Revised version
LU TP 01-07
November 28, 2001

# Enumerating Designing Sequences in the HP Model

Anders Irbäck and Carl Troein[1]

Complex Systems Division, Department of Theoretical Physics
Lund University, Sölvegatan 14A, S-223 62 Lund, Sweden
http://www.thep.lu.se/complex/

Submitted to *J. Biol. Phys.*

Abstract:
The hydrophobic/polar HP model on the square lattice has been widely used to investigate basics of protein folding. In the cases where all designing sequences (sequences with unique ground states) were enumerated without restrictions on the number of contacts, the upper limit on the chain length $N$ has been 18–20 because of the rapid exponential growth of the numbers of conformations and sequences. We show how a few optimizations push this limit by about 5 units. Based on these calculations, we study the statistical distribution of hydrophobicity along designing sequences. We find that the average number of hydrophobic and polar clumps along the chains is larger for designing sequences than for random ones, which is in agreement with earlier findings for $N \leq 18$ and with results for real enzymes. We also show that this deviation from randomness disappears if the calculations are restricted to maximally compact structures.

Key words: Hydrophobicity correlations, hydrophobic/polar lattice model, exact enumeration, protein sequence analysis, folding thermodynamics, protein folding.

[1]E-mail: anders, carl@thep.lu.se

# 1 Introduction

Coarse-grained models are an important tool in theoretical studies of protein folding, for computational as well as conceptual reasons, and have been used to gain insights into the physical principles of folding (for a recent review, see Ref. [1]). These models are often lattice based. The main advantage of using a discrete conformational space is that exact calculations can be performed for short chains, by exhaustive enumeration of all possible conformations. One model that has been extensively studied this way is the minimalistic hydrophobic/polar HP model of Lau and Dill [2] on the square lattice. In previous work on this model, all sequences with unique ground states were determined for chains with up to $N = 18$–$20$ monomers [3–5]. Such sequences are called designing; they design their ground state conformations.

In this paper, we show how a few optimizations make it possible to extend these calculations to $N = 25$, which corresponds to a 4000-fold increase in the number of possible sequence, conformation pairs.[2] We then use this data set to address the question of how designing sequences differ from random ones statistically.

By analyzing the behavior of block variables, it has been found that designing $N \leq 18$ HP sequences [6] as well as real (globular) protein sequences [6, 7] show negative hydrophobicity correlations. Therefore, one expects to find an increased number of hydrophobic and polar clumps along these chains. In this paper, we show that the average number of clumps is indeed larger for designing HP sequences than for random sequences. In particular, this implies that the earlier finding that designing sequences show negative hydrophobicity correlations remains unaffected when increasing $N$ to 25. This provides a non-trivial test of the robustness of this property.

In lattice model studies it is not uncommon to consider only maximally compact conformations, which for $N = 25$ are confined to a $5 \times 5$ square. This drastic reduction of conformational degrees of freedom leads to a sharp rise in the number of designing sequences. An interesting question is whether this reduction also affects the statistical properties of designing sequences. To address this question, we repeat the same statistical analysis for sequences that are designing when only maximally compact conformations are used. The results turn out to be qualitatively different in this case. In particular, this means that the agreement with the results obtained for real sequences gets lost when this reduced conformational space is used.

[2]The complete list of all designing $N \leq 25$ sequences and the corresponding conformations will be made electronically available at `http://www.thep.lu.se/complex/wwwserver.html`

Finally, we also study the character of the folding transition for one of the designing $N = 25$ sequences, which was selected by an optimization procedure. The thermodynamic behavior of this sequence is studied using Monte Carlo simulations.

The paper is organized as follows. In Section 2 we define the model and describe the algorithm and optimizations used for finding all designing sequences with $N \leq 25$. Our results are discussed in Section 3, which contains sequence and structure statistics, the statistical analysis of designing sequences, and the thermodynamic study of an optimized sequence. A summary is given in Section 4.

# 2 Enumerating designing sequences

In lattice models of proteins it is common to use a contact potential. This means that the energy that a sequence gets with a certain conformation, is given by what contacts exist in that conformation. That is,

$$E = \sum_{i<j} C_{ij} U(\sigma_i, \sigma_j) \tag{1}$$

where the contact map $C_{ij}$ is defined as

$$C_{ij} = \begin{cases} 1 & \text{if monomers } i,j \text{ are neighbors on the lattice but } |i-j| \neq 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

and $U(\sigma_i, \sigma_j)$ is the interaction matrix. In the HP model there are two amino acids, hydrophobic (H) and polar (P). The interaction matrix is

$$U = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix} \tag{3}$$

with H in the first row and column, so the energy is determined exclusively by the number of HH contacts.

Two conformations that have identical contact sets will have the same energy for all sequences, so they can be represented by a single contact set which is marked as *degenerate*. What this means is that a large number of conformations can be reduced to a smaller number of contact sets. A contact set that corresponds to a single conformation will be called *unique*.

