# Assignment 4: Computational methods in *Protein Folding*

Felix Finger[*]

December 19, 2019

## Introduction

For 50 years, the *protein folding problem* has been a major problem in biological research due to its computational complexity. It is useful to analyze the way proteins fold in order to determine how a protein molecule encodes the functions of living organisms, how our muscles exert force or how our immune systems reject pathogens. One problem that needs to be addressed is the vast number of possible three-dimensional conformations of protein molecules. A conformation is a self-avoiding walk on a 2D square lattice. A self-avoiding walk is a sequence of moves on a lattice that does not visit the same point more than once. Applied to protein molecules again, they could hypothetically fold in so many various ways. Yet, there is a systematic behind it as proteins fold quickly into unique, sequence-specific native structures. So, how can the folding process be so determined, fast and reliable? To answer this question, it is interesting to analyze how proteins fold up, for instance to improve protein structure prediction and therefore have important practical implications. An active research area is the development of fast and straightforward folding algorithms. In this summary, three different ideas for a computation of the optimal conformations of a protein are described.

## The HP Lattice Model

Many algorithms are explained by using a toy example, the so called HP lattice model as an environment. In this model, proteins are short chains having access to an ensemble of conformations on a 2-dimensional square lattice. Each HP sequence consists of two kinds of monomers, hydrophobic (H) and polar (P), and each monomer is represented as a single bead on a lattice site. The method was able to quickly give an estimate of protein structure by representing proteins as "short chains on a 2D square lattice". One example for 3 different conformations

[*] gusfinfe@student.gu.se

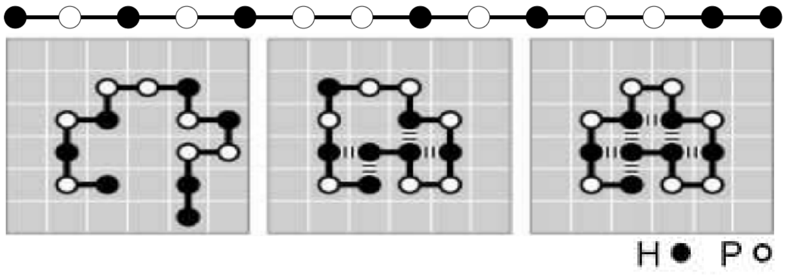

Figure 1: 3 different conformations in the HP Model

for one exemplary chain is shown in Figure 1. The energy of a conformation is determined by the contacts between two H monomers that are not adjacent in the sequence. Each HH-contact contributes -1 to the energy. Therefore, the third conformation in Figure 1 is the optimal one with the lowest energy.

## Solution 1: The CKY algorithm

Hockenmaier et al. [1] proposed a greedy search algorithm under the assumption that folding routines of proteins have an underlying tree structure. This allows us to apply the CKY algorithm, a dynamic programming technique that is normally used to identify the grammatical structure of natural language sentences, to protein folding. To understand this algorithm it is helpful to look at an example from natural language processing first.

In order to understand and distinguish the meaning of sentences such as *We eat sushi with tuna* and *We eat sushi with chopsticks*, it is necessary to parse them, that is to identify their correct syntactic structure. Therefore, a grammar is applied that contains the grammatical syntactic category of every word, for example, S (sentence), NP (noun phrase), VP (verb phrase), or PP (prepositional phrase). The grammar also contains the possible orders of the categories in the respective language. In the case of the above example, the parsing would result in respectively two trees as the sentence is ambiguous.

In the adaption of the CKY parsing to proteins, the sentence is represented by the chain of hydrophobic and polar monomers. This sequence is then split into $n$ sub strings and stored in the diagonal in a so called parse chart. Then, for every cell $i, j$ in the chart, any possible conformation of these sub strings, that is not a rotational variant of an already inserted item, is added to any cell with the adjacent cells $i-1$ and $j-1$. In addition, the folding process is also guided by a greedy search in which each cell in the tree identifies those conformations whose energies are locally optimal among all the possible conformations and only keeps those.

