A Systematic Mapping Study on Applications for Multi-core and Many-core Architectures for Protein Structure Prediction

Gesiel Rios Lopes[1], **Alexandre C. B. Delbem**[1]

[1]Institute of Mathematical and Computer Sciences
University of São Paulo (ICMC-USP)
São Carlos-SP, Brazil

gesielrios@usp.br, acbd@icmc.usp.br

Abstract. *Proteins are the most abundant organic compounds of living matter and perform essential functions to the life's process. Given a protein's amino acid sequence, the Protein Structure Prediction (PSP) problem is to find a three-dimensional structure that has the native energy level. It can help in the design of new drugs and medicine. However, despite advances made in recent years, the development of methodologies capable of achieving a high degree of predictability and accuracy remains a major challenge. This systematic mapping aims to find related studies and research opportunities of how multi-core and many-core architectures have been used to solve the PSP problem. We have defined a systematic mapping process and applied it to complete a systematic mapping study. Thirty-two primary studies were selected for discussions on advances and opportunities for further investigations. The results show that there is an increasing interest to apply solutions based on multi-core and many-core architectures for this computing hard problem.*

1. Introduction

Proteins are biological macromolecules, consisting of one or more long chains of a linear chain of amino acid residues, called a polypeptide, which performs essential functions for the creation and maintenance of life. Proteins perform a vast array of functions, such as enzymatic actions, catalyzing metabolic reactions, DNA replication, responding to stimuli, and transporting molecules from one location to another. Many proteins are extracellular signals such as insulin, transmitting signals to distant tissues, or are binding proteins, which carry biomolecules to different places in the body, performing vital organic functions. The execution of all these activities depends exclusively on the protein having an active biological function, which in turn depends on its native state, which is closely associated with its three-dimensional structure [Anfinsen 1973, Webster 2000].

The Protein Structure Prediction (PSP) has been a tool for development in aid of research in areas that require the determination of structures of new proteins by computational simulation. The PSP problem consists of determining the tertiary structure from a sequence of amino acids (primary sequence) of a given protein. The tertiary structure that the protein takes depends on the interactions between their atoms and those with the atoms of the environment (solvent) [Webster 2000]. Finding the tertiary structure of a protein means knowing the relative position of atoms in three-dimensional space. Therefore, this problem has been considered one of the main challenges of Cellular Molecular Biology [Webster 2000, Moult et al. 2014].

Most of the protein structures already determined were obtained using experimental methods such as X-ray Crystallography (CRX) and Nuclear Magnetic Resonance (NMR). Such methods, in general, require relatively significant financial resources and development time. Furthermore, it is not always possible to determine the structure by these methods. Therefore, it is indispensable to develop computational methods to predict 3D protein structures from protein sequences, in a faster, reliable and inexpensive way, since the number of these sequences increases each day.

According to the Levinthal's paradox [Levinthal, Cyrus 1968], there is a vast number of possible conformations to reach the correct native state, even considering the most powerful computers available, the time required to calculate the energy of all these conformations is comparable to the age of the universe. However, in nature, the time of folding of proteins take only a few seconds or less to reach their native state, and therefore the folding process can't occur through a random search for all possible conformations. In addition, Anfinsen's thermodynamic hypothesis [Anfinsen 1973] states that, at least for small globular proteins, the native structure is a unique, stable, and kinetically accessible minimum of the free energy.

Levinthal's paradox and Anfinsen's hypothesis allow us to formulate PSP problem as an optimization problem and in computational complexity theory is classified as NP-complete problems [Baker 2000]. Therefore, due to the complexity presented by the PSP problem and the inefficiency of the use of exact methods to solve NP-Complete problems, the use of approximate strategies, such as the metaheuristics[1], allied with the application of high performance computing techniques (HPC) has been a compelling alternative to obtain acceptable solutions to the PSP problem [Dorn et al. 2014, Llanes et al. 2016].

In this sense, despite advances made in recent years, the development of methodologies and use of appropriate techniques capable of achieving a high degree of predictability and accuracy remains a major challenge. For that reason to solve the PSP problem, if imperative to investigate the application of parallel processing in the energy potentials used to differentiate structures potentially near or far from the native state or in the optimization process used to find the native state, considering the use of more robust and comprehensive programming models, mainly those considering the use of heterogeneous processing, with CPUs and processing accelerators such as Graphics Processing Units (GPU), and Xeon Phi.