Table 1: Contact set terminology.

| | |
|---|---|
| Unique | Corresponds to a single self-avoiding walk |
| Degenerate | Corresponds to more than one self-avoiding walk |
| Maximal | Not a subset of any other realizable contact set |
| Redundant | Can be ignored in the search for designing sequences |

## 2.1 Compressing conformational space

To find all designing sequences, we first determine all relevant conformations, which then are combined with sequences in Section 2.2. The conformation search is sometimes simplified by only considering maximally compact conformations, where the number of contacts is maximal. Looking only at those conformations corresponds to shifting the energies by adding a large negative term to all elements of the interaction matrix. An efficient method for enumerating compact conformations was recently proposed by Kloczkowski and Jernigan [8].

In this paper, we consider all possible self-avoiding walks, and not only those that are maximally compact. The space of all possible self-avoiding walks grows rapidly with $N$. Using contact sets rather than self-avoiding walks gives a significant reduction of the conformational space (see Table 2 below), but this reduction is not sufficient for our purposes; memory limitations prevented us from storing the complete list of all possible contact sets for $N \geq 24$. To be able to go to larger $N$, we therefore developed two procedures for reducing the number of contact sets to be stored. These procedures are based on the observation that as long as there are no repulsive forces (that is, as long as the elements of the interaction matrix are all non-positive), it is never energetically disadvantageous to add contacts as long as no existing contacts are broken.

Before discussing these two procedures, it is helpful to introduce some terminology. A contact set will be called *maximal* if it is not a proper subset of any other (realizable) contact set. A conformation that is designed by at least one sequence is designable. It can be readily seen that every designable self-avoiding walk corresponds to a unique and maximal contact set. Another important class are contact sets that can be safely ignored in the search for designing sequences. Such contact sets will be called *redundant*. A summary of our contact set terminology can be found in Table 1.

### 2.1.1 Eliminating contact sets: Step 1

As mentioned above, it was impossible for us to store the complete list of all contact sets for $N \geq 24$. To circumvent this problem, we used a program that for each self-avoiding walk tries a carefully selected, predefined set of (local) moves. If any of these moves can be performed without destroying any existing contact (new contacts may form), the self-avoiding walk is discarded. All possible self-avoiding walks surviving this test are converted to contact sets. It is important to stress that the moves are chosen so that the resulting reduced list of contact sets has the property that any realizable contact set is a subset of some set in this list. In particular, the move set does not contain the inverse of any of its elements.

The list of contact sets obtained this way was indeed small enough to be stored up to $N = 25$, but the procedure has the disadvantage that it may eliminate non-redundant contact sets. This is an unwanted property, but the problem is easy to solve. The solution is that once the sequences that have non-degenerate ground states *with respect to the non-discarded conformations* have been found, each sequence is combined with its conformation, and the chain is tested with the opposite of the tests used to discard conformations. That is, the program tests whether any of the opposite moves can be performed without breaking any of the existing HH contacts. By using the fact that the forces are repulsive, it can be seen that if no such move is possible, then the conformation has to be a unique ground state of this sequence. Hence, by performing this test, one can make sure that no sequence is falsely declared to be designing, even though there are non-redundant contact sets that are missing in the list used.

It is also important to note that all the self-avoiding walks with a given contact set are never discarded if the set is maximal. This means that all the contact sets that correspond to designable conformations are included in the list generated by this procedure. This is important because it implies that this reduced list of contact sets can be used without losing any designing sequence. What could go wrong when using this list is that some non-designing sequences were classified as designing, but this is avoided by performing the test discussed above.

### 2.1.2 Eliminating contact sets: Step 2

Our second procedure removes contact sets from the list produced by the first procedure. This is done to speed up the calculations. All contact sets that are removed

in this second step are redundant.

The procedure relies on the fact that all pair energies are non-positive. To see how that can be used, consider one set of contacts $A$ which is a subset of another contact set $B$. It is then clear that $A$ cannot represent a unique ground state, because for any sequence, $B$ has the same or lower energy. The set $A$ is nevertheless non-redundant in case $B$ is a unique contact set that would be falsely classified as the unique ground state of some sequence if $A$ were removed. If, on the other hand, the set $B$ is degenerate, then it follows that $A$ must be redundant.

Suppose now that $A$ is a subset of two other sets $B$ and $C$. Then, provided that $C$ is kept, there can be no sequence such that $A$ is needed in order to decide whether or not this sequence has $B$ as a unique ground state. This follows immediately from the fact that for any given sequence, both $B$ and $C$ have energies at least as low as that of $A$. Hence, if a contact set is a subset of two or more sets, then it has to be redundant. In particular, this is true if the set is a subset of a subset.

This reasoning gives us the following simple procedure for elimination of redundant contact sets.

- For each set $A$, find all sets of which $A$ is a subset.
- Keep $A$ if
    1. no such sets are found, or
    2. *one* set is found, and this set corresponds to *one* conformation.
- Otherwise discard $A$.

Note that those of the surviving contact sets that meet condition 1 are maximal.

