Sequence design is a necessary tool for the investigation of sequence-structure relations. Insights into such fundamental properties will aid to understand protein folding, their evolution, and drug design. The HP-model by Lau and Dill [1] mimics globular water-soluble proteins. It is lattice based and focuses on hydrophobic forces. Even in this coarse-grained model, structure prediction and sequence design is NP-complete [2]. Nevertheless, Backofen and Will introduced a Constraint-based Protein Structure Prediction (CPSP) approach [3] that allows the enumeration of all optimal structures.

HPdesign uses the CPSP approach to solve the inverse folding problem for three-dimensional lattices. Here, a sequence \( X \) is searched that adopts a given structure \( S \) as its single optimal one.

### Preliminaries

**Energy and Optimality of a Structure:** The contact energy in the HP-Model is the negated sum over all non-successive H-monomer contacts. A structure with minimal energy (i.e., maximal H-H contacts) is called optimal and has usually a globular shape as in nature [4].

**H-cores:** The placing of the H-monomers in a structure is called H-core [3]. For a fixed sequence, the energy is completely determined by the H-core internal contacts. This is visualized in Fig. 1 by two structures with energy \(-3\) and \(-1\) (left/right) and the corresponding H-core with 4 contacts. The optimal H-cores are independent of a concrete sequence and can be precalculated in advance [3].

![3D Lattice protein](image)

**Protein-like Sequences:** In contrast to random sequences proteins adopt only one stable optimal structure. Therefore for simplicity, HP-sequences are regarded protein-like only if they have exactly one (or only a few) optimal structure [5].

### Method

The algorithm is a Generate-and-Test method that allows, in contrast to existing methods [6, 7], a systematic and complete enumeration of target sequences within user defined limits. First, a good set of candidate sequences is generated that have a high chance to form the given structure as an optimal one. Afterwards, these sequences are checked if this is true and if they are protein-like.

**Step 1: Candidate Set Generation**

In the HP-model, the number of possible sequences \( S \in S \) for a given structure \( L \) is \( 2^N \). To enable a Generate-and-Test approach we have to keep the number of sequences to test as small as possible. In HPDesign, this is done using a database of (sub-optimal) H-cores. As visible in Fig. 1, the placing of an H-core into a given structure determines a sequence. Following the constraint, that the sequences have to form \( L \) as optimal structure, we use optimal H-cores for sequence generation.

![Figure 2: Different matches and derived sequences for a structure and the H-core in Fig. 1.](image)

For each arbitrary optimal H-core \( H \) we shift the core through \( L \). If all positions of \( H \) can be mapped to positions of \( L \) a match is found and we store the resulting candidate sequence \( S \) in \( S \). This procedure yields a set of sequences \( S \) that can adopt \( L \) with a low energy and have high chance to form \( L \) as an optimal structure. The number of optimal H-cores is still exponential in the core size but increases much slower than the number of possible sequences (see Fig. 3).

**Step 2: Sequence Filtering**

CPSP: The Constraint-based Protein Structure Prediction (CPSP) approach [3] allows the optimal structure enumeration of 3D lattice proteins using Constraint Programming methods. Given a sequence \( S \) with \( k \) H's. For each H-core \( H \) of size \( k \) a CSP is formulated that constrains the monomer sequence \( S \) to form a self-avoiding walk in the lattice, placing all H-monomers on positions in \( H \). Starting with the optimal H-cores, this iterative process ensures optimality and allows further the complete enumeration of all optimal structures.

**Filtering:** To check each candidate sequence \( S \in S \) of step 1 to be proteinlike and to form the given structure stable we enumerate up to 2 optimal structures of \( S \) (CPSP). If there is only one, \( S \) forms only one stable structure \( L' \) and we check if \( L' \equiv L \). If \( S \) fulfills both criteria it is reported otherwise rejected.

### Conclusion

The presented method HPdesign is the first exact method that solves the Inverse Folding Problem for 3D lattice proteins in the HP-model. Using HPdesign one can generate HP-sequences that adopt a given structure as their optimal one. Further the number of optimal structures they can adopt, an important measure for protein-like sequences, can be constrained.

The Generate-and-Test approach is based on a precalculated database of optimal and suboptimal H-cores and the fast and exact CPSP-method by Backofen and Will [3]. It is currently implemented using the cubic and more complex face-centered-cubic (FCC) lattice (see Fig. 4).

The free CPSP-tools package including HPdesign and other tools is accessible at [http://www.bioinf.uni-freiburg.de/sw/cpsp/](http://www.bioinf.uni-freiburg.de/sw/cpsp/).

### References


### Acknowledgments

Martin Mann is supported by the EU project EMBO (EC contract number 02835).