Published in IET Systems Biology
Received on 14th December 2007
Revised on 14th April 2008
doi: 10.1049/iet-syb:20080076

Special Issue – Selected papers from the First q-bio Conference on Cellular Information Processing

ISSN 1751-8849

Satisfiability, sequence niches and molecular codes in cellular signalling

C.R. Myers

Computational Biology Service Unit, Life Sciences Core Laboratories Center, Cornell University, Ithaca, NY, USA
E-mail: crm17@cornell.edu

Abstract: Biological information processing as implemented by regulatory and signalling networks in living cells requires sufficient specificity of molecular interaction to distinguish signals from one another, but much of regulation and signalling involves somewhat fuzzy and promiscuous recognition of molecular sequences and structures, which can leave systems vulnerable to crosstalk. A simple model of biomolecular interactions that reveals both a sharp onset of crosstalk and a fragmentation of the neutral network of viable solutions is examined as more proteins compete for regions of sequence space, revealing intrinsic limits to reliable signalling in the face of promiscuity. These results suggest connections to both phase transitions in constraint satisfaction problems and coding theory bounds on the size of communication codes.

1 Introduction

The functioning of complex biochemical pathways hinges on conveying molecular signals reliably in the stochastic and evolving milieu of living cells. These signals are mediated by molecular interactions that distinguish physiological binding partners from myriad other cellular constituents: this ability to distinguish functional signals from molecular noise is ultimately the source of information processing in cellular networks. But molecular recognition is subtle: many of the molecular interactions involved in cellular regulatory and signalling pathways do not involve highly specific 'lock and key' binding, but instead are characterised by more fuzzy and promiscuous recognition of families of sequences and configurations [1–3]. Furthermore, there are often paralogous copies of molecules within a cell that interact with similar and potentially overlapping sets of substrates. This fuzzy recognition reflects a tradeoff between specificity and robustness, allowing systems to be more robust to genetic mutations [4]. But while interaction promiscuity may provide robustness against mutations, as well as opportunities for different modes of regulatory control, it introduces fragilities elsewhere, leaving systems more vulnerable to potentially disadvantageous crosstalk among reactants. Therefore a basic question concerning cellular signalling in crowded sequence spaces, where multiple proteins bind to similar families of molecular sequences and structures is: under what circumstances can crosstalk be avoided in such a system? This paper investigates a simple null model, associated with random molecular sequences, that is amenable to analysis and suggests connections to recent work on phase transitions in combinatorial NP-complete problems. While not directly applicable to the evolved molecular sequences found in nature, this model serves as a useful first step in defining the landscape of constraint satisfaction in cellular signalling.

The theory of communication in noisy channels, dating back to the seminal work of Shannon [5, 6], also provides a useful framework in which to interpret cellular signals. Engineered error-correcting codes embed messages in higher-dimensional spaces (e.g. via encoded checks on the message integrity), to insulate each possible codeword within a sphere in the embedding space. By packing such spheres so that they are disjoint, any corrupted word in a message can (up to some defined number of errors) be uniquely associated with an original code word. In molecular signalling, sequence recognition volumes play a similar role: these volumes describe the sets of sequences recognised (i.e. bound with significant probability) by different molecules. In molecular signalling, however, overlapping recognition of sequences precludes the sort of disjoint sphere packings found in engineered codes. Instead of asking, therefore, whether all messages can be

doi: 10.1049/iet-syb:20080076

communicated through a molecular interaction channel, we focus here instead on whether any message can be so conveyed (under the assumption that evolutionary selection might find such a solution if it does in principle exist). A central result presented here, which establishes limits on the number of proteins that can compete for regions in sequence space before crosstalk becomes likely, is akin to a bound on the size of a code in a communication system.

This problem – molecular discrimination in the face of potential crosstalk – arises in a variety of contexts. A classic problem in immunology is the ability of antibodies to discriminate between 'self' and 'nonself' antigens, with much work focused on identifying how large a recognition region needs to be in order to reliably perform this discrimination [7, 8]. In gene regulation, transcription factors (TFs) that control gene expression by binding to DNA are organised in families that often recognize similar sorts of sequences. Recent work in that area has explored tradeoffs between binding TF specificity and system robustness [4], balances between selection and mutation of TFs [9], evolutionary divergence of competing TF-binding sequence pairs to avoid crosstalk [10] and the application of ideas from coding theory to understand limits on the size of TF families [11]. Signal transduction is mediated largely by protein–protein interactions. In bacteria, this involves two-component systems with sensor kinases that activate response regulators, and active research is focused on how specificity is maintained among sensor–regulator pairs and to what extent there is crosstalk and cross-regulation within larger sets [12–15]. In eukaryotes, signalling often involves modular protein domains (e.g. SH2, SH3, WW) that recognise characteristic peptide motifs in partners [16]. There can be tens or even hundreds of proteins within a paralogous family in a given organism that must discriminate among sets of potential interaction partners, and inferring such interactions and their specificity as a basis for developing predictive models of signalling pathways is a crucial task in systems biology.

The problem of molecular discrimination provides the broad backdrop for this work, but the role of sequence niches in particular was crystallised in a set of elegant experiments on SH3-mediated signalling in yeast (*Saccharomyces cerevisiae*), by Zarrinpar *et al.* [17]. SH3 domains are known to bind to a set of proline-rich peptide sequences (the so-called 'PXXP' motif) [2, 18]. Zarrinpar *et al.* probed the yeast high-osmolarity signalling pathway, which involves the interaction of Sho1 (a protein with an SH3 domain) and Pbs2 (containing a PXXP motif). By making chimeric versions of Sho1 containing different SH3 domains, they demonstrated that no native yeast SH3 domains other than that in Sho1 were capable of interacting with Pbs2, but that half of the metazoan SH3 domains they tested were able to do so. They surmised that there has been an evolutionary selection against crosstalk with that pathway in yeast, with protein sequences having co-evolved such that the Pbs2 ligand lies in a niche in sequence space where it is recognised by only the Sho1 SH3 domain. Since there has been no such selection pressure to avoid crosstalk in other organisms, the Pbs2 motif bound to non-native SH3 domains with greater probability. (See supplementary text and Fig. S.1 for further discussion.) It is the structure of these sorts of sequence niches that form the core of this paper.

2 Results

2.1 Sequence niche question

We begin by distilling the central question to be considered here: under what conditions does a unique sequence niche exist so that signalling without crosstalk might be possible? To address this question, a highly abstracted model of molecular interaction is adopted, in which sequences are represented by binary strings of length L, as opposed to the 4-letter nucleotide alphabet relevant for protein–DNA interactions or the 20-letter amino acid alphabet for protein–protein binding. (Binary sequence models, such as the HP model, have been used in the study of protein folding [19], although it remains an open question as to whether there is an appropriate coarse-grained alphabet capable of capturing the essential biochemistry of protein–protein interactions involved in signalling [20].) In this model, binding of a sequence to a protein is achieved if the sequence is sufficiently close to the optimal sequence recognised by the protein, with Hamming distance used as a measure of closeness: two sequences bind if they differ in at most R positions, given some promiscuity radius R. Given this representation, this paper can pose the sequence niche question (SNQ), phrased and typeset in the canonical style of Garey and Johnson [21] and illustrated schematically in Fig. 1:

Sequence niche

Instance: Binary sequence T of length L, a set of binary crosstalk sequences C_i, for $i = 1, \ldots, N$, each of length L and an integer R, $0 \le R \le L$.

