Macromolecular Crowding as a Determinant of Protein Viability – a Lattice Model Study

Daniel Nilsson

February 21, 2017

Department of Astronomy and Theoretical Physics,
Lund University

LUND
UNIVERSITY

Master thesis supervised by Anders Irbäck

We study the effects of macromolecular crowding with interacting crowders in the simple HP-model of protein folding, by performing Monte Carlo simulations of a single flexible HP protein in the presence of folded rigid crowders. In contrast to the well-known stabilization of steric crowders, we find that interacting crowders usually destabilize the test protein, often significantly so. As such, careful design of the crowder surfaces is necessary, and we find that the size and geometry of single hydrophobic patches are more important than the total amount of hydrophobicity on crowder surfaces. In addition, test proteins are found to be more stable if they have fewer native contacts between hydrophobic and polar residues. We also investigate the effects of crowding on evolutionary processes, finding that incorporating constraints based on crowding noticeably affects the type of fold-switches that are allowed. Finally, the effects of crowding on binding and aggregation were investigated. In "realistic" situations we find that crowding weakly destabilizes both processes.

Contents

1 Introduction

Cells are often called the building blocks of life. As such, understanding the inner workings of cells is important both for practical applications e.g. within medicine and for gaining a deeper understanding of life itself. The interior of a cell is typically filled with a large number of biological macromolecules. One of the most common classes of macromolecules found in cells are proteins. Proteins consist of a polymer chain of so-called amino acids.

Proteins are usually divided into a few different categories, one of the major ones being globular proteins. Globular proteins typically fold to a unique native state, corresponding to the free energy ground state of the protein [1]. The three-dimensional structure of a protein is often important for it to be able to successfully perform its biological function.

There are several different physical forces involved in the folding of proteins. One of the most important driving forces involved in protein folding is the hydrophobic force [2]. The hydrophobic force is an effective force which causes polar and non-polar substances to separate if possible. Since some of the amino acids present in a protein chain are polar while others are non-polar, the non-polar, or hydrophobic, residues tend to form a hydrophobic core which is shielded from the surrounding water.

Because of their importance for the understanding of living cells, proteins have been the subject of a long range of studies, both experimental and computational. Due to the high computational cost associated with detailed simulations of realistic proteins, one approach has been to study simplified models of proteins. One example of such a model is the so-called HP model [3]. In studies of simplified models, it has been common to distinguish "protein-like" polymers from general ones by requiring that they have a unique ground state.

While the existence of a unique ground state is certainly a sensible requirement, it is unlikely to be sufficient for determining the viability of real proteins. For instance, the ground state has to be both kinetically accessible and thermally stable at physical temperatures. Evolutionary dynamics can also restrict the set of viable sequences to contain only sequences with high mutational tolerance.

Another restriction on the set of viable proteins, which has been garnering an increasing amount of attention recently, is the effects of macromolecular crowding. This effect refers to the fact that the cell interior consists of a large number of macromolecules, and that viable proteins must be able to coexist

both with each other and with other types of molecules. While recent years have seen a number of experimental and computational studies of the effects of crowding [4–6], many of them have been performed using realistic models and so suffer from high computational costs. Attempts to use simplified models, where the surrounding molecules are represented as hard spheres, to explain crowding phenomena have also been common [7]. These models will, however, almost invariably predict a stabilization of the protein folds, while experiments show that both stabilization and destabilization can occur [5].

The results presented in this thesis represent an attempt to study the impact of macromolecular crowding with interacting crowders in the simple HP-model. We focus on the ability of proteins to fold in the presence of crowders, although other aspects such as binding and aggregation are also considered.

2 Model and Methods

2.1 The HP-model

Theoretical and computational studies of protein folding are faced with the challenge of having to consider systems with an often insurmountable number of degrees of freedom. This frequently makes all-atom studies of real proteins infeasible. In order to get a better insight into the forces which govern protein folding, it is often useful to consider models with a simplified representation of both the protein chain and the forces which stabilize its fold. One minimalist model which has often been used for this purpose is the hydrophobic-polar (HP) model [3].

As for real proteins, an HP protein is represented as a sequence of amino acids. Unlike real proteins, HP proteins consist of only two amino acid types, hydrophobic (H) and polar (P).

Geometrically, the chain is represented as a self-avoiding walk on a lattice. Throughout this work a two-dimensional square lattice is used. Each amino acid occupies a single lattice point, and each lattice point can contain at most one amino acid. Amino acids which are neighbours along the chain are also restricted to occupying adjacent lattice points.

The energy of a given conformation is determined as follows. All energy contribution comes from pairs of amino acids which are adjacent on the lattice but not along the chain. In the most common version of the HP model, the

contribution of each such pair is given by

$$E = \begin{cases} -\epsilon & \text{if both amino acids are hydrophobic} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where ϵ is an arbitrary positive energy scale. For simplicity, we will use units in which $\epsilon = 1$.

Typically, we consider the behaviour of an HP protein at a given finite temperature T. As stipulated by elementary statistical mechanics, the probability of finding a protein, or a system of proteins, in a given configuration $\mathbf{r}$, is Boltzmann-distributed,

$$P(\mathbf{r}) = \frac{1}{Z} e^{-\beta E(\mathbf{r})}, \tag{2}$$

with $\beta = 1/kT$, k the Boltzmann constant, and the partition function, $Z = \sum_{\mathbf{r}} e^{-\beta E(\mathbf{r})}$, is a normalizing factor.

HP-chains have often been characterized as protein-like or not depending on whether they have a unique ground state, which is then considered the native state of the protein. All such "protein-like" chains have previously been determined for sequence lengths up to $N = 30$, with full energy histograms for $N = 27$ [8]. Unless otherwise mentioned, we will be using sequences of length 27 throughout this work.

The model as described above has typically been used only for systems containing a single protein. The generalization to systems of several proteins should be mostly straightforward, with the observation that the boundary conditions of the lattice must be specified. In this work we use a lattice with periodic boundary conditions. In order to ensure that the problem remains computationally feasible, we typically allow conformational updates only for a single "test" protein, keeping the crowder structures fixed.

2.2 Methods

In order to study the thermodynamic behaviour of HP proteins, we need to calculate the Boltzmann probabilities (2) for all conformations. Analytically, this is only feasible for very short sequences. One possible way to circumvent this restriction is to use Monte Carlo simulations to sample the correct distributions.

Monte Carlo methods are a wide range of methods based on the use of random numbers. For the type of simulations used in this work, algorithms

such as the Metropolis algorithm [9] or variants thereof are the most useful ones.

The Metropolis algorithm works by realizing a Markov chain on the set of all conformations. A Markov chain is a discrete time stochastic process where the state at a given time t is sampled from a probability distribution depending only on the state of the system at time $(t-1)$. The dynamics of the Markov chain can be specified using a so-called transition probability,

$$W(\mathbf{r} \to \mathbf{r}') = P(\mathbf{r}'; t|\mathbf{r}; t-1), \tag{3}$$

which provides the conditional probability of visiting state $\mathbf{r}'$ given that the last visited state was $\mathbf{r}$.

The probability distribution of a Markov process at time t will converge to the correct probability distribution if two conditions hold:

1. The chain is ergodic, meaning that each state can be reached from any other state.

2. The condition of detailed balance holds.

Note that the second condition is sufficient but not necessary.

2.2.1 Detailed Balance

The condition of detailed balance states that, when the process has relaxed to the equilibrium distribution, the transition probability, $W(\mathbf{r} \to \mathbf{r}')$, fulfils

$$W(\mathbf{r} \to \mathbf{r}')P(\mathbf{r}) = W(\mathbf{r}' \to \mathbf{r})P(\mathbf{r}'). \tag{4}$$

In other words the probability of moving from state $\mathbf{r}$ to $\mathbf{r}'$ at a given time step is equal to the probability of moving from state $\mathbf{r}'$ to state $\mathbf{r}$.

In the Metropolis algorithm, the transition probabilities are usually split into proposal probabilities $F(\mathbf{r} \to \mathbf{r}')$, and acceptance probabilities $A(\mathbf{r} \to \mathbf{r}')$,

$$W(\mathbf{r} \to \mathbf{r}') = F(\mathbf{r} \to \mathbf{r}')A(\mathbf{r} \to \mathbf{r}'), \qquad \mathbf{r} \neq \mathbf{r}'. \tag{5}$$

As the name indicates, the proposal probabilities are used to propose an update. They can in principle be chosen freely. This update is then accepted

with a probability $A(\mathbf{r} \to \mathbf{r}')$, which is chosen so as to preserve the condition of detailed balance,

$$A(\mathbf{r} \to \mathbf{r}') = \min\left(1, \frac{F(\mathbf{r}' \to \mathbf{r})}{F(\mathbf{r} \to \mathbf{r}')} \frac{P(\mathbf{r}')}{P(\mathbf{r})}\right). \tag{6}$$

If the update is rejected, the process will remain in the same state during the next time step.

For the case when the proposal probabilities are symmetric, $F(\mathbf{r} \to \mathbf{r}') = F(\mathbf{r}' \to \mathbf{r})$, and the sampled probability distribution is a Boltzmann distribution, the acceptance probabilities (6) take the simple form

$$A(\mathbf{r} \to \mathbf{r}') = \min\left(1, e^{-\beta \Delta E}\right) \tag{7}$$

where $\Delta E = E(\mathbf{r}') - E(\mathbf{r})$.

2.2.2 Proposal Probabilities

The proposal updates used in the work leading up to this thesis can be split into three categories: local conformational updates, global "pivot" conformational updates, and rigid-body updates of whole chains.

Local Updates

Local updates refer to updates where only a limited number of residue positions, located next to each other along the chain, are updated. In this work, we use single- and double-residue updates, where the residues are chosen with a uniform distribution. The chosen residues are then moved to new random positions, compatible with their neighbours along the chain. An illustration of the available moves for internal residues can be seen in figure 1. For beads at the end of the chain, more moves are available. If the move results in a collision with a distant section of the chain, or with a crowder molecule, the update is immediately rejected.

Pivot Updates

In contrast to the local update, pivot updates lead to large-scale global deformations of the protein chain. This is done by randomly choosing one residue as the "pivot point." The part of the chain stretching from one end of the chain to the pivot point is kept fixed while the other part is either rotated or reflected in some plane, so that its direction relative to the rest of the system is changed.

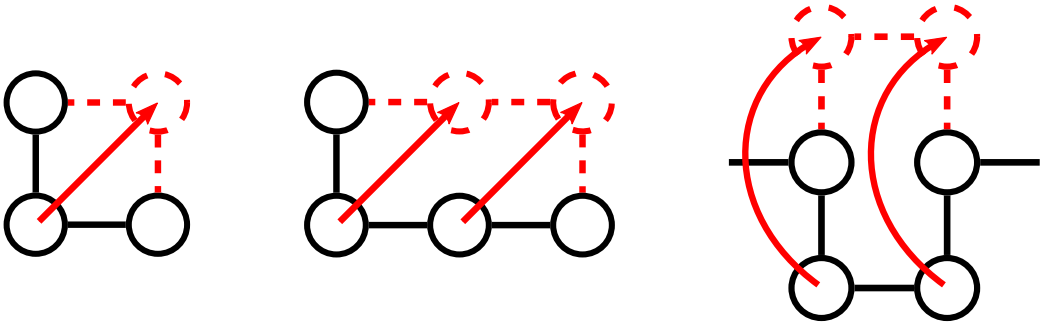

Figure 1: Illustration of the local updates used in the algorithm.

Rigid-Body Updates

Rigid-body updates refer to updates of the relative position of the test protein and crowder molecules, without conformational change. There are a couple of different schemes that can be used to perform these updates. The most simple method is to simply use a Metropolis style update where the suggested updates are rigid-body translations, rotations or reflections of individual molecules. This method is not necessarily suitable for a crowded environment however, since the close contact with other proteins will block many of the suggested moves.

One way to circumvent this is to build clusters of nearby proteins and update all of these at the same time. In its most simple version, a cluster can be built simply by picking a random protein and adding all proteins with at least one H-H contact to any protein in the cluster. In order to preserve detailed balance in the updates, the cluster must be the same before and after an update, so the update must not induce any energetic change.

Further improvement can be achieved by using a probabilistic cluster algorithm, such as the Swendsen-Wang cluster algorithm [10,11]. The Swendsen-Wang algorithm proceeds as follows:

1. Choose a random protein i, add it to the cluster.

2. For each protein j, which is in contact with protein i, add it to the cluster with probability $1 - e^{\beta E_{ij}}$, where E_{ij} is the energy involved in the interaction between the two proteins.

3. For each protein added to the cluster, continue adding proteins as per step 2.

For the Swendsen-Wang update, no acceptance update is needed, since detailed balance is already built into the method. To see this, consider two states differing only by the position/orientation of a single cluster. Since the internal structure of this cluster is the same in both cases, the probability of adding these proteins to the cluster, P_{cluster}, is the same in both cases. The probability of choosing exactly that cluster in the two cases therefore only differs due to contacts with proteins not part of the cluster.

For protein i, the probability of not being added to the cluster through protein j (in the cluster) is given by $e^{\beta E_{ij}}$, giving a total probability of not being added to the cluster of

$$\prod_{j \text{ in cluster}} e^{\beta E_{ij}} = e^{\beta \sum_{j \text{ in cluster}} E_{ij}}. \tag{8}$$

We can now write the total probability of building the required cluster as $P_{\text{cluster}} e^{\beta \sum_{j \text{ in cluster}} E_{ij}}$, while the Boltzmann probability of the state is given by

$$\frac{1}{Z} \exp\left(-\beta \sum_{i,j \text{ in cluster}} E_{ij} - \beta \sum_{\substack{i,j \text{ not in} \\ \text{cluster}}} E_{ij} - \beta \sum_{\substack{i \text{ in cluster} \\ j \text{ not in} \\ \text{cluster}}} E_{ij} \right), \tag{9}$$

giving

$$W(\mathbf{r} \to \mathbf{r}')P(\mathbf{r}) = \frac{1}{Z} \exp\left(-\beta \sum_{i,j \text{ in cluster}} E_{ij} - \beta \sum_{\substack{i,j \text{ not in} \\ \text{cluster}}} E_{ij} \right) P_{\text{cluster}}, \tag{10}$$

which is the same for two states only differing by the location of the cluster in question.

Since it is still possible to perform single-molecule updates in the Swendsen-Wang algorithm, the system is also ergodic.

In order to maximize the efficiency of the simulations, we performed test simulations using all three methods. Since the crowding environment can have a significant effect on the efficiency, we perform simulations in ten different homogeneous crowding environments. The choice of test protein should be less important, and we therefore use only the crowding protein with the highest melting temperature. For each crowding environment, we perform 10

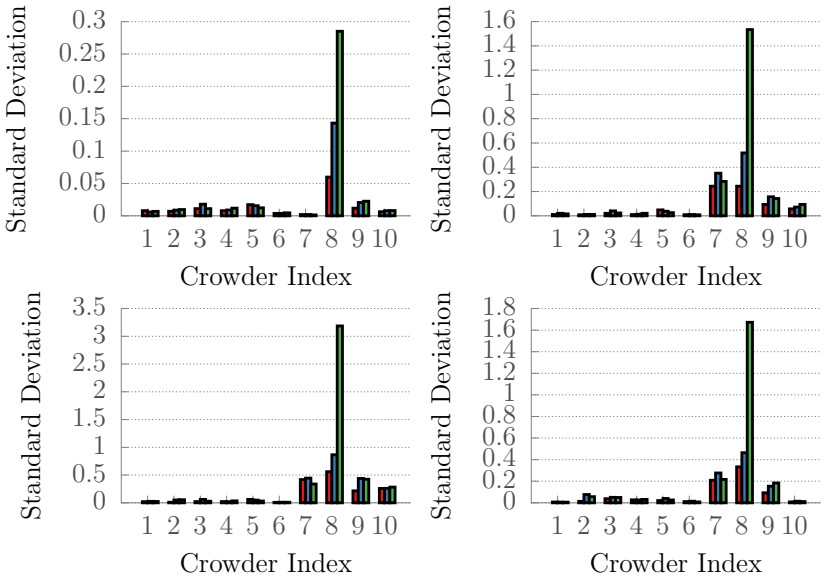

Figure 2: The estimated standard deviation of various observables when using different types of rigid-body updates, in ten different crowding environments. The different subfigures show the variances of the native probability (top left), the test protein energy (top right), the test protein-crowder interaction energy (bottom left), the interaction energy between the crowders (bottom right). The red bars show results with the Swendsen-Wang algorithm, the blue with the simple cluster algorithm, and the green with the single-molecule updates. The standard deviations when using the different algorithms are comparable in size for all but one of the environments, where the single-molecule update performs significantly worse. Information on the crowder indices can be found in the appendix.

simulations. Note that this comparison was performed using the simulated tempering algorithm, described below.

Figure 2 shows how the standard deviations of four different observables vary between the different methods. We see that for most of the environments, the choice of simulation method is largely irrelevant, though on average the Swendsen-Wang algorithm seems to perform slightly better. Mainly, we find that for one of the environments, the naïve method, and to a lesser extent the simple cluster method, perform significantly worse. Plots showing examples of the runtime evolution of the three different methods for this system can be seen in figure 3. The main cause of this decreased performance appears to be that the hydrophobic residues on these crowders are distributed in such a way that they are able to form strong bonds with sev-

eral other crowders at the same time. This causes the proteins to form large aggregates, which apparently slows down the simulations considerably when using the simpler methods.

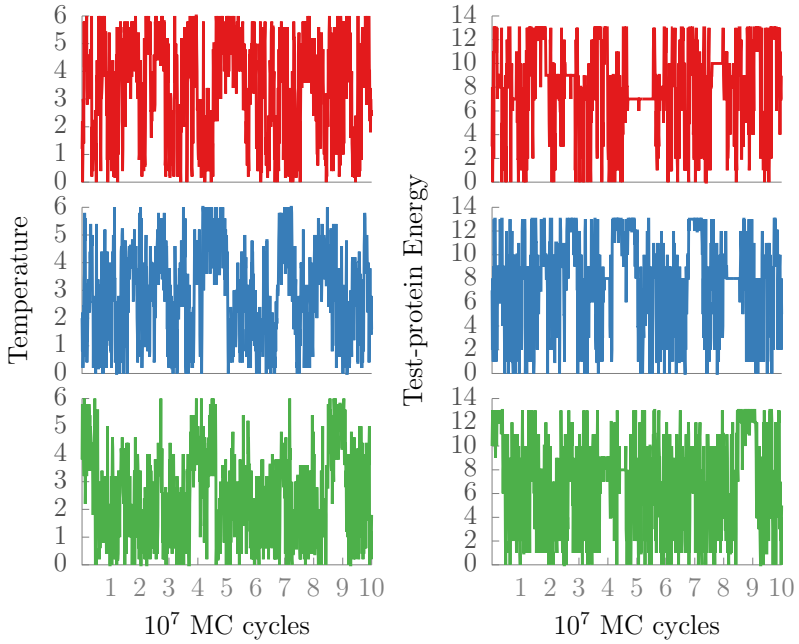

Figure 3: The runtime development of temperature and test-protein energy in simulations with the three rigid-body update types in the environment which is most difficult to sample. The red curves (top) are from a simulation with the Swendsen-Wang algorithm, the blue (middle) with the simple cluster algorithm, and the green (bottom) with the single-molecule updates. As can be seen, the Swendsen-Wang covers the allowed range of values more effectively than the other algorithms, especially for the temperature.

In order to ensure that the calculations performed are accurate for as large a set of crowding environments as possible, we use the Swendsen-Wang method in all further simulations. It should be noted, that in most of the systems of interest, we do not expect the proteins to aggregate, and as such the methods probably give comparable results for most of the simulations. It should also be mentioned that the simulation time of the cluster algorithms is typically slightly longer (up to a factor of 2) than the single-molecule update.

2.2.3 Simulated Tempering

There are various methods devised for improving the performance of the Metropolis algorithm. One commonly used scheme is to expand the ensemble

sampled. In this work we use simulated tempering [12, 13], which expands the ensemble by making the temperature a dynamic variable. Thus, instead of sampling the Boltzmann distribution, $P(\mathbf{r}) = e^{-\beta E(\mathbf{r})}/Z(\beta)$, we sample the joint probability distribution

$$P(\mathbf{r}, \beta) = \frac{1}{\Xi} e^{-\beta E(\mathbf{r}) + g(\beta)}, \tag{11}$$

with $\Xi = \sum_{\mathbf{r},\beta} e^{-\beta E(\mathbf{r})+g(\beta)} = \sum_{\beta} Z(\beta) e^{g(\beta)}$. This form of the probabilities ensures that the conditional probability distribution, $P(\mathbf{r}|\beta)$, is the Boltzmann distribution at inverse temperature β.

The free parameters $g(\beta)$ determine the marginal probability of visiting a given temperature,

$$P(\beta) = \frac{e^{g(\beta)}}{\Xi} \sum_{\mathbf{r}} e^{-\beta E(\mathbf{r})} = \frac{Z(\beta) e^{g(\beta)}}{\Xi}. \tag{12}$$

To ensure a high acceptance probability for updates of the temperature, $g(\beta)$ must be carefully chosen. A natural, although not necessarily optimal, choice is $g(\beta) = -\log Z(\beta)$, giving a uniform distribution in β.

Determining the partition function analytically is typically not possible, so the above choice is normally made using an iterative procedure. Taking the logarithm of equation 12, and noting that $Z(\beta)$ is independent of the choice of $g(\beta)$, we find that we can relate the probability distributions for two choices of $g(\beta)$,

$$\log P(\beta) - g(\beta) + \log \Xi = \log \tilde{P}(\beta) - \tilde{g}(\beta) + \log \tilde{\Xi}. \tag{13}$$

For the case when $\tilde{P} = \text{constant}$, we get

$$\tilde{g}(\beta) = g(\beta) - \log P(\beta) + \text{constant}. \tag{14}$$

Since the constant term does not affect the probability distribution, updating the probability distribution according to

$$g(\beta) \to \tilde{g}(\beta) = g(\beta) - \log P(\beta) \tag{15}$$

will give the correct probability distribution. Note that, since the sampling usually does not yield exact results, this procedure is typically repeated until the simulations are observed to yield an approximately flat temperature distribution.

2.3 Observables

In order to measure the relationship between sequence, thermodynamic and mutational stability, and crowding properties, we use a variety of different observables for each.

For thermodynamic stability, a natural measure is the melting temperature, T_{melt}, defined as the temperature at which the probability of finding the protein in its native state is 1/2. If we consider a system of proteins existing at a given temperature, it might be more useful to consider the probability of finding the protein in its native state at that temperature as a measure of thermodynamic stability.

For mutational stability, the natural measure would be to determine the number of neutral mutations that a sequence can tolerate. A neutral mutation is defined as a mutation which leaves the native conformation of the protein unchanged. The set of all proteins with the same fold is called a neutral set, and a connected component of such a set is called a neutral net. The sequence which can tolerate the largest number of neutral mutations in a neutral set is called the prototype sequence. If there are several sequences with the same number of neutral mutations, the prototype sequence is the sequence with the smallest average Hamming distance to other proteins in the set. (The Hamming distance between two sequences is defined as the minimum number of point mutations required to transform one to the other.) If we restrict ourselves to the set of all prototype sequences, the size of the neutral set, also called the designability, can also be used as a measure of mutational stability. This measure has the advantage of not only considering single-point mutations [14].

