Highly specific protein-protein interactions, evolution and negative design

Richard P. Sear

Department of Physics, University of Surrey,
Guildford, Surrey GU2 7XH, United Kingdom‡ and
The Isaac Newton Institute for Mathematical Sciences,
University of Cambridge, 20 Clarkson Road, Cambridge CB3 0EH, United Kingdom§
email: r.sear@surrey.ac.uk

Abstract. We consider highly specific protein-protein interactions in proteomes of simple model proteins. We are inspired by the work of Zarrinpar *et al.* [2003 *Nature* **426** 676]. They took a binding domain in a signalling pathway in yeast and replaced it with domains of the same class but from different organisms. They found that the probability of a protein binding to a protein from the proteome of different organism is rather high, around one half. We calculate the probability of a model protein from one proteome binding to the protein of a different proteome. These proteomes are obtained by sampling the space of functional proteomes uniformly. In agreement with Zarrinpar *et al.* we find that the probability of a protein binding a protein from another proteome is rather high, of order one tenth. Our results, together with those of Zarrinpar *et al.*, suggest that designing, say, a peptide to block or reconstitute a single signalling pathway, without affecting any other pathways, requires knowledge of all the partners of the class of binding domains the peptide is designed to mimic. This knowledge is required to use negative design to explicitly design out interactions of the peptide with proteins other than its target. We also found that patches that are required to bind with high specificity evolve more slowly than those that are required only to not bind to any other patch. This is consistent with some analysis of sequence data for proteins engaged in highly specific interactions.

Keywords: Specific binding, protein evolution, negative design.

‡ Permanent address
§ Address until end of June 2004

1. Introduction

Inside cells there is a constant flux of signals, for example a change in the cell's environment may trigger a signal that is transmitted across the cytoplasm to the nucleus to turn on transcription of a specific gene. Accurate transmission of a signal requires specific protein-protein interactions. A protein must bind strongly to one other protein to trigger the appropriate response, and not bind to any of a thousand other proteins, to avoid triggering inappropriate responses. We would like to understand how this high specificity binding has been achieved by evolution. Recent work by Lim and coworkers [1] has probed these interactions by replacing a protein domain involved in the osmoresistance signalling pathway in yeast by domains of the same class of protein domain but from other species. Inspired by this work, we consider partial proteomes, i.e., sets of protein domains of the same class, of simple model proteins [2, 3] and look at how proteins interact both with other proteins of the same proteome and with proteins from another proteome. We generate the proteomes by imposing the requirement that the model proteins bind with high specificity and then we uniformly sample the partial proteomes that satisfy this requirement. This is a crude model of what evolution might produce. This is in contrast to, for example, Havranek and Harbury [4], Shifman and Mayo [5] and Kortemme *et al.* [6], who all explicitly design proteins. We interpret our results in terms of the selection of proteins not only for the strength of their binding to their partner but also for the weakness of their binding to other proteins. This form of selection has parallels to what scientists who design proteins [4, 7, 8] call negative design: the design of proteins to *not* do something deleterious. Understanding how high-specificity binding is achieved is essential for developing drugs that have few side effects, i.e., whose action is highly specific. To achieve this the peptide would need to block or reconstitute one and only one specific signalling pathway without affecting other pathways. See Lazar *et al.* [9] for a recent review of drug design.

The outline of the paper is as follows. In the next section we set down what we require of the matrix of interactions between the members of a set of proteins which interact in pairs with high specificity. This requirement defines a functional set of interactions, or interactome. In the third section we define a simple model protein, similar but not identical to that of earlier work [2]. The fourth section is a compact description of how we sample the model proteomes, and the fifth section describes our results and compares them to the experimental results of Lim and coworkers [1]. We find qualitative agreement. In this section we also compare our results for the rate of neutral evolution with analysis of sequence data [10–17]. Again the results for our model agree with the data on real proteins, although some of this data is controversial. We find that the constraint on a binding site to participate in high specificity binding slows its rate of neutral evolution. The final section is a conclusion.