Question: Is there a binary sequence s of length L such that $H(T, s) \le R$ and $H(C_i, s) > R$ for $i = 1, \ldots, N$, where $H(x, y)$ is the Hamming distance between sequences x and y?

SNQ is an example of the distinguishing string selection problem (DSSP), as defined by Lanctot *et al.* [22]. (The DSSP allows for S_c strings to be within Hamming distance k_c, and S_f strings to be at least Hamming distance k_f apart.) The DSSP was proven to be NP-complete [22]; the SNQ is the DSSP with $S_c = 1$ and $R = k_c = k_f - 1$, but the computational complexity of the DSSP does not depend on the values of these parameters, so the SNQ is also NP-complete. The SNQ is similar in spirit to the well-known computer science problem SAT (and its specialisation

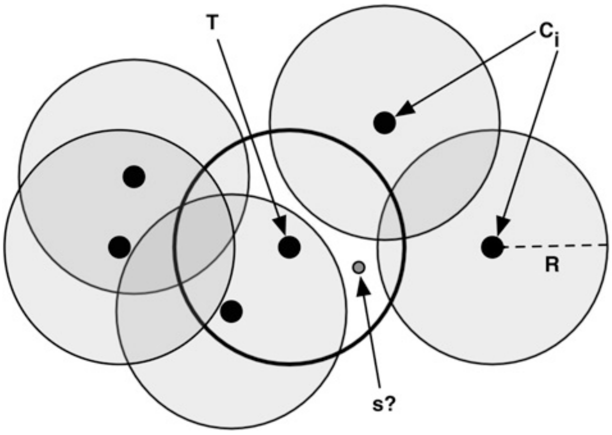

Figure 1 *Sequence niche question: given a target protein sequence T and a set of N crosstalking protein sequences {C}, is there a sequence s that is bound by T but not by any of the proteins C_i*

In this model, sequences are binary strings of length *L*, and two sequences bind if the Hamming distance between them is less than or equal to *R*

K-SAT), in that these problems ask whether there exists a solution that satisfies a set of (potentially conflicting) constraints [21]. Borrowing from the language of SAT, we say a particular instance of the SNQ is 'satisfiable' when a solution *s* exists, and 'unsatisfiable' otherwise. The SNQ asks whether discrimination of one target protein from a background of crosstalking proteins is possible. A symmetric generalisation of this problem would ascertain whether every protein in a collection is distinguishable, that is, whether there is a separate sequence niche for each of the *N* proteins; a problem of this sort was investigated previously by Sear [23, 24]. The generalised SNQ is presumably in the same complexity class as the single-target SNQ, since deciding it simply involves deciding *N* separate SNQs.

2.2 *Satisfiability of random sequence niches*

The NP-completeness of the SNQ is a statement about its worst-case complexity, but there has been increasing interest in recent years in quantifying the typical-case complexity of NP-hard problems. A common strategy is to examine ensembles of random instances of such problems, investigating how solution complexity depends upon parameters that characterise those random instances. A similar strategy is adopted here.

Multiple random instances of the SNQ were examined (with uniform equal probability of 0s and 1s in the sequence strings), for various values of the problem parameters *L*, *R* and *N*. A recursive algorithm proposed by Gramm *et al.* [25] was used to determine whether a given instance had a solution; see supplementary text for further details. Fig. 2*a* shows the average unsatisfiable fraction of random SNQ instances as a function of the number of crosstalking proteins *N*, averaged over an ensemble of 100 random instances for each *N*. Data are shown here only for $R = 2$ and $R = 8$; for intermediate values of *R*, the satisfiability data interpolate between these extremes. In addition, Fig. 2*b* shows the median solution time τ required for determining whether or not an instance is satisfiable. Similar to as is done for *K*-SAT, solution times are measured in units of the number of recursive calls to the solution algorithm [25]. (Since the distribution of solution times over random instances often show heavy tails [26], the median solution time is a better estimate of typical complexity than is the mean.) Fig. 2*a* demonstrates a transition from satisfiability (SAT) to unsatisfiability (UNSAT) as the number of crosstalking proteins is increased. Rather than a gradual diminution in the capacity for reliable signalling, the SNQ exhibits a relatively abrupt switch as log *N* increases. Fig. 2*b* reveals, for the same set of parameter values, that the solution time of the algorithm peaks near the point of the SAT-UNSAT transition, that is, it becomes significantly more difficult to decide if a given instance is satisfiable when that instance lies near the transition. The characteristic scales of the random SNQ are seen to vary over orders of magnitude. For the solution times, this is perhaps not surprising: since the SNQ is NP-complete, we expect the worst-case run time of the solution algorithm to be exponential in the size of the problem.

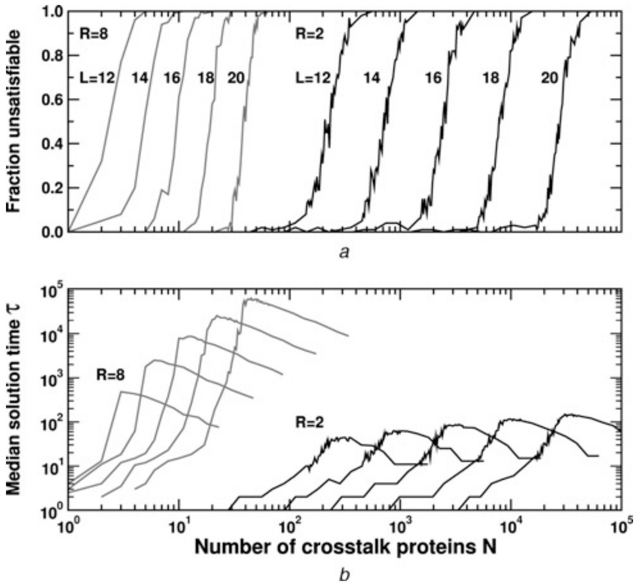

Figure 2 *Satisfiability and solution time data*

a Average fraction of unsatisfiable instances of the random SNQ as a function of *L*, *R* and *N* [(*L*, *R*) specified in figure legend, *N* varying along *x*-axis]

b Median solution time τ of the SNQ decision (number of recursive calls in the solution algorithm) for the same instances depicted in *a*

Averages in *a* and medians in *b* are for 100 instances of the SNQ for each (*L*, *R*, *N*) set