When considering the effects of crowding on thermodynamic stability, quantifying the effect using a single number is not entirely straightforward. When comparing the effects of varying crowding environments, the native probability, P_{nat}, constitutes an informative choice. We define P_{nat} as being the probability of the protein being folded to its native state. Usually we compare this probability at a single temperature, e.g. the temperature where $P_{\text{nat}} = 80\%$ without crowding.

When comparing different test proteins, the problem becomes somewhat more difficult, since care has to be taken so that the measure is largely independent of the melting temperature. To achieve this, the test protein is considered as being either folded (to its native state), or unfolded (i.e. in any other state). To each of these two states, we can associate a free energy

F. The probability of finding the protein in its native state is then

$$P_{\text{folded}} = \frac{e^{-\beta F_{\text{folded}}}}{e^{-\beta F_{\text{folded}}} + e^{-\beta F_{\text{unfolded}}}} = \frac{1}{1 + e^{\beta \Delta F}}, \tag{16}$$

where the free energy difference $\Delta F = F_{\text{folded}} - F_{\text{unfolded}}$. This allows us to calculate the free energy difference, which provides a measure of how likely the ground state is to be populated,

$$\Delta F = \frac{1}{\beta} \log \frac{1 - P_{\text{folded}}}{P_{\text{folded}}}. \tag{17}$$

The free energy difference is dependent on the surroundings of the protein, with a positive change corresponding to destabilization and a negative one to stabilization. Thus the change in the free-energy difference,

$$\begin{aligned} \Delta(\Delta F) &= \Delta F^{(\text{crowded})} - \Delta F^{(\text{uncrowded})} \\ &= \frac{1}{\beta} \left(\log \frac{1 - P^{(\text{crowded})}_{\text{folded}}}{P^{(\text{crowded})}_{\text{folded}}} - \log \frac{1 - P^{(\text{uncrowded})}_{\text{folded}}}{P^{(\text{uncrowded})}_{\text{folded}}} \right), \end{aligned} \tag{18}$$

can be used to provide a measure of the effects of crowding on protein stability.

3 Results

3.1 Effects of Crowding on Test Protein Stability

We begin our study by considering the effect of crowding by random proteins. To this end, we select a set of 30 random proteins of length $N = 27$, which are to be used as test proteins. In addition we generate 20 random crowding environments, each consisting of 10 proteins of length $N = 27$. All proteins were selected from the set of sequences folding to a unique ground state. Figure 4 shows some examples of the different behaviours which can arise from these simulations. As can be seen, the presence of crowders can have a wide range of different effects on the test proteins, both stabilizing and destabilizing.

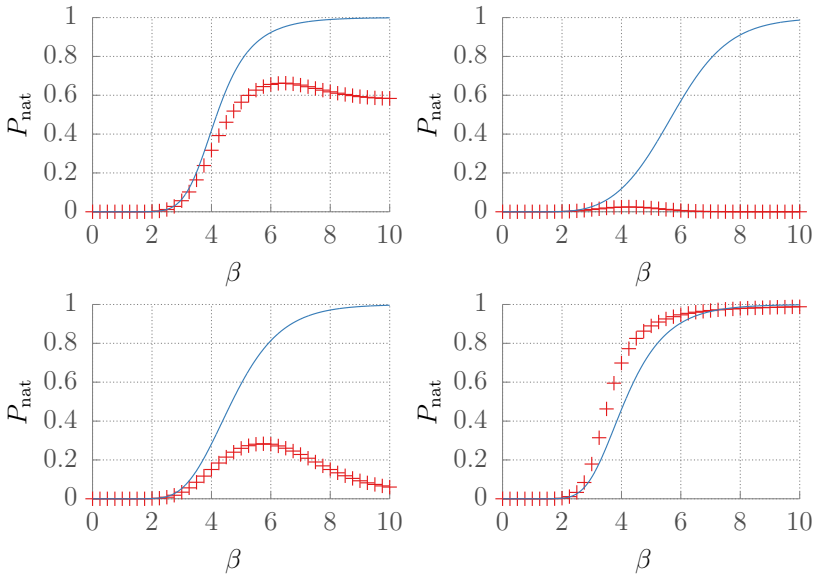

Figure 4: Some examples of how crowding affects the native stability of various proteins. The results in a crowded environment (red crosses) is compared to the behaviour in the absence of crowders (blue lines). The effects of crowding range from stabilization to almost complete destabilization of the native fold. Note that the different behaviours are not equally frequent. Statistical errors are estimated to be at most the same size as the points, usually smaller.

In order to be able to systematically compare the behaviour of all the test proteins and crowding environments investigated, we consider the be-

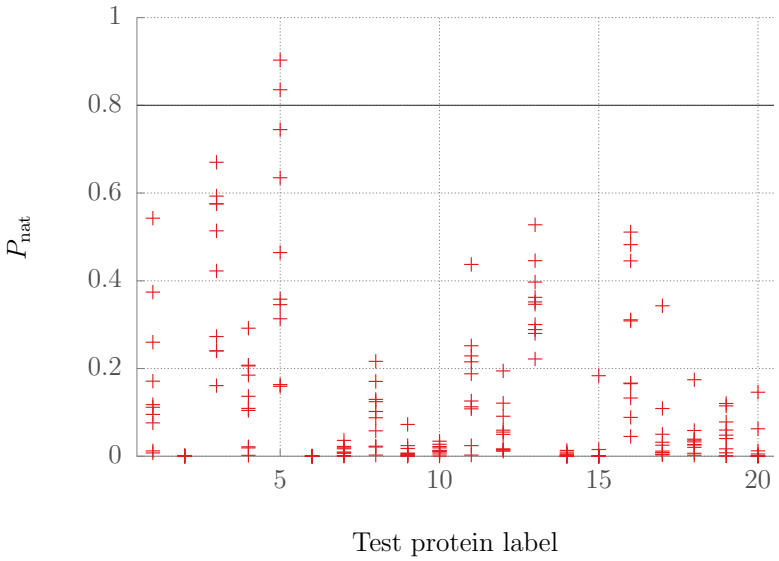

Figure 5: The native probability of a variety of random test proteins in a number of different random crowding environments. Each protein was simulated at the temperature at which the native probability equals 80% in the absence of crowders. Crowding is typically destabilizing, and often to such a large degree that it would likely impact on biological functions. For a key to the test protein labels, please refer to the appendix.

haviour of each test protein at the temperature where $P_{\text{nat}} = 80\%$ in an uncrowded environment. The choice of $P_{\text{nat}} = 80\%$ ensures that the HP proteins are marginally stable, as they typically are in biological systems. The exact choice of temperature does not impact the conclusions drawn. At this temperature, we determine the native probability for each crowding environment. As seen in figure 5, the most common outcome is a significant destabilization of the test protein.

The destabilization observed in figures 4 and 5 indicates that most of the test protein/crowding environment combinations constitute poor approximations of the contents of a living cell. Their melting temperatures will typically be so shifted that they could not fold at "physical" temperatures, if they do fold to their native structures at all. There are two possible ways in which this can be alleviated

1. Mutational pressure could ensure that sequences which are highly destructive and/or very sensitive to crowding are disfavoured.

2. The destabilizing interactions could be "insulated" by some type of specific molecules.

We start by considering the second possibility. One way in which this could potentially happen, would be by adding a number of very small peptides. Because of their sizes they would hopefully be able to bind non-specifically to a diverse set of hydrophobic surfaces.

To test this, we added small crowder particles to the systems investigated. A few different crowder geometries were considered, in this report results will be shown for a three-residue peptide where the central peptide is hydrophobic and he two inter-chain bonds are kept fixed at right angles to each other. Other types of short peptides tested give similar results. As can be seen in figure 6, this addition results in a slight increase in the number of surviving proteins. Nevertheless, the typical result of adding crowders remains a significant destabilization.

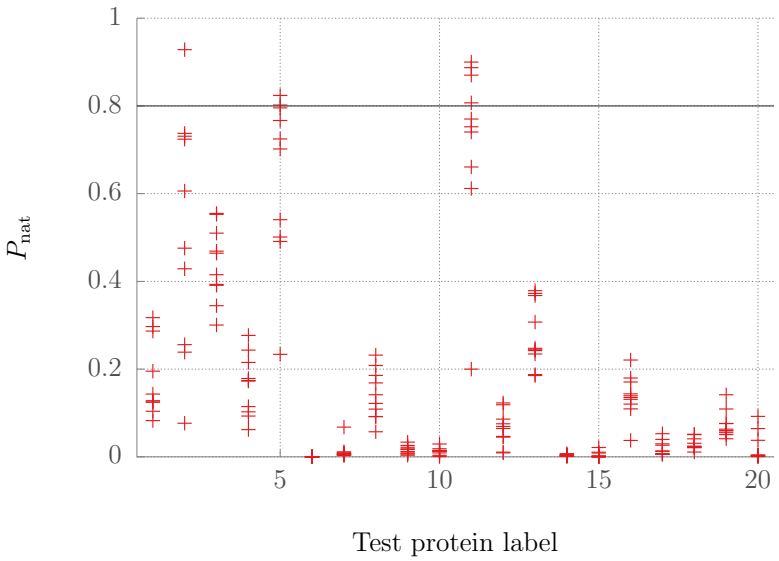

Figure 6: The native probability of a variety of random test proteins in a number of different random crowding environments, where small peptides have been added in the hope that they would shield the test protein from some of the destabilizing effects. Each protein was simulated at the temperature at which the native probability equals 80% in the absence of crowders. Although there are more stable proteins than in figure 5, the typical behaviour remains significantly destabilizing. For a key to the test protein labels, please refer to the appendix.

While the above results are not enough to discount the possibility that there exists some molecule which can prevent the destabilization seen, they are a strong indication that finding a single molecule for this function will be difficult. In order to prevent the destabilization of a particular crowder, it is of course often easy to find another protein which can bind to the destabilizing region(s) of the protein, thus forming a dimer. In order to make the crowding environment as a whole functional however, this would require the design of a particular binding partner for each crowder contributing to the destabilization.

We conclude that the crowding environment has a significant effect on the conformational stability of HP model proteins, and that the destructive effects of crowding must probably be kept in check by careful selection of proteins based both on the destabilizing effect they have on other proteins, and on the effects that are inflicted on them by other proteins.

3.2 Correlations between Crowder Destructiveness and Mutational or Thermodynamic Stability

Of course, crowding is not the only factor restricting the set of allowed sequences. As has been mentioned, both mutational and thermodynamic stability play a role. It has previously been shown that there is a correlation between the mutational and thermodynamic stability of HP model proteins [8]. Furthermore, there are reasons to believe that thermodynamically and mutationally stable proteins are more viable as crowders. For instance, highly stable proteins are characterized by the formation of a hydrophobic core with a mainly polar surface.

In this section, we consider only the prototype sequences for each structure. As a measure of thermodynamic stability, we use the (inverse) melting temperature, while the designability of the structure will be used to quantify the mutational stability (see section 2.3). Figure 7 shows the relation between these two variables. The points are coloured according to the number of hydrophobic residues on the surface. Since these residues can interact favourably with the residues in the core of a test protein, this measure can potentially be used as an approximate measure of the destructiveness of the crowder.

We note that a high designability guarantees a high melting temperature and a relatively low number of hydrophobic surface residues. The converse

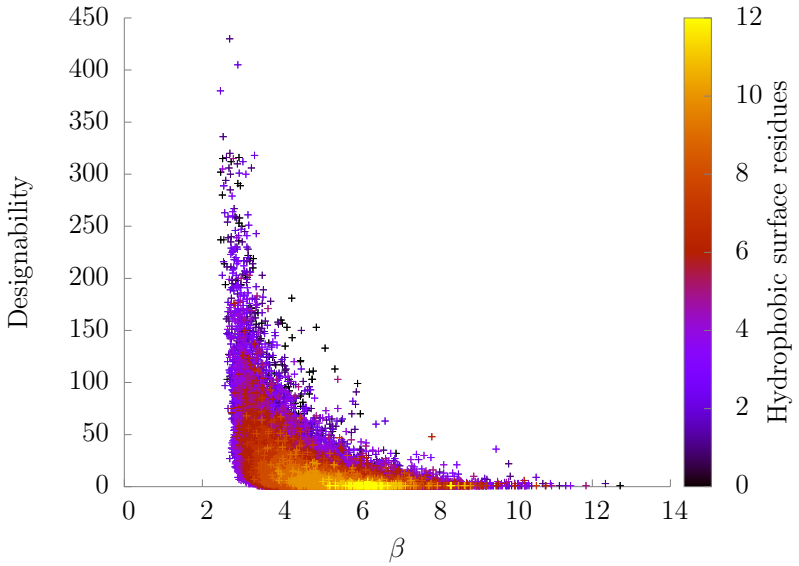

Figure 7: The relationship between inverse melting temperature, designability and number of hydrophobic surface residues for all prototype sequences of length 27. We note that high designability results in favourable values for the other two variables as well, while the reverse is not necessarily true. Also note that, due to the large number of points, the points corresponding to higher surface hydrophobicity are displayed in front of the others.

is not true; thermodynamically stable sequences with a highly polar surface may still be sensitive to mutations. Further, the sequences with the most hydrophobic surface tend to have low designability and a low but not minimal melting temperature.

In order to more directly probe the effects of mutational and thermodynamic stability on crowding properties, we also performed simulations to determine the crowder properties of proteins with different designability and melting temperature. In order to achieve this, we ordered the set of all designing sequences by either designability or melting temperature. Then we selected groups of proteins from five different regions of this list, corresponding to the top 1%, the bottom 1% and three groups with about 1% spread at the first, second and third quartiles. In total, ten groups were thus created, five based on designability, and five based on melting temperature. Due to the granularity of the designability, the two sets with the lowest mutational stability both end up being structures with designability 1. A summary of the properties of the groups can be seen in table 1. Once a group of crow-

Table 1: The values of the designability and melting β in the ten groups of sequences from which the crowding environments were constructed, as described in this section.

Quartile	Designability	Inverse melting temperature
0	1	7.2760822-12.71400
1	1	5.1926214-5.2243132
2	4	4.5676071-4.5899949
3	11	4.0122822-4.0360043
4	83-430	2.4651168-3.0751005

ders was selected, five random sets of ten crowders were chosen from each group. Ten test-proteins were used, these were chosen from the set of the most designable proteins.

The simulations were then performed similarly to those with random protein sequences. The measure of stability was again the native probability at the temperature where $P_{\text{nat}} = 80\%$ for uncrowded proteins. The results can be seen in figures 8 and 9.

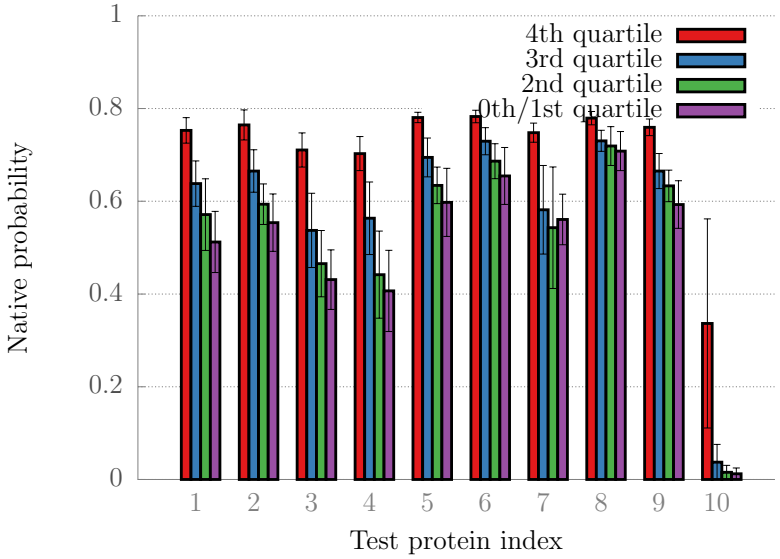

Figure 8: The native probability of ten different test proteins in crowding environments with crowders of varying designability. For each crowder designability, ten randomly chosen environments were used. The more designable crowding environments tend to have less of an effect on the test proteins. Note that error bars indicate standard deviations rather than errors of the mean.

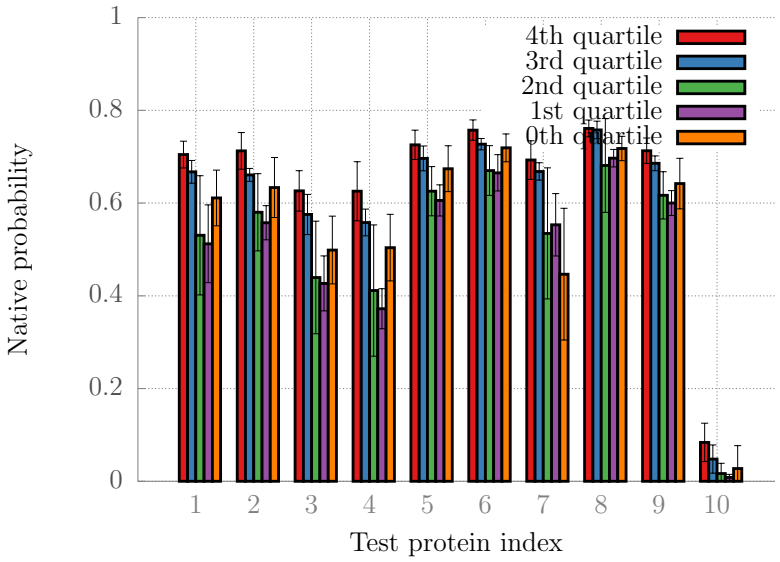

Figure 9: The native probability of ten different test proteins in crowding environments with crowders of varying melting temperature. For each melting temperature, ten randomly chosen environments were used. The correlation with crowder destructivity seems to be less marked than in figure 8. Note that error bars indicate standard deviations rather than errors of the mean.

The results show that while good crowder properties are promoted by both high mutational and high thermodynamic stability, the correlation with mutational stability is somewhat stronger. Of interest is also that sequences with very low melting temperature seem to have better crowder properties than those with more moderate melting temperatures. This is in qualitative agreement with the results of figure 7, where the sequences with really low melting temperature appear to have fewer hydrophobic surface residues.

In conclusion, we find that proteins with otherwise favourable properties tend to be more viable as crowder proteins. This tendency also seems to be in good agreement with the naïve expectation that crowder destructiveness is dependent on the amount of hydrophobic residues on the surface of the test protein.

3.3 Influence of Crowder Destructiveness on the Viablity of Proteins

3.3.1 Relations Between Surface Properties and Crowder Destructiveness

We now turn more directly to the question of how the properties of a protein determine its viability as part of a crowded environment. In this section, we will consider how the surface of a protein influences the destructiveness of the crowder, i.e. how much it will denature the surrounding proteins. (Naturally, the intrior of a folded protein cannot influence its surroundings.) A naïve guess would be that the destructiveness of a crowder is correlated with the number of hydrophobic residues on the surface, since these are the ones able to interact favourably with the core of the test protein.

In order to investigate what factors determine the effects of crowders on conformational stability, we considered crowding in a set of homogeneous crowding environments. The structure of the crowders was kept the same in all simulations, while the sequence was allowed to vary. The chosen crowder structure can be seen in figure 11. It was chosen because of its high designability, and because its surface is geometrically diverse. The chosen structure had a total of 122 allowed surface configurations. Simulations were performed with four different test proteins, randomly chosen from among the set of highly designable proteins used in section 3.2. Since the results were all similar, we only show results for two of them. As in the earlier sections, we determine the thermal stability of the test protein at the temperature where $P_{\mathrm{nat}} = 80\%$ in the absence of crowders.

Figure 10 shows how the native stability of the test proteins varies with the number of hydrophobic surface residues per crowder. As expected, the average native stability decreases with the number of hydrophobic residues on the crowder surfaces. On the other hand, individual data points show considerable deviation from the averages. For one of the test proteins, the crowder environments seem to form two groups, with relatively similar stabilities within each group. The average trend deems to be mainly driven by the fact that more of the highly hydrophobic environments fall in the more destructive group. For the other test protein the destabilization is more gradual, but even here, there is significant variation among residues with similar surface hydrophobicity. These observations indicate that the degree to which destabilization happens is highly dependent on the geometry of the surface around hydrophobic patches.

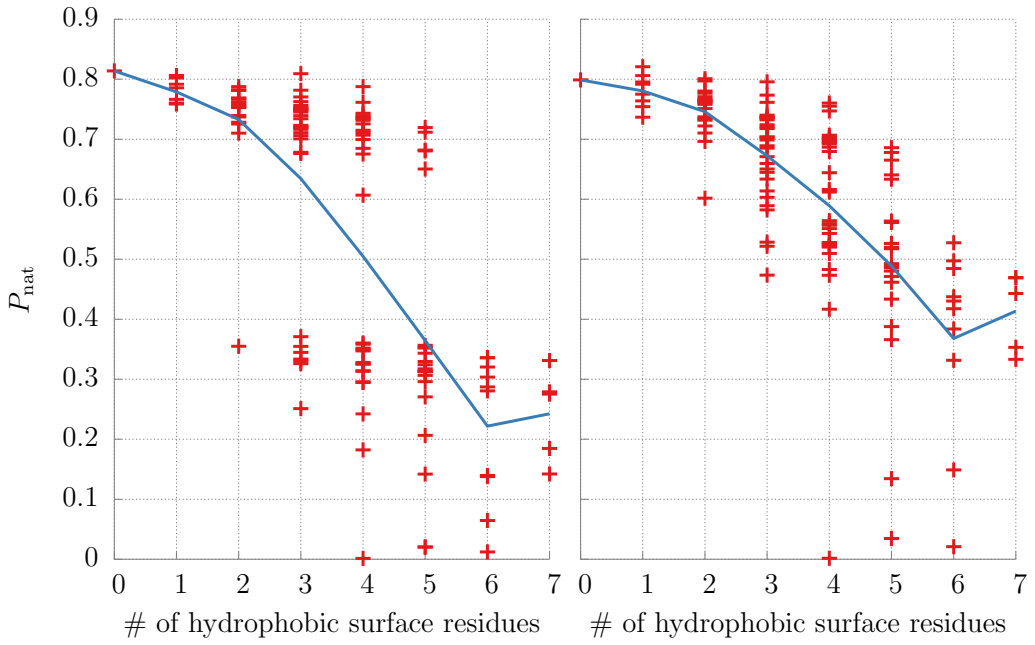

Figure 10: Native stability of the test protein as a function of surface hydrophobicity on the crowders. The two plots show the behaviour for two different test proteins. While the average destabilization (blue lines) increases with increasing surface hydrophobicity, the variation for proteins with similar levels of surface hydrophobicity is often larger than the average change.

To further test how geometric considerations influence crowder destructiveness, we compare crowder sequences differing only in a single surface residue. Figure 12 shows how the stability changes due to such "point flips". As can be seen, the site which is flipped has a significant influence on how much the destructiveness changes. Some sites are highly destructive while others actually increase the stability. We also note that while the destructiveness of any given position varies depending on the test protein, positions which are destructive for one test protein tend to be so also for the other.

We can also note that for some positions, most notably the ones labelled 3 and 4, the degree to which a protein is destabilized tends to fall within one of two groups. Which group a flip ends up in depends on whether the other of these two residues is hydrophobic or not. In other words, we see considerable destabilization if both residues are hydrophobic, but not so much if only one of them is.

