Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in Theoretical Computer Science 306 (2014) 45–59

Electronic Notes in Theoretical Computer Science

www.elsevier.com/locate/entcs

MASTERS: A General Sequence-based MultiAgent System for Protein TERtiary Structure Prediction

Thiago Lipinski-Paes[1]

Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas - LABIO
Programa de Pós-Graduação em Ciência da Computação, Faculdade de Informática
Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre, Brazil

Osmar Norberto de Souza[2,3]

Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas - LABIO
Programa de Pós-Graduação em Ciência da Computação, Faculdade de Informática
Programa de Pós-Graduação em Biologia Celular e Molecular, Faculdade de Biologia
Programa de Pós-Graduação em Biotecnologia Farmacêutica, Faculdade de Farmácia
Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre, Brazil

Abstract

As of May 12, 2014, there are approximately 39 million of non-redundant (nr) protein sequences in the NCBI nr database. The Protein Data Bank just now surpassed 100,000 three-dimensional (3-D) protein structures representing 1,393 different SCOP protein folds. Clearly, there is a huge gap between our capacities to generate protein sequences and to determine their experimental 3D structures. Structural bioinformatics, which addresses the problem of how a protein attains its 3-D structure starting only from its amino acid sequence, can reduce this gap. This is the so called the Protein Structure Prediction (PSP) problem. Thermodynamics considerations presented by Christian Anfinsen and co-workers in 1973 stated that a protein native structure is the one that minimizes its global free energy. Hence, we can treat the PSP problem as a minimization one within an NP-complete class of computational complexity. Several techniques have been proposed to predict the 3-D structure of proteins. In this work, we supplement these techniques by adding artificial intelligence concepts still not well explored in this scenario. More specifically, to address the PSP problem, we propose an ab initio-based framework for a cooperative hierarchical multiagent system guided by combined Simulated Annealing/Monte Carlo simulations. The framework was implemented in Netlogo, a widely used multiagent platform. MASTERS' main idea is to provide the user with the freedom to choose both the abstraction level and the energy function/force field model to perform the simulation. To demonstrate a typical MASTERS' application, we present a simple construct to the PSP problem, analyze its behavior and compare the results obtained with state-of-the-art optimization methods for equivalent coarse-grained abstractions.

Keywords: PSP Problem, Multiagent System, Monte Carlo, Simulated Annealing, Off-lattice, AB Model

[1] Email: thiagopaes@gmail.com

[2] Email: osmar.norberto@pucrs.br

[3] We thank the reviewers for their comments which helped improve the manuscript. We also thank CNPq,

http://dx.doi.org/10.1016/j.entcs.2014.06.014

1 Introduction

Proteins are polymers formed from a sequence of 20 different residues (19 amino acids and 1 imino acid). The physicochemical interactions between these residues within an appropriate environment create a unique spatial conformation for a protein [19]. The increase in the number of DNA sequences made available a large number of protein sequences. These sequences, translated from DNA, can be obtained from GenBank [3]. Protein 3-D structures, in turn, may be obtained from the Protein Data Bank or PDB B [5]. Currently, there are about 39 million non-redundant protein sequences. However, in the PDB, there are approximately 100,000 3-D structures of proteins. Eliminating redundancy by filtering very similar structures, we get only 1,393 different folds or topologies. Consequently, there is a huge gap between our capacities to produce protein sequences and to determine 3-D structures of new proteins with yet unknown folds [22]. Structural bioinformatics can help reduce this gap. The Protein Structure Prediction (PSP) problem emerged in the 60's and even today its solution remains a major challenge of molecular biology [11]. The PSP problem emerged in the 60's and even today its solution remains a major challenge of molecular biology [11]. Limitations of the 3-D structure experimental determination techniques, such as X-ray diffraction crystallography and nuclear magnetic resonance, highlight the importance of computational methods to predict the structure of proteins. Advances in handling the PSP problem will allow us to predict the 3-D structure of proteins with relevant applications in the biopharmaceutical industry. It will also improve our understanding of proteins involved in vital processes, including diseases such as cancer [12]. Considering the difficulties faced by traditional approaches (*in vitro* and *in vivo* experiments) in the treatment of problems concerning biological systems, the use of computer simulation becomes an attractive alternative, making possible the execution of low-cost and faster *in silico* experiments. An application that involves PSP must consider the system's real time adaptability, i.e., the modification of parameters such as the temperature of the thermal bath surrounding the protein.There is a clear need for *in virtuo* experiments: experiments performed via a computer simulation, susceptible to perturbations during their execution. While the easy modification of parameters is a typical property of all computer simulations (*in silico* experiments), the easy modification of the experiment itself is a specific property of multiagent systems or MAS, resulting in *in virtuo* experiments [1,28]. It is important to note that *in virtuo* experiments can be performed without employing agents. However, the use of MAS provides technical support particularly suitable to address problems such as the PSP one [1]. Here, we present a general tool that allows addressing PSP according to the user needs. The user is free to choose both the force field and abstraction level to guide the simulation of a protein. The agents have a hierarchical organization and the Searching Agents have the mission of exploring the conformational space. These agents have their own movement scheme, intimately linked to the optimization algorithm. Optimization is done by a Monte Carlo/Simulated Annealing simulation

CAPES and FAPERGS for financial support. ONS is a CNPq Research Fellow.

and the user can modify its parameters and even the optimization method itself. MASTERS can be obtained at labio.org.