2. Model of the cytoplasm of a cell

Inside a prokaryote cell or a compartment of a eukaryote cell there are many proteins that have to interact with high specificity [18]. Zarrinpar *et al.* [1] considered high specificity binding between the SH3 domains and their partners in yeast. The SH3 domain is one of many types of domains whose function is to bind with high specificity [19, 20]. Yeast has 27 different SH3 domains, each of which binds to a different one of a class of proteins. Motivated by this relatively small number of proteins, of order 10 not the many thousands present *in vivo*, we consider model proteomes of N proteins, where N is a few tens. Essentially, we assume that the SH3 domains and the partners do not bind to the other proteins in the cell. As we are interested solely in the specific interactions between proteins, we neglect the parts of the protein other than the patch on their surface responsible for the interaction. Our model proteome is simply N surface patches interacting in $N/2$ pairs.

This is a model of a small subset of all the proteins inside a cell. Now, the full complement of proteins of the cell is called the proteome. So the mixture of N patches is only a partial proteome, but, the sake of brevity we will call it a proteome here. We will use the same model for all proteins. This is rather different from the typical situation for real proteins where the binding is highly asymmetric: an SH3 domain binds to a characteristic partner domain, but there is no possibility of one SH3 domain binding to another SH3 domain. However, this does not change the essential feature of high specificity binding, which is to pick out one protein from many, and it does mean we only need one type of model protein.

In a partial proteome of N binding patches, each odd numbered patch i is required to bind to the $(i+1)$th patch, with a large equilibrium constant K_b. The ith (i odd) and $(i+1)$th proteins must bind only to each other, it must be specific, and so in addition to binding to each other we require that the equilibrium constant between any odd numbered patch i and any patch other than patch $i+1$ be less than $K_s \ll K_b$. Similarly, the equilibrium constant between patch $i+1$ and any protein other than protein i must also be less than K_s. Thus if we denote the binding constant between proteins i and j by K_{ij} we require

$$\begin{aligned} K_{ij} &> K_b \quad ij = 12, 21, 34, 43, \ldots, (N-1)N, N(N-1) \\ K_{ij} &< K_s \quad \text{otherwise.} \end{aligned} \tag{1}$$

In the language of genomics, this defines a functional partial interactome. Zarrinpar *et al.* [1] measured *in vitro* dissociation constants, $K_d = K^{-1}$, K being the equilibrium constant. The wild-type dissociation constant for the interaction between the osmoresistance SH3 domain and its partner is $K_d = 1.3\mu$M. Thus, we set $K_b = 10^6\text{nm}^3$. A dissociation constant of 1μM is approximately equal to an equilibrium constant of 10^6nm^3. Zarrinpar *et al.* found that osmoresistance was restored for dissociation constants of order $K_d = 40\mu$M or less. Thus, we set the maximum equilibrium constant between non-binding patches to be well below this:

$K_s = 10^4 \mathrm{nm}^3$.

Having defined a functional partial interactome we need a model protein in order to calculate the equilibrium constants. We also need some method of sampling functional proteomes of these model proteins. Of course, many different proteomes will satisfy equation (1). We will sample uniformly the proteomes that are functional according to equation (1), and assume that the proteomes produced by evolution are a relatively weakly biased set of the possible functional proteomes. This would be the case if there is extensive neutral evolution. Neutral evolution, as proposed by Kimura [21], and by King and Jukes [22] is the change of sequences with time without change in fitness; see for example Bastolla *et al.* [23] for recent work.

3. Model protein

Our model protein is similar to that used in our earlier work [2, 3] but there patches were specified by sequences of only hydrophobic and hydrophilic, polar, monomers. Here each patch has interactions that depend on a sequence $\mathbf{s}$ of n_M monomers, each of which is one of four types: hydrophobic (H), hydrophilic (P), positive (+) and negative (−). We use the four-letter alphabet of Backofen *et al.* [24]. Table 1 lists the interaction energies of all combinations of these four types of monomers, in units of $-\epsilon$. ϵ is positive and sets the energy scale of our interactions; we will work with $\epsilon = 2k_BT$. k_B and T are Boltzmann's constant and the absolute temperature respectively. If we write the sequences of these monomers of two interacting patches as the vectors $\mathbf{s}$ and $\mathbf{s}'$, then the interaction energy is $v(\mathbf{s}, \mathbf{s}')$, and their equilibrium constant is

$$K(\mathbf{s}, \mathbf{s}') = \left(\frac{1}{2}\right)^{\delta_{ij}} a \exp\left[-v(\mathbf{s}, \mathbf{s}')/k_BT\right], \tag{2}$$

where a is a measure of the phase volume accessible to two interacting patches (it would also absorb the effect of any non-specific binding), in units of nm^3. We take $a = (1/6)\mathrm{nm}^3$, as in earlier work [2, 3]. The prefactor is a symmetry factor that halves the equilibrium constant between identical patches. Note that although the monomers determine the interactions in our model and the residues determine the interactions in a real protein, there is no one-to-one relationship between one of our monomers and an amino acid.