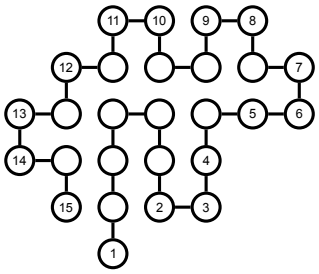

Figure 11: The structure used as crowder in simulations aimed at determining the determining factors of crowder destructiveness. Numbers shown on the surface residues serve as keys for figure 12.

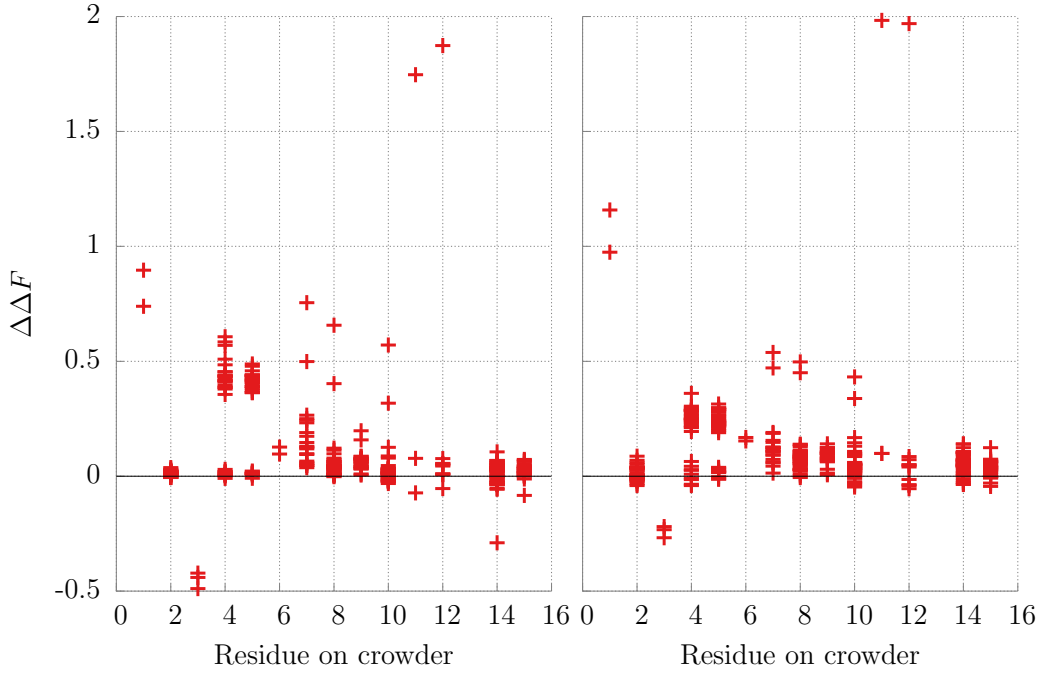

Figure 12: The effects of point mutations on the crowder surface in a homogeneous environment, on the test protein stability. The x-axis shows the position of the changed residue, while the y-axis shows the change in free-energy difference between the folded and unfolded states. As can be seen, the sensitivity to a mutation depends on the position of the mutated amino acid.

In summary, the results of this section imply that the geometry of the hydrophobic patches on a crowder play an important role in determining how much it affects the thermal stability of other proteins. We also find that more hydrophobic residues on the surface results in a more significant destabilization. This can quite possibly be because a more hydrophobic crowder is more likely to contain patches with a destabilizing geometry.

3.3.2 Constructing a Viable Crowding Environment

Knowing what the main factors determining crowder destructiveness are, it is natural to ask whether we can use this knowledge to select a set of proteins which could coexist at some "physical" temperature. In order to make such a determination at all possible, it is necessary to restrict ourselves to globular, monomeric proteins. In this section we attempt to construct such a set of proteins by imposing a set of restrictions on the constituents.

The first restriction imposed is naturally the existence of a unique ground state, as otherwise it would make little sense to talk about a globular protein. For this construction, we also want the proteins to be globular at a specific temperature. We impose this by requiring that the native probability at the "physical" temperature exceeds 80%. The choice of physical temperature is by necessity somewhat arbitrary, but we want to choose a temperature for which there are a reasonably large number of stable proteins, but where most are only marginally stable, as is the case for real proteins. We find that $\beta = 5$ is a reasonable choice fulfilling these criteria.

As a second restriction, we also include a cut-off for the mutational stability. Previous studies have also shown that evolutionary dynamics will favour proteins with the ability to withstand a large number of mutations without impeding their function [14]. In addition, imposing this restriction reduces the number of allowed proteins, making the following analysis a lot more manageable. For these reasons, we limit the allowed protein sequences to those that can withstand at least 10 point mutations without altering their structure.

Finally, we want to incorporate crowding considerations. As shown in the preceding section, the destabilizing effects of a given protein depends mainly on the local geometry of the surface around a hydrophobic patch. In order to better categorize the destructiveness of each patch, we introduce a set of motifs for the surface patches.

We define a motif as all adjacent hydrophobic residues along the surface,

as well as the two bordering polar ones. Examples of how this looks in practice can be seen in figure 13. The motif is thus defined both by the number of hydrophobic residues involved, and the shape in which these and the surrounding polar residues are arranged.

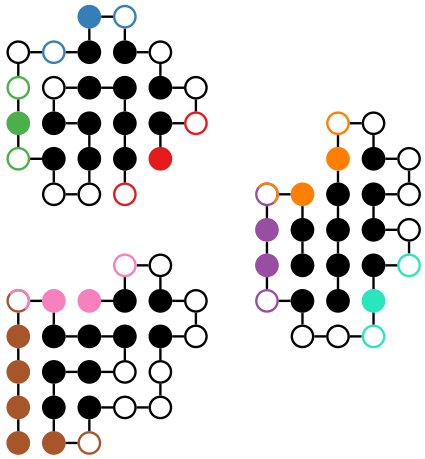

Figure 13: Illustration of how the motifs defined for surface patches work. Each motif is considered to consist of the hydrophobic residues in a patch, as well as the two nearest polar ones. In the figure, the residues belonging to each motif is shown in a specific color. Note that since the shape of the motif is important, all of the motifs shown coloured in the figure are considered to be of different types.

Among all HP proteins with a unique ground state, a total of 248 motifs were found to exist. In the set of proteins considered after imposing thermodynamic and mutational stability constraints as per the above, only 19 of these remained. While the reduction from 248 to 19 motifs may at first glance appear to be highly restrictive, it should be noted that most of the excluded ones consist of abnormally large patches. Since the shape of the whole patch is considered for this classification, the number of possible motifs increases exponentially with patch size. In addition, large patches should destabilize other proteins significantly, making their viability unlikely.

In order to determine the destructiveness of each of the 19 motifs, we selected 19 crowder proteins, each with a single hydrophobic surface patch in the shape of one of the motifs. Simulations were then performed with one test protein and a single of the crowder proteins, ensuring that no effects due to crowder-crowder interaction are present. As test proteins, we used 195 sequences randomly chosen among those with high thermodynamic and mutational stability as discussed above.

As a measure of whether a given test protein is able to survive a given crowder, we determine its native stability at $\beta = 5$. If it is higher than 80% we consider the combination as being able to coexist. Figure 14 shows the fraction of test proteins that are able to coexist with each motif. We see that there are significant differences in the destructiveness of the various motifs, with only a few being able to coexist with a significant number of test proteins. Finally, we can note that the size of a motif is highly correlated with its destructiveness.

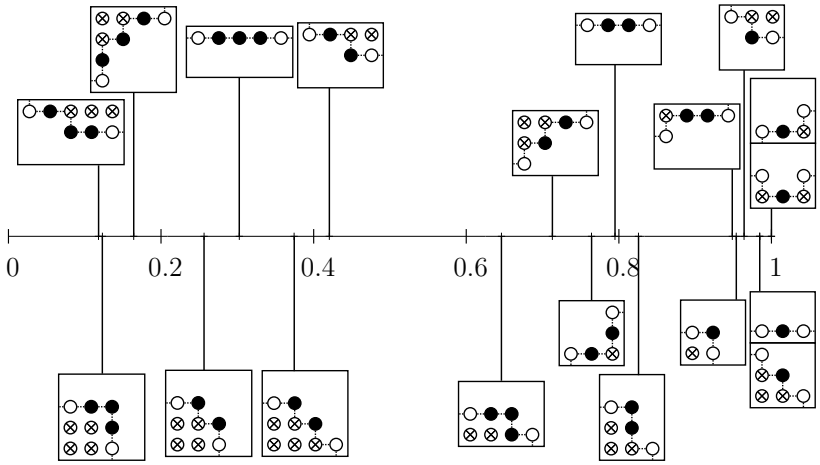

Figure 14: The motifs present on mutationally and thermodynamically stable proteins, rated by how large a fraction of the test proteins are able to survive their presence. Internal beads are marked with a cross to indicate that their hydrophobicity does not affect the surface.

In order to determine which proteins are viable in a larger crowding environment, we should consider not only how destructive a given motif is, but also how common it is. These two measures are shown in figure 15. We find that there are only five different motifs which are present on a larger number of proteins than the number which are denatured by their presence. In the construction of a viable crowding environment, we therefore exclude any protein containing motifs other than those five, or that is denatured by those five. There exist a few additional proteins where the number of denatured proteins is only slightly higher than the number of proteins with the motif,

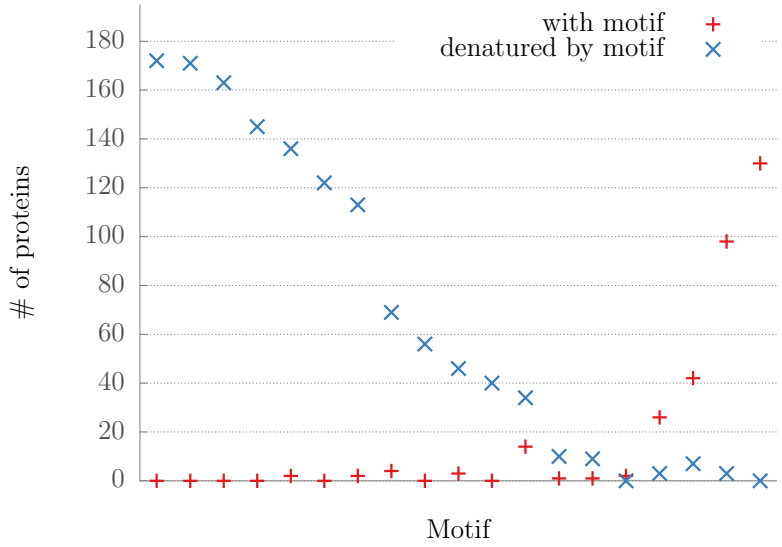

Figure 15: The number of proteins with (red) and denatured by (blue) the presence of the various motifs. We note that there are only five motifs which are present on more proteins than they denature.

usually because the number of sequences affected is small. Since the number of sequences affected is small, the choice of whether to include them or not is largely irrelevant.

3.3.3 Implications of Crowding on the Set of Allowed Proteins

After imposing all previously mentioned constraints, 4660 viable protein sequences remain. This set of Sequences with Viable Surfaces (SVS) will be of some interest going forward. A summary of the various constraints involved in constructing the SVS set can be found in table 2. When studying how a reasonable crowding environment affects various processes, it is useful to have a single, representative, environment. We created such an environment by simply choosing 20 proteins randomly from the SVS set. This smaller set will be denoted by SVS20. It is worthwhile to note that the SVS20 set contains hydrophobic patches of all five allowed motif types.

In the SVS20 environment, we simulated each of the 4660 proteins. 4004 of them were able to exist in this crowding environment while maintaining a native probability of at least 80%. Table 2 summarizes how the various

Table 2: The number of eligible protein sequences and structures after limiting the allowed proteins in various ways. Palindromic sequences are counted only once for this comparison. The bolded entry corresponds to the SVS set of proteins.

Criterion	# of sequences	# of structures
Unique ground state	1485131	204600
Thermodynamically stable at $\beta = 5$	158340	42986
At least 10 neutral mutations	5822	2884
Thermodynamically + mutationally stable	5597	2796
Above and no dangerous motifs	**4660**	**2422**
Above and survives crowding at $\beta = 5$	4004	2183

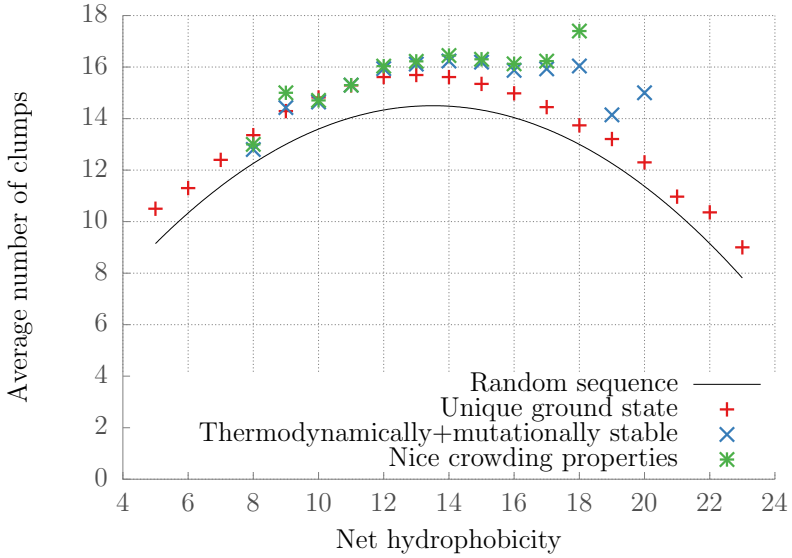

Figure 16: The number of hydrophobic and polar "clumps" along the chain for various sets of HP-sequences as defined in table 2. Nice crowding properties implies both a viable surface and the ability to survive crowding. Harsher demands on the sequences increase the number of clumps, particularly for sequences with a lot of hydrophobic residues.

requirements affect the number of functioning sequences and structures.

It should be noted that not all of the 20 proteins in the SVS20 environment survive crowding by that environment. Since there is, as far as we can tell, no specific properties of the surfaces of those proteins that do not survive, we still keep the same environment for future purposes. This also ensures that results can be compared between different simulations.

We also compare how the number of hydrophobic and polar clumps along the chain is affected by the various restrictions of sequences. The distributions can be seen in figure 16. It has previously been found that HP proteins folding to a unique ground state have a larger number of clumps [15]. Introducing additional restrictions further increases this number. This effect is most notable for sequences with a high net hydrophobicity, indicating that large hydrophobic regions may be the ones which are most harmful. The increase due to incorporating effects of crowding is modest compared to the effects of incorporating the requirement of thermal stability. Note however that the number of sequences excluded by the crowding considerations is comparatively small.

3.4 Properties of the Test Protein Influencing Its Stability in Crowded Environments

Thus far, we have primarily considered how the folded shape of a protein influences the stability of other proteins in its surroundings. In this section, we will consider what properties of a protein determines whether it is able to survive the crowding of other proteins. It should be clear that finding criteria for when a protein can survive crowding or not is a more difficult problem than determining its viability as a crowder. This is due to the fact that these effects are not solely dependent on the native shape of the protein, but also on all other possible conformations.

When studying the effects due to properties of the test protein, we restrict ourselves to the SVS set of proteins as defined in section 3.3.2. Thus, we exclude proteins with significant destabilizing effects on other proteins due to large hydrophobic regions. We also restrict ourselves to the SVS20 crowding environment created in that section. As a measure of how destabilized a certain protein is, we use the change in free energy difference as described in section 2.3.

One obvious candidate for being a destabilizing effect is the presence of hydrophobic residues which do not, or cannot, form bonds. Figure 17 shows how the destabilization depends on three different measures of this number. Figure 17b uses the most strict definition, counting only those hydrophobic residues which are adjacent to a polar one. Figure 17a also includes hydrophobic residues on the surface (counted double if more than one contact faces outwards). Finally, figure 17c counts all H-P contacts, as well as those

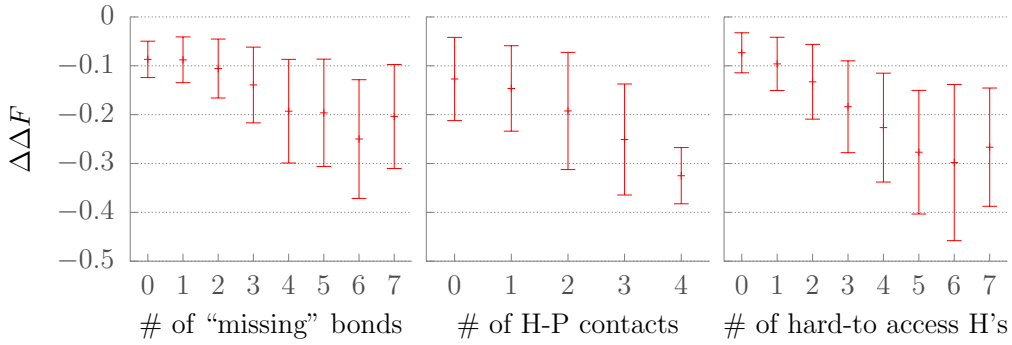

Figure 17: The change in free energy difference as a function of different measurements of the number of inaccessible non-bound hydrophobic residues. We find a clear correlation between the number of such residues and the amount of destabilization, although the variation is considerable. Furthermore, we see that this effect is not sensitive to the exact measure used. Note that the error-bars indicate standard deviations rather than error of the mean.

surface residues which are not "easily accessible" by the crowders. A surface residue is considered easily accessible if the position which is in contact with the residue is freely accessible from all other directions.

For all three measures, we find that the degree of destabilization is correlated with the number of unformed bonds, perhaps most clearly so in figure 17c. This indicates that unbound hydrophobic residues on the surface of the protein are less destabilizing than internal, especially if they can easily form bonds with other proteins.

Finally, we note that, while there is undoubtedly a trend, there are also large deviations for individual sequences. Note however that the error bars shown in figure 17 indicate standard deviations rather than the standard error on the mean. Since we are interested in the deviations for individual sequences, this appears to be the more relevant measure.

3.5 Implications of Crowding for Evolutionary Processes

In determining what sequences are viable in a crowded environment, we have considered only proteins with high mutational stability. While the majority of proteins found in living organisms are likely to have a high mutational stability, sequences with lower stability will have to have been visited during

some points in order to allow evolution to progress. In order to further the understanding of protein evolution it is thus of interest to also study crowding effects on the larger set of all proteins.

The analysis of surface motifs can easily be generalized to encompass all proteins. While the previous analysis focused only on a small number of motifs, most of the excluded ones can simply be considered as larger versions of those already tested. Since the largest motifs used there were found to be highly damaging, almost all the excluded motifs will also be so. While we might exclude one or two motifs which should have been included using this method, that will likely not account for more than a fraction of a percent of the total number of sequences.

For a complete analysis of the effects of crowding, we should also determine what proteins would be able to successfully fold in a crowded environment. Performing this rather large set of simulations would require a significant amount of computer time. We therefore do not consider this aspect in this section, excluding proteins only based on their surface and their native probability in an uncrowded environment.

Earlier studies have found that for RNA models, the sizes of neutral sets appear to be approximately log-normally distributed [16]. Whether this also holds for proteins is thus far unknown. Figure 18 shows the neutral set size distribution for all sequences with a unique ground state, and for thermodynamically stable sequences with non-destructive surfaces. We find that for the HP-model, a fitted log-normal curve approximates the distribution well. The mean of the distribution is, however, so small that there is no chance of seeing the left-most part of the distribution where the probability starts to decrease.

There are several possible reasons for why the neutral-set size distribution for RNA is supressed for small set sizes while that for HP proteins is not. One is that there is a higher degree of redundancy in the RNA case, since a binding base pair can in most situations be any of 4 pairs. In the HP-model on the other hand, there is only one binding pair. Second, there is some difference in how a structure is defined. The results for RNA are based on considering any set of base pair bonds as a unique structure. This means that structures with a number of unbound base pairs will also be present. These structures will typically be designed by more sequences, since there is less correlation between would-be bound pairs. There will also be more of them, since it will typically be possible to break pairs in several different locations. Nevertheless, the log-normal behaviour is still a good fit in the

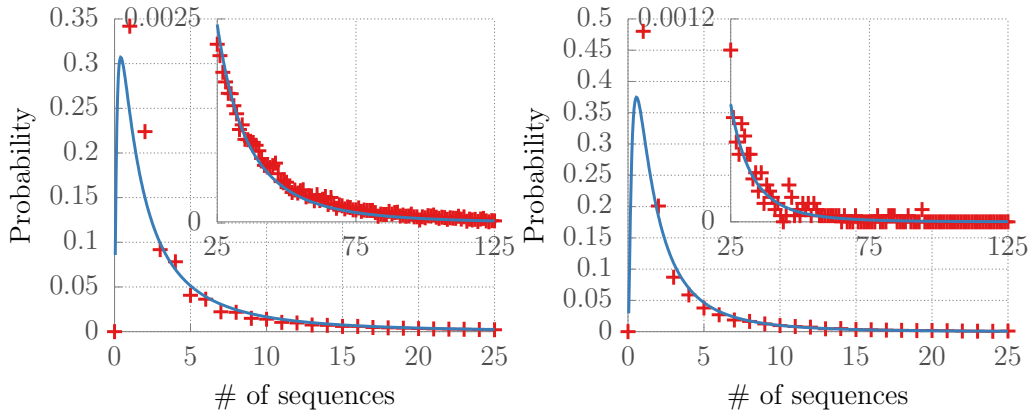

Figure 18: The distribution of sizes of the neutral nets for structures, considering all sequences folding to a unique ground state (left) and sequences which are thermodynamically stable at $\beta = 5$ and have non-destructive surfaces in their folded states (right). The solid lines shows the probability density of a log-normal distribution with the same mean and standard deviation. Insets show the tails of the distributions.

case of the HP model, particularly in the right tail.

Previous studies have shown that for HP proteins with a unique ground state, point mutations appear to be a viable method for exploring sequence space, with the neutral nets of various structures often being connected by so-called fold-switches [8]. This results in a sequence space which can be traversed by sequences with well-defined native states.

In order to quantify how easily traversed the sequence space is, we consider the mutational path between any pair of the top 100 most highly designable structures. Figure 19 shows the distribution of the number of fold-switches necessary to go from one of the structures to the other using only single-point mutations. As can be seen, the introduction of constraints on the protein structure makes it more difficult to traverse the sequence space, meaning that the number of fold-switches required increases. It should also be noted that when both thermodynamic stability and non-destructive crowder properties are required, the set of functional sequences is split into several components which are not connected to each other by point mutations.