2 Background

2.1 Proteins

Proteins are the most abundant biological macromolecules and are present in all cells and all parts of those cells. All proteins, either from the oldest bacteria or the most complex forms of life, are built from the same set of 20 amino acids. The amino acids differ in their side chains, also known as R groups. Each R group is covalently bound to the protein backbone (alpha carbon) and has a particular size (ranging from 1 to 11 non-hydrogen atoms), structure and polarity. The linear sequence of amino acids residues of a protein is called its primary structure [7].

2.2 The PSP Problem

The PSP problem is the problem of predicting the 3-D structure of a protein starting from its primary structure or amino acid sequence. The physical process by which a polypeptide folds into a functional protein is an old question (reviewed by Snow [25]) and continues to be one of the biggest challenges in current structural bioinformatics [11]. Knowing the tertiary structure of a protein is crucial since it is directly related to its function, therefore, allowing the identification of areas known as catalytic sites, sites of allosteric change and others [19]. Most of the drugs currently on the market act by interacting with enzymes, so the study of the sequence-structure-function relationship is vital for the creation of new drugs. Bioinformatics has an important role in accelerating this knowledge [39].This paper's approach is based on Anfinsen's proposal which says that, at the environmental conditions (temperature, solvent concentration and composition) at which folding occurs, the native structure is a unique, stable and kinetically accessible minimum of the protein free energy. However, finding this structure is not trivial and even simplified methods have NP-Complete complexity [9]. Still regarding the inherent difficulty of the problem we can cite the Levinthal's paradox [20], which states that for a 100-length chain there will be at least 2^{100} possible conformations (considering only two degrees of freedom), characterizes it as an intractable problem [30].]. In the last five decades different algorithmic approaches have been tested and, although there has been progress, the problem remains unsolved even for small proteins. While the ultimate goal is to predict the 3-D or tertiary structure from the primary structure, the current knowledge and computing power is insufficient to handle a problem of such complexity [15].

2.3 Multiagent Systems

Multiagent systems (MAS) are part of the Artificial Intelligence field and refer to the modeling of autonomous agents in a common universe. MAS is a relatively new subfield of Computer Science - it has been studied since about 1980 - and

the field has only gained widespread recognition since about the mid 1990 [36,35]. Agents are computational entities that interact with the environment and are goal-directed, having a body and a location in time and space. An agent does not exist without an environment to act on and the environment can be of numerous types and complexities. How the agent recognizes the environment strongly depends on the dwell capabilities, so the environment classification must be made from the agent standpoint. The complexity of the environment depends on the complexity of the agent. Typically, each agent can be described by a set of behavioral skills that define its jurisdiction, a set of objectives and the necessary autonomy to use their skills and achieve their goals. An agent is an autonomous computational entity that decides its own actions. The agents' autonomy means that they have an existence independent of other agents, and have to achieve their own goals [13,33]. A set of agents acting in an environment characterizes a MAS.

Netlogo

MASTERS was built using Netlogo (v5.0) [29]. Netlogo is a very popular agent-based modeling tool and is particularly well suited for modeling complex systems that take time into account and where hundreds or thousands of agents can be programmed and interact independently, making possible to explore the connection between micro- and macro-levels of behavioral patterns. Netlogo runs on the Java virtual machine, thus it works on all major platforms (Mac, Windows and Linux). One of the remarkable advantages of Netlogo is its embedded tools, and BehaviorSpace is one of them. BehaviorSpace offers the possibility of automatically performing a large set of experiments by changing simulation parameters' values. In MASTERS, due to the BehaviorSpace capability, it is possible to explore more resourcefully the conformational space in PSP predictions and tune MASTERS' parameters to improve the prediction results.

3 The MASTERS Framework

MASTERS was developed based on a command line framework for protein prediction supported by a hierarchical MAS and written in prolog and c++ [6]. A strong Netlogo's peculiarity is that it was developed for educational purposes, providing a rich modeling environment along with a very simple modeling language. It allows MASTERS' users to create, prepare and run simulations through an intuitive interface. It also lets users to edit the code and analyze the execution results (plots and 3-D visualization), on the fly.

Fig 1 illustrates MASTERS' interface running an AB model simulation [27]. In order to run the simulation, the user needs to follow a few steps, namely starting up the simulator (loading the main parameters), create and execute the agents. Once running the simulation, it is possible to verify its progress through the monitors (e.g. "Step" and "Current Energy") and plots (e.g., "Energy vs. Time" and "Acceptance Ratio vs. Time"). However, before being able to run simulations, one needs to be aware of some of MASTERS' core concepts.