When two patches with sequences $\mathbf{s}$ and $\mathbf{s}'$ interact every monomer of the sequence $\mathbf{s}$ interacts with only one monomer of the sequence $\mathbf{s}'$. As an example of the interaction between two patches with short sequences, consider the two patches of $n_M = 5$ monomers: $\mathbf{s}$ = (H, +,P,P−) and $\mathbf{s}'$ = (−,P,H, −,H). The interaction energy for this pair is then obtained by considering the interactions between the first monomer of $\mathbf{s}$ and the last monomer of $\mathbf{s}'$, the second monomer of $\mathbf{s}$ and the last-but-one monomer of $\mathbf{s}'$ and so on. For the sequences above the five interactions are, in order, HH, +−, PH, PP and −−. Thus the total interaction

	H	P	+	−
H	4	0	0	0
P	0	0	0	0
+	0	0	−1	1
−	0	0	1	−1

Table 1. The interaction free energies between the four different monomer types, in units of $-\epsilon$, for the 4-letter alphabet of Backofen *et al.* (1999).

energy is -4ϵ: -5ϵ from the HH and $+-$ interactions and ϵ from the $--$ interaction. Note that the first monomer of one sequence interacts with the last, not the first, monomer of the other sequence, the second with the last-but-one and so on. This is done as if each element i interacted with the element i on the other patch, then when a patch interacted with another patch with the same sequence any H monomers would always interact with themselves making the interaction between two identical patches always strongly attractive. Thus patches would always stick to themselves unless they had only a couple of H monomers. This would severely restrict the possible sequences. Also, our model protein patch has a primary structure, the sequence $\mathbf{s}$, and the interactions drive the binding together of specific pairs, quaternary structure, but it lacks secondary and tertiary structure.

There has been extensive previous work on simple models of proteins where only a single isolated protein is considered, for example work on protein folding [25, 26]. For example the neutral evolution of single proteins has been considered [23], i.e., evolution via mutations that still allow the protein to fold. There has also been some work on protein-ligand binding [27]. Proteins typically operate inside the crowded environment inside cells [18] surrounded by thousands of other proteins, and many proteins have to interact with other proteins in order to function. Thus there is a clear need to study the function and evolution of proteins interactions.

4. Methods

We will need to sample many proteomes that are functional according to equation (1). To do this we first require a functional proteome of N patches. This is generated as follows. Each of the $N/2$ pairs of patches is obtained by starting with a pair of patches all of whose monomers are polar, and then making enough of the corresponding pairs of elements either both hydrophobic, or positive and negative, to make their equilibrium constant larger than K_b. Having obtained $N/2$ pairs, we check the proteome for patches which violate equation (1) by sticking ($K_{ij} > K_s$) to patches other than their partner. Any pair of patches where either patch violates equation (1) is then removed and replaced with another pair of patches

selected as above. Having done this the proteome is again checked for violations of equation (1). The two steps of replacing pairs in which a patch sticks to another and rechecking for violations of equation (1) are then repeated a number of times until either a viable proteome is obtained or it is concluded that a viable proteome either does not exist at this value of N or that it is prohibitively difficult to locate such a proteome.

