Model-based evolutionary analysis: the natural history of phage-shock stress response

Maxime Huvet*†1, Tina Toni*†1, Hui Tan*, Goran Jovanovic‡, Christoph Engl‡, Martin Buck‡ and Michael P.H. Stumpf*‡2

*Centre for Bioinformatics, Division of Molecular Biosciences, Imperial College London, South Kensington, London SW7 2AZ, U.K., ‡Institute of Mathematical Sciences, Imperial College London, South Kensington, London SW7 2PG, U.K., and †Division of Biology, Imperial College London, South Kensington, London SW7 2AZ, U.K.

Abstract
The evolution of proteins is inseparably linked to their function. Because most biological processes involve a number of different proteins, it may become impossible to study the evolutionary properties of proteins in isolation. In the present article, we show how simple mechanistic models of biological processes can complement conventional comparative analyses of biological traits. We use the specific example of the phage-shock stress response, which has been well characterized in *Escherichia coli*, to elucidate patterns of gene sharing and sequence conservation across bacterial species.

Introduction
Differential reproductive success among individuals in a species underlies all evolutionary processes [1]. The expected number of viable offspring can be directly related to Darwinian or Malthusian fitness, and is one of the central quantities in population genetics. The notion of chance versus deterministic effects, in particular for natural selection, have been explored in great detail. These analyses have, however, focused generally on the rather special cases of one locus (or a small number of loci) and an infinite number of loci respectively. For both extremes, beautiful mathematical frameworks exist which have, even at the single-gene level, vastly increased our understanding of evolution and biology in general [2].

Real biological systems are, however, mesoscopic; this means that they contain only a finite number of constituent parts (genes and their protein products); furthermore, the interactions among these molecules exert considerable influence over the organism’s form, function and therefore viability and fitness [3]. The mere number of these interactions probably barely begins to describe the complexity of an organism: it is the controlled manner and dynamics in which these interactions are invoked and terminated under different environmental, physiological, (and in multicellular organisms) developmental conditions, which is important for survival and successful reproduction [4].

Recent years have seen a rise in empirical evolutionary analyses of biological networks. In particular, PINs (protein interaction networks) and metabolic networks have been scrutinized for evidence that properties of the network (or of the nodes in the network) affect Darwinian fitness. These attempts at beginning to understand evolutionary processes at the system level have recently been summarized for PINs and metabolic networks. The concept of fitness used is not straightforwardly related to the Darwinian fitness used in population genetics, but rather focused on essentiality (i.e. the inability of an organism to survive if a gene is knocked out or otherwise severely impaired) and related properties [5]. Moreover, these studies took a genome-wide perspective. Such evolutionary studies, almost necessarily, become data-mining exercises where different properties of the network or the proteins are cross-correlated in order to find relationships that could relate to fitness. This typically involves a search for aspects of a protein (e.g. its function, network characteristics or expression level), which might influence or correlate with some measure for the evolutionary rate of a protein or its corresponding gene. This is a statistically and biologically challenging problem, as it is notoriously difficult [6] to disentangle the interplay among all of these factors (e.g. expression can be closely related to gene function).

In the present article, we take a different approach where we pay close attention to only a small set of proteins, which are involved in modulating a biologically important stress response in the bacterium *Escherichia coli*: the phage-shock stress-response machinery [7]. The PspS (phage-shock proteins) contribute crucially to repair damage to inner cell membrane and the consequent PMF (protonmotive force) dissipation, and are highly conserved among enterobacteria [8]. Below we briefly introduce the Psp components before highlighting features characterizing the evolutionary history of this important signalling and stress-response system. Our analysis differs from previous approaches by taking an explicitly mechanistic hypothesis-driven perspective: all observations are related to a mathematical model of the Psp
response system. We will conclude by reviewing the advantages of this model-based approach and by outlining how it can be generalized to other molecular networks.