2.3 Scaling of the SNQ transition: a satisfiability bound on the number of crosstalking proteins

We can develop a simple scaling theory to describe the transition from satisfiability to unsatisfiability as we vary parameters L and R. A given instance is unsatisfiable if the target volume (i.e. the Hamming sphere of radius R surrounding the target sequence T) is completely covered by the union of the crosstalk volumes (centred about the crosstalk sequences $\{C\}$), a process illustrated schematically in Fig. 3*a*. We can estimate the critical number of crosstalk proteins N_c needed to cover the sequence volume of the target protein. The full derivation (along with extensions) is provided in the supplementary text, but essentially the bound stems from estimating the average number of sequences in the Hamming sphere of volume $V(L, R)$ centered about the target T remaining uncovered after N crosstalk proteins have been deposited at random in a sequence space of volume $V_0(L)$, which is modelled as a binomial process. When there are $O(1)$ sequences left uncovered, it is expected that the target volume to be covered with probability 1/2, such that $V(1 - V/V_0)^{N_c} = 1$, implying:

$$N_c = \frac{\log(1/V)}{\log(1 - V/V_0)} \qquad (1)$$

where $V_0(L) = 2^L$ is the total number of possible binary sequences of length L, and $V(L, R) = \sum_{n=0}^{R} \binom{L}{n}$ is the number of binary sequences in a ball of Hamming radius R about a given sequence. As discussed in more detail below, this can be interpreted as a random satisfiability bound on the approximate number of randomly distributed proteins that can coexist without crosstalk.

With this critical protein number N_c, the raw satisfiability and solution time data of Fig. 2 can be rescaled. These rescaled data are shown in Fig. 3, where we show results for $R = 2$, 3, 4, 6, 8 (not just $R = 2$ and 8 as plotted previously). In Figs. 3*b* and 3*c* the protein number (x-axis) is scaled as $N \to (N - N_c)/N_c$, and in Fig. 3*c*, the solution time data (y-axis) are scaled by the exponentially growing number of sequences in the search tree $V(L, R)$ that in principle need to be considered. The collapse of each set of unscaled data onto a reasonably compact scaling form suggests this simple description is approximately correct, although there is clearly some systematic variation with Hamming radius R. The scaling collapses for the solution time data are more variable than for the satisfiability fraction. The variability in the scaled solution time data indicates differences in efficiency of pruning the search tree, for the heuristics used in the recursive solution algorithm [25]. Closer examination of the data (not shown) suggests this efficiency is dependent approximately on the ratio L/R.

2.4 Fragmentation of the solution space

Previously it was considered whether there is any solution to a given instance of the SNQ. Here the structure of the space of all satisfying solutions for an instance is examined, as determined via exhaustive enumeration.

Consider a fixed target sequence T and a set of potential crosstalk sequences $\{C\}$. Imagine introducing crosstalk sequences one at a time, and identifying the set of all sequences $\{s_N\}$ that satisfy the SNQ for that instance with N crosstalk sequences. Of particular interest here is the size and structure of the solution set $\{s_N\}$ as a function of the number of proteins N. For each set, a graph is assembled whose nodes are sequences s that satisfy the SNQ and whose edges connect satisfying sequences if they are neighbours on the hypercube, that is, if their Hamming distance from each other is 1. This graph represents the neutral network of all solutions to a given instance of the SNQ, along which single point mutations to the solution string (bit flips) can be made without producing crosstalk. For various N, we compute the set of connected

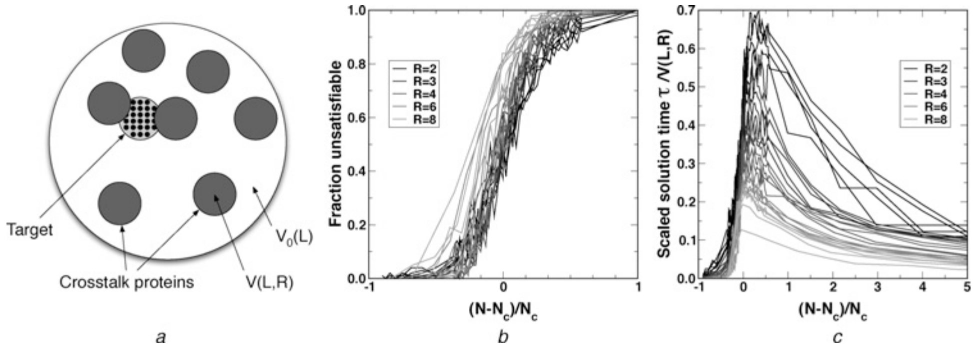

Figure 3 *Scaling description of the SAT–UNSAT transition in the SNQ*

a Schematic depiction of the covering of available sequences (black dots) in the target volume as crosstalk proteins (grey circles) are laid down randomly

b and *c* Scaling of the satisfiability and run time data in Fig. 2 based on the scaling theory presented: (*b*) the number of crosstalk proteins N are scaled by $N \to (N - N_c)/N_c$, and (*c*) in addition to scaling N, the run times τ are scaled by the number of sequences in the target volume $V(L, R)$ that must be considered components of the resulting graph. The change in the structure of the neutral network of satisfying solutions is illustrated, for a particular family of problem instances with $L = 16$ and $R = 6$, in Fig. 4. For small numbers of proteins (Fig. 4*a*), there are many possible solutions to the SNQ, and those solutions all coalesce into one connected cluster, such that any solution can be reached from any other via a succession of single-bit flips to the solution string. As N increases (Fig. 4*b*), the number of satisfying solutions decreases, and the connected cluster of solutions is fragmented into many disjoint sets (still dominated by a central core). This fragmentation and evaporation of the sequence clusters continue for larger N (Fig. 4*c*), until finally all solutions disappear, and unique signalling is no longer possible. While the neutral networks shown reveal the effects of mutations in the solution string s, it should be noted that single point mutations in the sequences representing the centres of the proteins T and $\{C\}$ – that is, mutations in the SNQ instance itself – can result in drastic changes in the neutral network topology, for example, by fragmenting a single large cluster into a set of smaller ones.

A summary of these trends is shown in Fig. 4*d*, by averaging over many SNQ instances (for $L = 16$ and $R = 6$). This reveals that the size (i.e. the number of nodes) of the largest cluster (solid line) decreases roughly exponentially with crosstalk number N. We can understand this decrease in part by considering the geometric argument summarised in Fig. 3*a*, which suggests that the size of the largest cluster should decrease approximately as e^{-qN}, where $q \equiv V(L, R)/V_O(L)$ (see the supplementary text for details). Also shown in Fig. 4*d* is the number of disjoint clusters (dashed line); this is seen to initially increase with N – as the single satisfying solution cluster is fragmented – and then decrease – as small sequence clusters evaporate in the presence of new crosstalk proteins. Fig. 4 reveals a number of isolated clusters of size 1, but these problem sizes are rather small (given the computational burdens of exhaustive enumeration). It is an open question whether nontrivial cluster size distributions will reveal themselves as larger problem sizes are considered.