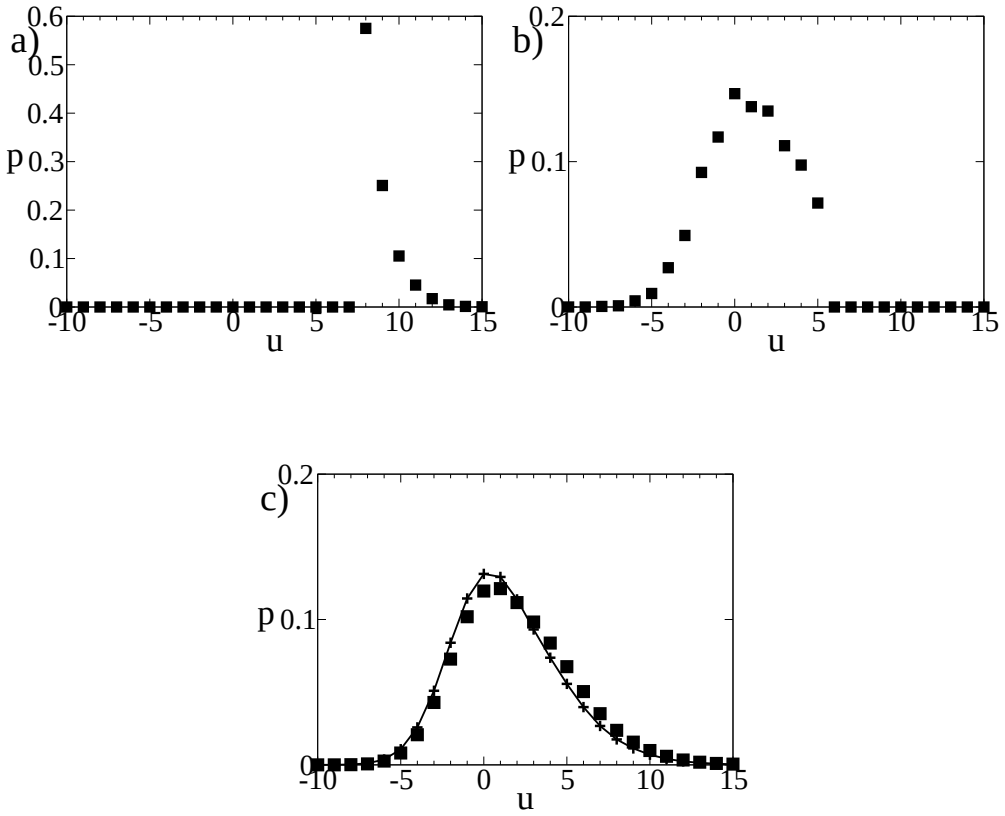
Having obtained a viable proteome, averages are obtained by simply mutating monomers at random, and accepting the mutation if the mutated proteome still satisfies equation (1) and rejecting it and returning to the original proteome if the mutated proteome violates equation (1). This is a common simple model of neutral evolution used frequently to model the neutral evolution of single proteins [23]. It generates a simple random walk in the space of the proteome of the N proteins. This is similar to protein space [28] but of course it is the space of a set of proteins, not that of a single protein. It is a proteome space not a protein space. Also, after we generated a new viable proteome and before we start accumulating averages, we evolve the proteome for a large number of attempted mutations. This procedure removes any biases in the sampling due to the way we generate our starting proteomes. To evolve the proteome we need to select the probabilities $p_{\alpha\beta}$ of a mutation converting a monomer of type α into one of a different type β. We take all the probabilities of type $\alpha = H, P, +, -$ mutating to a monomer of a different type β, $p_{\alpha\beta}$ to satisfy $p_{\alpha\beta}/p_{\beta\alpha} = 1$ except that $p_{H\beta}/p_{\beta H} = 1/2$. In the absence of selection for binding this makes H monomers twice as common as the other three types.

For each proteome we looked at the interaction both between binding pairs and between pairs of patches that do not bind to each other. In each case we determined the probability distribution function $p(u)$ of the interaction energy between the patches. It is convenient to use the reduced interaction energy $u = v(\mathbf{s}, \mathbf{s}')/(-\epsilon)$ — u is positive for attractive interactions. Also, we took many pairs of proteomes and for each pair we calculated the interactions of a patch from one proteome with the patches of the other proteome. From this we were able to calculate the probability distribution function for the interactions $p(u)$ of a patch from one proteome with the patches of another.

5. Results and discussion

Lim and coworkers [1] studied a cell function called osmoresistance, which relies on binding between the specific SH3 domain belonging to a protein called Sho1 and a domain on a protein called Pbs2. They studied this in *Saccharomyces cerevisiae* (yeast), in which 27 distinct SH3 domains have been identified. They took 12 SH3 domains from organisms *other* than *S. cerevisiae* as well as the other 26 SH3 domains of *S. cerevisiae*, and used each domain in turn to replace the SH3 domain that in *S. cerevisiae* binds to Pbs2 to trigger osmoresistance. Six of the 12 domains from other organisms bound sufficiently strongly to