It is clear from the above that not all fold-switches survive the introduction of more restrictions on the sequences. It is natural to ask which fold-switches survive and which do not. To investigate this, we consider all possible fold-switches for a given sequence set. For each fold-switch, we

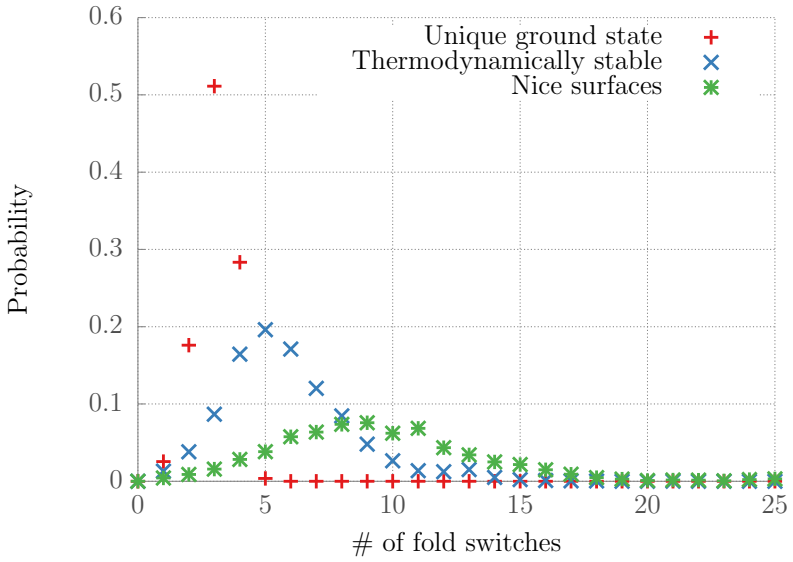

Figure 19: The distribution of the number of fold-switches necessary to move from one highly designable structure to another, for all sequences with unique ground states, all sequences which are thermodynamically stable ($P_{\text{nat}} > 80\%$) at $\beta = 5$, and sequences which are thermodynamically stable and have non-destructive surfaces. We see that imposing stricter conditions tend to increase the number of fold-switches necessary. Although it is not apparent from the figure, it should also be noted that for the most stringent set of conditions there are a number of sequences which cannot be reached from the others.

determine which amino acid contacts are present in each of the two conformations. The number of shared contacts can then be used as a measure of how similar the two structures are.

Figure 20 shows how the number of shared native contacts of the two structures involved in a fold switch are distributed. We can see that the typical structural difference between the two conformations is smaller when imposing restrictions on the proteins. It should be noted that the total number of possible fold-switches decreases significantly, so that even switches with many conserved contacts are typically broken. Nevertheless, the surviving fold switches all involve an increased amount of of conserved contacts.

3.6 Applications to Protein-Protein Interactions

Not only the fold is important for understanding the biological function of proteins, but also their interactions with one another. These interactions can

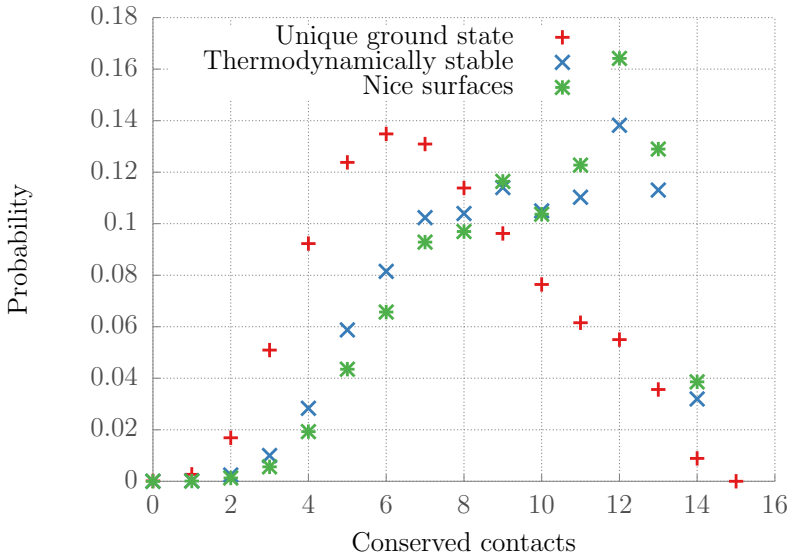

Figure 20: The distribution of the number of conserved native contacts in a fold switch, for all sequences with unique ground states, all sequences which are thermodynamically stable ($P_{\text{nat}} > 80\%$) at $\beta = 5$, and sequences which are thermodynamically stable and have non-destructive surfaces. Imposing a minimum thermodynamic stability highly favours those switches for which the similarity is high. Restricting the surfaces on the other hand has only a rather minor effect. Note that the distributions are normalized as probabilities.

either be protein-protein bonds, where two or more proteins form a tightly-bound complex, or transient interactions with the cytoplasmic environment [17]. Another type of protein-protein interaction is so-called aggregation, which is implicated as a mechanism in several diseases [18]. In this section, we investigate how crowding affects various types of protein-protein interaction.

3.6.1 Transient Interactions

Most proteins will have at least some residues capable of interaction with other proteins on their surfaces. This can result in temporary bonds formed between proteins close to each other. Often, these bonds will be non-specific, and as such they presumably do not have a major impact on the folding properties of the protein. They could, however, have significant impact on transport properties, e.g. the diffusion rate, of the protein. As such it is of interest to investigate how the frequency of bonds depend on both test protein and environment [17].

For this investigation we restrict ourselves to the SVS set of protein sequences as defined in section 3.3.2. We also consider both the test protein and the crowders only in their native states (a test simulation with flexible test proteins indicated that this does not significantly affect the results at temperatures of interest).

First, we consider only the effects of the test protein surface. A naïve view of these effects would be to assume that each hydrophobic patch on the test protein binds independently of the others. To test the veracity of this assumption, we perform simulations of proteins with 1-3 patches, where all patches on the same protein have the same motif. Three different motifs were used. For each motif type and each patch number, up to five different test protein sequences were simulated. (In some cases there are not five proteins with three patches of the same motif.) These were placed in the SVS20 crowding environment designed in section 3.3.2. The results of these simulations can be seen in figure 21.

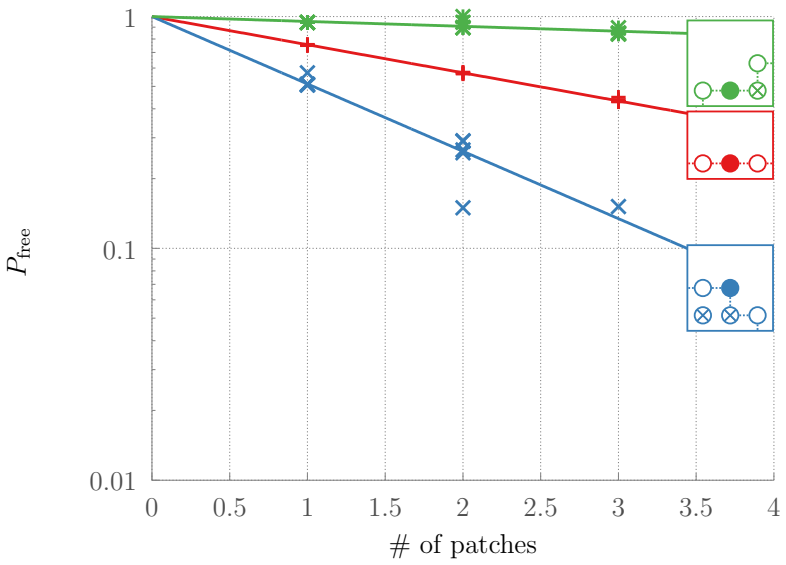

Figure 21: The probability for a test protein to stay unbound in a crowded environment, against the number of hydrophobic patches on its surface. The colours of the points indicate the motif of the patches on the test protein, which are all the same. Lines show an exponential fit to the points of the same color. The exponential dependence indicates that the surface patches to a large extent bind independently of each other.

There are two things of interest to note in figure 21. First, the naïve

hypothesis appears to hold up well. There are a few exceptions, but these can be explained by the fact that the definition of motifs does not capture the full shape of a patch, and, in one case, the test protein is able to form two bonds simultaneously with a single crowder. While not surprising, the results show that single hydrophobic patches can typically be considered independently of the rest of the protein.

The second thing to note is that patches corresponding to the three different motifs studied show significantly different affinities for binding. Furthermore, this happens despite the fact that they are similar in size, each consisting of a single exposed hydrophobic residue. Again this result is not particularly surprising, but it shows that, just as when it comes to the degree of destabilization (see section 3.3.1), the surface hydrophobicity is a limited predictor of the effects, and for a more complete picture the geometry of each patch has to be taken into account.

Turning to the surrounding crowders, we investigate how the amount of hydrophobicity on the crowder surfaces affects the probability of binding. When doing this, we construct crowding environments with varying amounts of total hydrophobicity, but the same relative frequency of different motifs by the following procedure.

1. The relative frequency of different motifs in the SVS set was determined.

2. For each environment, the total number of hydrophobic patches of each type of motif in that environment was chosen. These numbers were chosen so that the relative frequencies were the same as for the SVS set.

3. Each hydrophobic patch was placed on a random crowder. In this way the total number of patches of each motif type on each crowder was determined.

4. Each crowder was chosen from among all those with the designated set of patches.

5. If there were no proteins with the desired set of patches, one of them were randomly removed and placed on another crowder.

We used three different test proteins, each with only a single hydrophobic patch.

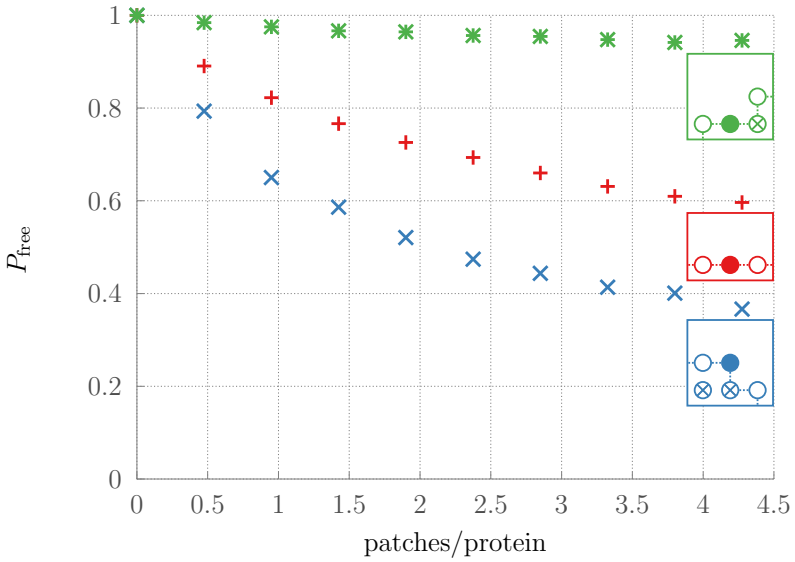

Figure 22: The probability for a test protein with a single hydrophobic patch to stay unbound in a crowded environment, against the number of hydrophobic residues on the surfaces of the crowder proteins. Colours indicate the motif of the patch on the test protein. The chance of binding seems to increase with the number of patches on the crowders, but with an upper limit.

As can be seen in figure 22, the test proteins show an increased binding affinity as the number of hydrophobic patches is increased. We can note that there seems to be an upper bound to how often the protein will bind. This could be due to the fact that the number of patches will at some point be placed so densely that not all of them can be bound simultaneously. Thus the effective number of patches available for the test protein to bind to would reach a maximum level.

The existence of non-specific interactions in a cell is probably impossible to escape. We find that the probability of a given protein being involved in transient interactions depends on the amount of hydrophobicity present on both the protein itself, and on the surrounding crowders. While different patches on the test protein seem to bind independently of each other, the dependence on the environment is not as obvious.

3.6.2 Dimeric Binding

Many of the proteins present in real cells are so-called oligomers, i.e. they consist of two or more amino-acid chains which are tightly bound to each other. These bonds are typically important for the protein to function properly, and as such it is of interest to investigate how crowding may affect their formation.

When studying this type of protein binding, we make a couple of restrictions. First off, we consider only dimeric bonds. Further, we take one of the binding partners to have a fixed conformation, and study the behaviour of the other partner. For the flexible partner, we restrict ourselves to proteins whose conformation when bound is identical to their unique ground state in isolation. With these restrictions, it is possible to study the effects of crowding on dimer formation in a systematic manner.

As a first step, we investigate what bonds can be formed using only mutationally stable proteins in their ground states. To do this, we search through all protein pairs, where both proteins have at least 10 allowed neutral mutations each. Each pair is considered as being able to form a bond if they can be placed in such a way that they form at least three intermolecular HH contacts. The resulting interaction graph is shown in figure 23. We note that the graph consists of a few clearly visible components, corresponding to only a few distinct motifs at the binding site. One of these is by far the most common, and we choose to focus only on proteins binding using this binding site.

We select 21 different proteins, chosen so as to have a wide range of different values of their thermal stability at $\beta = 5$. Note that in order to find proteins with sufficiently low thermal stability, we do not apply any requirement for the mutational stability. As mentioned, each test protein is simulated along with a binding partner with fixed structure. The binding partner was chosen to have no patches other than the binding site.

To start, we perform simulations with no other proteins present. Figure 24 shows how melting of the test protein and binding between the two proteins are related. We see that the bond with the binding partner typically stabilizes at about the same temperature as the protein melts. On the other hand, the two proteins often tend to form non-specific bonds with each other at significantly higher temperatures. This temperature tends to stay rather constant for all of the different test proteins, indicating that the binding process is mostly independent of the specifics of the folding path of

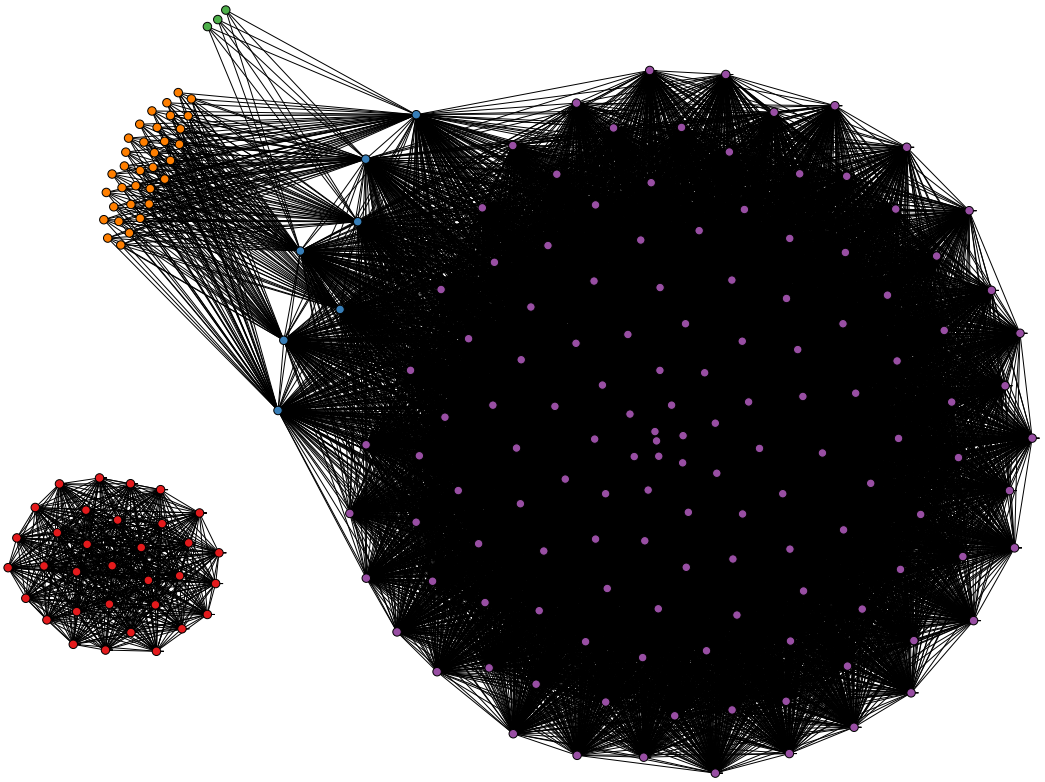

Figure 23: Graph showing the interaction network for all sequences able to tolerate at least 10 different point mutations. The nodes represent sequences, and two nodes are connected if the corresponding sequences can bind to each other in such a way that they form at least three H-H contacts while in their native forms. The colors of the nodes denote the motif of the hydrophobic patch involved in binding. Nodes without connections are excluded. Since each purple node is connected to all purple and blue nodes, individual edges are hard to discern. Figure generated using Gephi [19].

the protein.

To investigate the joint effects of binding and crowding on the protein, we simulate the pair in free form, with steric crowders, and with the interacting crowder environment designed in section 3.3.2. As can be seen in figure 25, we find that steric crowders have only a modest effect. In addition, we note that while the folded and bound state is usually stabilized, there are examples where this is not the case. Presumably, this is because the folded/bound state is actually not the most compact conformation for these sequences. Interacting crowders, on the other hand, result in a clear destabilization, both of the internal structure of the flexible chain, and of the binding.

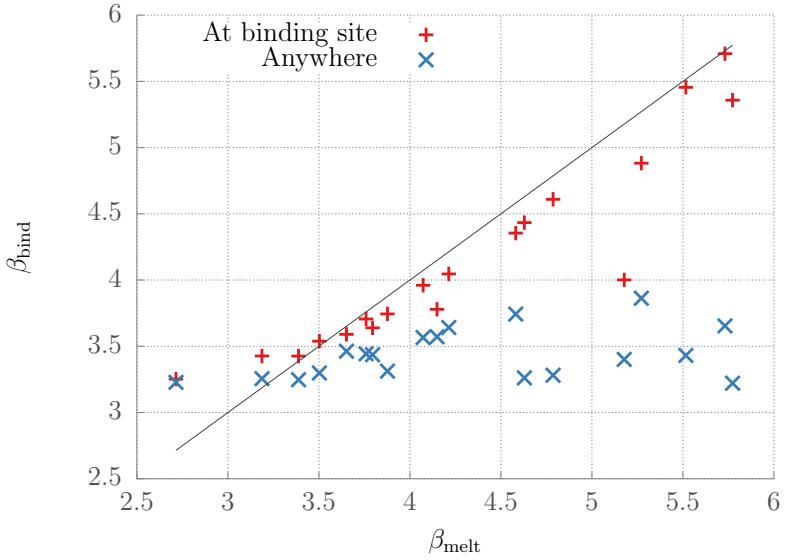

Figure 24: The correlation between inverse melting temperature and inverse binding temperature for a variety of test proteins in the absence of crowders (apart from the binding partner). The binding temperature is shown both for site specific (red pluses) and non-specific (blue crosses) binding. We note that the proteins tend to bind to their partner non-specifically at about the same temperature, regardless of their stability. Furthermore, the nonspecific binding temperature is typically higher than the folding temperature. On the other hand, forming bonds at the specific binding spot tends to happen at about the same temperature as the folding.

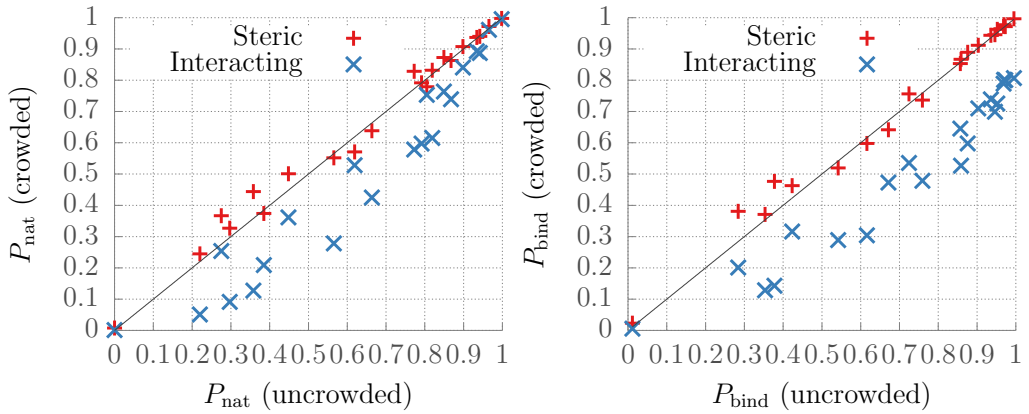

Figure 25: The effects of crowding on dimeric proteins at $\beta = 5$. The leftmost plot shows the native probability while the rightmost shows the binding probability, both as functions of the same probabilities in the absence of crowders. Steric crowders tend to have little effect, while interacting crowders often produce a clear destabilizing effect.

It should perhaps be noted that there are a few reasons for why the effects of these binding simulations can not be seen as completely representative of real protein binding. Primarily, we note that, due to the coarseness of the model, the shape of the binding spot normally will not be as specific as they would be in a more realistic model. The two binding proteins may thus be more prone to misinteraction than a pair with a more specific binding interface would. This could potentially result in a greater effect from crowders, as well as a higher temperature for non-specific interaction between the two proteins involved in the binding.

3.6.3 Peptide Aggregation in Crowded Environments

One important type of protein-protein interaction is protein aggregation. One type of protein aggregation is fibril formation. During fibril formation, strands of different protein molecules bind alongside each other to form elongated aggregates. Protein fibril formation is thought to play a potentially key role in several diseases.

In order to study aggregation in the HP-model we use chains of length five, where the central three residues are hydrophobic, while the two at the edges are polar. To study the effects of crowding on protein aggregation, we simulate 20 of these peptides, each with full conformational freedom. These are placed in three different environments: without crowders, with fully steric crowders, and in the representative crowding environment determined in section 3.3.2. The level of aggregation was measured as the number of HH-contacts between the aggregating peptides.

The results of the simulations are shown in figure 26. For steric crowders, we see a weak stabilizing effect from the crowding environment, resulting in a slightly higher aggregation temperature. For interacting crowders, this observation remains valid at high temperatures, while at lower temperatures, the energetic interactions seem to negate the excluded volume effects, resulting in a slight destabilization of the aggregate.

Experimental studies have shown that the time-evolution of the total fibril mass typically follows a sigmoidal curve [20]. This abrupt onset of aggregation indicates that the system is trapped in a meta-stable free-energy minimum corresponding to the solution phase. In simulations, we would then expect a bimodal energy distribution near the transition temperature. As evidenced by figure 27, this is not observed in the simulations. This is not unexpected, since fibril formation is a one-dimensional process in the studied

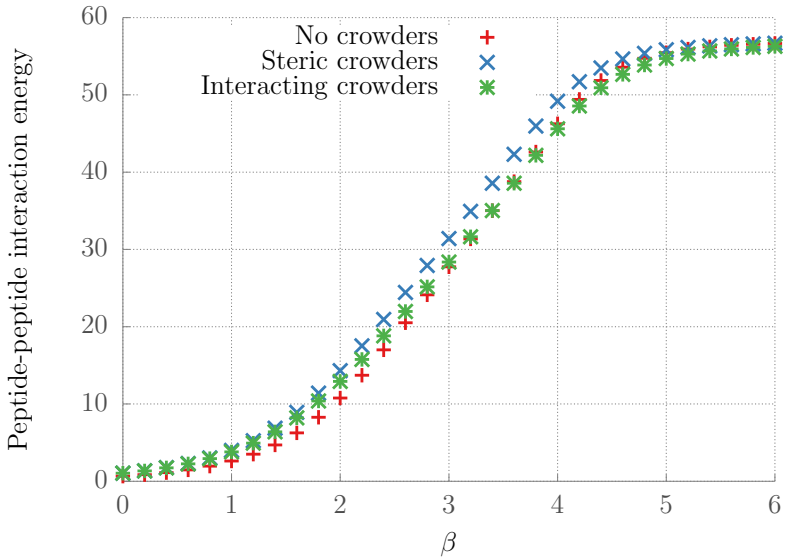

Figure 26: The level of aggregation in various crowding environments, as measured by the number of HH-contacts between the peptides forming aggregates. We note that steric crowders increases the tendency to form aggregates. The interaction crowders mostly cancels this effect, slightly favouring aggregation at high temperatures, and disfavouring it at low temperatures.