Despite advances made in recent years, the development of methodologies capable of achieving a high degree of predictability and accuracy remains a major challenge and, for that reason, new computational techniques have been investigated for PSP problem. In this paper, a Systematic Mapping Study (SMS) is presented in order to provided an overview of state of the art for the manner that solutions based on multi-core and many-core architectures are been used to solve the PSP problem.

The remaining of this paper is structured as follows. Section 2 presents a brief overview of the aspects related to PSP problem; Section 3 discusses the phases of the SMS, including its objective and the study selection criteria; Section 3.3 discusses the results; the conclusions are summarized in Section 4.

[1]Metaheuristic it is a general rule architecture that, formed from a common theme, can serve as a basis for solving generically optimization problems (usually from the area of combinatorial optimization)

2. Protein Structure Prediction Problem

Proteins are basic structures of all living beings made up from 20 L-α-amino acids which fold into a particular 3D structure that is unique to each protein. The folding of proteins has a compact form in relation to polypeptides, also generating the structural diversity necessary for proteins to perform and acquire a specific set of biological functions. It is known that better understanding the protein folding process results in medical advancements and development of new drugs [Anfinsen 1973]. A polypeptide is a continuous structure of many amino acid sequences which are bonded with peptide bond. An amino acid unit in the polypeptide chain is called residue.

The 3D structure is responsible for the structure of the protein. There are four levels of protein structure, like 'primary', 'secondary', 'tertiary' and 'quaternary'. The primary structure of the protein is also the amino acid sequence itself that differentiates one protein from another, this level of molecular organization is the simplest and most important because it originates the spatial arrangement of the molecule. The conformation that the protein will assume depends primarily on that amino acid sequence. The secondary structure refers to segments of the primary structure that form isolated folds. There are three main types of secondary structures: α-helix, β-sheet and the loops (also called coils). The combination of all secondary structures of a protein forms a tertiary structure, which corresponds to the final form that a protein will assume. Several tertiary structures bind to other tertiary structures to form a more complex structure, known as quaternary structures [Baker 2000].

The PSP problem refers to the determination of the 3D configuration of given protein from its amino acid sequence. This problem is not trivial [Brasil et al. 2013]. The three-dimensional structure that the protein assumes depends on the interactions between its atoms and these with the atoms of the medium (solvent). Therefore, finding the tertiary structure of the protein means knowing the relative position of atoms in three-dimensional space, a task that requires a high computational effort. Despite the significant advances in scientific computing in recent years, mainly regarding to the enlargement of the processing capacity of parallel computers at relatively low costs, the development of methodologies capable of achieving a high degree of predictability and accuracy in PSP remains one of the main challenges of Molecular and Cellular Biology [Verli 2014, Moult et al. 2014].

The first methods developed for the PSP problem were organized according to three main groups: Comparative Modeling, Fold Recognition (or Threading) and first principle (or *ab initio*. These methods differ in the use of information available in the databases of experimentally resolved three-dimensional structures of proteins. Comparative modeling is the methodology most dependent on this information, with *ab initio* being totally independent.

Despite the recent advances in the area of PSP, it can be noted that the separation among these three methods is increasingly tenuous. Besides, a quick found to the latest CASP[2] shows that many of the methods can be included in more than one category. For example, the separation between the prediction of protein folding and comparative mod-

[2]Critical Assessment of Protein Structure Prediction - CASP is a world-wide biannual meeting that aims at establishing the current state of the art in protein structure prediction of the different methodologies developed, identifying what progress has been made, and highlighting where future effort may be most productively focused [Moult et al. 2014].

eling is increasingly difficult, and the use of some structural/experimental information is widely observed even in so-called first principles. Therefore, the classification below is currently used when evaluating and comparing methods objectively [Verli 2014]:

1. **Template-free:** uses only information from an amino acid sequence of a protein and a force field that models the interactions among atoms, in order to restrict the dihedral angles for values that correspond to feasible folds [Webster 2000];
2. **Template-based:** prioritize structures (conformations) for a protein by considering its similarity to protein sequences obtained by other methods such as for example, CRX, and NMR [Verli 2014].