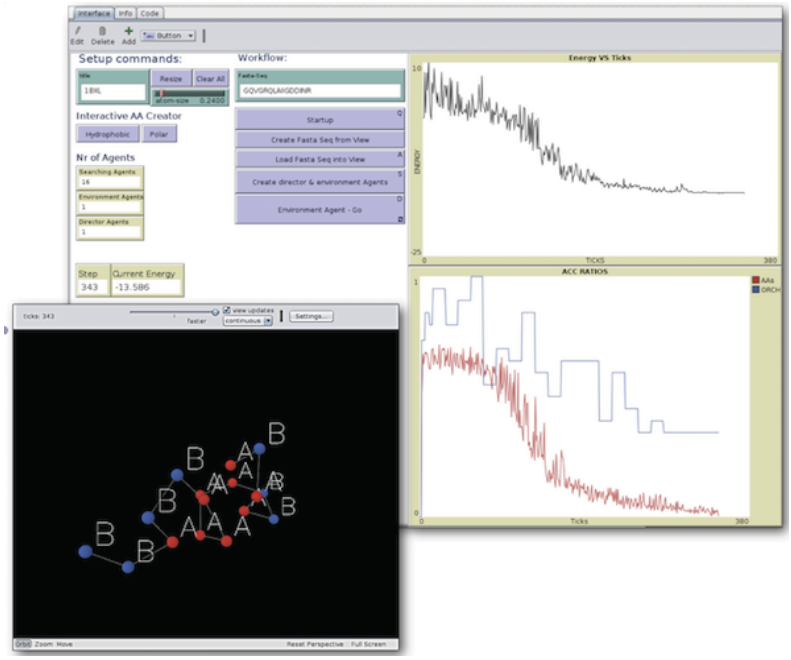

Fig. 1. MASTERS' interface. MASTERS provides the user with an interactive interface, where the target sequence can be pasted and loaded. It is also possible to tune the simulation by altering its parameters. It also displays real-time plots during the simulation and a 3-D visualization window with the actual 3-D structure achieved.

3.1 Choosing the Energy Function/Abstraction Level

MASTERS is orthogonal with respect to the simulation flow and the energy function/abstraction level used (Fig 2). The user must choose which energy function/force field to use and this is a primordial MASTERS step within the framework's generality. MASTERS was not built exclusively to a particular energy function. The framework can be applied to a wide range of optimization problems that involve Cartesian coordinates (2-D or 3-D). Regardless of the problem, the chosen energy function will affect directly the number of searching agents and their movements. Once the user concludes these steps, it can proceed with the simulation.

3.2 Monte Carlo/Simulated Annealing Scheme

All movements are controlled by a Monte Carlo (MC) criterion which defines two conformational states S_1 and S_2, each one with its correspondent energy E_1 and E_2. If $E_2 < E_1$ the move is accepted. If $E_2 > E_1$ there is still the possibility of accepting the move. However, in such cases, the probability of accepting a move turning turning S_1 into S_2 follows the equation 1, where k is the Boltzmann constant and T, the temperature.

$$e^{-(E2-E1)kT} = e^{-(\Delta 21)kT} \tag{1}$$

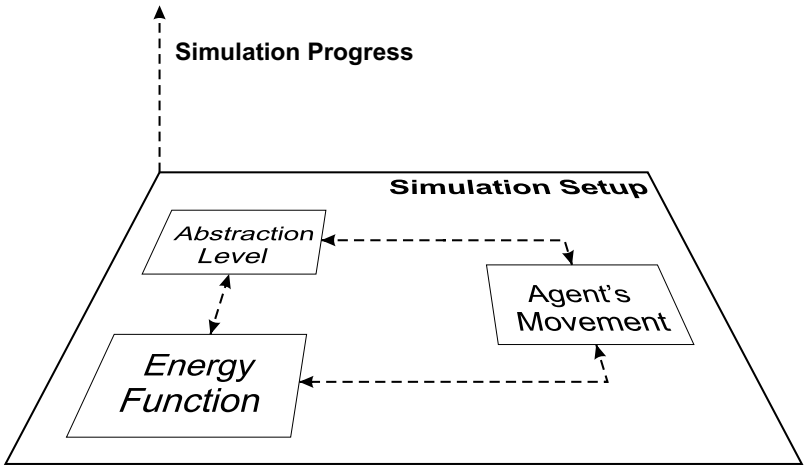

Fig. 2. MASTERS' perspective view. The abstraction level, energy function and type of moves are core features that must be set before the simulation starts. These features are orthogonal to the simulation progress.

The MC method has drawbacks though. In the PSP problem, for example, the temperature of the system determines the size of energy barriers that could potentially be overcame. When dealing with temperatures that are too low, MC will not explore far from the minimum energy found, leading to local minima. Simulated Annealing (SA) is a simple MC modification that turns it into a global optimizer. At the beginning of the simulation the system is set to a high temperature, allowing it the overcome fairly high-energy barriers. Then the system is gradually cooled, eventually causing the system to confine to a single energy well. Due to the gradual cooling rate (logarithmic), the system ends up spending more time in low energy regions of the conformational space. This may increase the chances to find the lowest energy state although there is no assurance [39]. The way the temperature is reduced is critical for the SA performance, given that convergence is guaranteed only if the temperature is reduced to zero logarithmically. In MASTERS the temperature is gradually decreased according to equation 2, where $\alpha = 0.98$:

$$T_{k+1} = T_k * \alpha \tag{2}$$

3.3 Hierarchical Cooperation

A core concept of MASTERS is the hierarchical organization of the agents. Based on [24], the agents are divided into different levels. Higher-level agents have the role of coordinating the actions of lower-level agents. We have essentially three types/levels of agents (Fig 3), in a bottom-up hierarchical order:

Searching Agents

The Searching Agents have the mission of exploring the conformational space. Usually, one or more searching agents are associated to each amino acid in the protein, depending on the chosen abstraction. The agents' position is expressed as

Cartesian coordinates, resulting in movements that are local, i.e., they do not affect the position of other agents.