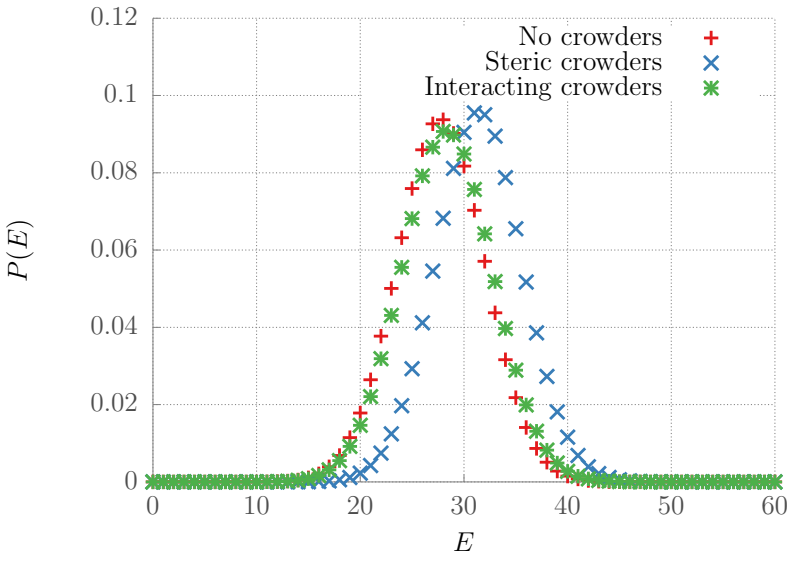

Figure 27: The histogram for peptide-peptide interaction energy at inverse temperature $\beta = 3$ in various crowding environments. The distribution shows a single peak, indicating that there is only a single free-energy minimum.

sytem [21]. The lack of a bimodal energy distribution indicates that further study of aggregation processes in the HP-model are probably uncalled for.

4 Discussion

The results presented in this report show that crowding is an important factor in determining the viability of a given HP protein. We see that protein environments where no consideration has been taken to the crowding properties of the proteins tend to be highly destabilizing for the test protein, and that this destabilizing effect persists even when steps are taken to shield the protein (see figures 4-6). To the extent that these conclusions are valid for real proteins, the properties of the protein surfaces appear to be of considerable importance for the viability of a protein in a crowded environment.

Furthermore, our results show that, while there is a clear correlation between the amount of hydrophobicity present in a given crowding environment and the destabilization on a test protein in this environment, there are considerable variations, and the use of surface hydrophobicity as a proxy-variable for destabilization is probably not warranted. When considering the effects of surface geometry, we find that the size and shape of a single hydrophobic patch plays an important role in the amount of destabilization experienced. In summary: a small amount of hydrophobic surface residues placed close together are typically more destructive than a large amount of evenly spread ones.

The facts that sequences need to be thermodynamically stable and have non-destructive surfaces have been accounted for in previous studies of evolution in the HP-model. We find that more strict requirements on the surface result in a fitness landscape that is less easy to navigate. While requiring thermodynamic stability has a more pronounced effect than the restraints on the surfaces (see e.g. figure 20), this may be due to the fact that it is the more severe restriction. We also note that the size distribution of neutral nets seem to follow a log-normal distribution reasonably well, in agreement with data for RNA. In comparison with the RNA models, where most sequences correspond to a specific structure, the number of HP proteins folding to a unique structure is small, and made even smaller by requiring that the protein have reasonable crowding properties. This also results in the mode of the distribution appearing at less than one sequence per structure.

We note that the effects of crowding seem to provide two qualitative constraints on the distribution of hydrophobic and polar residues in the protein. As previously described, the requirement that a protein should not cause significant destabilization of other proteins requires that the surface does not contain any large patches of hydrophobicity. In order to make sure

that the protein does not get significantly destabilized by the surrounding crowders, it needs to have a hydrophobic core that is relatively free from polar residues (see e.g. figure 17). To a lesser degree it also decreases the amount of hydrophobicity that can be present on the surface of the protein, especially if it is located at an inaccessible part of the surface. In conclusion, crowding strengthens the tendency for the protein to have a hydrophobic core surrounded by a polar surface.

Finally, we also found that crowding affects both binding and aggregation, slightly destabilizing both processes. When studying binding in particular, it should be noted that we were forced to make some more restrictions than would be optimal. In particular, we had to restrict ourselves to binding through docking. It can also be noted that the granularity of the model means that bonds tend not to be very specific. With regards to aggregation we found that in the model there was no sign of a phase transition, possibly due to the fact that the investigation was carried out in two dimensions.

The results presented in this report have all been found using a very simple protein model, and it is natural to ask to what extent they hold for more realistic scenarios. Among the most significant approximations made in the model are the restriction to two dimensions, the highly simplified energy function, and the coarse-grained representation of the chain arising from the requirement that each residue be placed on a lattice point. In the following paragraphs, we will discuss the impact of these approximations in some more detail.

One of the most significant approximations is the limitation to two dimensions. While the conclusion that it is the shape of local hydrophobic patches that are important rather than the total amount of surface hydrophobicity on the crowders should still hold, the size of patches which are allowed could very well be affected significantly. A monomeric protein with chain length N will in its native form fold to a roughly spherical shape with radius R. A hydrophobic patch on the surface could also be described as mostly circular, with radius r. We expect that a patch will be destructive once the energy of binding to it exceeds some fraction of the total binding energy. In two dimensions we would then have

$$\frac{E_{\text{patch}}}{E_{\text{internal}}} \propto \frac{r}{R^2} \propto \frac{r}{N}, \tag{19}$$

while in three

$$\frac{E_{\text{patch}}}{E_{\text{internal}}} \propto \frac{r^2}{R^3} \propto \frac{r^2}{N}. \tag{20}$$

For the maximum radius we would then have in two dimensions $r_{\max} \propto R^2 \propto N$, and in three $r_{\max} \propto R^{3/2} \propto N^{1/2}$. Thus, longer chains will be able to tolerate larger patches, but the increase will be more significant for two-dimensional chains. It is in other words likely that three-dimensional HP proteins of similar size would be more sensitive to crowding than two-dimensional ones.

Another major simplification in the model is that the energy only includes hydrophobic interactions, and thus destabilizing effects due to other types of forces are absent. The fact that the hydrophobic effect is one of the most important driving forces in protein folding suggests that it should also be highly important for crowding. There are also some other reasons, detailed below, to think that these effects may be less pronounced than that which comes from hydrophobic interactions.

Some of the forces involved in protein binding, such as van der Waals forces and hydrogen bonds are non-specific, meaning that any amino acid residue pair has similar chances of interaction. This means that there is typically little gain in exposing buried parts of the chain, and thus presumably little destabilizing effects. The exception would be structures where there are a lot of shielded residues which do not form these kinds of bonds with any other residue, but this kind of shape would presumably be rare.

There may also exist electrostatic interactions between various residues. These are obviously specific and thus could be expected to potentially have destabilizing effects. However, due to the fact that like charges repel, there is less chance of forming large patches on the surface. Since these tend to be the most destructive elements of the crowders, the electrostatic residues should result in modest destabilizing effects. Thus, while the model obviously does not catch all the nuances of protein folding in the presence of crowding, it may well provide an at least somewhat reasonable approximation of the destabilizing forces.

Finally, there is the fact that the model has a highly simplified geometry, with residues represented as single "beads" on a lattice. This results in a smaller number of possible local patch shapes, and also results in an artificial distinction between horizontal/vertical surfaces and diagonal ones. It also obviously makes it difficult to use for representing highly specific surfaces. For representing non-specific interactions between a protein and its surroundings, this approximation should still be able to provide some insight.

References

[1] Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).

[2] Dill, K. A. Dominant forces in protein folding. *Biochemistry* **29**, 7133–7155 (1990).

[3] Lau, K. F. & Dill, K. A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* **22**, 3986–3997 (1989).

[4] Smith, A. E., Zhang, Z., Pielak, G. J. & Li, C. NMR studies of protein folding and binding in cells and cell-like environments. *Current Opinion in Structural Biology* **30**, 7–16 (2015).

[5] Zhou, H.-X. Influence of crowded cellular environments on protein folding, binding, and oligomerization: Biological consequences and potentials of atomistic modeling. *FEBS Letters* **587**, 1053–1061 (2013).

[6] Feig, M. & Sugita, Y. Reaching new levels of realism in modeling biological macromolecules in cellular environments. *Journal of Molecular Graphics and Modelling* **45**, 144–156 (2013).

[7] Zhou, H.-X., Rivas, G. & Minton, A. P. Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annual review of biophysics* **37**, 375–397 (2008).

[8] Holzgräfe, C., Irbäck, A. & Troein, C. Mutation-induced fold switching among lattice proteins. *The Journal of Chemical Physics* **135** (2011).

[9] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**, 1087–1092 (1953).

[10] Swendsen, R. H. & Wang, J.-S. Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58**, 86–88 (1987).

[11] Frenkel, D. & Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*. Computational science series (Elsevier Science, 2001).

[12] Marinari, E. & Parisi, G. Simulated tempering: A new Monte Carlo scheme. *EPL (Europhysics Letters)* **19**, 451–458 (1992).

[13] Lyubartsev, A. P., Martsinovski, A. A., Shevkunov, S. V. & Vorontsov-Velyaminov, P. N. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *The Journal of Chemical Physics* **96**, 1776–1783 (1992).

[14] Bornberg-Bauer, E. & Chan, H. S. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proceedings of the National Academy of Sciences* **96**, 10689–10694 (1999).

[15] Irbäck, A. & Troein, C. Enumerating designing sequences in the HP model. *Journal of Biological Physics* **28**, 1–15 (2002).

[16] Dingle, K., Schaper, S. & Louis, A. A. The structure of the genotype–phenotype map strongly constrains the evolution of non-coding RNA. *Interface Focus* **5** (2015).

[17] Wirth, A. J. & Gruebele, M. Quinary protein structure and the consequences of crowding in living cells: Leaving the test-tube behind. *BioEssays* **35**, 984–993 (2013).

[18] Chiti, F. & Dobson, C. M. Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* **75**, 333–366 (2006).

[19] Bastian, M., Heymann, S. & Jacomy, M. Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media* (2009).

[20] Hellstrand, E., Boland, B., Walsh, D. M. & Linse, S. Amyloid β-protein aggregation produces highly reproducible kinetic data and occurs by a two-phase process. *ACS Chemical Neuroscience* **1**, 13–18 (2010).

[21] Irbäck, A. & Wessén, J. Thermodynamics of amyloid formation and the role of intersheet interactions. *The Journal of Chemical Physics* **143**, 105104 (2015).

A Used Sequences

This appendix contains listings of all sequences used in the simulations used in creating the figures, where the sequences were randomly chosen.

Figure 2

Table 3: The protein sequences used as crowders when producing figure 2. The sequence marked with an asterisk was also used as test protein.

Index	Sequence
1	HPHHHPPPPPHHPHHHHHHHPPPHPHP
2	HHHPHPPHPHHHPHHHPPPHHPPPHPH*
3	HHPHPPPHPHPHHPHPPPHHHPHPHPH
4	PHPPPPHHHHPHPPPPHPHHHHPHPHH
5	HHHHPPHPPHHPPPHHPPHPHHPHPHH
6	HPHPPPHPHPHHPPHPPHPPPHPHPHH
7	HHHHHHHPHPHHPPHHHPPHHHPHPPH
8	HHHHHPHHHHPHHPPHHPHHPPHPHPH
9	PHHHHPPPHPHPPHPPHHHHPPHPHHH
10	HHPHPPPPPHPHPPPHPPPHHHHPPHH

Figures 4-6

Table 4: The sequences used as test proteins in the results behind figures 5 and 6. The sequences marked by asterisks were used in creating the example curves in figure 4.

Index	Sequence
1	HHHHPHHPHHPHPPHHHHPPHPPHPPH
2	HHHPHPHHHPHHHHPPHPPHPPHHHHH*
3	HHHPHPPPHPPPHPPHPPHHPHPHHHP*
4	HHPHPPHHPHPHPPHPPHPPHHPPHPH*
5	HHPPHHPPPHPPHHHPPHPHPHHHPHH*
6	HHPPHPHPPPHPPPPHHPHHPHPHPPH
7	HHPPPHHHPHPPPPPPPHPPHHHPHHH
8	HHPPPHHPHPPHHPHPPHPHHPPHHPH
9	HHPPPHPPPHPHPPHHHHHPHPHPPHH
10	HPHHHHPHPHHPPPPHPHHPPHHHPPH

11	HPHHHPHPPPHHPPHHPHHHPHPPHHH
12	HPHHPPHPHHHHHHHHPPHPPHPHPHH
13	HPPHHHHHHPPPHHHHPHPHPPHPHHP
14	HPPHPPHPPHPHHHPPHHPHPPPHPPH
15	HPPPHHHHHHHPHPHPHPPPPHPHHHH
16	HPPPHHHHHPPHPHPHPPHPPHHHPHH
17	PHHHHHHHPHHPHPHPHPHHPHHPPHH
18	PHHHHHHHPHPHHPPHPHHHPPHPHPH
19	PHPHPPHPPHHPPHHPPHHHHPHPHPH
20	PHPPHPHPPHHPHHHHHPPPHPHHPHH

Table 5: The sequences used as crowders in the results behind figures 5 and 6. The sequence marked by an asterisk was used in creating the example curves in figure 4.

Environment	Sequence
1	PHHHHPHPHHPHHHPHPHPPPPHPPHP
1	HHHHHPHHPPHHPPHHHPPHPPPHPHP
1	HHPHHPPHHPHHPHHPPHHPHHHPHPH
1	HHPPHHHPPHHHHHHPPHPHPPHHHHP
1	PHHPHPHHHPHHPHHHHPHHPPHPPHP
1	HPPHPHHHPHHPHHPPPHPHPPHHHHP
1	HHPPHHHPPHHPPPPHPHPPPHPPHHH
1	HPPPHPHPPPPHPPPHHHPHPPPHHHH
1	PPHHPHHPPHPPHHPHPPPHPHHPPHH
1	HHHPPHPHPHPHHPHPHPPHHHHHPPH
2	HPPHHHPHPHPHPHPHHPHPPPHPHHP
2	HHHHPPPPHPPPHHPHPHPPHPHHHPH
2	HHHPPPPPHPHHHHHPHHPPPPHPPHP
2	HPPHPHPPHHPHHHPPHHHHPPPHHHH
2	HHPHPPHPPHPPPHPPHHHHHHHHPPH
2	HHHPPHPPHPPHPPPHHPHPPHHPHHH
2	HPHPHPPPHHHHHHPHPHPHPHPPHPH
2	HPPPPHPHHHHPPHHPHHPHHHPPHPH
2	PHPPHHHPHPPPHHPHHPPHHPHPPHP
2	HPPHPPHHHHPPHPPHHHHHPHPPHPH
3	HHHHPHHPHPHPPHHHPHHPPPPHPPH
3	HPPHPHPHHPPHPHHHHPPHPPHPHHH
3	HHHPHPPHHHPPPHHPPPHPPHPHPHP

3	HPPHPHPHPHHHPPHPPHPPHPPPHHH
3	PHPPHHPHPHPHPPHPPPHPPPPHHPH
3	PHHHHPPPHPPHPPHHHPPHPPHHPPH
3	HHPPHPPHHHHPHHHPHPPHHPHHPHH
3	PHPPHHPHPHHHPHPHHHHPHPHPHPH
3	PPHHHHHPHPPHHHPPHHHPHPHPHHP
3	HHPHPHPHHHPPHHHHHHHPPPHPHPH
4	HHPPHHPHPHPPHPHHHHHPHHHPPHH
4	PHPPHPPPPHPPPHHHPPHPPHPPHHH
4	PHPHPHPPPHPPHPPPHPPHHHPPHHH
4	HPPHHHPPPHPHHHPPHPHPHPHPPHP
4	HHHPPHPPHPPPHHPHPHHHHPHHHPH
4	HHHHHPPHHPHPHPHHPHHPPPHPPHP
4	HHHHHPPHHPPPHPPPHHPHHHHPHPH
4	PHHPHHPPHPHHPPHHPHHHPHHPHHH
4	PHPHPPHPPPPHPPPHPHHPPPPHHPH
4	HHPHPHPPHPHPHHPHPPPHPPPHHHH
5	HHPHPHPPPPHHHHPHPHPHHPHPHHH
5	PHPPHPHHHHHHHHPPHHPHPPHPPHP
5	HPPHPHHHHPPPHPHHPPPHHHHPPHP
5	HPHHPPHPPHPHHPPHPHPHPHHPHHP
5	HPPHHHHPHHHPHHPPPPHPPHPPHHHP
5	PHHPHPPHHHPHHPPPHPHPPHPPHHH
5	PHPPHPPHPPHHHPPPHHPPHPPPPPH
5	HHHPPPHHHPPHHHPPPPPPHHHPHHH
5	HPHPPPHPPPPHPHPPHPHPPHPHPHH
5	PHHPPHHPPPHPHPPHPPPPPPHPPHP
6	PHHHHPHPHHHPPHPHPPHHPHHHHPP
6	HHHPHPHHHPPHPPPHPHHHHPHPPPH
6	PPPHHHHHHHPHPHPHPPPPPPPPPPH
6	HPHPHPHPHHPHHPPHHPHHHPHHHHP
6	PHHPPPHHPPHHPHPPHHHHHPHPHHH
6	HHPHPHHPHHPPHPHHHPPPPHHPPHP
6	PPHHHPPPHHHHPHPHPPPPPHPPPHP
6	PHPPPHPHPPHHHHHPHHHPPHPPPHH
6	PHPHPPHPHPHPPPPPPHHHPHPHHHH
6	HHHPPPHPHHHPHPPHPHHHPPHHPHH
7	HHPPHPPPHPHHHPHHHHPPHHPPPHH

7	HHPPHPHHHPHHHHPHHPPPHPHPPHP
7	HHHPHPPHPHHPPPHPHHHPHPPHPHH
7	PHPHPPPHHHHHPHPPHPHHHHHPPHH
7	HHHPPPHHHPHHHHPHPPPHHHHPHPH
7	PHHPHHHPHPHPPHHHPPHHHHHPPPH
7	HHHHPPPPHPPPPHHHPPPHPHPHHPH
7	HHPPHPHPHPPPPHPPHHPHHPPPHHP
7	PHPPHHHPHPPHHPPPHHHPPHHHPPH
7	PHPHPPPHHHPHHPPPHPPHPPHHPHH
8	PHHHHHPPHPHPPPHHPPPHPPHHPHP
8	HPPHPHHHHPPHPPPHHPHHPPPHHHH
8	HHHHHPHPPPPHHPPPPHPHPHPHHHH
8	PHPHHHHHHHHHPHPHHHHPHPHPHHH
8	HHHHHHPPHHPPHPHPPPHPPHPPPHH
8	HHPPHHPHHHHPPPHHHHPHHHHPPHP
8	PHHPHPHHPPHPPHHHPPPHHPPHPPH
8	HHPHPHHHHPPPHHPHPPPHHPPPHPH
8	HPPHHHHPPHPPHHHPPHHPHHPPHHP
8	HPPHHPPPHPHPPPPHPPHPHHPHHHP
9	PHPPHHHHHHHHPHPHHHPPHHPPPHP
9	PPHHHPHPHPPHHPHHHPPHPHPHPHP
9	HHHHHPHPHHPPHPPHHPPHPHHPPHH
9	HHHPHPPHPHPPPPHHPHPPPHPHHHH
9	PHPHHHHPHHHPHHPHHPHHPPHPHPH
9	HHPPHHPHHHPHHHHPPPHHPPPPHHH
9	HHPPPHPHHHHHHPPHPHPPPHHHPHH
9	HHHHPPHHHHPHHHPPHHPPPHHPHPH
9	HHPHPHPHPPPPHHHHPHPHHHPPPHH
9	HPHHPHPHPPHHPPPHHHHHPPHPPHP
10	PHHPPHPHPHPPPPHHHPPHPHHPPHH
10	HPPPPPHPHHHHPHHHPHHHHHPHPHH
10	PHPPHHHPHHHHHPHHHPHHPPHHHHP
10	HPPPHHPPPPHPHPHHHHPHPHPPHHH
10	PHPHHPPPHPPPHHHPHPPPHPPHHHH
10	PHPPHPPHHPPPPHHHHPPPHPPHPHP
10	HPPHHHPPHHPPPHPHHPPHPHHPHHH
10	HHPHPHPHPHPPPPHPPHHPPPPPHHP
10	HPHHHHPHPHHHHPHPPPHPHHHPPPH

10	PHHHHPHPPHPHHPHPPPPPHHPHPHP

Figures 8-9

Table 6: The sequences used as test proteins in the results behind figures 8 and 9.

Index	Sequence
1	HPHPPPHPHPPHHPPHHHPHHPPHPHP
2	PHHHHPHPHPPHHPPHHPPHPHPPHHH
3	PHHHHPHPHPPHPHPHPHPPHPPHHPH
4	PHPHPHPHPHHHHPPPHPHPHPHHHHH
5	PHPPHHHPPHPPHHPPPPHPPHHPPHH
6	PHPPHHHPPHPPPPHHPPHHPPHPPHH
7	PHPPHPHHHHHHPHPHPPHPHPPHHPP
8	PHPPHPHPPHPHPHPHPPPPHPHHHHH
9	PHPPHPHPPPPHHPPHPPHHPHPPHHH
10	PHPPPHHPPPPPHPPPHPHHPHPHHHH

Table 7: The sequences used as crowders in the results behind figures 8.

