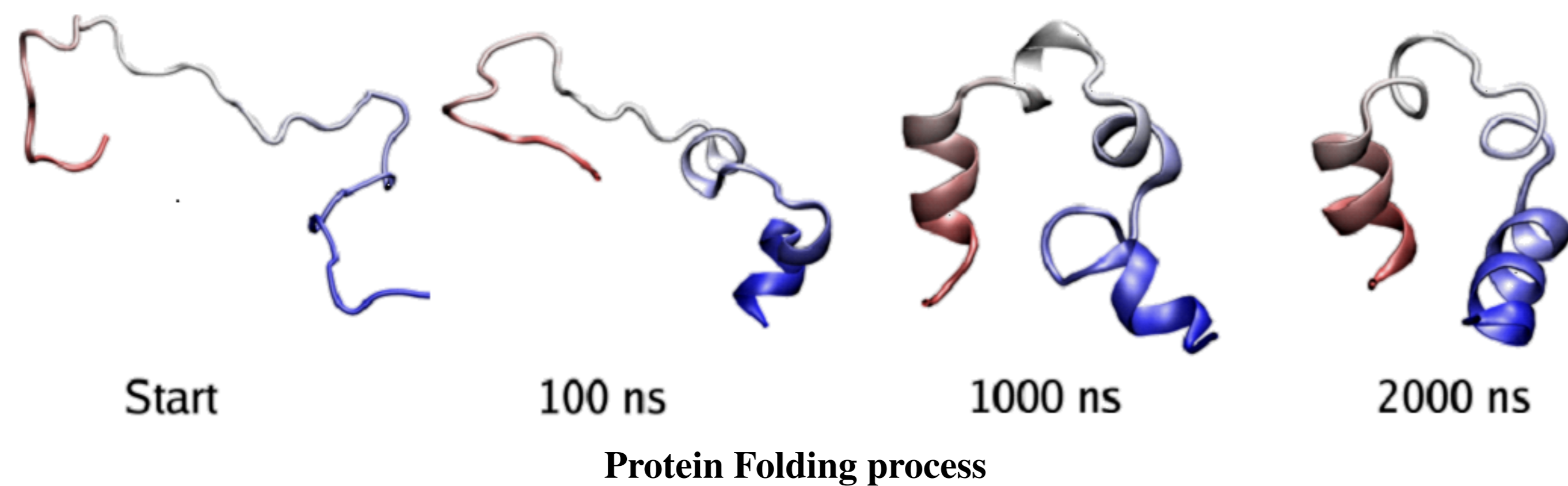


Overview

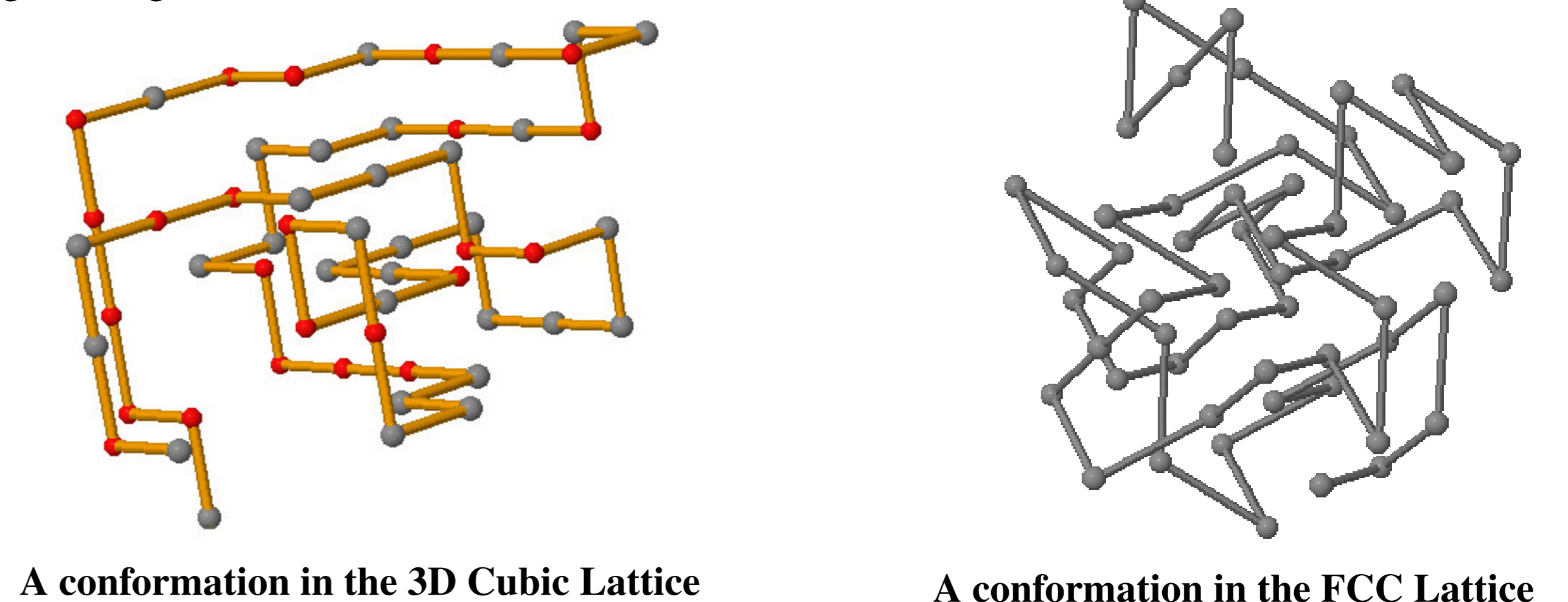
Proteins are believed to settle into three-dimensional structures of minimum energy, uniquely determined by the primary structure, i.e., amino-acid sequence. This minimum energy structure, also known as *native state*, determines the functionality of a protein. Knowledge of protein native state is crucial to pharmacology and medicine.