The Psp stress-response machinery
The main components of the Psp response in *E. coli* are the seven *psp* genes, labelled A–G [9,10]; furthermore, *arcA* and *arcB* are crucially involved in mediating signals

---

**Figure 1 | Possible Petri net representation of the Psp stress response**

Only the network structure is shown, i.e. a complete Petri net is not specified. In this model, we allow for reaction arcs, inhibitory arcs and test-arcs that control or enable reactions/transitions. The pivotal interaction in this network appears to be the transition that releases *pspF* from the bound state to *pspA* (in green towards the bottom of the Figure) which initiates the regulatory response. But the importance of activating *pspB* and *pspC* and the feedback provided by the *arcA/arcB* system is already evident in this system. Such models can be used to collate available functional and mechanistic information for further mathematical analysis.
that trigger the response. The genes coding for PspA, PspB, PspC, PspD and PspE are in a single operon adjacent to that for *pspF*, which sits in its own single gene operon. *pspF* is the transcription regulator, which activates the transcription of *pspG* and *pspA–pspE*, dependent on the σ^54 factors. *pspF* autogenously regulates its own transcription [11], which is constant and stress-independent. In our current model, under normal conditions, PspF is bound to PspA, which in turn is attached to a PspB/C membrane-bound complex. Upon damage to the inner membrane, PspB/C appears to undergo structural changes, which leads to the release of PspF from PspA [12]. This in turn triggers co-regulated expression of the genes in the *pspABCDE* operon and *pspG*. Of these, *pspA* continues to play an important role as the major Psp effector of membrane repair to recover from stress to the membrane and conserve PMF [13,14]. The quantitative, and at some level presumably also qualitative, nature of the stress response is influenced further by *arcA* and *arcB*: knockout mutants of these genes show drastically reduced stress responses, but the mechanism by which they and *pspE* and *pspG* enter the response is only beginning to emerge [9,15].

Figure 1 shows a graphical representation of the mechanism by which the Psp response is elicited, affects adaptation to inner membrane damage and ultimately repairs such damage. We describe this in terms of an SPN (stochastic Petri net) [15,16]; this allows us both to model (qualitatively as well as quantitatively) the Psp response, and to depict mechanistic hypotheses diagrammatically. This dual use can greatly facilitate dialogue between ‘wet’ and ‘dry’ scientists in systems biology.

In an SPN, circles denote molecular species, environmental conditions or physiological status and boxes denote biochemical reactions or transitions between different states. Edges with an arrow from a circle to a box indicate the educts of a reaction; edges pointing from a box to a circle denote the flux of products; bi-directional arrows, so-called test-arcs, specify conditions under which a reaction is possible, i.e. a reaction is only enabled if the species from which the test-arc originates is instantiated or present.

**The evolution of the Psp stress-response system**

Of the myriad questions we may ask in this context, we will address only two related issues: first, we are interested in the extent to which the constituent genes of the Psp system are shared across bacterial, in particular enterobacterial, species [8]; secondly, we shall consider how conserved the corresponding protein products are at the amino acid sequence level.

The advantage of the mechanistic model for the Psp stress response in Figure 1 is that we can begin to study the interplay between the functionality of the response mechanism and the evolutionary history of its constituent components. This, of course, assumes independence of different signalling, metabolic, etc. interaction networks and ignores that many proteins will be involved in multiple response networks. This will, however, always be a trade-off that we have to make in evolutionary analyses of biological systems.

**Shared components of the Psp response system**

If more than one gene contributes to a given phenotype, then we expect to see these genes to be co-inherited more frequently than would be expected to occur by chance (according to some suitable definition of chance), but genes are not physically independent objects. The operon structure of bacterial genomes facilitates this co-inheritance in many bacteria. Indeed we find that, in *E. coli*, the *pspA–pspE* genes are in a single operon (29.4 min on the chromosomal map), *pspF* is in its own, directly adjacent to the *pspABCDE* operon, and *pspG* occurs on its own in a different locus (91.8 min) far away from the other two transcription units. The other involved genes positioned on two separate loci, *arcA* (99.9 min) and *arcB* (72.2 min), are also known to contribute to other important biological processes, and are known to be shared widely among bacteria.
Figure 3 | Evolutionary divergence of the proteins in the Psp system

For the nine genes known to play important parts in the Psp response in *E. coli*, levels of sequence similarity (compared to *E. coli* K12) are plotted next to the assumed species tree (we used 23S ribosomal RNA to obtain an estimate for the species tree). Horizontal axes at the bottom of each panel represent percentage sequence identity.