Designability quartile	Environment Index	Sequence
4	1	HPPHPPHPPPHPPHPHPHPPPHHHPHH
4	1	HHHPPPPHPPPPHPHPPHPHPHPHPHH
4	1	PHHPPHPPPPHPHHHPPHPHPPHPPHH
4	1	HHPHPPHPPHHPPHPHPPPHPPPPHHH
4	1	PHPPPPHPHPHPPHPHPHPHPPPHHHH
4	1	HHHPPPHPPHPPPPHPHPPPPHPPPHH
4	1	HPPHHPHHHHHPHPHPPHPHPHPPPHP
4	1	HHPHHHHPPHPPHPHPPHHPPHPHPHP
4	1	HHPPHHPPHHPPHPHPHHHHPPHHHHP
4	1	HHHPHPHHHHPPHHPPPHPHPPHPPHH
4	2	PHPHPHPHPPPPHPHPPPHHPPHHHHH
4	2	PHPPHHHPPHPPHHPPPPHPPHHPPHH
4	2	HPHPHPHPPHPHPPPHHPPHHHHHHHH
4	2	HPPHPPPHPPHHPPPPHPPHHPPHHHH
4	2	HHPPHHHHHHPPHPHPPHPPPHHPPHP

4	2	HHPPHHPPPPPPHPPHPHPPHHHPPHP
4	2	PHPPPHPHPPPHPPHPHPHHHPHHHHP
4	2	HHHPPHPHPPHPPHHHPPHPPHHPPHH
4	2	HPPPPPHPHHHHPPHPPHPHPPPHHHP
4	2	HPHHHPHPPHPPHHPHPPHPPHPPPHP
4	3	PHPHPPHPHPHHHHHPPHHPPHHPPHH
4	3	HHHPPPHPPHHHHPHPPHPPHHPPPHP
4	3	PHHPPHHPHHHHPPHPPHHPPHPPHHH
4	3	PHPHPHPPHPPHHHHPPHHHPPHPPHH
4	3	HHPHHHPPHHPHPPHHPPPPHPPPHPH
4	3	HHHHHHHHPPHHPPPHPPHPHPPHHPH
4	3	PHPPHHHHPPPHPPPHPHPPHPHPHHH
4	3	HPHPPHHPPHHPPHPHHPPHHHHHHHP
4	3	HHHPPHPHPHHPPHPPPHPPHPPHHPH
4	3	HHHHPHPHPPHHPHPHHPHPPHPPHHH
4	4	HHHPPHPPHPHHPPHHPHPPHHPPHHH
4	4	PHPPHPHHPPPHPPHPPPHPHHHHPHH
4	4	PHHHHHPPHPPHPPHHPPHPHPPHPHP
4	4	PHHPHPPHPHPHPPHPHPPHPPPPHHH
4	4	HPHPPHHPPHHHHHPPPHPHPHPPHHH
4	4	HHHHHHPPHPPHHPPHPPPHPHPPHPH
4	4	PHPHPHPHHPHPHHHHPPHHPPHHHHH
4	4	HPPHHHHPHPPHPPHHPHPHPHPHHHH
4	4	PHHHHHHPPHPPHPHPPHHPPHPHPHH
4	4	HHHPPHPPHPHHPPHHPHPPHHPPHHH
4	5	HHHHPHPHPHPPHHPPHPHPPPHHPHH
4	5	HPHPHPPHHPPHHHHHPHPHPHPPHHH
4	5	HHHHHPHPPHPHPHHPPHHPPPHPPHP
4	5	HHHPPHPPHPHPPPPHPHPPPHPPHHP
4	5	HPHPHHHHPPHPPHPHPHHPPHPPHHH
4	5	PHPPHHPPPHPPHPPHPHPPPPHPHPH
4	5	PHPHPPPHPPHPHPPHPHPPHHHHHPH
4	5	HHPHPHPHHHHHHHPPHHPPHPHPHPH
4	5	HHHPPPPHPPHPHPPHHPPHPPPHPHP
4	5	HHPPHPPHPHPHPPHHHHHPHPHPPHP
3	1	HHHHHHHHHPHPHHPPHPPPHHPHPHP
3	1	HHPHPHPHPPHPHPHHHPPHPHHPHHH
3	1	HPPPHPHPHHHHHPHPHPPHPHPHPHP

3	1	HPPPHPPHPHHHHHHHPHHPPHPHPPH
3	1	PHPHPPPPHHPPHHHHHHPHHPPHPPH
3	1	HPHPPHHHPHPPHHHHPHPHPPHHPPH
3	1	HHPPPHPPPPHPHPPHHPHHHHHHPPH
3	1	HPPPPHPPHHPHPHPPHHHHHPPPPHP
3	1	PHHPPHPHHPPHPPHPPPPHPPHPHHP
3	1	PHPHPPHPHHPPHPHPHPHPPHPHPHH
3	2	PHHHPPPPHPPHPHPHHPHPPHPPHHH
3	2	PHHPHHHHHHPPHPPHHPHPPPPPHHP
3	2	HPHPHPHHHHPHHPPHPPPPPPPHHPP
3	2	HHPHPHPHHPPHPPHPHPHPHPPHHPH
3	2	HPHPPPPPHPPHPPPPPHHPPHPHPHH
3	2	PHHHPPHPPPPPHHHPPHHPPHHPHHH
3	2	PHPPPPHHHPPHHHPHPHHHHPPHPHH
3	2	HPHHPHHPPPHPPHHPPPHPPHPHHHH
3	2	HHHPHHHHPPHHPHPPHPHPHPPHPHP
3	2	PHPPHPHPPHPHHHHPPHHPHPHPPHP
3	3	HHPHHHHPHPHHPHPHHPPHPPPPHPH
3	3	PHHPHHPHPHPPPHPHPPHHPPHPPHH
3	3	HHPPHPPPPHHPPHPPPPHPPPPHPHH
3	3	HPHPHHHHPPPHHPPHPHHHHPHPPHP
3	3	HPHHHPPHHPHPPHPHPHPPHPPHPHP
3	3	HPPHPHPPHPPPHPPPHPPHPHHPHHP
3	3	PHPHPHPPHPHHHPPPPHPPHPPHHPP
3	3	PHPPHHPHPHHHPHHHHPPHHPHPPPH
3	3	HHHPPHPHHHHPPPHPHHHHHPPHPHP
3	3	PHPPHHPPPPHHHHPHPHPPHPPPHPPH
3	4	HHHPPHHPPHHHPHPHHPPHPHHPPHP
3	4	PHPHHHPHPHHHHHHHHPHHHPPPHHH
3	4	HPHPHPPHPPHHPHPPPHPPPHPPHHP
3	4	HPPHHPHPPHPPPHHPPHPPHHPPHHP
3	4	PHHPPPHPHHPHHHHPPPPPHPPHHHH
3	4	HPHHHPPHHPHPPHPHHHPHPHPPPHP
3	4	HHHPHPHHPPHPHHPPPHHPHPPHPHP
3	4	HPHPHPPPPHPHPPHHPPPHPHHHHHH
3	4	PHHPPPPHPPHHPPHHPPHPPHHHHPH
3	4	HHPHPPHHHPHPPHPHHPHPPPPHHPH
3	5	PHHPHHHPPPPPHHPPPHPHPPPHPHH

3	5	PHHHHHHPHPPPPPHPHHHHPHHHPHP
3	5	HPPHHHPPHHPHPPHHHHHPHHPPHPH
3	5	PHHPPPHPPPPHHPHPHHHHHHHPPHH
3	5	HPPHHPPPPHPHPHPPPPHPPHHPHHP
3	5	PHHPHPPHPHPPPPHHPPHPHPPHHPP
3	5	HHHPHPPHPHPHHPHHHHPPPPHPHHP
3	5	HPHHPPHPPHPHPHPPHPPPHHPPPHP
3	5	HPPPHPPPPHPHHPPHHPPPHHHHPHH
3	5	HPPPPPHHPHPPHHPHPPHPPHHHHHP
2	1	HHPHPPHPHPPHHHPPPPHPPHPPPPH
2	1	HHHHHHPHHHPHPHPPHHHHPPHPHHH
2	1	HPHHPPHPPHHPHPHPPHPHPHPHHHP
2	1	PHHPPHPPHPPHHPHPPHHHPHHPPHH
2	1	PHHPPHPPHPPPPHPPPHHPPHPHPHH
2	1	PHPHHPPPHPHPHHHPHPPPPHHHHP
2	1	HPPHPHPPHPPPHHHPHPPHPPHHPHP
2	1	HHPPHHPHPPPPHPPPHPPPPPHHPHP
2	1	PHPPHHHHPHHHHPHPHPPPPHPHHPH
2	1	HPPPHPHHPHHHPHPPHPPHPHHPHPP
2	2	HHHHPHHHHHPHPHPHHHHPPPPHHPH
2	2	HPPPHPPPHPPPPHHPPHPHPPPPHPHH
2	2	HPPHHHPHPPHHHHPPHPPHPHHHHHH
2	2	PHPPHHPPHHPPHPPPHHPHPHPPHPH
2	2	HHPPHHPHPPPHHHPHPHHHHHHPPHH
2	2	HHPPHPPHHHPPHHHHHPPHHHPPPHP
2	2	PHPHPPPPPPPHHPPPPHPPHHHHPPH
2	2	HHHHHHHPHHPHPPHHHPHPPPHHPHP
2	2	HHHPPHHHPPHHHHPPHHPHHHPHPHP
2	2	PHPPHHHHPPHPPHPPPHPPHPPPPPH
2	3	HHPHPHPHPPHPHHPPHPPHHHPPHPH
2	3	PHHHHHPPHPHHPPHHHPPHPHPPHPH
2	3	HPHPHPPPPHHHPPHHHHPHPHPPPHP
2	3	HPHPHHHPPHHHHHHHPHHHHHHHPHP
2	3	HPHPPHPHPPHPPHHHPHPHHPPHPHH
2	3	HPHHPPPHPHPPPHPHPPHHHHHPHHH
2	3	HPPHPPPHHPPHPPHHPHPPPPHPHHP
2	3	HPPHPHHHHPHPPHPHHPHPPPPHHPH
2	3	HHPPPHPHHHPHPHPPHPHHPPHPHPH

2	3	HPHPHPPHPHPHHHPPHHHPHHHPPHP
2	4	HHPHHPHHHHPPHPPPPHPHPPPHHPH
2	4	HHHPHHHHPPHHHHHPHPHHPPHHHHP
2	4	PHHHPHHHHPPHPHPHPPHPPPHPHPH
2	4	HPPHPHPHHPPHPPHHPPHPHHPPHHH
2	4	PHPHPPPHPPPPHPHHPHHHHPHHHPH
2	4	HPHPPHHHHPPHPHHHPPHPPHPPPPH
2	4	HPPHPPPPPPPHPPPHHHHPPPHPPPH
2	4	PHHPPHHHHHHHHPPPHPPPHHPHPHH
2	4	HPHPHPPHHHHHPHHHHHHHPPHHPHP
2	4	HPPHHHPHPPHHHHHHPPHHPHHPPHH
2	5	HHHPPHHHPHHHHPPPPHPPPPHHPHP
2	5	PHPPHPPHHPHHPPHPHPPHHHHPPHP
2	5	PHHPPPPHPPHHHPPHHHHHHPHPPHH
2	5	HHHHPPHHPHPPPHPHHHPPHHHHHHH
2	5	PHPPHPPHHHPPHHHPHPHHPPHPPHP
2	5	HPPPHHPPHPPHHHPHPPHHPPPHPPH
2	5	HPPHHHHPHPHPHPHPHPPHPHPHPHP
2	5	HHHHPPHHHHHHPPPPHPHPPPPHPHP
2	5	HPHHPHHPPHHHHHHHHPHHHHPPHHH
2	5	HHHHHHPHHHHHPPHPPHHHPPPHHPH
1/0	1	PHPPPHHPHHHHHPPHPPHHHPPHPHH
1/0	1	HHPPHHHPPHPHHPPPPHPPHPPPPHH
1/0	1	PHPPHPHHPHHHHHPPPHPPHPHHHHP
1/0	1	HHPHHHHPPHHHHPPHPHHHHPPPPHP
1/0	1	HHHHPHPHHHHPHHPPPPHPPPHPPPH
1/0	1	HHPPHHPHPPPHPPPPPHPHHHPPHPH
1/0	1	PHHHHPHPPPPHHPPPPPHHHPPPPHP
1/0	1	HHHPHPPHHHHHPHHHPPHHHPPPHHH
1/0	1	HHHPPHPPHPPHHPPHPHHPHPPHPHP
1/0	1	PHHPPHHPHPPHPHHPHPHHHHPPPHH
1/0	2	HPHPPHPPHPHPHHHHHHPHHPPHHHP
1/0	2	HHPHHHHPPHHHPPPHHPHHHHPPHHH
1/0	2	PHPHPPHHHPPHHPHHHHPHPHPPPPH
1/0	2	PHHPPPHHHHHHPHHHHHPHHHPPHPH
1/0	2	HHPHPHHHHHPPHPPPHPHPPPHPPHH
1/0	2	PHPHPPPPHPPPHHPHPHHPPHPHHPH
1/0	2	HHPPHPPHPHHPPHHPHHHHPPHHPHH

1/0	2	HHHPHHPHHPPHHHPPHPPHPHPPPHH
1/0	2	HPHHHHHHHHHPPHPHPHPHHHHPHHH
1/0	2	HPPHPHHPPHHHHHHHPHHPHHHHPHH
1/0	3	HPPHPHPPHHHPPPHPPHPHPPPPHPHH
1/0	3	HPHHPPHHHHHPHHHHPPPPHPPPPHH
1/0	3	HPPHPHHHPHHPPHPHPHPHHPHHHHH
1/0	3	HPPHHPHHPHPHPPPHPPPPHHHHPPH
1/0	3	HPHPPHPPHHHPPPHPHHHHPHHHHPH
1/0	3	HPPHHPHPPHPHHPPHPPHHHHPHPPH
1/0	3	HHPPHHHPPHPHPHHPHPHPPHPHHPH
1/0	3	HHHHPHPPPHHPHHPPHPHHHHPHPPH
1/0	3	HHPPHPPHPHHPPPHPPPPHPPPHPPH
1/0	3	HPPPHPPPHPPHPHPPHHHHPHPPHPH
1/0	4	PHHPPHHPPPHPPHPPHPPHHHPHPHH
1/0	4	HPHPHHHHPHPPHPPHPHPPPPHPPPH
1/0	4	HHPHPPHHPHHPPHHPPHPHHHHPHPH
1/0	4	HHHPHHPHPHPPPPHPPHHHPHHHPPH
1/0	4	HPPHPHHHPHPHPHHPHHPPHPPPHPH
1/0	4	PHPPHPHPPHHPHHHHHHPHHPPHPHH
1/0	4	HPHHPPHHPPHHHHPHHHPPHHPHPPH
1/0	4	PHPHPPPPHPPPPHPPPPHHPPHPPHH
1/0	4	HPPHHHPHPHHHHPHHHHHHPPPPHPH
1/0	4	PHPPPPHPHPHHHHPHHPHHPPHHHHP
1/0	5	HHHHPHPHPPPPHHPPPHHHPPHPHHH
1/0	5	HHPHPHPHHPPPPHPPHHHPHHHPPHP
1/0	5	HPPPHPPHPPPHHHPHHPPHPPHHPPH
1/0	5	HPPHHPPHPPHHPHPPPPHPHHPHPHH
1/0	5	PHHHHPHPPPPHPPHHHPHHHPPPPHP
1/0	5	HPPPHPPHPPHHPHPHPHPPHPPPHHH
1/0	5	PHHHHPHPPHHHPHHPPHHHHHHPHHH
1/0	5	PHHHHPPHHPHHHPPHPHPHHPHHHHH
1/0	5	HHPHHPHHPPPHPPHHHPPPHHHPPHH
1/0	5	PHPHPPHHPHHHHHHHPPHPHPHHPPH

Table 8: The sequences used as crowders in the results behind figures 9.

Inverse melting temperature quartile	Environment Index	Sequence

4	1	PPHHPPHPPHPHHPPHPHPPHPPHHHH
4	1	HPHPPHHHPHPPHHHHPPHPPHPHPPH
4	1	HHHHPPHPHPPHPHPPHHPHPHPPHPH
4	1	HPHPHPPHHHHHHHHPPHPHPHPHPPH
4	1	HHHPHHHHPHPHPPHHPHPHPPHPPHH
4	1	HHHHPPHHPHPHPHPPHPPHHHHPHPH
4	1	HPPHHHHHPHPPHPPHHPPHPHHPHHH
4	1	HHHPHHHPHHPPPHPPHPPHHPPHPHP
4	1	HHHHPHHHHHPHPHPHPHHPPPHPHPH
4	1	HPHPPHPPPHPHPHPPHPHPHHHHHHH
4	2	HHHPPHPHPPHPPHHHPPPPHPPHPHH
4	2	HHPPPPHPPHPHPHPPHPHPPPHHPHH
4	2	PHPPHHHHHHPPHHPPHPHPPHPHPPH
4	2	HPPHPHPHPPHHPPHPPHPHPPHHHHH
4	2	HHPPHHPHPPHHPPHPPHPHPPPPHHH
4	2	HHHPHHHHPPHPPHPHPPHPHHPHPHP
4	2	HPHHHHPHPPHPPHHPPHPHPPHPPHH
4	2	HHHHHHHPHPHPPHPPPHHPHPPHPHP
4	2	HHPPHPPHPHHPPHHPPHPHPHPPHHH
4	2	HHHHPHHPPHPPHHPHPPHPPHHPHPH
4	3	HPPHHHHPPHPPHPHPPHPHHPPHPHP
4	3	HPHPHPHPPHHHHHHHPPHPPHPHPHP
4	3	HPPPHPHHPPHPPHPHPHPHPHHHHHH
4	3	HHHHPPHPPHPHPHPHPHHPPHHPHHH
4	3	HHPHPHPPHHPPHHPPHPPHPHPPHHH
4	3	HHPHPHHHHPPHPPHPHHPHPHPPHHH
4	3	HHHHPPHPPHPHPPHHPHPHPHHPHHH
4	3	HHHPHHHHHPPHPPPHPPHPHPPHPHP
4	3	HHPPHPPHPHPPHHPHPPHHPHPPHHH
4	3	HHPHHPPPHPPHHPPHPHPHPPHPPHH
4	4	HHPHHHPPHPPPPHPHPHPPHPHPPPH
4	4	HHHPHPPHPHPPHHHHPHPPHPHHHHP
4	4	HHPHPHHPPHHPPHPPHHHPPHPPHHH
4	4	HHPHPPHPHPPHPHHHHHHPPHPHPPH
4	4	HPPHHHPPHPPHHPPHPHPPHHPPHHH
4	4	PHPHPPHPPHHPPHPPHHPHPPHPHHH
4	4	HPPHHHHHHPPHPHPPHPHPHHHPPHP

4	4	HPPHPHPHPHPHHHHHHHPPPPHPHPH
4	4	HPPHPHPHPPPHPPHPPHHPHHHPPHH
4	4	HPHPPHHPPHPHPPHPPHHHHPPHPHH
4	5	HHHPPHHHPHHPHPPPHPPHPHPHPHH
4	5	PHPPHHPPHPHPPHPHPPHHHPHHHHH
4	5	HHHHPHHHHPPHHPPHPHPPHHPHPPH
4	5	HPHPPHPHPPHPPPHHPHHHHHPPPHH
4	5	HHHPHPHPHHHHPHPHPHPPHPPHPHH
4	5	HHPPHHPHPPHHHPPPPHPHPPHPPHH
4	5	HPHHHHPPHPPHPHPHPPHPHHPHHHH
4	5	HHHPPHHPHPPHPHPHPHPPHPHPPHP
4	5	PHPPHHHHHHPPHHPPHPHPPHPHPPH
4	5	HHHHHPHHPHPPPHPHPHPHPHPPHPH
3	1	PHHHHHPHPHPPHHHHHHPPHPHPHPH
3	1	HPPHPHHPHHHPHHHHPPHHPPPHPHP
3	1	HHPHPPHHHHPPHPHHHPPHPPHPHHP
3	1	HHHPHHHPPHPHPHHHHPPHPHHPPHP
3	1	HHPHHHPPPHPPHPPHHHHPHHHPPHP
3	1	HPPHHHPHPPHHPPPPHHHPPPHPPHP
3	1	PHHPHPPHPPPHPHPPPHHHHPHPHHH
3	1	PHHPPHPPPHHPHHHPPPPHPPHHPPH
3	1	HHPHPHHPPPHPPHPHPHPPPPHPHHH
3	1	HPPPPHPPPHPHPHHHHPPHHHPHHHH
3	2	PHHHHPHHPPHHHPHPHHHHHPPHPHP
3	2	HHPPHHHHPHHPHHHHPHHPHPPPHHP
3	2	HPPHHHHPHPHHHPPHPHHPPHPHPHH
3	2	PHPPPHPPHPPHHHHHHPHHHPHHPPH
3	2	HPPHPHHPPPPHPPPPHPPHHPPPPHP
3	2	HHPHPPHPPHHPHPHPPHPHPHPHHPH
3	2	HPHPPPHHHHPPPHPHPHHHHPPHPHH
3	2	HPPHPPHHHPHPHHHHPPHHHPPHPHH
3	2	PHPHPPHHPPHHHPHPPPPHHHPHPPH
3	2	PHPHPPHPHPPPHPPHPHHHHHHPHHH
3	3	PHPHPHHHHPHPHHPPPHPPHPHPHHH
3	3	HPPHPHHHHPHHHHHPPHPHPHHPPHH
3	3	HPHPPHHHPPPHHPPHHPHPPHPHPPH
3	3	HPHHHPHPPHPPHHPHHPPHPHPPPHP
3	3	PHPHPPPPHPHPPHPHPPPPHHHPHHH

3	3	HHHPHHPHHHHHPPHHPPPPHPPHPPH
3	3	PHPPHHHPHHPPHPPHHPPHHPPPPHP
3	3	HPPHPHHHHPHPPHHPPPPHPPPPHHP
3	3	PHHPPHPPHPPPPHHPHHHPPPHPPHP
3	3	HHPPHHPPHHHPHPPPHHPPPHPPHHH
3	4	HPPHHHPHHPPHHHPHHPPPHHPPHPH
3	4	PHPHHPPHHPHPPHPHHHHHHPPPHHH
3	4	PPHHHHHPPHPHPPHPPPPPPHHPPHP
3	4	HPPHPHHPPHHHHPHHHPHHHHPHHHH
3	4	HHPPHHPPHHPPHPHHPHPHPPHHPPH
3	4	HHHPPHHHPHPPHHHHPHHPPHHHHHP
3	4	PHHPHPPHPHPHPHPPPPHPHPPHPHH
3	4	HHHPHPPHHHPHPPHHPHPHHHHPPHP
3	4	HPPPPHPPHPPPHHPHHHPPHHHPPHP
3	4	HHPPHHPHPHHHPHPPHPPHHHHPHHH
3	5	HPPPHPHPPHHHHHPHHHPHHHPHPPH
3	5	HHPHHPPHPPHHPPHHPPPHPPPPHHH
3	5	PHPHHPHPPHPPHPPHHPHPPHPPHHH
3	5	PHHPPPPPHHPHHHPPHHPPHPPHPPH
3	5	HHHHPPPHHPPPPHPHPPHHHPHPPHP
3	5	PHPPPHPPHHHPPHPHHHPPHHHPHHH
3	5	PHPPHPHHHPHPPHHPHPHHHHPPPHP
3	5	HPHHPHPPHPHPHHHHPHHHHPPPPHP
3	5	PHPPPHHHHHHPHPPHPHHPHHPPHHH
3	5	PHHHHPPPPHPPHHPPHHPHPPPHPPH
2	1	HPHPPPPHHHHPPHHHPPHHPHPPHHH
2	1	HHPPHHPHPPHHHHHPHPHHHHHPHHP
2	1	PHHPPPPPHPHPHPPHPHPHHPPPPHP
2	1	HPPPHPHHHHPPPPHPPPPHPHPPHHP
2	1	PHPPHPHHPPHHPHHPPHPPHPHPPHP
2	1	HPPHHPPHPPHHPPHHHPPHHHPHHHH
2	1	HPPHPHPHPHPPHPHPHHHPHPPHPHP
2	1	PHPPHPHPPHPPHHPHHPPHHPHPPHP
2	1	PHPPHPPHPHHHPHPHPPHPPPHPHPH
2	1	PHHHPHPPHPPHHHPHHPPHPPHPHHH
2	2	HPPPHHPHPHHHHPPHPHHPHPPHHHH
2	2	PHPHHHPHPPPPHPPHHPHHPPHHHHH
2	2	PHHPPHPHPPPPHHPHPHPPHPHPHPH