Figure 1. The probability distribution function, p, for u the interaction energy of a pair of patches in units of $-\epsilon$. a) is for a protein and its binding partner, and b) is for a protein and any other protein that is not its binding partner; both are for pairs of proteins from the same proteome. c) is for a protein introduced into another proteome; the squares are for any two patches from different proteomes, and the curve with crosses is for pairs of patches with random sequences. The number of monomers $n_M = 16$, the number of patches $N = 32$, and the interaction energy scale $\epsilon = 2k_BT$. The binding strength is $K_b = 10^6\text{nm}^3$, and $K_s = 10^4\text{nm}^3$.



the Pbs2 domain to allow at least partial osmoresistance. None of the 26 other SH3 domains from *S. cerevisiae* did so. This second observation makes sense of course: if SH3 domains other than the correct one trigger the osmoresistance pathway, then this pathway will be triggered by mistake, which is undesirable. However, as Zarrinpar *et al.* [1] noted, the fact that an SH3 domain from another organism sometimes binds to the osmoresistance Pbs2 of yeast while the yeast's own SH3 domains do not, suggests that the sequences of the SH3 domains are not only under selective pressure to bind to their partners they are also under selective pressure not to bind to proteins other than their partner.

In order to compare with the experimental data, we have generated many proteomes and have calculated the probability distribution function $p(u)$. u is the interaction energy between pairs of patches, in units of $-\epsilon$. We performed calculations for a proteome of $N = 32$ patches each with sequences $n_N = 16$ monomers long; the minimum equilibrium constant for binding $K_b = 10^6\text{nm}^3$ and the maximum equilibrium constant allowed between patches that are not partners $K_s = 10^4\text{nm}^3$. These values for K_b and K_s are comparable to those in *in vitro* measurements [1]. A binding strength of $K_b = 10^6\text{nm}^3$ requires an interaction energy of at least $u = 8$, and the upper limit $K_s = 10^4\text{nm}^3$ means that non-binding pairs must have $u \leq 5$.

Figures 1a) and 1b) show the results for proteins in a functional proteome. In a) $p(u)$ is plotted for pairs of proteins that are binding partners, whereas in b) $p(u)$ is plotted for proteins that are not binding partners. Note that the data in b) do not fall on a smooth curve because the number of ways of achieving a given value of u is not a smooth function of u. The curve is reproducible. We see that for binding partners $p(u)$ is cut off below $u = 8$, while for other pairs $p(u)$ is cut off above $u = 5$. These two cutoffs are essential for the proteome to be functional. We need $u \geq 8$ for a strong interaction in the protein interaction network of our model proteome, but to avoid 'cross-talk' we need all other interactions to be much weaker. When proteins are designed by protein engineers to bind with high specificity, interactions with the wrong other proteins have to be explicitly designed out [4]. This is often called negative design. For example Kortemme *et al.* [6] redesigned an existing pair of binding proteins and used specific negative design to prevent the redesigned proteins binding to the original proteins. Essentially, they designed in an upper cutoff of the form seen in figure 1b).

Figure 1c) shows results for the interactions between a foreign patch that has been introduced into a proteome and the N patches of this proteome. This distribution has no cutoffs. Collectively the sequences of all the patches of a single viable proteome must be such that equation (1) is satisfied, and so there are cutoffs present in the $p(u)$'s, but when proteins from different proteomes are considered, this is not the case. The interactions are not constrained. We have also calculated the probability distribution function for u assuming complete randomness. We generated sequences of $n_M = 16$ monomers that were random but on average had the same fractions of polar, positive etc. monomers. For the parameter values we are working with, the average fractions of monomers that are hydrophobic, polar, positive, and negative are 0.17, 0.28, 0.27 and 0.27, respectively. The distribution function for the interactions between patches with these sequences is shown as the crosses joined by lines in figure 1c). We see that it describes the probability distribution function for the interaction energy u between patches from different proteomes (the squares) quite well. Also, we have shown the results for only one pair of values of n_M and N but the results are not very sensitive to variations in either parameter. In particular, the qualitative shapes of the

$n_M =$	12	16
$N = 16$	0.15	0.21
$N = 32$	0.10	0.15
$N = 48$	0.07	0.13

Table 2. The probability p_r that a patch from one proteome will bind to a specific patch in another proteome sufficiently strongly to act as a binding partner. Sufficiently strongly means $u \geq 6$, equivalent to an equilibrium constant of $2.7\times10^4\text{nm}^3$ or a dissociation constant $\simeq 40\mu$M. The energy scale $\epsilon = 2k_BT$. Results are given for varying numbers of monomers, n_M, and of patches in the proteome, N.

curves is always as in figure 1.