Figure 2 shows how frequently each combination of genes occurs across a diverse set of 129 bacterial species (out of a total 696 species, only those with at least two clearly identifiable homologues of genes in the Psp system are included). What we do observe almost reflects an all-or-nothing scenario: either most of the Psp system is present or hardly anything. A host of patterns is discernable: *pspA* and *pspF* tend to occur together, as is expected from the mechanistic hypothesis in Figure 1; *pspD*, *pspE*, and *pspG* are absent most frequently, but, notably, the set of species which have only two readily identified orthologues of the genes in *E. coli* tend to contain orthologues of *pspC* and *pspE*. The latter fact may be partly explained by differences in the genome sequence of *pspE* and its (inferred) orthologues, which will be touched upon below.

Figure 2 does not, of course, take into account the phylogenetic context, which is intricately linked to the patterns in which genes and biological characteristics are shared
across different organisms. Such evolutionary relationships have been widely exploited in bioinformatics and functional genomics [17]. In a systems biology context, where function is mediated by groups of genes, this can give rise to almost bewilderingly complex processes as we illustrate in the next section.

**Sequence conservation**

Figure 3 shows the relative conservation of the amino acid sequences of the nine principal known genes in the Psp system in relation to the species tree, which was taken to correspond to the 23S ribosomal phylogeny. The horizontal bars to the right of the tree indicate percentage sequence identity compared with the experimental K12 strain of *E. coli*. We observe that sequence identity decays as expected with evolutionary distance from *E. coli*. Furthermore, it appears that an involvement in shared pivotal interactions, in particular that of *pspA* with *pspF* influences correlations in these sequence identity characteristics more profoundly than does membership of the same operon. The similarity in the relative patterns of sequence similarity to the experimental *E. coli* strain in the *pspA* and *pspF* genes is particularly noteworthy, although not unexpected from the patterns already seen in Figure 2 and predictable from the model described in Figure 1.

This pattern is, however, not uniform across the five genes, which must reflect different roles that different genes have in the different species. *pspC* and *pspE*, for example, are shared more widely among the 129 species than the other three genes in the same operon, but their patterns of divergence from *E. coli* K12 are very different indeed. This reinforces the notion that evolutionary analyses and the insights that can be gained from them depend on the comparisons being made.

**Discussion**

Our results clearly contradict the long-held and admittedly attractive notion of a cell or an organism as simply a bag of genes [18], where the evolution of systems has generally been studied, ignoring interaction between genes or proteins and other epistatic interactions. We find, however, that having complete and fully functional sets of genes appears to be vital for the organism’s functioning and thus ultimately affects the organism’s Darwinian fitness. By the same token, there appears to be little point in retaining a subset of genes which does not fulfil the required function [19]. A gene may be lost, particularly if a different gene can substitute for it. Whenever we observe that a gene that is vitally important for fully functional Psp response in *E. coli* is lacking in a given organism, a range of scenarios are plausible: (i) the response may involve different genes which substitute for the missing gene’s function; (ii) there may be substantial differences in the molecular machinery involved in the Psp response; or (iii) bacteria living under different conditions may not even require the Psp response in the same way that *E. coli* and many related organisms appear to do.

The mathematical model corresponding to Figure 1 can be used to explore these possibilities in silico, and such an approach can serve as an important hypotheses generation tool. Once a functioning model is at hand which explains the observations in extensively studied *E. coli*, we can generate in silico knockouts and study their properties. If these artificial constructs still have characteristics of a functional Psp response, we can test this experimentally. When a functional response is absent from the model, then this can either mean that the response machinery in species lacking a particular gene has recruited other (potentially diverged) genes, or that the required response is sufficiently different from that of *E. coli* under standard experimental conditions.

Without further experimental work (or improved in silico methodologies [20–22]), we are, unfortunately, unable to determine whether genes missing in a given organism have been substituted for by other genes, or whether the response to the Psp-equivalent stress is not as important as in *E. coli* or is mediated by an entirely different system.

What should have become clear, however, is the value of considering such evolutionary data in the light of mechanistic models that are amenable to mathematical analysis. It is notoriously easy to find patterns in sufficiently large and complex datasets, and these mathematical models can force us to evaluate and re-evaluate the nature of such patterns in the light of functioning biological systems.

**Funding**

M.H. gratefully acknowledges financial support through a Wellcome Trust VIP award. T.T. is funded through a Medical Research Council priority studentship. Research activities in the M.B. and M.P.H.S. groups are funded through the Biotechnology and Biological Sciences Research Council and the Wellcome Trust.

**References**


© The Authors Journal compilation © 2009 Biochemical Society

Received 21 January 2009
doi:10.1042/BST0370762