Director Agent

The Director Agent has total knowledge about the protein actual spatial conformation and its function is to coordinate searching agents, aiming at a more efficient spatial exploration. The Director agent has no representation in the Cartesian space. It is an agent that acts on the 3-D space from outside. Director Agents perform global moves on the searching agents. It is mandatory to have at least one Director Agent in the simulation.

Environment Agent

There is only one Environment Agent in the simulation. Its role is to control all other agents, as well as the simulation flow, simulated annealing scheme, number of movements per time/temperature step, real time plots, outputs and also the simulation ending.

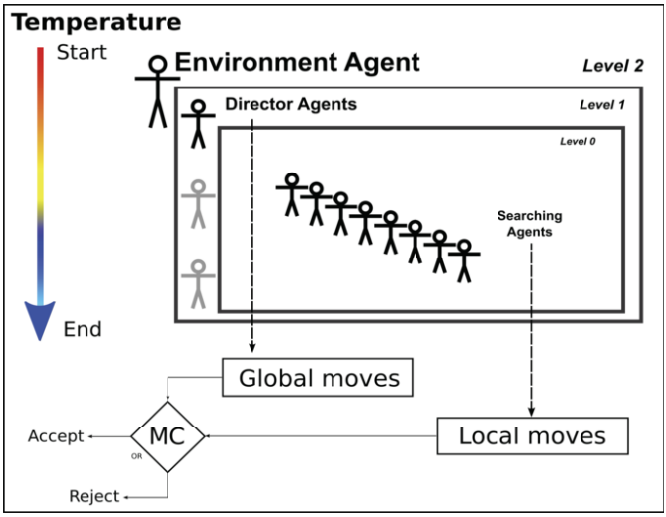

Fig. 3. MASTERS' behavior. The temperature starts high and is cooled according to the Simulated Annealing parameters. The Environment Agent controls the Director and Searching Agents. The moves are accepted/rejected based on the Monte Carlo criterion.

3.4 Finding the minima

The simulation relies on accounting each agent's movement attempts. To achieve minima at a given temperature, the system should explore the conformational space in a large number of times. Counters are used to summarize the average number of move attempts per agent type (both Searching and Director Agents). Every time step is related to a specific temperature (following a Simulated Annealing scheme). The system will be stuck in each temperature until an average number of movement

attempts has been attained. The simulation ends when the temperature reaches a particular value. This variable is set by the user but has a default value of 0.0099.

4 Results

To examine MASTERS' behavior and effectiveness, we adopted a simplified model called AB Model as a case of study. The simplest and most conventional model among all applied abstractions used in PSP is the HP Model [10]. The HP model divides all 20 amino acids in two different groups, the hydrophobic (H) and the hydrophilic (P) ones. The amino acids are placed at an on-lattice grid, and the energy computation at each conformation takes into account only interactions between next-neighbored nonadjacent hydrophobic amino acids [10]. The energy of a conformation is the number of hydrophobic-hydrophobic contacts that are adjacent on the lattice, but not adjacent in the string (sequence). The main idea is to force the establishment of a compact hydrophobic core as observed in real proteins [34]. Lattice models have proven to be useful tools for reasoning about the complexity of PSP problems [10] and despite of its high abstraction level, the PSP problem with HP models is still an NP-Complete challenge [4]. The AB model is also an on-lattice model in which the amino acids are once again divided into two groups: hydrophobic amino acids are marked as A while the hydrophilic ones are marked as B. The AB model, in comparison with the HP model, has the additional capability of collect information about local interactions that might be significant for the local structure of protein chains. This allows finding compact, well-defined native structures that would not be found if these local interactions were neglected [17]. Unlike the HP model, the interactions considered in the AB model include both sequence independent local inter-actions and the sequence dependent Lennard-Jones term that supports the energy convergence to a hydrophobic core [27,26].

4.1 Energy Function

The AB off-lattice energy model is described by equation 3, where θ is the bend angle between the two bonds defined by three consecutive residues and r_{ij} is the distance between residues i and j [26]. The first sum, the backbone bending potential, calculates the bending angle energy of the protein chain. The double sum is the Lennard-Jones potential. It calculates the long-range interaction energy, which is attractive for pairs of the same amino acids (AA or BB) and repulsive for AB pairs. The residue specific prefactor C is given by the equation 4.

$$E = \sum_{i=2}^{n-1} \frac{1}{4}(1 - cos\theta) + \sum_{i=1}^{n-2} \sum_{j=i+2}^{n-1} [r_{ij}^{-12} - C(\xi_i, \xi_j) r_{ij}^{-6}] \tag{3}$$

$$C(\xi_i, \xi_j) = \left\{ \begin{array}{l} +1, \quad \xi_i \xi_j = A \\ +1/2, \; \xi_i \xi_j = B \\ -1/2, \; \xi_i \neq \xi_j \end{array} \right\} \tag{4}$$

4.2 Target Sequences

We used several sequences as simulation targets to investigate the MASTERS' performance. The sequences were chosen from the literature to allow further comparisons. They are classified in two groups: Fibonacci sequences and real protein sequences. Table 1 shows their primary structures.

Table 1
MASTERS' target sequences and their sizes. The sequences with four alphanumeric IDs are from PDB. As in [32], the * indicates the chain B of the protein, and "X" in the chain represents a non-standard residue.