Zarrinpar *et al.* [1] found that proteins that interacted with a dissociation constant of around $K_d = 40\mu$M (measured *in vitro*) or less were able to at least partially reconstituted osmoresistance. For our model, $u = 6$ is enough to give a dissociation constant this small. We define p_r as the probability of two patches interacting with $u \geq 6$ and so binding sufficiently strongly to act as binding partners. The probability of two patches from different proteomes interacting with $u \geq 6$, is $p_r = 0.15$. This is rather lower than the 50% of foreign domains that reconstituted osmoresistance in yeast [1], but given the simplicity of our model we would not have expected quantitative agreement. Zarrinpar *et al.* [1] also present indirect evidence that the foreign SH3 domains interfere with signalling pathways other than the osmoresistance pathway. Within our model, with $u = 6$ considered enough to bind, on average a foreign patch strongly affects $32 \times 0.15 \simeq 5$ pathways.

In addition to our calculations with $N = 32$ patches with sequences $n_M = 16$ long, we have performed calculations with other values of N and n_M and tabulated the results for the probability of reconstituting a pathway in Table 2. The probability of reconstitution increases with the number of monomers and decreases with the number of patches in the proteome. Let us consider the dependence on N, the number of patches. We found above that the distribution function $p(u)$ for a protein in a foreign proteome could be reproduced quite accurately by simply modelling the sequences by random sequences with the appropriate fractions of H monomers, P monomers etc., see figure 1c). Now, for proteomes with $n_M = 16$ and $N = 64$, the average fractions of monomers that are hydrophobic, polar, positive and negative are 0.16, 0.29, 0.28 and 0.28, respectively. These figures are rather similar to those for $n_M = 16$ and $N = 32$, 0.17, 0.28, 0.27 and 0.27, respectively. However, for the larger value of N there are slightly fewer H monomers, and slightly more of the charged and polar monomers. The H monomers mediate the strongest attraction so if we consider the interactions between random sequences we see that decreasing the number of H monomers while increasing the fraction of P, + and − monomers, will on average weaken

the interactions. Varying the fraction of + and − monomers will have no effect when the sequences are random as there will as many ++ and −− interactions as +− interactions. Thus, as N increases the patches become a little less hydrophobic. Then two patches with uncorrelated sequences are less likely to interact with an interaction energy $u \geq 6$ and so reconstitute a pathway. Recall that the interaction between patches from different proteomes can be modelled quite accurately by the interaction between uncorrelated sequences, see figure 1c).

It is also true that p_r is sensitive to the alphabet of monomers used. Work with a simple 2-letter alphabet version of the model, that of [3], gives significantly smaller values of p_r [29]. The larger, 20-letter, alphabet of real proteins may partially explain the larger values of p_r found by Zarrinpar *et al.* [1]. We should also say that all our N model patches are equivalent, whereas presumably there was some variation in the SH3 domains studied by Zarrinpar *et al.* [1], particularly as they are from a variety of different species. Finally, when we evolved our proteomes, we took all the probabilities of type $\alpha = H, P, +, -$ mutating to a monomer of a different type β, $p_{\alpha\beta}$ to satisfy $p_{\alpha\beta}/p_{\beta\alpha} = 1$ except that $p_{H\beta}/p_{\beta H} = 1/2$. In the absence of selection for binding, this makes H monomers twice as common as the other three types. This was just a simple choice; calculations done where the probabilities $p_{\alpha\beta}$ are different changes the values of p_r somewhat.

Our proteomes are a uniform, i.e., unbiased, sample of all proteomes satisfying our requirement for viability, equation (1), Proteomes can be considered as points in proteome space. A point in this space is defined by the values of the elements of the sequences of all the patches. Proteome space is just a generalisation of protein space [28]. We sample the functional parts of this proteome space uniformly. Evolution will introduce at least some bias, i.e., favour some possible fit proteomes over others. However, although the curves of figure 1 are averages over many proteomes, the curve for a single protein inserted into a single foreign proteome typically looks rather similar to figure 1, i.e., most foreign proteins inserted into most proteomes have results that resemble our averaged findings. We conclude that biases are unlikely to change our results greatly.