The process of a protein acquiring its minimum energy structure is called *Protein Folding*. This process happens spontaneously in nature under certain physical conditions and is completed within a timescale, varying from nanoseconds to milliseconds. Nevertheless, the huge amount of possible structures for a specific amino-acid sequence implies that Protein Folding is not an exhaustive search procedure, but rather a directed search towards the global minimum. The driving mechanism of Protein Folding, though, remains an open problem.



Algorithmic Aspects

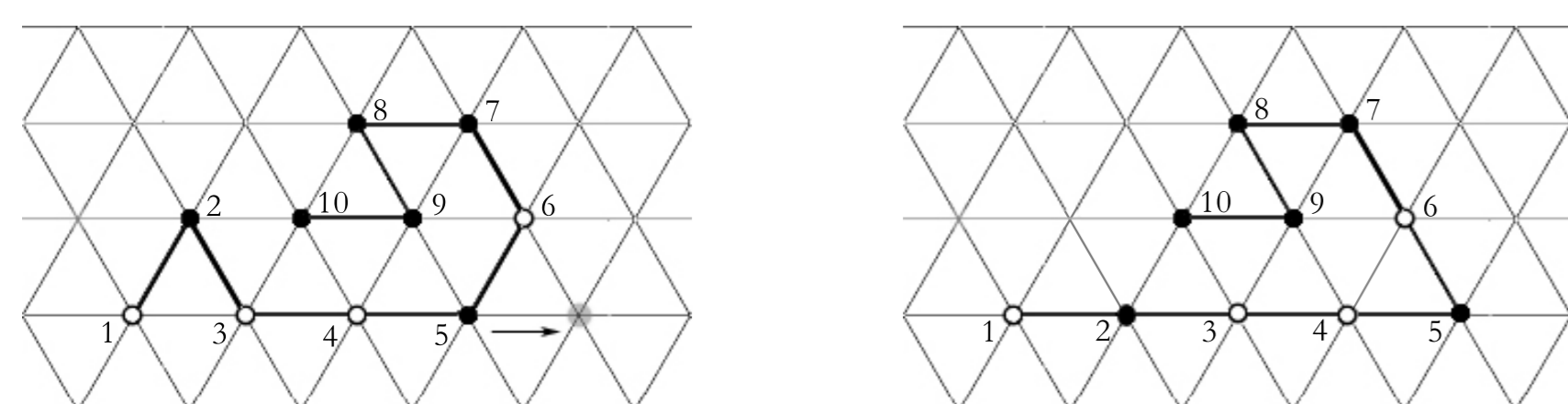
The problem of finding the global energy minimum of a protein is a NP-hard optimization problem. Accuracy of embedding the amino-acid sequence in simplified lattice structure and the complexity of the energy function are two major factors to distinguish the global minimum energy structure from the rest whilst keeping the computation feasible. Rectangular(cubic) and Triangular(FCC) are two popular lattice models. Consequently energy functions are proposed that take into account pairwise interactions among amino acids. HP energy model is the simplest, considering only two categories of amino acids. A more elaborate pairwise energy function is the Miyazawa-Jernigan matrix (MJ), which considers all of the 20 amino acids. The apparent intractibility of the problem enticed scientists to apply different approaches to solve it involving approximation algorithms, stochastic search algorithms and constraint programming.