ID	Size	Sequence
13	13	ABBABBABABBAB
21	21	BABABBABABBABBABABBAB
34	34	ABBABBABABBABBABABBABABBABBABABBAB
55	55	BABABBABABBABBABABBABABBABBABABBABBABABBABABBABBAB ABBAB
1CB3	13	XIDYWLAHKALAX
1BXL	16	GQVGRQLAIIGDDINR
1EDP	17	CSCSSLMDKECVYFCHL
2H3S*	25	PVEDLIRFYNDLQQYLNVVTRHRYX
2KPA*	26	VSVDPFYEMLAARKKRISVKKKQEQP
1TZ4	37	YPSKPDNPGEDAPAEDLAQYAADLRHYINLITRQRYX
1TZ5	37	APLEPVYPGDNATPEQMARYYSALRRYINMLTRPRYX
1AGT	38	GVPINVSCTGSPQCIKPCKDQGMRFGKCMNRKCHCTPK
1AHO	64	VKDGYIVDDVNCTYFCGRNAYCNEECTKLKGESGYCQWASPYGNACYCYK LPDHVRTKGPGRCH

Fibonacci Sequences

The Fibonacci sequences are defined recursively by $F_0 = A, F_1 = B$, $F_{i+1} = F_{i-1} * F_i$ where $*$ represents concatenation. For instance, if F_2 is a sequence formed by 'AB', F_3 is a sequence formed by 'BAB' and F_4 = 'ABBAB' and so forth [23]. Stillinger [27] has first used the Fibonacci sequences in 1993 and since then they are commonly used by researchers in simplified protein modelling. The Fibonacci sequences have the following intrinsic properties: (i) all A residues are isolated along the backbone, (ii) the Bs appear only isolated or in pairs, never as longer or continuous B strings and (iii) the molecules have a hierarchical string structure. Although this model has neither explicit representation of side chains nor hydrogen bonds and offers only an abstraction to the complexities of real proteins, its off-lattice phase space described by both Lennard-Jones and bending energy contributions makes it capable of capturing some of the essential features of protein structures [26]. For evaluation purposes, we chose target sequences with lengths ranging from 13 to 55 residues. Table 2 presents the lowest energy values in comparison with the literature. The results were acceptable for all four Fibonacci target sequences. None of the energy values was lower than the lowest results of the current literature, which indicates possible limitations of our method. However, even without any specificity with respect to the chosen model, the framework was able to obtain acceptable results. In addition to the energies, we analyzed the 3-D structures and it also showed expected features, such as the persistent formation of hydrophobic cores.

The generated 3-D structures of Fibonacci sequences are in the PDB format and can be obtained from the supplementary digital contents.

Table 2
Benchmark of the total energy results for AB model Fibonacci sequences. Comparison between MASTERS and the literature. $[16]_b$ is a related optimization method also presented in [16].

ID	[26]	[16]	$[16]_b$	[2]	[18]	[8]	[21]	[37]	[31,38]	MASTERS
13	-3.223	-3.973	-4.962	-4.967	-4.975	-4.975	-4.974	-6.569	-6.954	-4.869
21	-5.288	-7.686	-11.524	-12.316	-12.327	-12.062	-12.247	-12.327	-14.797	-11.786
34	-8.975	-12.860	-21.568	-25.476	-25.511	-23.044	-24.812	-25.511	-27.990	-22.182
55	-14.409	-20.107	-32.884	-42.428	-42.342	-38.198	-42.518	-42.342	-42.475	-36.202

Real Protein Sequences

To expand the evaluation of our framework we applied MASTERS to real protein sequences. In Table 3 we present the results for nine protein targets and compare them with the literature. The amino acids A,C,G,I,L,M,P and V are set to hydrophobic or class A and the other 12 remaining residues D,E,F,H,K,N,Q,R,S,T,W and Y are set to hydrophilic or class B. As can be seen, the results for real protein sequences ffollow the pattern of those for the Fibonacci sequences. In most cases, the energies are very close to those found in the literature. However, note that for the protein 1AGT the lowest energy found by MASTERS is significantly better than the one found by the other methods. Since low energies do not necessarily lead to acceptable 3-D structures, we further analyzed the structures (Fig 4). It is evident from the figure the correct hydrophobic packing (in red) of the structures.

Table 3
Benchmark results for AB model real protein sequences. Comparison between MASTERS and the literature.

ID	[37]	[38]	[32]	MASTERS
1BXL	-15.717	-15.825	-	-15.039
1EDP	-12.839	-13.777	-	-12.651
1AGT	-44.266	-46.084	-	-52.142
1CB3	-	-	-8.250	-7.145
2H3S*	-	-	-18.160	-17.500
2KPA	-	-	-25.100	-22.979
1TZ4	-	-	-39.340	-35.281
1TZ5	-	-	-45.300	-37.629
1AHO	-	-	-69.025	-64.472

Fig 4 shows the 3-D structure of 1CB3, 1EDP, 1AGT and 1AHO. It is possible to note the formation of one hydrophobic core for each protein. Even for 1AGT and 1AHO, earlier authors ([38,32]) state that they had not been able to obtain such compact cores.