In this classification, methods called of *de novo* are those that use some structural information, such as protein fragments, secondary structure prediction and statistical potentials, derived from proteins. In this way, what will indicate the choice of the method to be applied is the presence or not of structures solved experimentally, and deposited in banks of structures such as the Protein Data Bank (PDB), that can be used as a template for the modeling of the target sequence. The choice of a method is intrinsically related to the obtained from the alignment among the target sequence and possible template candidates.

The conformations associated with the global minimum of an energy function are considered the probable native conformations that the protein adopts under physiological conditions. Thus, methods of protein structure prediction must have, in their methodologies, the following common characteristics [Verli 2014, Llanes et al. 2016]:

(i) One representation of the protein structure and a set of degrees of freedom that define the search space of conformations;
(ii) Energy functions compatible with the representation of protein structure;
(iii) Search algorithms to efficiently select an energetically favorable conformation.

3. The Systematic Mapping Study

A Systematic Mapping Study (SMS) is a methodology to allow the organization of research reports and their results, given the necessary support to categorize and give a visual summary of them [Petersen et al. 2008].

The SMS presented in this paper was conducted considering three main phases: (i) planning, (ii) conducting and (iii) reporting. The next sections address these phases and their obtained results.

3.1. Planning

In this phase, the review protocol contains (i) research questions, (ii) search strategy, (iii) inclusion and exclusion criteria and (iv) data extraction process and methodology for the synthesis of the data were defined. Its main goal is provided an overview of the multi-core and many-core architectures, which are used to solve the PSP problem. Two research questions (RQ) were defined:

RQ*$_1$*: What parallel programming models are used to solve the PSP problem with multi-core or many-core architectures?

RQ*$_2$*: What mathematical models are used to solve the PSP problem with multi-core or many-core architectures?

A search string and the electronic databases were also defined. The search string was elaborated and refined according to an initial set of key papers selected and based on opinion of experts in such area.

The search string (Figure 1) was defined in [Lopes et al. 2019], considering the following keywords: prediction, protein structure, PSP, tertiary structure, computing method, computational model, optimization, optimizing, high performance, HPC, parallelization, parallel, concurrent, graphics processing unit, GPU, graphic accelerator, FPGA, field programmable gate array and Boolean operations.

```
(( "prediction" ) AND ( "protein structure" OR "psp" OR "tertiary structure" OR "ab initio" OR "template
free" ) AND ( "computing method" OR "computational model" OR "optimizing" OR "optimization" OR
"high performance" OR "hpc" OR "parallelization" OR "parallel" OR "concurrent" OR "graphics
processing unit" OR "gpu" OR "graphic accelerator" OR "field programmable gate array" OR "fpga" ))
```

Figure 1. Search String.

With the purpose of selecting the most adequate databases for our search, we considered the criteria discussed by Dieste and Padua [Dieste and Padua 2007], and selected five databases, namely: ACM Digital Library, IEEE Xplore, Scopus, and Web of Science.

Relevant primary studies were selected based on the following inclusion and exclusion criteria (respectively IC and EC). Not all inclusion criteria should be satisfied for each primary study. An inclusion criterion was defined for the verification of importance and usefulness for each study; exclusion criteria were defined for the refinement of the results. Our IC and EC were: (**IC-1**) Techniques or approaches of multi-core and many-core architectures that are used for the prediction of protein tertiary; (**EC-1**) The written language is not English; (**EC-2**) In the case of duplicate studies, the most complete one is considered; (**EC-3**) The study does not relate techniques or approaches of multi-core and many-core architectures to solve PSP problems; (**EC-4**) The study does not relate prediction of protein tertiary; (**EC-5**) The study does not relate details about implementation; (**EC-6**) The primary study is a table of contents, short course description, tutorial, keynotes, copyright form or summary of an event (e.g., a conference or a workshop); (**EC-7**) The study is published in the grey literature[3].