It should be stressed that, because all the contact sets removed by this procedure are redundant, one can still use the test in Section 2.1.1 to make sure that no sequence is falsely classified as designing.

Implementing the rules described above requires some care, since the number of sets grows rapidly with $N$. The solution we used is based on storing the sets in a tree, where each node in the tree represents the question of whether or not a certain contact is included. Before building the tree, it is useful to sort the sets in descending order by the number of contacts they contain. This sorting ensures that a set, once added

to the tree, never has to be removed, since all sets of which it can be a subset are already added when it is examined. Having access to the tree, it is straightforward to search for supersets of a given set. One starts at the root of the tree and each time the tree branches one has to consider either one of the branches or both. The procedure is memory-consuming for large $N$ (see example below), but the CPU time required is relatively modest.

### 2.1.3 Examples

To get an idea of the sizes of these different lists of contact sets, let us consider the case $N = 18$. For this $N$, there are 5808335 self-avoiding walks and 170670 contact sets. After applying the redundancy test in Section 2.1.2 to the list of all possible contact sets, we are left with 33223 maximal contact sets (condition 1) and 6609 contacts sets with one superset corresponding to one conformation (condition 2). Among the 33223 maximal contact sets, there are 6181 sets representing more than one conformation. Subtracting the degenerate contact sets, we are left with $33223-6181 = 27042$ contact sets corresponding to possibly designable conformations. Of those 27042 sets, 5660 sets have a total of 6609 listed subsets. These subsets are needed because they may degenerate an otherwise unique ground state corresponding to one of the 5660 sets.

Here, the redundancy test was applied to the complete list of all possible contact sets. Alternatively, we may first use the program described in Section 2.1.1, which for $N = 18$ generates a list of 51373 contact sets (which contains the 33223 maximal ones). When applying the redundancy test in Section 2.1.2 to this reduced list, we end up with $33223 + 449$ contact sets. Note that with this approach, we find only 449 of the 6609 non-redundant contact sets meeting condition 2. As a result, some conformations are erroneously found to be unique ground states. The test discussed in Section 2.1.1 removes these false unique ground states.

The CPU times required to generate these different lists on a Pentium III 800 MHz were as follows. Generating the list of all possible contact sets, by exhaustive enumeration, took 6 seconds, and reducing this list from 170670 to $33223 + 6609$ contact sets required 7 seconds. The time needed to run the program that generates the list described in Section 2.1.1 was 3 seconds, and reducing this list from 51373 to $33223 + 449$ contact sets took 1 second. The corresponding two numbers for $N = 25$ were 40 and 60 minutes, respectively. In this case, building the tree used in the redundancy test required 220 megabytes of internal memory.

## 2.2 Searching sequence space

We now turn our attention to the sequence space. For each of the $2^N$ sequences we wish to determine what set of contacts gives the lowest energy. If this ground state energy can be accomplished only by a single contact set, and if that set corresponds to a single conformation, the sequence designs that conformation.

### 2.2.1 Organizing the search

The most straightforward approach to finding all sequences with unique ground states is to go through all the sets of contacts for each sequence, and calculate the energy for each of the sets. By only storing the differences between consecutive sets of contacts, and by representing the sequences and contacts as numbers with one bit per position, the number of operations required for each combination can be kept small.

It is, however, also possible to use a very different approach. Represent each sequence by a binary number, and consider any given set of contacts. Between two consecutive sequences two bits are toggled on the average, which indicates that using information about the previous sequence and its energy will be a lot faster than recalculating the energy from scratch. This approach can be used if the whole sequence space is examined for one contact set at a time. The downside to doing this is that information needs to be stored for every sequence until all the contact sets have been considered.

Neither of the methods described above is bad, but they each contain an optimization that the other does not have. It is desirable to utilize both the similarity of consecutive sequences and the similarity of consecutive contact sets. The solution is to divide the sequence space into a number of blocks of fixed size, and apply the second method to each of those blocks. A block consists of $2^M$ sequences that have their first $N - M$ residues in common. This part of the sequence will be referred to as the *fixed part*, and the remaining $M$ positions make up the *variable part*. We call a contact *active* if it connects two H monomers, and note that for each contact there are three possibilities, depending on the position and type of the monomers it connects:

- If both monomers are in the fixed part, the contact gives a contribution of $-1$ to the energy for *all* the sequences of this block, if and only if both monomers are H.
- If at least one of the monomers is a P in the fixed part, the contact cannot be

active for any sequence in the block.

- If both monomers are in the variable part, or if one of them is in the variable part and the other is an H in the fixed part, whether or not the contact is active depends on the variable part.