4.3 CPU Time

Tables 4 and 5 show the CPU time spent by MASTERS to compute the 3-D structure of the Fibonacci and real protein sequences. The authors of [21] provide exact values for CPU times. In [16] the authors do not provide exact values. However,

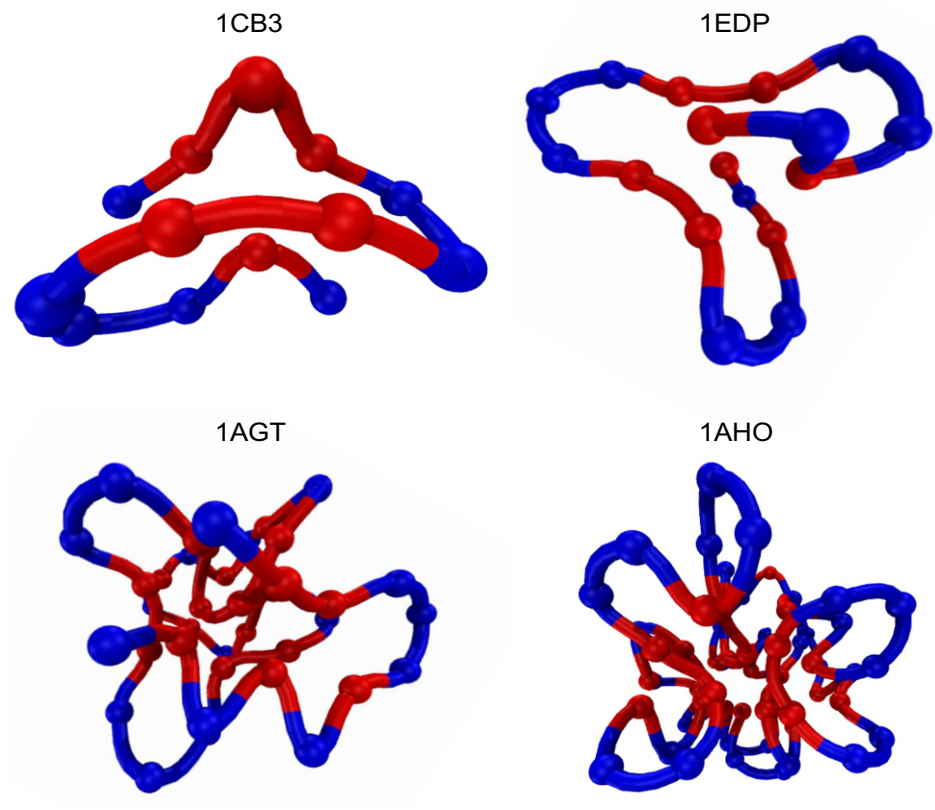

Fig. 4. 3-D structures of 1CB3, 1EDP, 1AGT and 1AHO obtained by MASTERS. Red means hydrophobic or A and blue means hydrophilic or B.

they mention that the results for the four Fibonacci sequences were obtained on a Unix workstation with up to 2 days of CPU time. The authors of [8] mention only that 2 hours of CPU time in a single 2.4GHz core PC was necessary to achieve their results. Our simulations were done in a single core 2.8 GHz machine at LABIO. As can be seen by the values in Table 4 and the information provided about other methods, MASTERS' CPU time requirements is substantially lower than those given in the literature. In addition, it is important to stress that in MASTERS, at each time step, the number of movement attempts for each amino acid should not be less than 100, which means the CPU time grows with the sequence length. Tables 4 and 5. No information was found about CPU time for real protein sequences. We present our values to allow future benchmarks.

4.4 Reliability

MC simulations are non-deterministic. Therefore, it is important to look to the method's precision. The following figures show the fluctuation of the total energy obtained for each sequence. Aiming at obtaining meaningful statistical values, MASTERS ran 40 independent random simulations for each sequence. Fig 5 shows the

Table 4
Benchmark results for AB model Fibonacci sequences CPU time, in seconds. Comparison between MASTERS and [21].

ID	[21]	MASTERS
13	4874.19	231.53
21	7985.34	518.73
34	23565.70	1287.93
55	45234.09	3445.30

Table 5
MASTERS' results for AB model for real protein sequences CPU time, in seconds.

ID	MASTERS
1BXL	324.89
1EDP	361.43
1AGT	1571.11
1CB3	232.66
2H3S∗	707.66
2KPA	765.69
1TZ4	1482.74
1TZ5	1494.93
1AHO	4699.08

results.The 3-D plots show the change in energy as a function of run number and its respective time step (related also to the temperature). The plots show if the results converge or not due to the vast conformational space covered by the simulations. For all sequences the system proved to have consistent behavior, except for a few instances like the anomalous value pointed in the 1AHO plot. It is also possible to notice an emergent pattern from the plots, which we named the waterfall phenomenon. The waterfall phenomenon emerged in all simulations. We presented here just 1CB3, 1EDP, 1AGT and 1AHO for the sake of brevity. It is widely believed and experimentally established the premise that realistic short single-domain proteins are two-state folders. In other words, these proteins can be only in two states: folded or unfolded (denatured). Between these two states should be a transition that triggers the folding/unfolding [14]. Since it does not occur at a specific critical temperature during the simulation, the waterfall emergent phenomenon points in a direction that reinforces that the AB model is suitable and useful as a simplified protein heteropolymer model. Furthermore, we observed that the waterfall phenomenon occurs whenever the energy falls below zero (transitions from positive to negative energy). This seems to be intrinsic to the protein AB model abstraction and energy function.