3 Discussion

The goal of this paper has been to examine the limits of crosstalk-free communication in a simple model of competitive molecular interactions, as a first step towards developing a more comprehensive and realistic theory applicable to protein–protein and protein-DNA interactions involved in regulation and signalling. The numerical experiments presented were motivated by phase transitions observed in the random K-SAT problem [27–30], where a SAT–UNSAT transition occurs as the ratio of constraints to variables is increased. The numerical results presented for the SNQ demonstrate something similar: a relatively sharp transition from satisfiability to unsatisfiability with increasing competition for sequence space, along with an increase in computational complexity near the transition. Phase transitions have been studied in a number of NP-hard problems, although applications to biological problems have been scant and generally at coarser levels of biological description [31–33]. A second phase transition has more recently been identified in K-SAT, lurking near the SAT–UNSAT phase boundary, involving the fragmentation of the set of satisfying solutions [34–36]. We find evidence for such a fragmentation transition in small instances of the SNQ, although further theoretical and computational work is needed to fully characterise these transitions, which are only strictly defined in the limit of infinite system size.

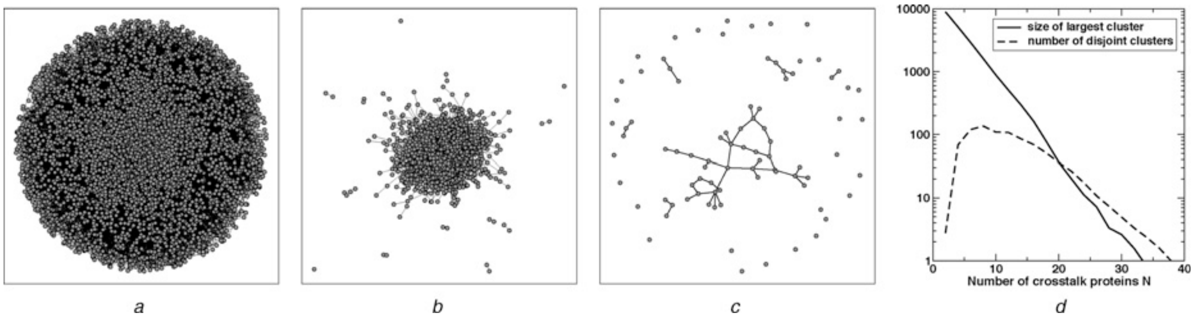

Figure 4 *Fragmentation of the solution space as the SAT–UNSAT transition is approached*

The neutral network of satisfying solutions $\{s_N\}$ for one particular problem instance ($L = 16$, $R = 6$), as a function of number of crosstalking proteins N

Satisfying sequences (nodes) are connected by edges (lines) in a network if they are separated by Hamming distance 1

The spatial layout of nodes has no meaning; all sequences are vertices on an L-dimensional hypercube

a $N = 4$: there are 5786 satisfying solutions in one large connected component. This cluster is broken up into multiple pieces as N increases

b $N = 12$: 1226 sequences are distributed among 18 connected components

c $N = 20$: only 85 sequences remain viable, scattered across 38 disjoint components

d For $L = 16$, $R = 6$, average values of the size of the largest connected sequence cluster (solid line) and the number of disjoint clusters (dashed line) as a function of N, averaged over 100 SNQ instances for each value of N

doi: 10.1049/iet-syb:20080076

The scaling of the SNQ transition – embodied in the critical number of crosstalking proteins in (1) – can be interpreted as a type of bound on the size of a molecular interaction code. Such a code is envisioned as relating two sets of molecules (e.g. proteins and their substrates), and the fidelity of communication in a molecular channel is a function of how reliably discrimination of signals can be achieved [37]. Bounds of this sort are common in coding theory, the most well-known being the sphere-packing bound derived by Shannon [5, 6]. The sphere-packing bound identifies the maximal number of spheres of Hamming radius R and dimension L that can be packed without overlap, such that a word with up to R errors can be unambiguously associated with a code word. Using the notation developed here, that implies that no more than the integer part of $V_0(L)/V(L, R)$ spheres can be disjointly arranged. The random satisfiability bound derived in (1) allows for denser, overlapping packings, since it considers only whether any message can be unambiguously associated with the target protein. Fig. 5 compares the sphere packing and random satisfiability bounds for some representative parameter values. The bound presented in (1) is explicitly applicable to binary sequences without reverse-complement symmetry. It is straightforwardly generalisable (see supplementary text), within the assumption that binding is entirely dictated by the Hamming distance between two sequences, to sequences with larger alphabets (e.g. 20 amino acids) or to sequences with reverse-complement symmetry (e.g. as has been done for other code bounds treating DNA sequences [11, 38]).

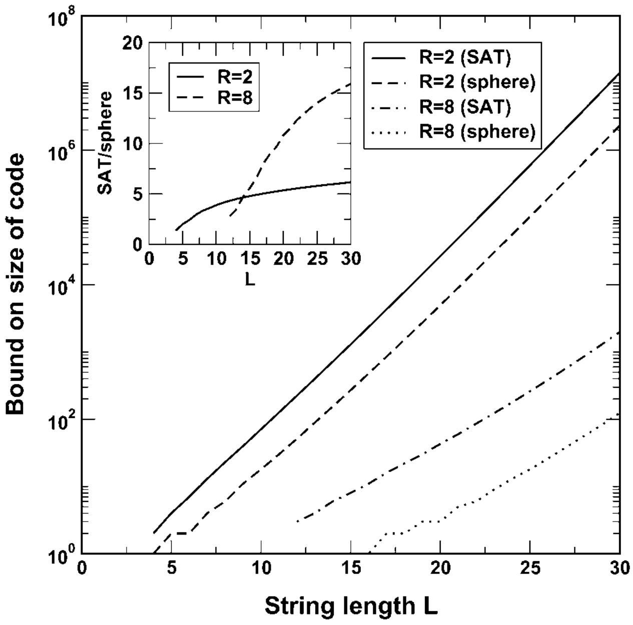

Figure 5 *Comparison of sphere packing (sphere) and random satisfiability (SAT) bounds on the size of a molecular code, for various L and R*

Inset: the ratio of SAT/Sphere bounds for the data shown
The SAT bound allows for more dense packing of spheres since not all sequences need to be disambiguated