### 2.2.2 The cutoff energy

This leads us to the next optimization, which has to do with the possible ground state energies, and can be seen as a sequence-dependent reduction of the conformational space. Clearly, a non-degenerate ground state with $N \geq 3$ cannot have an energy of 0. More generally, it is unreasonable that a small number of active contacts should be enough to give an arbitrarily long polymer a unique ground state. To see how this can be used to speed up the calculations, consider some contact set and sequence block, such that there are $p$ active contacts in the fixed part and $q$ contacts whose state depends on the variable part. If $-(p+q)$ is larger than some cutoff value $E_{\max}$, none of the sequences in this block can have a unique ground state for this contact set. The major problem that arises with this optimization is to know what value to use for $E_{\max}$; the algorithm will find only those unique ground states that have energies $E \leq E_{\max}$. For $N \leq 20$ all energies have been considered in the calculations, and it turns out that there are no unique ground states with $E > -4$ for $15 \leq N \leq 20$ (see Table 3 below). For $N > 20$ we have not proven that there can be no unique ground states with $E > -4$, but it seems very reasonable that *if* there are any, most or all of them would have $E = -3$. Therefore we have used a cutoff energy $E_{\max} = -3$ for $N > 20$. It turns out that for $20 < N \leq 25$ there are no unique ground states with energy $-3$, and this strongly indicates that $-4$ is the highest possible ground state energy for any HP chain with $N \geq 15$.

To illustrate how the computer time required varies with $E_{\max}$, let us again take $N = 18$ as an example. For this $N$, the sequence search is found to be about four times faster with $E_{\max} = -4$ than with $E_{\max} = -1$. With $E_{\max} = -4$, the total time needed to find the 6349 designing $N = 18$ sequences, including the time required to generate the conformations, is a few minutes on a Pentium III 800 MHz.

# 3 Results

Using the algorithm and optimizations described above, it was possible to determine all designing sequences for $N \leq 25$ within a reasonable amount of time. Previous work has covered $N \leq 18$ for the normal HP model [3] and $N \leq 20$ for shifted HP models [4, 5]. The increase in $N$ corresponds roughly to a 100-fold increase in the number of known designing sequences and conformations. This gives better confidence when doing statistics on the designing sequences, and it makes it possible to study how properties of the model depend on protein size.

## 3.1 Sequence and structure statistics

Sequence and structure statistics for $4 \leq N \leq 25$ are summarized in Tables 2 and 3. Column three of Table 2 shows the total number of contact sets, which has been studied before [9]. It was estimated to grow as $\mu^N$ with $\mu = 2.29 \pm 0.02$ for large $N$ [9]. The number of maximal contact sets, column four of Table 2, appears to grow exponentially too, but slightly slower. A fit of our data for $15 \leq N \leq 25$ to the form $\mu^N$ yields $\mu \approx 2.07$. This growth with $N$ is considerably slower than that of the number of self-avoiding walks, for which the best available estimate is $\sim N^{\gamma-1}\mu^N$ with $\gamma = 43/32$ and $\mu = 2.6381585$ [10].

The fifth column of Table 2 shows the number of designing sequences. It turns out that the fraction of designing sequences varies between 2.27 and 2.57% for $19 \leq N \leq 25$, which is in line with previous results for smaller $N$ [11]. The last column of Table 2 shows the number of designable conformations. The designability of a conformation is the number of sequences that designs it. From Table 2 it can be seen that the average designability of the conformations that are designable grows with $N$ and is $765147/107336 \approx 7.1$ for $N = 25$.

In Table 3 we show the distribution of ground state energies for different $N$, which is crucial for the optimization discussed in Section 2.2.2.

Figure 1 shows three designable $N = 25$ conformations. The conformation (a) is the most designable one for $N = 25$, with a designability of 326, whereas conformation (c) is designed by one sequence only.

For comparison, we also determined the sequences that are designing when only

Table 2: The number of designing sequences $S_N$ and designable conformations $D_N$ for the HP model on a square lattice. The third column shows the total number of contact sets, as obtained by exhaustive enumeration of all possible self-avoiding walks. Memory requirements (see Sec 2.1.1) prevented us from counting them for $N \geq 24$. Column four shows the number of maximal contact sets.

| $N$ | Conformations | Contact sets | Maximal contact sets | $S_N$ | $D_N$ |
|---|---|---|---|---|---|
| 4 | 5 | 2 | 1 | 4 | 1 |
| 5 | 13 | 3 | 2 | 0 | 0 |
| 6 | 36 | 8 | 4 | 7 | 3 |
| 7 | 98 | 14 | 9 | 10 | 2 |
| 8 | 272 | 41 | 20 | 7 | 5 |
| 9 | 740 | 78 | 39 | 6 | 4 |
| 10 | 2034 | 212 | 95 | 6 | 4 |
| 11 | 5513 | 424 | 174 | 62 | 14 |
| 12 | 15037 | 1113 | 420 | 87 | 25 |
| 13 | 40617 | 2309 | 779 | 173 | 52 |
| 14 | 110188 | 5953 | 1818 | 386 | 130 |
| 15 | 296806 | 12495 | 3409 | 857 | 218 |
| 16 | 802075 | 31940 | 7810 | 1539 | 456 |
| 17 | 2155667 | 67389 | 14717 | 3404 | 787 |
| 18 | 5808335 | 170670 | 33223 | 6349 | 1475 |
| 19 | 15582342 | 363010 | 63434 | 13454 | 2726 |
| 20 | 41889578 | 910972 | 140939 | 24900 | 5310 |
| 21 | 112212146 | 1953847 | 273049 | 52183 | 9156 |
| 22 | 301100754 | 4868343 | 599821 | 97478 | 17881 |
| 23 | 805570061 | 10513774 | 1174460 | 199290 | 31466 |
| 24 | 2158326727 | | 2561884 | 380382 | 61086 |
| 25 | 5768299665 | | 5057733 | 765147 | 107336 |