5 Conclusion and Future Work

MASTERS is effective in capturing information about the simple off- lattice models investigated. The chosen models demonstrated a considerable capacity to simulate folding by basic energy contributions without being knowledge-based. In spite of the achievement of slightly higher values of energy in most of the cases compared to the literature, the A residues fold into hydrophobic cores for all targets. Even

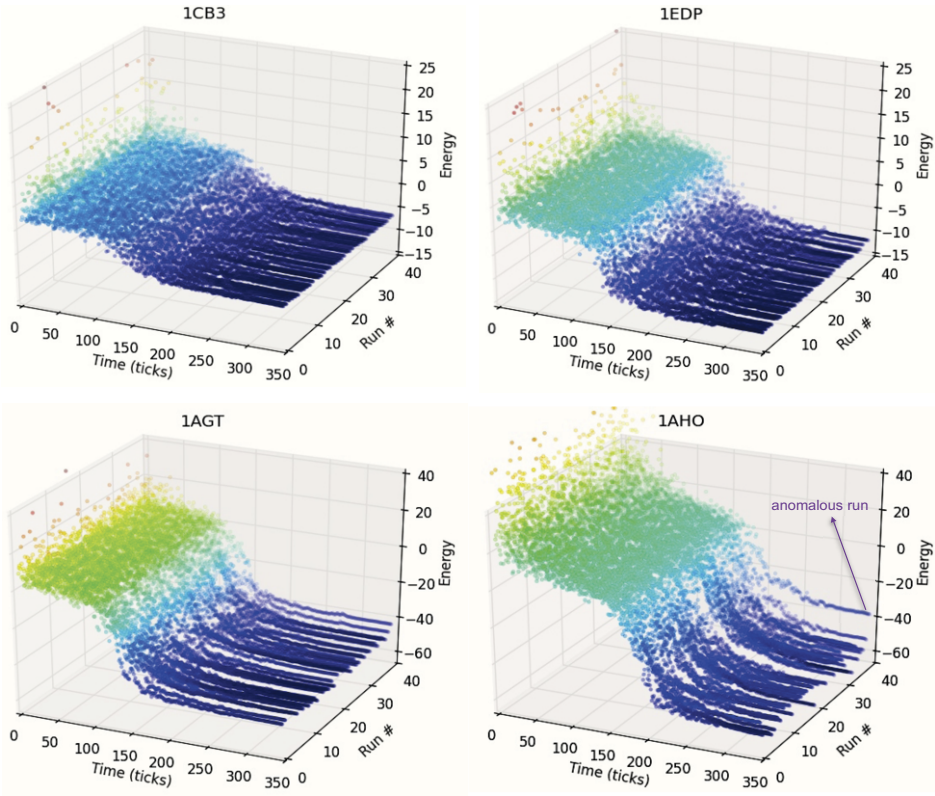

Fig. 5. MASTERS' reliability. Results for 1CB3, 1EDP, 1AGT and 1AHO. For better understanding the points are colored by energy. There are 2 scales for energy, given the size difference between the chains.

without specific treatments for the energy function and abstraction level and starting from a stretched chain (different from some techniques that apply optimizations to achieve promising initial configurations e.g [8]), MASTERS was able to achieve better results than the current literature in one example. We found lower energy results than those found in previously published methods for 1AGT and 1AHO, with both folding into compact hydrophobic cores. The latter findings were a result not accomplished by previous approaches [32,37,38]. The reliability of the method was tested in 40 independent runs for each target sequences and, aside a few anomalous values, the non-deterministic Monte Carlo/Simulated Annealing scheme was sufficiently stable. Regarding CPU time performance, the results were better than all other methods with published CPU times, highlighting a characteristic that could be truly explored in future works, e.g., the application of MASTERS to more complex systems such as all atom models and real energy functions. Further studies are necessary to detail the relationship between the AB model abstraction, the energy function and the observed waterfall phenomenon. Given its exequibility, MASTERS is very suitable to continue this exploration.

References

[1] F. Amigoni and V. Schiaffonati. Multiagent-based simulation in biology. In L. Magnani and P. Li, editors, *Model-Based Reasoning in Science, Technology, and Medicine*, volume 64 of *Studies in Computational Intelligence*, pages 179–191. Springer Berlin Heidelberg, 2007.

[2] M. Bachmann, H. Arkin, and W. Janke. Multicanonical study of coarse-grained off-lattice models for folding heteropolymers. *Physical Review E*, 71(3), 2005.

[3] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Research*, 41(D1):D36–D42, 2013.

[4] B. Berger and T. Leighton. Protein folding in the hydrophobic-hydrophilic (hp) model is np-complete. *Journal of Computational Biology*, 5(1):27–40, 1998.

[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[6] L. Bortolussi, A. Dovier, and F. Fogolari. Agent-based protein structure prediction. *Multiagent and Grid Systems - Multi-agent systems for medicine, computational biology, and bioinformatic*, 3(2), 2007.

[7] J. Boyle. Lehninger principles of biochemistry (4th ed.): Nelson, d., and cox, m. *Biochemistry and Molecular Biology Education*, 33(1):74–75, 2005. 1539-3429.