Pull Move-Based Local Search

Hard optimization problems are often tackled with stochastic *Local Search* algorithms. In order to apply such algorithms, it is necessary to employ a move set. The move set defines a neighborhood relationship among conformations. It is essential that a move set is *reversible* and *complete* to ensure reachability of global minimum.

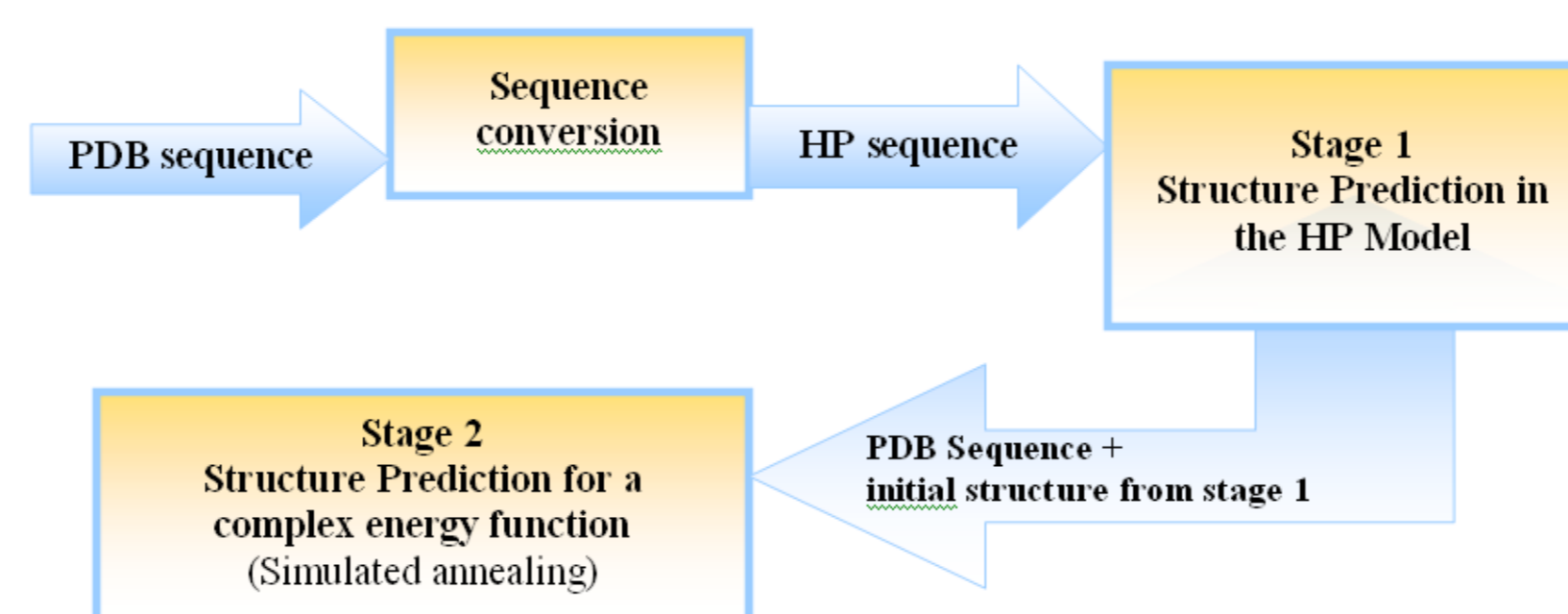
The *pull moves*, introduced by us, is an example of such a move set in triangular lattices. *Pull-moves* transform one conformation to another by creating or erasing a special substructure called *loop*. A loop is formed by three consecutive amino acids in the sequence positioned in three vertices of a triangle in triangular grid, which increases the chance of creating pairwise interaction between the first and third amino acid.



We ensure reversibility and completeness of our move set. FCC lattice is isomorphic to the 2D triangular lattice, therefore the pull move set is also extendible to more realistic FCC lattice. We incorporated pull moves in tabu search and simulated annealing and found optimal conformations as well as new lower energy conformations for several HP benchmarks in both triangular and FCC lattice.