2	2	HPHHPPHHHHHPHPPPHPPPPHPHPHH
2	2	PHPHPPHPHHPHPHPPHHHHPPHPHHH
2	2	PHHHHPHPPHHHHHPPHHPHHHHPPHP
2	2	PHPPHPHHPPHPPPPHPPPPHHPPPHH
2	2	HPPPPHHPPHPHHHHPHPHHPPPHHHH
2	2	HHPPPHHHPHPHHPHHHHHPHHHHHHH
2	2	HPHPPHHHHHPHPPPPPHPPPHPHPPH
2	3	HPPHPPPHHPHPPHHPPHPPPPHHPHH
2	3	HPPPHPPHHHHPPPHPHHPHHHHPHHP
2	3	HHHHPHHHHHHPPPPHPPHHPHPHPPH
2	3	HPHHHHPPPHPHHHHPHHHHHPHPHPH
2	3	HHPHPPPPHPHPHHPPPPHPPHHHHHH
2	3	HPPPHPPHHHPHPHHPPPHHPPHHPHH
2	3	HPHPPHPHHHHPHHPPHHPHPPHPHHP
2	3	PHPPHPHPHPHHHPHHPHPPPHPPHHP
2	3	PHPPPHPHPPPPHPPHHHHPHPPHPPH
2	3	PHPPHPHPHHPPPPHPHPPHHPPHPHP
2	4	HHPHPHPPHPHHPHHHHHHPHPPHPHH
2	4	HPPHPHHPHHHHHPHHHHHHHPPHPHP
2	4	HPHPPPHPPPPHPPHPPHPHHPHHPHP
2	4	PHHHHHPPHHHPPHPHPPHPPPPPHHH
2	4	PHHHHPHHPHHPPPHHPHPPHPHHPHH
2	4	HHPPPPHPPHHPPHHHPHHHHPPHHPH
2	4	HPHHPPHPHHPHPPPHPPHPPPPHPHP
2	4	HPPPHHHPPHPHPHHHHPHHPHHHPHH
2	4	HHPHHHHHHPHHPHPHHPHPPPPHHHH
2	4	PHPPHHPPPHPPPPHHPHPPHHPPHHH
2	5	HPPPHPHPHHHHPHPPHPHHPPPHPHH
2	5	HPPHPHHHHHPHHHPPHHPHHPHHHHH
2	5	HPPHHPPHPHHHHHHPPHPHPHHPHHH
2	5	HPPHHHPPPHHPPHPPPPPPHHPPHPH
2	5	HHHPHHPHPPPPHPPPPHHHHPHHPHP
2	5	HHPPHHPPPPHHHHPHHPHHHHHPHHP
2	5	HPPHPHPHHHPPHHPPHPPPPHHHPPH
2	5	PHHPHHHHHPHHPPHHHHPPHPPHHHH
2	5	HPPHPHPPHPHPHHHPPHPHPPPHPHH
2	5	PHHPHHHHHPHPHHHHHPPHPHHPPHH
1	1	HHHPHHHPPHHHHPHHHHPHHHHHHHP

1	1	HPHPPHPHPHHPHPPPPPHHPHPHHHP
1	1	PHPHHPHPPHHPPHHPHPPHPPHHHPH
1	1	PHPPHHPHHPHHHPHHPPHHHHPPHPH
1	1	HHPHHHPPHPPPHPPPPHPHHPHHPHP
1	1	HHPHPPHHPHHHHPPPHPHHHHHPHHP
1	1	HHPPPHPHHHHHHHHPPHPPHHHHPHH
1	1	HPPHPHHPPPHPPHPPHPPPPHHHPPH
1	1	HHHPHHHHPPHHPPHHHPHHHPPPPHP
1	1	HPHHPHPPHHPPHHPHHPPHHPPHPHP
1	2	PHHPHHHHHPHPHHPPPHPPPHPPHHP
1	2	HPHHPHHHHHPHHPHHHPPPHPPPPHP
1	2	PHPPPPHHHHHPPHHPHHHHHPHHPHP
1	2	PHPPHPPHHHPHHHHPPHHPHPPHHPP
1	2	PPHHPPHPHPPHPHPHPPPPHPHPPHH
1	2	HHPHPHHPHHHPHPHHPPPHHPHPPHP
1	2	HPHHPHPHHHHPHHPPHHHPHHPPPHP
1	2	HPPHHPPHHHPHHHHHHHPHPPPHHHH
1	2	HHHHHPHHPHPPHHHPHHHHHHHHPHP
1	2	HPPPHPHPHHPPHPHHPHPPHHPHPHP
1	3	HHPHPPHPHPHPHHPHHHPPHPPPPHP
1	3	PHHHPPPPHHPPHPPHPPHPHPPHHPH
1	3	PHPHHPPHPHHPPHHPPPHPHPPHPHH
1	3	HHHPHHHHHPHHPPHHHHPHHHPHPHP
1	3	HHHHHHHPHHPPPPHHHPPHHHHPPHHH
1	3	PHHHPHPPHPHPPHPPHPPPPHHHHPH
1	3	HHHPPHHHHPHHPPHPHPPHPHHHPHH
1	3	PHHPHHHHHPHPHHHHPHHHHPPPHHP
1	3	PHHPHPPHPHPHHHPPHPHHHHPHHHH
1	3	PHHHHPPPPHPHHPPPPPHHHHPHPPH
1	4	HHHPPPPHPHPHHHPPHHHHPPHPPPH
1	4	HPPHPPPPHHPHPPHHHHPHPPPHPPH
1	4	HPHPPHPPPPPHHPHPPHPHPHPPHHP
1	4	PHHPHPPHPPPHPPHPHPPHHHHHPPH
1	4	HPHPPHPHHPHHPHHHHPPPHPPPPHP
1	4	PHHHPHPPHHHHHHPHHPHPPHHPPPH
1	4	PHHPHHPHHHPHPPHHHPPPHPHHPPH
1	4	PHPPHHHHHHPPHHPHPPHPHPHPHPH
1	4	HHHHHHPHPHHPPHHHPPHHPHPPPHP

1	4	HHPPHPPPPHHHHPHHPPHHHHHHPHH
1	5	HHPPHPPHHHHHPHPPHHHHPPHPPPH
1	5	PHHPHHPHPHPPPPPHPHPPHHPPHHH
1	5	HPPPPHPHPHHHHPHPHPPPHPHHPHH
1	5	HPHPPPPHPHHPHHPHPPPHPHPPHPH
1	5	HPHPPHHPHPPPHPPPPHHHHHPPPHP
1	5	PHHHHPPHPPPPHHHPPPPPHHHPPHP
1	5	HPHHHHHHHPPHPHPPHPPHPPHPPPH
1	5	HHPHPHHHPHHPPPHHPHHPHPPHHHH
1	5	HPPPHPHHHHHHPPPHPHPPPPHPPPH
1	5	HHHHPHHPPHHHPPPHPPHHHHPPHHH
0	1	PHPPHHPPPHHHHPPPPPHPPPPPHPH
0	1	HPPHHHPPPHPPPPPPPHPPHPPPPHP
0	1	HPPHPHPPHPHPPHPPPPHPPPPHPPH
0	1	PHHHPHHPPPHPHHPHPHPPPPHHPPH
0	1	HHPPHPHPPHPHHHPHPPPPPHPPPHP
0	1	PHPPPHPPPHHPPHHHPPHHPHHHPPH
0	1	HPHHPPPHHHPHPPHHPPPPHPHPHHP
0	1	PHHHHPHHPPPPPPPHPHHPPHHPHPH
0	1	HHPPPPPHHHPHHHHHPPPHHHHPHHH
0	1	HPPHHPPHPPPPPHPPPHHPHPHPPHH
0	2	PHPHPHPPHPHPPPPPPHPPHPPHHHP
0	2	PPHPHHHHPHPPPPPPPHPPPPHHPHH
0	2	HPPHPPHHPPHHPHPPPPPPPPHHPHPP
0	2	HPPHPHHPPPHHPPPPPHPHPHPHPPH
0	2	HPPPPHPPPPPPPHHHHHHHHHPHHPH
0	2	PHPPHHHHPPPPPPPPHHHPHHHHPHPP
0	2	HPPHPHPHHHPHHPPPHPPPPHPPPHP
0	2	PHPPHHPHHHHHPPHHHHPPPPPPPHH
0	2	HPPHHHPHPHHHHPHPPHPHPHPPPHP
0	2	PHHHHPHHHHHPPHHHPPPHHPPPPPH
0	3	HHPHHHHHHPPPPPPPHPPPHPHHPHP
0	3	HHHPHPHHPPHPPPHPPPPHPPPPPHP
0	3	PHPHPPHHPPHHPPHPHPPPPHPPPHP
0	3	HPPHHPPPHPPHPHPPHPPHPPPPPHP
0	3	HHPPHHPHPPPPPHHHPPHPPHPPPHP
0	3	PHPPPHPHPHHPHPHPPPPHPPPHHHP
0	3	HHPHPPHHPHPHPHPPPPPPPHHPHPP

0	3	HPPPHPPPPHPHPHPPHPHHPPHHHHHP
0	3	HPPHPHPPPPPHHPHPPHPHPHPHHPP
0	3	HPPHHPPPPPPPHPHHPPPHPHHHPPP
0	4	PHPHPPPPPHPPPPHHHPPHPHHHPPH
0	4	PHPPPHPPPPPHPHPHHPHHPPHHHPP
0	4	PPHHPPHPHPPHPPPHPPPHPPHHPHH
0	4	HPPHHHPPHHPPPHPPHPPPPHHPHPP
0	4	PPHPHPPHPPPPPPPHHHPHPPHPPHP
0	4	HPPPPHPPPPPPPPPHPHHPHPHPPHP
0	4	PHPPPPHPPPHPHHPPPHPPPPPHHPH
0	4	HHHPPPPPPPPHHPHHPPHHHPPHHHP
0	4	PHPHPPHPPPPPHHPPPHPHHHPHPPH
0	4	HPPPHPPPHPHPHPPHPHHHHHPPPHH
0	5	PHPPPPPPPPPHPPPHHPPHPPPPHPHH
0	5	HPHPPHPPPHHHHHHPPHHPPPPPPPH
0	5	HPHHPPHPPPPPHPPHPHPHHPPPPPHP
0	5	PHPPPHHPHHPPHPHPPHPPHPPPPPH
0	5	HPPPPPPPHHHPPHPPPPPHHPPHHPH
0	5	HHHPHPPPPPHPPPHHHHPHPHHPPHH
0	5	HPPPHPHPHPPHPHHHPPPHPPPPHPP
0	5	PHPPPHPPPPPPPHHPHPHPHPHPPHP
0	5	HPPPHHPHPPPPPHPHPPHHHPHPHPH
0	5	PPHPHHPHPPHPPHPPHPPHPPPHPPH

Figures 14-15

Table 9: The sequences used as test proteins in the results behind figures 14-15.

Index	Sequence
1	PHHHHHHPHPHPHPPPPPHPPHPHPHP
2	PHHHPHPHHHHPHPHPPPHPPPHPPHH
3	PHHPPHHPPHPPHPPHHPPPPPHPPHP
4	PHPPHPHHHHPHPHPHPPPHPPHHHPH
5	HPHHHPPHHPPHPHHHHPPHPHPPPHP
6	HHHPHPPHPPHHPPPHPHPHHHHPPHP
7	PHPPPHPPPPPHPPHPHPPPHHHPPHH
8	HHHHHHPHPHPPHPHPHPPPPHPHPHP

9	HHHPHHHHPHPPPHPHPPPHPPHPHPH
10	HHHPHPHPPHPHPHPPHPPHHHHPPHP
11	PHPPPPPPHPPPPHPHPPHPHHHPPHH
12	PHPPPHPPHPHPPHPHHHHPHHHPPHH
13	HPHHHPHPPHHPPHHHHPPHPHPPPHP
14	PHHHHHPPHPHHPHPHPPHPPPPPPHPH
15	HPPPHPHPPHHHPHHHHPPHPPHHPHP
16	HPPPHPHPPHHHPHHHPHPPHHHPPHP
17	HHHHPPHPHHPHPHPPHPPHHPPHPPH
18	PHPHPHPHPPHPHPHPPPPHHPHHHHH
19	PHPHPPHPPPPPHPPPHHPPHPHHHHP
20	HHHHHPHPHPHHPPHHPHPHPPPPHHHH
21	PHPPHHPPHPPHPHPPHHPPHPPPHHH
22	PHPPHHPPHPPHHHHHHPPHPHPPHPH
23	HHPPHPHPHHPPPHPHPPHHHPHHHHP
24	PHPPHHHPHHPPHHPPPHPPHPPHHPP
25	HHHPHPHPPHHPPHPPPHHPPHPHHHH
26	HHHHPHPPHPHPPHPHPHPPPHPPHHH
27	PHHPHHHHHHHPPHHPHPHPHPPHPHP
28	HHHPPHHPPPPPPHPPPPHPHPPHPHP
29	HHHPPHPHPHPHPPHPHHPPHPPHHHH
30	HHHPPHPHPPHPHHHHPPHPPHHPPHP
31	HPHHHHHPPHPHPPHPHPPHPPPPHHHP
32	HPPHPHPPHPPHHHHPPHPHHPHPPHH
33	HPPHPPHPPHHPPHPHPPHPPPPHPHH
34	HHPHPHHPPPPPHPPHPHPPPPHHHHP
35	HHPHHPPHHPHPPPHPPHPPHHPPHHP
36	HHHHPPHPHHHHPPHHPPPHPPHHPPH
37	PPHHHHPPHPHPPHPPPPPPPHHPPHH
38	HHHPPHHHHPHPHHPPPHPHPPHPPHH
39	HHPPHPHPPHPPHPPHPHPHPPHPPHP
40	HHPPPHPPHPPHHPPHPHPHPPPPHHH
41	PHPHHPPHPPPPHPPPHPPHPPHHPHH
42	HPHPPPHHHHPPHPPHPHPHPHPHHHH
43	PHPHPHPPPPPHPPHPPPHPHHHPHHH
44	PHPHPHPHHHHPPPHPHPHHHHPPHHH
45	HPHPHPPPPPHPPHPHPHPPHHHHHHP
46	PHPHPPPHPPHPHPPPHHPPHPHHHHH

47	PHPHPHPPPPHPPHHPPPHPHPHHHHH
48	HPHPHPHHPPPPPHPPPHPHHHHPHHH
49	HHHHHPHPHPHPPHPHHPPHPPHHPHP
50	PHPHHPPHPPHHPPHPHPPPPHPPHHH
51	HPHPPHHPPHPHPPHPPHHHHPPHPHH
52	HHHHPHPHPHPHPHPPHPHPHPPPHHH
53	PHPPHHPPHPPHPPPHHPPPPHPPHHH
54	HHPHHHHPHPPHPHPPHHPPHPPHHHH
55	PHPPHPHHHHHPPHPHPPPPHPHPPHP
56	PHPHHHHPPHPPHPHPHHHHPPHPPHH
57	HHHPHPHHHPHPPHPPPHPPPHPHPHP
58	HPHHHPPHPPHPHHHHPHPHPHPPPHP
59	HHHPHHPHPPHPPHPPPHPHPHPPHPH
60	HHHPHHHPHPHPPHPHPPHHPPPHPHP
61	PHPPPPHHPPPPHPPHPHPPHPPPHHH
62	HHHPPPHPPHPPPPHHPPPPHPPHPHP
63	HPHHPPHPPHPPPHPHHPPHPHPPHHH
64	HHHPPPPHPHPPHPPPPHHPPPHPPHP
65	HHHPPPPHPHPPPHPPHHPPHPHPPHP
66	HPHPPHHPPPPPHPPHPHPHPHHHHHP
67	HPPHPHPHPPHPHPPHPPHHHHHPPHH
68	HPPHPPHHPPHPPHPHHPPHHPPHHHH
69	HPPPHPHHHHHHPHPHPHPPHPHPPHP
70	HHHHPPHPHHHPPHPPPHPHPPHPHHH
71	PPHHHHHHPPHPPHHPHPPHPPHPPHP
72	PHHHHHPHPHHPPPHPPPHPHPHHPPH
73	HHHPPHPPPPHHPPHPPHPHPHHPPHH
74	HHPPPPHPPHPHHPPHPHPPHPPPHHH
75	PHHHHPPHPHHPPHHPPPPPHPHPHHP
76	PHHHHPHHHHHPHPPHPPPPPHPPHPH
77	PHHHHHPHPPPHPHPPHPPPPPHPPPH
78	HHPHHPPPHPPHPHPPHPHHHHPHPHH
79	HHPHHPHPPPHPPHPPPHHHHPHPPHP
80	PPHHPPHPPHHPHPPHPHHHHPHPPHP
81	HHPHHHHPHPHPPPHPPHPPPHHPPHP
82	HHPPHPPPPHHPPHHPPHPPPHPPHHH
83	PHPPPHHPPHPPHHPPHHHHPPHPHHH
84	PHPHPHPPHHPHPHPPHPHPHPPPHHH

85 PHPPPHPPHHPPHPPHHHHPHPPPHHH
86 HHHHPPHPHPPHHPPHPPHHPHPHPPH
87 PHPHPHPPHHHHHHPPHPHPHPHPPHP
88 HPHPHPPHPPHHHHHHPPHPHHPHPHP
89 PHPHHPHPHPHPPPPPPPHHPPPHHHH
90 PHPHPPHPHPPHPHPHPHPPHHHPHHH
91 PHPHPPPHPPHHPHPPHPHPPHHHHHH
92 HHHPHPPPHPHPPHHHPPPHPPPHHHH
93 HHHHHPPHHPPPHPHPHPHHPHPHPHP
94 HHHHHPPPPHHPHPHPHPPHPHPHPHP
95 HHHHHHPPHPHPPHPPPHPHPPHPHHH
96 PHPPHHPPHPPHHHPHHPPHPPHPPPH
97 PHPPHPHPHPPHPPPPHHHHHPHHPHP
98 HHPPHPHPHPPHHPPHPHPHPHHHHHH
99 PHPPHPHPPHPHHPPHPPHHHPPPHHH
100 HHPPHHPPHHPPHPHPHHHHPPHHHHP
101 HHHPHHHHPPHHPPPHHPPPPHHPHPH
102 PHPPHHHPPHPPHPHHPPHHPPHPPHP
103 HHHPPPHHPHHHPHPHPPHPPPHPPHP
104 HHHHPHPPHPPPHHPPHHPHPHPPHHH
105 HHHPPHHPHPHHPPHPPHPPHHPHPHP
106 PHPPPPHPPPHHPPHPPHHPPHHHHPP
107 PHPPPPHHHPPHPPHHPPHPHPPHHHP
108 HHHHPPHHHPPHPHHPPHPHPPHPHHH
109 HHHPPHPHPPPHPPHHPPHPHPPPPHP
110 HPPHHPPHPPHPPHPPHPPPHHHHHHP
111 HHPHPPHHHHPPHHPPHPPHHHPPHPH
112 HPPHPPHPPHHPHPPHPHPHPHPHHHH
113 HPPHPPHPPHHPPPPHPHPPHPPHPHH
114 HHHHHPHPHPPHHPPHPHPPHPHPHPH
115 HHHPHPHHHHPPHHPPPHPHPPHPPHH
116 PHHHHHPPPHPPHPPPHHHPHPPPPPH
117 HHHPPHHPPHHHHPHPPHPHHPHHPHH
118 PHHHHHHHPHPHPHPPHPHHHPPPPPH
119 HHHPPHPPHHPPHHPPHHHPPHHPPHH
120 PHHHHHHPPHPPHPPHPPPPHPHPPHH
121 PHHHHHPHPPPPHPHPPPPPHPPHPHP
122 PHHHHPPHPHPPPHHPPPPPHPHPPHP

123 HPPPHPPHPPPPPPHPPHPHPHHHPPH
124 HHPHHPHPHHPHPPPHPHHHHHHHHHH
125 PHHHHPHPHPPHPHPHPHPHPPHPHPH
126 HHHHHHPPHPHPPHPHHPPPPHPPPHP
127 HHPHHHHHHPPPHPPPHPHPHPPHHHP
128 HHPPHHPHPHPPHHHPPHPHPHHPPHH
129 HHPPHPPHHPHHHHPPHHPHPHPPHHH
130 PHHHHPHPPHPHPHHHHPHPPHPHPPH
131 HHPPHHPPPHPPHPPHPPHPPHHHHPP
132 PHPPPPPHPPHPHPHHPHPPHHHHHPH
133 PHPHHPPHPHPPPHPPHPPPHHHPHHH
134 PHPHHPPHHPPHPPPHPHPPHHHHHPH
135 HHHHPPHPHHPHPHHPHPPHHPPHPPH
136 HHPHPHHHHPPPPHPHPPHPHPPHHHH
137 PHPHPHHPPPPPHPHPPPHPHHHHHHP
138 PHPHPHPPPHPHHPHHHPHPPHPPHHH
139 PHPHPHPPPHPPHPHPHPPPHHHHHHH
140 PHHHPHPPHPHPPHPHHHHHHPHPHHH
141 HPHPHPPHHPPHHPPHPHPHPHHHHHH
142 HPHPHPPHPHPPHPPHHHHHHHPHPHP
143 PHPHHPHPHPHPHHHHHHHPPHPPHHP
144 HHHHHPPHHPPPHPHPHPHHPPPHPHP
145 PHPHPPPHPHPPHPHPHHPHPPHHHHH
146 HHHHHPPPPHPPHPHPHPHHPHPHHHP
147 HHHHHHPPHHPPHPHPHPHHPPPHPHH
148 HHHHHHPPHHPHHHPHPPHPHPHPPHH
149 HHHHHPPHHHPHPPHHPPHPHPHPHHP
150 PHPHHPPHPPHHPHHPHPHPHPHHHHH
151 PHHHHHPHPPHPHPHHHHPHPPHPHPH
152 HPHPPHPHHPPHPHPHPHPPHPPHHHH
153 PHPPHPPHPPPPHPHPPHHHHHHPPHH
154 HHHHHHHHPHPPHHPHPHPHHHPPHPH
155 PHHPHPHPHPHHHPPHHHHHHPPHPHH
156 HHPPHPHPHPPHPHHHPHHPPHHHHHH
157 HHPPHPHPHPPPPPPPHHPPPHHHHHP
158 PHPPHPHPPHPPPHPPHHPHHPHHHPP
159 PHPPHPHPPHPHHPPHPHHPHPHHHHH
160 PHHPHPHPPHPHPPPHHPPHHHPHHH

```
161 PHPPHPHPPHPHHPHPPHHPPHHHHPP
162 HHPPHPPHHHHPPHPPHPPPHHPHPHP
163 PHPPHHHHPPHPPHPPPHHPPPPHPPH
164 HHPPHHPHPPHHHHHHPPHPHPHPHHP
165 HHHPHHHPHPPPHHHHPHPHPHPPPHP
166 HHHPHHHPHPPHPPPHPHPHPHHPPHH
167 PHPPHHHPHPPPHPPHPHPHPPHPPHH
168 HHHPHHHHPPHHHPPHPPPHPHPHPHP
169 HHHPHHHHPPPHPHPHPHPPPHPPHPH
170 HHHHHHHPPHPPPHHHPPHPHHPPHPH
171 PHPPPHPHPPPHHHHPPHPPHPHPPHH
172 HHHHHHHPHPHPPHPPPHHPHPPHHHP
173 HHHPPHPPHPHHPHPHPPHHPHPHHHH
174 HHPPPPHPPPPHPPHHPPPPHPPHPHP
175 HPHHHHPPHPPHPHPPHHPPHPHHPHH
176 HHHPPHHHHPPHPPHPHPHPPHPHPHP
177 HHHPPHHHHPPHHPHPPHHPPHPPPHP
178 HHPPPPHPHPPHPHPPHHHHHHPPPHH
179 HHHPPPHPPPHPPHHPHPPHPHPHPHP
180 HHHPPHHPHHPPHHHHPPHPPHPHPHP
181 HPHHHPHPHHPPPHPHPPPHHHHHPPH
182 HHHPPPPHPPPHHHPPHHPPHPPHPHP
183 HHHHPPHHHHPHPHPHPPPPHPHHPPH
184 HHHPPHPHPHPHPPHPHPHHPPPPHHH
185 HHHPPHPHPHPHHPHPHHPPHPPHHHH
186 HHHPPHPHHPPHHPPHHHHPHPHPPHP
187 HHPHHPPPHPPHHPPHPHPHPPHPPHH
188 PHHPPHPPHPPHPPHHPPHHHHHHPPH
189 HPPHPHPPHHHHPHPPHPHPHHPHHHH
190 HPPHPHPPHPHHHHHHPPHPPHPHPHH
191 HHPHPPHPPHHPPHHPPHPHPPPPHHH
192 HHPHPPHPPPHHHHPPPHPHPHPPHHH
193 HHPHPPHPPPHPHHPHHHHPPHPPHHP
194 HPPHPPPHHHHHPPHPPHPHPPPPHPH
195 PHHPPPPHPPPPHPPPHPHHPHHHPPH
196 PHHPPPPPHPPHHPPHPHPHHHHPPHH
197 HHPPPHHPPPPPHPPHPHPPPPHHHHP
198 HHPPPHHHPPHPPHHPPHHPHPHHHHH
```

199	HHPHPHPHPHPPHPPHHPPPHPHHHHH
200	HHPHHPPHHPHHHPPHPPHPPPHPPHP

Table 10: The sequences used as crowders in the results behind figures 14-15.