3.2. Conducting

In this phase, the primary studies were identified from databases and analyzed. Scopus returned a larger set of studies (686). Web of Science, IEEE Xplore and ACM Digital Libary returned 646, 198 and 165, respectively (Figure 2). The duplicated studies (306 papers in all) were identified and removed. A set of 691 papers was selected according to the inclusion and exclusion criteria, in the selection phase, based on the partial reading (titles and abstracts). After a full reading, only 32 papers were selected. We aimed to be conservative as possible and, therefore, the search string has become generic to retrieve many studies from electronic databases, even if it would give us more effort in the selection process. Many papers were introduced as primary studies, but only few of them had more contributions and/or larger impacts.

[3]Grey literature uses materials/research made available by organizations not belonging to the academic or traditional commercial publishing

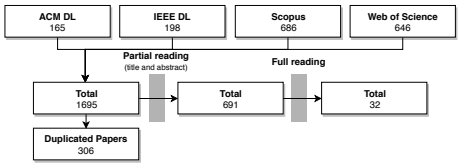
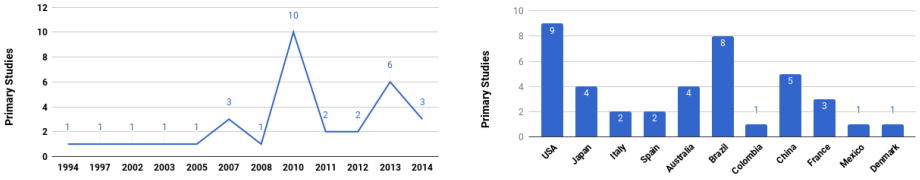


Figure 2. Distribution of papers (conduction phase).

3.3. Reporting

This section discusses an overview of multi-core and many-core architectures based on selected primary studies. Figure 3(a) shows the distribution of primary studies selected per year used in our SMS. Although the automatic search in search engines has not been limited to a specific period, the last decade concentrates the most of the papers. This points out that research aiming to make possible the resolution of the PSP problem through multi-core or many-core architecture has increased in the last years.



(a) (b)

Figure 3. (a) Distribution of primary studies selected over the years. (b) Quantity of primary studies selected per countries.

Our SMS counted 70 authors in the 32 primary studies. Researchers are located in 11 different countries, as shows Figure 3(b). The sum of publications in each country exceeds the number of studies selected because some studies were developed in cooperation among different researchers and countries.

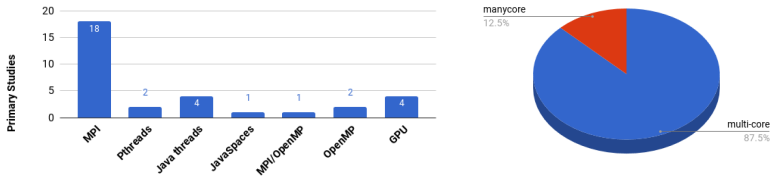
The remainder of this Section discuss the main idea of the selected primary studies considering each research questions proposed.

1) RQ_1: **What parallel programming models are used to solve the PSP problem with multi-core or many-core architectures?**

Figure 4(a) presents the distribution of programming models used in the solutions for the PSP problem. Observe that the most of the adopted solutions found in the primary studies selected, corresponding to 59.3%, use Message Passing as the programming model through MPI (56,2%) and JavaSpaces (3,1%). Already Shared Memory as the programming model corresponds to 25% with Pthreads (6,2%), OpenMP (6,2%) and Java threads (12,4%). We still have the use of the many-core architectures corresponding to 12,4% through GPU and one case of the hybrid programs using both MPI and OpenMP, corresponding to the remaining 3,3%.

Figure 4(b) shows the disposal of the primary studies selected from each architecture.

2) RQ_2: **What mathematical models are used to solve the PSP problem with multi-core or many-core architectures?** This research question seeks to discover what



(a) (b)

Figure 4. (a) Distribution of primary studies per programming models. (b) Distribution of primary studies per architectures

Mathematical Models have been used to simplify both the polypeptide chains and the residue positions of an atomic model. Figure 5 shows the disposal of the primary studies selected from mathematical models. We can observe that HP Model is the most used model with 19 paper (59,4%), indicating that it is the most conventional and has been widely used in PSP, according to already indicated by [Sar and Acharyya 2014]. AB-OFF Lattice Model and the Atomic model based on the dihedrals angle base between the C-alpha also has been found, with 8 papers (25,0%) and 5 papers (15,6%) respectively.