Two-Stage Optimization

The Constraint-based Protein Structure Prediction (CPSP) approach enables the calculation of optimal structures in 3D HP-models within very short computational time. Nevertheless, it is computationally intractable for more elaborate energy functions such as a 20 amino acid pairwise interactions energy function. Local search approaches, on the other hand, work well in practice for elaborate energy functions, despite the large number of iterations required.

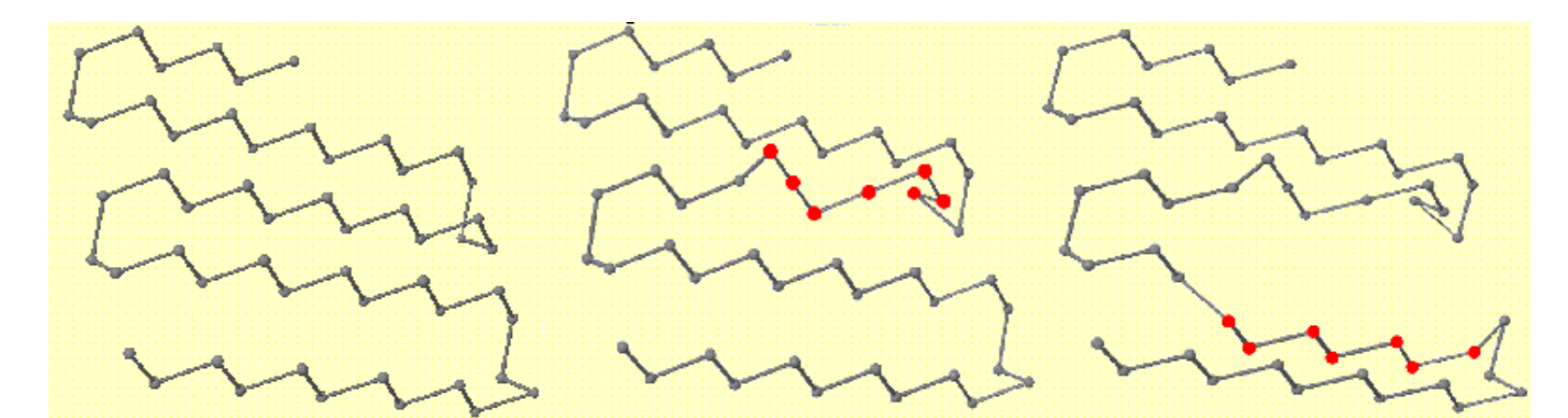


An outline of the two-stage optimization

We introduce a protein folding simulation procedure that employs two stages of optimization in order to find structures of minimum energy for elaborate energy functions. The input protein sequence first collapses to a compact structure and then a slower annealing procedure follows to find the minimum energy structure. Specifically we employ the CPSP tool to obtain an minimum energy HP model conformation in the FCC lattice. Then the CPSP output is given as input to a simulated annealing-based local search procedure which employs the elaborate energy function.

Hybrid Local Search

Local search approaches are usually faster but do not guarantee optimal solutions in polynomial time since the search space is randomly explored and tend to stuck in local minima. Constraint Programming techniques, on the other hand, guarantee to provide optimal solutions if the problem is modeled correctly. But as the solution space grows exponentially according to the problem size, the exponentially-increasing execution time is always a huge concern. We combine local search and constraint programming approaches aiming the expected outcome of better quality solutions in acceptable execution time.



Constraint Solver On Lattices (COLA) is a public domain solver that can model the protein folding as a Constraint Satisfaction Problem (CSP) on 3D lattices and produce acceptable quality solutions for larger proteins. We follow a protein folding simulation procedure on FCC lattice that employs COLA solver to generate neighbourhood states for a simulated annealing-based local search method. In each iteration of the local search, a randomly chosen smaller subchain is allowed to change its orientation and COLA solver outputs the best of all possible conformations considering that change.

Results and Analysis

Sequence	Length	2D Tri.			3D FCC		
		E_{HGA}	E_{TS}	E_{LSA}	E_{HGA}	E_{TS}	E_{LSA}
10100110100101100101	20	-15	-15	-15	-29	-23	-23
110010010010010010010011	24	-13	-17	-17	-28	-23	-23
0010011000011000011000011	25	-10	-12	-12	-25	-17	-17
0001100110000011111	36	-19	-24	-24	-51	-38	-38
111001100001100100							
001011101110000011111111	48	-32	-40	-43	-69	-74	-74
110000001100110010011111							
110101010111101000100010001	54	-23	-31	-40	-59	-77	-77
000010001000101111010101011							
0011101111111000111111111010	60	-46	-70	-70	-117	-130	-130
001111111111100001111101010							
11111111111101010011001100100110	64	-46	-50	-74	-103	-132	-132
011001001100110010101111111111							

E_{HGA} : Minimum energy found by Hybrid Genetic Algorithm with Twin Removal.

