485–493.

Volz, K. and Matsumura, P. (1991). Crystal structure of *Escherichia coli* CheY refined at 1.7 resolution. *J. Biol. Chem.*, **266**, 15511–15519.

Wadhams, G. H. and Armitage, J. P. (2004). Making sense of it all: bacterial chemotaxis. *Nature Reviews Molecular Cell Biology*, **5**, 1024–1037.

Wadhams, G. H., Warren, A. V., Martin, A. C., and Armitage, J. P. (2003). Targeting of two signal transduction pathways to different regions of the bacterial cell. *Mol Microbiol*, **50**(3), 763–770.

Walderhaug, M. O., Polarek, J. W., Voelkner, P., Daniel, J. M., Hesse, J. E., Altendorf, K., and Epstein, W. (1992). KdpD and KdpE, proteins that control expression of the kdpABC operon, are members of the two-component sensor-effector class of regulators. *J. Bact.*, **174**(7), 2152–2159.

Wehland, M. and Bernhard, F. (2000). The RcsAB box. Characterization of a new operator essential for the regulation of exopolysaccharide biosynthesis in enteric bacteria. *J. Biol. Chem.*, **275**(10), 7013–7020.

White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. (2007). Features of protein-protein interactions in two-component signaling deduced from genomic libraries. *Methods in Enzymology*, **422**, 75–101.

Whitworth, D. E. and Cock, P. J. A. (2008a). Myxobacterial two-component systems. In D. Whitworth, editor, *Myxobacteria: Multicellularity and Differentiation*, chapter 10. ASM Press, Washington, DC, USA.

Whitworth, D. E. and Cock, P. J. A. (2008b). Two-component systems of the myxobacteria: Structure, diversity and evolutionary relationships. *Microbiol.*, **154**(2), 360–372.

Whitworth, D. E., Millard, A., Hodgson, D. A., and Hawkins, P. F. (2008). Protein-protein interactions between two-component system transmitter and receiver domains of *Myxococcus xanthus*. *Proteomics*, **8**(9), 1839–1842.

Wolfe, A. J. and Stewart, R. C. (1993). The short form of the CheA protein restores kinase activity and chemotactic ability to kinase-deficient mutants. *PNAS*, **90**(4), 1518–22.

Wolfe, A. J., McNamara, B. P., and Stewart, R. C. (1994). The short form of CheA couples chemoreception to CheA phosphorylation. *J. Bact.*, **176**(15), 4483–91.

Wu, J., Ohta, N., and Newton, A. (1998). An essential, multicomponent signal transduction pathway required for cell cycle regulation in *Caulobacter*. *PNAS*, **95**(4), 1443–1448.

Wu, J., Ohta, N., Zhao, J. L., and Newton, A. (1999). A novel bacterial tyrosine kinase essential for cell division and differentiation. *PNAS*, **96**(23), 13068–13073.

Wurgler-Murphy, S. and Saito, H. (1997). Two-component signal transducers and MAPK cascades. *Trends Biochem. Sci.*, **22**, 172–176.

Xu, J., Chiang, H. C., Bjursell, M. K., and Gordon, J. I. (2004). Message from a human gut symbiont: sensitivity is a prerequisite for sharing. *Trends Microbiol.*, **12**(1), 21–28.

Xu, Q. and West, A. (1999). Conservation of structure and function among histidine-containing phosphotransfer (Hpt) domains as revealed by crystal structure YPD1. *J. Mol. Biol.*, **292**(5), 1039–1050.

Xu, Q., Porter, S. W., and West, A. H. (2003). The yeast YPD1/SLN1 complex: Insights into molecular recognition in two-component signaling systems. *Structure*, **11**, 1569–1581.

Yaku, H., Kato, M., Hakoshima, T., Tsuzuki, M., and Mizuno, T. (1997). Interaction between the CheY response regulator and the histidine-containing phosphotransfer (HPt) domain of the ArcB sensory kinase in *Escherichia coli*. *FEBS Letters*, **408**(3), 337–340.

Yamamoto, K., Hirao, K., Oshima, T., Aiba, H., Utsumi, R., and Ishihama, A. (2005). Functional characterization in vitro of all two-component signal transduction systems from *Escherichia coli*. *J. Biol. Chem.*, **280**(2), 1448–1456.

Yonekawa, H., Hayashi, H., and Parkinson, J. S. (1983). Requirement of the *cheB* function for sensory adaptation in *Escherichia coli*. *J. Bact.*, **156**(3), 1228–1235.

Zapf, J., Sen, U., Madhusudan, M., Hoch, J. A., and Varughese, K. I. (2000). A transient interaction between two phosphorelay proteins trapped in a crystal lattice reveals the mechanism of molecular recognition and phosphotransfer in signal transduction. *Structure Fold. Des.*, **8**, 851–862.

Zhang, W. and Shi, L. (2005). Distribution and evolution of multiple-step phosphorelay in prokaryotes: lateral domain recruitment involved in the formation of hybrid-type histidine kinases. *Microbiol.*, **151**, 2159–2173.

Zhang, W., Culley, D. E., Wu, G., and Brockman, F. J. (2006). Two-component signal transduction systems of desulfovibrio vulgaris: Structural and phylogenetic analysis and deduction of putative cognate pairs. *J. Mol. Evol.*, **62**(4), 473–487.

Zhu, Y., Qin, L., Yoshida, T., and Inouye, M. (2000). Phosphatase activity of histidine kinase EnvZ without kinase catalytic domain. *PNAS*, **97**(14), 7808–7813.