Given the extreme simplicity of the model studied here, it is reasonable to ask whether the phenomena reported are relevant to the biology of protein–protein and protein–DNA interactions in cellular regulation and signalling. Those interactions are of course not dictated by Hamming distances and sharp cutoffs, but rather by dynamic and thermodynamic processes with softer thresholds determining the probability of interaction. In addition, interactions that might in principle be possible often do not occur in practice because they are outcompeted by other higher-affinity reactions, or even by a broad background of non-specific interactions. In this case, we should consider a sequence recognition volume as probabilistically defined, and not intrinsic to a given protein but dependent upon the context in which that protein finds itself. Despite these differences, molecular discrimination formally remains a constraint satisfaction problem regardless of the underlying details of representation and interaction, and phase transitions should in principle be possible. The larger question, in some sense, is whether biological systems actually do butt up against such constraints in their function and evolution. Clearly more work is needed to answer this, in part to identify the number of specificity-determining bases and/or residues in various protein–DNA and protein–protein interactions, as well as the effective alphabet size contributing to molecular discrimination. The experimental work reported in [17] demonstrated an increase in cross-reactivity among yeast SH3 domains and single-base-pair missense Pbs2 mutants, suggesting that the Pbs2 ligand lies near the periphery of a sparse and tenuous sequence niche. In related computational work motivated by sequence niches in SH3 signalling, Sear introduced a model based on a four-letter amino acid alphabet (hydrophobic, polar, positively and negatively charged) and equilibrium-binding kinetics to demonstrate that the mutual discrimination of a set of proteins and their substrates was possible [24]. Molecular modelling of protein–protein and protein–DNA interactions is not yet a broadly practical tool, and many computational predictions are instead based on sequence similarity with training data from experiments [39–41] or from comparative sequence analysis [15, 42]. One interesting question is whether analysis of cross-reactivity and sequence niches can provide more sensitive tests of the accuracy of predicted interactions. Also of interest is the geometry of recognition domains in real biological systems. The Hamming spheres considered here are compact, but it is unknown whether regulatory and signalling proteins recognise more convoluted sets of sequences, which could introduce even more geometric structure into the problem of mutual discrimination.

The biological implications of these sorts of constraints and transitions are also of interest. Nature has of course not produced random sequences, and a central question is what sorts of molecular codes has evolution uncovered to achieve reliable signalling. Have evolutionary innovations – such as novel interaction domains [11] or scaffolds that localise signalling proteins and confer context-dependent specificity in addition to the intrinsic sequence [43–46] – arisen to rescue cellular networks from the precipice of crosstalk? Fragmentation of the network of satisfying solutions of the sort demonstrated here leads to complex neutral network topologies. The extent to which neutral network topology influences evolution remains an open question [47, 48]. Neutral network fragmentation could lead to biological systems becoming frozen in local regions of sequence space, unable to mutate to other satisfactory configurations far away. This could produce a sort of speciation at the molecular scale, perhaps shedding light on phylogenetic relationships among related protein interaction domains. Larger-scale genomic rearrangements, such as homologous recombination and horizontal transfer, may play a role in helping biological communication systems become unstuck from a glassy, fragmented phase where single-point mutations are unable to do so. Addressing the question of evolving sequence niches, however, requires an appropriate definition of fitness. If discrimination among different sequences were the only determinant of fitness, we might expect encodings to more closely resemble sphere packings, with recognition volumes maximally distinct from one another. Other determinants could alter such packings, however; a fitness advantage from some weak crosstalk, perhaps as a form of degeneracy or functional redundancy [49], might keep recognition volumes from diverging too far from one another. And of course evolutionary mutation itself plays a central role in posing these constraint satisfaction problems, in that gene duplication leads to the creation of homologous proteins that recognise similar substrates. The random limit considered here, while useful for analysis, is not directly relevant to the biology of duplicated proteins that may diverge from one another just far enough to be distinguishable [10].

4 Acknowledgments

This work was supported by USDA-ARS project 1907-21000-027-03. I would like to thank Jim Sethna, Bart Selman, Carla Gomes, Walter Fontana, Marc Mézard, Sue Coppersmith, Chris Henley, Bistra Dilkina, David Schneider, David Krakauer and Bill Bialek for their useful inputs.

5 References

[1] PTASHNE M., GANN A.: 'Genes & signals' (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2002)

[2] MAYER B.: 'SH3 domains: complexity in moderation', *J. Cell Sci.*, 2001, **114**, (7), pp. 1253–1263

[3] CASTAGNOLI L., COSTANTINI A., DALL'ARMI C., *ET AL.*: 'Selectivity and promiscuity in the interaction network mediated by protein recognition modules', *FEBS Lett.*, 2004, **567**, (1), pp. 74–79

[4] SENGUPTA A., DJORDJEVIC M., SHRAIMAN B.: 'Specificity and robustness in transcription control networks', *Proc. Natl. Acad. Sci. USA*, 2002, **99**, (4), pp. 2072–2077

[5] SHANNON C.E.: 'A mathematical theory of communication', *Bell Syst. Tech. J.*, 1948, **27**, pp. 379–423; 623–656

[6] SHANNON C.E.: 'Communications in the presence of noise', *Proc. IRE*, 1949, **37**, pp. 10–21

[7] PERCUS J., PERCUS O., PERELSON A.: 'Predicting the size of the T-cell receptor and antibody combining region from consideration of efficient self-nonself discrimination', *Proc. Natl. Acad. Sci. USA*, 1993, **90**, (5), pp. 1691–1695

[8] FIGGE M.T.: 'Statistical model for receptor-ligand binding thermodynamics', *Phys. Rev. E*, 2002, **66**, (6), p. 061901

[9] GERLAND U., HWA T.: 'On the selection and evolution of regulatory DNA motifs', *J. Mol. Evol.*, 2002, **V55**, (4), pp. 386–400

[10] POELWIJK F.J., KIVIET D.J., TANS S.J.: 'Evolutionary potential of a duplicated repressor-operator pair: simulating pathways using mutation data', *PLoS Comput. Biol.*, 2006, **2**, (5), pp. 467–475 (e58)

[11] ITZKOVITZ S., TLUSTY T., ALON U.: 'Coding limits on the number of transcription factors', *BMC Genomics*, 2006, **7**, (1471–2164 (Electronic)), p. 239

[12] BIJLSMA J.J.E., GROISMAN E.A.: 'Making informed decisions: regulatory interactions between two-component systems', *Trends Microbiol.*, 2003, **11**, (8), pp. 359–366

[13] HELLINGWERF K.J.: 'Bacterial observations: a rudimentary form of intelligence?', *Trends Microbiol.*, 2005, **13**, (4), pp. 152–158

[14] LAUB M.T., BIONDI E.G., SKERKER J.M., MELVIN I., SIMON B.R.C., CRANE A.: 'Phosphotransfer profiling: systematic mapping of two-component signal transduction pathways and phosphorelays' (Academic Press, 2007), vol. 423, pp. 531–548

[15] BURGER L., VAN NIMWEGEN E.: 'Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method', *Mol. Syst. Biol.*, 2008, **4**, article id: 165

[16] CESARENI G., GIMONA M., SUDOL M., YAFFE M., *ET AL.*: 'Modular protein domains' (Wiley-VCH Verlag GmbH, Weinheim, 2005)

[17] ZARRINPAR A., PARK S.-H., LIM W.: 'Optimization of specificity in a cellular protein interaction network by negative selection', *Nature*, 2003, **426**, pp. 676–680

doi: 10.1049/iet-syb:20080076

[18] CESARENI G., PANNI S., NARDELLI G., CASTAGNOLI L.: 'Can we infer peptide recognition specificity mediated by SH3 domains?', *FEBS Lett.*, 2002, **513**, (1), pp. 38–44