E_{TS} : Minimum energy found by Tabu Search.

E_{LSA} : Minimum energy found by Logarithmic Simulated Annealing.

Result: Pull Move Set based local search in triangular-HP energy model

Id	Length	Enumeration		BBF heuristic		LSA		Two-stage		Hybrid	
		Energy	Time	Energy	Time	Energy	Time	Energy	Time	Energy	Time
4RXN	54	-14.52	5h	-41.21	5h	-165.401	-167.781	10m 51s	-168.076	1h 05m	
1ENH	54	-24.058	5h	-41.854	5h	-152.747	-153.098	2m 33s	-157.062	1h 02m	
4PTI	58	-22.811	5h	-52.775	5h	-215.698	-212.500	6m 21s	-213.778	1h 20m	
2IGD	61	-19.598	5h	-47.589	5h	-180.893	-183.205	2m 37s	-186.696	55m	
1YPA	64	-22.831	5h	-61.464	5h	-256.017	-257.81	16m 54s	-258.709	42m	
1R69	69	-20.716	5h	-57.491	5h	-215.166	-219.402	14m 42s	-222.317	35m	
1CTF	74	-21.503	5h	-30.697	5h	-228.921	-233.86	11m 12s	-233.764	1h 36m	

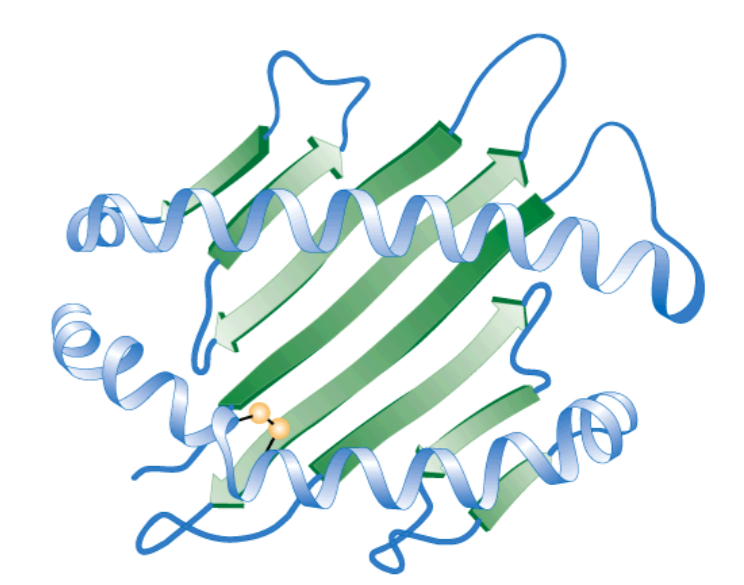
Analysis: Different approaches for FCC-MJ model

Conclusions

We propose three different non-deterministic approaches to protein folding problem. First, we introduce a unique pull move set to be integrated in local search algorithms that works for any types of contact based energy model. Then, we use a constraint programming tool, that claimed to give optimal result for HP model but not for elaborate energy functions, to generate high quality initial states for move set based local search algorithms. An intelligent sequence conversion technique is employed during the process. Finally, we modify an existing ad-hoc constraint solver for protein folding so that it can work as neighbourhood state generator for local search algorithms. Results show gradual improvement in terms of final energy value or runtime for these approaches. We also compare hybrid local search approach with pure constraint programming (CP) approach and found out that hybrid approach outclass CP significantly in terms of final energy and runtime. Though the first two approach proposed by us are problem specific, i.e., applicable to protein folding only, hybrid approach can be applicable to any hard combinatorial optimization problem.

Future Works

So far, we have avoided secondary structure information, i.e., rigid local substructures like α -helix and β -sheet, of protein for the sake of simplicity. But those are absolutely necessary to obtain realistic prediction of final conformation. Our future works will center around extending pull move set based local search and hybrid local search implementation to incorporate secondary structure information.



Acknowledgement

I would like to thank my supervisor for her continuous encouragement and guidance for this poster presentation. My gratitude also extends to my colleague Leonidas for technical assistance. Finally, special thanks to my wife for helping me with the graphics and design.