maximally compact conformations are used. In this case, it turns out that there are 6181800 designing $N = 25$ sequences. The corresponding number is 765147 when the full conformational space is used (see Table 2), so the ratio between the number of designing sequences in the maximally compact ensemble and the number of truly designing sequences is $6181800/765147 \approx 8.1$, for $N = 25$. This ratio has previously been shown [12] to fluctuate between approximately 4 and 11 for $N = 11$ through $N = 18$.

It is worth noting that among the 765147 $N = 25$ sequences that are designing

Table 3: The number of designing sequences for different $N$ and ground state energies.

| $N$ | Energy | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | -1 | -2 | -3 | -4 | -5 | -6 | -7 | -8 | -9 | -10 | -11 | -12 | -13 | -14 | -15 | -16 |
| 4 | 4 | | | | | | | | | | | | | | | |
| 5 | | | | | | | | | | | | | | | | |
| 6 | 0 | 7 | | | | | | | | | | | | | | |
| 7 | 0 | 10 | | | | | | | | | | | | | | |
| 8 | 0 | 0 | 7 | | | | | | | | | | | | | |
| 9 | 0 | 0 | 8 | | | | | | | | | | | | | |
| 10 | 0 | 0 | 1 | 5 | | | | | | | | | | | | |
| 11 | 0 | 0 | 6 | 54 | 2 | | | | | | | | | | | |
| 12 | 0 | 0 | 2 | 27 | 49 | 9 | | | | | | | | | | |
| 13 | 0 | 0 | 0 | 78 | 54 | 41 | | | | | | | | | | |
| 14 | 0 | 0 | 2 | 53 | 110 | 132 | 89 | | | | | | | | | |
| 15 | 0 | 0 | 0 | 58 | 88 | 355 | 330 | 26 | | | | | | | | |
| 16 | 0 | 0 | 0 | 43 | 158 | 250 | 638 | 417 | 33 | | | | | | | |
| 17 | 0 | 0 | 0 | 33 | 160 | 662 | 882 | 1337 | 330 | | | | | | | |
| 18 | 0 | 0 | 0 | 24 | 149 | 623 | 1431 | 2021 | 1676 | 425 | | | | | | |
| 19 | 0 | 0 | 0 | 8 | 154 | 971 | 1936 | 4996 | 3324 | 2007 | 58 | | | | | |
| 20 | 0 | 0 | 0 | 13 | 147 | 955 | 2573 | 5582 | 7665 | 5415 | 2481 | 69 | | | | |
| 21 | | | 0 | 17 | 134 | 1312 | 3116 | 11670 | 12132 | 13917 | 8898 | 987 | | | | |
| 22 | | | 0 | 12 | 120 | 1116 | 3802 | 11672 | 22386 | 24171 | 22394 | 10610 | 1195 | | | |
| 23 | | | 0 | 26 | 92 | 1547 | 4204 | 21050 | 29944 | 56902 | 44940 | 31961 | 8118 | 506 | | |
| 24 | | | 0 | 17 | 134 | 1321 | 4916 | 21096 | 48017 | 78496 | 100746 | 75346 | 40376 | 9596 | 321 | |
| 25 | | | 0 | 20 | 64 | 1708 | 5270 | 32484 | 59470 | 158044 | 159704 | 191377 | 102944 | 46386 | 7688 | 6 |

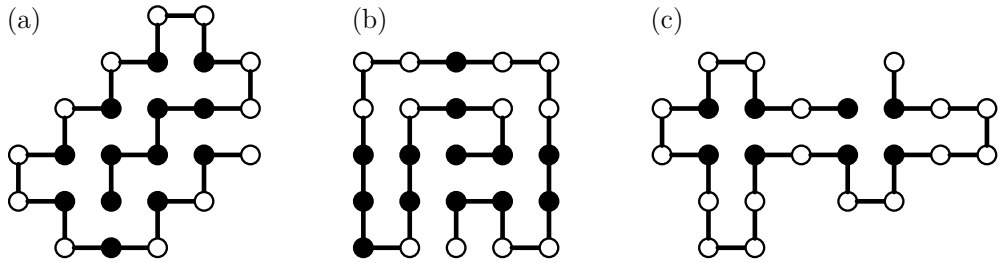


Figure 1: Some designable structures with $N = 25$: (a) the most designable structure, (b) a maximally compact structure, and (c) a structure with few contacts. Filled circles represent H.

when the full conformational space is used, there are only 605 sequences that design maximally compact conformations. Furthermore, it turns out that no maximally compact conformation is designed by more than 10 sequences, whereas the most designable conformation, as mentioned above, has a designability of 326. In fact, there are 19360 conformations that are more designable than the most designable one among the maximally compact conformations.