[8] M. Chen and W.Q Huang. Heuristic algorithm for off-lattice protein folding problem. *Journal of Zhejiang University SCIENCE B*, 7(1):7–12, 2006. J. Zhejiang Univ. - Sci. B.

[9] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis. On the complexity of protein folding. *Journal of Computational Biology*, 5:597–603, 1998.

[10] K.A. Dill. Theory for the folding and stability of globular-proteins. *Biochemistry*, 24(6):1501–1509, 1985.

[11] K.A. Dill and J.L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.

[12] Y. Duan and P. A. Kollman. Computational protein folding: from lattice to all-atom. *IBM Systems Journal*, 40:297–309, 2001. 0018-8670.

[13] J. Ferber. *Multi-Agent Systems - An Introduction to Distributed Artificial Intelligence*. Addison-Wesley, 1999.

[14] D. Gorse. Application of a chaperone-based refolding method to two- and three-dimensional off-lattice protein models. *Biopolymers*, 64(3):146–160, 2002.

[15] G. Helles. A comparative study of the reported performance of ab initio protein structure prediction algorithms. *Journal of the Royal Society Interface*, 5(21):387–396, 2008. 1742-5689.

[16] H.P. Hsu, V. Mehra, and P. Grassberger. Structure optimization in an off-lattice protein model. *Physical Review E*, 68(3), 2003.

[17] A. Irback, C. Peterson, F. Potthast, and O. Sommelius. Local interactions and protein folding: A three-dimensional off-lattice approach. *Journal of Chemical Physics*, 107(1):273–282, 1997.

[18] S.Y. Kim, S.B. Lee, and J. Lee. Structure optimization by conformational space annealing in an off-lattice protein model. *Physical Review E*, 72(1), 2005.

[19] A. M. Lesk. *Introduction to bioinformatics*. Oxford University Press, Oxford ; New York, 3rd edition, 2008.

[20] C. Levinthal. Are there pathways for protein folding? *Journal of Medical Physics*, 65(1):44–45, 1968.

[21] J.F. Liu and W.Q. Huang. Quasi-physical algorithm of an off-lattice model for protein folding problem. *Journal of Computer Science and Technology*, 22(4):569–574, 2007.

[22] A. Pavlopoulou and I. Michalopoulos. State-of-the-art bioinformatics protein structure prediction tools (review). *International Journal of Molecular Medicine*, 28(3):295–310, 2011.

[23] A.N. Philippou and F.S. Makri. *Longest Circular Runs with an Application in Reliability Via the Fibonacci-Type Polynomials of Order K*, book section 31, pages 281–286. Springer Netherlands, 1990.

[24] A. Roli and M. Milano. Magma: A multiagent architecture for metaheuristics, 2004.

[25] C. D. Snow, E. J. Sorin, Y. M. Rhee, and V. S. Pande. *How well can simulation predict protein folding kinetics and thermodynamics?*, volume 34 of *Annual Review of Biophysics*, pages 43–69. Annual Reviews, Palo Alto, 2005.

[26] F.H. Stillinger and T. Head-Gordon. Collective aspects of protein-folding illustrated by a toy model. *Physical Review E*, 52(3):2872–2877, 1995.

[27] F.H. Stillinger, T. Head-Gordon, and C.L. Hirshfeld. Toy model for protein-folding. *Physical Review E*, 48(2):1469–1477, 1993.

[28] J. Tisseau. *Virtual reality, in virtuo autonomy*. Accreditation to Direct Research, Universit de Rennes 1, 2001.

[29] S. Tisue and U. Wilensky. Netlogo: A simple environment for modeling complexity. 2004.

[30] A. Tramontano. Integral and differential form of the protein folding problem. *Physics of Life Reviews*, 1(2):103–127, 2004. 1571-0645.

[31] T. Wang and X. Zhang. 3d protein structure prediction with genetic tabu search algorithm in off-lattice ab model. *Knowledge Acquisition and Modeling, 2009. KAM '09. Second International Symposium on*, 1:43–46, 2009.

[32] T. Wang and X. Zhang. A case study of 3d protein structure prediction with genetic algorithm and tabu search. *Wuhan University Journal of Natural Sciences*, 16(2):125–129, 2011. Wuhan Univ. J. Nat. Sci.

[33] G. Weiss. *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. MIT Press, 1999.

[34] E. H. William and N. Alantha. *Protein Structure Prediction with Lattice Models*, pages 30–1–30–24. Chapman & Hall/CRC Computer & Information Science Series. Chapman and Hall/CRC, 2005.

[35] M. Wooldridge. *An Introduction to MultiAgent Systems*. Wiley Publishing, 2nd edition, 2009.

[36] M. Wooldridge and N.R. Jennings. Intelligent agents - theory and practice. *Knowledge Engineering Review*, 10(2):115–152, 1995.

[37] X. Zhang and W. Cheng. Protein 3d structure prediction by improved tabu search in off-lattice ab model. *Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. The 2nd International Conference on*, pages 184–187, 2008.

[38] X.L. Zhang, T. Wang, H.P. Luo, J.Y. Yang, Y.P. Deng, J.S. Tang, and M.Q. Yang. 3d protein structure prediction with genetic tabu search algorithm. *Bmc Systems Biology*, 4, 2010.

[39] M. Zvelebil and J. Baum. *Understanding Bioinformatics*. Garland Science, 2007.