5.1. Rate of evolution

It has been suggested [10] that analysis of protein sequence data shows that there is a negative correlation between the number of other proteins a given protein interacts with and its rate of evolution. This suggestion is controversial [10–15]. There is also other work [16, 17]. that has found that the evolution of residues in proteins that participate in specific binding is slower than that of other residues. We look at whether neutral evolution of the sequences of our patches depends on whether they are required to bind to specific other patches or not.

We consider only neutral evolution, not the evolution of new interactions, and assume that the rate of neutral evolution is proportional to the fraction of mutations that are neutral

not deleterious. The idea is that mutations arise spontaneously at a constant rate; those that are neutral are retained and contribute to the measured rate of change of the sequence, while those that are deleterious are eliminated. Thus, the higher the fraction of possible mutations that are neutral the faster the rate of neutral evolution.

When we evolve our proteomes they undergo a random walk in proteome space. For the system of figure 1, i.e., $N = 32$ patches, on average a fraction 0.21 of mutations of a functional proteome result in another functional proteome, as defined by equation (1). The remaining fraction 0.79 of mutations result in proteomes that violate equation (1). In contrast if we consider a proteome of $N = 32$ patches that are simply required not to be too sticky, i.e., we simply require that all $K_{ij} < K_s$, a fraction 0.54 of mutations of one functional proteome result in another functional proteome. In the absence of any requirement to engage in high-specificity protein-protein interactions, the proteome explores proteome space at a rate that is a factor of $0.54/0.21 = 2.6$ higher. Thus, within this simple calculation for our simple model, the requirement to bind reduces the rate of evolution by almost a factor of three.

We can look at why mutations of a proteome with high-specificity binding are deleterious. We find that a fraction 0.27 of mutations are deleterious because they weaken the binding between a patch and its partner enough to make the equilibrium constant drop below K_b, but a much higher fraction, 0.67 of mutations are deleterious because they cause a patch to bind to a patch other than its partner with an equilibrium constant greater than K_s. Thus most mutations are deleterious because they violate the negative-design requirement to avoid sticking to the wrong other patch. Selection eliminates sequences because they bind too strongly to a protein other than their partner. Note that 0.67 and 0.21 do not add up to 0.79 because some mutations cause both requirements to be violated. Within our model, the functional requirement to bind with high specificity reduces the fraction of neutral mutations more than does the requirement not to stick to other patches. For these reasons binding patches evolve much more slowly. In real proteins, other factors will affect the rate of evolution: the evolution may not be completely neutral; there are constraints on the amino-acid sequences other than those due to protein-protein interactions, the requirement to fold for example, and so on. These other factors may be obscuring the effect of binding on the rate of evolution, making the evidence for it ambiguous.

6. Conclusion and outlook

The work of Zarrinpar *et al.* [1] suggests that if we want to synthesise a peptide that can block one and only one signalling pathway, not only do we need to design the peptide to bind to a specific domain but we need to design it to specifically avoid all other members of this class of domains. This is certainly true for the model proteomes studied here, and of

course it requires knowledge of the sequences of all these other proteins. Thus, if we want to develop a peptide drug that is highly specific, i.e., had few or no side effects, then we need to explicitly use negative design to design out unwanted interactions in order to avoid side effects. Without negative design the equilibrium constants for the interactions of the peptide drug with other domains, and hence potential side effects, may lie anywhere within a very wide range. Indeed the distribution of interaction energies can be accurately modelled by the interaction energies between purely random sequences, with the same average fraction of hydrophobic, positive, etc. monomers.

Half of the 12 foreign SH3 domains were found to bind to the osmoresistance SH3 partner (Pbs2) in yeast whereas none of yeast's other 26 SH3 domains did so. This is clear evidence: i) that the evolution of the 27 SH3 domains is that the proteome level, i.e., it is not the case that one of the 27 pairs evolves independently of the other 26 pairs, and ii) that the evolution has required selection against binding to the wrong other protein as well as for binding to the right other protein. Also, maintaining this functional set of proteins must involve the elimination of mutations that cause inappropriately strong interactions between a domain and another domain that is not its partner. Certainly within our model, during our simple model of neutral evolution, more mutations are rejected because they cause interactions between domains that are not partners than because they weaken a binding interaction. There is evidence that proteins whose function requires a number of specific interactions evolve slowly [10–17]. If this is so our calculations suggest that the slow rate of evolution is to a great extent due to many mutations being rejected because they cause inappropriate binding. Studies on yeast with mutant forms of binding proteins may be able to observe this effect.