It is interesting to compare these results to those of Shahrezaei and Ejtehadi [5], who studied a shifted HP model where the interaction matrix is given by $U(\mathrm{H},\mathrm{H}) =$

$-2-\gamma-E_c$, $U(\mathrm{H},\mathrm{P})=-1-E_c$ and $U(\mathrm{P},\mathrm{P})=-E_c$. Using the full conformational ensemble, these authors found that the set of highly designable conformations was independent of the parameters $\gamma$ and $E_c$. In particular, this suggests that the set of highly designable conformations should remain the same when only maximally compact conformations are considered. Our results show that this conclusion does not hold in the original, unshifted HP model.

## 3.2 Statistical properties of designing sequences

In this section, we study the statistical properties of designing sequences by monitoring two different quantities. The first one is the total hydrophobicity $M$, defined as

$$M=\sum_{i=1}^{N}\frac{1+\sigma_i}{2}\,, \tag{4}$$

where $\sigma_i=1$ (H) or $\sigma_i=-1$ (P). Our second quantity is the number $j$ of hydrophobic and polar clumps along the chain [13], which can be written as

$$j=\frac{N+1}{2}-\frac{1}{2}\sum_{i=1}^{N-1}\sigma_i\sigma_{i+1}\,. \tag{5}$$

A similar analysis has previously been performed for $N\leq 18$ [6]. By analyzing the fluctuations of block variables, evidence was found for negative $\sigma_i,\sigma_j$ correlations. Therefore, we expect the average $j$ to be larger for designing sequences than for random ones.

In Figure 2a we show the relative abundance of hydrophobic amino acids, $\langle M\rangle/N$, as a function of $N$, where $\langle\cdot\rangle$ denotes an average for fixed $N$. For the sequences that are designing when the full conformational space is used, we see that the $N$ dependence of $\langle M\rangle/N$ is quite weak if $N$ is not too small, which confirms a trend seen earlier [6]. Furthermore, we see that these sequences, as expected, are more hydrophobic than those that are designing when only maximally compact conformations are used. Finally, it can also be seen that the sequences that design maximally compact conformations differ greatly from those that are designing when only such conformations are used. Figure 2b shows the frequency of different $M$ for $N=25$.

In Figure 3 we show the results of our clump analysis for $N=25$. The average number of clumps for fixed $M$ (and $N$), $\langle j\rangle_M$, is indeed found to be larger for designing sequences than for random ones, which is in nice qualitative agreement with previous results for real protein sequences and model sequences with smaller $N$ [6,7]. Sequences

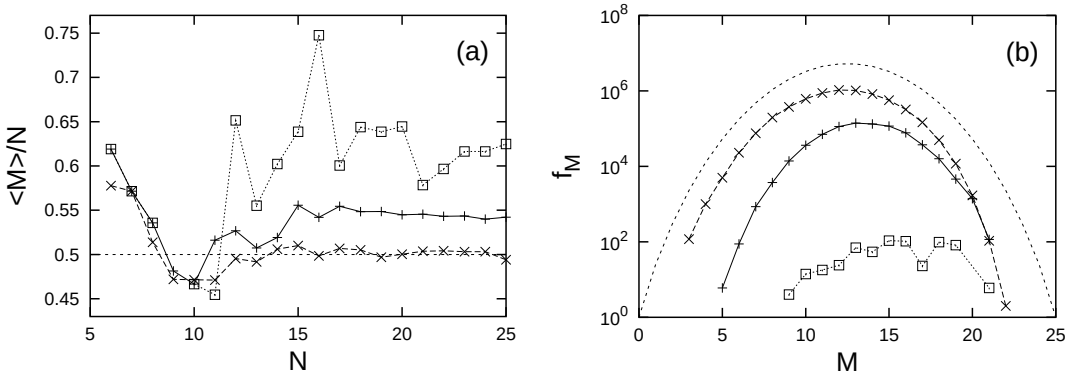


Figure 2: (a) The average hydrophobicity $\langle M \rangle / N$ as a function of $N$ and (b) the frequency $f_M$ of different $M$ for $N = 25$. Shown are the results for all designing sequences (+), all sequences that are designing when only maximally compact conformations are considered ($\times$), and sequences that design maximally compact conformations when all conformations are considered ($\Box$). The dashed lines represent random sequences.