[19] LAU K., DILL K.: 'A lattice statistical mechanics model of the conformation and sequence spaces of proteins', *Macromolecules*, 1989, **22**, pp. 3986–3997

[20] NOIREL J., SIMONSON T.: 'Neutral evolution of protein-protein interactions: a computational study using simple models', *BMC Struct. Biol.*, 2007, **7**, article id: 79

[21] GAREY M.R., JOHNSON D.S.: 'Computers and intractability: a guide to the theory of NP-completeness' (W.H. Freeman, 1979)

[22] LANCTOT J., LI M., MA B., WANG S., ZHANG L.: 'Distinguishing string selection problems', *Inf. Comput.*, 2003, **185**, (1), pp. 41–55

[23] SEAR R.P.: 'Specific protein-protein binding in many-component mixtures of proteins', *Phys. Biol.*, 2004, **1**, (2), pp. 53–60

[24] SEAR R.P.: 'Highly specific protein-protein interactions, evolution and negative design', *Phys. Biol.*, 2004, **1**, (3), pp. 166–172

[25] GRAMM J., NIEDERMEIER R., ROSSMANITH P.: 'Fixed-parameter algorithms for CLOSEST STRING and related problems', *Algorithmica*, 2003, **37**, (1), pp. 25–42

[26] GOMES C.P., SELMAN B., CRATO N., KAUTZ H.: 'Heavy-tailed phenomena in satisfiability and constraint satisfaction problems', *J. Autom. Reason.*, 2000, **24**, (1–2), pp. 67–100

[27] MITCHELL D.G., SELMAN B., LEVESQUE H.J.: 'Hard and easy distributions of SAT problems'. Proc. 10th Natl. Conf. on Artif. Intell. (AAAI), 1992, pp. 459–465

[28] KIRKPATRICK S., SELMAN B.: 'Critical behavior in the satisfiability of random Boolean expressions', *Science*, 1994, **264**, (5163), pp. 1297–1301

[29] MONASSON R., ZECCHINA R., KIRKPATRICK S., SELMAN B., TROYANSKY L.: 'Determining computational complexity from characteristic 'phase transitions'', *Nature*, 1999, **400**, (6740), pp. 133–137

[30] FRIEDGUT E.: 'Sharp thresholds of graph properties, and the k-sat problem', *J. Am. Math. Soc.*, 1999, **12**, (4), pp. 1017–1054

[31] CORREALE L., LEONE M., PAGNANI A., WEIGT M., ZECCHINA R.: 'Core percolation and onset of complexity in Boolean networks', *Phys. Rev. Lett.*, 2006, **96**, (1), p. 018101-4

[32] COPPERSMITH S.N.: 'Complexity of the predecessor problem in Kauffman networks', *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.*, 2007, **75**, (5), p. 051108–7

[33] GRAVNER J., PITMAN D., GAVRILETS S.: 'Percolation on fitness landscapes: effects of correlation, phenotype, and incompatibilities', *J. Theor. Biol.*, 2007, **248**, (4), pp. 627–645

[34] MEZARD M., ZECCHINA R.: 'Random K-satisfiability problem: from an analytic solution to an efficient algorithm', *Phys. Rev. E*, 2002, **66**, p. 056126

[35] MEZARD M.: 'Physics/computer science: passing messages between disciplines', *Science*, 2003, **301**, (5640), pp. 1685–1686

[36] MEZARD M., MORA T., ZECCHINA R.: 'Clustering of solutions in the random satisfiability problem', *Phys. Rev. Lett.*, 2005, **94**, (19), p. 197205

[37] TLUSTY T.: 'Rate-distortion scenario for the emergence and evolution of noisy molecular codes', *Phys. Rev. Lett.*, 2008, **100**, (4), p. 048101-4

[38] MARATHE A., CONDON A.E., CORN R.M.: 'On combinatorial DNA word design', *J. Comput. Biol.*, 2001, **8**, (3), pp. 201–219

[39] LANDGRAF C., PANNI S., MONTECCHI-PALAZZI L., *ET AL.*: 'Protein interaction networks by proteome peptide scanning', *PLoS Biol.*, 2004, **2**, (1), pp. 94–103 (e14)

[40] DJORDJEVIC M., SENGUPTA A.M., SHRAIMAN B.I.: 'A biophysical approach to transcription factor binding site discovery', *Genome Res.*, 2003, **13**, (11), pp. 2381–2390

[41] BRANNETTI B., VIA A., CESTRA G., CESARENI G., CITTERICH M.H.: 'SH3-SPOT: an algorithm to predict preferred ligands to different members of the SH3 gene family', *J. Mol. Biol.*, 2000, **298**, (2), pp. 313–328

[42] RAMANI A.K., MARCOTTE E.M.: 'Exploiting the co-evolution of interacting proteins to discover interaction specificity', *J. Mol. Biol.*, 2003, **327**, (1), pp. 273–284

[43] PAWSON T., SCOTT J.D.: 'Signaling through scaffold, anchoring, and adaptor proteins', *Science*, 1997, **278**, (5346), pp. 2075–2080

[44] BURACK W.R., SHAW A.S.: 'Signal transduction: hanging on scaffold', *Curr. Opin. Cell Biol.*, 2000, **12**, (2), pp. 211–216

[45] MORRISON D.K., DAVIS R.J.: 'Regulation of MAP kinase signaling modules by scaffold proteins in mammals', *Annu. Rev. Cell Dev. Biol.*, 2003, **19**, (1), pp. 91–118

[46] MCCLEAN M.N., MODY A., BROACH J.R., RAMANATHAN S.: 'Cross-talk and decision making in MAP kinase pathways', *Nat. Genet.*, 2007, **39**, (3), pp. 409–414

[47] VAN NIMWEGEN E., CRUTCHFIELD J.P., HUYNEN M.: 'Neutral evolution of mutational robustness', *Proc. Natl. Acad. Sci. USA*, 1999, **96**, (17), pp. 9716–9720

[48] CILIBERTI S., MARTIN O.C., WAGNER A.: 'Innovation and robustness in complex regulatory gene networks', *Proc. Natl. Acad. Sci. USA*, 2007, **104**, (34), pp. 13591–13596

[49] EDELMAN G.M., GALLY J.A.: 'Degeneracy and complexity in biological systems', *Proc. Natl. Acad. Sci. USA*, 2001, **98**, (24), pp. 13763–13768

doi: 10.1049/iet-syb:20080076

Supplementary material for C.R. Myers, "Satisfiability, sequence niches, and molecular codes in cellular signaling"