Acknowledgements

It is a pleasure to acknowledge J. Doye, D. Frenkel, A. Louis and M. Vendruscolo for discussions. I would like to thank The Isaac Newton Institute for Mathematical *Science*s for its hospitality during my stay there as a participant in the Statistical Mechanics of Molecular and Cellular Biological Systems programme, and T. McLeish and W. Poon for the opportunity to participate. This work was supported by The Wellcome Trust (069242).

Glossary

Interactome. A complete set of interactions in an organism, typically specific functional interactions between proteins.
Negative design. The design of an object not to do something, e.g., a protein can be designed not to bind to another protein. This is contrast to positive design, where a protein might be

designed to do something, e.g., to bind to another protein.
Proteome. A set of proteins, especially either the complete set of proteins encoded for by the genome of an organism or that present in a cell at a particular time.
Specific binding. A pair of proteins in a mixture of many proteins, such as that inside cells, binding to each other and only to each other, not to any other protein present.

References

[1] Zarrinpar A, Park SH and Lim WA 2003 *Nature* **426** 676-670
[2] Sear RP 2004 *J. Chem. Phys.* **120** 998-1005
[3] Sear RP 2004 *Phys. Biol.* **1** 53-60
[4] Havranek JJ and Harbury PB 2003 *Nature* Struct. Biol. **10** 45-52
[5] Shifman J M and Mayo S L 2003 *Proc. Nat. Acad. Sci.* 2003, 100:13274-13279
[6] Kortemme T, Joachimak LA, Bullock AN, Schuler AD, Stoddard BL and Baker D 2004 *Nature Struct. Mol. Biol.* **11** 371-379
[7] Hecht MH, Richardson JS, Richardson DC and Ogden RC 1990 *Science* **249** 884-891
[8] Hellinga HW 1997 *Proc. Nat. Acad. Sci.* **94** 10015-10017
[9] Lazar GA, Marshall, SA, Plecs JJ, Mayo SL and Desjarlais JR 2003 *Curr. Opin. Struct. Biol.***13** 513-518
[10] Fraser H B, Hirsh A E, Steinmetz L M, Scharfe C and Feldman M W 2002 *Science***296** 750-752
[11] Jordan I K, Wolf Y I and Koonin E V 2003 *BMC Evol. Biol.* **3** 1
[12] Fraser H B, Wall D P and Hirsh A E 2003 *BMC Evol. Biol.* **3** 11
[13] Bloom JD and Adami C 2003 *BMC Evol. Biol.* **3** 21
[14] Fraser HB and Hirsh AE 2004 *BMC Evol. Biol.* **4** 13
[15] Bloom JD and Adami C 2004 *BMC Evol. Biol.* **4** 14
[16] Caffrey DR, Somaroo S, Hughes JD, Mintseris J and Huang ES 2004 *Prot. Sci.* **13** 190-202
[17] Teichmann S A 2002 *J. Mol. Biol.* **324** 399-407
[18] Alberts B, Bray D, Lewis J, Raff M, Roberts K and Watson JD 1994 *Molecular Biology Of The Cell* (3rd Edition, Garland Publishing, New York).
[19] Mayer BJ 2001 *J. Cell. Sci.* **114** 1253-1263
[20] Panni S, Dente L and Cesareni G 2002 *J. Biol. Chem.* **277** 21666-21674
[21] Kimura M 1968 *Nature* **217** 624-626
[22] King JL and Jukes TH 1969 *Science* **164** 788-798
[23] Bastolla U, Porto M, Eduardo Roman H and Vendruscolo M 2003 *J. Mol. Evol.* **57** S103-S119
[24] Backofen R, Will S and Bornberg-Bauer E 1999 *Bioinformatics* **15** 232-242
[25] Onuchic JN, Luthey-Schulten Z and Wolynes PG 1997 *Ann. Rev. Phys. Chem.* **48** 545-600
[26] Pande VS, Grosberg AYu and Tanaka T 2000 *Rev. Mod. Phys.* **72** 259-314
[27] Williams PD, Pollock DD and Goldstein RA 2001 *J. Mol. Graphics Model.* **19** 150-156
[28] Maynard-Smith J 1970 *Nature* **225** 563-564
[29] Sear RP unpublished.