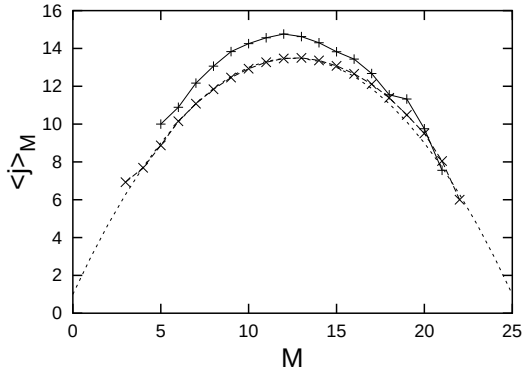


Figure 3: The number of clumps, $\langle j \rangle_M$, as a function of the total hydrophobicity $M$ for $N = 25$. Shown are the results for sequences that are designing when all conformations (+) and only maximally compact ones ($\times$), respectively, are considered. The dashed line represents random sequences.

that are designing when the interaction energies are shifted so far that only the 1081 maximally compact conformations need to be considered, have, by contrast, a $\langle j \rangle_M$ close to that of random sequences. Hence, the results obtained from studying only maximally compact conformations seem to be of less relevance, with respect to the applicability to real proteins, than the results obtained when considering all conformations.

Finally, we note that Buchler and Goldstein have compared various lattice models and found arguments against the use of only two letters [14,15], as in the HP model.

These findings were all based on calculations for maximally compact conformations. However, it is not known to what extent the results of such calculations remain valid when the full set of conformations is used. Our results show that, in the HP model, both the set of highly designable conformations and the statistical properties of designing sequences depend strongly on which of the two conformational ensembles is used.

### 3.3 The character of the folding transition

The model studied in this paper has the important feature that there exists a significant number of designing sequences (this is not true on the triangular [16] and cubic [17] lattices), and that the corresponding conformations tend to show protein-like regularities [18]. However, as in other two-dimensional models, the folding transition is not protein-like in character for the typical sequence; the folding process is not cooperative enough [19]. On the other hand, the folding behavior is, at least to some extent, sequence dependent, and therefore we decided to look into the thermodynamic behavior of a carefully chosen sequence.

This sequence was obtained by applying a Monte Carlo-based sequence design algorithm [20] to the 326 sequences that design the most designable $N = 25$ conformation (see Figure 1a). The design algorithm maximizes the stability of a given conformation with respect to sequence at a fixed non-zero temperature. The sequence we obtained by using this method is shown in Figure 1a. Subsequently, this sequence was subjected to Monte Carlo simulations at different temperatures. In Figure 4 we show the temperature-dependence of the specific heat, which is found to exhibit a pronounced peak. Also shown is the distribution of energy at $kT = 0.479$, which is just above the specific-heat maximum. The energy distribution has one peak corresponding to the ground state, at $E = -13$, and another, broader peak centered at $E \approx -8$. The coexistence of these states implies that the folding transition is much more cooperative for this sequence than for the typical sequence in this model.

## 4 Summary

By greatly reducing the conformational space, and by carefully optimizing the sequence space exploration, we were able to decrease the time needed to exhaustively search for designing sequences with $N = 18$ roughly thousandfold compared to the

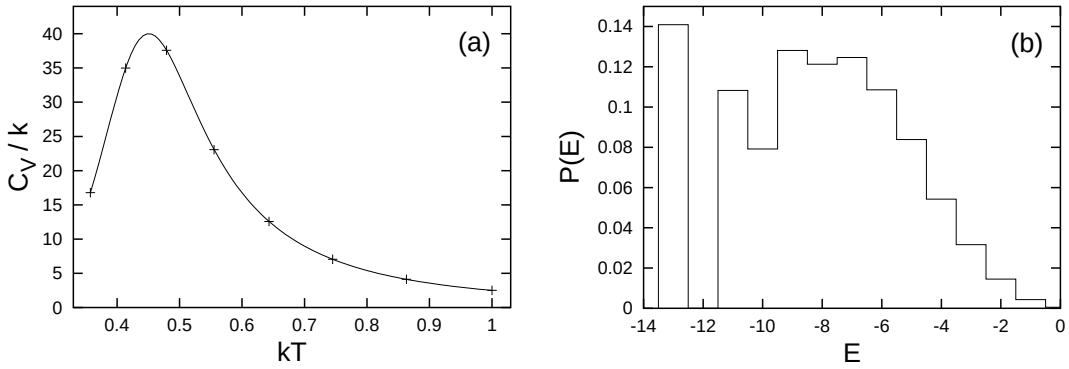


Figure 4: Results from Monte Carlo simulations of the sequence shown in Fig. 1a. (a) Temperature dependence of the specific heat $C_v = (\langle E^2 \rangle - \langle E \rangle^2)/kT^2$. The line is an extrapolation obtained by umbrella sampling [21]. (b) Histogram of energy at $kT = 0.479$.

most naïve methods. It became feasible to find all designing sequences for $N$ as large as 25 using only a small number of workstations. The results obtained by doing this were used to look at the statistical properties of designing sequences. We found that the average number of hydrophobic and polar clumps along the chains is larger for designing sequences than for random ones. In particular, this means that the finding that designing HP sequences, like real enzymes, show negative hydrophobicity correlations [6, 7] remains unaffected when increasing $N$ from 18 to 25. By contrast, qualitatively different results were obtained when discarding conformations that are not maximally compact. This is of interest because restrictions to compact structures are common in both lattice model studies and determinations of statistical potentials from known protein structures. Finally, we saw an example of a folding behavior that is more cooperative than for the typical sequence in this model.