1. Derivation of critical number of crosstalking proteins (random satisfiability bound)

Here we derive the result stated in eq. (1) of the main text, the critical number of crosstalking proteins N_c for a given sequence length L and promiscuity radius R, which we can interpret as a random satisfiability bound for the size of the protein-protein interaction code. A given instance of the SNQ is unsatisfiable if the target volume (i.e., the Hamming sphere of radius R surrounding the target sequence T) is completely covered by the union of the crosstalk volumes (centered about the crosstalk sequences $\{C\}$), a process that is illustrated schematically in the main text in Fig. 3(a). We can estimate the critical number of crosstalk proteins N_c needed to cover the sequence volume of the target protein. For a given binary string of length L, the number of sequences $V(L, R)$ in a ball of Hamming radius R is

$$V(L,R) = \sum_{n=0}^{R} \binom{L}{n} \tag{S1}$$

and the total possible number of sequences $V_0(L)$ is

$$V_0(L) = 2^L \tag{S2}$$

Let q be the ratio of these sequence volumes:

$$q \equiv V/V_0 \tag{S3}$$

We consider depositing at random sequence volumes of size $V(L, R)$ in a space of volume $V_0(L)$. From the binomial distribution, the probability that a given point in sequence space is covered n times after N proteins have been deposited is

$$P_q(n|N) = \binom{N}{n} q^n (1-q)^{N-n} \tag{S4}$$

Therefore the probability $U_q(N)$ that a given point in sequence space is left *uncovered* by N proteins is

$$U_q(N) = P_q(0|N) = (1-q)^N \tag{S5}$$

We can thus estimate the average number of sequences $S_u(V, q, N)$ in the target volume V left uncovered by N proteins to be

$$S_u(V,q,N) = V(1-q)^N \tag{S6}$$

We wish to estimate the critical number of proteins N_c required to cover the target volume; since the sequence space is discrete, we estimate N_c as the number of proteins for which there is $O(1)$ remaining uncovered sequence in the target volume. This yields

$$V(1-q)^{N_c} = 1 \tag{S7}$$

which implies

$$N_c = \frac{\log(1/V)}{\log(1-V/V_0)} \tag{S8}$$

The estimate (S8) appears to adequately describe the SNQ simulation data presented in the main text, as indicated by the scaling collapses shown in Fig. 3 of the main text. We expect the quality of the estimate to degrade, however, as the discrete nature of the sequence space becomes more important, i.e., as the number of sequences in the target volume $V(L,R)$ becomes small (of $O(1)$). Indeed, for the situation $R=0$, where there is only one sequence in the target volume to be covered (namely the target sequence T), the estimate (S8) yields $N_c = 0$. For this case, however, we can independently estimate the number of randomly situated crosstalking sequences required to insure that the target sequence T is covered with probability 1/2:

$$1-(1-q)^{N_c^{R=0}} = 1/2 \implies N_c^{R=0} = \log(1/2)/\log(1-q) = \log(1/2)/\log(1-1/V_0) \tag{S9}$$

The result (S8) assumes an alphabet size $A=2$ (i.e., binary sequences). We can generalize the satisfiability bound in a straightforward manner, if we assume that binding of two sequences continues to be dictated by a maximal Hamming distance, i.e., two sequences s_1 and s_2 will bind if $H(s_1,s_2) \leq R$. In this case, the form of the bound (S8) remains unchanged, and we need simply redefine the relevant sequence volumes corresponding to an alphabet of size A:

$$V(L,R) = V(L,R,A) = \sum_{n=0}^{R} \binom{L}{n} (A-1)^n \tag{S10}$$

$$V_0(L) = V_0(L,A) = A^L \tag{S11}$$

In the case of reverse complement symmetric (RCS) sequences (e.g., for binding of protein to DNA in the regulation of gene transcription), the bound is reduced because each sequence in the target volume can be covered either by a ball centered within Hamming distance R

of the sequence, or by a ball centered within distance R of the reverse complement of that sequence. This has the effect of doubling the coverage ratio q: $q \equiv 2V/V_0$. As a result,

$$N_c^{RCS} = \frac{\log(1/V)}{\log(1 - 2V/V_0)} \tag{S12}$$

which is only valid for $R < L/2$. For $R \geq L/2$, $N_c^{RCS} = 1$.

The main text alludes to a symmetric generalization of the SNQ that asks whether every protein in a collection is distinguishable, that is, whether there is a separate sequence niche for each of N proteins. While we do not have a general estimate for the critical number of proteins N_c for this problem, we can produce such an estimate for the special case of $R = 0$, where crosstalk occurs only if two sequences are exactly the same (no mismatches). In that limit, the question boils down to this: For binary sequences of length L, how many randomly chosen sequences must be chosen for there to be a probability of at least 1/2 that two sequences are identical? This is just the classic "birthday problem" of probability theory, for a system where a "year" contains $V_0 = 2^L$ possible days (see, e.g., `http://en.wikipedia.org/wiki/Birthday_problem`). The probability $p(n)$ that two sequences out of n will match is:

$$p(n) = 1 - \frac{V_0!}{(V_0 - n)!\; V_0{}^n} \tag{S13}$$

so, for a given sequence length L, we can find the number N_c for which this probability exceeds 1/2 to arrive at an estimate for the $R = 0$ bound of the generalized SNQ.

2. Size of the largest solution cluster

Fig. 4(d) of the main text demonstrates that the size S_0 of the largest cluster (solid line) decreases roughly exponentially with crosstalk number N. From the geometric argument illustrated in Fig. 3(a) in the main text, we might expect

$$S_0 \sim (1 - q)^N \approx \exp(-qN) \quad \text{for small q} \tag{S14}$$

where $q \equiv V(L, R)/V_0(L)$. For $L = 16, R = 6$, $q \approx 0.23$, and a fit to the cluster size data in Fig. 4(d) reveals $S_0 \sim \exp(-0.29N)$. The exponential approximation to the power law in eq. (S14) would be more accurate for smaller q, but part of the discrepancy between the predicted and measured decay rate is due to the fact that the geometric argument only describes the elimination of viable sequences by crosstalk proteins, and not the fragmentation of clusters. Some of the decrease in S_0 is due to the latter effect.

3. Review of results from Zarrinpar, Park and Lim

We describe here in slightly more detail the experimental results of ref. [1] (ref. [17] in main text). Zarrinpar *et al.* investigated SH3-mediated signaling in yeast (*Saccharomyces cerevisiae*), probing in particular the signaling pathway involved in a high-osmolarity response, predicated on the interaction of the Sho1 protein (containing an SH3 domain) and the Pbs2 protein (with an exposed proline-rich, PXXP, peptide sequence). Experimentally, they created chimeric versions of the Sho1 protein, replacing the native SH3 domain with each of the other 26 SH3 domains found in yeast. (Three of the Sho1 chimeras were insoluble, however, so they could not be assayed *in vivo*.) They then sought to determine whether any of those domains could reconstitute the function of the high-osmolarity pathway, and found that none of the other yeast domains could so function. *In vitro* peptide binding assays also carried out revealed a similar lack of interaction from any but the Sho1-Pbs2 pair. When SH3 domains from 12 metazoan proteins were tested (both *in vivo* and *in vitro*), however, it was discovered that 6 of those were able to reconstitute the function of the high-osmolarity pathway. Their interpretation was that there has been an evolutionary selection against crosstalk in yeast, whereby domains and peptides have evolved such that the Pbs2 PXXP motif lies in a niche in sequence space where it is recognized by only the Sho1 SH3 domain, as is illustrated schematically in Fig. S.1(a). Since there has been no such selection pressure in other organisms, it was perhaps not surprising that the Pbs2 motif overlaps with the recognition volumes of many of non-yeast SH3 proteins, as is illustrated in Fig. S.1(b).