In Figure 2, a result of the algorithm applied on the chain of monomers shown in Figure 1 can be explored for a better understanding. One can clearly see that in the upper right cell the optimal structure is found that can be build

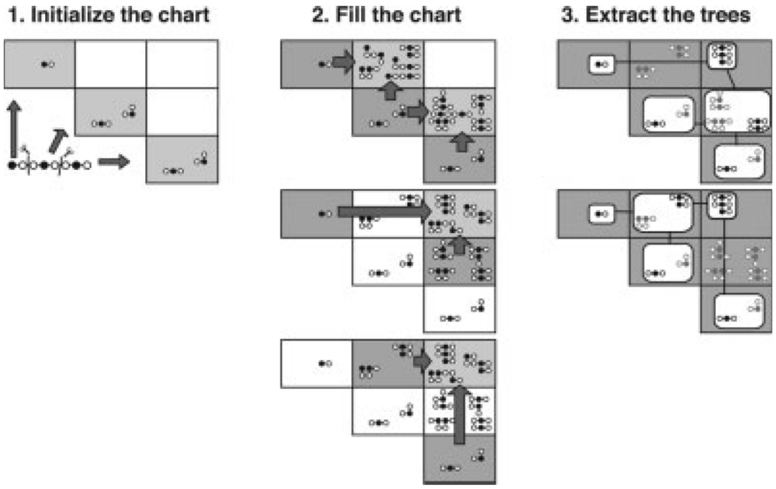

Figure 2: The CKY algorithm applied on Protein folding

up by the two variants of trees shown on the right hand side of the figure. Using this approach, the computational complexity to solve the problem can be drastically reduced.

## Solution 2: Monte Carlo Simulation

There is another solution to the protein folding problem, namely the Monte Carlo (MC) Simulaton proposed by Unger and Moult [2]. The MC simulation for protein folding can be described with the following algorithm. (1) Start with a random conformation for instance a straight chain. (2) From this initial conformation $S_1$ with energy $E_1$ make a single random change of the conformation to conformation $S_2$ and evaluate its energy $E_2$. If $E_2 <= E_1$ then accept the change to the conformation $C_2$. The random change is performed by randomly selecting any amino acid in the chain and rotating the attached chains around the amino acid. This is done until convergence meaning until the model with the lowest energy is found. The disadvantage of this method is the high computational cost and it cannot be guaranteed that it converges at the global minimum.

## Solution 3: Genetic Algorithms

As an improved model to the MC simulation, genetic algorithms are proposed for solving the problem by Unger and Moult [2]. The process starts with N extended structures that represent different generations. In each generation each structure already was subject to a number of mutations that correspond to a single MC step. Then, a pair of structures is selected. For this pair, a random point is chosen along the sequence and the N-Terminal portion of the first structure is connected to the C-Terminal of the second structure. Moreover, it needs to be tested how the 2 structures can be attached such that for instance, no residue from one structure occupies a lattice point from the other structure. This operation is performed until N-1 new structures are created, that become the population of the next generation. This goes on until a certain threshold is reached for instance a certain energy value or a defined number of populations. In comparison to the MC simulation, the genetic algorithm found the optimal conformation in an average of 35.000 energy evaluations of single structures, whereas the MC simulation took in average 4.000.000 energy evaluations.

# Conclusion

The three solutions to the protein folding problem described in this summary demonstrate, that over time more efficient computational models were invented. It is important to mention that it is not clear whether the functional conformation of every category of proteins is actually at the global minimum of energy within the conformation, yet all these models are optimized in regard to this goal.

# References

[1] Julia Hockenmaier, Aravind K Joshi, and Ken A Dill. Routes are trees: the parsing perspective on protein folding. *Proteins: Structure, Function, and Bioinformatics*, 66(1):1–15, 2007.

[2] Ron Unger and John Moult. Genetic algorithms for protein folding simulations. *Journal of molecular biology*, 231(1):75–81, 1993.