## Acknowledgements

We would like to thank Erik Sandelin for fruitful discussions, Björn Samuelsson for generously providing the program discussed in Section 2.1.1, and two anonymous referees for useful remarks. This work was in part supported by the Swedish Foundation for Strategic Research.

# References


[1] Chan, H.S., Kaya, H. and Shimizu, S.: Computational Methods for Protein Folding: Scaling a Hierarchy of Complexities, to appear in T. Jiang, Y. Xu and M.Q. Zhang (eds.) *Current Topics in Computational Biology* (MIT Press, Cambridge, Massachusetts, USA)

[2] Lau, K.F. and Dill, K.A.: A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins, *Macromolecules* **22** (1989), 3986–3997.

[3] Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D. and Chan, H.S.: Principles of Protein Folding — A Perspective from Simple Exact Models, *Protein Sci.* **4** (1995), 561–602.

[4] Hirst, J.D.: The Evolutionary Landscape of Functional and Model Proteins, *Protein Eng.* **9** (1999), 721–726.

[5] Shahrezaei, V. and Ejtehadi, M.R.: Geometry Selects Highly Designable Structures, *J. Chem. Phys.* **113** (2000), 6437–6442.

[6] Irbäck, A. and Sandelin, E.: On Hydrophobicity Correlations in Protein Chains, *Biophys. J.* **79** (2000), 2252–2258.

[7] Irbäck, A., Peterson, C. and Potthast, F.: Evidence for Nonrandom Hydrophobicity Structures in Protein Chains, *Proc. Natl. Acad. Sci. USA* **93** (1996), 9533–9538.

[8] Kloczkowski, A. and Jernigan, R.L: Transfer Matrix Methods for Enumeration and Generation of Compact Self-Avoiding Walks. I. Square Lattices, *J. Chem. Phys.* **109** (1998), 5134–5146.

[9] Vendruscolo, M., Subramanian, B., Kanter, I., Domany, E. and Lebowitz, J.: Statistical Properties of Contact Maps, *Phys. Rev.* **E 59** (1999), 977–984.

[10] Madras, N. and Slade, G.: *The Self-Avoiding Walk* (Birkhauser, Boston, 1993).

[11] Chan, H.S. and Dill, K.A.: Transition States and Folding Dynamics of Proteins and Heteropolymers, *J. Chem. Phys.* **100** (1994), 9238–9257.

[12] Chan, H.S. and Dill, K.A.: Comparing Folding Codes for Proteins and Polymers, *Proteins Struct. Funct. Genet.* **24** (1996), 335–344.

[13] White, S.H. and Jacobs, R.E.: Statistical Distribution of Hydrophobic Residues along the Length of Protein Chains. Implications for Protein Folding and Evolution, *Biophys. J.* **57** (1990), 911–921.

[14] Buchler, N.E.G. and Goldstein, R.A.: Effect of Alphabet Size and Foldability Requirements on Protein Structure Designability, *Proteins Struct. Funct. Genet.* **34** (1999), 113–124.

[15] Buchler, N.E.G. and Goldstein, R.A.: Surveying Determinants of Protein Structure Designability across Different Energy Models and Amino-Acid Alphabets: A Consensus *J. Chem. Phys.* **112** (2000), 2533–2547.

[16] Irbäck, A. and Sandelin, E.: Local Interactions and Protein Folding: A Model Study on the Square and Triangular Lattices, *J. Chem. Phys.* **108** (1998), 2245–2250.

[17] Yue, K., Fiebig, K.M., Thomas, P.D., Chan, H.S., Shaknovich, E.I. and Dill, K.A.: A Test of Lattice Protein Folding Algorithms, *Proc. Natl. Acad. Sci. USA* **92** (1995), 325–329.

[18] Chan, H.S. and Dill, K.A.: Origins of Structure in Globular Proteins, *Proc. Natl. Acad. Sci. USA* **87** (1990), 6388–6392.

[19] Abkevich, V.I., Gutin, A.M. and Shakhnovich, E.I.: Impact of Local and Non-Local Interactions on Thermodynamics and Kinetics of Protein Folding, *J. Mol. Biol.* **252** (1995), 460–471.

[20] Irbäck, A., Peterson, C., Potthast, F. and Sandelin, E.: Design of Sequences with Good Folding Properties in Coarse-Grained Protein Models, *Struct. Fold. Des.* **7** (1999), 347–360.

[21] Torrie, G.M. and Valleau, J.P.: Nonphysical Sampling Distribution in Monte Carlo Free-Energy Estimation: Umbrella Sampling, *J. Comput. Phys.* **23** (1977), 187–199.