Zarrinpar *et al.* also sought to characterize the nature of protein-protein interactions in the sequence space surrounding the wild-type Pbs2 motif, which they did by assaying a library of 19 single-base-pair missense mutations to the native yeast Pbs2 motif (leaving the core prolines of the PXXP motif unchanged). While some mutations resulted in increase affinity for Sho1, and some resulted in decreased affinity, all mutations resulted in an increased cross-reactivity with other yeast SH3 domains. This suggests that the wild-type Pbs2 is optimized not for affinity, but for discrimination among different SH3 domains.

4. Methods

To ascertain whether a given instance of the SNQ was satisfiable or not, I implemented the algorithm by Gramm *et al.* [2] ("Algorithm D" in [2], modified as described to treat the Distinguishing String Selection Problem). This is a recursive, backtracking algorithm in the style of Davis-Putnam(DP)-type methods used in the study of other NP-complete problems (e.g., $k-$SAT [3]). Algorithm D in [2] implements heuristics to prune the search tree, tailored to the Distinguishing String Selection Problem (DSSP). DP-type algorithms are known to be significantly slower in practice for $k-$SAT than other algorithms (e.g., WalkSAT [4] or survey propagation [5]), but have the advantage of being *complete*, i.e., able to determine whether any instance is satisfiable or not, given sufficient computer time. (Incomplete algorithms can typically find a solution if there is one, but are not guaranteed to stop if there is no solution.) For forays into a newly-identified NP-complete problem such as this, complete algorithms are a useful first step. For each SNQ instance, it was determined whether the instance was satisfiable, and how long it took to decide that question. Since DP-type methods are recursive, it is conventional to measure algorithm run times in units of number of calls to the recursive core, which is what we have done here.

The SNQ, as stated, applies to any set of sequences T and $\{C\}$. This paper has focused on random instances of the SNQ, where the relevant sequences are sampled uniformly at random from the set of all binary sequences of length L, with equal probabilities of 0 and 1 in the sequences T and $\{C\}$. Simulations of random instances of the SNQ were carried out, for various values of the relevant control parameters: the string length L, the Hamming radius R, and the number of crosstalk proteins N. Average satisfiability and median solution time were computed from 100 random SNQ instances for each set of L, R, and N.

To explore the full solution space of SNQ instances, exhaustive examination was carried out. For each of the possible 2^L sequences, it was determined whether that sequence satisfied the given SNQ. The set of valid solutions was assembled to form an undirected graph, whose nodes were SNQ solutions and whose edges joined nodes with sequences that differed by Hamming distance of 1, i.e., by 1 bit flip. The network analysis package NetworkX [networkx.lanl.gov] was used to compute connected components of the resulting graphs, and to generate layouts for visual display. This work motivated a contribution on my part to the NetworkX source code repository [networkx.lanl.gov/changeset/223], using tuples of index coordinates to label grid graphs, such as would be used to represent an L-dimensional hypercube. This representation is natural for graphs connecting nodes in sequence space. A spring force layout algorithm was used to generate the images in Figs. 4(a)-(c) in the main text, whereby connected nodes are attracted to each other to produce compact representations of connected components. As noted, however, the positions of the graph nodes in Figs. 4(a)-(c) have no intrinsic meaning, as all nodes are vertices on the L-dimensional hypercube. The problem of usefully visualizing complex network structures in high-dimensional sequence spaces is an ongoing challenge in computational biology.

[1] Zarrinpar, A., Park, S.-H., and Lim, W. ‘Optimization of specificity in a cellular protein interaction network by negative selection’. *Nature*, **426**:pp. 676–680, 2003.

[2] Gramm, J., Niedermeier, R., and Rossmanith, P. ‘Fixed-Parameter Algorithms for CLOSEST STRING and Related Problems’. *Algorithmica*, **V37**(1):pp. 25–42, 2003.

[3] Davis, M. and Putnam, H. ‘A Computing Procedure for Quantification Theory’. *J. ACM*, **7**(3):pp. 201–215, 1960.

[4] Selman, B., Kautz, H. A., and Cohen, B. ‘Local Search Strategies for Satisfiability Testing’. In M. Trick and D. S. Johnson, editors, ‘Proceedings of the Second DIMACS Challange on Cliques, Coloring, and Satisfiability’, Providence RI, 1993.

[5] Mezard, M. and Zecchina, R. ‘Random K-satisfiability problem: from an analytic solution to an efficient algorithm’. *Physical Review E*, **66**:p. 056126, 2002.

FIGURES

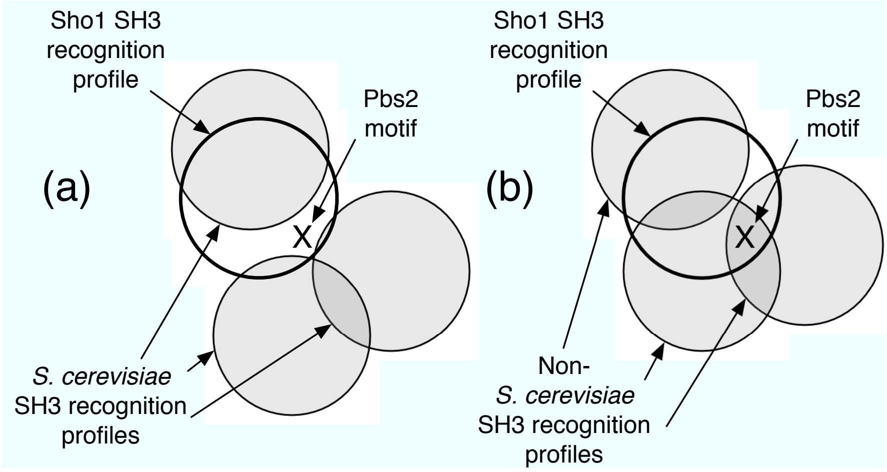
Sho1 SH3 recognition profile
Pbs2 motif
(a)
X
S. cerevisiae SH3 recognition profiles
Sho1 SH3 recognition profile
Pbs2 motif
(b)
X
Non-S. cerevisiae SH3 recognition profiles

FIG. S.1: The interpretation offered by Zarrinpar, Park and Lim to describe (a) the lack of crosstalk among *S. cerevisiae* SH3 domains and (b) the presence of crosstalk among non-*S.cerevisiae* SH3 domains. [Adapted from [1].] (a) In *S. cerevisiae*, evolutionary selection against crosstalk has driven the proline-rich Pbs2 motif to a niche where it is recognized only by the Sho1 SH3 domain. (b) There is no such selection pressure in other organisms, so domains introduced from elsewhere can bind Pbs2.