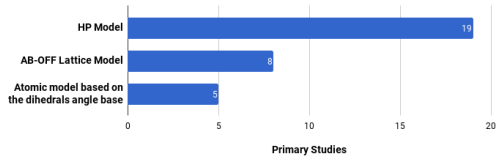


Figure 5. Distribution of primary studies per mathematical models.

4. Conclusion

This paper has addressed a systematic mapping study to find evidences about the application of solutions based on the multi-core and many-core architectures applied to the protein structure prediction problem. In order to provide a mapping of research topics, a set of 4.969 primary studies was analyzed, in which 32 were selected for discussions. We have defined three research questions that reflect the scope of the study to map the contributions and challenges.

Based on our results, it is possible to identify a trend in the use of distributed memory with MPI as programming model to solve PSP problems. However, this may indicate an underutilization of the available resources of the architecture used and makes the solutions non-scalable. This situation leads to the development of hybrid programming model, as proposed by [Gang et al. 2006]. The HP Model was the most widely used in the primary studies selected. Although the HP Model has shown impressive results, such results do not adequately evaluate potential solutions by using only the usual metric of hydrophobic contacts, hamming the performance of the algorithm. So, many other models need be proposed, which consider other interactions levels. It is worth mentioning that solving the problem of the prediction of protein structure, it is not an easy task, and the use of multi-core and many-core architectures will be an decisive resource to solve this mission-critical science problem.

References

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(4096):223–230.

Baker, D. (2000). A surprising simplicity to protein folding. *Nature*, 405(6782):39.

Brasil, C. R. S., Delbem, A. C. B., and da Silva, F. L. B. (2013). Multiobjective evolutionary algorithm with many tables for purely ab initio protein structure prediction. *Journal of computational chemistry*, 34(20):1719–1734.

Dieste, O. and Padua, A. G. (2007). Developing search strategies for detecting relevant experiments for systematic reviews. In *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*, pages 215–224.

Dorn, M., e Silva, M. B., Buriol, L. S., and Lamb, L. C. (2014). Three-dimensional protein structure prediction: Methods and computational strategies. *Comp. biology and chemistry*, 53:251–276.

Gang, W., Xiaoguang, L., and Jing, L. (2006). Parallel algorithm for protein folds prediction. In *2006 International Conference on Computational Intelligence and Security*, volume 1, pages 470–473.

Levinthal, Cyrus (1968). Are there pathways for protein folding? *Journal de chimie physique*, 65:44–45.

Llanes, A., Muñoz, A., Bueno-Crespo, A., García-Valverde, T., Sánchez, A., Arcas-Túnez, F., Pérez-Sánchez, H., and M Cecilia, J. (2016). Soft computing techniques for the protein folding problem on high performance computing architectures. *Current drug targets*, 17(14):1626–1648.

Lopes, G. R., de Souza, P. S. L., and Delbem, A. C. B. (2019). A systematic mapping on high-performance computing for protein structure prediction. In Senger, H., Marques, O., Garcia, R., Pinheiro de Brito, T., Iope, R., Stanzani, S., and Gil-Costa, V., editors, *High Performance Computing for Computational Science – VECPAR 2018*, pages 77–91, Cham. Springer International Publishing.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (casp) - round x. *Proteins: Structure, Function, and Bioinformatics*, 82:1–6.

Petersen, K., Feldt, R., Mujtaba, S., and Mattsson, M. (2008). Systematic mapping studies in software engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, EASE'08, pages 68–77, Swindon, UK. BCS Learning & Development Ltd.

Sar, E. and Acharyya, S. (2014). Genetic algorithm variants in predicting protein structure. In *2014 Int. Conf. on Comm. and Signal Proc.*, pages 321–325.

Verli, H. (2014). *Bioinformática: da biologia à flexibilidade molecular*. Sociedade Brasileira de Bioquímica e Biologia Molecular, São Paulo, SP, BR, 1ª edition.

Webster, D. M. (2000). *Protein structure prediction: methods and protocols*, volume 143. Springer Science & Business Media.