Index	Sequence
1	HHHHPPHPPHPHPPPHPHPPHPHPHHH
2	HHHPHPPHPHPHPPHPPHHHHPPHPHH
3	HHHHHHHPHPPHPHPPPHPHPPHPHHH
4	HHHHHHPHPHPPPHPHPPHPHPHPPHH
5	HHHPHHHPPHPPPHPPHPPHHPPHPHH
6	HHHHPHHPHHHPPPHPHPPHPPPHPHP
7	HHHPPHHHPPHPPHPHPHPPPPHPPHH
8	HHHPPHHPHPHPPHPHPHPPHPPHPHH
9	HHHHHPPHPPHPHPHPHHHPPHPHPHP
10	HHHPHPPHHHHPPHPPHHPPHPPHPHH
11	HHHPHHHPHPHPPHPHPHPPHPPHPHH
12	HHPHPHPPHPPHPHPPHHPPHPPHPPH
13	HHHHHPPHPPHPHPHPPHHHPHPHPHP
14	HHHHHPPHPPHPHPHPPHHPPHPHPHP
15	HHHHHPPHPHPPPHPPHHPPHPPPHHH
16	HHHHHPPHPPHPHPHPPHPHPPHHHHP
17	HHHHHPPHPPHHPPHPHPPHPHPPHHH
18	HHHPPHHHPPHPPHHPPHPHPPHPPHH
19	HHHPPHPHPPHPPHHHPPHPPHPPPHH

Table 11: The sequences comprising the SVS20 environment.

Index	Sequence
1	HHPHPPHHHHPPHHPPHPPHHHPPHPH
2	PHPPHPHHHHPHPHPHPPPHPPHHHPH
3	PHPHPPPHPPHPHPPPHHPPHPHHHHH
4	PHPHHPPHPPHHPHHPHPHPHPHHHHH
5	HPPHHPPHPPHPPHPPHPPPHHHHHHP
6	HPPHPPHHPPHPPHPHHPPHHPPHHHH
7	HPHPHPHHPPPPPHPPPHPHHHHPHHH
8	PHPPPPPPHPPPPHPHPPHPHHHPPHH
9	PHPPPHPPHHPPHPPHHHHPHPPPHHH

10	HPPHPPHPPHHPPHPHPPHPPPPHPHH
11	HHHPPHPPHPHHPHPHPPHHPHPHHHH
12	PHHHHHPHPPPHPHPPHPPPPPHPPPH
13	HHHPPHPHPPHPHHHHPPHPPHHPPHP
14	HPPPHPPHPPPPPPHPPHPHPHHHPPH
15	PPHHHHPPHPHPPHPPPPPPPHHPPHH
16	PHPHHHHPPHPPHPHPHHHHPPHPPHH
17	HHHPHPPPHPHPPHHHPPPHPPPHHHH
18	HPHHPPHPPHPPPHPHHPPHPHPPHHH
19	PHHPHHHHHHHPPHHPHPHPHPPHPHP
20	HHHHPHPPHPPPHHPPHHPHPHPPHHH

Figures 21-22

Table 12: The sequences used as test proteins in the results behind figure 21. Astersks mark those sequences used as test proteins in the results behind figure 22. Motifs are specified by what colour they are marked with in the figure.

Motif	# of patches	Sequence
Red	1	PHPHPHPHPPHPHPHPHHPPHPPHPPH*
Red	1	PPHHHHPHPPHPHHPPPHPPHPHPPHP
Red	1	PHHHPPHPPPPPHPPHHPHPPHPHPHP
Red	1	HHPPHPPPPHPPHPPPPHPHPHPPHHH
Red	1	PHPPPHHPPHPPHPPPHPPHPHHHPHH
Red	2	PHPHPHPHPPPHPPHPHPPHHHHPHHH
Red	2	PHPHPHPPHPHHPPHPPHPPHHPHHHH
Red	2	PHPHPHPHPPPPPPHPPPHHHHPHHHH
Red	2	PHPHPHPHPHPPHPHPHPHPPHHHHHH
Red	2	PHPHPHHHHPPHPHPPHPHPHPPHHHH
Red	3	HHHHPHHPHPPPHPHPPHHPPHPHPHP
Red	3	HHPHHPHPHPPPHPPHPHPHPHPPHHH
Red	3	HHPHHPHPHPPPHPPHHPHPPHPHPHP
Red	3	HHPHHPPPHHHPHPHPPPHPPHPHPHP
Red	3	PHPHPHPHPPPHPPPPHPHPHHHPHHH
Blue	1	PHHHHPPHPHHPHPPPPPHPHPPPPHP*
Blue	1	PHPPPPPHPPHPHPHHPHPPHHHHHPH
Blue	1	PHPHPPPHPPPPHPHPHPHHPPHHHHH
Blue	1	PHHPHPHPPHPHPHPPHPPHHHPPHHH

Blue	1	PHPHPPPHPPHPHPHPHHPPPPHHHHH
Blue	2	PHPHPPHHHHHPHHPPPPHPHPHPHHP
Blue	2	PHPHHPHHHHHPPPPHPPHPHPHPHHP
Blue	2	HPHPHPPHPPPPHHPPPPHPHHHHPHHH
Blue	2	PHPHPPHPHPHHPPPPPHHPHHHPHHH
Blue	2	HHPPPHPHHPHHPHHPPHPPHPHPPHHH
Blue	3	HPHPHPHHPPPPHHPPPPHPHHHHPHHH
Green	1	PHPPPHHPPHHHHPPHPPHPHPPPPHHH*
Green	1	HHPHHPPHPPHPPPHPPHPPPHHHPPH
Green	1	PHHHHPHPPPPPHPPHPHPHPPHPPHH
Green	1	PHPHPPHPHPPHHPPPHHHHHHHPPHP
Green	1	PHHHHPHPPHPPHPHPPHPPHHHPPHP
Green	2	HPPHPHHHHPPHPPHPHPPHPHHHPPH
Green	2	PHPPHHHHPPHHPHPPHPHPPHPHHHH
Green	2	HHPPHHPPPPHHPPHPPHPPHHPHPPH
Green	2	PHHHHPPPPHPHPPHPPPPPHHPPPHH
Green	2	PHPPHPHPPHPHHPPHPPPPHHHPHHH
Green	3	PHPPHHHHPHHPPHPHPPHPPHHHHPH
Green	3	PHPHPPHHHHPHPPHPHHHHPPHPHHH
Green	3	PHPPHHHPHPHPPHHPPHPHHPPHHHH
Green	3	HHHPHHPPHPPHHHHPPHPPHHPPHHP
Green	3	PHHPHPPHPHPPPPHHPPHHHHHHPPH

Table 13: The sequences used as crowders in the results behind figure 22.

Environment	Sequence
1	PHHHHPPHPHPPPHPPHPHPHPPPPHP
1	PHPPPHHPPHPPPHPPHPPPHPHPHHH
1	PHPHPPHHHHHPPHHPPPHPHPHPPHP
1	PHPHPPHPHPPHPPHHHHPPPPPHPPH
1	PHPHPPHPPHPPHPHPPHPPHHHHHHP
1	PPHHHHHPPHPPHHPPHPPHPPHPPHP
1	HHPHHPPPPHPPPHPPHPPPHPPHPPH
1	PHHHHHPHPPPHPHPPHPPPPPHPPPH
1	PHPHPPHPHHHHPPHPPHHPPHPPHHH
1	PHPPPHPPPPPHPPHPHPPHHHHPHHH
1	PHPPHPHPPHPHPHHPHPPHHHHPHHH
1	PHPPHPHPPHPHPPPPPPPPHHHHPPH

1	PHPPPHHHHPPHPPHPHHPPHPPHPPH
1	PPHHPPHPPHPPHHPPPHPPPPHPPHH
1	HHPHHPHPPHPPHPPPHPPPPHPPPHP
1	PHPPHHHPPHPHPPPHPPPPHPPHHPH
1	HHHPHPPHPHPHPPPPHHPPHPPPPHP
1	HHHPPHPPHHPPPHPPHPPHHPPPPHP
1	PHPHPHPPHPHHPPHPHPPHPPPPHHH
1	HHPPPHPPHPPHHPPHPHPPPHPPHHH
2	PHHPPHPPHPPPPHPPHPPHPHPHHPH
2	PPHHHHPHPPHHPPPHPPHPHPPPPHP
2	PHHPHPHPPHHPPPPHPPHHPHPPHHH
2	PHHPHHHHHHHPHPHPPPHHPPPPHH
2	PHPHPPHHHHHPPHPPPPHPHPHPPHP
2	HHPHHPPHPPHPPPHPPHPHPHPPPHH
2	PHPHPHPPPHHHHPPHPPHPHPPPPHH
2	HPHPPHHPPHPHPPHPPHHHHHPHPPH
2	HPHPPHPPPPPHPPPHHHHPHHHHPPH
2	PHPPHPHPPHPHPPPHPPPPHPHHHHH
2	HHHHHPPHPHPHPHPPHPHPHPPPPHP
2	HHHHHHPHPPPHPPHPHPHPPPHPHPH
2	PHPPHPPPHPPHPPHPHHPHHPPPPHP
2	HHHPPHPPHPHPPHPHPPHHHPPPPHP
2	PPHHPPHPPHPHHHHHPPHPPHPHPPH
2	PHPPHPHPHPPHHPPHPHHHHHHPPHH
2	HHPPPHPPHPPHHPPHPPPPHHHPPHP
2	PHPPPHHPPHPHPPPPHPHPPHPPHHH
2	HHPPHPPHHPPHPPPHPHPPHPPHHHH
2	HHPHHPHPPHPPHPPPHPPPPHPHPHP
3	PHPHHHHPPHPPHPHPPPPHHPPHPPH
3	HHPPHPPHHPPHPHPPHPPHHHHPPHP
3	HHPHHPHHPPHHPPHPPPHPPHPHPHP
3	PHPPPHPPHPHPHPPHPHPPPHHHHHH
3	HPPHHHHPHPPHPPPHHHHPPHPHPHH
3	HHPPPHPPHPPPPHHPPHPHPHPPHHH
3	PHPPHPHPPPPHPHPPHHHHHPHPPHP
3	PHHHHHHPHPHPHPPPPPHPPHPHPHP
3	PHPHPPHHPPHHHPPHPPHPHPPPHHH
3	HHHHHHHPHPHPHPPHHHPHPHPPHHP

3	PHPHPPHHHHHPPPPHPPHPHPHPPHP
3	PHPHPHPPPHPPHPHPHHPHPPHHHHH
3	HHHHPPHPPPPPPPHPHPHHPPPHPHP
3	HPHPHPPHHPPHPHPHPPPHPPHHHHH
3	PHPHHPHHHHPHHPHPHPPPHPHPPHP
3	HHHHPHPHPPHPPHHPHPPHHPPHPPH
3	PHPHPPHPPHPPHPHPPHPPHHHHHHP
3	HHPHPPPPHPPHPPHHPPHPPPPPHHP
3	PHPPPHHPPHPPHHPPHHHHPPHPHHH
3	HPPHPPHPPPHPPHPPPHHHPHHHPPH
4	HHHPHPHHHHPPHHPHPPHPPHHPHHP
4	HHHHPHHPPPPPHPHPPHHPPHPPPHP
4	HHHHPHPPHPHPPHPHPPHHHPPHPPH
4	HHHHPPHPPHHPPHPPPHHHPPHPHHH
4	PPHHPPHPPHPHHHHHHPHPHPPHPHP
4	PHPPHPPPHPPHPPHPHHPHHPPPPHP
4	PHHHHPPPPHPHPPHPPPPPHPHPHHP
4	HHHPPHHPPHHHHPHPPHPHPPPPHHH
4	PHPHPPHPHPPPPHHPPHPPHPHPHHP
4	HHPHHPPPHHHPHPHPPPHPPHPHPHP
4	HHPPHPHPPPPHPHPPHHHHHHPPPHH
4	PHPHPPHHPPPPHPHPPPPHPPPPHHH
4	PHPPHPPPHPPHPPHHPPPHPPHHHHP
4	HHPHHPPPHPHPPHPHPPPPPPPHPPH
4	HHPHHHPPHHPPPHPPHPPPHPPHPPH
4	HHHHHHPPHHPPHHPHPHPHHPHPHPH
4	PHPPHHHPHHPPHPPPPHPPPHPHPPH
4	PHHHHPHPHPHHHPHHHPPHPHPPPHP
4	PHPHPHPPHHHPPHPPHPHPPHHPHPH
4	PHPHPPHHHHPHPPHPPHHPPPHPHHH
5	HHPHHPHHPPHPPHPHPPHPPPHPPHP
5	PHPHPPPHPPHHPPHPPPHHHHPPHHH
5	HHHPPHPPPPHHPPHPPPPHPPHPPHP
5	HHHHHHHPHPHPHPPHPHPHPHPHPHH
5	HHPPHHHPPHHPPHPHPPHPPPPPPHP
5	PHHHHPPHPPHPHPPPPHPHPHPPHPH
5	PHPHPHPHPPPHPPHHHPPHHHHPHHH
5	PHPPHPHPPHPHPHPPPPPPHHHHPPH

5	HHHHHHHPHPHPHPHHPPPHHPPHPHH
5	PHPHPHHHHHHHHPHHPPHHPPHPHPH
5	PHPPHHPHHHHPPHPHPPPHPPHHHPH
5	PHPPHPPPHPPHPPHHPHHPPHHHPPH
5	PHPHPPHPPHHPPPPHPPPPHPPPHHH
5	HPHPHPHHHPHHHHPPPHPPPHHPHHP
5	PHPHPHHHHHPHHPPHPHPPHPHPHHH
5	HHPPHPHPPHPHPPPPHPHHHHPPHHH
5	HHHHPHPPHPPPHHPPPHPHPHPPHHH
5	PHHHPHPPHPHPHHHHPHPPHPPPPPH
5	HHPPHHPPHPHHHPHHHPPHPHPPPHP
5	PHHHHHHHHPHPPHPHPPHHPHHPHPH
6	PHPHHPHHPPHHHPHPHPHHHHPPHHH
6	PHPHPPHHHPHPHPHHHHPPHHHHPPH
6	PHHHHPHHPPHPHPPHPPHHHHPHPPH
6	PHPHPHHPHPPHPPHHHHPPPPHPHHH
6	PHHHHPHPPHPPHHPHPPHHPPHPHHH
6	PHPHPPHHPPHPHPPHPPHHHPPHHPH
6	HPHPPHPHHHHHHPPHHPPHHHHPHHP
6	PHHHPHPHPHHPHPHPHPHPHHHHHHH
6	HPPPHPHPHPHHPPPHHHHHPHHPHHP
6	PHPHPPPPHHPHPPPPHPHPPHPPHHH
6	PHHHPHPHPHHPHPHPHPHPHHHHHHH
6	HHHPPHHPPHPHPPPPHHPPHPPHPHH
6	PHHPHHHHPHPHPHPPPHPPHPPPPPH
6	HHPPHPPHPHPHHPPHPPHPPPPHHPP
6	PPHHHPHPHHPPHPPHPPPHPPPHPHP
6	HHHHPHPHPHPHPPHHPPHPPPHPHHH
6	PPHHHHHPPHPPHHPPHHPPHPHPPHP
6	PHPPHHHPPHPHPPHPHPPHHPPHHHH
6	PHHHHPPHPPHPHPPPPHHPHHHPPHP
6	PHHHHHHPHPHPHPPPPPHPPHHHPHP
7	HPPPHPHPHPHHPPPHHHHHPHHPHHP
7	HHHPHHPPHHPPHHPPHPPHPHPPHHH
7	HPHPHPHHHHHHPPHHPHPHPHPHHHH
7	HHPPHHHPPHPHPPHPPPPPPHPPPHH
7	PHPHPPHPHPPHPPPPPHPPHHHPHHH
7	PHHHHPPHPHHPPHHPPPPPHPHPPHP

7	PHHPHHHPHPPHPHPPPPHPPHPHPHH
7	HHHHPHPPHPHPPPPHHHPHPHPPHHH
7	PHPHPHPHHHHPPHHPHPHHHHPPHHH
7	HPHPHPHHPPHPPHPHHHPHPHHHHHH
7	PHPHHPPHHHHHHPPHHPHPHPHPPHH
7	HHHHPPPHHPPHPHHPHHPPHPHPHHH
7	PHPHPPHPPPHPHHPHHHPPPPHPHHH
7	HPHPHPHPHHPHPPHHPHPHPHHHHHH
7	HHHPHPPHPHPHHPPHHPPHPPHHHHP
7	PHPPHHPPPPHHPPPPHPPHPPHHPPH
7	PHPHPHPHHHHPPHHPPPHHHHPPHHH
7	HHHPHPHPHPPHPPHPPHPHPHPPHHP
7	PHPHPHHHHPHPHPHPHHHHHHHPHHP
7	HHHHHPHPHPPHHPPHHPPHPHPHPHH
8	HHHPHPHPHHHHPPHHPHHHPPPHPPH
8	HHPHHPHPPPHHPHPHPHPHHHHPPHH
8	PHPHPHPHPPHPHHHPPPPHPPHHHHH
8	PHPHPHHHHHHPHPHPHPHHHPPPPHP
8	PHHHHHHPHPHPHHPPHPPPPPHHPHP
8	PHPHPPPPHPHPPPHHPPHHPPHHPHH
8	HHHHHHPPHHPPHHPHPHPHHPHPHPH
8	HPHHHPHPHHPPPPHPHPHPHHPPHHH
8	HHHHPHHPHPHPPPPHPHPPHPHPHHP
8	PHHHHPPPPPPHHPHPHPHPPPPHPHH
8	HHPHHPHPPPHHPHPHPHPHHHHPPHH
8	HHHHHPHPHPHPPHPPPHHPPHPHHHH
8	PHHHHPPHPPHPHPPPPPPHPPPPHHP
8	PHHPPHHPPHHPPHHPHPHHHHPPHHH
8	HPPPHPPPHPPHHHPHHHHPHPPHPHP
8	PHPHPHPHHPHPHPHPPPHHPPHHHHH
8	PHHHHPHHPPHPHPPHPHHHHHHPPHH
8	HPPPHPHHPHHHPPPPHPPHPHHPPHH
8	PHHPHHPHPPHHHPHPHPHHHHPPHHH
8	PHPHHPPHPHPPHPHPHPHHPPHHHHH
9	HHHHHHPPHHPPHHPHPHPHHPHPHPH
9	PHHHPPHHPPHHPPHPHPHHHHPPHHH
9	PHPHPHHHHPPHPPHPHPPPHPHHPHH
9	HHPPHPHPPPHHPHPHPHHHHPPHHHH

9	HPHPHPHHPHPPHHHHPHPPHHPHHHH
9	HHPHHPHPPPHHPHPHPHPHHHHPPHH
9	HHHPPHHHHPHPHHPPPHHPPPHHHHP
9	PHPHHPHHPPHPHPHHHHPPHHHHPPH
9	PHHHPPHPHHPPHHHHHHHPHPHPPHP
9	PHPHPPHPHPHPHHPHPHHPHHHHHHH
9	PHPPHHHPPHHHPPHPHHPPHPPHHHH
9	HHHHHPHPHPHPHHPHPHPHPHPHHHH
9	HHHPHPHPHHHHPPHHPHHPPHPPPHH
9	HPHHPHPPPHPPHHHHHPPHPHPHPHH
9	PHPHHPPPHPHPPPHPPPHHPHPHHHH
9	HHHHHPHPHPHPHHPHPHPHPHPHHHH
9	PHPHPHPHHHHPPHHHHPPPHHPPHHH
9	HHPHHPHPHPPPHPHHHPHPPHPHPHP
9	HHHPPHHHHPHPHHPPPHHPHPHHHHP
9	HHHHPPHHPPHHHHPHPPHHPHHPHPH

Figures 24-25

Table 14: The sequences used as test proteins in the results behind figures 24-25.

Index	Sequence
1	HHHHPPHHHPHPHPPHPHPHPHPPHHH
2	HHHPHHHHHPHPHHPHPPHPPHHPPHP
3	HHPHPPPHHHHHHHPPHHHPHPHPHPH
4	PHPHHHHHHHPPHPHPHPHPHHHPPHH
5	HHHHPHHHPHPPHHPHPPHHHPPPPHP
6	HPHHHHPPHHHPHHPPHHPPHPPHPHH
7	HHHHHHPHPPPHPHHPPHPHPPHHPHH
8	HHHPHHHPPHPHHPPPPHHPPHHHPPH
9	PHHHPHPPHPPHHHHPHHHPHPPPPPH
10	HPHPHHHHHHPPPPHPPPHHPHPHPPH
11	HPHHHPHHHHPPHPPPPHPPHPPHPHH
12	HHHPHPHHPHPHPHHHHHHPHPHPHPH
13	HPHPPPHHPHPHPPPHPPHPHHPPHHH
14	HHPPHHHPHPHHHHPPHHPPHPPPHHP
15	PHHPPHHPHPHHHPPPPHHPHHHPHHH

16	HHHHHPHPPPHPPHHPPPHHHHPHPPH
17	HPHPPPPHPHHHHPPPPPHHPHHHPPH
18	PHPPPHHPPHHHHPHPHPHHHHHPHPH
19	HPPPPHHPHPHPHHHPPHPHHPHPPHP
20	HHHPHPHHPHPHHPPPPPPPHPPPPHP
21	PPHHPPPHHPPPPPPPHPPPHPPPPPH